Module 1: Simple Linear Regression

1.1 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. It has wide applicability to fields like healthcare policy, finance marketing, elections, movie rating, bike share, you name it. The two set of models you will learn in this course include* ***standard linear regression*** *and* ***generalized linear regression.*** *We'll begin today with the simplest regression analysis, which is simple linear regression.*

Slide 3:

Throughout the course, we will regularly come back to a set of data examples to help demonstrate in practice the theories taught in the course. Our first example is a company which sells medical supplies to hospitals and is considering the effectiveness of new advertising program. It wants to assess whether there is a relationship between sales and their expenditure on advertising.

Slide 4:

We have data for 25 offices and for each office we begin our analysis with the two variables of interest; one is sales, measured in thousands. And the other one is advertising expenditure, measured in hundreds. If you look at the first row of these data, you have a value of 963 in sales, which means, for that specific office, the amount of sales is $963,000. For advertising expenditure, the amount of the expenditure is $37,427.

Slide 5:

The second example is related to economic theory. Here, we are interested to evaluate and test 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 rates between two countries. In a perfect world, meaning an efficient international economy, 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 changes in currency rates and inflation rates across multiple countries.

Slide 6:

For this example, we have data across 41 countries, including developed and developing countries. We will study the relationship between inflation difference and exchange rate change.

Slide 7:

In the third example, we will 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 those three states was Florida, which really mattered in the final count. As such, in Florida, there was a recount of votes for weeks after the election date. After this recount, the court decision was that George W. Bush won Florida by a margin of 537 votes, just a very, very small number of additional votes. In this analysis, we will study whether there is a specific aspect in this election that could have overturned the decision on which candidate could have won the presidential election in 2000. 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. In this data example, we will assess the relationship between the votes of the independent candidate and the republican candidate and identify whether Palm Beach County was indeed an outlier**.**

For these 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.

Slide 8:

When we speak of regression data, what do we mean?

We have one 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, the number of bikes rented, the election count across counties in a state. This variable is often referred to as the **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 as the response variable and we will represent it by Y.

We also have a set of other variables or factors that we think might be useful in modeling the response variable. Say, the price of a 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. Those variables are usually represented by Xs, X_1, X_2, and so on. For consistency in this course, I will refer to those variables as predicting, or explanatory variables.

Slide 9:

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 the predicting variable, or variables.

The response variable is a **random variable,** because it varies with changes in the predicting variable, or with other changes in the environment. Whenever we're going to see Y in our notation, it means that it is the response random variable. This is particularly important in the context of statistical inference on the regression. I will add here that in experimental or observational studies from which we derive the data for the regression analysis, we observe the response variable and hence we have observations of the response random variable.

On the other hand, the predicting variable is a **fixed variable.** It is set fixed, before the response is measured. That is, we first set the predicting variable to a fixed value then given this value, we next observe the response. 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 on advertisement at the beginning of that year and then they will observe the sales at the end of that year. Thus first, they fix the advertising expenditure. But then, the sales will depend not only on what they spent on advertising, but also on other factors.

Slide 10:

Here are a few examples, just to give you a feel for the difference between the response and predicting variables.

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 to control the LDL level. And we'll control it with, for example, different types of medications. 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, 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.

In this example, 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 the high school. 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 it may predict how students perform in a college, measured by the GPA, or the college grade point average.

Slide 11:

A simple deterministic relationship between two factors, herein X and Y, is a linear relationship. 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.

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

Further, we can take this to extend 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.

With simple linear regression we estimate the relationship between the response and the predicting variable as one straight line. But that line doesn't fit perfectly the points. We can see the points are around the line.

In the second model, multiple linear regression, we can have a plane for a linear regression model with two predictions. Where, in the last case, 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 as two different predicting variables, and model using a linear regression.

What you need to remember is that the linear regression iis a very general model. Practically, most of the regression models are some variations from linear regression.

Slide 12:

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.

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 carrying out statistical modeling in general, 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.**

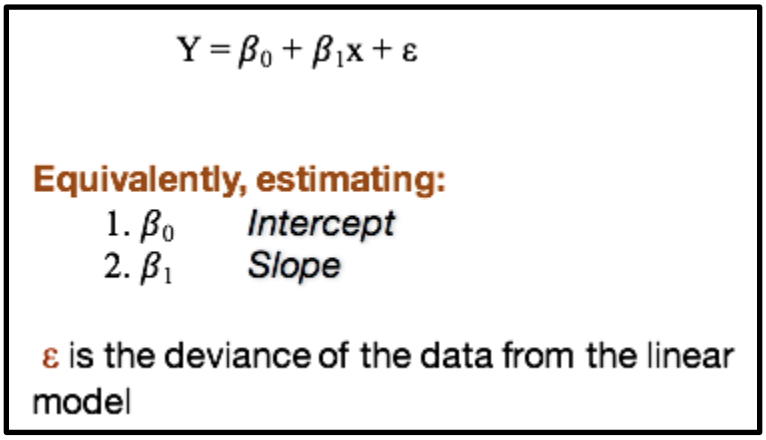
Another useful piece of advice comes from another very famous statistician, John Tukey. "Embrace your data, not your models." While simple regression is a simple model, it has very wide applicability, and it can be generalized to much more complex models.

1.2. Estimation Method

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

Slide 3:

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 β_0 and β_1, where β_0 is the intercept parameter, or the parameter that the value at which the line intersects the y-axis, and β_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.



Slide 4:

The goal is to find the line that describes a linear relationship, that is, to find β_0 and β_1, such that we fit this model. How can we do that? If plot the x and y using a scatter plot, we would like to identify a line that fits the scattered data points. But we can fit many such lines. The question is, how to find the best line? What criterion to use to find the best line? What means “best line”? Based on what criterion shall we identify a best line?

Slide 5:

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 model. 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): $\{(x_1, Y_1), \ldots, (x_n, Y_n)\}$

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, \ldots, n$

Another important aspect in the modeling framework is clearly stating the model assumptions. For simple linear regression, the assumptions 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 error terms or deviances is constant for the given population.
- Independence assumption which means that the deviances are independent random variables.

Later we'll also assume that epsilon's 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 β_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 population. 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 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 at 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. Violation of this assumption can lead to misleading assessments of the strength of the regression.

The errors are assumed normally distributed. This is needed for statistical inference, for example, confidence or prediction intervals, and hypothesis testing. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be misleading.

Slide 6:

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 β_0** and the **slope parameter β_1**, there is a **third parameter: the variance of the error terms.** Thus 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** approximate of those parameters.

Slide 7:

How can we get estimates of the regression coefficients or parameters in linear regression analysis? One approach is to minimize the sum of squared residuals or errors with respect to β_0 and β_1. Specifically, we minimize the sum of the squared differences between data and the model as provided on the slide. T

This translated into finding the line such that the total squared deviances from the line is minimum.

Slide 8:

This is an optimization problem with respect to β_0 and β_1, and the solution to this optimization or minimization problem gives the estimators for β_0 and β_1:

To estimate (β_0,β_1), we find values that minimize squared error:

$$\sum_{i=1}^{n}\left(y_i-(\beta_0+\beta_1 x_i)\right)^2 \Longrightarrow \begin{array}{l}\widehat{\beta_0}= \\ \bar{y}-\widehat{\beta_1}\bar{x} \\ \widehat{\beta_1}=\dfrac{S_{xy}}{S_{xx}} \equiv \dfrac{\Sigma_{i=1}^{n}\, y_i(xi-\bar{x})}{\Sigma_{i=1}^{n}\,(xi-\bar{x})^2}\end{array}$$

We put a **"hat"** ("^") on top of those estimators to differentiate between the estimates and the true but unknown parameters being estimated. So β_0 hat is different than β_0, β_1 hat is different than β_1.

Slide 9:

Let's take a closer look at the derivation of those estimators. Again, what we're interested is to minimize the so called sum of least squares or the sum of squared errors with respect to β_0 and β_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.

Once we take the first order derivatives, we'll have a set of two linear equations with two parameters. Solving these equations will give us β_0 hat, and β_1 hat.

Begin with the minimization problem:

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

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

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n} \left(y_i-(\beta_0+\beta_1 x_i)\right)^2 =0$$

$$\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n} \left(y_i-(\beta_0+\beta_1 x_i)\right)^2 =0$$

$$\widehat{\beta_0}=\bar{y}-\widehat{\beta_1}\bar{x}$$

$$\widehat{\beta_1}=\frac{S_{xy}}{S_{xx}} \equiv \frac{\Sigma_{i=1}^{n} y_i(xi-\bar{x})}{\Sigma_{i=1}^{n} (xi-\bar{x})^2}$$

→ Result into a system of linear equation in β_0 and β_1

→ Solve using linear algebra

→ Solutions to the system are $\widehat{\beta_0}$ and $\widehat{\beta_1}$

Slide 10:

We define the fitted values to be the regression line where the parameters are replaced by the estimated values of the parameters. The residuals are simply the difference between observed response and fitted values, and they are proxies of the error terms in the regression model. The estimator for sigma square is sigma square hat, and is the sum of the squared residuals, divided by n - 2. This is also called the mean squared error or abbreviated MSE.

Slide 11:

The **sampling distribution** of the estimator of the variance is chi-square, with n - 2 degrees of freedom (more on this in a moment). This under the assumption of normality of the 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 β_0 and β_1. But if we replace β_0 and β_1 with β_0 hat and β_1 hat, 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.**

Slide

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

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

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

↑

Estimating σ^2 ← Sample variance

Slide 12:

I will review here the sample variance estimation approach.

What is the sample variance estimation?

Basic statistic concept:

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

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

Why n-1?

We loose a degree of freedom because we replace $\mu \leftarrow \bar{Z}$

Now, going back to $\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2} \sim \chi^2_{n-2}$

This looks like the sample variance estimates except we use n-2 degrees of freedom. **Why?**

We start with Z's that are normally distributed variables with mu (the mean) and sigma square (the variance).

We often use the notation S square, representing the sum of the Z_i minus their average, squared, divided by n -1. From basic statistics, the sampling distribution of S square is the chi-square with n -1 **degrees of freedom.**

Why n -1 degrees of freedom? We lose a degree of freedom because we replace the true parameter mu with z bar.

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. Why is that?

Slide 13:

This is because we've replaced the deviances or the error terms with the residuals. We lose two degrees of freedom because we replaced the two parameters β_0 and β_1 with their estimators to obtain the residuals. 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.

Thus the variance estimator is MSE with a chi-square distribution with n-2 degrees of freedom. This is called the sampling distribution of the variance estimator.

Slide 14:

In simple linear regression, we're interested in the behavior of β_1. We can expect β_1 to be positive, negative, or in fact, close to zero.

If we have a **positive value** for β_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 β_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 β_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.

Slide 15:

We interpret the least squares estimated coefficients as follows:

- **β_1 hat** is the estimated expected change in the response variable associated with one unit of change in the predicting variable.
- **β_0 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 use β_1 hat. However, when we make statistical statements about the relationship, we always have to mention the statistical significance, whether statistically significantly positive, statistically significantly negative, or no statistical significance.

1.3. Estimation Data Examples

In this lesson we'll see examples of the implementation of the estimation concept. I will illustrate the implementation of the linear regression model using the R statistical software. I will also expand on the output of the linear regression model fit in R.

Slide 3:

In this illustrative 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 is: Which is the response and which is the predicting variable?

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. Thus, in this example, the response variable consists of the sales and the predicting variable consists of the advertising expenditure.

Slide 4:

Here is what we will study for this example:

- fit a linear regression.
- identify the estimated regression coefficients β_0 hat and β_1 hat.
- interpret the coefficients.
- quantify 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.
- Predict sales for a very large amount of advertising expenditure, for example $100,000.

Slide 5:

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 columns in the data file have names or not. We also want to extract

specific variables from these data. In this case sales is in the first column, and advertising expenditure in the second column.

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 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 β_0 hat and β_1 hat, and the estimated variance.

```
## Read Data in R
data = read.table("meddcor.txt", sep="", header = FALSE)
## Response & Predicting Variable
sales = data[,1]
adv = data[,2]
## Fit a linear regression model
model = lm(sales ~ adv)

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
Multiple R-squared:  0.8106,Adjusted R-squared:  0.8024
F-statistic: 98.43 on 1 and 23 DF,  p-value: 8.873e-10
```

Estimated Model Parameters:

$\hat{\beta}_0 = -157.3301$

$\hat{\beta}_1 = 2.7721$

$\hat{\sigma} = 101.4$

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

Slide 6:

Let's go back to the questions we wanted to address. The estimated regression coefficients from the output are -157.3,2.77. Based on those estimates we can now write the regression equation:

*sales = estimated intercept + estimated slope * 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. Thus, we 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 used for expenditure, which is equivalent to 10 units, or 10 hundreds; to address question C, we plug in expenditure as 10 units in the prediction of sales, specifically, 10 times the estimated value for the slope gives 27.7 units of sales, which is equivalent to $27,700 increase in sales. Thus the increase in sales with every additional thousand dollars in expenditure is $27,700.

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

Slide 7:

If we were interested to predict the sales for an advertisement expenditure of $30,000, again we have to convert $30,000 into original units, corresponding to 300 units. 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 estimated standard deviation of the error terms, defined in the R output as the residual standard error**. To obtain the estimate of the variance, we need to take the square of the residual standard error: 101.4^2 = 10,281.96

What could we say about the sales for an advertisement expenditure of $100,000? This expenditure corresponds to 1000 units and it's a very large value, way out of the range of the observed advertising expenditure. This is the histogram of the advertisement expenditure and you can see that the maximum value we've observed is 650, or $65,000.

Because the value of advertising expenditure $100,000 for which we would like to predict sales is much larger than the values we observed, this is what we call **extrapolation.** In this case we cannot 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 advertising expenditure increases at the specific value, for example.

1.4. Statistical Inference

We'll now move from estimation to statistical inference. In this 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.

Slide 3:

We will begin with deriving the statistical properties of the estimator for β_1.The expectation of β_1 hat is equal to β_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 β_1, $E(\widehat{\beta_1}) = \beta_1$

we can show $Var(\widehat{\beta_1}) = \sigma^2 / S_{xx}$

I will show you a brief derivation of the expectation of β_1 hat and we can use the same sequence of derivations for the variance of β_1 hat.

Let's go back to the formula of the estimator for β_1. We could write that as the sum of a constant times the random variable Yi where that constant is c_i:

$$\widehat{\beta_1} = \frac{\Sigma(x_i - \bar{x})Y_i}{S_{xx}} \text{ but } x_i \text{ fixed} \rightarrow \frac{x_i - \bar{x}}{S_{xx}} = c_i \text{ fixed}$$

What that means is that β_1 **hat 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.

$$E[\widehat{\beta_1}] = E\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i E[Y_i]$$

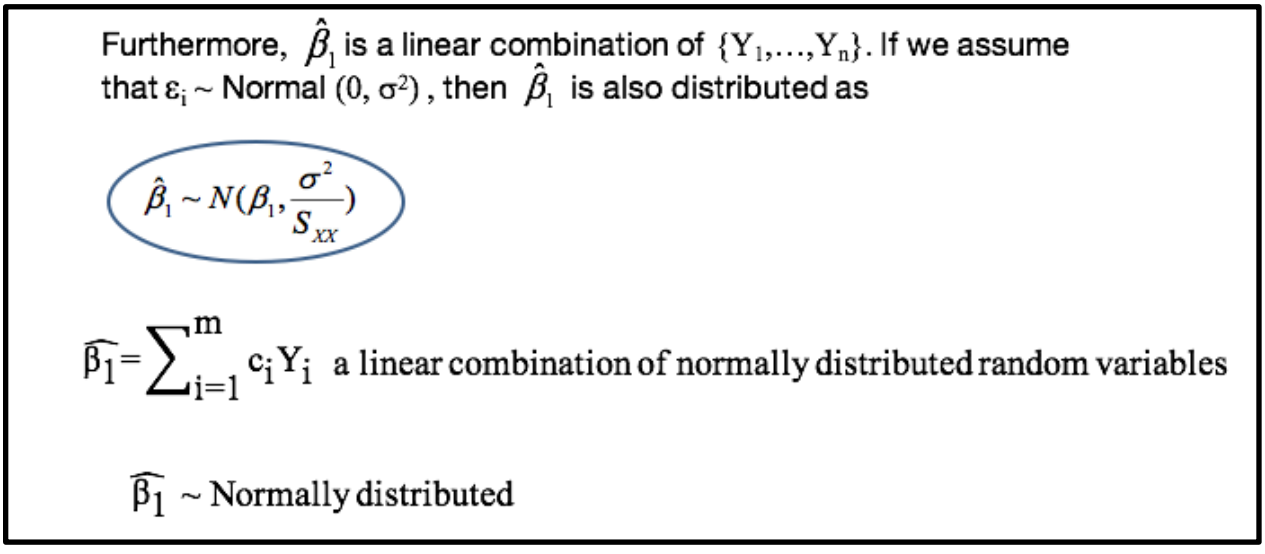
Now we replace the expectation of the random variable of the response variable Yi with the linear relationship in X. We divide that into two sums and now the first sum is equal to zero, because the sum of the constants c_i's is zero, and the sum of the constants time Xi is equal to one.

Thus the expectation of the estimator for the slope parameter is exactly β_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 β_1 hat is an unbiased estimator for β_1. Thus although β_1 is unknown, we have that in expectation the slope estimator is equal to the true parameter. This is an important statistical property.

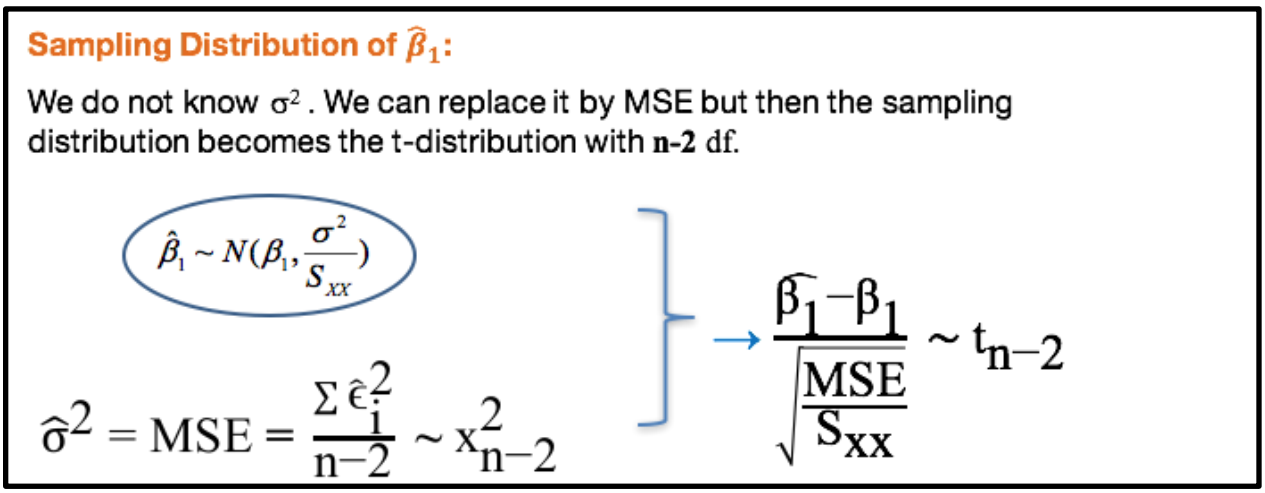
Slide 4:

Let's dive more in into the properties of this estimator. I will remind you that β_1 hat is a linear combination of the response variables, Yi. Under the normality assumption, β_1 hat is thus a linear combination of normally distributed random variables, and thus β_1 hat is also normally distributed. Along with the expectation from the previous slide as well as a similar derivation for the variance of the estimator, now we get the distribution for β_1 hat.



However, the sampling distribution of β_1 hat is not useful because sigma squared or the variance of the error terms 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 squared error,** or the sum of squared residuals divided by N - 2.

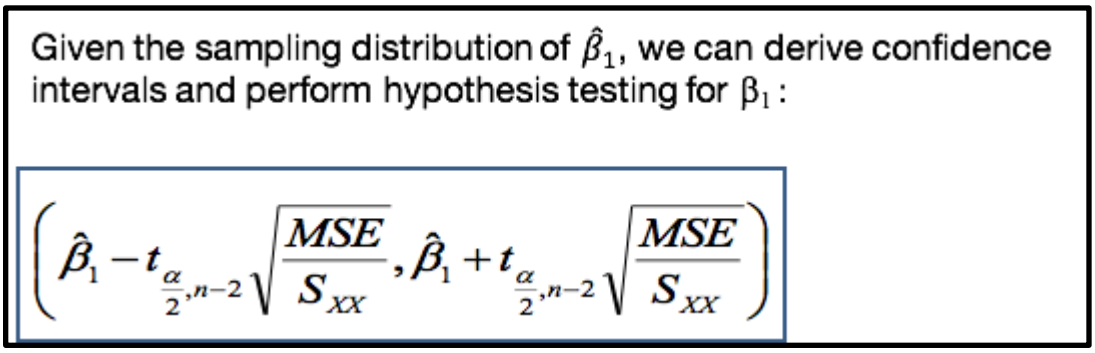
Because this estimator has a chi-square distribution with N - 2 degrees of freedom, the sampling distribution of β_1 hat becomes a t distribution with N - 2 degrees of freedom. The N - 2 degrees of freedom come from the fact that the distribution of the variance of the estimator variance is a chi-square distribution with N - 2 degrees of freedom.



We will use this sampling distribution to derive confidence intervals, and also to perform hypothesis testing with respect to β_1.

Slide 5:

This is the confidence interval for β_1 based on the normality assumption:



Slide 6:

Let's digest a bit this formula. Again, the sampling distribution of the estimator β_1 hat is a T distribution.

$$\frac{\widehat{\beta_1}-\beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2} \qquad t - \text{interval for } \beta_1$$

If we want to obtain a confidence interval with (1-alpha)% confidence level, then we can center the confidence interval at the estimated value for β_1, plus or minus the (1-alpha **critical point** T. This critical point comes from the fact that the sampling distribution of β_1 hat is a T distribution with N – 2 degrees of freedom. The critical point also incorporates information about the confidence level, specifically it is the alpha/2 critical point for a 1-alpha confidence interval. Last, to account for the variability of the estimator, we multiply this critical point with the standard deviation of β_1 or the squared root of the estimated variance of β_1 hat provided in the previous slide.

$$1\text{-}\alpha \text{ Confidence interval} \rightarrow \underbrace{\widehat{\beta_1}}_{\text{Estimate of } \beta_1} \pm \underbrace{t_{\frac{\alpha}{2},n-2}}_{\text{t-critical point}} \underbrace{\sqrt{\frac{MSE}{S_{xx}}}}_{\text{Standard Deviation of } \widehat{\beta_1}}$$

Slide 7:

If we want to perform hypothesis testing on β_1, for example, test whether β_1 is equal to zero versus the alternative that β_1 is not equal to zero, we again use statistical inference. The hypothesis testing 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, the estimated value for β_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 β_1 is equal to zero. If the null hypothesis is rejected, we interpret this that β_1 is statistically significant.

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

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

We reject H_o if |t-value| is large. If the null hypothesis is rejected, we interpret this as β_1 being **statistically significant**.

Statistical significance means that β1 is statistically different from zero. We will use this concept throughout the entire course.

Slide 8:

But what if we want to change the procedure to test whether β1 is equal to a constant versus β1 not equal to that constant, where that constant may be a different value from 0?

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

Remember, this is the T value. We can replace zero from the previous t- statistic with C, depending on the specified null value in the null hypothesis. If the T value is larger than its critical point in absolute value, we say that the slope coefficient is statistically significantly different from c.

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

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

We can also make this decision based on a P value, which is going to be the sum of the tails of the distribution of β_1 hat 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 new hypothesis.

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

If P-value small (p-value<0.01)

Slide 9:

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 β1 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 whether β_1 is greater or less than zero as the null hypothesis. In this case the P value will change in the sense that we're interested in only one of the tail. For β_1 greater than zero we're interested on the right tail of the distribution of the β_1 hat. For β_1 smaller than zero we're interested on 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 = $P(T_{n-2} > \text{t - value})$

What if we want to test for negative relationship
$H_o: \beta_1 \geq 0$ versus $H_A: \beta_1 < 0$?
P-value = $P(T_{n-2} < \text{t - value})$

Slide 10:

The inference for the intercept parameter is going to be similar to the inference for the slope parameter. β_0 hat is also linear combination of random variables, a linear combination of Y's.

We can derive similarly that the expectation of this estimator is equal to the true parameter, thus β_0 hat is an unbiased estimator of β_0, and the variance of this estimator is on the slide.

With this information, and with the fact that β_0 hat is a linear combination of normally distributed, the sample distribution is also a T distribution. Thus the confidence intervals is going to be very similar to the confidence interval for $\beta 1$ hat. It's center of the estimator β_0 plus or minus the critical point from the T distribution times the standard error.

Confidence interval:

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

1.5. Statistical Inference Data Examples

The topic of this lesson is illustrating statistical inference for simple linear regression with a data example using the R statistical software. We're going to use the output from the regression fit to make statistical inferences on the estimated coefficients.

Slide 3:

We are returning now to the example in which we are interested in the relationship between advertising expenditure and sales.

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

Slide 4:

To perform statistical inference, we need to find:

- the estimated coefficient β_1 and its variance along with the sample distribution of β_1.
- the estimated coefficient for the intercept β_0 and its variance along with the sample distribution
- whether the coefficient β_1 is statistically significant.
- whether β_1 is statistically positive.

We'll also learn how to derive a 99% confidence interval for β1. I will conclude with an interpretation of the p-value in the general context in a general hypothesis testing procedure context.

Slide 5:

This is the output of the regression model (from a previous lesson) is on the slide. 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 of the regression coefficients.

The estimated value for β_1 is (2.7721)

If we want the estimated variance (of β_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

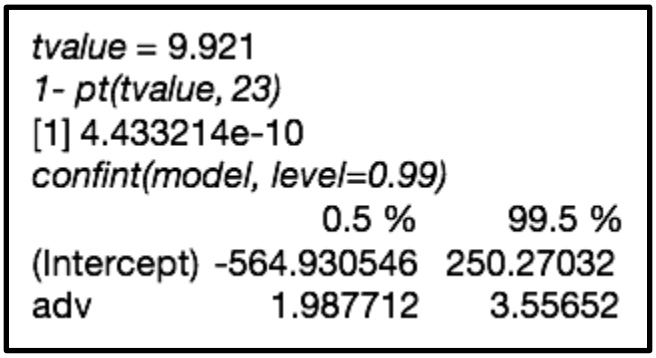
that in order to get the variance. To remember, the R output provides standard deviances, not the variance. That is the square root of the variance not the variance itself.

If we want to test whether β_1 is 0, we can use the p-value provided by R, which approximately equal to 0, meaning that we reject the null hypothesis that β_1 is equal to zero, and conclude that β_1 is statistically significant.

Slide 6:

To test whether β_1 is statistically positive, now we have to change the alternative hypothesis to β_1 greater than zero.

For this we can use the output but we need to adjust the p-value. The difference between the p-value of the two sided test (with alternative that the coefficient is non-zero) vs the one-sided test with the alternative that the coefficient is positive is that the p-value will only be based on both the right and left tails for the former and only on the right tail of the latter.



We can compute the p-value of the one-sided test 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 quantile equal to the t-value. In order to get the right tail of that distribution we'd have to take one minus that probability. That p-value is again very small, leading us to conclude that β_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).

The lowest value of the 99% confidence interval is 1.9 and the highest is 3.5. The interpretation of a confidence interval is tricky. The interpretation is in a frequentist sense; specifically, if we have a 99% confidence interval it means that one out of 100 times, the interval may miss including β_1, i.e. β_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#.Wlka6FQ-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.

1.6 Regression Line: Estimation & Prediction

The topic of this lesson is simple linear regression, and I'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.*

Slide 3:

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 due to the uncertainty level of the predicted 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 the newness of the observation.

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

Let's say we're interested on the mean response evaluated at this predicting value x*. Under estimation, x* can be one of the observed values that we fitted 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*. Thus it's really an average across all possible settings when we could observe x*.

Prediction at this value x* is when x* is considered an 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*. While in estimation, we're averaging across all *possible* settings for prediction, in prediction, we focus on one particular setting.

Slide 4:

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 of the regression line. Because the estimators of those two coefficients, the slope and the intercept, are normally distributed, so is y-hat.

Slide 5:

Y-hat has a normal distribution and the expectation of y-hat is very easy to derive: it's equal to β_0 plus $\beta_1 x^*$, so this is actually the true regression line; thus the estimated regression line is an unbiased estimator just like the estimators of the regression coefficients. The variance is as on the slide.

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

$$E(\hat{y} \mid x^*) = \beta_0 + \beta_1 x^*$$

$$Var(\hat{y} \mid x^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right)$$

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 $x^* = \bar{x}$

This means that 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 away from the average.

Slide 6:

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 it at the estimated regression line, plus or minus the t-critical point times the standard deviation of the estimated regression line. Similarly, because of the variance changes with x*, the confidence interval also will be wider the further x* is from the mean (x-bar):

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

- ✓ 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 this course).

Slide 7:

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 are predicting the response under a new setting.

Slide 8:

So how does that translate into uncertainty of the prediction of a new response? The variation due to the estimation is similar to the variation of the estimated regression line. But now we adding the variation due to new measurement, which is sigma squared.

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: σ^2

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

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

Practically, the difference is that 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 squared divided by Sxx. As a reminder, the formula for Sxx is the sum of the differences between the X's and the average squared differences.

Slide 9:

If we want to obtain a prediction interval for a future value y-bar, 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-α)% ***prediction*** interval for a future y* (at x*) is

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

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

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

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

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*.

1.7 Regression Line: Estimation & Prediction Examples

This lesson covers the implementation of estimation and prediction of the regression line using the R statistical software.

Slide 3:

We are returning 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.

Slide 4:

This is a set of questions that we will address in this lesson:

- 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 intervals differ for the two confidence levels?

Slides 5-6:

Let's go back to the summary of the model. This is what we've seen 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 units. Remember the units for the advertising expenditure is $100. Thus what we input in the regression line is 300 units, 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 values for SXX, X*, and the average of the X. The estimated variance is the square of the residual standard error from the R output, which is equal to the mean square error **(mse=101.4^2)**. The sample size is the degrees of freedom which is 23 plus two (the lost DFs due to the estimation of the two regression coefficients) thus **(n=23+2)**. Remember that the degrees of freedom are equal to n - 2.
To get the value for SXX, we will use the formula from a previous lesson, the formula for the variance of the estimator for β_1. Recall that variance equals sigma square divided by S_{xx} or Sxx = MSE/ V(β_1). So I can use that formula to get S_{XX}, which is going to be the estimated variance of the error terms or MSE divided by the variance for β_1. Both values in this formula are available in the output.

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

Now you plug in all these values into the formula for the variance of the predicted sales, and the value of the variance is 14286.26.

b. The variance of the predicted sales is

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

Slides 7-8:

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 derive confidence intervals and prediction intervals for a linear model, we will use the function predictLM().

We first need to set the new data, in this case, the advertising expenditure equals to 300 units, or $30,000. We also need to specify whether we want a prediction interval corresponding to prediction or a confidence interval corresponding to estimation. Please

note that while R differentiates between confidence intervals and prediction intervals, in statistics, we refer to confidence intervals in both the estimation and prediction contexts. Thus, if you are asked to derive a confidence interval, you will need to first differentiate whether you are deriving under estimation or prediction setting, then use the appropriate inputs.

The predict() R function also takes as the input the confidence level of the confidence interval, the default being 95%.
Remember, we are interested in comparing 99% and 95%. So this will be the two intervals, the 99% and 95%, that I provided with these lines of codes. If we want to derive confidence intervals for estimation, the only thing that's going to change is to specify the type of interval as “confidence”.

Let's compare the confidence intervals under estimation versus prediction. You can see that the prediction interval is significantly wider. This is because we have additional uncertainty due to predicting under a new setting whereas the confidence intervals under estimation 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.

Note again that we need to be careful about whether we want a confidence interval estimation vs prediction. **Just to wrap up the comparison, the confidence intervals under estimation are *narrower* than the prediction intervals because the prediction intervals have additional variance from the variation of a new measurement**. We also interpret those intervals differently. The prediction intervals are for one specific setting, whereas confidence intervals are average across settings that have the same value for the predicting variable.

1.8 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.

Slide 3:

Let's go back to the model framework for simple linear regression. The data consist of bivariate data of a 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 and 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.

Slide 4:

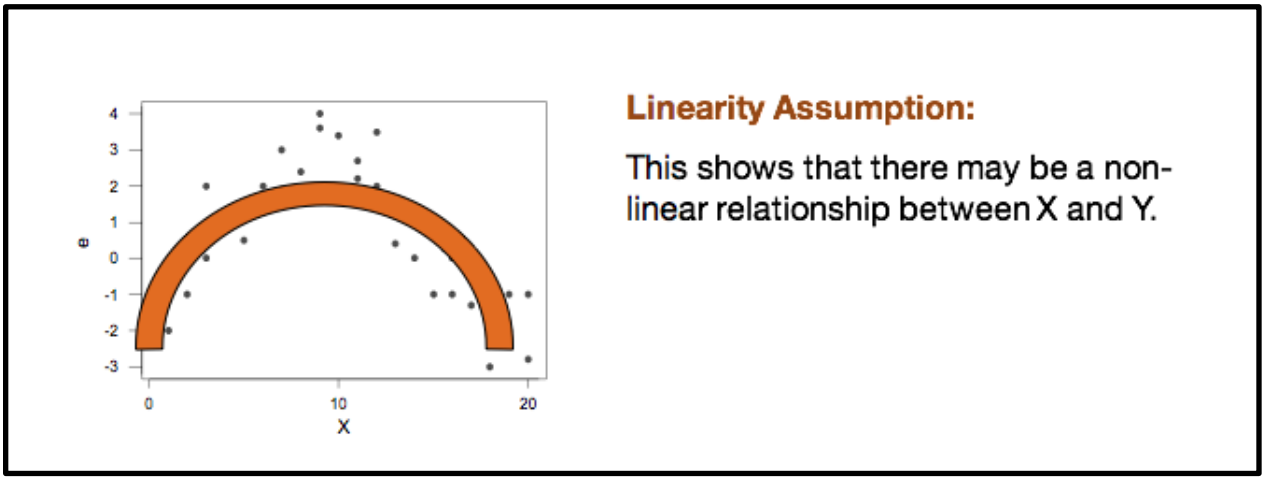
The approach for diagnosing this assumption is to evaluate the **residuals.** We're not going to evaluate the assumptions of the error terms directly because we do not know β_0 and β_1 and thus we don't have the error terms. Instead, we evaluate the assumptions on the residuals, which are the differences between the observed responses and the fitted responses.

Specifically, we plot the 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.

Slide 5:

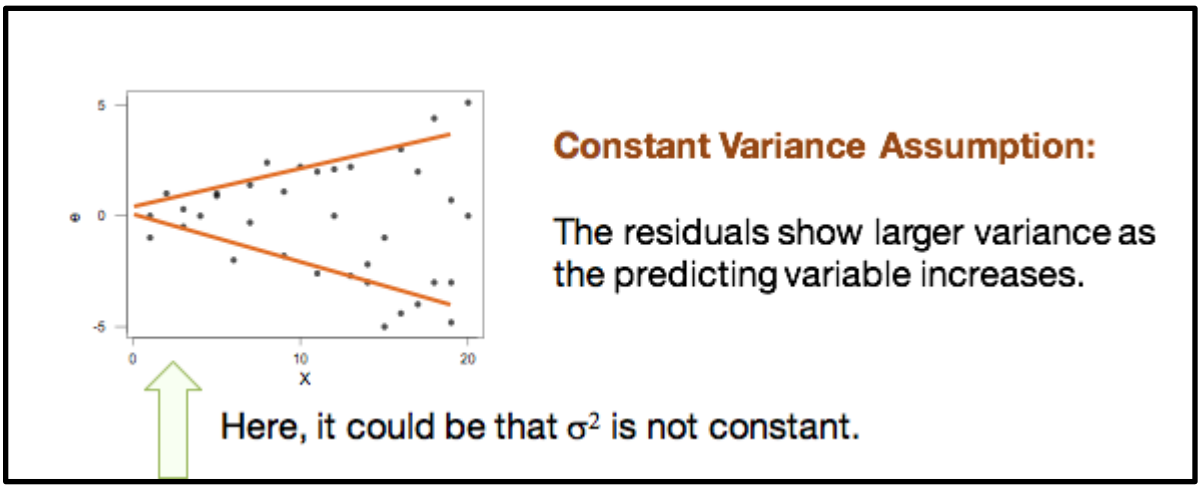
This is an example with residuals plotted against the predicting values:



As you see, there is a curvature between the relationship between residuals and the predicting variable showing their relationship is not linear. It's a departure from the linearity assumption.

Slide 6:

A second example is this plot on the slide, showing a megaphone effect on the residuals, in the sense that the residuals increase with increasing fitted values, which means that the constant variance assumption does not hold.

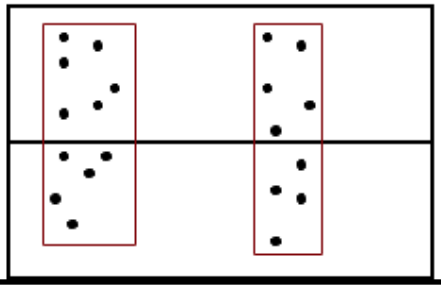


Slide 7:

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

Independence Assumption:

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



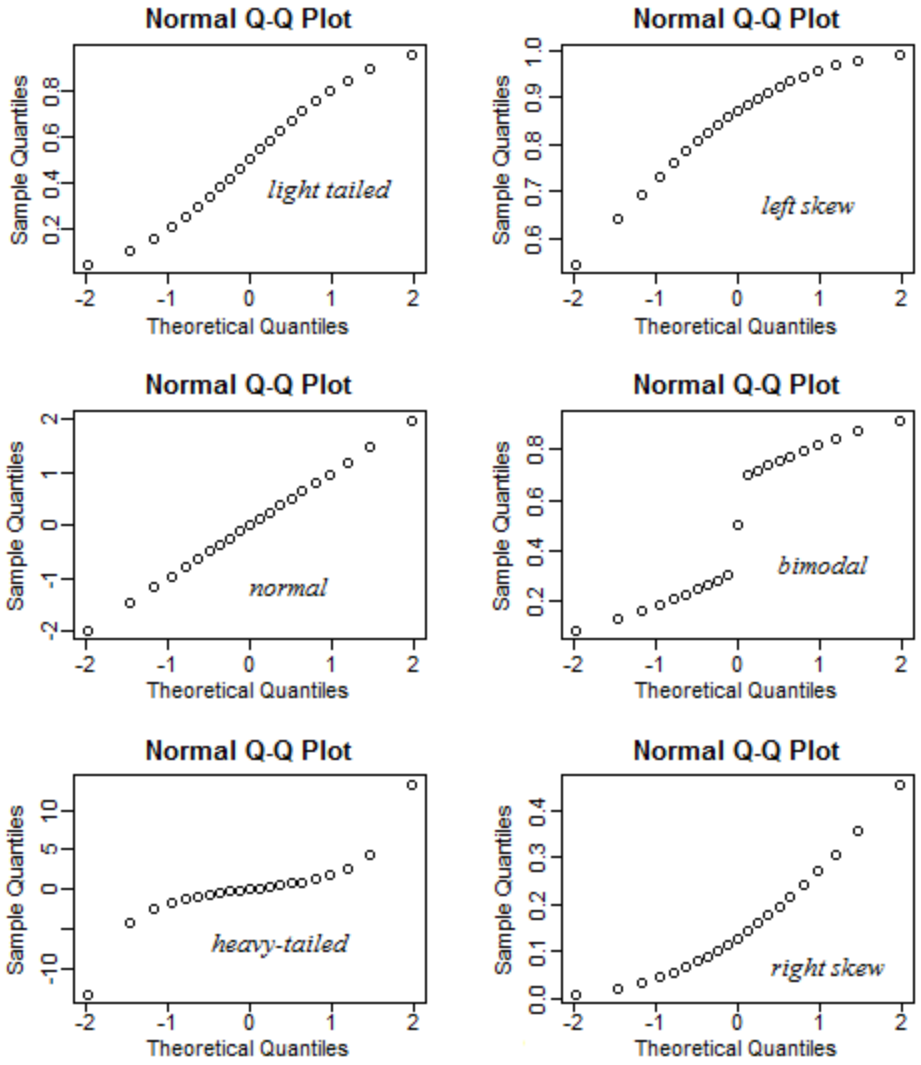
You can see here that the residuals now are clustered in two separated clusters which means that the residuals may be correlated due to some clustering effect, for example proximity in geography where the observed responses may have been observed.

Keep in mind that **residual analysis cannot be used to check for the independence assumption.** Recall, the assumption is independence, not uncorrelated errors. But all we can assess with the 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 be used to assess uncorrelated errors, not independent errors.

Slide 8:

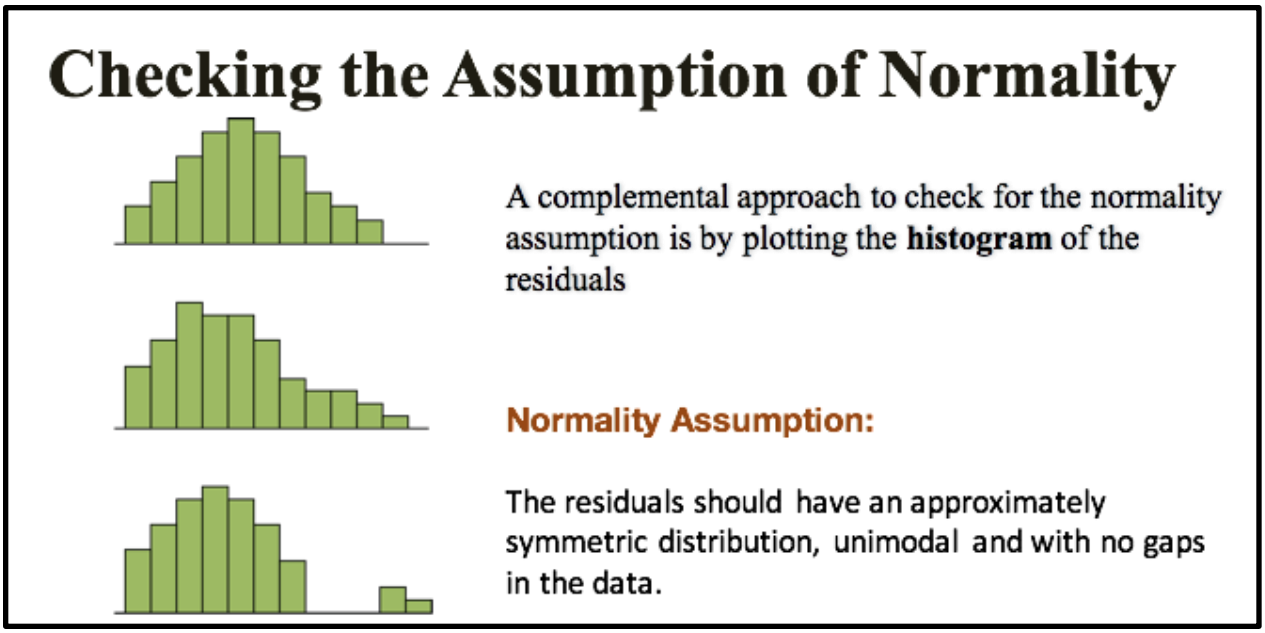
For checking normality, we can use the **quantile plot, or normal probability plot**, on which data are plotted against a theoretical normal distribution in such a way that the points should form a straight line. The x-axis of the normal probability plot is formed by the normal or statistic medians, and the y-axis is the ordered 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 quintiles of the normal distribution. If the residuals are normal, then the quantiles of the residuals will line up with the normal quantiles, thus we should expect that they follow a straight line. Departure from a straight line could be in the form of a tail, which is an indication of either a skewed distribution, or heavy-tail distribution. Do not attempt to do your own implementation of this plot—use a 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.)



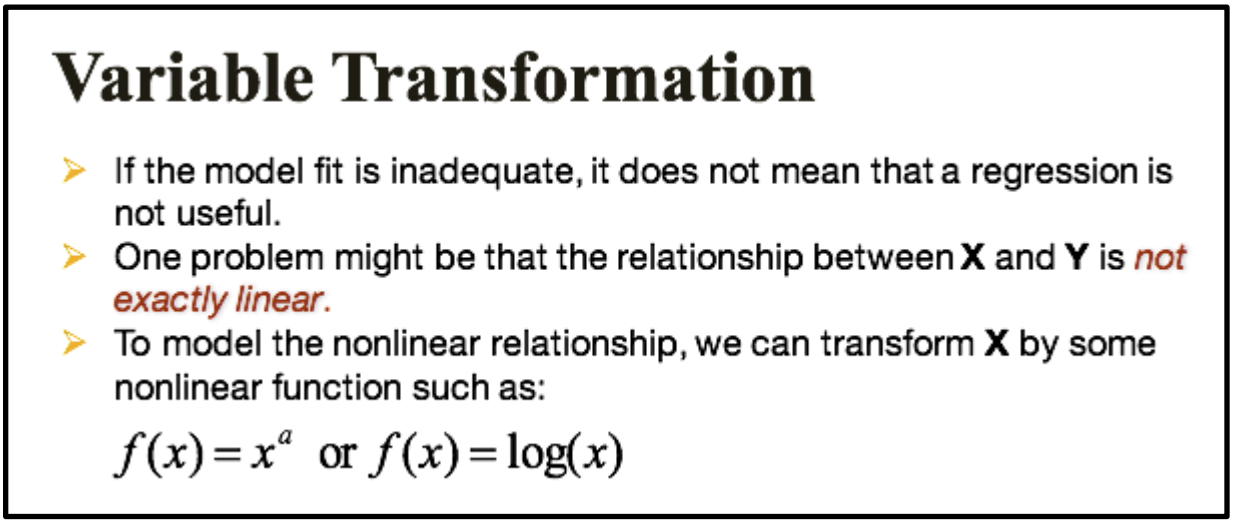
Slide 9:

Another approach to check for the normality is using the histogram plot. 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. Examples of departures from the normality assumption is if we see skewedness in the shape of the distribution, modality, that is, 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.



Slide 10:

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 in simple linear regression, 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.



Slide 11:

What if the normality or constant variance assumption does not hold? Often we use a transformation that normalizes or variance-stablizes the response variable.

That common 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.

after transforming Y, you will need to fit the model again and evaluate the residuals for departures from assumptions. If the transformation(s) did not address these departures from assumptions, you will need to consider other transformations. If you cannot identify the appropriate transformation, it may that you will need to consider a different modeling approach, some illustrations are discussed in the last unit (also an optional unit) of this course.

1.9. Outliers and Model Evaluation

In this lesson, I will continue by considering other aspects of model fit and also approaches to evaluate the model performance.

Slide 3:

An important aspect in regression is the presence of **outliers**, **which are data points far from the majority of the data in x and/or y**. 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 the x's and/or the y's is called **influential point** if it influences the regression model fit significantly. They can change the value of the estimated parameters, the statistical significance, the magnitude the estimated parameters, or even the sign. It is important to note that an outlier, including a leverage point, may or may not impact the regression fit significantly, thus it may or may not be an influential point.

It is tempting to just discard outliers. But sometimes the outliers belong to 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 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 how an outlier influences the regression fit.

Slide 4:

To check outliers, a very simple approach is to use the standardized residuals, and then compare the standardized residuals to the -2 and 2 band or even tighter, the -1 and 1 band. Statistical packages usually compute the standardized residuals and or point to outliers. (We'll learn about other ways to evaluate outliers in a different lesson in Unit 3 when I introduce multiple linear regression.)

Slide 5:

Once we establish the goodness of fit of the model by evaluating the model assumption, we also want to see also whether the linear model is useful to predict. One approach to quantify the predictive power is using the coefficient of determination, a statistic that efficiently summarizes how well the predicting variable x can be used to linearly predict the response variable. This is the so-called R-square, which is 1 minus the ratio between the sum of squared errors and sum of square total.

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.

Slide 6:

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.** One could also use the coefficient correlation to evaluate various transformations of X and Y to improve the linearity assumption in the simple linear regression. Relevant in the context of evaluating the explanatory power is how the correlation coefficient relates to R-squared in simple linear regression. The relationship is that the square of the correlation coefficients is actually the R squared.

1.10. Diagnostics and Model Evaluation Examples

In this lesson, 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.

Slide 3:

We'll return now to the example in which 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?

Slide 4:

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

Slide 5:

To review, the assumptions are **linearity**, **constant variance**, **independence** and **normality**. Do the assumptions hold? One way to evaluate the linearity is using the scatterplot of the predicting variable versus sales. This is how to plot the scatter plot of the predictive variable (advertising) versus the response variable (sales) in R using the 'plot' R command – this gives us the first plot. To evaluate the constant variance and the assumption of uncorrelated errors, we can use a scatter plot of the residuals vs fitted values, which is the second plot. To evaluate the normality we can use the normal probability plot. There are multiple ways you can display the normality plot; here I am using an R command from the 'car' library. The resulting plot is the third one.

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

Slide 6:

Do we identify any outliers? We can also identify outliers using the plots from previous slide -- 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.

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 from the model output using the command: 'summary(model)$r.squared'.

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.

1.11 Data Example: 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.

Slide 3:

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.

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

Slide 4:

The average annual change in exchange rate is the response variable expressed as US dollar per unit of a country's currency. It is calculated as a difference in natural logarithm divided by the number of years and multiplied by 100, to create percentage change as shown on the slide. 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 US versus another country as shown on this slide. 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.

Slide 5:

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

We'll first read the data with the **read.table** command in R where the input is the data file called 'ppp.dat' for this data asset. We also specify the separator of the data values in the file as a tab delimiter, we specify that the data files has a header, specifically, that the columns have names. You should check whether the data matrix coincides with

the data you read from the file. We can find how many columns and rows are in a data using the **dim** command in R. For this example, we have 40 rows, which means we have data for 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. I'm adding now these data to the initial data matrix by using the 'rbind' command, and I'm converting now the matrix into a data frame. Last I attached these data using 'attach' command in order for the columns in the data matrix to be recognized by R as individual vectors.

I will relabel also the column corresponding to type of the country using their initial denomination. 1 indicates "developed" and 0 indicates "developing." The reason I'm doing that is because when we do exploratory analysis, is better to refer to the names of those categories, the denomination of those categories.

Slide 6:

Next, we'll perform exploratory data analysis for this dataset. One visual approach is by plotting x vs y. In this case, x is inflation difference, and y is the 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. We also can evaluate how would exchange rate varies with the categorical (qualitative) 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. The two plots are in the next slide.

Slide 7:

What you see from the scatter plot of inflation difference against the exchange rate change is that there is a clear linear relationship between those two. From the box plot, we can see that there is a significant difference in exchange rate change amount between the two types of countries developed and developing.

Slide 8:

Let's ignore for now the presence of the outlier and the fact that there are differences in the response variable across developed and developing countries as we've learned from the boxplot. We're now going to perform the simple linear 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. Part of the output is on the slide.

Let's dissect the part of the output that provide information about the coefficients.

The estimated β_0 is -1.5193 and its standard error is 0.2941. The intercept says that that if the analyzed difference in inflation between a country and the U.S. is zero, then 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 estimated β_1 slope is 0.961 and the standard error is 0.0178. The slope coefficient says that a 1% point change in analyzed difference in inflation rate is associated with the estimated expected value of 0.962% of point change in exchange rates.

The other portion of the summary output that is of interest is the residual standard error and multiple R squared. This summary gives estimated standard deviation of the estimated sigma. The number of degrees of freedom is 39, thus n is 41. Also, the R squared score is high meaning 98.7% of the variability 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 to 2% points roughly 95% of the time.

Slide 9:

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.

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 PPP theory. We'll need to use statistical inference such as hypothesis testing.

Testing for β_0 equal to zero means that testing for statistical significance. From previous slide we find that β_0 is statistically different from zero because the p-value provided in the output is approximately equal to 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 β_1 is equal to one is a slightly different test than the test for statistical significance since the null 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 and for this has a t-value of 2.148. I computed the p-value similarly as for the test with the null value equal to zero; we take 2 times 1 minus the left tail of the t distribution with 39 degrees of freedom, for the t-value equal to 2.14; with that the p value 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 a the significance level 0.05 but we reject it at the level 0.01.

Thus, we see violations of the purchasing power parity with respect to both the intercept and the slope.

Slide 10:

Let's review again how we can perform the hypothesis testing procedure for β equal to 1 using 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.

We are using the command **linearHypothesis**. The input in this command is the model called PPPA along with the vector of null values for the regression coefficients, c(0,1), corresponding to a null value equal to 0 for the intercept and a null value equal to 1 for the slope coefficient. Last, the option *rhs=1* specifies the alternative hypothesis.

We can also the compute the t-value directly by computing the t.value as the β_1 hat minus one, divided by the standard deviation of β_1 hat. We can compute the p-value as the formula I provided in the previous slide, which is 2 x (1- pt(tvalue,39), providing the probability of a t-value distribution where we input the t-value and the degrees of freedom.

If you would like to learn more about how to use the pt() command as well as other commands, you can use the help() command, which will take you a help page describing the command.

1.12. 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.

Slide 3:

To get confidence or prediction bands around the estimated regression line you can use the R function *regplot.confbands.fun* provided in the R code for this data example. This R function is an example of you can write your own functions in R. 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. Similar to this example, I give the function a name to refer to it later along with the input parameters, for example X and Y, which are the predicting and the response variables in this case. Other inputs in this function are the confidence level for the confidence band, then we can use this same R function to produce the simple regression scatter plot along with the confidence bands. This figure illustrates the output of this R function for the relationship between inflation difference and exchange rate change across the 41 countries.

The fitted line plot shows several lines, the continuous line is the fitted regression line, the wider interrupted line band is the prediction confidence band and the narrower interrupted line band is the confidence band. The circles correspond to outliers.

First, the confidence band represented by the inner pair of the lines is much narrower than the prediction band represented by the outer pair of lines. While a 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, hence wider. 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. This type of analysis allows us to identify potential outliers.

Slide 4:

To compute confidence and prediction intervals for new observations, we use the **predict()** command in R. Using this command, we need to specify the value or values we want to predict and create separately a data frame of the data used for prediction. 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.

In this example, we are estimating the prediction and confidence intervals for a data point corresponding to -0.68 in inflation difference, which is roughly the value for Norway. For that I created a new data frame and then input this into the predict() command. Estimating both intervals we see that the two intervals are different; 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? The confidence interval provides our estimate 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 estimate for what the exchange rate change will be for one country with inflation difference equal to -0.68. The first interval includes only negative values whereas the second one includes both negative and positive values and is much wider. The prediction interval includes the zero value.

Slide 5:

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.

Here is some R code example for four residual analysis 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 scatterplot of the fitted values versus the residuals. The third one is the normal probability plot along with the histogram in the fourth plot. 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.

Slide 6:

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. For this example, there's no specific pattern, thus the linearity assumption holds. We do identify one outlier, again which corresponds to Brazil.

The second plot can be used to evaluate the assumptions of constant variance and 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, meaning that the constant variance does not hold. We do not see a grouping of the residuals, meaning that the assumption of uncorrelated error possibly holds.

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 in the histogram of the residuals.

Slide 7:

Let's discuss the outlier that we've noted in the two graphs, specifically, the isolated point in residual plot corresponding to Brazil. Observations for which the predicting variable is away from that range is called a leverage point. 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 results of the regression, so we cannot simply ignore it. We can remove it from the data, and analyze the data without it, while informing the reader about the removal of the outlier and its implications. That is, we can present the 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.

Slide 8:

Next we perform a linear regression for the data without Brazil. 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 R output 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 aprox 0.

- Testing the null hypothesis that the slope 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).

Thus 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.

Slide 9:

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.

Slide 10:

These are the four plots.

Similarly to the model with Brazil, the linearity assumption holds. There is no grouping in the residuals. However, the constant variance assumption still does not hold. A different outlier appears, 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 in terms of the purchasing power parity theory.

In fact, you can perform the analysis separately for developed and developing countries. Since Brazil and Indonesia are both developing countries, the question is, will the purchasing power parity theory hold differently for the two sets of countries?

Slide 11:

Those are the findings of this example. 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.*

1.13. 2000 Elections in Florida

The topic of this last lesson in Unit 1 is the application of simple linear regression to look into 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.

Slide 3:

In the presidential elections in 2000, during the election night, the electoral votes for the two candidates, George W Bush and Al Gore, 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 votes 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. Particularly we're going to look at one county, the Palm Beach County, where the number of votes for Buchanan was very large.

Slide 4:

A first step in a data analysis using R is to read the data file in R, and in this case we're using the read.table() command in R. To use 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; here, we're checking the first four rows.

In this example, you can see that the data consist of many more factors, but we will only look at the vote counts for Bush and Buchanan.

Slide 5:

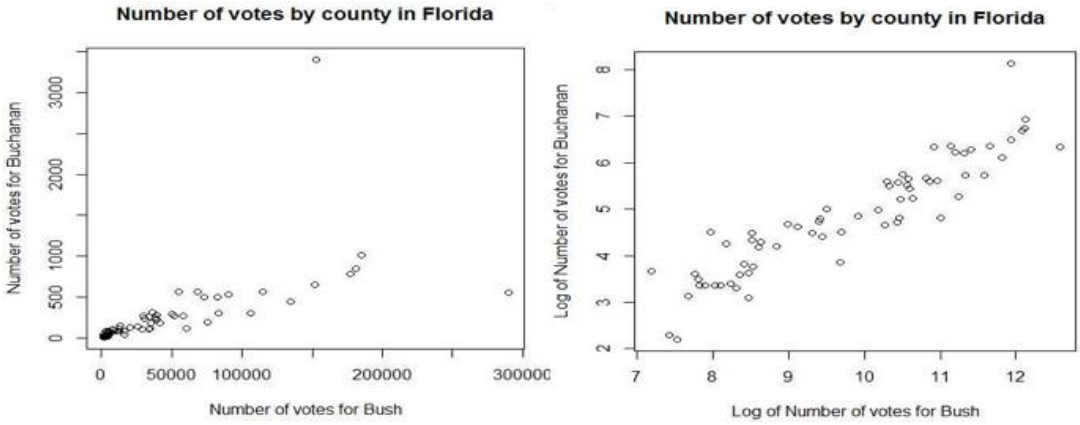
In this analysis, we're interested in the vote counts for Buchanan and for Bush; we extract those factors from the data matrix that we run in R. Next, we can use again the plot() R command for the scatter plot of two variables, in this case, the vote counts for Bush and the votes count for Buchanan. The scatter plot of those two variables is on the slide.

We can identify two specific outliers, one corresponding to the Palm Beach county, but we also can note that the relationship between the number of votes for Bush and number of votes for Buchanan is a curvature, thus not a linear relationship. Thus we'll

have to perform some transformations on X or/and Y in order to fix this nonlinearity between the two factors.

Slide 6-7:

Here we compare the scatterplots of the number of votes for Bush versus the number of votes for Buchanan (left), versus the log of the number of votes for Bush and log of number of votes for Buchanan (right). We can see that one of the outliers doesn't seem to be an outlier anymore. However, Palm Beach which is a large value in the votes for Buchanan is still present.



We can also see that the linearity assumption has improved significantly. With these transformations, the correlation has increased from 0.625 to 0.922. 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. Here you can play with multiple transformations of the two factors and see which one most improve or increases the correlation coefficient.

Slide 8:

This is the model output from the regression of the log-count of votes for Bush onto the log-count of votes for Buchanan.

The slope coefficient is 0.756 and it is statistically significantly different from zero because the pvalue of the t-test for slope coefficient is approximately zero.

Moreover, the model performance in terms of capturing the linear relationship between the response and predicting variables is high with an R-squared of 0.85, meaning that 85% of the variability in the log-count of votes for Buchanan is explained by the log-count of votes for Bush, thus the assertation I made earlier the votes for the two candidates would be similar is supported.

Slide 9:

Next, we'll perform the regression analysis using the *transformed data*. The residual analysis is similar as before.

These are the resulting four plots. Based on this residual analysis, we learn that the assumption of constant variance holds because we don't see a change in the variability of the residuals. We also can assess the linear assumption using first plot, and because there's no pattern, we conclude that the linearity assumption holds also. There is also not a clustering among the residuals, indicating that the assumption of uncorrelated errors holds. For normality, we can use the bottom plots to evaluate normality and while the QQ plot looks reasonably well, that is the quantiles of the residuals line up with the quantiles of the normal distribution, the histogram tells us the residuals have a skewed distribution.

Slide 10:

We can further interpret 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. This command will give you the confidence intervals for both the intercept and the slope.

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.

Slide 11:

Is Palm Beach an outlier?

In order to evaluate whether the vote counts for Buchanan in Palm Beach county 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. Thus we remove the 50th observation from both the number votes for Buchanan and for Bush. Those are the estimated coefficients, along with statistical inference for the regression coefficients from the summary output.

Now we're going to predict the vote counts for Palm Beach County for Buchanan based on the model that omitted the number of votes for Palm Beach. Then we're comparing what we predict with what we observed. Because the predicted value is on the log-

scale, we need to take the exponential of the predicted value to compare it to the observed value. The difference 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.

```
## 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
```

We can also look at the prediction intervals for the number of votes for Buchanan in Palm Beach. 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 transform the lower and upper bounds of the prediction interval back to the original scale by taking the exponential function. From the output, the lower bound of the number of votes for Buchanan is 252 and the upper bound is 1412. We compare the lower and upper bound with what we observed, the observed vote count for Buchanan in Palm Beach, which is 3,407. The observed value is much, much larger than even the upper bound of the prediction interval, suggesting that this is an outlier.

Slide 12:

Interpreting the results, we find that that 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 the 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 an analysis of this kind, of even a simple linear regression, could have overturned elections in 2000.