

# ISYE 6414 Lecture Transcripts

## Important notes on adding content -- read this first!

Thank you to everyone who has added and cleaned text and imported images into this transcript document. There are a few best practices you should keep in mind while you are contributing: **Replace errors and poor phrasing entirely.** If you come across transcription errors or clunkily phrased language, change them. You don't have to -- and should not -- employ strikethrough, brackets, highlighting, comments, or any other annotation. Just **be bold** and make the changes.

**Don't duplicate or replace text with images.** The document is beginning to respond sluggishly because of its size. Instead of adding screenshots of text, reproduce it using native formatting (bolding, bullets, underlines, etc.) Also, replacing text with images makes it unsearchable. The big exception to this is mathematical formulas, which are much easier to understand in proper notation.

**Crop images before adding them.** Google Docs holds on to the entire image file, which again adds to file size and sluggishness.

**Import from PowerPoint where possible.** Not only are images crisper than from screenshots of the lecture videos, but you can resize images and delete elements that might obscure the graphic you are importing.

**Use <sup>superscript</sup> and <sub>subscript</sub> as warranted.** Press [Control] (or [Command] on a Mac) + [Comma] to toggle superscript and [Control] + [Period] to toggle subscript.

**Thanks again for all of your help!**

*For use as study material for GTx 6414x. Text and images copied directly from the course videos, with light editing for readability. Not intended for broad distribution or general use.*

*Content is the intellectual property of Nicoleta Serban and Georgia Tech.*

# About This Document

This document was originally created in the spring of 2018 and is maintained collaboratively through the efforts of ISYE 6414 students using transcripts and screenshots from the video lectures. You are strongly encouraged to improve the formatting, layout, add or adjust images, bold key words, and even condense copy.

It is expected that sections may be added, removed, or modified -- in which case, again, you should please take the liberty of adjusting this document to match.

Some tips on formatting: the Weeks (e.g., Week 1), Modules (e.g., 1. Introduction), and Sub-Modules (e.g., 1.3 (C): What is Modeling?) are formatted respectively as Heading 1, Heading 2, and Heading 3. These can be adjusted in the format menu, or by pressing Ctrl + 1, Ctrl + 2, and Ctrl + 3, respectively, while the cursor is on the line you wish to change.

Pressing Ctrl + , will toggle <sub>subscript text</sub> on and off.

Important notes on adding content -- read this first!	1
About This Document	2
1.1: Estimation & Statistical Inference	8
1. Basics	8
Case Studies	8
The Objectives of Linear Regression	15
2. Estimation Method	17
Identifying the Model Parameters	19
Estimating Sigma Squared	21
3. Estimation Data Examples	25
R Function: lm()	25
Mean 4. Statistical Inference	29
5. Statistical Inference Data Examples	37
R Functions: pt() and confint()	38
Knowledge Check	41
1.2 Prediction and Model Evaluation	42

1. Regression Line: Estimation & Prediction	42
2. Regression Line: Estimation & Prediction Examples	47
R Function: predictLM()	49
Knowledge Check	51
3. Diagnostics	52
Checking For Normality	54
4. Outliers and Model Evaluation	58
5. Diagnostics and Model Evaluation Examples	61
1.3 Data Examples	65
1. Testing the Theory of Purchasing Power Parity (Part 1)	65
2. Testing the Theory of Purchasing Power Parity (Part 2)	74
3. 2000 Elections in Florida	83
<b>Unit 2: Basics of ANOVA</b>	<b>91</b>
1.1 Analysis of Variance (ANOVA)	91
1. Basics of ANOVA	91
Example: Choir voices and height	92
Example: Keyboard layout efficiency	92
Primary Objectives of ANOVA	93
2. Estimation Method	94
Determining sampling distribution for the pooled variances	95
Estimation of Mean Parameters	95
3. Estimation Data Examples	98
4. Test for Equal Means	100
SST Decomposition	101
Testing Equal Variances With The F Test	102
Example: Testing for Equal Means	103
Knowledge Check	106
2.2 Basic Concepts and Estimation	107
1. Comparing Pairs of Means	107
Confidence Intervals with Multiple Populations	108
R Function: TukeyHSD()	108
2. Model Fit Assessment	111
Assumptions of ANOVA	111
Types of Diagnostic Plots in ANOVA	112

Example: Voice Pitches and Model Fit Assessment	113
Knowledge Check	116
3. ANOVA vs. Simple Linear Regression	117
Decomposition of ANOVA into a linear regression model	117
4. Data Example	120
Example: Cancer Survival	120
R Functions: <code>hist()</code> and <code>log()</code>	121
Estimation of Parameters	123
<b>Unit 3: Multiple Linear Regression</b>	<b>128</b>
3.1 Basic Concepts and Estimation	128
1. Objectives and Data Examples	128
Example 1: Medical Supply Advertising	128
Example 2: SAT Scores by State	128
Example 3: IMDb	129
The Uses of Regression Analysis	130
2. Basic Concepts	131
Assumptions of Multiple Linear Regression	131
Identifying Model Parameters	132
Four Basic Regression Approaches	133
1. First-Order Model	133
2. Second-Order Model	134
3. First-Order Interaction Model	134
Quantitative and Qualitative Variables	136
Multiple Linear Regression Concepts Applied	137
Example: Advertising and Sales	137
Example: IMDb Data	137
3. Estimation Method	139
Parameter Estimation	139
4. Model Interpretation	142
When to Use Multiple Linear Regression	143
Causality Versus Association	143
Example: SAT Scores and College GPAs	143
Roles of Predicting Variables: Controlling, Explanatory, Predictive	144
5. Estimation Data Examples	146

Example 1: Advertisement Expenditure and Sales	146
Using R for Multiple Linear Regression	146
R Code Used In This Example	149
Example 2: SAT Scores By State	149
Explanatory and Controlling Variables	149
Lecture 3.1 Knowledge Check	151
3.2 Statistical Inference and Prediction	153
1. Statistical Inference	153
Properties of Regression Estimators	153
Estimating Sigma Square	154
Confidence Interval Estimation	155
Testing Statistical Significance	156
Testing for Statistical Positivity or Negativity	157
2. Testing for Subsets of Coefficients	158
Testing Subsets of Coefficients	159
Generalizing the Test for Coefficients	161
3. Statistical Inference Data Examples	163
Knowledge Check	167
4. Regression Line: Estimation & Prediction	168
Estimating The Regression Line	168
5. Regression Line: Estimation & Prediction Data Examples	173
Knowledge Check	176
Homework 2	177
3.3 Model Diagnostics, Evaluation, and Multicollinearity	180
1. Assumptions and Diagnostics	180
Properties of Errors and Residuals	181
The Residual Analysis	182
Residual Analysis: Linearity Assumption	183
Residual Analysis: Constant Variance Assumption	183
Residual Analysis: Independence Assumption (Uncorrelated Errors)	184
Residual Analysis: Normality	185
Transforming Predicting Variables	186
2. Assumptions and Diagnostics Data Examples	190
3. Model Evaluation and Multicollinearity	195

4. Multicollinearity Data Examples	200
Knowledge Check	203
3.4 Case Study: Ranking States by SAT Performance	204
1. Exploratory Analysis	204
2. Regression Analysis	210
3. Ranking States by SAT	214
4. Model Fit Assessment	216
3.5 Case Study: Prediction of IMDb Movie Ratings	221
1. Exploratory Analysis	221
2. Regression Analysis	228
3. Prediction and Findings	237
<b>Unit 4: Generalized Linear Models</b>	<b>242</b>
4.1 Logistic Regression: Basic Concepts and Estimation	242
4.1.1. Introduction	242
4.1.2. Data Example	246
4.1.3. Model Description and Estimation	248
4.1.4. Model Estimation Data Example	251
Knowledge Check	254
4.2 Logistic Regression: Statistical Inference, Model Assessment, and Classification	255
4.2.1. Statistical Inference	255
4.2.2. Statistical Inference Data Example	260
Knowledge Check 1	262
The 4.2.3. Model Fit Assessment	263
4.2.4. Model Fit Assessment Data Example	269
4.2.5. Classification	275
Knowledge Check 2	278
4.3 Case Study: The Demographics of Obesity	279
4.3.1. Exploratory Data Analysis	279
4.3.2. Modeling and Prediction	281
4.3.3. Goodness of Fit	286
4.4 Poisson Regression: Basic Concepts and Estimation	291
4.4.1. Introduction	291
4.4.2. Data Examples	296

4.4.3. Model Description and Estimation	299
4.4.4. Model Estimation Data Example	302
Knowledge Check	306
4.5 Poisson Regression: Statistical Inference, Model Assessment	307
4.5.1. Statistical Inference	307
estimated regression coefficients 4.5.2. Statistical Inference Data Example	312
Knowledge Check 1	315
4.5.3. Model Fit Assessment	316
4.5.4. Model Fit Assessment Data Example	320
Knowledge Check 2	323
Practice Midterm 2	324
<b>Unit 5: Variable Selection</b>	<b>330</b>
5.1 Basics of Variable Selection	330
5.1.1. Introduction	330
Lecture 5.1.2 – Data Examples	333
Lecture 5.1.3 – Prediction Risk Estimate	340
Lecture 5.1.4 – Model Search	346
Lecture 5.1.5 – Model Search Data Examples	350
LECTURE 5.2: REGULARIZED REGRESSION	359
Lecture 5.2.1 – Regularized Regression: Penalties	359
Lecture 5.2.2 – Regularized Regression: Approaches	363
Lecture 5.2.3 – Regularized Regression: Data Examples	370
LECTURE 5.3: DATA ANALYSIS EXAMPLE	379
Lecture 5.3.1 – Emergency Department Healthcare Costs	379
Lecture 5.3.2 – Exploratory Data Analysis	385
Lecture 5.3.3 – Multiple Regression: Fitted Model and Residual Analysis	393
Lecture 5.3.4 – Variable Selection	399
Lecture 5.3.5 – Findings	409
<b>UNIT 6 – OTHER REGRESSION MODELS</b>	<b>418</b>
Lecture 6.1.1 – Weighted Least Squares Regression	418
Lecture 6.1.2 – Robust Regression	422
Lecture 6.1.3 – Nonlinear & Nonparametric Regression	424
LECTURE 6.2: OTHER REGRESSION MODELS (PART 2)	427
Lecture 6.2.1 – Time Series Regression	427

Lecture 6.2.2 – Spatial Regression	430
Lecture 6.2.3 – Mixed Effects Models	432
Lecture 6.2.4 – Regression Analysis: Overview	435

## 1.1: Estimation & Statistical Inference

### 1. Basics

**Regression analysis** is one of the simplest ways we have in statistics to investigate the relationship between two or more variables in a non-deterministic way. And it has very wide applicability to fields like healthcare policy, finance, marketing, elections, you name it. The two models we will cover in this course are **standard linear regression** and **generalized linear regression**. What we'll begin today is with the simplest regression analysis, which is simple linear regression.

#### Case Studies

### Example 1

A company, which sells medical supplies to hospitals, clinics, and doctor's offices, had considered the effectiveness of a new advertising program.

Management wants to know if the **advertisement** is related to **sales**.

This company intends to increase the sales with an effective advertising program.



Throughout the course, we will regularly come back to a set of example scenarios to help demonstrate in practice the theories taught in the course. Our first example is a company which sells medical supplies to hospitals, clinics, and is considering the effectiveness of a new advertising program. They want to identify whether there is a relationship between sales and their expenditure on advertising. We have data for 25

offices, and for each office we have two variables of interest. One is sales, measured in thousands. And the other one is advertising expenditure, measured in hundreds.

## Data Example 1

The company observes for **25 offices** the yearly sales (**in thousands**) and the advertisement expenditure for the new program (**in hundreds**)

Sales	ADV
963.50	374.27
893.00	408.50
1057.25	414.31
1183.25	448.42
1419.50	517.88
...	

So if you look at the first row of this data, you have a value of 963.5 in sales, which means, for that specific office, the amount of sales is 963,500. For advertising expenditure, the amount of the expenditure is 37,400.

## Example 2

- The principle of purchasing power parity (**PPP**) states that over long periods of time **exchange rate** changes will tend to offset the differences in **inflation rate** between two countries.
- In an efficient international economy, exchange rates would give each currency the same purchasing power in its own economy. Even if it does not hold exactly, the **PPP** model provides a benchmark to suggest the levels that exchange rates should achieve.

The second example is related to economic theory. Here, we are interested in evaluating and testing the principle of purchasing power parity, which states that over long periods of time, exchange rate changes will tend to offset the difference in inflation rate between two countries. In a perfect world, if we were to have an efficient international economy, what that means is that exchange rates would give each currency the same purchasing power in its economy. What we're trying to identify here

is the relationship between the currency exchange rates and inflation rates across multiple countries.

## Data Example 2

The data are recorded for 41 countries, including both developed and developing countries. The data include the following columns.

Country	Inflation.difference	Exchange.rate.change	Developed
Australia	-1.2351	-3.1870	1
Austria	1.5508	1.4781	1
Belgium	1.0371	0.0395	1
Canada	0.0461	-1.6416	1
Chile	-18.4126	-20.6329	0

For this example, we have data across 41 countries, including developed and developing countries. And we're going to study the relationship between annual inflation difference and exchange rate change.

## Example 3

- In 2000 **Bush** and **Gore** were the main candidates for President in the U.S. Buchanan, a strongly conservative candidate, was also on the ballot. In the **state of Florida**, **Bush** and **Gore** essentially tied, hence the counts were examined carefully county by county.
- **Palm Beach County** exhibited strange results. Even though the people in this county are not conservative, many votes were cast for **Buchanan**. Examination of the voting ballot revealed that it was easy to mistakenly vote for **Buchanan (a conservative candidate)** when intending to vote for **Gore**. We will thus predict whether those who voted for **Buchanan** were indeed going for a conservative candidate.

In the third example, we are going to study data related to the presidential elections in 2000. State results, tallied on election night, gave 246 electoral votes to the republican candidate, George W. Bush, and 255 to democratic candidate, Al Gore, with three states too close to call that evening. Among the states, was Florida, that really mattered in the final count. As such, there was a recount of votes, in Florida, for weeks after the election date. After an intense recount process, and the court decision, George W. Bush won Florida by a margin of 537 votes, just a very, very small number of votes. We want to see whether there is a specific aspect in this election that could have overturned the decision on which candidate the court would decide to be the president. Particularly, in one specific county, Palm Beach County, the count for the independent candidate, Buchanan, was much higher than expected. And the reason is that Buchanan was a conservative candidate and we would have expected that those that voted for Buchanan could have voted for Bush as well. Why don't you see whether there is a relationship between the votes of the independent candidate and the republican candidate, and identify whether Palm Beach County was indeed an outlier. For this data, we have many more variables. What we're going to study is only the relationship between the votes of Buchanan and George W. Bush.

## Variables in Regression

The regression framework is characterized by the following:

1. We have one particular variable that we are interested in understanding or modelling, such as sales of a particular product, or the stock price of a publicly traded firm. This variable is called the **response (dependent) variable**, and is usually represented by Y.
2. We have a set of other variables that we think might be useful in predicting or modelling the response variable (say the price of the product, the competitors' price, and so on; or the profits, revenues, financial position of the firm, and so on). These are called the **predicting or explanatory (independent) variables**, and are usually represented by x<sub>1</sub>, x<sub>2</sub>, etc.

When we speak of regression data, what do we mean? The regression framework is characterized by the following.

We have one particular variable that we're interested in understanding or modeling, or testing. For example, sales of a particular product, the stock price of a publicly traded

firm. And this variable is often referred to as the target, or response variable, because this is a variable we're interested to model. Some textbooks will refer to this variable as the dependent variable as well. But, for consistency in this course, we're going to refer to the variable of interest--and regression--as the response variable. And we will represent it by Y.

We also have a set of other variables that we think might be useful in modeling the response variable. Say, the price of the product, if we're interested in modeling the sales, or the revenue financial position of the firm profits, if we're interested to model the stock price. These are called predicting, or explanatory variables. Often, textbooks refer to these variables as independent variables. And those variables are usually represented by Xs,  $X_1$ ,  $X_2$ , and so on. For this class, for consistency, I will refer to those variables as predicting, or explanatory variables.

One of the first things that you will need to learn, when you perform regression analysis, is to correctly identify which is the response variable, and which is your predicting variable, or variables.

## Variables in Regression

### RESPONSE VARIABLE versus PREDICTING VARIABLE?

**Response Variable:** It is a **Random Variable**. It varies with changes in the predictor/s along with other random changes.

**Predicting Variable:** It is a **Fixed Variable**. It does not change with the response, but it is set fixed before the response is measured.

You have to keep in mind that the response variable is a **random variable**, because it varies. It varies with changes in the predicting variable, or with other changes in the environment where we observe the response variable. So, it is a random variable. Whenever we're going to see Y in our derivations, we will use it as a random variable.

On the other hand, the predicting variable is a **fixed variable**. It does not change with the response, and it is set fixed, before the response is measured. So, we first observe the predicting variable. And then for those variables, for those values of the predicting

variable, then we will observe the response variable. For example, if we take the example, where we're interested in a relationship between sales and advertising expenditure, the company will first set what they would spend for advertisement at the beginning of that year. And then they will observe the sales at the end of that year. So first, they fix the advertising expenditure. But then, the sales will depend not only on what they spent, what they spent on advertising expenditure, but also on other factors.

So, here are a few examples, just to give you a feel for the difference between **response** and **predicting** variables:

The **effect** of several types of cholesterol medications on LDL levels in humans.

- Response variable: Change in LDL levels
- Predicting variable: Type of medication

The **relationship** between driving habits and fuel efficiency

- Response variable: miles per gallon of fuel
- Predicting variable: average driving speed

The **relationship** between college grade point averages and scores on the SAT

- Response variable: GPA
- Predicting variable: SAT score

In the first example, we're interested in the effect of several types of cholesterol medication on LDL levels. In this example, we are interested in controlling the LDL level. And we'll control it with, for example, different types of medications. So, what is fixed here is the type of medication. What varies, is the change in the LDL level. In the second example, we're interested in our relationship between driving habits and fuel efficiency. Again, here, what we're interested in, is the fuel efficiency. And that could be different, from one car brand to another. It could be different whether one drives on a highway, or on surface streets. And definitely will depend on the driving habits, like average driving speed.

Here, the response variable is the fuel efficiency measured by miles per gallon, where the predicting variable is the average driving speed.

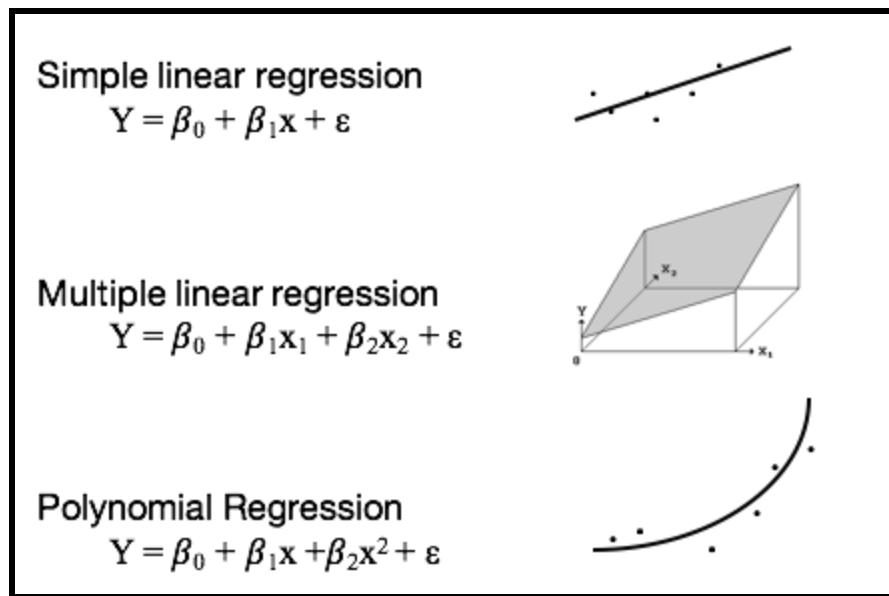
In the third example, we are interested in a relationship between college GPAs and SAT scores. The score of the SAT is the score of a test that students take at the end of high

school. And many colleges, if not all, are admitting students based on their SAT scores. The SAT score is the predicting variable, because it predicts the admission. And also, down the road, it's going to predict how students perform in a college. And that can be measured by the GPA, or the college grade point average.

A simple deterministic relationship between two factors, here in X and Y, is a linear relationship. But, we can generalize that to a relationship between two factors, in a non-deterministic way, as the first model, which is the simple linear regression (below).

We can extend that to a model where we include more than one predicting variable, which is called multiple linear regression. And this is going to be covered in a different topic of this course.

Further, we can take this to polynomial regression, or more complex relationships between the predicting variable and the response. For example, what we have on the bottom is the relationship between Y and a quadratic relationship with X.



We're going to see with simple linear regression that we have is one straight line. But that line doesn't fit perfectly the points. We can see the points are around the line.

In the second example, multiple linear regression, we can have a plane, if we have two predictions. Where, in the last case, you can see we are capturing a nonlinear relationship. In fact, all these models fall under the same framework of linear regression, even the last one, because the last one can be translated into a linear

regression. We can think of X and X-squared are two different predicting variables, and model those using a linear regression.

What you need to remember is that the simple linear regression, and linear regression in general, is a very general model. Practically, almost all of the models are some variations from linear regression.

## The Objectives of Linear Regression

There are three objectives in regression:

1. **Prediction:** We want to see how the response variable behaves in different settings. For example, for a different location, if we think about a geographic prediction, or in time, if we think about temporal prediction.
2. **Modeling:** Modeling the relationship between the response variable and the explanatory variables or predicting variables.
3. **Testing hypotheses** of association relationships.

Among the three examples, for the election example, we're interested in seeing whether the predicted vote count for Palm Beach County is similar to what we observe. If it's not similar, that means that the vote count for that county could be an outlier. And in terms of modeling, in the first example we discussed, we're interested in the relationship between sales and advertising expenditure. Do sales change with advertising expenditure and by how much? In the last example I presented, where we're interested in testing the purchasing power parity theory, we're interested to test whether the relationship between exchange rate and inflation is what we expect, according to this theory.

Why restrict ourselves to linear models? Well, they are simple to understand and they're simpler, mathematically. But most importantly, they work well for a wide range of circumstances (though definitely not for all). It's a good idea, when pursuing this, and actually in the statistical model, to remember the words of the famous statistician, George Box:

*"All models are wrong, but some are useful."*

**We do not believe that the linear model represents a true representation of reality. Rather, we think that, perhaps, it provides a useful representation of reality.**

Another useful piece of advice comes from another very famous statistician, John Tukey:

*"Embrace your data, not your models."*

While simple regression is a simple model, it has very wide applicability, and it can be generalized to much more complex models.

## 2. Estimation Method

We'll cover a simple linear regression topic. We're going to begin with the modeling framework, particularly the data structure, the model formulation, and assumptions. We'll also learn about the estimation approach of the simple linear regression.

In simple linear regression, the objective is to fit a non-deterministic linear model between the predicting variable  $x$  and the response variable  $y$ , which is equivalent to estimating the parameters  $\beta_0$  and  $\beta_1$ , where  $\beta_0$  is the intercept parameter, or the parameter that the value at which the line intersects the  $y$ -axis, and  $\beta_1$ , the slope parameter, which is the slope of the line we are trying to fit. In this model formulation, epsilon is the deviance of the data from the linear model. Again, the goal is to find the line that describes a linear relationship, that is, to find  $\beta_0$  and  $\beta_1$ , such that we fit this model.

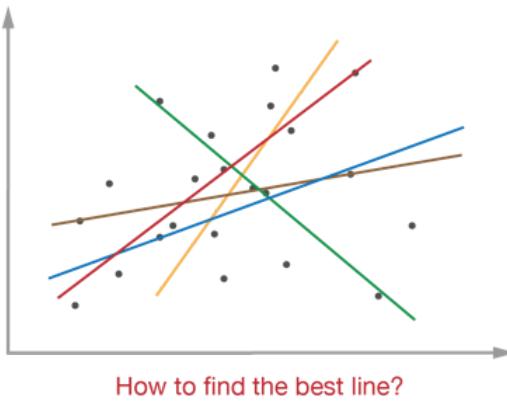
$$Y = \beta_0 + \beta_1 x + \epsilon$$

**Equivalently, estimating:**

1.  $\beta_0$       Intercept
2.  $\beta_1$       Slope

$\epsilon$  is the deviance of the data from the linear model

What does that mean? Let's plot the  $x$  and  $y$ . Now we can fit lines across those points -- many lines:



The question is, how to find the best line? What criterion to use to find the best line?

Here, we'll learn about the modeling framework for the simple linear regression. This is a general framework that you should use for other models, not only for the simple linear regression.

First we start with identifying the data structure. For simple linear regression, we have pairs of data consisting of a value for the response variable, and a value for the predicting variable. And we have  $n$  such pairs. The model formulation relates  $x$  and  $y$ , in this case, in a linear fashion.

**Data (pairs):**  $\{(x_1, Y_1), \dots, (x_n, Y_n)\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$

The assumptions of linear regression consist of:

- linearity/mean zero assumption, which means that the expectation of the deviances is zero.
- constant variance assumption, which means that the variance (represented in statistics by the Greek letter sigma squared) of the residuals is consistent for the given population.
- Independence assumption which means that the deviances are independent random variables.

Later we'll also assume that  $\varepsilon_i$  are normally distributed.

Let's go back to the assumptions, and digest each one at a time. **The linearity/mean zero assumption** means that the expected value of the errors is zero. That is, it cannot be true that for certain subgroups in the population, the model is consistently too low, while for others, it's consistently too high. A violation of this assumption will lead to difficulties in estimating  $\beta_0$ , and means that your model does not include a necessary systematic component.

**Constant variance assumption** means that it cannot be true that the model is more accurate for some parts of the population, and less accurate for other parts of the populations. A violation of this assumption means that the estimates are not as efficient as they could be in estimating the true parameters and better estimates can be calculated, it also results in poorly-calibrated prediction intervals.

**The assumption of independence** means that the deviances, or in fact the response variables (y's), are independently drawn from the data-generating process. That is, it cannot be true that knowing that the model under-predicts y for one particular case tells you anything or all about what it does for any other case. This violation most often occurs in data that are ordered in time, like in time series data, where areas that are near each other in time are similar to each other. Such time-related correlation is often called autocorrelation. Violation of this assumption can lead to very misleading assessments of the strength of the regression.

**The errors are assumed normally distributed.** This is needed if we want to do any confidence or prediction intervals, or hypothesis tests, which we usually do. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be very misleading.

## Identifying the Model Parameters

The goal of modeling is identifying the model parameters that are to be estimated using the observed data. In the linear regression model, in addition to the **intercept parameter  $\beta_0$**  and the **slope parameter  $\beta_1$** , there is a **third parameter: the variance of the deviances**. So in simple linear regression, we have three parameters to estimate.

What do we mean by parameters in statistics? Model parameters are **unknown quantities**, and they stay unknown regardless how much data are observed. We estimate those parameters given the model assumptions and the data, but through

estimation; we're not identifying the true parameters. We're just estimating approximations of those parameters. What does that mean?

Minimizing this problem is an optimization problem with respect to  $\beta_0$  and  $\beta_1$ , and the solution to this problem consists of estimators for  $\beta_0$  and  $\beta_1$ :

To estimate  $(\beta_0, \beta_1)$ , we find values that minimize squared error:

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \longrightarrow$$

$$\begin{aligned}\hat{\beta}_0 &= \\ \bar{y} - \hat{\beta}_1 \bar{x} & \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \equiv \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

We put a "**hat**" ("^") on top of those estimators to differentiate between the estimates and the true but unknown parameters being estimated. So  $\hat{\beta}_0$  is different than  $\beta_0$ ,  $\hat{\beta}_1$  is different than  $\beta_1$ , etc.

Let's take a closer look at the derivation of those estimators. *Again, what we're interested in is to minimize the sum of least squares, with respect to  $\beta_0$  and  $\beta_1$ .*

We will perform this optimization problem by taking the first-order derivatives of the objective function, and equate those to zero. We now have a system of two equations and two unknowns. And once we take the first order derivatives, we'll have a set of two linear equations.  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

Begin with the minimization problem:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

To solve we will take the first order derivatives of the function to be minimized and equate to 0:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \equiv \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 &= 0 \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 &= 0\end{aligned}$$

- Result into a system of linear equation in  $\beta_0$  and  $\beta_1$
- Solve using linear algebra
- Solutions to the system are  $\hat{\beta}_0$  and  $\hat{\beta}_1$

## Estimating Sigma Squared

The estimator for sigma-squares is sigma-squared-hat, and is the sum of the squared residuals, divided by  $n - 2$ . (BELOW)

The **sampling distribution** of this estimator is chi-square, with  $n - 2$  degrees of freedom (more on this in a moment). This is under the assumption of normality of error terms.

We use the epsilon i hat as proxies for the deviances or the error terms. We don't have the deviances because we don't have  $\beta_0$  and  $\beta_1$ . But if we replace  $\beta_0$  and  $\beta_1$  with  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , we get the deviances with a hat, and now we are estimating sigma square, based on those residuals. The estimator of the variance of the error terms is now the sample variance.

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi_{n-2}^2$$

(chi-squared distribution with n-2 degrees of freedom)

Assuming  $\hat{\epsilon}_i \sim \epsilon_i \sim N(0, \sigma^2)$



Estimating  $\sigma^2 \leftarrow$  Sample variance

And let me review the sample variance estimation approach.

### What is the sample variance estimation?

**Basic statistic concept:**

Consider  $Z_1, \dots, Z_n \sim N(\mu, \sigma^2)$  with  $\mu$  and  $\sigma^2$  unknown

The sample variance estimator:  $s^2 = \frac{\sum (Z_i - \bar{Z})^2}{n-1} \rightarrow \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$

We start with zs that are normally distributed with mu (the mean of the population) and sigma square (the variance). Both parameters are unknown, and the sample variance estimator, we often use the notation s square, is the sum of the  $Z_i$  minus their average, squared, divided by n-1. And from basic statistics, the estimated sampling distribution of  $s^2$  is a chi-square distribution with n-1 **degrees of freedom**.

Why n-1 degrees of freedom? We lose a degree of freedom because we replace the true parameter **μ (Actual Mean)** with **Z (Sampling Mean)**.

Recall that  $\epsilon_i = (y_i - (\beta_0 + \beta_1 x_i))$

↑ Replaced by

$$\hat{\epsilon}_i = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

} We use two degrees of freedom because  
 $\beta_0 \leftarrow \hat{\beta}_0$   
 $\beta_1 \leftarrow \hat{\beta}_1$

Thus, assuming that  $\epsilon_i \sim N(0, \sigma^2)$

$$\rightarrow \hat{\sigma}^2 = \text{MSE} \sim \chi_{n-2}^2$$

(This is called the sampling distribution of  $\hat{\sigma}^2$ )

Now let's go back to the estimator of sigma square, under simple linear regression. Our estimator looks just like the sample variance estimator, except that we use  $n - 2$  degrees of freedom because we've replaced the deviances with the residuals. And to do so, we lose two degrees of freedom because we replaced the two parameters  $\beta_0$  and  $\beta_1$ , with their estimators. So in this case, we are using the two degrees of freedom, each one degree of freedom for each parameter. Under the normality assumption, that the sample distribution of the variance estimator is chi-square with  $n - 2$  degrees of freedom.

We also use the notation or the equivalent formulation of this estimator as the mean squared error.

In simple linear regression, we're interested in the behavior of  $\beta_1$ . We can expect  $\beta_1$  to be positive, negative, or in fact, close to zero.

If we have a **positive value** for  $\beta_1$ , then that's consistent with a **direct relationship** between the predicting variable  $x$  and the response variable  $y$ . For example, higher values of height are associated with higher values of weight. A **negative value** of  $\beta_1$  is consistent with an **inverse relationship** between  $x$  and  $y$ . For example, lower inflation rate is associated with a higher savings rate.

But we also have situations when the value of  $\beta_1$  is **close to zero**. In that case, we interpret that there is not a significant association between predicting variables, between the predicting variable  $x$ , and the response variable  $y$ .

We interpret the least squares estimated coefficients as follows:

- $\beta_1$  is the estimated expected change in the response variable associated with one unit of change in the predicting variable.
- $\beta_0 \text{ hat}$  is the estimated expected value of the response variable, when the predicting variable equals zero.

When we interpret whether the relationship between  $x$  and  $y$  is positive, negative, or there is no relationship, we can also use  $\hat{\beta}_1$ . But when we make statements about the relationship, we always have to mention the statistical significance, whether statistically significantly positive, statistically significantly negative, or no statistical significance. So the relationship  $\hat{\beta}_1$ , can capture the relationship between  $x$  and  $y$ , but in a statistical statement framework.

### 3. Estimation Data Examples

In this lesson we'll see examples of the implementation of the estimation concept. I will show you using this R statistical software the implementation of the linear regression model. And we will discuss the output of the linear regression model fit in R. We'll return to the first example that I illustrated, a simple linear regression model.

In this example, we are interested in a relationship between sales and advertising expenditures under a new advertising program. The first question we'd like to address: which is the response and which is the predicting variable within this example? In this example, we're interested in modeling the sales, and again the sales can vary with the expenditure in advertisement but also with other factors. So, in this example, the response variable consists of the sales and the predicting variable consists of the advertising expenditure.

In this example, we may want to:

- fit a linear regression.
- identify the estimated regression coefficients  $\beta_0$  hat and  $\beta_1$  hat.
- interpret the coefficients.
- see whether the relationship between sales and advertising expenditure.
- identify to quantify with the amount of increase in the sales for each additional sales in dollars invested in advertising expenditure.
- predict the sales for a specific value of an advertising expenditure, for example \$30,000.
- estimate the error variance, or even estimate sales for a very large amount of advertising expenditure, p

The first step in using R statistical software is to read the data from a file into R. One common function used in R is the command `read.table()`, with which we will need to specify the file where the data are, the separator of the data values in the file, and whether the data file the columns in our data file have names or not. We also want to extract specific predictors variables from this data. In this case sales will be in the first column, and advertising expenditure going to be in the second column.

R Function: `lm()`

Fitting a linear regression model is very simple in R. We can use the **function `lm()`**. We will need to specify the response variable, in this case sales, on the left, and the

predicting variable, which is  $\text{Sales}$  the advertising expenditure on the right, separated by a tilde (~).

What I'm showing you here is the summary of the model:

It's a portion of that summary from which we can obtain the estimated coefficients, the estimated regression coefficients  $\beta_0$  hat and  $\beta_1$  hat, and the estimated variance.

From this output the estimated intercept is -157.33. The estimated slope is 2.7721 and the estimated standard deviation, not the variance, it's 101.4, and you can see those values highlighted in the output.

Let's go back to the questions we wanted to address. We already fit a linear regression. Now we can identify the estimated regression coefficients from the output which are already provided.

- a. Fit a linear regression. What are the estimated regression coefficients and the estimated regression line?

**Solution:** Estimates  $(\beta_0, \beta_1)$  are (-157.33, 2.77) and the *regression equation is*:

$$\text{Sales} = -157.33 + 2.77 \text{ Adv Expenditure}$$

- b. Interpret the coefficients.

**Solution:** The sales increase by \$2770 with each \$100 additional expenditure in advertisement. Or the sales increase with \$27.7 with each dollar invested in advertisement expenditure.

- c. What does the model predict as the advertisement expenditure increases for an additional \$1,000?

**Solution:** The increase in sales is  $10 \times 2.77 = 27.7$  thousands.

I

Based on those estimates we can now write the regression equation:

$$\text{sales} = \text{estimated intercept} + \text{estimated slope} * \text{advertising expenditure}$$

To interpret the coefficients, we have to keep in mind the fact that sales and advertising expenditure are measured in different units. Sales are measured in thousands, and advertising expenditure are measured in hundreds. So using this difference between the units, we want to interpret that the sales increased by \$2,770 with each additional \$100 in expenditure for advertisement. Or, we can also interpret that the sales increase with \$27.7 with each dollar invested in advertising expenditure.

If we want to quantify how much more we would derive in sales with an additional \$1,000 in advertising expenditure, again, we have to convert \$1,000 in advertisement expenditure into the units that we measured the data, hundreds. So we will have 10 units, 10 hundreds, and when we obtain, when we plot this number in our model, it's not a thousand, we plug in as a 10. 10 times the estimated value for the slope gives us 27.7 thousand. Again, we measure sales in thousand. So the increase in sales with every additional thousand dollars is 27.7 thousand.

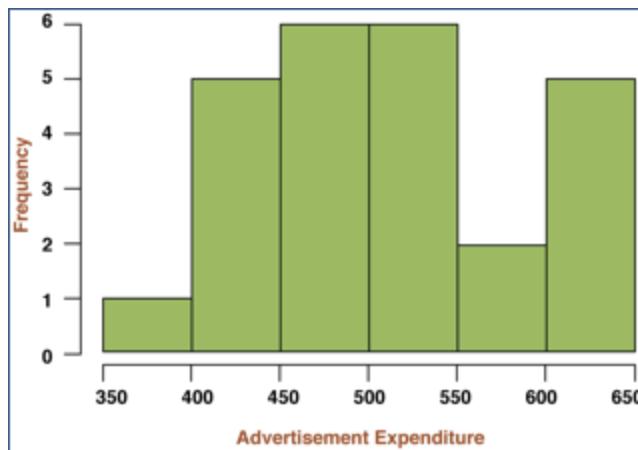
**Pay particular attention to the units of both the response and the predicting variables for correct interpretation of the model.**

If we were interested in predicting the sales for an advertisement expenditure of 30,000, again we have to convert 30,000 into original units, so that is 30,000 original units will be 300. We plug this in the estimated regression line, and we obtain 673,000 in sales.

What is the estimate of the error variance? **Remember what we get from the output is not the variance, it's the standard deviation. So what we'll need in order to obtain the estimate of the variance is take the square of that value (sigma hat or residual square error = 101.4;  $101.4^2 = 10,281.96$ )**

What could we say about the sales for an advertisement expenditure of 100,000? It's a very large value, way out of the range and 1,000 units away from the maximum value of advertising expenditure we observe.

This is the histogram of the advertisement expenditure and you can see that the maximum value we've observed is 650, and, which is 65,000.



Because the value that we would like to predict sales of advertising expenditure 100,000 is much larger than the values we observed, this is what we call **extrapolation**. And in this case we cannot really say much about the sales because it's not within the range of the observed axis. We cannot assume that a relationship between sales and advertising expenditure is the same beyond the range of our axis. It's possible that sales will tip off once we, once the sales, once advertising expenditure increases up to a specific value, for example.

## 4. Statistical Inference

We'll now move from estimation to statistical inference. In the lesson we'll learn about statistical properties of the estimated coefficients and how to use statistical properties in making inferences such as confidence intervals and hypothesis testing.

We will begin with deriving the statistical properties of the estimator for  $\beta_1$ .

The expectation of  $\hat{\beta}_1$  is equal to  $\beta_1$  which is the true parameter, and the variance of this coefficient of this estimator is sigma square divided by  $S_{XX}$ :

For the slope parameter  $\beta_1$ ,  $E(\hat{\beta}_1) = \beta_1$   
we can show  $\text{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}$

I will show you a brief derivation of the expectation of  $\hat{\beta}_1$  and we can use the same sequence of derivations for the variance of  $\hat{\beta}_1$ .

Let's go back to the formula of the estimator for  $\beta_1$ . We could write that as the sum of a constant times the random variable  $Y_i$  where that constant is  $c_i$ :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}} \text{ but } x_i \text{ fixed} \rightarrow \frac{x_i - \bar{x}}{S_{xx}} = c_i \text{ fixed}$$

What that means is that  $\hat{\beta}_1$  is a **linear combination of random variables**. We know from basic statistics that the expectation of a linear combination of random variables is equal to the linear combination of the expectations.

$$E[\hat{\beta}_1] = E\left[ \sum_{i=1}^n c_i Y_i \right] = \sum_{i=1}^n c_i E[Y_i]$$

Now we replace the expectation of the random variable of the response variable  $Y_i$  with the linear relationship in  $X$ . We divide that into two sums and now the first sum is equal to zero because the sum of the constant is equal to zero, and the sum of the constant times  $x_i$  is equal to one.

$$= \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i$$

||      ||

What we get is that the expectation of the estimator for the slope parameter is exactly  $\beta_1$ .

$$= \beta_1 \rightarrow E[\hat{\beta}_1] = \beta_1$$

This property, the fact that the expectation of the estimator is exactly the true parameter that we're estimating, is called **unbiasedness**. What that means is that  $\hat{\beta}_1$  is an unbiased estimator for  $\beta_1$  and that is a very important statistical property of an estimator.

Let's dive more into the properties of this estimator. I will remind you that the  $\hat{\beta}_1$  is a linear combination of the response variables,  $Y_i$ . **Under the normality assumption,  $\hat{\beta}_1$  is thus a linear combination of normally distributed random variables, and thus  $\hat{\beta}_1$  is also normally distributed.** So along with the expectation of variance from the previous slide, now we get the distribution for  $\hat{\beta}_1$ .

Furthermore,  $\hat{\beta}_1$  is a linear combination of  $\{Y_1, \dots, Y_n\}$ . If we assume that  $\epsilon_i \sim \text{Normal}(0, \sigma^2)$ , then  $\hat{\beta}_1$  is also distributed as

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$\hat{\beta}_1 = \sum_{i=1}^m c_i Y_i$  a linear combination of normally distributed random variables

$\hat{\beta}_1 \sim \text{Normally distributed}$

However, the sampling distribution of  $\beta_1$  hat is not useful because sigma squared is unknown. In order to get a full specification of this distribution, we must replace sigma squared with an estimator. We can use the estimator we discussed in the previous lesson, the **mean square error**, or the sum of square residuals divided by  $N - 2$ .

Because this estimator has a chi-square distribution with  $N - 2$  degrees of freedom, the sampling distribution of  $\beta_1$  hat is a t distribution with  $N - 2$  degrees of freedom. The  $N - 2$  degrees of freedom come from the fact that the distribution of the variance of the estimator variance is a chi-square distribution with  $N - 2$  degrees of freedom.

### Sampling Distribution of $\hat{\beta}_1$ :

We do not know  $\sigma^2$ . We can replace it by MSE but then the sampling distribution becomes the t-distribution with  $n-2$  df.

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\hat{\sigma}^2 = \text{MSE} = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi^2_{n-2}$$

$$\left. \right\} \rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSE}}{S_{xx}}}} \sim t_{n-2}$$

We will use this sampling distribution to derive confidence intervals, and also to perform hypothesis testing procedures with respect to  $\beta_1$ . This is the confidence interval for  $\beta_1$

based on the normality assumption—that the deviances are normally distributed, or the data with the Y's are normally distributed.

Given the sampling distribution of  $\hat{\beta}_1$ , we can derive confidence intervals and perform hypothesis testing for  $\beta_1$ :

$$\left( \hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{MSE}{S_{xx}}} \right)$$

Let's look a little bit at this confidence interval. Again, the sampling distribution of the estimator  $\beta_1$  hat is a T distribution.

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2} \quad t - \text{interval for } \beta_1$$

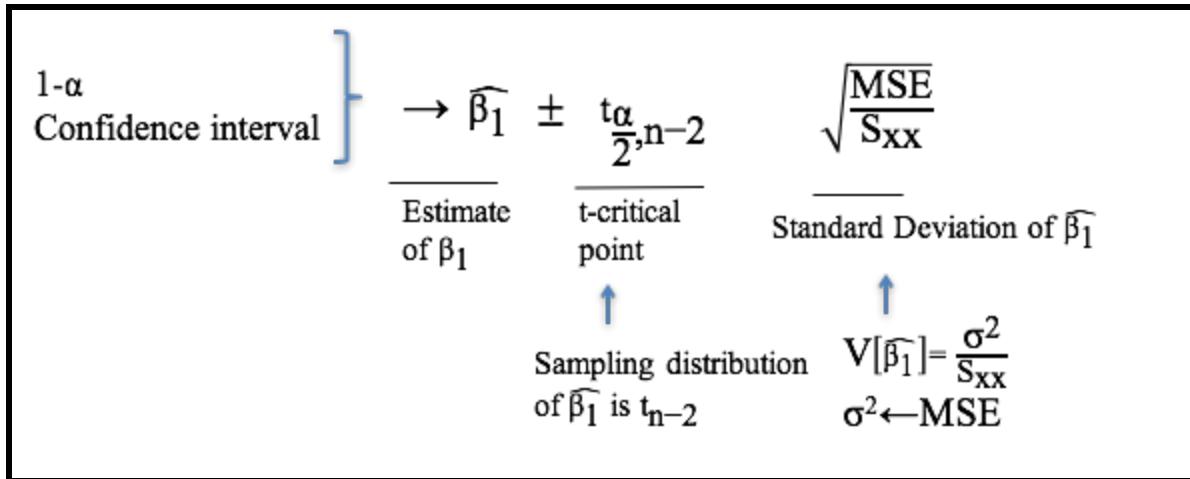
If we want to obtain a one minus alpha of a confidence interval then we can center the confidence interval at the estimated value for  $\beta_1$ , plus or minus the **critical point T**. This comes from the fact that the sampling distribution is a T distribution, and we carry over the degrees of freedom  $N - 2$ .

Alpha over two comes from the fact that we want a one minus alpha confidence interval, and we multiply this critical point with the standard deviation of  $\beta_1$  hat.

$$\text{Confidence interval } \left[ \hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{MSE}{S_{xx}}} \right]$$

Estimate of $\beta_1$	$t_{\frac{\alpha}{2}, n-2}$	$\sqrt{\frac{MSE}{S_{xx}}}$
	t-critical point	Standard Deviation of $\hat{\beta}_1$

Again, the T critical point comes from the fact from the sampling distribution of  $\hat{\beta}_1$ , and the standard deviation of  $\hat{\beta}_1$  comes from the variance of  $\hat{\beta}_1$ , which I shared with you previously, but replacing sigma squared with the estimator mean, the mean squared error.



If we want to perform hypothesis testing on  $\beta_1$ , we may want to test whether  $\beta_1$  is equal to zero versus the alternative that  $\beta_1$  is not equal to zero. The procedure is going to be very similar to testing for the mean parameter in a standard normal distribution problem. The T value now—that is, the difference between the data and the null hypothesis that is here—is the estimated value for  $\beta_1$ , minus the null value, in this case is zero, divided by the standard error of the estimator. **If that T value is large, reject the null hypothesis that  $\beta_1$  is equal to zero. If the null hypothesis is rejected, we interpret this to mean that  $\beta_1$  is statistically significant.**

One way we can test statistical significance is to use the t-test for  
 $H_0: \beta_1 = 0$  vs.  $H_a: \beta_1 \neq 0$

$$t\text{-value} = \frac{\hat{\beta}_1 - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\hat{\sigma}}$$

We reject  $H_0$  if  $|t\text{-value}|$  is large. If the null hypothesis is rejected, we interpret this as  $\beta_1$  being **statistically significant**.

In this course, I will use this terminology across many contexts. **Statistical significance means that  $\beta_1$  is statistically different from zero.** But what if we want to change the procedure to test whether  $\beta_1$  is equal to a constant versus  $\beta_1$  not equal to that constant, where that constant can be zero, but can also be a different value as well?

**How will the procedure change if we test:**  
 $H_0: \beta_1 = c$  vs.  $H_a: \beta_1 \neq c$  for some known  $c$ ?

Remember, this is the T value. We can replace zero with C, depending on the constant, the null value, in the null hypothesis. When we reject this is when a T value is larger than its critical point in absolute value.

$$t\text{-value} = \frac{\hat{\beta}_1 - c}{\text{se}(\hat{\beta}_1)} \text{ how large to reject } H_0: \beta_1 = c ?$$

For significance level  $\alpha$ , Reject if  $|t\text{-value}| > t_{\frac{\alpha}{2}, n-2}$

We can also make this decision based on a P value, which is going to be the sum of the tails on the left and on the right of the T value. If the P value is small, for example, smaller than .01, we would reject the null hypothesis.

Alternatively, compute P-value =  $2P(T_{n-2} > |t\text{-value}|)$

If P-value small ( $p\text{-value} < 0.01$ )

What if we were to change the procedure and were interested in testing whether the regression coefficient is positive or negative? That means we'll change the alternative hypothesis from  $\beta_1$  different from zero.

How will the procedure change if we test:

$H_o: \beta_1 = 0$  vs.  $H_a: \beta_1 > 0$

OR

$H_o: \beta_1 = 0$  vs.  $H_a: \beta_1 < 0$

Now we're interested in whether  $\beta_1$  is greater or less than zero. In this case the P value will change in the sense that we're interested in only one of the tails. For  $\beta_1$  greater than zero we're interested in the right tail. For  $\beta_1$  smaller than zero we're interested in the left tail.

What if we want to test for positive relationship

$H_o: \beta_1 \leq 0$  versus  $H_A: \beta_1 > 0$ ?

P-value =  $P(T_{n-2} > t\text{-value})$

What if we want to test for negative relationship

$H_o: \beta_1 \geq 0$  versus  $H_A: \beta_1 < 0$ ?

P-value =  $P(T_{n-2} < t\text{-value})$

The inference for the intercept parameter is going to be similar to the inference for the slope parameter.  **$\beta_0$  hat is also a linear combination of random variables, a linear combination of Y's.**

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} = \beta_0$$

$$Var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

We can derive similarly that the expectation of this estimator is equal to the true parameter, thus  **$\beta_0$  hat is an unbiased estimator of  $\beta_0$** , and the variance of this estimator is on the slide (above).

With this information, and with **the fact that  $\beta_0$  hat is a linear combination of normally distributed random variables, the sample distribution is also a T distribution**. Thus the confidence interval is going to be very similar to the confidence interval for  $\hat{\beta}_1$ . It's the center of the estimator  $\beta_0$  plus or minus the critical point from the T distribution times the standard error (below).

### Confidence interval:

$$\left( \hat{\beta}_0 - t_{\frac{\alpha}{2}, n-2} \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}, \hat{\beta}_0 + t_{\frac{\alpha}{2}, n-2} \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \right)$$

## 5. Statistical Inference Data Examples

The topic is simple linear regression and what I'm going to show you next is an example using the R statistical software to look into statistical inference of the regression coefficients. We're going to use the output from the regression feed to make statistical inferences on the estimated coefficients using the example of a company that wants to know if advertisement expenditure is related to sales.

The question we want to address: **what inferences can be made on the regression coefficients?**

To answer that, we need to find:

- the estimated coefficient  $\beta_1$  and its variance.
- the sample distribution of  $\beta_1$  and the estimator.
- the estimated coefficient for the intercept  $\beta_0$  and its variance.
- whether the coefficient  $\beta_1$  is statistically significant, which we will do using p-value of the hypothesis testing procedure.
- whether  $\beta_1$  is statistically positive, which we will determine by performing our hypothesis test, and we'll draw our conclusion based on a p-value.

We'll also learn how to derive a 99% confidence interval for  $\beta_1$ . And I will conclude with an interpretation of the p-value in the general context in a general hypothesis testing procedure context.

This is the output of the regression model (from a previous lesson):

*summary(model)*

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-157.3301	145.1912	-1.084	0.29
adv	2.7721	0.2794	9.921	8.87e-10
Residual standard error: 101.4 on 23 degrees of freedom				

- a. The estimate for  $\beta_1$  is 2.7721. The variance estimate is 0.2794<sup>2</sup>. The sampling distribution is a t-distribution with 23 degrees of freedom.
- b. The estimate for  $\beta_0$  is -157.3301. The variance estimate is 145.1912<sup>2</sup>.
- c. The estimate for  $\beta_1$  is statistically significant, as evidenced by a p-value of  $8.87 \times 10^{-10}$

Using this output, we not only can get the estimated values of the coefficients, but also the standard errors and the p-values for the statistical significance. This is how we can extract the information on the estimated values for  $\beta_1$  (2.7721). If we want the estimated variance (of  $\beta_1$ ) we need to take the square of the standard error, so this value (0.2794) squared.

The sample distribution is a t-distribution with 23 degrees of freedom. This is available also in the output. The estimated value for the intercept coefficient is (-157.3301), and the variance estimator (std. error of intercept) is provided (145.1912), and we take the square of that in order to get the variance. This is a standard error, not the variance, again.

If we want to test the procedure whether  $\beta_1$  is statistically significant, then the p-value is going to be (8.87e-10). We can see the p-value is very small, which means that we reject the null hypothesis that  $\beta_1$  is equal to zero, and conclude that  $\beta_1$  is statistically significant.

To test whether  $\beta_1$  is statistically positive, now we have to change the hypothesis, the alternative hypothesis, to  $\beta_1$  greater than zero.

R Functions: `pt()` and `confint()`

For this we can use the output but we need to adjust the p-value, the t-value stays the same. For the p-value we will need to compute that p-value as only taking into account only the right tail. And we can do that using the **function 'pt()'**, which stands for the probability of a t-distribution. That function is going to give us the left tail of evaluating the

```
tvalue = 9.921
1 - pt(tvalue, 23)
[1] 4.433214e-10
confint(model, level=0.99)
              0.5 %      99.5 %
(Intercept) -564.930546  250.27032
adv           1.987712   3.55652
```

quantile equal to t-value. In order to get the right tail of that distribution we'd have to take one minus that probability. And that p-value is again very small ( $4.433214e-10$ ), leading us to conclude that  $\beta_1$  is statistically positive.

In order to estimate a confidence interval, we can use the **function 'confint()'**. This is the confidence interval estimation and the values for this function we get confidence intervals for both the intercept (-564.930546 to 250.27032) and the slope (1.987712 to 3.55652).

Here we're only interested in the confidence interval for  $\beta_1$ , and the values for the lower and upper bound (answer f. is below)

e.  $\beta_1$  statistically positive:  $H_A: \beta_1 > 0$

We accept the alternative hypothesis because p-value is  $4.43 \times 10^{-10}$ . (The test statistic is 9.921.)

f. The 99% confidence interval for  $\beta_1$  is (1.988, 3.557)

g. The p-value is a *measure of how rejectable the null hypothesis is*. The smaller the p-value, the more rejectable the null hypothesis is for the observed data.

The lowest value of the 99 confidence interval is 1.9, the highest is 3.5. The interpretation of a confidence interval is a little tricky. You have to remember that if we have a 99% confidence interval it means that one out of 100, the interval will miss including  $\beta_1$ , i.e.  $\beta_1$  will not be in the confidence interval one time out of a hundred.

So what is a p-value? **P-value is a measure of how rejectable the null hypothesis is.** The smaller the p-value is, the more rejectable the null hypothesis is for the observed data, given the observed data. It's not the probability of rejecting the null hypothesis, nor is it the probability that the null hypothesis is true.

**The ASA's Statement on p-Values: Context, Process, and Purpose**

<http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.WIka6FQ-cUw>

What is a p-value?

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

- 1.P-values can indicate how incompatible the data are with a specified statistical model.
- 2.P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- 3.Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- 4.Proper inference requires full reporting and transparency
- 5.A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- 6.By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

## Knowledge Check

1. The estimators for the regression coefficients are:
  - A. Biased but with small variance
  - B. Unbiased under normality assumptions but biased otherwise.
  - C. Unbiased regardless of the distribution of the data.
  
2. The assumption of normality:
  - A. It is needed for deriving the estimators of the regression coefficients.
  - B. It is not needed for linear regression modeling and inference.
  - C. It is needed for the sampling distribution of the estimators of the regression coefficients and hence for inference.
  - D. It is needed for deriving the expectation and variance of the estimators of the regression coefficients.<sup>1</sup>

---

<sup>1</sup>1. C  
2. C

## 1.2 Prediction and Model Evaluation

### 1. Regression Line: Estimation & Prediction

*The topic is simple linear regression, and we'll cover estimation and prediction of the regression line.* Specifically, we will learn to:

- differentiate between estimation and prediction
- estimate confidence intervals,
- obtain the expectation and the variance under estimation and prediction
- derive confidence intervals.

Prediction is often a main objective of regression analysis. Prediction can be in time, geography, or just simply for other settings of the predicting variable. **But prediction is not the same as estimation.** This is not only due to the interpretation, but also in the uncertainty level of the prediction mean response. Particularly, **the uncertainty in estimation comes from estimation alone; whereas for prediction the uncertainty comes from the estimation of the regression parameters and from a newness of the observation.**

Let's distinguish between the concepts of prediction and estimation.

Let's say we're interested in the mean response evaluated at this predicting value  $x^*$ . Under estimation,  $x^*$  can be one of the observed values that we fitted to the model. The interpretation at this value  $x^*$  is that the estimated regression line is the average estimated mean response for all settings under which the predicting variable is equal to  $x^*$ . So it's really an average across all possible settings when we could observe  $x^*$ .

Whereas for prediction at this value  $x^*$ , when  $x^*$  is considered a new observation under a new setting, the predicted regression line is interpreted as the estimated mean response for one setting under which the predicting variable is equal to  $x^*$ . **So in estimation, we're averaging across all possible settings for prediction. In prediction, we focus on one particular setting.**

When we estimate the regression line, it's a very simple formula. We plug in  $x^*$  in the regression line, and this is going to be the estimator because the estimators of those

two coefficients of the slope and the intercept are normally distributed; so is  $\hat{y}$ .

At some selected value of  $x$  (say  $x^*$ ), we estimate the “mean response” of  $Y$  (or the regression line) via

$$\hat{y} | x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Because the estimators of  $\beta_0$  and  $\beta_1$  are normally distributed, so is  $\hat{y}$ . That means we can draw inference using  $\hat{y}$  if we know expected value and variance.

$\hat{y}$  is going to have a normal distribution and the expectation of  $\hat{y}$  is very easy to derive: it's equal to  $\beta_0$  plus  $\beta_1 x^*$ , so this is actually the true regression line. The variance is as on the slide.

$\hat{y}$  has a normal distribution with

$$E(\hat{y} | x^*) = \beta_0 + \beta_1 x^*$$
$$Var(\hat{y} | x^*) = \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$$

We can see the variances increase as  $x^*$ , the value of the predictive value, moves away from the range of the average of the predicting variable values.

**Note:** variability is smallest if we check the regression line at, the middle of the  $X$ 's; i.e., at  $x^* = \bar{x}$

So the variance is going to be smaller at the center of the average and is going to increase as we go away from the average.

The uncertainty in the estimated regression line is going to be higher as the predicted value  $x^*$  is further away from the average.

If we want to construct a confidence interval for the mean response, it's very similar to what we did before for the estimated coefficients.

We center the estimated regression line, plus or minus the critical point for T times the standard error. Similarly, because of the variance changes with  $x^*$ , the confidence interval also will be wider the further  $x^*$  is from the mean ( $\bar{x}$ ):

$$\hat{y} | x^* \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

- ✓ Interval length depends on  $x^*$
- ✓ As  $x^*$  changes, we can construct a confidence band for
- ✓ Confidence bands show why extrapolation fails

If we take several values of  $x^*$  and we construct such confidence intervals, we get what we call the **confidence band** (which will be covered later).

In contrast to estimation, prediction contains two sources of uncertainty:

1. The new observation and
2. the parameter estimation.

The second source of uncertainty comes also in estimation of the response for  $x^*$ , but the first source of uncertainty is because we observe (we are predicting their response under a new setting).

So how does that translate into uncertainty of the prediction of a new response? A variation due to the estimation is similar to the variation of the estimated regression line. We already saw this variance, but now we're adding this variation of a new measurement, which is the sigma squared, just a variance of the deviances.

If you add those up, this is what the formula looks like:

1. Variation of the estimated regression line:  $\sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$
2. Variation of a new measurement:  $\sigma^2$

The new observation is independent of the regression data, so the total variation in predicting  $y | x^*$  is

$$\sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) + \sigma^2 = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)$$

So we're adding the sigma squared to the variance of the estimated regression line.

Now you see that if we combine these two, we'll get a sigma squared times one plus one over N plus the difference between  $x^*$  minus  $x\bar{x}$  squared divided by  $S_{xx}$ . As a reminder, the formula for  $S_{xx}$  is the sum of the differences between the X's and the average squared differences.

If we want to obtain a prediction interval for a future value  $y^*$ , this is again very similar for the estimation of a new response. It's centered at the predicted response plus or minus the critical point times the standard error.

A  $100(1-\alpha)\%$  ***prediction*** interval for a future  $y^*$  (at  $x^*$ ) is

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

The standard error is different from the standard error from the confidence interval for the regression line because we have this additional 1 (circled above).

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$  is the same as the line estimate, but the interval is wider than the confidence interval for the mean response.

Note: the predicted regression line is the same as the estimated regression line. What is different, again, is this standard error.

Additional from ppt notes:

The prediction interval should not be confused with a confidence interval for a fitted value, which will be narrower. The prediction interval is used to provide an interval estimate for a prediction of  $y$  for one member of the population with a particular value of  $x^*$ ; the confidence interval is used to provide an interval estimate for the true average value of  $y$  for all members of the population with a particular value of  $x^*$ .

## 2. Regression Line: Estimation & Prediction Examples

This unit covers the implementation of estimation and prediction of the regression line using the R statistical software. I'm going to show you with R how to estimate and predict the regression line, and how to interpret the R output.

And we are going to return now to the example of the relationship between the sales and advertising expenditure to determine what inferences can be made on the prediction of the sales, given a targeted advertisement expenditure.

This is a set of questions that we may address in this context:

- What sales would you predict for an advertisement expenditure of 30,000?
- What is the variance estimate of the estimated predicted sales for this value?
- Can we get a lower and upper bound for the sales under this value of advertisement expenditure with 99% confidence level, or with a 95% confidence level?
- How do the confidence levels differ?

We also will compare the confidence intervals, and interpret those when we compare the confidence bands versus the prediction bands.

So let's go back to the summary of the model. This is what we've seen before in a different lesson. This time I added some additional R code.

```
summary(model)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -157.3301 145.1912 -1.084   0.29
adv          2.7721   0.2794  9.921 8.87e-10
---
Residual standard error: 101.4 on 23 degrees of freedom
xbar = mean(ADV)
n = 23+2
mse = 101.4^2
var.beta1 = 0.2794^2
sxx = mse / var.beta1
pred.var = mse * (1 + 1/n + (xbar - 300)^2 / sxx)
pred.var
[1] 14286.16
```

For the advertising expenditure of 30,000, the predicted sales is replacing X with 300. Remember the units for the advertising expenditure is \$100. So what we input in the regression line is \$300, not \$30,000. Predictive sales are going to be 673,000.

If we want to compute the variance of the predicted sales, this is the formula that we saw before from a different lesson.

In this formula, we need to input the estimated variance, the sample size, the S<sub>XX</sub>, X\*, and the average of the X. For the estimated variance, this is the square of the mean square, of the standard residual error, which is equal to the mean square error

**(mse=101.4^2)**. For the sample size it's going to be the degrees of freedom which is 23 plus two (**n=23+2**). Remember that the degrees of freedom are equal to n - 2, so in order to get n, we just add the two to the degrees of freedom.

To get the value for S<sub>XX</sub>, we will use the formula from a previous lesson, the formula for the variance  $\sigma$  of the estimator for  $\beta_1$ . Recall that variance equals  $\sigma$  divided by S<sub>xx</sub> **[S<sub>xx</sub> = MSE/ V( $\beta_1$ )]**. So I can use that formula to get S<sub>xx</sub>, which is going to be the variance or the estimated variance divided by the variance for  $\beta_1$ . Both values in this formula are available in the output.

You don't have to do extra work, you just get those two values from the output, and then you will get the value for S<sub>XX</sub>.

For the average across X (**xbar = mean(ADV)**), you just take the average across the values you observed for the predicting variable .

And now you plot all these values, and this is what the line could shows you on the bottom. I'm really writing the formula for the variance of the predictive sales, and the value of the variance is 14,286.

**b. The variance of the predicted sales is**

$$\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right) = 14286.16$$

## R Function: predictLM()

What are the lower and upper limits of predicted sales for an advertisement expenditure of \$30,000 at 99% confidence interval? How will the limits change if we lower the confidence level to 95%? To predict or how to derive confidence intervals and prediction intervals for a linear model, we will use the **function predictLM()**.

Here is the new data: advertising expenditure equal to 300 units, or \$30,000. And we need to specify whether we want a prediction interval and we need to specify the confidence level. Remember, we are interested in comparing

```
new = data.frame(adv = 300)
predict.lm(model, new, interval = "predict", level = 0.99)
  fit     lwr      upr
1 674.3047 338.712 1009.897
predict.lm(model, new, interval = "predict", level = 0.95)
  fit     lwr      upr
1 674.3047 427.0146 921.5948
```

99% and 95%. So this will be the two intervals, the 99% and 95%, that I have provided with these lines of code. If we want to do confidence intervals, the only thing that's going to change is we need to specify the type of interval as confidence. And those are going to be the values for the confidence intervals.

c. A 99% **prediction** interval at an advertisement expenditure of \$30,000 is (338.712, 1009.897). A 95% interval is (427.014, 921.594).

- d. Compare the confidence bands of the estimated regression line versus the predicted regression line. Interpret.

Let's compare the confidence intervals versus prediction intervals. You can see that the prediction interval is significantly wider than the confidence intervals.

```

new = data.frame(adv = 300)
predict.lm(model, new, interval = "predict", level = 0.99)
  fit     lwr      upr
1 674.3047 338.712 1009.897
predict.lm(model, new, interval = "predict", level = 0.95)
  fit     lwr      upr
1 674.3047 427.0146 921.5948
predict.lm(model, new, interval = "confidence", level = 0.99)
  fit     lwr      upr
1 674.3047 496.6497 851.9596
predict.lm(model, new, interval = "confidence", level = 0.95)
  fit     lwr      upr
1 674.3047 543.395  805.2143

```

And this is because we have additional uncertainty level because we are predicting under a new setting. Where the confidence intervals are reflecting an average across all settings for that specific value. We also see that when we compare the 99% versus 95 intervals, we can see that the higher the confidence, the wider the confidence intervals.

d. A 99% **confidence** interval at an advertisement expenditure of \$30,000 is (496.649, 851.959). A 95% interval is (543.395, 805.214).

An important thing to remember is the specification of the data that we need to predict. And it's simple to do it for a simple integration model. We see that this becomes more complicated if we have multiple predictors.

Again, we need to be careful about whether we want a confidence interval, or a prediction interval. **Just to wrap up the comparison, the confidence intervals are narrower than the prediction intervals because the prediction intervals have additional variance from the variation of a new measurement.** From the variation of numerous measurements, we also interpret those intervals differently. One, the prediction intervals are for one specific setting, whereas confidence intervals are average across settings that have the same value for the predicting variable.

## Knowledge Check

The estimated versus predicted regression line for a given  $x^*$ :

- A. Have the same variance
- B. Have the same expectation
- C. Have the same variance and expectation
- D. None of the above

The variability in the prediction comes from:

- A. The variability due to a new measurement.
- B. The variability due to estimation.
- C. The variability due to a new measurement *and* due to estimation.
- D. None of the above.<sup>2</sup>

---

<sup>2</sup> 1: B

2: C

### 3. Diagnostics

The topic of this lesson is simple linear regression, with a focus on the model assumption and diagnostics. Particularly, we'll overview the assumptions of the simple linear regression and graphical approaches to assess those assumptions.

Let's go back to the model for the simple linear regression. The data consists of bivariate data of response variable  $y$  and a predicting variable  $x$ . The relationship between those two variables is a linear relationship plus an error term. The assumptions in the simple linear regression are:

- Linearity (or the mean zero assumption): the expectation of the error terms is equal to 0.
- Constant variance assumption: the variance of the error terms is equal to sigma squared, is the same across all error terms.
- Independence assumption: error terms are independent random variables.
- Normality assumption: the error terms are normally distributed. This assumption is needed for statistical inference.

In the next few slides, I'll show you and I'll discuss how to assess those assumptions.

The approach for diagnosing this assumption is to evaluate the **residuals**. We're not going to evaluate the assumptions in the error terms because we do not know  $\beta_0$  and  $\beta_1$ . Instead, we're going to evaluate the assumptions on the residuals, which are the differences between the observed responses and the fitted responses.

## Residual Analysis

$$\text{Residual Values: } e_i = \hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

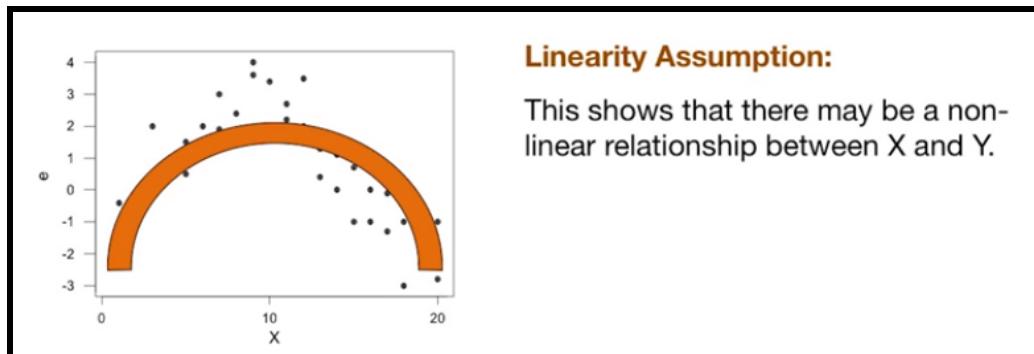
Graphical display: **Plot of the residuals  $e_i$**

If the scatter of  $e_i$  is **not random around zero line**, it could be that

- ✓ The relationship between  $X$  and  $Y$  is not linear
- ✓ Variances of error terms are not equal
- ✓ Response data are not independent

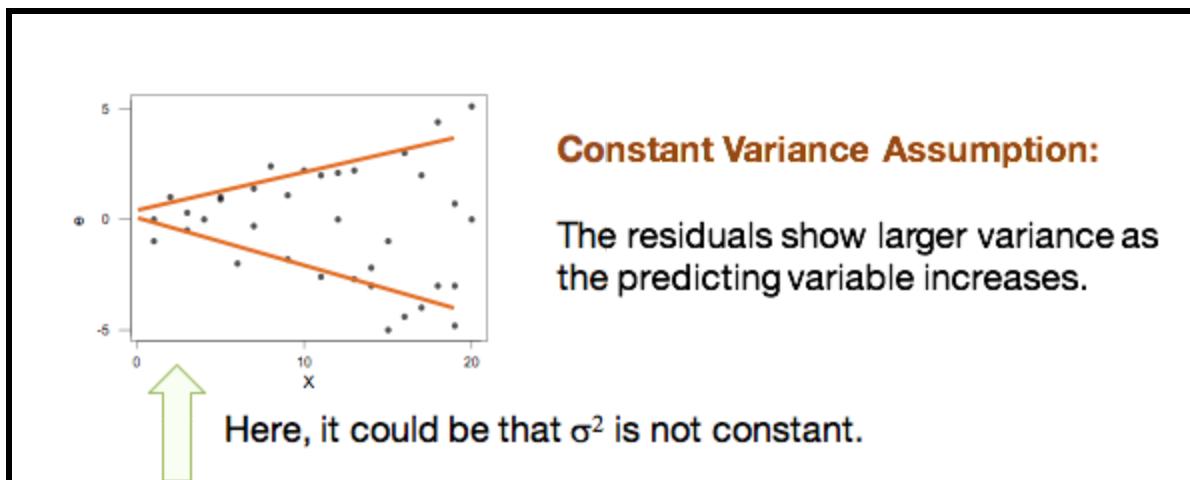
We will plot residuals against the fitted values and against the predictive values. If the scatter plot of the residuals is not random around zero line, the relationship between x and y may not be linear, or the variances of the error terms are not equal, and the response data or the error terms are not independent.

Let's look at the few examples of departures from the assumptions when using this approach. This is an example with residuals plotted against the predicting values:



As you see, there is a curvature between the relationship between residuals and x, showing their relationship is not linear. So it's a departure from the linearity assumption, and this means that the relationship between X and Y is non linear.

A second example, when you see this plot (below), it's a megaphone effect in the sense that the residuals increase with increasing values of the predicting variables, which means that the constant variance assumption does not hold.

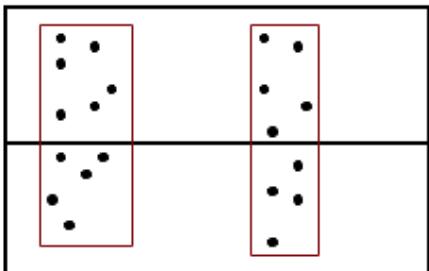


This is the third example (below) where we see a departure from the assumption.

You can see here that the residuals now are clustered in two separated clusters which means that the residuals are correlated.

### Independence Assumption:

There are clusters of residuals: the independence assumption does not hold.

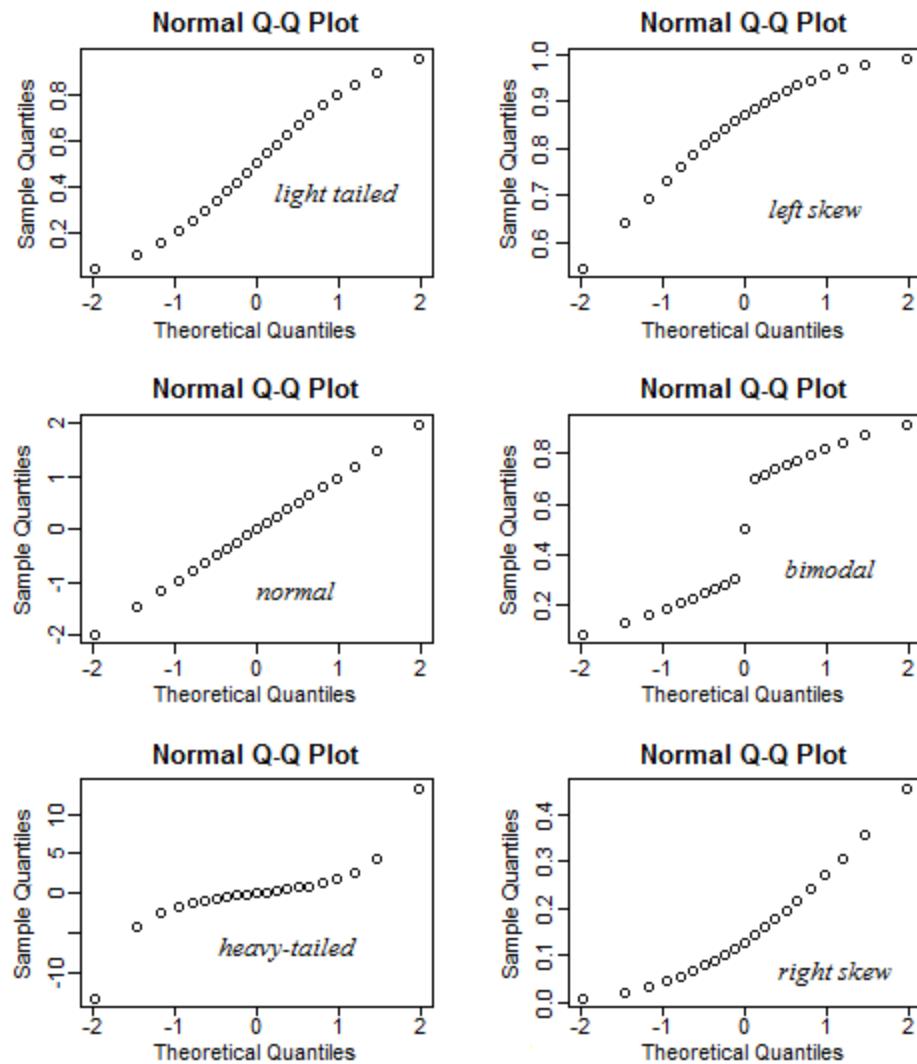


Keep in mind that **residual analysis does not check for the independence assumption**. Remember, the assumption is independence, not uncorrelated errors. But all we can assess with residual analysis is uncorrelated errors. Independence is more complicated to evaluate. If the data are from a randomized trial, the independence is established. But most data you're going to apply regression on are from observational studies and thus independence does not hold. In those cases, residual analysis is going to assess uncorrelated errors, not independent errors.

### Checking For Normality

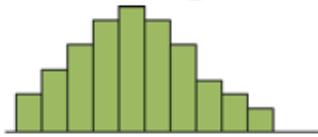
For checking normality, we can use the **quantile plot, or normal probability plot**, on which data are plotted against a theoretical normal distribution in such a way that the points should form a straight line. The x-axis of the normal probability plot is formed by the normal or statistic medians, and the y-axis is the order residual values. Departures from the straight line indicate departures from normality. The intuition behind this plot is that it compares the quartile of the residuals against quantiles of the normal distribution. If the residuals are normal then the quantiles of the residuals will line up with the normal quantiles, and as we should expect that they follow a straight line. Departure from a straight line could be in the form of a tail at the end, which is an indication of either a skewed distribution, or heavy-tail distribution. Do not attempt your own implementation of this plot—use the statistical software to do it for you. (We

will see many examples of the q-q norm plot in this class; do not worry if you do not quite get the concept right now.)

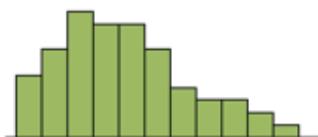


Another approach to check for the normality is using the histogram. Histograms are often used to evaluate a shape of a distribution. In this case, we would plot a histogram of the residuals and will identify departures from normality if we see skewness in the shape of the distribution, or by modality, when we have two or more modes in the distribution, gaps in the data, and so on. I suggest using both the normal probability plot and the histogram approaches to evaluate normality.

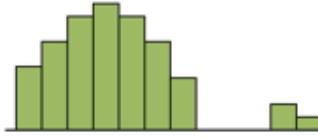
## Checking the Assumption of Normality



A complementary approach to check for the normality assumption is by plotting the **histogram** of the residuals



### Normality Assumption:



The residuals should have an approximately symmetric distribution, unimodal and with no gaps in the data.

If some of the assumptions do not hold, then we interpret that the model fit is inadequate, but it does not mean that the regression is not useful. For example, if the linearity does not hold, then we could transform Y or X to improve the linear assumption. This is generally a trial and error exercise, although sometimes you may just need to fix a curvature in the relationship, which could be done through using a power transformation or the classic log transformation.

## Variable Transformation

- If the model fit is inadequate, it does not mean that a regression is not useful.
- One problem might be that the relationship between **X** and **Y** is *not exactly linear*.
- To model the nonlinear relationship, we can transform **X** by some nonlinear function such as:

$$f(x) = x^a \text{ or } f(x) = \log(x)$$

**What if the normality assumption does not hold?** Often we use a transformation that normalizes the response variable.

That transformation is a power transformation of  $y$ . If  $\lambda$ , for example, the power is equal to 1, we do not transform. If  $\lambda$  is equal to 0, we actually use the normal logarithmic transformation. If  $\lambda$  is equal to -1 use the inverse of  $y$ , this is called the Box-Cox Transformation.

## Normality Transformations

**Problem:** Normality assumption does not hold.

**Solution:** Transform the response variable from  $y$  to  $y^*$  via

$$y^* = y^\lambda$$

where the value of  $\lambda$  depends on how  $\text{Var}(Y)$  changes as  $x$  changes.

$$\sigma_y(x) \propto \text{const} \quad \lambda = 1 \quad (\text{don't transform})$$

$$\sigma_y(x) \propto \sqrt{\mu_x} \quad \lambda = 1/2$$

$$\sigma_y(x) \propto \mu_x \quad \lambda = 0 \quad y^* = \ln(y)$$

$$\sigma_y(x) \propto 1/\mu_x \quad \lambda = -1$$

**This is called Box-Cox Transformation: The parameter  $\lambda$  can be determined using R statistical software.**

## 4. Outliers and Model Evaluation

An important aspect in regression is the presence of **outliers**, which are data points far from the majority of the data in both x and y or just x. Data points that are far from the mean of the x's are called **leverage points**. A data point that is far from the mean of both the x's and the y's are called **influential points**, because they're influencing the fit of the regression. *They can change the value of the estimated parameter, the statistical significance, the magnitude of the estimated parameters, or even the sign.*

It is tempting to just discard such points. But sometimes the outliers belong in the data. The elephant may be an outlier in terms of its size, but it's a real mammal nonetheless. Excluding an elephant from an analysis would skew or bias your conclusions.

Other times, there are good reasons for excluding a subset of points when there are errors in a data entry or in the experiment. *When outliers belong in the data, you will have to perform the statistical analysis with and without the outliers and inform the reader about the differences.*

To check outliers, a very simple approach is to use the standardized residuals, and then compare the standardized residuals to the minus 2 and 2 band or even tighter, the minus 1 and 1 band.

### Checking for Outliers

If we look at the **standardized residuals**

$$r_i^* = \frac{y_i - \hat{y}_i}{\sqrt{MSE}}$$

- Standardized residuals bigger than one are large; bigger than two extremely large.
- Most statistics packages will calculate these automatically.

Statistical packages usually compute the standardized residuals and or point to outliers. (We'll learn about other ways to evaluate outliers in different lectures, and different lessons in this class.)

One approach is to evaluate the predictive power of the model. Once we've established the goodness of fit of the model, **we want to see also whether the model is useful to predict**. One approach to quantify the predictive power is using the coefficient of variation or coefficient of determination, a statistic that efficiently summarizes how well the x can be used to predict the response variable. This is the so-called R-squared, which is 1 minus the ratio between the sum of squared errors and sum of square total.

## Coefficient of Variation/Determination

A statistic that efficiently summarizes how well the X's can be used to predict Y is the R-square:

$$R^2 = 1 - SSE / SST$$

which is interpreted as

**R<sup>2</sup> = Proportion of total variability in Y that can be explained by the regression (that uses X)**

$$SSE = \sum_{i=1}^n r_i^2$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

The interpretation of the R-square is the proportion of total variability in the response variable Y that can be explained by the linear regression that uses X.

Another approach to establish the linear relationship between two variables, for example the predicting variable and the response, is through computation of the **correlation coefficient**.

# Correlation Coefficient

A statistic that efficiently summarizes how well the **X's** are linearly related to **Y** is the correlation coefficients:

$$\rho = \text{cor}(X, Y) = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}} \sqrt{S_{YY}}} = \hat{\beta}_1 \sqrt{\frac{S_{XX}}{S_{YY}}}$$

Correlation coefficient and coefficient of variation:

$$\rho^2 = R^2$$

One could use the coefficient correlation to also evaluate various transformations of X and Y to improve the linearity assumption in the simple linear regression. All that is relevant in the context of explanatory power of the regression is that the relationship between the correlation coefficient and R-square. The relationship is that the square of the correlation coefficients is indeed the R2.

## 5. Diagnostics and Model Evaluation Examples

In this example, I will show you how to evaluate the assumptions of the simple linear regression, and how to quantify using R square, the predictive power of the model. We'll return now to the model where we're interested in evaluating the relationship between sales and advertising expenditure to address the question: do the assumptions of the linear regression model hold? And what is the explanatory power of the model?

First, we will review the assumptions of the linear regression, and then evaluate those assumptions using graphical displays and also identify outliers using the residual plots. And then we're going to compute the variability, the R squared, to evaluate the variability in the sales explained by the advertising expenditure.

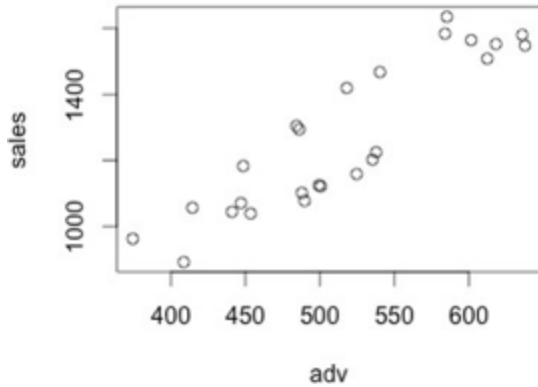
- a. What are the assumptions of linear regression?
- b. Do the assumptions hold? Provide the graphical displays needed to support the diagnostics. Interpret.
- c. Do you identify any outliers?
- d. How much variability in sales is explained by the advertisement expenditure?

To review, (a) the assumptions are **linearity, constant variance, independence** and **normality**.

(b.) Do the assumptions hold? Provide graphical displays...

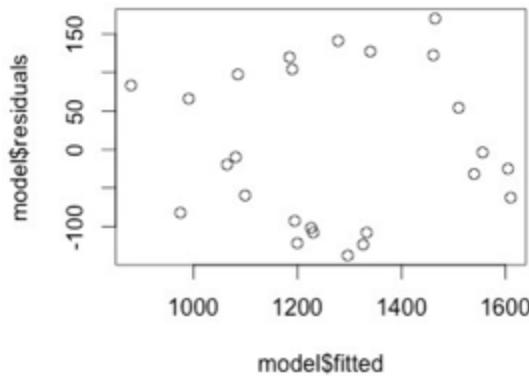
One way to evaluate the linearity is using the scatterplot of the predicting variable versus sales. This is the example of how to plot the scatter plot of the predictive variable (advertising) versus the response variable (sales) in R:

```
plot(adv, sales)
```



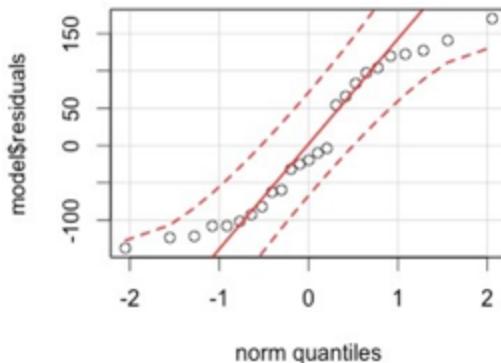
To evaluate the constant variance and independence assumption, we can use a scatter plot of the fitted values against the residuals, which is a second plot:

```
plot(model$fitted, model$residuals)
```



And to evaluate the normality we can use the normal probability plot:

```
library(car); qqPlot(model$residuals)
```



We can see from those three graphical displays that the linearity assumption between advertising expenditure and sales holds. Also, the residuals are scattered around the 0 line, which indicates that we do have constant variance. And independence were

uncorrelated errors. We do see some departure from normality especially in the tail, which could be an indication that the distribution of the residuals is heavy tail?. *But overall, based on those plots, the assumptions appeared to hold.*

(c.) Do you identify any outliers? From those plots, we can also identify outliers and, in fact, we do not see anything outside of the range of the residuals. Which is an indication that there do not appear to be outliers in the data.

(d.) How much variability in sales is explained by the advertisement expenditure? To quantify R square, we can use the summary of the model and extract the R square:

**summary(model)\$r.squared**

[1]0.8105919

The value of the R square for this example is 0.81, which means that 81% of the variability in the sales is explained by advertising expenditure alone. This is a very large R squared; we will rarely see such large R squared in real practice.

### Question 1

1/1 point (ungraded)

Which one is correct?

- Independence assumption can be assessed using the residuals vs fitted values.
- Independence assumption can be assessed using the normal probability plot.
- Residual analysis can only be used to assess uncorrelated errors. ✓
- None of the above

### Question 2

1/1 point (ungraded)

We detect departure from the assumption of constant variance

- When the residuals vs fitted values are larger in the ends but smaller in the middle. ✓
- When the residuals vs fitted are scattered randomly around the zero line.
- When the histogram does not have a symmetric shape.
- All of the above.

### Question 3

1/1 point (ungraded)

Which one is correct?

- If a departure from normality is detected, we transform the predicting variable to improve upon the normality assumption.
- If a departure from the independence assumption is detected, we transform the response variable to improve upon this assumption.
- The Box-Cox transformation is commonly used to improve upon the linearity assumption.
- None of the above ✓

### Question 4

1/1 point (ungraded)

In evaluating a simple linear model

- There is a direct relationship between coefficient of variation and the correlation between the predicting and response variables.
- The coefficient of variation is interpreted as the percentage of variability in the response variable explained by the model.
- Residual analysis is used for goodness of fit assessment.
- All of the above. ✓

## 1.3 Data Examples

### 1. Testing the Theory of Purchasing Power Parity (Part 1)

In this lesson, I'll introduce one specific example to which we'll use to practice the concepts of simple linear regression. In this example, we will study the relationship between inflation rates and exchange rates to evaluate the economic theory of the Purchasing Power Parity. The principle of Purchasing Power Parity (PPP) states that, over long periods of time, exchange rate changes tend to offset the differences in inflation rate between two countries. In an efficient national economy, exchange rates would give each currency the same purchasing power in its own economy. Even if it does not hold exactly, the purchasing power parity model provides a benchmark to suggest the levels that exchange rates should achieve.

## Regression Variables

**Response Variable:** Average annual change in the exchange rate

$$\frac{\ln(\text{Exchange Rate for 2012}) - \ln(\text{Exchange Rate for 1975})}{\text{no. years}} \% = \text{Annualized Percentage Change}$$

**Predicting Variable:** Average of the *difference in annual inflation rates* for a country vs U.S.

$$\frac{1}{\text{no. years}} \sum_{y=1975}^{2012} (\text{Inflation}_y(\text{U.S.}) - \text{Inflation}_y(\text{Country}))$$

The average annual change in exchange rate is the response variable expressed as US dollar per unit of the countries currency. It is calculated as a difference in natural logarithms divided by the number of years and multiplied by 100, to create percentage change as shown on the slide (above). This is approximately equal to the proportional change in exchange rate over all 37 years, producing an annualized change in exchange rate.

The predicting variable is the estimated average annual rate of change of the differences in a wholesale price index values for the United States versus the country as shown on this slide (below). We'll analyze data for 41 countries including both developed and developing countries, covering the years 1975 to 2012. The data

columns include country and the inflation difference on the exchange rate change over a period of time.

We also have a column specifying whether a developed or a developing country. We'll explore the purchasing power theory using simple linear regression.

Country	Inflation.difference	Exchange.rate.change	Developed
Australia	-1.2351	-3.1870	1
Austria	1.5508	1.4781	1
Belgium	1.0371	0.0395	1
Canada	0.0461	-1.6416	1
Chile	-18.4126	-20.6329	0

This example was made available by Dr. Jeffrey Simonoff, of the New York University.

Let's begin with the first step in data analysis using R.

```
# ASCII Data files use read.table R command#
ppp = read.table("ppp.dat",sep="\t", header=T, row.names=NULL)
## Check to make sure you read the data in R correctly
ppp[1:2,]
  Country Inflation.difference Exchange.rate.change Developed
1 Australia           -1.2351          -3.1870            1
2 Austria             1.5508           1.4781            1
## How many countries?
dim(ppp)
[1] 40 4
## Brazil is an outlier and it was not included in the data set initially; I am adding it back as follows
Addp = data.frame("Brazil",-76,-73,0)
names(addp) = names(ppp)
## Save the data variables to be recognized by R as separate variables
ppp = data.frame(rbind(ppp,addp))
attach(ppp)
## Re-label the 'Developed' column to differentiate between Developed and Developing countries
Developed[Developed==1] = "Developed"
Developed[Developed==0] = "Developing"
```

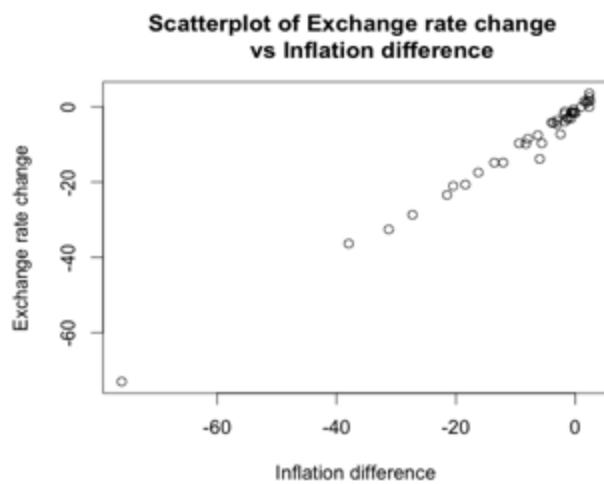
We'll first read the data with **read.table** command in R where the input as the data file which is called ppp.dat for this data asset. We also specify the separator of the data values in the file on this here is a tab delimiter , we specify the data files has a header and then the columns have names. And the real names have no names. You should check whether the data matrix coincides with the data you read from the file. In this case we check the first two rows of the data, and we can compare this with the data from the data file.

We can find how many columns and rows are in a data using the **dim** command in R. You see here that we have 40 rows, which means we have 40 countries. However initially I said that we have 41 countries. That is because we will now add another country, Brazil, which is an outlier not included in the dataset initially. And the data points for inflation difference, exchange rates, and whether it's developed or not are provided with the name of the country. And I'm adding now this data to the initial data on that matrix by using the **R bind** command, and I'm converting now the matrix into a data frame. And I attached this data in order for the columns in a data matrix to be recognized by R, as individual vectors. I will relabel also the column corresponding to developed using their initial denomination. 1 indicates "developed" 0 indicates "developing." And the reason I'm doing that is because when we do exploratory analysis, is better to refer to numerical values to the names of those categories, the denomination of those categories.

Next, we'll perform exploratory data analysis for this dataset.

One visual approach is by plotting the x and y. In this case, x is inflation difference, and y is the exchange rate change.

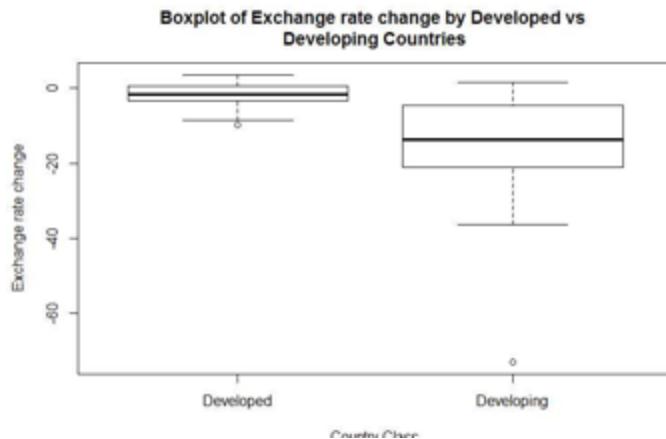
```
# Evaluate the Linear Relationship: Perform a scatter plot of the two variables
plot(Inflation.difference, Exchange.rate.change, main="Scatterplot of Exchange rate change vs Inflation difference", xlab="Inflation difference", ylab="Exchange rate change")
```



I recommend labeling both the title, the labels for x and y axis accordingly because it's much easier to interpret the graphics. What you see here is the plot of inflation difference against the exchange rate change, and I can see that there is a linear relationship between those two.

We also can evaluate how would exchange rate varies with the categorical (qualitative) x-variable Developed, and can use the box plot command in R to provide a side by side box plot for development of the exchange rate change.

```
# Evaluate differences between developed and developing countries
boxplot(Exchange.rate.change~as.factor(Developed), main="Boxplot of Exchange rate change by Developed vs Developing Countries",xlab="Country Class",ylab="Exchange rate change")
```



And from this box plot (above), we can see that there is a significant difference in exchange rate change amount between the two types of countries developed and developing.

Let's ignore for now the presence of the outlier and the fact that there are differences in response variable across developed and developing countries as we've learned from the previous slide.

We're now going to perform least squares regression with Exchange Rate Change as the response variable and Inflation Difference as the predicting variable. The R command is **LM** and for this command the response variable Exchange Rate Change is provided on the left, separated by a tilde from the predicting variable, which is Inflation Difference. The output of this function:

```
pppa = lm(Exchange.rate.change ~ Inflation.difference) ## regression model
summary(pppa)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.51930	0.29415	-5.165	7.43e-06
Inflation.difference	0.96185	0.01781	53.991	< 2e-16
---				

Residual standard error: 1.646 on 39 degrees of freedom

Multiple R-squared: 0.9868, Adjusted R-squared: 0.9865

F-statistic: 2915 on 1 and 39 DF, p-value: < 2.2e-16

Let's dissect the part of the output that provides information about the coefficients. The estimated  $\beta_0$  and order intercept is -1.5193 and the standard error is 0.2941. The estimated  $\beta_1$  slope is 0.961 and the standard error is 0.0178.

$$\hat{\beta}_0 = -1.5193, \text{ se}(\hat{\beta}_0) = 0.2941$$

$$\hat{\beta}_1 = 0.9618, \text{ se}(\hat{\beta}_1) = 0.0178$$

Test for statistical significance:

$$\beta_0 \text{-t-value} = -5.165, \text{ p-value} \approx 0$$

$$\beta_1 \text{-t-value} = 53.991, \text{ p-value} \approx 0$$

The intercept says that given that the analyzed difference in inflation between a country and the U.S. is zero, that is the country has the same inflation experience as does the U.S.. However, the estimated expected analyzed change in exchange rates between the two countries is not 0, it's -1.52%. That is, the currency becomes devalued relative to the US dollar.

The slope coefficient says that a 1% point change in analysed difference in inflation rate is associated with the estimated expected value of 0.962% of point change and analyse change in exchange rates. Based on the output the p value of both test statistical significance of the coefficient is approximately 0.

How do we interpret this? This means that both coefficients are statistically significant different from 0.

The other portion of the summary output that is of interest is the residual standard error and multiple R squared.

$\hat{\sigma} = 1.646$ ,  $n-2 = 39$   
 $R^2 = 98.7\%$  variability explained

This summary gives estimated standard error of the estimated sigma. The numbers of degrees of freedom which is 39, thus n is 41. Also our (multiple R squared) score is high meaning that 98.7% of the variability in exchange rate change is explained by the inflation difference.

Although this model would not be used to trade currency, the estimated standard error of 1.6 tells us, that this model could be used to predict annualized changes in exchange rates to within 3.2% points roughly 95% of the time.

Let's go back to the validity of the purchasing power parity theory which says that, in an efficient market the intercept is 0, and the slope is 1.

## Does the Theory Hold?

The principle of purchasing power parity (PPP) states:

$$\text{Average annual change in the exchange rate} = \text{Difference in average annual inflation rates} + \text{Random error}$$

The economic theory says that  $\beta_0 = 0$ ,  $\beta_1 = 1$

VS.

The estimates for these coefficients are:  $\hat{\beta}_0 = -1.519$ ,  $\hat{\beta}_1 = -0.961$

### Testing the theory:

$\beta_0 = 0$ : Based on the t-test of statistical significance we find that  $\beta_0$  is statistically different from zero.

$\beta_1 = 1$ : We need to perform a t-test with this as the null hypothesis

$$T\text{-value} = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = \frac{0.9618 - 1}{0.0178} = 2.1448 \text{ & p-value} = 2(1 - P(T_{39} < 2.1448)) = 0.038$$

However, we see that the estimates for these coefficients are not quite what this theory says. However looking only at estimated values it's not sufficient to make statistical statements about the theorem. We'll need to use the statistical inference such as hypothesis testing. Testing for  $\beta_0$  equal to zero means that testing for statistical significance. From previous slide we find that  $\beta_0$  is statistically different from zero, thus the theory does not hold with respect to the intercept. That is, the foreign currencies

appear to have appreciated less than would be predicted by the purchasing power theory, since the intercept is negative.

Testing whether  $\beta_1$  is equal to one is a slightly different test than the test for statistical significance since the known value is one and not zero. For this test, the test value is 2.14, and the p-value is 0.038. In this test, we compute the T-value by replacing 0 with 1. As we use for the statistical significance. And for this has a t-value of 2.148.

And we computed the p-value similarly as for statistical significance we take 2 times 1 minus the left tail of the t distribution with 39 degrees of freedom, at 2.14, and the p value, again is 0.038. This p-value is small but not very small. We would like to see a p-value that is smaller than 0.01. However, this p-value is smaller than 0.05, which means that we do not reject the null hypothesis. We reject the hypothesis at 0.05, but we do not reject the null hypothesis at 0.01. *Thus, we see violations of the purchasing power parity with respect to both the intercept and the slope.*

Let's see how we can perform the hypothesis testing procedure for  $\beta$  equals to 1, with R. We're going to use a function in the library car. (Don't forget to install the package before using the library.) You can use the command `install.packages()` to install a package, and the `library()` command to upload the library into R.

```
# Perform the hypothesis test for slope coefficient H0: slope=1
# use the library 'car' available in R (you need to install this library first then download it)
install.packages("car")
library(car)
```

We are using the **function linear hypothesis** (below). And the input in this command is the model called PPPA, and we're providing `c(0,1)` as the vector telling us on which coefficient we're interested to perform this linear hypothesis, and the right hand side tells us where we want to look. The right hand side which specifies how we want the alternative hypothesis.

```
linearHypothesis(pppa,c(0,1),rhs=1)
```

We can also the compute the t-value directly (below) by specifying the  $\beta_1$  hat minus one, divided by the standard error. We can compute the p-value as the formula I provided in the previous slide, which is **2 x (1- pt(tvalue,39)** which provides the **probability of a t-value distribution where we input the t-value and the degrees of freedom.**

```

## Alternatively, you can compute the t-value and p-value as follows:
tvalue = (0.9618-1)/0.01781
pvalue = 2*(1-pt(tvalue,39))

```

Based on this the t value is again provided below.

$$\text{P-value} = 2P(T_{n-2} > |t\text{-value}|)$$

where

$$t\text{-value} = \frac{\hat{\beta}_1 - 1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

This is what we compute in using R, and the p-value is as on this formula but we can use R again to do 2 times 1 minus PT because PT keys as the left tail.

Let's look a little closer at the PT function, which finds the probability for a t-distribution. We will see that we can use other similar functions for different kinds of distributions.

```

## Use the help menu to learn more about the functions used above:
help(pt)
help(linearHypothesis)

```

To learn more about this command we can use their help menu. And for this command, we need to specify the t-value, which is the quantile, and the number of degrees of freedom to get the probability of that quantile. Here are some examples:

```

# Any distribution in R has several functions available that start with p, q, r, d
# To compute the probability of a t-distribution with 39 degrees of freedom for the quantile of 2.145
pt(2.145,39)
# # To compute the probability of the normal distribution with mean 1 and variance 2 for the quantile of 2.145
pnorm(2.145,1,sqrt(2))
# To compute the quantile of a t-distribution with 39 degrees of freedom for the probability of 0.95
qt(0.95,39)
# # To compute the quantile of the normal distribution with mean 1 and variance 2 for the probability of 0.95
qnorm(0.95,1,sqrt(2))

```

For example, **function pt** with a value of 2.145 and 39 degrees of freedom will give us the probability of a t-distribution with 39 degrees of freedom for the quantile 2.145.

We can use a similar function for the normal distribution with the corresponding **function pnorm**. Again, this is going to give us the probability of a normal distribution. Here for this, we need to specify again the quantile which is 2.145. We also need to specify the mean and the variance. The other functions we could use is the quantile for this distribution, or the quantile of other distributions, like normal. We can have the other functions that are related to distributions, like the density, the sampling from a

distribution which is **function RT** or **function Rnorm** and so on. I recommend you would practice with those functions because they're very useful for this class.

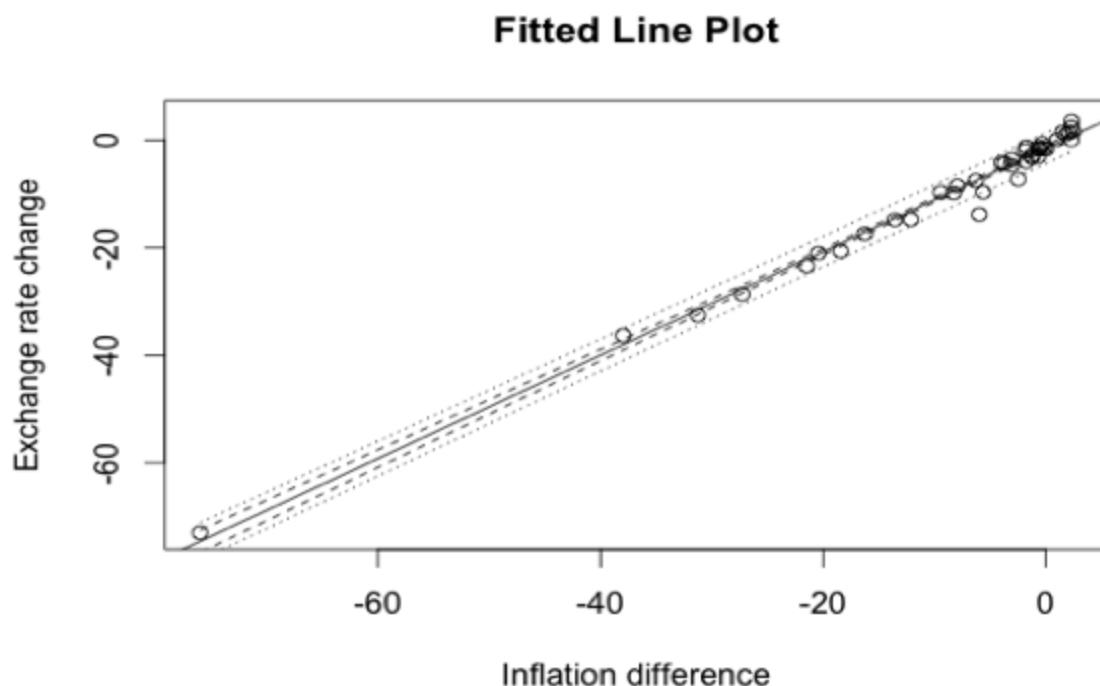
## 2. Testing the Theory of Purchasing Power Parity (Part 2)

In this lesson, we'll go back to example one, testing the theory of purchasing power parity. We'll focus on statistical inference, and we will also study the impact of one outlier in this data, particularly the outlier corresponding to Brazil.

To get confidence in prediction bands around the estimated regression line you can use the following R function:

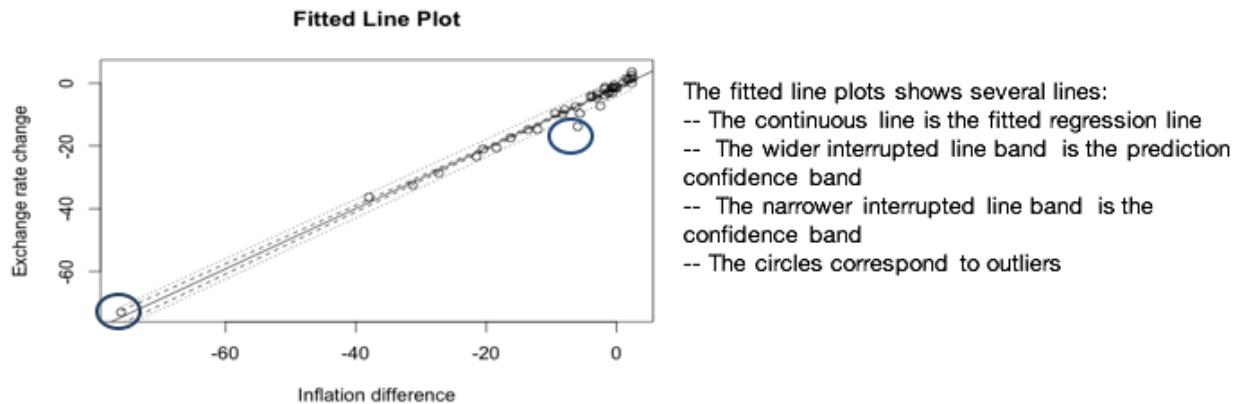
```
# Function for fitted line plot: See ppp-revised.R for this function
#regplot.confbands.fun = function(x, y, confidencelevel=.95, Clmean=T, PI=T, Clregline=F, legend=F){
  ##### Modified from a function written by Sandra McBride, Duke University
  ##### For a simple linear regression line, this function
  ##### will plot the line, CI for mean response, prediction intervals,
  ##### and (optionally) a simultaneous CI for the regression line.
...}
regplot.confbands.fun(Inflation.difference,Exchange.rate.change)
```

When we want to write more general R code that can be used in many other settings, we'll do so in the form of R functions. Like in this case, we give it a name to refer to it later and we provide the input parameters, for example X and Y, which are the predicting and the response variables. Along with the confidence level and other input, then we can use this R function to produce the simple regression scatter plot. In this figure, which illustrates the use of confidence intervals and prediction intervals:



Consider again the estimated regression line with  $x$  and  $y$  from our data example to produce this plot. First, the pointwise confidence interval represented by the inner pair of the lines is much narrower than the pointwise prediction interval represented by the outer pair of lines. While the confidence interval takes into account only the variability and estimation, the prediction band takes into account both the variability due to the estimation and the variability due to the uncertainty in the data.

In this plot, we can also identify two outliers:



One is on the left, which corresponds to Brazil. The other one is a point outside both the prediction interval and the confidence interval. This point corresponds to Mexico. The prediction interval again is wider than the confidence interval. This type of analysis allows us to identify outliers.

This analysis, this approach, doesn't allow us to compute confidence and prediction intervals for new observations so we need to use the **predict function** in R.

For this function we need to specify the value or values we want to predict and create separately a data frame. We also need to specify whether we want to estimate a confidence versus a prediction interval, so we need to specify that in the predict function.

```

# Confidence and prediction intervals for new observation
# Create new data point
newppp = data.frame(inflation.difference = c(-0.68))
# Specify whether a confidence or prediction interval
predict(pppa,newppp,interval=c("confidence"))
  fit      lwr      upr
1 -2.173351 -2.756818 -1.589884
predict(pppa,newppp,interval=c("prediction"))
  fit      lwr      upr
1 -2.173351 -5.554071  1.207369
## Why are the intervals different?

```

In this example, we've estimated the prediction and confidence interval for a point corresponding to -0.68 which is roughly the value for Norway. For that we created a new data frame and then we input this into the predict function. Estimating both intervals we see that the two intervals are different, they're both the same value, but the prediction interval is wider than the confidence interval, since uncertainty in the prediction interval is higher than for the confidence interval.

How do we interpret the intervals?

**Interpretation of the two intervals:**

- The 95% confidence limits of the average exchange rate change for all countries inflation difference equal to -0.68 are (-2.757,-1.590);
- The 95% confidence limits for the exchange rate change for one country with inflation difference equal to -0.68 are (-5.554,1.207).

The confidence interval provides our guess for what the average exchange rate change would be for all countries with inflation difference equal to -0.68 and this confidence interval is between -2.75, and -1.59. The prediction interval provides our guess for what the exchange rate change will be for one country with inflation difference equal to -0.68. And this interval is very large and in fact it includes zero, so you can see the difference between those confidence intervals. The first one includes only negative values whereas the first one includes both negative and positive values and is much wider.

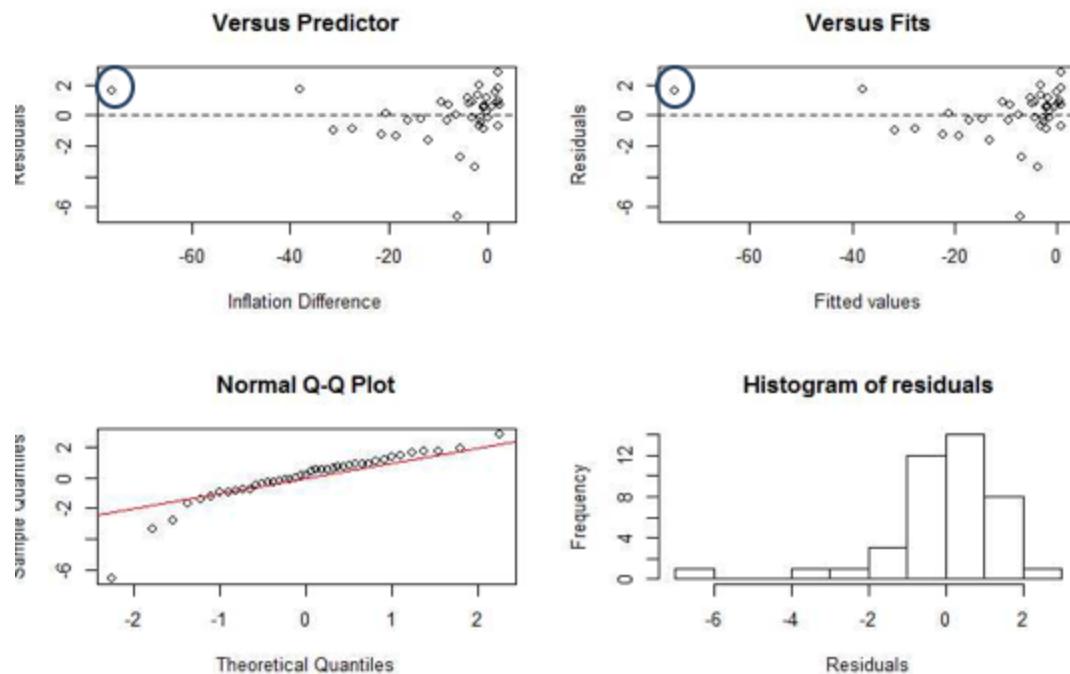
All the inference provided so far on the regression coefficients relies on the fact that the model assumptions hold. What are the assumptions of the simple linear regression, again? Linearity, constant variance, independence, and normality. We'll have to check those assumptions in order to rely on the inference result I provided so far.

```

par(mfrow=c(2,2))
plot(inflation.difference, residuals(pppa), xlab="Inflation Difference", ylab="Residuals", main="Versus Predictor")
abline(h=0, lty=2)
plot(fitted(pppa), residuals(pppa), xlab="Fitted values", ylab="Residuals", main="Versus Fits")
abline(h=0, lty=2)
qqnorm(residuals(pppa))
abline(0,1, lty=1, col="red")
hist(residuals(pppa), main="Histogram of residuals", xlab="Residuals")

```

Here are the four plots that I often use to evaluate assumptions to do this residual analysis:



The first one is the scatter plot of the predicting variable, in this case inflation difference versus the residuals.

The second one provides the scatter-plot of the fitted values versus the residuals.

The third one is the normal probability plot along with the histogram (fourth plot). As I mentioned in previous lessons, it's good practice to correctly label the X axis and Y axis. For example, for the first plot, the X axis corresponds to inflation difference and the Y axis corresponds to the residuals.

Let's take a closer look at those four plots. The first one, again, is the residuals against the inflation difference, which is used to evaluate whether we have a linearity assumption. If there's no pattern in this plot, we conclude the linearity assumption holds. So in this example, there's no specific pattern, we do identify one outlier, again

which corresponds to Brazil. But we do have the linearity assumption, and the linearity assumption holds.

The second plot can be used to evaluate the constant variance and were uncorrelated errors. In fact, the first plot can be used for that as well. What we can learn from this plot is that there is a difference in variability of the residuals with increasing fitted values. And which means that the constant variance does not hold. We also do not see a grouping of residuals, which means that the assumption of uncorrelated error is possible to hold.

The two graphs on the bottom can be used to evaluate the normality. You can see that we do have a slight tail in the normal quantile plot that's also reflected also in the histogram of the residuals.

So let's discuss the practical outlier that we've noted in the two graphs. Observations for which the predicting variable is away from that range again is called the leverage point and the isolated point in residual plot is Brazil.

And why is Brazil a leverage point? Brazil had a period of hyperinflation from 1980 to 1994, a time period during which prices went up by a factor of roughly 1 trillion. This hyperinflation was caused by an expansion of the money supply. The government financed projects not through taxes or borrowing, but simply by printing more money, a crisis triggered by the worldwide energy crises of the 1970s and political instabilities of the Brazilian military dictatorship.

What should we do about this case? The unusual point has the potential to change the result of the regression, so we can simply ignore it. We can remove it from the data analyze the data without it being sure to inform the reader about what we're doing. That is, we can present results without Brazil while making clear that the implications of the model do not apply to Brazil or probably to other countries with a similar unsettled economic situation.

**Leverage Points:** Observations for which the predicting variable is away from the range.

The isolated point in residual plots is Brazil. Why is Brazil a leverage point?

-- Brazil had a period of hyperinflation from 1980 to 1994, a time period during which prices went up by a factor of roughly 1 trillion.

Why do we care about leverage points?

-- It can have a strong effect on the fitted regression, drawing the line away from the bulk of the points. It also can affect measures of fit like R-squared and t-statistics.

Here we perform a linear regression for the data without Brazil:

```
##### Repeat Analysis: Omit Brazil #####
## remove the data row corresponding to Brazil
newppp = ppp[ppp$Country!="Brazil"]
attach(newppp)
## Fit Linear Regression
pppn = lm(Exchange.rate.change ~ Inflation.difference)
summary(pppn)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.37222	0.30517	-4.497	6.31e-05
Inflation.difference	0.99152	0.02626	37.757	< 2e-16
Residual standard error:	1.62	on 38 degrees of freedom		
Multiple R-squared:	0.974,	Adjusted R-squared:	0.9734	

We first remove the data point from the data and then reattach a new data in R such that R will recognize the columns from this new data set. We next run the linear model using the LM command, and a portion of the summary of the estimated coefficient is provided here. We note that the estimated intercept has changed from about -1.59 to -1.37 although that is still statistically different from 0 since the p-value is approximately 0.

On the other hand, testing the null hypothesis that the intercept is equal to 1, the p value is now 0.748. Indicating that we do not reject the null hypothesis and it is possible for the intercept to be equal to one (below).

$\hat{\beta}_0 = -1.372$ ,  $se(\hat{\beta}_0) = 0.305$   
Statistical significance for  $\beta_0$ :  
 $t\text{-value} = -4.497$ ,  $p\text{-value} \approx 0$

What about the slope? (below)

```

## Test whether the slope is equal to 1 (PPP theory)
tvalue = (0.9915-1)/ 0.02626
pvalue = 2*(1-pt(tvalue,39))

```

$\hat{\beta}_1 = 0.9915, \text{se}(\hat{\beta}_1) = 0.02626$   
Test the null hypothesis  $\beta_1 = 1$ :  
p-value = 0.748

**So we conclude that we're seeing violations of the purchasing power parity theory, with respect to intercept only, when we omit Brazil from the dataset.**

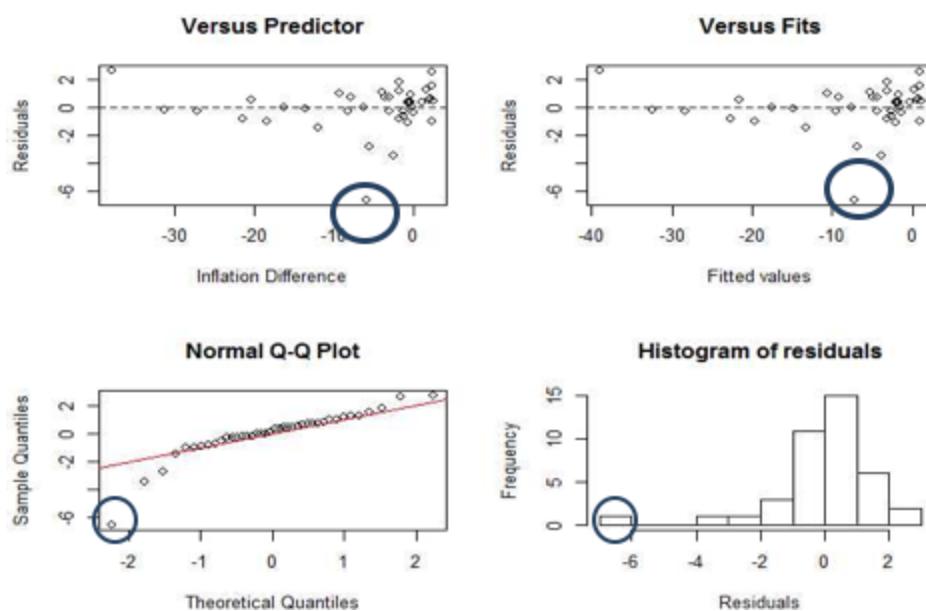
#### RESIDUAL ANALYSIS: MODEL WITHOUT BRAZIL

We're going to redo again the residual analysis without Brazil and we're going to use the same four graphical displays in order to evaluate the four assumptions.

```

par(mfrow=c(2,2))
plot(inflation.difference, residuals(pppn),xlab="Inflation Difference",ylab="Residuals",main="Versus Predictor")
abline(h=0,ity=2)
plot(fitted(pppn),residuals(pppn),xlab="Fitted values",ylab="Residuals",main="Versus Fits")
abline(h=0,ity=2)
qqnorm(residuals(pppn))
abline(0,1,ity=1,col="red")
hist(residuals(pppn),main="Histogram of residuals",xlab="Residuals")

```



And now you can see that the outlier corresponding to Brazil is not there because we omitted Brazil, however, we can still see that we still do not have constant variance. We see now a different outlier corresponding to Indonesia. Shall we remove the outlier and run the model again and test the purchasing power parity theory? We can do that, in fact, we'll find that there's still going to be violations with respect to the intercept.

*So in conclusion we have linearity, constant variance, uncorrelated errors, normality, and those are the assumptions and those can be checked with the four plots.* In this case, the only assumption that we need to be concerned with is the constant variance. The outliers we identify, we identify another outlier Indonesia and you can round the regression analysis by omitting Indonesia. And you can evaluate to see how the model changes as you omit this as well.

**Assumptions:**

**Linearity:** No pattern in the residuals with respect to the predicting variable.

**Constant Variance:** The variance is higher for higher fitted values. Does not hold.

**Uncorrelated Errors:** No grouping of the residuals

**Normality:** Except for the presence of an outlier, it is reasonably symmetric.

**Outliers:** Observations for which the residual value is away from the range.

The isolated point in the residual plots is Indonesia. Would omitting Indonesia change anything? The strength of the relationship would increase, but so the rejection of PPP.

In fact, you can also perform the analysis separately for developed and developing countries. You can see that Brazil and Indonesia are developing countries, and the question is, will the purchasing power parity theory hold differently for the two countries?

So those are the findings by performing multiple analyses.

**Findings:**

- Support is decidedly mixed;
- Developed countries:
  - Changes in inflation difference do seem to be balanced by exchange rate changes;
  - One outlier: Greece;
- Developing countries:
  - The case for PPP is considerably weaker;
  - Brazil and Indonesia
- PPP is not robust to unusual economic or political conditions

The support of the theory is decidedly mixed. By separating developed countries from developing countries we found that changes in inflation difference do seem to be balanced by exchange rate changes with one outlier Greece. For developing countries the case for this theory is much weaker and we can see Brazil and Indonesia are two of the outliers. *So we conclude that the purchasing power parity is not robust to unusual economic or political conditions.*

### 3. 2000 Elections in Florida

The topic of the lesson is Simple Linear Regression using R. And we'll focus on one particular example on the Presidential Elections of 2000 in Florida. In this example, we'll particularly focus on identifying one outlier, among the vote counts in one county in Florida and we see with our analysis could have suggested an overturn of the election results in 2000.

In this example in the presidential elections in 2000, during the election night, the two presidents, George W Bush and Al Gore, the results, the electoral votes were very tied. George W Bush had 246 electoral votes and Al Gore had 255 with three states too close to call that night. The state that really mattered was Florida, so, weeks after the election night, there was an intense recount of the vote in Florida. In this example, we're going to analyze the vote counts for Bush and for the independent candidate Buchanan, because we would expect that the votes for those two candidates would be similar, since Buchanan was an independent candidate that was more conservative. And particularly we're going to look at one county, the Palm Beach County, where the number of votes for Buchanan was very large.

A first step in a data analysis using R is to read the data file in R, and in this case we're going to use a `read.table` command in R and for using this command we need to specify the name of the file along with information whether the columns in the data file have a header. We need to check the data content by looking at the first few columns of the data content and here, we're looking at the first four rows. In this example, you can see that the data consists of many more factors for the counties in Florida, but we practically only looked at the vote counts for Bush and Buchanan.

```

## Read data with read.table R command which is used for reading ASCII files
elections = read.table("elections.txt",header=TRUE)

## Check the data content elections[1:4,]
  co lat lon npop whit blac hisp o65 hsed coll inco bush gore brow
1 1 29.7 82.4 198326 74.4 21.8 4.7 9.4 82.7 34.6 19412 34124 47365 658
2 2 30.3 82.3 20761 82.4 16.8 1.5 7.7 64.1 5.7 14859 5610 2392 17
3 3 30.2 85.6 146223 84.2 12.4 2.4 11.9 74.7 15.7 17838 38637 18850 171
4 4 29.9 82.2 24646 76.1 22.9 2.6 11.8 65.0 8.1 13681 5414 3075 28
nade harr hage buch mcre phil moor
1 3226 6 42 263 4 20 21
2 53 0 3 73 0 3 3
3 828 5 18 248 3 18 27
4 84 0 2 65 0 2 3

```

*The data file includes many other variables characterizing the counties.  
We will focus only on the number of votes in this analysis.*

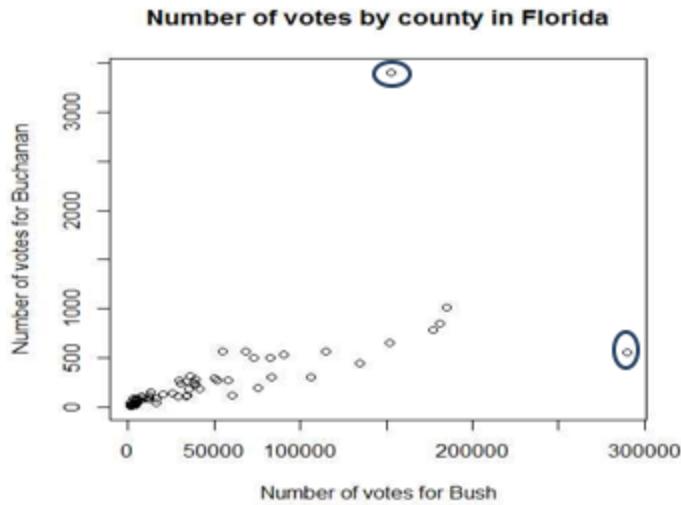
So, our first step in a data analysis is to explore for the factors that we're interested in at analysis. In this case, we're interested in a vote counts Buchanan at for Bush to extract those factors from the data matrix that we run in R. We can use the command where we specify the name of the file in this case elections. The name of the column separated by a dollar sign.

```

#### Extract number of votes for each candidates
buch = elections$buch
bush = elections$bush
gore = elections$gore
#### Visualize the relationship between number of votes between Buchanan and Bush
plot(bush,buch,xlab="Number of votes for Bush",ylab="Number of votes for Buchanan",
main="Number of votes by county in Florida")
cor(buch,bush)

```

Next, we can use again the plot R command that provides us the scatter plot of two variables in this case the votes counts for Bush and the votes count for Buchanan. This is how the scatter plot of those two variables looks like.



We can identify (above) two specific outliers, but we also can note that the relationship between the number of votes for Bush and number of votes for Buchanan is a curvature so it's not a linear relationship.

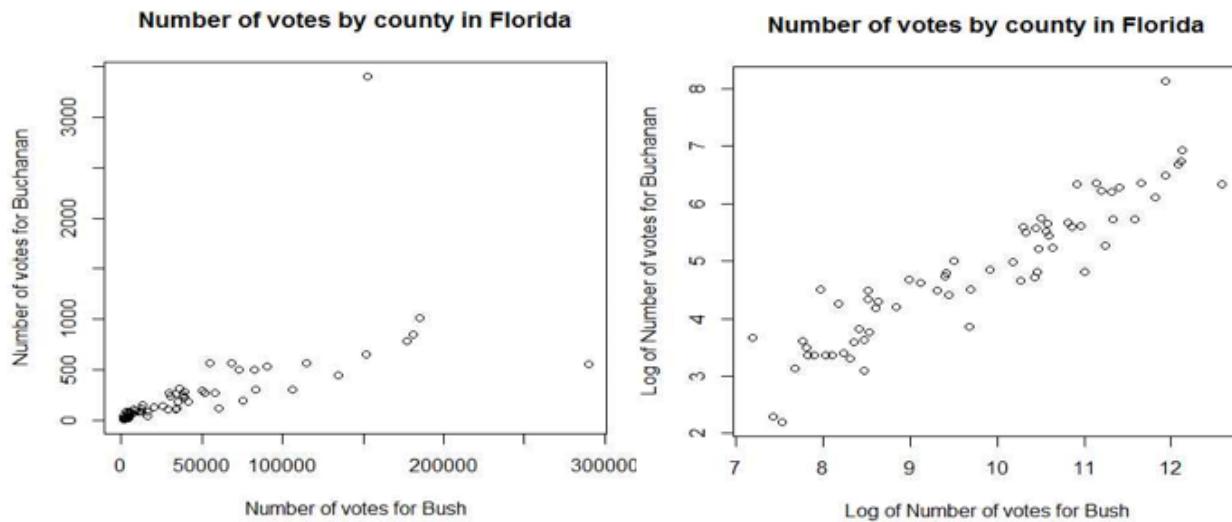
**Linearity Assumption:**

- The scatterplot shows a strong positive relationship between the number of votes for the two candidates except for two outliers, one corresponding to the Palm Beach county. The correlation is high also (0.625).
- Curvature in the relationship – consider transformations

So in this context we'll have to perform some transformations on X or and Y in order to fix this nonlinearity between the two factors and this is what I'm showing you on the slide.

```
### Transform both variables using the log-transformation
plot(log(bush),log(buch),xlab="Log of Number of votes for Bush",ylab="Log of Number of votes for Buchanan",
main="Number of votes by county in Florida")
cor(log(bush),log(buch))
```

We compare the scatterplots of the number of votes for Bush versus the number of votes for Buchanan (below left), versus the log of the number of votes for Bush and log of number of votes for Buchanan (below right).



So now we can see that the linearity assumption has improved significantly. We can also see that one of the outliers doesn't seem to be a outlier anymore. However, Palm Beach which is a large value in the votes for Buchanan is still present.

Now, we also see that the correlation has increased from 0.625 to 0.922 (below).

And I mentioned in a different lesson that an approach to identify a transformation that will improve the linearity between two factors is using the correlation. So, here you can play with multiple transformations of the two factors and see which one most improves or increases the correlation coefficient.

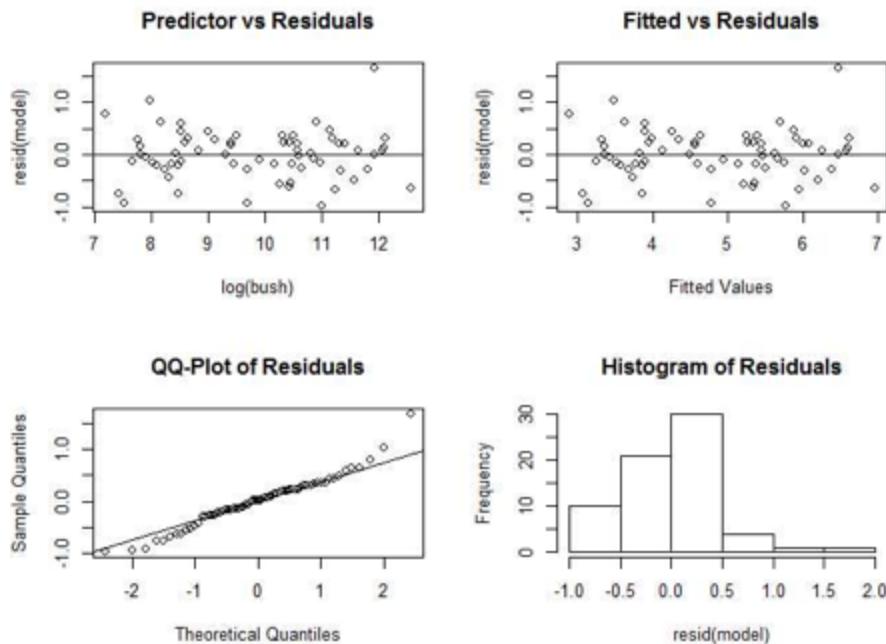
***Linearity Assumption:***

- The linear relationship has improved with the transformations
- The correlation has increased from 0.625 to 0.922
- We will perform the regression analysis using the transformed data

Next, we'll perform the regression analysis using the *transformed data using log* of the votes for both Bush and Buchanan. It is good practice in any regression analysis is to assess goodness of fit through evaluating the assumptions of the linear regression model. And the four assumptions that again we are assessing is linearity, constant variance, independence or uncorrelated error, and normality. And those plots can be used to assess such assumptions. The first one is the scatter plot of the predicting variable versus the residuals. The second one is the scatter plot of the fitted values

versus the residuals and that the last two plots are used to assess normality using the quantile, quantile normal plot and the histogram.

```
## Perform Residual Analysis
par(mfrow=c(2,2))
plot(log(bush),resid(model), main="Predictor vs Residuals")
abline(0,0)
plot(fitted(model),resid(model),main="Fitted vs Residuals",
      xlab="Fitted Values")
abline(0,0)
qqnorm(resid(model),main="QQ-Plot of Residuals")
qqline(resid(model))
hist(resid(model),main="Histogram of Residuals")
```



And this is the output of this R code, based on the residual analysis plots, we learned that the assumption of constant variance holds because we don't see a change in the variability of residuals. We also can assess the linear assumption using the first plot, and because there's no pattern, we conclude that the linearity assumption holds. There is also not a clustering among the residuals, and that indicates that the assumption of uncorrelated errors holds as well. For normality there, we can use the bottom plots to evaluate normality and while the QQ plot looks reasonably well the histogram tells us the residuals have a skewed distribution.

We can further look at the estimated regression coefficients and provide confidence intervals. The function that you can use to estimate confidence intervals in R is **confint**,

which stands for confidence intervals. Then, it will give you the confidence intervals for both the intercept and the slope.

```
## Estimated Regression Coefficients  
betas = coef(model)  
Betas  
(Intercept) log(bush)  
-2.5507857 0.7561963  
## Confidence intervals for the coefficients  
confint(model)  
2.5 % 97.5 %  
(Intercept) -3.3277351 -1.7738363  
log(bush) 0.6776289 0.8347638
```

**Interpretation:**

- As number of votes for Bush increase by 1% the expected % increase of votes for Buchanan is 0.756.
- The maximum % increase is 0.677 and the minimum % increase is 0.834

The way we interpret this output is that as a number of log of votes for Bush increases by 1% the expected increase of the log votes for Buchanan is 0.756. This interpretation is on the log scale but it is better practice to provide such interpretation on the original scale.

For the confidence interval for the slope, the confidence interval is between 0.677 for the lower bound, and 0.834 for the upper bound.

Is Palm Beach an outlier? In order to evaluate whether the vote counts for Buchanan and Palm beach is an outlier, we're going to omit the Palm Beach from the analysis, and perform the linear regression model without the value of the votes for Palm Beach.

```
## Omit Palm Beach  
model.red = lm(log(buch)[-50] ~ log(bush)[-50])  
summary(model.red)  
Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) -2.31657 0.35470 -6.531 1.23e-08 ***  
log(bush[-50]) 0.72960 0.03599 20.271 < 2e-16 ***
```

So we remove the 50th observation from both the number votes for Buchanan and for Bush. And those are the estimated coefficients, along with statistical inference for the

regression coefficients from the summary output. And now we're going to predict the vote counts for Palm Beach County for Buchanan based on the model that where we omitted the number of votes for Palm Beach. And then we're going to compare what we predict with what we observed.

```
## Obtain the predicted vote count for Palm Beach given the fitted model without
new = data.frame(bush = bush[50])
## The difference between predicted on the original scale and the observed vote count
buch[50] - exp(predict(model.red,new))
[1] 2809
```

So, in order to use the predict function we need to specify the new data point that we'll use for prediction. In this case, is the number of votes for Bush for Palm Beach and we'll use the model where we omitted Palm Beach. And the difference between the observed and predicted, again predict in order to get the predicted number of votes we need to bring it to the original scale. So we need to take the exponential function and that difference again between observed and predicted is 2,809 votes. This is not a large number given the number of votes in the entire US or even in Florida.

We can also look at the prediction intervals for the number of votes for Buchanan in Palm Beach. And we can use the predict function again, but now we need to specify the type of interval we want to use, in this case 'prediction'.

```
## Prediction Confidence Interval for log(vote count)
predict(model.red,new,interval='prediction',level=.95)
## Prediction Confidence Interval on the original scale
exp(predict(model.red,new,interval='prediction',level=.95))
  fit      lwr      upr
597.5019 252.738 1412.564
## Is the observed vote count in the prediction interval?
buch[50]
[1] 3407
```

To obtain the lower and upper bound for the predicted number of votes for Buchanan, we need again to bring it back to the original scale by taking the exponential function. And from this output, the lower bound of the number of votes for Buchanan is 252, the upper bound is 1412. We compare the lower and upper bound with what we observed, the observed vote count for Buchanan and Palm Beach, which is 3,407 and this value is much, much larger than even the upper bound of the prediction interval. Which means that it's an indication that this is an outlier.

Interpreting the results, we find that the difference between predicted and observed vote count for Buchanan in the Palm Beach County is 2,809. The upper bound of the prediction confidence interval for the vote count is 1,412 which is much lower than observed vote count.

While a difference of 2,800 votes is not large given the total US votes or total Florida votes, this was particularly decisive for the 2000 election. To recall the court decision on George W Bush winning Florida was by a margin of 537 votes. This is much smaller than the difference we identify here of 2,800 votes. This analysis indicates that analysis of this kind, of even a simple linear regression, could have overturned elections in 2000.

# Unit 2: Basics of ANOVA

## 1.1 Analysis of Variance (ANOVA)

### 1. Basics of ANOVA

The topic of this lesson is Analysis of Variance (ANOVA). And in this lesson we'll learn about the data structure in the simplest ANOVA model, so called **one-way ANOVA**. And I will illustrate this model with two examples.

The data in the ANOVA model consist of multiple samples of data for a response variable of interest, differentiated in means of groups or populations described by a categorical variable or a label.

The simple model is the so-called one way ANOVA. As shown (below) in one way ANOVA we have  $k$  different populations and for each population we observe a sample of data for the response variable  $y$ .

Population 1:  $(\mu_1, \sigma_1^2) \rightarrow$  Sample 1:  $(Y_{1,1}, \dots, Y_{1,n_1}) \rightarrow (\bar{Y}_1, s_1^2)$

Population 2:  $(\mu_2, \sigma_2^2) \rightarrow$  Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2}) \rightarrow (\bar{Y}_2, s_2^2)$

.....

Population  $k$ :  $(\mu_k, \sigma_k^2) \rightarrow$  Sample  $k$ :  $(Y_{k,1}, \dots, Y_{k,n_k}) \rightarrow (\bar{Y}_k, s_k^2)$

We assume that the true mean and variance for the response variable are  $\mu_1$  and  $\sigma_1^2$  for the first population,  $\mu_2$  and  $\sigma_2^2$  squared for the second population, and  $\mu_k$  and  $\sigma_k^2$  for  $k^{\text{th}}$  population.

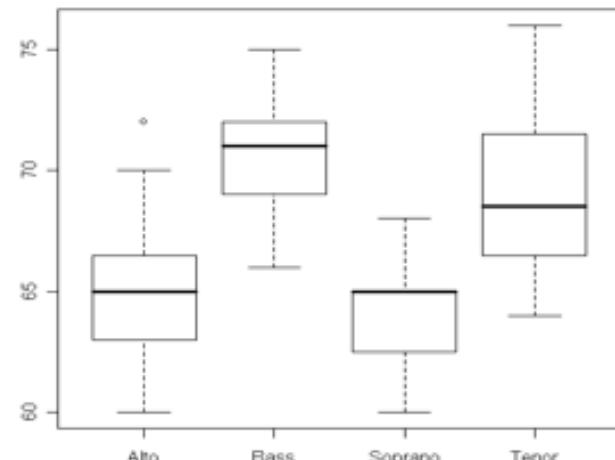
Based on the sample of data for the response variable for each population, we can obtain estimates for the mean and variance parameters for each population. The mean estimate is the sample mean denoted as  $\bar{Y}$ , and the variance estimate is the sample variance denoted as  $s^2$ . **The overarching objective in the ANOVA is to compare the means across the  $k$  populations, are the means equal? Which pairs of the means are different?**

## ANOVA: Comparing the means of multiple samples

### Example: Choir voices and height

In the first illustrative example of one-way ANOVA, the population of interest consists of singers in the New York choral society, and the response variable of interest is their height to the nearest inch. The singers are grouped by their voice pitch, from highest pitch to lowest pitch including soprano, alto, tenor, and bass. The first two are typically sung by female voices and the last two by male voices. One can examine how height varies across voice pitches.

One approach to evaluate or study the response variable with respect to a categorical variable (in this case) voice pitch is to use the side by side boxplot. So what that means is that for each category we plot a single box plot of the response variable and we compare the box plots across all categories.



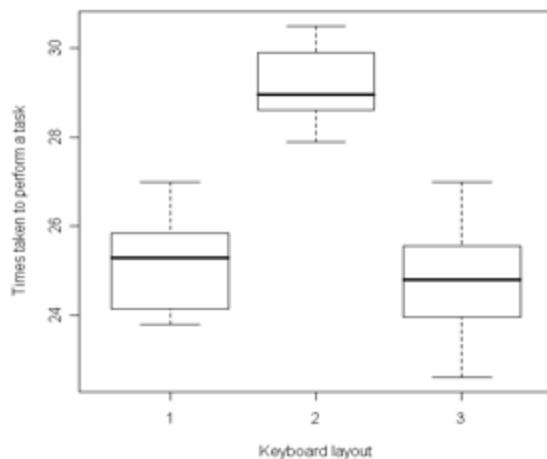
So in this case, because we have four categories, we have four box plots. In this example, we may be interested to address the following question, is there a difference, in the height by voice pitch? Which singers are taller?

To do so, we compare whether the means across all four groups are different. And particularly we're going to compare in ANOVA the within-variability which is the variability within voice pitches, visually we can assess that by the variability within each box. Versus the between-variability, which is the variability between the means which are the lines. *So we will identify differences across means if the between-variability is larger than within-variability.*

### Example: Keyboard layout efficiency

In the second example we're interested in the typing speed of three computer keyboard layouts by determining whether there are differences in their respective mean speeds.

Is there a difference in the time taken to perform a test across the three layouts? Which layout is more effective?



If we compare the three layouts, we can see the second layout has a much higher mean speed as compared to layouts one and three, which means we may detect at least visually, differences across the means at least between layout one and layout three. But are those differences statistically significant?

## Primary Objectives of ANOVA

In ANOVA, the primary objectives are to:

- Analyze the variability in the data using the ANOVA table.** That means we compare the variability within each group to the variability between the means.
- Use this analysis of variance** in order to test whether the means are equal.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

- Estimate confidence intervals** for all the pairs of means, in order to identify which of the means are not equal, or which of the means are statistically significantly different.

$$\mu_i - \mu_j \text{ for } i \text{ and } j = 1, \dots, k$$

The second and third objectives are **statistical inference problems**, and we'll use hypothesis testing and confidence intervals to evaluate statistical difference between the means.

## 2. Estimation Method

The topic of this lesson is analysis of variance, and we're going to focus on parameter estimation. We'll learn about the estimation of both the means and the variance, which is assumed constant across the populations.

As provided in the previous lesson of ANOVA, the data in the ANOVA model consist of a response variable of interest observed for multiple populations differentiated by a categorical variable or a label. For example, for the analysis of suicide rates, the categorization could be instead age group, year, weather type etc. Let's begin with the model. In our notation,  $Y_{ij}$  are the response data differentiated across the  $k$  categories.  $j$  is the index within group and  $i$  is the index across groups. The model is:  $Y_{ij} = \mu_i + \epsilon_{ij}$  (mean of the group  $i$ ) +  $\epsilon_{ij}$  (the error term  $\epsilon_{ij}$ ).

**Data:**  $Y_{ij}$  for  $j = 1, \dots, n_i$ ;  $i = 1, \dots, k$

**Model:**  $Y_{ij} = \mu_i + \epsilon_{ij}$  where  $\epsilon_{ij}$  = error term

The assumptions on ANOVA are with respect to the error term:

- Constant variance assumption. Recall that we assume the response data has constant variance across all groups, and thus the variance of the error terms is constant, equal to sigma square:  $\text{Var}(\epsilon_{ij}) = \sigma^2$
- Independence assumption. Response data and error terms (" $\epsilon$ ") are independent.
- Normality assumption. Error terms are normal, and thus the  $y_{ij}$  response data are normal as well.

In ANOVA, we assume that the variance of the response variable is the same across all populations and equal to sigma square. Thus, we compare the means, assuming the variances are the same, and estimate the variance across all samples using the so-called **pooled variance estimator (aka mean squared error MSE)**.

### Pooled Variance Estimator:

$$S_{\text{pool}}^2 = \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{\sum_{i=1}^k (n_i - 1)} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2}{N - k}$$

In the formula for the pooled variance,  $S^2$  are the sample variances of individual samples of data. And the  $n$ 's --  $n_1, n_2, n_3, n_k$ , and so on -- are the sample sizes of the individual samples. By adding up the weighted sample variances, we get the sum of the pool variance estimator. The big  $N$  is the total number of samples.

The degrees of freedom in the pooled variance estimator is  $N - k$  because we replace  $k$  different means with their sample means, and thus we lose  $k$  degrees of freedom. I'll come back to this aspect when we're going to discuss the sampling distribution for the pooled variance estimator.

The formula above for the pooled variance estimator is also called the **mean square error in ANOVA**.

The sum of squares of the responses minus the sample means of the individual samples form what we call the sum of square error:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \text{Sum of Squares of Error} = \text{SSE}$$

If we divide the sum of squared errors by  $N - k$  we end up with a mean square error.

Determining sampling distribution for the pooled variances

The individual sample variances for the  $k$  samples have a chi-square distribution because we assume that the data are normally distributed. An important property of

the chi-square distribution is that, **if we have independent chi-squared random variables, their sum is also a chi-square distribution.**

$s_1^2, \dots, s_k^2$     *The sum of independent Chi-square random variables is also Chi-square*

So now, if we sum the individual sample variances multiplied by  $N_i - 1$  divided by sigma squared, the result is a chi-square distribution where the number of degrees of freedom is the sum of the degrees of freedom across the k chi-squared distributions:

$$\frac{SSE}{\sigma^2} = \frac{(n_1-1)s_1^2}{\sigma^2} + \dots + \frac{(n_k-1)s_k^2}{\sigma^2} \sim \chi_v^2 \text{ where } v=N-k$$

**In a nutshell, the sampling distribution of the pooled variance is a chi-square distribution with  $N - k$  degrees of freedom.**

### Estimation of Mean Parameters

Let's go back to the estimation of mean parameters. We use the sample mean of individual samples to estimate the mean parameters:

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

But what is the sampling distribution of those estimated means?

Remember that we assume data are normally distributed for each sample -- that is, they have a normal distribution with mean  $\mu_i$  and sigma square:

$$\text{If } Y_{1i}, \dots, Y_{ni} \sim N(\mu_i, \sigma^2)$$

And we estimate  $\mu_i$  with  $\hat{\mu}_i$  which is the average across the responses:

$$\hat{\mu}_i = \bar{Y}_i = \frac{Y_{1i} + \dots + Y_{ni}}{n_i} \sim N(\mu_i, \frac{\sigma^2}{n_i})$$

And we know from basic statistics that the sampling distribution of the sample mean is also normal within  $\mu_i$  and variance  $\sigma^2$  divided by the sample size  $n_i$ .

However, we do not know  $\sigma^2$ , so we replace  $\sigma^2$  with the pooled variance estimator, the mean squared error.

So in this context now the sampling distribution would change, it will be a t distribution:

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\frac{MSE}{n_i}}} \sim t_{N-k}$$

But we have a t distribution with  $N - k$  degrees of freedom because the number of degrees of freedom of the chi-square distribution of  $\sigma^2$ -hat is  $N - k$ .

**Why  $N - k$ ?**

$$MSE = \hat{\sigma}^2 \sim \chi^2_{N-k}$$

So the  $N - K$  from the t distribution, comes from the number of degrees of freedom of the distribution for the estimated variance, the pooled variance estimator.

We can use the estimated sample means

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \text{ for } i = 1, \dots, k$$

and the estimated variance

$$\hat{\sigma}^2 = MSE$$

to calculate  $(1 - \alpha)$  confidence intervals for the treatment means:

$$\left( \hat{\mu}_i - t_{\alpha/2, N-k} \sqrt{MSE/n_i}, \hat{\mu}_i + t_{\alpha/2, N-k} \sqrt{MSE/n_i} \right)$$

With the sample estimated mean and estimated variance, we now can get the confidence intervals for each individual means. Just like confidence intervals for sample means, we center the confidence intervals at the sample means, plus or minus the key critical points with N, where the t distribution has n minus k degrees of freedom, multiplied by the standard error of the  $\mu_i$  hat which is square root of the mean square error divided by  $n_i$ . Again, the mean square error replaces sigma square. The mean square error is the estimator for the variance.

### 3. Estimation Data Examples

The topic of this lesson is analysis of variance, and I'll illustrate parameter estimation using an R example. Particularly, I will show you with R, how to obtain the estimated means and how to interpret them.

Let's go back to the example where we are interested in differences in height among singers with different voice pitches. The question that we wanted to address is,

*what are the estimates for the mean height for the different groups of singers?*

To run an ANOVA model in R, we can use the **R command aov()**. What we need to input in this command is the response variable height which is the response variable consisting of all the samples, all the case samples stacked in one vector. And on the right is the vector of labels, in this case is voice pitch, telling R which of the labels correspond to what responses.

To get the estimated means, we can use the model.table command in R. But you need to specify what kind of summary you want to obtain from the parameters, means or medians, for example. In this case we're interested in estimated means.

```
model = aov(height ~ pitch)
model.tables(model, type = "means")
```

Tables of means

Grand mean

67.11538

pitch

	Alto	Bass	Soprano	Tenor
rep	35.00	39.00	36.00	20.00
64.89	70.72	64.25	69.15	

Overall Mean: 67.11536

$\hat{\mu}_{\text{alto}} = 64.89$

$\hat{\mu}_{\text{bass}} = 70.72$

$\hat{\mu}_{\text{soprano}} = 64.25$

$\hat{\mu}_{\text{tenor}} = 69.15$

The output is provided and what the output provides is the grand mean, the overall mean across all samples, which in this case is 67.11, as well as the means of the individual samples. For example the height for singers with alto voice pitch, the sample mean is 64.89, where for bass the sample mean is 70.72.

So those are the estimated parameters that we would like to compare. We see there are some differences across the means of the four samples. But the question we will address in the next lesson is are those differences statistically significantly different?

In the second example, we're interested in the typing speed across three different keyboard layouts. **And we want first to estimate the mean typing speeds across the three keyboard layouts.** The same as in the previous example, we can use the aov command, along with a model.tables command provides us the estimating means.

```
model = aov(speed ~ layout)
model.tables(model, type = "means")
```

	Tables of means												
Grand mean	26.21212												
layout	<table><thead><tr><th></th><th>1</th><th>2</th><th>3</th></tr></thead><tbody><tr><td>rep</td><td>25.12</td><td>29.11</td><td>24.76</td></tr><tr><td>12.00</td><td>10.00</td><td>11.00</td><td></td></tr></tbody></table>		1	2	3	rep	25.12	29.11	24.76	12.00	10.00	11.00	
	1	2	3										
rep	25.12	29.11	24.76										
12.00	10.00	11.00											

Overall Mean: 26.21212
$\hat{\mu}_{\text{layout1}} = 25.12$
$\hat{\mu}_{\text{layout2}} = 29.11$
$\hat{\mu}_{\text{layout3}} = 24.76$

For layout 1, the sample mean is 25.12. For the second layout it's 29.11. For the third layout it's 24.76. This is in seconds.

What this means is that there are differences we can see when we compare the three means, there are differences. Particularly the second layout has a much larger typing speed than the other two, but are those, again, statistically significantly different? We will have to perform a hypothesis testing procedure in order to address that question.

## 4. Test for Equal Means

The topic of this lesson is Analysis of Variance. And we're going to focus on hypothesis testing procedures for equal means. We will learn about how to perform the statistical test, particularly how to derive the test statistic and how to decide based on a test statistic. And we also demonstrate with two examples how to implement the test and how to draw inferences based on the test.

Using the **hypothesis testing procedure for equal means**, we test:

**Null hypothesis:** the means are all equal ( $\mu_1 = \mu_2 = \dots = \mu_k$ ) versus

**Alternative hypothesis:** some means are different.

Not all means have to be distinct for the alternative hypothesis to be true -- at least one pair of the means needs to be different.

Under the null hypothesis, we can combine all  $k$  samples into one big large sample because assuming all means are equal implies that observations have a normal distribution with a common mean  $\mu$  and a common variance  $\sigma^2$ . We can estimate this common mean by pooling all the case samples into one sample and estimating the mean with the sample mean of this combined case samples:

$$\bar{Y}_{..} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

If we want to estimate the variance, similarly we're going to use the sample variance of the entire combined case samples. The difference between this variance estimator and the mean square error or the pool variance estimator is that now we are replacing the mean with the overall mean, not with the individual sample means. Because we are replacing only one parameter, the common parameter with a sample mean, we're now only losing one degree of freedom.

We can rewrite this estimate as sum of square total divided by big  $N - 1$ :

$$S_0^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}{N-1} = \frac{SST}{N-1}$$

- **SST = Sum of Squares Total**

Again, N is the sum of all samples. Because we only have to estimate one mean, we only lose 1 degree of freedom above, and thus the denominator is N-1, not N-k.

So now, the **sample distribution of the variance estimator** is the chi-square distribution with N- 1 degrees of freedom -- note, not N – k as we had for the pooled variance estimator:

$$\frac{(N-1)S_0^2}{\sigma^2} = \frac{SST}{\sigma^2} \sim \chi_{N-1}^2$$

#### SST Decomposition

The sum of squares total can be decomposed into two components, the sum of square of error, plus the sum of square of treatments:

$$\mathbf{SST = SSE + SST_R}$$

The sum of square of errors (SSE) is the sum of square differences between the observations and the individual sample means:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

The sum of square treatment (SSTr) is the sum of the square difference between the sample means of the individual samples minus the overall mean:

$$\text{where } SST_R = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2 \text{ and } \bar{Y}_i = i^{\text{th}} \text{ sample mean}$$

The interpretation of the decomposition is as follows:

1.  $MSE = SSE/(N-k) = \text{within-group variability}$
2.  $MSST_R = SST_R/(k-1) = \text{between-group variability}$
3. ANOVA: comparing between to within variability
4.  $F = \text{between-group variability}/\text{within-group variability}$

1. The ratio between the sum of squared errors (SSE) divided by  $N - k$  is called **mean sum of squared errors (MSE)**. It's a measure of the within-group variability. Remember that we used this to estimate the pool variance estimator.
2. The **mean sum of square treatments (MSSTr)** is the sum of square treatment ( $SSTr$ ) divided by  $k - 1$ , where  $k$  is the number of samples. And this is a measure of the between-group variability.
3. One of the main purposes of ANOVA is to compare the variability between samples to the variability within a sample.
4. **F-test** is the ratio of between-group variability and within-group variability

#### Testing Equal Variances With The F Test

The **F-test** is the ratio between the sum of square treatments divided by  $k - 1$ , divided by the sum of square of errors divided by  $N - k$ :

$$\frac{SST_R/(k-1)}{SSE/(N-k)}$$

This is equivalent to the mean sum of square treatments (MSSTr) divided by the mean sum of square of error MSE:

$$\frac{SST_R/(k-1)}{SSE/(N-k)} \equiv \frac{MST_R}{MSE} = F_0 \sim F_{k-1, N-k}$$

The F-test compares variability *between* groups against  $k$ - variability within groups.

If the F-test is large, variability between groups is larger than variability within groups, and thus we reject the null hypothesis that the means are equal. We make this

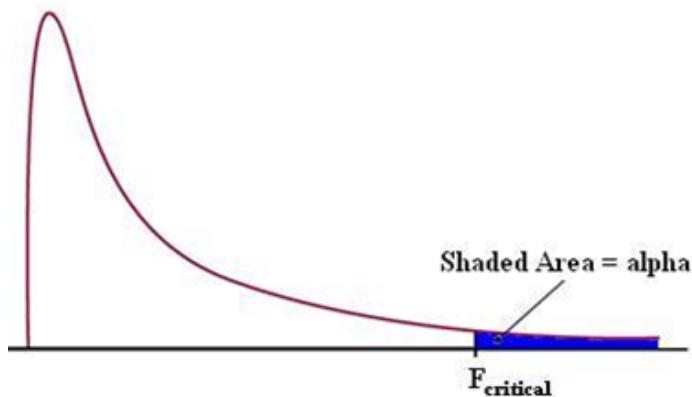
determination by comparing the F-test with the **F critical point** where the F critical point is for F distribution, with  $k-1$  and  $N-k$  degrees of freedom:

Reject  $H_0$  if  $F_0 > F_{\alpha}(k-1, N-k)$ , which is the upper  $\alpha^{\text{th}}$  quantile of the F distribution.

We can also make a decision based on the p value:

$$\text{P-value for the F-test} = P(F > F_0), \text{ where } F \sim F_{(k-1, N-k)}$$

which is the area under the right tail of the F distribution, shown here in blue (and the F-critical value is the F statistic):



### Example: Testing for Equal Means

Let's go back to the first example where we are interested in comparing the height of singers that have different voice pitches. **Are the mean heights for the four groups of singers statistically different?**

To perform the testing procedure for equal means, we can use the **AOV() Command in R**. And to extract the ANOVA table, we can use the `summary()` command. What I'm showing you here the output is what we call the ANOVA table:++

```
summary(aov(height ~ pitch))
      Df Sum Sq Mean Sq F value Pr(>F)
pitch      3   1058.5    352.8   55.8 <2e-16 ***
Residuals 126   796.7     6.3
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSTR = 1058.5, k-1 = 3  
SSE = 796.7 , N-k = 126  
F-value = 55.8  
P-value ≈ 0

P-value ≈ 0 : Reject the null hypothesis of equal mean heights

In the ANOVA table, we have two important rows, one corresponding to the treatments (voice pitches) and the one corresponding to the residuals or the errors.

- The first column provides the degrees of freedom for the treatment versus the residuals, there are three degrees of freedom for treatments because we have k groups. So k- 1 is going to be 3. We have 126 degrees of freedom for residuals, which corresponds to big N- k.
- The second column provides the values for the sum of squares, and the sum of square treatment is 1058.5. The sum of square of error is 796.7.
- The third column provides the mean sum of squares. That means we take the sum of square treatment, and divide it by the corresponding degrees of freedom. So we take 1,058.5 and divide it by 3, and we're going to get 352.8.
- The fourth column gives us the F value which is the ratio between the mean sum of square treatments divided by mean sum square of errors.
- The last column provides the P value of the F test. What we learn from here is that because the P value is approximately equal to zero, we reject the null hypothesis of equal mean height across the four groups differentiated by voice pitches.

In the second example, we're interested to compare the means of the typing speed between the three keyboard layouts. **Are the mean typing times for the three keyboard layouts statistically different?** Similar to the previous example, we perform an ANOVA using R. The summary of this ANOVA is going to be the ANOVA table, and here we have the treatments are the keyboard types:

```
summary(aov(speed ~ keytype))
   Df Sum Sq Mean Sq F value Pr(>F)
keytype     2   121.24    60.62   52.84 1.48e-10 ***
Residuals  30   34.42     1.15

```

--  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**SSTR = 121.24, k-1 = 2**  
**SSE = 34.42 , N-k = 30**  
**F-value = 52.84**  
**P-value ≈ 0**

**P-value ≈ 0 : Reject the null hypothesis of equal mean typing times**

We have two degrees of freedom for the treatment because we have three groups. We have N- k is 30, the sum of square of treatments is 121.24, the sum of square of error is 34.42, the F-Value is 52.84. And because the P-Value is approximately equal to zero we conclude that we reject the null hypothesis of equal mean typing speeds.

## Knowledge Check

### Question 1

1/1 point (ungraded)

The total sum of squares divided by N-1 is

- The mean sum of squared errors
- The sample variance estimator assuming equal means and equal variances ✓
- The sample variance estimator assuming equal variances.
- None of the above.

### Question 2

1/1 point (ungraded)

The mean squared errors (MSE) measures:

- The within-treatment variability. ✓
- The between-treatment variability.
- The sum of the within-treatment and between-treatment variability.
- None of the above.

### Question 3

1/1 point (ungraded)

Which is correct?

- If we reject the test of equal means, we conclude that all treatment means are not equal.
- If we do not reject the test of equal means, we conclude that means are definitely all equal
- If we reject the test of equal means, we conclude that some treatment means are not equal. ✓
- None of the above.

## 2.2 Basic Concepts and Estimation

### 1. Comparing Pairs of Means

The lecture is on Analysis of Variance. In this lesson, I'll focus on comparing pairs of means. This will apply this approach after we reject the null hypothesis of equal means. I'll use a hypothesis testing procedure.

**One primary goal of ANOVA might be to determine which treatment means are bigger or smaller.** One way to do this is to compare all possible pairs, which are  $k(k-1)/2$  pairs of treatments. We do so using confidence intervals for differences in means. This is called **pairwise comparison**. That is, we estimate the difference in the means (for example, a pair:  $\text{mean}_i$  and  $\text{mean}_j$ ) as a difference between their corresponding means. And the **confidence interval** is going to be centered at the estimated mean difference, plus or minus a critical point times the standard deviation of the estimated difference in means:

$$(\hat{\mu}_i - \hat{\mu}_j) \pm q_{\alpha, k, N-k} \sqrt{\frac{MSE}{2} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}$$

The estimate of the difference in means      The  $\alpha$  percentile of the “studentized range” distribution.      The standard deviation of the estimator

Here, the critical point is now the alpha percentile of the studentized range distribution. why do we use the q critical point rather than the t critical point? The reason is that we want to correct for simultaneous or joint inference. The q critical point allows for correcting for multiplicity, which means that the intervals are wider, to compensate for the fact that we're making simultaneous comparisons

## Confidence Intervals with Multiple Populations

*When we perform as a joint inference, confidence will decrease if we do not perform a correction.*

Consider calculating a 95% confidence interval for two population means based on two independent samples. But here, we're not interested in confidence intervals for each individual population separately; we are interested in simultaneous inferences on both. This means that their significance level is  $(.95)(.95)$ , or about 90%, not 95% as we initially wanted.

Now imagine we are analyzing three populations, and want 95% confidence intervals. If we multiply 0.95, with 0.95, and 0.95, we get 0.86, not 95%.

The correction using q critical points for the purest comparison in ANOVA is through this q value, which is not easy to compute. There are some tables in some of the textbooks, but I would suggest you use statistical software to get the confidence intervals. And the R statistical software does have a command that you can use called a **Tukey method**, developed by statistician John Tukey.

### R Function: TukeyHSD()

Let's illustrate the implementation of the comparison with the first example: which mean heights for the four groups of singers are statistically different?

Recall that we rejected the hypothesis that the means are equal, so what that means is that at least two of the means are statistically different. The function you can use to obtain the confidence interval is called **TukeyHSD()** and what you need to input is the fitted ANOVA model.

If you want to use a confidence level different than 95% which is the default, you need to specify that, as well. The output of this command is:

*TukeyHSD(aov(height ~ pitch))*

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = height ~ pitch)

\$pitch

	diff	lwr	upr	p adj
Bass-Alto	5.8322	4.3078	7.3566	0.0000
Soprano-Alto	-0.6357	-2.1898	0.9184	0.7114
Tenor-Alto	4.2642	2.4290	6.0995	0.0000
Soprano-Bass	-6.4679	-7.9811	-4.9547	0.0000
Tenor-Bass	-1.5679	-3.3686	0.2327	0.1113
Tenor-Soprano	4.9000	3.0740	6.7259	0.0000

So, the way you interpret this output is as follows:

- The first column provides the pairs of the means that we compare. For example, the first pair is the mean height between those with Bass voice pitch versus those with Alto voice pitch.
- The second column will provide the differences in the estimated means.
- The next columns, lower and upper, provide the lower and upper bounds of the 95% confidence intervals for the differences in the means.
- The last column is the **adjusted p-value**, which gives us information whether we reject or not the null hypothesis for when we test whether the means of the pairs are equal or not.

So how can I use this output? First, you must look at the lower and upper bounds, and identify the confidence intervals that include the zero values. For those confidence intervals, it's plausible for the difference to be zero.

So for example, if you look at the difference between the mean of mean height of Soprano versus Alto. The lower bound is -2.18, the upper bound is 0.91. So this confidence interval includes zero, which means that the means of those two groups could be possibly equal to zero. You can also see the probability of p-value adjusted is 0.71, which is a large p-value. And at this large p-value, we do not reject the null hypothesis of equal means.

Another aspect that we'll look into when evaluating the pairwise comparison is identifying the confidence levels that have only positive values or only negative values. For example, in the mean height for Tenor and Alto, all the values in the confidence

levels are positive, which means that the mean height of singers with a tenor voice pitch is statistically larger than the mean height of the singers with an alto voice pitch.

So here's a conclusion based on this output:

- *Singers with bass or tenor pitch are statistically significantly taller than those with alto pitch, in average.*
- *Singers with soprano pitch are statistically significantly shorter than those with a bass pitch, in average.*
- *Singers with tenor pitch are statistically significantly taller than those with a soprano pitch, in average.*
- *Those with soprano and alto pitch may plausibly have similar heights, in average.*
- *Those with tenor and bass pitch may plausibly have similar heights, in average.*

Let's go back to the second example where we want to compare the means of the typing speed of the three keyboard layouts. Which mean typing speeds for the three keyboard layout are statistically different? Again, we can perform a pairwise comparison using the Tukey method.

We have only three pairs: 2 and 1, 3 and 1, and 3 and 2. For the pair 3 and 1, the third and the first layout, the confidence interval includes zero, which means that the means of the typing speed for the two layout is plausibly similar, plausibly equal.

When we compare layouts 1 and 2 and layouts 3 and 2 respectively, we can see statistically significant difference, a large differences.

`TukeyHSD(aov(speed ~ keytype))`

Tukey multiple comparisons of means  
95% family-wise confidence level  
Fit: aov(formula = speed ~ keytype)

\$keytype

	diff	lwr	upr	p adj
2-1	3.9850	2.8543	5.1156	0.0000
3-1	-0.3613	-1.4635	0.7408	0.7008
3-2	-4.3463	-5.5000	-3.1926	0.0000

So in conclusion while we learn from this example is that the keyboard type 2 has statistically significantly higher typing speed or time than keyboard layouts 1 and 3 on average. And it's plausible that keyboard layout 1 and 3 have similar typing speed on average.

## 2. Model Fit Assessment

The topic of this lesson is analysis of variance with a focus on model fit assessment. Particularly, we're going to overview the ANOVA model and the model assumptions. We'll learn about simple ways to diagnose these assumptions using graphical displays.

As provided in the first lesson of ANOVA, the data in ANOVA modeling consists of a response variable of interest observed for multiple populations differentiated by a categorical variable or a label. For example, the voice speech could also be colors, seasons, or other type of categorizations.

**Data:**  $Y_{ij}$  for  $j = 1, \dots, n_i; i = 1, \dots, k$

**Model:**  $Y_{ij} = \mu_i + \varepsilon_{ij}$  where  $\varepsilon_{ij}$  = error term

In our notation,  $y_{ij}$  are the response data differentiated across the  $k$  categories.  $j$  is the index within group and  $i$  is the index across groups. The model is:  $Y_{ij} = \mu_i$  (mean of the group  $i$ ) +  $\varepsilon_{ij}$  (the error term epsilon<sub>ij</sub>).

### Assumptions of ANOVA

The assumptions on ANOVA are with respect to the error term:

- **Constant variance assumption.** Recall that we assume the response data has constant variance across all groups, and thus the variance of the error terms is constant equal to sigma square:  $\text{Var}(\varepsilon_{ij}) = \sigma^2$
- **Independence assumption.** Response data and error terms (" $\varepsilon$ ") are independent.
- **Normality assumption.** Error terms are normal, and thus the  $y_{ij}$  response data are normal as well.

To diagnose these assumptions, we did not diagnose assumptions under error terms because we did not know the means. Instead, we're going to diagnose the assumptions on the residuals. The residuals are the difference between the responses minus the estimated means of the individual samples:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\mu}_i$$

**If the model fit is a good fit, then the residual should be scattered around zero (randomly) if we look at the plot of the residuals against either fitted or just their order.**

### Types of Diagnostic Plots in ANOVA

When we evaluate the assumptions in ANOVA, here are the type of residual plots we consider:

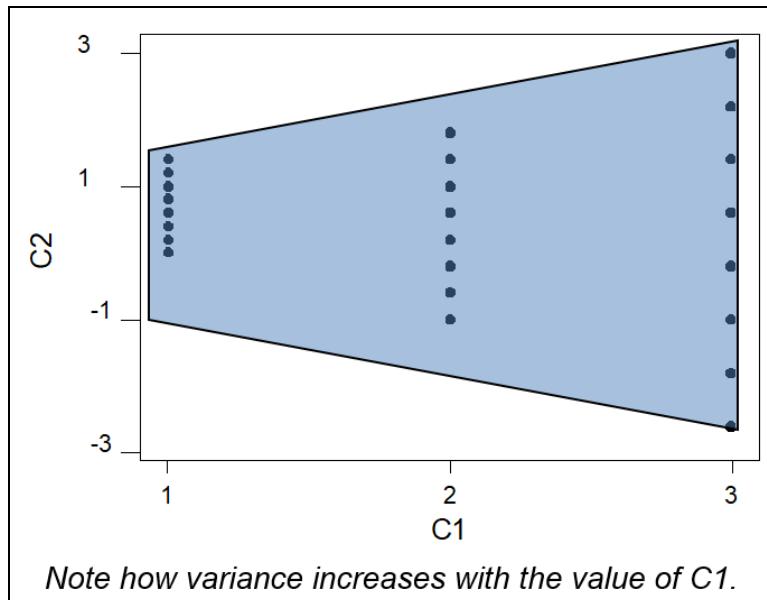
- We can plot the **residuals for each treatment group** to evaluate whether there is a different variability across the groups.
- We can plot the quantile-quantile normal plot to evaluate normality
- We can also plot a histogram of the residuals, similarly used to evaluate normality.

If the scatter plot of the residuals ( $\epsilon_{ij}$ ) is **NOT random**:

- The sample responses are not independent, or
- the variances of responses are not equal

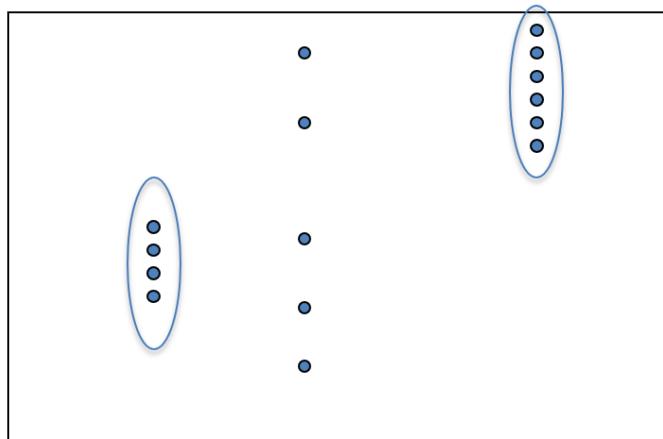
**If the quantile-quantile normal plot and the histogram show departure from normality, you may consider a transformation in order to normalize the data.**

Here is one example for departure from the assumption of *non-constant variance*:



You can see here that the variability of the residuals changes from group one to group three. We have a much smaller variability for group one than for group three.

This is another example of departure from one of the assumptions - we can see the residual clusters, which means that we have *correlated errors*:

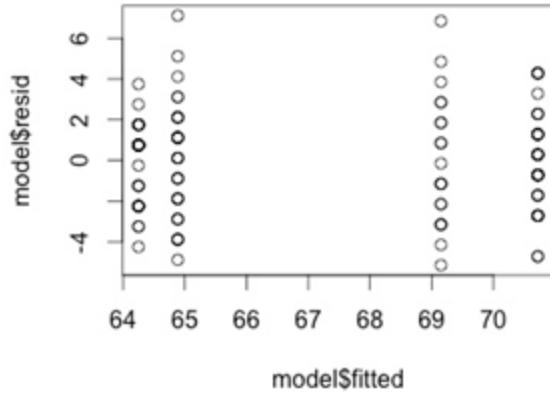


*Clusters of residuals: correlated errors*

### Example: Voice Pitches and Model Fit Assessment

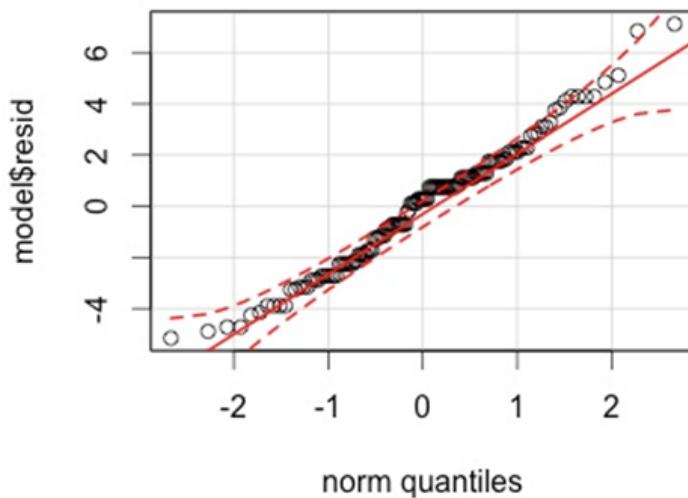
So let's go back to the first ANOVA example, where we're comparing the mean height across different voice pitches. Are the inferences on the difference in height means reliable?

In order to address this question, we will perform a residual analysis using three plots. The first one plots residuals against the fitted means, the estimated means of the four groups, by voice range:



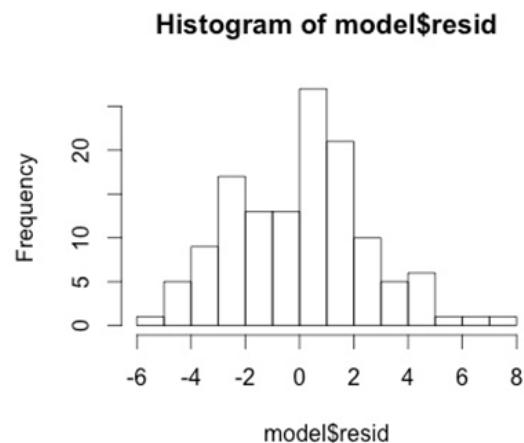
We're going to use this plot in order to evaluate whether the residuals are scattered around the zero line, and whether the variances are different across the four groups or whether there are clusters of individuals which is an indication of analytical errors.

The second plot is the quantile-quantile normal plot:



Here, we expect the residuals to line up on the line which means that the residuals will have a similar distribution to the normal distribution.

The third plot is a histogram:



What we expect to see here is it's approximately symmetric distribution. For this example, all three assumptions hold.

## Knowledge Check

### Question 1

1/1 point (ungraded)

The objective of the residual analysis is

- To evaluate departures from the model assumptions ✓

- To evaluate whether the means are equal.

- To evaluate whether only the normality assumptions holds.

- None of the above.

### Question 2

1/1 point (ungraded)

The objective of the pairwise comparison is

- To find which means are equal.

- To identify the statistically significantly different means. ✓

- To find the estimated means which are greater or lower than other.

- None of the above.

### 3. ANOVA vs. Simple Linear Regression

The topic of this lesson is the comparison between the two models we've learned so far, analysis of variance and simple linear regression. Particularly, we'll learn that ANOVA is a particular case of linear regression.

In other regression models examined so far, both the response and predictive variables have been quantitative. Can this be generalized to analyzing the variability in a response variable with different groups of predicting variables? For example:

- Does knowing the education level of a person, say high school, college, have predictive power for their annual salary?
- Is a return on a stock related to the industry group of the company?
- Is the height of a singer related to the voice pitch?

This is a special kind of regression question in this context: if group membership has predictive power for the response, then the average mean of the response variable is different for different groups. This is thus actually a comparison of means, as we learned in analysis of variance. So ANOVA is a linear regression model where the predicting factor is a categorical variable.

#### Decomposition of ANOVA into a linear regression model

The data in the ANOVA is a response variable  $Y_{ij}$ , but we can write the  $Y_{ij}$  as the sum between the mean of the group  $i$  plus an error term. We can write this further. We can write  $\mu_i$  as  $\mu$  plus  $\tau_i$ , where  $\mu$  is the overall mean and  $\tau_i$  are the so-called treatment effects, where the sum of  $\tau_i$  is equal to 0:

#### **ANOVA:**

**Data:**  $Y_{ij} = 1, \dots, n_i; i = 1, \dots, k$

**Model:**  $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$  where  $\sum_{i=1}^k \tau_i = 0$

$\mu_i$  =  $i$ -th group mean decomposed into  $\mu_i = \mu + \tau_i$

So let's see how we can write this model into a regression model. So now I take all the  $Y_{ij}$ 's and I stack them up into a big vector. The first  $n_1$  values correspond to the first group, the first sample, the next  $n_2$  values correspond to the second group. The last  $k$

values in this vector will correspond to the kth sample. So this is going to be the Y response variable:

**Define Y be the response variable:**

$$Y = (Y_{11}, \dots, Y_{1n_1}, Y_{21}, \dots, Y_{2n_2}, \dots, Y_{k1}, \dots, Y_{kn_k})$$

Now, I define another variable. I call it a label or categorical variable. Again, I stack those values, Ls, which are the labels. The first  $n_1$  values in this vector corresponds to label 1, the label of the first category. The next  $n_2$  values correspond to the label of the second category. The last  $kn_k$  values correspond to the label of the kth sample:

**Define L be the label/categorical variable:**

$$L = (l_{11}, \dots, l_{1n_1}, l_{21}, \dots, l_{2n_2}, \dots, l_{k1}, \dots, l_{kn_k}) \text{ Where } l_{ij}=I$$

So now ANOVA is a linear regression where we regress the label onto Y:

### **Linear Regression: $Y \sim L$**

When we have a categorical or labeled predicting variable with k different labels, we then convert those into what we call **dummy variables**, labeled  $x_1$  to  $x_k$ . The  $x_1$  for example have one for the first  $n_1$  values and 0 for the rest of the values. The kth dummy variable has 0s at the beginning but the last  $n_k$  values are 1's. All of those dummy variables have the same length, and the length is N, which is the sum of all sample sizes across the k samples:

**Categorical Variables in Linear Regression:**

•**Transform categories into dummy variables**

$$x_1 = (1, \dots, 1, 0, 0, \dots, 0)^T; \dots; x_k = (0, 0, 0, \dots, 1, \dots, 1)^T$$

When we model a regression analysis with those variables, now those k dummy variables become the predicting variables, where Y (presented in a previous slide) is the response variable. If the model has an intercept we only include  $k-1$  dummy variables because of the linear dependence between the x's:

- If intercept in the model, only k-1 dummy variables because of linear dependence:  $(1,1,\dots,1)^T = x_1 + x_2 + \dots + x_k$
- Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{k-1} x_{ik-1} + \varepsilon_i, i=1,\dots,n$

If the model does not have an intercept, then we'll include all k dummy variables in the model:

- If no intercept in the model, all k dummy variables
- Model:**  $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, i=1,\dots,n$

But most importantly, what you can see here is that I wrote the **ANOVA as a multiple linear regression model where the predicting variables are the dummy variables and the response variable is the response data stuck into one vector.**

So ANOVA is a particular case of a multiple regression model. And in fact,  $\beta_1$ ,  $\beta_2$  and  $\beta_k$  in a model with no intercept correspond to the mean parameters, mu 1, mu 2 and mu k.

## 4. Data Example

The topic of this lesson is analysis of variance and our straight analysis of variance with a part one R data example. What we're going to learn in this example is how to implement ANOVA in R and how to draw inferences based on the R output.

### Example: Cancer Survival

In this example, we're interested in the number of survival days for patients with different types of cancers that were treated with ascorbate. The response variable is the number of survival days for each patient for different types of cancers. J again is the index for a patient and i is the index for the type of cancer or the group.

We have five different groups ( $k = 5$ ). And each group responds to a cancer of a different organ (stomach, bronchus, colon, ovary, breast). There are different sample sizes across the five groups. In fact, the group with the ovarian cancer has only six observations:

**Response Variable:**

$Y_{ij}$ = The number of survival days for the  $j^{\text{th}}$  patient with  $i^{\text{th}}$  type of cancer

**Categories:**

Cancer type i for  $i = 1,2,3,4,5$

Stomach	Bronchus	Colon	Ovary	Breast
124	81	248	1234	1235
42	461	377	89	24
25	20	189	201	1581
45	450	1843	356	1166
412	246	180	2970	40
51	166	537	456	727
1112	63	519		3808
46	64	455		791
103	155	406		1804
876	859	365		3460
146	151	942		719
340	166	776		
396	37	372		
	223	163		
	138	101		
	72	20		
	245	283		

A first step in any data analysis is to perform an exploratory data analysis. In ANOVA, we are interested in evaluating how the variability in the response variable changes across the different groups.

First, we're going to read the data in R using the **read.table()** command, where we need to specify the name of the file and whether the columns have a header or not. Then we need to extract the individual variables of interest. The response variable

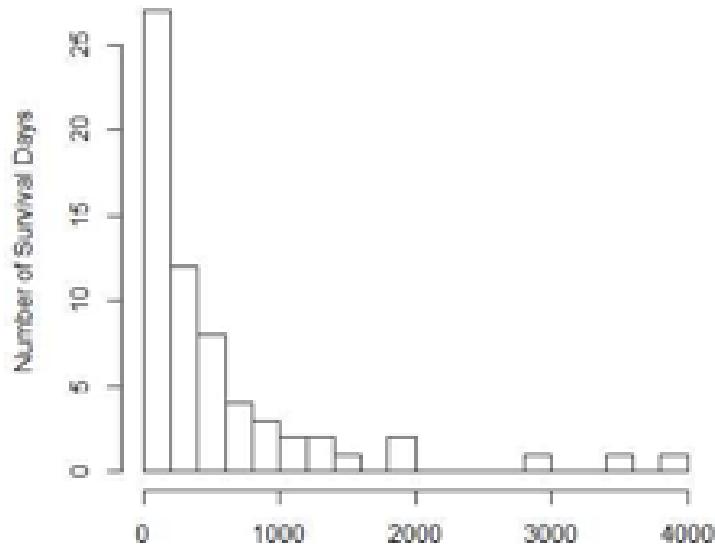
survival and the way we can extract it from the data is by specifying the name of the matrix, cancer\_data. The name of the column with a dollar sign in-between.

## R Functions: hist() and log()

Our first step is to analyze whether the survival S has an approximately normal distribution using a histogram with **R function hist()**. Do not forget to specify the X axis and the Y axis. X axis are going to be the range of the survival data, but most importantly the Y axis, the label for the Y axis is the number of survival days:

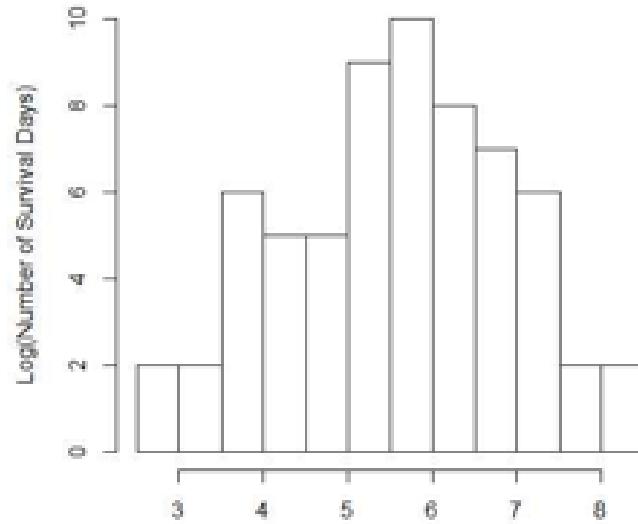
```
## Read data with 'read.table' R command for reading ASCII files
cancer_data=read.table("CancerStudy.txt",header=T)
## Response Variable
survival = cancer_data$Survival
## Explore the shape of the distribution of the response variable
hist(survival,xlab=" ", ylab = "Number of Survival Days",main=" ",nclass=15)
```

Here is the histogram of the survival response data:



What we see here is that the response variable has a skewed distribution and that's an indication that we do not have the normality assumption as required in ANOVA. We will need to transform the data to normalize it. A common transformation to normalize when there is a strong skewness is to use a **log() transformation**:

```
## Transform due to skewness of the distribution
hist(log(survival),xlab=" ", ylab = "Number of Survival Days",main=" ",nclass=15)
```



Now you can see that we don't have that long tail on the right anymore and the distribution looks a lot more symmetric.

We're not looking for a perfect normal, symmetric distribution here. In fact, if you remember the data come from multiple normal distributions. What we should expect is to see a multi-mode distribution where the modes correspond to each individual group. What we see here is that we do see two modes in this distribution. But the shape of the distribution is symmetric.

For the ANOVA, we're going to use the (log) transformed data, not the initial data. For the ANOVA, we need two variables. One is going to be the response variable in this case it's going to be log of the survival number of days and the second variable is going to be the label. Remember the response variable consists of all the stacked response data and now we have to tell R which response corresponds to what category. And for that we'll create this vector of labels. In this case, we take the label, our cancer type, extracted from the data matrix cancer\_data. The column is organ, and we separate it out by using the dollar sign.

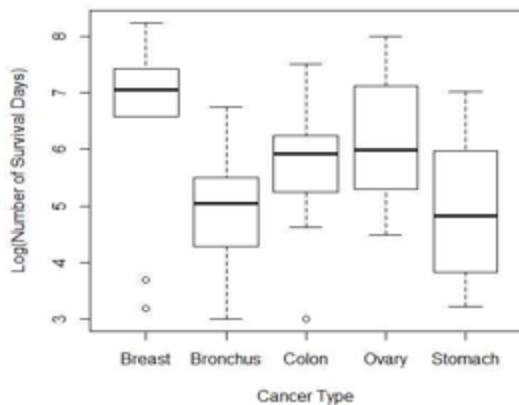
However, we need to tell R that this is a categorical variable. And we need to do so by transforming it using the **as.factor() command**. By doing so, R will know that this is a categorical variable and when you apply the side-by-side boxplot R will plot the survival for each category separately. Remember again you need to provide a label and for both the X and Y axis:

```

## Need to specify Response & Categorical Variables
survival = log(survival)
cancertype = cancer_data$Organ
## Convert into categorical variable in R
cancertype = as.factor(cancertype)
## Explore relationship visually
boxplot(survival~cancertype, xlab = "Cancer Type", ylab = "Log(Number of Survival Days)")

```

This is the side by side boxplot of the log of survival data versus the cancer type:



So you quickly see here that there are differences in the means between the groups. And particularly for example if you compare those with breast cancer versus other cancers, the number of survival days will be much larger for those NOT with breast cancer compared to breast cancer.

We can see that the variability within each group is slightly different. But that may be also because some groups have more data than others. For example, the group of patients that have ovary cancer has a very small sample size, and thus will have much larger variability also.

**When we study the side-by-side boxplot, we compare the within-variability to the between-variability.** The within-variability is the variability within the group, each group, and we see that some groups have higher variability than others. The between-variability is the variability of the means between the groups.

And the question we want to address, is the between-variability significantly larger than the within-variability?

## Estimation of Parameters

We'll next perform an **Analysis of Variance** (AOV) on the log survival rate versus the cancer type. The output of the ANOVA (AOV command in R) consists of the ANOVA

table, which provides information about the sum of square of residuals as well as sum of square of treatments, the degrees of freedom, mean sum of squares, the F test for equals means:

```
## ANOVA in R: Is the between-variability significantly larger than within-variability?
model = aov(survival ~ cancertype)
summary(model)
```

Df	Sum Sq	Mean Sq	F value	Pr(>F)
cancertype	4	24.49	6.122	4.286 0.00412 **
Residuals	59	84.27	1.428	
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *.
	0.1 ' '	1		

$$\begin{aligned} SSTR &= 24.49, k-1 = 4 \\ SSE &= 84.27, N-k = 59 \\ F\text{-value} &= 4.286 \\ P\text{-value} &= 0.00412 \end{aligned}$$

The first column in this table are the degrees of freedom for each source of variability, the treatment (cancer type) and the residuals. There are four degrees of freedom for the cancer type, because there are five different cancer types. There are 59 degrees of freedom for the source of variability due to the residuals.

The next column corresponds to the sum of squares. The sum of square treatments is 24.49 and the sum of square of residuals is 84.27.

The F-value is used for the F-test for performing the test for equality of the means. So the F-value is 4.286. And the P-value of the F-test is 0.004.

The other set of output is provided by the **model.tables command** where we input the model that we fitted before the ANOVA model. We need to specify what kind of summaries we want to provide, in this case for the means. This output provides us the overall mean across all values, all the survival rates, the log survival rate, along with the means for individual groups:

```
## Obtain estimated means
model.tables(model, type = "means")
Tables of means
Grand mean
5.555785
cancertype
  Breast Bronchus Colon Ovary Stomach
  6.559  4.953  5.749  6.151  4.968
rep 11.000 17.000 17.000 6.000 13.000
```

$$\begin{aligned} \hat{\mu}_{\text{Breast}} &= 705.6, n_{\text{Breast}} = 11 \\ \hat{\mu}_{\text{Bronchus}} &= 141.6, n_{\text{Bronchus}} = 17 \\ \hat{\mu}_{\text{colon}} &= 313.9, n_{\text{colon}} = 17 \\ \hat{\mu}_{\text{ovary}} &= 469.2, n_{\text{ovary}} = 6 \\ \hat{\mu}_{\text{stomach}} &= 143.7, n_{\text{stomach}} = 13 \end{aligned}$$

For example, the estimating mean for the group that has breast cancer, the average log survival rate is 6.56 and the number of patients in this group is 11. For patients with colon cancer, the average of the log survival rate is 5.75 with a sample of size of 17.

Recall that we are performing ANOVA in order to test whether the means are equal across the groups, across the cancer types:

**Are the means statistically significantly different?**

P-value = 0.0041: Reject the null hypothesis of equal means

**Since we reject the null hypothesis of equal means...we can ask which means are statistically significantly different?**

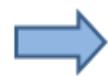
#### PAIRWISE COMPARISON / STATISTICAL INFERENCE

To answer that we perform a pairwise comparison. We'll use the **F command** in R which will provide us the estimated confidence intervals to estimate the difference and means for all possible combinations of the pairs of the means:

```
## Which means are statistically significantly different? Pairwise Comparison
TukeyHSD(model)
Tukey multiple comparisons of means
 95% family-wise confidence level
```

Fit: aov(formula = survival ~ cancertype)

```
$cancertype
    diff      lwr      upr   p adj
Bronchus-Breast -1.6054 -2.9067 -0.3041 0.0083
Colon-Breast     -0.8094 -2.1107  0.4918 0.4119
Ovary-Breast      -0.4079 -2.1147  1.2987 0.9615
Stomach-Breast    -1.5906 -2.9683 -0.2129 0.0158
Colon-Bronchus     0.7959 -0.3575  1.9494 0.3072
Ovary-Bronchus     1.1974 -0.3994  2.7943 0.2296
Stomach-Bronchus   0.0147 -1.2242  1.2537 0.9999
Ovary-Colon        0.4014 -1.1954  1.9984 0.9540
Stomach-Colon      -0.7812 -2.0202  0.4578 0.3981
Stomach-Ovary       -1.1826 -2.8424  0.4770 0.2766
```



**Statistically significant:**  
 $\log(\hat{\mu}_{\text{Bronchus}}) - \log(\hat{\mu}_{\text{Breast}})$   
 $\log(\hat{\mu}_{\text{Stomach}}) - \log(\hat{\mu}_{\text{Breast}})$

We see that there are only two pairs of means that have statistically significantly different means. The first one, when we compare the means of the log of the number survival of days between patients with Bronchus cancer versus those with Breast cancer.

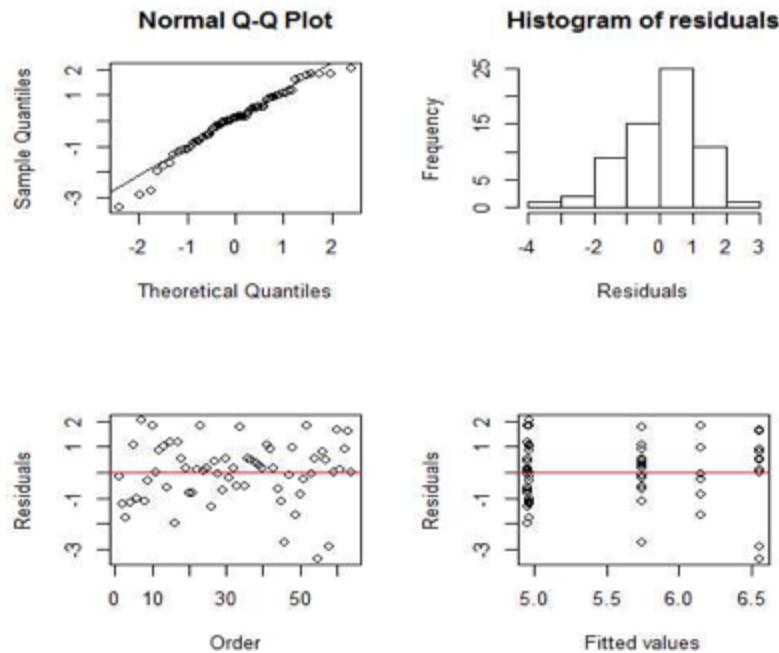
The other one is the pair of means between Stomach cancer and Breast cancer. In both cases, the confidence interval includes only negative values, which means that the log mean of the number of survival days for those with breast cancer is significantly larger than the log mean of the number of survival days for those with Bronchus or Stomach cancer.

All the other pairs are not statistically significantly different, which means that it's possible that the number of survival days across all the other cancers are similar.

## RESIDUAL ANALYSIS

It's important also to evaluate the assumptions (via residual analysis) because otherwise we cannot rely on the statistical inference we made (using the hypothesis testing on the pairwise comparison I presented previously):

```
par(mfrow=c(2,2))
qqnorm(residuals(model))
qqline(residuals(model))
hist(residuals(model),main="Histogram of residuals",xlab="Residuals")
plot(residuals(model),xlab="Order",ylab="Residuals")
abline(0,0,ity=1,col="red")
plot(fitted(model), residuals(model),xlab="Fitted values",
      ylab="Residuals")
abline(0,0,ity=1,col="red")
```



I'll remind you that the three assumptions in ANOVA are the constant variance, independence, and normality. The first two plots, the quantile normal plot and the histogram are used to evaluate normality. The next two plots can be used to evaluate whether the other two assumptions, constant variance, and independence. The normality plot and the histogram looks reasonably well, which is an indication that the distribution of the residuals is approximately normal. Also, we don't see any pattern in the residuals.

In terms of non-constant variance or correlation among the residuals, which means that the two assumptions hold as well.

In conclusion, from this residual analysis:

- the quantiles align on the line and the histogram is approximately symmetrical thus the normality assumption holds
- the residuals are scattered around zero line with no pattern thus both the constant variance and uncorrelated errors hold
- In fact, all three assumptions hold.

Using the Analysis of Variance, we've learned from this example that:

- There is strong evidence for differences in the survival time across the five types of cancer
- The survival time is particularly statistically different for those patients with Breast cancer versus those with Bronchus or Stomach cancer
- Note:this example is from 1978 and those results may not hold today.

# **Unit 3: Multiple Linear Regression**

## **3.1 Basic Concepts and Estimation**

### **1. Objectives and Data Examples**

This unit will regularly return to three examples to illustrate regression concepts.

#### **Example 1: Medical Supply Advertising**

The first example builds on the example from the simple linear regression sessions which studied the relationship between advertisement expenditure and medical supply sales under a new advertising program, assessing whether an increase in advertising expenditure would be expected to lead to an increase in the sales and by how much.

We will now include other predicting variables such as:

- total amount of bonuses paid
- market share in the territory where the company's offices are located
- largest competitor's sales
- the region in which the office is located.

In this example we have both quantitative and qualitative predictive variables. The indicator of the region in which each office is located is a qualitative or categorical variable. It is important to account for this addition of predicting variables in addition to advertisement expenditure since they control for factors that may impact sales.

#### **Example 2: SAT Scores by State**

SAT is a standardized test widely used for college admissions in the United States. But back in 1982, SAT was not widely used for college admission. In this example, the average SAT scores by state for all the states in the United States. The average SAT scores varied considerably by state with mean scores falling between 790 for South Carolina to 1,088 for Iowa. The researchers examined compositional and demographic variables to study to what extent the statistics were tied to SAT scores.

The response variable is the mean SAT score, which is the verbal and quantitative combined. The predicting variables are:

- **X1: "takers."** The percentage of total eligible students, high school seniors, in the state who took the exam.
- **X2: "income."** The median income of families of test takers in hundreds of dollars.
- **X3: "years."** The average number of years that test takers had in social sciences, natural sciences, and humanities combined.
- **X4: "public."** A percentage of test takers who attended public schools.
- **X5: "expend."** A state expenditure on secondary schools in hundreds of dollars per student.
- **X6 "rank."** The median percentile of ranking of test takers.

Research questions to be addressed in these examples are:

- Which variables are associated with SAT scores?
- How do the states rank?
- Which states perform best for the amount of money they spend?

### Example 3: IMDb

The Internet Movie Database, or IMDb, is the most used website to access comprehensive information for movies and TV content. With over 65 million registered users, it remains one of the most trustable sources of data. It allows users to submit ratings and reviews as well as make customized lists. The most popular feature of the site are its movie ratings, and they heavily influence movie-goers deciding whether to watch a particular movie. IMDb is speculated to sway director castings and is becoming a popular metric to compare movie stars.

Analyzing IMDb ratings and seeing what factors will have an impact of predicting the ratings has been a popular problem over the past couple of years. The data set provides many different qualitative and quantitative variables, including:

- |   |  |
|---|--|
| <ul style="list-style-type: none"> <li>● number of votes</li> <li>● Duration</li> <li>● gross earnings</li> <li>● total budget</li> <li>● release year</li> </ul> | <ul style="list-style-type: none"> <li>● Genre</li> <li>● Language</li> <li>● director rating</li> <li>● movie awards</li> </ul> |
|---|--|

... and many more.

## The Uses of Regression Analysis

Typically, a regression analysis is used for following purposes:

- Prediction of the target or response variable,
- Modeling the relationship or association between predicting variables and the response variable,
- Testing hypothesis.

Why restrict ourselves to linear models? Well, they are simpler to understand, and they're simpler mathematically. But most importantly, they work well for a wide range of circumstances. Of course, not all of them. It's a good idea when considering this kind of model, or in fact any statistical model, to remember the words of a famous statistician, George Box: "All models are wrong, but some are useful." We do not believe that a linear model will provide a *true* representation of reality, rather we think that perhaps it provides a *useful* representation of reality.

Another useful piece of advice comes from another very famous statistician, John Tukey. "Embrace your data, not your models."

## 2. Basic Concepts

This unit introduces the model structure of the data in multiple linear regression, along with variations from standard multiple regression including models with interaction and understanding about different roles that variables take.

In multiple linear regression, the data are the response variable given the values of the predicting variables. More specifically, we observe "n" realizations of the response variable along with the corresponding predicting variables:

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

The relationship captured is the linear relationship between the response variable and the predicting variables.

### Assumptions of Multiple Linear Regression

Deviances, or epsilons (also called "error terms") are the difference between the response variable and the linear function of the x's. For multiple linear regression, assume that the deviances have a **zero mean, constant variance**, and are **independent**.

#### Assumptions:

- Linearity/Mean Zero Assumption:  $E(\varepsilon_i) = 0$
- Constant Variance Assumption:  $\text{Var}(\varepsilon_i) = \sigma^2$
- Independence Assumption:  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- (Later we assume  $\varepsilon_i \sim \text{Normal}$ )

For estimation we only need these assumptions. However, for statistical inference we also need to assume that the deviances are normally distributed.

**Zero mean** assumption: the expected value of the errors is zero across all errors, meaning also that the linearity assumption holds.

**Constant variance assumption:** it can not be true that the model is more accurate for some parts of the population and less accurate for other parts. A violation of this assumption means that the estimates are not as efficient as they could be in estimating the true parameters, and would also result in poorly calibrated prediction intervals.

- **Independence assumption:** the response variables are independently drawn from the data-generating process. Violation of this assumption can lead to a misleading assessment of the strength of the regression.

**Normally distributed errors assumption:** If this assumption is violated, hypothesis tests and confidence prediction intervals can be misleading.

### Identifying Model Parameters

In the linear regression model, the parameters defining the regression line, the betas,  $\beta_0, \beta_1$  through  $\beta_p$ , are **parameters**. We also have an additional parameter, the **variance of the deviances**, denoted with sigma squared.

Model parameters are unknown regardless of how much data we observe. But we can derive some

**The model parameters are:  $\beta_0, \beta_1, \dots, \beta_p, \sigma^2$**   
**• Unknown regardless how much data are observed**  
**• Estimated given the model assumptions**  
**• Estimated based on data**

approximations or estimates given the data. The parameter estimates will take different values if one uses different data sets.

In multiple linear regression, the model can be written in the matrix form. We define the **design matrix** as a matrix consisting of columns of predicting variables including the column of ones corresponding to the intercept:

Design Matrix	Response	Error	Coefficients
$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$	$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}$	$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$	$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{pmatrix}$

We'll also stack up all the values of the response variable into a one-column matrix. The same for the parameters the betas and the error terms.

The matrix formulation of the model is:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

## Four Basic Regression Approaches

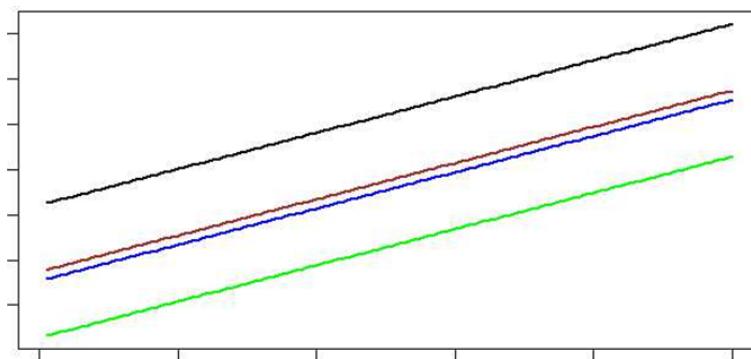
Even with only a small number of variables, there are a number of different approaches to regression that will yield different results, and may be more or less useful in different scenarios. Below are four basic approaches demonstrating the flexibility of linear regression. To keep the examples simple, we will demonstrate each with just two predicting variables, but all can use many more.

### 1. First-Order Model

Let's start by examining a simple "first-order" model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The model gives the equation of a two-dimensional plane or surface, plus some disturbances due to the error. It states that for a fixed value of A, the predicting variable, the expected value of x is a linear function of the other variable. If we graph a regression function as a function of only one variable, say  $x_1$ , for several values of  $x_2$ , we obtain as contours of the regression function a collection of lines:



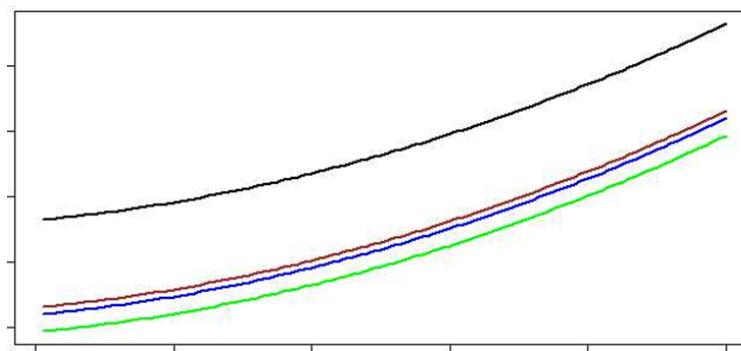
## 2. Second-Order Model

It is often desirable for some predicting variables to be a mathematical function of others in a sense that the resulting model will be much more successful in explaining variation in the response variable than any model without such predictors.

The "linear" in linear regression refers to fitting the response as a linear function of the observed data, but it doesn't necessarily mean that the "linear" is a linear relationship in the individual predictors. In fact, we can extend this model to a Second Order model where we include the square of the predictors, so we include an  $x_1$  squared and an  $x_2$  squared as additional predictors:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

For this model, when we fix  $x_2$ , the expected change in  $y$  for one unit increase in  $x_1$  is not  $\beta_1$ , but  $\beta_1$  plus  $\beta_3 x_1$ . If we graph a regression function as a function of only one variable (say,  $x_1$  for several different values of  $x_2$ ), we obtain as contours of the regression function a collection of curves rather than lines:



Thus, while the estimation of the model is the same as that for a linear model, the interpretation is not.

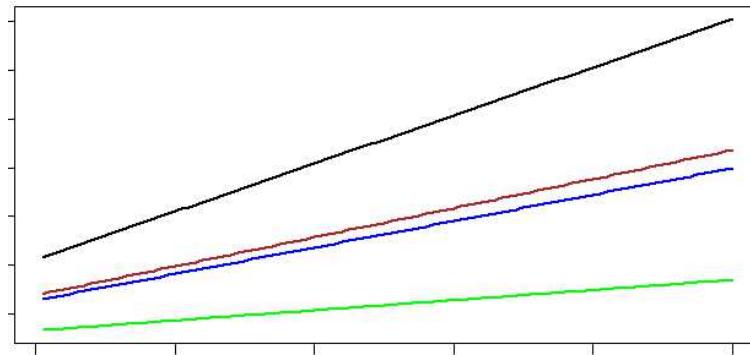
## 3. First-Order Interaction Model

Not only can predictor variables have multiple discrete parameters -- different parameters can also interact with each other. In the third model, First Order *Interaction* model, the contours of the regression function are

### 1<sup>st</sup> Order Interaction Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

non-parallel straight lines for any interaction model. Here, when  $x_1$  is increased by 1, the expected change in Y is  $\beta_1$  plus  $\beta_3$  times  $x_2$  thus depending on  $x_2$ .

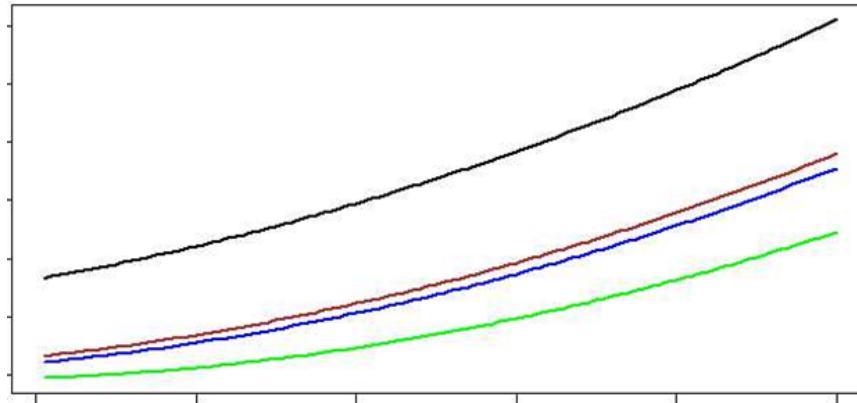


Our final example, the Second Order Interaction Model, combines aspects of these previous examples. Here, the model includes both second-order and interaction terms:

### 2<sup>nd</sup> Order Interaction Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

In this model, a plot of the contours of the regression function yields non-parallel curves:



## Quantitative and Qualitative Variables

Earlier, we contrasted the simple linear regression model with the ANOVA model. In simple linear regression, we consider modeling variation in the response with respect to a *quantitative* variable whereas in ANOVA, we consider modeling variation in the response with respect to one or more *qualitative* variables.

Multiple regression is a generalization of both models.

When we have both quantitative and qualitative variables in a multiple

regression model, we need to understand how to interpret the model and how to model qualitative variables.

Assume a model with both quantitative and qualitative variables where the qualitative variable has three levels. Remember: when we have qualitative variables and regression model with  $k$  levels, we only include  $k-1$  dummy variables if the model has an intercept.

**Simple Linear Regression:** Linear regression with one quantitative predicting variable.

**ANOVA:** Linear regression with one or more qualitative predicting variables.

**Multiple Linear Regression:** Multiple predicting variables which could be quantitative, qualitative, or both.

$$\text{Model: } Y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \epsilon \rightarrow \text{Intercept varies}$$

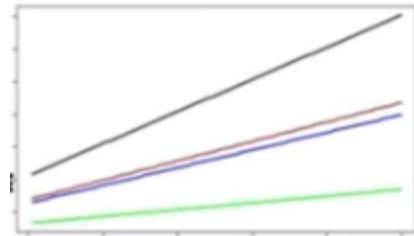
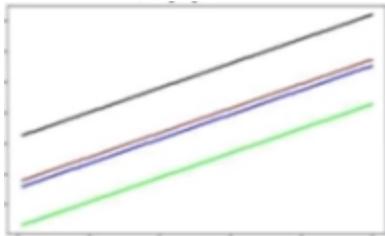
Thus for this example, we include the two dummy variables  $d_1$  and  $d_2$  but not  $d_3$ . The presence of the dummy variable impacts the model in that the intercept will vary depending on the label of the response variable. So for example, for those three models if we have  $d_1$  equal to 0,  $d_2$  equal to 0, that means we consider the response variable for the third category. The model is  $\beta_0$  plus  $\beta_1 x_1$ . The intercept is  $\beta_0$ .

$$\text{Model: } Y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \varepsilon \rightarrow \text{Intercept varies}$$

If  $d_1=0, d_2=0: \beta_0 + \beta_1 x_1$   
If  $d_1=1, d_2=0: \beta_0 + \beta_2 + \beta_1 x_1$   
If  $d_1=0, d_2=1: \beta_0 + \beta_3 + \beta_1 x_1$

**Parallel regression lines**

Now if  $d_1$  is equal to one and the other two are zero, then the intercept is going to be  $\beta_0$  plus  $\beta_2$ . If  $d_2$  is equal to one then the intercept is  $\beta_0$  plus  $\beta_3$ , so what we obtain are parallel regression lines (figure below on left).



As we discussed in the previous slide, if we include an interaction term between the qualitative variables and the quantitative variables, the regression lines are non-parallel as shown in the figure (above on right).

### Multiple Linear Regression Concepts Applied

Let's revisit our examples to see how these concepts can be applied.

#### Example: Advertising and Sales

Our example of the relationship between the sales and advertisement expenditure includes both quantitative predicting variables -- bonuses, the market share, their largest competitors -- and a qualitative variable: the region in which the office is located.

#### Example: IMDb Data

In our IMDb example, we are interested in differentiating between qualitative variable and quantitative variables that impact movie ratings. We have four quantitative predicting variables: number of votes, duration of the movie, total budget in millions and several qualitative predicting variables.

#### Quantitative Predicting Variables:

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999  
 $X_2$  = the total amount of bonuses paid in 1999  
 $X_3$  = the market share in each territory  
 $X_4$  = the largest competitor's sales

#### Qualitative Predicting Variable:

$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

I want to highlight particularly the release year. At first glance, this may seem obviously to be a quantitative variable, and indeed, measures of time certainly can be. But in cases where there are only a few years across many observations (in this case we have five different years of data), it may be better to consider release year as a *qualitative* variable. If the observations are made over many years, then we considering release year as a quantitative variable might be more appropriate.

**Generally, we can transform a quantitative variable into a qualitative, categorical variable,** particularly when we see that there are non-linear relationships of that predictive variable with respect to the response.

### 3. Estimation Method

The topic of this lesson is parameter estimation. We'll overview the approach for estimating the regression coefficients and also the variance parameter of the error terms, along with the statistical properties of the estimator for the variance.

#### Parameter Estimation

Finding our parameters using the **Parameter Estimation approach** is similar to the approach we learned in estimating the parameter for a linear regression model with a single predicting variable. We again want to find the estimates for our parameters that minimize the sum of least squares. Here, the least squares are the square differences between the observed responses  $y_i$  and the expected responses, which are  $\beta_0$  plus  $\beta_1x_1 + \dots + \beta_px_p$ .

We then square the difference, and add up across all possible observations:

We can write this, the least sum of squares of error, into a matrix format as  $Y$ , which is a stacked vector of responses, minus  $X$  times  $\beta$  ( $X$  is a design matrix,  $\beta$  is a stacked vector of parameters) transpose times  $Y - X\beta$ :

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ , we find values that minimize squared error:

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1x_{i1} + \dots + \beta_px_{ip}))^2 = (Y - X\beta)^T(Y - X\beta)$$

If we use linear algebra to minimize this sum of least squares error, we can obtain the equality of estimating equations:

$$X^T X \hat{\beta} = X^T Y$$

In order to solve this for beta, we need to assume that  $X^T X$  is invertible because we need to multiply with the inverse of this matrix on both the right and the left-hand sides of the estimating questions. And if this matrix is invertible, then the estimator is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

The fitted values are the differences between  $\hat{y}$  and the fitted. The fitted here are  $X$  times  $\hat{\beta}$ . Again  $X$  is the design matrix.  $\hat{\beta}$  is the vector of all the estimated

regression coefficients. We can plug in for  $\beta$  hat and we get what we call  $H * Y$  where  $H$  is what we call the **hat matrix**. For residuals, we take the difference between observed and fitted:

Observed values are  $Y$  and fitted are  $X$  transpose  $\beta$  hat. We can rewrite this as  $I - H$  ( $I$  is the identity matrix,  $H$  again is the hat matrix) \*  $Y$  :

**The fitted values are:**  $\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$   
**Where  $H$  is called the hat matrix.**

The estimated variance is now the sum of squared errors divided by  $n-p-1$ .

So this is very similar estimator as for simple regression except that now we're using  $n-p-1$  on the denominator rather than  $n-2$ :

$$\hat{\epsilon} = Y - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{H})Y \rightarrow \hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-p-1}$$

Assuming that the error terms are normally distributed, the sampling distribution of the estimated variance, or so-called mean square error, is a chi-squared distribution with  $n-p-1$  degrees of freedom:

**The estimator of  $\sigma^2$  is MSE:**

**•Assuming  $\varepsilon_1, \dots, \varepsilon_n$  are normally distributed, then**  
**MSE  $\sim \chi^2$  with  $n-p-1$  degrees of freedom (Why  $n-p-1$ ?)**

Why  $n-p-1$ ?

So let's assume again that the error terms are normally distributed. We do not have the error terms because we don't know the betas. But we can replace them with the residuals, the epsilon hats, which are also going to be normally distributed. And thus we can use a sample variance of the residuals in order to estimate sigma squared:

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-p-1} \sim \chi^2_{n-p-1}$$

(chi-squared distribution with  $n-p-1$  degrees of freedom)

Assuming  $\hat{\varepsilon}_i \sim \varepsilon_i \sim N(0, \sigma^2)$

↑  
Estimating  $\sigma^2$  — Sample variance

But for the sample variance we actually use  $n-p-1$  degrees of freedom because, when we replace the error terms with the residuals, we replaced  $p+1$  coefficients parameters -- we replaced  $\beta_0$  with  $\hat{\beta}_0$ ,  $\beta_1$  with  $\hat{\beta}_1$ , and so on. So we lose now  $p+1$  degrees of freedom:

<p>Recall that <math>\varepsilon_i = (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))</math></p> <p>Replaced by <math>\hat{\varepsilon}_i = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}))</math></p>	<p>Use <math>p+1</math> degrees of freedom because</p> <p><math>\beta_0 \leftarrow \hat{\beta}_0</math>  <math>\beta_1 \leftarrow \hat{\beta}_1</math>  <math>\dots</math>  <math>\beta_p \leftarrow \hat{\beta}_p</math></p>
---	---

Thus, **the sampling distribution of the variance is Chi-square with  $n-p-1$  degrees of freedom.** And this is so called sampling distribution of the estimated variance:

Thus, assuming that

$$\rightarrow \hat{\sigma}^2 = \text{MSE} \sim \chi^2_{n-p-1}$$

(This is called the sampling distribution of  $\hat{\sigma}^2$ )

## 4. Model Interpretation

The topic of this lesson is model interpretation for multiple linear regression.

Particularly, I'm going to touch upon several aspects of model interpretation including interpretation of the regression coefficients and aspects related to differences in terms of causality versus association, as well as different roles that predicting variables can take.

The interpretation of the estimated regression coefficients in multiple linear regression is similar to that in simple linear regression, except that we need to consider multiple predicting variables in the model. As before, the estimated intercept is an estimate of the expected value of the response variable when the predictors equal zero. Only now there are several predictors, the estimated value for  $\beta_i$  (through)  $\beta_j$ , **representing the estimated expected change in y associated with one unit of change in the corresponding predicting variable, holding all else in the model fixed.**

**Thus now, we need to specify that there are other predictive variables in the model held fixed while we vary one of the predictors.** This is a very important aspect of the interpretation of the regression coefficients as I'll explain next.

Modeling the relationship of a predicting variable to a response variable can be done with a simple linear regression model as we did so far, such as when we fit the relationship between advertising expenditure and sales. But also it can be done in a multiple regression context when we add other predicting variables in the model. Thus we differentiate between so-called marginal and conditional models.

The **marginal model**, or **simple linear regression**, captures the association of one predicting variable to the response variable marginally, that means without consideration of other factors.

The **conditional or multiple linear regression model** captures the association of a predictor variable to the response variable, conditional on other predicting variables in the model.

Importantly, the estimated regression coefficients for the conditional and marginal relationships can be different, not only in magnitude but also in sign or direction of the relationship. Thus, the two models used to capture the relationship between a predicting variable and a response variable will provide different estimates of the relationship.

## When to Use Multiple Linear Regression

But why do we need multiple linear regression when we can use simple linear regression? Often, the relationship between a response and predicting variable is dependent on other factors, and cannot be singled out to be estimated using a simple linear regression. Multiple linear regression allows for quantifying the relationship of a predicting variable to a response when other factors vary.

## Causality Versus Association

One of the dangers of using multiple linear regression without much knowledge of fundamentals about regression is the interpretation of the results from the regression. This is particularly prevalent in a context of making causal statements when the setup of the regression does not allow so. Causality statements can only be made in a controlled environment such as randomized trials or experiments. In experimental situations, analysts can change the setting of one particular factor in the environment, holding others fixed thereby isolating its effect.

But such isolation is not possible with observational data. Most of the data to which you will apply regression analysis will likely come from observational studies which generate data without the ability to control biases and correlations among the observations. Multiple regression provides a statistical version of this practice through its ability to statistically represent a conditional action that would otherwise be impossible.

However, interpretation of relationships under multiple regression need to be fully considered as part of the entire multiple regression model.

### Example: SAT Scores and College GPAs

Let's look at a very specific example. Say we take a sample of college students and determine their college GPA as well as their high school GPA and their SAT score. We then build a model of college GPA as a function of high school GPA and SAT:

$$\text{COLGPA} = 1.3 + 0.7 \text{ HSGPA} - 0.0003 \text{ SAT}$$

It is tempting to say that the coefficient for SAT must have the wrong sign, because it seems to say

**Incorrect Interpretation:**  
Higher values of SAT are associated with lower values of college GPA.

**Correct Interpretation:** higher values of SAT are associated with lower values of college GPA, *holding high school GPA fixed.*

The coefficients of a multiple regression must not be interpreted marginally!

that higher values of SAT are associated with lower values of college GPA. What it says is that higher values of SAT are associated with lower values of college GPA, *on the condition that high school GPA is held fixed*. High school GPA and SAT are absolutely correlated with each other. So changing SAT by one unit, holding high school GPA may not actually even happen, it may not be possible. **The coefficient of multiple regression thus must be interpreted in the context of other predictors in the model. So we do not want to interpret them marginally.**

If you really are interested in the relationship between college GPA and SAT only, you should perform a simple linear regression of college GPA on SAT. However, that will provide an incomplete understanding of these relationships since high school GPA is an important factor in explaining the variability in college GPA.

This simple example illustrates the fact that we cannot make direct or causal statements about how SAT impacts college GPA. We can only say that there is an associative relationship. But we need to be careful in interpreting the regression coefficients when there are other predictive factors in a model that are correlated to SAT, such as high school GPA.

### Roles of Predicting Variables: Controlling, Explanatory, Predictive

So more explicitly, multiple linear regression allows including variables to explain the variability in the response variable, taking different roles. Particularly, I differentiate factors into controlling, explanatory, or predictive factors.

**Controlling variables** can be used to control for bias selection in a sample. They're used as default variables to capture more meaningful relationships. They are used in regression for observational studies, for example, when there are known sources of bias selection in the sample data. They are not necessarily of direct interest, but once a researcher identifies biases in the sample, he or she will need to correct for those, and will do so through controlling variables.

**Explanatory variables** can be used to explain variability in the response variable. They may be included in the model even if other similar variables are in the model.

**Predictive variables** can be used to best predict variability in the response regardless of their explanatory power. Thus when selecting explanatory variables, the objective is to explain the variability in the response. Whereas when selecting predictive variables, the objective is to predict the response.

As we'll learn in the model selection lectures later in this course, the selective variables depending on the objective, whether predictive or explanatory, could be different.

## 5. Estimation Data Examples

*The topic of this lesson is parameter estimation with data example. We'll learn using R, how to implement a multiple linear regression model in R, and how to interpret it.*

### Example 1: Advertisement Expenditure and Sales

Let's return to the example of the relationship between advertisement expenditure and sales. And again, this is a set of predictive variables divided into quantitative and qualitative predicting variables.

This is a set of questions we may be interested to address in such example. We want to:

- A. Fit a linear regression with all predictors and estimate the regression coefficients.  
What are the estimated regression coefficients and the estimated regression line?
- B. Interpret and compare the estimated coefficients from the conditional model versus the marginal model. Having analyzed the estimated regression coefficient for the advertisement expenditure variable under simple regression, we want examine it under the conditional and the marginal models.
- C. See how the predictions of those two models will differ, and which of the predictions are more meaningful.
  - a. What does the model predict as the advertisement expenditure increases for an additional \$1,000 using the full regression model?
  - b. Is the prediction different when compared to the prediction from the simple linear model with just the advertisement expenditure variable?
- D. Compare the estimated error variance under the conditional versus the marginal model.
  - a. What is the estimate of the error variance?
  - b. Is it different from the simple linear regression model? Why?

### Using R for Multiple Linear Regression

Let's use R to see what we can determine about our data. First, we will read in the data:

```
meddcor = read.table("meddcor.txt", sep="", header = FALSE)
```

Next, because this data set does not have a header, we provide the names of the columns in the data based on what each column represents.

```
colnames(meddcor) = c("sales", "advertising", "bonuses",
"marketshare", "largestcomp", "region")
```

We will then use "**as.factor()**" to convert the column corresponding to the region into a categorical variable in order to tell R that this is not a quantitative but a qualitative variable:

```
meddcor$region =
as.factor(meddcor$region)
```

The R command to create the model is **lm()**, the same command as used for simple linear regression. Again, on the left we use the response variable, and on the right we use the predictors. This time, however, instead of specifying each variable explicitly, I put a dot:

```
model = lm(sales ~ ., data = meddcor)
```

This tells the **lm()** function to take all of the columns (except the response) as being predictors. Now, all of the columns in the data will be considered as predicting variables. Next, we can use **summary()** to view information about our model:

summary(model)					
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	117.0200	192.9732	0.606	0.5518	
advertising	1.4092	0.2687	5.244	5.49e-05	
bonuses	1.0123	0.4641	2.181	0.0427	
marketshare	3.1548	2.9802	1.059	0.3038	
largestcomp	-0.2354	0.2338	-1.007	0.3275	
region2	53.6285	34.7359	1.544	0.1400	
region3	267.9569	47.5577	5.634	2.40e-05 ***	
---					
Residual standard error: 55.57 on 18 degrees of freedom					

The data in the blue rectangle above are the estimated regression coefficients. The estimated  $\beta$  coefficient for the advertising expenditure in the blue circle is 1.4092. We would interpret this number to be the expected additional gain in sales, in thousands of dollars for each additional \$100 expenditure in advertisement, *while holding all other*

### Recall:

#### Quantitative Predicting Variables:

- $X_1$  = the amount (in hundreds of dollars) spent on advertising
- $X_2$  = the total amount of bonuses paid
- $X_3$  = the market share in each territory
- $X_4$  = the largest competitor's sales

#### Qualitative Predicting Variable:

- $X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

*predictors fixed.* A reminder here that while the units for sales is thousands, the units for advertising expenditure is \$100, so be careful in the interpretation.

If you contrast this with the marginal model, recall the estimated regression coefficient was 2.772, significantly larger than the estimated coefficient on the conditional model (1.4092).

Comparing the predictions of these two approaches, the conditional model predicts an additional \$1,000 in advertising expenditure will yield \$14,000 in additional sales, while the marginal model predicts a much larger \$27,700 revenue increase for the same increase in advertising expenditure.

**b. Conditional model:**

$$\hat{\beta}_{adv} = 1.4092$$

The expected additional gain in sales in thousands for \$100 additional expenditure in advertisement **while holding all other fixed.**

**Marginal model:**

$$\hat{\beta}_{adv} = 2.772$$

The expected additional gain in sales in thousands for \$100 additional expenditure in advertisement **not accounting for other predicting variables.**

So which one is more meaningful? Because sales vary with other factors, the interpretation based on a multiple regression is more meaningful. :

- c. An additional **\$1,000** in advertising expenditures results in **\$14,092** additional sales under full model and **\$27,720** additional sales under simple linear model.

**Which is more meaningful?** Because sales vary with other factors, the interpretation based on the multiple regression is more meaningful.

- D. Under the full model, the estimated variance is 55.57<sup>2</sup>. This value, the estimated variance, which appears in the summary output (above), is the squared residual standard error. Under the simple linear model, the variance estimate was 101.4 squared.

The variance under the full model is thus much smaller than the estimated variance on the model with one predictor. This is because, when we include multiple variables in a

model, the model better explains variability in the response as compared to the model when we include only one variable.

### R Code Used In This Example

Above, we used the following code to build and analyze our model:

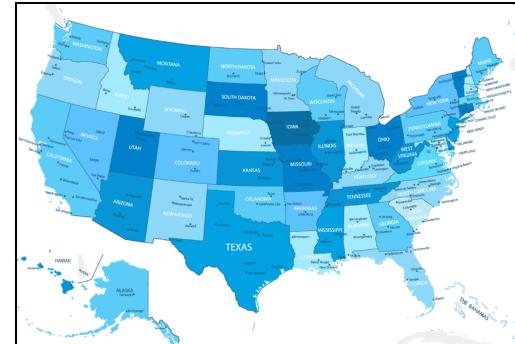
```
meddcor = read.table("meddcor.txt", sep = "", header = FALSE)
colnames(meddcor) = c("sales", "advertising", "bonuses", "marketshare",
"largestcomp", "region")
meddcor$region = as.factor(meddcor$region)
model = lm(sales ~ ., data = meddcor)
summary(model)
```

### Example 2: SAT Scores By State

In our second example, we will examine compositional and demographic variables to determine to what extent these characteristics impact SAT scores.

#### Explanatory and Controlling Variables

All of the variables in the study are used as **explanatory factors** except for two which are used as **controlling variables**. If we would like to rank states by mean response (the mean SAT score), we'll need to first control for these two factors. Moreover, if we want to study the impact of the other explanatory factors, again, we need to control for these two factors.



#### Controlling factors:

$X_1$  = % of total eligible students in the state who took the exam

$X_6$  = median percentile of ranking of test takers within their secondary school classes

#### Explanatory Factors:

$X_2$  = median income of families of test takers, in hundreds of dollars

$X_3$  = average number of years that test takers had in social sciences, natural sciences, and humanities

$X_4$  = % of test takers who attended public schools

$X_5$  = state expenditure on secondary schools, in hundreds of dollars per student

The first controlling variable is **ranking**, the rank of those students taking SAT to control for this bias across all the states. The second controlling variable is **takers**. Doubters of the study from which this data is derived noted that states with high average SAT scores had low percentage of students taking the exam.

One reason: Midwest states used to administer different tests to students going to college

in-state. Only their best students planning to attend out-of-state colleges took the SAT. As a percentage of takers increase for other state, so does the likelihood that the takers include the lower qualified students.

## Lecture 3.1 Knowledge Check

*Answers in the footnote below.*

1. The objective of multiple linear regression is:
  - A. To predict future new responses
  - B. To model the association of explanatory variables to a response variable accounting for controlling factors.
  - C. To test hypotheses using statistical inference on the model.
  - D. All of the above.
2. Which is correct?
  - A. A multiple linear regression model with  $p$  predicting variables but no intercept has  $p$  model parameters.
  - B. The interpretation of the regression coefficients is the same whether or not interaction terms are included in the model.
  - C. Multiple linear regression is a general model encompassing both ANOVA and simple linear regression.
  - D. None of the above.
3. Which is correct?
  - A. The regression coefficients can be estimated only if the predicting variables are not linearly dependent.
  - B. The estimated regression coefficient  $\hat{\beta}_i$  is interpreted as the change in the response variable associated with one unit of change in the  $i$ -th predicting variable.
  - C. The estimated regression coefficients will be the same under marginal and conditional model; only their interpretation is not.
  - D. Causality is the same as association in interpreting the relationship between the response and predicting variables.
4. Which one correctly characterizes the sampling distribution of the estimated variance?
  - A. The estimated variance of the error term has a chi-squared distribution regardless of the distribution assumption of the error terms.

- B. The number of degrees of freedom for the chi-squared distribution of the estimated variance is  $n - p - 1$  for a model without an intercept.
- C. The sampling distribution of the mean squared error is different from that of the estimated variance.
- D. None of the above.<sup>3</sup>

---

<sup>3</sup>

1: D.

2: C.

3: A.

4:D.

## 3.2 Statistical Inference and Prediction

### 1. Statistical Inference

*The topic of this lesson is inference for regression parameters on the multiple linear regression. What we learn in this lesson is the statistical properties of the estimated regression coefficients, along with procedures on how to estimate confidence intervals and also to perform hypothesis testing for the regression coefficients.*

#### Properties of Regression Estimators

Let's begin first with the properties of the regression estimators. Just like in a simple linear regression, the expected values of the estimators are equal to the true parameters.

The variance of the estimated regression coefficients is:

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \Sigma$$

Recall that X is the **design matrix**.

So what we see here is that the variance of  $\beta$  hat is now a variance. Remember:  $\beta$  hat is a vector of estimated regression coefficients. If we have 10 predictors, the vector will have a length of 11.

Because this is a vector, the variance is a matrix. On the diagonal we'll include the variances of the coefficients, and on the off-diagonal will be the covariances between the estimated regression coefficients:

# Covariance Matrix

- Representing Covariance between dimensions as a matrix e.g. for 3 dimensions:

$$C = \begin{pmatrix} \text{cov}(x,x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(y,x) & \text{cov}(y,y) & \text{cov}(y,z) \\ \text{cov}(z,x) & \text{cov}(z,y) & \text{cov}(z,z) \end{pmatrix}$$

**Variances**

- Diagonal is the **variances** of x, y and z
- $\text{cov}(x,y) = \text{cov}(y,x)$  hence matrix is **symmetrical** about the diagonal
- N-dimensional data will result in **NxN covariance matrix**

lest

Furthermore,  $\hat{\beta}$  is a linear combination of the  $y_i$ s. Assuming the error terms are normal (i.e., the response variables are normally distributed), then  $\hat{\beta}$  is also distributed normally, with  $\beta$  as the mean and  $\Sigma$  as the covariance matrix:

$$\hat{\beta} \sim N(\beta, \Sigma)$$

In short, the estimator  $\hat{\beta}$  is unbiased for  $\beta$  because the expectation of the estimator is equal to the true parameter.

Also, distribution of the  $\hat{\beta}$  is normal. But the covariance matrix depends on  $\sigma^2$ , which we do not know. What should we do?

## Estimating Sigma Square

We can replace sigma square with the mean square error, which is equal to the sum of squared of residuals divided by n- p- 1:

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_j^2}{n-p-1} \sim \chi_{n-p-1}^2$$

Now, when we deploy sigma square with estimator, the **sampling distribution** for individual  $\beta$  hat is a t-distribution with n- p - 1 degrees of freedom where n - p- 1 comes from the degrees of freedom of the variance estimator:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{se(\hat{\beta}_j)}} \sim t_{n-p-1}$$

## Confidence Interval Estimation

Now we can derive confidence intervals for  $\beta_j$  using this t sampling distribution:

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p-1} se(\hat{\beta}_j)$$

We use this confidence interval to answer whether  $\beta_j$  is statistically significant by checking whether 0 is in a confidence interval. And if it is not, we conclude that  $\beta_j$  is statistically significant.

But why is this a t-interval? As shown above, the sampling distribution for the individual regression coefficients is the t-distribution with n- p- 1 degrees of freedom:

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{se(\hat{\beta}_j)}} \sim t_{n-p-1} \rightarrow t\text{-interval for } \beta_j$$

To build a  $1 - \alpha$  confidence interval, then, I center it at  $\hat{\beta}_j$   $\pm$  the t critical point. Again, I'm using here  $n-p-1$  because the sampling distribution is  $T_{n-p-1}$  multiplied by the standard deviation:

$$\text{Confidence interval } 1-\alpha \rightarrow \frac{\hat{\beta}_j}{\text{Estimate of } \hat{\beta}_j} \pm \frac{t_{\alpha/2, n-p-1}}{\text{t-critical point}} \sqrt{\text{se}(\hat{\beta}_j)}$$

Standard Deviation of  $\hat{\beta}_j$

### Testing Statistical Significance

We can use statistical inference based on hypothesis testing to test for statistical significance of individual  $\beta_j$ , again using the t-test. Importantly, this test measures the statistical significance of  $\beta_j$  *given all other predicting variables in the model* and not in isolation.

Here, the null hypothesis is that the coefficient is 0 versus the alternative hypothesis that it is not:

$$H_0: \beta_j = 0 \text{ vs. } H_a: \beta_j \neq 0$$

Similar to simple linear regression, the t-value is the estimator minus 0 divided by the standard error:

$$t\text{-value} = \frac{\hat{\beta}_j - 0}{\text{se}(\hat{\beta}_j)}$$

If this t-value is large, we reject the null hypothesis and conclude that the coefficient is statistically significant.

How will the procedure change if we test whether the coefficient is equal to a constant? Again, the t-value looks very similar, but we replace 0 with the new value b:

$$H_0: \beta_j = b \text{ vs. } H_a: \beta_j \neq b$$

We reject the null hypothesis if the t-value is larger than  $t_{\alpha/2, n-p-1}$ , [which is to say], when the absolute value of t-value, is larger than the critical point:

For significance level  $\alpha$ , Reject if  $|t\text{-value}| > t_{\frac{\alpha}{2}, n-p-1}$

Alternatively, we can compute the p-value:

$$\text{P-value} = 2P(T_{n-p-1} > |t\text{-value}|)$$

If the p-value is small, for example smaller than .01, we reject the null hypothesis that  $\beta_0$  is equal to 0:

### Testing for Statistical Positivity or Negativity

How does the procedure change if we test whether the coefficient is statistically positive or statistically negative? If we want to test for positive relationship, then the p-value is the probability of the t-distribution with  $n-p-1$  greater than t-value. The key differences are that we are not using the absolute value, and we're not using the two in front:

#### Test for positive relationship:

$$H_0: \beta_j \leq 0 \text{ versus } H_A: \beta_j > 0?$$

$$\text{P-value} = P(T_{n-p-1} > t\text{-value})$$

If we want to test the negative relationships, the p-values is the probability of the left tail of the t-distribution, when it's smaller than the t-value. Again, we no longer use an absolute value or double probability:

#### Test for negative relationship:

$$H_0: \beta_j \geq 0 \text{ versus } H_A: \beta_j < 0?$$

$$\text{P-value} = P(T_{n-p-1} < t\text{-value})$$

## 2. Testing for Subsets of Coefficients

We're going to first learn about the hypothesis testing procedure for overall regression. And then we extend that procedure to a test where we're interested on the SUBSET of multiple regression coefficients.

Testing Overall Regression n-p-1

To perform a test for overall regression, we'll use **Analysis of Variance for multiple regression**. This is similar to what we learned in Analysis of Variance model. Here we divide the variability in the response variable between the variability due to the regression and the variability due to the errors.

Just like the ANOVA table here, we have multiple columns corresponding respectively to the degrees of freedom, the sum of squares, and the mean sum of square and f statistic:

Source	DF	Sum of Sq	Mean SS	F-statistic
Regression	p	SSReg	SSReg/p	MSSReg/MSE
Residual	n-p-1	SSE	SSE/n-p-1	
Total	n-1	SST		

$$\text{Where } \text{SSReg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \text{ and } \text{SST} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

- **DF Regression:** The degrees of freedom for the variability due to the regression is equal to P, where P is the number of predictive variables.
- **DF Residual:** The degrees of freedom corresponding to source of variability due to the residuals or error is n-p-1.
- **DF Total:** And the total degrees of freedom is the sum across those two, which is n-1.
- **The sum of squared regression:** the sum of the square differences between the fitted values and the average across all the responses.
- **Sum of squared total:** the sum of all the square differences between observations and the average.
- And **sum of squared residuals** is the sum of squared error divided by (n - p - 1).

We will use analysis of variance (ANOVA) to test the hypothesis that the regression coefficients are zero:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

We are only interested in the regression coefficients corresponding to the predictors, not on the intercept. The alternative hypothesis is that at least one of the regression coefficients is not equal to zero, meaning that at least one of the predictors included in the model has predictive power.

To perform this test using ANOVA, we can use the F-test just like the analysis of variance model. The f statistic is going to be the ratio between the mean sum of square regression and mean sum of square of error. We reject the null hypothesis if the F-statistic is larger than the critical point, with alpha being the significance level of the test. This means, again, that at least one of the coefficients is different from 0 at the alpha significance level.

We can also use the p-value, which is the probability of the left tail of the F-distribution, the evaluated F-statistic, and if this p-value is small, then we reject again the null hypothesis:

### Testing Subsets of Coefficients

When we test for subsets of coefficients, we will decompose the variability in smaller portions. First, we decompose the sum of square total as the sum of the square regression of the full model plus the sum of square *error* of the full model:

$$SST(X_1, X_2, \dots, X_p) = SSReg(X_1, X_2, \dots, X_p) + SSE(X_1, X_2, \dots, X_p)$$

In fact, we can decompose the sum of square regression into multiple small parts:

$$\text{SSReg}(X_1, X_2, \dots, X_p) = \text{SSReg}(X_1) + \text{SSReg}(X_2|X_1) + \text{SSReg}(X_3|X_1, X_2)$$

$$+ \dots + \text{SSReg}(X_p|X_1, \dots, X_{p-1})$$

$\text{SSReg}(X_1)$  = sum of squares explained by using only  $X_1$  to predict Y

$\text{SSReg}(X_2|X_1)$  = **extra sum of squares** explained by using  $X_2$  in addition to  $X_1$  to predict Y

$\text{SSReg}(X_3|X_1, X_2)$  = **extra sum of squares** explained by using  $X_3$  in addition to  $X_1$  and  $X_2$  to predict Y

$\text{SSReg}(X_p|X_1, \dots, X_{p-1})$  = **extra sum of squares** explained by using  $X_p$  in addition to  $X_1, X_2 \dots X_{p-1}$  to predict Y

First, ( $\text{SSReg}(X_1)$ ) is going to be the sum of square regression due to  $X_1$ , which for the regression explained by using  $X_1$  to predict Y.

The next, ( $\text{SSReg}(X_2|X_1)$ ) is the extra sum of squares explained by using  $X_2$  in addition to  $X_1$  to predict Y which means that we already have  $X_1$  in the model, now we adding  $X_2$  to the model and this is the extra sum of square for adding  $X_2$  to the model.

The next part, ( $\text{SSReg}(X_3|X_1, X_2)$ ) is the extra sum of square explained by using  $X_3$  in addition to  $X_1, X_2$  to predict Y.

The last predictor,  $\text{SSReg}(X_p|X_1, \dots, X_{p-1})$ ) is the extra sum of squares explained by using  $X_p$  in addition to all the other predictors in a model to predict Y.

You can see here that it's important to take into account the order, we first add the  $X_1$ , then  $X_2, X_3$ , and so on.

Let's see how we can use to test for subset of coefficients. For example, if we want to address the question whether  $X_1$  alone significantly aids in predicting Y, we can compare the sum of square regression of the model including  $X_1$  versus the sum of square of error of the model including  $X_1$ :

**SSReg( $X_1$ ) vs. SSE( $X_1$ ): Does  $X_1$  alone significantly aid in predicting Y?**

If we want to answer whether the addition of  $X_2$  significantly contributes to the prediction of Y after we account (or control) for the contribution of  $X_1$ , we compare the sum of square regression, which is the additional sum of square of regression

generated by adding  $X_2$  to the model that already has  $X_1$  versus the sum of square error of the model including both:

**SSReg( $X_2|X_1$ ) vs SSE( $X_1, X_2$ ): Does the addition of  $X_2$  significantly contribute to the prediction of Y after we account (or control) for the contribution of  $X_1$ ?**

We would use the same method to determine whether the addition of  $X_3$  contributes to the prediction of Y after we control for the contribution of  $X_1$  and  $X_2$ , and so on for  $X_p$  variables we might want to add to the model:

**SSReg( $X_p|X_1, \dots, X_{p-1}$ ) vs. SSE( $X_1, X_2, \dots, X_p$ ): Does the addition of  $X_p$  significantly contribute to the prediction of Y after we account (or control) for the contribution of  $X_1, \dots, X_{p-1}$ ?**

### Generalizing the Test for Coefficients

We can also use this idea more generally. Consider this full model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_q z_q + e$$

We can divide the predictors in two groups, Xs, and Z's. We have  $\beta$  coefficients for X, and we have alpha coefficients for Z's. For example, the X's can be controlling factors, and Z's can be additional explanatory factors.

### PARTIAL F-TEST

And we may want to test the null hypothesis that all the alpha coefficients corresponding to the z predictors are zero:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_q = 0 \text{ versus } H_A : \text{at least one is not zero}$$

For this, we can use what we call the **Partial F-test**, which means that we're going to compare the sum of square regression (the extra sum of square regression by adding the Z predictor to the model that already has the X's) versus the sum of square of error for the full model.

Remember, for the F-statistic, we divide the mean sum of square regression by the mean sum of square of error. Here we can compare this statistic with the critical point of the F distribution with P and  $n-p-1$  degrees of freedom. (Remember that P here is the number of predictors that are already in the model.)

**Partial F-test:**  $F_{partial} = \frac{SSReg(Z_1, \dots, Z_q | X_1, \dots, X_p) / q}{SSE(Z_1, \dots, Z_q, X_1, \dots, X_p) / (n-p-q-1)}$

We reject  $H_0$  if F-statistic is large ( $F\text{-statistic} > F_{\alpha, q, n-p-q-1}$ ); Which means that at least one of the coefficients is different from zero at the  $\alpha$  significant level.

### 3. Statistical Inference Data Examples

The topic of this lesson is the implementation of statistical inference with a data example. And I'm going to illustrate that example using the R statistical software.

We're going to use the second example where we're interested in the variation in the mean SAT score. And we'll learn how to include controlling factors and explanatory factors in the model.

This is a set of questions we may want to address in statistical inference:

- A. What is the estimate of the coefficient of  $\beta_1$  or other coefficients and its variance? What is its sampling distribution?
- B. Is this coefficient  $\beta_1$  statistically significant? (We can draw this conclusion based on a P-value.)
- C. What is the F-statistic for the overall regression? Do we reject the null hypothesis that all regression coefficients are zero (the test for overall regression)?
- D. Obtain the 99% confidence interval for  $\beta_1$
- E. Given  $X_1$  and  $X_6$  are controlling factors, test the null hypothesis that the coefficients of the rest of predictors are zero (will adding exploratory factors will have a predictive power and will improve the predictive power of the regression model). Clearly state the hypothesis test. Show how you perform the test.  
Interpret the results.

As always, first we must read the data into R:

```
data = read.table("SATData.txt", header = TRUE)
```

This data has a header, meaning the columns already have names, so we set header to TRUE. We then use the R function "attach()" to use the columns in the table as individual objects:

```
attach(data)
```

#### Controlling factors:

$X_1$  = % of total eligible students in the state who took the exam

$X_6$  = median percentile of ranking of test takers within their secondary school classes

#### Explanatory Factors:

$X_2$  = median income of families of test takers, in hundreds of dollars

$X_3$  = average number of years that test takers had in social sciences, natural sciences, and humanities

$X_4$  = % of test takers who attended public schools

$X_5$  = state expenditure on secondary schools, in hundreds of dollars per student

Once again, we use the `lm()` function to build a model from the data:

```
regression.line = lm(sat ~ takers + rank + income + years + public +  
expend)
```

Note that SAT is the response variable. You will recall that the first two predictive variables, takers and rank, are **controlling factors**.

We then use `summary()` to view information about our model:

```
summary(regression.line)
```

This series of R commands produces the following output:

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-94.659109	211.509584	-0.448	0.656731
takers	-0.480080	0.693711	-0.692	0.492628
rank	8.476217	2.107807	4.021	0.000230 ***
income	-0.008195	0.152358	-0.054	0.957353
years	22.610082	6.314577	3.581	0.000866 ***
public	-0.464152	0.579104	-0.802	0.427249
expend	2.212005	0.845972	2.615	0.012263 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.34 on 43 degrees of freedom  
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618  
F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

As with previous examples, here T-value we have information about estimated regression coefficients, standard errors, and the probability of the statistical significance test.

Let's focus first on the first predictor: takers. As can be seen in the second column above, the estimated coefficient for takers is -0.48, and the standard error is 0.693. And the sampling distribution has a t-distribution with 43 degrees of freedom. (If you remember, 43 corresponds to  $n - p - 1$ .)

We can use the P-value in this output to measure the likelihood of statistical significance. The P-value of takers is greater than 0.1, which means that we do not reject the null hypothesis that the coefficient corresponding to this predictor is zero.

However, as this is one of the controlling factors, we would expect this predictor to impact the variation in the mean SAT scores.

Takers and rank are highly correlated. You will recall we interpret statistical significance in a context of a multiple linear regression. That means, we conclude that the coefficient is not statistically significant given that there are other predictors in a model, given for example that rank, the second predictor is in the model. So this is why we find that the P-value is large in this context and this draws our attention to the idea of marginal versus conditional model.

If we want to test for overall regression we can use the F test. The F-value is 51.91 and the P-value is approximately zero, which means that there is at least one of the predictive variable that has predictive power among the predictive variables we include in the model.

To estimate the confidence intervals for the individual regression coefficients, we use the confint() function, specifying the model, the predictor for which we want to estimate a confidence interval, and the level if different than the default of 0.95:

```
confint(regression.line, "takers", level = 0.99)
```

This code produces:

	0.5 %	99.5 %
takers	-2.349701	1.389541

With estimated coefficients between -2.3 and 1.3, it is plausible the interval includes zero given all other predictor variables are in the model.

To test whether the model is better with only the controlling factor or with the controlling factors plus the other four explanatory variables is better, we can perform the partial F test using the R command Anova(). First, we create a reduced model with only the controlling factors: takers and rank:

```
regression.line.reduced = lm(sat ~ takers + rank)
```

We then use anova() to compare the two models:

```
anova(regression.line.reduced, regression.line)
```

The result:

### Analysis of Variance Table

Model 1: sat ~ takers + rank

Model 2: sat ~ takers + rank + income + years + public + expend

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	53778			
2	43	29842	4	23935    8.6221	3.35e-05 ***

And the output provides the partial F-value, which is 8.6221 and the P-value. Because the P-value is approximately zero, we reject the null hypothesis that the coefficients corresponding to the four predictors we're adding to the model are zero. So at least one predictive variable among the four has improved the predictive power of the model versus the model that doesn't include the four predictors:

This is the more formal form of test that we just performed:

**Test H0 :  $\beta_{income} = \beta_{public} = \beta_{year} = \beta_{expend} = 0$**

**How was the if F-statistic computed:**

$$\text{F-statistic} = \frac{\text{SSReg}(Income, public, Years, Expend | Takers, Rank) / 4}{SSE / (50 - 6 - 1)}$$

**The p-value is computed as**

$$P(F_{4,43} > \text{F-statistic}) = 1 - P(F_{4,43} < \text{F-statistic})$$

**Interpretation: The p-value is approximately 0 thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expend) will be significantly associated to the state-average SAT score.**

We wanted to test whether the coefficients of the four predictors we're adding to the model are equal to zero.

The F-statistic is the extra sum of square regression by adding the four predictors to the model that already has the controlling factors, divided by the mean sum of square of errors of the full model.

The P-value is the probability that F with a 4 and 43 degrees of freedom is greater than F-statistic. Because the P-value is approximately zero, we reject the null hypothesis. We conclude that at least one of the predictor among the four predictors: income, years, public, and expenditure, will be significantly associated to the state average SAT score.

## Knowledge Check

The sampling distribution of the estimated regression coefficients is:

- A. Centered at the true regression parameters.
- B. The t-distribution assuming that the variance of the error term is unknown and replaced by its estimate.
- C. Dependent on the design matrix.
- D. All of the above.

The estimators for the regression coefficients are:

- A. Biased but with small variance
- B. Unbiased under normality assumptions but biased otherwise
- C. Biased regardless of the distribution of the data.
- D. Unbiased regardless of the distribution of the data.

We can test for a subset of regression coefficients:

- A. Using the F-statistic test of the overall regression.
- B. Only if we are interested in whether additional explanatory variables should be considered in addition to the controlling variables.
- C. To evaluate whether all regression coefficients corresponding to the predicting variables excluded from the reduced model are statistically significant.
- D. None of the above.

**Answers:** D, D, D

## 4. Regression Line: Estimation & Prediction

The lecture is multiple linear regression and we'll focus on estimation and prediction of the regression line. In this lesson, we'll learn about the estimation approach and estimation of the confidence intervals for the regression line along with prediction and prediction intervals for new responses.

### Estimating The Regression Line

Just like a simple linear regression, we will begin with an  $x^*$  which in this case is a vector of values consisting of values for all individual predictors. And we would like to estimate the mean response,  $y$  given this  $x^*$ . The estimated regression line is the regression line where we replace the beta coefficients with the estimated regression coefficients.

So it's going to be  $\hat{\beta}_0$  plus  $\hat{\beta}_1$  times  $x_1^*$  and so on. And in matrix format, we can write this as  $x^*$  transposed times the vector of the estimated coefficients  $\hat{\beta}$ :

**At some selected value of  $x$  (say  $x^*$ ), we estimate the "mean response" of  $Y$  (or the regression line) via**

$$\hat{Y} | x^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \dots + \hat{\beta}_p x_p^* = x^{*T} \hat{\beta}$$

Because the estimators of beta are normally distributed, so is  $\hat{y}$ .  $\hat{y}$  is thus normally distributed and we can use this normal distribution in order to make statistical inferences on  $\hat{y}$ . But we also need to specify the expected value and the variance of  $\hat{y}$ :

**Because the estimators of  $\beta$  are normally distributed, so is  $\hat{Y}$ . That means we can draw inference on the regression line using  $\hat{Y}$  if we know the expected value and variance.**

Similar to simple linear regression, the expectation of the estimated mean response is the linear combination of the expectations of the estimators of betas, which we know are equal to the true parameters. Thus the expectation of the mean response, or estimated regression line, is the regression line itself. And that's an unbiased estimator. The variance is thus provided on the slide. The variance depends on the variance of the

error terms ( $\sigma^2$ ), on the design matrix (X), and the values  $x^*$  at which we evaluate the regression line:

$\hat{Y}$  has a normal distribution with

$$E(\hat{Y} | x^*) = x^{*T} \beta = \beta_0 + \beta_1 x_1^* + \dots + \beta_p x_p^*$$

$$Var(\hat{Y} | x^*) = \sigma^2 x^{*T} (X^T X)^{-1} x^*$$

Similar to simple linear regression, if  $x^*$  is on the edge of the space of the predicting variables, the variance of the estimator will be large.

The variance also depends on the design through the inverse of  $x$  transpose  $x$ . We'll discuss this in more detail in a different lesson. But if there is a strong correlation between the predictors, then the values in  $(X^T X)^{-1}$  can be very large. Thus the estimated regression line will have high uncertainty under strong correlation or near linear dependence among the predictive variables.

Since the variance also depends on sigma squared, which is unknown, we need to replace it. We're going to replace it with its estimator the mean square error. And by doing so, now the sampling distribution of  $\hat{y}$  becomes a t-distribution with  $n-p-1$  degrees of freedom, where the degrees of freedom will come from the sampling distribution of the estimated variance, which I'll remind you is a chi-squared distribution with  $n-p-1$  degrees of freedom:

If we replace the unknown variance with its estimator ( $\hat{\sigma}^2 = MSE$ ), the sampling distribution becomes a t-distribution with  $n-p-1$  degrees of freedom.

#### CONFIDENCE INTERVAL FOR REGRESSION LINE

Similar to the derivation of the confidence interval for the regression coefficients, a confidence interval for the mean response is centered at the estimator plus or minus the t-critical point and times the standard deviation of the estimator. The interval length

depends on  $x^*$ , but also on the design matrix through the inverse of the matrix  $X^T X$ . Thus, when we have correlation among the predictors this variance could be large and could lead to wide confidence intervals for the estimated regression line:

The  $(1 - \alpha)$  Confidence interval for the *regression line or mean response* for one instance of predicting variables  $x^*$  is:

$$\hat{y} | x^* \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 x^{*T} (X^T X)^{-1} x^*}$$

Furthermore, if we are interested in estimating a confidence interval for all observed  $x^*$ s then we need to correct for joint statistical inference. That is, for jointly or simultaneously estimating the confidence values for all  $x^*$ s. In this case, we'll use a different critical point. We'll replace the t critical point with the critical point based on the f-distribution which is meant to correct for the simultaneous inference across all  $x^*$ s:

The  $(1 - \alpha)$  Confidence surface for all possible instances of the predicting variables :

$$\hat{y} | x^* \pm \sqrt{(p+1)F_{\alpha, p+1, n-p-1}} \sqrt{\hat{\sigma}^2 x^{*T} (X^T X)^{-1} x^*}$$

## PREDICTING A NEW RESPONSE

Prediction is one of the objectives we perform regression analysis. That is to predict future or different setting responses. While the predicted response is derived similarly to the estimated regression line, prediction is not the same as estimation. This is not only due to the interpretation, but also in the uncertainty level of the predicted mean response. Specifically, the uncertainty in the estimation, in the estimated regression line comes from the estimation alone, from the estimation of the regression coefficients. Whereas for prediction, the uncertainty comes from **the estimation of the regression coefficients** and from **the newness of the observation**:

One of the primary motivations for regression is to use the regression equation to predict future responses. The prediction is the same as the estimator for the mean response.

---

But a prediction is not the same as the line estimate. The prediction contains two sources of uncertainty:

1. Due to the new observation/s
2. Due to parameter estimates (of  $\beta$ 's)

How does this translate in terms of the variance of the predicted mean response? The variance will consist of two components, one coming from the estimation, from the estimated regression line. The other one due to the new measurement  $x^*$ . And this variance will be equal to the variability of  $y$  given  $x^*$ , which is sigma squared under the assumption of constant variance:

1. Variation of the estimated regression line:  $\sigma^2 x^{*T} (X^T X)^{-1} x^*$
2. Variation of a new measurement:  $\sigma^2$

If we add those two variances together, we obtain the variance of the predicted regression line. The difference between the estimated regression line and the predicted line is in the sigma squared which is, again, due to the variability of a new measurement:

The new observation is independent of the regression data, so the total variation in predicting  $Y^* | x^*$  is

$$\sigma^2 x^{*T} (X^T X)^{-1} x^* + \sigma^2$$

The confidence interval for the predicted mean response looks very much like the confidence interval for the estimated mean response, except that now we have an additional value 1 here, which stands for an additional sigma square hat. And this is, again, is because we have these two component of the variability:

A  $(1-\alpha)$  ***prediction*** interval for one new future  $y^*$  (at  $x^*$ ) is

$$x^{*T} \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 (1 + x^{*T} (X^T X)^{-1} x^*)}$$

Note that the predicted regression line is the same as the estimated regression line at  $x^*$ . However, the prediction confidence interval is wider than the estimation confidence interval because of the higher variability in the prediction:

$\hat{y} = x^{*T} \hat{\beta}$     is the same as the line estimate, but the  
interval is wider than the confidence interval  
for the mean response.

If we're interested in prediction intervals for  $m$  different new  $x^*$ s, then we'll need to adjust the critical point for the joint prediction intervals. This adjustment will make the prediction intervals much wider than if we were to consider only one prediction:

A  $(1-\alpha)$  ***prediction*** interval for  $m$  new future  $y^*$  (at  $x^*$ ) is

$$\hat{y} | x^* \pm \sqrt{m F_{\alpha, m, n-p-1}} \sqrt{\hat{\sigma}^2 (1 + x^{*T} (X^T X)^{-1} x^*)}$$

## 5. Regression Line: Estimation & Prediction Data Examples

*Today, I will illustrate the estimation and prediction of the regression line with a data example. Particularly, I will show you in this lesson how to do estimation prediction of the regression line using R statistical software.*

We will return to the example of the relationship between ad expenditure and sales in the presence of other variables that impact sales:

The topic questions we will address are:

- What are the average (mean) estimated sales and the standard deviation for all offices with the same characteristics as those for the first office? What is the 95% confidence interval for this mean response?
- What sales would you predict for the first office if its largest competitor sales would increase at \$303,000 assuming everything else is fixed? What is the standard deviation of this prediction? What is the 95% prediction interval?

In order to address the first set of questions, we will need to first set up the data for the first office, which is the first row in our data matrix. We are interested in extracting the predicting variables:

### **Quantitative Predicting Variables:**

- **X<sub>1</sub>:** the amount (in hundreds of dollars) spent on advertising in 1999.
- **X<sub>2</sub>:** the total amount of bonuses paid in 1999.
- **X<sub>3</sub>:** the market share in each territory
- **X<sub>4</sub>:** the largest competitor's sales (in thousands of dollars)

### **Qualitative Predicting Variable:**

- **X<sub>5</sub>:** A variable indicating the region in which the territory is located (1: South, 2: West, 3: Midwest)

For estimating the standard deviation, we will use the formula introduced in a previous lesson. We will need the estimated variance of the residuals ( $\sigma^2$ ) which we can extract from the summary of the model. We also need the  $x^*$ , the new data

where the values of the characteristics of the predicting values of the first office, and also X, which is a design matrix consisting of all values of the predicting variables for all offices. We then put this together in the next row and, based on the formula of the variance, we take the square root of this value to get the standard deviation.

To obtain the confidence intervals and the estimated mean response, we use the **predict() R command**. In this command, we need to input information from the fitted model, the new data, and we must specify what interval we want. In this case, we want a "confidence" interval estimation.

```
## Data for the first office
newdata = meddcor[1,2:6]

## Estimate standard deviation
s2 = summary(model)$sigma^2
xstar = as.double(newdata)
X = data.matrix(meddcor[,2:6])
predvar = s2*(xstar%*%solve(t(X)%*%X)%*%xstar)
sqrt(predvar)
[1]
[1,] 22.98872

## Confidence Interval
predict(model, newdata, interval="confidence")
fit      lwr      upr
1 934.776 865.0446 1004.509
```

a. Average estimated sales or mean response for sales:

$$\hat{y} = 934.77$$

Estimated standard deviation:

$$se(\hat{y}) = 22.988$$

95% Confidence Interval:

$$(865.04, 1004.51)$$

**Interpretation:** For other offices with the same characteristics as the first office, the average estimated sales are \$934,770 with a lower bound of \$865,040 and upper bound of \$1,004,510.

Going back to the questions: if we want the estimated mean response for sales, the value is 934.77 (units are \$1,000). We can get the estimated standard deviation from the Summary, and the lower and upper bound confidence interval from the predict() R command.

How do we interpret these values? For other offices with the same characteristics as the first office, the average estimated sales will be in a range of 934,770 with a lower bound of 865,000 and an upper bound of 1.00451M.

To address the second set of questions, we will perform a similar exercise, except we will make sure to take into account our predictions. For the prediction, we need to change the competitors' sales:

```

## Change the competitor's sales
newdata[4] = 303

## Estimate standard deviation
s2 = summary(model)$sigma^2
xstar = as.double(newdata)
X = data.matrix(medddcor[,2:6])
predvar = s2*(1+xstar%*%solve(t(X)%*%X)%*%xstar)
sqrt(predvar)
[1]
[1,] 62.62246

## Prediction Interval
predict(model, newdata, interval="prediction")
  fit      lwr      upr
1 911.0569 775.9446 1046.169

```

**b. The predicted sales of the office given the higher competitor's sales:**

$$\hat{y} = 911.05$$

**Estimated standard deviation:**

$$se(\hat{y}) = 62.6224$$

**95% Confidence Interval:**

$$(775.94, 1046.16)$$

**Interpretation:** If the competitor's sales would increase at \$303,000, the predicted sales reduce with \$23,719. Since this is prediction, the standard deviation increases.

We will change the values in newdata to 303 because their sales increased to \$303,000. To estimate the standard deviation, we use a very similar approach, except we now must take into account that for prediction, we need to add one sigma squared hat, but the rest will be the same. We will use the **predict()** R command for the prediction interval, with the interval type of "prediction". This yields a mean of 911.05, the standard deviation is 62.62, and the confidence interval is 775.94 - 1,046.16.

This means that, if the competitors' sales were to increase by \$303,000, predicted sales would reduce with \$23,719. (Because this is prediction, the standard deviation would increase as well.)

## Knowledge Check

The estimated versus predicted regression line for a given  $x^*$ :

- A. Have the same variance
- B. Have the same expectation
- C. Have the same variance and expectation
- D. None of the above

Which one is correct?

- A. The prediction intervals need to be corrected for simultaneous inference when multiple predictions are made jointly.
- B. The prediction intervals are centered at the predicted value.
- C. The sampling distribution of the prediction of a new response is a t-distribution.
- D. All of the above.

**Answers:** B, D

## 3.3 Model Diagnostics, Evaluation, and Multicollinearity

### 1. Assumptions and Diagnostics

*The lecture today is multiple linear regression, and we're going to focus on assumptions and diagnostics. For this lesson, we will learn how to evaluate the assumptions of multiple linear regression. We'll also learn about statistical properties of the residuals in contrast to those of the error terms and I will illustrate departures from the assumptions and we'll discuss the transformation of the model in order to improve the fit of the regression.*

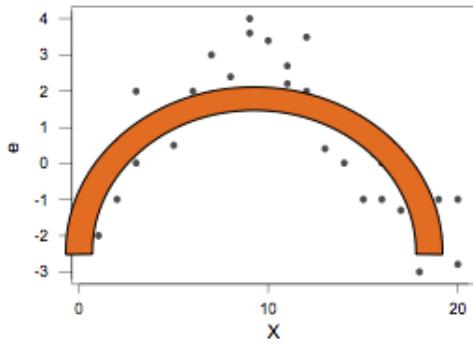
In multiple linear regression, the data consists of response variable Y, and the set of P predicting variables. The model is a linear relationship with respect to predicting variable X, plus the error term epsilon:

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

The assumptions in multiple regression are:

- Linearity assumption: the relationship between Y and  $X_j$  is linear for all predicting variables.
- Constant variance assumption: the variance of the error terms is the same (sigma squared), across all error terms.
- Independence assumption: the error terms are independent random variables.
- Normality assumption: error terms are normally distributed.



### Linearity Assumption:

This shows that there may be a non-linear relationship between X and Y.

## Properties of Errors and Residuals

We'll contrast here the properties of the **error terms** and **residuals**. For the error terms (the true errors or the epsilon) the expectation of the error terms is 0, and the variance is sigma squared. If we stack up the error term into a vector, epsilon, the expectation is a vector of zeros and the variance is going to be sigma squared times the identity matrix:

## Properties of (true) errors:

- $E(\varepsilon_i) = 0$
- $\text{Var}(\varepsilon_i) = \sigma^2$
- $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$

Estimated residuals are the difference between observed and fitted values. The expectation of the residuals are still a vector of zeros just like for the error terms. However, the variance or the covariance matrix of the residuals is sigma squared times I minus H, where I is the identity matrix and H is the hat matrix. This means that the variance of each individual predictor, epsilon hat i, is sigma squared times 1 minus hii

where this hii here, an h with the double index i, is the ith element on the diagonal of the hat matrix:

## Properties of the (estimated) residuals:

- $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
- $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$  (or  $E(\hat{\varepsilon}_i) = 0$ )
- $V(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$  (or  $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{i,i})$ )
  - Where  $\mathbf{H}$  is the hat matrix, and  $h_{i,i}$  is the  $i$ -th element on its diagonal

While the error terms have constant variance, the estimated residuals do not. The variance of epsilon\_i-hat again is sigma squared times 1 minus hii, which depends on i. So it changes from one residual to another so which means that if we want to use the residuals for evaluating the model assumptions we need to standardize them - so we need to take the residuals and divide them by their standard deviation:

- **While the true errors have constant variance, the estimated residuals do not.**
- **To use the estimated residuals for assessing the model assumptions, we need to standardize:**

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

## The Residual Analysis

To use the residual analysis for evaluating the assumptions, we can use various graphical displays: plots of residuals against each individual predicting variable to evaluate linearity, The residuals against the fitted values to evaluate a constant variance of independence, or/and the normality plot and the histogram to evaluate the normality:

$$\text{Standardized Residual Values: } r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 (1-h_{ii})}}$$

Graphical display: **Plot of the residuals  $r_i$**

- Versus each predictor → Linearity
- Versus fitted values → Constant Variance & Independence
- QQ normal plot & histogram → Normality

I'll point out a few facts here that you need to remember when you evaluate assumptions. **We evaluate a normality assumption using the residuals, not the response variable.** So we're not plotting the histogram of the response variable in order to evaluate normality. We plot the histogram of the residuals, the QQ normal plot of the residuals. It is possible for example that if you were to use a histogram of the response variable you see modality, which is not necessarily an indication that the normality assumption does not hold as an indication that the response variable can be explained by a categorical variable, for example.

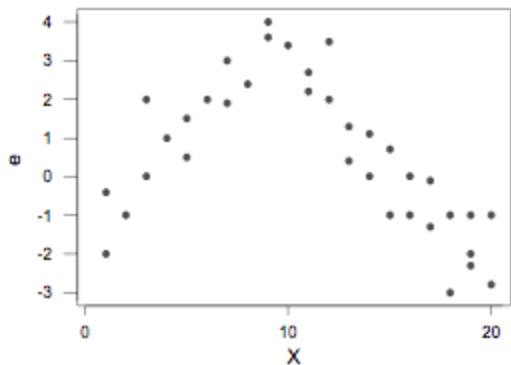
We do not check the predicting variables for normality; we do not assume that the predicting variables are normally distributed. However, if the distribution of a predicting variable is highly skewed, it is possible that the linearity assumption with respect to that variable will not hold. And thus you'll have to consider transformations in order to improve the linearity. This is an example of a departure from the linearity assumption:

- **We evaluate the normality assumption using the residuals not the response variable.**
- **We do not check the predicting variables for normality; however, if the distribution of a predicting variable is strongly skewed, it is possible that the linearity assumption with respect to that variable will not hold.**

#### *Residual Analysis: Linearity Assumption*

If this is a plot of the residuals against one predicted variable, you can see that we have a pattern here, and this shows that there may be a nonlinear relationship between X and Y:

**Linearity:** Plot the residuals against each predicting variable.

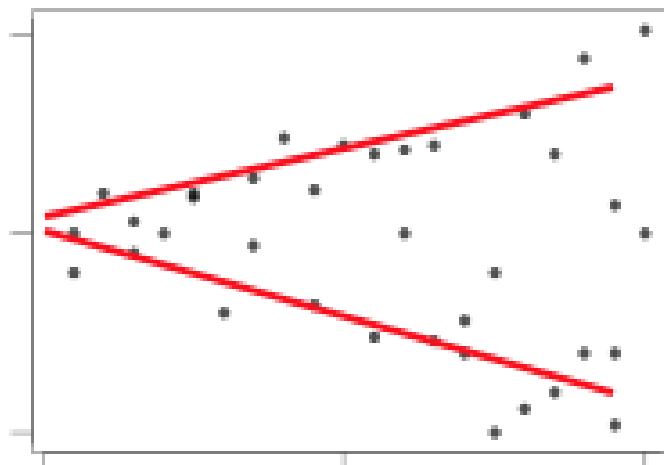


This shows that there may be a non-linear relationship between  $X$  and  $Y$ .

#### *Residual Analysis: Constant Variance Assumption*

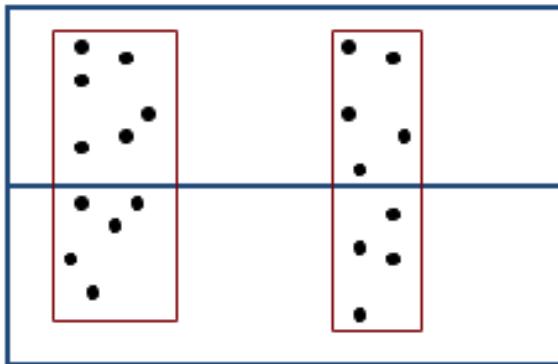
Here is another departure from the assumptions, the constant variance assumption.

In this plot, if this is the residuals against the predicting variables or the fitted variables, what we see here, instead of the residuals show larger variance as a predicting variable or the fitted values increase, which means that the sigma squared is not constant:



#### *Residual Analysis: Independence Assumption (Uncorrelated Errors)*

This is a third example where, of departure from the assumptions, here what we have here is a clusters of residuals which means that the error terms are correlated:



To highlight here is that using residual analysis, we're checking for uncorrelated errors, not independence. Independence is a more complicated matter. If the data are from a randomized experiment, then independence holds. But most data are from observation studies. We commonly correct for selection bias or potential correlation in observation studies using controlling variables - this is one of the benefits of using multiple linear regression.

#### *Residual Analysis: Normality*

To check normality, we can use the quantile plot or the normal probability plot. In such a plot, the data are plotted against theoretical normal distribution in such a way that the points should form an approximate straight line. The intuition behind this plot is that it compares the quantiles of the residuals against quantiles of the normal distribution:

One way to check this assumption in a regression is using  
**a Normal Probability Plot**

y-axis:	$e_i$
x-axis:	$\Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$

$r_i$  = rank of  $e_i$  (between 1, n)  
 $\Phi$  = CDF of Normal Distribution

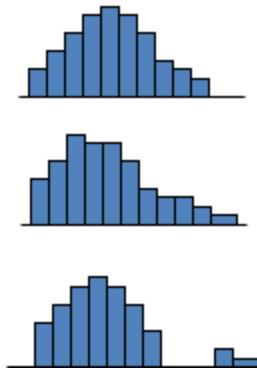
- Let the R statistical software do this for you!
- A straight line in normal probability plot implies that the assumption is valid
- **Curviture (especially at the ends) shows non-normality**

If the residuals are normally distributed, then the quantiles of the residuals will line up with the normal quantiles and thus we should expect that they follow a straight line. Departures from a straight line could be in the form of tail at the end which is an

indication either a skewed distribution or half tailed distribution thus a departure from the normality assumption.

Do not attempt to do this plot with your own implementation. Use the R statistical software to do it for you.

Another complementary approach to evaluating normality is to use a histogram. The histogram is often used to evaluate the shape of a distribution. What we should expect when we plot the histogram residuals is that we should have an approximately symmetric distribution, unimodal and with no gaps in the data.



**Normality Assumption:**  
The residuals should have an approximately symmetric distribution, unimodal and with no gaps in the data.

## Transforming Predicting Variables

What if one or more of the assumptions do not hold? It doesn't mean that the regression is not useful. We can transform predictive variables or/and the response variable in order to improve the fit.

For example, if the assumption does not hold with respect to one or more predicting variables, then we can transform the predicting variable with using some nonlinear function. The common transformations are the **power transformation**, the **log transformation**, and the **polynomial transformation**:

- If the model fit is inadequate, it does not mean that a regression is not useful.
- One problem might be that the relationship between one or more X's and Y is not linear.
- To model the nonlinear relationship, we transform X by some nonlinear function such as:

$$f(x) = x^a \text{ or } f(x) = \log(x)$$

#### NORMALITY TRANSFORMATION

What if the normality or the constant variance do not hold? In this case, we would need to transform the response variable. A common transformation is the power transformation y to the lambda. Also called the Box-Cox transformation, this is used to improve the normality and/or the constant variance assumption. Examples here would be when lambda is equal to 1 we do not transform, when lambda is equal to 0 we use the log transformation. When lambda is equal to 1 over 2 we use a square root transformation. You can identify lambda using the R statistical software there is a command in R called Box-Cox where you need to input the model, fit and will output the estimated value for lambda:

**Problem:** Constant variance or/and normality assumption

**Solution:** Transform the response variable from y to  $y^*$  via

$$y^* = y^\lambda$$

where the value of  $\lambda$  depends on how  $\text{Var}(Y)$  changes as  $x$  changes.

$$\sigma_y(x) \propto \text{const} \quad \lambda = 1 \quad (\text{don't transform})$$

$$\sigma_y(x) \propto \sqrt{\mu_x} \quad \lambda = 1/2$$

$$\sigma_y(x) \propto \mu_x \quad \lambda = 0 \quad y^* = \ln(y)$$

$$\sigma_y(x) \propto \mu_x^2 \quad \lambda = -1$$

## OUTLIERS IN REGRESSION

Another important aspect in regression is the presence of outliers. Any data point that is far from the majority of the data in both x and y is called an outlier.

Data points that are far from the mean of the x's are called **leverage points**.

A data point that is far from the mean both x and y are **influential points** and change the value of the estimated parameters significantly. It can change the statistical significance, it can change the magnitude. It can change even the sign.

It is tempting to just discard such points, however, there is an upshot to this quick fix. Sometimes the outlier belongs in the data. Excluding outliers from an analysis would skew or bias conclusions. But sometimes there are good reasons for excluding subset of data points: when there are errors in the data entry, when there are errors in experiment. **It's good practice to perform the regression analysis with and without the outliers and evaluate the differences.**

You also need to warn the reader of the results or of your findings of the impact of the removal of the outliers may have on the final findings.

**Any data point that is far from the majority of the data (in x's and y) is called an outlier.**

- Data points that are far from the mean of the x's or near the edge of the observation space are called *leverage points*.
- A data point that is far from the mean of both the y and the x's are *influential points* and can change the value of the estimated parameters significantly.

**The upshot:** Sometimes there are good reasons for excluding subsets (there were errors in the data entry; there were errors in the experiment). Sometimes - the outlier belongs in the data. Outliers should always be examined.

## CHECKING FOR OUTLIERS

In order to identify outliers, we can compute Cook's distance. This is the distance between the fitted values of the model with all the data versus the fitted values of the model discarding the  $i^{\text{th}}$  observation:

# Checking for Outliers

$$\textbf{Cook's Distance: } D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{(k + 1)\hat{\sigma}^2}$$

where  $\hat{Y}_{(i)}$  are the fitted values from the model fitted without the  $i^{\text{th}}$  observation (i.e., excluding the  $i^{\text{th}}$  observation from the data) and  $\hat{Y}$  are the fitted values from the model fitted with the  $i^{\text{th}}$  observation (i.e., including all observations).

**Cook's Distance measures how much the estimated parameter values in the regression model change when the  $i^{\text{th}}$  observation is removed.**

**Rule of Thumb:**  $D_i > 4/n$ ,  $D_i > 1$ , OR any “large”  $D_i$  should be investigated.

So the idea here is that the Cook's distance will measure how much all of the values in the regression model change when the  $i^{\text{th}}$  observation is removed. A rule of thumb is that when the Cook's distance for a particular observation is larger than  $4/n$ , it could be an indication of an outlier. Often, we can, if  $D$ , the Cook's distance is larger than 1, it is a clear indication of an outlier. But sometimes we just look for very large values for the Cook's distance as compared to the Cook's distance values for other observations.

## 2. Assumptions and Diagnostics Data Examples

The lecture is Multiple Linear Regression. We're going to focus on implementation of diagnostics of the assumptions in multiple linear regression. We'll illustrate the diagnostics of the assumptions using a data example and using the R statistical software.

We'll return to the example, we're interested in the relationship between advertising expenditure and sales in the presence of other predictive variables that impact sales:



### Quantitative Predicting Variables:

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

### Qualitative Predicting Variable:

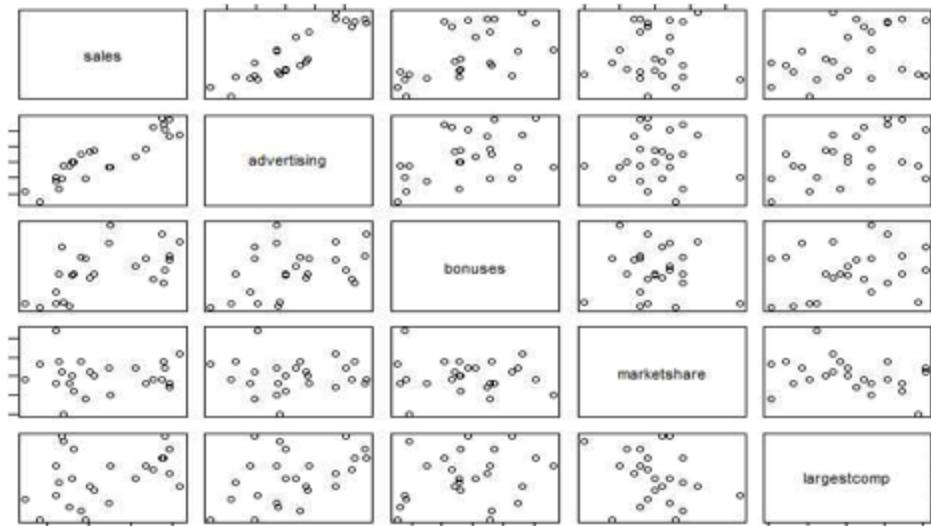
$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

The questions that we'll address in this example are as follows:

- Do the assumptions of multiple linear regression hold? And we want to provide the graphical display that will support those diagnostics.
- If one or more assumptions do not hold, what can we do about it?
- Do we identify any outliers?

To validate our assumption, one way, one approach is to plot the scatter plot of all possible values. The response versus the predicting variables and the scatter plot of all the predicting variables. And we can do that through using the command **plot()**. And the input now is a matrix, not an x and a y is a matrix of variables, which consist of the column of the response variable and the columns of the predicting variables. And the output will look just like:

```
## Scatter plot matrix of sales and numeric predicting variables  
plot(meddcor[,1:5])
```



What we see in this output in the first row of scatter plots are the scatter plots of the sales versus all four quantitative predicting variables: sales versus advertising, sales versus a bonus amount, sales versus market share, or sales versus the largest competitor of sales.

The other set of plots are the scatter plots of the predicting variables: advertising versus the bonuses, market share, and largest competitor, and so on.

So what we can learn from this in terms of linearity when we look at the first row of plots:

- Sales versus advertising expenditure - a strong linear relationship
- Sales versus amount of bonuses - a weaker linear relationship
- Sales versus market share - scattered
- Sales versus largest competitor sales - scattered

So it's possible that we do not have a strong linear relationship with sales with respect to those predicting variables (market share and largest competitor sales).

The other plots, the scatter plots of the predicting variables, can be used to evaluate correlation between predicting variables. We're going to return to correlation between predicting variables later in the next lesson.

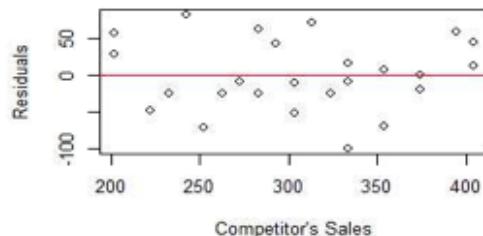
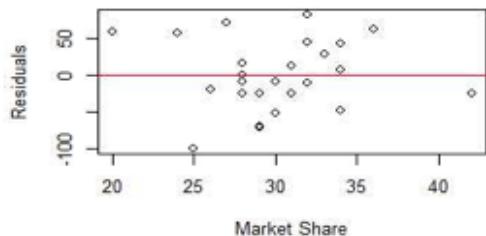
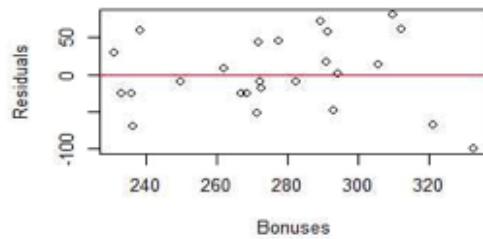
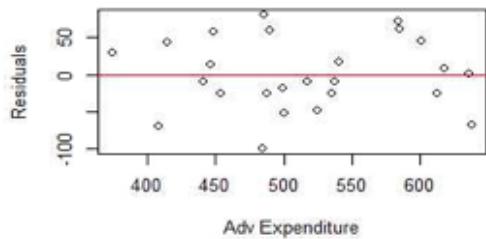
According to this set of scatter plots, we conclude the linearity assumption holds for all predicting variables. And that most of the variables have some relationship with sales, or some of the relationship are stronger than others:

- **Linearity assumption holds for all predicting variables;**
- **For advertisement expenditure, bonus amount and competitor's sales, the relationship with sales is strongly linear.**

Another way to evaluate the linearity assumption is through the residual plots. We'll plot the residuals versus individual predicting variables. So here, I'm extracting the residuals from the model fit using **model\$resid**. And then I'm dividing the display into four different quadrants. And I'm going to plot each individual plot of predicting variable against the residuals. And for each plot, I'm adding the 0 line in order to compare the residuals with the 0 line.

What we're looking for is that the residuals must be spread randomly across the 0 line. This is what we get:

## Linearity Assumption

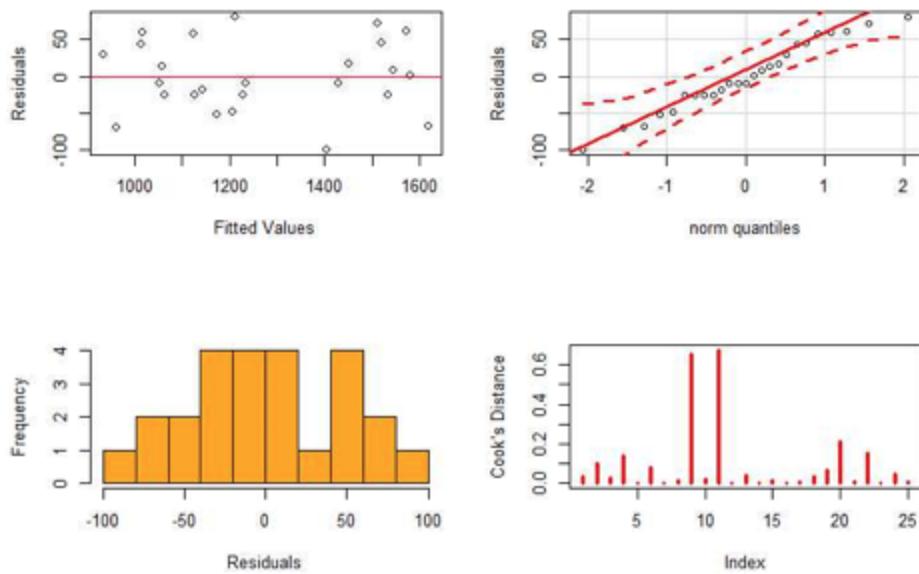


So we can see that for across all four predicting variables, that the residuals are scattered around the 0 line, which is an indication that the assumption of linearity holds.

If we want to evaluate other assumptions, we can use the plot for evaluating normality. We can also use the plot of fitted versus residual step can be used to evaluate uncorrelated errors in constant variance. And here, I'm also providing how to obtain the Cook's distances and how to plot those in order to evaluate whether we have outliers or not:

```
library(car)
fits = model$fitted
cook = cooks.distance(model)
par(mfrow =c(2,2))
plot(fits, resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
qqPlot(resids, ylab="Residuals", main = "")
hist(resids, xlab="Residuals", main = "",nclass=10,col="orange")
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's Distance")
```

So distance will be the set of plots based on our set of commands:



The first one are going to be the residuals versus fitted. And what we see in the first plot is that the residuals are spread around the 0 line, which is an indication that both the constant variance and the uncorrelated errors assumptions will hold.

The next plot is the normal probability plot, and we can see that the points do line up pretty well on a straight line.

The histogram shows that the residuals are symmetric, except that we see a gap somewhere around 40.

The last plot is the plot of the Cook's distances. When you evaluate this plot—[looking for] values that are much larger than the other Cook's distances—we see that two values are somewhat larger than the other values. And it would be important here to evaluate whether those are influential points or not.

### box3. Model Evaluation and Multicollinearity

The lecture is Multiple Linear Regression, and this lesson is about Model Evaluation and Multicollinearity. What we learn in this lesson is how to evaluate the Multiple Linear Regression model. And also, we'll learn about the concept of Multicollinearity, which is particularly specifically important for Multiple Linear Regression when we have multiple predictors in a model.

#### MODEL EVALUATION: R SQUARED, THE COEFFICIENT OF DETERMINATION

Just like in a simple linear regression, an approach for evaluating the Multiple Linear Regression, is the Coefficient of Variation/Coefficient of Determination. This is the so called  $R^2 = 1 - \frac{SSE}{SST}$ :

A statistic that efficiently summarizes how well the X's can be used to explain Y is the R-square:

$$= 1 - \frac{SSE}{SST}$$

where the sum of squares are provided:

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

And we interpret  $R^2 = \text{proportion of total variability in } Y \text{ that can be explained by the linear regression model}$ :

which is interpreted as

**$R^2 = \text{Proportion of total variability in } Y \text{ that can be explained by the regression model}$**

## R<sup>2</sup>: NOTATION AND TERMINOLOGY

The R-squared formula involves the so-called sum of squares. Here I will recap the sum of squares differentiated into **sum of squared errors**, **sum of squares total**, and **sum of squares for regression**. Unfortunately, the field of statistics abounds in inconsistent terminology and notation. This slide provides an account of different ways these sums squares are denoted or defined. For consistency, I will try to stay with the same notation throughout the course although do keep in mind all this other notation.

- **SSE**: sometimes denoted  $\text{SS}_{\text{error}}$  or  $\text{SS}_{\text{err}}$ , is also known as **RSS** (residual sum of squares) and  $\text{SS}_{\text{res}}$  (sum of squared residuals, sometimes **SSR**).
- **SST**: sometimes written as  $\text{SS}_{\text{total}}$  or  $\text{SS}_{\text{tot}}$ . It is also called total sum of squares and written as **TSS**.
- **SSR**: also called the sum of squares due to regression, and it is sometimes written as  $\text{SS}_{\text{reg}}$ . It's also called explained sum of squares (**ESS**). Don't confuse ESS with SSE, and, for R<sup>2</sup>, remember that SSR is SS regression, not SS residuals!

## MODEL EVALUATION: F-TEST

Another approach to evaluate the model is through the overall regression test or the F-test. For this test, the null hypothesis is that all the regression coefficients except the intercept are 0. Versus the alternative that at least one is not 0. What this says is that, if we reject the null hypothesis, we will conclude that at least one of the predicting variables has explanatory power for the variability in the response:

1. **F-test for  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$  vs  $H_A: \text{at least one is not zero}$**

In order to perform this test, we can use the  $F_0 = \text{MSR} / \text{MSE}$ . Just like the ANOVA model, the F-statistic has  $F(p, n-p-1)$  of freedom. We reject the null hypothesis if the F-statistic is larger than the F quantile, the 1 minus alpha quantile of the F distribution:

$$\begin{aligned} F_0 &= \text{MSR} / \text{MSE} \sim F(p, n-p-1) \\ \text{MSR} &= \text{SSR}/p \quad \text{MSE} = \text{SSE}/n-p-1 \end{aligned}$$

Again, another, we discussed that the coefficient of determination is R squared, another way to evaluate the model. However what I would like to point out here is that **R<sup>2</sup> will always increase if we add more predicting variables**. So when we want to compare models with different numbers of predicting variables, we should use the **adjusted R<sup>2</sup>**, because the adjusted R<sup>2</sup> is adjusted for the number of predictive variables. So it's not going to increase as we add more predictive variables:

2. Coefficient of variation/determination:

$$R^2 = \frac{SSR}{SST}$$

- R<sup>2</sup> will always increase if we add more predicting variables

3. Adjusted Coefficient of variation:

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)(1-R^2)}{(n-k-1)}$$

*When we're interested in explaining the variability explained by the model, we use the R<sup>2</sup>, when we want to compare models with different numbers of predicting variables, we're going to use the adjusted R<sup>2</sup>.*

## CORRELATION COEFFICIENT

Another evaluation approach is the Correlation Coefficient. The Correlation Coefficient is used to evaluate the linear dependence, the linear relationship between two variables, It could be between X and Y, could be between two X's:

A statistic that efficiently summarizes how well **ONE** of the X's is linearly related to Y (or to another X) is the p, the (Pearson) correlation coefficient:

$$\rho = \text{cor}(X_j, Y) = \frac{\sum_{i=1}^n y_i (x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- It can be used to evaluate the linear relationship between the response variable and the predicting variables;
- It can also be used to evaluate the correlation between the predicting variables for detecting (near) linear dependence among the variables (or multicollinearity).

Since we can use the Correlation Coefficient in order to evaluate whether we have a linear relationship between the response variable and the predicting variables, we can use this coefficient in order to find a good transformation in order to improve the linearity assumption. So we would try several transformations for X, for the predictive variable, and we'll choose the transformation that will most improve the Correlation Coefficient.

We also can use the Correlation Coefficient to evaluate the correlation between the predicting variables, for detecting (near) linear dependence among the variables (or multicollinearity) as I'll discuss on the next slide.

#### DIAGNOSING MULTICOLLINEARITY

How can we diagnose Multicollinearity? An approach to diagnose collinearities through the computation of the variance inflation factor, which you will compute for each predicting variable. If we considered a j predicting variable, the variance inflation factor or  $VIF_j = 1 / 1 - R^2_j$  where this  $R^2_j$  is the coefficient of variation or the  $R^2$  of the regression of the variable  $X_j$  regressed on all other predicting variable:

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ : find values minimizing squared error:

if  $\mathbf{X}^T \mathbf{X}$  invertible 
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$\mathbf{X}^T \mathbf{X}$  is not invertible if columns of  $\mathbf{X}$  are linearly dependent, i.e. one variable is a linear combination of the others

- Implication: the standard error of  $\hat{\beta}$  is infinite.

**Near Collinearity:** columns of  $\mathbf{X}$  are approximately linearly dependent

- The standard error of  $\hat{\beta}$  is artificially large;
- If one value of one of the predicting variables is changed only by slightly, the fitted regression coefficients can change dramatically;
- The overall F statistic may be significant, yet each of the individual t-statistics is not significant.
- Prediction is also affected since the relationship to the response likely will change widely in the presence of collinearity.

How can I diagnose Multicollinearity? On  $X_j$  regressed on all other predictive variables:

The variance inflation factor(VIF) for each predicting variable:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where  $R_j^2$  is the coefficient of variation of the regression of the variable  $X_j$  on all other predicting variables.

How big of a VIF or variance inflation factor indicates a problem? We can evaluate this condition where the  $VIF < \max(10, 1 / 1 - R^2 \text{ model})$ . Where the  $R^2$  model is the coefficient of variation, or the  $R^2$  of the regression model.

How do we interpret the VIF? VIF measures the proportional increase in the variance of beta hat j compared to what it would have been if the predictive variables had been completely uncorrelated:

**Interpretation:** VIF measures the proportional increase in the variance of  $\hat{\beta}_j$  compared to what it would have been if the predicting variables had been completely uncorrelated. How big a VIF indicates collinearity:

$$VIF < \max(10, \frac{1}{1 - R_{model}^2})$$

Where  $R_{model}^2$  is the coefficient of variation of the regression model.

So what we want to see is that the variance of beta hat j is not significantly larger or, when we have correlation among the predictive variables versus when we don't have correlation among the predictive variables, which means that Multicollinearity will not cause a problem in the regression. So we will check this VIF, we'll check this condition for every single predicting variable. If this condition holds for all the predicting variables, it means that the estimated coefficients are not likely to be unstable, so collinearity is not a problem.

Again, it could be that the predicting variables are correlated, but it doesn't necessarily mean that that will lead to a problem in the stability of the estimated regression coefficients.

## 4. Multicollinearity Data Examples

The lecture is multiple linear regression with a focus on multicollinearity with a data example. In this lesson, we'll learn how to identify multicollinearity.

### **Quantitative Predicting Variables:**

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

### **Qualitative Predicting Variable:**

$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

arity in the R statistical software

with a data example.

We'll return to the example, where we're interested in a relationship between the advertising expenditure and sales in the presence of other predicted variables that impact sales:

The questions I will address are:

- What are the correlation coefficients between the quantitative predicting variables? Is there any potential multicollinearity?
- Can we obtain the variance inflation factors? How? Is there multicollinearity?
- What is the coefficient of variation, and how do we interpret it?

In order to compute the correlation between the predictive variables, we can use the **cor()** Command in R which stands for correlation. Here we're going to input the matrix where the columns of the matrix are the four predicted variables. And what we see here is the matrix of the correlation values:

```
cor(meddcor[2:5])
      advertising    bonuses   marketshare largestcomp
advertising 1.00000000  0.41868215 -0.02029937  0.4524897
bonuses     0.41868215  1.00000000 -0.08484673  0.2286563
marketshare -0.02029937 -0.08484673  1.00000000 -0.2872159
largestcomp  0.45248974  0.22865628 -0.28721592  1.0000000
```

**The maximum correlation between predicting variables is 0.452.**

For example, the values 0.418 is the correlation between advertising expenditure and amount of bonuses. The maximum correlation between predictive variables and across all four predictors is 0.452. So it's not a very strong correlation among the predictive variables.

If we want to compute the VIFs for each individual predictor, we can use the VIF command in R, where the input is the fitted model, and where you need to focus is on the first column of values, which are the VIF values:

```
vif(model)
      GVIF   Df  GVIF^(1/(2*Df))
advertising 3.081657  1    1.755465
bonuses     1.359601  1    1.166019
marketshare 1.311265  1    1.145105
largestcomp 1.569851  1    1.252937
region       3.784660  2    1.394783
```

**None of the VIFs are greater than max(10, 1/(1-R<sup>2</sup>)) = 10**

Now, we can compare the VIF values with the threshold which is maximum of 10, maximum between (10 and 1/(1- R<sup>2</sup>)) = 10. And we can see that none of the values, the VIF values are larger than ten, which is an indication that we don't have multicollinearity in this example.

If we were interested to obtain the R squared, we can use the summary of the fitted model and specify that we want R squared from this summary:

```
summary(model)$r.squared
```

**The coefficient of variation is 0.955. Thus the model explains 95.5% of the variability in the sales.**

And the analysis is used for goodness of fit assessment.

- A. All of the above.

In the presence of near multicollinearity:

- A. The coefficient of variation decreases.
- B. The regression coefficients will tend to be identified as statistically significant even if they are not.
- C. The prediction will not be impacted.
- D. None of the above.

When do we use transformations?

- A. If the linearity assumption with respect to one or more predictors does not hold, then we use transformations of the corresponding predictors to improve on this assumption.
- B. If the normality assumption does not hold, we transform the response variable, commonly using the Box-Cox transformation.
- C. If the constant variance assumption does not hold, we transform the response variable.
- D. All of the above.

Which one is correct?

- A. The residuals have constant variance for the multiple linear regression model.
- B. The residuals vs. fitted can be used to assess the assumption of independence.
- C. The residuals have a t-distribution if the error term is assumed to have a normal distribution.
- D. None of the above.

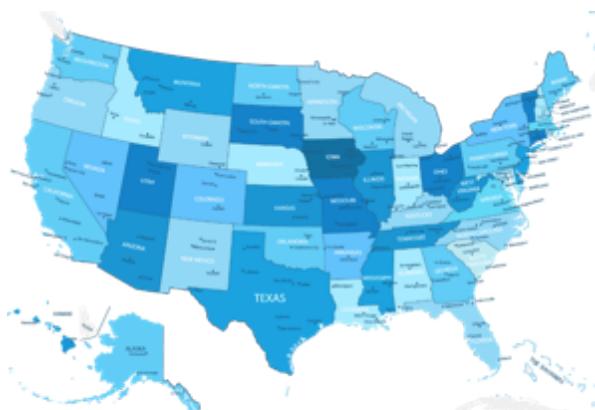
**Answers:** D, D, D, D

## 3.4 Case Study: Ranking States by SAT Performance

### 1. Exploratory Analysis

*This is the multiple regression lecture and I will illustrate multiple regression with an example related to ranking states by SAT performance. In this lesson I'll focus on introduction of the example as well as I will focus on bias selection and exploratory data analysis.*

In 1982, average SAT scores were published with breakdown of state by state performance of SAT in the United States. The average state SAT scores vary considerably by state with mean scores falling between 790 for South Carolina to 1,088 for Iowa:



SAT Mean Score by State – Year 1982

790 (South Carolina) - 1088 (Iowa)

**Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?**

Two researchers examined compositional and demographic variables to understand to what extent those characteristics were tied to SAT scores. Their research questions addressing this example are:

Which variables are associated with the state SAT scores?

How do the states rank with respect to the SAT performance?

Which states perform best for the amount of money they spend?

**The response variable is:**

**Y** = the state average SAT score (verbal and quantitative combined).

### **The predicting variables are as follows.**

- **takers**, the percentage of total eligible students in the state who took the exam.
- **rank**, the median percentile of ranking of test takers within their secondary school classes.
- **income**, the median income of families of test takers, in hundreds of dollars.
- **years**, the average number of years that a test taker has had in social sciences, natural sciences, and humanities.
- **public**, a percentage of test takers who attended public schools.
- **expend**, a state expenditure on secondary school and hundreds of dollars per student.

Back in 1982, not all colleges required SAT for admission, particularly those in the Midwest. This resulted in the states with high average SAT scores had low percentages of takers. The reason is that only the best students planning to attend college out of state took the SAT exams. As the percentage of takers increase for other states, so does the likelihood that the takers include lower-qualified students. Thus, in this example, two variables can be used to control for this bias selection, which are the percentage of students taking SAT, or takers and the median percentile of ranking of test takers within their secondary school classes which is rank:

### **Selection Bias:**

The states with high average SAT scores had low percentages of takers. Those taking the test will tend to be in the higher median percentile of ranking of test takers within their secondary school classes.

### **Controlling Factors:**

- **takers**: % of total eligible students in the state who took the exam
- **rank**: median percentile of ranking of test takers within their secondary school classes

We'll first read the data in R using a **read.table() command** in R where we need to input the file name, and we need to specify that a filename that columns in a filename have a header. To check the data I read out the first four rows of the data and you can see the columns in the data matrix:

```

## Read the data using the 'read.table()' R command because it is an ASCII file
data = read.table("SATData.txt", header = TRUE)
## Check data to make sure correctly read in R
data[1:4,]
  state      sat  takers income years public expend rank
1  Iowa     1088      3    326  16.79   87.8   25.60  89.7
2 SouthDakota 1075      2    264  16.07   86.2   19.95  90.6
3 NorthDakota 1068      3    317  16.57   88.3   20.62  89.8
4  Kansas    1045      5    338  16.30   83.9   27.14  86.3

## Check dimensionality of the data file
dim(data)

## Attach data to automatically recognize the columns in the data as individual vectors
attach(data)

```

**The data consist of 50 rows, each corresponding to a U.S. state.**

We can check the dimensionality of the data file which consists for 50 rows, each row corresponding to a US state. We will attach the data here, in order to automatically recognize the columns in the data as individual vectors.

## EXPLORATORY DATA ANALYSIS in R

Exploratory data analysis allows us to look at the variables containing the data set before beginning any formal analysis.

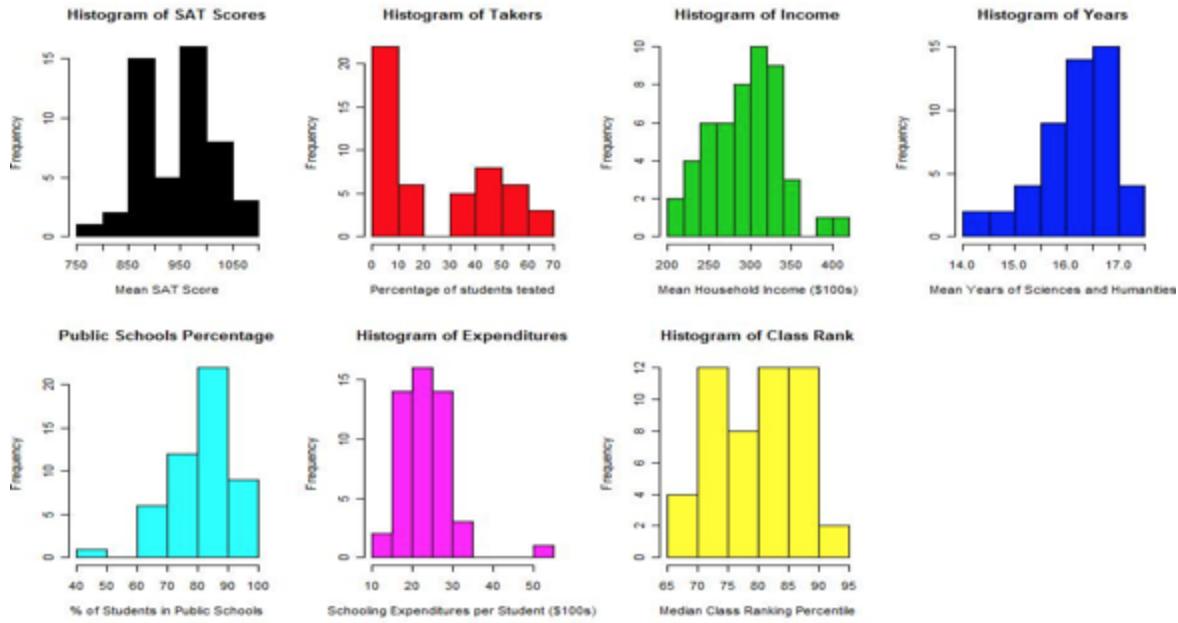
For exploratory data analysis in this example, we'll first examine the variables individually through histograms. For each variable in the data, we plot its histogram. With a histogram, we can see the general range of the data, shape such as skewness, as well as any other potential trends:

```

## Evaluate the shape of the distribution of each predicting variable and of the response variable
par(mfrow = c(2, 4))
hist(sat, main = "Histogram of SAT Scores", xlab = "Mean SAT Score", col = 1)
hist(takers, main = "Histogram of Takers", xlab = "Percentage of students tested", col = 2)
hist(income, main = "Histogram of Income", xlab = "Mean Household Income ($100s)", col = 3)
hist(years, main = "Histogram of Years", xlab = "Mean Years of Sciences and Humanities", col = 4)
hist(public, main = "Public Schools Percentage", xlab = "Percentage of Students in Public Schools", col = 5)
hist(expend, main = "Histogram of Expenditures", xlab = "Schooling Expenditures/Student ($100s)", col = 6)
hist(rank, main = "Histogram of Class Rank", xlab = "Median Class Ranking Percentile", col = 7)

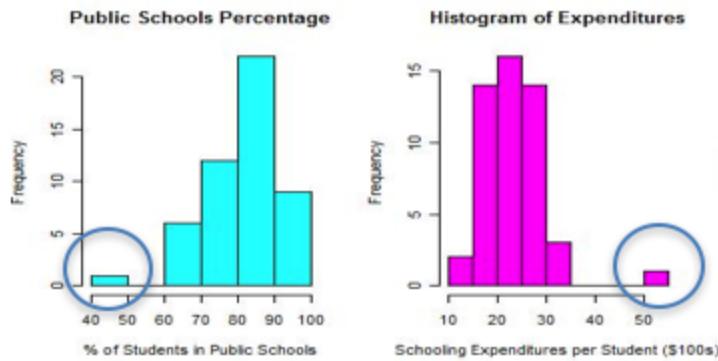
```

These are the histograms for the SAT scores in black and for all the other six predictors:



From these histograms we note that the state average SAT score has a bi-modal distribution. And this is potentially indicating the clustering of states depending on whether the colleges in the states require SATs or not. Thus potentially due to the bias selection.

Following on this line of thought, we can see the takers, the next (red) histogram, that this variable has clearly two clusters which may explain the modality in the SAT score average, as well. We can note also, that one state, Alaska, has almost double the amount of secondary schooling expenditure compared to all the other states:

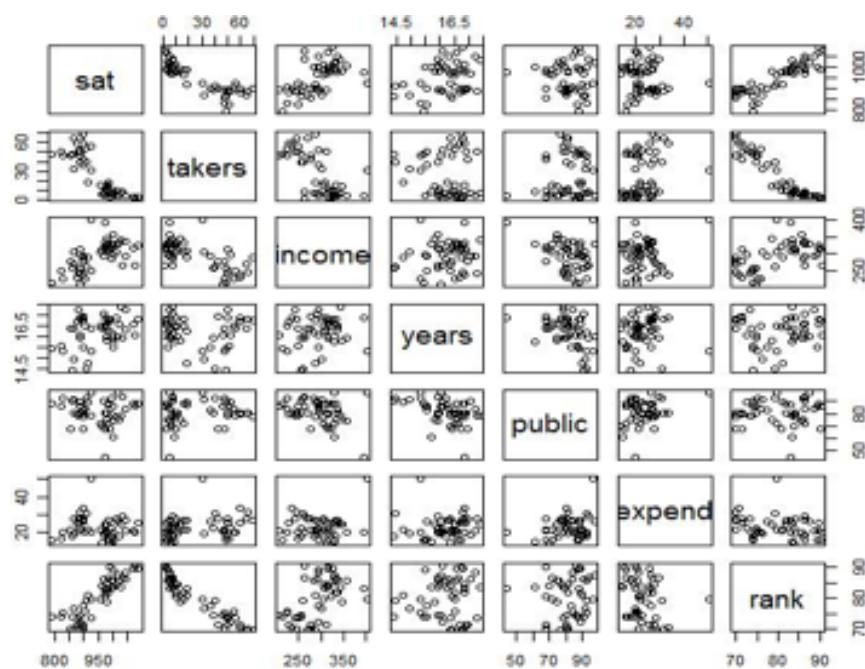


Similarly, the state of Louisiana has very few students coming from public schools taking the SAT score.

We can also look at all the variables together by using a pair-wise correlation by using the correlation command in R. We can also make a scatter-plot matrix of all the variables including the response and the predictive variables. The scatter plot command that I showed you in a previous slide outputs a scatterplot matrix showing the relationship between the variables:

```
## Evaluate the scatter plot matrix of the data ', ignoring the first column
par(mfrow = c(1, 1))
plot(data[,-1])

## Explore the correlation coefficients
round(cor(data[,-1]), 2)
```



Generally, we're looking for trends here. *Does the value of one variable tend to affect the value of another? If so, is their relationship linear?* This type of question helps us think of what type of interaction in higher order terms we may want to include in the regression model.

We can also evaluate whether the linearity assumption holds and whether there is strong linearity among the predictive variables.

The scatter plot matrix that we see on the slide shows a clear relationship between SAT, the response variable, and takers and rank, which are the two controlling variables. Interestingly Alaska shows up as a high value in expenditure. And we can see that Alaska has a rather average SAT score despite its very high levels of spending. We'll

leave Alaska in the model, in the data set but a more complete analysis would seek to remove outliers and high influential points.

In fact, this data contains two rather obvious outliers. One feature visible in both the scatter plot and the histogram in the gap in the distribution of takers. When there is such a distinct gap in a variable distribution, sometimes it's a good idea to consider transformation from a continuous variable to an indicator variable.

Since subtle trends are often difficult to identify in the scatter plot matrices, sometimes a correlation matrix can be useful:

	<b>sat</b>	<b>takers</b>	<b>income</b>	<b>years</b>	<b>public</b>	<b>expend</b>	<b>rank</b>
sat	1.00	-0.86	0.58	0.33	-0.08	-0.06	0.88
takers	<b>-0.86</b>	1.00	-0.66	-0.10	0.12	0.28	<b>-0.94</b>
income	0.58	-0.66	1.00	0.13	-0.31	0.13	0.53
years	0.33	-0.10	0.13	1.00	-0.42	0.06	0.07
Public	-0.08	0.12	-0.31	-0.42	1.00	0.28	0.05
expend	-0.06	0.28	0.13	0.06	0.28	1.00	-0.26
rank	<b>0.88</b>	-0.94	0.53	0.07	0.05	-0.26	1.00

From the correlation matrix for this data, we note that both the income and the years variables have moderately strong positive correlations with the response variable SAT. Their respective correlations are 0.58 and 0.33, indicating that higher levels of income and years of education in science and humanities are generally associated with higher trends with the SAT scores. However, this does not imply causation. And each of these trends may be nullified or even reversed when accounting for the other variables in the model.

A variable such as years may be of particular interest to researchers. Although neither science nor humanities are directly tested on the SAT, researchers may be interested in whether an increase in the number of years of such classes is associated with a significant increase in a SAT score. I'll come back to this point in a different lesson.

## 2. Regression Analysis

*This is the multiple linear regression lecture and the focus of this lesson is regression analysis. For the example, we're interested in ranking states by SAT performance. In this lesson, I'll focus on the implementation of the multiple linear regression and inference based on the statistical and the hypothesis testing procedure for subset of the regression coefficients.*

### FIT THE MULTIPLE REGRESSION MODEL: LM()

The R command used to fit a multiple linear regression model is **lm()**. Where we input a response variable, in this case, the state average SAT score, denoted with SAT, on the left of the tilde, and the predicting variables joined by the plus sign on the right of the tilde. This is a portion of the output of the model fit with information on the estimated coefficients:

```
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-94.659109	211.509584	-0.448	0.656731
takers	-0.480080	0.69371	-0.692	0.492628
rank	8.476217	2.10780	4.021	0.000230 ***
income	-0.008195	0.15235	-0.054	0.957353
years	22.610082	6.31457	3.581	0.000866 **
public	-0.464152	0.57910	-0.802	0.427249
expend	2.212005	0.84597	2.615	0.012263 *

---

Residual standard error: 26.34 on 43 degrees of freedom

Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618

F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

What we can learn from this output is that it's not only the estimated coefficients, but also statistical inference on the statistical significance of the coefficients:

```
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-94.659109	211.509584	-0.448	0.656731
takers	-0.480080	0.69371	-0.692	0.492628
rank	8.476217	2.10780	4.021	0.000230 ***
income	-0.008195	0.15235	-0.054	0.957353
years	22.610082	6.31457	3.581	0.000866 ***
public	-0.464152	0.57910	-0.802	0.427249
expend	2.212005	0.84597	2.615	0.012263 *
---				

Test for statistical significance:  
 $\hat{\beta}_{takers}$ : t-value = -0.692, p-value > 0.1  
 $\hat{\beta}_{rank}$ : t-value = -0.692, p-value < 0.01  
 $\hat{\beta}_{income}$ : t-value = -0.054, p-value > 0.1  
 $\hat{\beta}_{years}$ : t-value = 3.581, p-value < 0.01  
 $\hat{\beta}_{public}$ : t-value = -0.802, p-value > 0.1  
 $\hat{\beta}_{expend}$ : t-value = 2.615, p-value = 0.012

For example, among the coefficients, those that are statistically significant are for the rank-predicting variable, for year's predicting variable, and potentially for expenditure.

In the lower part of the output:

---

Residual standard error 26.34 on 43 degrees of freedom  
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618  
F-statistic 51.91 on 6 and 43 DF, p-value: < 2.2e-16

$\hat{\sigma} = 26.34$ ,  $n-2 = 43$   
 $R^2 \sim 87.8\% \text{ variability explained}$

We can see information about the estimated standard deviation of the error terms, which is 26.34 with 43 degrees of freedom, which corresponds to  $n-p-1$ .

We can also see information on the F test for the overall regression, which is 51.91. The F-statistic and the p-value is very small, indicating that at least one of the predicting variables has explanatory power on the variability of the SAT scores.

We also see the R squared, which is 0.87, meaning 87.8% of the variability in the SAT is explained by the model.

#### TESTING FOR SUBSETS OF COEFFICIENTS: ANOVA()

To use the ANOVA command to the decomposition of the sum of regression into extra sums of squared of regression, due to adding one predictive variable at a time to the model, as we learned in a lesson on testing for subset of regression coefficients.

Note that the order in which the predicted variables enter the model is important here. I recommend to review the lesson where I introduce the testing procedure for subset of coefficients.

```
## Compare models: reduced with controlling variables only vs. full with all variables  
anova(regression.line)
```

Analysis of Variance Table

Response: sat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
takers	1	181024	181024	260.8380	< 2.2e-16 ***
rank	1	11209	11209	16.1512	0.0002313 ***
income	1	2858	2858	4.1182	0.0486431 *
years	1	16080	16080	23.1701	1.86e-05 ***
public	1	252	252	0.3631	0.5499447
expend	1	4745	4745	6.8369	0.0122629 *
Residuals	43	29842	694		

The anova command (above) gives the sum of squares explained by the first variable. Then the second variable conditional on including the first. Then the third variable conditional on the first and second and so forth. For example, the extra sum of squares of adding income to the model, which includes takers and rank is 2,858. Another example, the extra sum of squares for adding the predictive variable years to the model that includes takers, rank and income is 16,080, and so on.

For the SAT data, we would like to test whether dropping income, years, public and expenditure is better than the model with these variables. That is we would like to test whether any of these variables will improve the predictive power of the model, when added to takers and rank, the controlling factors.

For this, we compute the F value of the ratio between the sum of the extra sum of squares of regressions due to adding these four variables to the model, divided by four, which is the number of predictive variables we're adding. And we divide this by the mean sum of squares of the full model. We compute the p-value as the left tail of the distribution, with 4 and 40 [43?] degrees of freedom, evaluated this F statistic, F value:

```
## compute partial-F statistic  
fstat = ((2858+16080+252+4745)/4)/(29842/43)  
pvalue = 1-pf(fstat,4,43)  
pvalue  
[1] 3.349778e-05
```

So let's overview once more this test:

**Test:**  $H_0: \beta_{income} = \beta_{public} = \beta_{years} = \beta_{expend} = 0$

How was the F-statistic computed?

$$\text{F-statistics} = \frac{\text{ssReg}(Income, Public, Years, Expend | Takers, Rank) / 4}{SSE / (50 - 6 - 1)}$$

The p-value is computed as

$$P(F_{4,43} > F - \text{statistic}) = 1 - P(F_{4,43} < F - \text{statistic})$$

**Interpretation:** The p-value is approximately 0 thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expenditure) will be significantly associated to the state-average SAT score.

What we're testing here is whether, in the null hypothesis, the regression coefficients corresponding to the four predictors are 0. The alternative hypothesis would be that at least one of those coefficients is not 0.

How was the F statistic computed again? We divide the sum of square of regressions by, adding extra sum of square regressions, by adding income, public, years, and expenditure to the model that already includes takers and rank, divided by 4. The denominator is the mean sum of square error of the full model. The p-value is computed as the left tail for the F distribution with 4 and 43 degrees of freedom, evaluated F statistic.

The way we interpret this is as follows: the p-value is approximately 0, thus we reject the null hypothesis. And we conclude that unless one other predictor among the four predictors: income, years, public and expenditure, will be significantly associated as state-average SAT score.

### 3. Ranking States by SAT

*In this lesson we'll learn how to perform ranking by controlling for bias selection.*

We would like to rank the states by the SAT performance. Their raw ranking without bias selection correction will put the states with lower percentage of takers and higher median class rank at the top, but it doesn't necessarily mean that these states perform best in terms of state average SAT because of the bias selection of the students taking SAT. This is because some state universities require the SAT and some require a different exam. States with a high proportion of takers probably have in state requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing this bias selection.

#### USING RESIDUALS TO CREATE BETTER RANKINGS

Thus, instead of ranking by actual SAT score, we rank the schools by how far they fall above or below their fitted regression line value, using the residuals from the reduced module with only the two controlling factors for the bias selection.

In this example we're going to fit the smaller model where we include only the controlling factors to correct for the bias which are takers and rank. Then we'll obtain the order of states by the residuals of this model and put this information into a bigger table where we add the states, the information on the residuals, along with the old ranking:

**Bias Selection:** Some state universities require the SAT and some require a competing exam. States with a high proportion of takers probably have "in state" requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing a bias.

```
## Consider the model with the two controlling factor to correct for bias
reduced.line = lm(sat ~ takers + rank)
## obtain the order of states by the residuals of the reduced model
order.vec = order(reduced.line$res, decreasing = TRUE)
## Re-order the states and create a table including state name, new and old order.
states = factor(data$order.vec, 1)
newtable = data.frame(State = states, Residual = as.numeric(round(reduced.line$res[order.vec], 1)),
oldrank = (1:50)[order.vec])
```

This slide compares the ranking with and without the correction of the bias selection:

	State	Residual	Oldrank	After controlling for selection bias, Connecticut moved from 35 <sup>th</sup> to 1 <sup>st</sup> .				
1	Connecticut	53.9	35					
2	Iowa	53.5	1	43	Arkansas	-31.2	12	
3	New Hampshire	45.8	28	44	West Virginia	-38.9	25	
4	Massachusetts	41.9	41	45	Nevada	-45.4	30	
5	New York	40.9	36	46	Mississippi	-49.3	16	
6	Minnesota	40.6	7	47	Texas	-50.3	45	
After controlling for selection bias, Mississippi moved from 16 <sup>th</sup> to 46 <sup>th</sup> .				48	Georgia	-63.0	49	
				49	North Carolina	-71.3	48	
				50	South Carolina	-98.5	50	

Note how dramatically the ranking shifts once we control for the variables takers and rank. On the left I instructed the ranking of the top six states. The old rank column provides the ranking without the correction for the bias selection. For example, after controlling for bias selection, Connecticut moved from 35th to 1st. And Massachusetts moved from 41 to 4th.

On the right, I provide the bottom eight states in the ranking using the residuals, or the correction for the bias selection. For example, after controlling for the selection bias, Mississippi moved from 16th to 46th and Arkansas slid from 12th to 43th. We could further analyze the ranks by accounting for such things as expenditure to get a sense of which states appear to make efficient use of their spending, for example.

## 4. Model Fit Assessment

*So far, we studied the estimation and statistical inference for the SAT example. To reliably make inferences on the regression coefficients and on the regression line, we need to also insure the model fit. In this lesson, we'll perform the residual analysis for this example.*

We can obtain the residuals as provided on the first R command line in which the model object reduced.line followed by the dollar sign, and then specifying the residuals.

The next command line obtains the Cook distances used to identify outliers.

The set of plots of interest are the scatter plot of the response variable, or fitted values versus residuals along with the 0 line. And now, there's also the scatter plot of the quantitative predicted variables versus the residuals, the histogram and normal probability plot, and last the plot of the Cook distance.

```
## Residual analysis for the reduced model
res = reduced.line$res
cook = cooks.distance(reduced.line)
par(mfrow = c(1,3))
plot(sat, res, xlab = "SAT Score", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(takers, res, xlab = "Percent of Students Tested", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(rank, res, xlab = "Median Class Ranking Percentile", ylab = "Residuals", pch = 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnorm(res)
qqline(res)
plot(cook,type="h",lwd=3, ylab = "Cook's Distance")
```

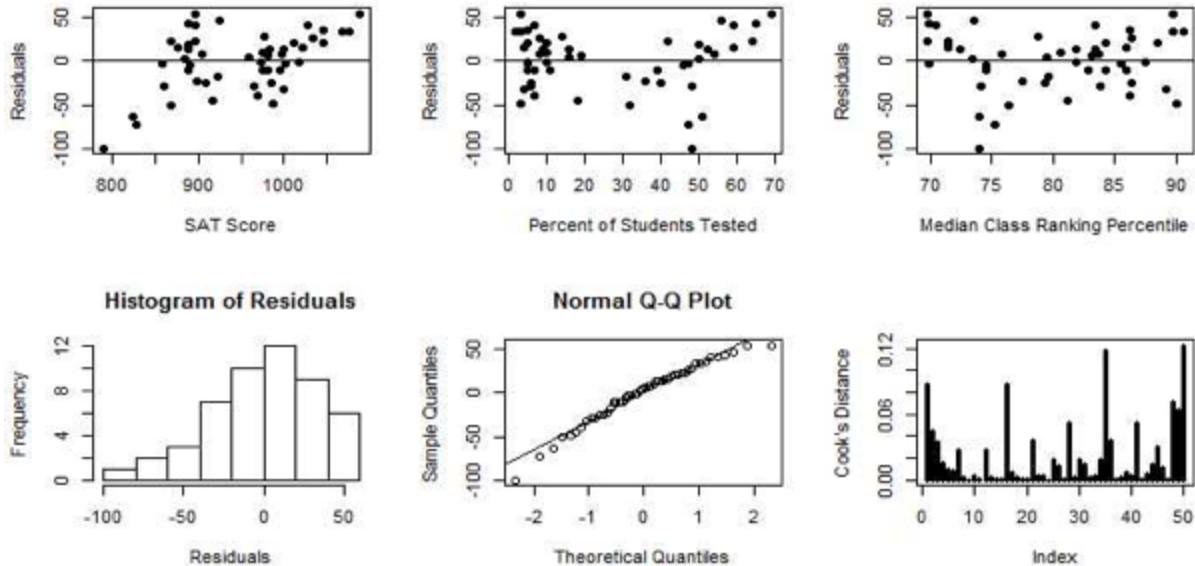
To review, we evaluate the following assumptions graphically:

**Constant variance and uncorrelated errors:** plotting the response or fitted values versus residuals.

**Linearity:** the predicting variables versus the residuals in all the scatter plots. We seek around the pattern around the 0 line.

**Normality:** histogram and the normal probability plot.

**Outliers:** we used the cook distance plot.



In the (first) plot of the residuals versus SAT scores, there is a clustering in the residuals with a grouping into two clusters, possibly an indication of correlation. It is possible that the controlling factors may not have control for the bias selection fully using the linear model.

(Second plot) Moreover, the plot of takers or the percentage of students tested, this versus residuals has higher residuals on the edges and low residuals in the center. This is an indication of nonlinearity with respect to this predictor. Thus, we'll need to transform this predictor.

Note the separation in the residual in the first plot could be a product of the fact that there is nonlinear relationship with the respect to the predictor takers, which is a controlling factor of the bias selection.

The QQ plot indicates the residual in our regression model have [heavy] tails on both the negative and the positive side observed sample quantiles. The quantiles of the residuals are larger than theoretical quantiles, the histogram points to this as well.

The residuals do not show outliers, as we might have expected. Recall that Alaska had a large expenditure, but it does not show as being an outlier influential point based on this model.

- **Transform the predicting variable: Percent of Students Tested (takers)**

- **Heavy tailed residuals**

Next, we're going to transform the predicting variable, takers, and we redo the analysis using with a log transformation:

```
regression.line = lm(sat ~ log(takers) + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	407.53990	282.76325	1.441	0.15675
log(takers)	-38.43758	15.95214	-2.410	0.02032*
rank	4.11427	2.50166	1.645	0.10734
income	-0.03588	0.13011	-0.276	0.78407
years	17.21811	6.32007	2.724	0.00928 **
public	-0.11301	0.56239	-0.201	0.84168
expend	2.56691	0.80641	3.183	0.00271 **

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.86 on 43 degrees of freedom

Multiple R-squared: 0.8919, Adjusted R-squared: 0.8769

F-statistic: 59.15 on 6 and 43 DF, p-value: < 2.2e-16

Without the transformation, Takers was not statistically significantly associated with SAT score given all other predictor variables in a model. Now with a transformation, it is statistically significant at the significance level of 0.05. The p-value is 0.02. However, now the predicting variable rank is not statistically significant anymore as compared to the model without transformation. The R-squared improves slightly from 87.8% to 89%. The standard deviation of the error term decreases slightly also:

```
regression.line = lm(sat ~ log(takers) + rank + income + years + public + expend)
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	407.53990	282.76325	1.441	0.15675
log(takers)	-38.43758	15.95214	-2.410	0.02032*
rank	4.11427	2.50166	1.645	0.10734
income	-0.03588	0.13011	-0.276	0.78407
years	17.21811	6.32007	2.724	0.00928 **
public	-0.11301	0.56239	-0.201	0.84168
expend	2.56691	0.80641	3.183	0.00271 **

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 ' 1

Residual standard error: 24.86 on 43 degrees of freedom  
 Multiple R-squared: 0.8919, Adjusted R-squared: 0.8769  
 F-statistic: 59.15 on 6 and 43 DF, p-value: < 2.2e-16

Test for statistical significance:

$\hat{\beta}_{\text{takers}}$ : p-value = 0.02

$\hat{\beta}_{\text{rank}}$ : p-value > 0.1

$\hat{\beta}_{\text{income}}$ : p-value > 0.1

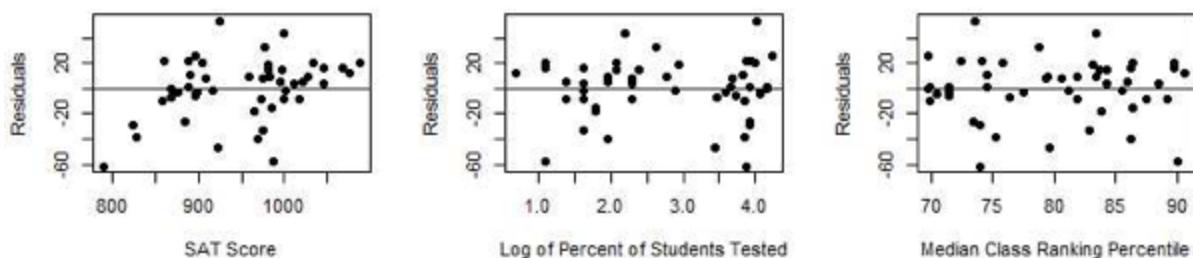
$\hat{\beta}_{\text{years}}$ : p-value < 0.01

$\hat{\beta}_{\text{public}}$ : p-value > 0.1

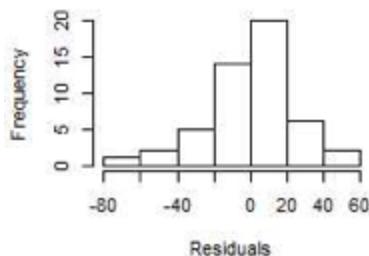
$\hat{\beta}_{\text{expend}}$ : p-value < 0.01

$$\hat{\sigma} = 24.86, n-2 = 43$$

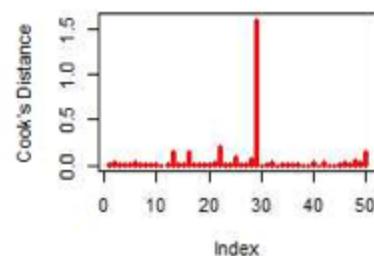
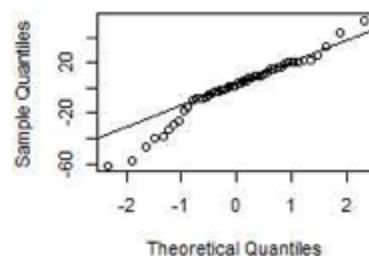
$R^2 \sim 89\%$  variability explained



Histogram of Residuals



Normal Q-Q Plot



The linearity assumption does hold now for the true predicting variables.

Now we can look at the log of percent of student tested versus residuals. And we don't see the pattern from the model without transformation anymore.

However, we still see some clustering in the residuals versus SAT scores, although it's less so than for the model without transformation.

The distribution of the residuals is still heavy tailed.

Now we also see one Cook Distance standing out: the value corresponding to Alaska. So now for this model with transformation, we see with the outlier, the values for the Alaska is influential point. And the reason is that, the predicting variable expenditure is now strongly associated to the SAT score, because the p-value is small. So we see a change also in the statistical significance for expenditure.

To review:

- **Transformation has improved on the linearity assumption**
- **Heavy tailed residuals remains**
- **Cook's Distance: Alaska is an outlier/influential point for the model**

The transformation has improved the linearity assumption. We still have heavy tailed residuals, and the cook distance shows Alaska is an outlier and influential point for the model. You would need to study this outlier further for a more comprehensive analysis.

Let's now review the findings based on this analysis:

### **State SAT Performance: Findings**

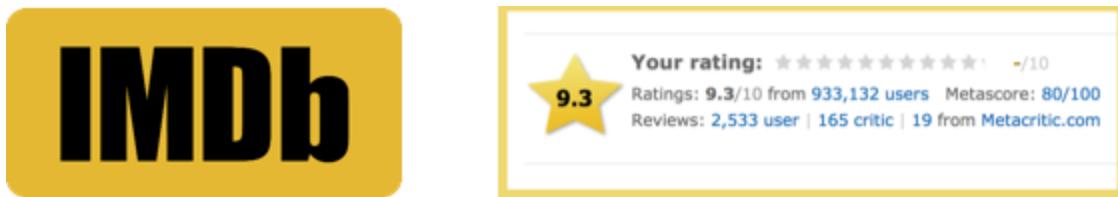
- Given all other predictors in the model, percent of students taking SAT from a public school and family income of test takers, are not statistically significantly associated to SAT score.
- Given all other predictors in the model, with 100,000 increase in the expenditure on secondary school results, we see only a 2.56 points increase in the SAT score.
- Given all other predictors in the model, one additional year that test takers had in social sciences, natural sciences, and humanities leads to 17.2 points increase in SAT score.
- The predictors in the model explain close to 90% of the variability in SAT score.
- We find that relationship between state-average SAT score and the percentage of students taking SAT to be nonlinear for this example.
- Ranking changes significantly after controlling for the bias selection factors. For example, Connecticut moves up to be the 1st from 35th, Massachusetts to 4th from 41st, and New York to 5th from 36th.

## 3.5 Case Study: Prediction of IMDb Movie Ratings

### 1. Exploratory Analysis

*In this lesson, we'll illustrate multiple linear regression with a prediction of IMDb movie ratings. In this lesson, I will introduce this example, along with exploratory analysis based on visual analytics and numerical summaries for both quantitative and qualitative variables.*

The Internet Movie Database, or IMDb for short, is the most used website to access comprehensive information for movies and TV content:



With over 65 million registered users, it remains one of the most trustable sources of data. It allows users to submit ratings and reviews. The most popular feature of this site is its movie ratings, being heavily influenced by the number of people watching movies on the screen or rent out DVDs. Analyzing the IMDb ratings, particularly identifying the factors predicting the ratings, has been a popular problem over the past couple of years. The source of the data is IMDb's website. The response variable in this example is the rating provided by the IMDb, scaled with values between 0 to 100. The data set /website provides many different qualitative and quantitative predicting variables. I selected the most relevant ones. Among the quantitative variables, including this analysis, include number of votes for the movie on the IMDb platform, duration of the movie, gross earnings scaled in 1000's, and total budget in millions:

**The response variable is:**

**Y** = the rating provided by the IMDb, scaled with values between 0 to 100

**Quantitative predicting variables are:**

**X<sub>1</sub>** = Number of votes for the movie on the IMDb platform

**X<sub>2</sub>** = Duration of the movie

**X<sub>3</sub>** = Gross earnings scaled in 1000's

**X<sub>4</sub>** = Total budget in millions

Several qualitative variables are also considered: including release year, rating of the movie's suitability for a certain audience based on its content, language, genre, director rating, actor rating, movie awards:

**Qualitative predicting variables are:**

**X<sub>5</sub>** = Release year (between 2010-2014)

**X<sub>6</sub>** = Rating of a film's suitability for certain audiences, based on its content (G, PG, PG-13, R)

**X<sub>7</sub>** = Language: English (1) and Other languages (0)

**X<sub>8</sub>** = Genre: Action (1), Documentary (2), Comedy (3), Horror, Sci-Fi (4)

**X<sub>9</sub>** = Director Rating: Awarded (1), Nominated (2), None (3)

**X<sub>10</sub>** = Actor Rating: Classified into two groups based on their performance, with 0 for low ranking and 1 for high ranking.

**X<sub>11</sub>** = Movie Awards: Awarded (1), Nominated (2), None (3)

Why do we consider 'Year' as a qualitative variable?

I considered year as a qualitative variable since we have only few years of data. In the case when there would be more years, then this could be considered as a quantitative variable. Generally, when a predicted variable is not very granularly observed, like year in this case, it can be transformed into a qualitative variable.

This slide shows the data processing of the quantitative variables. We begin with reading the data from the file training.csv which, includes a header and a separator between observations is comma:

```
## Read data using read.csv  
data = read.csv("training.csv", header=TRUE, sep=",")
```

For this example, I consider only a subset of the movies available on IMDb website, only 100 movies. To have a rating between 0 and 100, I multiply the response data values with 10:

```
## how many observations?  
dim(data)[1]  
[1] 100  
  
## Response Variable: scaled between 0 and 100  
imdb=data$imdb*10
```

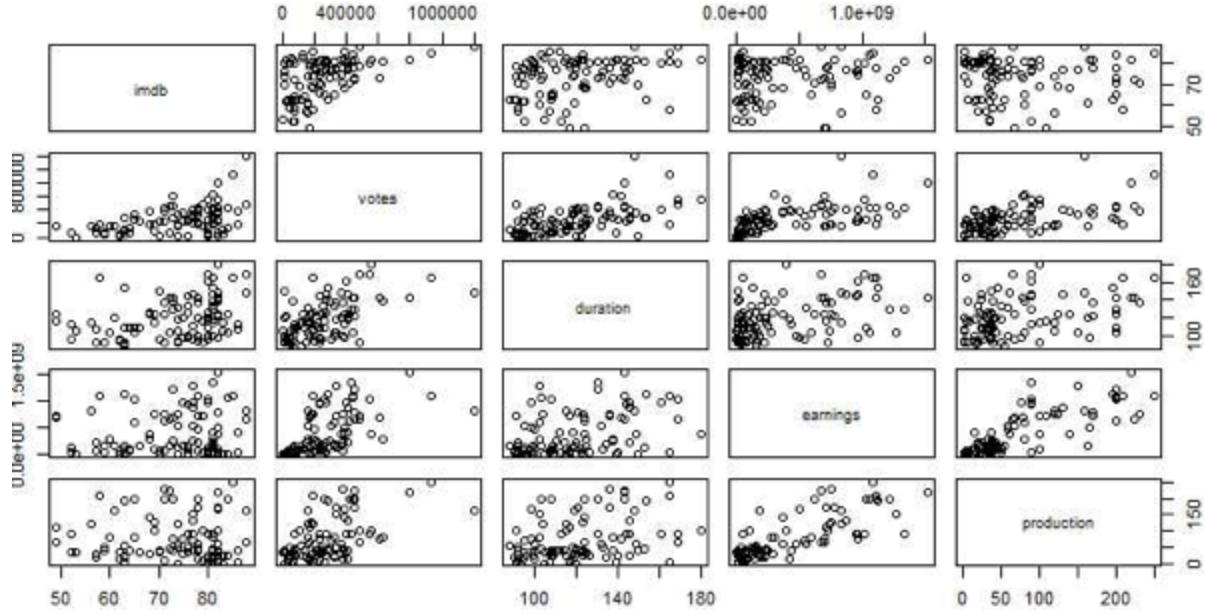
Furthermore, I extracted the quantitative variables from the data matrix using the dollar sign:

```
## Quantitative Predicting Variables  
# Number of imdb user votes for the movie  
votes=data$votes  
# The duration of the movie  
duration=data$time  
# Gross earnings in 1000's  
earnings=data$boxoffice/1000  
# Total budget in millions  
production=data$productionbudget
```

I then created a data frame which is used to obtain the scatter plot matrix of the response variable and the four predicting variables:

```
## Exploratory Data Analysis for Quantitative Data  
# Scatter plot matrix  
pqmat=as.data.frame(cbind(imdb,  
votes,duration,earnings,production))  
plot(pqmat)
```

This is the scatter plot matrix:



And what I'm presenting here on the first row are the scatter plots of the response variable, the IMDb writings versus the four predicting variables, along with the scatter plots of pairs of the predicting variables. There's some clear patterns of the relationship between the IMDb ratings and the predicting variables.

There is a strong relation between rating and number of votes.

There's also a weak relationship with duration of the movie.

There are some correlation among the predicting variables, for example, between the number of votes and duration, and between earnings and total budget, as actually expected.

Processing the qualitative variables is more work as you can see, the code for R is long:

```



```

First, if the variables are coded numerically rather than their descriptive labels, I recommend to use the descriptive labels. This is what I did for this R code. For example for genre, the variable is coded with numerical values 1 to 4. But I replaced this labeling with the description of the genre including action, documentary, comedy, and horror, science fiction.

If you want to explore whether this kind of quantitative variables explain some of the variability in the response variable, you can use a side by side boxplot similar to the analysis of variance to ANOVA. We compare the variability of the response variable across their labels of each of this qualitative variables. The box plot commands of this slide do just that:

```

## Exploratory Data Analysis for Qualitative Data
par(mfrow=c(2,3))
boxplot(imdb~year,col="blue",main="Year")
boxplot(imdb~rating,col="red",main="Rating for Audience")
boxplot(imdb~language,col="green",main="Language")
boxplot(imdb~genre,col="purple",main="Genre")
boxplot(imdb~rtdirector,col="purple",main="Director Awards")
boxplot(imdb~rtactor,col="grey",main="Actor Performance Rating")

```

To review:

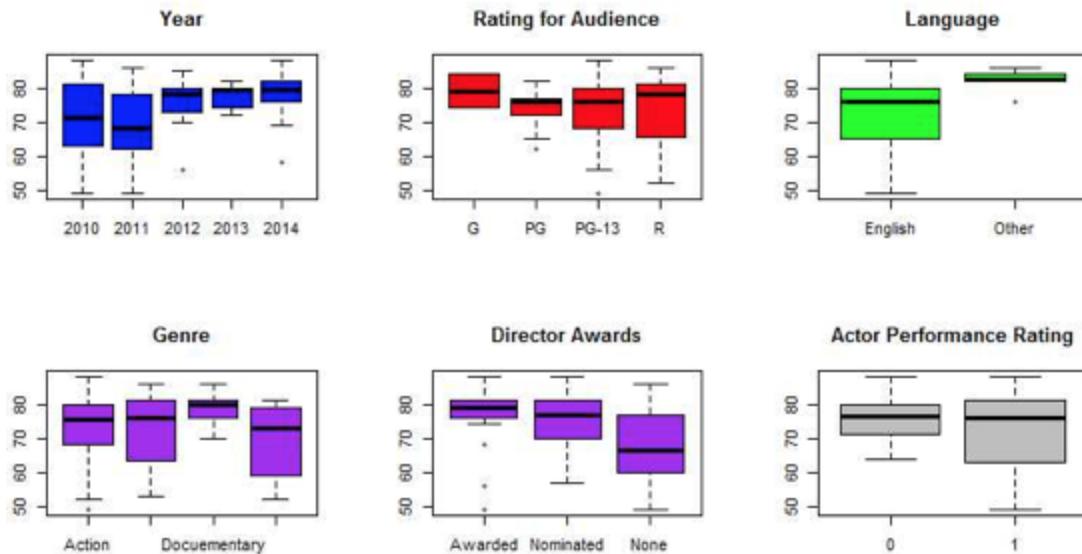
## Qualitative Variables

- For better visual analytics and reference, inform quantitative variables using their specific class names
- Transform into a categorical variable using the `as.factor()` R command

First, for better visual analytics and reference, remember to inform the qualitative variables using very specific class or label names, rather than numeric or other less meaningful labeling.

Second, do not forget to transform each one of it into a categorical variable using the `factor` command in R. These are important first steps in dealing with qualitative variables.

Here are the box plots for six qualitative variables in our analysis:



The response variable does vary with year and the award number of directors.

There is less between group variability with respect to other predicting factors.

We can complement the visual analytics with numerical summaries of the data. For qualitative variables, we can use the correlation analysis, the `cor()` command in R can be used for obtaining the correlation coefficients for multiple factors for multiple variables by using a matrix as an input in this command. I use here the command `round` so that I only up the correlation coefficients with only two digits:

```

## Correlation, Multicollinearity
## Quantitative data
round(cor(pqmat),2)
      imdb  votes duration earnings production
imdb    1.00   0.46    0.33    0.09   -0.04
votes    0.46   1.00    0.52    0.61    0.56
duration  0.33   0.52   1.00    0.42    0.38
earnings  0.09   0.61    0.42   1.00    0.80
production -0.04   0.56    0.38    0.80   1.00

```

As pointed out before, there is a linear correlation between rating and number of votes, and between rating and the duration of the movie. There is also a strong linear correlation between earnings and total budget for the movie production.

For numerical summaries between the response and individual qualitative predicting variables, we can use the ANOVA to test whether the means in the response are statistically different across the categories of the quantitative variable:

```

## Qualitative data: Response vs. Predicting Variables
summary(aov(imdb~year))
  Df Sum Sq Mean Sq F value Pr(>F)
year     4 1513    378.3  4.931  0.00118 **
Residuals 95 7288    76.7

```

For example, the p-value for equal means of the rating across years is small, suggesting that the means of the IMDb ratings are statistically significantly different across years. And thus, a difference in rating can be due to the release year of the movie.

For numerical summaries between any two qualitative predicting variables, we can use the table() command in R, which gives us the number of observations by groups:

```

## Qualitative predicting variables
table(rtdirector,awards)
  awards
  rtdirector Awarded Nominated None
Awarded        13       10      4
Nominated      10       23     10
None           1        10     19

```

For example, there are 13 movies in our sample with both movie and director awards. But there are ten movies where the director was nominated but not awarded and the movie was awarded.

To quantify statistically whether there are differences in the proportion across the cells in this table, we can use the chi-squared test:

```
chisq.test(rtdirector,awards)
Pearson's Chi-squared test

data: rtdirector and awards
X-squared = 26.192, df = 4, p-value = 2.895e-05
```

For this example, the p-value of the test for the two categorical variables is very small, indicating that there are differences in the proportion across the cells in the table.

In summary:

#### **Exploratory Analysis using Numerical Summaries:**

- Correlation captures linear dependence between quantitative variables
- ANOVA can be used to assess whether the means of the response variable are statistically different with respect to a quantitative variable
- The 'table' command in R provides the contingency table between any two quantitative variables
- The Pearson chi-square test can be used to test for 'correlation' between quantitative variables

If we use the correlation coefficient to evaluate the relationship between two quantitative variables, we need to keep in mind that the correlation coefficient only quantifies linear dependence, linear relationships.

To evaluate whether a qualitative variable has explanatory power for the response variable, we can use the simple ANOVA model.

Moreover, we can also evaluate the relationship between any two qualitative variables using the table command and the Pearson chi-squared test.

## 2. Regression Analysis

*The lecture is multiple linear regression, and I'll provide a regression analysis on the prediction of IMDB movie ratings. Particularly, we'll see in this lesson the implementation of the statistical analysis with the focus on interpretation.*

To implement the regression model in R with IMDB rating data, we are going to use the LM() command in R with the response variable IMDB rating on the left, and the

predicting variables summed up on the right. Note that the qualitative variables have already been converted into categorical variables. And thus, the LM command will consider only the first **k-1** dummy variables for a categorical variable with K levels.

The output is too long and I'm only providing the estimated regression coefficient and statistical inference on those coefficient for the four qualitative variables for the year factor variable. Please practice with this data example to see the entire output:

```
fit=lm(imdb ~ votes+duration+earnings+production+year+rating+language+
       genre+rtdirector+ractor+awards)
summary(fit)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.466e+01	5.65e+00	13.212	<2e-16 ***
votes	2.672e-05	4.27e-06	6.250	1.95e-08 ***
duration	5.219e-03	3.495e-02	0.149	0.881673
earnings	-6.462e-09	2.549e-09	-2.535	0.013209 *
production	-1.819e-02	1.580e-02	-1.151	0.253341
year2011	-3.958e-01	1.500e+00	-0.264	0.792608
year2012	2.452e+00	1.899e+00	1.291	0.200333
year2013	4.696e+00	1.734e+00	2.707	0.008308 **
year2014	7.212e+00	1.841e+00	3.918	0.000189 ***

Test for statistical significance:  
 $\hat{\beta}_{\text{votes}}$ : p-value < 0.01  
 $\hat{\beta}_{\text{duration}}$ : p-value > 0.1  
 $\hat{\beta}_{\text{earnings}}$ : 0.01 < p-value < 0.1  
 $\hat{\beta}_{\text{production}}$ : p-value > 0.1  
 $\hat{\beta}_{\text{year 2011 vs 2010}}$ : p-value > 0.1  
 $\hat{\beta}_{\text{year 2012 vs 2010}}$ : p-value > 0.1  
 $\hat{\beta}_{\text{year 2013 vs 2010}}$ : p-value < 0.01  
 $\hat{\beta}_{\text{year 2014 vs 2010}}$ : p-value < 0.01

Residual standard error: 5.125 on 79 degrees of freedom  
Multiple R-squared: 0.7642, Adjusted R-squared: 0.7045  
F-statistic: 12.8 on 20 and 79 DF, p-value: < 2.2e-16

$$\hat{\sigma} = 5.125, n-2 = 79$$

$R^2 \sim 76.4\%$  variability explained

Among the four quantitative variables only number of votes and earnings have statistically significant coefficient given all of the predictive factors in the model.

For the categorical variable year the coefficient for 2013 and 2014 - only those coefficients are statistically significant. Thus the means of these two years are statistically significantly higher than the baseline year of 2010 and it's higher because the coefficients for those two years are positive. Generally, when we have a qualitative predicting variable with multiple levels and only some of the regression coefficients corresponding to its dummy variables are statistically significant, it is good practice to include all dummy variables of that categorical variable. As an example, Year 2011 and the year 2012, the coefficient of those two dummy variables are not statistically significant but you should include all four variables 2011, 2012, 2013 and 2014 in order to account for the variability of how year impacts or explains the variability in IMDB rating.

The R square is 0.76. Thus, about 76% of the variability in IMDB rating is explained by the predicting variables considered for this model.

I'll remind you once more that when you perform a linear regression analysis with multiple qualitative predictors, you'll end up with a large number of predicting variables. For example, here we have 7 categorical predictive variables, but that translates into 16 predicting variables or 16 dummy variables. Thus, it can be challenging to perform a regression analysis with a large number of qualitative variables. One approach would be to group some of the dummy variables, some of the groups, in order to reduce the number of dummy variables.

### Linear Regression in R:

- If you have qualitative variables you must either convert them using the 'as.factor' command or you will need to specify dummy variables and add them all (except one if the model has intercept) to the model
- Models with many qualitative variables have many parameters because each one of it will introduce several dummy variables as predicting variables

In the next slide, I will provide more insight in coding dummy variables, or coding qualitative variables in the regression analysis in R.

### CODING DUMMY VARIABLES IN R

So let's take a closer look at the multiple ways on how to model a qualitative predictive variable in R.

The first approach on this slide is by converting the qualitative variable into dummy variables:

```
### Create Dummy Variables
genre = data$genre
genre.1 = rep(0,length(genre))
genre.1[genre==1] = 1
genre.2 = rep(0,length(genre))
genre.2[genre==2] = 1
genre.3 = rep(0,length(genre))
genre.3[genre==3] = 1
genre.4 = rep(0,length(genre))
genre.4[genre==4] = 1
## Include all dummy variables without intercept
fit.1 = lm(imdb~genre.1+genre.2+genre.3+genre.4-1)
```

For example, for genre we have 4 different [labels] for dummy variables. If we do not include an intercept, as in the first model fit (fit.1), then we can include all four dummy variables as predicting variables, the fitted model is provided:

`summary(fit.1)`

	Estimate	Std. Error	t value	Pr(> t )
genre.1	72.524	1.433	50.62	<2e-16
genre.2	78.923	2.575	30.65	<2e-16
genre.3	73.051	1.487	49.13	<2e-16
genre.4	69.500	3.791	18.33	<2e-16

As you see in the R output, each dummy variable has its individual role in the output since it's a predicting variable on its own. The output does not provide a row for the intercept, since we specified, since this is a no intercept model.

A second approach is to consider a model with intercept and these will all include only three dummy variables (fit.2):

`## Include 3 dummy variables with intercept  
fit.2 = lm(imdb~genre.1+genre.2+genre.3)`

The output of this model includes an intercept, and the first three dummy variables as provided in the model. In this example, we chose to have the last dummy variable, or the fourth genre type as the baseline:

`summary(fit.2)`

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	69.500	3.791	18.334	<2e-16
genre.1	3.024	4.052	0.746	0.4574
genre.2	9.423	4.583	2.056	0.0425
genre.3	3.551	4.072	0.872	0.3853

A third approach (fit.3) is to convert the genre categorical variable into a factor in R, and fit the model with a genre factor rather than individual dummy variables. For this R select genre 1 as the baseline:

`## Use categorical variable  
genre = as.factor(data$genre)  
fit.3=lm(imdb~genre)`

As we can see in the output, we have genre 2, 3, and 4 dummy variables in the model, but not genre 1:

`summary(fit.3)`

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	72.5238	1.4328	50.619	<2e-16
genre2	6.3993	2.9470	2.171	0.0324
genre3	0.5275	2.0648	0.255	0.7989
genre4	-3.0238	4.0524	-0.746	0.4574

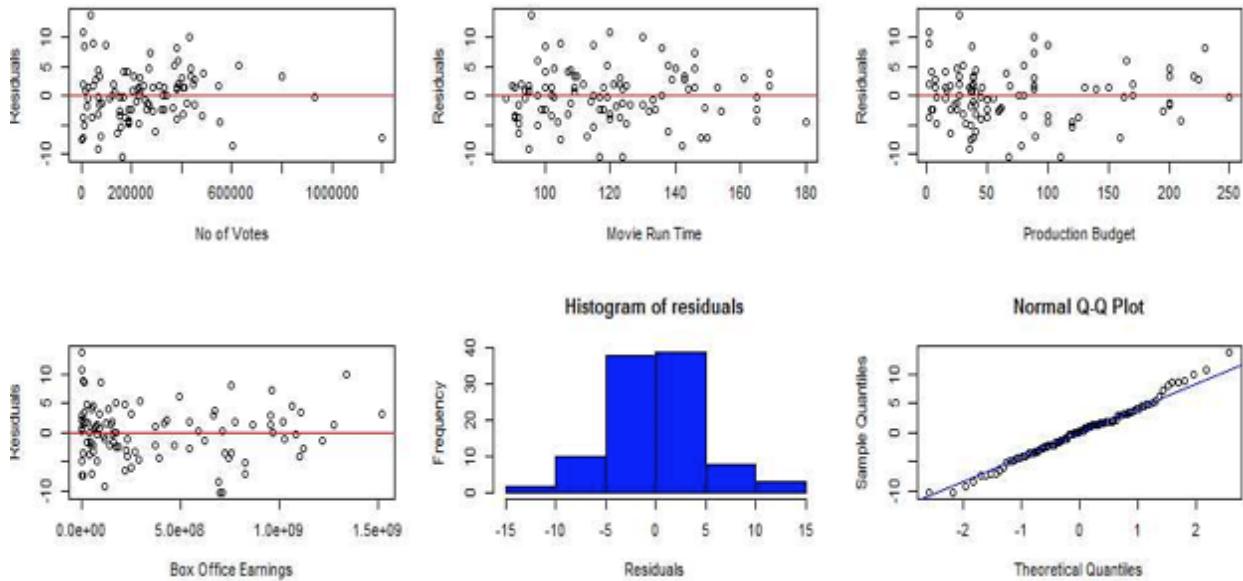
It is important to remember that R chooses the first label as the baseline category or label and compare the last K - 1 dummy variables to the baseline, when you input the categorical variable as a factor. If a different category should be the baseline, then you will need to either define the dummy variables and include them as separate predicting variables in the LM command, or you may change the labeling in such a way that the first label [will] be corresponding to the baseline.

Also be careful when using a model without the intercept - interpreting the regression coefficients for qualitative variables will be different since there is no baseline comparison.

### Coding Dummy Variables

- R sets the “first” class as being the baseline; if a different class is the baseline, either use dummy variables or change ‘contr.treatment’
- Be careful when using a model without intercept in R!

Using a similar R code as for the SAT example, I'm performing here a residual analysis for this data example. I'm not providing the R code here, but it is available in the lecture model for you to practice. The resulting plots are on the slide:



The first four plots present a scatter plot of each of the predicting variables versus the residuals, used to evaluate linearity.

These four plots do not show any specific trend: an indication that the linearity assumption holds.

We also do not see any clustering of the residuals thus the assumption of uncorrelated errors holds. The variance of the residuals don't vary with changes in the predicting variables.

This plot should be complemented by the scatter plot of the fitted values, or response, versus residuals to evaluate constant variance.

Last, there are no departures from normality assumption, as provided by the histogram or the normality probability plot.

Let's interpret the model, particularly the estimated regression coefficients. The general interpretation is as follows:

```

summary(fit)
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.466e+01 5.651e+00 13.212 <2e-16 ***
votes        2.672e-05 4.275e-06 6.250 1.95e-08 ***
duration     5.219e-03 3.495e-02 0.149 0.881673
earnings     -6.462e-09 2.549e-09 -2.535 0.013209 *
production   -1.819e-02 1.580e-02 -1.151 0.253341
year2011     -3.958e-01 1.500e+00 -0.264 0.792608
year2012      2.452e+00 1.899e+00 1.291 0.200333
year2013      4.696e+00 1.734e+00 2.707 0.008308 **
year2014      7.212e+00 1.841e+00 3.918 0.000189 ***
.....
Residual standard error: 5.125 on 79 degrees of freedom
Multiple R-squared:  0.7642, Adjusted R-squared:  0.7045
F-statistic: 12.8 on 20 and 79 DF, p-value: < 2.2e-16

```

- **Coefficient Interpretation:** the expected change in the response for a one unit change in the predictor while holding all other predictors fixed.
- **Example:** The coefficient for 'votes' is .00002672. This means that if we fix the other predictors, for each additional vote, we expect the IMDB rating to increase by .00002762 points.

An estimated coefficient is the expected change in the response for one unit change in the predictor variable while holding all other predictors fixed.

For example, if we take voters as a predictive variable of interest, the estimated coefficient is as provided in output. This means that if we fix all other predictors for each additional vote, we expect the IMDB rating to increase by 0.00002762 points.

Here I'm highlighting once more that **in a context of multiple regression we always have to add the interpretation of the coefficients in the context that other predictors are in the model.**

#### STATISTICAL SIGNIFICANCE: MARGINAL VS CONDITIONAL

Let's take a closer look at the interpretation between marginal and conditional models. We'll further explore the difference between condition and marginal statistical inference using three of the variables.

We begin by fitting the full model, called here the conditional model, since we focus on the statistical inference on the regression coefficients given other predictors variables in a model:

```

Estimate Std. Error t value Pr(>|t|)
.....
duration      5.219e-03 3.495e-02 0.149 0.8816
earnings     -6.462e-09 2.549e-09 -2.535 0.0132
rtdirectorNominated -5.389e-01 1.443e+00 -0.373 0.7098
rtdirectorNone   -1.398e+00 1.678e+00 -0.833 0.4070
.....
summary(aov(imdb~rtdirector))
  Df Sum Sq Mean Sq F value Pr(>F)
rtdirector  2  1353   676.3   8.807 0.000306***
Residuals  97  7449    76.8
summary(lm(imdb ~ duration))
Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.46913  4.94644 11.416 <2e-16 ***
duration    0.14132  0.04066  3.476  0.00076***
summary(lm(imdb ~ earnings))
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.259e+01 1.313e+00 55.285 <2e-16 ***
earnings   2.057e-09 2.379e-09  0.865  0.389

```

The focus is going to be on three variables: duration of the movie, earnings, and the movie director award factor.

From the conditional model, duration and the movie director award factor are not significantly associated to the movie rating given all other predictors in the model. Earning is statistically significant for significance level of 0.05 but not for 0.01, the p value is 0.013:

**Conditional:**

$\hat{\beta}_{duration}$ : p-value > 0.1  
 $\hat{\beta}_{earnings}$ : 0.01 < p-value < 0.1  
 $\hat{\beta}_{director\ nominated\ vs\ awarded}$ :  
p-value > 0.1  
 $\hat{\beta}_{director\ none\ vs\ awarded}$ : p-value > 0.1

If we were to consider the marginal model with only the movie director award variable we can fit the regression model with one categorical variable using the ANOVA or the aov R command. Marginal, there is a significant statistical difference in the mean ratings across different types of awarding of the movie director because the p value of the F-test is very small:

**Marginal:** F-test: p-value < 0.01  
At least one statistically significant:  
 $\beta_{nominated\ vs\ awarded}$  ,  $\beta_{none\ vs\ awarded}$  ,  
 $\beta_{nominated\ vs\ none}$

Note that this variable is not statistically significant in the conditional model. Similarly, when fitting the duration and earnings, particularly variables alone, so in a marginal models, we see a difference of the statistical significance from the conditional model. For duration, the margin association with a movie rating is statistically significant whereas the conditional association is not. In contrast, for earnings, marginal association with a movie rating is not statistically significant whereas the conditional association is:

**Marginal: F-test: p-value <0.01**

At least one statistically significant:

$\beta_{\text{nominated vs awarded}}$ ,  $\beta_{\text{none vs awarded}}$ ,

$\beta_{\text{nominated vs none}}$

Thus:

- If the association of a predicting variable to the response is statistically significant under conditional model, it does not mean that it is so under the marginal model and vice versa.
- Moreover, if a qualitative variable shows no difference in mean response under the conditional model, it does not mean that it is so under the marginal model.
- To remember, always interpret the statistical significance in a multiple regression model conditionally.

### 3. Prediction and Findings

Summary: I presented prediction of IMDb ratings for additional set of movies based on the model fitted on the initial movie data we began with. I also presented different measures for prediction accuracy, and I provided some insights on utility of those measures. And I concluded with findings on the study.

*In this lesson, we'll focus on the prediction of the IMDb movie ratings.*

Particularly for this lesson, we'll use a set of 21 movies for which we'll predict the ratings given the fit of the model based on a data of 100 movies.

So in addition to the training data used to fit the model, now we have a set of 21 movies for which we would like to predict the rating using the fitted model. For this, we first read the testing data including information of predicting variables of the 21 movies.

Next, we'll need to prepare this new data for the 21 movies in a similar fashion as we did for the training data. I'm not repeating the commands used to process the training data here since they are the same for the testing data. Note that the R code is available as part of this lecture:

```
## Read New Data (Test Data)
testdat = read.csv("Testing.csv",header=TRUE,sep=",")
dim(testdat)[1]
[1] 21
## Prepare the new data the same as the training data
## Response Variable: scaled between 0 and 100
nimdb=testdat$imdb*10
## Quantitative Predicting Variables
# Number of imdb user votes for the movie
nvotes=testdat$votes

.....
## Qualitative Predicting Variables
# Rating of a film's suitability for certain audiences, based on its content
nrating=as.factor(testdat$rating)

.....
## Write the new data into a data frame
newdat = data.frame(votes=nvotes, duration=nduration, earnings=nearnings, production=nproduction,
year=nyear, rating=nrating, language=nlanguage, genre=ngenre, rtdirector=nrtdirector,
ractor=nrtactor, awards=nawards)
# Specify whether a confidence or prediction interval
predict(fit,newdat,interval=c("prediction"))
```

Once we process the testing data, we next create a data frame for the testing data to input in the **predict command** in R. Note that we need to process the testing data to contain the same set of predictors as the training data used for fitting the model to be

used in the predict R command. If the set of predicting variables are not the same as the ones that we used for the fitted model, you'll get an error message from R.

The output of the predict command is as provided here:

Prediction Output:			
	fit	lwr	upr
1	76.29411	64.27794	88.31028
2	62.73839	50.21367	75.26311
3	72.03305	59.41553	84.65057
4	76.75066	64.15607	89.34525
5	80.41676	67.90456	92.92896
6	72.58612	59.66305	85.50919
7	62.13416	49.43550	74.83281
8	73.93309	61.64041	86.22576
9	71.05206	58.81055	83.29357
10	61.86840	49.15973	74.57707
11	74.57625	62.18013	86.97238
.....			

I'm only presenting the prediction output for 11 movies out of 21 due to the space limitation on this slide. The output consists of three columns. The predictions are in the first column, and the lower and upper bounds for the prediction intervals are in the second and third column. For example, for the first movie, the predicted rating is around 76 with a lower bound of approximately 64 and an upper bound of roughly 88.

#### EVALUATING PREDICTION ACCURACY: COMMONLY REPORTED MEASURES

But how good are those predictions? We can compare the predictions with observed responses, with the observed ratings. In the real world, we do not have the observed responses at that time of predictions, and thus we cannot evaluate the prediction accuracy of a model as we will do here. But here, we first pretend we do not have the observed responses, the observed ratings, and predict given the values of the predicting variables for the 21 movies. Now we assume that we've seen the ratings and will compare the predicted ratings with those that are observed for evaluating the accuracy of the prediction.

Generally, the question of how good is the prediction comprises two separate aspects. Firstly, measuring predictive accuracy per se as we'll do in this example.

Secondly, comparing various forecasting models, which we'll not do in this example. The most common reported measures of predicting accuracy are:

- MSPE = mean squared prediction error = the sum of the square differences between predicted and observed
- MAE = mean absolute prediction errors = the sum of the absolute values of the differences
- MAPE = percentage measure such as the mean absolute percentage error = the sum of the absolute values of the differences scaled by the observed responses
- PM = precision error = the ratio between MSPE and the sum of square differences between the response and the mean of the responses

The three measures can be computed using or as provided on the slide:

```
## Save Predictions to compare with observed data
predicttestdata = predict(fit,newdat,interval=c("prediction"))
imdb.pred = predicttestdata[,1]
imdb.lwr = predicttestdata[,2]
imdb.upr = predicttestdata[,3]

### Mean Squared Prediction Error (MSPE)
mean((imdb.pred-nimdb)^2)
### Mean Absolute Prediction Error (MAE)
mean(abs(imdb.pred-nimdb))
### Mean Absolute Percentage Error (MAPE)
mean(abs(imdb.pred-nimdb)/nimdb)
### Precision Measure (PM)
sum((imdb.pred-nimdb)^2)/sum((nimdb-mean(nimdb))^2)
```

<b>Prediction Accuracy:</b> MSPE = 23.69 MAE = 3.96 MAPE = 0.055 PM = 0.058 All new observations fall within the prediction CI
---

Just to give you some insights on which one are better than others:

MSPE is appropriate for evaluating prediction accuracy for a linear model estimated using least squares, but it depends on the scale of the response data, and thus is sensitive to outliers.

MAE is not appropriate for evaluating prediction accuracy of a linear model estimated using least squares, and depends on scale, but it is robust to outliers.

MAPE is not appropriate to evaluate prediction accuracy of a linear model estimating using least squares, but it does not depend on scale and it is robust to outliers.

Last, precision error is the best of all because it is appropriate for evaluating prediction accuracy for the linear models estimating using least squares and it does not depend on

scale. The precision measure is reminiscent of the regression r squared. It can be interpreted as a proportion of the variability in the prediction versus the variability in the new data. *While MAE and MAPE are commonly used to evaluate prediction accuracy, I recommend using the precision measure.*

#### EVALUATING PREDICTION ACCURACY: FIT WITHIN PREDICTION INTERVALS

Another approach for evaluating a prediction is by checking whether the observed values fall within the prediction intervals, as in the last R command provided here. For this data example, the accuracy measures are provided here:

```
### Does the observed data fall in the prediction intervals?  
sum(nimdb<imdb.lwr)+sum(nimdb>imdb.upr)
```

The precision measure is 0.058, which means that the proportion between the variability in the prediction and the variability in the new data is 0.058. That is the variability in the prediction is significantly smaller than the variability in the data. The closer this is to zero, the better the prediction is.

*Last, we also note that all observed responses fall within the prediction interval.*

So let me summarize the findings based on this study:

## Prediction of IMDb Movie Ratings: Findings

- Duration and Budget ~ IMDb Rating;
- Each additional 10,000 votes => an increase of 2.67 in IMDb score.
- Movies released in 2013 are rated 4.696 points > movies produced in 2010. Similarly, movies released in 2014 are rated 7.212 points > those released in 2010. Other years show no statistically significant difference from 2010. This seems to indicate a recency bias in how movies are rated on IMDb;
- The predicting variables in the model explain close to 76% of the variability in IMDb scores.
- The model also provides good predictions.

Given all other predictors in the model, the duration of the movie and the total budget for production of the movie are not statistically associated to IMDb rating.

Given all of the predictors in a model, each additional 10,000 votes result in an increase of 2.67 in IMDb score. That is, popular movies tend to be voted more often.

Given all other predictors in the model, movies released in 2013 are rated 4.696 points higher than movies released in 2010. Similarly, movies released in 2014 are rated 7.212 point higher than those released in 2010. Other years show no statistically significant difference from 2010. This seems to indicate a bias in terms of the time since release in how movies are rated on IMDb.

The predictors in a model explain close to 76% of the variability in IMDb scores.

We'll also find that the model provides good predictions of the IMDb rating.

# Unit 4: Generalized Linear Models

## 4.1 Logistic Regression: Basic Concepts and Estimation

### 4.1.1. Introduction

This lesson introduces the logistic regression model, which is commonly used for modeling binary response data. We will focus on the basics of this model, particularly the definition of the model, and its assumptions.

Regression models are usually thought of as only being appropriate for continuous response variables. Is there any situation where we might be interested in prediction of a categorical response variable? The answer is most definitely yes. Here are a few examples:

- How likely is it that users will like a new layout of our website?
- Will customers leave a wireless service at the end of their subscription?
- What financial characteristics can be used to predict whether or not a business will go bankrupt?

In all these examples, we'd like to explain, or predict, yes/no questions—that is, response variables that are binary, zero or one, yes or no, winter or summer, small or big, leave or not leave. Binary response variables are very common in practice, and in this lecture, we'll learn how to model to explain or to predict binary response variables.

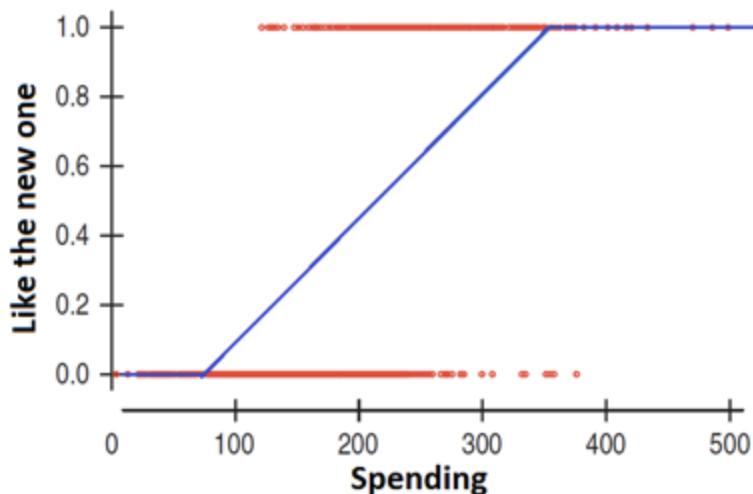
There is a fundamental difference between the yes/no questions and the kind of regression questions we have used to asking so far. With modeling the value of a response variable, we model the probability of yes. What is the same simple way to do that? We could simply do an ordinary least squares regression, as we did so far, treating the zero/one variable as the target, the response variable, but does this make sense? Let's review again the linear regression model we learned in the previous lectures. In this model, data consist of observations for a response variable and a set of predicted variables. The model has a linear relationship in the predicted variables, plus an error term.

The assumptions are that the error terms are normally distributed with mean zero, and constant variance, and that they are independent. The normality assumption also

implies that the response variable is normally distributed. But, in the examples I provided in a previous slide, the response variable is a binary variable, and thus, not normally distributed. Thus, we'll not be able to apply the regression model we learn in the previous lectures, because we don't have the normality assumption.

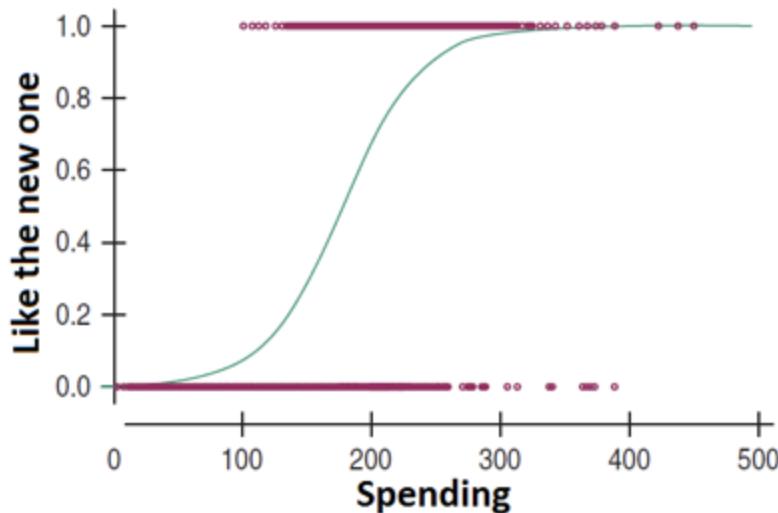
Let's go back to another yes/no example: Uber. Uber recently changed its logo. And you are asked to model whether Uber users will like the new logo based on how much they spent in the last three months using Uber. The scatter plot of the response variable, whether a user likes the new logo versus the spending of that user, is provided in the red, in this plot. If you were to fit a linear regression model to this data, then we would fit the blue line.

## What is Wrong with Linear Regression?



But the customers will not behave like this. They will behave more like an s-shape. For example, for lower than 200, spending of 200, we may believe that each additional dollar in the spending is associated with the fixed constant increase in the probability of liking the new logo.

## S-shaped Curve



However, for spending around 300, this is no longer true. At that point, the probability of liking the new logo is so high that an additional dollar on the spending adds little to the probability of liking the logo. In other words, the probability curve, as a function of spending, levels off for high values of spending.

This situation is similar for low spending. At around \$100 in spending, the probability of liking the logo is so low, that one dollar lower on the spending subtracts little from the probability of liking the logo. The logo does not matter at low spending. In other words, the probability curve as a function of spending levels off for low values of spending. These three patterns together suggest that the probability curve is likely to have an s-shape.

Let's review the model that is commonly used to model binary responses, according to the examples I provided before, in previous slides. Again, the response variable is a binary response. The common model used to model such s-shaped patterns I described before, for binary response data, is called the logistic regression model. In logistic regression, we model the probability of a success, not the response variable, given the predicting variables.s

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

$$p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$$

link  $p$  to the predicting variables through a nonlinear *link function*

$$g(p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

There is no error term! What are the model assumptions?

We furthermore link the probability of a success to the predicting variables using the  $g$  link function, in a way that this  $g$  function of the probability of success is a linear model of the predicting variables. The  $g$  function is the s-shape function that models the probability of success with respect to the predicting variables, as I showed you in the Uber example. But in this model, we do not have an error term.

What are the model assumptions? How can we define the model assumptions?

- 1) A first assumption is the linearity of the predicted variables. Similar to the regression model we have learned in the previous lectures, the relationship we assume now, between the link, the  $g$  of the probability of success and the predicted variable, is a linear function. This is a slightly different linearity assumption than in a standard regression model under normality, although I'm still going to refer to this assumption, as a linearity assumption. The relationship may not be linear, however, and transformation may improve the fit.
- 2) Similar to the standard regression model, we also assume independence in the response data.
- 3) The third assumption is specific to the logistic regression model. The logistic regression model assumes that the link function is a so-called logit function, provided here on the slide. **The link function  $g$  is the log of the ratio of  $p$  over one minus  $p$ , where  $p$  is the probability of success.** This is an assumption since the logit function is not the only function that yields s-shaped

curves. And it would seem that there is no reason to prefer the logit to other possible choices.

There are other s-shaped functions that are used in modeling binary responses, under a more general model framework called the binomial model. We'll learn about other shape functions in a different lesson.

### 4.1.2. Data Example

In this lecture I'll introduce a data example that I will use throughout this lecture. In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle, United Kingdom. Twenty years later a follow-up study was conducted. Among the information obtained originally was whether a person was a smoker or not. And it was found that twenty years later, 76.12% of the 582 smokers were still alive with only 68.58% of 732 nonsmokers were still alive. That is, smokers had a higher survival rate than non-smokers. That will make the story for Philip Morris smoking leads to a longer life span.

This example was provided by Dr. Jeffrey Simonoff from New York University. The data in the file used as the input in the `read.table` command in R. Note the specification of the data separator in this function. Incorrectly specifying the type of separator can lead to incorrect reading of the data. The data file does not provide names for the columns. I provided them using the `names` command in R. I also attach this data in order to be able to refer to the columns in this data individually.

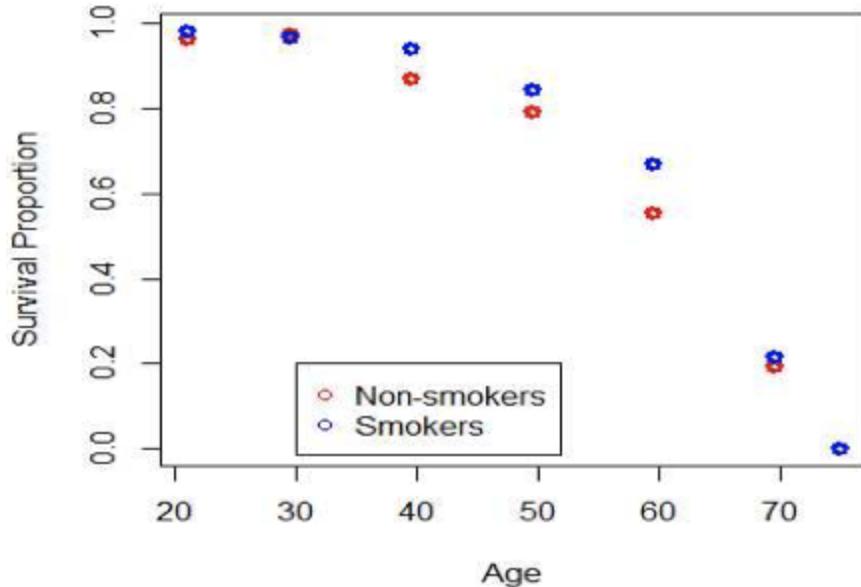
```
## Read data in R
smoking = read.table("CIGARETT.dat",sep="",row.names=NULL)
names(smoking)=c("Age","Smoker","Survived","At.risk")
attach(smoking)
```

Here is the code for plotting age versus the proportion of those that survived. We want to look at the relationship between age and proportion of survival. What we see here is that there is a non-linear relationship between age and survival proportion. In fact, this looks more like an S shape. As I mentioned in the lesson where I introduced the logistic regression model.

```

## Plot proportion of survival
plot(Age, Survived/At.risk, xlab="Age", ylab="Survival Proportion",
col=c("red","blue"), lwd=3)
legend(30,0.2, legend=c("Non-smokers","Smokers"), pch=1, col=c("red","blue"))

```

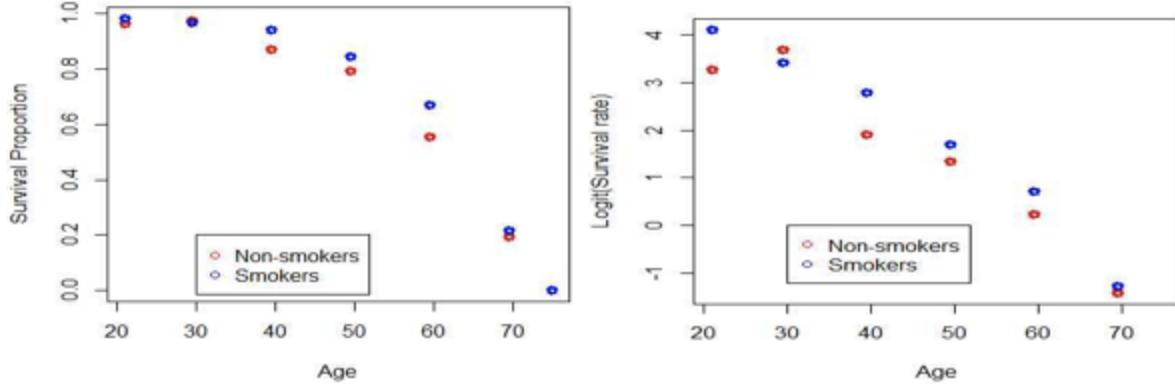


Next, I'm going to transform survival proportion using the logit function which is the log of the ratio between the proportion of survival divided by one minus the proportion of survival. So here I'm plotting the age versus the logit of the proportion of survival and I'm contrasting the plot that you saw in a previous slide on the left with the plot of the age versus logit of the transformed survival rate. Note the relationship between age and the transformed survival rate improved compared to the un-transformed survival proportion. We still see a slight curvature, and I'm going to discuss this when we're going to perform the logistic regression analysis on this example.

```

## Plot of logit transformation of the proportion survival
prop.survival=Survived/At.risk
plot(Age,log(prop.survival/(1-prop.survival)), col=c("red","blue"), xlab="Age",
ylab="Logit(Survival Proportion)", lwd=3)
legend(30,0, legend=c("Non-smokers","Smokers"), pch=1, col=c("red","blue"))

```



#### 4.1.3. Model Description and Estimation

This lecture is logistic regression and in this lesson, I focus on the model estimation and approach used to estimate logistic regression model and also on the interpretation of the regression coefficients.

Logistic regression is the generalization of regression that is used when the response variable  $y$  is binary or binomial. Assume that  $Y_i$  takes 0 or 1 values, thus binary, and we want to regress  $Y$  onto some predictive variables  $X$ . We model the probability of success using the logit link function as presented in a previous lesson. That is the logit function of the probability of success is a linear model of the predicting variables.

We can rewrite this as the probability of success equal to the ratio between the exponential of the linear combination of the predicting variables over 1 plus this same exponential. The two formulations are equivalent and we will use them interchangeably throughout this lecture.

**Model:** Probability of success given predictor(s)

$$p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$$

link  $p$  to the predicting variables through **logit link function**

$$g(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

OR

$$p(x_1, \dots, x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}$$

Let's consider a model with only one predicting variable for ease of interpretation. The logit function which is the log of the ratio between the probability of a success and the probability of a failure is called the log odds function, so the ratio between the log of P over 1- p, is the log odds function. Taking the exponential of the logit function, we have the ratio between the probability of success and the probability of failure called the odds of success given the predicting variable.

Probability of success given predicting variable X = x:

$$p = p(x) = P\{Y=1|x\}$$

- The logit function  $\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 x$  is the *log odds function*.
- Exponential of the logit function  $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 x}$  is the *odds* of Y = 1 at X = x
- The odds at X = A versus X = B are equal to the *odds ratio*:

$$\frac{e^{\beta_0 + \beta_1 A}}{e^{\beta_0 + \beta_1 B}} = e^{\beta_1(A-B)}$$



Furthermore, if we compare the odds for two different values of the predicting variable A and B, we have the odds ratio as provided on the slide. If we replace A with B + 1, then we interpret the regression coefficient beta as the log of the odds ratio for an increase of one unit in the predicting variable.

Note that we do not interpret beta with respect to the response variable but with respect to the odds of success. This is one important difference between the standard regression model and the logistic regression model. In the model described so far, the model parameters are the regression coefficients. We do not have an additional parameter to model the variance since there's no error term. Thus, for P predictors, we have P + 1 regression coefficients for a model with intercept with beta 0.

We estimate the model parameters using the maximum likelihood estimation approach. Assuming that the response data are Bernoulli or binomial with one trial with probability of success depending on the predictor variables then the likelihood function is on this slide. We can further take of the log-likelihood function and obtain the log-likelihood functions as on the slide. We want to maximize the likelihood function or the log-likelihood function with respect to the model parameters, or the regression coefficients. The resulting log-likelihood function to be maximized, is very complicated

and it is non-linear in the regression coefficients beta 0, beta 1, and beta p which we need to estimate using the maximum likelihood estimation.

**Approach:** Maximum Likelihood Estimation:

$$\max_{\beta_0, \beta_1, \dots, \beta_p} L(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^n p(x_{1i}, \dots, x_{pi})^{y_i} (1-p(x_{1i}, \dots, x_{pi}))^{1-y_i}$$

OR

$$\max_{\beta_0, \beta_1, \dots, \beta_p} l(\beta_0, \beta_1, \dots, \beta_p) = \log(L(\beta_0, \beta_1, \dots, \beta_p)) =$$

$$\sum_{i=1}^n \left\{ y_i \log \left( \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \right) + (1 - y_i) \log \left( \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \right) \right\}$$

Maximizing likelihood is challenging. In order to maximize, in order to get the beta 0 hat, beta one hat, beta p hat, we need to choose this numerical algorithm to estimate the regression coefficients. The estimated parameters and their standard errors are approximate estimates.

#### 4.1.4. Model Estimation Data Example

We'll return to the data example where we'll model whether smokers had a higher survival rate than non-smokers for the study population in the survey.

In this example, we have binomial data, which means binary data with repetitions. Thus, the response variable has a binomial distribution where  $p_i$  is the probability of a success, and  $n_i$  is the number of trials for the  $i$ th response. Within this context of this data example, the response  $Y_i$  is the number of people who survived, or the number of successes, and  $n_i$  is the number of people at risk for the  $i$ th response. The response coded in R is "survived" and the number of trials is "at.risk".

Whether the data are binary without repetitions as in an example we'll explore in a different lesson or with repetitions as in this example the logistic regression applies generally.

The R command to fit a logistic regression is **glm()**, which stands for generalized linear model, so we put a g in front of lm(). The response variable is the proportion of those who survived, and the predictive variable is whether smoker or not. When fitting a logistic regression for binary data with repetitions or binomial data with  $n_i$  larger than 1, as in this example, the response input is the proportion of survival, provided in the left of the tilde. In our notation, the input would be  $y_i$  divided by  $n_i$  as the response.

We also need to specify the weight in this case. Specify by the vector at risk or the annotation  $ni$  which is number of trials, or number of repetitions. The predictive variable is provided on the right of tilde, and it is the proportion of people for the  $i$ th response who are smokers.

When using the GLM command, it is also important to specify that we fit a binomial model by specifying family equal binomial. This means that we fit a logistic regression model. We'll learn in a different lecture that this R command is more general along to fit more other models, not only logistic regression.

**Data:**  $Y_i$  binary responses  $\sim \text{Binomial}(p_i, n_i)$

- $Y_i$  number of people at risk who survived (Survived)
- $n_i$  number of people at risk (At.risk)

## Fit a logistic regression model

```
smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk, family=binomial)  
summary(smoke1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.78052	0.07962	9.803	< 2e-16 ***
Smoker	0.37858	0.12566	3.013	0.00259 **

$\hat{\beta}_{\text{smoker}} = 0.378$ : For smokers versus non-smokers, the log odds of survival increases by 0.378 OR the odds of survival increase by 1.459.

The coefficient for a smoker is significantly positive. This positive coefficient says that being a smoker is associated with higher survival. To be more specific in our interpretation for smokers versus non-smokers, the log odds of survival increases by 0.378 or the odds of survival increases by 1.459. In the exploratory analysis for this data example, we learn that the age of the person 20 years ago is strongly and negatively associated with them surviving 20 years later. This is not surprising of course, thus we should expect that age would explain some of the variables routine survival proportion.

We next consider the model with both the smoker and age variables. The R command is similar as the one in the previous slide, except that we are adding an additional predictor: age. A portion of the R output is provided on this slide. We'll focus on the estimated regression coefficient for the smoker factor. For this model, the smoking variable has a negative coefficient, in contrast to the previous model where it had a positive coefficient.

```

## Fit a logistic regression model
smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk, family=binomial)
summary(smoke2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 7.785001 0.454999 17.110 < 2e-16 ***
Smoker      -0.240831 0.167885 -1.435    0.151
Age         -0.127419 0.007397 -17.227 < 2e-16 ***

```

$\hat{\beta}_{smoker} = -0.24$ : For smokers versus non-smokers, the log odds of survival decreases by 0.24 OR the odds of survival decreases by 1.27.

We interpret coefficient as follows, the log odds of survival decreases by 0.24 or the odds of survival decreases by 1.27 ( $\exp[0.24] = 1.27$ ) when comparing smokers versus non-smokers. We can see that the addition of the age variable reverses the sign of the coefficient corresponding to smoker variable. We'll discuss about this so called Simpson's paradox in a different lesson.

## Knowledge Check

1. Logistic regression is different from standard linear regression in that:
  - A. It does not have an error term
  - B. The response variable is not normally distributed.
  - C. It models probability of a response and not the expectation of the response.
  - D. All of the above.
2. Which one is correct?
  - A. The logit link function is the only link function that can be used for modeling binary response data.
  - B. Logistic regression models the probability of a success given a set of predicting variables.
  - C. The interpretation of the regression coefficients in logistic regression is the same as for standard linear regression assuming normality.
  - D. None of the above.
3. In logistic regression,
  - A. The estimation of the regression coefficients is based on maximum likelihood estimation.
  - B. We can derive exact (close form expression) estimates for the regression coefficients.
  - C. The estimations of the regression coefficients is based on minimizing the sum of least squares.
  - D. All of the above.
4. Using the R statistical software to fit a logistic regression,
  - A. We can use the lm() command.
  - B. The input of the response variable is exactly the same if the binary response data are with or without replications.
  - C. We can obtain both the estimates and the standard deviations of the estimates for the regression coefficients.
  - D. None of the above.

*Answers appear in the footnote below.<sup>4</sup>*

---

<sup>4</sup> 1 = D, 2 = B, 3 = A, 4 = C

## 4.2 Logistic Regression: Statistical Inference, Model Assessment, and Classification

### 4.2.1. Statistical Inference

This is the logistic regression lecture and in this lecture we'll focus on statistical inference. So we'll move from estimation of the regression coefficients to the statistical inference of the regression coefficients.

#### MODEL ESTIMATION

We learned that for estimating a logistic regression model, we use maximum likelihood estimation or abbreviated MLE. Using this approach, we cannot derive the estimated parameters or estimated regression coefficients in an exact form. For logistic regression, thus we need to use a numeric algorithm, which provides approximate estimated parameters.

MLE is a common estimation approach for statistical models. The reason is that the MLE has good statistical properties under the assumption of a large sample size that means a large N. I would say that MLE is the most applied estimation approach. In fact, the least square estimation for the standard regression model that we've learned in a previous lecture is equivalent with MLE under the assumption of normality.

#### STATISTICAL INFERENCE

Given that estimators for the regression coefficients and logistic regression are MLEs, we can use the large samples of the statistical properties of MLEs, specifically for large sample data for large N.

The sampling distribution of MLEs can be approximated by a normal distribution. Similar to the standard regression the estimate is for the regression co-efficiencies in logistic regression are unbiased and that's the mean of the approximate normal distribution is theta. The variance of the estimator does not have a close form expression and thus I suggest using a software to obtain this variance variance matrix for the estimators beta hat.

It's also important to note that approximately normal distribution relies on a large sample of data. Using this approximate normal distribution we can further derive confidence intervals. Since the distribution is normal, the confidence interval is the z

interval, as provided on the slide. Centered are the estimated regression coefficient plus or minus the z quantile, the z critical point, times the standard error, or the square root of the variance of the estimator.

$$\text{1-\alpha Approximate Confidence interval} \quad \left[ \hat{\beta}_j \pm \frac{z\alpha}{2} \sqrt{V(\hat{\beta}_j)} \right]$$

To perform hypothesis testing, we can use again the approximate normal sampling distribution. The resulting hypothesis test is also called the Wald test since it relies on the large sample normal approximation of MLEs. So if you want to test whether the coefficient  $\beta_j = 0$  or not, we can use the z-value. The z-test value is the ratio between the estimated coefficient minus 0, which is the null value, divided by the standard error.

$$z\text{-value} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

We reject the null hypothesis that the regression coefficient is 0 if the z value is larger in absolute value than the z critical point. Or the 1- alpha over 2 normal quanta. We interpret this that the coefficient is statistically significant.

Furthermore, if we want to test a more general hypothesis that the regression coefficient is equal to this constant b. Then the z-value changes in that we subtract b from the estimated coefficients of the numerator. We can make a decision whether to reject also using the P-value, which is 2 times the left tail of the standard normal of the quantile provided by the absolute value of the z-value.

$$z\text{-value} = \frac{\hat{\beta}_j - b}{se(\hat{\beta}_j)} \text{ how large to reject } H_0: \beta_j = b?$$

For significance level  $\alpha$ , Reject if  $|z\text{-value}| > \frac{z\alpha}{2}$

Alternatively, compute P-value =  $2P(Z > |z\text{-value}|)$

If we're interested in the hypothesis testing for statistical significant positive or negative regression coefficient then the z-value is the same but the P-value will change as on the slide. These derivations are similar to those for the standard regression model, except that we use the normal, not the T distribution in making the statistical inference.

Most importantly for standard regression analysis under the assumption of normality, the statistical inference relies on the T-distribution that applies under both small and large samples. On the other hand, for logistic regression, the statistical inference based on the normal distribution applies only under large sample data.

If the sample size  $n$  is small, then the statistical inference is not reliable. For example, the hypothesis testing procedure will have a probability of type I error larger than the significance level; that is, more type I errors than expected. This is an important aspect to keep in mind when reporting results based on logistic regression. If the sample size is small, you need to warn on the lack of the reliability of the results.

#### TESTING FOR SUBSETS OF COEFFICIENTS

Similar to the standard regression under normality, we can also test for subsets of regression coefficients under logistic regressions. We begin with a full model where the predicting variables divide into a set defined by  $x$  and a set defined by  $z$ , so we have  $p$   $x$  predictor variables and  $q$   $z$  predicting variables. The regression coefficients for the  $X$ s are the beta coefficients, and the regression coefficients for the  $Z$ s, are the alpha coefficients.

*Full model:*

$$\text{Logit}(p("x_1", \dots, "x_p; z_1, \dots, "z_q")) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + a_1 z_1 + a_2 z_2 + \dots + a_q z_q$$

*Reduced model:*

$$\text{Logit}(p("x_1", \dots, "x_p")) = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

*The hypothesis test:*

$H_0 : a_1 = a_2 = \dots = a_q = 0$  versus  $H_A : \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\alpha}_1, \dots, \hat{\alpha}_q)$
- Maximize the likelihood function under reduced model:  $L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$
- Test Statistic:

$$\text{Deviance} = \log(L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)) - \log(L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\alpha}_1, \dots, \hat{\alpha}_q)) \approx \chi_q$$

$$\text{P-value} = P(\chi_q > \text{Deviance})$$



For example, the  $X$ s, the first set of predictors, can be controlling variables for bias selection in a sample and the  $z$  factors can be additional explanatory variables. We want

to compare the reduced model assumed in the null hypothesis to the full model. The hypothesis testing procedure is testing the null hypothesis that all alpha coefficients are 0, versus alternative that at least one of the coefficients is not 0, the approach for performing this test is as follows.

We estimate the regression coefficients under the full, and reduced models using MLE. Then the test statistic is the difference of the log likelihood under the reduced model and the log likelihood under the full model, this difference is called deviance. For large sample size data, the distribution of this test statistic, assuming the null hypothesis is true, is a chi square distribution. With Q degrees of freedom where Q is the number of regression coefficients discarded from the full model to get the reduced model or the number of Z predicting variables.

The P-value of the test is computed as the left tail [correction: RIGHT tail] of the chi-square distribution with Q degrees of freedom of the test value. [Note: the pchisq command returns the left tail by default, so subtract it from 1 to get right tail i.e. p-value] Which is the deviance in this case.

Two aspects to keep in mind. First, just like other statistical inference for logistic regression, this test relies on large sample data and thus reliable only for large N, and second this test is not a goodness of fit test. It simply compares two models and decides whether the larger model is statistically significantly better than the reduced model.

However, this comparison can apply to models that do not fit the data well so you can compare two bad models is still a comparison. We'll come back to this concept later, goodness of fit versus comparison models. Since it is important to differentiate between comparing models versus goodness of fit, particularly for logistic regression.

## TESTING FOR OVERALL REGRESSION

We can use a similar approach to test for the overall regression. Recall that for the standard regression model under normality we use the F test to test for the overall regression. The null hypothesis here is similar but the test is different. The null hypothesis is that all regression coefficients except intercept are 0 versus the alternative that at least one is not 0. Meaning that the overall regression has statistically significant power in explaining the response variable.

*Full model:*

$$\text{Logit}(p("x_1", \dots, "x_p; z_1, " \dots, "z_q")) = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p$$

*Reduced/Null model:*

$$\text{Logit}(p("x_1", \dots, "x_p")) = b_0$$

The test statistic is the difference in the log likelihood function of the model under the null hypothesis, also called a null-deviance, and the log likelihood of the full model.

*The hypothesis test:*

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$  versus  $H_A$ : at least one is not zero

- Maximize the likelihood function under full model:  $L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$
- Maximize the likelihood function under reduced/null model:  $L(\hat{\beta}_0)$
- Test Statistic:

$$Dev = \log(L(\hat{\beta}_0)) - \log(L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)) \approx \chi_p$$

$$\text{P-value} = P(\chi_p > Dev)$$

Similar to the test for subsets of regression coefficients, the distribution of the test statistic is approximate chi-squared with  $p$  degrees of freedom where  $p$  is the number of predicting variables. The approximation is again assuming large sample data, we reject the null hypothesis when the P-value is small, indicating that the overall regression has explanatory power.

## 4.2.2. Statistical Inference Data Example

We'll return to the data example where we will model whether smokers had a higher survival rate than nonsmokers for the study population in the survey.

Let's review the model where only the smoker variable was included in the logistic regression without the age variable. According to this model, the smoker variable not only has a wrong sign but is also statistically significant. The p-value of the test, for the statistical significance of this regression coefficient is about 0.002 and that's very small.

```
## Fit a logistic regression model
smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk, family=binomial)
summary(smoke1)
Coefficients:
              Estimate Std. Error z value   Pr(>|z|)
(Intercept)  0.78052  0.07962  9.803 < 2e-16 ***
Smoker       0.37858  0.12566  3.013  0.00259 **
Null deviance: 641.5 on 13 degrees of freedom
Residual deviance: 632.3 on 12 degrees of freedom
1-pchisq(9.2,1)
[1] 0.002420151
```

Let's also evaluate the overall regression. The test value is the difference between the null deviance and the residual deviance provided in the R output. The degree of freedom is one, since we have only one predicted variable for which we tested the statistical significance of the overall regression. We compute the p-value of the test using the chi-square distribution with one degree of freedom: `1-pchisq(9.2,1)`. In R, use the p chi-square command which gives us the LEFT tail. Since we want RIGHT **upper** [lower?] tail, we'll take one minus this probability. The p-value for this test is 0.0024, thus the overall regression is statistically significant.

Let's now consider the second model including both the smoker and age variables. The smoking variable is now not significantly different from zero.

```
## Fit a logistic regression model
smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk, family=binomial)
summary(smoke2)
Coefficients:
              Estimate Std. Error z value   Pr(>|z|)
(Intercept)  7.785001  0.454999 17.110 < 2e-16 ***
Smoker      -0.240831  0.167885 -1.435    0.151
Age         -0.127419  0.007397 -17.227 < 2e-16 ***
Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 43.459 on 11 degrees of freedom
```

The p value for the test of statistical significance of the smoker variable is 0.151. However, the regression coefficient for the age variable is statistically significant because the p-value is approximately equal to zero. This means that age contributes significantly to their survival whereas smoking does not when we take age into account, at least according to this model.

## Knowledge Check 1

1. Logistic regression is different from standard linear regression in that:
  - A. The sampling distribution of the regression coefficient is approximate.
  - B. A large sample data is required for making accurate statistical inferences.
  - C. A normal sampling distribution is used instead of a t-distribution for statistical inference.
  - D. All of the above.
2. In logistic regression,
  - A. The hypothesis test for subsets of coefficients is a goodness of fit test.
  - B. The hypothesis test for subsets of coefficients is approximate; it relies on large sample size.
  - C. We can use the partial F test for testing whether a subset of coefficients are all zero.
  - D. None of the above.

**Answers:** D, B

### The 4.2.3. Model Fit Assessment

In regression analysis the goodness of fit and diagnosis of the model assumptions are important aspects of modelling for improving the model fit. In this lesson we learn about goodness of fit for logistic regression; in particular about residuals for logistic regression and how to evaluate assumptions using the residual analysis and other visual analytics and also using hypothesis testing procedures.

#### LOGISTIC REGRESSION MODEL

We'll return now to the representation or definition of the logistic regression model. The assumptions in linear logistic regressions are as follows:

##### **Assumptions:**

- *Linearity Assumption:*  $g\{p(x_1, \dots, x_p)\} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- *Independence Assumption:*  $Y_1, \dots, Y_n$  are independent random variables
- *Logit link function:*  $g(p) = \ln\left(\frac{p}{1-p}\right)$

First, we assume that the logit transformation of the probability of success is a linear combination of the predicting variables. I refer to this assumption as the linearity assumption, although it is different from the linearity assumption for the standard linear regression model.

Second, we assume that the response binary variables are independently observed. This is a similar assumption to that of the standard normal regression.

Third, unique to logistic regression, we assume that the link function,  $g$ , is the logit function. The logit function is not the only function that yields the s-shaped kind of curve. There are other shaped s-shaped functions that are used in modeling binary responses.

Note: There is no error term. How can we evaluate these assumptions or goodness of fit if we do not have error terms? Recall that for the linear regression model under normality, we use the residuals as proxies for the error terms to evaluate the model assumptions.

#### RESIDUALS IN LOGISTIC REGRESSION

For logistic regression, we also can define residuals for evaluating model goodness of fit although with one caveat. We can only define residuals for binary data with replications. In logistic regression, we differentiate between binary data without replications and binary data with replications.

Let's clearly understand the difference. For each unique set of the observed predicting variables, we can observe binary data with no repeated trials. That is a binomial distribution with one trial where  $n_i = 1$ . In contrast, we can observe binary data for repeated trials. That is a binomial distribution with more than one trial or  $n_i$  greater than 1. The data example used in this lecture to illustrate logistic regression is for binary data with replications. We will review another example in a different lesson for a data with binary responses without replications.

As I mentioned in the previous slide, residuals can be only defined for logistic regression with replications. Generally, we perform goodness of fit only for logistic regression with replications under the assumption that  $Y_i$  is binomial with  $n_i > 1$ .

### **Logistic Regression With Replications:**

$$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Binomial}(n_i, p(x_{i1}, \dots, x_{ip})), n_i > 1$$

- Estimated probabilities are:

$$\hat{p}_i = \hat{p}(x_{i1}, \dots, x_{ip}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}$$

- Pearson Residuals:  $r_i = \frac{Y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$

And given that the estimated probabilities of success  $\hat{p}_i$ . We define the Pearson residuals as the standardized difference between the  $i$ th observed response and estimated expected response, which is  $n_i$  times the probability of success. I will note that we need to standardize the difference between observed and expected response, as the responses have different variances.

### **Deviance Residuals:**

$$d_i = \text{sign}(Y_i - \hat{Y}_i) \sqrt{2Y_i \log[Y_i/\hat{Y}_i] + 2(n_i - Y_i) \log[(n_i - Y_i)/(n_i - \hat{Y}_i)]}$$

Another type of residuals are so called deviance residuals. The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model when we assume that the estimate expected response is the observed response versus the fitted model. Because of this definition, deviance has played a role of squared differences of observed minus fitted, in the sum of least squares, in a linear model.

From the binomial approximation with a normal distribution using the central limit theorem, the Pearson residuals have an approximately standard normal distribution. From the properties of the likelihood function, deviance residuals also have a standard normal distribution if the model assumptions hold. That is, if the model is a good fit.

### MODEL GOODNESS OF FIT (GOF)

To evaluate whether the model is a good fit or whether the assumptions hold, use the Pearson or Deviance residuals to evaluate whether they are normally distributed. If they're normally distributed, then conclude that the model is a good fit. If not a good fit what can go wrong? One, it can be that the linearity assumption as defined above doesn't hold. The linearity assumption can be evaluated by plotting the logit of the success rate versus the predicting variables. If there's a curvature or some non-linear pattern, it may be an indication that the lack of fit may be due to the non-linearity with respect to some of the predicting variables.

#### GOF Visual Analytics:

- Normal Probability plot & Histogram of the Residuals
- Residuals vs predictors: Linearity & Independence Assumption
- Logit of success rate vs predictors: Linearity Assumptions

Another approach to evaluating goodness of fit is through hypothesis testing. In the goodness of fit test, the null hypothesis is that the model fits well. So, the null hypothesis is that the model fits well. And the alternative is that the model does not fit well. The test statistic for the goodness of fit test is the sum of squared deviances.

#### Hypothesis Testing Procedure:

$H_0$ : the logistic model fits the data

$H_A$ : the logistic model does not fit the data

Deviance test statistic:  $D = \sum_{i=1}^n d_i^2$   
Under null hypothesis,  $D \sim \chi_{df}^2$  with  $df = n-p-1$

Under the null hypothesis of good fit, the test statistic has a Chi-Square distribution with  $n-p-1$  degrees of freedom. Very important to remember that if the p-value is small, we reject the null hypothesis of good fit, and thus we conclude that the model is not a good fit. This is the only time when we want large p-values, since a large p-value indicates that it's possible for the model to be a good fit. Thus, remember, that what we

want in a goodness of fit test is a large p-value for goodness of fit. This is again the only time we want a large P value.

## GOODNESS OF FIT vs. PREDICTIVE POWER

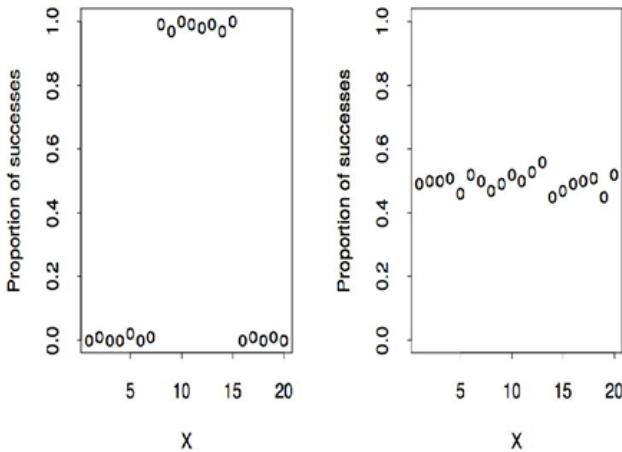
In a previous lesson we learned how to test for a subset of regression coefficients, for comparing a full versus a reduced model. Above, we learned how to evaluate a goodness of fit. In both testing procedures, we use a difference in the log-likelihood of two models. For the testing procedure for subsets of coefficients, we compare the likelihood of a reduced model versus a full model. For goodness of fit test, we compare the likelihoods of the saturated model versus the fitted model.

These two tests provide different inferences about the model. The former (testing for a subset of coefficients) provides inferences on the predictive power of the model whereas the latter (saturated vs fitted model) provides inferences on the goodness of the model.

- **Goodness of fit** means that the model assumptions hold. For example, that the s-shaped logic functions fits the data.
- **Predictive power** means that the predicting variables predict the data even if one or more of the assumptions do not hold.

It is thus important to remember that the logistic model is a sensible one for probabilities, but is not necessarily appropriate for any particular data set. This is not the same thing as saying that the predicting variables are not good predictors for the probability of success.

Consider an example with two plots. The variable  $x$  is a potential predictor for the probability of success. While the vertical axis keeps observed proportions of successes in samples taken of those values of  $x$ .



So for example,  $x$  could be the dosage for particular drug and the response variable is the proportion of people in the trial that were cured when given that dosage. In the plot on the right,  $x$  is not a useful predictor, since the probability of success appears to be unrelated to  $x$ . But the logistic regression model fits the data as a very flat S-shaped curve through the observed proportions of success reasonably well. This illustrates that the model may fit well but will not have predictive power. In the left plot,  $x$  is very useful for predicting successes but the logistic regression model does not fit the data, since the probability of success is not a monotone function of  $x$ . This illustrate that the model might predict well but will not be a good fit.

#### WHAT IF NO GOODNESS OF FIT?

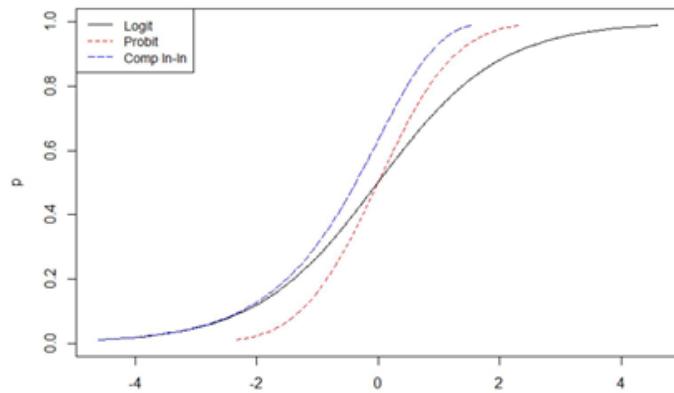
What if the model is not a good fit? One reason why the logistic model may not fit is that there may be other variables that should be included in the model and/or the relationship between logit of the expected probability and predictors might be multiplicative rather than additive. For example, if a predictor is right long tailed you might find that using a log transformation of this predictor is more effective than not using a log-transformed predictor. Generally it may be that normality transformation of the predictive variable will improve the fit. Such transformation can be identified by comparing the logit of the success rate, versus the predicted variables. For various transformations of the predicted variables.

Initial observation outliers and leverage points are also still an issue for this model. The model should be fitted with and without outliers.

Another source of lack of fit of a logistic regression model is that the binomial distribution isn't appropriate. This can happen for example, if there's correlation among the responses or there's heterogeneity in the success that hasn't been modeled. Both of

these violations can lead to what we call overdispersion, where the variability of the probability estimates is larger than would be implied by a binomial random variable. In this case, we would need to use methods that correct for overdispersion.

Another reason may be that the logit function does not fit well with the data. There are other S-shape functions such as probit or complementary log-log. The difference in the S-shape functions across these three link functions is provided here.



What I'm providing here is not the link function but the inverse of the link functions. As you may see from this plot, the c-log-log function has very long tails, meaning that it works best in extremely skewed distributions. The probit function is the inverse of the CDF of a standard normal distribution. This fits data with least-heavy tails among the three S shaped functions. Thus, it would work well when the probabilities are all concentrated within a small range.

The use of the logit function has several advantages however over other methods. The logit function is what is called the canonical link function, which means that parameter estimates under logistic regression are fully efficient and tests on those parameters are better behaved for small samples. Moreover, the interpretation of regression coefficients in terms of log odds is possible with a logit function but not other S-shape functions. Because of these reasons, logit function has become the most popular link function starting around 1970s, and it seems that it's still the default in most regression analysis for binary responses.

#### 4.2.4. Model Fit Assessment Data Example

The lecture is logistic regression and this lesson will illustrate how to apply goodness of fit for logistic regression with a data example using the R statistical software. We'll return to the data example modeling whether smokers had a higher survival rate than nonsmokers for the study population in the survey. And we'll return to the model with smokers and age included in the model.

##### DATA EXAMPLE: SMOKING - GOF HYPOTHESIS TEST

For this model, extract the sum of squared deviance residuals using the deviance command. Compute the p-value for the chi-square test for goodness-of-fit, using the p, standing for probability, chi square command where we put the test value and the number of degrees of freedom. Since we want the upper tail, we take one minus this probability. Based on this test, the p-value is small and we reject the null hypothesis of good fit. Thus, not a good fit.

```
## Deviance Test for GOF (using deviance residuals)
c(deviance(smoke2), 1-pchisq(deviance(smoke2),11))
[1] 4.345918e+01 9.033325e-06
```

We can also perform the goodness-of-fit test using the Pearson residuals. We can obtain the residuals from the fitted model by using the residuals command and specifying that we want the Pearson residuals. Further, we can compute the sum of squared residuals and compute the p-values similarly as for the deviance residuals.

```
## GOF test using Pearson residuals
pearres2 = residuals(smoke2,type="pearson")
pearson.tvalue = sum(pearres2^2)
c(pearson.tvalue, 1-pchisq(pearson.tvalue,11))
[1] 36.751889370 0.000126796
```

##### Test for goodness-of-fit:

- Using deviance residuals: p-value  $\approx 0$
- Using Pearson residual: p-value = 0.0001
- Reject the null hypothesis of good fit. Thus NOT a good fit.

The p-value using Pearson residuals is also very small. Thus, based on this test, we conclude that the model is not a good fit.

##### LINEARITY ASSUMPTION

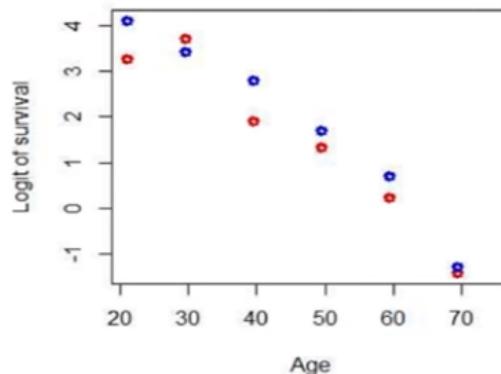
One reason for not having a good fit is a departure from the linearity assumption. We can evaluate this by plotting the predicting variables, or predicting variable age versus the logit of the success rate in this case, logit of survival. From this plot, we learn that

the relationship between the logit of survival and age is more quadratic than linear. That suggests we may improve the fit if we transform this predicting variable], age.

## Is it a linear fit?

```
plot(Age,log((Survived/At.risk)/(1-Survived/At.risk)), ylab="Logit of survival",
main="Scatterplot of logit survival rate vs age", col=c("red","blue"), lwd=3)
```

The relationship between the logit of survival and age is more quadratic than linear.



## IMPROVE THE FIT

Since the relationship looks quadratic we'll add age.square to the model to account for the quadratic relationship. The model fit is as before except that we're adding a third predicting variable, on the right of tilde in the glm function. A portion of the output from this fit is provided below. From the output, smoking is now significantly associated with survival because the probability is 0.015, in contrast to the model without age squared, where it was not statistically significant. Moreover, the regression coefficient responding to age squared is also statistically significant given the other two variables in the model.

```
## Fit a logistic regression model
Age.squared = Age*Age
smoke3 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk,
family=binomial)
summary(smoke3)

Estimate      Std. Error    z value   Pr(>|z|)
(Intercept)  2.5190783  1.0248206  2.458    0.0140 *
Smoker       -0.4284561  0.1770581 -2.420    0.0155 *
Age          0.0951102  0.0430095  2.211    0.0270 *
Age.squared -0.0021673  0.0004309 -5.030  4.91e-07 ***
```

Null deviance: 641.496 on 13 degrees of freedom  
Residual deviance: 19.808 on 10 degrees of freedom

**Test for significance  $\beta_{smoker}$ :** p-value=0.015, statistically significant at 0.05  
**Test for significance  $\beta_{age.sq}$ :** p-value ≈ 0, statistically significant

## GOF TEST FOR IMPROVED MODEL

We can evaluate the goodness of fit of the improved model, which models the linearity with respect to age, using the deviance and Pearson residuals. The R commands are the same, except that for a different model from the output. We still reject the null hypothesis of good fit with respect to the deviance residuals but not using the Pearson residuals. Thus, we have mixed results on goodness of fit.

```
## Test for goodness of fit
pearres3 = residuals(smoke3,type="pearson")
pearson = sum(pearres3^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.79 0.14
round(c(deviance(smoke3), 1-pchisq(deviance(smoke3),10)),2)
[1] 19.80 0.03
```

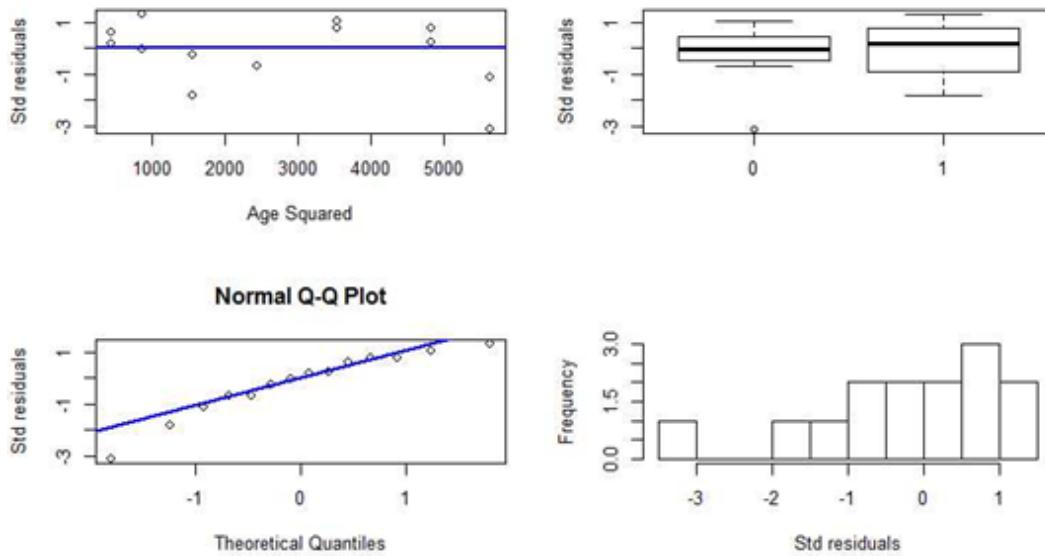
### Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.03
- Using Pearson residual: p-value = 0.14
- Do not reject the null hypothesis of good fit using Pearson residuals but do reject using Deviance residuals at the significance level 0.03 or higher.

## DATA EXAMPLE CONTINUED - RESIDUAL PLOTS

Note that the model still fits moderately well even after including the quadratic term for the age factor. The Q-Q plot and the histogram of the residuals look reasonable in the sense that they do not point to a depart from normality. Moreover, we can evaluate uncorrelated errors using the bottom plots, which again point out that the assumption of uncorrelated errors holds as well.

```
## Residual Plots
res = resid(smoke3,type="deviance")
par(mfrow=c(2,2))
plot(Age.squared,res,ylab="Std residuals",xlab="Age Squared")
abline(0,0,col="blue",lwd=2)
boxplot(res~Smoker,ylab = "Std residuals")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```



## HIGHER ORDER NONLINEARITY

One possible reason for this mixed result on goodness of fit is that a relationship with age, while being non-linear in terms of the logit, is not quadratic. We can investigate that by entering age into the model as a categorical (factor) variable rather than a numerical one. This allows for any relationship with age. A portion of the model output is provided on the slide for this particular model.

```
## Fit a logistic regression model with Age as a factor
smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age), weights=At.risk,
family=binomial)
summary(smoke4)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.8601   0.5939  6.500 8.05e-11 ***
Smoker      -0.4274   0.1770 -2.414 0.015762 *
factor(Age)29.5 -0.1201   0.6865 -0.175 0.861178
factor(Age)39.5 -1.3411   0.6286 -2.134 0.032874 *
factor(Age)49.5 -2.1134   0.6121 -3.453 0.000555 ***
factor(Age)59.5 -3.1808   0.6006 -5.296 1.18e-07 ***
factor(Age)69.5 -5.0880   0.6195 -8.213 < 2e-16 ***
factor(Age)75    -27.8073 11293.14 -0.002 0.998035

Null deviance: 641.4963 on 13 degrees of freedom
Residual deviance: 2.3809 on 6 degrees of freedom
```

Based on this model, the regression coefficients for the age dummy variables are all statistically significant except one, an indication that a high order nonlinear relationship fits better. Second, the estimated coefficient for the smoker indicated variable is virtually unchanged. So the basic implications remain the same: given age, a smoker is estimated to have lower odds of having survived 20 years later.

**Test for significance  $\beta_{smoker}$ :** p-value=0.015, statistically significant at 0.05  
**Test for significance:** Not all regression coefficients for the dummy variables for age are statistically significant

## HIGHER ORDER NONLINEARITY: GOF

The goodness of fit test for this model (logistic with Age as factor) shows significant improvement for both the deviance and Pearson residuals. The p-value is large, indicating possible good fit.

```
## Test for goodness of fit
pearres4 = residuals(smoke4,type="pearson")
pearson = sum(pearres4^2)
round(c(pearson, 1-pchisq(pearson,6)),2)
[1] 2.37 0.88
round(c(deviance(smoke4), 1-pchisq(deviance(smoke4),6)),2)
[1] 2.38 0.88
```

### Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.88
- Using Pearson residual: p-value = 0.88
- Do not reject the null hypothesis of good fit using Pearson residuals or using Deviance residuals.

## DIFFERENT LINK FUNCTION

Although we will learn that a higher order, non-linear of age will fit the model better, let's explore the possibility that a different link function might fit the data better.

```
## Use probit link function
smoke5 = glm(Survived/At.risk ~ Smoker + Age + Age.squared, weights=At.risk,
family=binomial(link = probit))
summary(smoke5)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.1033963 0.4904877 2.250   0.02447 *
Smoker      -0.2277451 0.0970191 -2.347   0.01890 *
Age         0.0681279 0.0213095  3.197   0.00139 **
Age.squared -0.0013767 0.0002173 -6.335 2.37e-10 ***
Null deviance: 641.496 on 13 degrees of freedom
Residual deviance: 18.233 on 10 degrees of freedom
```

**Test for significance  $\beta_{smoker}$ :** p-value=0.018, statistically significant at 0.05  
**Test for significance  $\beta_{age.sq}$ :** p-value ≈0, statistically significant

To use a different link function in the `glm` function, change the specification of `family`. Here, specify `family binomial` and in parentheses, the link function as the probit link function. Thus, use the binomial model but with the probit link function. The output looks the same as the logit link function, the statistical inference will also be the same.

For example, according to this model, the regression coefficients for both smoker and age.squared are statistically significant at the level 0.05. However, we cannot interpret the coefficients in terms of log odds ratios as we did using the logit link function.

```
## Test for goodness of fit
pearres5 = residuals(smoke5,type="pearson")
pearson = sum(pearres5^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.00 0.17
round(c(deviance(smoke5), 1-pchisq(deviance(smoke5),10)),2)
[1] 18.23 0.05
```

### Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.17
- Using Pearson residual: p-value = 0.05
- Do not reject the null hypothesis of good fit using Pearson residuals or using Deviance residuals at the significance level 0.05.

Similarly for the model with the logit link function, where we changed the link to probit, we find mixed results on the goodness of the model. Thus, the probit link function has not improved the model fit.

## SIMPSON'S PARADOX

How does this reversal of the age effect from the marginal to the conditional relationship happen? This is called Simpson's paradox, and it refers to reversal of an association when looking at a marginal relationship versus a partial or conditional one. This is a situation where the marginal relationship has a wrong sign. The reason that Simpson's Paradox occurred here is that being a smoker was associated with elderly people (who naturally have low survival 20 years later) being more likely to be non smokers. Thus, including age in the model reverses the sign of smoking since the two are correlated.

### Marginal versus Conditional relationship:

- Marginal:** Capturing the association of a predicting variable to the response variable marginally, i.e. without consideration of other factors.
- Conditional:** Capturing the association of a predicting variable to the response variable, conditional of other predicting variables in the model.

Once again, this contrast of marginal and conditional models as well as the fact that we need to be careful in interpreting models for observational studies and for conditional versus marginal models.

#### 4.2.5. Classification

In this lesson, I introduce a concept that is unique to logistic regression, in particular classification. Classification is the prediction of binary responses. To review, the response data are binary, and we estimate the probability of a success in logistic regression. Classification is nothing more than a prediction of the class of your response,  $y^*$  (y star), given the predictor variable,  $x^*$  (x star). We can predict the class 0 or 1 of the new response based on the probability of success given the new predicting variables  $x^*$ . If the predicted probability is large, then classify  $y^*$  as a success.

But how good the classification or the prediction is? As mentioned before, goodness of fit is different from predictive power and thus we cannot extrapolate that if a model is a good fit then it would predict well also. If we have many models for classification, how do we choose among them? As with regression curve fitting, we want a model that fits well but doesn't overfit the data. Such a model will predict the future well.

Given  $x_1, \dots, x_p$ , the predicted probability is:

$$\hat{p}(x_1, \dots, x_p) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p}}$$

Classifier:  $h(x_1, \dots, x_p) = \begin{cases} 1 & \hat{p}(x_1, \dots, x_p) > r \\ 0 & \text{otherwise} \end{cases}$

where  $r$  is a classification threshold, e.g.  $r = 1/2$ .

Classification error rate:  $L(h) = P(Y \neq h(x_1, \dots, x_p))$

How to quantify the error rate?

Specifically, we would like to have a classifier "h" with a low classification error rate. Given a set of predictive variables, we define the classifier as follows. It takes value 1 if the predictive probability is larger than some threshold value R, taking values between 0 and 1. Most common value for R is 0.5 however a different R can be used to improve the prediction accuracy.

Using this classifier, we define the classification error rate as the probability that the new response is equal to the classifier. But how to compute the classification error? One way is to simply use the data to fit the model then compute the classifier from each response in the data and take the proportion of the responses we misclassified. This is called a training error, however, we cannot use the training error rate as an estimate of the true error classification error rate because it is biased downward. And the bias comes from the fact that we use the data twice. First, we used it for fitting the model and the second time to estimate the classification error rate.

How else can we estimate the classification error without the need of observing new data? The answer involves a trick called cross validation. No analysis of prediction model is complete without evaluating the performance of the model using this technique. The basic idea of cross validation is to leave out some of the data when fitting a model. Split the data into two parts, first part called the training data will be used to fit the model and thus get the estimated regression coefficients. The second portion of the data also called the testing or validation data will be used to predict or classify the responses for this portion of the data, then compare to the observed response to estimate the classification error, one can repeat the process several times.

## Cross Validation

Split the data  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  into:

- **Training set:** Used to fit the model, i.e. estimate  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$
- **Testing/Validation set:** Used to estimate the classification error rate

$$\hat{L}(h) = \frac{1}{m} \sum_{i \in \text{Validation Set}} I(h((x_{i1}, \dots, x_{ip})) = Y_i),$$

where  $m$  is the size of the validation set.

In an ideal world, we'd have so much data that we would not mind setting some aside for validation or testing. In this way, we could obtain an unbiased estimate of how well we predict future data. In reality, we rarely have enough data to spare any of that, moreover there is something quite arbitrary about a choice of data for validation. If we do it only once, we can split a data only once and evaluate the classification error rate based on one data split. We will get an unbiased estimate of the risk. However, it's going to be quite variable, depending on how the data are split.

In practice, we used one of the three options for splitting the data:

1. Random sampling
2. K-fold cross-validation
3. Leave-one-out cross-validation which is a particular case of the K-fold cross-validation.

The simplest version of cross validation involves randomly splitting the data in two pieces, thus using **random sampling**. With random sampling, we randomly split the data into two portions, over and over again. Training the model on one portion and validating, or testing on the other portion. The risk is then calculated by averaging the risk over all the validation set to evaluate the risk of the model.

The second approach is to divide the data with **K-folds** or subset of approximately equal sizes, for each fold of data. We take that fold out of the fitting data and use that data without that fold for training them on. Then we classify the responses in the fold based on the fold that we took out based on the fitting model. On the corresponding training data to obtain the classification for that fold. Applying this to all folds of data, we get the classification for all responses and thus we can compute the classification error rate based on the this classifications.

But which one to choose from these two approaches? Random sampling is computationally more expensive than the K-fold cross validation, with no clear advantage in terms of the accuracy of the estimation classification error rate. Since K fold is preferred at least from a computation standpoint which K to choose.

**Leave-one-out cross validation is a K-fold cross validation with  $K = n$**  thus it's the extreme K cross validation. The larger K is, the larger the number of folds, the less bias the estimate of the classification error is but has higher variability. The rule of thumb for choosing K is about K equal to ten, it has been shown empirically that it provides good approximation to the classification error rate.

## Knowledge Check 2

1. In logistic regression:
  - A. We can perform residual analysis for response data with or without replications.
  - B. Residuals are derived as the fitted values minus the observed responses.
  - C. The sampling distribution of the residual is approximately normal distribution if the model is a good fit.
  - D. All of the above.
2. Which one is correct?
  - A. We can evaluate the goodness of fit a model using the testing procedure of the overall regression.
  - B. In applying the deviance test for goodness of fit in logistic regression, we seek large p-values, that is, not reject the null hypothesis.
  - C. There is no error term in logistic regression and thus we cannot perform a goodness of fit assessment.
  - D. None of the above.
3. Which is correct?
  - A. Prediction translates into classification of a future binary response in logistic regression.
  - B. In order to perform classification in logistic regression, we need to first define a classifier for the classification error rate.
  - C. One common approach to estimate the classification error is cross-validation.
  - D. All of the above.
4. Comparing cross-validation methods,
  - A. The random sampling approach is more computational efficient than leave-one-out cross validation.
  - B. In K-fold cross-validation, the larger K is, the higher the variability in the estimation of the classification error is.
  - C. Leave-one-out cross validation is a particular case of the random sampling cross-validation.
  - D. None of the above.

**Answers:** C, B, D, B

## 4.3 Case Study: The Demographics of Obesity

### 4.3.1. Exploratory Data Analysis

The lecture is logistic regression and in this lesson I will introduce a data example that we'll use to illustrate logistic regression. Obesity among adults is a serious emerging health problem in the United States that is producing immense financial strain on the healthcare system. About 35.7% of adults age 20 or older were obese and the prevalence of obesity in 2009-2010 has doubled since 1976-1980's. For advancing interventions to reduce the incidents of obesity, a first question would be: Where are the communities with most need of intervention?

To address this question we need to estimate the prevalence of obesity at the community level. While there are nationwide surveys to estimate obesity prevalence for the whole country, for small geographic areas there are sparse resources to derive an obesity prevalence estimate. This is when we turn to statistical modeling. In the study, we plan to explore a method to estimate prevalence of adult obesity.

#### CASE STUDY OVERVIEW

The data for the study was acquired from the Centers for Disease Prevention and Control, in short, CDC. The survey providing the data is called National Health and Nutrition Examination Survey, or NHANES. The data are properly collected all over from CDC. The objective is to evaluate how well we can predict using a logistic regression model with a reduced set of predicting variables available from NHANES.

In this model, the response variable is whether a person is obese or not; a binary response. The selected predicting variables are age, education level, and gender. There are many more predicting variables available in NHANES, but we'll focus only on these predictive variables since they allow prediction for small areas with data on demographics that are available from the U.S. Census Bureau.

I divided the data into training and testing to evaluate the prediction accuracy. I also transformed the age predicting variable into age groups to allow for non-linearity in the regression model with respect to age.

## EXPLORATORY DATA ANALYSIS

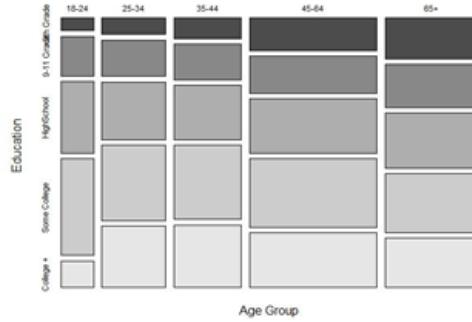
Begin with the reading the data from the file. The file is obesitydata.txt and has a header. Attach the data to refer to the variables in this data matrix as individual vectors. Then, convert the response variable into a vector with labels obese and not obese for ease of reference. Similarly, relabel the qualitative predicting variables using their meaningful labels. For example, for gender I used a factor command to specify this variable as being categorical with categories male and female.

```
## Read data in R  
(This code is hidden behind the images in the powerpoint file. It's also in R code file posted in edX.)  
obdata = read.table("obesitydata.txt", h=T)  
attach(obdata)  
obesityind = factor(Obesity, labels=c("NotObese", "Obese"))  
agegr = factor(AgeGroup, labels=c("18to24", "25to34",  
    "35to44", "45to64", "65+"))  
gender = factor(Gender, labels=c("Male", "Female"))  
edu = factor(Education, labels=c("<9thGrade", "9to11Grade",  
    "HighSchool", "SomeCollege", "College+"))
```

Next is an approach on how to visualize the association between two qualitative variables. First, create the contingency table of the two categorical variables, education and age group, using the x tabs command.

```
## Exploratory data analysis: Categorical Predictors  
tb_ageedu = xtabs(~agegr+edu)  
library(vcd)  
mosaicplot(tb_ageedu, xlab="Age Group", ylab="Education",  
color=TRUE, main="")
```

Then, I used the mosaic command to plot the data in this table using a mosaic of color ranges depending on the range of values in the cells of the table, and this is how the mosaic plot will look like. From this plot, we conclude that there is not a clear, strong relationship between those two variables. In this plot, darker colors are for the lower education levels (9-11 Grade at the top).

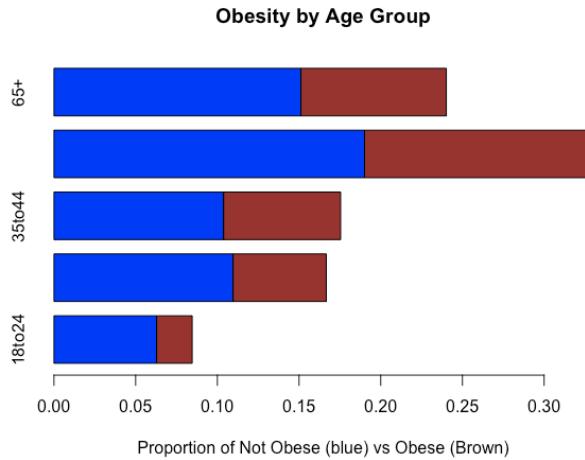


Another approach to visualize a relationship between two categorical variables is using a bar plot. In this set of plots we'll visualize the relationship between the response binary variable and the three predicting variables. We'll first create the contingency tables using the x tabs command. Then, use these tables as inputs in the bar plot

command. The inputs are actually the proportions rather than the counts in the f-test contingency tables. We convert into proportions using the prop.table command.

```
## ## Exploratory data analysis: Response vs Predictors
tb_obage = xtabs(~obesityind+agegr)
tb_ogender = xtabs(~obesityind+gender)
tb_oedu = xtabs(~obesityind+edu)
barplot(prop.table(tb_obage),axes=T,space=0.3,
        xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
        horiz=T, col=c("blue","brown"),main="Obesity by Age Group")
barplot(prop.table(tb_ogender),axes=T,space=0.3,
        xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
        horiz=T, col=c("blue","brown"),main="Obesity by Gender")
barplot(prop.table(tb_oedu),axes=T,space=0.3, horiz=T,
        xlab="Proportion of Not Obese (blue) vs Obese (Brown)",
        col=c("blue","brown"),main="Obesity by Education Level")
```

The bar plots are here. In this plot dark red corresponds to the proportions for obese and blue for not obese. We see from this plot we see differences in the proportions for each group in each of the three predicting variables.



#### 4.3.2. Modeling and Prediction

The lecture is logistic regression. In this lesson, I will illustrate logistic regression, particularly estimation, statistical inference, and prediction with the obesity prevalence estimation example.

We begin with the estimation, we use the glm R command to fit a logistic regression.

```
## Fit a logistic regression modelFlinear
```

```
model = glm(Obesity~agegr+gender+edu, family=binomial)
summary(model)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.2058	0.1573	-7.666	1.78e-14
agegr25to34	<u>0.4727</u>	0.1443	3.276	0.00105
agegr35to44	0.7649	0.1420	5.388	7.13e-08
agegr45to64	0.8482	0.1324	6.406	1.49e-10
agegr65+	0.6009	0.1375	4.370	1.24e-05
genderFemale	<u>0.2304</u>	0.0636	3.621	0.00029
edu9to11Grade	0.0563	0.1223	0.461	0.64511
eduHighSchool	-0.0344	0.1144	-0.301	0.76358
eduSomeCollege	0.1395	0.1104	1.264	0.20630
eduCollege+	-0.4008	0.1176	-3.409	0.00065
Null deviance:	5739.9	on 4313	degrees of freedom	
Residual deviance:	5641.3	on 4304	degrees of freedom	

The difference from the smoking example is that the data for the obesity example consists of individual-level observations of obesity. And thus, it is recorded as a binary data without repetitions.

For such data, the response variable, obesity, is a factor variable, rather than the standard success or survival rates as in the smoking example. The predicting variables are all factors or qualitative variables, as provided in one of the previous slides. A portion of the output of the model is provided on this slide. How do we interpret the coefficients?

Let's take one specific example, the coefficient for the age group 25 to 34. Based on its estimate, the log odds of non-obese increases by 0.4727, Or the odds of non-obese increases by 1.604 versus the age group 18 to 24, given that all the other predicted variables are fixed are in the model.

If we take another example, for females, the log odds of non-obese increases by 0.2304. Or the odds of non-obese increases by 1.259 versus males, given that all the other the other predicting variables are fixed are in the model.

## STATISTICAL INFERENCE

For the test for the overall regression, we can use a difference in the null deviance and the residual deviance provided in our output to derive the test statistic. Under the null hypothesis that all coefficients except the intercept are zero. The test statistic has an approximate chi-squared distribution with degrees of freedom provided by the number of predicting variables. The p-value is thus computed as 1 minus the `pchisq`, the

probability of a chi square evaluated in the test value, defined as gstat here on the slide. And the number of predictive variables, which is the length of the vector of coefficients minus 1 to account for the intercept, not including the null hypothesis.

```
## Test for overall regression
gstat = model$null.deviance - deviance(model)
cbind(gstat, 1-pchisq(gstat,length(coef(model))-1))
[1, ] 98.636 0
```

Because our p-value is approximately zero, we conclude that least one predicting variable has explanatory power. The last R command provided in this slide outputs the p-value of the regression coefficients obtained from the summary of the model.

```
round(coefficients(summary(model))[4],4)
(Intercept) agegr25to34 agegr35to44 agegr45to64 agegr65+
0.0000 0.0011 0.0000 0.0000 0.0000
genderFemale edu9to11Grade eduHighSchool eduSomeCollege eduCollege+
0.0003 0.6451 0.7636 0.2063 0.0007
```

I also rounded the p-values to four digits. The output is provided below this command.

From this output, we learn that education regression coefficients, for example, are all except one not statistically significant, given that we account for age and gender.

Because we cannot say anything about the goodness of fit of the model, since the data are binary responses without replications, that means one response for each individual. We'll instead evaluate a prediction power of the model using cross validation. So thus, we use cross validation to estimate the classification error rate. For this, we'll use the cv.glm R command available from the library boot in R. Don't do to install this library before using the command library boot.

For using the cv.glm function in R, we will need to first define what is called the cost function, which, in terms of classification, is the classifier that we defined when we defined a classification error rate.

```
## Prediction Accuracy
library(boot)
cost0.5 = function(y, pi){
  ypred=rep(0,length(y))
  ypred[pi>0.5] = 1
  err = mean(abs(y-ypred))
  return(err)}
obdata.fr = data.frame(cbind(Obesity,agegr,gender,edu))
## classification error for 10-fold cross-validation
```

```

cv.err = cv.glm(obdata.fr,model,cost=cost0.5, K=10)$delta[1]
.....
cv.err = c(cv.err0.35, cv.err0.35, cv.err0.4, cv.err0.45, cv.err0.5,
           cv.err0.55, cv.err0.6, cv.err0.65)
## Smallest prediction error is 0.3824
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), cv.err,
      type="l", lwd=3, xlab="Threshold", ylab="CV Classification Error")

```

The cost function in the function cost0.5() takes the value 1 if the predicting probability of success for a new response is larger than 0.5, and 0 otherwise.

Next, I define the data frame of the data, including the response variable and the three predictive variables, called obdata.fr. To obtain the K fold classification error rate with K = 10, ten folds. I'm using the cv.glm command with the input consisting of the data frame of the response in the predictive variables, the fitted model.

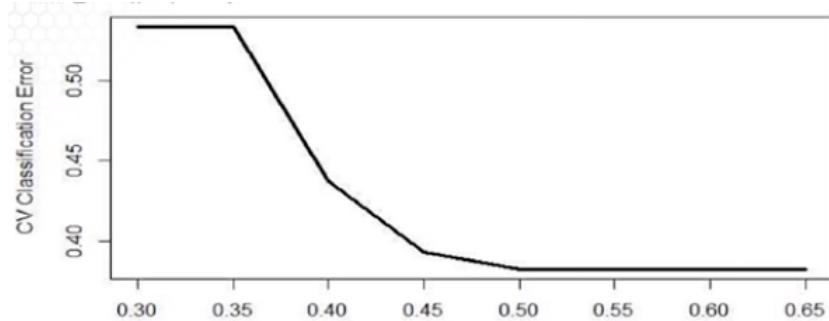
The cost function, defined cost 0.5, and K, the number of folds. Since we're interested only in the error term, we can only save the value by using the dollar sign followed by delta. Note that we estimate a classification error rate for any cost function.

For example, I defined in the R code available with this lecture the cost functions for different thresholds for the predicted probability of success, including 0.35, 0.4, up to 0.65. That is, I changed the cost 0.5 into, say, cost 0.35, where we predict a success if the predicted probability of a success is greater than 0.35.

So the only time we're gonna only change the value in this function, we're in the line where we have pi greater than 0.5. Next I get all the classification error rates for different ratios into one vector, and plot it against the threshold values. The plot of the threshold versus the classification error rates is provided here.

We see that the classification error is high for small thresholds, and decreases for higher thresholds. In fact, prediction accuracy is highest and equal for thresholds higher than 0.5, why is that?

It is the same as the prediction accuracy if we were to replace all predictions with 0, that means predict everyone not to be obese. Thus, the model does not have predictive power since it performs the same as the prediction without modeling using the regression with the three predicting variables.



The message here is that we should not forget to compare the classification error of the model to the classification error obtained when all response are predicted to be 1s or 0s, depending on which category is larger. Will we have a better or worse prediction if we were to test accuracy using a test data, rather than cross validation?

For this data example, I set aside 1,000 individuals whose data are the test data.

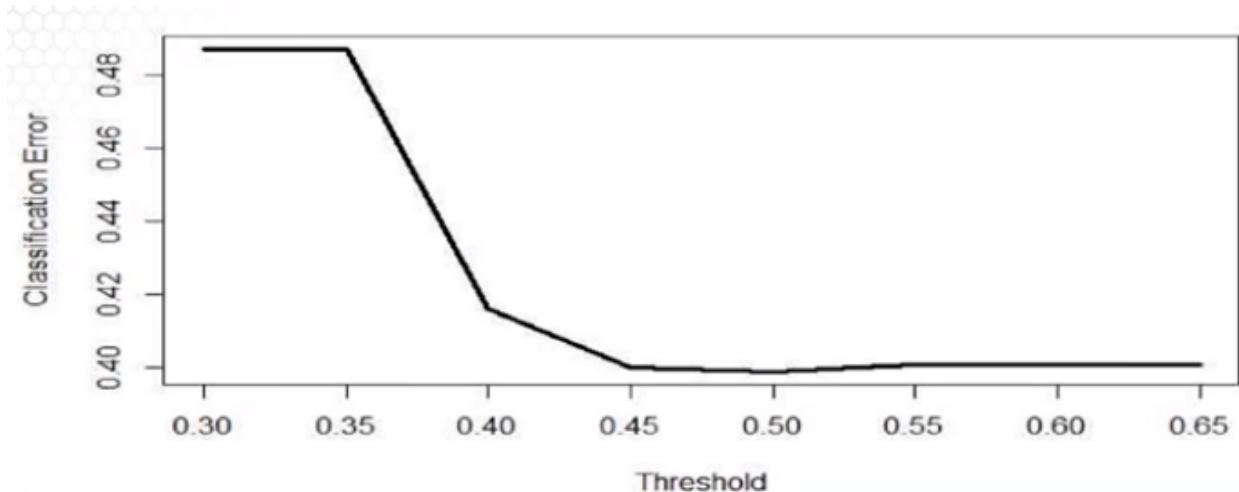
In order to predict whether non-obese versus obese, given the new predicting variables of the 1,000 individuals, we'll need to first create the pred.data frame in this prediction data, in a similar way we processed the training data. We first read the test data in R, then process the response variable and the three predictive variables using the factor command in R, along with the specification of the labels of each of those variables.

```

## Prediction given a set of new observations
## Prepare the test data
testobdata = read.table("testobesitydata.txt", h=T)
agegr.t = factor(testobdata$AgeGroup, labels=c("18to24", "25to34",
                                              "35to44", "45to64", "65+"))
gender.t = factor(testobdata$Gender, labels=c("Male", "Female"))
edu.t = factor(testobdata$Education, labels=c("<9thGrade", "9to11Grade",
                                             "HighSchool", "SomeCollege", "College+"))
pred.data = data.frame(agegr=agegr.t, gender=gender.t, edu=edu.t)
#### Predict
predict.glm(model, pred.data, type="response")
#### Prediction Accuracy for multiple thresholds
err0.3 = cost0.3(testobdata$Obesity, pred.test)
...
err = c(err0.35, err0.35, err0.4, err0.45, err0.5, err0.55, err0.6, err0.65)

```

Next, we apply the predict command in R where the input is the fitted model and the data frame with the test data. Third, we can use the same set of cost functions as before to give the classification error rate for the testing. Fourth, we can plot the classification error rates versus the thresholds in the cross function similar as I did in the previous slide.



And this is a plot of the classification error rates versus the thresholds. We see a very similar pattern as for the classification error rates estimated using the cross-validation approach. A large classification error for small thresholds, and a smaller classification error for larger threshold. We thus find the prediction accuracy is higher at 0.5, which is similar as a prediction accuracy if we were to predict everyone not to be obese.

Overall, the prediction accuracy using the fitted model did not improve for the test data.

#### 4.3.3. Goodness of Fit

In this lesson, we continue the example related to the obesity estimation data. However, in this lesson, we aggregate this data and provide not only estimation but also goodness of fit of the model based an aggregated response. Here on this slide I will demonstrate how to aggregate the response data in a way that we have only unique sets of predicting variables.

```

### Aggregate data for Logistic Regression with repetitions
obdata.agg.n = aggregate(Obesity~agegr+gender+edu,FUN=length)
obdata.agg.y = aggregate(Obesity~agegr+gender+edu,FUN=sum)
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
                       Total = obdata.agg.n$Obesity, agegr = obdata.agg.n$agegr,
                       gender=obdata.agg.n$gender, edu=obdata.agg.n$edu)

```

So we'll aggregate as follows. We take the 30 categorical predictors, and aggregate all binary responses for the same predicting values. The aggregate function in R allows us to compute the number of samples which is the number of responses with the same set of predictive values by specifying the function equal length in the first command line. The aggregate command also allows to compute the number of observed successes for the same set of predicting values by specifying the function equals sum in the second command line. These two together will give us a number of successes and the number of replications for each aggregated response defined as obesity and total here in the data frame.

We can also get the count for the predictive variables with each unique combination of the predicting variables which will now be aggregated values of the predicting variables. In other words, this code allows us to aggregate binary data without repetitions, without replications, into binary data with replications. This will be possible only in cases when all the predicting variables are categorical, thus allowing you to find a final subset of unique combinations of the predicting variables.

```

## Fit a logistic regression model
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
                 data = obdata.agg, family=binomial)

## Test for GOF: Using deviance residuals
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
[1] 29.0640209 0.8996714

```

Next, we'll apply the logistic model on the aggregated data. Recall from the smoking example that for binary response data with repetitions that input for the response needs to specify information about both the number of successes and the number of repetitions. Here I will use a different implementation of the model with repetitions in R from the one we practiced in the smoking example. Instead of providing the input of success rate and weights, as we did in the smoking example, here in this example, the

response consists of two columns. The first one is the number of successes, the vector in the data frame is obesity. The second column is the number of trials or repetitions the vector in the data frame is totaled. When we specify both columns, we do not need to specify the weights anymore, since they are specified in the second column.

The two implementations, the one we use for the smoking sample and the one we're using in this example, will provide exactly the same fitting model. It is only a difference on the input of the response data.

Because the response data are with replications, we can now perform the deviance test for goodness of fit. The p-value of the test is large, which indicates possibly a good model fit.

Coefficients:		Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.2058	0.1573	-7.666	1.78e-14	
agegr25to34	0.4727	0.1443	3.276	0.00105	
agegr35to44	0.7649	0.1420	5.388	7.13e-08	
agegr45to64	0.8482	0.1324	6.406	1.49e-10	
agegr65+	0.6009	0.1375	4.370	1.24e-05	
genderFemale	0.2304	0.0636	3.621	0.00029	
edu9to11Grade	0.0563	0.1223	0.461	0.64511	
eduHighSchool	-0.0344	0.1144	-0.301	0.76358	
eduSomeCollege	0.1395	0.1104	1.264	0.20630	
eduCollege+	-0.4008	0.1176	-3.409	0.00065	

Null deviance: 127.701 on 49 degrees of freedom  
 Residual deviance: 29.064 on 40 degrees of freedom

This slide shows the output of the model fit using the aggregated data. I'll highlight here that the output on the regression coefficients is the same as for desegregated data as the output I provided for the model I fitted in a previous lesson. However, what is different between the two outputs is the null and residual deviances and their degrees of freedom. Why would be that the case? Why are those different? The reason is that the log likelihood functions for the model with aggregated data versus the model with individual data are computed. The likelihood function is computed differently for those two type of responses for these two data sets.

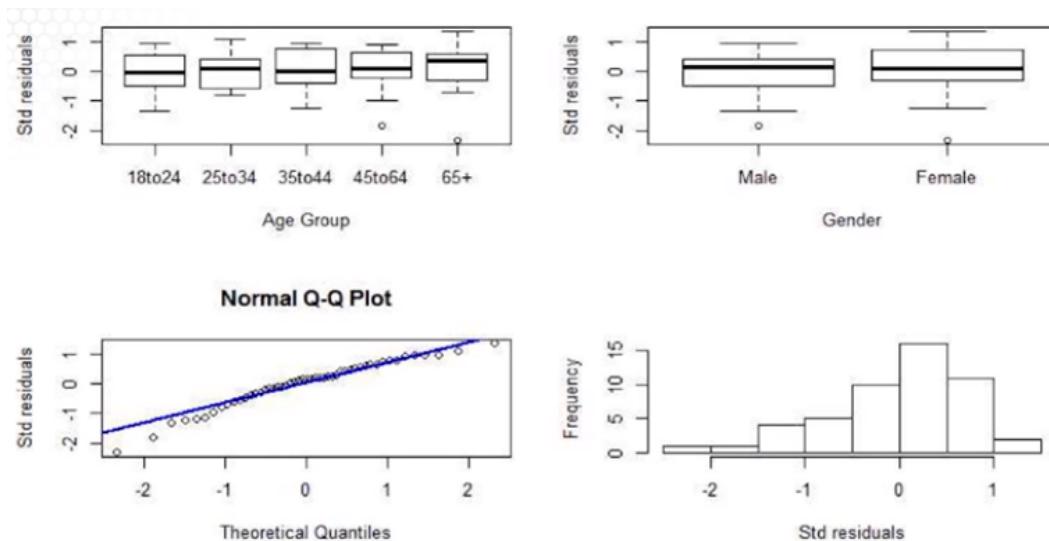
So you should use the deviances from the aggregated model for goodness of fit, not based on the individual level data. For this data, for example, the p value for the goodness of fit test is large, indicating a good fit.

```

res = resid(model.agg,type="deviance")
par(mfrow=c(2,2))
boxplot(res~agegr,xlab="Age Group",ylab = "Std residuals",data =
obdata.agg)
boxplot(res~gender,xlab="Gender",ylab = "Std residuals",data = obdata.ag
qnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")

```

We'll next perform our residual analysis using the deviance residuals derived from the aggregated data model. We do not have any quantitative variables and thus, there is no need to evaluate the assumption of linearity. We could instead evaluate the residuals versus the qualitative predictive variables using the side by side box plot. The R code for providing this analysis is here. We're only plotting the side by side box plot for age group for gender. We also plot the normal probability plot and the histogram to evaluate normality and hence goodness of fit.



These other plots we see that there is not a significant variability between the group means for age groups and for gender. Thus the model explains the variabilities due to these factors. As for normality, the distribution of the residuals is somewhat skewed potentially indication of some departures from a good fit. Let's overview the conclusions of the study.

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity. But the fitted model with education, gender, and age group does not improve prediction.

- After factor aggregation, goodness of fit can be performed.
- The p-value of the deviance test for goodness of fit is high, indicating good fit. But the residual analysis suggests that there may be some departures from normality, and thus from goodness of fit.
- Models with different link functions were including interaction terms have not shown improvement. I'm not showing the results in this example but you can practice with different link functions. You can add interaction terms to expand on this analysis.
- The sample size is large enough for trusting the statistical significance for reliable statistical inference.

What can be done to improve the model fit and the predictive power that I haven't tried? One thing to keep in mind is that we may miss some important factors that explain the variability in obesity response factor. Such factors could be income level, unemployment rates, ethnicity, among others. You may also consider interaction terms between age group, education, and gender with other factors. Those may improve the model fit and also the predictive power of the model.

## 4.4 Poisson Regression: Basic Concepts and Estimation

### 4.4.1. Introduction

In this lesson, I'll introduce a new regression model that is commonly used for modeling rate and count data. And I'll introduce this model in the context of a more general framework called Generalized Linear Models. We have learned so far about regression model for normally distributed responses and for binary responses.

Are there any other situations where we might be interested in prediction or explaining other types of responses? The answer is definitely yes. And here are few examples.

- What drives the rate of phone calls per day in a calling service center?
- What does predict the density per mile of trees in a forest?

In these examples, we would like to explain or predict the rate of an event, that being a phone call or a tree in those three examples, within a specific unit, that being the day in the first example or mile in the second example.

Such count or rate responses data are common in practice. In such examples, the underlying assumption is that the response variable has a Poisson distribution. In other examples, responses could be the wait time for a well-visit at the physician office or the time until a severe event happens for example, an asthma attack, attack for patients with asthma. In such examples the underlying assumption is that the response variable has an exponential distribution.

In this lecture, we will learn how to model, to explain, to predict count or rate response variable under more general modeling framework, the generalized linear model. What is a sensible way to model data from other distributions than normal distribution? We could simply do an ordinary least squares regression treating the count variable as the response, and assuming normality.

Let's review again the linear regression model we learned in the previous lectures. In this model, data consists of observations of response variables, and a set of predicting variables. The model is a linear relationship in the predicting variables plus an error term. The assumptions are that the error terms have mean zero, constant variance and that they are independent. The Normality Assumption also implies that the response variable is normally distributed.

But in the examples I provided before on the previous slide, the response variable does not follow a normal distribution. To generalize the standard regression model to response data that do not have a normal distribution, this so-called generalized linear model, or abbreviated, GLM, generalizes the linear model to response data coming from other distributions.

In GLM or generalized linear models, the response  $Y$  is assumed to have a distribution from the exponential family of distributions. Under this model, we model a transformation  $g$  of the expectation of  $Y$ , given the predicting variables as a linear combination of the predicting variables. Equivalently, we can write the expectation as the inverse of the  $g$  transformation of the linear combination of the predicting variables.

**Model:** Model the conditional expectation:

$$g(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

OR

$$E(Y|x_1, \dots, x_p) = g^{-1}(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

where  $g()$  is a *link function* and  $g^{-1}()$  the *inverse link function* depending on the distribution of  $Y$ .

In this model link framework, the transformation  $g$  is called a link function since it links the expectation of the response to the predicting variables. The transformation  $g$  depends on the distribution of the response variable as we'll see in the next slide. But what is the exponential family of distributions? It encompasses several distributions that have the probability density function, in short PDF, or the Probability Mass Function, in short, PMF, with the following format. In this formulation, as provided on the slide, we have the density function as  $f$  and is equal to  $h(y)$ , this is a function that depends on  $e$  to the power of  $g(\theta)$ , which is a function of the parameter  $\theta$  only, times  $T(y)$  which is a function of  $y$  only, minus  $B(\theta)$  which again is a function of  $\theta$  only. So, the function  $g(\theta)$  represents in this case what we call the link function, the  $g$  transformation I mentioned on the previous slide.

$Y \sim$  distribution in the exponential family if its density function can be written as:

$$f(y; \theta) = h(y) e^{g(\theta)T(y) - B(\theta)}$$

where  $\theta$  is the parameter of the distribution and  $g(\theta)$  is the link function.

Distribution	Link	Regression Function
Normal	$g(\mu) = \mu$	$\mu = x^T \beta$
Poisson	$g(\mu) = \log(\mu)$	$\mu = e^{x^T \beta}$
Bernoulli	$g(\mu) = \log(\mu/1-\mu)$	$\mu = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}$
Gamma	$g(\mu) = 1/\mu$	$\mu = \frac{1}{x^T \beta}$

Examples of a distribution from this family are classic and well-known distributions such as normal, binomial, Poisson, Gamma among others. For all such distributions, we can apply the generalized linear model. This table provides the  $g$  function, the link function in the definition provided in definition of the density function on the slide for normal distribution. For example, assuming sigma squared, the variance to be fixed, the link function is the identity function. For the Poisson distribution, the  $g$  function the transformation  $g$  is the log function. For the binomial distribution, the  $g$  function is the logit function. For gamma distribution, the  $g$  function is the inverse function.

So again what is this  $g$  function? The  $g$  transformation, the  $g$  function is the link function mentioned in the previous slide. To recall for logistic regression, the link function is the logit function as provided by the representation of the binomial distribution using the formula of the density function for the exponential family of distributions. For normal, since the  $g$  function is the identity function, the GLM really boils down to the standard regression model assuming normality. Thus GLM encompasses a standard linear regression and normality logistic regression, as well as other models such as Poisson regression discussed in this lecture. The G-Link function is also called the canonical link function.

The response  $Y$  in Poisson regression is assumed to have a Poisson distribution, and this is commonly used for modeling count or rate data. The common model used to model Poisson data links the expectation of the response variable to the predicting variables using the log function, which is the canonical link function as I described in a previous slide. This is equivalent with modeling the expectation of the response variable, as the exponential of the linear combination of the predicting variables, where the betas are

the model parameters' order regression coefficients. Just like in a standard regression and normality, and in logistic regression.

### **Model:** Model the conditional expectation:

$$\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

OR

$$E(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$$

What is the difference between using Poisson regression versus the standard regression with, say, with the log transformation of the response variable? If we would consider the standard regression with the log transformation, we estimate the expectation of the log of the response. So again this is, we'll model the expectation of the log of Y. The variance under the standard regression is assumed constant. That was one of the assumption for the standard regression model. For Poisson regression we estimate the log of the expectation of the response variable. More importantly, the variance of the response is assumed to be equal to the expectation, since for the Poisson distribution, the variance is equal to the expectation. Thus the variance is not constant.

### **Standard Linear Regression with log-transformation:**

- $E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $V(\log(Y)|x_1, \dots, x_p)$  constant

### **Poisson Regression:**

- $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR  
 $\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

So what we gather from comparing those two models is that using the standard linear regression with log transformation instead of Poisson regression, will result in violations of the assumption of constant variance. One could transform the response using a variance stabilizing transformation instead of using the log transformation. A classic transformation for count data is the square root of the response plus 3 over 8. This transformation will work well for large count when the data, the response data are large counts. Generally, I suggest using the Poisson regression instead of standard regression

with this transformation especially when the response data are small counts. For example, 1, 2, 3, 4, 5 counts per unit.

#### 4.4.2. Data Examples

In this lesson, I'll illustrate the applicability of the Poisson regression with two data examples. And I'll use those two data examples to illustrate the Poisson regression throughout this lecture. In the first example, we have data for the number of awards for several high schools. This data have been provided by the Digital Research and Education at University Center of at University of California, Los Angeles.

In this example, the response variable is the number of awards for each high school in the data set. Specifically, it indicates the number of awards earned by students at a high school within a year. There are also two predicting variables. Math is a continuous predictor variable and represents students' scores on their math final exam. And prog is a categorical variable with three levels indicating a type of program in which the students were enrolled.

```
## Read data in R
awardsdata = read.csv("students_awards.csv", header=T)
## Convert qualitative variable in the data into factor in R
awardsdata = within(awardsdata, {
  prog = factor(prog, levels=1:3, labels=c("General", "Academic", "Vocational"))
  id = factor(id)})
```

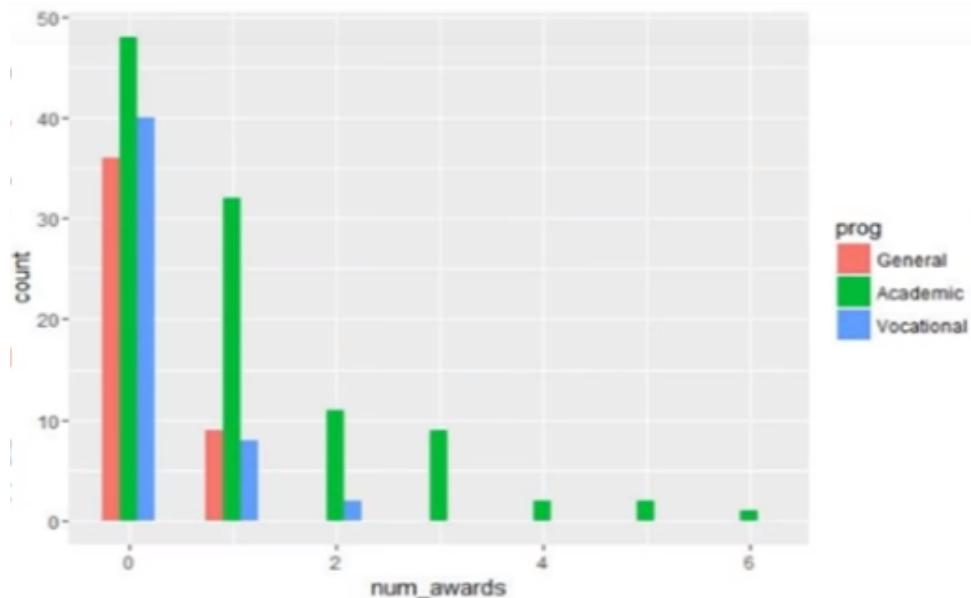
The data are in the files students awards.csv which is the input in the read.csv command, because the data file provides names for columns we also need to specify header equal to true. To convert the variable program in the data from awards data into a factor, we can use the within command that allows changing directly the type of the column prog in the awards data and also we can changed its labels.

```
## Conditional histograms
library(ggplot2)
ggplot(awardsdata, aes(num_awards, fill = prog)) +
  geom_histogram(binwidth=.5, position="dodge")
```

In order to visualize the relationship between the number of awards and the categorical variable prog. We can use a so called conditional histogram or conditional bar plot. The R command is ggplot variable, available in a ggplot2 library make sure to install this library before using it. The use of this command is a bit more complicated.

First, you need to specify the variable that needs to be plotted. In this case, number of awards and differentiate by the variable prog. The geom\_histogram command allows

input from the user on how far should the bars be from each other or how large the width of the bars, along with the position type of the bars. I recommend reading the help menu for assistance in using this function.



The output of this conditional bar plot is here. You see that there is one bar for each program differentiated by the count of awards. The maximum number of awards per high school is six. And only one school with an academic program has six awards, most schools have no awards. We see more academic programs with no awards because there are more schools with an academic program.

In the second example, we have data for the number of claims for car accidents or events leading to car damage submitted to an insurance company. The response variable is the number of car insurance claims per policyholder. Thus, the unit for the rate of events is a policyholder for the insurance company. In order to specify their response, we have information on the number of policyholders and the number of claims across all of the responses. Taking the ratio between the two will get the rate of claims per policy holder.

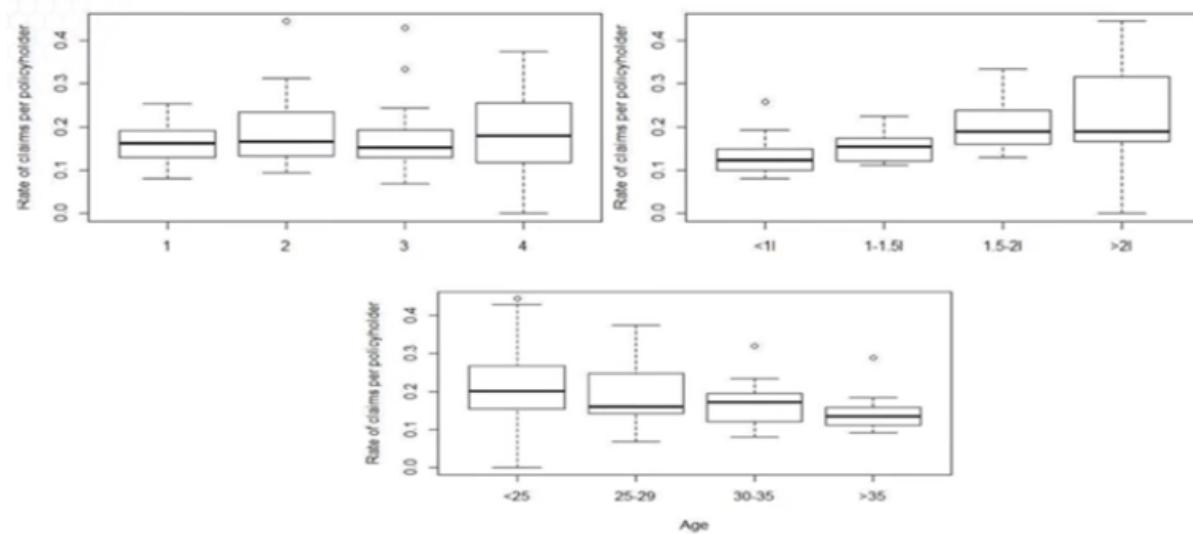
The set of predicting variables includes district of residence of the policy-holder taking values between 1 to 4, where 1 is a rural district and 4 is for major cities. Classification of cars with four levels, differentiated by the type of the engine. And age group of the policyholder is differentiated into four age groups. Less than 25 years old, between 25 and 29, between 30 and 35, and older than 35.

The data are available at R library MASS and the data file is called Insurance. In order to download this data, you only need to access the library MASS through the command library(MASS). To learn more about the data content, you can type summary(Insurance).

```
## Data in the R library MASS
library(MASS)
summary(Insurance)

## Relationship between rate of claims and predictors
boxplot(Claims/Holders~District, xlab = "District", ylab = "Rate of claims per policyholder", data=Insurance)
boxplot(Claims/Holders~Group, xlab = "Group", ylab = "Rate of claims per policyholder", data=Insurance)
boxplot(Claims/Holders~Age, xlab = "Age", ylab = "Rate of claims per policyholder", data=Insurance)
```

To visualize the relationship between three category variables and the rate of claims per policyholder, we can use a side by side boxplot. The variable of interest is the rate of claims which is computed as the ratio between the number of claims and the number of policyholders. And the R code for the side by side boxplot is here. The resulting boxplots are here.



There are small differences in the means of the rate of claims per policy holders with respect to the district but there are large differences with respect to the type of car and the age group.

#### 4.4.3. Model Description and Estimation

In this lesson I will provide the estimation approach for the Poisson regression along with the interpretation of the regression coefficients.

Poisson regression is a generalized model that is used when the response variable is a count or rate, or more specifically, when the response variable has a Poisson distribution. To overview, a random variable  $Y$  has a Poisson distribution with rate lambda, if its probability mass function is as provided on the slide. To note, the mean and the variance are both equal to the rate parameter, lambda.

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  are event count data per observation unit with a Poisson distribution

**Poisson Distribution:**  $Y \sim \text{Poisson}(\lambda)$ :  $P(Y=y) = \frac{e^{-\lambda} \lambda^y}{y!}$

$$E(Y) = V(Y) = \lambda$$

Extending to Poisson regression, we assume that the  $i$ -th response  $Y_i$  has a Poisson distribution, with rate lambda  $i$ . Where the rate parameter is the expectation of the response  $Y_i$ , given the predicting variables, which is modeled as the exponential of the linear combination of the predicting variables since the link function between expectation and the predicting variables is the log function as provided in the first lesson of this lecture. Equivalently, we can write log of the rate lambda  $i$  is equal to the linear combination of the predicting variables.

**Model:** Model the conditional expectation:

$$Y_i | x_{i1}, \dots, x_{ip} \sim \text{Poisson}(\lambda_i) \text{ with}$$
$$\lambda_i = E(Y_i | x_{i1}, \dots, x_{ip}) = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

OR

$$\log(\lambda_i) = \log(E(Y_i | x_{i1}, \dots, x_{ip})) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

Let's consider the model with only one predicting variable for ease of interpretation, the log function of the expected value of the response is called the log rate. Taking the ratio of the rate with an increase of one unit in the predicting  $x$ . When  $x$  is numeric, it's a quantitative variable, then we obtain exponential of the beta one, the regression coefficient corresponding to the  $x$  predicting variable. Thus, the regression coefficient is

interpreted as the log ratio of the rate with an increase of one unit in the predicting variable.

The rate of event occurrence given predicting variable  $X = x$ :

$$\lambda = \lambda(x) = E(Y|x) = e^{\beta_0 + \beta_1 x}$$

- The log function  $\ln(\lambda(x)) = \beta_0 + \beta_1 x$  is the *log rate*.
- The *ratio of the rates* with an increase with one unit in  $x$  is

$$\frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

A similar interpretation can be provided for categorical predicting variables, except that the comparison is with respect to a baseline group. Also note that we do not interpret beta with respect to the response variable but with respect to the ratio of the rate. This is one important difference between the standard regression model and under normality and the Poisson regression model. Furthermore, if we have multiple predictive variables then we need to make the interpretation assuming that all other predictor variables are fixed.

In the model described so far the model parameters are the regression coefficients, the betas. Note that we do not have an additional parameter due to the error variance, since there is no error term. Thus, for  $p$  predictors, we have  $p + 1$  regression coefficients for a model with intercept. But we have  $p$  regression coefficients for a model without intercept.

**Approach:** Maximum Likelihood Estimation:

$$L(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

$$\max_{\beta_0, \beta_1, \dots, \beta_p} l(\beta_0, \beta_1, \dots, \beta_p) = \log(L(\beta_0, \beta_1, \dots, \beta_p)) =$$

$$\sum_{i=1}^n \{y_i \log \lambda_i + \lambda_i\} = \sum_{i=1}^n \{y_i (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) - e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}\}$$

We estimate the model parameters using the maximum likelihood estimation or abbreviated MLE. Assuming that the response data have a Poisson distribution with a rate depending on the predictive variables then likelihood function is as on the slide. In MLE, we maximize the likelihood function with respect to the model parameters or in this case, the regression coefficient. We can further take the log of the likelihood

function since the log of a product is the sum of the logs. The resulting log likelihood functions to be maximized is on the slide.

From this derivation, the objective function or the log likelihood that needs to be maximized is highly non-linear. In the regression coefficient beta and thus we can not derive a closed form expression of the estimates. We cannot derive an exact expression of the estimate. We need to use a numeric algorithm to maximize the log likelihood, the estimated regression coefficients are thus not obtained exactly.

The upshot is that the estimated parameters and their standard errors are approximate estimates. Do not attempt to do the estimation all by yourself. Use the statistical software, like the R statistical software, to derive the estimated regression coefficients. Furthermore, this approximation will have implications on the statistical inference, as we'll see in a different lesson.

#### 4.4.4. Model Estimation Data Example

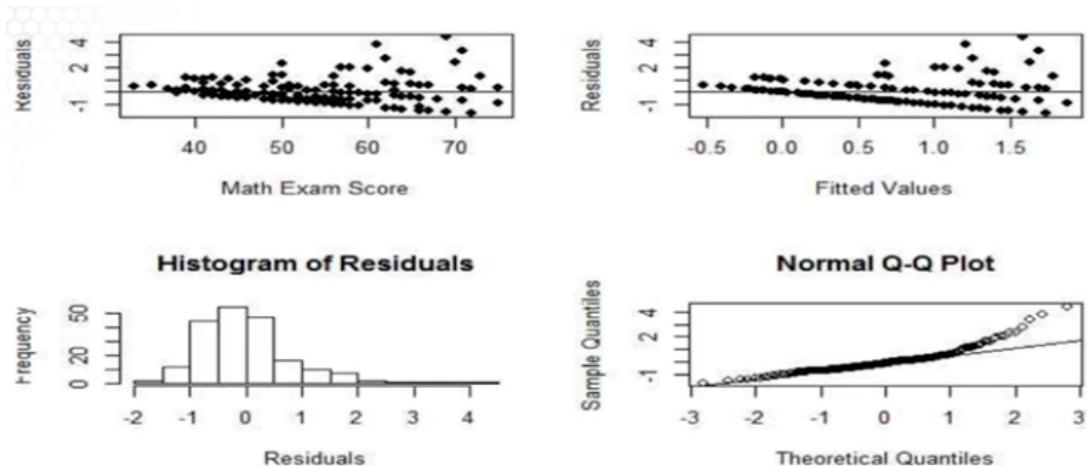
And in this lesson, I illustrate the estimation of the Poisson Regression with two data examples.

Let's first fit a standard regression model under the assumption of normality. We can use the LM command to fit the model where the response variable is the number of awards. And the predicting variable is the math score in the program. We will return to the awards data example where we will model the number of awards per high school with respect to two predicting variables.

```
## Fit a standard regression model
m0 = lm(num_awards ~ prog + math, data=awardsdata)
## Residual Analysis for Goodness of Fit
par(mfrow = c(2,2))
plot(awardsdata$math, res, xlab = "Math Exam Score", ylab = "Residuals",
pch = 19)
abline(h = 0)
plot(fitted(m0), res, xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnorm(res)
qqline(res)
```

We evaluate the goodness of fit for the model. Recall that one problem with fitting a normal regression model. Poisson data is the departure from the assumption of constant variance. We'll perform a residual analysis including the scatter plot of the math score versus residuals. The scatter plot of the fitted values versus residuals, and the normal probability plot and the histogram.

And the residual plots are here.



It is clear from both plots in the first row that the variance of the residuals is not constant motivating the need of using Poisson Regression instead of the regression under normality. Note that for this example, the number of awards per school takes values between zero and six. And thus the number of counts per response is small. This is one case where Poisson regression will perform much better than the standard normal regression model. Even with the transformation of the response variable.

The command in R used to fit a poisson regression is GLM, which stands for Generalized Linear Models. The response variable is the number of awards, and the predictive variables are the math and type of program. When using the GLM command, it's also important to specify that we fit a Poisson model by specifying family equal poisson. This means that we fit a Poisson regression model. We use this command for fitting a logistic regression model in a different lecture.

```
m1 = glm(num_awards ~ prog + math, family="poisson", data=awardsdata)
summary(m1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15 ***
progAcademic	1.08386	0.35825	3.025	0.00248 **
progVocational	0.36981	0.44107	0.838	0.40179
math	0.07015	0.01060	6.619	3.63e-11 ***

$\hat{\beta}_{math} = 0.07$ : the expected ratio of count of awards per year for one unit increase in the math final exam score is  $\exp(0.07) = 1.072$  given the program.

$\hat{\beta}_{academic} = 1.084$ : the expected ratio of the counts of awards per year for an academic program vs a general program is  $\exp(1.084) = 2.956$

In fact, we can use this command to fit any generalized linear model. That is a model with the response variable following a distribution from the exponential family of distributions. The output of the model is not much different than for the regression model under normality, fitted using the lm R-command, except for the statistical inference that we'll discuss in a different lesson.

The coefficient for the math predicting variable is positive and equal to 0.07. We interpret this coefficient as the expected count for awards per year for one unit increase in the math finals exam scores, which is exponential 0.07 equal to 1.072 ( $\exp[0.07] = 1.072$ ) given the program. The coefficient for the academic program is 1.084. Which is interpreted as the expected ratio of the counts of awards per year for an academic program, versus a general program. And this is equal to exponential of 1.084, which is the regression coefficient, and it's equal to 2.956 ( $\exp[1.084] = 2.956$ ).

We'll next fit a Poisson regression model to the rate of claims per policy holders given the three predicting qualitative variables. For this example, I will only focus on the implementation of the Poisson regression. The R command is this GLM, although for this example, we'll need to consider what is called exposure. Poisson regression is also appropriate for rate data where the rate is a count of the events occurring for a particular unit of observations. Which means that a rate is the count of events divided by the number of units.

```
m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)), data = Insurance, family = poisson)
```

For example here, the response variable consists of two components. The number of claims and the number of policyholders. The number of claims are the number of events, and the number of policyholders represents the number of units. To get the rate of claims per policy holder, we take the ratio between the number of claims and the number of policyholders. Thus, we may model the rate of claims per policy holder

where the number of policyholders is what we call the exposure of the response variable.

In Poisson regression, this is handled using an off-set. Where the exposure variable enters in the linear combination of the predicting variables. But with the coefficient for log of exposure constrained to one. That is, in the `gstat = smoke2$null.deviance - deviance(smoke2)` for the log of the expectation of the response data, we're adding another term, which is log of exposure.

This translates in the R implementation as specified as an offset. The offset option in R allows us to include log of exposure in the model without estimating a regression coefficient for the exposure. In this example the number of policy holders is the exposure thus the offset is equal to the log of the number of policyholders. Please do remember to account for this offset when the number of units is different across the observed responses as in this example.

## Knowledge Check

1. Poisson regression can be used:
  - A. To model count data.
  - B. To model rate response data.
  - C. To model response data with a Poisson distribution.
  - D. All of the above.
2. Which one is correct?
  - A. The standard normal regression, the logistic regression and the Poisson regression all fall in; under the generalized linear model framework.
  - B. If we were to apply a standard normal regression to response data with a Poisson distribution, the constant variance assumption would not hold.
  - C. The link function for the Poisson regression is the log function.
  - D. All of the above.
3. In Poisson regression,
  - A. We model the log of the expected response variable not the expected log response variable.
  - B. We use the ordinary least squares to fit the model.
  - C. There is an error term.
  - D. None of the above.
4. Which one is correct?
  - A. The estimated regression coefficients and their standard deviations are approximate not exact in Poisson regression.
  - B. We use the `glm()` R command to fit a Poisson linear regression.
  - C. The interpretation of the estimated regression coefficients is in terms of the ratio of the response rates.
  - E. All of the above.

**Answers:** 1: D, 2: D, 3: A, 4: E

## 4.5 Poisson Regression: Statistical Inference, Model Assessment

### 4.5.1. Statistical Inference

And in this lesson, we'll move from estimation to statistical inference for the Poisson regression model.

We learned that for estimating a Poisson regression model, we use maximum likelihood estimation or abbreviated MLE. Using this approach, we cannot derive the estimated parameters or regression coefficients in exact form. And thus, we need to use a numeric algorithm which provides approximate estimated parameters.

MLE is a common estimation approach for statistical models. The reason is that even for more complicated models, such as logistic regression or Poisson regression, we can use large sample size properties of the estimators for statistical inference. Given that estimators for the regression coefficients in the Poisson regression are MLEs, we can use the large sample statistical properties of MLEs. Specifically, for large sample data, the sampling distribution of MLEs of the maximum likelihood estimators, can be approximated by a normal distribution.

#### Statistical Properties of MLEs:

- Approximate Sampling Distribution:  $\hat{\beta} \approx N(\beta, V)$
- The normal approximation relies on the assumption of large sample size  $\Rightarrow$  Statistical inference is not reliable for small sample data

Similarly to the standard regression, the estimators for the regression coefficients in the Poisson regression are approximately unbiased. And thus the mean of the approximate normal distribution is beta. The variance of the estimator does not have a close form expression and thus, I suggest using a software to obtain this V, the variance covariance matrix for the estimator beta hat.

$$1-\alpha \text{ Approximate Confidence interval} \quad \left\{ \quad \hat{\beta}_j \pm z_{\frac{\alpha}{2}} \sqrt{v(\hat{\beta}_j)} \right.$$

It is important to note the approximate normal distribution, this approximation relies on the large sample data. Using this approximate normal distribution, we can further derive confidence intervals. Since the distribution is normal, the confidence interval is the Z intervals as provided on the slide. To perform hypothesis testing, we can use again the approximate normal sampling distribution. The resulting hypothesis test is also called the Wald test, as it's realized on the large sample normal approximation of MLEs.

$$H_0: \beta_j = 0 \text{ vs. } H_a: \beta_j \neq 0$$

$$z-value = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

If we test where the regression coefficient is 0, then the z-value is the ratio between the estimate and the standard deviation. We reject the null that the regression coefficient is 0 if the z-value is larger in absolute value than the z critical point or the 1 minus alpha over 2 normal quantile. We interpret this as that the coefficient is statistically significant.

$$z-value = \frac{\hat{\beta}_j - b}{se(\hat{\beta}_j)} \quad \text{how large to reject } H_0: \beta_j = b?$$

For significance level  $\alpha$ , Reject if  $z-value > z_{\frac{\alpha}{2}}$

Furthermore, if we want to test a more general hypothesis, that the regression coefficient is equal to this constant  $b$ , then the z-value changes in that we subtract  $b$  from the estimate of the coefficient of the denominator. We can make a decision whether to reject using the p-value which is 2 times the left tail of the standard normal of the quantile provided by the absolute value of the z-value.

Alternatively, compute P-value =  $2P(Z > |z\text{-value}|)$

*What if we want to test for positive relationship?*

$H_0: \beta_j \leq 0$  versus  $H_A: \beta_j > 0$ ?

P-value =  $P(Z > z\text{-value})$

*What if we want to test for negative relationship?*

$H_0: \beta_j \geq 0$  versus  $H_A: \beta_j < 0$ ?

P-value =  $P(Z < z\text{-value})$

If we're interested in that of testing for statistical significant positive or negative regression coefficients, then the z-value is the same, but the p-value will change as on the slide. These derivations are similar to those for the standard regression model under normality except that we use a normal, not the t-distribution, in making the statistical inference.

Most importantly, for standard regression analysis under the assumption of normality, the statistical inference relies on a t-distribution that applies on the small and large sample. On the other hand, for a Poisson regression, the statistical inference based on a normal distribution applies only on the large sample data.

If the sample size or n is small, the statistical inference is not reliable. For example, the hypothesis testing procedure will have a probability of type 1 error larger than the significance level, that is more type 1 errors than expected. This is an important aspect to keep in mind when reporting results based on the Poisson regression. If the sample size is small, you need to warn the reader on the lack of reliability of the results.

---

*Full model:*

$$\log(E(Y|x_1, \dots, x_p; z_1, \dots, z_q)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_q z_q$$

*Reduced model:*

$$\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

*The hypothesis test:*

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_q = 0 \text{ versus } H_A: \text{at least one is not zero}$$

Similar to the standard regression under normality, we can also test for subset of regression coefficients under the Poisson regression model. Specifically, we begin with a full model where the predicting variables divide into a set defined by x's and a set defined by z's. The regression coefficients for the first set are the beta coefficients. And

the regression coefficients for the second set are the alpha coefficients. For example, the x's can be controlling variables for bias selection in the sample, and the z's are the additional explanatory variables.

We want to compare the reduced model assumed in the null hypothesis to the full model. The hypothesis testing procedure is testing the null hypothesis that all alpha coefficients are zero, versus the alternative that at least one alpha coefficient is not zero. The approach for performing this test is as follows. We estimate the regression coefficients under the full and reduced models using MLE. Then the test statistics is a difference of the log likelihood under the reduced model and the log likelihood under the full model. This difference is called deviance.

For large sample size data, the distribution of the test statistic, assuming the null hypothesis, is a chi-squared distribution, with the q degrees of freedom, where q is the number of regression coefficients discarded from the full model to get the reduced model, or the number of z predicting variables. The p-value of the test computed is the **left tail** [correction: right tail] of the chi-squared distribution with q degrees of freedom, at the test value which is the deviance. Just like all the statistical inference for a Poisson regression, this test relies on large sample data, and thus is reliable only for large n.

We can use a similar approach to test for the overall regression. Recall that for the standard regression model under normality, we used the F-test to test for the overall regression. The null hypothesis here is similar but the test is different. The null hypothesis that all regression coefficients except intercept are 0, versus the alternative that this one is not zero. Meaning that the overall regression has statistically significant power in explaining the response, the variability in the response variable. The test statistic is the difference in the log likelihood function of the model under the null hypothesis, also called the null deviance, and the log likelihood of the full model.

*Full model:*

$$\log(E(Y|x_1, \dots, x_p; z_1, \dots, z_q)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

*Reduced model:*

$$\log(E(Y|x_1, \dots, x_p)) = \beta_0$$

*The hypothesis test:*

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$  versus  $H_A: \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$
- Maximize the likelihood function under reduced model:  $L(\hat{\beta}_0)$
- Test Statistic:

$$Dev = \log(L(\hat{\beta}_0)) - \log(L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)) \approx \chi_p$$

$$\text{P-value: } P(\chi_p > Dev)$$

Similar to the test for subset of regression of coefficients as provided in the previous slide, the distribution of the test statistic is approximately chi-squared with  $p$  degrees of freedom, where  $p$  is the number of predicting variables. The approximation is again assuming large sample data. We reject the null hypothesis if the p-value is small, indicating that the overall regression has explanatory power.

estimated regression coefficients

#### 4.5.2. Statistical Inference Data Example

And in this lesson, I will illustrate statistical inference for the poisson regression model using the two data examples. We'll return to the awards data example where we'll model the number of awards per high school with respect to two predicting variables.

```
m1 = glm(num_awards ~ prog + math, family="poisson", data=awardsdata)
summary(m1)

Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.24712 0.65845 -7.969 1.60e-15 ***
progAcademic 1.08386 0.35825 3.025 0.00248 **
progVocational 0.36981 0.44107 0.838 0.40179
math          0.07015 0.01060 6.619 3.63e-11 ***
Null deviance: 287.67 on 199 degrees of freedom
Residual deviance: 189.45 on 196 degrees of freedom
1-pchisq((287.67-189.45),(199-196))
[1] 0
```

This is the model fit, with the number of awards as a response variable, and the two predicting variables, math score and type of program. The p-value for the test of the statistical inference of the regression coefficient corresponding to the math score is approximately zero. Thus, we reject the null hypothesis and conclude the math score, thus have explanatory power for the number of awards.

Let's also evaluate the overall regression. The test value is a difference between the null deviance and the residual deviance provided in the R output. The degree of freedom is three since we have three predicting variables. One, numerical variable, the math score and two dummy variables for the program type. We compute the p-value of the test using the chi-squared distribution with three degrees of freedom. In R, we can use the pchisq command which gives us the LEFT tail. Since we want the RIGHT or upper tail, we will only take one minus this probability. The p-value of this test is approximately zero that's the overall regression is statistically significant.

We'll next perform the statistical inference for a poisson regression model for the rate of claims per policy holder given three predicting qualitative variables. This is the R output for the model where the response variable is the rate of claims for car damage for policy holder. It's important to note that we do not have a row for the log of exposures

in the R output. Since this is not a predictor in a model, but the opposite. We do not estimate a regression coefficients for the log of the exposure.

```
m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)),  
data = Insurance, family = poisson)  
summary(m.ins)  
Coefficients:  
Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.810508 0.032972 -54.910 < 2e-16 ***  
.....  
Age.L -0.394432 0.049404 -7.984 1.42e-15 ***  
Age.Q -0.000355 0.048918 -0.007 0.994210  
Age.C -0.016737 0.048478 -0.345 0.729910  
Null deviance: 236.26 on 63 degrees of freedom  
Residual deviance: 51.42 on 54 degrees of freedom  
# test for overall regression  
1-pchisq((236.26-51.42),(63-54))
```

The baseline for the age of group is the young group up to 25 years old. The age.L corresponds to the age group of between 25 and 29. When comparing the age group of 25 to 29, versus the baseline of the age group up to 25 years old. The regression coefficient is negative and statistically significant. Thus, the ratio of the rate claims per policy holder for age group 25 to 29 versus the young ones, the age group that are younger than 25 is 0.67 or exponential of the beta coefficient which is -0.394, suggesting a lower rate for those of age 25 to 29, versus those younger ones, given all of the predictor variables fixed in the model.

However, when comparing other age groups to the baseline group, the regression coefficients are not statistically significant thus the rates of the rate of claims may be possible or similar when comparing those younger of that 25 versus those of age 30 or older.

Let's also evaluate the overall regression. We compute the p-value of the test, using the chi square distribution with 9 degrees of freedom. In R, we can use the pchisq command. The p-value for this test is approximately 0. Thus, the overall regression is statistically significant.

Let's address the following question, is the district of residence of policyholder a statistically significant predictor given all other predicting variables in the model? For this, we compare the reduced model with only age and group, versus the full model

including also the dummy variables corresponding to the district predicting variable. For this test, we will use the wald.test function in the library aod in R. Remember to install the package corresponding to this library before using this library.

```
library(aod)
wald.test(b=coef(m.ins), Sigma=vcov(m.ins), Terms=2:4)
Wald test:
-----
Chi-squared test:
X2 = 14.6, df = 3, P(> X2) = 0.0022
```

The Wald test is exactly the test for subset of regression coefficients. The input consists of the estimated regression coefficients, the beta hats, the variance covariance matrix of the estimator of the estimated beta hats along with the coefficients in the vector beta that needs to be tested. The vector of regression coefficients include the intercept, three dummy variables for the district variable, three dummy variables for the group of cars, and three dummy variables for the age factor. Since the reduced model does not include the district variable, thus discarding all fake dummy variables corresponding to the variables, this reduces to the regression model, to testing for the regression coefficients to be equal to zero that are in the position 2 to 4 in the vector of the regression coefficients. Thus, we need to specify in the wald.test function the terms the coefficients in the vector b that need to be tested are in the position 2 to 4. The output of this test is provided as on the slide.

X2 is a test value and 0.022 is the p-value of the test. Since the p-value is small, we reject the null hypothesis. And thus conclude that the district predicting variable has explanatory power for the rate of claims per policy holder.

## Knowledge Check 1

1. In Poisson regression:
  - A. We make inference using t-intervals for the regression coefficients.
  - B. Statistical inference relies on exact sampling distribution of the regression coefficients.
  - C. Statistical inference is reliable for small sample data.
  - D. None of the above.
  
2. Which one is correct?
  - A. We use a chi-square testing procedure to test whether a subset of regression coefficients are zero in Poisson regression.
  - B. The test for subsets of regression coefficients is a goodness of fit test.
  - C. The test for subsets of regression coefficients is reliable for small sample data in Poisson regression.
  - D. None of the above.

**Answers: D, A**

### 4.5.3. Model Fit Assessment

In any regression analysis, an important part of the analysis is the assessment of the goodness of fit of the model and particularly through hypothesis testing or residual analysis. In this lesson, I will provide the goodness of fit and residual analysis for the Poisson regression model.

We'll return now to the representation or definition of the Poisson regression model, the response data or seem to have a Poisson distribution along with the p predicting variables. The assumptions in Poisson Regression are as follows.

- First, we assume that the log transformation of the rate is a linear combination of the predicting variables. I'll refer to this assumption as a linearity assumption, although this is different from the linearity assumption from the standard linear regression model under normality.
- Second, we assume that the response variables are independently observed. This is a similar assumption to that of the standard normal regression.
- Third, we assume that the link function  $g$  is the log function, but for logistic regression there are other linked functions that are commonly used. For Poisson regression, the log link function is almost always used, unless the counts are large and modeled using the standard regression model, assuming normality using a transformation of the response, for stabilizing the variance.

However, how can we evaluate these assumptions or goodness of fit, if we do not have error terms. Recall that for the linear regression model under normality, we use the residuals as proxies for the error terms to evaluate the model assumptions, we cannot do that anymore here. For Poisson regression, we also can define residuals for evaluating model goodness-of-fit although there is not an error term. Under the assumption that  $Y$  has a Poisson distribution, and given the estimated rates,  $\lambda_i$ .

## Poisson Regression:

$$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Poisson}(\lambda(x_{i1}, \dots, x_{ip}))$$

- Estimated rates are:

$$\hat{\lambda}_i = \hat{\lambda}(x_{i1}, \dots, x_{ip}) = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}}$$

- Pearson Residuals:  $r_i = \frac{Y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$

We defined the Pearson residuals as a standardized difference between the  $i$ th observed response, and estimated expected rate of events  $\lambda_i$  divided by the square root of the variance where the variance is equal to  $\lambda_i$ . Note that we need to standardize the difference between observed and expected response since their responses have different variances. And that the type of residuals is so called deviance residuals. The deviance residuals are the sign square root of the likelihood evaluated the saturated model when we assume that the estimated expected response is the observed response versus the fitted model.

$$\text{Deviance Residuals: } d_i = 2 \sum_{i=1}^n \left\{ Y_i \log \left( \frac{Y_i}{\hat{\lambda}_i} \right) - (Y_i - \hat{\lambda}_i) \right\}$$

Because of this definition, deviances play the role of the squared differences observed minus fitted in the sum of least squares in the linear model under normality. From the Poisson approximation with the normal distribution, you see the central limit theorem the Pearson residuals have an approximate standard normal distribution from the properties of the likelihood function the deviance results also have a standard normal distribution if the model assumptions hold, that is if the model is a good fit.

To evaluate whether the model is a good fit or whether the assumptions hold, we can use the Pearson or deviance residuals to evaluate whether they are normally distributed, if they are normally distributed then we conclude that the model is a good fit. If not a good fit, the linearity assumption as defined in a previous slide could be evaluated by plotting the log of the event rate versus the predicting variables. If there is a curvature or some nonlinear product, and may be an indication that a lack of fit may be to the nonlinearity, with respect to some of the predictor variables.

We can also evaluate linearity on the assumption of uncorrelated responses using the scatter plot for the residuals versus the predicting variables. Another approach to evaluate goodness of fit is through hypothesis testing. In a goodness of fit test, the null hypothesis is that the model fits well and the alternative that the model does not fit well.

**Hypothesis Testing Procedure:**

$H_0$ : *the Poisson model fits the data*

$H_A$ : *the Poisson model does not fit the data*

Deviance test statistic:  $D = \sum_{i=1}^n d_i^2$

Under null hypothesis,  $D \sim \chi_{df}^2$  with  $df = n-p-1$

Reject the null that the model is correct if  $p\text{-value} = P(\chi_{df}^2 > D)$  small.

Note that for this test, we want large p-values!!!!

The test statistic for the goodness of fit test is the sum of square root deviances. Under the null hypothesis of good fit, the test statistic has a chi-squared distribution of  $n-p-1$  degrees of freedom. Very important to remember if the p-value is small. We reject the null hypotheses of good fit and thus, we conclude that the model is not a good fit. This is the only time that we want large p's as a large p value indicates that it's possible for the model to be a good fit, thus, remember that we want large p value for the goodness of fit test.

What if the model is not a good fit? One reason why the process regression model might not fit is that there may be other variables that should be included in the model. Or, and the relationship between the log of the expected rate and the predicting variables might be not linear. Thus, it may be that non-linear transformations of the predicting variables would improve the fit. Such transformations can be identified by comparing the log of the estimated rate versus the predicting variables. Unusual observations, outliers, leverage points are also still an issue for these models. The model should be fitted with and without outliers.

Another source of lack of fit of a Poisson regression model is that the Poisson distribution is inappropriate. This can happen, for example, if there is correlation among the responses. Or if there is heterogeneity in the event of rates that hasn't been modeled. Both of these violations can lead to overdispersion, where the variability of the rate estimates is larger than would be implied by a Poisson model. In this case, you would need to use methods that correct for overdispersion. So what is overdispersion?

Overdispersion is a general concept for generalized linear models, not only for Poisson regression. And this happens when the variability of the response variable is larger than estimated by the model. This is common when the variability of a function of the expectation, like in logistic regression and Poisson regression. For example, in logistic regression, the variance of the response variable given the predicting variables is  $n$  times the probability of a success times 1 minus the probability of a success. Thus, once we estimate the probability to success or the expected value, we automatically obtained the variance also.

**Overdispersion:** the variability of the response variable is larger than would be implied by the model

**Binomial regression model:**

- $V(Y_i|x_1, \dots, x_p) = n_i p(x_{i1}, \dots, x_{ip})(1-p(x_{i1}, \dots, x_{ip}))$
- Overdispersed Binomial:  $V(Y_i|x_1, \dots, x_p) = \phi n_i p(x_{i1}, \dots, x_{ip})(1-p(x_{i1}, \dots, x_{ip}))$

**Poisson regression model:**

- $V(Y_i|x_1, \dots, x_p) = \lambda(x_{i1}, \dots, x_{ip})$
- Overdispersed Poisson:  $V(Y_i|x_1, \dots, x_p) = \phi \lambda(x_{i1}, \dots, x_{ip})$

Under overdispersion, the variance of the response is, in fact, large, larger than implied by the model. And thus, we estimate the model where the variance has an additional multiplicative factor phi, allowing for larger variance than otherwise estimating using the logistic regression. For Poisson regression, the variance as provided by the model is simply the rate lambda. Again under overdispersion we added multiplicative factor phi, allowing for a larger variance than otherwise estimated using the Poisson regression.

**Overdispersion Parameter:  $\phi$**

- Estimate:  $\hat{\phi} = \frac{D}{n-p-1}$  where D is the sum of the squared deviances
- If  $\hat{\phi} > 2$  then overdispersed model

How can we identify overdispersion? We can estimate the overdispersion parameter, which is the deviance, or the sum of the squared deviance residuals, divided by the degrees of freedom,  $n - p - 1$ . If the estimated overdispersion parameter is larger than two, then an over-dispersed model will fit better. To note that this overdispersion impacts the estimated variance. It will also impact statistical inference if overdispersion is not accounted for, statistical inference will not be as reliable.

#### 4.5.4. Model Fit Assessment Data Example

The lecture is Poisson Regression, and in this lesson I will illustrate the assessment of goodness-of-fit with two data examples. We'll return to the awards data example, where we'll model the number of awards per high school with respect to two predicting variables.

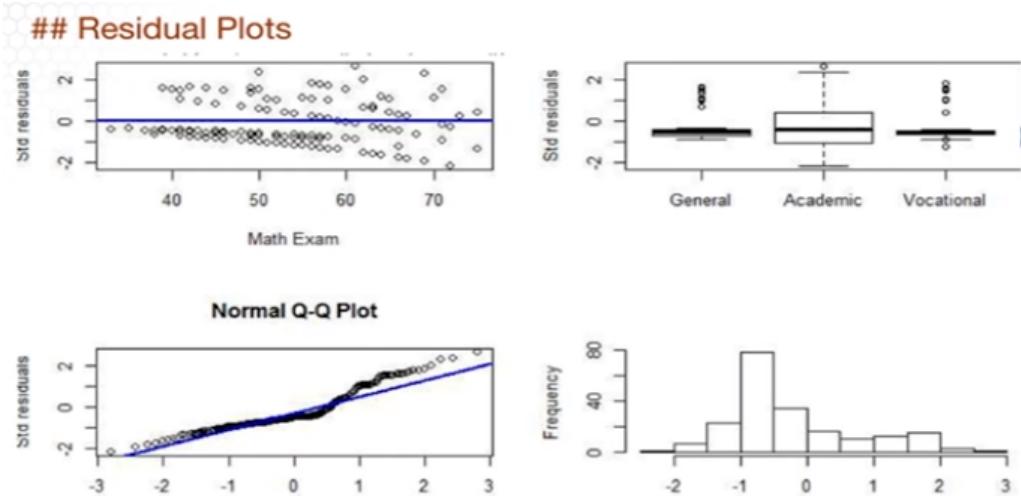
Let's go back to the fitted model for this example. For this model, we can extract the sum of square deviance residuals using the deviance command. We can further compute the p-value for the chi square test for goodness-of-fit using the pchisq command where we input the test value and the number of degrees of freedom. Since we want the upper tail, we take one minus this probability. Based on this test, the p-value is large, concluding that we do not reject the null hypothesis of good fit.

```
## Deviance Test for GOF
with(m1, cbind(res.deviance = deviance, df = df.residual,
               p = 1 - pchisq(deviance, df.residual)))
res.deviance   df          p
[1,]    189.4496  196  0.6182274
```

Test for goodness-of-fit:

- Using deviance residuals: p-value = 0.61
- Do not reject the null hypothesis of good fit.

Further, we study the residuals, specifically, I plotted the residuals versus the math score variable. I also provided the side-by-side box plot with respect to the R program, along with the normal probability plot and the histogram. The r code for this plot is provided with this lecture.



While based on the goodness-of-fit test we concluded that it is possible that the model to be a good fit, it seems that there is a nonlinear relationship with respect to the Math score. The normality assumption also does not hold since the distribution of the residuals is skewed. However, note that the normal distribution is only an approximation in the Poisson regression.

One approach to considering transformations of the predicting variables is by trial and error. Try to fit the Poisson regression model for multiple transformations of the predicting variable and choose the one that provides the best fit.

Alternatively, we can instead fit a model where we assume a non-parametric transformation of the math score predicting variable. That is, let the data tell us which transformation is best. For that we can use the `gam` function in the library(`mgcv`). It is useful to remember this function, it stands for generalized additive models, and it applies to a response with a distribution from the exponential family distributions including normal binomial Poisson and others.

```
## Fit a logistic regression model with math nonlinearly associated to awards count
library(mgcv)
m2 = gam(num_awards ~ prog + s(math), family="poisson", data=awardsdata)
```

The difference from the `glm` function is that the `gam` function allows for considering non-parametric transformations of the quantitative predicting variables like this example. In order to consider such transformation I specify `s` of math. Which means that the function or the transformation, the `gam` function will fit a smooth

non-parametric transformation of the math score. It would be good practice to learn more about this R command using the help menu.

- The residuals vs math: downward trend: Consider a **non-parametric** transformation of ‘math’ predicting variable
- *Nonparametric association*: not specifying the transformation but allowing the data to best identify/fit the transformation
- For this example, we do not see an improvement in the fit.

What do we conclude by fitting this model? For this example we did not see an improvement in the fit. Thus a transformation of the math score will not improve the fit of the relationship between the number of words and the math score.

## Knowledge Check 2

1. Residual analysis in Poisson regression can be used:
  - A. To evaluate goodness of fit of the model.
  - B. To evaluate whether the relationship between the log of the expected response and the predicting variables is linear.
  - C. To evaluate whether the data are uncorrelated.
  - D. All of the above.
2. When we do not have a good fit in generalized linear models, it may be that:
  - A. We need to transform some of the predicting variables or to include other variables.
  - B. The variability of the expected rate is higher than estimated.
  - C. There may be leverage point that need to explored further.
  - D. All of the above.

**Answers: D, D**

## Practice Midterm 2

1. Cross-validation using random sampling is less computationally efficient (more computationally expensive) in estimating the model error rate than the K-fold cross validation.

**True**

2. If the non-constant variance assumption does not hold in multiple linear regression, we apply a transformation to the predicting variables.

**False**

3. The prediction of the response variable has higher uncertainty than the estimation of the mean response.

**True**

4. In Poisson regression, we also need to check for the assumption of constant variance of the error terms.

**False**

A 5

The canonical link function for Poisson regression is the logit link function.

**False**

A 6

The linear regression model under normality is also a generalized linear model with link function the identity link function.

**True**

A 7

The regression coefficients for the Poisson regression can be estimated in exact/closed form.

**False**

A 8

The sampling distribution of the predicted response variable used in statistical inference is normal in multiple linear regression under the normality assumption.

**False**

A 9

The sampling distribution for the estimated regression coefficients under logistic regression is approximately t-distribution.

**False**

A 10

We can perform goodness-of-fit analysis through residual diagnosis for a logistic regression without replications.

**False**

A 11

We can perform goodness-of-fit analysis for a Poisson regression.

**True**

A 12

The logit link function is not the only S-shape function that can be used to model binary response data.

**True**

A 13

An approximate test can be used to test for the overall regression in Poisson regression.

**True**

A 14

When a Poisson regression does not fit well the data, it may that there may be more variability in the estimators than provided by the model.

**True**

A 16

The statistical inference for logistic regression relies on large size of the sample data.

**True**

A 17

The statistical inference for linear regression under normality relies on large size of sample data.

**False**

A 18

Other link functions for the Poisson regression model are c-log-log and probit.

**False**

A 19

**False**

A 20

If you apply linear regression under normality to count data, the assumption of constant variance still holds.

**False**

Part 2: Multiple Choice

Which answer is correct?

B 1

Which of the four answers is correct?

Let  $Y = 1$  if a driver gets a ticket and 0 otherwise. Consider a categorical predictor indicating which town a driver is in: slow town or fast town. You get this R output from fitting using the `glm` command with 'family=binomial'.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.6931	1.2247	0.566	0.571
townslow	-1.3863	1.7321	-0.800	0.423
(Dispersion parameter for binomial family taken to be 1)				

Null deviance: 8.3178 on 5 degrees of freedom

Residual deviance: 7.6382 on 4 degrees of freedom

AIC: 11.638

The link function used for fitting this model is:

- A. **Logit link function**
- B. Log link function

- C. Complementary log-log link function
- D. Probit link function

B 2

What is the odds ratio for getting a ticket while in town fast?

- A. -1.37
- B. 4**
- C. 0.25
- D. None of the above

B 3

What is the expected probability of getting a ticket while in fast town?

- A. 0.199
- B. 0.667**
- C. 0.8
- D. None of the above

B 4

What is the 95% confidence interval for the regression coefficient corresponding to slow town? (The z-critical point to be used in the computation of the confidence interval is 1.96.)

- A. [0.0083, 0.882]
- B. [-4.781,2.0086]**
- C. [-0.985, 4.449]
- D. None of the above

B 5

What is the 95% confidence interval for the probability of getting a ticket in slow town?

- A. [0.0083, 0.882]**
- B. [-4.781,2.0 086]
- C. [-0.985, 4.449]
- D. None of the above



# Unit 5: Variable Selection

## 5.1 Basics of Variable Selection

### 5.1.1. Introduction

In this lesson I introduce the general concept of variable selection, or model selection along with the common objectives in variable selection.

#### OBJECTIVES (PROBLEMS THAT VARIABLE SELECTION TRIES TO MINIMIZE)

- High Dimensionality: In linear regression, when the number of predicting variables  $P$  is large, we might get better predictions by omitting some of the predicting variables. **Models with many predictors have low bias but high variance. Models with few predictors have high bias but low variance.** We seek to balance a trade off between bias and variance.
- Multicollinearity: When there's multicollinearity among the predicting variables, we can reduce its impact by considering a subset of the predicting variables rather than the full model. This can significantly improve prediction and statistical inference.
- Prediction vs Explanatory: Consider the purpose of the analysis. If the purpose is to simply come up with accurate predictions for the response, researchers tend to simply look for variables that are easily obtained that account for a high degree of variation to response. Most commonly, one would consider smaller number of predicting variables for prediction versus explanation of the response variable. When the objective is to explain the relationship to the response, one might consider including predicting variables which are correlated. For prediction, this should be avoided.

Overall, such objectives are achieved using rigorous variable selection.

#### IMPLICATIONS AND WORDS OF CAUTION

- Confounding vs Explanatory Variables: When selecting variables for a model, consider the research hypothesis, as well as any potential confounding variables to control for. For example, in most medical studies, age and gender are always included in the model since they're common confounders. Researchers are

looking for the effect of other predictors on the response once age and gender have been accounted for.

- Targeted Predicting Variables: If your research hypothesis specifically addresses the effect of a variable, either include it in your model or show explicitly in your analysis why the variable does not belong.
- Over-Interpretation: Be wary of over-interpretation of the model in a multiple regression setting for these reasons:
  - The selected variables are not necessarily special. Variable selection methods are highly influenced by correlations between variables. When two predictors are highly correlated, usually one will be omitted despite the fact that the other may be a good predictor on its own. The problem is that since the two variables contain overlapping information once we include one, the second variable accounts for very little additional variability in the response
  - +e.
  - If we have a regression coefficient of 0.2 for variable A, for example, the interpretation is as follows. While holding the values of other predictors constant, a one unit increase in the value of A is associated with the increase of 0.2 in the expected value of the response.
  - For observation studies, causality is rarely implied.

## NO MAGIC BULLET

There have been many approaches and advancements toward variable selection. Still, variable selection for a large number of predicting variables is an unsolved problem in statistics. While some of the modern approaches we will learn at the end of this lecture attempt to address this problem, variable selection is an art by itself.

In some sense model selection is data mining. Data miners, machine learners often work with many predictors. I recommend against blindly and automatically applying variable selection without the learning of the problem at hand, objective of the analysis, variables of interest, [underlying] hypothesis, among others. All such considerations would lead to a meaningful model and interpretation.

Generally variable selection approaches need to be tailored to the problem at hand. There's no magic bullet, there are no magic procedures to get you the best model. I highlighted this quote in the first lecture of this class and I come back to this because

it's relevant in the context of variable selection. "All models are wrong but some are useful."

## NOTATION

To begin the introduction of the variable selection methods, here is some notation.

Given  $p$  predicting variables, we can have  $2^p$  combinations of the variables, and thus,  $2^p$  models to choose from.

Denote  $S$  a subset of indices in the set 1 to  $p$  with the subset of predictors among the  $p$  variables with these corresponding indices. Use  $\hat{\beta}(S)$  to be the estimated regression coefficients based on the model fitted with the predicting variables with indices in  $S$  or with  $X_S$  being the design matrix. Similarly  $\hat{Y}(S)$  are the fitted values for the same model. I refer to this model as well as its output as the  $S$  submodel. Again we can have  $2^p$  such submodels.

## Notation

Given  $S \subset \{1, \dots, p\}$  a subset of indices and  $(x_j \text{ for } j \in S)$  the subset of predicting variables with indices in  $S$ :

- $\hat{\beta}(S)$  estimated regression coefficients for the submodel with the  $X_S = (x_j \text{ for } j \in S)$  predicting variables
  - $\hat{Y}(S)$  fitted values for the submodel with the  $X_S = (x_j \text{ for } j \in S)$  predicting variables (e.g. for regression assuming normality  $\hat{Y}(S) = X_S \hat{\beta}(S)$ )
- I will refer to this model as the **S submodel**

In this lesson, I provided the overall framework, the general concept of variable selection.

## Lecture 5.1.2 – Data Examples

In this lesson, I illustrate variable selection with two data examples. I introduced this first example in the lectures for multiple linear regression.

### EXAMPLE: RANKING STATES BY SAT PERFORMANCE

Researchers examined compositional and demographic variables to understand to what extent these characteristics were tied to state level average SAT scores. The research questions to be addressed are: Which variables are associated with state SAT scores? How do the states rank? Which states perform best for the amount of money they spend? We'll return to this example to perform variable selection, where the focus is on how we can perform variable selection in the presence of confounding variables.

#### **The response variable is:**

**Y** = State average SAT score (verbal and quantitative combined)

#### **The predicting variables are:**

- “**takers**”: % of total eligible students (high school seniors) in the state who took the exam
- “**rank**” median percentile of ranking of test takers within their secondary school classes
- “**income**”: median income of families of test takers, in hundreds of dollars
- “**years**”: avg number of years takers had in social sciences, natural sciences, & humanities
- “**public**”: % of test takers who attended public schools
- “**expend**”: state expenditure on secondary schools, in hundreds of dollars per student

The first two predicting variables are the control for other variables via selection in this example. Here's the output of the regression model fitted with the response and the predicting variables. In a different lecture we learned that the takers variable had a non-linear relationship with SAT variable. Thus we transformed this predicting variable using the log transformation.

# Regression Analysis

```
regression.line = lm(sat ~log(takers) + rank + income + years + public + expend)  
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	407.53990	282.76325	1.441	0.15675
log(takers)	-38.43758	15.95214	-2.410	0.02032 *
rank	4.11427	2.50166	1.645	0.10734
income	-0.03588	0.13011	-0.276	0.78407
years	17.21811	6.32007	2.724	0.00928 **
public	-0.11301	0.56239	-0.201	0.84168
expend	2.56691	0.80641	3.183	0.00271 **

Test for statistical significance:

$\hat{\beta}_{takers}$ : p-value=0.02

$\hat{\beta}_{rank}$ : p-value>0.1

$\hat{\beta}_{income}$ : p-value>0.1

$\hat{\beta}_{years}$ : p-value<0.01

$\hat{\beta}_{public}$ : p-value>0.1

$\hat{\beta}_{expend}$ : p-value<0.01

Shall we discard the predicting variables with regression coefficients that are not statistically significant?

NO. Perform variables selection

From this output, we find that some of the regression coefficients are statistically significant and some are not. I have seen some practitioners fitting such models discard those predicting variables corresponding to regression coefficients that are not statistically significant. This is not a good practice.

It is possible and often the case that once the predictive variable is discarded, there will be a change in what is statistically significant and what is not. A regression coefficient may not be statistically significant in the full model, but once another predicting variable is discarded, it may become statistically significant and vice versa.

It's also possible to select a model to include variables that are not statistically significant, even though that model will provide the best prediction, for example. Moreover, when performing variable selection, one has to take into account confounding variables as in this data example.

## INFERENCE ON A SUBSET OF COEFFICIENTS

In a previous lecture, we compared a full model (including the confounding variables, takers and rank, along with the four explanatory variables) to the reduced model (including only the confounding variables.) To do so, we used ANOVA command in order to obtain a decomposition of the sum of regression into extra sums of regressions.

Using this command, we test whether dropping income, years, public, and expend is better than the model with these variables. That is, we test whether any of these variables will improve the **predictive** [explanatory?] power of the model when added to takers and rank, and the controlling factors.

```
regression.red = lm(sat ~ log(takers) + rank)
anova(regression.red, regression.line)

Model 1: sat ~ log(takers) + rank
Model 2: sat ~ log(takers) + rank + income + years + public + expend

  Res.Df    RSS   Df Sum of Sq    F    Pr(>F)
1     47 45530
2     43 26585   4    18945 7.6604 9.42e-05 ***
```

- Testing for a subset of regression coefficients:
  - $H_0$ : Reduced Model (takers and rank only) vs.  $H_A$ : Full Model
  - Partial F Test: F-value = 7.6604; P-value  $\approx 0$
- **Confounding and explanatory variables:** log(takers) and rank need to be in the model.
- **Partial F test for explanatory variables:** at least one predicting variable has c power. Which ones? Perform variable selection!!!

Since the P-value is small, we conclude that the test of the subset of variables not included in the reduced model collectively contain valuable information about the relationship with the SAT score. But we don't know which are important. But the P-value indicates that removing them all would be unwise. In order to identify the variables that are important explaining the variability within SAT score, we need to perform variable selection.

#### EXAMPLE: PREDICTING BANKRUPTCY

In the second data example, we analyze factors that are associated to bankruptcy.

Understanding and predicting bankruptcy has always been important and now more than ever given the failure of multi billion dollar enterprises like Enron, KMart, Lehman Brothers and others. Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm's prospects. Roughly 40 years ago, Ed Altman

showed that publicly available financial ratios can be used to distinguish between firms that are about to go bankrupt and those that are not.

In this example, we'll address the following question. Which financial indicators are associated with bankruptcy for telecommunication firms? This data example was provided by Dr. Jeffrey Simonoff from New York University. The data consists of the 25 telecommunications firms that declared bankruptcy between May 2000 and January 2002, and that had issued financial statements for at least 2 years along with 25 other telecommunications firms selected from the December 2000 financial statements that did not declare bankruptcy. The last set of firms were selected to match the other set of 25 firms that declared bankruptcy by asset sizes.

The idea of matching is common practice in statistics in such analyses. This is motivated by the intent to replicate an experimental data setting. For this particular example, we can only conclude where the financial indicators point to bankruptcy if the firms that declare bankruptcy are compared to those which didn't. However, the comparison has to be among firms that are similar with respect to some characteristics. In this example, asset size, if the matching is performed vigorously, such analysis could allow for causal reference.

In this example, the response variable is binary: bankrupt or not. The predicting variables are financial indicators, derived from the financial statements as follows.

- *Working capital* as a percentage of total assets, or abbreviated WC.TA, expressed in percentages. Working capital is the difference between current assets and liabilities and is a measure of liquidity. Bankruptcy could be associated to less liquidity.
- *Retained earnings* as a percentage of total assets or abbreviated RE.TA, expressed again in percentages. This is a measure of cumulative profitability over time and is an indicator of profitability depending on age. Both youth of a firm or less profitability would be associated with an increased risk of insolvency, and thus possible bankruptcy.
- *Earnings before interest and taxes* as a percentage of total assets, or abbreviated EBIT.TA, expressed again in percentages. This is a measure of the productivity of a firm's assets, with higher productivity expected to be associated with a healthy firm.

- *Sales as a percentage of total assets*, or abbreviated as S.TA, expressed in percentage. It indicates the ability of a firm's assets to generate sales. Lower sales would be expected to be associated with unhealthy prospects for a firm.
- *Book value of equity divided by book value of total liabilities*, or abbreviated BE.TL, a smaller value is indicative of the decline of a firm's assets relative to its liabilities, presumably, an indicator of unhealthiness.

Because the response variable is binary, we'll explore the relationship of the predicting variables to the response variable using box plots. The box plots can be used to see the separation between the response variable groups on the predicting variables. This does not take into account the variables having joint effects. And it doesn't necessarily imply that a linear logistic model is appropriate. But it's still helpful.

To obtain the side by side box plot, we first read from the data file. Then we apply the boxplot R command as provided on the slide. And here are the plots.

```
# boxplots
par(mfrow=c(2, 3))

boxplot(split(WC.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="WC.TA",main="Boxplot of WC/TA")

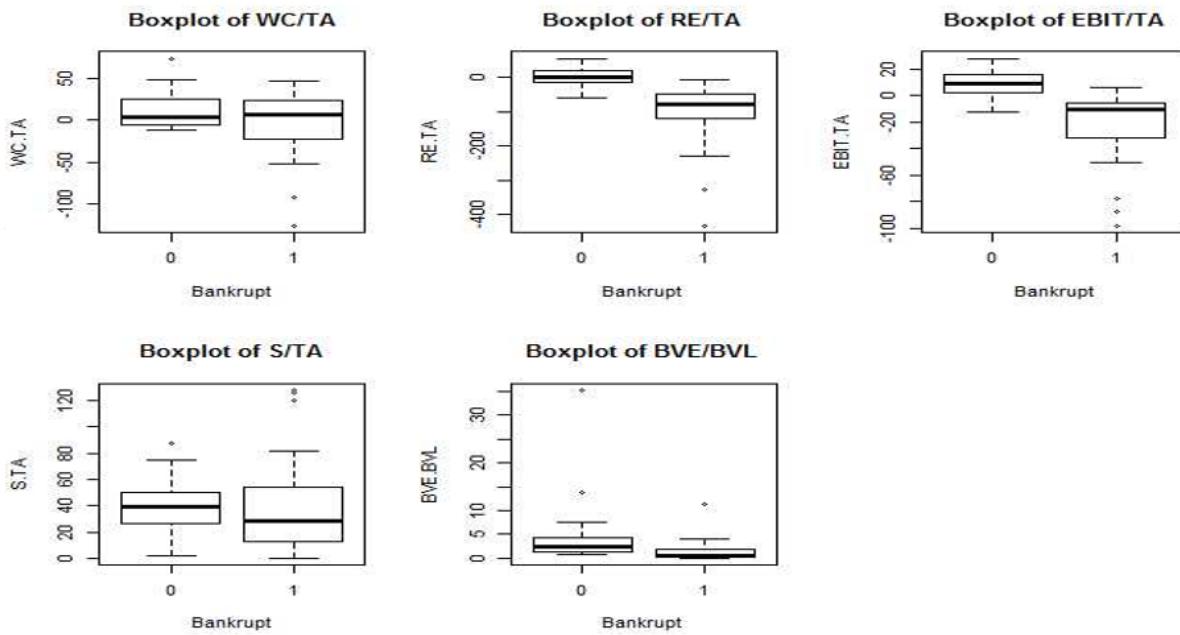
boxplot(split(RE.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="RE.TA",main="Boxplot of RE/TA")

boxplot(split(EBIT.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="EBIT.TA",main="Boxplot of EBIT/TA")

boxplot(split(S.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="S.TA",main="Boxplot of S/TA")

boxplot(split(BVE.BVL,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="BVE.BVL",main="Boxplot of BVE/BVL")
```

# Exploratory Data Analysis



The working capital, retained earnings, and earnings (EBIT) variables all show clear separation between bankrupt and non-bankrupt firms, in the way that would have been expected. The sales (S/TA) variable shows less **predictive** [explanatory?] power with the bankrupt firms actually having the highest values of sales as the percentage of assets. Non-bankrupt firms have generally higher equity to liabilities ratio. Although the long tail of the variable makes this a little harder to see.

Here's an attempt to fit a logistic regression model to this data. Because we fit a logistic regression, we use the `glm` R command with the specification of `family= binomial`. A portion of the output is on the slide.

# Regression Analysis

```
bank1 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL, family=binomial)  
summary(bank1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	7.42646	6.35770	1.168	0.243
WC.TA	-0.15587	0.12208	-1.277	0.202
RE.TA	-0.07605	0.06311	-1.205	0.228
EBIT.TA	-0.49111	0.32260	-1.522	0.128
S.TA	-0.08040	0.09216	-0.872	0.383
BVE.BVL	-2.07764	1.47488	-1.409	0.159

Test for statistical significance:  
All p-values > 0.1: none of the coefficients is statistical significant

Null deviance: 69.315 on 49 degrees of freedom  
Residual deviance: 11.847 on 44 degrees of freedom

```
gstat = bank1$null.deviance - deviance(bank1)  
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-1))  
gstat
```

[1,] 57.46799 4.049594e-11

Test for overall regression:  
p-value ≈ 0: The overall regression has predictive power

The overall regression pvalue is 4.049594e-11.

Interestingly, the output points out that no regression coefficient is statistically significant, since the p-values are all larger than 0.1. However, **for the test of the regression** we reject the null hypothesis that all regression coefficients are zero, indicating that the regression has some explanatory **or predicting** [?explanatory only?] power for the response variable.

This is an extreme example where none of the variables would be selected to be included in the model if we were to discard those that are not statistically significant. While it's clear from the test of the overall regression that *some* of the predictors should be considered to be kept in the model. But which ones? To address this question, we perform variable selection.

### Lecture 5.1.3 – Prediction Risk Estimate

The topic of this lesson is prediction risk estimation. I introduce several criteria that can be used to select among models with different combinations of the predicting variables. I will also illustrate how to compute those criteria using the R statistical software with a specific example.

#### BIAS-VARIANCE TRADEOFF

The goal of a regression analysis is to build a model that explains and/or predicts well. An important aspect in predictions is how it performs in new settings, thus we would like to have a prediction with low uncertainty for new settings. This means that we're willing to give up some bias to reduce the variability in the prediction.

From [Wikipedia](#):

- The *bias* is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The *variance* is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

Generally, models with many covariates have low bias but high variance. Models with few covariates have high bias but low variance. The best predictions come from balancing these two extremes. This is called the bias-variance tradeoff.

A measure of the bias-variance tradeoff is the prediction risk. The prediction risk varies across the set of submodels  $S$ , thus we write the prediction risk as a function of  $S$ . In the regression analysis assuming normality, the prediction risk is the sum of expected squared differences between fitted values by the model  $S$  and future observations.

#### Prediction Risk: Measure of the Bias-Variance Tradeoff

$$R(S) = \frac{1}{n} \sum_{i=1}^n E(\hat{Y}_i(S) - Y_i^*)^2$$

for a submodel  $S$ , with  $\hat{Y}_i(S)$  the fitted response for model  $S$  and  $\hat{Y}_i^*$  the future observation.

We want to minimize the risk of making poor predictions based on the fitted model. For generalized linear regression, the prediction risk is the minus expected log likelihood

function. However, we cannot obtain the prediction risk because we do not have the future observations at the time of prediction.

## TRAINING RISK

So how to estimate the prediction risk? One approach is to compute the prediction risk for the observed data and take the sum of squared differences between fitted values for sub model S and the observed values. This is called the training risk.

### Training Risk

- Replace with actual observations

$$R_{tr}(S) = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i(S) - Y_i)^2$$

for a submodel S, with  $\hat{Y}_i(S)$  the fitted response for model S and  $Y_i$  the future observation.

- Use data twice (data snooping): upward bias in the estimate of the risk
- Always prefers larger/more complex model

→ **Correct for the bias**

$$R_{tr}(S) + Complexity\ Penalty$$

However, this is a biased estimate of the prediction risks since we used the data twice: once for fitting the model S and once for estimating the prediction risk. Thus, the training risk is biased upward. In fact, the training risk increases with the number of variables included in the model. The larger the number of variables is, the larger the training risk is.

So what shall we do in this case? We need to correct for this bias. More specifically, we need to penalize the training risk, in such a way that we'll not always prefer complex models. This means we'll add a complexity penalty to the training risk to correct for the bias. In this formulation, the first term increases while the second term increases with model complexity. This formulation is at the basis of all variable selection approaches.

## VARIABLE SELECTION CRITERIA (Mallows, AIC)

The oldest approach for variable selection based on this idea is the so-called Mallow's Cp for which the complexity penalty is two times the size of the model (the number of variables in the submodel) times the estimated variance divided by n.

## Variable Selection Criteria

→ **Correct for the bias:**  $R_{tr}(S) + \text{Complexity Penalty}$

- Mallow's Cp: Complexity Penalty =  $\frac{2|S|\hat{\sigma}^2}{n}$

where  $|S|$  is the model size (number of predictors) and  $\hat{\sigma}^2$  is the estimated variance based on the full model.

- Akaike Information Criterion (AIC): Complexity Penalty =  $\frac{2|S|\sigma^2}{n}$  where  $\sigma^2$  is the true variance.

→ **For AIC, we need to replace  $\sigma^2$  with an estimate (from the full model or from the submodel S).**

Thus this approach assumes that we can estimate the variance from the full model. This is not the case when  $p$  is larger than  $n$ . This estimate is named in honor of Collin Mallows who invented it.

Another criteria is the Akaike information criterion (AIC) which is a more general approach. For linear regression under normality this becomes the training risk plus penalty that looks just like the Mallow's Cp except that the variance is the true variance and not its estimate.

Most statistical software, including R, replaces true variance with the estimated variance of the sub-model  $S$ , which is different from the variance from the full model as specified by the Mallow's Cp .

Another criteria for variable selection is cross validation which is a direct measure of predictive power. Interestingly it can be shown that the leave-one-out cross validation score can be approximated by the sum between the training risk plus the complexity penalty that is just like the one for Mallow's Cp. Except that the variance is for the  $S$  submodel rather than the full model because the variability of a submodel  $S$  is smaller than that for a full model. Then leave-one-out cross validation penalizes complexity less than Mallow's Cp.

While the training risk and the variable selection criteria I provided are for the regression model assuming normality, we can define similarity criteria for any generalized linear model, including logistic regression and poisson regression. Specifically, the training risk is the sum of square deviances for the submodel  $S$ .

To correct for complexity, we'll use similar approaches, as provided before. The most common approaches are AIC and BIC, since their core definition is defined as a function of the local likelihood function.

## Variable Selection Criteria (cont'd)

→ **Correct for the bias:**  $R_{tr}(S) + \text{Complexity Penalty}$

- *Bayesian Information Criterion (BIC):*

$$\text{Complexity Penalty} = \frac{|S|\sigma^2 \log(n)}{n}$$

where  $\sigma^2$  is the true variance.

- For BIC, we need to replace  $\sigma^2$  with an estimate (from the full model or from the submodel S).
- BIC penalizes complexity more than other approaches and thus preferred in model selection for prediction.

## Variable Selection Criteria (cont'd)

→ **Correct for the bias:**  $R_{tr}(S) + \text{Complexity Penalty}$

- *Leave-one-out Cross Validation:*

$$\hat{R}_{CV}(S) = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_{(i)}(S) - Y_i)^2$$

where  $\hat{Y}_{(i)}(S)$  is the i-th predicted value from the submodel S without i-th observation.

- *Leave-one-out Cross Validation Approximation:*

$$\hat{R}_{CV}(S) \approx R_{tr}(S) + \frac{2|S|\hat{\sigma}^2(S)}{n}$$

where  $\hat{\sigma}^2(S)$  is the estimated variance based on the S submodel.

- Leave-one-out CV is approximately AIC when the true variance is replaced by the estimate of the variance from the S submodel.
- Leave-one-out CV penalizes complexity less than Mallow's Cp since  $\hat{\sigma}^2(S) \leq \hat{\sigma}^2(\text{full})$

## Generalized Linear Models

- Logistic regression & Poisson regression
- Training Risk

$$R_{tr}(S) = \frac{1}{n} \sum_{i=1}^n 2Y_i \log[Y_i/\hat{Y}_i(S)] + 2(n_i - Y_i) \log[(n_i - Y_i)/(n_i - \hat{Y}_i(S))]$$

for a submodel S, with  $\hat{Y}_i(S)$  the fitted response for model S and  $Y_i$  the future observation.

### → Correct for the bias

$$R_{tr}(S) + Complexity\ Penalty$$

- AIC & BIC are commonly used for model selection for GLMs

Let's now illustrate how we can compute all this criteria in R with the SAT example.

There are many libraries and commands in R that can be used. I'm providing here only those from the library CombMSC.

The R commands are Cp, for the Mallow's Cp, and then AIC for both AIC and BIC. The Mallow's Cp statistics can be obtained using the Cp command with the model as the input. It also requires the input of the estimated variance of the full model which is the squared residual standard deviation from the model output.

# Model Selection Criteria Using R

```
library(CombMSC)
n = nrow(datasat)
## full model
c(Cp(regression.line, S2=24.86), AIC(regression.line, k=2), AIC(regression.line, k=log(n)))
[1] 7 472 487
## reduced model
c(Cp(regression.red, S2=24.86), AIC(regression.red, k=2), AIC(regression.red, k=log(n)))
[1] 29 491 498
```

- Mallow's Cp:  $\hat{\sigma} = 24.86$  is the estimated standard deviation for the full model
- BIC: It is similar to AIC except that the AIC complexity is further penalized by  $\log(n)/2$
- The values of the three criteria are different and not comparable
- The full model is better according to all three criteria

```
library(lars)
object = lars(x = predictors, y = log(EDCost.pmpm))
plot(object)
object$Cp

plot.lars(object, xvar="df", plottype="Cp")
```

[3rd number is BIC even though it uses the AIC function]

For AIC, we need to specify k equal to 2. And for BIC, we need to specify log of n, where n is the sample size. Note that the values for the three criteria are different and not comparable for a given model. Based on this output, the full model is better, according to all three criteria since the value are smaller for the full model.

To summarize, in this lesson I provided the most common criteria for comparing models, thank you.

## Lecture 5.1.4 – Model Search

In this lesson, I will describe the several approaches used for model search, given a selected criterion.

An important aspect in prediction is how it performs in new settings. Thus, we'd like to have a prediction with low uncertainty for new settings. This means that we're willing to give up some bias, but to reduce the variability in the prediction.

Generally, models with many covariates, with many predicting variables, have low bias but high variance. And models with few predicting variables have high bias but low variance. Variable selection reduces to balancing these two extremes.

Thus, we first need to choose a model selection criteria, or estimate of the prediction risk to balance the Bias-Variance Tradeoff. Once we chose model selection criteria such as cross-validation or AIC, we then need to search through all models, assign a score to each model, and then choose the model with the best score.

But how to search over all models? For  $p$  predicting variables, there are  $2^p$  different submodels. For example, for  $p$  equal to six, there's 64 different models. For  $p$  equal to ten, there are 1,024 models. Thus, if  $p$  small, in a range of up to six or seven predicting variables, we can fit all models and compare them with respect to the criteria I use for estimating the prediction of risk, say, AIC, then select the model, the submodel with the smallest AIC.

But if  $p$  is large, it's infeasible to fit a large number of submodels. Instead, we can perform a heuristic search, such as a stepwise regression, which is a greedy search algorithm, where greedy means we always take the biggest jump, up or down, in the selected criterion.

# Model Search

- If  $p$  is the number of predicting variables, there are  $2^p$  possible submodels;
  - If  $p$  small, fit all submodels
  - If  $p$  large, search using heuristics/greedy search
- *Stepwise Regression:*
  - Forward: Start with no predictor and one at a time
  - Backward: Start with all predictors and drop one at a time
  - Forward-Backward: Add and drop one variable at a time iteratively

→ Stepwise regression is a greedy algorithm; it does not guarantee to find the model with the best score  
→ Forward stepwise regression is preferable over backward stepwise regression  
→ It does not necessarily select the same model as the one selected using backward stepwise regression

There are three types of stepwise regression. *Forward*, meaning that we start with no predictor or with a minimum model, and add one predictor at a time. *Backward*, meaning we start with all predictors, the full model and drop one predictor at a time. And *Forward-Backward* stepwise regression, meaning adding and discarding one variable at a time iteratively.

## Forward Stepwise Regression

- Select criterion for model selection (e.g. AIC)
- Fit  $p$  marginal regressions for all  $j=1,\dots,p$ :
  - $C_j$  the criterion value for the model with the  $j$ -th predictor;
  - Select  $j_1$  predictor with the smallest criterion value if  $C_{j_1}$  smaller than the criterion value for the model without any predictor;
- Fit  $p-1$  regressions with  $j_1$  predictor in the model and adding another predictor for all  $j=1,\dots,j_1-1, j_1+1, p$ :
  - $C_j$  the criterion value for the model with the  $j$ -th predictor
  - Select  $j_2$  predictor with the smallest criterion value to add to the model including  $j_1$  predictor if  $C_{j_2}$  smaller than  $C_{j_1}$ ;
  - if  $C_{j_2}$  larger than  $C_{j_1}$  then stop; the selected model includes only the  $j_1$ -th predictor;
- Continue adding predictors until the criterion does not improve

A few important aspects to remember. Stepwise regression is a greedy search algorithm. It does not guarantee to find the model with the best score. Forward stepwise regression starts with no predictors in the model and usually tends to select smaller models. Because of that, it's also preferable over Backward stepwise regression, which starts with a full model. To recall, the three stepwise regression approaches do not necessarily select the same model, especially when  $p$  is large.

## Backward Stepwise Regression

- Select criterion for model selection (e.g. AIC)
- Fit *full* model and discard one predictor for all  $j=1, \dots, p$ :
  - $C_j$  the criterion value for the model with the  $j$ -th predictor discarded
  - Select  $j_1$  predictor to be discarded with the smallest criterion value if  $C_{j_1}$  smaller than the criterion value for the full model.
- Fit the regressions without  $j_1$  predictor and discarding another predictor for all  $j=1, \dots, j_1-1, j_1+1, p$ :
  - $C_j$  the criterion value for the model with the  $j$ -th predictor discarded
  - Select  $j_2$  predictor with the smallest criterion value to discard from the model if  $C_{j_2}$  smaller than  $C_{j_1}$ ;
  - if  $C_{j_2}$  larger than  $C_{j_1}$  then stop; the selected model discards only the  $j_1$ -th predictor;
- Continue discarding predictors until the criterion does not improve
 

- ➔ It cannot be performed if  $p$  larger than  $n$
  - ➔ More computationally expensive than forward stepwise regression
  - ➔ It will select larger models if  $p$  large

How does the Forward Stepwise Regression work? First, select a criterion for the model selection, say, AIC. Then feed the model with no predictors in the model, or with predictors that will be always in the model, so that's not selected. And then compute the AIC score for this model.

Next, feed all models with one predictor by adding one predictor to the minimum model, a total of  $p$  models. For each model, compute AIC and compare to the AIC from the model with no predictor. Select the predictor with the smallest AIC, say this is the  $j_1$  predictor.

The next step, explore models with two predictors if the starting model is there are no predictors, by adding one additional predictor to the  $j_1$  predictor, selected [in] the previous step. Again fit  $p-1$  models. Compute AIC for each model and compare.

If the AIC of each of these  $p-1$  submodels is larger than the one with the  $j_1$  predictor, only then, stop, and the selected model is the one including one predictor only.

If not, select to add a second predictor to the model that has a smallest AIC among the  $p-1$  model. Then, consider models with three predictors by adding a third predictor to the two selected variables in the previous step.

Continue adding one predictor at a time until the AIC score does not improve or does not decrease anymore.

In the Backward Stepwise Regression, first fit the model with all predictors and then compute AIC for this model. Next, fit all models by dropping, by discarding one predictor, a total of  $p$  models. For each model compute AIC and compare to the AIC from the model with all predictors.

If the full model has a smaller AIC than all AIC scores from discarding one predictor, then the selected model is the full model. Otherwise, discard the predictor with the smallest AIC. Say, this will be the  $j_1$  predictor.

The next step, explore models by discarding two predictors, excluding  $j_1$ , which was discarded at the previous step, and a second predictor among the predictors that were not considered. Again, fit this model's computed AIC for each submodel, and compare . If the AIC of each of these models is larger than the one without  $j_1$  predictor, then stop. And the selected model is the one discarding  $j_1$  predictor.

If not, select to discard a second predictor, the one that has the smallest AIC. Then consider models by discarding three predictors.

Continue discarding one predictor at a time until the AIC score does not improve or does not decrease anymore. To note, that we cannot apply backward stepwise regression if the number of predictors is larger than the sample size, since we cannot fit the full model.

Moreover, backward stepwise regressions is more computationally expensive than forward stepwise regression, since it fits larger models. You'll also prefer larger models.

In summary in this lesson, I introduced stepwise regression, one of the most common approaches used in variable selection.

## Lecture 5.1.5 – Model Search Data Examples

In this lesson, I will illustrate model search with two data examples using the R statistical software.

### **Approach 1: Mallow's Cp (useful when there are no control variables)**

We'll return to the SAT example. We'd like to identify or select explanatory variables that explain the variability in the state-level average SAT score. For 6 predictors, there are 64 possible models. The **leaps** command in R from the **leaps library** allows comparing all possible models using various variable selection criteria although not AIC or BIC.

### Compare All Models

```
library(leaps)
out = leaps(datasat[,-c(1,2)], sat, method = "Cp")
cbind(as.matrix(out$which),out$Cp)
  1 2 3 4 5 6
1 0 0 0 0 0 1 34.026
1 1 0 0 0 0 0 47.639
1 0 1 0 0 0 0 187.387
1 0 0 1 0 0 0 269.647
1 0 0 0 1 0 0 306.188
1 0 0 0 0 1 0 307.076
.....
best.model = which(out$Cp==min(out$Cp))
cbind(as.matrix(out$which),out$Cp)[best.model,]
```

The output includes all 64 combinations of predictors with specification of which predictors are in the model and the Cp score value for each model.

The best model with respect to Mallow's Cp criterion: Years, Public, Expend, Rank (last four predictors in the input dataset)  
**Does not allow for specification of confounding variables!!!**

In this command, we only need to specify the matrix of predicting variables, the response variable, and the criteria, Mallow's Cp. The output includes all 64 combinations of predictors with specification of which predictors are in the model, and the Cp score value for each model. I'm not providing on this slide the entire output since it has 64 rows, 1 for each model.

So let's learn how we can read this output. The ones in the output correspond to the variables included in the model, the zeros to variables not included in the model. The last column corresponds to the values of the CP statistic.

For example, the first row corresponds to the model where only the sixth predictor corresponding to the sixth column in the matrix of the predictive variables is in the model. In this case, the sixth predictor is rank. When only this predictor is in the model, the Cp statistic is 34.026. The last row provided on the slide, corresponds to the model where the fifth predictor is the only predictor included in the model. The Cp statistic is 307 for this model.

How can we use this output to find the model with the smallest Cp statistic? We can find first the index of the row with smallest Cp using the command “which” then we can output that row.

```
Best.model = which(out$Cp==min(out$Cp))
```

The model with the smallest Cp has ones for variables three to six. These variables are years, public, expend, and rank.

Note that two of the predictors ~~are confounding~~, are the controlling variables, and should be included in the model [takers and rank?][confounding & controlling are 2 different ideas, verify she means controlling]. Variable selection should be performed only for the four predictor explanatory variables. However, the “leaps” function does not allow for such variable selection.

#### **[Please ignore this section**

To perform a stepwise regression, we can use a step command in R. This function in R does allow specification of a minimum model. In this case, the minimum model includes the two confounding variables.

To do so, the first input is the model with the confounding variables only. Then the scope option allows to specify a starting model, in this case, the one with the two confounding variables, and the full model. The direction specified here is forward.

The output shows each step taken to reach the selected model. I'm providing here most of the output to illustrate how stepwise regression is performed. I'm providing, first, the smallest model.

I have to redo this

#### **[Please resume here]**

### **Approach 2: Stepwise Regression Approach (when there are control variables)**

To perform stepwise regression, we can use a `step` command in R. This function does allow specification of a minimum model. In this case, the minimum model includes the two **controlling** variables.

To do so, the first input is the model with the-control variables only. Then the scope options allows us to specify a starting model, in this case, the one with the controlling variables, and a full model. The directions specified here is forward. The output shows each step taken to reach the selected model. I'm providing here most of the output to illustrate how stepwise regression is performed.

## Stepwise Regression

```
# Forward Stepwise Regression
step(lm(sat~log(takers)+rank), scope = list(lower=sat~log(takers)+rank,
                                              upper = sat~log(takers)+rank+expend+years+income+public), direction = "forward")

Start: AIC=346.7
sat ~ log(takers) + rank
  Df Sum of Sq RSS   AIC
+ expend 1  13149.5 32380 331.66
+ years   1    9827.2 35703 336.55
<none>          45530 346.70
+ income  1   1305.3 44224 347.25
+ public   1      15.9 45514 348.69

Step: AIC=323.9
sat ~ log(takers) + rank + expend + years
  Df Sum of Sq RSS   AIC
<none>          26637 323.90
+ income  1   26.6165 26610 325.85
+ public   1    4.5743 26632 325.89

Call:
lm(formula = sat ~ log(takers) + rank + expend + years)

Coefficients:
(Intercept) log(takers)      rank     expend     years
            388.425    -38.015     4.004     2.423    17.857
```

- Stepwise regression in R allows for specification of a reduced model, including confounding variables
- Selected model: `expend & years` with confounding variables `log(takers) & rank`

First, the smallest model is estimated and AIC computed, which is 346.7, then each of the 4 predictors is added as shown with a plus. For example, when years is added the AIC is 336. All AIC values including the one, when no additional value is added

corresponding to the role with none are provided. Since the smallest AIC is for the model adding expend, this is the first variable added to the model.

In the next portion of the output, we're adding one variable among the three left. AIC has provided and it is smallest when years is added to the model. But the third iteration the two variables left are considered to be added but since the AIC for adding any of those two variables is not smaller than without the two, we select the model from the last step that includes log of takers, rank, expenditure, and years.

Thus, an advantage of stepwise regression over "Leaps" which compares all the models, is that it allows for confounding variables to be part of the starting model. The selected model is the one with expend and years along with the confounding variables for this particular example. We'll compare this model selection with the one provided by other approaches in a different lesson.

We can also perform a backwards stepwise regression using the step command. For this, we need to specify the full model to begin with, along with the scope option, just like the implementation of the forward stepwise regression. We also need to specify the direction backward.

As provided on this slide, the first model fitted is the full model on the AIC value also computed. Next, we discard one variable as indicated by minus and the AIC values are computed for each model. The model, without discarding any of the variables is also included in the rows corresponding to none.

# Stepwise Regression (cont'd)

# Backward Stepwise Regression

```
full = lm(sat~log(takers)+rank+expend+years+income+public)
```

```
minimum = lm(sat~log(takers)+rank)
```

```
step(full, scope = list(lower=minimum, upper = full), direction = "backward")
```

Start: AIC=327.8

sat ~ log(takers) + rank + expend +  
years + income + public

	Df	Sum of Sq	RSS	AIC
- public	1	25.0	26610	325.85
- income	1	47.0	26632	325.89
<none>		26585	327.80	
- years	1	4588.8	31174	333.77
- expend	1	6264.4	32850	336.38

Step: AIC=325.85

sat ~ log(takers) + rank + expend + years + income

	Df	Sum of Sq	RSS	AIC
- income	1	26.6	26637	323.90
<none>		26610	325.85	
- years	1	5452.8	32063	333.17

Step: AIC=323.9

sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
<none>		26637	323.90	
- years	1	5743.5	32380	331.66
- expend	1	9065.8	35703	336.55

- Selected model: `expend & years` with confounding variables `log(takers) & rank`
- The same model was selected using forward regression; generally for a large number of predictors the two methods will select different models

Comparing the AIC values, the model that discards `public` variable has the smallest AIC. Next, we discard one of the three other predicting variables and compute the AIC values. The model without `income` variable has the smallest AIC and thus this variable is discarded. Last, we discard one of the two other predictive variables left but AIC does not improve, does not decrease anymore and thus we stop here. The model selected is the one with `expenditure` and `years` along with the confounding variables [`takers` and `rank`]. This is the same auto selective using forward regression. Generally, for a large number of predictors, the two methods will select different models.

Let's now provide a variable selection analysis for the bankruptcy data example. For this example, we like to identify which financial indicators are associated with bankruptcy for telecommunications firms. Let's begin with comparing all models.

We can, again, use the `leaps` R command. The output is similar to the one provided for the SAT example. The selected models, the one with variables 2, 3 and 5, specified as TRUE in the output.

## Compare All Models

```
out = leaps(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL), Bankrupt)
best.model = which(out$Cp==min(out$Cp))
as.matrix(out$which)[best.model,]
  1   2   3   4   5
FALSE TRUE TRUE FALSE TRUE
```

The best model selected with respect to Mallow's Cp: RE.TA, EBIT.TA, BE.BVL

Thus, the best model selected with respect to the Mallow's Cp includes retained earnings as a percentage of total asset (RE.TA), earnings before interests and taxes as a percent of total asset (EBIT.TA), and book value of equity divided by the book value of total liabilities (BE.BVL). The fitted model, using these predictors, is provided here.

## Compare All Models

```
out = leaps(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL), Bankrupt)
best.model = which(out$Cp==min(out$Cp))
as.matrix(out$which)[best.model,]
  1   2   3   4   5
FALSE TRUE TRUE FALSE TRUE
bank2 = glm(Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family=binomial, x=T)
summary(bank2)
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.29478  1.12317 -0.262  0.7930
RE.TA       -0.05627  0.02745 -2.050  0.0404 *
EBIT.TA     -0.16763  0.09269 -1.808  0.0705 .
BVE.BVL    -0.62975  0.39429 -1.597  0.1102
...etc...
```

The best model selected with respect to Mallow's Cp: RE.TA, EBIT.TA, BE.BVL

- RE.TA is now statistically significant at  $\alpha = 0.05$
- Not all coefficients are statistically significant

To remind you, when I fitted the full model with all the five predictors, none of the regression coefficients was statistically significant while the overall regression was statistically significant.

We see now that as we discarded two predictors, two coefficients are statistically significant at the level 0.1 and one coefficient, RE.TA, at the level 0.05. The RE.TA coefficient says that an increase of 1 percentage point in the retained earnings as a percentage of total assets (RE.TA), is associated with a decrease in the odds of going bankrupt in the next year by 5.6% holding all else fixed. The EBIT.TA coefficient says

that an increase of 1 percentage point in the earnings before interest and taxes (EBIT.TA) is associated with a decrease in the odds of going bankrupt by 17%.

Moreover, if we perform a test for subset regression coefficients comparing the reduced model with the selected predictive variables, in this case, the hypothesis versus the full model, the p-value is large, indicating that we do not reject the null hypothesis of the reduced model.

```
gstat = deviance(bank2) - deviance(bank1)                                Statistical significant
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-length(coef(bank2))))  
gstat  
[1] 4.040336 0.1326332
```



The null (reduced model) not rejected

There's one outlier among the companies included in this analysis, and this is 360 Networks, which is the first observation in this data set. This firm was in the business of building computer networks and was one of the only two firms that ultimately went bankrupt that had positive earnings the year before insolvency. Its value of RE.TA was also not negative, but part of this could be from the nature of its business. The thousands of miles of cable that it owned resulted in the firm having 6.3 billion in total assets only 3 months before it declared bankruptcy, making RE.TA less negative.

If we omit this observation and try to fit a model using all of the predicting variables, we get the results on the slide.

# Remove Outlier

```
bankrupt2 = bankruptcy[-1,]  
attach(bankrupt2)  
bank3 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL,  
family=binomial, data=bankrupt2)  
summary(bank3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	265.467	576281.709	0	1
WC.TA	-4.297	12439.717	0	1
RE.TA	-1.516	5131.146	0	1
EBIT.TA	-17.043	35543.170	0	1
S.TA	-2.859	7408.747	0	1
BVE.BVL	-77.540	184903.000	0	1

The model fits perfectly. This is complete separation, and the solution is to simplify the model if that is possible.

From this output, all p-values are one. What is going on in this model? The model fits perfectly. This is what is called complete separation.

The solution, again, is to simplify the model if that is possible. And I will note that complete separation is not a bad thing in and of itself. It just indicates that the possibility of a simpler model being good enough should be explored.

Performing against the best subset selection, the selected predicting variables are the same as for the model with outliers.

## Compare All Models: Without Outlier

```
out = leaps(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL),Bankrupt)
best.model = which(out$Cp==min(out$Cp))
as.matrix(out$which)[best.model,]
  1   2   3   4   5
FALSE TRUE TRUE FALSE TRUE
bank4 = glm(Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family=binomial, x=T)
summary(bank4)
  Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09166  1.47135 -0.062  0.9503
RE.TA       -0.08229  0.04230 -1.945  0.0517
EBIT.TA     -0.26783  0.15854 -1.689  0.0912
BVE.BVL     -1.21810  0.76536 -1.592  0.1115
exp(coef(bank2)[-1])
  RE.TA   EBIT.TA   BVE.BVL
0.9452862 0.8456655 0.5327273
exp(coef(bank4)[-1])
  RE.TA   EBIT.TA   BVE.BVL
0.9210091 0.7650371 0.2957930
```

Now, the statistical significance has changed with RE.TA and EBIT.TA being statistically significant at levels 0.1 but not at the levels 0.05.

Comparing the estimated regression coefficients from the reduced models with and without an outlier, we can see that the estimate regression are smaller for the model without the outlier, but they have not changed sign. Furthermore, we can perform stepwise regression using the step command in R. And the output is provided here.

The interpretation of the output is the same as for the SAT example. The stepwise regression in steps selected four predictive variables in comparison to the best subsets which selects three. The additional variable selected is WCTA.

In summary, in this lesson, I provided an implementation in R of stepwise regression with two particular examples.

## LECTURE 5.2: REGULARIZED REGRESSION

### Lecture 5.2.1 – Regularized Regression: Penalties

In this lesson, I'll introduce a different approach for variable selection from those ones that we've learned in the previous lessons. Those approaches are introduced in this lesson. Perform variable selection and estimation simultaneously, and they're referred to as penalized or regularized regression.

Let's return to the concept of bias-variance tradeoff. As presented in the previous lesson, the prediction risk is a measure of the bias-variance tradeoff. If we decompose the prediction risk, we can rewrite it as the sum between three components. One is the variance of a future observation, or sigma squared, which is an irreducible error and thus cannot be controlled. The other two components are the bias squared and the variance of the prediction, which is also called the mean. Their sum is also called the mean square error.

Thus, the prediction risk is the sum between the irreducible error and the mean square error, where only the latter can be controlled. Mean square error is commonly used in statistics to obtain estimators that may be biased, but less uncertain than unbiased ones. And that's preferred.

### ***Prediction Risk: Measure of the Bias-Variance Tradeoff***

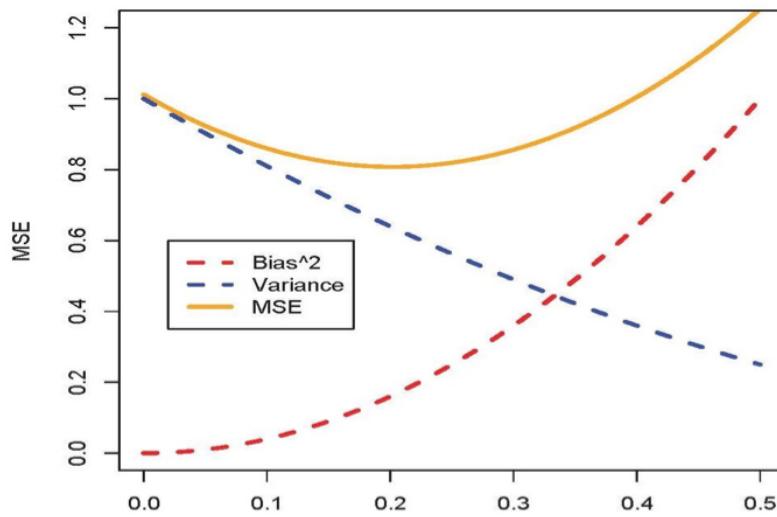
$$\begin{aligned} R(S) &= \frac{1}{n} \sum_{i=1}^n E(\hat{Y}_i(S) - Y_i^*)^2 \\ &= V(Y_i^*) + \text{Bias}^2(\hat{Y}_i(S)) + V(\hat{Y}_i(S)) \end{aligned}$$

Similarly, in variable selection, it is possible to find a model with lower mean square error than the unbiased or full model. It is generic in statistics, almost always introducing some bias yields a decrease in mean square error followed by a latter increase.

This figure depicts the bias-variance tradeoff. We cannot have both low bias and low variance. Instead, when the variance is high, the bias is low and vice versa. There is a

point where the two are low, although not at their lowest levels. If we add the bias squared and the variance, we get the mean square error and thus the yellow line. We can see that the mean square error is minimized at a value that does not correspond to the lowest bias.

## Bias-Variance Tradeoff



In variable selection, the x axis will correspond to the number of predictive variables but starting with the largest number of predictors on the left to the smallest on the right. When we want to trade in some bias for less uncertainty, not all biased models are better. We need a way to find good biased models.

In the approaches introduced next the basis of variable selection is that we would like to penalize large values of beta's jointly. This should lead to multivariate shrinkage of the vector of regression coefficients. Particularly, this translates into penalizing large models with many predictive variables where large units of complex model [?]. However, you will need to keep in mind that this approach will not always work. There are situations when a complex model is a better fit.

Let's begin with the ordinary least squares where we minimize the sum of square differences between observed and the expected. We minimize with respect to the regression coefficients betas. With penalization, we add a penalty for complexity lambda times the penalty where lambda is a constant that balances the tradeoff between the lack of fit measured by the sum of the squares and the complexity

measured by the penalty which depends on the regression coefficients. The bigger lambda is, the bigger the penalty for model complexity.

## Regularized Regression

**Without Penalization:** Estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  by minimizing the sum of squared errors

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2$$

**With Penalization:** Estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  by minimizing the penalized sum of squared errors

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \text{Penalty}(\beta_1, \dots, \beta_p)$$

The bigger  $\lambda$ , the bigger the penalty for model complexity.

Three penalties are as follows.

- L0 penalty, which is the number of nonzero regression coefficients. So when we apply this penalty to the vector of the coefficients, this would be equal to the number of nonzero regression coefficients. Using this penalty, the penalized least squares is equivalent to searching over all models and thus not completely viable for a large number of predictive variables.
- The next penalty is so-called L1 penalty, which applied to the vector of regression coefficients, is equal to the sum of the absolute values of the regression coefficients to be penalized.

## Comparing Penalties

- $L_0$  penalty: provides the best model given a selection criterion but it requires fitting all submodels
- $L_1$  penalty measures sparsity

Example: Consider the following two vectors of length p

$$u = (1, 0, \dots, 0) \text{ & } v = \left( \frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}} \right)$$

(u is sparse since it contains many zeros)

$$\|u\|_1 = 1 \text{ & } \|v\|_1 = \sqrt{p} \text{ versus } \|u\|_2 = 1 \text{ & } \|v\|_2 = 1$$

- $L_2$  penalty is easy to implement but it does not do variable selection

- Minimizing the penalized least squares using this penalty will force many betas, many regression coefficients to be 0s. The resulting regularized regression is the so called LASSO regression.
- The last penalty is the L2 penalty, which apply to a vector, to the vector regression coefficients, is equal to the sum of the squared regression coefficients to be penalized. Minimizing the penalized least squares using this penalty accounts for multicollinearity, but does not perform variable selection. The result in regularized regression is a so-called ridge regression. We'll review LASSO and ridge regression in the next lesson.

## Regularized Regression (cont'd)

The penalized sum of squared errors:

$$Q(\beta_1, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \text{Penalty}(\beta_1, \dots, \beta_p)$$

We consider three choices for the penalty:

$L_0$  penalty:  $\|\beta\|_0 = \#\{j: \beta_j \neq 0\}$   $\Rightarrow$  Maximizing Q means searching through all submodels.

$L_1$  penalty:  $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$   $\Rightarrow$  Maximizing Q forces many  $\beta_j$ 's to be zeros. (*LASSO Regression*)

$L_2$  penalty:  $\|\beta\|_2 = \sqrt{\sum_{j=1}^p \beta_j^2}$   $\Rightarrow$  Maximizing Q accounts for multicollinearity. (*Ridge Regression*)



In this slide, I'll provide some insights on the three penalties.  $L_0$  penalty provides the best model given a selection criteria, but it requires fitting all submodels.

$L_1$  penalty measures sparsity as illustrated with the following example. Consider two vectors, U and V. The vector U has only one non-zero value and thus sparse. The second vector V has no non-zero values and thus not sparse. If we take the L1 norm of both vectors, in other words, take the sum of the absolute values in the vector, then the L1 norm is 1 for U and square root of p for the second vector V. Thus L1 norm for U is much smaller for the sparse vector.

If we take the L2 norm, that is take the sum of the squared values in the vector, then the L2 norm is 1 for both vectors, not distinguishing between a sparse and non sparse vector. Last, L2 penalty is easy to implement, but again, it does not measure sparsity. And thus it does not perform variable selection.

To summarize in this lesson, I introduced a concept of penalization and regression. In the next lesson, I will introduce the Ridge and LASSO of regression, the common approaches for penalized regularized regression.

## Lecture 5.2.2 – Regularized Regression: Approaches

In this lesson I'll illustrate how to use the penalties we learn about in the previous lesson, in a context of penalization, regularization, of regression, for variable selection. And I'll focus particularly on the advantages and limitations of the two most common approaches for regularized regression, Ridge regression and LASSO regression.

I'll begin by pointing out that, in regularized regression for variable selection, we need to first rescale all the predicting variables in order to be comparable on the same scale. I recommend also rescaling the response variable, if numeric, although it is not required for the implementation of regularized regression. Last after selecting the variables for the final model, you may fit the model using the variables on the original scale, for easy interpretation of the regression coefficient and the estimate regression line.

## Variable Standardization & Notation

For regularized regression,

- Rescale the  $j$ -th predicting variable  $x_j$  for  $j=1, \dots, p$  as follows:

$$\frac{1}{n} \sum_{i=1}^n x_{ij} = 0, \frac{1}{n} \sum_{i=1}^n x_{ij}^2 = 1$$

It is recommended to also

- Rescale the response variable as follows:

$$\frac{1}{n} \sum_{i=1}^n Y_i = 0, \frac{1}{n} \sum_{i=1}^n Y_{ij}^2 = 1$$

→ Use the original scale when fitting the selected model for interpretation of the regression coefficients.

Let us now overview Ridge Regression. For this regression model, the penalty is the sum of square regression coefficients times the lambda constant. Minimizing those

provides a closed-form expression for the estimated regression coefficients as provided on the slide. The formula looks similar to that of the estimated coefficients under the ordinary least squares that we learned in a lecture about estimation of the multiple linear regression model.

- Minimize

$$SSE_{\lambda}(\beta) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \sum_{i=1}^n \beta_i^2$$

- The estimated regression coefficients:

$$\hat{\beta} = (\mathbf{X}\mathbf{X}^T + \lambda I)^{-1} \mathbf{X}^T \mathbf{Y} \text{ where } I \text{ is the identity matrix}$$

When  $\lambda = 0$  we get the least squares estimate (low bias, high variance). When  $\lambda = 1$  we get  $\hat{\beta} = 0$  (high bias, low variance).

- Commonly used to fit a regression model under multicollinearity
- Not used for model selection: it does not “force” any  $\hat{\beta}_j = 0$

The only addition is this term, lambda times identity matrix, which is due to the penalty. If lambda is 0, we have the estimated regression coefficients for the ordinary least squares regression without the penalty, without penalization. This estimator has low bias but high variability. On the other hand, if lambda is equal to 1, then the estimated coefficients are 0. Thus the fitted regression has high bias but low variability.

Ridge regression has been developed to correct for the impact of multicollinearity, if there is multicollinearity in the model, but all predicting variables are considered to be included in the model, then ridge regression will allow for re-weighting the regression coefficients in a way that those corresponding to correlated predictor variables share their explanatory powers and thus minimizing the impact of multicollinearity on the estimation and statistical inference of the regression coefficients.

However, ridge regression does not perform, and thus cannot be used for, variable selection. It only shrinks coefficients to zero but it does not force coefficients to be zero, as needed in variable selection. On the other hand, the Lasso regression, where the penalty is the sum of absolute values of the regression coefficients except for intercept, that force coefficients to be zero.

- Lasso (Least Absolute Shrinkage and Selection)

- Normal Linear Regression: Minimize

$$SSE_{\lambda}(\beta_0, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- Generalized Linear Model: Minimize

$$SSE_{\lambda}(\beta_0, \dots, \beta_p) = -l(\beta_0, \dots, \beta_p) + \lambda \sum_{j=1}^p |\beta_j|$$

where  $l(\beta)$  is the log-likelihood function.

- The estimated regression coefficients: Use numerical algorithms since there is not a close form expression
- Used for model selection: it does “force” any  $\hat{\beta}_j = 0$

For regression analysis under normality assumption, the penalized least squares problem is as on this slide. However, for generalizing your model would replace the sum of least squares by [minus] local likelihood function. The penalty for complexity is the same thus, we can apply Lasso to standard linear regression, logistic regression, poisson regression, and other linear models.

Unlike the Ridge regression, we do not have a closed-form expression for the estimated regression coefficients under this model. Thus, numerical algorithms need to be employed to minimize with respect to the regression coefficients. Because the minimization problem is convex we only have a unique solution for the regression coefficients. The [existing] numerical algorithms used for the minimization problem provide accurate estimates.

I will note that Lasso performs estimation of variable selections simultaneously. However, the estimated regression coefficients from Lasso are less efficient than those provided by the ordinary least squares. And thus, once lasso regression has provided a selected model, the selected predicting variables, the ordinary least squares should be used to estimate the regression coefficients for the model with the selected predicting variables.

One important aspect in all regularized regression problems, whether ridge regression, lasso regression, or other types of regularization is determining the penalty constant lambda. This constant has the role of balancing the tradeoff between lack of fit and

model complexity. And thus, different lambdas will provide different models, as explained up in the Ridge regression example.

But how do we choose lambda? The answer involves a trick called cross-validation. The basic idea of cross-validation is to leave out some of the data when fitting a model, that is, split the data into two parts. One part, also called a training data, will be used to fit the model given a specific lambda, and thus give the estimated regression coefficients given that lambda constant.

Split the data  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  into:

- **Training set:** Fit the penalized model given  $\lambda$ , i.e. estimate  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$
- **Testing/Validation set:**
  - Estimate the mean squared error for normal regression
  - Estimate the classification error rate for logistic regression
  - Generally, estimate a scoring rule depending on the regression problem

The second portion of the data, also called the testing or validation data, will be used to compare the error rate, which will be dependent on lambda since the fitted model depends of lambda. The error rate depends on the problem at hand. If the regression analysis under normality, then we seek to minimize the mean squared error.

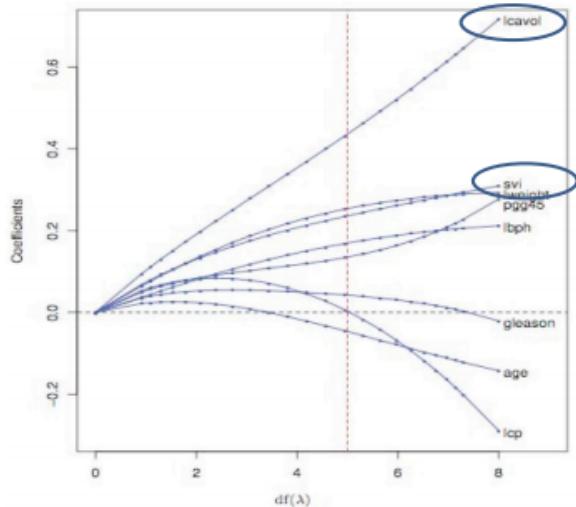
If logistic regression or poisson regression, we can minimize the deviance, the sum of square deviances or for [other?] regression, we can also minimize the classification error rate. One can repeat the process for multiple lambdas.

But how to split the data? The common approach is to divide the data into k folds or subsets approximately of equal sizes. For each fold of data, we take that fold of the data out and use the data without that fold for training the model for fitting the model. Then we predict the response in the fold that we took out based on the fitted model and then obtained the predictions for that fold. Applying those to all folds of the data, we get the predictions of all responses and thus we can compute the error rate based on those predictions.

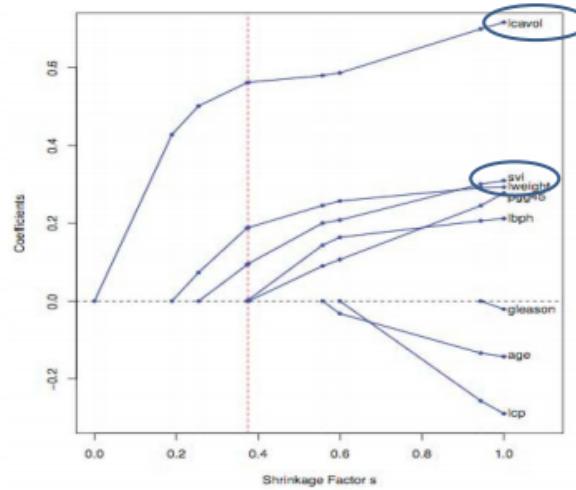
Because the fitted model depends on lambda, so does the error rate. Thus this procedure would be applied for different lambda values, say lambda 1 to lambda v within a specified range. Then select the lambda constant that minimizes the error rate, meaning that best balances the trade-off between bias and variance.

To see the difference in shrinkage and selection of the predicting variables in Ridge vs Lasso Regression, compare the following two figures that were provided from the book acknowledged on the slide. These two figures show the path of each regression coefficient varying by lambda, or some other shrinkage factor depending on lambda, for example, effective degrees of freedom for each regression.

## Lasso vs Ridge Regression



**FIGURE 3.8.** Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter  $\lambda$  is varied. Coefficients are plotted versus  $df(\lambda)$ , the effective degrees of freedom. A vertical line is drawn at  $df = 5.0$ , the value chosen by cross-validation.



**FIGURE 3.10.** Profiles of lasso coefficients, as the tuning parameter  $t$  is varied. Coefficients are plotted versus  $s = t / \sum_j |\hat{\beta}_j|$ . A vertical line is drawn at  $s = 0.36$ , the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.

**Acknowledgement:** From Hastie, T., Tibshirani, R., Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics.

Each point on the coefficient path corresponds to the estimated coefficient for a different lambda. On the left side, they all start at 0, then for each value of lambda, the coefficient changes its value. For ridge regression, in the left plot the path of some of the regression coefficients may intersect the 0 line for large effective degrees of freedom but for others, the regression coefficients increase. For lasso regression, once a coefficient is non-zero, then it does not go back to zero.

For the chosen, for the selected lambda in this example, where the lambda is selected by cross-validation and marked by the vertical dotted line. Only three of the variables are included in the final model using the Lasso variable selection. Those are highlighted.

In contrast, because Ridge Regression is not a model selection procedure, all of the predicting variables are included in a model. However, the three selected variables by Lasso have the larger coefficients in the Ridge model. The path for the regression coefficients, for the predicting variables not selected by Lasso are close to the zero line or across the zero line in the Ridge regression graph. However, they are not forced to be zero again.

While Lasso has been extensively used for variable selection, it does have a series of limitations. In the case, where  $p$  the number of predictors is larger than  $n$  the number of observations, that is, more variables than observations, the Lasso selects, at most,  $n$  variables because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method for the usual case where  $N$  is larger than  $P$ .

If there exists high correlation among predictors it has been empirically observed that the prediction performance of the Lasso is dominated by ridge regression. Last, if there is a group of variables among which the correlation are very high, then the Lasso tends to select only one variable from that group and does not care which one is selected.

One method to overcome this problem is elastic net. Similar to the lasso, the elastic net simultaneously performs variable selection and continues shrinkage and can select groups of correlated variables. Elastic net often out performs the lasso in terms of prediction accuracy.

# Elastic Net

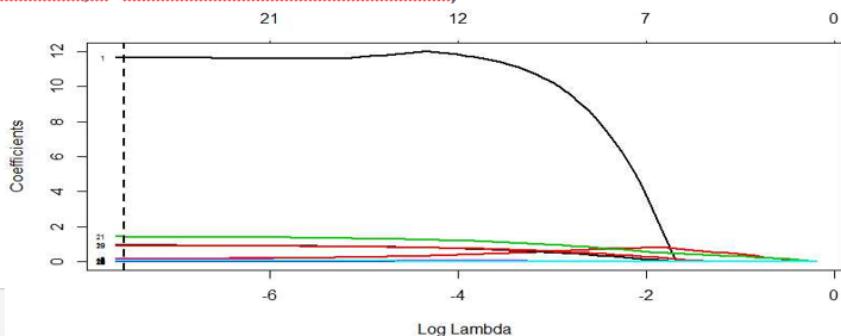
- Minimize

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j|$$

- $L_1$  penalty generates a sparse model.
- $L_2$  penalty
  - Removes the limitation on the number of selected variables;
  - Encourages group effect;
  - Stabilizes the  $L_1$  regularization path

## Elastic Net Regression

```
# 10-fold CV to find the optimal lambda
enetmodel.cv=cv.glmnet(predictors.log(EDCost.pppm),alpha=0.5)nfolds=10)
## Fit lasso model with 100 values for lambda
enetmodel = glmnet(predictors.log(EDCost.pppm),alpha = 0.5) nlambd = 100
## Plot coefficient paths
plot(enetmodel,xvar="lambda",label=TRUE, lwd=2)
abline(v=log(enetmodel.cv$lambda.min),col='black',lty = 2,lwd=2)
## Extract coefficients at optimal lambda
coef(enetmodel,s=enetmodel.cv$lambda.min)
```



- RankingsSocial** is not selected according to Lasso & penalty selected using 10-fold CV or Mallow's Cp;
- High-coefficient path corresponds to HO variable;
- Other large-coefficient paths correspond to State dummy variables

sparse

The difference between lasso and elastic net is the additional penalty just like the one used in ridge regression. By considering both penalties we have the advantages of both

lasso and ridge regression. The L1 penalty generates a sparse model that enforces some of the regression coefficients to be 0. Just like last regression, the L2 penalty removes the limitation of the number of selected variables, encourages group effect, stabilizes the L1 regularization path. There is a reference on the bottom providing extensive details on this approach.

In summary, in this lesson, I covered three methods that are commonly used in regression, where two of them are used for model selection. The third one, Ridge regression, is used for dealing with multicollinearity.

### Lecture 5.2.3 – Regularized Regression: Data Examples

In this lesson, I illustrate the implementation of the regularized regression approaches using two data examples.

#### EXAMPLE: RANKING STATES BY SAT PERFORMANCE

In the SAT example we try to identify or select explanatory variables that explain the variability in the state level average SAT score. The Ridge Regression R function is available in the MASS library; thus you have to upload this library to perform ridge regression. I created the matrix of predicting variables and applied the *scale* function.

Recall that for ridge regression the variables need to be scaled. R does not perform this step automatically; thus the variables need to be rescaled before using them in the ridge regression implementation. I also scaled the response variable; however, that is not necessary.

#### Approach 1: Ridge Regression - SAT Example

```
library(MASS)
## Scale the predicting variables and the response variable
ltakers = log(takers)
predictors = cbind(ltakers, income, years, public, expend, ranks)
predictors = scale(predictors)
sat.scaled = scale(sat)
```

Next, I specify a range of penalty constants from 0-10 with a difference of 0.25 between the consecutive values, using a SEQ command to get a total of 41 lambda values. Next, fit the ridge regression with the scaled predicting variables and the response variable and for the different values of lambda.

```

## Apply ridge regression for a range of penalty constants
lambda = seq(0, 10, by=0.25)
out = lm.ridge(sat.scaled~predictors, lambda = lambda)
round(out$GCV, 5)
which(out$GCV == min(out$GCV))
2.25
10

```

The ridge regression outputs estimates for each lambda in the considered range

The lambda is selected to minimize the (generalized) CV score.

The output of the `lm.ridge` R command consists of estimates for each lambda in the considered range. However, we'd like to extract the estimates for the lambda with the smallest cross-validation score.

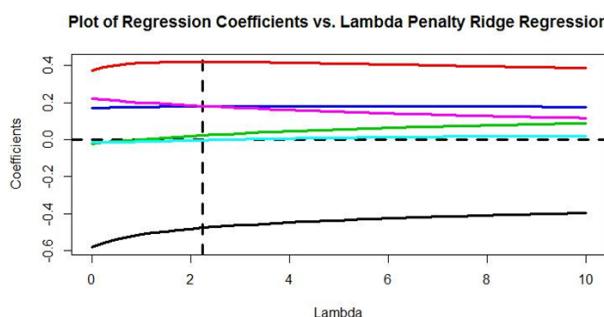
In the next two R commands, the first output is the cross-validation scores for all lambda values (`round(out$GCV, 5)`), and the second command identifies the lambda value with the smallest cross-validation score (`which(out$GCV == min(out$GCV))`). For this example, it is the tenth value in the lambda vector, and it's equal to 2.25. To extract the regression coefficient for this lambda, we specify the tenth row of the coefficient output matrix in which each row corresponds to the estimates. The estimates of the regression coefficients given one of the lambda constants. These are the estimated coefficients provided by the smallest lambda.

```
round(out$coef[,10],4)
```

predictorslakers	predictorskrank	predictorsincome	predictorsyears
-0.4771	0.4195	0.0223	0.1796
predictorpublic	predictorsexpend		
-0.0028	0.1808		

These coefficients are not comparable with the estimated coefficients from the original fitted model because we applied the regression on the scaled data.

Here's the plot of the path of the regression coefficients, along with the vertical line corresponding to the optimal lambda.



For the optimal lambda, there are four regression coefficients that are away from the 0 line and two that are close to the 0 line. Those two correspond to income and public. However, as pointed in a previous lesson, ridge regression does not force coefficients to be 0. Thus these two coefficients, although close to 0, are not forced to be 0.

### **Approach 2: Lasso Regression - SAT Example**

Here I show one implementation of Lasso regression. The next slide shows a different implementation. For this implementation, I use the `lars` command in the **lars Library**. The Lars requirement requires the input of the scaled predicting variables and the response variable. The output of Lars provides the order of how the coefficients are added to the model. The order of the variables entering the model are log(Takers), rank, years, expend, income, and public.

```
library(lars)
object = lars(x = predictors, y = sat.scaled)
object
```

Sequence of LASSO moves:

	Itakers	rank	years	expend	income	public
Var	1	2	4	6	3	5
Step	1	2	3	4	5	6

- After Lasso variable selection, apply ordinary least squares with the selected predicting variables. The selected model according to Mallow's Cp is at the fourth variable introduced in the model. (See more below)

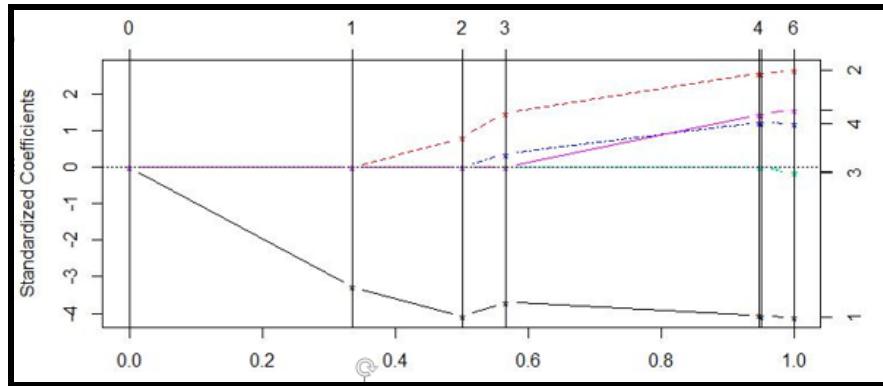
### **Approach 3: Use Mallow's Cp statistic to select best predictors. Can force the controlling variables into the model.**

Next, I provide the values of the Mallow's Cp for each of the six steps, where each step corresponding to the introduction when we introduce one variable. For example, the smallest value among the Cp values is 3.10 and it corresponds to the fourth step. Which means that the best model, according to Cp, is the one including the first four variables in order, added to the model.

```
round(object$Cp,2)
      0      1      2      3      4      5      6
349.91 103.40 46.89 35.64  3.10  5.09  7.00
plot.lars(object)
plot.lars(object, xvar="df", plottype="Cp")
```

Furthermore, we can plot the path of the regression coefficients using the `plot.lars` R command, where the input in this command is the fitted model.

The second command using the `plot.lars` function plots the values of the Cp statistic for each of the six steps and here is the plot of the estimated regression path.



From this plot, we see that we start with a path with the first predictor entering the model which is takers in this example. Then, the next path starts for the rank predictor, corresponding to the vertical line of the second step. The third predictor is years, corresponding to the vertical line at step three, and so on to conclude based on the output.

The order of selected predictors is: log takers, rank, years, expend, income and public and only the first four are selected. Once the predicting variables are selected, don't forget to apply the ordinary least squares with the selected predicting variables, instead of relying on the estimated model using Lasso. [What does this mean?]

#### **Approach 4: Lasso with cross-validation to find optimal lambda**

A second implementation in R uses the functions in the `glmnet` library. This is a modular implementation since it not only allows finding a Lasso regression, but also the more general elastic net regression. It also can be used for regression analysis and the normality, but also for other models under the framework of generalized linear models.

For fitting lasso regression, first obtain the optimal lambda value using the `cv.glmnet` which takes as input: the matrix of predicting variables X, the response Y, the specification of the alpha value that indicates the type of method, and the number of folds for the k-fold cross-validation used to determine the penalty constant lambda.

```
library(glmnet)
Xpred= cbind(ltakers, rank, income, years, public, expend)
# Find the optimal lambda using 10-fold CV
```

```

satmodel.cv=cv.glmnet(Xpred, sat, alpha=1,nfolds=10)
  ## Fit lasso model with 100 values for lambda
satmodel = glmnet(Xpred, sat, alpha = 1, nlambda = 100)
  ## Extract coefficients at optimal lambda
coef(satmodel, s=satmodel.cv$lambda.min)
(Intercept) 478.328624
Itakers      -37.572757
rank         3.587894
income        .
years        15.028032
public        .
expend       1.899913
  ## Plot coefficient paths
plot(satmodel, xvar="lambda")
abline(v=log(satmodel.cv$lambda.min),col='black',lty = 2)

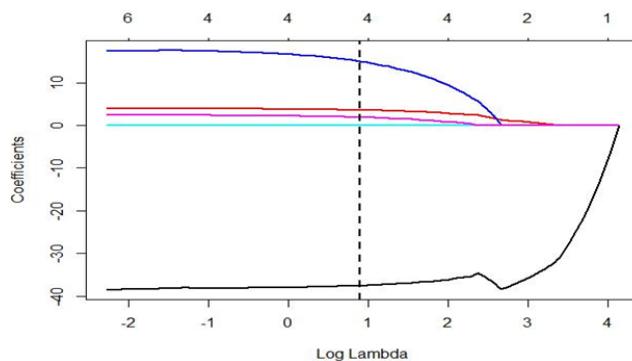
```

Next, use a `glmnet` function to fit the penalized regression for multiple lambda values.

To plot the path of the regression coefficients, we can use the `plot` function. The `abline` adds the vertical line corresponding to the optimal lambda. For this example, I'm using alpha equal to one, which correspond to the Lasso method.

If we fit a Ridge regression, we would specify alpha equal to 0. If we use elastic net, the value of alpha could be between 0 and 1.

This is the graph of the regression coefficients provided only for the four coefficients given the set of lambda selected. The vertical line corresponds to the optimal lambda and shows that we include four predictors in the model because it's close to the value four on the top. The corresponding predictors are log of takers, rank, years & expenditure.



Again, this is based on the Lasso method and penalty selected using the 10-fold cross-validation. This is the different from the Lasso provided on the previous slide because we used a Cp statistic on the previous implementation (Approach 3). We can

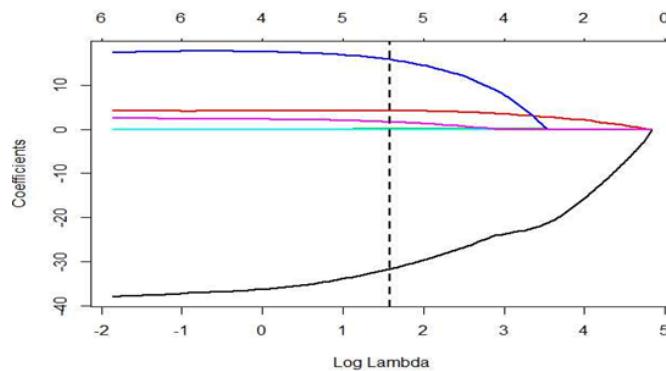
use a similar implementation as on previous slide but for elastic net and the only difference now is in the specification of alpha.

### Approach 5: Elastic Net with cross validation

Here I use alpha = 0.5, as opposed to alpha = 1 for Lasso. Alpha = 0.5 says that I give equal weight to the two penalties, the L2 and L1 penalties, or the ridge and Lasso penalties.

```
library(glmnet) ## alpha = 1 lasso, alpha=0 ridge
Xpred= cbind(ltakers, rank, income, years, public, expend)
      # Find the optimal lambda using 10-fold CV
satmodel.cv=cv.glmnet(Xpred, sat, alpha=0.5, nfolds=10)
      ## Fit lasso model with 100 values for lambda
satmodel = glmnet(Xpred, sat, alpha = 0.5, nlambda = 100)
      ## Plot coefficient paths
coef(satmodel,s=satmodel.cv$lambda.min)
ltakers      -31.62400226
rank         4.22409311
income        0.02588644
years         15.81282685
public        .
expend        1.65644751
      ## Extract coefficients at optimal lambda
plot(satmodel,xvar="lambda", lwd=2)
abline(v=log(satmodel.cv$lambda.min),col='black',lty = 2, ,lwd=2)
```

This is the graph of the regression coefficients provided only for the regression coefficients and the vertical line corresponds to the optimal lambda.



Based on this approach, the selected predictors are rank, takers, income, years, and expenditures. So we select five out of six predictors with elastic net. The penalty constant is selected using the 10-fold cross-validation.

### Comparing the above 5 approaches

Compare the set of selected predicting variables for all approaches considered across the lessons in these lectures.

	Log(Takers)	Rank	Income	Years	Public	Expend
Best subset & Mallow's Cp		x		x	x	x
Stepwise & AIC	x	x		x		x
Lasso & Mallow's Cp	x	x		x		x
Lasso & 10-fold CV	x	x		x		x
Elastic Net & 10-fold CV	x	x	x	x		x

- Rank, Years & Expend are selected by all approaches
- Takers is not selected by best subset only
- Income is not selected by any approach

**Approach 1:** The first approach we implemented was the best subset using the Mallow's Cp statistic. For this approach, we selected rank, years, public, and expenditure.

**Approach 2:** The second approach with Stepwise regression, we use the AIC, but in this approach, we specify that we force the two controlling variables in a model, takers and rank. Then we selected only two additional variables, years and expenditure.

**Approach 3:** The third approach was Lasso, using the Mallow's Cp statistic. And for this approach, although we did not force the two confounding variables to be in a model, this approach selected those controlling variable to be in a model, along with years and expenditure.

**Approach 4:** The next approach was Lasso, but with the 10-fold cross-validation approach for obtaining the optimal lambda and the set of predictors was the same as the previous approach.

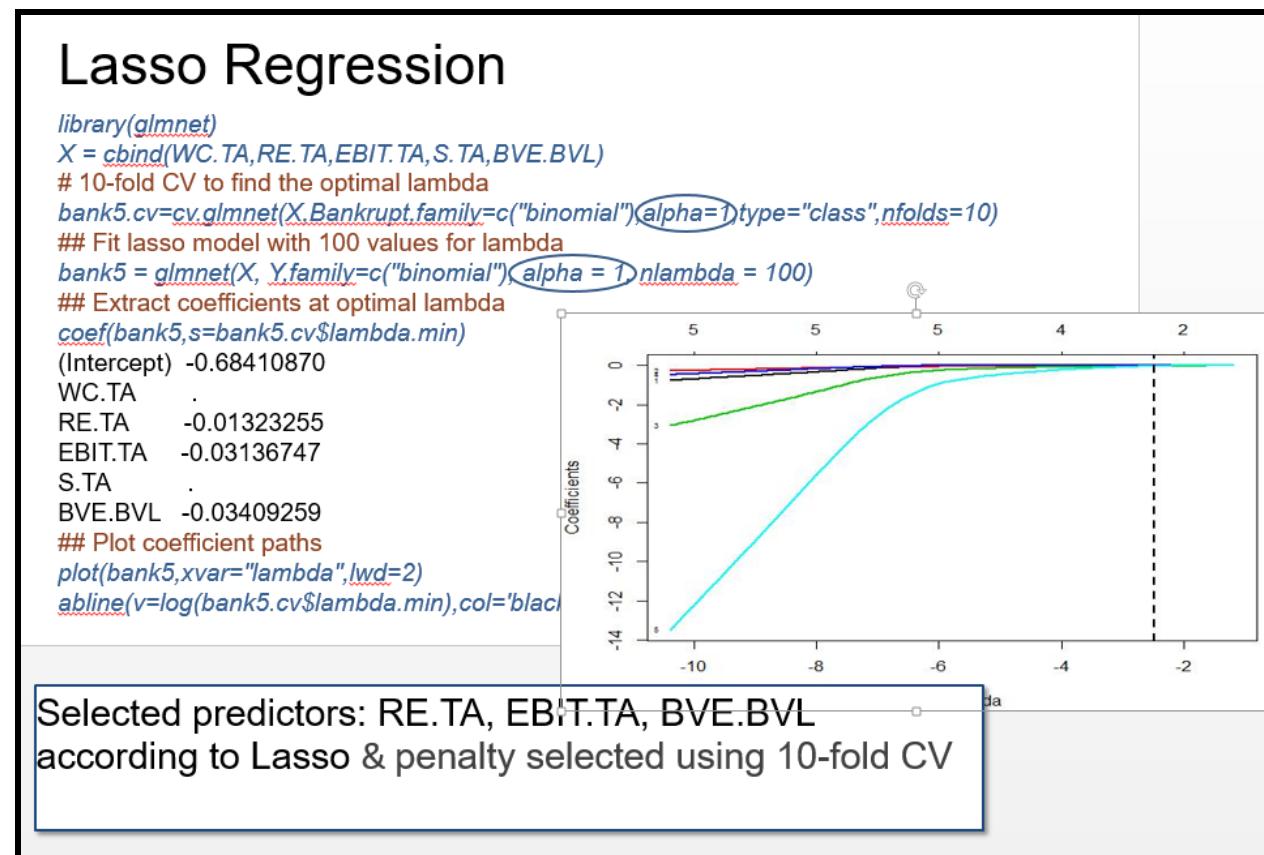
**Approach 5:** The last approach was elastic net with the 10-fold cross-validation used to obtain the optimal lambda. This approach selected an additional predictor which corresponds to income.

Overall we see that rank, years, and expenditure are selected by all approaches, except takers is not selected by best subset only. Income is selected by elastic net and public

is selected by the best subset using the Mallow's Cp. So those two factors have the least explanatory power, they explain the least variability and SAT score.

#### EXAMPLE: PREDICTING BANKRUPTCY

Let's now return to the example we're interested in explaining, whether a company went bankrupt or not. And the companies, the firms that we consider, are telecommunication firms. And we're going to begin with Lasso because the regression model is the logistical regression. We use the more general implementation provided by GLM Net. In this implementation, we need to specify family equal binomial for the logistical regression model.



Similar to the previous implementation for the SAT example we specify alpha equal to 1 to fit the Lasso regression. The plot of the path of the regression coefficients are here according to the optimal lambda we used. The model selected includes 3 predictors RE.TA, EBIT.TA, and BVE.BVL. Again, the optimal lambda is selected using the ten fold cross validation.

Elastic net is implemented similarly, except that we're using alpha is equal to 0.05. And those were gonna be the, the path of the estimated regression coefficients.

For this example we select four predictors for the optimal lambda and those selected predictors are WC.TA, RE.TA, EBIT.TA and BVE.BVL. And again, the optimal lambda is selected using the 10-fold cross validation.

Let's compare the set of selected predicting variables for all approaches considered in this lecture.

## Overview of All Selection Approaches

	WC.TA	RE.TA	EBIT.TA	S.TA	BVE.BVL
Best subset & Mallow's Cp	✗	✗	✗		✗
Stepwise & AIC		✗	✗		✗
Lasso & 10-fold CV		✗	✗		✗
Elastic Net & 10-fold CV	✗	✗	✗		✗

The first approach was the best subset using the Mallow's Cp. And for this approach, we selected working capital as a percentage of total assets. Retained earnings a percentage of total assets, earnings before interest and taxes, as a percentage of total assets and book value of equity divided by book value of total liabilities. In fact, only sales in this example is not selected by any of the approaches.

The three of the predictors, the retained earnings, the earnings before interest and taxes, and the book value of equity divided by book value of total liabilities are selected by all approaches. The working capital as a percentage of total assets in selected by best best subset and by elastic net only.

To summarize in this lesson, I provided implementation of the least [?] regression, Lasso regression, and elastic net regression. With [with] two examples, one [. One] example was using the standard linear regression and a [under] normality. The second example was a logistic regression model[.]

## LECTURE 5.3: DATA ANALYSIS EXAMPLE

### Lecture 5.3.1 – Emergency Department Healthcare Costs

In this lesson, I will introduce and apply a problem that is of interest to the health policy in the United States, specifically, the healthcare costs for the emergency department.

Emergency department healthcare has been a topic of many research studies, generally, in the United States. Healthcare provided in the emergency department is significantly more expensive than regular care in the physician office. Emergency department, or abbreviated ED in this lecture, is the place for emergency care, as well as regular care, for many people.

It is believed that many of the ED encounters can be preventable if regular healthcare is provided or/and those with chronic conditions keep up with their treatment to control their health conditions. In this study, we'll analyze a real data example related to the cost of the ED health care. Particularly, we'll identify factors that explain the variability in the cost of the ED healthcare to potentially suggest interventions that can be targeted to reduce the cost.

One particular targeting intervention is improving access to primary care, since people who have access to primary care might have less severe health outcomes, leading to less emergency department encounters.

Thus, the research questions to be addressed in this study are as follows:

What factors impact the cost of healthcare due to emergency department encounters?

Is access to primary care providers associated to healthcare costs due to emergency department encounters?

If access to primary care improves, can we predict a reduction in the cost of the ED healthcare?

## Emergency Department Healthcare Costs

Figure 1. Percentage who had selected reasons for last emergency room visit among adults aged 18-64 whose last visit in past 12 months did not result in hospital admission: United States, January-June 2011

Reason	Percentage
Services of medical provider	66.0
Only hospital could help	54.5
Problem too serious for doctor's office	42.5
Health provider said to go	20.1
Arrived by ambulance	8.9
Lack of access to other healthcare	79.7
Doctor's office not open	48.0
No other place to go	48.3
Emergency room is closest provider	45.8
Most care is at emergency room	17.7

**Research Question 1:**  
What factors impact the healthcare cost due to emergency department encounters?

**Research Question 2:**  
Is access to primary care providers associated to healthcare costs due to emergency department encounters?

## Emergency Department Healthcare Costs

**Study population:** Medicaid-enrolled adults in four southeast states: Alabama, Arkansas, Louisiana, and North Carolina in 2011

- Medicaid is a health insurance program for the low-income population

**Data Source:** The Medicaid Analytic eXtract (MAX) claims files available from the Centers of Medicare and Medicaid Services (CMS)

- Disclaimer: The research on healthcare cost for the Medicaid population using the MAX claims data has been approved by the Georgia Tech Internal Review Board and by CMS; Do NOT use the data provided for this analysis for other purposes beyond the study in this lecture.
- Additional data sources: US Bureau Census, Health Analytics Group at GT, Robert Wood Johnson Foundation among others.

In this study, we'll focus on the cost of the emergency department healthcare for the population of adults enrolled in the Medicaid health insurance program in four states in the United States, including Alabama, Arkansas, Louisiana, and North Carolina.

Medicaid is a health insurance program for the low-income population, covering more than 50 million adults nationwide. The data source is the Medicaid Analytic eXtract, or in short, MAX, claims data. The MAX data are collected by the Centers for Medicare and Medicaid Services, in short, CMS. Each state reports information on Medicaid enrollment and medical records to CMS through the Medicaid Statistical Information System. This data consists of all medical records for all enrollees, all adults and children, in fact, enrolled in Medicaid across multiple years.

We're going to focus on one year of data, we're going to focus on the 2011 MAX data. The purpose of the MAX data that is disseminated through CMS is to provide data to support research on the Medicaid population, such as policy, on actuarial analysis.

One example is this study on ED healthcare cost. This kind of analysis on healthcare cost complies with the study protocol approved by CMS and the Institute Review Board at Georgia Tech. Please do not use this data for other purposes than this study. Since the study's based on the Medicaid Analytic eXtract data available from Georgia Tech, you need to have prior CMS and IRB approval from the Review Board at Georgia Tech.

The aggregated cost data, along with healthcare utilization characteristics of the Medicaid population, were derived from this, from the MAX data. The variables in the study were acquired from publicly available data sources. For example, census demographics and socioeconomic factors were acquired from the American Community Survey conducted by the US Census Bureau. Healthcare access measures were provided by the research group, the Health Analytics Group at Georgia Tech. County level covariants are also taken from the county health rankings published by the University of Wisconsin Population Health Institute and the Robert Woods Johnson Foundation.

You can see that for this data study, we used data from multiple sources and with different levels of granularity, requiring extensive data processing, and also data knowledge. While I provided the data needed for this study, extensive efforts went toward the data acquisition and data processing. Data acquisition and data processing are commonly part of any data analysis. Many of the courses in this MS analytics program will provide you with the skills for data processing. But the data acquisition relies primarily on the knowledge of the applied problem.

The primary variable of interest in the study is the aggregated cost at the census tract level for emergency department, in contrast for the Medicaid enrollees in the year 2011. And this is defined as ED cost in the study and this analysis. Census tracts, we

estimate, we aggregate the cost at the census tract. The reason is that census tracts are proxies of communities in the United States. It's a contiguous division of all states in the United States.

## Response & Predicting Variables

### **Response variable:**

- Emergency Department cost aggregated at the census tract level (*EDcost*)
- Number of member months aggregated at the census tract level (*PMPM*)

### **Predicting variables are:**

**Location:** state (*State*) and census tract identification (*GEOID*)

**Utilization:** three predicting variables measuring the number of claims for the Emergency Department (*ED*), of hospitalizations (*HO*) and physician office (*PO*)

**Population characteristics:** percentages of Medicaid-enrolled adults who are black (*BlackPop*), white (*WhitePop*) or other race/ethnicity (*OtherPop*); percentages of Medicaid-enrolled adults who are health (*HealthyPop*), with chronic conditions (*ChronicPop*) or with complex health problems (*ComplexPop*)

**Socio-Economic and Health Environment Factors:** 13 predicting variables including unemployment, median income, urbanicity of the census tract, access to primary care, and health rankings among others.

It's also important to note that the aggregated cost will depend on the number of Medicaid-enrolled adults who are using the healthcare system, and the length of their enrollment. For example, some adults may be enrolled in Medicaid for two months, whereas others will be enrolled for the whole year. Or, some states will have a lot of Medicaid enrollees, where others will have less.

In order to account for this, the total number of enrollment months in the year 2011 is provided for each census tract. The variable name in this study is *PMPM*, which stands for per member per month. This variable should be used to rescale the cost for comparison across census tracts with different levels of enrollment and different number of adults enrolled in the Medicaid program.

We differentiate a set of factors to be included in a model in this study as follows. We have location information, referring to the census tract ID, or *GEOID*. And the state, which is differentiated into Alabama, abbreviated *AL*, Arkansas, abbreviated *AR*, Louisiana, abbreviated *LA*, and North Carolina, abbreviated *NC*. The variable name is *state*. We'll only use the *state* variable in this study.

Healthcare utilization factors are also provided, and they include the number of claims for ED encounters, the number of claims for the physician office visits. And the number of claims for hospitalizations, or so-called inpatient care. And we provide the number of

claims for each of those three types of healthcare services separately. These three variables can be viewed as proxies for utilization of each service type.

In order to compare the level of utilization across census tracts with different Medicaid population and with different enrollment, we'll also need to scale these variables by the PMPM scaling factor, just like the ED cost.

We also have study population characteristics, including race, the race variable, such as the percentage of the Medicaid adult population who are black, white, and other. I'm also providing the percentage of adults on Medicaid who are classified as healthy with chronic conditions, or with complex conditions, such as cancer and other life-threatening conditions.

In addition, we have a set of predicting variables referring to the socioeconomic and health environment factors, a total of 13 total variables in this group. And including unemployment rate, median household income, percentage of families below the poverty level. Percentage of the population that have a bachelor's degree or higher, urbanicity factor, differentiating census tracts into urban, suburban, and rural. Access to primary care, measured as the mean travel distance, called accessibility in this study. And access to primary care, measured as the mean waiting time for appointment, called availability in this study.

And also, have a set of county health ranking factors, like primary care physician rate, food environment index, housing problems, exercise access, social environment. And also, last, we have a factor that measures the provider density, in a sense, measures the density of the healthcare infrastructure.

It is important to first establish whether there are factors that could lead to selection bias across the population within the census tracts in relation to the cost of ED healthcare.

# Controlling Variables

## **Selection Bias:**

- The utilization of healthcare emergency services is directly driven by the health status of the population utilizing the system. Adults with multiple chronic conditions and/or with complex health problems tend to need emergency healthcare services more than the healthy population.
- Controlling factors: Percentage of population with chronic conditions (ChronicPop) or with complex health problems (ComplexPop)

## **Confounding Variable:**

- The number of ED claims is a confounding variable not an explanatory factor for ED cost because it is a measure of utilization of the emergency department which leads to ED healthcare cost; it correlates with both the response and the predicting variables.
- Such confounding variables should not be included in the model.

Specifically, one bias selection is due to the fact that the utilization of healthcare emergency services is directly driven by the health status of the population utilizing the system.

Adults with multiple chronic conditions and/or with complex health problems tend to need emergency healthcare services more than the healthy population. For example, if a community has a large number of adults on Medicaid with complex conditions, this could lead to higher utilization and higher ED cost, as compared to a community with a healthy population. Thus, the controlling factors in the study are the percentage of population with chronic conditions and percentage of population with complex health problems.

Moreover, the utilization of ED, measured by the number of ED claims, is a confounding variable in this study, and not an explanatory factor. Because it's a measure of utilization of the emergency department, which leads to ED healthcare cost. It correlates it with both the response and the predicting variables. Such confounding variables should not be included in the model. So for this analysis, we're not gonna include, in the model, the number of ED claims as a predicting factor.

To summarize, in this lesson, I introduced a applied problem related to the emergency department healthcare cost. In the next lessons, we will analyze data for this study[.]

## Lecture 5.3.2 – Exploratory Data Analysis

In any statistical analysis, it's important to first get a feel for the data through the exploration of the data. In this lesson, I will perform an exploratory data analysis for the study of the healthcare cost for the emergency department.

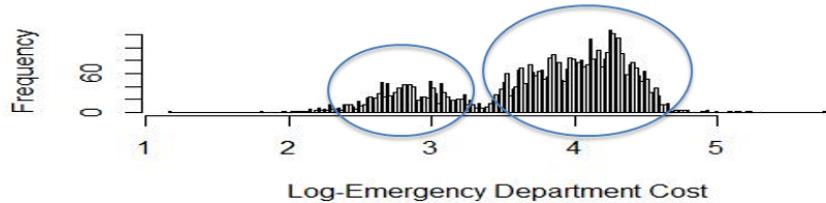
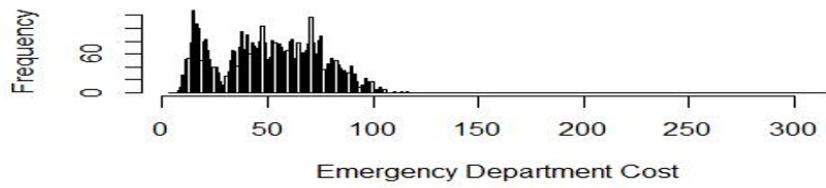
The data file is DataADULT.csv, read in R using the read.csv R command. The EDCost and the utilization of the physician office and hospitalization are scaled by the scaling factor per member per month. I will refer to the scaled measure as ED costs per member, per month, and utilization per member, per month.

### Exploratory Data Analysis: Response Variable

```
## Read the data using the 'read.csv()' R command
dataAdult=read.csv("DataADULT.csv", header=TRUE)
attach(dataAdult)
## Outcome/Response Variable
EDCost.pppm = EDCost/PMPM
## Rescale utilization
dataAdult$PO = PO/PMPM
dataAdult$HO = HO/PMPM
#Histogram of the response variable
par(mfrow=c(2,1))
hist(EDCost.pppm,breaks=300, xlab="Emergency Department Cost", main="")
hist(log(EDCost.pppm),breaks=300, xlab="Log-Emergency Department Cost", main="")
log.EDCost.pppm = log(EDCost.pppm)
```

To explore the outcome variable, the response variable, which is the ED cost per member, per month, I'm using the histogram. I'm comparing the histogram of the response variable and its transformation using the log function. The histogram of the data is provided first.

## Response Variable



The shape of the distribution is greatly skewed with the data concentrated into a rather small range. Commonly when we see such a shape, it is an indication that the normality assumption might not hold.

Generally, you should perform a regression analysis with the untransformed response variable before considering a transformation. For this example, I did so, the residuals based on the regression analysis with untransformed response were skewed as expected.

Hence, I'm suggesting here a transformation of the response variable. It seems that the log transformation does well in centering the data while spreading the data over a wider range. But what we see here is that, we have two clear modes in the distribution of the response variable after transforming the data. It is our hope that this bimodality will be explained by the predicting variables considered in this study.

In order to assess the relationship between the response variable and qualitative predicting variables, we can use the side by side boxplots, as in the analysis of variance (ANOVA), where we compare the within to the in-between variability.

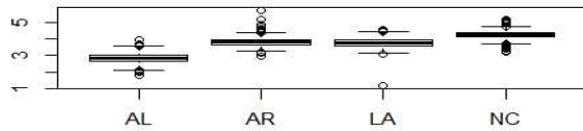
# Exploratory Data Analysis: Response vs Qualitative Predictors

```
## Response variable vs categorical predicating variables  
par(mfrow=c(2,1))  
boxplot(log.EDCost.pppm ~ State, main = "Variation of log of ED costs by state")  
boxplot(log.EDCost.pppm ~ Urbanicity, main = "Variation of log of ED costs by urbanicity")
```

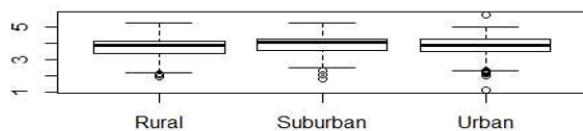
In this example, **we have two qualitative predicting variables, state and urbanicity level of the census tracts.** The box plots are here, when comparing the ED cost per member per month across state we see clear differences in the medians, with North Carolina having the higher median and Alabama having the lowest median.

# Exploratory Data Analysis: Response vs Qualitative Predictors

Variation of log of ED costs by state



Variation of log of ED costs by urbanicity



For urbanicities the differences among the medians across the three different urbanicity levels (rural, suburban, and urban) are not strikingly different. However, if we perform an ANOVA analysis, we reject the null hypothesis of equal means.

To explore the relationship between the response and quantitative predictive variables, it is common practice to consider the matrix plot, which is a matrix of scatter plots between, all/any pair of variables. Since we have a total of 25 variables, we'll actually consider the study on 23. It will not be visually pleasant to look at a 25 by 25 matrix of scatter plots. For this analysis, we can instead group the quantitative variables **ensue** [to ensure?] meaningful clusters of predicting variables.

# Exploratory Data Analysis: Response vs Qualitative Predictors

```
## Scatterplot matrix plots
library(car)
## Response vs Utilization
scatterplotMatrix(~log(EDCost.pmpm)+HO+PO,smooth=FALSE)
## Response vs Population Characteristics
scatterplotMatrix(~log(EDCost.pmpm)+WhitePop+BlackPop+OtherPop+HealthyPop+
ChronicPop+ComplexPop,smooth=FALSE)
## Response vs Social and Economic Environment Characteristics
scatterplotMatrix(~log(EDCost.pmpm)+Unemployment+Income+Poverty+Education+
Accessibility+Availability+ProvDensity,smooth=FALSE)
## Response vs County Health Rankings
scatterplotMatrix(~log(EDCost.pmpm)+RankingsPCP+RankingsFood+RankingsHousing+
RankingsExercise+RankingsSocial,smooth=FALSE)
```

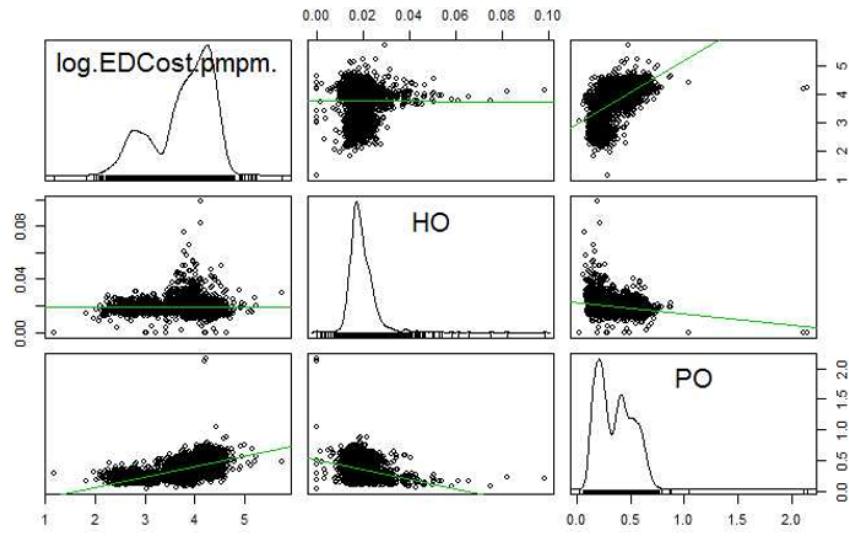
We also use a new R command available in the *car* library. **This command is `scatterplotMatrix`** where the input is the set of variables to be considered in a scatterplot matrix. Do not forget to install the **package car** before using this library.

For example, if we will consider utilization variables including utilization of physician office and hospitalization and the ED cost. The command line is **scatterplotMatrix ~ log of EDCost.pmpm plus hospitalization (HO) and physician office (PO) utilization variables**. We apply the same command for the variables characterizing the population characteristics (WhitePop, BlackPop, etc.). And for the predicting variables referring to social economic characteristics (Unemployment, Income, etc.) and last, all the predictive variables referring to the county health rankings.

The first matrix plot is for utilization.

# Response vs Quantitative Predictors

**ED Cost vs Utilization Measures:** *number of claims for HO and PO*



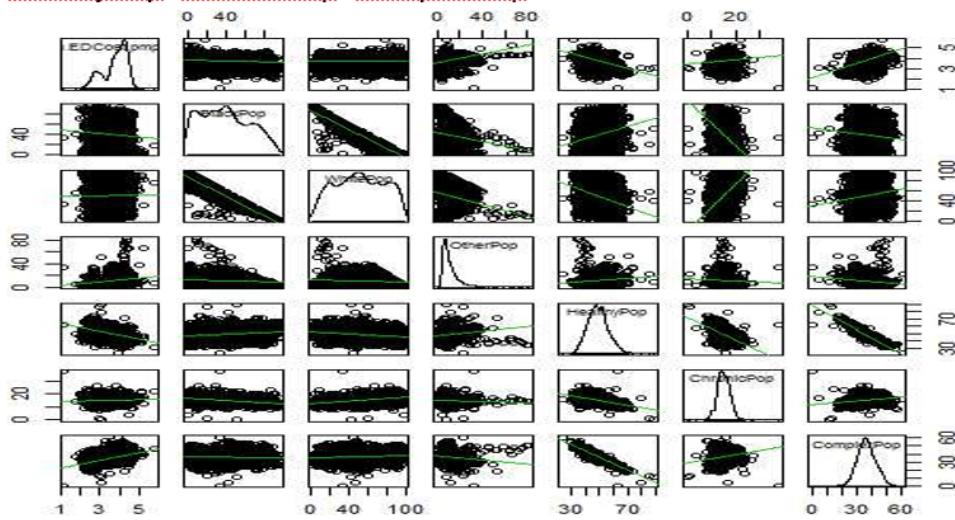
On the diagonal, we can see an estimate density function for the distribution of each variable. For example, the first one is for the log of the ED cost, and then the density shows a bimodal distribution, as we actually identify using the histogram of the response of the log of ED cost.

The other plots are the scatter plots for any pair of the variables. From this plot, we infer that there is a strong positive correlation between the ED cost and the utilization of physician office, but not a strong relationship with (hospital) in-patient utilization. Moreover, there is only a weak correlation between the utilization of physician office and in-patient utilization, thus not a reason for concern with respect to co-linearity.

The next set of plots are for population characteristics.

# Response vs Quantitative Predictors

**ED Cost vs Population Characteristics:** BlackPop, WhitePop, OtherPop, HealthyPop, ChronicPop, ComplexPop



We have several population characteristics to consider here, there are six in total. The percentage of black population, percentage of white population, percentage of other population, the percentage of adults that are healthy, the percentage of adults that have chronic conditions and the percentage of adults that have complex conditions a  $7 \times 7$  matrix plot.

Similarly, on the diagonal, we can see an estimated density function for the distribution of each variable. The names of the variables are not clearly seen from this plot, but order is as provided on [top] above the plot.

From this set of plots, there is a weak relationship of the response with respect to a black and white population percentages, but a stronger positive relationship with a percentage of Other population. There is also a negative relationship with the percentage of adults that are considered healthy, a weak positive relationship with the person of adults with chronic conditions, and last, a strong positive relationship with a percentage of population with complex conditions.

I'll note here that the percentage of OtherPopulation is linearly dependent with the other two race variables- the percentage of black, and the percentage of white populations. Moreover, the percentage of the healthy population is also linearly dependent with the percentage of population that have chronic conditions and the percentage of population that have complex conditions. This linear dependence is also

apparent in the scatterplot as we see a strong relationship between the three race variables and a strong relationship among the three health condition variables.

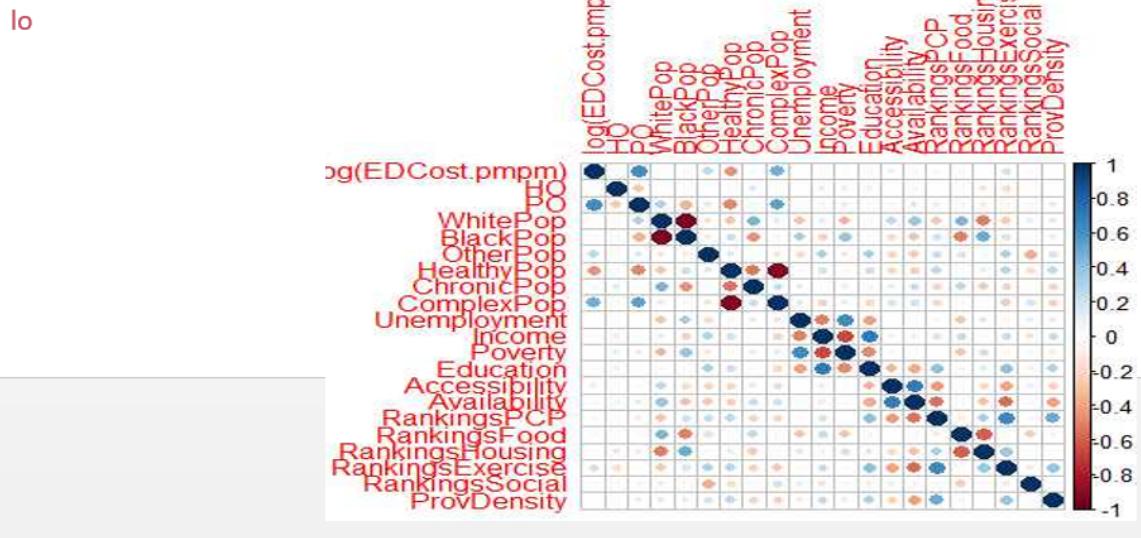
I will note here that when we have such a linear dependence, one of the variables among the race variables will need to be discarded and one of the variables referring to the condition of the population will need to be discarded as well. Last, I will note that there's a strong negative relationship between percentage of black population and percentage of white population, an indication of segregation with respect to these two races in these four states.

I will now be providing the scatter-plot for the other two sets of variables. You can practice the analysis for these two other groups by yourself, since studying all such scatter-plot matrices is cumbersome, especially for a large number of predictors. And because they do not capture the relationship among variables across different groups of variables.

Alternatively, we can consider the Correlation Matrix Plot. It can use the command `corrplot` from the library `corrplot` where the input in this command is simply the correlation matrix of all variables.

# Response vs. Predicting Variables: Correlation Matrix Plot

```
## Correlation matrix plot  
library(corrplot)  
corr = cor(cbind(EDCost.pppm,dataAdult[,-c(1,2,3,18)]))  
corrplot(corr)
```



The correlation plot is here, first you need to pay attention to the scaling. Dark blue correspond to high positive correlation, and dark red corresponds to high negative correlation. Light shades of colors correspond to low correlation. The size of the circle or the dot within each cell is also an indication of the magnitude of the correlation. Its cell in the matrix, is a visual representation of the correlation between a pair of variables where the names of the variables are on the margins of the plot.

For example, we see a large positive correlation between log of ED cost and utilization of physician office and a large positive correlation between the log of ED cost and the percentage of adults with complex conditions. We also see a large negative correlation between the log of ED cost and the percentage of population that are considered healthy.

Among the predicting variables, we see a strong negative correlation between the percentage of black and percentage of white populations, and between percentage of adults that are considered healthy and percentage of adults with complex conditions.

Among the variables from different groups, there is a high correlation between county rankings with respect to food and housing, and variables representing percentage of black and percentage of white population. Density is also correlated with the two measures of access.

To summarize, in this lesson, I provided an exploratory analysis on the variables included in this study of the healthcare costs for the emergency department.

### Lecture 5.3.3 – Multiple Regression: Fitted Model and Residual Analysis

In this lesson, I'll perform the multiple regression analysis along with the residual analysis for the data problem related to the emergency department health care costs.

To fit the regression model, we will need to discard from the set of predicting variables the GEOID, which is the ID of the tracts, the PMPM scaling factor, the ED utilization, which is a confounding factor, and the percentage of Other population due to the linear dependence with the black and white population percentages. And also discarding **healthy** [correction: complex] population percentages due to the linear dependence with the percentage of adults with chronic and with **complex** [correction: healthy] conditions. After excluding these variables, we get the reduced data set.

## Multiple Linear Regression Model

```
## Exclude GEOID, scaling factor (PMPM); confounding factor (ED)
## Exclude OtherPop & ComplexPop because of linear dependence
dataAdult.red = dataAdult[,-c(1,3,4,5,10,13)]
fullmodel = lm(log(EDCost.pppm)~ ., data = dataAdult.red)
summary(fullmodel)
```

Now, in the LM model fit, we can simply use all predicting variables in the data matrix by using this [INAUDIBLE] [??] implementation. The summary of the model is here.

# Multiple Linear Regression Model

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.208e+00	1.175e-01	18.788	< 2e-16 ***
StateAR	9.235e-01	1.610e-02	57.353	< 2e-16 *
StateLA	9.081e-01	1.358e-02	66.853	< 2e-16 *
StateNC	1.418e+00	1.650e-02	85.909	< 2e-16 *
HO	1.168e+01	7.587e-01	15.401	< 2e-16 *
PO	1.378e-01	4.114e-02	3.350	0.000815 *
WhitePop	4.416e-03	5.800e-04	7.614	3.16e-14 *
BlackPop	4.894e-03	5.824e-04	8.403	< 2e-16 *
HealthyPop	-9.044e-04	8.160e-04	-1.108	0.267751
ChronicPop	-5.949e-03	2.052e-03	-2.899	0.003760 *
Unemployment	4.390e-04	7.377e-04	0.595	0.551797
Income	-2.556e-07	2.774e-07	-0.922	0.356769
Poverty	-3.306e-04	4.460e-04	-0.741	0.458529
Education	-1.447e-03	3.296e-04	-4.390	1.16e-05 *
UrbanicitySuburban	-4.565e-04	1.369e-02	-0.033	0.973406
UrbanicityUrban	2.067e-02	1.269e-02	1.629	0.103356
Accessibility	-1.965e-03	7.094e-04	-2.770	0.005623 *
Availability	8.037e-02	1.975e-02	4.068	4.81e-05 *
RankingsPCP	7.596e-04	1.819e-04	4.175	3.03e-05 *
RankingsFood	6.586e-03	5.203e-03	1.266	0.205642
RankingsHousing	-4.642e-03	1.562e-03	-2.973	0.002967 *
RankingsExercise	3.993e-04	2.332e-04	1.712	0.086907
RankingsSocial	-3.895e-04	1.347e-03	-0.289	0.772497
ProvDensity	6.042e-02	1.573e-02	3.841	0.000124 *
---				
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'

Residual standard error: 0.2321 on 4995 degrees of freedom  
 Multiple R-squared: 0.8486, Adjusted R-squared: 0.8479  
 F-statistic: 1218 on 23 and 4995 DF, p-value: < 2.2e-16

**Socio-economic** predicting variables (unemployment, median income, % below the poverty level and rankings with respect social environment) are **not** statistically significant given other predicting variables in the model.

**Access** to primary care (accessibility and availability) is statistically significantly associated to ED cost.

**85%** of the variability in the ED cost is explained.

For this model fit, we find that socioeconomic predicting variables including unemployment, median income, percentage of population below the poverty level, and rankings with respect to social environment, are not statistically significant given other predicting variables in the model.

The variables of interest in the second research policy question presented in the first lesson, are the two access measures. Access to primary care including the measure of accessibility and measure of availability, is statistically significantly associated to ED cost, given other predicting variables in the model. Last, we can see that the set of predicting variables considered in the study explained approximately 84% of the variability in the log of ED cost or rather large R-squared.

Next, we'll perform a residual analysis toward validating the model assumptions. First, two lines of R code on this slide extract the residuals from the model output and compute the Cook's distances.

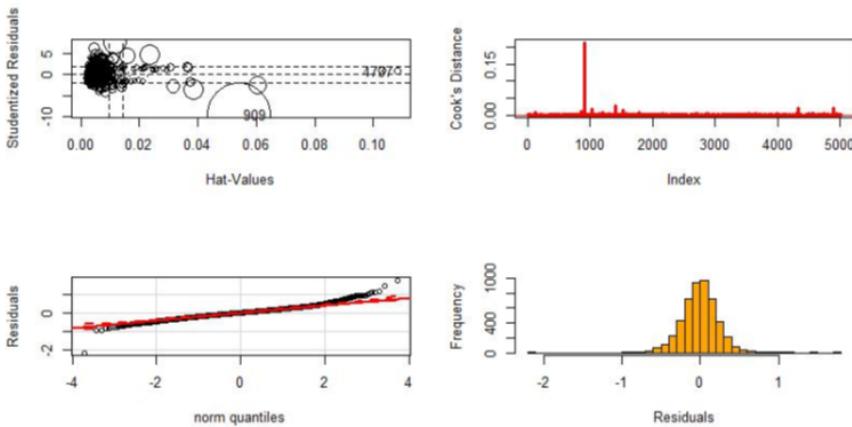
# Residual Analysis: Outliers & Normality

```
## Residuals versus individual predicting variables
full.resid = residuals(fullmodel)
cook = cooks.distance(fullmodel)
par(mfrow=c(2,2))
## Check outliers
influencePlot(fullmodel)
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's Distance")
## Check Normality
abline(0,0,col="red")
qqPlot(full.resid, ylab="Residuals", main = "")
hist(full.resid, xlab="Residuals", main = "",nclass=30,col="orange")
```

Next, I'm using a new R command **influencePlot** with the input being the fitted model. This command along with the plot of the Cook's distances can be used to identify outliers, to identify influential points.

The last two R commands on this slide are for evaluating the normality assumption using the normal probability plot and the histogram.

# Residual Analysis: Outliers & Normality



**Outliers:** Observation 909 stands out

**Normality:** Symmetric but heavy tails

Here are the four plots. The first plot is the influence plot which creates a bubble plot of standardized residuals by Hat values with areas of the circles representing the

observations proportional to Cook's distances. Both these plots and the Cook's distances plot point to one clear outlier, the observation 909. We'll investigate the influence of this observation on the model fit in the following lesson.

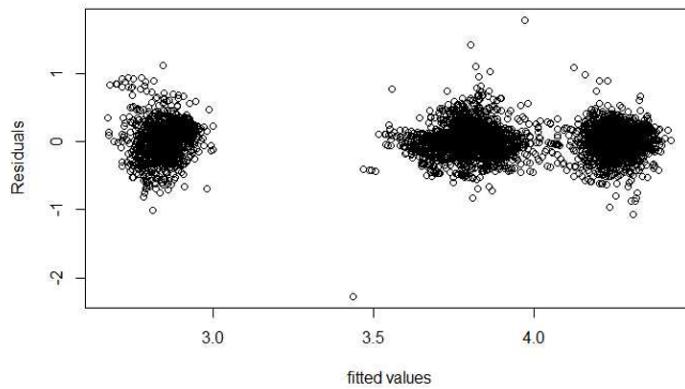
The bottom plots show the shape of the distribution of the residuals and that distribution is symmetric, but with heavy tails, possibly, more of a T distribution than a normal distribution. We cannot do much about these heavy tails. Alternatively, one could use the robust regression instead of a least squares regression, but this is a more advanced modeling techniques.

We'll next evaluate the assumptions of constant variance and uncorrelated errors by plotting the fitted values versus residuals using the classic scatter plot.

## Residual Analysis: Constant Variance & Uncorrelated Errors

```
## Check Constant Variance & Uncorrelated Errors  
full.fitted = fitted(fullmodel)  
par(mfrow=c(1,1))  
plot(full.fitted,full.resid, xlab="fitted values", ylab="Residuals")
```

## Residual Analysis: Constant Variance & Uncorrelated Errors



**Constant Variance:**  
No pattern

**Uncorrelated Errors:**  
Three well defined clusters

In terms of constant variance, we do not see a change in variables across the residuals. However, we do see three clear clusters of the residuals pointing out to a departure from the independence assumption or more precisely, the uncorrelated errors

assumption. The correlation might be driven by a factor that we do not include in our model.

I'll also point out that our data are geographically correlated. The cost of ED care is measured at the community level. We should expect ED cost to be more similar for communities in near proximity than for communities far away. This is called the first law of geography and leads to the so-called spatial dependence in the data.

To rigorously model spatial dependence, we will need models that account for this type of dependence in the error terms or/ and allow for the regression coefficients to vary smoothly over space or geography. This is a topic of a field called, in statistics, spatial statistics. Again, spatial statistical modeling is a more advanced modeling technique.

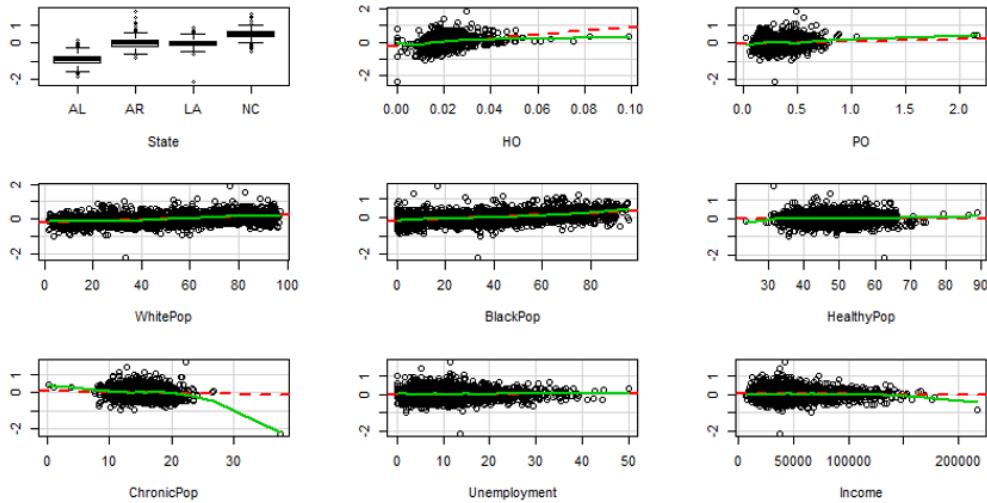
Last, we'll evaluate the assumption of linearity. The common approach is to plot the residuals versus each predicting variable. But because we have 23 variables, it would require 23 lines of code to do 1 by 1. Alternatively, you can use the crPlots R command with the input, the model fit. This allows to do all plots with only one line of code.

## Residual Analysis: Linearity

```
## Check Linearity  
crPlots(fullmodel,ylab="")
```

Here are the plots for the first nine predicting variables.

# Residual Analysis: Linearity

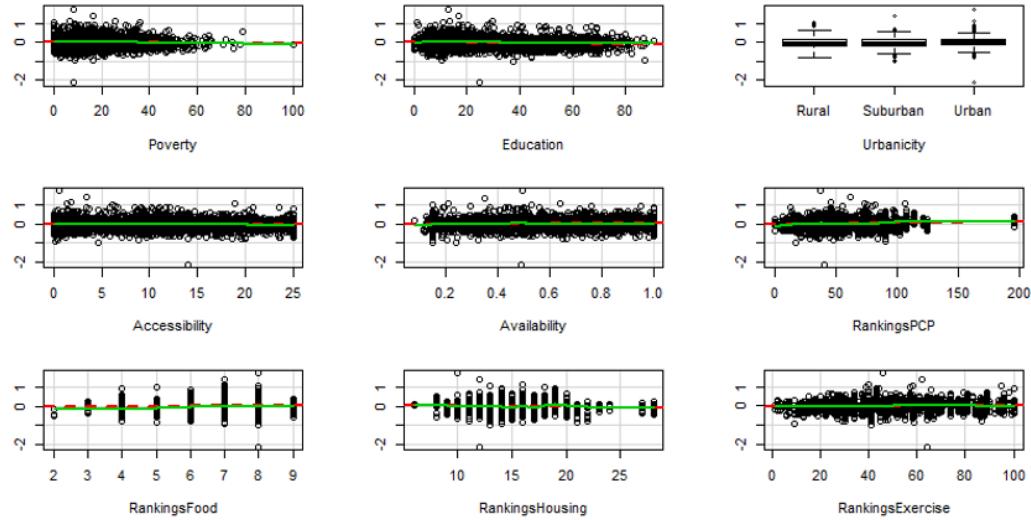


Note that for qualitative variables, we have side by side boxplots and for quantitative variables, we have scatter plots. The first is the side by side boxplots for the residuals versus state. While, we see some variation across the states, the variation in the medians is small.

The next plot is the scatter plot of the in-patient utilization per member per month versus residuals. The slight increase in the trend is due to the fact that there are a few large values in the inpatient utilization that pull the trend upwards. We note the same pattern for the next plot for the relationship between the utilization of the physician's office and log of ED costs. In fact, for all scatter plots on this slide, where we see that there is an upward trend with some outliers, we see that this upward trend is due to the outliers except we see a trend in the relationship between the residuals and the percentage of black population, represented white population. And this trend is slowly going upward.

The next set of slides are for linearity assumptions with respect to nine more predicting variables. The rest are not included.

# Residual Analysis: Linearity (cont'd)



For all those nine predicting variables, there is not a significant trend left in the residuals with respect to the predicting variables, indicating that the linearity assumption holds with respect to these predicting variables.

In this lesson, I provided the regression analysis and the residual analysis for the exploration of the data for the ED healthcare cost.

## Lecture 5.3.4 – Variable Selection

In this lesson, I'll continue with variable selection. And I will address the following questions: Which of the variables provide most of the explanatory power for the log of ED cost? Or which of the variables in this study are discarded and why?

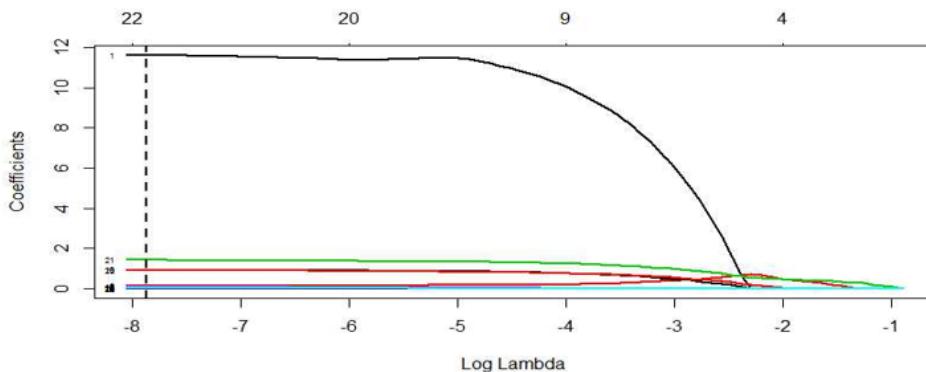
Note that, for this study we have 23 predicting variables, which is equivalent to about 8 million combinations of predicting variables, more than 8 million models. Thus, we cannot estimate all 8 million models and compare them to choose the best model selection.

Instead, we're gonna apply regularized regression and stepwise regression to perform variable selection. I'll begin with variable selection via regularized regression using the

functions in the `glmnet` library. This is a more general implementation, since it not only allows fitting a lasso regression, but also the more general elastic net regression.

## Lasso Regression

```
# 10-fold CV to find the optimal lambda  
lassomodel.cv=cv.glmnet(predictors.log(EDCost.pppm),alpha=1)nfolds=10)  
## Fit lasso model with 100 values for lambda  
lassomodel = glmnet(predictors.log(EDCost.pppm), alpha = 1) nlambda = 100  
## Plot coefficient paths  
plot(lassomodel,xvar="lambda",label=TRUE,lwd=2)  
abline(v=log(lassomodel.cv$lambda.min),col='black',lty = 2,lwd=2)
```



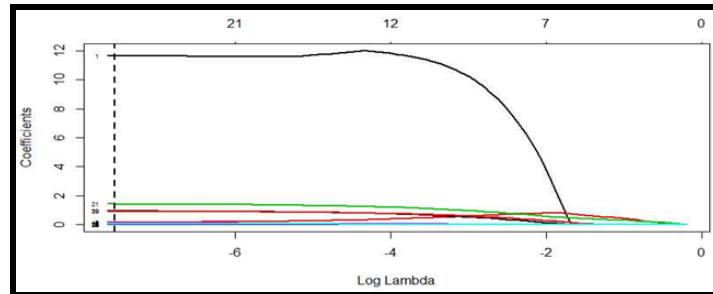
- RankingsSocial and Suburban dummy variables are not selected according to Lasso & penalty selected using 10-fold CV or Mallow's Cp;
- High-coefficient path corresponds to HO variable;
- Other large-coefficient paths correspond to State dummy variables

Alternatively, you could use the `lars` library with the `lars` command, which only allows fitting a lasso regression. For fitting the lasso regression, we first obtain the optimal lambda using the `cv.glmnet`, **would take** [which takes] as input the matrix of the predicting variables, the response, along with the specification of the alpha, which indicates the type of method, of the regularization method, and the number [of] folds for the k-fold cross validation used to determine the penalty constant lambda.

Next, we can use a `glmnet` function to fit the penalized regression for multiple lambda values. To plot the path of the regression coefficients, we can use the `plot` function. And we can add the vertical line corresponding to the optimal lambda, using the `abline`

command. For this example, I'm using alpha = 1, which corresponds to lasso. If we were to fit ridge regression, we would specify alpha equal to 0.

This is the path of the regression coefficients.



The vertical line corresponding to the optimal lambda points to a selection of 21 out of 23 predicting variables. RankingSocial variable and the Suburban dummy variables are the two variables not selected by lasso.

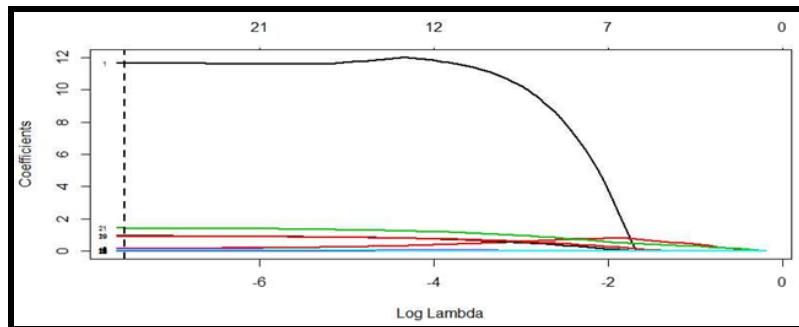
The same selection will be provided by the lasso method using the Cp statistics for the optimal value, optimal lambda, via the lars R command implementation as we learned in the previous lecture on variable selection. Among the coefficient paths that stand out, the black line corresponds to the first variable in the matrix of the predicting variables which is actually the utilization of hospitalization or in-patient utilization.

We can use a similar implementation, as used in the previous slide, but now for more general variable selection approach, the elastic net. The only difference is the specification of alpha. Here I'm using alpha equal to 0.5 as opposed to alpha equal to one for lasso. The alpha is 0.5 says that I'm giving equal weights to the two penalties, the **I2** [L2] or the ridge regression penalty and the **I1** [L1] which is the lasso penalty.

## Elastic Net Regression

```
# 10-fold CV to find the optimal lambda
enetmodel.cv=cv.glmnet(predictors,log(EDCost.pppm),alpha=0.5,nfolds=10)
## Fit lasso model with 100 values for lambda
enetmodel = glmnet(predictors,log(EDCost.pppm), alpha = 0.5, nlambda = 100)
## Plot coefficient paths
plot(enetmodel,xvar="lambda",label=TRUE, lwd=2)
abline(v=log(enetmodel.cv$lambda.min),col='black',lty = 2,lwd=2)
## Extract coefficients at optimal lambda
coef(enetmodel,s=enetmodel.cv$lambda.min)
```

This is the path of the regression coefficients.



The vertical line corresponds to the optimal lambda using elastic net. We're using elastic net, we select 22 variables out of 23. And the variable that is not selected is the county rankings by social environment. Using ridge regression, we did not reduce the model significantly, only with one or two variables.

Let's next perform variable selection using stepwise regression. To perform stepwise regression, we can use the `step` command in R. This function allows specification of a minimum model in this case, the minimum model includes only the two controlling variables. To do it so, the first input is the model with the controlling variables, [.] **thus**, [Thus,] the scope option **allows to** [lets us] specify a starting model and the full model.

## Stepwise Regression

```
full = lm(log(EDCost.pppm) ~ HealthyPop + ChronicPop + State + Urbanicity + HO + PO +  
BlackPop + WhitePop + Unemployment + Income + Poverty + Education +  
Accessibility + Availability + ProvDensity +  
RankingsPCP + RankingsFood + RankingsExercise + RankingsSocial)  
minimum = lm(log(EDCost.pppm) ~ HealthyPop + ChronicPop)  
# Forward Stepwise Regression  
forward.model = step(minimum, scope = list(lower=minimum, upper = full),  
direction = "forward")  
summary(forward.model)  
# Backward Stepwise Regression  
backward.model = step(full, scope = list(lower=minimum, upper = full),  
direction = "backward")  
summary(backward.model)  
# Forward- Backward Stepwise Regression  
both.min.model = step(minimum, scope = list(lower=minimum, upper = full),  
direction = "both")  
summary(both.min.model)
```

The step R function allows performing stepwise regression with different directions, forward, backward and both, forward and backward. Stepwise regression selects the same set of predicting variables for all directions. The variables not selected are unemployment, income, poverty, ranking of the counties by exercise, access, and ranking of counties by social environment. These were predicting variables that were not statistically significant in the full model.

- **Variables not selected by all methods: Unemployment, Income, Poverty, RankingExercise, RankingSocial**
- **State dummy variables followed by number of claims per-member per-month are first selected by forward stepwise regression**

Forward selection, this is forward stepwise regression, also points to the order of variables with the highest explanatory power. State dummy variables followed by the number of claims per-member per-month are first selected by forward stepwise regression.

The regression model with the variables selected by stepwise regression is here.

Stepwise Regression Model					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.0271089	0.0995378	20.365	< 2e-16 ***	
HealthyPop	-0.0005092	0.0007837	-0.650	0.515917	
ChronicPop	-0.0051250	0.0020252	-2.531	0.011418 *	
StateAR	0.9324593	0.0155667	59.901	< 2e-16 ***	
StateIA	0.9003846	0.0118631	75.898	< 2e-16 ***	
StateNC	1.4268425	0.0157605	90.533	< 2e-16 ***	
HO	12.0476486	0.7237072	16.647	< 2e-16 ***	
Education	-0.0016689	0.0002312	-7.218	6.08e-13 ***	
ProvDensity	0.0605923	0.0156154	3.880	0.000106 ***	
RankingsPCP	0.0007885	0.0001577	5.000	5.94e-07 ***	
Availability	0.0756249	0.0191618	3.947	8.03e-05 ***	
Accessibility	-0.0019930	0.0007001	-2.847	0.004433 **	
PO	0.1232428	0.0406869	3.029	0.002466 **	
UrbanicitySuburban	-0.0017746	0.0136754	-0.130	0.896758	
UrbanicityUrban	0.0226383	0.0124409	1.820	0.068870 .	
BlackPop	0.0050790	0.0005596	9.076	< 2e-16 ***	
WhitePop	0.0046371	0.0005522	8.398	< 2e-16 ***	
RankingsFood	0.0158764	0.0040770	3.894	9.98e-05 ***	
---					
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *	. 0.1 ' '
Residual standard error:	0.2322	on 5001 degrees of freedom			
Multiple R-squared:	0.8483	Adjusted R-squared:	0.8478		
F-statistic:	1645	on 17 and 5001 DF,	p-value:	< 2.2e-16	

Compared to the full model, the two dummy variables corresponding to the urbanicity level, which differentiate communities into **rurals** [correction: urban], suburban and

rural, are not statistically significant for the reduced model. But [, but] **they're worst to discuss** [they were statistically] significant for the full model.

Access measures remain statistically significantly associated to log of ED healthcare cost. Moreover, we find out we did not lose explanatory power when discarding the five variables that were not selected by stepwise aggression. Furthermore, we can compare the full model versus the reduced model with the variables selected by stepwise regression using the partial F-test.

## Stepwise Regression Vs Full Models

```
## Compare full model to selected model
reg.step = lm(log(EDCost.pmpm)~HealthyPop+ChronicPop+State+HO+
               Education+ProvDensity+RankingsPCP+Accessibility+Availability+
               PO+Urbanicity+BlackPop+WhitePop+RankingsFood)
anova(reg.step, full)
  Res.Df   RSS   Df Sum of Sq    F    Pr(>F)
  1     5001 269.56
  2     4996 269.46  5     0.10406  0.3859  0.8588
```

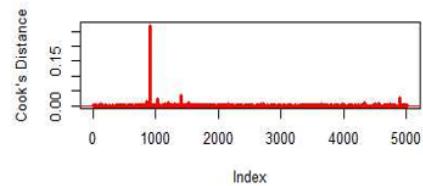
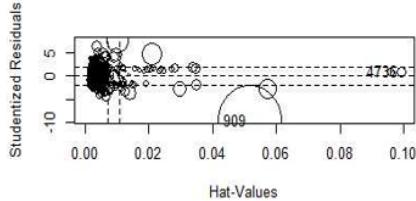
- P-value large  $\Rightarrow$  Do not reject the null hypothesis (reduced model)
- The reduced model is plausibly as good in terms of explanatory power as the full model.

The R command is ANOVA, with the reduced and full model as inputs. The output of this command is here. Based on this output, the p-value of the test is large, indicating that we do not reject the null hypothesis corresponding to the reduced model.

Thus, we conclude that the reduced model is plausibly as good in terms of explanatory power as a full model. And thus, we prefer the reduced model because it includes a smaller number of predicting variables.

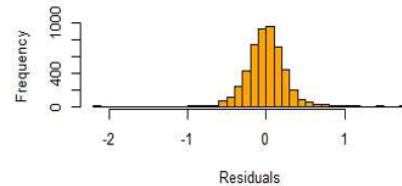
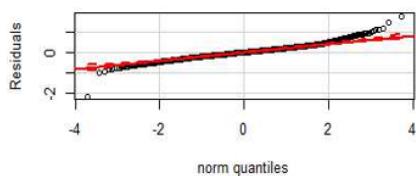
Performing a similar residual analysis as for the full model, we find that the observation 909 is again an outlier, just like for the full model.

# Residual Analysis: Outliers & Normality



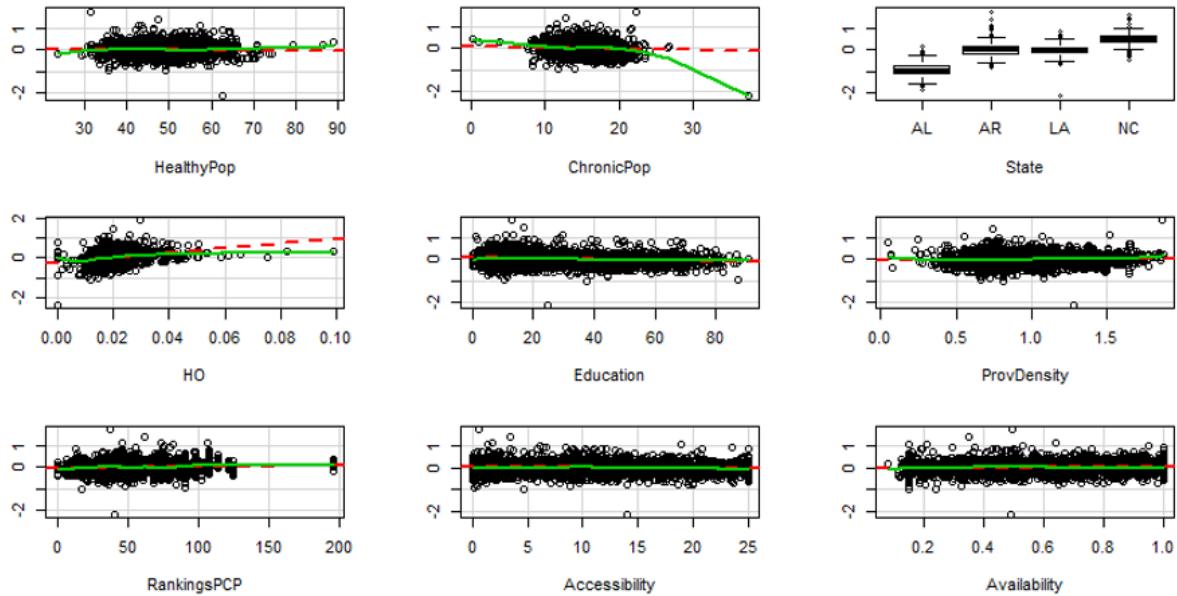
**Outliers:** Observation 909 stands out

**Normality:** Symmetric but heavy tails



The distribution of the residuals is symmetric but with heavy tails. As far as the linearity assumption, nothing changed as compared to the full model.

# Residual Analysis: Linearity



**Linearity:** Some nonlinearity with respect to HO, BlackPop & WhitePop

**Transformations:** Not an improvement in the fit

We find that some nonlinearity with respect to the e-patient utilization, and respect to the percentage of population, black and white population. I've tried to transform these variables, but there was no improvement in the fit.

Let's now **to** see the impact of the outlier 909. So here I'm comparing the output of the model with and without this observation that we identify as being an outlier.

# Removing Outlier

## With Outlier

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.0271089	0.0995378	20.365	< 2e-16 ***
HealthyPop	-0.0005092	0.0007837	-0.650	0.515917
ChronicPop	-0.0051250	0.0020252	-2.531	0.011418 *
StateAR	0.9324593	0.0155667	59.901	< 2e-16 ***
StateLA	0.9003846	0.0118631	75.898	< 2e-16 ***
StateNC	1.4268425	0.0157605	90.533	< 2e-16 ***
HO	12.0476486	0.7237072	16.647	< 2e-16 ***
Education	-0.0016689	0.0002312	-7.218	6.08e-13 ***
ProvDensity	0.0605923	0.0156154	3.880	0.000106 ***
RankingsPCP	0.0007885	0.0001577	5.000	5.94e-07 ***
Availability	0.0756249	0.0191618	3.947	8.03e-05 ***
Accessibility	-0.0019930	0.0007001	-2.847	0.004433 **
PO	0.1232428	0.0406869	3.029	0.002466 **
UrbanicitySuburban	-0.0017746	0.0136754	-0.130	0.896758
UrbanicityUrban	0.0226383	0.0124409	1.820	0.068870 .
BlackPop	0.0050790	0.0005596	9.076	< 2e-16 ***
WhitePop	0.0046371	0.0005522	8.398	< 2e-16 ***
RankingsFood	0.0158764	0.0040770	3.894	9.98e-05 ***
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *

Residual standard error: 0.2322 on 5001 degrees of freedom  
 Multiple R-squared: 0.8483, Adjusted R-squared: 0.8478  
 F-statistic: 1645 on 17 and 5001 DF, p-value: < 2.2e-16

## Without Outlier

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.9356344	0.0991296	19.526	< 2e-16 ***
HealthyPop	0.0003798	0.0007824	0.485	0.627430
ChronicPop	-0.0010849	0.0020519	-0.529	0.597031
StateAR	0.9379139	0.0154403	60.745	< 2e-16 ***
StateLA	0.8989533	0.0117596	76.444	< 2e-16 ***
StateNC	1.4282364	0.0156224	91.422	< 2e-16 ***
HO	11.5397384	0.7193214	16.043	< 2e-16 ***
Education	-0.0017147	0.0002292	-7.480	8.72e-14 ***
ProvDensity	0.0654339	0.0154862	4.225	2.43e-05 ***
RankingsPCP	0.0007560	0.0001564	4.835	1.37e-06 ***
Accessibility	-0.0018658	0.0006940	-2.688	0.007205 **
Availability	0.0755848	0.0189930	3.980	7.00e-05 ***
PO	0.1338608	0.0403440	3.318	0.000913 ***
UrbanicitySuburban	-0.0006647	0.0135555	-0.049	0.960895
UrbanicityUrban	0.0222961	0.0123314	1.808	0.070654 .
BlackPop	0.0050502	0.0005547	9.105	< 2e-16 ***
WhitePop	0.0044178	0.0005478	8.064	9.14e-16 ***
RankingsFood	0.0162198	0.0040412	4.014	6.07e-05 ***
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *

Residual standard error: 0.2301 on 5000 degrees of freedom  
 Multiple R-squared: 0.8504, Adjusted R-squared: 0.8499  
 F-statistic: 1672 on 17 and 5000 DF, p-value: < 2.2e-16

And the output does not show any striking differences in estimated coefficients, except for the variable corresponding to the percentage of population with chronic conditions.

Now in the model without the outlier, this variable is not statistically significant associated to log of ED cost, as compared to the model with the outlier. The R-squared only changes slightly. Thus, the influence of this observation on a model fit is not substantive. Nonetheless, we'll continue with the interpretation of the model, based on the model on the data without this observation.

## Model Interpretation: Access to Care

### Access to primary care:

- Availability – proxy measure of wait times for appointment, measured in level of congestion and takes values between 0 and 1 (the higher the value, the worse the wait time)
- Accessibility – travel distance to primary care providers, measured in miles

### Interpretation:

- An increase of 1% in lack of availability of primary care providers results in 0.0755 unit increase in log(ED cost PMPM) given all other predictors fixed
- A reduction of 1 mile in travel distance to primary care providers results in 0.001 unit increase in log(ED cost PMPM) given all other predictors fixed
- The correlation between the two measures is 0.695: If Availability is discarded from the model, Accessibility is not statistically significant.

I'll first provide the interpretation of the regression coefficients corresponding to the dummy variables of the state predictor. Since the regression coefficient are statistically significant, we conclude that there are differences in the cost for ED healthcare per-member, per-month across the four states. The baseline is Alabama. When comparing Alabama to Arkansas, the estimated regression coefficient is 0.938939, which means that the ED cost per-member per-month is exponential of this value higher in Arkansas versus Alabama. Or we can translate this that ED cost is \$30 per-member, per-year higher in Arkansas versus Alabama, controlling for utilization access, access, and socio-economics.

## Model Interpretation: State Differences

### **Location: Comparing ED Costs for AL, AR, LA and NC in 2011**

- ED cost PMPM is  $\exp(0.938)$  higher in AR versus AL, or ED cost is \$30.65 per-member per-year higher in AR versus AL controlling for utilization, access and socio-economics;
- ED cost PMPM is  $\exp(0.899)$  higher in LA versus AL, or ED cost is \$29.48 per-member per-year higher in LA versus AL controlling for utilization, access and socio-economics;
- ED cost PMPM is  $\exp(1.428)$  higher in NC versus AL, or ED cost is \$50.04 per-member per-year higher in NC versus AL controlling for utilization, access and socio-economics;

**Overall interpretation:** Controlling for many potential factors contributing to ED cost, North Carolina pays significantly more while Alabama pays significantly less per-member than other states on emergency care.

Similarly, the coefficient for Louisiana is 0.899, which means that the ED cost per-member per-month is the exponential of this coefficient higher in Louisiana versus Alabama. Equivalently, we can say that the ED cost is about \$30 per-member per-year higher in Louisiana versus Alabama while controlling for utilization, access, and socio-economics.

The same for North Carolina. The estimated coefficient is 1.428, which means that the ED cost is \$50 per-member per-year higher in North Carolina versus Alabama controlling for utilization, access, and socio-economics. Overall, controlling for many potential factors contributing to ED cost, we find that North Carolina pays significantly more while Alabama pays significantly less per-member, than other states on emergency care.

Last, I'll interpret the coefficients for the utilization of physician office and for utilization of inpatient care, measured as a number of claims for physician office and for

hospitalization scaled by the PM, PM [PMPM] scaling factor, or the number of member month [months].

## Model Interpretation: Utilization

### Healthcare Utilization:

- PO – number of claims reimbursed for care in the physician office, a proxy of utilization of regular care
- HO – number of claims reimbursed for hospital care, a proxy of utilization of inpatient care

### Interpretation:

- An increase of one claim PMPM for regular care results in 0.133 increase in log of ED cost PMPM given all other predictors fixed
- An increase of one claim PMPM for inpatient care results in 11.54 increase in log of ED cost PMPM given all other predictors fixed

Because the estimated regression coefficient for physician office variable is 0.133, we interpret this as an increase with one claim per-member per-month for regular physician care, results in a 0.133 **decrease** [correction: increase] in log of ED cost per-member per-month given all other predictors fixed.

Moreover, because the estimated regression coefficient for inpatient variable for the hospitalization variable is 11.54, we interpret this as an increase of one claim per-member per-month for inpatient care, results in 11.54 units increase in the log of ED cost per-member per-month, given all other predictors fixed.

In summary, in this lesson, I provided a variable selection analysis for the study on healthcare cost for emergency department.

## Lecture 5.3.5 – Findings

In this lesson, I'll return to the research question related to The [the] relationship of access on ED healthcare cost. But I also wrap up this data example with some findings related to the analysis we have performed in the previous lessons.

First, we found that access particularly availability measure is statistically significantly associated to the emergency department healthcare cost.

# Access to Care: Intervention

## Access to primary care:

- Availability – proxy measure of wait times for appointment, measured in level of congestion and takes values between 0 and 1 (the higher the value, the worse the wait time)

## Interpretation:

- An increase of 1% in lack of availability of primary care providers results in \$1.078 unit increase in ED cost PMPM given all other predictors fixed

## Policy Research Question:

- Does improvement in availability of primary care providers reduce the cost of ED care?

The more specific interpretation **of** [after] transforming back the ED cost is as follows.

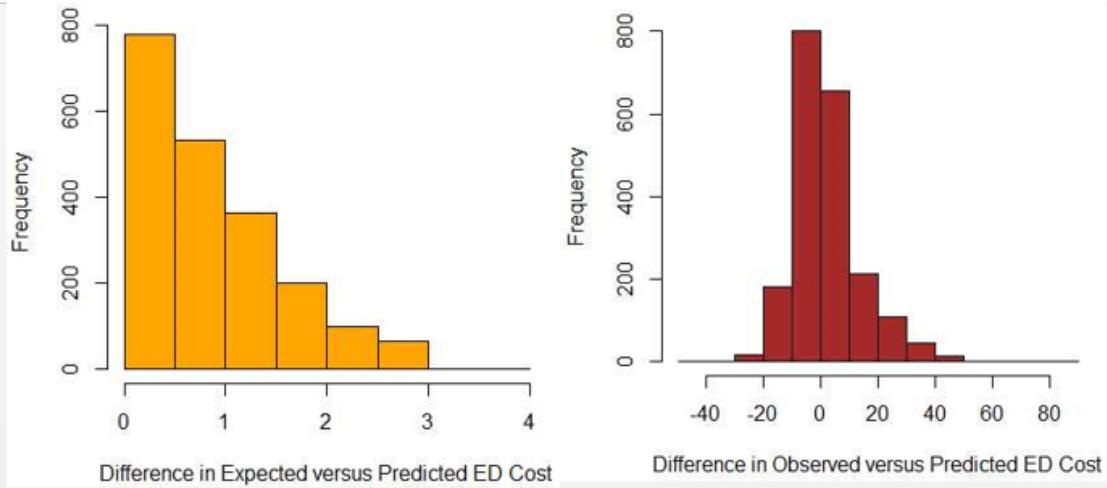
An increase of 1% in lack **fo** [of] availability of primary care providers results in \$1.078 increase in **80** [ED] cost per member per month given all other predictors fixed.

Because we do see an association, the question is, does improvement in availability of primary care providers reduce the cost of ED care?

This is a second research question posed in their introduction of the study. For this, we'll change the values **of** for the availability **predict in verbal** [predicting variable] but keep everything else fixed and then predict. Particularly, I defined, the **availability that intervention** [availability.interv] vector **Which** [which] takes the same values as the availability vector except for those values larger than 0.5. If the availability measure is larger than 0.5[,] I'll replace that with 0.5.

# Findings: Access Intervention

```
Availability = dataAdult.no.out$Availability  
# Improve Availability to less than 0.5 congestion experienced by all communities  
Availability.interv = Availability  
Availability.interv[Availability>=0.5] = 0.5  
newdata=dataAdult.no.out  
newdata$Availability=Availability.interv  
index = which(Availability>=0.5)  
# Predict by changing availability with all other predictors fixed  
EDCost.predict = predict(reg.step.no.out, newdata,interval="prediction")[, 1]  
# Compare predicted to fitted for those communities with intervention  
EDCost.diff.fitted = exp(fitted(reg.step.no.out)) - exp(EDCost.predict)  
hist(EDCost.diff.fitted[index],xlab="Difference in Expected versus Predicted ED Cost")  
# Compare predicted to observed for those communities with intervention  
EDCost.diff.observed = EDCost.pppm[-909] - exp(EDCost.predict)  
summary(EDCost.diff.observed[index])  
hist(EDCost.diff.observed[index],xlab="Difference in Observed versus Predicted ED Cost")
```



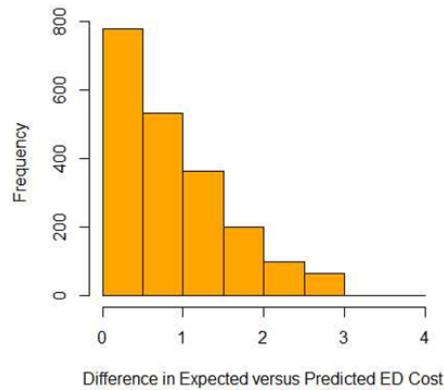
In terms of interpretation of this change, I'm assuming that I can design an intervention that targets those communities with an availability measure larger than 0.5. In [in] a way that I can reduce the congestion experienced by the population in this [those] communities to 50% of the provided capacity. Resulting [resulting] in high availability of the primary care providers or less wait time for employment [an appointment] to see a primary care provider.

After changing the availability variable in this way, I create a new data where I replace only the column of the availability variable and everything else stays the same. This is the new data for which I predict the ED healthcare cost.

In the last our [R] command in the last set of our [R] commands I compare the predicted cost. When [, when] I change the availability of some of the communities[,] with the expected cost or the Timated [estimated] expected cost. And I also compare the predicted cost with the observed cost. And I compare this, I'm [I] perform this comparison. Only [only] for those committee's that [communities where] I've changed [the] availability measure and [. And] I the [INAUDIBLE] [indices] for those committee [communities are provided by] for [the] index vector. Which [, which] is equal to which availability greater than 0.5 so [. So] that's how I get that index using the which command. .And [and] I compared a cost only for those communities that [where] I'm changing the availability value.

So let's look at the histogram between the difference in expected versus predicted. And a difference between observed versus predicted ED cost[.]

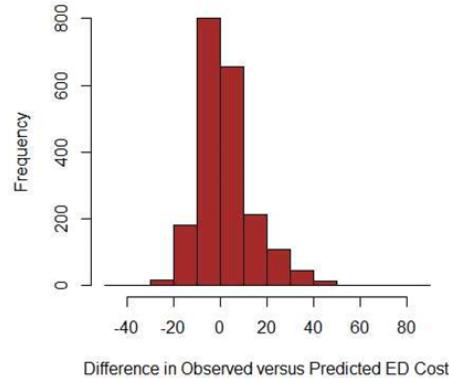
Note here that I used the exponential value of the predict cost. And I used the exponential value of the expected cost and take the difference, because our model, models log of ED costs. When I look at the difference between observed versus predicted, I only check the exponential of the predicted ED cost.



The of the first the orange histogram corresponds to the difference in expected versus predicted. We can see that the difference is always positive and for some communities the difference in cost is of almost \$3 per member, per month. For other communities, the difference in the cost is smaller than \$1 per member, per month.

Now, well, [while] this doesn't sound a large If [, if]we multiply the cost with 12 to get the cost difference per year. And [, and] we multiply that with a [the] number of adult [adults] within each census track [tract], the difference in the cost could be quite large.

The histogram on the right **correspond** [corresponds] to the differences between observed **Versus** [versus] predicted ED cost.



And for this, we can see that we have both positive, and large [correction: negative] differences.**and**, [and] this is because observed is the expected plus **sum** [some] error term. And, that error term **Makes** [makes] some of the differences to be positive and **or** [others] negative.

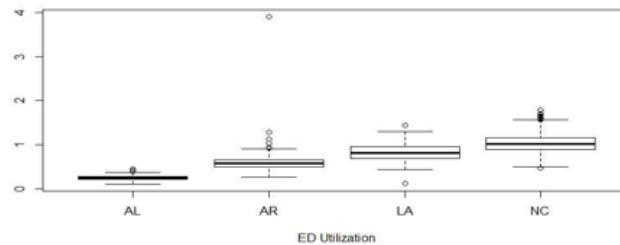
Well, we see that, there is a longer tail on the right, on the positive side which says that there **is**. **There** are more communities with a positive difference than with a negative difference.

In the previous lesson, the results from **and** the regression modeling **show** [showed] that there are large variations in healthcare cost for the ED care across the four states. **With** [, with] North Carolina being the leading state and Alabama being the trailing state in cost of ED care.

What would be the case of such significant differences that could result in millions of dollar difference in ED costs? First Medicaid programs are operated by States. Thus, they do not operate under the same health policies and under the same reimbursement of health care services. For example, one state may apply higher reimbursement for primary care services than other states or they may apply **low** [lower] reimbursement for emergency **permanent** [department] care than other states.

## Findings: State Variations

- There are large variations in healthcare cost for the ED encounters across the four states, with North Carolina being the leading state and Alabama being the trailing state in cost of ED care; *Why?*
- Medicaid programs vary by states, with different health policies and reimbursements levels.
- ED utilization PMPM is also highest in North Carolina and lowest in Alabama



The correlation between ED cost and ED utilization is 0.899

Such differences in what the Medicaid programs pay for the emergency care services may lead to differences in ED costs per member per year as we see here. However, another plausible explanation is that the **adult** [adults] in North Carolina **utilizing** [utilize the] emergency department more than those in Alabama for example.

We see this when we compare utilization level across the four states, **we** [. We] see that Alabama has the lowest utilization level per adult and North Carolina has the highest utilization. In fact, the correlation between ED costs and ED utilization is close to 0.9 **because** [. Because] of this explanation a next step in such analysis would be to study the factors influencing ED utilization.

To **reminded that one over** [remind you that one of our overarching objectives] of improving access to primary care is to increase utilization of primary care over the utilization of the emergency care. Thus we also included in the model **of** [a] variable which is a proxy for non-emergency care, particularly, number of physician office claims per adult per month. This variable is positively associated to ED cost of care given the other predictor variables fixed in the model[.]

## Findings: Utilization

- Utilization of physician office is positively associated to ED cost of care given the other predicting variables fixed in the model;
- Correlation between utilization of physician office and ED is high (0.54) thus the positive relationship to ED cost may because there are communities with higher utilization of healthcare in general and thus higher ED costs.
- Utilization of inpatient care (hospitalizations) is positively associated to ED cost of care given the other predicting variables fixed in the model;
- There is a very weak correlation between utilization of ED and utilization of inpatient care; further investigation is needed.

, this [This] is an unexpected result. We would expect for the relationship to be negative **that is** [. That is,] the higher the, the physician office utilization, the lower the ED cost for a community.

Why do we see such a **positive**, positive association? The **correction** [correlation] between utilization of physician office and the utilization of the emergency department is actually high, **it's, it's** [. It's] 0.54. Thus the possible relationship to ED cost **maybe** [may be] because there are **quantities** [communities] with higher utilization of healthcare in general regardless **for the** [whether it's] primary care, emergency or other services. This may be due to the [fact that in] access to healthcare **it** [. It] is common **that** [for] healthcare providers, to collocate themselves. And thus close to a primary care provider, we may also have an emergency department.

**Well**, [While] we control for the level of **organicity** [urbanicity], we'd expect better access in urban areas than in rural areas, for example. Thus higher utilization for any type of care, **we** [. We] also estimate a positive association of inpatient utilization to ED cost of care given the other predicting variables fixed in the model. The estimated coefficient for the inpatient utilization is highest among all coefficients.

What is interesting is that when one considers marginally[,] there's little correlation between the ED utilization or log of ED cause and inpatient utilization. However, when considered in the presence of other predictors, there is a stronger association between ED cost and hospitalization utilization.

**Full** [Further] investigation in the **direction** [connection] between this variable and other predicting variables **where it's** [is] needed in order to understand this. **This** difference between the marginal and conditional relationship[.]

as [As] far as the other predictive variables. For [, for] example, socio-economic variables[,] all expect [except] for education are not selected to be in the reduced model. That [that] is selected by the stepwise regression thus [. Thus,] these variables do not add any additional explanatory power in addition to the other variables included in the model.

## Findings: Other Variables

- Socio-economic variables except for Education are not selected to be included in the reduced model; they do not add additional explanatory power given the other predicting variables in the model;
- Availability of primary care providers is statistically significantly associated to ED cost of care, and intervening to improve availability will show a reduction in the expected ED cost of care according to the fitted model; such analysis however relies on causal inference and thus the regression model not appropriate to address such research question.
- Whether living in urban or rural communities is not statistically significantly associated to ED cost of care given other predicting variables in the model.

In fact, when you consider each [one] of it marginally[,] so we consider the **margin model**. [marginal models with ] Respect [respect] to the log of ED cost[,] we see that there are the correlation to the log of ED cost is rather low, lower than 0.1 absolute value. Thus this **variables s** [variable is] not associated to ED cost neither conditionally nor marginally.

I'll again reiterate our findings in terms of our attempt to use the model to evaluate the impact of interventions for improving availability of care. All we found here is that the availability of primary care providers is statistically significantly associated to ED cost of care. But also intervening to improve availability will reduce the **ED cost of care**, expected ED cost of care but not necessarily the **upserved** [observed] ED cost.

What I'm going to point out here is that this prediction exercise **is** [has] assumed that we model a causal relationship, thus, through measuring the impact of **the liability** [availability] onto ED cost. **But**, [But] this is not possible with the model we considered here.

First, we base our model on **a** [an] observation study. Second, there are predicting variables in a model that are correlated to availability. For example, provider density

has a direct relationship with availability of the providers. Provider density is high in the community if there are many healthcare providers, in that community or that community has access to many healthcare. **Care**-providers including primary care, specialist, dentist, hospitals and so on.

However, healthcare providers tend to cluster and we should expect that where we have many healthcare providers **will** [we'll] also have many primary care providers. Thus altering **changing** the availability **Factor** [factor] implicitly will change the provider density as well. **The such** [Thus, such] analysis of the impact of intervention cannot be performed with regression models unless in an experimental setting.

Last, an important finding is that living in a rural community **Is** [is] not statistically significantly associated to ED cost of care given other predicting variables in a model. However, performing an ANOVA for differences in ED cost means across the three groups of urbanicity. We reject the null hypothesis of equal means and thus there is a marginal relationship with respect to ED cost. It is possible that the provider density again might explain some of the variability due to **urban or city** [urbanicity] since there are statistically significant differences. **In** [in] the means of the provider density for **urban, suburban**, and rural communities, as provided by an ANOVA analysis.

In this lesson, I provided an overall summary of the findings based on this data analysis. The summary and those findings do not only rely on the regression analysis but also on understanding the applied problem.

This example is a common practice in regression analysis, where we begin with exploratory data analysis. We perform regression analysis in the context of the problem, we evaluate goodness of fit. We perform variable selection and we conclude with findings in the context of the applied problem.

# UNIT 6 – OTHER REGRESSION MODELS

## LECTURE 6.1: OTHER REGRESSION MODELS (PART 1)

### Lecture 6.1.1 – Weighted Least Squares Regression

In the lessons in this lecture, I'll introduce briefly other approaches in regression analysis to give you a sense of the scope of regression analysis. In this lesson, I will focus on one particular assumption, the constant variance assumption. And I'll introduce a method that is commonly used to address departures from this assumption.

Let's review the multiple regression model. In linear regression, the data are the response variable, given the values of the predicting variables. More specifically, you observe N realizations of the response along with the corresponding predicting variables. The relationship **capture** [captured] is a linear relationship between the response and the predictors.

## Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**What if the variance is not constant?**

- Transform the response variable using a variance-stabilizing transformation
- Weighted Least Squares Regression

- Constant Variance Assumption:  $\text{Var}(\varepsilon_i) = \sigma^2$

- Independence Assumption:  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables

- Normality Assumption:  $\varepsilon_i \sim \text{Normal}$

The assumptions describe the probability distribution of the data. For multiple linear regression, we have seen that the deviances or error terms, denoted as epsilon i and defined by the difference between the response and the linear function in x, have 0 mean and constant variance, and they are independent. For estimation, we only need

those assumptions. However, for statistical inference we also need to assume that the error terms are normally distributed.

In this lesson, I'll focus on the assumption of constant variance. Constant variance assumption means that it cannot be true that the model is more accurate for some parts of the population, and less accurate for other parts. A violation of this assumption means that estimates are not as efficient as they could be in estimating the true parameters. It also results in poorly calibrated prediction intervals.

What if the assumption does not hold? Since this is an important assumption for well-calibrated predictions and efficient estimation, it's important to consider modelling approaches that address this. We can transform the response variable using a variance stabilizing transformation. This is how we dealt with this departure from this assumption in the previous lectures.

However, we can also use a different regression approach called weighted least squares as I'll briefly introduce in this lesson. I'll introduce here one particular example of regression analysis when we should expect to have a non-constant variance.

Assume that we observe binomial data, but let's also assume that we only recorded the frequencies rather than recording both the number of successes and the number of trials. For example,  $D_i$  on the slide, is the number of individuals in a population with a specific condition or disease. And  $m_i$  is the total number of individuals with or without the condition.

Instead of decoding  $D_i$  and  $m_i$ , we have only the prevalence of the disease  $Y_i$ . If we know that  $m_i$  is large, as commonly in such data examples, then we can use a normal approximation of the binomial distribution according to the central limit theory and assume an approximate normal distribution for the frequency  $Y_i$ . But although we can approximate  $Y_i$  with a normal distribution, we cannot assume constant variance because the variance of  $Y_i$  depends on  $m_i$  as we learned in previous lessons. This is a classic example where we can assume normality but not constant variance.

What is weighted least squares? It's a multiple regression model. But the difference is that we assume that the variance of the errors is not constant. In fact, weighted least squares, also abbreviated as WLS, can be general in the sense that the vector of errors can be assumed to have a covariance-variance matrix  $\sigma^2$ , thus allowing for correlated errors. We have the independence assumption only when  $\sigma^2$  matrix is a

diagonal matrix, which is in fact the common implementation of the weighted least squares regression.

# Weighted Least Regression (WLS)

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

**Assumptions:** For the vector of errors  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon) = 0$
- **Covariance-Variance Assumption:**  $V(\varepsilon) = \Sigma$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables if  $\Sigma$  is a diagonal matrix
- *Normality Assumption:*  $\varepsilon \sim \text{Normal}$

How does the estimation of the weighted least squares regression change? We still use a least squares approach. But this time, we need to account for the variance-covariance matrix sigma in weighting the error terms by their standard errors. And when there is correlation, in other words, sigma is not diagonal, we need to decorrelate the errors. Thus, we'll add the inverse of the sigma matrix to the least squares objective function as in this slide. The resulting estimated beta is as here.

The estimated regression coefficients still have a closed form expression, however, we correct for the variance-covariance matrix of the error terms. The estimated regression coefficients remain unbiased, but the variance of the coefficients changes. The sample distribution of beta estimators also remains to be normal under the normality assumption. And practically all the statistical inference we learned about in the lectures on multiple linear regression will be the same. The upshot is that we need to know the covariance-variance matrix sigma, not only for statistical inference, but also for obtaining the estimate of the regression coefficients, [.]

thus [Thus,] how to get sigma. [?] The most common implementation of the **the** weighted least square, which is also the implementation in R, is when we assume that the error terms **aren't correlated** [are uncorrelated] by the variance **by** [of the] ith error

term is  $w_i$  times sigma squared, where  $w_i$  has a role of the weight specifying the importance of the  $i$ th response in the estimation of betas, of the regression coefficients.

The example I provided before when we begin with frequency data **approximate** [approximated] by normal using the central limit theorem has such formulation. The weight for the  $i$ th response is 1 over  $m_i$  which, **In** [in] this case is known for this particular example.

What if the weights  $w_r$  are not known? How can we obtain the weights? One option is to use external information, **there** [. There] are some cases where all the information on the variance is available. For example, when we know the measurement errors. That is [a] very rare case. When replications are available, that is for each set of predicting variable  $x_i$ , we observe multiple responses. **Then** [, then] the weights can be estimated by estimating the variance of the **replicate** [replicates]. **Last** [The last] option is to estimate the variance as a function of the predicting variable or variables using a nonparametric regression.

How to implement weighted least squares in R, [?] **we** [We] can use this same R command, **lm**, that we use for fitting the multiple linear regression with a specification of the weight vector in order to estimate the variance. Given the predicting variables, you can use many R commands. But the simplest one for one predicting variable is the **lowest** ["lowess"] R command.

In summary, in this lesson I presented a regression approach called weighted least squares, which is in nutshell, the multiple linear regression assuming non-constant variance. I will not expand on this approach farther because it's an advanced statistical modeling approach. But I just wanted to give you a sense of other linear regression models that you could use when the variance is not constant.

## Lecture 6.1.2 – Robust Regression

In this lesson, I'll introduce a regression approach that is robust to outliers. That is, we approach where when we can estimate the regression coefficients in the presence of outliers.

Let's return to the slide on multiple regression once more. While the presence of outliers or lack of it, is not an explicit assumption in multiple linear regression. We [, we] have learned in previous lectures that it's important to explore the influence of outliers on the estimation, and the statistical inference of a regression analysis.

In fact, when there are outliers in the data, we'll have departures from the normality assumption, even if the distribution of the residuals is symmetric. If there are outliers, the distribution has heavy tails and that's not normal. The assumption **of** [that] the errors are normally distributed is needed if we want to do any confidence or prediction intervals or hypothesis tests, which we usually do. If this assumption is violated, hypothesis test and confidence and **predication** [prediction] intervals can be very misleading. Thus, it is important to address outliers in any regression analysis.

How do we deal with outliers? If one or two, remove the outliers and fit again. Then compare the models with and without outliers. This is how we dealt with outliers in previous lectures.

But, if we have many outliers then again, the distribution of the error terms and hence, of the residuals, is **have it** [heavy] tailed. Moreover, we cannot remove the outliers one by one and re-fit to compare the models when we have a lot of outliers. It would be too complicated in terms of interpretation and comparison. This is when we need a different approach that would provide estimated regression coefficients and statistical inference that is robust to outliers. **When** [when] we have a **tail** [heavy-tailed] distribution that **is** [has] **symmetry** [symmetric] **when we** [we need to] replace the normal distribution with **the distribution with** the **probably** [probability] density function as on the slide.

The main difference between this distribution and that of a normal distribution is that we have now the absolute value of the difference between  $y$  and the centrality parameter  $\mu$ , **or as** [whereas] for normal distribution, we had  $y - \mu$  squared.

This distribution has heavier tails than the normal. If we estimate  $\mu$  using the maximal likelihood estimation approach, it reduces to minimizing with respect to  $\mu$ , the sum of the absolute values between observations and the parameter  $\mu$ .

In fact, the parameter mu is not the mean or the expectation anymore, but it **was** [is] the median of a distribution. Moreover, the estimated mu using this approach is the sample median.

How do you translate this in linear regression analysis? **Would he place** [We replace?] the sum of squared errors with the sum of absolute errors, thus **estimate** [estimating] the regression coefficients by minimizing the sum of absolute errors.

Let's review the comparison between the ordinary least squares approach, abbreviated OLS, and the robust regression approach, abbreviated RR. In OLS, we minimize the sum of square root errors to estimate the expectation of Y, given the predictor variables. In robust regression, minimize the sum of absolute errors to estimate the median of Y, given the predictor variables.

Both the expectation and the median are measures of centrality of the distribution. However, median is robust to **appliers** [outliers], whereas the mean or the expectation is not. Since one would prefer robust regression over ordinary least squares since it's robust to potential **informational** [influential] points, why not using [use] the robust regression **win** [in] all **contests** [contexts].[?]

First, we cannot obtain **close** [closed] or exact expressions for the estimated regression coefficients as in OLS. Thus, we need to employ a numeric algorithm. In fact, the numeric algorithm uses the idea of the weighted least squares and applies weighted least squares iteratively. **Until** [until] the estimated regression coefficients do not change much from one iteration to another. Thus, the estimated coefficients are approximate estimates.

**Since this** [Statistical] inference relies on an approximation [of] the **distributionary regression's coefficients** [distribution of regression coefficients] also. Moreover, the estimated confidence intervals are wider **from** [for] robust regression **then** [than] **from** [for] ordinary least squares. And thus, the statistical inference is more conservative.

In summary, in this lesson, I introduced a regression approach that is robust to outliers. Specifically, we estimate with robust regression the estimated regression coefficients in the presence of outliers.

### Lecture 6.1.3 – Nonlinear & Nonparametric Regression

In this lesson, I'll introduce two different approaches that are commonly used to deal with nonlinearity. Let's return to this slide on multiple linear regression once more.

In multiple linear regression, you assume that the expected value of the errors is zero, also meaning that the linearity assumption holds. What if this assumption does not hold? Here are three situations.

1. If it does not hold for a few quantitative predicting variables, we can assess multiple transformations of the predicting variables to improve the **fate** [fit]. This is how we deal with departures from linearity in previous lectures.
2. However, there are situations when the relationship between the response variable and the predicting variables is known but it cannot be expressed as a sum of transformed predicting variables. In this case, we can use the so called nonlinear regression.
3. More generally, if the linearity assumption does not hold for many variables, or and it is difficult to identify, transformations that improve the **fate** [fit], we can use the so-called the Generalized Additive Regression, which is a non parametric model.

Here's an example from my research where a nonlinear regression applies. In mass spectroscopy or nuclear magnetic resonance frequency data, often the data can be modeled using this sum of components where each component in the sum is a function of the **finity** [affinity] variable  $x$ , the so called frequency, and a series of parameters.

The relationship between  $y$ , the response variable, and the frequency  $x$ , looks like in this plot. **Where** [where] each large peak corresponds to one component in the sum, with  $\mu_L$  being the center of the  $L$  peak, and  $A_L$  being its amplitude. Clearly, we cannot express a relationship between  $y$  and  $x$  using a transformation. This is an example where nonlinear regression can be applied.

As illustrated in this example, in nonlinear regression, the regression function has a known structure given the predicting variables, or variable. And the regression function depends on a series of parameters. More generally, similar to the regression models we learned in this class, the data consists of a response variable and a set of predicting variables. The difference is that response is the sum of the regression function plus the

error, where the regression function is a function of the predicting variables and the parameter theta, which often is a vector of parameters.

In nonlinear regression, we know the function F up to the parameters theta. That is, we know the function F except for the theta. We estimate the model by minimizing the sum of squared errors, the difference between the response, and the regression coefficient. We minimize this with respect to theta. When the regression function can not be expressed as a sum of transformed predicting variables as a multiple linear regression, we need to apply numeric algorithms to obtain the estimate for theta.

What are the differences and similarities between multiple linear regression and nonlinear regression? Both methods use the least squares approach for estimation, and assume the same assumptions on the error terms **his** [ . The] goodness of fit can be performed similarly.

In terms of differences, the regression function is nonlinear versus linear in the parameters. And the estimation of the parameters provides estimates that are not in **close** [closed] form for the nonlinear regression. Moreover, if we were to implement a nonlinear regression model in R, in the R statistical software, we cannot use the lm command anymore.

One implementation of a nonlinear regression in R, is the R command nls. Generally, nonlinear regression models are more challenging to implement. But what if we do not know the relationship between the response and predicting variables.[?] Thus, we do not know the regression function F

One could use a non parametric **model** [regression] assuming that the function **does** [is] not dependent on parameters. In non parametric **model** [regression], the regression function has an unknown structure given the predicting variables, and the regression function does not depend on any parameters.

Using non parametric **model** [regression] with an increasing number of predicting variables p, there are many, many possible regression **function** [functions] F, which leads to the so called, curse of dimensionality. To maintain a given degree of accuracy of an estimator, the sample size must increase exponentially with the dimension p. This is what is called the curse of dimensionality.

For example, if we were to have a regression model with p equal to five predictors, we will need about 30,000 data points, **the** 30,000 observations[,] to get the same accuracy for a model with one predictor when n is equal to 300.

A compromise is to decouple the interaction between the predictive variables and use a sum of individual regression functions where each regression function is a non parametric transformation of one predicting variable. Thus, we write the relationship between the response and the predicting variables as a sum of p functions each one, corresponding to one predicting variable.

Each function is assumed to be smooth. That is, it doesn't change significantly with the predicting variable. That means it changes just slowly as a [the] predicting variable changes. This leads to a so-called **the additive or the** generalized additive regression model.

The common estimation algorithm is the so-called back fitting algorithm which reduces to estimating a non parametric model with one predicting variable at each iteration. It starts with some initial estimates for the functions  $f_1$  to  $f_p$  and then at each step it subtracts from the response all estimated functions except for one, which is re-estimated and updated, then continued with the estimation of another function. The algorithm iterates over all functions for several times until there is no significant improvement to the fit.

To overview, in the generalized additive model, versus a multiple regression model, the relationship of predicting variable to the response is assumed unknown. But the estimation for both is still, through minimizing the sum of squared errors. However, the estimation **that** [of] the parameters in the gam, or Generalized Additive Model, do [does] not have a **close** [closed] form expression.

**To commonly use the** [The most commonly used] implementations of the gam are the gam R command from two different libraries. One is the gam library, the other one is the mgcv library. As you gather from this brief introduction, the gam order requires more advanced knowledge of statistical modeling. **Particularly** [,particularly] an understanding of non parametric smoothing and other advanced statistical concepts. You should implement this model after acquiring additional knowledge and such advanced statistical concepts.

To summarize in this lesson, I introduced two approaches that can be used for dealing with **a purchase** [appearance] of nonlinearity.

## LECTURE 6.2: OTHER REGRESSION MODELS (PART 2)

### Lecture 6.2.1 – Time Series Regression

In this lesson, I will discuss another assumption of regression models. Particularly, the assumption of uncorrelated errors and I will brief [briefly] introduce an approach **that is** that deals with correlation[,] particularly, correlation in time.

Let's return to the slide on Multiple Linear Regression once more. In multiple linear regression, the assumption of independence means that the deviances, or in fact the response variables **wise** [Yi's], are independently drawn from the data generating process. Generally, we only care about whether the errors are uncorrelated, rather than independent in regression analysis. Independence can only be guaranteed in experimental designs.

What if the assumption of uncorrelated errors does not hold? Violations of this assumption can lead to misleading assessment of the strength of the regression. This is because the degrees of freedom are not equal to the sample size. In fact, there are less degrees of freedom due to the correlation. Moreover, not accounting for correlation will result in **horrible ability** [higher variability] or uncertainty **that** [in the] estimate thus less reliable statistical inference. A classic correlation in data is the correlation in time[.] **there** [There] are many many examples in **which with** [where] data are collected in time. Time series arise in any field, from economic data to finance, to healthcare, and climate among many others.

The level of time granularity from yearly to minute or to second depends on the objective of the problem at hand. The time granularity also depends on resources needed to observe the time processes of interest and on how smoothly the time processes change over time.

Here are several examples, **our** [. Our] first example is the US yearly GDP, the gross domestic product. The second example is the monthly sales of wine. A third example is the monthly accidental death in the US. A fourth example is the monthly interest rates.

Often, yearly and monthly data are observed over several years, often tens of years. A fifth example is the daily average temperature. The sixth example is the daily stock price of IBM.

Ideally, we would observe daily data so it allows us analysis and multiple other resolutions since daily data can be aggregated into weekly[,] monthly[,] or yearly data. Depending [, depending] on the objective of the analysis but [. But] even more, sometimes, we have data, more granular a time scale, for example, the minute, or even at the second level. For example, the one-minute inter-day S&P 500 return.

More generally, the level of time granularity and the time series modeling to that we'll need [that needs] to be considered in drawing in making inferences on a time process[,] depends on a series of characteristics.

Trend [Trends] can be long-term or long term increase or decrease in the data overtime [over time] or in can be or it can fluctuates [fluctuate] mostly overtime [over time].

Seasonality is influenced by seasonal factors, quarter of the year, month, or day of the week. Once seasonality repeats exactly at the same time in 2,000[?] with exactly the same regular pattern, we have periodicity. But they [there] could be other cyclical trend [trends] that do not repeat over similar or the same period of time. Cyclical trends are when data exhibit rises and falls that are not of a fixed period.

For some time series, we can observe so-called heteroskedasticity, which means the variability varies with time. For example, financial indicators of companies that have been around for the case [decades] would be least so vulnerable [less variable] in the 1980s in comparison to 2010 years.

Last in [is] correlation where time [which at times] can be positive, in other words successive observations are similar or negative, in other words successive observations are dissimilar.

Let's look at a few examples of time series, here's [. Here's] a plot of gross domestic product for the US. The trend is mostly [clearly] increasing monotone over the period of time. There's no seasonality per a deciduous sequitur [, periodicity, or cyclical] of trends, the [. The] variability of the observations does not vary over time.

Here is a plot for the close price of the stock for IBM where IBM stands for International Business Machines. The company was initiated in 1911 but we only have data since 1960's. The trend is overall monotone increasing with fluctuation in the later years. There is no clear seasonality that can be evaluated visually. However, we could identify

such seasonalities if we were to look at less general data, for example, weekly, or monthly data.

What is very clear here is the fact that the variance, the variability in the stock price shows a big change from being statistical significantly smaller in the early years versus much larger in the later years. The so-called Heteroskedasticity, [.]

theta [Data] on a much lower granular hierarchy that [than] day, week, or month is, is, is common in some fields. Including [including] an high frequency financial data or for example, in patient monitoring such as cardiac monitoring using EKG technology.

This on the [On this] slide is an example where I plotted the S&P500 Intraday stock return. I am showing here the data for 9 different days. We can see the intraday data is much different from one day to another, with much higher variability on some days than in others.

What [Why] do we need another set of statistical modeling tools to model time series data.[?] As I pointed out at the beginning of this lesson, time series response data are correlated. This correlation results in a much smaller number of degrees of freedom than otherwise assumed under independence. More over [Moreover,] because of the correlation the data are concentrated into a smaller part of the probability space where the data lie.

Ignoring dependence leads to inefficient estimates of regression parameters, [and] leads to poor predictions. Standard errors are unrealistically small. In other words, too narrow confidence intervals thus improper statical inferences.

This slide provides a snapshot of the basic modeling of time series for the simplest time analysis or  $Y_t$ , the time varying observations with  $T$  is the time index. For [, for] example, day[.] In order to account for trend and seasonality or periodicity, we recompose a time series into three components,  $m_t$ ,  $s_t$ , and  $X_t$ . Where  $m_t$  is the trend,  $s_t$  is the seasonality on [, and]  $X_t$  is the time process after counting for tran [accounting for trend and seasonality].

$X_t$  is often as seem [assumed] to be as [a] stationary process, in other words it's [its] probability distribution does not change when shifted in time. Most classical time series models assume that the time processes are stationary and thus, we need to first estimate a trend and a seasonality component. So struct [, subtract] them from  $Y_t$ , then to model  $X_t$ . Note that  $X_t$  is not an error term, time [Time] series analysis is a topic of [an] entire course in this masters program.

In summary, in this lesson I introduced another regression analysis approach, the time series analysis. In this regression model, we take into account the correlation in the response data[, in] particular[,] correlation in time.

## Lecture 6.2.2 – Spatial Regression

In this lesson, I'll introduce another regression approach which deals with collated data, particularly I'll focus on correlation in space.

This is again the slide on multiple linear regression. In multiple linear regression the assumption of independence means that the error terms or the response variable Y's are independently drawn from the data generative process. Generally, we only can evaluate whether errors are uncorrelated.

What if the assumption of uncorrelated errors does not hold? Violation of this assumption can lead to misleading assessments of the strength of the regression. In the previous lesson, we learned about correlation in time. In this lesson, we'll learn about another classic correlation in data, the correlation in space.

Spatial data arise in many fields, including social economics, demographics, surveillance, health care, brain imaging, other imaging data, movement data, forestry, environmetrics, among many others. Here are a few examples.

In the data analysis example introduced in the previous lecture of this course, one of the predicting variable was the average travel distance to primary care providers for adults estimated at the community level. We should expect that this healthcare access measure will be similar in neighboring communities, since they have access to a similar pool of primary care providers instance. They will probably have similar access barriers to care.

A second example for spatial process is a number of emergency department visits per member per month, the response variable in the regression analysis in the previous lecture of this course. **While in** [In the] analysis I provided in the previous lecture, I ignored the spatial correlation in the response. We saw that the **residual**, [residuals] **they'd show** [showed] clear correlation, which should be modeled rigorously.

A third example, is the worldwide locations of outbreaks of a disease or condition, for example, Zika, which has been of a great concern in the past year. In order to

accurately predict future outbreak locations, we need again to account for the **spatia** [spatial] dependence in the past outbreaks.

A fourth example is functional magnetic resonance imaging or fMRI used to capture brain activity. Our fifth example is the trajectory of the movement of bison in the landscape.

Modern spatial processes requires an understanding of the characteristics of special [spacial] data. Here are a few important considerations.

Thus the spatial process have a trend. The trend can be long-distance, a long-distance increase or a decrease in the data over space. Periodicity or seasonality is not common for spatial processes. However, we can also have heteroskedasticity which means the variability varies with space. For example, that we need more variability in urban areas than in rural areas.

Another important aspect in the observation of spatial process is whether observations are coming from a continuous or a discrete distribution. For example, the healthcare access measure is a continuous spatial process, where the outbreak for Zika is a point process.

Last, spatial processes can be observed over **a** regular **grades** [grid] such as images, **a** [. A] regular **grade** [grid] is also called the lattice. Spatial processes can also be observed on irregular **grades** [grids], for example, much of the geographic data are observed irregularly, communities have different sizes and different shapes.

Let's next look at a few examples of spatial processes. Here is the map of the percentage of children, in the left map, versus adults, in the right map, enrolled in Medicaid who have had at least one emergency department visit in 2012 in the state of Georgia. The darker the red[,] the **high** [higher] the percentages.

For the adult population, we see a trend with high percentages in [the] north of Georgia but lower percentages in the **Atalanta** [Atlanta] area. For children, we identify several high percentages of children utilizing ED, spread across the entire state.

Thus, we see more of a spatial trend for adults **then** [than] for children, but there is **high** [higher] variability in rural communities than urban communities in the utilization of the ED for children.

This is an example of images using fMRI. The plot shows the brain surfaces showing differences in two parts of the brain in a response **of** [to] a given task. Performing

spatial analysis of the fMRI images, we can highlight areas of activation as shown in the yellow on this example.

Here is the map of the estimated smooth probability for the trajectory of bison in the landscape, highlighted in red, characterizing the movement of bison. This estimated trajectory can be used to identify preferred trajectories when the bison moved [move] between foraging areas. It can also be used to assess mode, inhabited features providing valuable information on less [landscape] connectivity.

This slide provides a snapshot of the basic modeling of spatial processes. For the simplest spatial data analysis, the response data are  $Y_s$ , the space variant observations, the space variant responses, where  $s$  is the index for space. For example,  $s$  could index communities in geographic data, or voxels in imaging data. We generally use different modeling approach if  $Y_s$  is a continuous spatial process as opposed to a point process.

Two assumptions in modelling spatial processes are stationarity, meaning that the probability distribution of the spatial process does not change when shifted in space. An [, and] isotropic, meaning that the distribution of  $Y_s$  is the same in all orientations. Much of the spatial advanced statistical modeling focuses on how to relax such assumption which are strictly very or [very restrictive for] spatial processes.

Last, it's also important to identify whether the spatial dependence is more localized, or so-called small-scale dependence, in comparison to strong dependence over large distances.

All these aspect are important considerations in modeling spatial data. The field of spatial statistic [statistics] is just like time scale [series] analysis, requires [requiring] an entire course for covering modelling approaches for spatial data.

In summary, in this lesson, I introduced yet another regression analysis approach that generally deals with correlated data, but the correlation is in space.

### Lecture 6.2.3 – Mixed Effects Models

In this lesson, I'll introduce a last regression analysis approach that deals with replications in the response data.

I'll return now to the one way ANOVA Model. In ANOVA, or Analysis of Variance, the data consists of multiple samples of data for response variable of interest, differentiating groups or populations, described by a categorical variable or label. In our notation,  $Y_{ij}$  are the response data differentiated across  $k$  categories.

We can write a response as a sum between the mean of the category from which the response is observed, plus an error term epsilon. We can decompose the means farther as the sum between an overall mean  $\mu$ , and the group **fx** [effects]  $\tau_i$  where the sum of all group **fx** [effects] is zero.

In many examples of such data, each sample consists of replications of the same measurement across subject or experimental settings. For example, for a set of subjects, we may observe their neurological response through tests of a repeated measurements in the same experimental setting. In this case, the group effect  $\tau_i$  is best thought of as random, because we only sample a subset of the possible outcomes.

How would the ANOVA Model change if we assume that the group effect is random? The common approach is to assume that not only the error term is **normal stability means 0 and cause of variance. But** [normally distributed with zero mean and constant variance, but ] that the group **effort** [effect] is also normally distributed with  $\mu_0$  **on** [and] constant variance, where the **variance is** [variances] of the [error] terms and of the group effect are two different parameters. In this model, we may be interested whether there is variability across subjects for example.

## ANOVA Model: Random Effects

**Data:**  $Y_{ij}$   $j = 1, \dots, n_i$ ;  $i = 1, \dots, k$

**Are the assumptions the same as in ANOVA with fixed effects?**

- **In random effects model, the observations are no longer independent (even if the error terms are independent).**

- $\epsilon_{ij} \sim N(0, \sigma^2)$
- $\tau_i \sim N(0, \sigma_r^2)$
- We might be interested in the in the variability across subjects, i.e.  $\sigma_r^2$ . Is it zero?

Are the assumptions the same as in ANOVA with fixed effects? In random **affects** [effects] model, the observations are no longer independent, even if the error terms are independent.

When to use a random effect model, versus the ANOVA model we've learned in one of the earlier lectures in this course.[?] Generally, a group effect is random. **If** [if] we can think of the responses we observe in that group to be samples from a larger population.

Have we observed all possible outcomes within the particular experimental setting.[?] Here are two examples. If collecting data from different medical centers, "center" might be thought [of] as random. If surveying students in different campuses, "campus" may be a random effect. We can not observe all medical centers or all campuses. We observe only a sample of the medical centers, or only a sample of the campuses.

We can **extent** [extend] this idea to the multiple regression model. In fact, even more **general** [generally], we can assume that some factors are random, whereas others are fixed, leading to the so-called mixed effects model. For example, suppose we study the effect of a blood pressure drug meant to lower blood pressure over time. And we studied several individuals. For each patient we record blood pressure at regular intervals over a week.

We can model blood pressure using a mixed effects model **or** [where] the group effect tau i is assumed random, since the drug for blood pressure will have bearing [?] efficacy in the population. Whereas, we can also model the relationship to the weak effect assuming a fixed effect.

Such models are again estimating using maximum **INAUDIBLE** [likelihood] estimation or a variation of this estimation approach. If not all the X's are the same for each subject, or some observations are missing, things are more complicated in terms of estimation. **More over** [Moreover, the] covariance matrix of Y is more complicated, since it is not a diagonal matrix as in multiple [linear] regression.

Overall such models are challenging to estimate, so use a computer to estimate the models. That is what the computers are for.

In summary in this lesson I introduced, the basics of mixed effects models commonly used to model data with replications.

## Lecture 6.2.4 – Regression Analysis: Overview

In this last lesson of the course, I'll overview all the regression approaches introduced in this course. I'll provide the roadmap of this course.

The first model introduced in this course is a simple linear regression, in which we're interested in the relationship between two quantitative variables. A response variable  $y$  and the predicting variable  $x$ . The predicting variable is assumed to be fixed, whereas the response variable  $y$  is random.

In simple linear regression, we model their relationship as a linear function in  $x$  plus an error term  $\epsilon$ . The assumptions are as follows.

- Linearity, meaning the relationship between the response and the predicting variable is linear, or that the expectation of the error term is 0.
- Constant variance, meaning that the variance of the error term is the same across all observations, [.]
- Independence or less **restrictive** [restrictively,] **on correlated** [uncorrelated] errors
- Normality of the error terms.

All in all, we assume that the error terms are independent and identically distributed from a normal distribution with mean 0 and variance  $\sigma^2$ .

The model parameters are the two regression coefficients, the intercept and the slope, along with the variance of the error term. These parameters are **known** [correction: unknown], but estimated based on the observed data using ordinary least squares approach.

Furthermore, we can perform statistical inference in the regression coefficients using the sampling distribution of the estimated regression coefficients. The t distribution with **n-1** [correction: n-2] degrees of freedom.

An extension to this very simple regression model is to consider the relationship between the response variable  $y$  and one or more qualitative or categorical variables. And of course, when considering one categorical predicting variable, the resulting model is the one-way ANOVA. When considering two categorical predicting variables, the resulting model is the two-way ANOVA, and so on.

Generally in ANOVA, the data consist of multiple samples of data for response variable of interest differentiated in groups or populations described by a categorical variable or

label. We can write the response as a sum between the mean of the category from which the response is observed plus an error term epsilon.

The assumptions on the error terms are **seen, which is also the** [the same as for] simple linear regression, except that we do not have the linearity assumption, since we're not considering a relationship with a quantitative variable.

The model parameters are the group mean parameters, along with the variance of the error terms. The mean parameters are estimated as the group sample means.

In the ANOVA lecture, I pointed out the equivalence between ANOVA and linear regression. Except that in ANOVA, we fit a multiple linear regression model, rather than a simple linear regression, even though we may have only one categorical predicting variable as in one-way ANOVA.

An extension of the simple linear regression in ANOVA is the multiple linear regression. In multiple linear regression, we observed  $n$  realizations of the response variable along with the corresponding predicting variables, which can be quantitative and/or qualitative. The relationship captures a linear relationship between the response and the predicting variables.

The assumptions described the probability distribution of the data. For multiple linear regression, we assume that the error terms **epsilon, I** [epsilon i] have 0 mean and constant variance, and they're independent. We also assumed that error terms are normally distributed.

In the linear regression model, the parameter is defining the regression line, beta 0, beta 1, to beta  $p$  are parameters. But we also have the additional parameter, the variance of the error terms, denoted with sigma squared.

These parameters are unknown, but estimated based on the observed data using the ordinary least squares approach. Furthermore, we can perform statistical inference on the regression coefficients using the sampling distribution of the estimated regression coefficients that **the** [is the T] distribution with  $n$  minus  $p$  minus 1 degrees of freedom, where  $P$  is the number of predicting variables in the model.

The normality assumption in the previous models also implies that the response model [variable] is normally distributed. But for binary response data, the response variable has a binomial distribution.

For binary data, we model the probability of a success given the predicting variable using the g link function in [such] a way. **The** [that the] g function of the probability of

the success is a linear model of the predicting variables. The g function is the S-shaped function that models the probability of success with respect to the predicting variables.

The model, where the link function g is the logit function, is the logistic regression. The link function g is the log of the ratio of p and 1- p.

**Whether** [What are] the model assumptions in logistic regression, **our** [? Our] first assumption is the linearity in the predicting variables. Similar to the standard regression model, we also need to assume independence in the response of observed data. The third assumption is specific to the logistic regression model. The logistic regression model assumes that the link function is the logit function provided here. This is an assumption since the logit function is not the only function that yields S-shaped curves. There are other shaped function that are used in modeling binary responses.

The parameters for logistic regression are beta 0, beta 1 and beta p. These parameters are unknown, but estimating based on the observed data using the maximum likelihood approach. Furthermore, we can perform statistical inference in the regression coefficient using an approximate sampling distribution of the estimated regression coefficients, where the approximate distribution is the normal distribution.

Other response data can be in the form of counts commonly modeled using the Poisson distribution leading to the Poisson regression model. The common model used to model Poisson data links the expectation of the response variable to the predicting variables using the log function. This is **the key only** [equivalent] with modeling the expectation of the response variable as the exponential of the linear combination of the predicting variables.

The assumptions in Poisson regression are as follows. First, we assume that a log transformation of the rate is a linear combination of the predicting variables. Second, we assume that the response count data are independently observed. This is **the** [a] similar assumption to that of the standard normal regression. Third, we assume that the link function data is the log function. **Well, for** [Similar to] logistic regression, there are other link functions that are commonly used for Poisson regression. The log link function is almost always used.

Similar to logistic regression, the only model parameters are the regression coefficients estimated using the maximum likelihood approach. We can perform statistical inference **in** [on] the regression **coefficient** [coefficients] using an **approximate** [approximation] of the sampling distribution, the normal distribution.

All the regression models discussed so far fall under [the] more general modeling framework called generalized linear models, or abbreviated GLMs. For GLMs, the response Y is assumed to have a distribution from the exponential family of distributions. Example of distributions in the exponential family of distributions are normal, binomial, Poisson, gamma, among others.

Under the GLM, we model a transformation  $g$  of the expectation of Y as a linear combination of the predicting variables. In this modeling framework, the transformation  $g$  is called the link function, since it links the expectation of the response data to the predicting variables. The transformations  $g$  depends on the distribution of the response data.

Model parameters, their estimation and statistical inference is similar as discussed for logistic regression, **Poisson regression** and Poisson regression.

So far, **we discuss** [we've discussed] how to deal with response data that are not normally distributed, but what if the assumption of constant variance does not hold? For example, one could use the regression model under normality for **con** [count] data or for frequency data instead of using the Poisson or the logistic regression model. For such data, we should expect to have non-constant variance.

In previous lectures, we used transformations for stabilizing the variance. But another common approach is to reweight the least squares and the sum of least squares via the weighted least squares approach, which is an extension of the multiple linear regression model when the variance of the error terms is not constant. The implementation of weighted least squares is very similar to that of multiple linear regression, except that you will need to specify the weight. Everything else is the same as in multiple linear regression.

Another assumption that can be challenging to address using transformations is the linearity assumption. This is particularly challenging when we have a large number of predicting variables. Linearity is assumed for all models introduced in this course, since we primarily focus on linear models.

Alternatively, in order to capture non-linearity, **but** [or?] abnormal [?] relationships between the response data and predictive variables, we can use the non-permitting model called generalized additive model. This model applies an unknown transformation to each predictor where the **transformation** [transformations] are estimated from the data. This model applies to response data following a distribution from the exponential family of distributions, including normal, binomial and Poisson. It is thus a

generalization of all models discussed, so far for assuming uncorrelated, so far assuming uncorrelated errors and constant variance.

In summary, in this lesson, I provided an overview of all the regression analysis models introduced in this class, beginning with a simple linear regression model, and concluding with an extension of all the models, the generalized additive model. While this course is having [heavy in] details, such details are paramount to the correct and rigorous implementation of the regression models.

### **Regression Analysis: Summary**

In this course, you have learned the basics of regression analysis such as linear regression, logistic regression, pearson [Poisson] regression, generalized linear models, and model selection. Throughout this course you have been exposed to not only fundamental concepts of regression analysis, but also many data examples using the R statistical software. Thus, you have the ability now to implement regression models using the R statistical software, along with interpretation of the results and findings derived from such implementations. Although this [This] course provides the basis for other more advanced statistical and machine learning modeling.

Thank you for taking this course.