

# Regression Analysis

## Simple Linear Regression

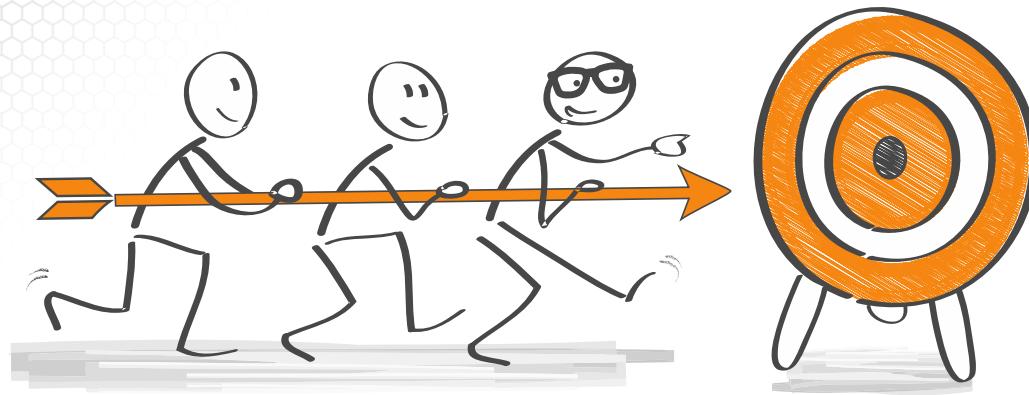
**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Regression Concepts: Basics***

# About this lesson



# Example 1

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

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

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

# Data Example 1

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

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

# Example 2

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

# Data Example 2

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

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

# Example 3

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

# Variables in Regression

The regression framework is characterized by the following:

1. We have one particular variable that we are interested in understanding or modelling, such as sales of a particular product, or the stock price of a publicly traded firm. This variable is called the *response (dependent) variable*, and is usually represented by Y.
2. We have a set of other variables that we think might be useful in predicting or modelling the response variable (say the price of the product, the competitors price, and so on; or the profits, revenues, financial position of the firm, and so on). These are called the *predicting or explanatory (independent) variables*, and are usually represented by x1, x2, etc.

# Variables in Regression

The regression framework is characterized by the following:

## **RESPONSE VARIABLE versus PREDICTING VARIABLE?**

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

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

called the *predicting* or *explanatory (independent) variables*, and are usually represented by  $x_1, x_2$ , etc.

# Response vs Predicting Variable

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

- **Response Variable:** Change in LDL levels
- **Predicting Variable:** Type of Medication

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

- **Response Variable:** Miles Per Gallon (**MPG**) of Fuel
- **Predicting Variable:** Average Driving Speed

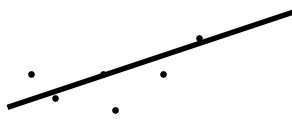
The **relationship** between college grade point average (**GPA**) and scores on the SAT

- **Response Variable:** GPA
- **Predicting Variable:** SAT score

# Linear Regression: General Model

Simple linear regression

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



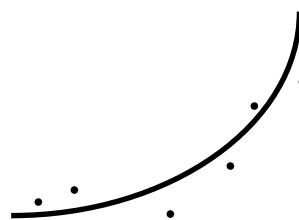
Whether a linear or polynomial model in X, we can estimate the relationship in x using linear regression.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$



Polynomial Regression

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \varepsilon$$



# Regression: Basics

A regression analysis is used for:

1. **Prediction** of the response variable;
2. **Modelling** the relationship between the response variable and the explanatory variables; or
3. **Testing** hypotheses of association relationships.

**Linear Regression:** The basis of what we will be talking about most of this course is the linear model. Virtually all other methods for studying dependence among variables are variations on the idea of linear regression.

“ *All models are wrong, but some are useful.*      George Box ”

“ *Embrace your data, not your models.*      John Tukey ”

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Regression Concepts:  
Estimation***

# Simple Linear Regression: Model

Our goal is to find the best line that describes a linear relationship  $(\beta_0, \beta_1)$  that is, find  $(\beta_0, \beta_1)$  where

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

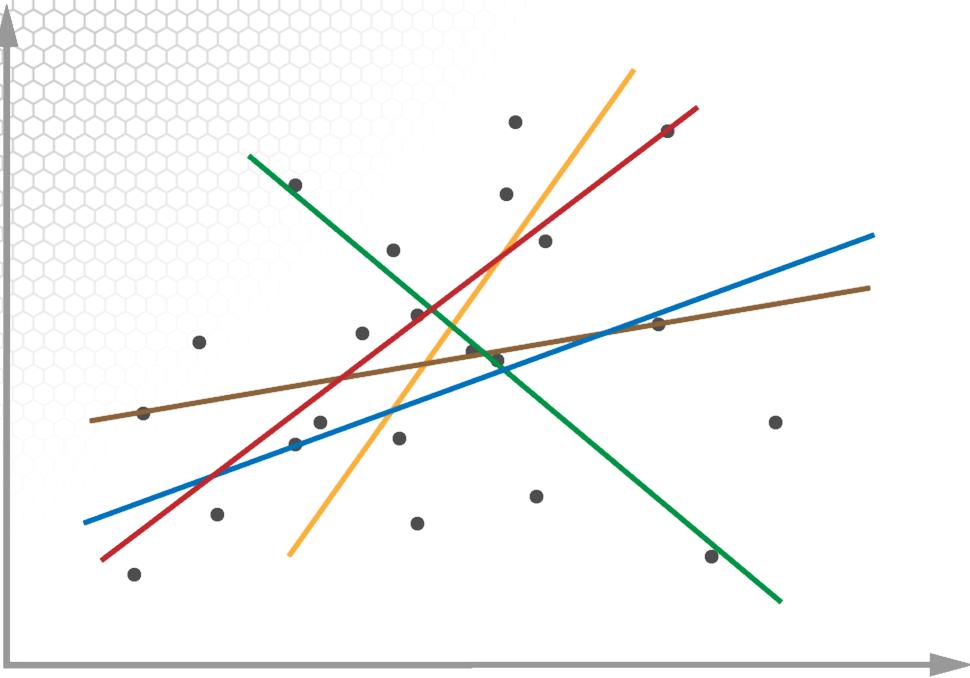
$$Y = \beta_0 + \beta_1 x +$$

Equivalently, estimating:

1.  $\beta_0$  Intercept

2.  $\beta_1$  Slope

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



How to find the best line?

Our goal is to find the line that describes a linear relationship; that is, find  $(\beta_0, \beta_1)$  where

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

# Simple Linear Regression: Model

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

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

**Assumptions:**

- *Linearity/Mean Zero Assumption:*  $E(\epsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\epsilon_i) = \sigma^2$
- *Independence Assumption*  $\{\epsilon_1, \dots, \epsilon_n\}$  are independent random variables
- *(Later we assume  $\epsilon_i \sim \text{Normal}$ )*

# Simple Linear Regression: Model

***The model parameters are:  $\alpha$ ,  $\beta$***

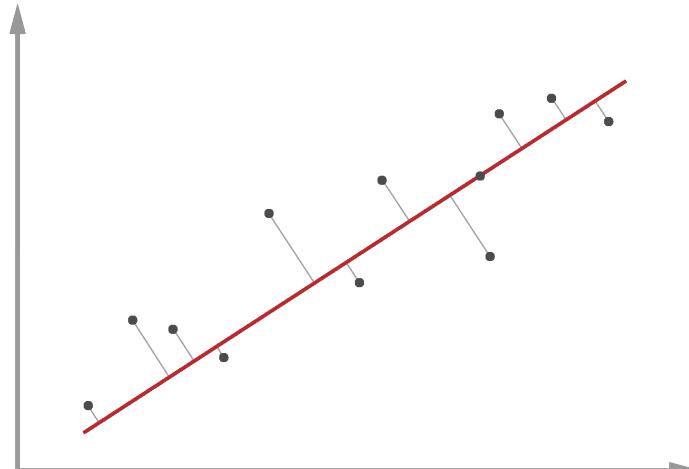
- ***Unknown regardless how much data are observed***
- ***Estimated given the model assumptions***
- ***Estimated based on data***

- *Linearity/Mean Zero Assumption :  $E(\epsilon_i) = 0$*
- *Constant Variance Assumption:  $\text{Var}(\epsilon_i) = \sigma^2$*
- *Independence Assumption  $\{\epsilon_1, \dots, \epsilon_n\}$  are independent random variables*
- *(Later we assume  $\epsilon_i \sim \text{Normal}$ )*

# Model Estimation: Approach

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

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



# Model Estimation: Approach

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

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

=

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \equiv \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Begin with the minimization problem:

=

=

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

- Result into a system of linear equation in and
- Solve using linear algebra
- Solutions to the system are and

# Fitted values and Residuals

Given the estimates of  $\beta_0$  and  $\beta_1$ , we define:

- *Fitted values:*  $\hat{y}_i = \beta_0 + \beta_1 x_i$
- *Residuals:*  $r_i = y_i - \hat{y}_i$
- *Mean squared error:* Estimator for  $\sigma^2$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n r_i^2$$

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

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



Estimating  $\sigma^2$   
variance

Sample

What is the sample variance estimation?

**Basic statistic concept:**

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

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

$$\rightarrow \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Why  $n-1$ ?

We lose a degree of freedom because we replace

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

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

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

↑ Replaced by



We use two degrees of freedom because

Thus, assuming that

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

$$\Rightarrow \hat{\sigma}^2 = \text{MSE} \quad \chi^2_{n-2}$$

(This is called the sampling distribution of )

# Model Parameter Interpretation

Commonly interested in the behavior of

- ✓ A positive value of  $\beta$  is consistent with a direct relationship between  $x$  and  $y$ ; **e.g.**, higher values of height are associated with higher values of weight, or lower values of revenue are associated with lower values of profit;
- ✓ A negative value of  $\beta$  is consistent with an inverse relationship between  $x$  and  $y$ ; **e.g.**, higher price of a product is associated with lower demand, or a lower inflation rate is associated with a higher savings rate;
- ✓ A close to zero value of  $\beta$  means that there is not a significant association between  $x$  and  $y$ .

# Model Estimate Interpretation

The Least Squares estimated coefficients have specific interpretations:

- ✓ is the estimated expected change in the response variable associated with one unit of change in the predicting variable;
- ✓ is the estimated expected value of the response variable when the predicting variable equals zero;

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Regression Concepts:  
Estimation Example***

# Example in R

A company, which sells medical supplies to hospitals, clinics, and doctor's offices, had considered the effectiveness of a new advertising program. Management wants to know if the advertisement is related to sales. This company intends to increase the sales with an effective advertising program.

**Which are the response and the predicting variables?**

**Y = Sales and X = advertisement expenditure**

# Example in R: Estimation

- a. Fit a linear regression. What are the estimated regression coefficients and the estimated regression line?
- b. Interpret the coefficients.
- c. What does the model predict sales as the advertisement expenditure increases for an additional **\$1,000**?
- d. What sales would you predict for an advertisement expenditure of **\$30,000**?
- e. What is the estimate of the error variance?
- f. What could you say about the sales for an advertisement expenditure of **\$100,000**?

# Example in R (cont'd)

```
## 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

Simple linear regression  
 $y = \beta_0 + \beta_1 x_1 + \epsilon$   
**Estimated Model**  
Parameters:  
Multiple linear regression  
 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$   
**= -157.3301**  
Polynomial regression  
 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$   
**= 2.7721**  
**= 101.4**

# Example in R (cont'd)

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

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

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

- b. Interpret the coefficients.

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

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

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

# Example in R (cont'd)

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

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

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

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

each dollar invested in advertisement expenditure.

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

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

# Example in R (cont'd)

- d. What sales would you predict for an advertisement expenditure of \$30,000?

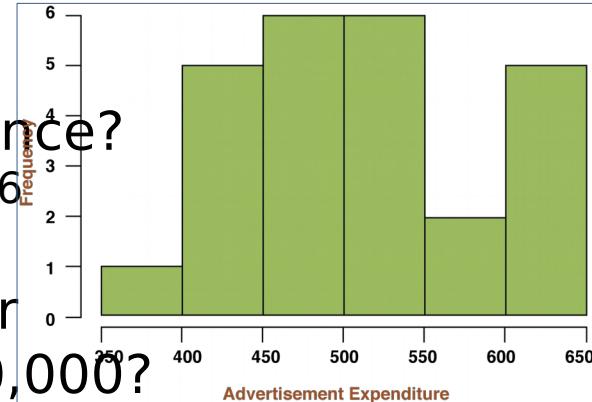
➤ **Solution:** The predicted sales is  $-157.33 + 300 \times 2.77 = 673.67$  thousands

- e. What is the estimate of the error variance?

➤ **Solution:** Estimate  $\sigma^2$  with  $MSE = 10,281.96$

- f. What could you say about the sales for an advertisement expenditure of \$100,000?

➤ **Solution:** An advertisement expenditure of \$100,000 or 1000 units is outside of the observed range and thus we cannot predict the sales since this is extrapolation.



# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

**Regression Concepts:**  
**Statistical Inference**

# Regression Estimators: Properties

For the slope parameter  $\beta_1$ , we can show

$$\text{but } \sum (\text{fixed } \bar{x}) y_i \rightarrow \text{fixed} \quad \text{but } x_i \text{ fixed } \rightarrow \frac{x_i - \bar{x}}{S_{xx}} = c_i \text{ fixed}$$

$$E(\hat{\beta}_1) = \beta_1$$

$$\text{Var}(\hat{\beta}_1) = \sigma^2 / S_{xx}$$

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum_{i=1}^n c_i y_i\right] = \sum_{i=1}^n c_i E[y_i] \\ &= \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \\ &= \beta_1 \Rightarrow E[\hat{\beta}_1] = \beta_1 \end{aligned}$$

||      ||

# Regression Estimators: Properties

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

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

$$\hat{\beta}_1 = \sum_{i=1}^m c_i Y_i$$

linear combination of normally distributed random variables

Simple linear regression  
 $Y = \beta_0 + \beta_1 X + \epsilon$

Multiple linear regression  
 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$

Polynomial Regression  
 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_p X^p + \epsilon$

$\hat{\beta}_1$  Normally distributed

# Regression Estimators: Properties

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

We do not know  $\sigma^2$ . We adapt it by  $MSE$ . Then the sampling distribution becomes the distribution with  $n-2$  df.

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

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

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

# Inference for Slope Parameter

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

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

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

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

1- $\alpha$   
Confidence  
interval

$$\left. \rightarrow \widehat{\beta}_1 \pm \frac{t_{\frac{\alpha}{2}, n-2}}{\sqrt{\frac{MSE}{S_{xx}}}} \right]$$

Estimate of  $\beta_1$  of point

$$\frac{t_{\frac{\alpha}{2}, n-2}}{\sqrt{\frac{MSE}{S_{xx}}}}$$

t-critical  
standard deviation of  $\widehat{\beta}_1$

$$\frac{\widehat{\beta}_{11} - \beta_{11}}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2} \rightarrow t = \text{interval for } \beta_1$$

1- $\alpha$   
Confidence  
interval

→  $\widehat{\beta}_1 \pm \frac{t_{\alpha/2, n-2}}{2}$

Estimate  
of  $\beta_1$   
of point

Sampling distribution  
of  $\widehat{\beta}_1$

$\sqrt{\frac{MSE}{S_{xx}}}$

Standard Deviation/Error  
of  $\widehat{\beta}_1$

$$\sqrt{V[\widehat{\beta}_1]} = \frac{\sigma^2}{S_{xx}}$$

$\sigma^2 \leftarrow \text{MSE}$

# Testing the Overall Regression

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

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

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

How will the procedure change if we test:

$H_0: \beta_1 = c$  vs.  $H_a: \beta_1 \neq c$  for some known  $c$ ?

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

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

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

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

How will the procedure change if we test:

$H_0: \beta_1 = 0$  vs.  $H_a: \beta_1 > 0$

OR

$H_0: \beta_1 = 0$  vs.  $H_a: \beta_1 < 0$ ?

What if we want to test for positive relationship

$H_0: \beta_1 \leq 0$  versus  $H_A: \beta_1 > 0$ ?

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

What if we want to test for negative relationship

$H_0: \beta_1 \geq 0$  versus  $H_A: \beta_1 < 0$ ?

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

# Inference for Intercept Parameter

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



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

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



Confidence interval:

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

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

**Regression Concepts:  
Statistical Inference  
Examples**

# Linear Regression: Example in R

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

Management wants to know if the advertisement is related to sales. This company intends to increase the sales with an effective advertising program.

**What inferences can be made on the regression coefficients?**

# Example in R: Inference

- a. What is the estimate of the coefficient  $\beta_1$  and its variance? What is its sampling distribution?
- b. What is the estimate of the coefficient  $\beta_0$  and its variance?
- c. Is the coefficient  $\beta_1$  statistically significant? What is the p-value of the test. Interpret.
- d. Is the coefficient  $\beta_1$  statistically positive? What is the p-value of the test. Interpret.
- e. Obtain the 99% confidence interval for  $\beta_1$
- f. What is the p-value of a hypothesis testing procedure?

# Example in R (cont'd)

*summary(model)*

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-157.3301	145.1912	-1.084	0.29
adv	2.7721	0.2794	9.921	8.87e-10

Residual standard error: 101.4 on 23 degrees of freedom

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

# Example in R (cont'd)

```
tvalue = 9.921
1- pt(tvalue, 23)
[1] 4.433214e-10
confint(model, level=0.99)
               0.5 %
99.5 %
(Intercept) -564.930546
250.27032
adv           1.987712
3 55652
```

Please read the P-value Statement by the American Statistical Association at:  
<http://amstat.tandfonline.com/doi/abs/>

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

$$\beta_1 > 0$$

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

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

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

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Basic Regression:***  
***Estimating Regression Line***  
***and Prediction***

# Estimation vs. Prediction

Interpretation of estimated mean response:

- If  $x^*$  is one of the observations for the predicting variable, then we use **estimation**. Estimated regression line for the value  $x^*$  is interpreted as the average estimated mean response for all settings under which the predicting variable is equal to  $x^*$ .
- If  $x^*$  is a new observations of the predicting variables, the we use **prediction**. Predicted regression line for the value  $x^*$  is interpreted as the estimated mean response for one setting under which the predicting variable is equal to  $x^*$ .

# Estimating the Regression Line

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

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

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

# Estimating the Regression Line

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

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

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

**If  $x^*$  is away from the range of  $x$ 's, how will be the impact on estimation?**

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

# Confidence Interval for Mean Response

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

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

# Predicting a New Response

One of the primary motivations for regression is to use the regression equation to predict future responses. The prediction is the same as the estimator for the “mean  $\hat{y}$  response”, which is .

But the prediction contains two sources of uncertainty:

1. Due to the new  $(n+1)$ th observation
2. Due to parameter estimates (of  $\beta_0$  and  $\beta_1$ )

# Predicting a New Response

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

The new observation is independent of the regression data, so the total variation in predicting

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

# Predicting a New Response

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

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

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

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Basic Regression:***  
***Estimating Regression Line***  
***and Prediction Example***

# Linear Regression: Example in R

A company, which sells medical supplies to hospitals, clinics, and doctor's offices, had considered the effectiveness of a new advertising program. Management wants to know if the advertisement is related to sales.

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

**What inferences can be made on the prediction of the sales given a targeted advertisement expenditure?**

# Example in R: Estimating Regression Line & Prediction

- a. What sales would you predict for an advertisement expenditure of \$30,000?
- b. What is the variance estimate of the estimated predicted sales for an advertisement expenditure of \$30,000?
- c. What are the lower and upper limits of predicted sales for an advertisement expenditure of \$30,000 at 99% confidence level? How will the limits change if we lower the confidence level at 95%?
- d. Compare the confidence bands of the estimated regression line versus the predicted regression line. Interpret.

# Example in R

```
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

$\bar{x}_{ADV} = mean(ADV)$

$n = 23 + 2$

$mse = 101.4^2$

$var.\beta_1 = 0.2794^2$

$s_{xx} = mse / var.\beta_1$

$pred.var = mse * (1 + 1/n + (x_{bar} - 300)^2)$

$pred.var$

[1] 14286.16

- a. For advertising expenditure of \$30,000, the predicted sales is -  $157.33 + 300 \times 2.77 = 673.67$  thousand.
- b. The variance of the predicted sales is  $\frac{1}{n} + \frac{(x_{bar} - \bar{x})^2}{S_{xx}^2}$  = 14286.16

# Example in R (cont'd)

```
new = data.frame(adv = 300)
predict.lm(model, new, interval =
"predict", level = 0.99)
```

	fit	lwr	upr
1	674.3047	338.712	1009.897

```
predict.lm(model, new, interval =
"predict", level = 0.95)
```

	fit	lwr	upr
1	674.3047	427.0146	921.5948

```
predict.lm(model, new, interval =
"confidence", level = 0.99)
```

	fit	lwr	upr
1	674.3047	496.6497	851.9596

```
predict.lm(model, new, interval =
"confidence", level = 0.95)
```

	fit	lwr	upr
1	674.3047	543.395	805.2143

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

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

The confidence intervals are narrower than the prediction intervals because the prediction intervals have additional variance from the variation of a new

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Basic Regression:  
Assumptions and  
Diagnostics***

# Simple Linear Regression: Model

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

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

**Assumptions:**

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normally Distributed}$

# Residual Analysis

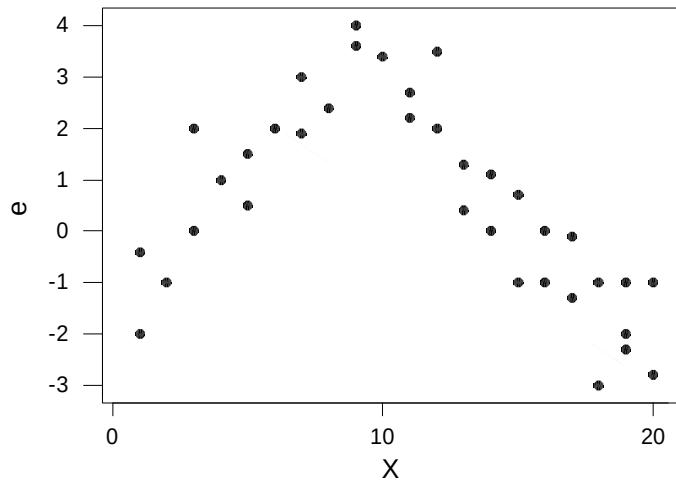
Residual Values:  $e_i = \hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

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

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

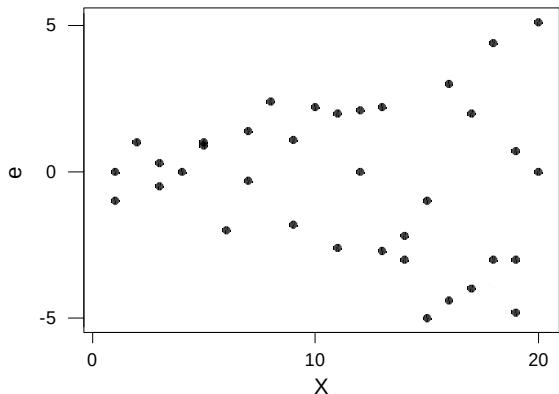
- ✓ The relationship between X and Y is not linear
- ✓ Variances of error terms are not equal
- ✓ Response data are not independent

# Checking Assumptions: Residual



**Non-linearity Assumption:**  
This plot shows that there may be a non-linear relationship between X and Y.

# Checking Assumptions: Residual Analysis



## Non-constant Variance Assumption:

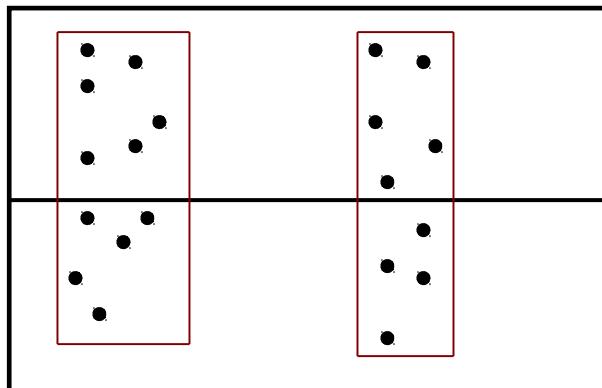
The residuals show larger variance as the predicting variable increases.

Here, it could be that  $\sigma^2$  is not constant.

# Checking Assumptions: Residual Analysis

## Independence Assumption:

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



# Checking Assumptions: Residual Analysis

## Independence Assumption:

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

- Using residual analysis, we check for uncorrelated errors but not independence;
- Independence is a more complicated matter; if the data are from a randomized trial, then independence is established but most data are from observational studies.

# Checking the Assumption of Normality

One way to check this assumption in a regression is using a Normal Probability Plot

x-axis  $\Phi^{-1} \left( \frac{r_i - 3/8}{n + 1/4} \right)$

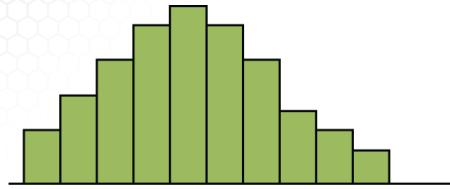
y-axis:  $e_i$

$r_i$  = rank of  $e_i$  (between 1, n)

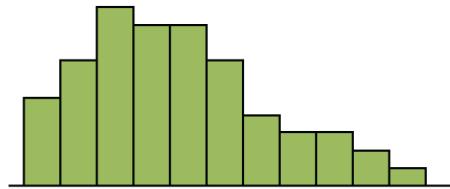
$\Phi$  = CDF of Normal Distribution

- ✓ Let the R statistical software do this for you!
- ✓ A straight line in normal probability plot implies assumption of normality is valid
- ✓ **Curvature** (especially at the ends) shows non-normality

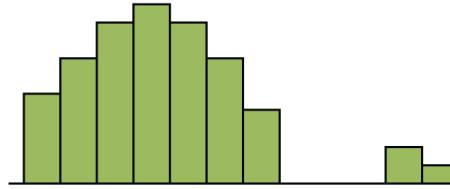
# Checking the Assumption of Normality



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



## Normality Assumption:



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

# Variable Transformation

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

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

# Normality Transformations

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

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

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

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

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

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

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

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

# Normality Transformations

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

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

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

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

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

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

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Basic Regression:***  
***Outliers and Predictive***  
***Power***

# Outliers in Regression

Any data point that is far from the majority of the data (in both x and y) is called an *outlier*.

- Data points that are far from the mean of the x's are called *leverage points*.
- A data point that is far from the mean of both the y's and the x's are *influential points* and can change the values of the estimated parameters significantly.

**The upshot:** Sometimes there are good reasons for excluding subsets (there were errors in the data entry; there were errors in the experiment). Sometimes - the outlier belongs in the data. Outliers should always be examined.

# Checking for Outliers

If we look at the **standardized residuals**

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

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

# Coefficient of Determination

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

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

which is interpreted as

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

(that uses X)

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

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

# Correlation Coefficient

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

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

Correlation coefficient and coefficient of variation:

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

# Regression Analysis

## Simple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

***Basic Regression: Model  
Diagnostic and Evaluation  
Example***

# Linear Regression: Example in R

A company, which sells medical supplies to hospitals, clinics, and doctor's offices, had considered the effectiveness of a new advertising program. Management wants to know if the advertisement is related to sales.

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

**Do the assumptions of the linear regression model hold? What is the explanatory power of the model?**

# Example in R: Residual Analysis

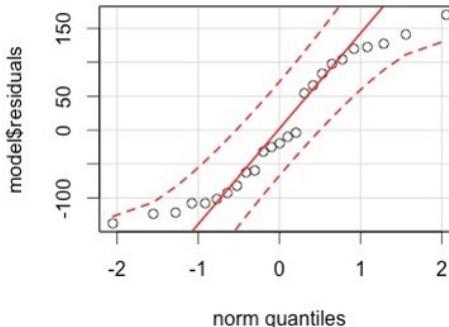
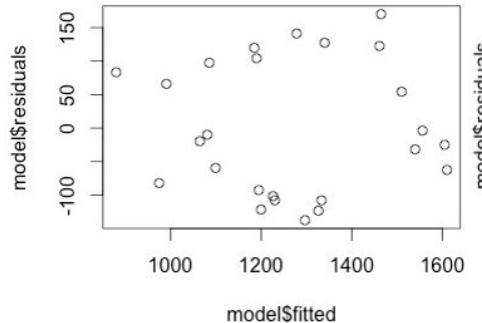
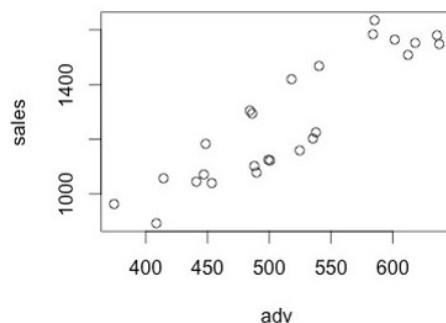
- a. What are the assumptions of linear regression?
- b. Do the assumptions hold? Provide the graphical displays needed to support the diagnostics. Interpret.
- c. Do you identify any outliers?
- d. How much variability in sales is explained by the advertisement expenditure?

# Example in R (cont'd)

- a. The assumptions are: **Linearity, Constant Variance, Independence, and Normality**

b. *plot(adv, sales)*

*plot(model\$fitted, model\$residuals)*  
*library(car); qqPlot(model\$residuals)*



Based on the above plots, the assumptions appear to hold.

# Example in R(cont'd)

c. Do you identify any outliers?

Based on the plots provided in part b, **there do not appear to be outliers**.

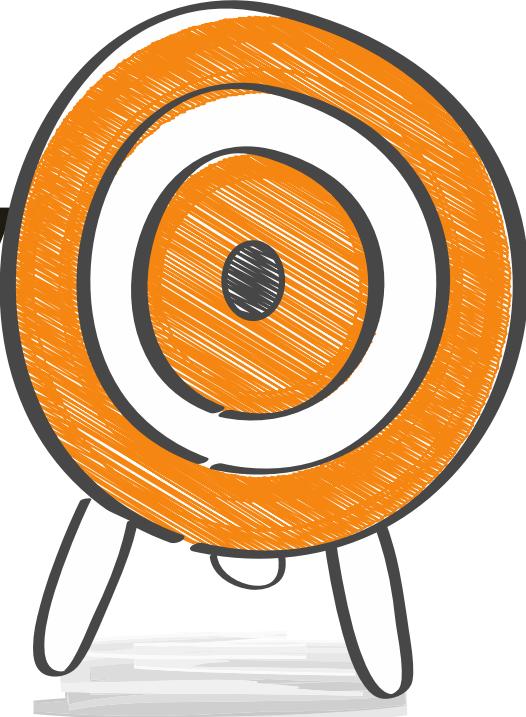
d. How much variability in sales is explained by the advertisement expenditure?

```
summary(model)$r.squared
```

```
[1] 0.8105919
```

Around **81%** of the variability in sales is explained by the advertising expenditure.

# Summary



# Regression Analysis

## Analysis of Variance

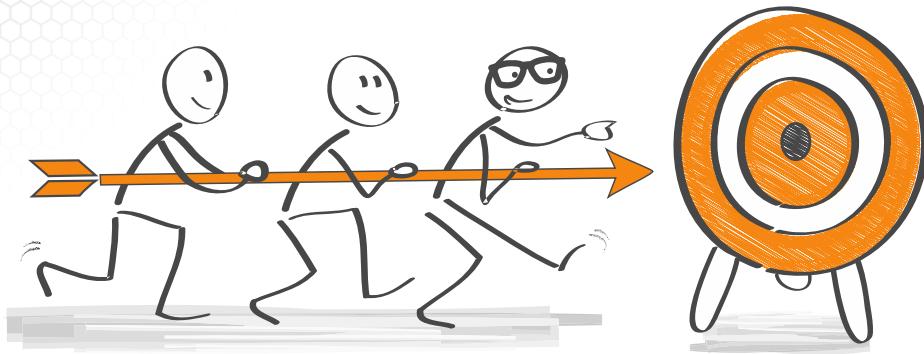
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA):  
Basics Concepts

# About this lesson



# ANOVA: Analysis of Variance

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

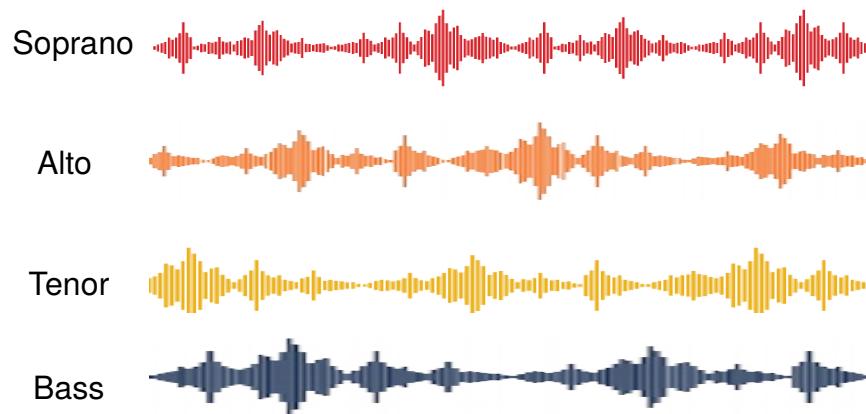
Population 2:  $(\mu_2, \sigma^2)$  → Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2})$  →  $(\bar{Y}_2, s_2^2)$

.....

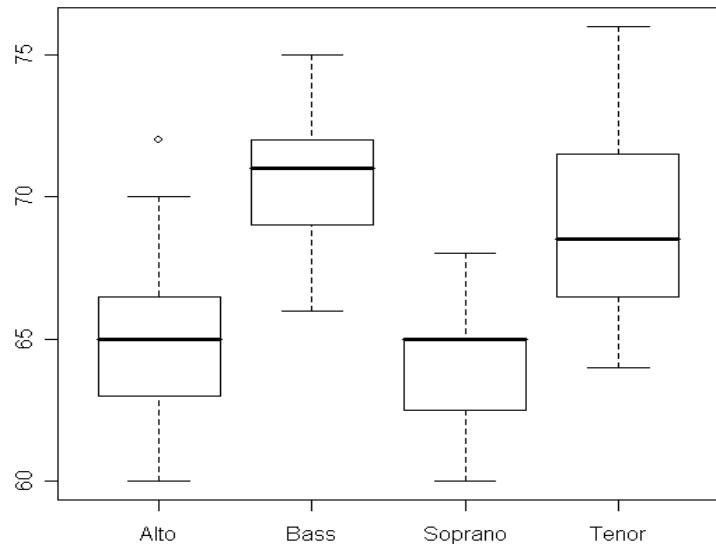
Population k:  $(\mu_k, \sigma^2)$  → Sample k:  $(Y_{k,1}, \dots, Y_{k,n_k})$  →  $(\bar{Y}_k, s_k^2)$

ANOVA: Comparing the means of multiple samples

# ANOVA Example 1: Voice Pitch and Height



# ANOVA Example 1: Voice Pitch and Height



1. Is there a difference in the height by voice pitch?
2. Which singers are taller?

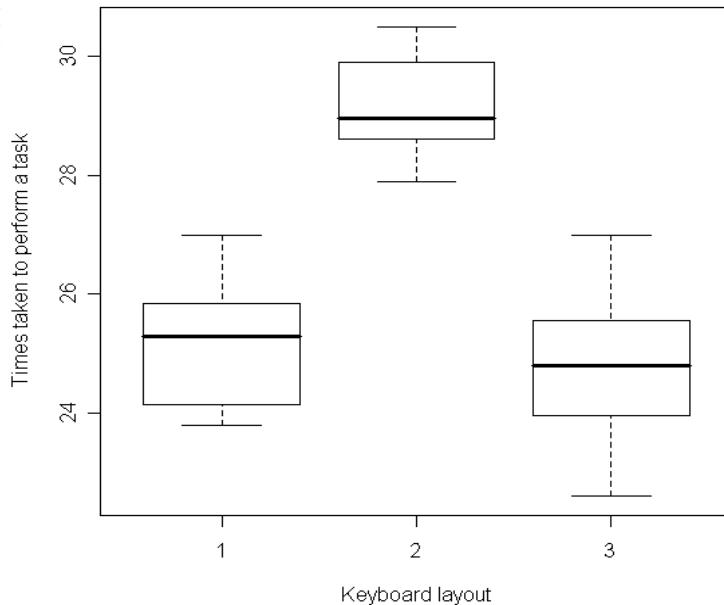
# ANOVA Example2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.



	Layout 1	Layout 2	Layout 3
	23.8	30.2	27.0
	25.6	29.9	25.4
	24.0	29.1	25.6
	25.1	28.8	24.2
	25.5	29.1	24.8
	26.1	28.6	24.0
	23.8	28.3	25.5
	25.7	28.7	23.9
	24.3	27.9	22.6
	26.0	30.5	26.0
	24.6	*	23.4
	27.0	*	*

# Operation Time by Keyboard Layout



1. Is there a difference in the time taken to perform a task?
2. Which layout is more effective?

# ANOVA: Objectives

*Primary objectives in ANOVA:*

1. Analysis of the variability in the data – the ANOVA table

2. Testing for equal means

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

3. Estimation of simultaneous confidence intervals for the mean differences

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

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA):  
Parameter Estimation

# ANOVA: Basics

Comparing means from multiple populations assuming the variances are the same and equal to :



*Pooled Variance Estimator:*

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

Where  $N$  = total number of samples ( $n_1 + \dots + n_k$ )

The degree of freedom is  $N-k$  because we replace  $S_i^2$  for  $i=1, \dots, k$  and thus losing  $k$  degrees of freedom

# Estimating the shared variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{\sum_{i=1}^k (n_i - 1)} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{N - k} = \mathbf{MSE}$$

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

# Mean Squared Error (MSE)

,..., *The sum of independent Chi-square random variables is also Chi-square*

where  $=N-k$

The sampling distribution of the pooled variance is a chi-square distribution with  $N-k$  degrees of freedom.

# Estimating Parameters in ANOVA

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

What is the sampling distribution?

But unknown, then replace with the pooled variance estimation:

$$\Rightarrow = =$$

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

MSE

Why  $N - k$ ?  
MSE =

# Confidence Intervals for the Means

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \quad \text{for } i = 1, \dots, k$$
$$\hat{\sigma}^2 = \text{MSE}$$

(1- ) confidence intervals for treatment means

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

# Regression Analysis

## Analysis of Variance

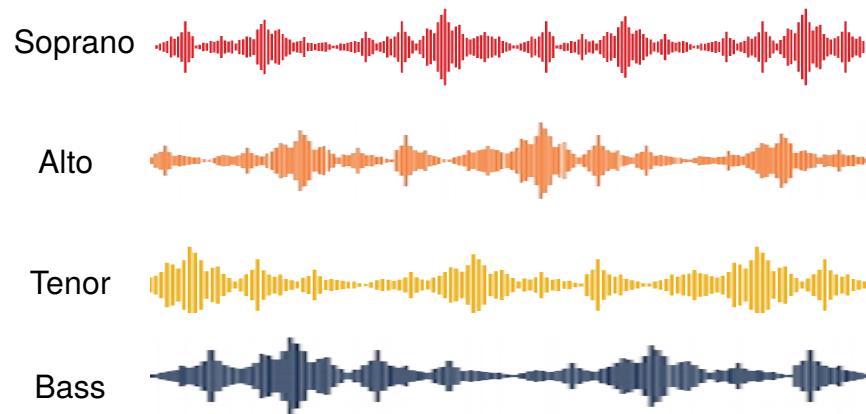
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA):  
Parameter Estimation Examples

# Example1: Voice Pitch and Height



What are the estimates for the mean heights for the different groups of singers?

# Parameter Estimation

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

Tables of means

Grand mean

67.11538

pitch

	Alto	Bass	Soprano	Tenor
rep	35.00	39.00	36.00	20.00

Overall Mean: 67.11536

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

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

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

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

tenor

# Example 2: Keyboard layout

Three different keyboard layouts are being compared in terms of typing speed.



	Layout 1	Layout 2	Layout 3
23.8	30.2	27.0	
25.6	29.9	25.4	
24.0	29.1	25.6	
25.1	28.8	24.2	
25.5	29.1	24.8	
26.1	28.6	24.0	
23.8	28.3	25.5	
25.7	28.7	23.9	
24.3	27.9	22.6	
26.0	30.5	26.0	
24.6	*	23.4	
27.0	*	*	

What are the estimates for the mean typing times for the different groups of keyboards?

# Parameter Estimation

```
model = aov(speed ~ layout)
```

```
model.tables(model, type = "means")
```

Tables of means

Grand mean

26.21212

layout

	1	2	3
rep	12.00	10.00	11.00
layout1	25.12	29.11	24.76
layout2			
layout3			

Overall Mean: 26.21212

$\frac{25.12}{MSE} \sim t_{n-2}$

$\frac{29.11}{S}$

$\sqrt{\frac{24.76}{S}}$

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA):  
Test for Equal Means

# Hypothesis Test for Equal Means

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

$H_a$ : some means are different

# Null Hypothesis

- If the null hypothesis is true, we combine  $k$  samples to estimate overall mean:

$$\text{Population 1: } \bar{Y} = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$$

- Base variance estimate on this mean:

$$S_0^2 = \frac{\sum_{i=0}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2}{N-1} = \frac{SST}{N-1}$$

- **SST** = Sum of Squares Total

$$\begin{aligned} & \text{Simple linear regression} \\ & Y = \beta_0 + \beta_1 x + \varepsilon \\ & (N-1)S_0^2 \sim \chi^2_{N-1} \\ & \text{Multiple linear regression} \\ & Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ & \sigma^2 \sim \chi^2_{N-1} \\ & \text{Polynomial Regression} \\ & Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \end{aligned}$$

# SST Decomposition

We can *partition* SST into two separate parts:

$$\text{SST} = \text{SSE} + \text{SST}_R,$$

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

1.  $\text{MSE} = \text{SSE}/N-k = \text{within-group variability}$
2.  $\text{MSST}_R = \text{SST}_R/k-1 = \text{between-group variability}$
3. ANOVA: comparing between to within variability
4.  $F = \text{between-group variability}/\text{within-group variability}$

# Testing equal variances with F-test

$$\frac{SST_R / (k-1)}{SSE / (N-k)} \text{ is } \frac{\text{MS}_R}{\text{MSE}} = F_0 \sim F_{k-1, N-k} \text{ if } H_0 \text{ is true}$$

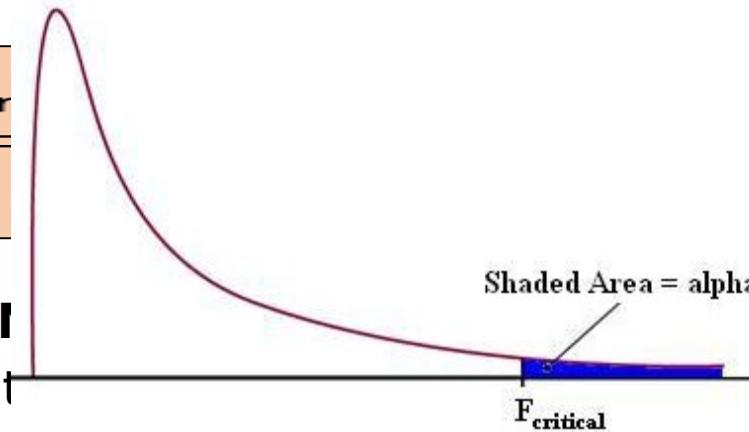
Reject  $H_0$  if  $F_0 > F_{\alpha}(k-1, N-k)$ , which is the upper  $\alpha^{\text{th}}$  quantile of the F distribution.

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

# Testing equal variances with F-test

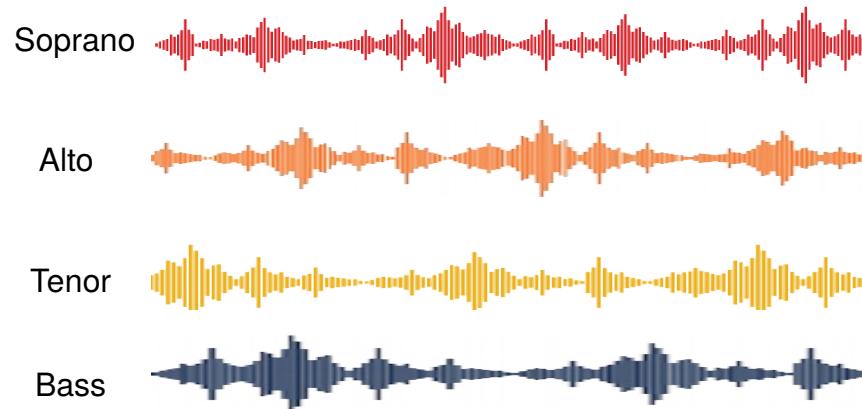
$$\frac{SST_R / (k-1) \text{ (Sampling MS)} }{SSE / (N-k) \text{ (MSE)}} \sim F_R$$

Reject  $H_0$  if  $F_0 > F_{\alpha}(k-1, N-k)$   
quantile of the F distribution



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

# Example 1: Voice Pitch and Height



Are the mean heights for the four groups of singers statistically different?

# Testing for Equal Means

*summary(aov(height ~ pitch))*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pitch	3	1058.5	352.8	55.8	<2e-16 ***
Residuals	126	796.7	6.3		

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

P-value  $\approx$  0 : Reject the null hypothesis of equal mean heights

# ANOVA Example 2: Keyboard layout

Three different keyboard layouts are being compared in terms of typing speed.



Are the mean typing times for the three keyboard layouts statistically different?

	Layout 1	Layout 2	Layout 3
23.8	30.2	27.0	
25.6	29.9	25.4	
24.0	29.1	25.6	
25.1	28.8	24.2	
25.5	29.1	24.8	
26.1	28.6	24.0	
23.8	28.3	25.5	
25.7	28.7	23.9	
24.3	27.9	22.6	
26.0	30.5	26.0	
24.6	*	23.4	
27.0	*	*	

# Testing for Equal Means

```
summary(aov(speed ~ keytype))
```

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

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SSTR = 121.24, k-1 = 2

SSE = 34.42 , N-k = 30

F-value = 52.84

P-value ≈ 0

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

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA)  
Comparing Pairs of Means

# Pairwise Comparison of Means

One primary goal of ANOVA might be to determine which treatment means are bigger or smaller. One way to do this is to compare all  $k(k-1)/2$  pairs of treatments.

For a  $(1 - \alpha)$  confidence interval for mean difference  $\mu_i - \mu_j$ :

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



The estimate of the difference in means



The percentile of the “studentized range” distribution.



The standard deviation/error of the estimator

# Difference between t and q

## Correct for simultaneous Inference:

- $q > t$  (at any fixed  $\alpha$  and df)
- intervals are wider to compensate for the fact that we are making simultaneous comparisons (multiplicity correction)

## Why?

95% CIs for two populations  $\Rightarrow (.95)(.95) \approx .90$   $\Rightarrow$  The simultaneous or joint confidence level for the two parameters is roughly **90%**.

95% CIs for three populations  $\Rightarrow (.95)(.95)(.95) \approx .86$   $\Rightarrow$  The simultaneous or joint confidence level for the three parameters is roughly **86%**.

# ANOVA Example 1: Voice Pitch and Height



Which mean heights for the four groups of singers are statistically different?

# Pairwise Comparison

*TukeyHSD(aov(height ~ pitch))*

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = height ~ pitch)

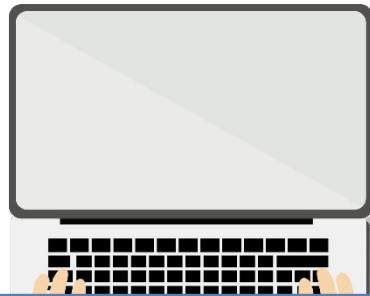
\$pitch

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

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

# ANOVA Example 2: Keyboard layout

Three different keyboard layouts are being compared in terms of typing speed.



Which mean typing times for the three keyboard layouts are different?

	Layout 1	Layout 2	Layout 3
23.8	30.2	27.0	
25.6	29.9	25.4	
24.0	29.1	25.6	
25.1	28.8	24.2	
25.5	29.1	24.8	
26.1	28.6	24.0	
23.8	28.3	25.5	
25.7	28.7	23.9	
24.3	27.9	22.6	
26.0	30.5	26.0	
24.6	*	23.4	
27.0	*	*	

# Pairwise Comparison

*TukeyHSD(aov(speed ~ keytype))*

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = speed ~ keytype)

\$keytype

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

- *Keyboard type 2 has a statistically significantly higher typing time than Keyboard type 1 and type 3, in average.*
- *It is plausible that keyboard types 1 and 3 have similar typing time in average.*

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Analysis of Variance (ANOVA)  
Model Fit Assessment

# ANOVA: Model & Assumptions

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

Population 2:  $(\mu_2, \sigma_2^2)$

Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2})$

Model:  $Y_{ij} = \mu_2 + \epsilon_{ij}$  where  $\epsilon_{ij}$  = error term

## Assumptions:

- Constant Variance Assumption:  $\text{Var}(\epsilon_{ij}) = \sigma^2$

- Independence Assumption:  $\{Y_{1,1}, \dots, Y_{1,n_1}, Y_{2,1}, \dots, Y_{2,n_2}\}$  are independent random variables

- Normality Assumption:  $\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$

# Residual Analysis

$$\bar{Y}_{ij} = \mu_i + \varepsilon_{ij}$$



In the model,  $\varepsilon_{ij}$  is the *error term*. We want  $\varepsilon_{ij} \sim N(0, \sigma^2)$ . To check to see if this is true, we examine the residual errors:



If the model fit is a good fit, then the residuals should be scattered around zero (randomly).

# Residual Analysis

Residual plots:

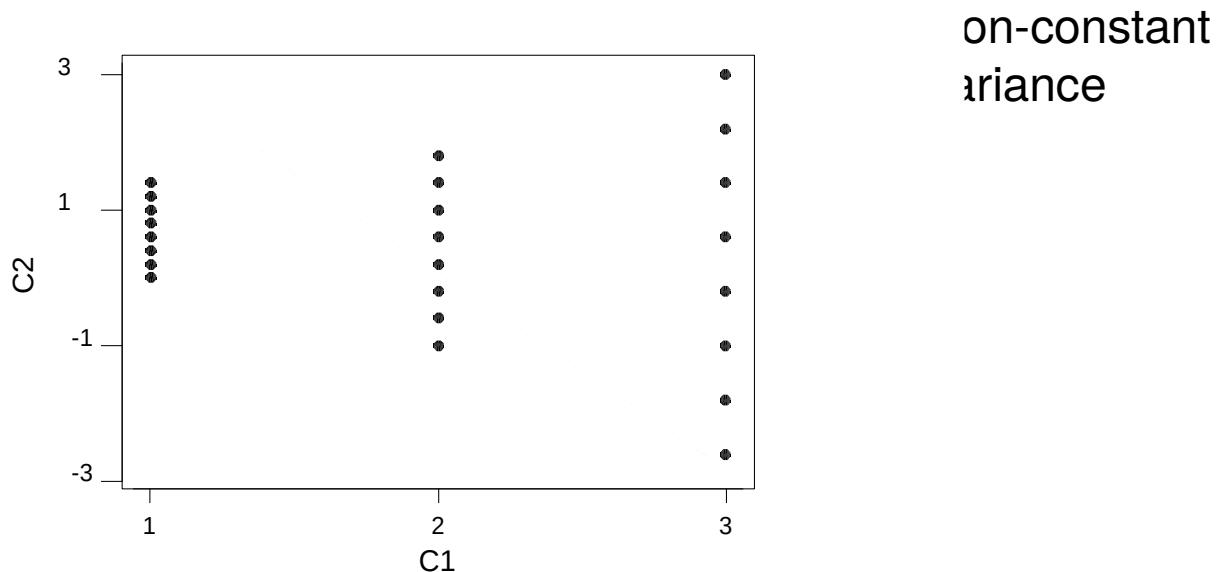
- plot for each treatment group
- plot the quantile-quantile normal plot Q-Q
- plot the histogram of

If the scatter of is not random, it could be that

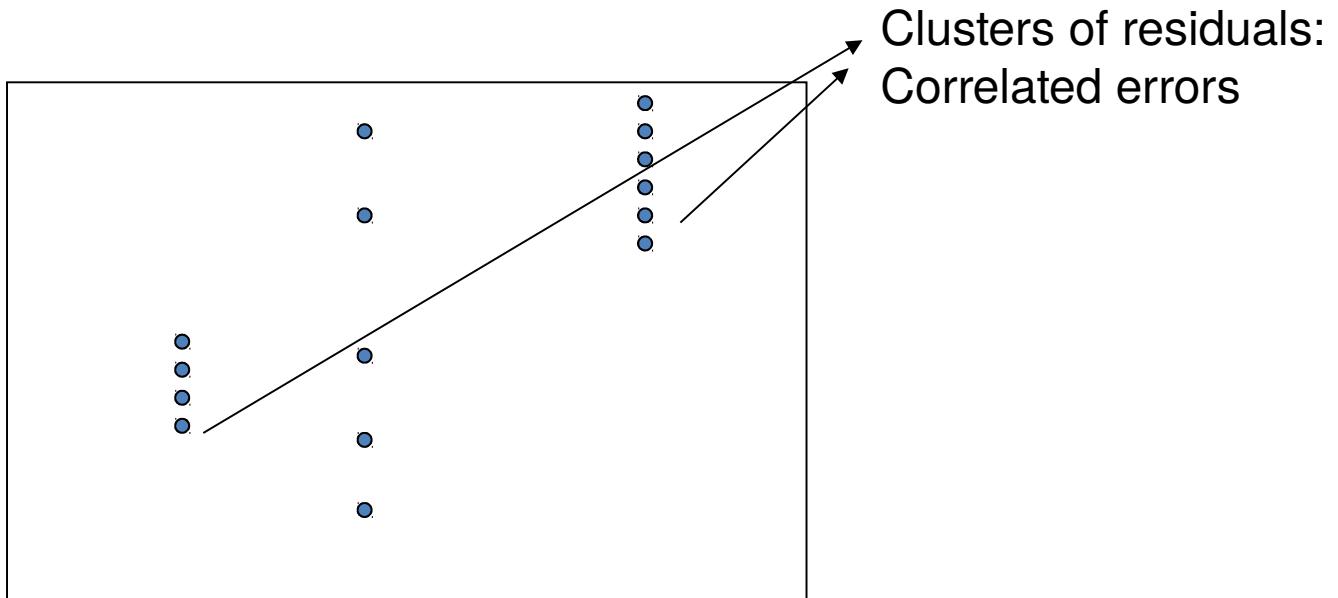
- sample responses are not independent
- variances of responses are not equal

If the quantile-quantile normal plot and the histogram show departure from normality, you may consider a transformation.

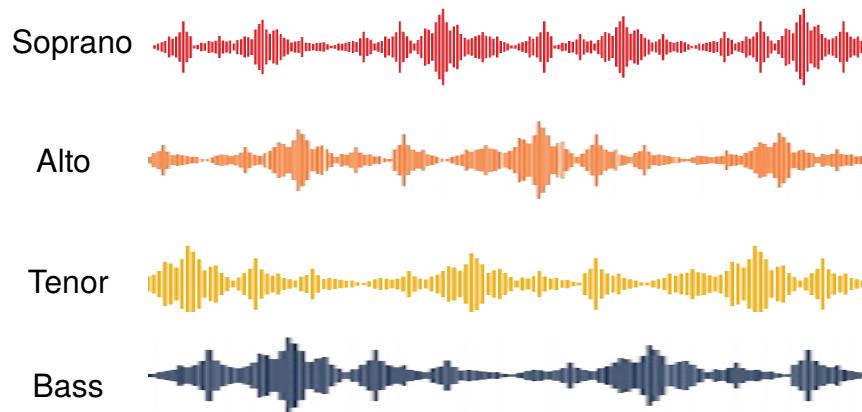
# Residual Plot Example 1



# Residual Plot Example 2

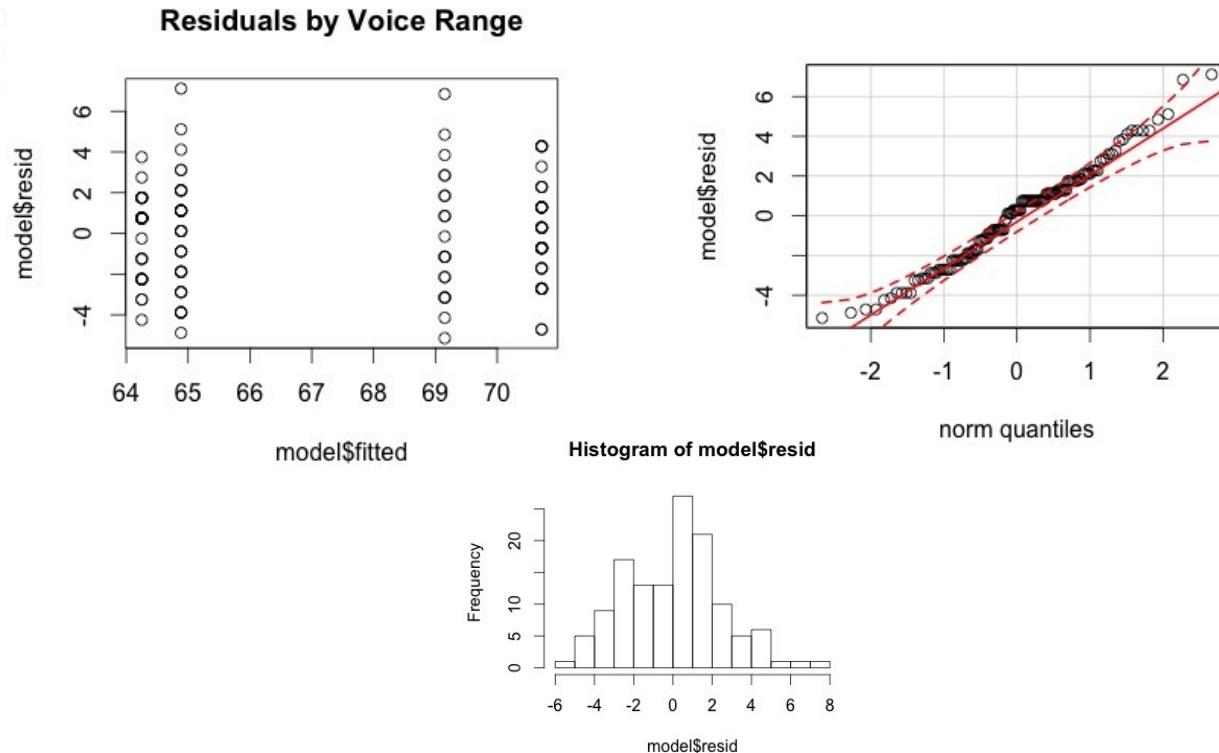


# ANOVA Example 1: Voice Pitch and Height



Are the inferences on the difference in height means reliable?

# Residual Analysis: Example 1



# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Analysis of Variance (ANOVA) vs  
Simple Linear Regression

# Linear Regression & ANOVA

Simple Linear Regression:

Population 2:  $(\mu_2, \sigma_2^2)$       Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2})$

Data:  $\{(x_i, Y_i), \dots, (x_n, Y_n)\}$

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

**ANOVA:** A linear regression model where the predicting factor is a categorical variable.

Population 1:  $(\mu_1, \sigma_1^2)$       Sample 1:  $(Y_{1,1}, \dots, Y_{1,n_1})$

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

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

=  $i$ -th group mean decomposed into =

# ANOVA & Linear Regression (cont'd)

ANOVA:

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

**Model:**  $Y_{ij} = \text{where}$

$\equiv$  i-th group mean decomposed into =

Define  $Y$  be the response variable

$Y = (Y_{11}, Y_{12}, \dots, Y_{1k}, \dots, Y_{21}, \dots, Y_{2k}, \dots, Y_{k1}, \dots, Y_{kk})$

Define  $L$  be the label/categorical variable:

$L = (l_{11}, \dots, l_{21}, \dots, \dots, l_{k1}, \dots, \dots, l_{kk})$ , Where  $l_{ij} = l$

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

# ANOVA & Linear Regression (cont'd)

Categorical Variables in Linear Regression:

- Transform categories into dummy variables

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

- If intercept in the model, only  $k-1$  dummy variables because of linear dependence:  $(1, 1, \dots, 1)^T = x_1 + x_2 + \dots + x_k$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{k-1} x_{ik-1} + \epsilon_i, i = 1, \dots, n$

- If no intercept in the model, all  $k$  dummy variables

**Model:**  $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i, i = 1, \dots, n$

ANOVA: A linear regression model with multiple predictors  Multiple Linear Regression

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

ANOVA R Example

# Cancer Survival



## Reference:

Cameron, E. and Pauling, L. (1978) Supplemental ascorbate in the supportive treatment of cancer: re-evaluation of prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Science USA*, 75, 4538-4542.

# ANOVA Example Data

Response Variable:

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

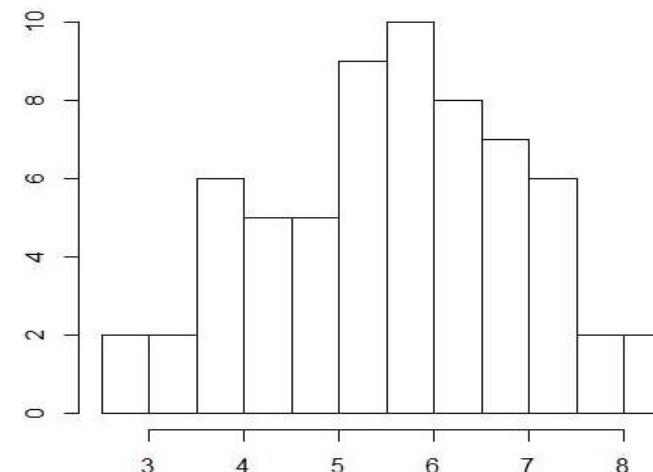
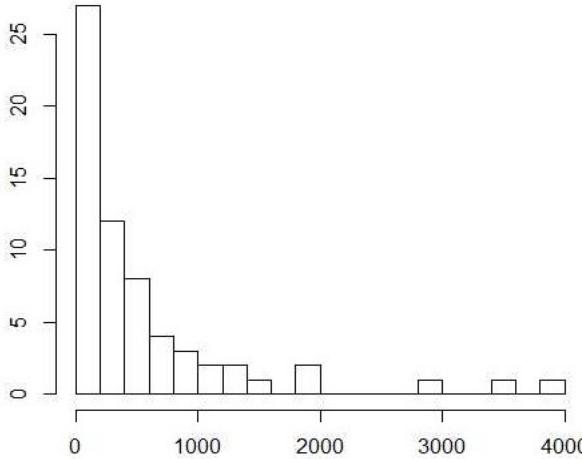
Categories:

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

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

# Exploratory Data Analysis in R

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



# ANOVA in R

```
## Need to specify Response & Categorical Variables
```

```
survival = log(survival)
```

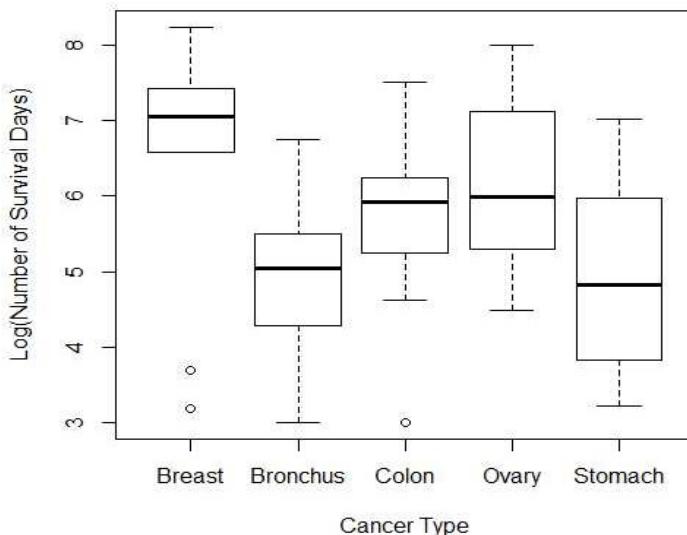
```
cancertype = cancer_data$Organ
```

```
## Convert into categorical variable in R
```

```
cancertype = as.factor(cancertype)
```

```
## Explore relationship visually
```

```
boxplot(survival~cancertype, xlab = "Cancer Type", ylab = "Log(Number of
```



**Between-variability** – there is some variability between the means of the five groups

*Is the between-variability significantly larger than the within-variability?*

# ANOVA in R (cont'd)

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cancertype	4	24.49	6.122	4.286	0.00412
Residuals	59	84.27	1.428		

---

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

```
## Obtain estimated means  
model.tables(model, type = "means")
```

Tables of means  
Grand mean  
5.555785  
cancertype

	Breast	Bronchus	Colon	Ovary	Stomach
rep 1	11.000	17.000	17.000	6.000	13.000



SSTR = 24.49, k-1 = 4  
SSE = 84.27, N-k = 59  
F-value = 4.286  
P-value = 0.00412



$$\begin{aligned} &= 705.6, n_1 = 11 \\ &= 141.6, n_2 = 17 \\ &= 313.9, n_3 = 17 \\ &= 469.2, n_4 = 6 \\ &= 143.7, n_5 = 13 \end{aligned}$$

Population 1:  $(\mu_1, \sigma_1)$

Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2})$

Are the means statistically significantly different?

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

# Pairwise Comparison in R

## Which means are statistically significantly different? Pairwise

Comparison

*TukeyHSD(model)*

*Tukey multiple comparisons of means*

*95% family-wise confidence level*

Fit: *aov(formula = survival ~ cancertype)*

*\$cancertype*

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

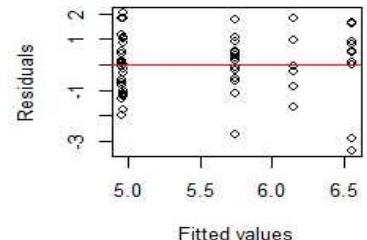
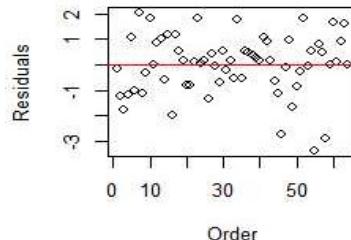
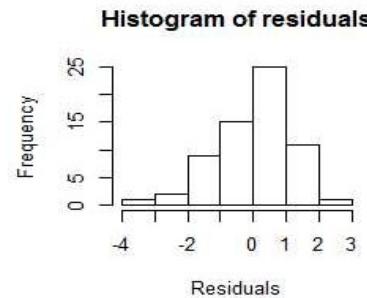
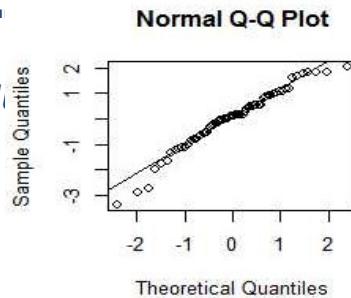


Statistically significant:  
 $t$  – interval for  $\beta_1$

# Residual Analysis in R

```
par(mfrow=c(2,2))
qqnorm(residuals(model))
qqline(residuals(model))
hist(residuals(model),main="Histogram of
residuals",xlab="Residuals")
plot(residuals(model),xlab="Order",ylab="Residuals"
abline(0,0,lty=1,col="red")
plot(fitted(model), residuals(model),xlab="Fitted val
ylab="Residuals")
abline(0,0,lty=1,col="red")
```

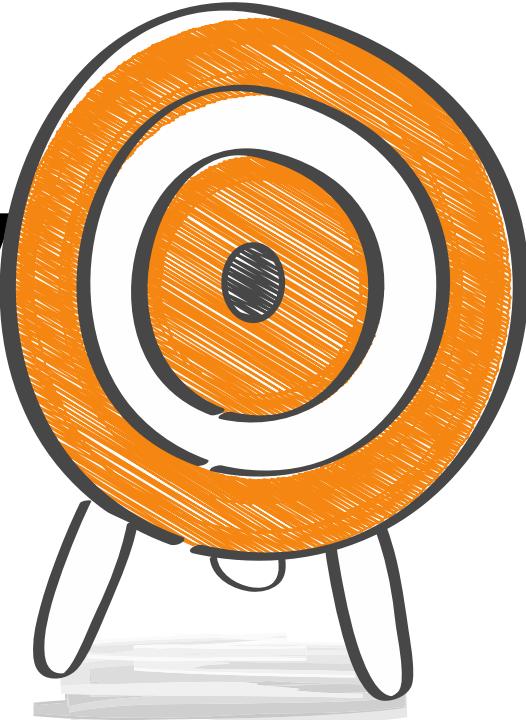
- The quantiles align on the line and the histogram is approx. symmetric thus normality assumption holds
- Residuals are scattered around zero line with no pattern thus both the constant variance and uncorrelated errors hold



# Cancer Survival: Findings

- There is strong evidence for the difference in the survival time across the five different types of cancer;
- Survival time: breast cancer vs. Bronchus or Stomach cancer.

# Summary



# Regression Analysis

## Multiple Linear Regression

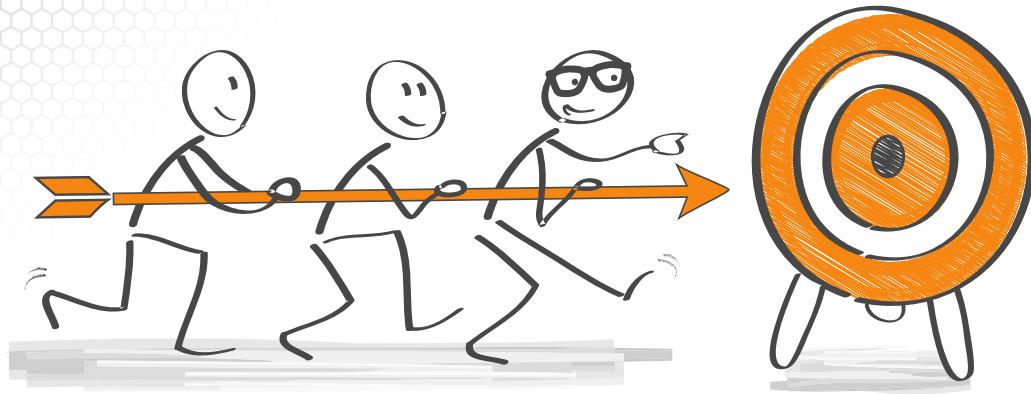
**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Objectives and Examples

# About this lesson



# Linear Regression: Example 1



# Linear Regression: Data Example :

**The response variable is:**

**Y** = Sales (in thousands of dollars)

**The predicting variables are:**

**X<sub>1</sub>** = the amount (in hundreds of dollars) spent on advertising

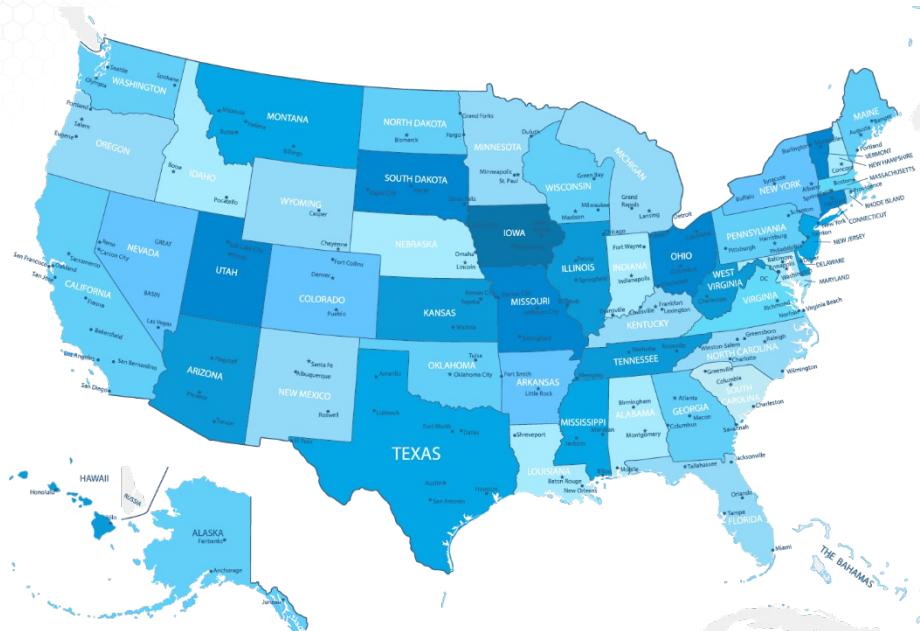
**X<sub>2</sub>** = the total amount of bonuses paid

**X<sub>3</sub>** = the market share in each territory

**X<sub>4</sub>** = the largest competitor's sales

**X<sub>5</sub>** = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

# Linear Regression: Example 2



SAT Mean Score by State - Year  
1982  
790 (South Carolina) - 1088 (Iowa)

# Linear Regression: Data Example

2

**The response variable is:**

**Y** = State average SAT score (verbal and quantitative combined)

**The predicting variables are:**

**X<sub>1</sub>** = % of total eligible students (high school seniors) in the state who took the exam

**X<sub>2</sub>** = median income of families of test takers, in hundreds of dollars

**X<sub>3</sub>** = average number of years that test takers had in social sciences, natural sciences, and humanities

**X<sub>4</sub>** = % of test takers who attended public schools

**X<sub>5</sub>** = state expenditure on secondary schools, in hundreds of dollars per student

**X<sub>6</sub>** = median percentile of ranking of test takers within their secondary school classes

# Linear Regression: Example 3



# Linear Regression: Data Example 3

**The response variable is:**

**Y** = the rating provided by the IMDb

**The predicting variables are:**

**X<sub>1</sub>** = Number of votes for the movie on the IMDb platform

**X<sub>2</sub>** = Duration of the movie

**X<sub>3</sub>** = Gross earnings scaled in 1000's

**X<sub>4</sub>** = Total budget in millions

**X<sub>5</sub>** = Release year (between 2010-2014)

**X<sub>6</sub>** = Rating of a film's suitability for certain audiences, based on its content

**X<sub>7</sub>** = Language: English (1) and Other languages (0)

**X<sub>8</sub>** = Genre: Action (1), Documentary (2), Comedy (3), Horror, Sci-Fi (4)

**X<sub>9</sub>** = Director Rating: Awarded (1), Nominated (2), None (3)

**X<sub>10</sub>** = Actor Rating (Low=0 and High=1)

**X<sub>11</sub>** = Movie Awards: Avarded (1), Nominated (2), None (3)

# Multiple Linear Regression: Objectives

A regression analysis is used for:

1. **Prediction** of the response variable;
2. **Modelling** the relationship between the response variable and the explanatory variables; or
3. **Testing** hypotheses of association relationships.

**Linear Regression:** The basis of what we will be talking about most of this course is the linear model. Virtually all other methods for studying dependence among variables are variations on the idea of linear regression.

“ ”

*All models are wrong, but some are useful.*    George Box

“ ”

*Embrace your data, not your models.*    John Tukey

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Basics of Multiple Regression

# Multiple Linear Regression: Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i$

$= 1, \dots, n$

**Assumptions:**

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- (Later we assume  $\varepsilon_i \sim \text{Normal}$ )

# Multiple Linear Regression: Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i$

**The model parameters are:**  $\beta_0, \beta_1, \dots, \beta_p, \sigma^2$

- Unknown regardless how much data are observed
  - Estimated given the model assumptions
  - Estimated based on data
- 
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
  - (Later we assume  $\varepsilon_i \sim \text{Normal}$ )

# Multiple Linear Regression: Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, i = 1, \dots, n$   
**Population:**  $(\mu_2, \sigma^2_2)$       **Sample:**  $(Y_{2,1}, \dots, Y_{2,n2})$   
**Model in Matrix Form:**  $\mathbf{Y} = \mathbf{X} \beta + \epsilon$   
+

**Design Matrix**      **Response**      **Error**      **Coefficients**

**Model in Matrix Form:**  $\mathbf{Y} = \mathbf{X} \beta + \epsilon$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

# Model Flexibility: Main Effects & Interactions

For  $k=2$  predicting variables, four useful regressions:

1. 1<sup>st</sup> Order Model:

$$Y = \beta_0 + \beta_1 + \beta_2 + \varepsilon$$

2. 2<sup>nd</sup> Order Model:

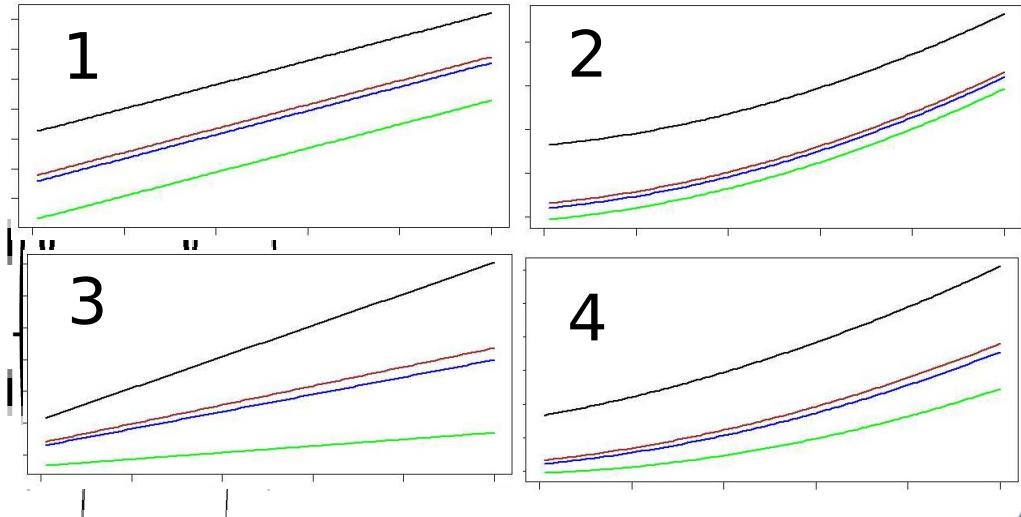
$$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \varepsilon$$

3. 1<sup>st</sup> Order Interaction Model:

$$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \varepsilon$$

4. 2<sup>nd</sup> Order Interaction Model:

$$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \varepsilon$$



# Quantitative and Qualitative Variables

**Simple Linear Regression:** Linear regression with one quantitative predicting variable

**ANOVA:** Linear regression with one or more qualitative predicting variables

**Multiple Linear Regression:** Multiple quantitative and qualitative predicting variables

# Quantitative and Qualitative Variables

**Multiple Linear Regression:** Multiple quantitative and qualitative predicting variable

$\mathbf{X}_1$  quantitative &  $\mathbf{X}_2$  qualitative with three levels:  $\mathbf{D}_1$ ,  $\mathbf{D}_2$ , and  $\mathbf{D}_3$  dummy variables

Model:  $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{d}_1 + \beta_3 \mathbf{d}_2 + \varepsilon$

**Intercept varies**

If  $\mathbf{d}_1=0, \mathbf{d}_2=0$ :  $\beta_0 + \beta_1 \mathbf{x}_1$

If  $\mathbf{d}_1=1, \mathbf{d}_2=0$ :  $\beta_0 + \beta_2 + \beta_1 \mathbf{x}_1$

**Parallel regression**

**lines**

If  $\mathbf{d}_1=0, \mathbf{d}_2=1$ :  $\beta_0 + \beta_3 + \beta_1 \mathbf{x}_1$

**If  $\mathbf{X}_1 \mathbf{X}_2$  interaction:** The regression lines are

# Quantitative and Qualitative Variables

**Multiple Linear Regression:** Multiple quantitative and qualitative predicting variable

$X_1$  quantitative &  $X_2$  qualitative with three levels:  $D_1$ ,  $D_2$ , and  $D_3$  dummy variables

Model:  $Y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \epsilon$

**Intercept varies**

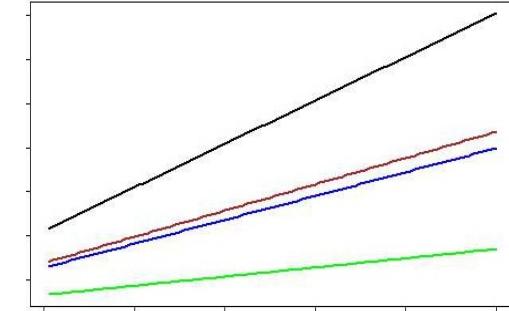
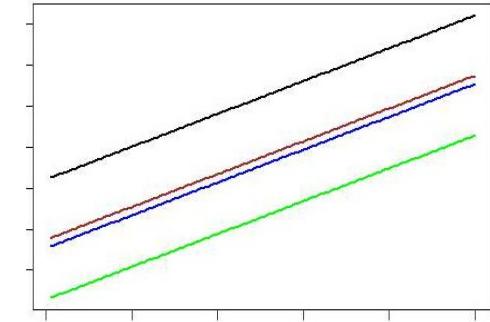
If  $d_1=0, d_2=0$ :  $\beta_0 + \beta_1 x_1$

If  $d_1=1, d_2=0$ :  $\beta_0 + \beta_2 + \beta_1 x_1$

**Parallel regression lines**

If  $d_1=0, d_2=1$ :  $\beta_0 + \beta_3 + \beta_1 x_1$

**If  $X_1, X_2$  interaction:** The regression lines are



# Linear Regression: Example

## 1



# Linear Regression: Example

## 1

### **Quantitative Predicting Variables:**

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

### **Qualitative Predicting Variable:**

$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

# Linear Regression: Example 3



# Linear Regression: Example

3

**Quantitative predicting variables are:**

$X_1$  = Number of votes for the movie on the IMDb platform

$X_2$  = Duration of the movie

$X_3$  = Gross earnings scaled in 1000's

$X_4$  = Total budget in millions

**Qualitative predicting variables are:**

$X_5$  = Release year (between 2010-2014)

$X_6$  = Rating of a film's suitability for certain audiences,  
based on content

$X_7$  = Language: English (1) and Other languages (0)

.....

# Linear Regression: Example

3

**Quantitative predicting variables are:**

$X_1$  = Number of votes for the movie on the IMDb platform

$X_2$  = Duration of the movie

**'Year' : A quantitative or qualitative predicting variable?**

- If observations are made over many years, then consider it a quantitative;**
- If observations are made over few years, then consider it as qualitative.**

$X_5$  = Release year (between 2010-2014),

$X_6$  = Rating of a film's suitability for certain audiences, based on content

$X_7$  = Language: English (1) and Other languages (0)

.....

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Regression Parameter Estimation

# Parameter Estimation $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ ,

$\sigma^2$

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ , we find values that minimize

squared error for:  $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 = (Y - X\beta)^T (Y - X\beta)$

$$X^T X \hat{\beta} = X^T Y \quad \text{if } X^T X \text{ invertible} \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

The fitted values are  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = H Y$

Where **H** is called the **hat matrix**.

The residual are  $\hat{\epsilon} := Y - X\hat{\beta} = (I - H)Y \quad \hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n - p - 1}$

# Parameter Estimation $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ ,

$\sigma^2$

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ , we find values that minimize

The estimator of  $\sigma^2$  is MSE:  $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 = (Y - X\beta)^T (Y - X\beta)$

Assuming  $\varepsilon_1, \dots, \varepsilon_n$  are normally distributed, then

MSE  $\sim \chi^2$  with n-p-1 degrees of freedom (Why n-p-1?)

The fitted values are  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = H Y$

Where **H** is called the **hat matrix**.

The residual are  $\hat{\varepsilon} := Y - X\hat{\beta} = (I - H)Y \rightarrow \hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p - 1}$

# Parameter Estimation

$\hat{\chi}^2 = \sum \hat{\varepsilon}_i^2$   
(Chi-squared distribution with  $n-p-1$  degrees of freedom  
(~~Chi-squared distribution with  $n-p-1$  degrees of freedom~~)

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



Estimating  $\sigma^2$  — Sample variance

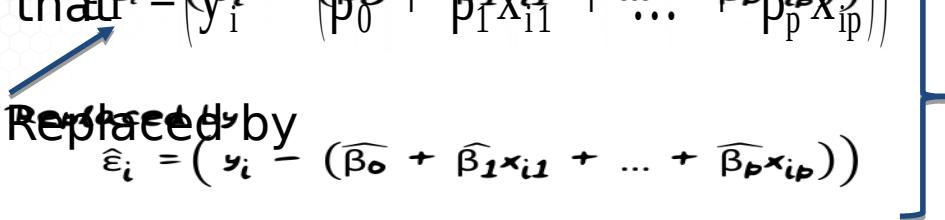
This is the sample variance estimator except we use  $n-p-1$  degree of freedom. **Why?**

# Parameter Estimation

Recall that  $\hat{\epsilon}_i = (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))$

Replaced by  $\hat{\epsilon}_i = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}))$

Use  $p+1$  degrees of freedom because  $\hat{\beta}_0 \leftarrow \beta_0$   
 $\hat{\beta}_1 \leftarrow \beta_1$   
....  
 $\hat{\beta}_p \leftarrow \beta_p$



Thus, assuming that

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

$$\hat{\sigma}^2 = \text{MSE} \sim \chi^2_{n-p-1}$$

(This is called the sampling distribution of )

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Model Interpretation

# Model Interpretation: Parameters

The Least Squares estimated coefficients have specific interpretations:

- ✓  $\hat{\beta}_0$  is the estimated expected value of the response variable when all the predictor variables equal zero;
- ✓  $\hat{\beta}_i$  is the estimated expected change in the response variable associated with a unit change in the  $i$ th predictor variable predicting all other predictors fixed at their mean values in the model for all ;

# Multiple Interpretation: Simple vs Multiple Regression

Marginal versus Conditional relationship:

- *Marginal*: Simple linear regression captures the association of a predicting variable to the response variable marginally, i.e. without consideration of other factors.
- *Conditional*: Multiple linear regression captures the association of a predicting variable to the response variable, conditional of other predicting variables in the model;
- Generally, the estimated regression coefficients for the conditional and marginal relationships can be different not only in magnitude but also in sign or direction

# Multiple Interpretation: Causality vs. Association

Causality Statements: Experimental Designs

Association Statements: Observational Studies

- *Example*: We take a sample of college students and determine their College grade point average (COLGPA), High school GPA (HSGPA), and SAT score (SAT). The estimated model is:

$$COLGPA = 1.3 + 0.7 HSGPA - 0.0003 SAT$$

- o **Incorrect Interpretation**: Higher values of SAT are associated with lower values of College GPA.
- o **Correct Interpretation**: higher values of SAT are associated with lower values of College GPA, holding High school GPA fixed.
- o The coefficients of a multiple regression must not be interpreted marginally!

# Different Roles of Predicting Variables

Predicting Variables can be distinguished as:

- **Controlling** – to control for bias selection in the sample.  
They are used as ‘default’ variables in order to capture more meaningful relationships.
- **Explanatory** – to explain variability in the response variable;  
they may be included in the model even if other “similar” variables are in the model;
- **Predictive** – to best predict variability in the response regardless of their explanatory power.

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Regression Parameter Estimation:

Data Example

# Linear Regression: Example 1



## **Quantitative Predicting Variables:**

$X_1$  = the amount (in hundreds of dollars) spent on advertising

$X_2$  = the total amount of bonuses paid

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

## **Qualitative Predicting Variable:**

$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

# Example 1: Estimation & Interpretation

- a. Fit a linear regression with all predictors. What are the estimated regression coefficients and the estimated regression line?
- b. Interpret the coefficients. Compare the estimated coefficient for the advertisement expenditure variable under the conditional (full) model vs. marginal (one predictor) model.
- c. What does the model predict as the advertisement expenditure increases for an additional \$1,000 using the full regression model? Is the prediction different when compared to the prediction from the simple linear model with the advertisement expenditure variable only?
- d. What is the estimate of the error variance? Is it different from the simple linear regression model? Why?

# Example 1: Estimation & Interpretation

```
meddcor = read.table("meddcor.txt", sep = "", header =  
FALSE)  
colnames(meddcor) = c("sales", "advertising", "bonuses",  
"marketshare", "largestcomp", "region")  
meddcor$region = as.factor(meddcor$region)  
model = lm(sales ~ ., data = meddcor)  
summary(model)
```

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	117.0200	192.9732	0.606	0.5518
advertising	1.4092	0.2687	5.244	5.49e-05
bonuses	1.0123	0.4641	2.181	0.0427
marketshare	3.1548	2.9802	1.059	0.3038
largestcomp	-0.2354	0.2338	-1.007	0.3275
region2	53.6285	34.7359	1.544	0.1400
region3	267.9569	47.5577	5.634	2.40e-05 ***
---				

Residual standard error: 55.57 on 18 degrees of freedom

a. Estimated Regression Coefficients  
Coefficients

b. Conditional model:  
b. Conditional2model:

The expected additional gain in sales in thousands for \$100 additional expenditure in advertising holding all other fixed

Marginal model holding all other fixed

The expected additional gain in sales in thousands for \$100 additional expenditure in advertising holding all other fixed

The expected additional gain in sales in thousands for \$100 additional expenditure in advertising holding all other fixed

additional expenditure in advertisement not accounting for other predicting variables.

# Example 1: Estimation & Interpretation

```
meddcor = read.table("meddcor.txt", sep = "", header = FALSE)
```

```
colnames(meddcor) = c("sales", "advertising",  
"bonuses", "marketshare", "largestcomp", "region")
```

```
meddcor$region = as.factor(meddcor$region)
```

```
model = lm(sales ~ ., data = meddcor)
```

```
summary(model)
```

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	117.0200	192.9732	0.606	0.5518
advertising	1.4092	0.2687	5.244	5.49e-05
bonuses	1.0123	0.4641	2.181	0.0427
marketshare	3.1548	2.9802	1.059	0.3038
largestcomp	-0.2354	0.2338	-1.007	0.3275
region2	53.6285	34.7359	1.544	0.1400
region3	267.9569	47.5577	5.634	2.40e-05 ***
---				

Residual standard error: 55.57 on 18 degrees of freedom

- c. An additional **\$1,000** in advertising expenditures results in **\$14,092** additional sales under full model and **\$27,720** additional sales under simple linear model.

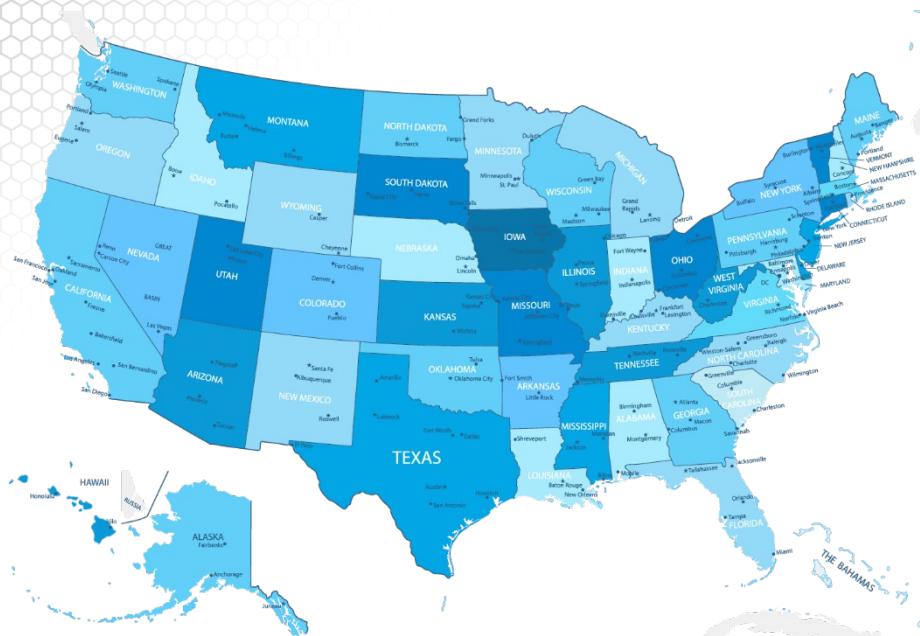
## Which is more meaningful?

Because sales vary with other factors, the interpretation based on the multiple regression is more meaningful.

- d. Under full model, the variance estimate is **55.57<sup>2</sup>**. Under the simple linear model, the variance estimate was **101.4<sup>2</sup>**.

**Why?** More variability in the response is explained when including multiple predicting variables

# Linear Regression: Example 2



## Controlling factors:

$X_1$  = % of total eligible students in the state who took the exam

$X_6$  = median percentile of ranking of test takers within their secondary school classes

## Explanatory Factors:

$X_2$  = median income of families of test takers, in hundreds of dollars

$X_3$  = average number of years that test takers had in social sciences, natural sciences, and humanities

$X_4$  = % of test takers who attended public schools

$X_5$  = state expenditure on secondary schools, in hundreds of dollars per student

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Inference for Regression Parameters

# Properties of Regression Estimators

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \Sigma$$

Furthermore,  $\hat{\beta}$  is a linear combination of  $\{Y_1, Y_2, \dots, Y_n\}$ . If we assume that  $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ , then  $\hat{\beta}$  is also distributed as  $\text{Normal}(\beta, \Sigma)$ .

$$\hat{\beta} \sim N(\beta, \Sigma)$$

# Properties of Regression Estimators

$$E(\hat{\beta}) = \beta$$

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \Sigma$$

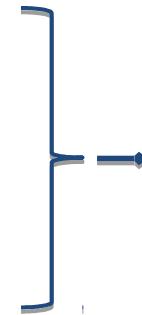
Furthermore,  $\hat{\beta}$  is a linear combination of  $\{Y_1, Y_2, \dots, Y_n\}$ . If we assume that  $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ , then  $\hat{\beta}$  is also distributed as  $\text{Normal}(\beta, \Sigma)$ . If we assume that  $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$ , then  $\hat{\beta}$  is also distributed as  $\text{Normal}(\beta, \Sigma)$ .

- The estimator  $\hat{\beta}$  is unbiased for  $\beta$ .
- The sampling distribution of  $\hat{\beta}$  is normal with the covariance matrix depending on the design matrix and  $\sigma^2$ . But we do not know  $\sigma^2$ !

# Properties of Regression Estimators

$$\hat{\beta} \sim N(\beta, \Sigma)$$

But <sub>unknown</sub> ←



(chi-squared distribution with  $n-p-1$  degrees of freedom)

**Model in Matrix Form:**  $\mathbf{Y} = \mathbf{X} \beta + \varepsilon$

$$\mathbf{x} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}}$$

$$\sim t_{n-p-1} \text{ GT}$$

# Confidence Interval Estimation

We can derive confidence intervals for  $\beta_j$

$$\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p-1} se(\hat{\beta}_j)$$

- **Is  $\beta_j$  statistically significant? Check whether zero is in the confidence interval**
- **Why is this a t-interval?**

# Confidence Interval Estimation

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{se}(\hat{\beta}_j)}} \sim t_{n-p-1}$$

Confidence interval

→ t - interval for

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

Estimate of

Overall Mean: 67.11536  
t critical = 2.2282  
 $\hat{\mu}_{\text{alto}} = 64.89$   
Standard Deviation/Err  
 $\hat{\mu}_{\text{bass}} = 70.72$

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

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

$$SST_R = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_R)^2$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$$

# Testing Statistical Significance

- Test for statistical significance of  $\beta_j$  given all other predicting variables in the model by using the t-test for

$$H_o: \beta_j = 0 \text{ vs. } H_a: \beta_j \neq 0$$

$$t\text{-value} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

We reject  $H_o$  if  $|t|$  gets too large. We interpret this as  $\beta_j$  being statistically significant if the null hypothesis is rejected.

# Testing Statistical Significance

t-value =  $\hat{\beta}_j$  how large to reject  $H_0: \beta_j = b$ ?

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

For significance level  $\alpha$ , Reject if  $|t\text{ value}| > \frac{\alpha}{2, n-p-1}$

Alternatively, compute P-value =  $2P(T_{n-p-1} > |t\text{ value}|)$   
value small (p-value < 0.01)  $\rightarrow$  Reject

How will the procedure change if we test:

$H_0: \beta_j = b$  vs.  $H_a: \beta_j \neq b$  for some known  $b$ ?

# Testing Statistical Significance

What if we want to test for a positive relationship?

$H_0: \beta_j \leq 0$  versus  $H_A: \beta_j > 0$ ?

P-value =  $\Pr(T_{h_{pp1}} > t \text{ value})$

What if we want to test for a negative relationship?

$H_0: \beta_j \geq 0$  versus  $H_A: \beta_j < 0$ ?

P-value =  $\Pr(T_{h_{pp1}} < t \text{ value})$

How will the procedure change if we test:

$H_0: \beta_j = 0$  vs.  $H_A: \beta_j > 0$

OR

$H_0: \beta_j = 0$  vs.  $H_A: \beta_j < 0$

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Testing for Subsets of  
Regression Parameters

# Testing Overall Regression

Analysis of Variance (ANOVA) for multiple regression:

Source	DF	Sum of Sq	Mean SS	F-statistic
Regression	p	SSReg	SSReg/p	MSSReg/ MSE
Residual	n-p-1	SSE	SSE/n-p-1	
Total	n-1	SST		

Where  $SS_{REG} = \sum_{i=1}^n (Y_i - \hat{Y})^2$  and  $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$ .

ANOVA is used to test  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$

We reject  $H_0$  if F-statistic is large ( $F_{p,n-p-1} > F_{\alpha, p, n-p-1}$ ); Which means that at least one of the coefficients is different from zero at the  $\alpha$  significant level. The p-value of the test is:  $P(F_{p,n-p-1} > F_{\text{statistic}})$  where  $F_{p,n-p-1}$  is the F-distribution with p and n-p-1 degrees of freedom.

The p-value of the test is:  $P(F_{p,n-p-1} > F_{\text{statistic}})$  where  $F_{p,n-p-1}$  is the F-distribution with p and n-p-1 degrees of freedom.

# Testing Subsets of Coefficients

## Analysis of Variance (ANOVA):

$$SST(X_1, X_2, \dots, X_p) = SSReg(X_1, X_2, \dots, X_p) + SSE(X_1, X_2, \dots, X_p)$$

$$SSReg(X_1, X_2, \dots, X_p) = SSReg(X_1) + SSReg(X_2|X_1) + SSReg(X_3|X_1, X_2) \\ + \dots + SSReg(X_p|X_1, \dots, X_{p-1})$$

$SSReg(X_1)$  = sum of squares explained by using only  $X_1$  to predict Y

$SSReg(X_2|X_1)$  = **extra sum of squares** explained by using  $X_2$  in addition to  $X_1$  to predict Y

$SSReg(X_3|X_1, X_2)$  = **extra sum of squares** explained by using  $X_3$  in addition to  $X_1$  and  $X_2$  to predict Y

$SSReg(X_p|X_1, \dots, X_{p-1})$  = **extra sum of squares** explained by using  $X_p$  in addition to  $X_1, X_2 \dots X_{p-1}$  to predict Y

# Testing Subsets of Coefficients

**SSReg( $X_1$ ) vs. SSE( $X_1$ ):** Does  $X_1$  alone significantly aid in predicting Y?

**SSReg( $X_2|X_1$ ) vs SSE( $X_1, X_2$ ):** Does the addition of  $X_2$  significantly contribute to the prediction of Y after we account (or control) for the contribution of  $X_1$ ?

**SSReg( $X_3|X_1, X_2$ ) vs. SSE( $X_1, X_2, X_3$ ):** Does the addition of  $X_3$  significantly contribute to the prediction of Y after we account (or control) for the contribution of  $X_1$  and  $X_2$ ?

**SSReg( $X_p|X_1, \dots, X_{p-1}$ ) vs. SSE( $X_1, X_2, \dots, X_p$ ):** Does the addition of  $X_p$  significantly contribute to the prediction of Y after we account (or

# Testing Subsets of Coefficients

**More generally**, consider the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_q Z_q + \varepsilon$$

With two sets of predictors  $(X_1, X_2, \dots, X_p)$  and  $(Z_1, Z_2, \dots, Z_q)$  for example, the first set can represent controlling factors and the second set can represent explanatory factors. We test:

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_q = 0 \text{ versus } H_A : \text{at least one is not zero}$$

**Partial F-test:**  $F_{partial} = \frac{SSE(Z_1, \dots, Z_q, X_1, \dots, X_p) / q}{SSE(Z_1, \dots, Z_q, X_1, \dots, X_p) / (n-p-q-1)}$

We reject  $H_0$  if F-statistic is large ( $F\text{-statistic} > F_{\alpha, q, n-p-q-1}$ ); Which means that at least one of the coefficients is different from zero at the  $\alpha$  significant level.

# Regression Analysis

## Multiple Linear Regression

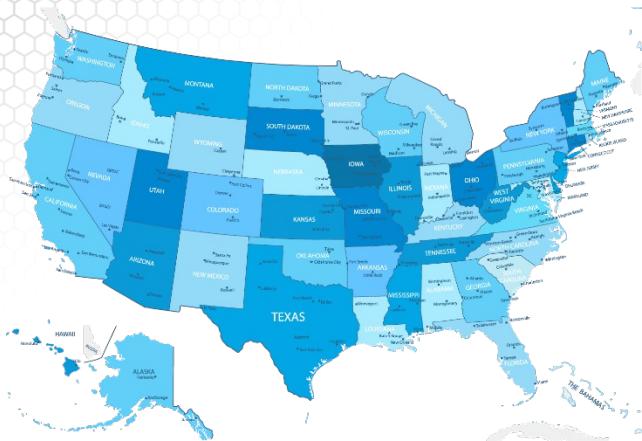
**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Statistical Inference: Data  
Example

# Linear Regression: Example 2



SAT Mean Score by State - Year  
1982  
790 (South Carolina) - 1088 (Iowa)

## Controlling factors:

$X_1$  = % of total eligible students in the state who took the exam

$X_6$  = median percentile of ranking of test takers within their secondary school classes

## Explanatory Factors:

$X_2$  = median income of families of test takers, in hundreds of dollars

$X_3$  = average number of years that test takers had in social sciences, natural sciences, and humanities

$X_4$  = % of test takers who attended public schools

$X_5$  = state expenditure on secondary

# Example 2: Inference on Coefficients

- a. What is the estimate of the coefficient  $\beta_1$  and its variance? Interpret. What is its sampling distribution?
- b. Is the coefficient  $\beta_1$  statistically significant? What is the p-value of the test. Interpret.
- c. What is the F-statistics. Do we reject the null hypothesis that all regression coefficients are zero (the test for overall regression)?
- d. Obtain the 99% confidence interval for  $\beta_1$
- e. Given  $X_1$  and  $X_6$  are controlling factors, test the null hypothesis that the coefficients of the rest of predictors are zero. Clearly state the hypothesis test. Show how you perform the test. Interpret the results.

# Example 2: Inference on Coefficients

```
## Read the data using the 'read.table()' R command
data = read.table("SATData.txt", header = TRUE)
attach(data)
regression.line = lm(sat ~ takers + rank + income + years + public + expend)
summary(regression.line)
```

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-94.659109	211.509584	-0.448	0.656731
takers	-0.480080	0.693711	-0.692	0.492628
rank	8.476217	2.107807	4.021	0.000230 ***
income	-0.008195	0.152358	-0.054	0.957355
years	22.610082	6.314577	3.581	0.000866 ***
public	-0.464152	0.579104	-0.802	0.427249
expend	2.212005	0.845972	2.615	0.012263 *
---				

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

Residual standard error: 26.34 on 43 degrees of freedom

Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618

F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

**a. Estimation & Distribution:**  
 $\hat{\beta}_{takers} = 0.480080$ ;  $t_{takers} = 0.692$ ; T-distribution with 43 degrees of freedom

**b. Test for statistical significance:**  
Sampling Distribution of  $\hat{\beta}_i$ : We do not know  $\sigma^2$ . We can replace it by MSE but then the sampling distribution becomes the t-distribution with 42 df.  
 $t\text{-value} = -0.692$ ,  $p\text{-value} > 0.1$

**c. Test for overall regression**  
 $F\text{-value} = 51.91$ ,  $p\text{-value} \approx 0$

# Example 2: Inference on Coefficients

`confint(regression.line, "takers", level = 0.99)`

	0.5 %	99.5 %
takers	-2.349701	1.389541

## d. Confidence Interval for Regression Coefficients:

$$\beta_{takers}: (-2,349, 1,389)$$

### Interpretation:

- The interval ~~includes zero, so~~ thus it is ~~not~~ plausible that the coefficient is zero given a coefficient of the predicting variable is ~~not~~ the other predicting variables in the model.

# Example 2: Inference on Coefficients

```
regression.line.reduced = lm(sat ~ takers + rank)  
anova(regression.line.reduced, regression.line)
```

Analysis of Variance Table

Model 1: sat ~ takers + rank

Model 2: sat ~ takers + rank + income + years + public + expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	53778				
2	43	29842	4	23935	8.6221	3.35e-05 ***

e. Testing for a subset of regression coefficient efficiencies:

Reduced Model (takers and rank only) Has Full Model

Model

Partial F Test: F-value = 8.6221; P-value  $\approx$  0

Partial F Test: F-value = 8.6221; P-value  $\approx$  0

# Example 2: Inference on Coefficients

Test  $H_0 : \beta_{income} = \beta_{public} = \beta_{year} = \beta_{expend} = 0$

How was the if F-statistic computed:

$$\text{F-statistic} = \frac{\text{SSReg}(Income, public, Years, Expend | Takers, Rank) / 4}{\text{SSE} / (50 - 6 - 1)}$$

The p-value is computed as

$$P(F_{4,43} > \text{F-statistic}) = 1 - P(F_{4,43} < \text{F-statistic})$$

Interpretation: The p-value is approximately 0 thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expend) will be significantly associated to the state-average SAT score.

Partial F Test: F-value = 8.6221; P-value  $\approx 0$

I  
the  
cantly

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Estimating the Regression Line and  
Predicting a New Response

# Estimating the Regression Line

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

$$\hat{Y} | x^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \dots + \hat{\beta}_p x_p^* = x^{*T} \hat{\beta}$$

Because the estimators of  $\beta$  are normally distributed, so is  $\hat{Y}$ .  
That means we can draw inference on the regression line using if we know the expected value and variance.

# Estimating the Regression Line

has a  $\underbrace{\beta_1}_{\text{normal}} + \underbrace{\beta_2}_{\text{normal}} + \dots + \underbrace{\beta_{n-2}}_{\text{normal}}$  distribution with

$$E(\hat{Y} | x^*) = x^{*T} \beta = \beta_0 + \beta_1 x_1^* + \dots + \beta_p x_p^*$$

$$\text{Var}(\hat{Y} | x^*) = \sigma^2 x^{*T} (X^T X)^{-1} x^*$$

If we replace the unknown variance with its estimator (= MSE), the sampling distribution becomes a t-distribution with  $n-p-1$  degrees of freedom.

$$\text{SSE} = \sum_{i=1}^k (y_i - \hat{y}_i)^2$$

# Confidence Interval for Regression Line

The  $(1 - \alpha)$  Confidence interval for the *regression line or mean response* for one instance of predicting variables  $x^*$  is:

$$\hat{y} | x^* \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 x^{*T} (X^T X)^{-1} x^*}$$

The  $(1 - \alpha)$  Confidence surface for all possible instances of the predicting variables :

$$\hat{y} | x^* \pm \sqrt{(p+1)F_{\alpha, p+1, n-p-1}} \sqrt{\hat{\sigma}^2 x^{*T} (X^T X)^{-1} x^*}$$

# Predicting a New Response

One of the primary motivations for regression is to use the regression equation to predict future responses. The prediction is the same as the estimator for the mean response.

---

But a prediction is not the same as the line estimate. The prediction contains two sources of uncertainty:

1. Due to the new observation/s
2. Due to parameter estimates (of  $\beta$ 's)

# Predicting a New Response (cont'd)

1. Variation of the estimated regression line:  $\sigma^2 x^{*T} (X^T X)^{-1} x^*$
  2. Variation of a new measurement:  $\sigma^2$
- 

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

$$\sigma^2 x^{*T} (X^T X)^{-1} x^* + \sigma^2$$

# Predicting a New Response (cont'd)

A  $(1 - \alpha)$  ***prediction*** interval for one new future  $y^*$  (at  $x^*$ ) is

$$x^{*T} \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 (1 + x^{*T} (X^T X)^{-1} x^*)}$$

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

A  $(1 - \alpha)$  ***prediction*** interval for  $m$  new future  $y^*$  (at  $x^*$ ) is

$$\hat{y} | x^* \pm \sqrt{m F_{\alpha, m, n-p-1}} \sqrt{\hat{\sigma}^2 (1 + x^{*T} (X^T X)^{-1} x^*)}$$

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

School of Industrial and Systems Engineering

Estimating Regression Line & Predicting  
a New Response: Data Example

# Linear Regression: Example 1



## Quantitative Predicting Variables:

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales (in thousands of dollars)

## Qualitative Predicting Variable:

$X_5$  = a variable to indicate the size of the territory in which the product is sold

# Example 1: Mean Response & Prediction

- What are the average estimated sales and the standard deviation for all offices with the characteristics as those for the first office? What is the 95% confidence interval for this mean response?
- What sales would you predict for the first office if its largest competitor sales would increase at \$303,000 assuming everything else fixed? What is its standard deviation? What is the 95% prediction interval for this prediction?

# Example 1: Mean Response Estimation

```
## Data for the first office  
newdata = meddcor[1,2:6]
```

```
## Estimate standard deviation  
s2 = summary(model)$sigma^2  
xstar = as.double(newdata)  
xstar = c(1,as.double(newdata))  
X = cbind(rep(1,n),data.matrix(meddcor[,2:6]))  
sqrt(predvar)
```

[,1]

[1,] 29.1095

```
## Confidence Interval  
predict(model, newdata,  
interval="confidence")
```

fit

lwr

upr

1 934.7767 865.0446 1004.509

**b. Average estimated sales  
or mean response for  
sales:**

$\hat{\beta}_1$

**Estimated standard  
deviation:**

se

**95% Confidence Interval:**  
(865.04, 1004.51)

**Interpretation:** For other  
offices with the same  
characteristics as the first  
office, the average estimated  
sales are \$934,770 with a  
lower bound of \$865,040 and  
upper bound of \$1,004,510.

MSE

S

GTx

# Example 1: Mean Response Prediction

```
## Change the competitor's sales  
newdata[4] = 303
```

```
## Estimate standard deviation  
s2 = summary(model)$sigma^2  
xstar = c(1,as.double(newdata))  
X = cbind(rep(1,n),data.matrix(meddcor[,2:6]))  
predvar = s2*(1+xstar%*%solve(t(X)%*%X)%*%xstar)  
sqrt(predvar)  
[1]  
[1,] 63.76933
```

```
## Confidence Interval  
predict(model, newdata, interval="prediction")  
fit lwr upr  
1 911.0569 775.9446 1046.169
```

a. The predicted sales of the office given the higher competitors' sales: The predicted sales of the office given the higher competitors' sales

Estimated standard deviation:

$se(\hat{y})$  Estimated standard deviation

95% Confidence Interval:  $se(\hat{y}) = 63.769$   
(775.94, 1046.16)

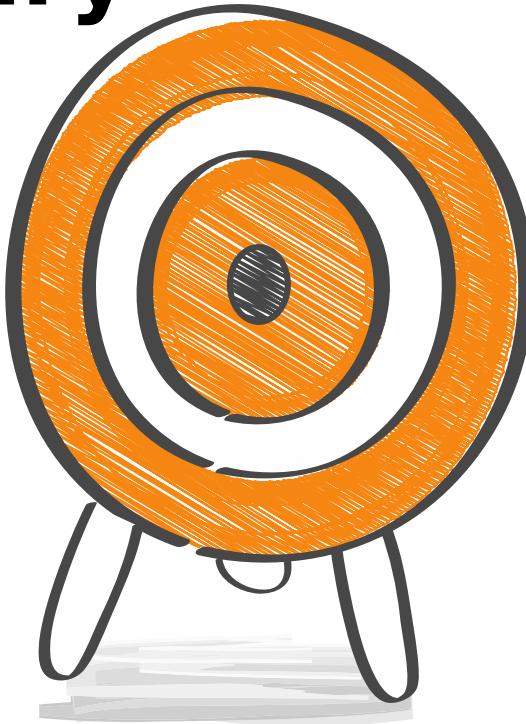
95% Confidence Interval:

Interpretation: If the competitor's sales would increase at \$303,000, the predicted sales reduce with \$23,719.

Since this is prediction, Interpretation: If the predicted sales reduce with the deviation increases.

sales would increase at \$303,000, the predicted sales reduce with the deviation increases. Since this is prediction, the

# Summary



# Regression Analysis

## Multiple Linear Regression

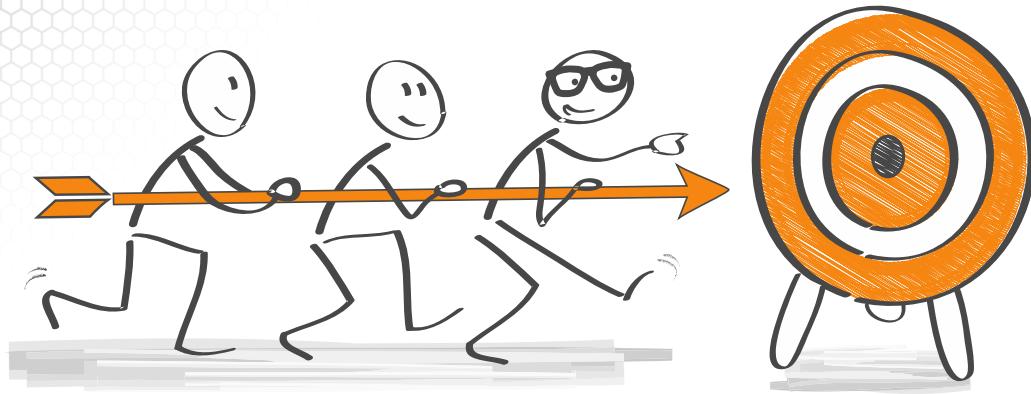
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Assumptions and Diagnostics

# About this lesson



# Multiple Linear Regression: Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

**Assumptions:**

- *Linearity Assumption:* The relationship between  $Y$  and  $X_j$  is linear for all predicting variables
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Properties of the Errors & Residuals

## Properties of (true) errors:

### Properties of (true) errors: $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 I$

### Properties of the (estimated) residuals: $\hat{\varepsilon} = Y - X\hat{\beta}$

- $E(\hat{\varepsilon}) = 0$  (or  $E(\hat{\varepsilon}_i) = 0$ )
- $V(\hat{\varepsilon}) = \sigma^2(I - H)$  ( $V(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$ )

Where  $H$  is the hat matrix and  $h_{ii}$  is the  $i$ th element on its diagonal

- While the true errors have constant variance, the estimated residuals do not.
- To use the estimated residuals for assessing the model assumptions, we need to standardize:

$H_a$  : some means are different

# Residuals Analysis

Standardized Residual Values:  $r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\hat{\sigma}^2 (1-h_{ii})}}$

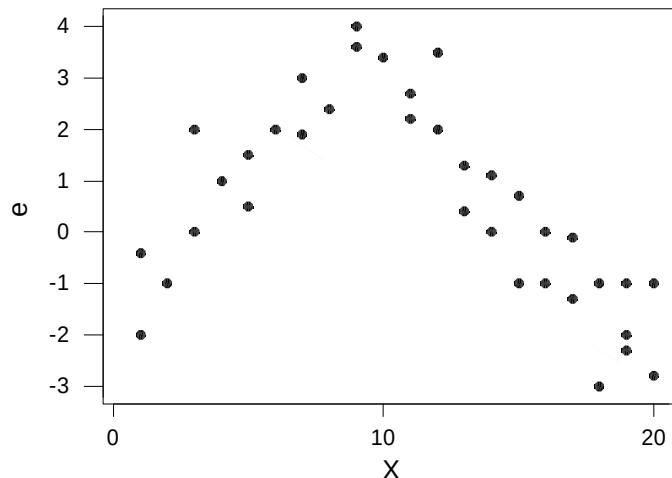
Graphical display: **Plot of the residuals  $r_i$**

- Versus each predictor  Linearity
- Versus fitted values  Constant Variance & Independence
- QQ normal plot & histogram  Normality

- **We evaluate the normality assumption using the residuals not the response variable.**
- **We do not check the predicting variables for normality; however, if the distribution of a predicting variable is strongly skewed, it is possible that the linearity assumption with respect to that variable will not hold.**

# Residual Analysis: Linearity Assumption

each predicting

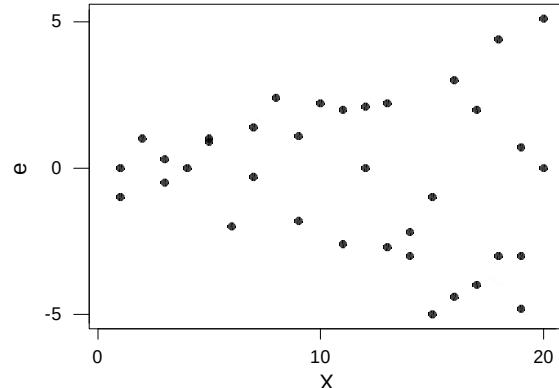


shows that there  
be a non-linear  
ionship between X  
Y.

# Residual Analysis: Constant Variance Assumption

Ch 10

iduals against fitted



residuals show  
increasing variance as the  
independent variable  
increases.

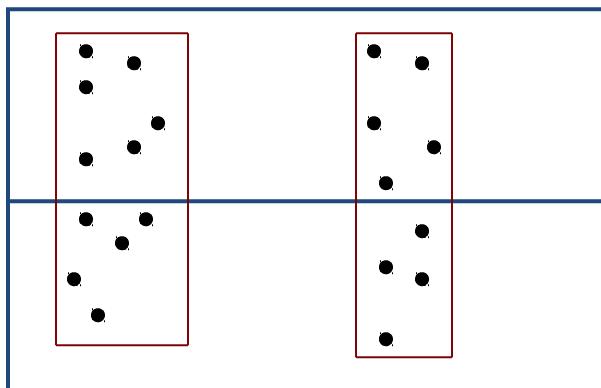


Here, it is an example for which  $\sigma^2$  is not constant.

# Residual Analysis: Constant Variance Assumption

## Independence Assumption:

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



# Residual Analysis: Constant Variance Assumption

## Independence Assumption:

- **Using residual analysis, we are checking for uncorrelated errors but not independence;**
- **Independence is a more complicated matter; if the data are from a randomized experiment, then independence holds; but most data are from observational studies.**
- **We commonly correct for selection bias in observational studies using controlling variables.**

# Checking the Assumption of Normality

One way to check this assumption in a regression is using a **Normal Probability Plot**

y-axis:

$e_i$

x-axis:

$\Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$

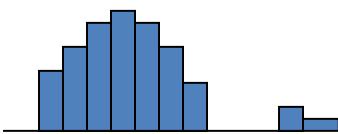
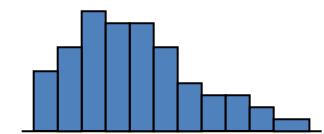
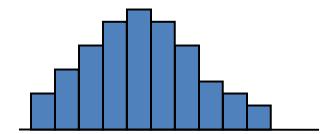
$r_i$  = rank of  $e_i$  (between 1,  $n$ )

$\Phi$  = CDF of Normal Distribution

- Let the R statistical software do this for you!
- A straight line in normal probability plot implies that the assumption is valid
- Curviture (especially at the ends)** shows non-

# Residual Analysis: Normality Assumption

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



## Normality Assumption:

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

# Predicting Variable Transformation

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

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

# Normality Transformation

Problem: Constant variance or/and normality assumption

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

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

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

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

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

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

$$\sigma_y(x) \propto \mu_x^2 \quad \lambda = -1$$

# Outliers in Regression

Any data point that is far from the majority of the data (in x's and y) is called an outlier.

- Data points that are far from the mean of the x's or near the edge of the observation space are called *leverage points*.
- A data point that is far from the mean of both the y and the x's are *influential points* and can change the value of the estimated parameters significantly.

**The upshot:** Sometimes there are good reasons for excluding subsets (there were errors in the data entry; there were errors in the experiment). Sometimes - the outlier belongs in the data. Outliers should always be examined.

# Checking for Outliers

**Cook's Distance:**

$$D_i = \frac{(Y_i - \hat{Y})^T (Y_{(i)} - \hat{Y})}{e_i^2}$$

Where  $\hat{Y}_{(i)}$  are the fitted values from the model fitted without the  $i$ -th observation (i.e. observations excluding  $i$ -th observation) and the fitted values from the model fitted with the  $i$ -th observation (all observations).

**It measures how much all the values in the regression model change when the  $i_{th}$  observation is removed.**

**Rule of Thumb:**  $D_i > 4/n$  or  $D_i > 1$  or any “large”  $D_i$  should be investigated.

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Assumptions and Diagnostics:

Data Example

# Linear Regression: Example 1



## Quantitative Predicting Variables:

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

## Qualitative Predicting Variable:

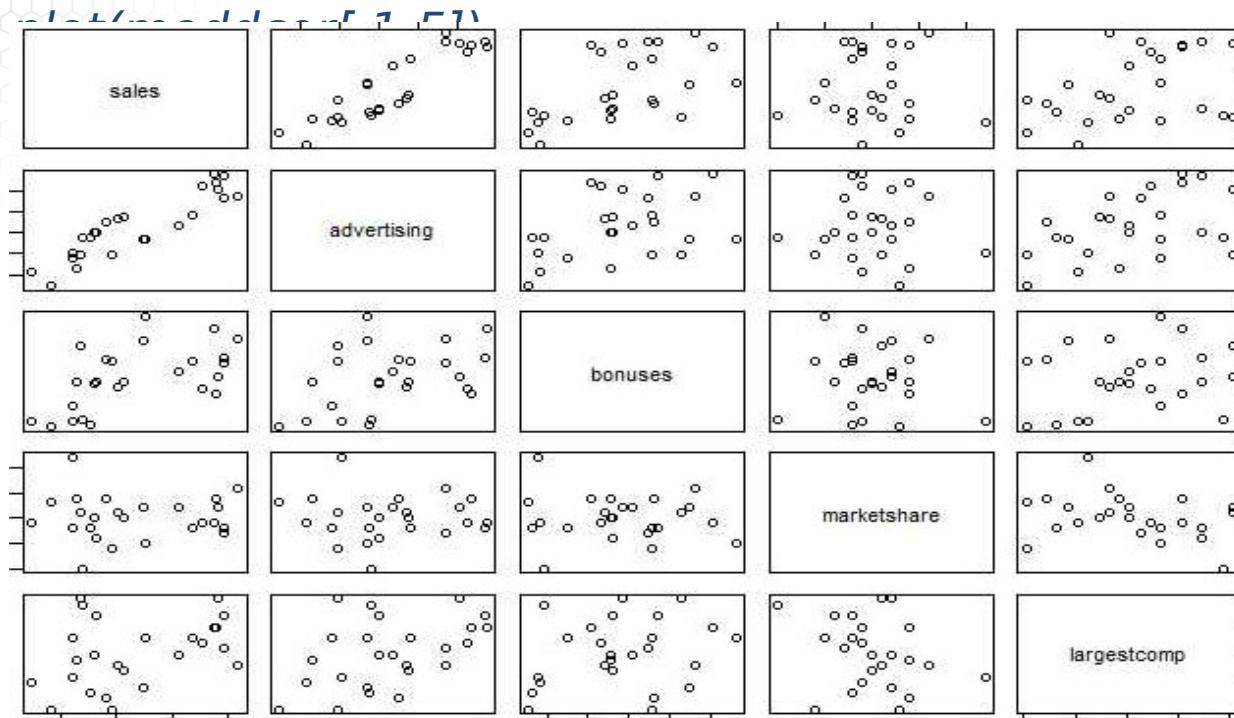
$X_5$  = a variable to indicate the region in which territory is located (1 = U.S., 2 = Europe, 3 = Asia)

# Residual Analysis: Example 1

- a. Do the assumptions hold? Provide the graphical displays needed to support the diagnostics. Interpret.
- b. If one or more assumptions do not hold, what transformations do you suggest? Did the residual diagnoses improved with the suggested transformations?
- c. Do you identify any outliers?

# Linearity Assumption

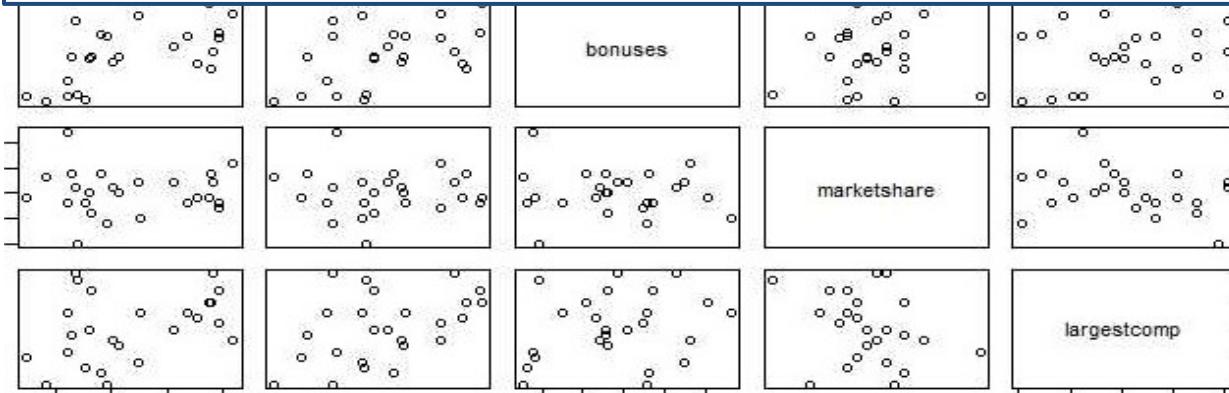
## Scatter plot matrix of sales and numeric predicting variables



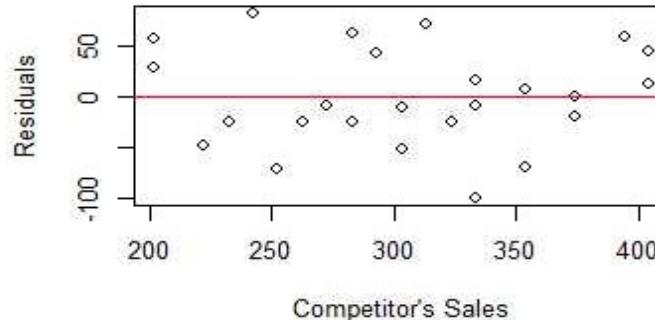
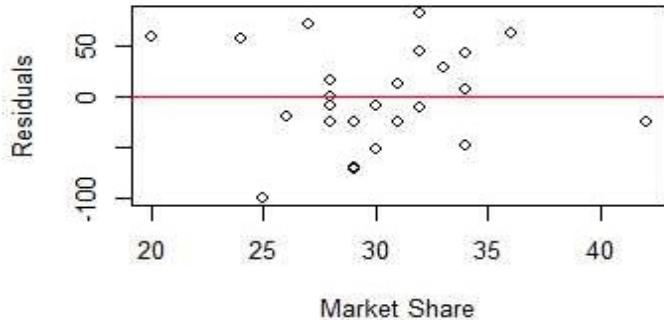
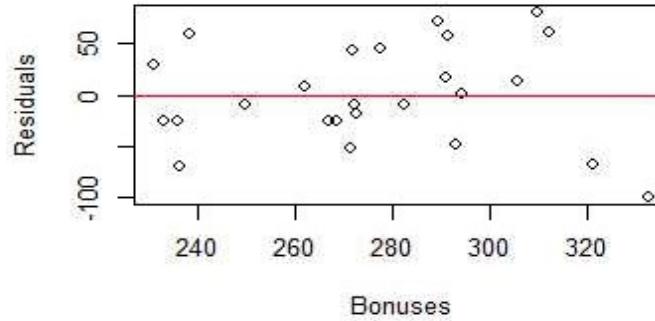
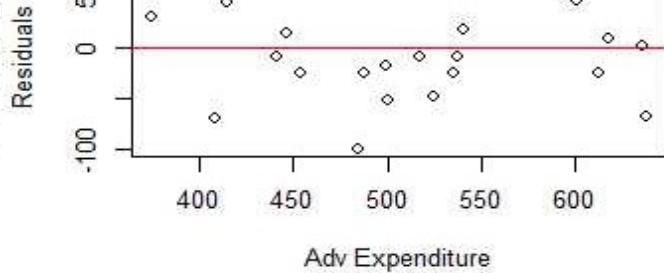
# Linearity Assumption

## Scatter plot matrix of sales and numeric predicting variables

- **Linearity assumption holds for all predicting variables;**
- **For advertisement expenditure, bonus amount and competitor's sales, the relationship with sales is strongly linear.**



# Linearity Assumption

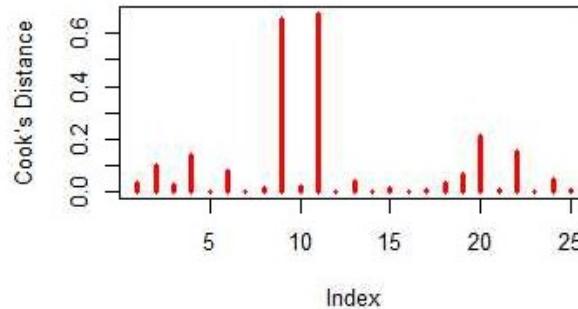
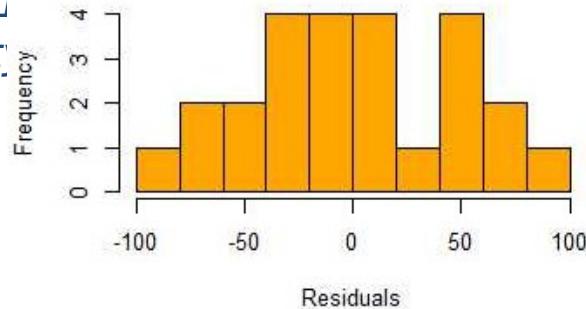
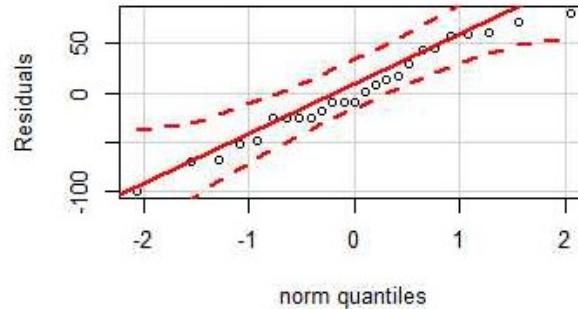
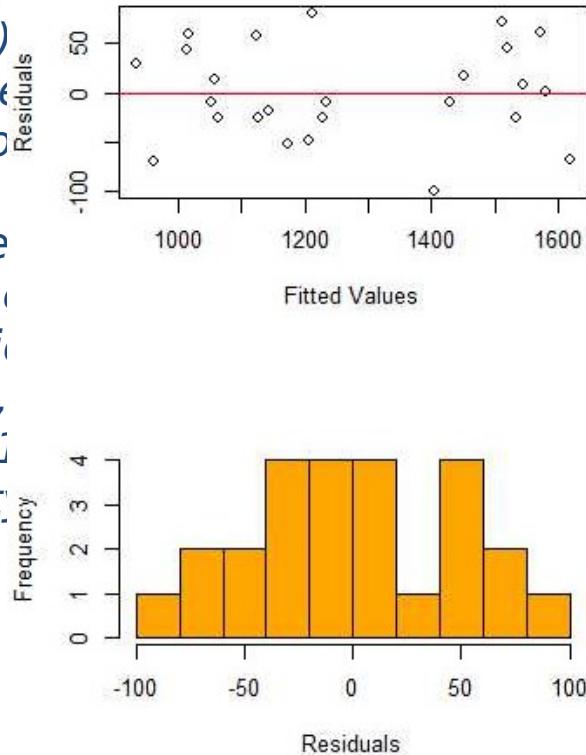


# Residual Analysis: Other Assumptions

```
library(car)
fits = model$fitted
cook = cooks.distance(model)
par(mfrow =c(2,2))
plot(fits, resids, xlab="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
qqPlot(resids, ylab="Residuals", main = "")
hist(resids, xlab="Residuals", main =
"",nclass=10,col="orange")
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's
Distance")
```

# Residual Analysis: Other Assumptions

```
library(car)
fits = model
cook = cook
par(mfrow
plot(fits, re
abline(0,0,
qqPlot(resi
hist(resids,
","",nclass=1
plot(cook,t
Distance")
```



# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Evaluation and Multicollinearity

# R<sup>2</sup> Coefficient of Determination

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

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

which is interpreted as

$R^2$  Proportion of total variability in  $Y$  that can be explained by the regression model

$$SSE = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

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

# Model Evaluation

1. *F-test* for  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

$$F_0 = MSR / MSE \sim F(k, n-k-1)$$

$$MSR = SSR/k \quad MSE = SSE/n-k-1$$

2. *Coefficient of determination:*

$$R^2 = SSR/SST$$

- $R^2$  will always increase if we add more predicting variables

3. *Adjusted Coefficient of determination:*

$$\text{adjusted } R^2 = 1 - (n-1)(1-R^2)/(n-k-1)$$

# Correlation Coefficient

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

$$\rho = \text{cor}(X_j, Y) = \frac{\sum_{i=1}^n y_i(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- It can be used to evaluate the linear relationship between the response variable and the predicting variables;
- It can also be used to evaluate the correlation between the predicting variables for detecting (near) linear dependence among the variables (or multicollinearity).

# Multicollinearity

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ : find values minimizing squared error:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

~~if  $\mathbf{X}^T \mathbf{X}$  invertible~~  
 ~~$\mathbf{X}^T \mathbf{X}$  is not invertible if and only if the columns of  $\mathbf{X}$  are linearly dependent, i.e. one variable is a linear combination of the others.  $\beta_0, \beta_1$~~

- Implication: the standard error of  $\beta_1$  is infinite.

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{x} + \varepsilon$$

**Near Collinearity:** columns of  $\mathbf{X}$  are approximately linearly dependent

- ~~Equivalently, estimating~~ The standard error of  $\beta_1$  is artificially large;
- ~~1. If the Intercept~~ If one value of one of the predicting variables is changed only by slightly, the fitted regression coefficients can change dramatically;
- ~~2. If the Slope~~  $\varepsilon$  is the deviance of the data from the linear model The overall F statistic may be significant, yet each of the individual t-statistics is not significant.
- Prediction is also affected since the relationship to the

# Multicollinearity Diagnosis

The variance inflation factor(VIF) for each predicting variable:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where  $R_j$  is the coefficient of variation of the regression of the variable  $j$  on all other predicting variables.

Interpretation: VIF measures the proportional increase in the variance of  $\hat{\epsilon}_j$  compared to what it would have been if the predicting variables had been completely uncorrelated. How small a VIF indicates the collinearity is not present?

- $E(\hat{\epsilon}) = 0$  (or  $E(\hat{\epsilon}_i) = 0$ )  $VIF < \max(10, \frac{1}{1 - R_{model}^2})$

Where  $R$  is the coefficient of variation of the regression model.

$$\bullet V(\hat{\epsilon}) = \sigma^2 (I - H) (V(\hat{\epsilon}_i) = \sigma^2 (1 - h_{ii}))$$

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Multicollinearity: Data Example

# Linear Regression: Example 1



## Quantitative Predicting Variables:

$X_1$  = the amount (in hundreds of dollars) spent on advertising in 1999

$X_2$  = the total amount of bonuses paid in 1999

$X_3$  = the market share in each territory

$X_4$  = the largest competitor's sales

## Qualitative Predicting Variable:

$X_5$  = a variable to indicate the region in which territory is located (1 = south, 2 = west, 3 = midwest)

# Model Evaluation: Example 1

- a. What are the correlation coefficients between the quantitative predicting variables? Any potential multicollinearity?
- b. Obtain the variance inflation factors for the quantitative predicting variables. Any potential multicollinearity?
- c. What is the coefficient of determination? Interpret.

# Model Evaluation: Example 1

`cor(meddcor[,2:5])`

	advertising	bonuses	marketshare
largestcomp			
advertising	1.00000000	0.41868215	-0.02029937
0.4524897			
bonuses	0.41868215	1.00000000	-0.08484673
0.2286563			
marketshare	-0.02029937	-0.08484673	1.00000000
0.2872159			
largestcomp	0.45248974	0.22865628	-0.28721592
1.0000000			

`vif(model)`

	GVIF	Df	$GVIF^{1/(2*Df)}$
advertising	3.081657	1	1.755465
bonuses	1.359601	1	1.166019
marketshare	1.311265	1	1.145105
largestcomp	1.569851	1	1.252937
region	3.784660	2	1.394783

**The maximum correlation between predicting variables is 0.452.**

**None of the vifs are greater than  $\max(10, 1/(1-)) = 22.22$ .**

**The coefficient of determination is 0.955. Thus the model explains 95.5% of the variability in the sales.**

Sam

GTx

# Regression Analysis

## Multiple Linear Regression

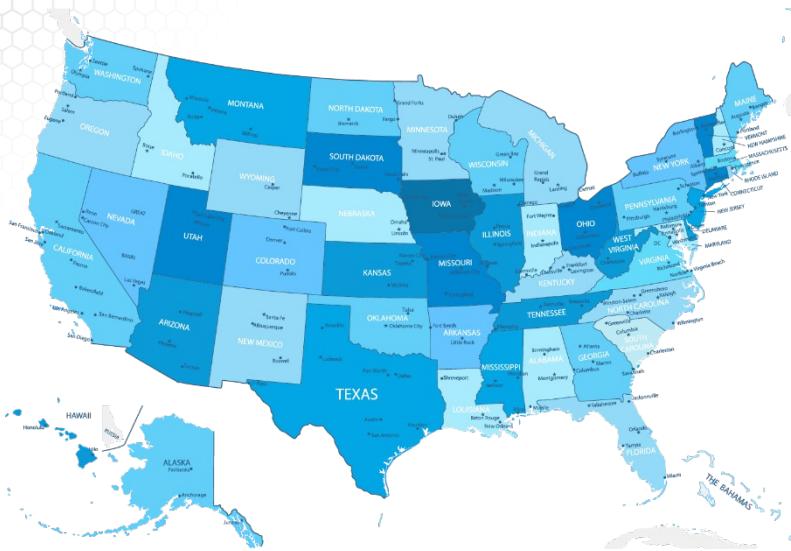
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Ranking States by SAT  
Performance: Exploratory Analysis

# Ranking States by SAT Performance



SAT Mean Score by State - Year 1982

790 (South Carolina) -1088  
(Iowa)

*Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?*

# Response & Predicting Variables

**The response variable is:**

**Y** = State average SAT score (verbal and quantitative combined)

**The predicting variables are:**

**“takers”**: % of total eligible students (high school seniors) in the state who took the exam

**“rank”** median percentile of ranking of test takers within their secondary school classes

**“income”**: median income of families of test takers, in hundreds of dollars

**“years”**: average number of years that test takers had in social sciences, natural sciences, and humanities

**“public”**: % of test takers who attended public schools

**“expend”**: state expenditure on secondary schools, in hundreds of dollars per student

# Controlling Variables

## Selection Bias:

The states with high average SAT score had low percentages of takers. Those taking the test will tend to be in the higher median percentile of ranking of test takers within their secondary school classes.

## Controlling factors:

**“takers”**: % of total eligible students in the state who took the exam

**“rank”**: median percentile of ranking of test takers within their secondary school classes

# Read the Data in R

```
## Read the data using the 'read.table()' R command because it is an ASCII file
```

```
data = read.table("SATData.txt", header = TRUE)
```

```
## Check data to make sure correctly read in R
```

```
data[1:4,]
```

	state	sat	takers	income	years	public	expend	rank
1	Iowa	1088	3	326	16.79	87.8	25.60	89.7
2	SouthDakota	1075	2	264	16.07	86.2	19.95	90.6
3	NorthDakota	1068	3	317	16.57	88.3	20.62	89.8
4	Kansas	1045	5	338	16.30	83.9	27.14	86.3

```
## Check dimensionality of the data file
```

```
dim(data)
```

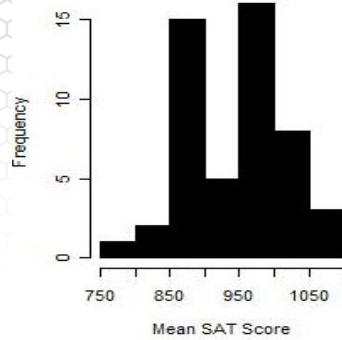
```
## Attach data to automatically recognize the columns in the data as individual vectors
```

```
attach(data)
```

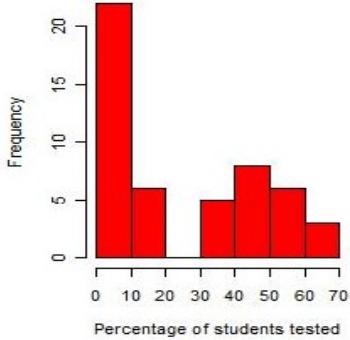
The data consist of 50 rows, each corresponding to a U.S. state.

# Exploratory Data Analysis in R

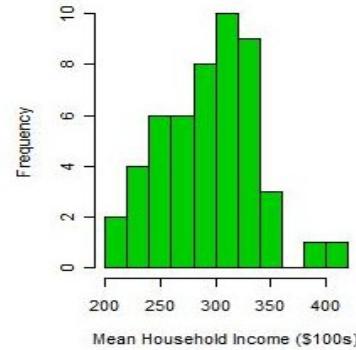
Histogram of SAT Scores



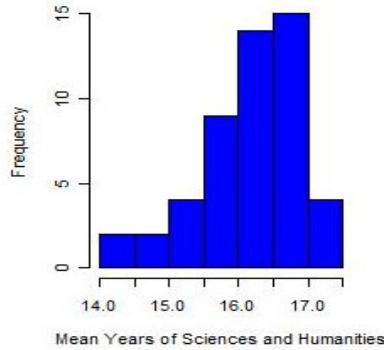
Histogram of Takers



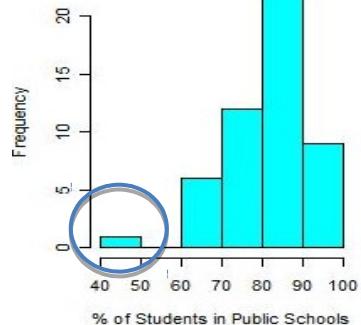
Histogram of Income



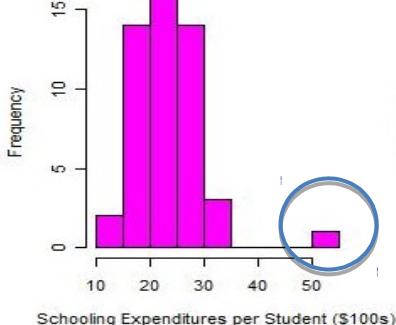
Histogram of Years



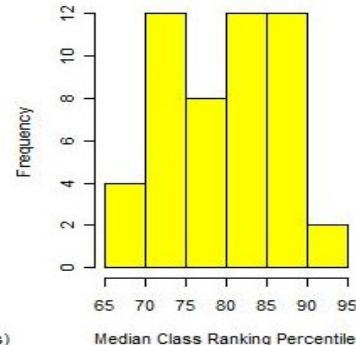
Public Schools Percentage



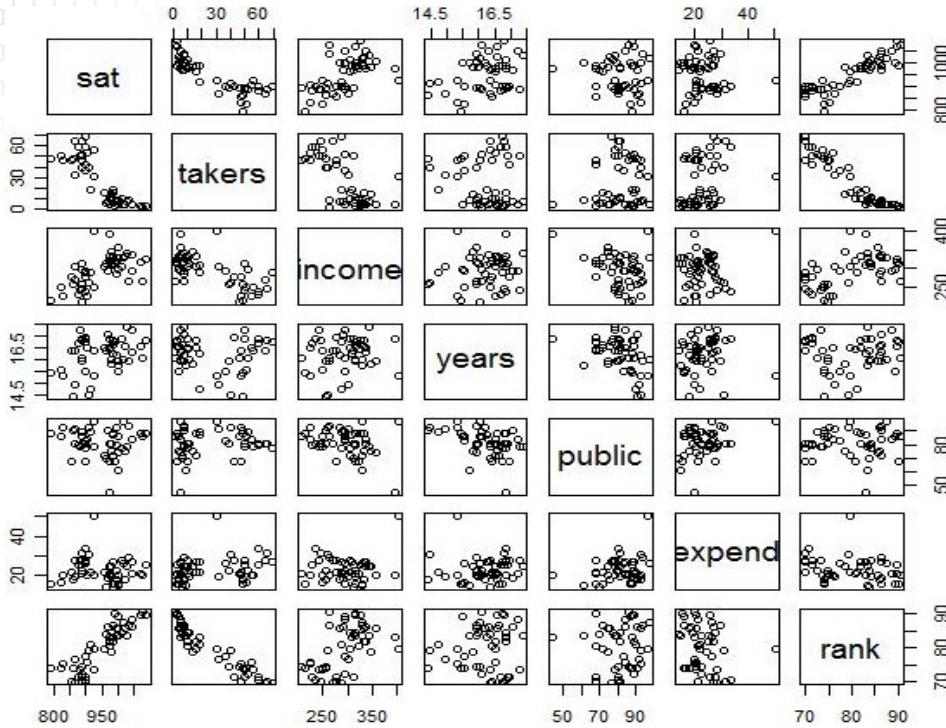
Histogram of Expenditures



Histogram of Class Rank



# Exploratory Data Analysis in R (Cont'd)



	sat	takers	income	years	public	expend	rank
sat	1.00	-0.86	0.58	0.33	-0.08	-0.06	0.88
takers	-0.86	1.00	-0.66	-0.10	0.12	0.28	-0.94
income	0.58	-0.66	1.00	0.13	-0.31	0.13	0.53
years	0.33	-0.10	0.13	1.00	-0.42	0.06	0.07
Public	-0.08	0.12	-0.31	-0.42	1.00	0.28	0.05
expend	-0.06	0.28	0.13	0.06	0.28	1.00	-0.26
rank	0.88	-0.94	0.53	0.07	0.05	-0.26	1.00

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Ranking States by SAT  
Performance: Regression Analysis

# Linear Regression Analysis in R

```
regression.line = lm(sat ~ takers + rank + income + years + public + expend)  
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-94.659109	211.509584	-0.448	0.656731
takers	-0.480080	0.69371	-0.692	0.492628
rank	8.476217	2.10780	4.021	0.000230 ***
income	-0.008195	0.15235	-0.054	0.957353
years	22.610082	6.31457	3.581	0.000866 ***
public	-0.464152	0.57910	-0.802	0.427249
expend	2.212005	0.84597	2.615	0.012263 *

Residual standard error: 26.34 on 43 degrees of freedom  
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618  
F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

Test for statistical significance:

- : t-value= -0.692, p-value>0.1
- : t-value= -0.692, p-value <0.01
- : t-value= -0.054, p-value>0.1
- : t-value= 3.581, p-value<0.01
- : t-value= -0.802, p-value>0.1
- : t-value= 2.615, p-value = 0.012

$$H_0: \mu = 26.34, H_1: \mu \neq 26.34$$

Multiple R-squared: 0.8787, Adjusted R-squared: 0.8618

Residual standard error: 26.34 on 43 degrees of freedom

F-statistic: 51.91 on 6 and 43 DF, p-value: < 2.2e-16

# Testing for Subsets of Coefficients

Test:  $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

How was the F-statistic computed?

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

F-statistics =

The p-value is computed as

$$P(\text{F-statistic} \geq \text{observed})$$

$H_0: \text{all means are equal}$

$H_a: \text{some means are different}$

Interpretation: The p-value is approximately 0, thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expend) will be significantly associated to the state-average SAT score.

pvalue

[1] 3.349778e-05

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Ranking States by SAT Performance  
Using Residuals

# Using Residuals to Create Better Rankings

**Bias Selection:** Some state universities require the SAT and some require a competing exam. States with a high proportion of takers probably have “in state” requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing a bias.

```
## Consider the model with the two controlling factor to correct for bias
reduced.line = lm(sat ~ takers + rank)
## obtain the order of states by the residuals of the reduced model
order.vec = order(reduced.line$res, decreasing = TRUE)
## Re-order the states and create a table including state name, new and old order.
states = factor(data[order.vec, 1])
newtable = data.frame(State = states, Residual = as.numeric(round(reduced.line$res[order.vec], 1)),
oldrank = (1:50)[order.vec])
```

# Using Residuals to Create Better Rankings

	State	Residual	Oldrank	After controlling for selection bias, Connecticut moved from 35 <sup>th</sup> to 1 <sup>st</sup> .				
1	Connecticut	53.9	35					
2	Iowa	53.5	1	43	Arkansas	-31.2	12	
3	New Hampshire	45.8	28	44	West Virginia	-38.9	25	
4	Massachusetts	41.9	41	45	Nevada	-45.4	30	
5	New York	40.9	36	46	Mississippi	-49.3	16	
6	Minnesota	40.6	7	47	Texas	-50.3	45	
After controlling for selection bias, Mississippi moved from 16 <sup>th</sup> to 46 <sup>th</sup> .				48	Georgia	-63.0	49	
				49	North Carolina	-71.3	48	
				50	South Carolina	-98.5	50	

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Ranking States by SAT Performance:  
Model Fit

# Residual Analysis

```
## Residual analysis for the reduced model
res = reduced.line$res
cook = cooks.distance(reduced.line)
par(mfrow = c(1,3))
plot(sat, res, xlab = "SAT Score", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(takers, res, xlab = "Percent of Students Tested", ylab = "Residuals", pch = 19)
abline(h = 0)
plot(rank, res, xlab = "Median Class Ranking Percentile", ylab = "Residuals", pch = 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnorm(res)
qqline(res)
plot(cook,type="h",lwd=3, ylab = "Cook's Distance")
```

# Residual Analysis

```
## Residual analysis for the reduced model
res = reduced.line$res
cook = cooks.distance(reduced.line)
par(mfrow = c(1,3))
plot(sat, res, xlab = "SAT Score", ylab = "Residuals", pch = 19)
abline(h = 0)
```

**Constant Variance & Uncorrelated Errors:** Response Variable or Fitted Values vs Residuals

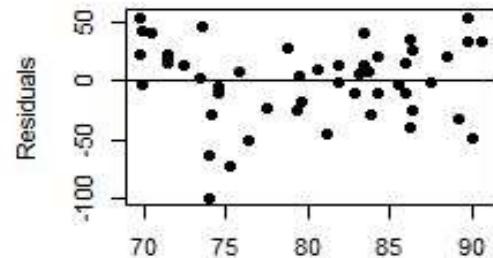
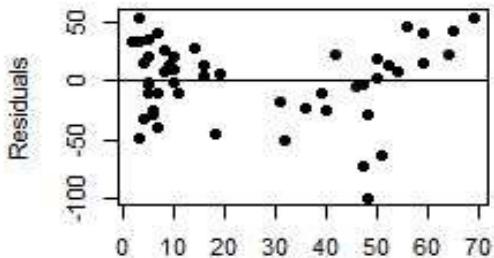
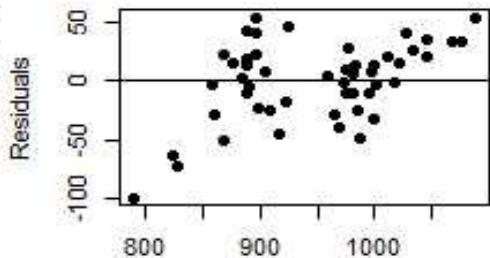
**Linearity:** Predicting Variables vs Residuals

**Normality:** Histogram and QQ normal plot

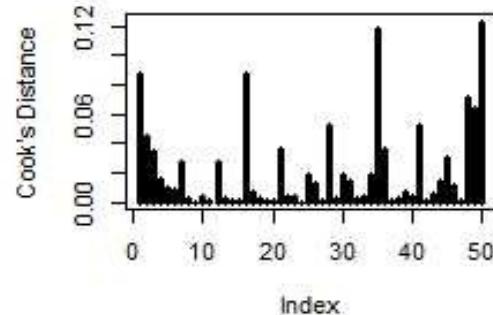
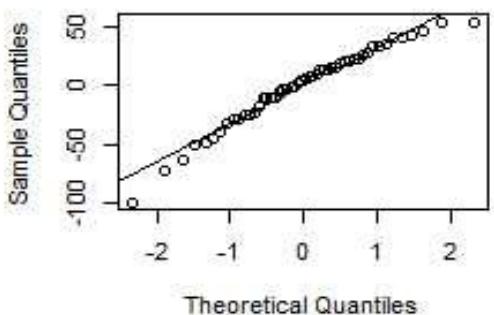
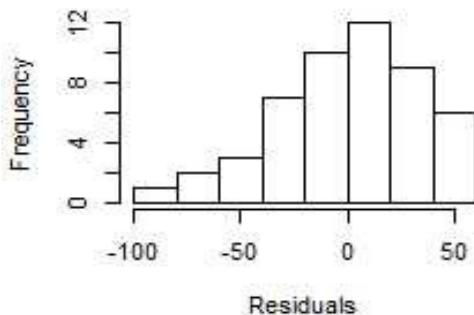
**Outliers:** Cook Distance Plots

```
plot(cook,type="h",lwd=3, ylab = "Cook's Distance")
```

# Residual Analysis



- Transform the predicting variable: Percent of Students Tested (takers)
- Heavy tailed residuals



# Linear Regression Analysis in R

```
regression.line = lm(sat ~ log(takers) + rank + income + years + public + expend)  
summary(regression.line)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	407.53990	282.76325	1.441	0.15675
log(takers)	-38.43758	15.95214	-2.410	0.02032 *
rank	4.11427	2.50166	1.645	0.10734
income	-0.03588	0.13011	-0.276	0.78407
years	17.21811	6.32007	2.724	0.00928 **
public	-0.11301	0.56239	-0.201	0.84168
expend	2.56691	0.80641	3.183	0.00271 **

---

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

Residual standard error: 24.86 on 43 degrees of freedom

Multiple R-squared: 0.8919, Adjusted R-squared: 0.8769

F-statistic: 59.15 on 6 and 43 DF, p-value: < 2.2e-16

Test for statistical significance:

• p-value = 0.02

Properties of true errors:  $E(\varepsilon) = 0, \text{Var}(\varepsilon) = \sigma^2 I$

• p-value > 0.1

Properties of the estimated residuals:  $\hat{\varepsilon} = Y - X\hat{\beta}$

• p-value > 0.1

•  $E(\hat{\varepsilon}) = 0$  (or  $E(\hat{\varepsilon}^2) = 0$ )

•  $\text{Var}(\hat{\varepsilon}) = (X^T X)^{-1} \sigma^2 I$

• p-value < 0.01

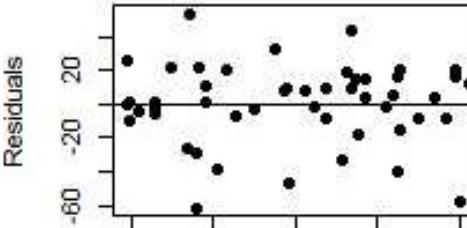
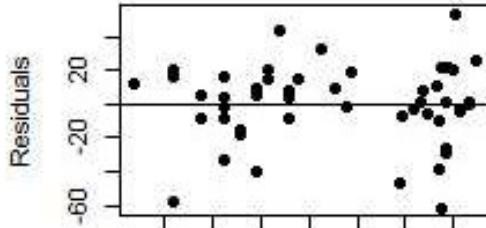
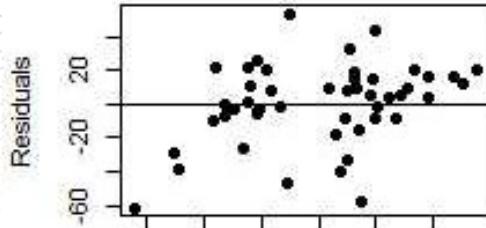
•  $\text{p-value} > 0.01$

•  $\text{p-value} < 0.001$

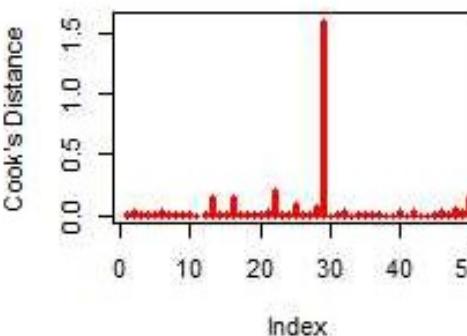
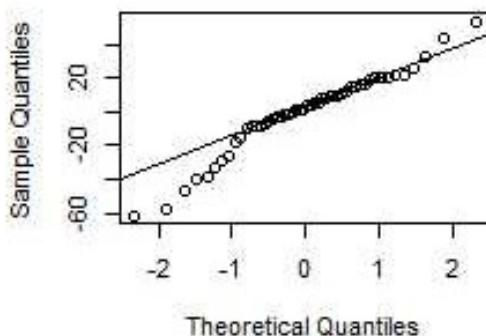
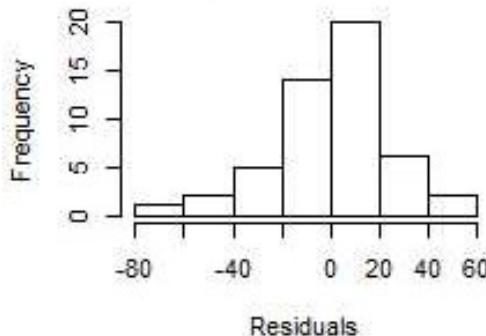
• Where  $X$  is the hat matrix and  $I$  is the  $i$ -th element on its diagonal

$\hat{\beta}_1 = \frac{24.86}{43}$ ,  $n-7 = 43$   
but  $x_1$  fixed  $\rightarrow \hat{\beta}_1 = c_1$  fixed  
variability explained

# Residual Analysis



- Transformation has improved on the linearity assumption
- Heavy tailed residuals remains
- Cook's Distance: Alaska is an outlier/influential point for the model



# State SAT Performance: Findings

Given all other predictors in the model, percent of students taking SAT from a public school and family income of test takers are not statistically significantly associated to SAT score;

- Given all other predictors in the model, with 100,000 increase in the expenditure on secondary school results in a 2.56 points increase in the SAT score;
- Given all other predictors in the model, one additional year that test takers had in social sciences, natural sciences, and humanities leads to 17.2 points increase in the SAT score;
- The predictors in the model explain close to 90% of the variability in SAT score;
- We find that the relationship between state- average SAT score and the percent of students taking SAT to be nonlinear;
- Ranking changes after controlling for the bias selection factors; Connecticut moves up to be first from 35<sup>th</sup>; Massachusetts to 4<sup>th</sup> from 41<sup>st</sup>; and New York to 5<sup>th</sup> from 36<sup>th</sup>.

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Prediction of IMDb Movie Ratings:  
Exploratory Analysis

# Prediction of IMDb Movie Ratings



# Response & Predicting Variables

**The response variable is:**

**Y** = the rating provided by the IMDb, scaled with values between 0 to 100

**Quantitative predicting variables are:**

**X<sub>1</sub>** = Number of votes for the movie on the IMDb platform

**X<sub>2</sub>** = Duration of the movie

**X<sub>3</sub>** = Gross earnings scaled in 1000's

**X<sub>4</sub>** = Total budget in millions

# Response & Predicting Variables (cont'd)

**Qualitative predicting variables are:**

$X_5$  = Release year (between 2010-2014)

$X_6$  = Rating of a film's suitability for certain audiences, based on its content (G, PG, PG-13, R)

$X_7$  = Language: English (1) and Other languages (0)

$X_8$  = Genre: Action (1), Documentary (2), Comedy (3), Horror, Sci-Fi (4)

$X_9$  = Director Rating: Awarded (1), Nominated (2), None (3)

$X_{10}$  = Actor Rating: Classified into two groups based on their performance, with 0 for low ranking and 1 for high ranking.

$X_{11}$  = Movie Awards: Awarded (1), Nominated (2), None (3)

Why do we consider 'Year' as a qualitative variable?

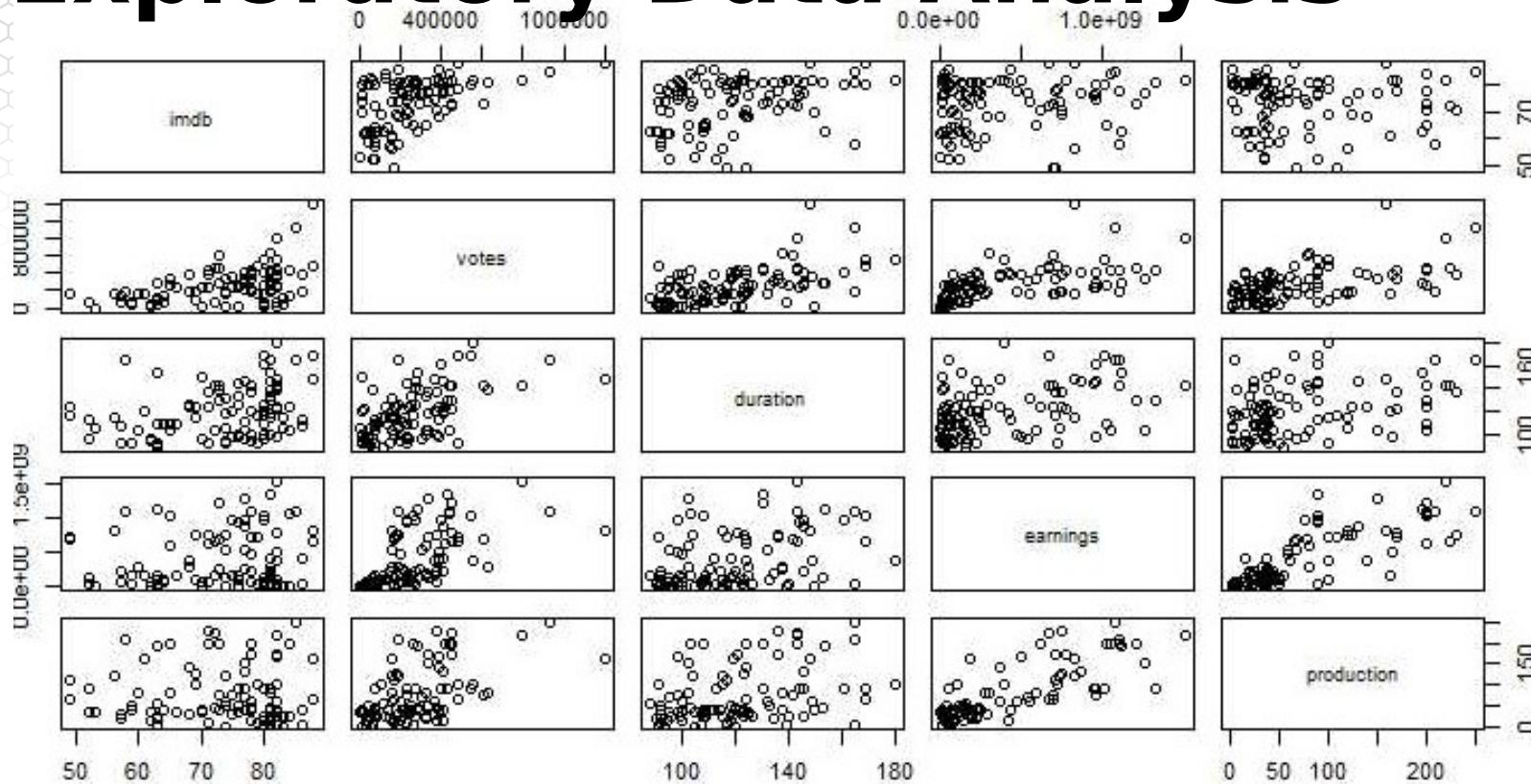
# Exploratory Data Analysis in R

```
## Read data using read.csv  
data = read.csv("training.csv",header=TRUE,sep=",")  
## how many observations?  
dim(data)[1]  
[1] 100
```

```
## Response Variable: scaled between 0 and 100  
imdb=data$imdb*10  
## Quantitative Predicting Variables  
# Number of imdb user votes for the movie  
votes=data$votes  
# The duration of the movie  
duration=data$time  
# Gross earnings in 1000's  
earnings=data$boxoffice/1000  
# Total budget in millions  
production=data$productionbudget
```

```
## Exploratory Data Analysis for Quantitative Data  
# Scatter plot matrix  
pqmat=as.data.frame(cbind(imdb,  
votes,duration,earnings,production))  
plot(pqmat)
```

# Exploratory Data Analysis



# Exploratory Data Analysis in R (cont'd)

```
table(data$year)
year=as.factor(data$year)
rating=as.factor(data$rating)
language=data$language
language[language==1] = "English"
language[language==0] = "Other"
language = as.factor(language)
genre=data$genre
genre[genre==1]= "Action"
genre[genre==2]= "Documentary"
genre[genre==3]= "Comedy"
genre[genre==4]= "Horror, Sci-Fi"
genre = as.factor(genre)
rtdirector=data$directorrating
rtdirector[rtdirector==1]="Awarded"
rtdirector[rtdirector==2]="Nominated"
rtdirector[rtdirector==3]="None"
rtdirector = as.factor(rtdirector)
```

```
rtactor=as.factor(data$actorrating)
awards=data$movieaward
awards[awards==1]="Awarded"
awards[awards==2]="Nominated"
awards[awards==3]="None"
## Exploratory Data Analysis for Qualitative Data
par(mfrow=c(2,3))
boxplot(imdb~year,col="blue",main="Year")
boxplot(imdb~rating,col="red",main="Rating for Audience")
boxplot(imdb~language,col="green",main="Language")
boxplot(imdb~genre,col="purple",main="Genre")
boxplot(imdb~rtdirector,col="purple",main="Director Awards")
boxplot(imdb~rtactor,col="grey",main="Actor Performance Rating")
```

# Exploratory Data Analysis in R (cont'd)

```
table(data$year)
```

```
rtactor=as.factor(data$actorrating)
```

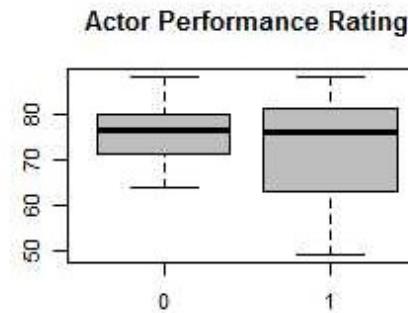
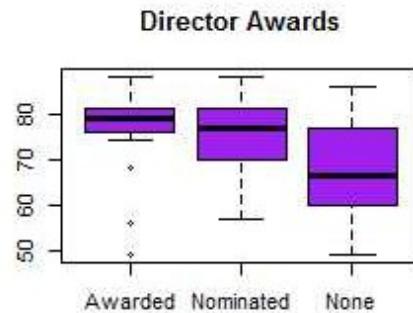
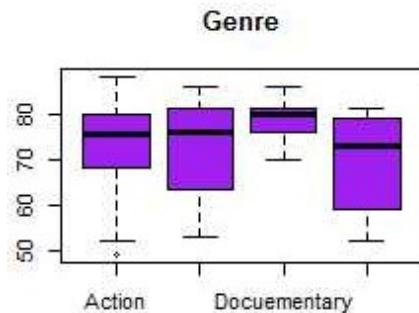
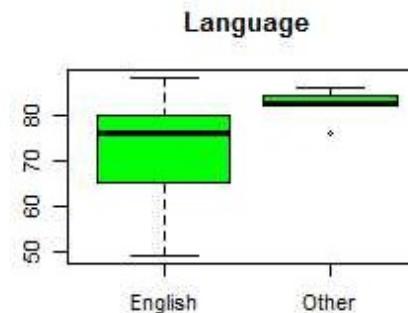
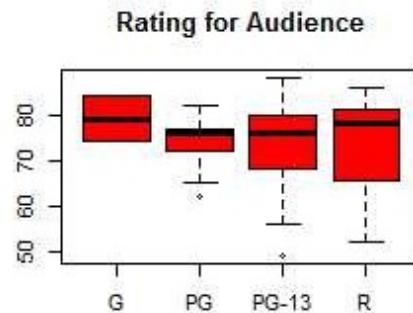
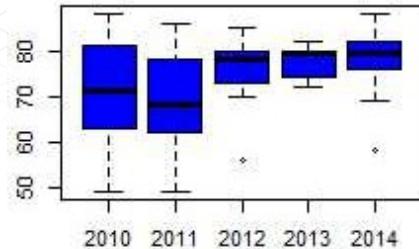
## Quantitative Variables:

- For better visual analytics and reference, inform quantitative variables using their specific class names
- Transform into a categorical variable using the as.factor() R command.

```
genre=data$genre
genre[genre==1]= "Action"
genre[genre==2]= "Documentary"
genre[genre==3]= "Comedy"
genre[genre==4]= "Horror, Sci-Fi"
genre = as.factor(genre)
rtdirector=data$directorrating
rtdirector[rtdirector==1]="Awarded"
rtdirector[rtdirector==2]="Nominated"
rtdirector[rtdirector==3]="None"
rtdirector = as.factor(rtdirector)
```

```
boxplot(imdb~year,col="blue",main="Year")
boxplot(imdb~rating,col="red",main="Rating for Audience")
boxplot(imdb~language,col="green",main="Language")
boxplot(imdb~genre,col="purple",main="Genre")
boxplot(imdb~rtdirector,col="purple",main="Director Awards")
boxplot(imdb~rtactor,col="grey",main="Actor Performance Rating")
```

# Exploratory Data Analysis (cont'd)



# Exploratory Data Analysis in R

```
## Correlation, Multicollinearity
```

```
## Quantitative data
```

```
round(cor(pqmat),2)
```

	imdb	votes	duration	earnings	production
imdb	1.00	0.46	0.33	0.09	-0.04
votes	0.46	1.00	0.52	0.61	0.56
duration	0.33	0.52	1.00	0.42	0.38
earnings	0.09	0.61	0.42	1.00	0.80
production	-0.04	0.56	0.38	0.80	1.00

```
## Qualitative data: Response vs. Predicting Variables
```

```
summary(aov(imdb~year))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
year	4	1513	378.3	4.931	0.00118 **
Residuals	95	7288	76.7		

```
## Qualitative predicting variables
```

```
table(rtdirector,awards)
```

		awards		
rtdirector	Awarded	Awarded	Nominated	None
Awarded	13		10	4
Nominated	10		23	10
None	1		10	19

```
chisq.test(rtdirector,awards)
```

Pearson's Chi-squared test

data: rtdirector and awards

X-squared = 26.192, df = 4, p-value = 2.895e-05

# Exploratory Data Analysis in R

## Exploratory Analysis using Numerical Summaries:

- Correlation captures linear dependence between quantitative variables;
- ANOVA can be used to assess whether the means of the response variable are statistically different with respect to a qualitative variable;
- The ‘table’ command in R provides the contingency table between any two qualitative variables.
- The Pearson chi-square test can be used to test for “correlation” between qualitative variables.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
year	4	1513	378.3	4.931	0.00118 **	data: rtdirector and awards
Residuals	95	7288	76.7			X-squared = 26.192, df = 4, p-value = 2.895e-05

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Prediction of IMDb Movie Ratings:  
Regression Analysis

# Linear Regression Analysis in R

```
fit=lm(imdb ~ votes+duration+earnings+production+year+rating+language+
```

```
genre+rtdirector+rtactor+awards)
```

```
summary(fit)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.466e+01	5.65e+00	13.212	< 2e-16 ***
votes	2.672e-05	4.27e-06	6.250	1.95e-08 ***
duration	5.219e-03	3.495e-02	0.149	0.881673
earnings	-6.462e-09	2.549e-09	-2.535	0.013209 *
production	-1.819e-02	1.580e-02	-1.151	0.253341
year2011	-3.958e-01	1.500e+00	-0.264	0.792608
year2012	2.452e+00	1.899e+00	1.291	0.200333
year2013	4.696e+00	1.734e+00	2.707	0.008308 **
year2014	7.212e+00	1.841e+00	3.918	0.000189 ***

Residual standard error: 5.125 on 79 degrees of freedom

Multiple R-squared: 0.7642, Adjusted R-squared: 0.7045

F-statistic: 12.8 on 20 and 79 DF, p-value: < 2.2e-16

Test for statistical significance:

$H_0: \beta_1 = 0$  p-value < 0.01

$H_0: \beta_1 \neq 0$  p-value > 0.1

$0.01 < \text{p-value} < 0.1$

p-value > 0.1

p-value > 0.1

p-value < 0.05

p-value < 0.01

some means are different

Use  $p+1$  degrees of freedom because

$\beta_0 \leftarrow \widehat{\beta}_0$

$\beta_1 \leftarrow \widehat{\beta}_1$

$\beta_p \leftarrow \widehat{\beta}_p$

# Linear Regression Analysis in R

```
fit=lm(imdb ~ votes+duration+earnings+production+year+rating+language+
genre+rtdirector+rtactor+awards)
summary(fit)
```

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

Test for statistical significance:

| | | | - | | | | p-value < 0.01

## Linear Regression in R:

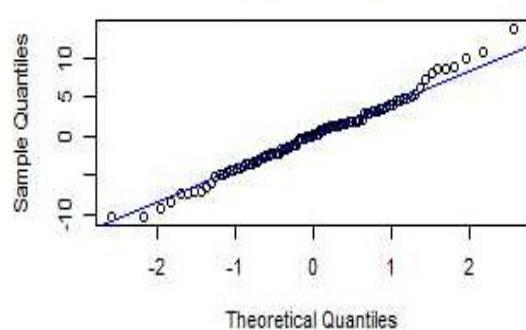
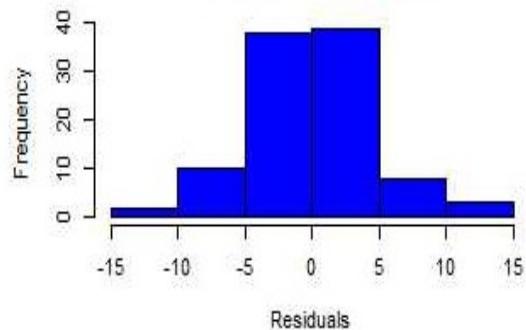
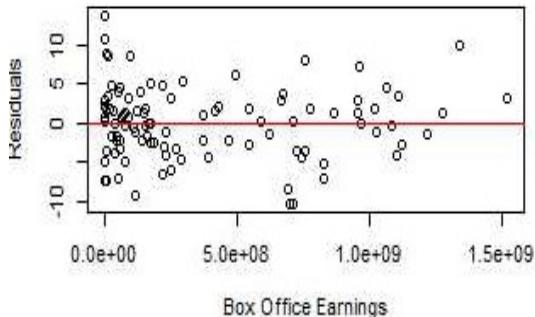
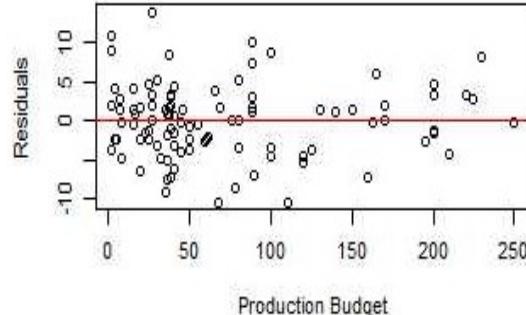
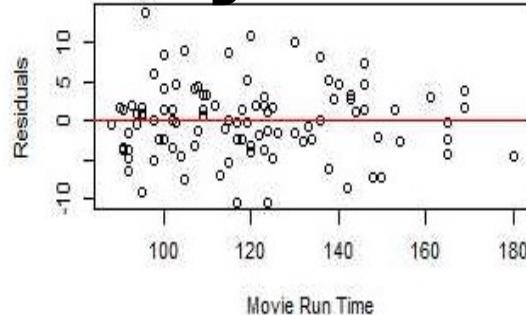
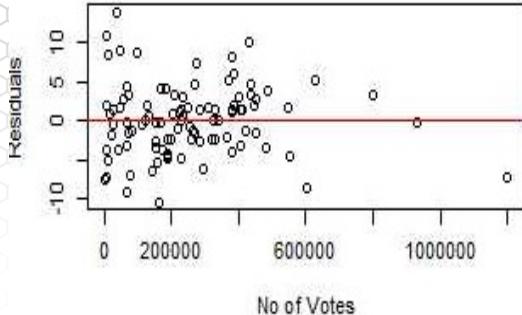
- If you have qualitative variables you must either convert them using the 'as.factor' command or you will need to specify dummy variables and add them all (except one if the model has intercept) to the model;
- Models with many qualitative variables have many parameters because each one of it will introduce several dummy variables as predicting variables.

year2014	7.212e+00	1.841e+00	3.918	0.000189	***
----------	-----------	-----------	-------	----------	-----

Residual standard error: 5.125 on 79 degrees of freedom  
Multiple R-squared: 0.7642, Adjusted R-squared: 0.7045  
F-statistic: 12.8 on 20 and 79 DF, p-value: < 2.2e-16

Use  $p+1$  degrees of freedom because  
 $n-2 = 79$   
 $\beta_0 \leftarrow \hat{\beta}_0$   
 $\beta_1 \leftarrow \hat{\beta}_1$   
 $\beta_p \leftarrow \hat{\beta}_p$   
variability explained

# Residual Analysis



# Coding Dummy Variables in R

```
### Create Dummy Variables
```

```
genre = data$genre
genre.1 = rep(0,length(genre))
genre.1[genre==1] = 1
genre.2 = rep(0,length(genre))
genre.2[genre==2] = 1
genre.3 = rep(0,length(genre))
genre.3[genre==3] = 1
genre.4 = rep(0,length(genre))
genre.4[genre==4] = 1
```

```
## Include all dummy variables without intercept
```

```
fit.1 = lm(imdb ~ genre.1 + genre.2 + genre.3 + genre.4-1)
```

```
## Include 3 dummy variables with intercept
```

```
fit.2 = lm(imdb ~ genre.1 + genre.2 + genre.3)
```

```
## Use categorical variable
```

```
genre = as.factor(data$genre)
```

```
fit.3=lm(imdb ~ genre)
```

*summary(fit.1)*

	Estimate	Std. Error	t value	Pr(> t )
genre.1	72.524	1.433	50.62	<2e-16
genre.2	78.923	2.575	30.65	<2e-16
genre.3	73.051	1.487	49.13	<2e-16
genre.4	69.500	3.791	18.33	<2e-16

*summary(fit.2)*

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	69.500	3.791	18.334	<2e-16
genre.1	3.024	4.052	0.746	0.4574
genre.2	9.423	4.583	2.056	0.0425
genre.3	3.551	4.072	0.872	0.3853

*summary(fit.3)*

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	72.5238	1.4328	50.619	<2e-16
genre2	6.3993	2.9470	2.171	0.0324
genre3	0.5275	2.0648	0.255	0.7989
genre4	-3.0238	4.0524	-0.746	0.4574

# Coding Dummy Variables in R

```
### Create Dummy Variables
```

```
genre = data$genre
```

```
genre.1 = rep(0,length(genre))
```

```
genre.1[genre==1] = 1
```

```
summary(fit.1)
```

	Estimate	Std. Error	t value	Pr(> t )
genre.1	72.524	1.433	50.62	<2e-16
genre.2	78.923	2.575	30.65	<2e-16

## Coding Dummy Variables:

- R sets the “first” class as being the baseline; if a different class is the baseline, either use dummy variables or change ‘contr.treatment’.
- Be careful when using a model without intercept in R!

```
## Include all dummy variables without intercept
```

```
fit.1 = lm(imdb ~ genre.1 + genre.2 + genre.3 + genre.4-1)
```

```
## Include 3 dummy variables with intercept
```

```
fit.2 = lm(imdb ~ genre.1 + genre.2 + genre.3)
```

```
## Use categorical variable
```

```
genre = as.factor(data$genre)
```

```
fit.3 = lm(imdb ~ genre)
```

	Estimate	Std. Error	t value	Pr(> t )
genre.2	9.423	4.583	2.056	0.0425
genre.3	3.551	4.072	0.872	0.3853

```
summary(fit.3)
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	72.5238	1.4328	50.619	<2e-16
genre2	6.3993	2.9470	2.171	0.0324
genre3	0.5275	2.0648	0.255	0.7989
genre4	-3.0238	4.0524	-0.746	0.4574

# Model Interpretation

*summary(fit)*

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	7.466e+01	5.651e+00	13.212	< 2e-16	***
votes	2.672e-05	4.275e-06	6.250	1.95e-08	***
duration	5.219e-03	3.495e-02	0.149	0.881673	
earnings	-6.462e-09	2.549e-09	-2.535	0.013209	*
production	-1.819e-02	1.580e-02	-1.151	0.253341	
year2011	-3.958e-01	1.500e+00	-0.264	0.792608	
year2012	2.452e+00	1.899e+00	1.291	0.200333	
year2013	4.696e+00	1.734e+00	2.707	0.008308	**

year2014 7.212e+00 1.841e+00 3.918 0.000189 \*\*\*  
.....

Residual standard error: 5.125 on 79 degrees of freedom  
Multiple R-squared: 0.7642, Adjusted R-squared: 0.7045  
F-statistic: 12.8 on 20 and 79 DF, p-value: < 2.2e-16

- **Coefficient Interpretation:** the expected change in the response for a one unit change in the predictor while holding all other predictors fixed.
- **Example:** The coefficient for 'votes' is .00002672. This means that if we fix the other predictors, for each

# Statistical Significance: Marginal vs Conditional

	Estimate	Std. Error	t value	Pr(> t )
.....				
duration	5.219e-03	3.495e-02	0.149	<b>0.8816</b>
earnings	-6.462e-09	2.549e-09	-2.535	<b>0.0132</b>
rtdirectorNominated	-5.389e-01	1.443e+00	-0.373	<b>0.7098</b>
rtdirectorNone	-1.398e+00	1.678e+00	-0.833	<b>0.4070</b>
.....				

`summary(aov(imdb~rtdirector))`

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rtdirector	2	1353	676.3	8.807	<b>0.000306 ***</b>
Residuals	97	7449	76.8		

`summary(lm(imdb ~ duration))`

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	56.46913	4.94644	11.416	< 2e-16 ***
duration	0.14132	0.04066	3.476	<b>0.00076 ***</b>

`summary(lm(imdb ~ earnings))`

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.259e+01	1.313e+00	55.285	< 2e-16 ***
earnings	2.057e-09	2.379e-09	0.865	<b>0.389</b>

Our goal is to find the best line that describes a linear relationship; that is, find  $(\beta_0, \beta_1)$  where

$$\begin{aligned} &: p\text{-value} \geq 0.1 \\ &: 0.01 < p\text{-value} < 0.1 \end{aligned}$$

**Equivalently, estimating:**

$$p\text{-value} > 0.1$$

$\beta_0$  is the intercept  
 $\beta_1$  is the slope  
 $\epsilon$  is the deviance of the data from the linear model

**Marginal: F-test: p-value < 0.01**

At least one statistically significant:  
 $\beta_{\text{nominated}} \text{ vs awarded}$ ,  $\beta_{\text{none}} \text{ vs awarded}$ ,  
 $\beta_{\text{nominated}} \text{ vs none}$

**Marginal:**

$\text{Population 1: } (\beta_0, \beta_1)$        $\text{Sample 1: } (Y_{1,1}, \dots, Y_{1,n})$   
 $\text{p-value } < 0.01$   
 $\text{p-value } > 0.1$

# Statistical Significance: Marginal vs Conditional

Estimate Std. Error t value Pr(>|t|)

.....

Our goal is to find the best line that describes a linear relationship; that is, find  $(\beta_0, \beta_1)$  where  
: p-value  $\leq 0.1$

## Linear Regression in R:

- If the association of a predicting variable to the response is statistically significant under conditional model, it does not mean that it is so under the marginal model and vice versa;
- If a qualitative variable shows no difference in mean response under the conditional model, it does not mean that it is so under the marginal model.
- Always interpret statistical significance in a multiple regression model conditionally!

duration 0.11102 0.01005 0.1170 0.00070

summary(lm(imdb ~ earnings))

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.259e+01	1.313e+00	55.285	< 2e-16 ***
earnings	2.057e-09	2.379e-09	0.865	<b>0.389</b>

### Marginal:

Population 1:  $\mu_1$   $\leq 0.01$  Sample 1:  $(Y_{1,1}, \dots, Y_{1,n})$   
: p-value  $> 0.1$

# Regression Analysis

## Multiple Linear Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Prediction of IMDb Movie Ratings:  
Prediction & Findings

# Prediction

```
## Read New Data (Test Data)
testdat = read.csv("Testing.csv",header=TRUE,sep=",")
dim(testdat)[1]
[1] 21
## Prepare the new data the same as the training data
## Response Variable: scaled between 0 and 100
nimdb=testdat$imdb*10
## Quantitative Predicting Variables
# Number of imdb user votes for the movie
nvotes=testdat$votes
.....
## Qualitative Predicting Variables
# Rating of a film's suitability for certain audiences, based on its content
nrating=as.factor(testdat$rating)
.....
## Write the new data into a data frame
newdat = data.frame(votes=nvotes, duration=nduration, earnings=nearnings, production=nproduction,
year=nyear, rating=nrating, language=nlanguage, genre=ngenre,rtdirector=nrtdirector,
rtactor=nrtactor,awards=nawards)
# Specify whether a confidence or prediction interval
predict(fit,newdat,interval=c("prediction"))
```

# Prediction

```
## Read New Data (Test Data)
testdat = read.csv("Testing.csv", header=TRUE, sep=",")
dim(testdat)[1]
[1] 21
## Prepare the new data the same as the training data
## Response Variable: scaled between 0 and 100
nimdb=testdat$imdb*10
## Quantitative Predicting Variables
# Number of imdb user votes for the movie
nvotes=testdat$votes
.....
## Qualitative Predicting Variables
# Rating of a film's suitability for certain audiences, based on i
nrating=as.factor(testdat$rating)
.....
## Write the new data into a data frame
newdat = data.frame(votes=nvotes, duration=nduration, earnin
year=nyear, rating=nrating, language=nlanguage, genre=nge
rtactor=nrtactor, awards=nawards)
# Specify whether a confidence or prediction interval
predict(fit,newdat,interval=c("prediction"))
```

## Prediction Output:

	fit	lwr	upr
1	76.29411	64.27794	88.31028
2	62.73839	50.21367	75.26311
3	72.03305	59.41553	84.65057
4	76.75066	64.15607	89.34525
5	80.41676	67.90456	92.92896
6	72.58612	59.66305	85.50919
7	62.13416	49.43550	74.83281
8	73.93309	61.64041	86.22576
9	71.05206	58.81055	83.29357
10	61.86840	49.15973	74.57707
11	74.57625	62.18013	86.97238

# Prediction Accuracy

## Save Predictions to compare with observed data

```
predicttestdata = predict(fit,newdat,interval=c("prediction"))
```

```
imdb.pred = predicttestdata[, 1]
```

```
imdb.lwr = predicttestdata[, 2]
```

```
imdb.upr = predicttestdata[, 3]
```

### Mean Squared Prediction Error (MSPE)

```
mean((imdb.pred-nimdb)^2)
```

### Mean Absolute Prediction Error (MAE)

```
mean(abs(imdb.pred-nimdb))
```

### Mean Absolute Percentage Error (MAPE)

```
mean(abs(imdb.pred-nimdb)/nimdb)
```

### Precision Measure (PM)

```
sum((imdb.pred-nimdb)^2)/(nimdb-mean(nimdb))^2)
```

### Does the observed data fall in the prediction intervals?

```
sum(nimdb<imdb.lwr)+sum(nimdb>imdb.upr)
```

Analysis of Variance (ANOVA) for multiple regression:  
**Accuracy Measures:**

MSPE =

MAE =

Where  $SSReg = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$  and  $SSRes = \sum_{i=1}^n (Y_i - \bar{Y})^2$ .

ANOVA is used to test:  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$   
PM =

We reject  $H_0$  if F-statistic is large ( $F\text{-statistic} > F_{\alpha, p, n-p-1}$ ); Which means that at least one of the coefficients is different from zero at the  $\alpha$  significant level.

The p-value of the test is:  $P(F_{p, n-p-1} > F\text{-statistic})$  where  $F_{p, n-p-1}$  is the F-distribution with  $p$  and  $n-p-1$  degrees of freedom.

**Prediction Accuracy:**

MSPE = 23.69

MAE = 3.96

MAPE = 0.055

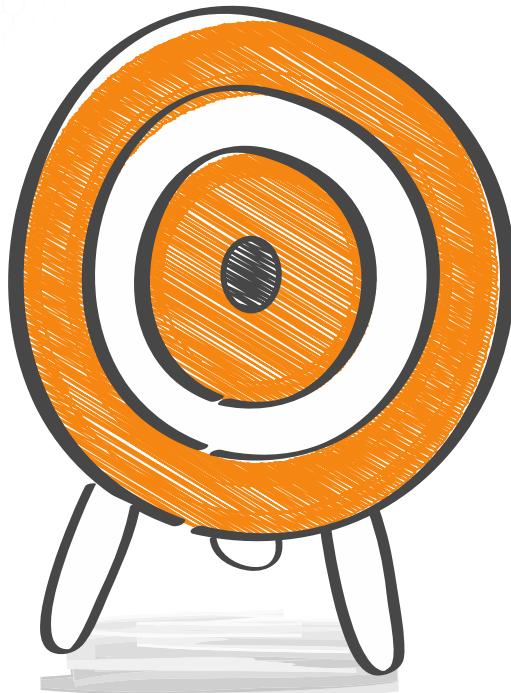
PM = 0.058

All new observations fall within the prediction CI

# Prediction of IMDb Movie Ratings: Findings

- Duration and Budget ~~X~~ IMDb Rating;
- Each additional 10,000 votes => an increase of 2.67 in IMDb score.
- Movies released in 2013 are rated 4.696 points > movies produced in 2010. Similarly, movies released in 2014 are rated 7.212 points > those released in 2010. Other years show no statistically significant difference from 2010. This seems to indicate a recency bias in how movies are rated on IMDb;
- The predicting variables in the model explain close to 76% of the variability in IMDb scores.
- The model also provides good predictions.

# Summary



# Regression Analysis

## Logistic Regression

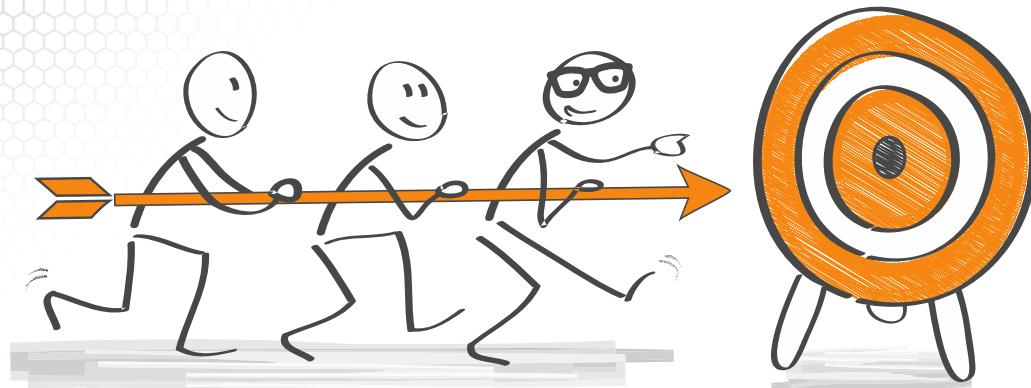
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Introduction

# About this lesson



# Yes/No Questions

- How likely is it that users will like a new layout of our website?
  - Will my customers leave my wireless service at the end of their subscription?
  - What financial characteristics can be used to predict whether or not a business will go bankrupt?
- Model the probability of 'Yes'

# Linear Regression

**Model:**  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{e_1, \dots, e_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

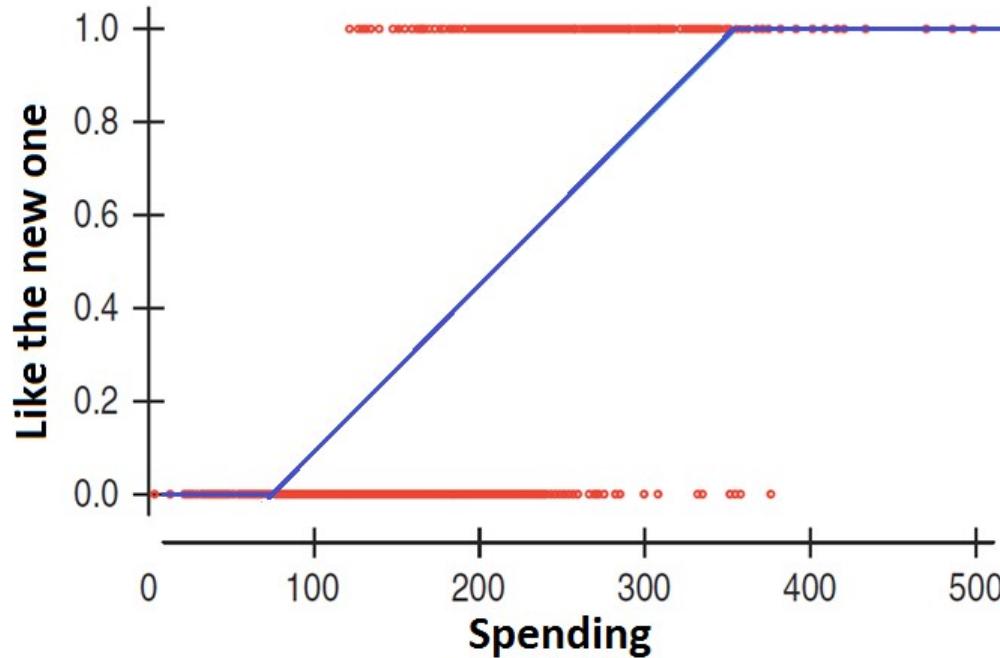
# Linear Regression for Yes/No Question?

- Uber recently changed their logo



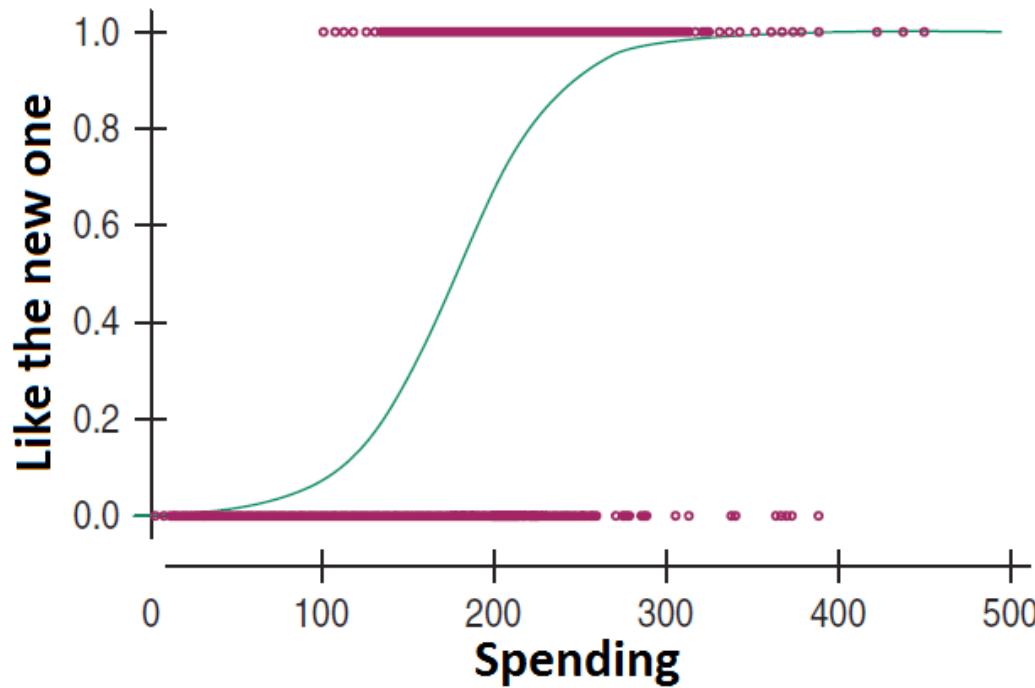
- You are asked to model whether Uber users will like the new logo based on how much they spent in the last 3 months using Uber

# What is Wrong with Linear Regression?



Customers will not behave like this!

# S-shaped Curve



# Logistic Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

$$p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$$

link  $p$  to the predicting variables through a nonlinear *link function*  
*function*  $g(p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

+

+

There is no error term! What are the  
model assumptions?

# Logistic Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

## Assumptions:

- *Linearity Assumption:*  $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \dots + \widehat{\beta}_p x_p$
- *Independence Assumption:*  $Y_1, \dots, Y_n$  are independent random variables
- *Logit link function:*  $\frac{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \dots + \widehat{\beta}_p x_p}{\sum_{i=1}^n (x_i - \bar{x})^2}$

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Data Example

# Data Example: Smoking

In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.

- Among the information obtained originally was whether a person was a smoker or not.

Twenty years later a follow-up study was conducted:

- 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers! Call Philip Morris -- smoking leads to a longer life span!

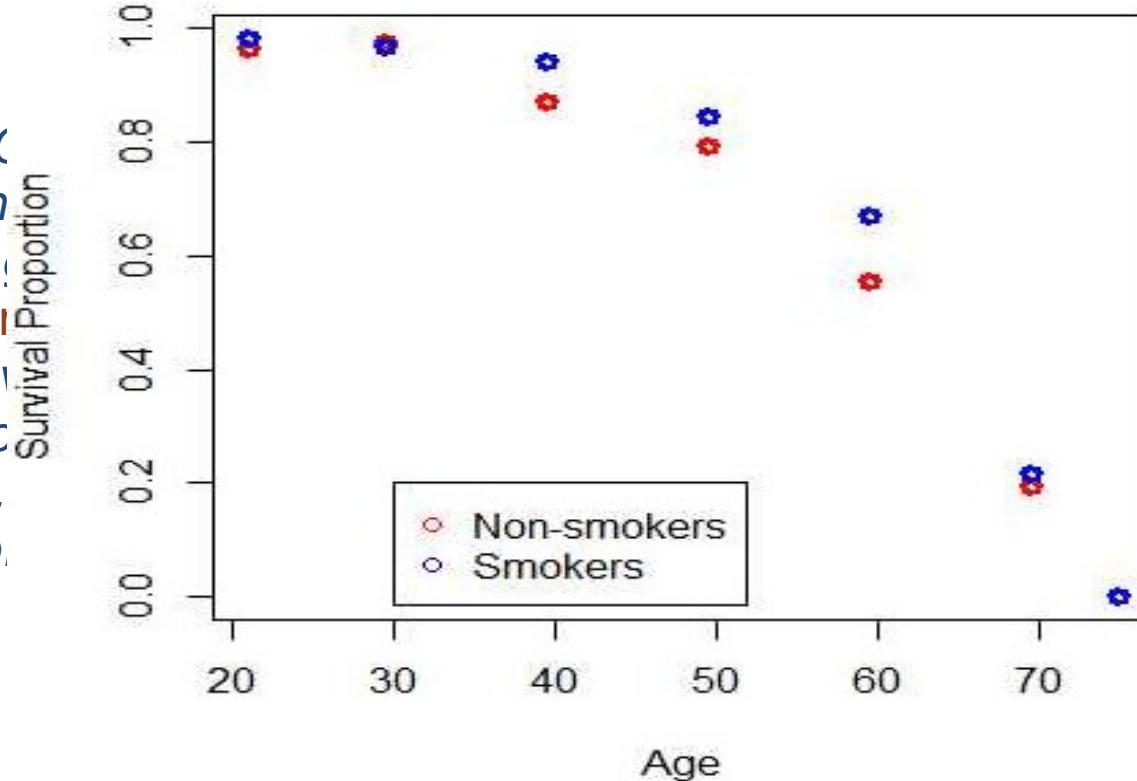
Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

# Data Example in R

```
## Read data in R
smoking =
read.table("CIGARETT.dat",sep="",row.names=NULL)
names(smoking)=c("Age","Smoker","Survived","At.risk")
attach(smoking)
## Plot proportion of survival
plot(Age,Survived/At.risk, xlab="Age", ylab="Survival
Proportion", col=c("red","blue"),lwd=3)
legend(30,0.2, legend=c("Non-smokers","Smokers"),pch=1,
col=c("red","blue"))
```

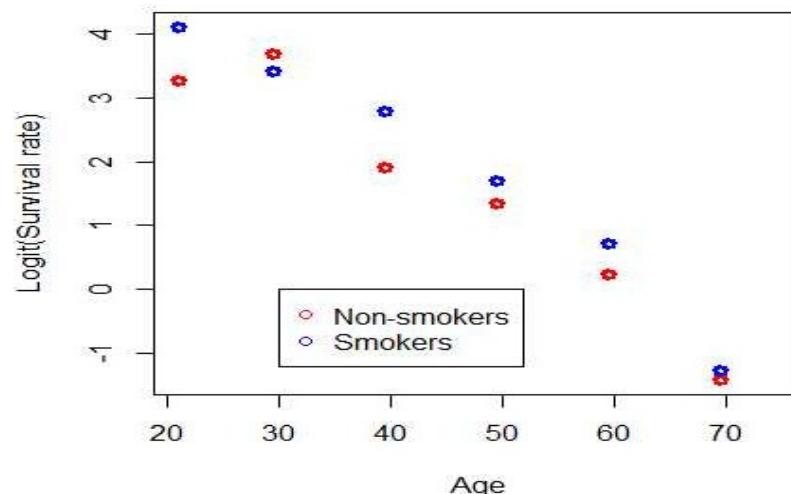
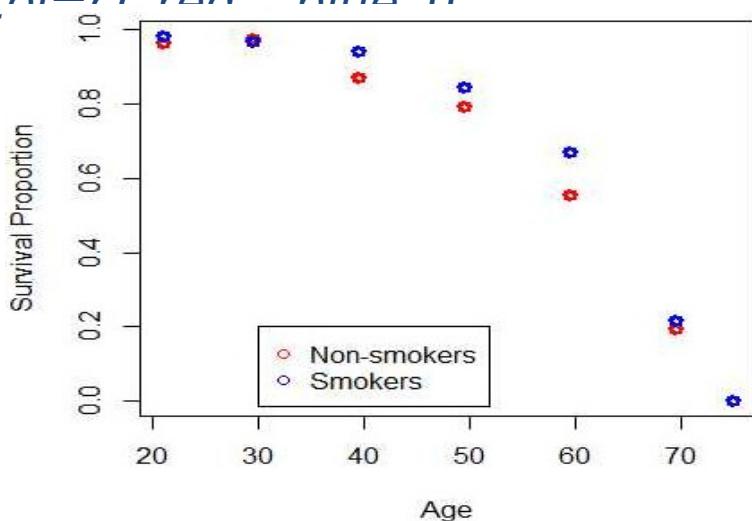
# Data Example in R

```
## Read data
smoking =
read.table("C:/.../smoking.txt")
names(smoking)
attach(smoking)
## Plot proportion
plot(Age,Survival
Proportion", cex=1.5, xlab="Age", ylab="Survival Proportion", col=c("red","blue"))
legend(30,0.2, c("Non-smokers", "Smokers"))
```



# Data Example in R (cont'd)

```
## Plot of logit transformation of the proportion survival  
prop.survival=Survived/At.risk  
plot(Age,log(prop.survival/(1-prop.survival)), col=c("red","blue"),  
xlab="Age", ylab="Logit(Survival Proportion)", lwd=3)  
legend(30,0, legend=c("Non-smokers","Smokers"), pch=1,  
col=c("red" "blue"))
```



# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Description and Estimation

# Logistic Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

$$\beta_1 - \beta_0$$

link  $p$  to the predicting variables through  ~~$t_{n-\log 2}$~~  link function

$$\sqrt{\frac{MSE}{S_{xx}}}$$

OR

# Model Interpretation

Probability of success given predicting variable  $X = x$ :

- The logit function is the *log odds function*.
- Exponential of the logit function = is the odds of  $Y = 1$  at  $X = x$
- The odds at  $X = A$  versus  $X = B$  are equal to the *odds ratio*:

=

# Model Estimation

**Model** the probability of success given predictor(s):

Logit(

**Parameters:** , ,...,

**Approach:** Maximum Likelihood Estimation:

$$\rightarrow \hat{\beta}_1 \pm \frac{ta}{2, n-2}$$
$$(\hat{\beta}_1, \dots, \hat{\beta}_k) = \log(L(\hat{\beta}_1, \dots, \hat{\beta}_k)) =$$

$$\sqrt{\frac{MSE}{S_{xx}}}$$

# Model Estimation (cont'd)

**Approach:** Maximum Likelihood Estimation

$$\max_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n \left\{ y_i \log \left( \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \right) + (1 - y_i) \log \left( \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \right) \right\}$$

- Maximizing the (log-)likelihood function with respect to  $\beta$  in close form
- Maximizing the (log-)likelihood function with respect to  $\beta_0, \beta_1, \dots, \beta_p$  in close form. expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters
- Use numerical algorithm to estimate  $\Rightarrow$ 
  - Use numerical algorithm to estimate  $\beta_0, \beta_1, \dots, \beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

**Upshot:** The estimated parameters and their standard errors are approximate

**Upshot:** The estimated parameters and their standard errors are approximate estimates

Do not attempt to do it yourself. Use a statistical software to derive the estimated regression coefficients.

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Estimation: Data Example

# Data Example: Smoking

In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.

- Among the information obtained originally was whether a person was a smoker or not.

Twenty years later a follow-up study was conducted:

- 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers! Call Philip Morris -- smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

# Data Example in R

**Data:**  $Y_i$  binary responses  $\sim \text{Binomial}(p_i, n_i)$

- $Y_i$  number of people at risk who survived (Survived)

```
## number of people at risk (At.risk)
```

```
smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk,  
family=binomial)
```

```
summary(smoke1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
--	----------	------------	---------	----------

(Intercept)	0.78052	0.07962	9.803	< 2e-16 ***
-------------	---------	---------	-------	-------------

Smoker	0.37858	0.12566	3.013	0.00259 **
--------	---------	---------	-------	------------

For smokers versus non-smokers, the log odds of survival increases by 0.378 OR the odds of survival increase by 1.459.

# Data Example in R

```
## Fit a logistic regression model
```

```
smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk,  
family=binomial)
```

summary(smoke2)

## Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	7.785001	0.454999	17.110	< 2e-16 ***
Smoker	-0.240831	0.167885	-1.435	0.151
Age	-0.127419	0.007397	-17.227	< 2e-16 ***

For smokers versus non-smokers, the log odds of survival decreases by 0.24 OR the odds of survival decreases by 0.2134.

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Statistical Inference

# Model Estimation

**Model** the probability of success given predictor(s):  
The Least Squares estimated coefficients have specific interpretations.

Logit(

✓  $\hat{\beta}_0$  is the estimated expected value of the response variable when all predicting variables equal zero;  
**Approach**: Maximum Likelihood Estimation:

✓  $\hat{\beta}_i$  is the estimated expected change in the response variable associated with one unit of change in the  $i$ -th predicting variable holding fixed all other predictors in the model for all  $i = 1, \dots, p$ ;  
 $I(\cdot, \dots, \cdot) \equiv \log(L(\cdot, \dots, \cdot))$  OR

# Statistical Inference

Maximum Likelihood Estimators (MLEs):

Statistical Properties of MLEs:

- Approximate Sampling Distribution:
- The normal approximation relies on the assumption of large sample size  $\Rightarrow$  Statistical inference is not reliable for small sample data

1- $\alpha$  Approximate  
Confidence  
interval



# Statistical Inference (cont'd)

Test for statistical significance of  $\beta_j$  given all other predicting variables in the model by using the z-test (Wald test) for

$$H_0: \beta_j = 0 \text{ vs. } H_1: \beta_j \neq 0$$

$$\text{z-value} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

We reject  $H_0$  if  $|z\text{-value}|$  gets too large. We interpret this as  $\beta_j$  being statistically significant if the null hypothesis is rejected.

# Statistical Inference (cont'd)

Approach: Maximum Likelihood Estimation

$$z\text{-value} = \frac{\hat{\beta}_1 - \beta_1}{\text{standard deviation of } \hat{\beta}_1}$$

- Maximizing the (log-)likelihood function with respect to  $\beta_0, \beta_1, \dots, \beta_p$  in close form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters

- Use numerical algorithm to estimate  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Upshot: The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use a statistical software to derive the estimated regression coefficients.

For significance level  $\alpha$ , Reject if  $|z\text{-value}| > z_{\alpha/2}$

Alternatively, compute  $P\text{-value} = 2P(Z > |z\text{-value}|)$

**Model** Probability of success given predictor(s)  
 $p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$

**versus**  
link  $p$  to the predicting variables through a nonlinear *link function*

$$P\text{-value} = P(Z > z\text{-value})$$

**What if we want to test for negative relationship?**

**versus**  
$$\frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

# Statistical Inference (cont'd)

Approach: Maximum Likelihood Estimation

$$z\text{-value} = \frac{\text{how large to reject } H_0: \beta_1 = b?}{\sqrt{\sum_{i=1}^n \left\{ y_i \left( \frac{\beta_0 + \beta_1 x_i + \beta_p x_p}{1 + e^{\beta_0 + \beta_1 x_i + \dots + \beta_p x_p}} \right) - b \right\}^2}}$$

- Maximizing the (log-)likelihood function with respect to  $\beta_0, \beta_1, \dots, \beta_p$  in close form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters

- Use numerical algorithm to estimate  $\beta_0, \beta_1, \dots, \beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Upshot: The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use a statistical software to derive the estimated regression coefficients.

For significance level  $\alpha$ , Reject if  $|z\text{-value}| >$   
Standard Deviation/Error of  $\hat{\beta}_1$

- Because the approximation of the normal distribution relies on large sample size, so the hypothesis testing procedures do.
- What if  $n$  is small? The hypothesis testing procedure will have a probability of type I error larger than the significance level; that is, more type I errors than expected.

**versus**

$$\text{P-value} = P(Z < z\text{-value})$$
$$\sum_{i=1}^n \frac{y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

# Testing for Subsets of Coefficients

*Full model:*

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_q z_q$$

*Reduced model:*

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

*The hypothesis test:*

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_q = 0$  versus  $H_A : \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L_f$  )
- Maximize the likelihood function under reduced model:  $L_r$
- Test Statistic:

$$Dev = \log(L_f) - \log(L_r)$$

$$P\text{-value} = P(Dev)$$

# Testing for Subsets of Coefficients

*Full model:*

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_q z_q$$

- **The hypothesis test for subsets of coefficients is approximate; it relies on large sample size.**
- **This is not a test for goodness of fit! It only compares two models.**

$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_q = 0$  versus  $H_A : \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L_f$
- Maximize the likelihood function under reduced model:  $L_r$
- Test Statistic:

$$Dev = -2 \log \left( \frac{L_r}{L_f} \right)$$

$$P\text{-value} = P(Dev)$$

# Testing for Overall Regression

Full model:

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Reduced/Null model

$$= \beta_0$$

The hypothesis test:

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$  versus  $H_A: \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L$ )
- Maximize the likelihood function under reduced/null model:  $L$ )
- Test Statistic:

$$Dev = \log(L)$$

$$P\text{-value} = P(Dev)$$

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Statistical Inference: Data Example

# Data Example: Smoking

In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.

- Among the information obtained originally was whether a person was a smoker or not.

Twenty years later a follow-up study was conducted:

- 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers! Call Philip Morris -- smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

# Data Example in R

```
## Fit a logistic regression model
smoke1 = glm(Survived/At.risk ~ Smoker, weights=At.risk,
family=binomial)
summary(smoke1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.78052	0.07962	9.803	< 2e-16 ***
Smoker	0.37858	0.12566	3.013	0.00259 **

Null deviance: 641.5 on 13 degrees of freedom

Residual deviance: 632.3 on 12 degrees of freedom

```
1-pchisq(9.2,1)
```

```
[1] 0.002420151
```

**Test for significance** p-value=0.0025 thus statistically

Significant

**Test for overall regression:** Null deviance -Residual Deviance = 9.2 with

**Test for overall regression:** Null deviance -Residual Deviance

= 9.2 with the p-value=P() = 0.0024

# Data Example in R (cont'd)

```
## Fit a logistic regression model  
smoke2 = glm(Survived/At.risk ~ Smoker + Age, weights=At.risk,  
family=binomial)  
summary(smoke2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	7.785001	0.454999	17.110	< 2e-16 ***
Smoker	-0.240831	0.167885	-1.435	0.151
Age	-0.127419	0.007397	-17.227	< 2e-16 ***

Null deviance: 641.496 on 13 degrees of freedom

Residual deviance: 43.459 on 11 degrees of freedom

**Test for significance** p-value = 0.151, not statistically significant

**Test for significance** Age: p-value = 0, statistically significant

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

# Logistic Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

# Assumptions

- Assumptions:**  ~~$\beta_1 - \beta_1$~~  • Linearity Assumption: ~~1...1~~  $\sim t_{n-2}$

- *Independence Assumption*:  $Y_1, \dots, Y_n$  are independent random variables

- *Logit link function:*

# There is no error term! How to check the assumptions?

# Residuals in Logistic Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

- **Logistic Regression Without Replications:** For each  $Y_i$  observe a unique set of predictors  $(x_{i1}, \dots, x_{ip})$  across all  $i=1, \dots, n$

$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Bernoulli}(\pi_i)$  or  $\text{Binomial}(1, \pi_i)$

- **Logistic Regression With Replications:** Observed  $n_i$  repeated responses  $Y_i$  for a set of predictors  $(x_{i1}, \dots, x_{ip})$  across all  $i=1, \dots, n$ .

$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Binomial}(n_i, \pi_i)$ ,  $n_i > 1$

# Residuals in Logistic Regression

## Logistic Regression With Replications:

$$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Binomial}(n_i, p_i), n_i > 1$$

- Estimated probabilities are:

- Pearson Residuals:

- Deviance Residuals:

$$\hat{p}_i \pm \frac{ta_{\alpha/2, n-2}}{\sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}^2}}$$

$$\sqrt{\frac{MSE}{S_{xx}}}$$

# Residuals in Logistic Regression

## **Logistic Regression With Replications:**

- Pearson's residuals follow directly a normal approximation to a binomial. Hence approximately  $N(0,1)$
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model. Thus approximately  $N(0,1)$  if the model is a good fit.

# Model Goodness of Fit

**Approach:** Maximum Likelihood Estimation

- Normal Probability plot & Histogram of the Residuals
- Residuals vs predictors: Linearity & Independence Assumption
- Logit of success rate vs predictors: Linearity Assumptions expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters

## Hypothesis Testing Procedure:

: the logistic model fits the data

: the logistic model does not fit the data. Use numerical algorithm to estimate  $\beta_0, \beta_1, \dots, \beta_p \Rightarrow \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Deviance test statistic:  $D =$

**Upshot:** The estimated parameters and their standard errors are approximate estimates.

Under null hypothesis,  $D \sim \chi^2$  with  $df = n-p-1$

Do not attempt to do it yourself! Use a statistical software to derive the estimated regression coefficients.

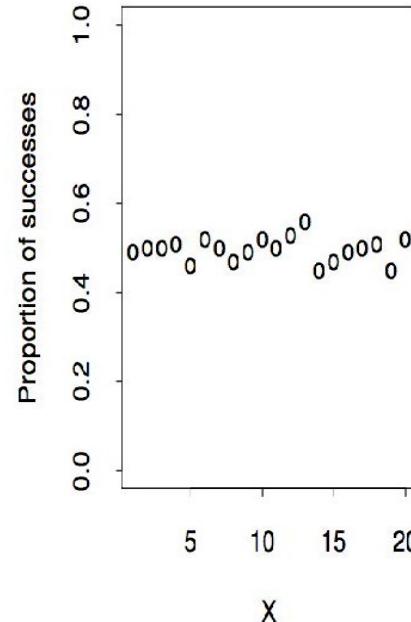
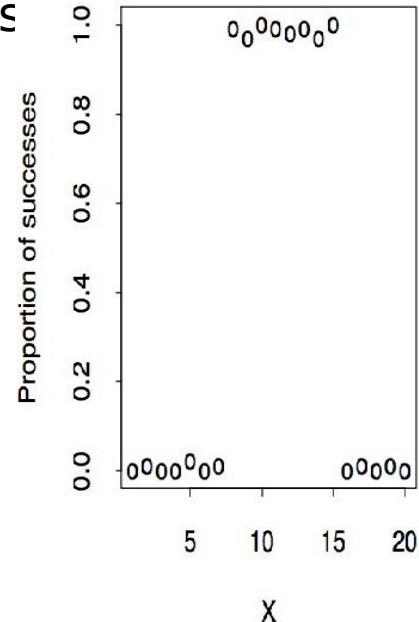
Reject the null that the model is correct if p-value =  $P(>D)$  small. Note that for this test, we want large p-values!!!!

# Goodness of Fit vs. Predictive Power

The logistic model is a sensible one for probabilities, but is not necessarily appropriate for a particular data set. This is not the same thing as saying that the predicting variables are not good predictors for the probability of  $s$ .

**Goodness of fit:** Model assumptions hold – e.g. the S-shape logit function fits the data.

**Predictive Power:** The predicting variables predict the data even if the one or more assumptions do not hold.

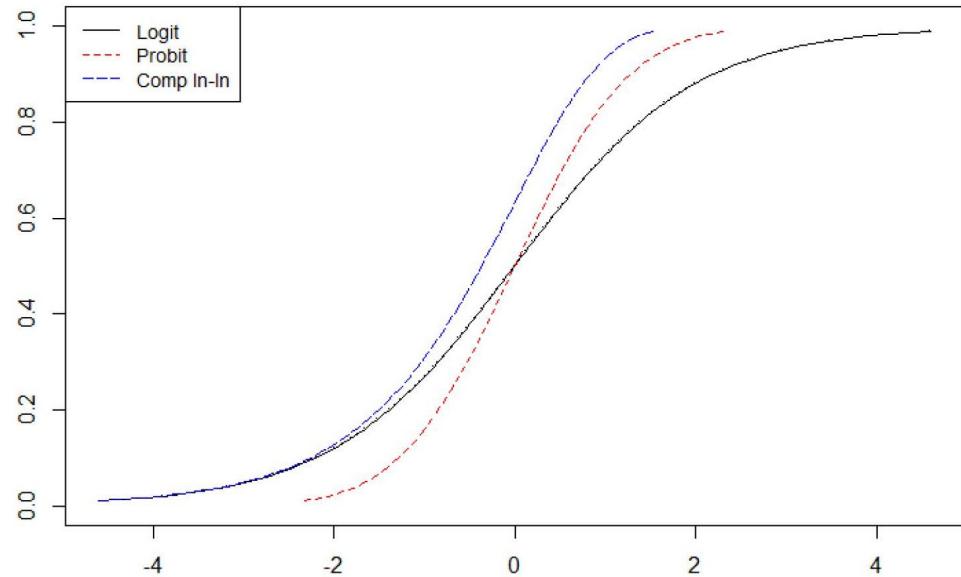


# What if No Goodness of Fit?

- Add predicting variables or/and transform predicting variables to improve linearity;
- Identify unusual observations (outliers, leverage points);
- The binomial distribution isn't appropriate:
  - Overdispersion: the variability of the probability estimates is larger than would be implied by a binomial random variable
    - Correlation in the observed responses
    - Heterogeneity in the success probabilities that hasn't been modeled
  - Logit function not appropriate
    - Other S-shape functions: probit, c-log-log

# What if No Goodness of Fit?

- Add predicting variables variables to improve line
- Identify unusual observations
- The binomial distribution
  - Overdispersion: the standard error of the estimates is large because the binomial random variable is not a good approximation of the observed data
  - Correlation in the data
  - Heterogeneity that hasn't been modeled
- Logit function not appropriate
  - Other S-shape functions: probit, c-log-log



# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment: Data Examples

# Data Example: Smoking

In 1972-1974 a survey was taken in Whickham, a mixed urban and rural district near Newcastle upon Tyne, United Kingdom.

- Among the information obtained originally was whether a person was a smoker or not.

Twenty years later a follow-up study was conducted:

- 76.12% of the 582 smokers were still alive, while only 68.58% of 732 nonsmokers were still alive.

Smokers had a higher survival rate than nonsmokers! Call Philip Morris -- smoking leads to a longer life span!

Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University.

# GOF Hypothesis Test

```
## Deviance Test for GOF (using deviance residuals)
c(deviance(smoke2), 1-pchisq(deviance(smoke2),11))
[1] 4.345918e+01 9.033325e-06
## GOF test using Pearson residuals
pearres2 = residuals(smoke2,type="pearson")
pearson.tvalue = sum(pearres2^2)
c(pearson.tvalue, 1-pchisq(pearson.tvalue,11))
[1] 36.751889370 0.000126796
```

## Test for ~~goodness-of-fit~~

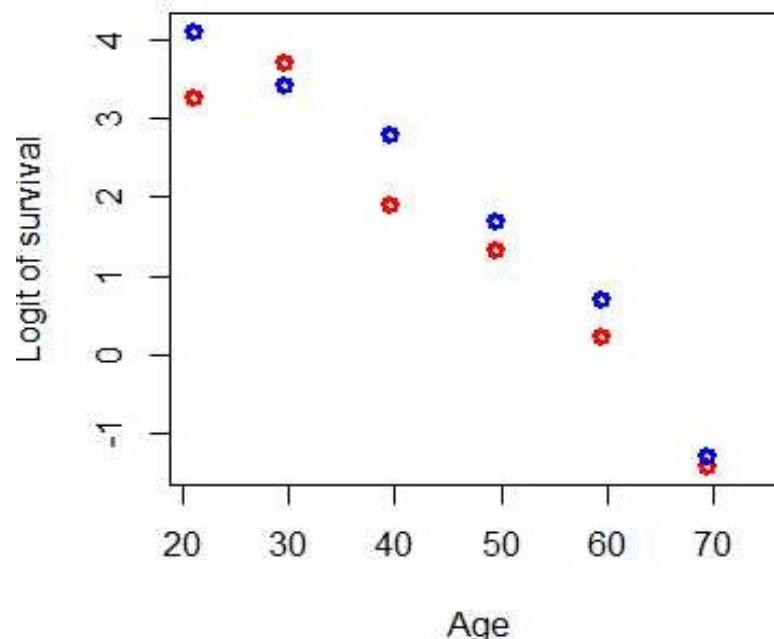
- Using deviance residuals: p-value 0
- Using Pearson residual: p-value = 0.0001
- Reject the ~~null hypothesis~~ of good fit. Thus NOT a good fit.

# Linearity Assumption

## Is it a linear fit?

```
plot(Age,log((Survived/At.risk)/(1-Survived/At.risk)), ylab="Logit of survival", main="Scatterplot of logit survival rate vs age", col=c("red","blue"), lwd=3)
```

The relationship between the logit of survival and age is more quadratic than linear.



# Improve the Fit

## Fit a logistic regression model

*Age.squared = Age\*Age*

*smoke3 = glm(Survived/At.risk ~ Smoker + Age +*

*Age.squared, weights=At.risk, family=binomial)*

*summary(smoke3)*

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	2.5190783	1.0248206	2.458	0.0140 *
Smoker	-0.4284561	0.1770581	-2.420	0.0155 *
Age	0.0951102	0.0430095	2.211	0.0270 *
Age.squared	-0.0021673	0.0004309	-5.030	4.91e-07 ***

Null deviance: 641.496 on 13 degrees of freedom

Residual deviance: 19.808 on 10 degrees of freedom

Test H0 :  $\beta_{income} = \beta_{public} = \beta_{year} = \beta_{expend} = 0$   
How was the F-statistic computed:  
F-statistic =  $\frac{SSReg(income\_public\_years\_Expend|Takers.Rank)}{SSRes(income\_public\_years\_Expend|Takers.Rank) / 4}$   
The p-value is computed as:  $P(F_{4,43} > F\text{-statistic}) = 1 - P(F_{4,43} < F\text{-statistic})$

Interpretation: The p-value is approximately 0 thus reject the null hypothesis. We conclude that at least one other predictor among the four predictors (income, years, public and expend) will be significantly associated to the state-average SAT score.

Test for significance p-value 0, statistically significant

# GOF Test for Improved Model

```
## Test for goodness of fit
pearres3 = residuals(smoke3,type="pearson")
pearson = sum(pearres3^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.79 0.14
round(c(deviance(smoke3), 1-pchisq(deviance(smoke3),10)),2)
[1] 19.80 0.03
```

## Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.03
- Using Pearson residual: p-value = 0.14
- Do not reject the null hypothesis of good fit using Pearson residuals but do reject using Deviance residuals at the significance level 0.03 or higher.

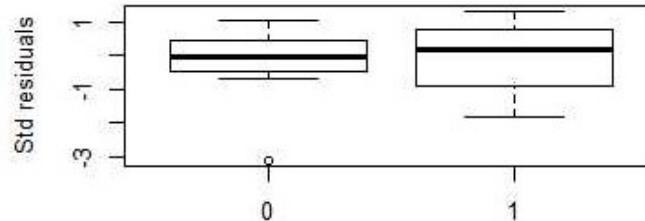
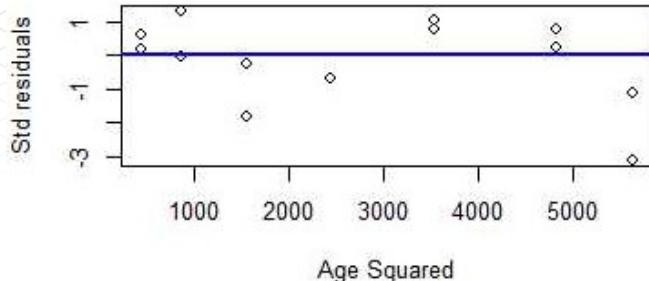
# Data Example in R

(cont'd)

```
## Residual Plots
res = resid(smoke3, type="deviance")
par(mfrow=c(2,2))
plot(Age.squared, res, ylab="Std residuals", xlab="Age Squared")
abline(0,0, col="blue", lwd=2)
boxplot(res ~ Smoker, ylab = "Std residuals")
qqnorm(res, ylab="Std residuals")
qqline(res, col="blue", lwd=2)
hist(res, 10, xlab="Std residuals", main="")
```

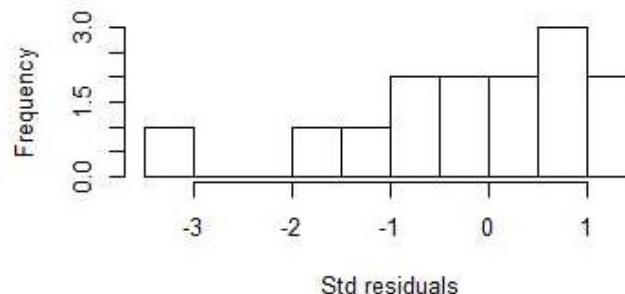
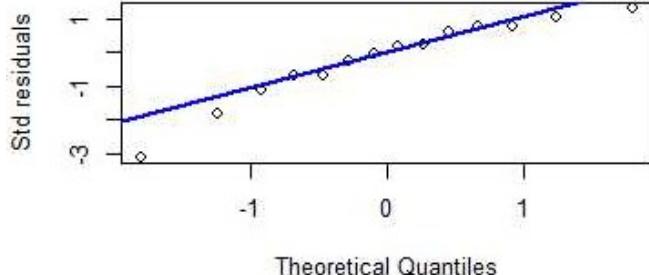
# Data Example in R

~~# Residual Plots~~



*(squared")*

Normal Q-Q Plot



# Higher Order Nonlinearity

```
## Fit a logistic regression model with Age as a factor
smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age),
weights=At.risk, family=binomial)
summary(smoke4)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.14	-0.002	0.998035

Null deviance: 641.4963 on 13 degrees of freedom

Residual deviance: 2.3809 on 6 degrees of freedom

# Higher Order Nonlinearity

## Fit a logistic regression model with Age as a factor

```
smoke4 = glm(Survived/At.risk ~ Smoker + factor(Age),
```

Approach: Maximum Likelihood Estimation

Test for significance p-value=0.015, statistically significant at 0.05

Test for significance: Not all regression coefficients for the dummy

variables for age are statistically significant

(Intercept)	3.8601	0.5939	6.500	8.05e-11 ***
Smoker	-0.4274	0.1770	-2.414	0.015762 *
factor(Age)29.5	-0.1201	0.6865	-0.175	0.861178
factor(Age)39.5	-1.3411	0.6286	-2.134	0.032874 *
factor(Age)49.5	-2.1134	0.6121	-3.453	0.000555 ***
factor(Age)59.5	-3.1808	0.6006	-5.296	1.18e-07 ***
factor(Age)69.5	-5.0880	0.6195	-8.213	< 2e-16 ***
factor(Age)75	-27.8073	11293.14	-0.002	0.998035

Null deviance: 641.4963 on 13 degrees of freedom

Residual deviance: 2.3809 on 6 degrees of freedom

# Higher Order Nonlinearity: GOF

```
## Test for goodness of fit
```

```
pearres4 = residuals(smoke4,type="pearson")
```

```
pearson = sum(pearres4^2)
```

```
round(c(pearson, 1-pchisq(pearson,6)),2)
```

```
[1] 2.37 0.88
```

```
round(c(deviance(smoke4), 1-pchisq(deviance(smoke4),6)),2)
```

```
[1] 2.38 0.88
```

## Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.88
- Using Pearson residual: p-value = 0.88
- Do not reject the null hypothesis of good fit using Pearson residuals or using Deviance residuals.

# Different Link Function

## Use probit link function

```
smoke5 = glm(Survived/At.risk ~ Smoker + Age + Age.squared,  
weights=At.risk, family=binomial(link = probit))
```

```
summary(smoke5)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.1033963	0.4904877	2.250	0.02447 *
Smoker	-0.2277451	0.0970191	-2.347	0.01890 *
Age	0.0681279	0.0213095	3.197	0.00139 **
Age.squared	-0.0013767	0.0002173	-6.335	2.37e-10 ***

Null deviance: 641.496 on 13 degrees of freedom

Residual deviance: 18.233 on 10 degrees of freedom

**Test for significance** p-value=0.018, statistically significant at 0.05

**Test for significance** p-value 0, statistically significant

# Different Link Function: GOF

```
## Test for goodness of fit
pearres5 = residuals(smoke5,type="pearson")
pearson = sum(pearres5^2)
round(c(pearson, 1-pchisq(pearson,10)),2)
[1] 14.00 0.17
round(c(deviance(smoke5), 1-pchisq(deviance(smoke5),10)),2)
[1] 18.23 0.05
```

## Does the goodness of fit improve?

- Using deviance residuals: p-value = 0.17
- Using Pearson residual: p-value = 0.05
- Do not reject the null hypothesis of good fit using Pearson residuals or using Deviance residuals at the significance level 0.05.

# Simpson's paradox

Marginal versus Conditional relationship:

- ***Marginal***: Capturing the association of a predicting variable to the response variable marginally, i.e. without consideration of other factors.
- ***Conditional***: Capturing the association of a predicting variable to the response variable, conditional of other predicting variables in the model.

**Simpson's paradox**: Reversal of an association when looking at a marginal relationship versus a conditional relationship.

- Smoking is statistically significant with a negative estimated coefficient under the marginal model.
- Smoking has a positive estimated coefficient under the conditional

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Classification

# Classification Objective

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

$$\beta_1 - \beta_0$$

**Objective:** Classify (predict) a new binary response  $Y^*$  based on observed predicting variables

- Predicted probability:
- If the predicted probability is large, then classify  $Y^*$  as a success.

# Classification Objective

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

$$\beta_1 - \beta_0$$

How good is the classification or prediction? based

- Predicted probability:
- If the predicted probability is large, then classify  $Y^*$  as a success.

MSE

$\frac{S_{xx}}{S_{xx}}$

# Classification Error Rate

- Given  $h$ , the predicted probability is:
- classifier:  $h$   
where  $r$  is a classification threshold, e.g.  $r = 1/2$ .
- Classification error rate:  $E(h) = P(Y \neq h)$

How to quantify the error rate?

# Cross Validation

Split the data  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  into:

- **Training set:** Used to fit the model, i.e. estimate
- **Testing/Validation set:** Used to estimate the classification

error rate

$$\rightarrow \hat{\beta}_1 \pm t_{\alpha/2, n-2}$$

where  $m$  is the size of the validation set.

How to split the data?

1. Random subsampling
2. K-fold cross-validation (KCV)
3. Leave-one-out Cross-Validation

$$\sqrt{\frac{MSE}{S_{XX}}}$$

# Cross Validation: How to Split Data?

## 1. Random subsampling

- Randomly split the data into two portion, training the model on one portion and validating (or testing) on the other portion
- Randomly split multiple times
- Average the classification error rate across all random splits
- Randomly split multiple times

## 2. Average the classification error rate across all random splits

- Randomly divide the data into K chunks of approximately equal size.

## 2. K-fold cross validation (KCV)

- Randomly divide the data into K chunks of approximately equal size.
  - The training data consist of data without the  $k$ -th fold of data and the testing data consist of the  $k$ -th fold;
  - Compute classification error rate  $\hat{L}_{(k)}$  for the  $k$ -th fold testing data
- For  $k = 1$  to  $K$ ,
  - The training data consist of data without the  $k$ -th fold of data and the testing data consist of the  $k$ -th fold;
  - Compute overall classification error:  $\hat{L}^{(h)} = \frac{1}{K} \sum_{k=1}^K \hat{L}_{(k)}$
  - Compute classification error rate for the  $k$ -th fold testing data.

# Cross Validation: How to Split Data?

## Random CV or K-fold CV?

- random subsampling is computationally more expensive than K-fold CV.

## How to choose K?

- Leave-one-out CV is KCV with  $K=n$ , less computationally efficient than KCV
- The larger K, the less bias but more variance

- Compute classification error rate for the k-th fold testing data.

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

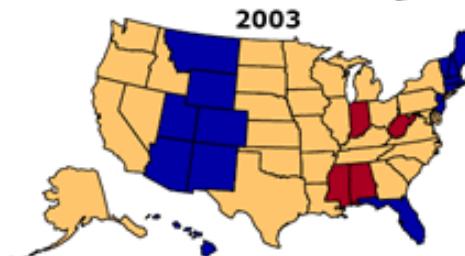
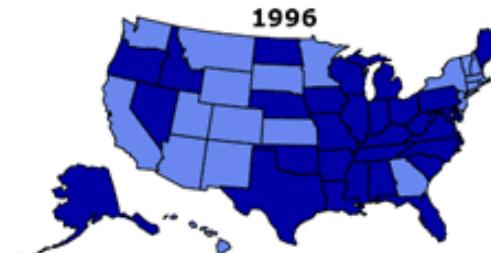
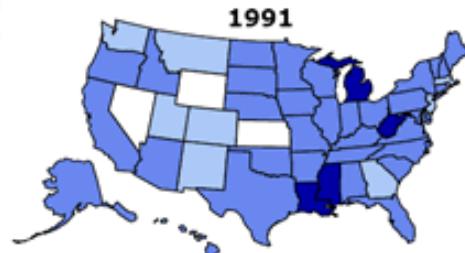
Case Study: The Demographics  
of Obesity

# Obesity in the US

## Obesity Trends\* Among U.S. Adults

BRFSS, 1991, 1996, 2003

(\*BMI  $\geq 30$ , or about 30 lbs overweight for 5'4" person)



Source: Behavioral Risk Factor Surveillance System, CDC.



# Case Study Overview

## Objective:

- Use National Health and Nutrition Examination Survey (NHANES) to create a model used to predict likelihood of obesity.
- Identify predicting factors with predictive power

## Variables:

- Response Variable: Whether an adult is classified as obese
- Predicting Variables: Age, Education Level, Gender

Data

Training Data (4314 Observations)

Testing Data (1000 Observations)

# Case Study Overview

## Objective:

- Use National Health and Nutrition Examination Survey (NHANES) to create a model used to predict life obesity.
- Identify predicting factors with predictive power

## Variables:

- Response Variable: Whether an adult is classified obese
- Predicting Variables: Age, Education Level, Gender

Data

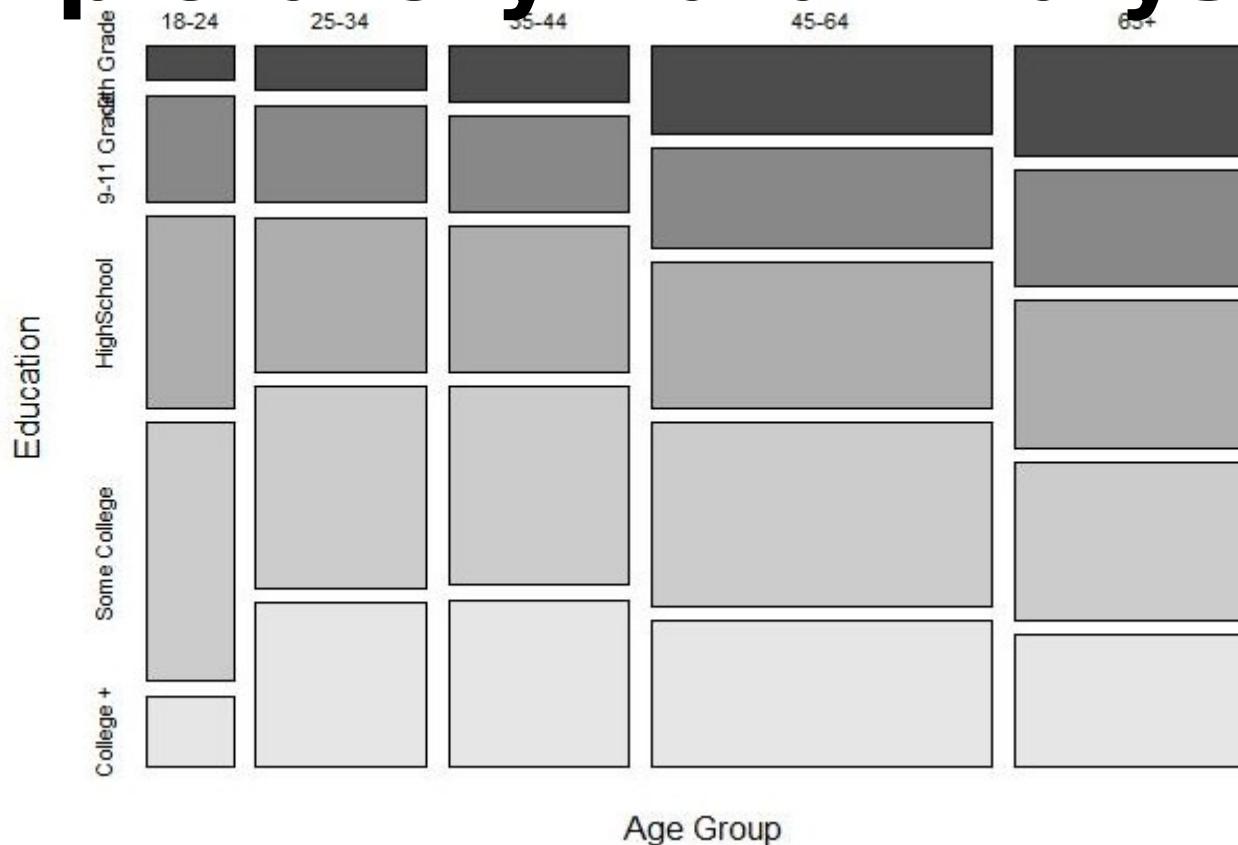
Training Data (4314 Observations)

Testing Data (1000 Observations)

Age variable

- present as a continuous variable in the data
- Recoded in classes (or ranges) like Class 1: 18-24 years Class 2: 25-34 years, etc

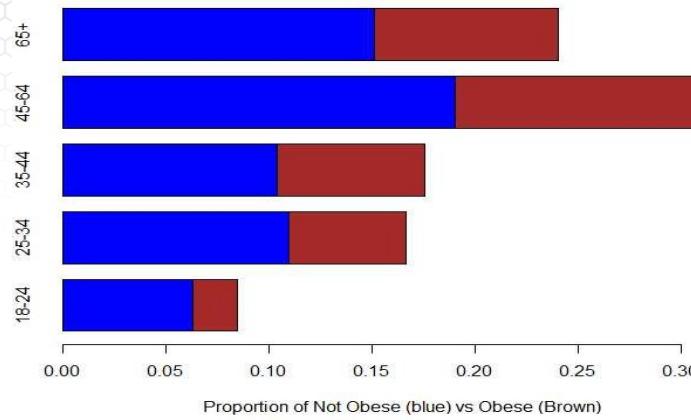
# Exploratory Data Analysis



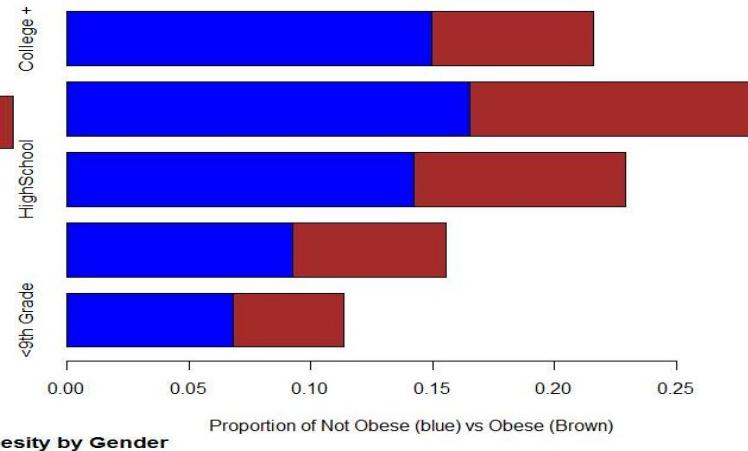
else))  
4",  
'Grade",  
"

# Exploratory Data Analysis

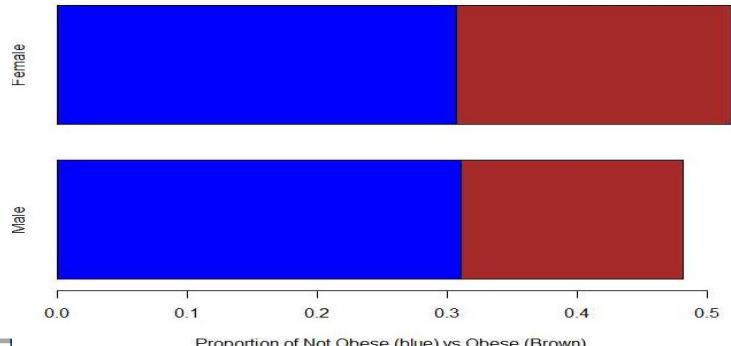
Obesity by Age Group



Obesity by Education Level



Obesity by Gender



Proportion of Not Obese (blue) vs Obese (Brown)

col=c("blue", "brown") *What is Obesity by Education*

7)",

7)",

GTx

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

The Demographics of Obesity:  
Modeling and Prediction

# Model Estimation

## Fit a logistic regression model

```
model = glm(Obesity~agegr+gender+edu,  
family=binomial)  
summary(model)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.2058	0.1573	-7.666	1.78e-14
agegr25to34	0.4727	0.1443	3.276	0.00105
agegr35to44	0.7649	0.1420	5.388	7.13e-08
agegr45to64	0.8482	0.1324	6.406	1.49e-10
agegr65+	0.6009	0.1375	4.370	1.24e-05
genderFemale	0.2304	0.0636	3.621	0.00029
edu9to11Grade	0.0563	0.1223	0.461	0.64511
eduHighSchool	-0.0344	0.1144	-0.301	0.76358
eduSomeCollege	0.1395	0.1104	1.264	0.20630
eduCollege+	-0.4008	0.1176	-3.409	0.00065

Null deviance: 5739.9 on 4313 degrees of freedom

Residual deviance: 5641.3 on 4304 degrees of freedom

# Model Estimation

## Fit a logistic regression model

```
model = glm(Obesity~agegr+gender+edu,  
family=binomial)  
summary(model)
```

	Estimate	Std. Error
(Intercept)	-1.2058	0.15
agegr25to34	0.4727	0.14
agegr35to44	0.7649	0.14
agegr45to64	0.8482	0.13
agegr65+	0.6009	0.13
genderFemale	0.2304	0.06
edu9to11Grade	0.0563	0.12
eduHighSchool	-0.0344	0.114
eduSomeCollege	0.1395	0.110
eduCollege+	-0.4008	0.111

Null deviance: 5739.9 on 4313 degrees of freedom

Residual deviance: 5641.3 on 4304 degrees of freedom

For age group 25 to 34, the log odds of obese increases by 0.4727 OR the odds of obese increases by 1.604 versus the age group 18 to 24 given that all other predicting variables fixed.

For females, the log odds of obese increases by 0.2304 OR the odds of obese increases by 1.259 versus males given that all other predicting variables fixed.

# Statistical Inference

## Test for overall regression

```
gstat = model$null.deviance - deviance(model)
```

```
cbind(gstat, 1-pchisq(gstat,length(coef(model))-1))
```

```
[1, ] 98.636 0
```

```
round(coefficients(summary(model)),4),4)
```

	(Intercept)	agegr25to34	agegr35to44	agegr45to64
agegr65+	0.0000	0.0011	0.0000	0.0000
0.0000				

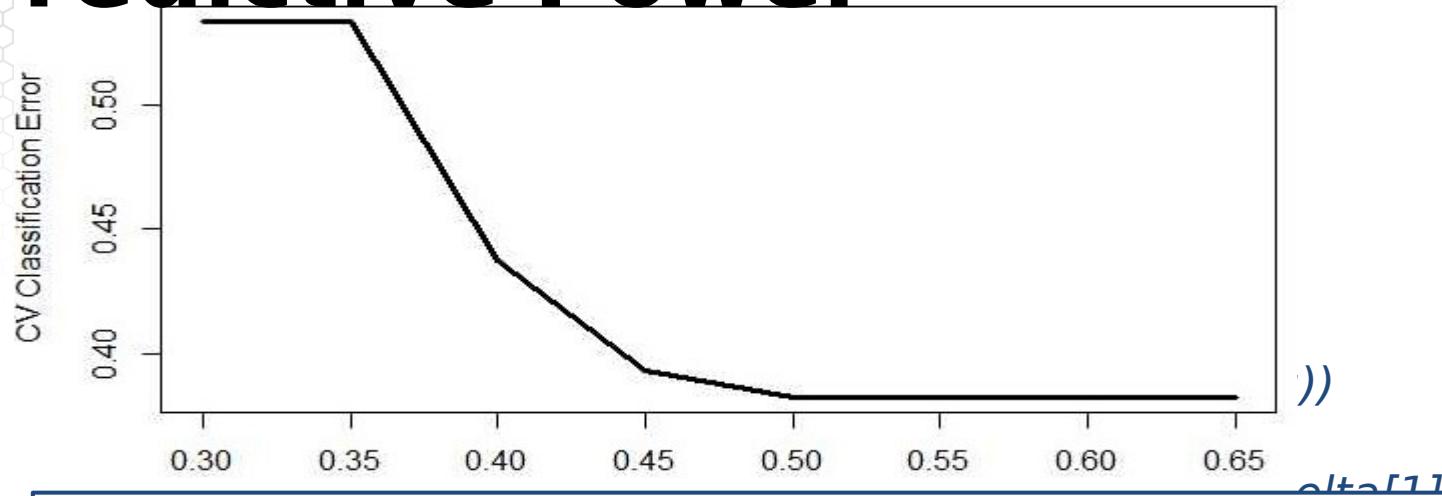
```
genderFemale edu9to11Grade eduHighSchool eduSomeCollege
```

- **Test for overall regression:** p-value<0.01 thus we reject the null hypothesis that all regression coefficients are zero. Thus there are predicting variables which will explain the variability in obesity.
- Education regression coefficients are all (except one) not statistically significant given that we account for age and

# Predictive Power

```
## Prediction Accuracy
library(boot)
cost0.5 = function(y, pi){
  ypred=rep(0,length(y))
  ypred[pi>0.5] = 1
  err = mean(abs(y-ypred))
  return(err)}
obdata.fr = data.frame(cbind(Obesity,agegr,gender,edu))
## classification error for 10-fold cross-validation
cv.err = cv.glm(obdata.fr,model,cost=cost0.5, K=10)$delta[1]
.....
cv.err = c(cv.err0.35, cv.err0.35, cv.err0.4, cv.err0.45, cv.err0.5,
           cv.err0.55, cv.err0.6, cv.err0.65)
## Smallest prediction error is 0.3824
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), cv.err,
      type="l", lwd=3, xlab="Threshold", ylab="CV Classification Error")
```

# Predictive Power



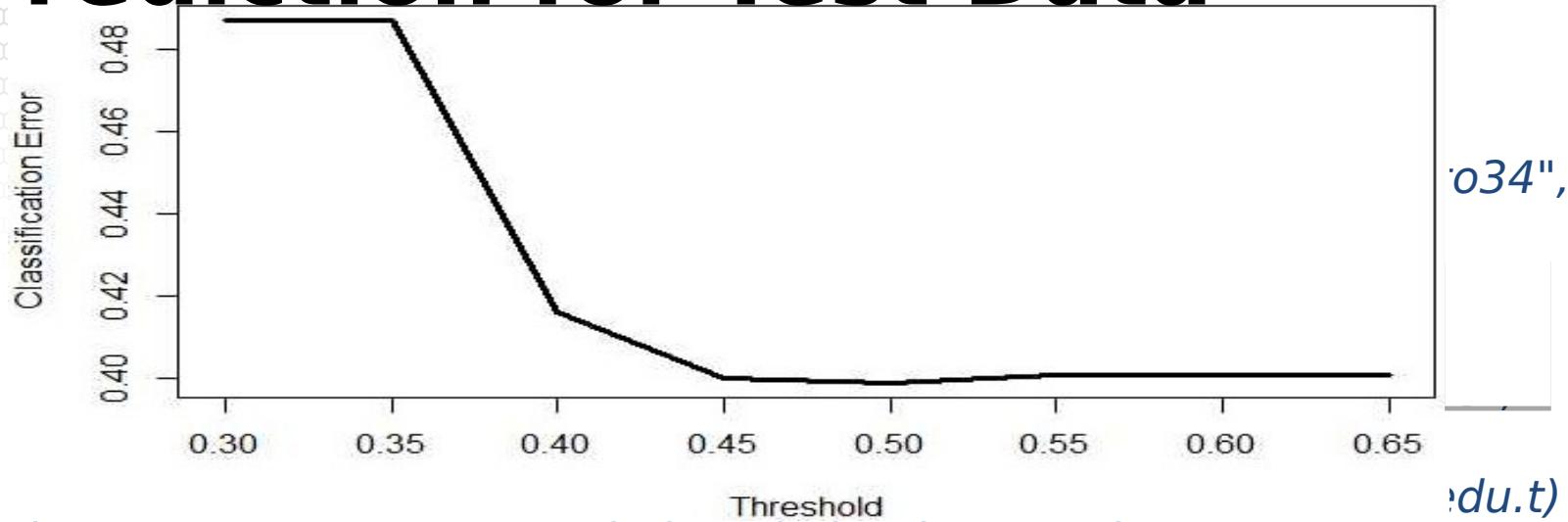
Prediction accuracy is highest and equal for thresholds higher than 0.5. *Why?*

- It is the same as the prediction accuracy if we were to replace all predictions with 0; that is predict everyone is not obese.
- Thus the model does not have any predictive power since it performs worse than the prediction without modeling.

# Prediction for Test Data

```
## Prediction given a set of new observations
## Prepare the test data
testobdata = read.table("testobesitydata.txt",h=T)
agegr.t = factor(testobdata$AgeGroup, labels=c("18to24", "25to34",
                                              "35to44", "45to64", "65+"))
gender.t = factor(testobdata$Gender,labels=c("Male","Female"))
edu.t =
factor(testobdata$Education,labels=c("<9thGrade","9to11Grade",
                                       "HighSchool","SomeCollege","College+"))
pred.data = data.frame(agegr=agegr.t,gender=gender.t,edu=edu.t)
#### Predict
predict.glm(model,pred.data,type="response")
#### Prediction Accuracy for multiple thresholds
err0.3 = cost0.3(testobdata$Obesity,pred.test)
.....
err = c(err0.35,err0.35,err0.4,err0.45,err0.5, err0.55,err0.6,err0.65)
```

# Prediction for Test Data



- Prediction accuracy is highest at 0.5; it is similar as the prediction accuracy if we were to predict everyone is not obese.
- The prediction accuracy using the fitted model did not improve for the test data.

# Regression Analysis

## Logistic Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

The Demographics of Obesity:  
Goodness of Fit

# Logistic Regression With Replications

```
### Aggregate data for Logistic Regression with repetitions
```

```
obdata.agg.n = aggregate(Obesity~agegr+gender+edu,FUN=length)
```

```
obdata.agg.y = aggregate(Obesity~agegr+gender+edu,FUN=sum)
```

```
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
```

```
    Total = obdata.agg.n$Obesity, agegr = obdata.agg.n$agegr,
```

```
    gender=obdata.agg.n$gender, edu=obdata.agg.n$edu)
```

```
## Fit a logistic regression model
```

```
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
```

```
    data = obdata.agg, family=binomial)
```

```
## Test for GOF: Using deviance residuals
```

```
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
```

```
[1] 29.0640209 0.8996714
```

# Logistic Regression With Replications

Coefficients:

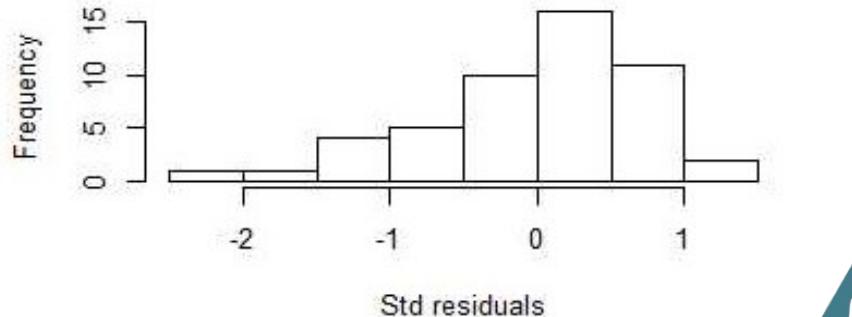
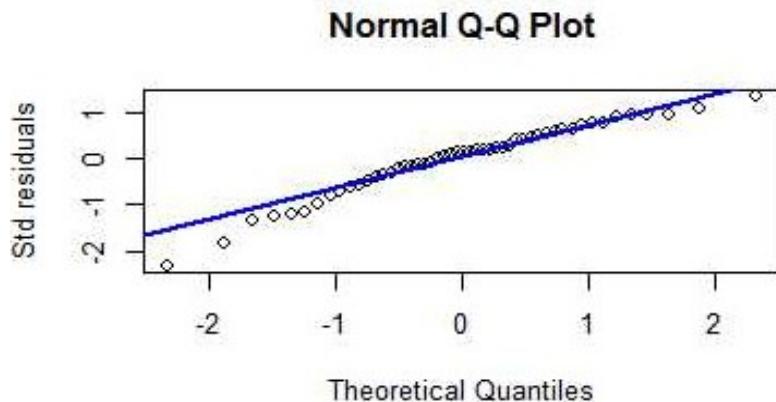
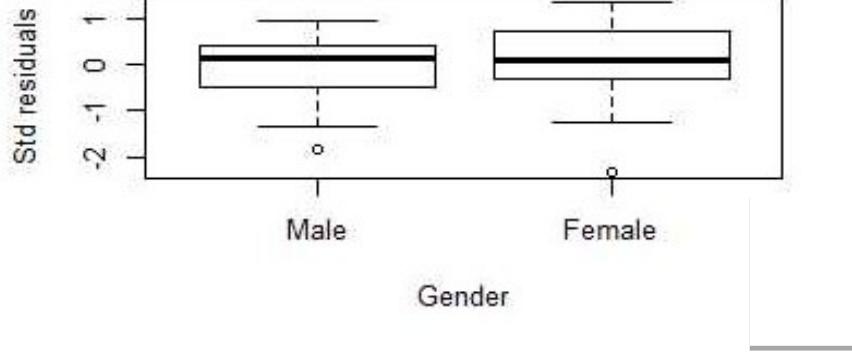
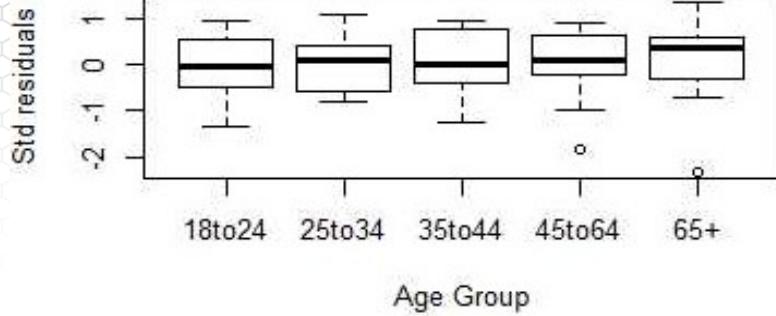
	Estimate	Std. Error	z value
Pr(> z )			
(Intercept)	-1.2058	0.1573	-7.666
1.78e-14			
agegr25to34	0.4727	0.1443	3.276
0.00105			
agegr35to44	0.7649	0.1420	5.388
7.13e-08			
agegr45to64	0.8482	0.1324	6.406
1.49e-10			
agegr65+	0.6009	0.1375	4.370
1.24e-05			
genderFemale	0.2304	0.0636	
3.621			
0.00029			
edu9to11Grade	0.0563	0.1223	0.461
0.64511			
eduHighSchool	-0.0344	0.1144	-0.301
0.76358			

- The output for the estimation and statistical inference on the regression coefficients is the same as for logistic regression without replications.
- The output of the null and residual deviances is different. *Why?*
- With replications, we can perform goodness of fit:  
 $p\text{-value} = 0.899$  thus a good fit

# Residual Analysis

```
res = resid(model.agg,type="deviance")
par(mfrow=c(2,2))
boxplot(res~agegr,xlab="Age Group",ylab = "Std residuals",data =
obdata.agg)
boxplot(res~gender,xlab="Gender",ylab = "Std residuals",data =
obdata.agg)
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```

# Residual Analysis



# Prediction of Adult Obesity:

## Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity BUT the fitted model with education, gender and age group does not improve prediction.
- After factor aggregation, goodness of fit can be performed.
- The p-value of the deviance test for goodness of fit is high, indicating good fit BUT the residual analysis suggests that there may be some departures from normality and thus from goodness of fit.
- Models with different link functions or including interaction terms have not shown improvement. (Results not shown in this lecture.)
- The sample size is large enough for reliable statistical inference.

# Prediction of Adult Obesity:

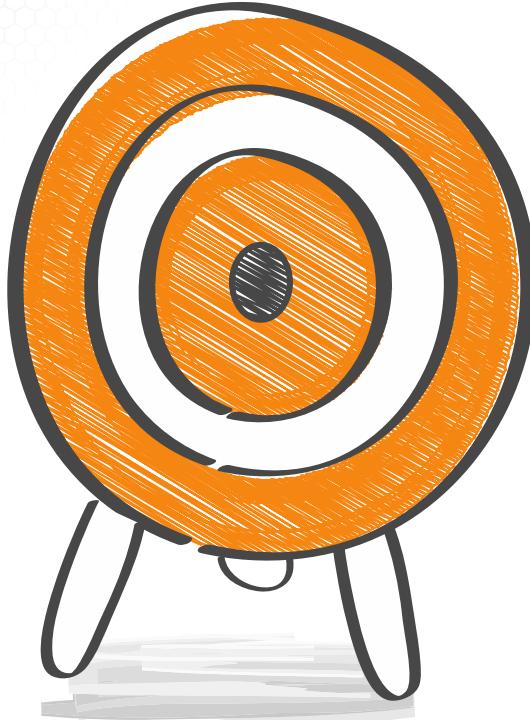
## Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity BUT the fitted model with education, gender and age group does not improve prediction.

### ***What can be done to improve the model fit and the predictive power?***

- Include other factors in the model, such as income level, unemployment, race, ethnicity among others
- Consider interaction terms between age group, education and gender with other factors.
- Models with different link functions or including interaction terms have not shown improvement. (Results not shown in this lecture.)
- The sample size is large enough for reliable statistical inference.

# Summary



# Regression Analysis

## Poisson Regression

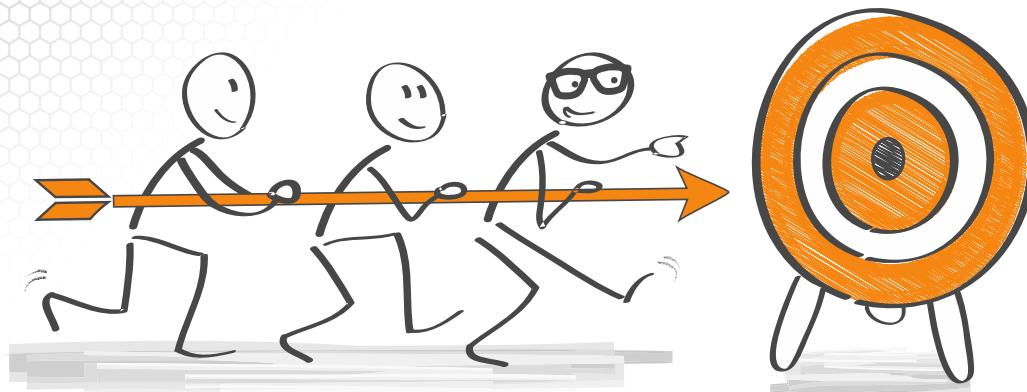
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Introduction

# About this lesson



# Other Distributions of the Response

- What drives the rate of phone calls per day in a calling service center?
- What predicts the density per mile of trees in a forest?
-  The response variable (e.g. rate) has a **Poisson distribution**
- What explains the wait time for a wellness visit at your physician offices?
-  The response variable (e.g. wait time) has an **exponential distribution**

# Standard Linear Regression

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

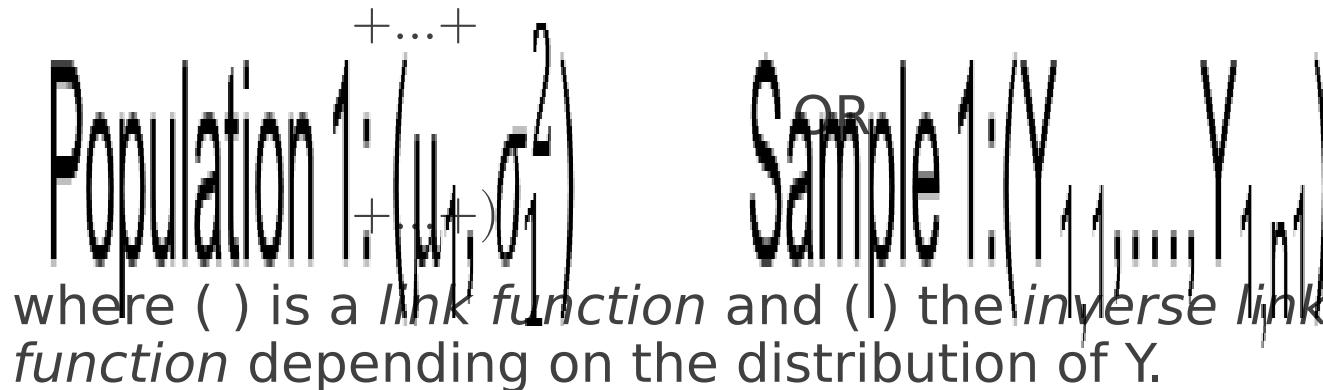
## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- **Normality Assumption:**  $\varepsilon_i \sim \text{Normal}$

# Generalized Linear Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  response variable with a **distribution from the exponential family**

**Model:** Model the conditional expectation:



Population      Sample  
+...+  
OR

where  $(\cdot)$  is a *link function* and  $(\cdot)$  the *inverse link function* depending on the distribution of  $Y$ .

# Generalized Linear Model

$Y \sim$  distribution in the exponential family if its density function

Population 2:  $(\mu_2, \sigma_2^2)$

Sample 2:  $(Y_{2,1}, \dots, Y_{2,n_2})$

where  $\mu$  is the parameter of the distribution and  $g$  is the link function.

Distribution	Link	Regression Function
Normal	$g(\mu) = \mu$	Population k: $(\mu_k, \sigma_k^2)$ Sample k: $(Y_{k,1}, \dots, Y_{k,n_k})$
Poisson	$g(\mu) = \log(\mu)$	Population k: $(\mu_k, \sigma_k^2)$ Sample k: $(Y_{k,1}, \dots, Y_{k,n_k})$
Bernoulli	$g(\mu) \equiv \log(\mu/1-\mu)$	Population k: $(\mu_k, \sigma_k^2)$ Sample k: $(Y_{k,1}, \dots, Y_{k,n_k})$
Gamma	$g(\mu) = 1/\mu$	$\mu =$
Gamma	$g(\mu) = 1/\mu$	Population k: $(\mu_k, \sigma_k^2)$ Sample k: $(Y_{k,1}, \dots, Y_{k,n_k})$

# Poisson Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  response variable with a **Poisson distribution**

**Model:** Model the conditional expectation:

$$\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

OR OR

$$E(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$$

# Standard Linear Regression versus Poisson Regression

## Standard Linear Regression with log transformation:

### transformation:

- $E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $+ \dots +$
- $V(\log(Y)|x_1, \dots, x_p)$  constant
- $V$  constant

## Poisson Regression:

### Poisson Regression:

- $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR  
 $\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Standard Linear Regression versus Poisson Regression

- ~~Model Probability of success given predictor(s)~~  
 $p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$   
Using Standard Linear Regression with log transformation instead of Poisson Regression will result in violations of the assumption of constant variance.
- ~~Alternatively, Standard Linear Regression model is used if the number of counts are large and with the variance stabilizing transformation.~~  
 $g(p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$   
Alternatively, Standard Linear Regression model is used if the number of counts are large and with the variance stabilizing transformation.
- $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR  
 $\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Data Examples

# Data Example 1: High School Awards

**Objective:** To model and predict the number of awards earned by students for multiple high schools.

**Response Variable:** The number of awards earned by students at a high school per year

## **Predicting Variables:**

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"; and
- The score on the final exam in math.

Acknowledgement: *This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.*

# Exploratory Data Analysis

```
## Read data in R
awardsdata = read.csv("students_awards.csv", header=T)
## Convert qualitative variable in the data into factor in R
awardsdata = within(awardsdata, {
  prog = factor(prog, levels=1:3, labels=c("General", "Academic",
"Vocational"))
  id = factor(id)})
```

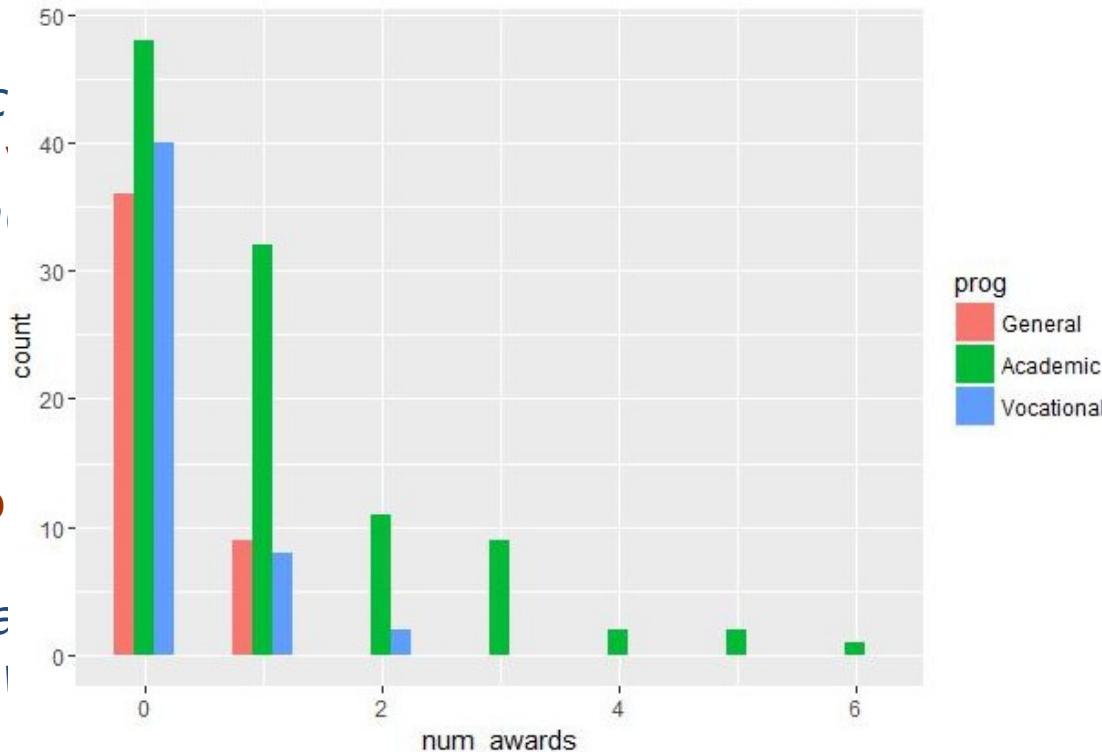
  

```
## Conditional histograms
library(ggplot2)
ggplot(awardsdata, aes(num_awards, fill = prog)) +
  geom_histogram(binwidth=.5, position="dodge")
```

# Exploratory Data Analysis

```
## Read data in R
awardsdata = read.c
## Convert qualitati
awardsdata = within(
  prog = factor(prog,
  "Vocational"))
id = factor(id)})
```

```
## Conditional histo
library(ggplot2)
ggplot(awardsdata, c
geom_histogram(bin
```



# Data Example 2: Insurance Claims

**Objective:** To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

**Response Variable:** The number of car insurance claims per policyholder:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

**Predicting Variables:**

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1-1.5 litre, 1.5-2 litre, >2 litre

# Exploratory Data Analysis

```
## Data in the R library MASS
```

```
library(MASS)
```

```
summary(Insurance)
```

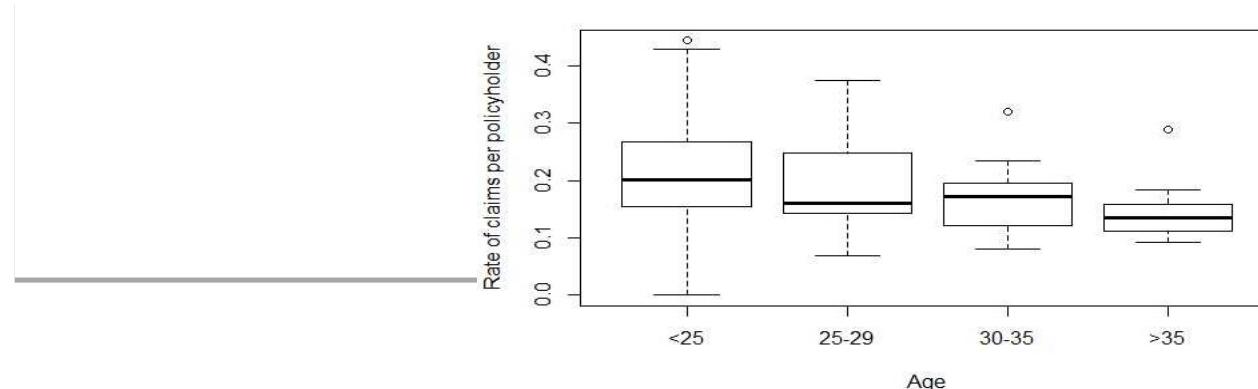
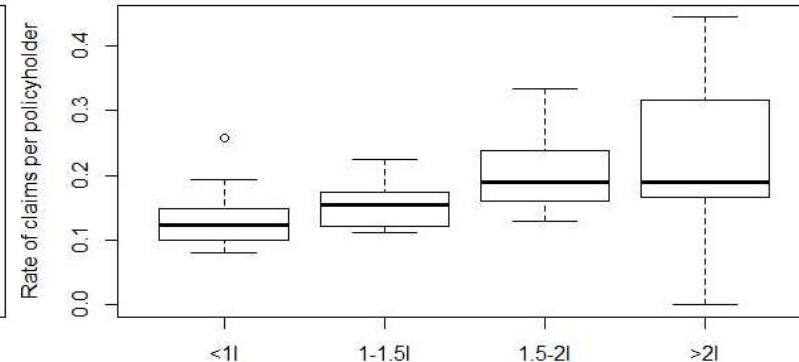
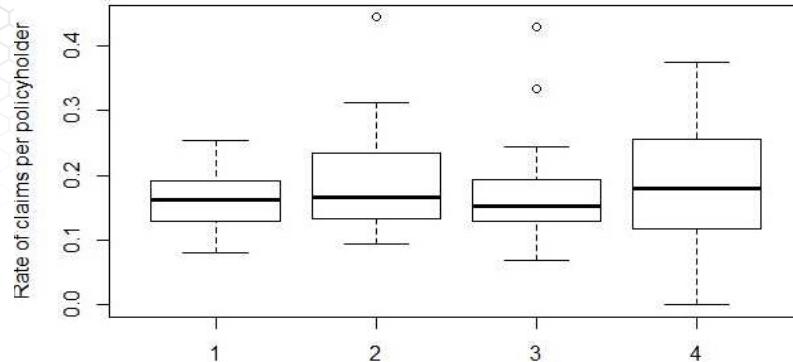
```
## Relationship between rate of claims and predictors
```

```
boxplot(Claims/Holders~District, xlab = "District", ylab = "Rate  
of claims per policyholder",data=Insurance)
```

```
boxplot(Claims/Holders~Group, xlab = "Group", ylab = "Rate of  
claims per policyholder",data=Insurance)
```

```
boxplot(Claims/Holders~Age, xlab = "Age", ylab = "Rate of  
claims per policyholder",data=Insurance)
```

# Exploratory Data Analysis



of

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Description and  
Estimation

# Poisson Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  are event count data per observation unit with a Poisson distribution.

**Poisson Distribution:**  $P(Y_i = y_i) = \text{Poisson}(y_i; \lambda)$

1. *Analysis of the variability in the data – the ANOVA table*

**Model:** Model the conditional expectation:

2. *Testing for equal means*

$Y_i \sim \text{Poisson}(\lambda)$  with

3. *Estimation of simultaneous confidence intervals for the mean differences*

$\mu_i = \log(\mu_j + \text{OR})$  for  $i$  and  $j = 1, \dots, k$

# Model Interpretation

The rate of event occurrence given predicting variable  $X = x$ :

- The log function is the *log rate*.
- The *ratio of the rates* with an increase with one unit in  $x$  is  $e^{\beta_1 + \beta_2 + \dots + \beta_k}$
- If other predicting variables are in the model, then we need to fix all other predicting variables.

# Model Estimation

**Model** the log rate given predictor(s):

$\log(+\dots+$

**Parameters**:  $\beta_0, \beta_1, \dots, \beta_p$

**Approach**: Maximum Likelihood Estimation:

$l(\beta_0, \beta_1, \dots, \beta_p) = \log(L(\beta_0, \beta_1, \dots, \beta_p)) =$

$=$

# Model Estimation (cont'd)

**Approach:** Maximum Likelihood Estimation

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Maximizing the (log-)likelihood function with respect to  $\beta_1$  in close form expression is not possible because the (log-)likelihood function is a non-linear function in the model parameters
- Use numerical algorithm to estimate  $\Rightarrow$

**Upshot:** The estimated parameters and their standard errors are approximate estimates. Do not attempt to do it yourself! Use a statistical software to derive the estimated regression coefficients.

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Estimation: Data  
Example

# Data Example 1: High School Awards

**Objective:** To model and predict the number of awards earned by students at one high school for multiple high schools.

**Response Variable:** The number of awards earned by students at a high school per year

## Predicting Variables:

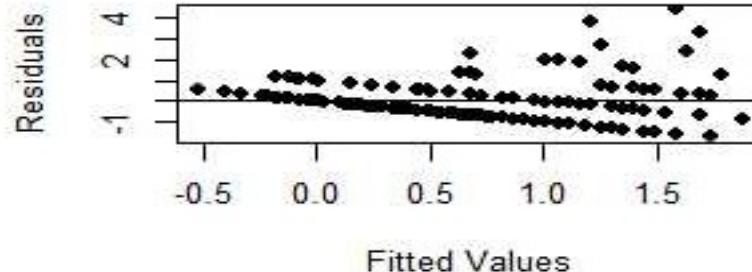
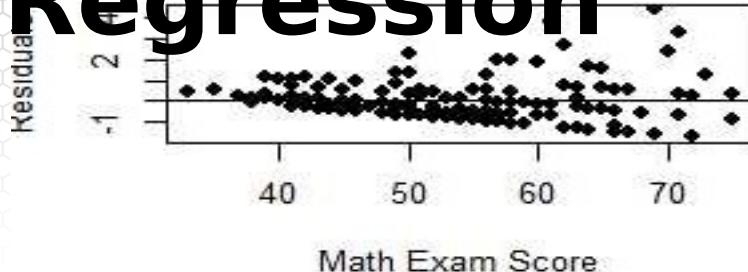
- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"; and
- The score on the final exam in math.

Acknowledgement: This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.

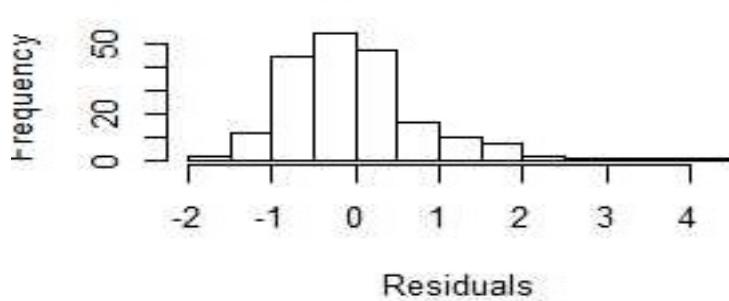
# GOF: Standard Linear Regression

```
## Fit a standard regression model
m0 = lm(num_awards ~ prog + math, data=awardsdata)
## Residual Analysis for Goodness of Fit
par(mfrow = c(2,2))
plot(awardsdata$math, res, xlab = "Math Exam Score", ylab =
"Residuals", pch = 19)
abline(h = 0)
plot(fitted(m0), res, xlab = "Fitted Values", ylab = "Residuals", pch
= 19)
abline(h = 0)
hist(res, xlab="Residuals", main= "Histogram of Residuals")
qqnorm(res)
qqline(res)
```

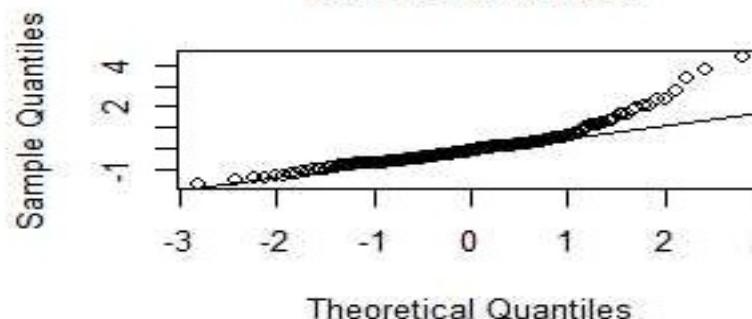
# GOF: Standard Linear Regression



**Histogram of Residuals**



**Normal Q-Q Plot**



44/115 (12%)

# Poisson Regression Estimation

```
m1 = glm(num_awards ~ prog + math, family="poisson",  
data=awardsdata)
```

```
summary(m1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15	***
progAcademic	1.08386	0.35825	3.025	0.00248	**
progVocational	0.36981	0.44107	0.838	0.40179	
math	0.07015	0.01060	6.619	3.63e-11	***

the expected ratio of count of awards per year for one unit increase in the math final exam score is  $\exp(0.07) = 1.072$  given the program

the expected ratio of the counts of awards per year for an academic program vs a general program is  $\exp(1.084) = 2.956$

# Data Example 2: Insurance Claims

**Objective:** To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

**Response Variable:** The number of car insurance claims per policyholder:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

**Predicting Variables:**

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1-1.5 litre, 1.5-2 litre, >2 litre

# Poisson Regression Estimation

*m.ins = glm(Claims ~ District + Group + Age + offset(log(Holders)),  
data = Insurance, family = poisson)*

## Important to note:

- Event rates can be calculated as events per units of varying size; the unit size is called **exposure**;
- In Poisson regression, exposure is accounted for using an **offset**, where the exposure variable enters in the linear combination of the predicting variables, but with the coefficient (for  $\log(\text{exposure})$ ) constrained to 1:
- In this example, the number of policyholders is the exposure since the rate of claims is per policyholder (hence the unit).

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Statistical Inference

# Model Estimation

- The estimator  $\hat{\beta}$  is unbiased for  $\beta$ .
- The sampling distribution of  $\hat{\beta}$  is normal with the covariance matrix depending on the design matrix and  $\sigma^2$ . But we do not know  $\sigma^2$ !

# Statistical Inference

Maximum Likelihood Estimators (MLEs):  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$

Statistical Properties of MLEs:

- Approximate Sampling Distribution:  $\hat{\beta} \approx N(\beta, V)$
- The normal approximation relies on the assumption of large sample size  $\Rightarrow$  Statistical inference is not reliable for small sample data

1- $\alpha$  Approximate Confidence interval

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

# Statistical Inference (cont'd)

Test for statistical significance of  $\beta_j$  given all other predicting variables in the model by using the z-test (Wald test) for

$$H_0: \beta_j = 0 \text{ vs. } H_1: \beta_j \neq 0$$
$$z\text{-value} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)}$$

We reject  $H_0$  if  $|z\text{-value}|$  gets too large. We interpret this as  $\beta_j$  being statistically significant if the null hypothesis is rejected.

# Statistical Inference (cont'd)

z-value =  $\frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$  how large to reject  $H_0: \beta_j = b$ ?

z-value =  $\frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$  how large to reject  $H_0: \beta_j = b$ ?

Analysis of Variance (ANOVA) for multiple regression:

For significance level  $\alpha$ , Reject if z-value >

ANOVA is used to test:  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$

We reject  $H_0$  if F-statistic is large ( $F$  statistic >  $F_{\alpha/2, n-p-1}$ ). Which means that at least one of the coefficients is different from zero at the  $\alpha$  significant level.

The p-value of the test is:  $P(F_{\alpha/2, n-p-1} \geq F\text{-statistic})$  where  $F_{\alpha/2, n-p-1}$  is the  $F$ -distribution with  $p$  and  $n-p-1$  degrees of freedom.

Alternatively, compute  $P\text{-value} = 2P(Z > |\text{z-value}|)$

What if we want to test for positive relationship?

versus



$P\text{-value} = P(Z > z\text{-value})$

What if we want to test for negative relationship?

What if we want to test for negative relationship?

$H_0: \beta_j \geq 0$  versus  $H_A: \beta_j < 0$

$P\text{-value} = P(Z < z\text{-value})$

# Statistical Inference (cont'd)

z-value =  $\frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$  how large to reject  $H_0: \beta_j = b$ ?

z-value =  $\frac{\hat{\beta}_j - b}{\text{se}(\hat{\beta}_j)}$  how large to reject  $H_0: \beta_j = b$ ?

Analysis of Variance (ANOVA) for multiple regression:

- Because the approximation of the normal distribution relies on large sample size, so the hypothesis testing procedures do.
- What if  $n$  is small? The hypothesis testing procedure will have a probability of type I error larger than the significance level; that is, more type I errors than expected.

What if we want to test for negative relationship?

What if we want to test for negative relationship?

$H_0: \beta_j \geq 0$  versus  $H_A: \beta_j < 0$ ?

P-value =  $P(Z \leq z_{\text{value}})$

# Testing for Subsets of Coefficients

*Full model:*

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_q z_q$$

*Reduced model:*

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

*The hypothesis test:*

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$  versus  $H_A : \text{at least one is not zero}$

- Maximize the likelihood function under full model:  $L_f$
- Maximize the likelihood function under reduced model:  $L_r$
- Test Statistic:

$$Dev = \log(L_f) - \log(L_r)$$

$$\text{P-value: } P(Dev)$$

# Testing for Overall Regression

Full model:

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Reduced model:

$= \beta_0$  Test for significance  $\beta_{smoker}$ : p-value=0.0025 thus statistically significant

The hypothesis test:

$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$  versus  $H_A$ : at least one is not zero

- Maximize the likelihood function under full model:  $L$
- Maximize the likelihood function under reduced model:  $L'$

the p-value  $P(\chi^2 > 9.2) = 0.0024$

$$Dev = \log(L')$$

$$P\text{-value. } P(Dev)$$

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Statistical Inference: Data  
Example

# Data Example 1: High School Awards

**Objective:** To model and predict the number of awards earned by students at one high school for multiple high schools.

**Response Variable:** The number of awards earned by students at a high school per year

## **Predicting Variables:**

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"; and
- The score on the final exam in math.

Acknowledgement: *This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.*

# Data Example 1: Statistical Inference

```
model1 = glm(awards ~ prog + math, family="poisson",  
            data=awardsdata)  
summary(m1)
```

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-5.24712	0.65845	-7.969	1.60e-15	***
progAcademic	1.08386	0.35825	3.025	0.00248	**
progVocational	0.36981	0.44107	0.838	0.40179	
math	0.07015	0.01060	6.619	3.63e-11	***

Null deviance: 287.67 on 199 degrees of freedom

Residual deviance: 189.45 on 196 degrees of freedom

$1 - pchisq((287.67 - 189.45), (199 - 196))$

H1.0  
**Test for significance**: p-value  $\approx 0$  thus statistically significant

**Test for overall regression**: p-value  $\approx 0$  thus at least one predicting variables significantly explains the variability in the number of awards

# Data Example 2: Insurance

## ~~Claims~~ **Claims**

**Claims**: To explain factors that are associated to car insurance claims due to accidents or other events leading to car damage.

**Response Variable**: The number of car insurance claims per policyholder:

- Holders: numbers of policyholders; and
- Claims: numbers of claims

## **Predicting Variables**:

- District of residence of policyholder (1 to 4): 4 is major cities.
- Classification of cars with levels <1 litre, 1-1.5 litre, 1.5-2 litre, >2 litre

# Data Example 2: Statistical Inference

*m.ins = glm(Claims ~ District + Group + Age +*

## Test for significance

$\beta_{age.L}$ : p-value  $\approx 0$  thus statistically significant

$\beta_{age.Q}$  &  $\beta_{age.C}$ : p-value  $> 0.1$  thus not statistically significant

**Test for overall regression**: p-value  $\approx 0$  thus at least one predicting variables significantly explains the variability in the number of awards

number of awards

.....

Age.L	-0.394432	0.049404	-7.984	1.42e-15	***
Age.Q	-0.000355	0.048918	-0.007	0.994210	
Age.C	-0.016737	0.048478	-0.345	0.729910	

Null deviance: 236.26 on 63 degrees of freedom

Residual deviance: 51.42 on 54 degrees of freedom

# test for overall regression

*1-pchisq((236.26-51.42),(63-54)) is approximately 0*

# Data Example 2: Statistical Inference

Is the district of residence of policyholder a statistically significant variable given all other predicting variables in the model?

Full model: District + Group + Age

Reduced model: Group + Age

*library(aod)*

*wald.test(b=coef(m.ins), Sigma=vcov(m.ins), Terms=2:4)*

Wald test:

-----  
Chi-squared test:

$\chi^2 = 14.6$ , df = 3,  $P(> \chi^2) = 0.0022$

**Test for subsets of coefficients:** p-value = 0.002 reject the null hypothesis and conclude that the District variable does have significant explanatory power

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment

# Poisson Regression Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  are event count data per observation unit with a Poisson distribution

## **Assumptions:**

- *Linearity Assumption:*  $+$  ...  $+$
- *Independence Assumption:*  $Y_1, \dots, Y_n$  are independent random variables
- *Variance Assumption:*

There is no error term! How to check the assumptions?

# Residuals in Poisson Regression

## Poisson Regression:

$$Y_i | (x_{i1}, \dots, x_{ip}) \sim \text{Poisson}(\lambda)$$

- Estimated rates are:
- **Test for significance  $\beta_{smoker}$ :** p-value=0.151, not statistically significant
- Pearson Residuals:
- **Test for significance  $\beta_{age}$ :** p-value  $\approx 0$ , statistically significant
- Deviance Residuals:

# Residuals in Poisson Regression

## Poisson Regression:

- Pearson's residuals follow directly a normal approximation to a binomial. Hence approximately  $N(0,1)$
- The deviance residuals are the signed square root of the log-likelihood evaluated at the saturated model vs. the fitted model. Thus approximately  $N(0,1)$  if the model is a good fit.
- Deviances play the role of sum of squares in a linear model.

# Goodness of Fit

## GOF Visual Analytics:

- Normal Probability plot & Histogram of the Residuals
- Residuals vs predictors: Linearity & Independence Assumption
- Log of the event rate vs predictors: Linearity Assumptions

~~t-value~~ ~~SE~~ ~~how large to reject  $H_0: \beta_1 = b$ ?~~

## Hypothesis Testing Procedure:

- : the Poisson model fits the data
- : the Poisson model does not fit the data

Deviance test statistic:  $D =$

Under null hypothesis,  $D \sim$  with  $df = n-p-1$

Reject the null that the model is correct if  $p$ -value =  $P(>D)$  small. Note that for this test, we want large  $p$ -values!!!!

# What if No Goodness of Fit?

- Add predicting variables, consider interaction terms, or/and transform predicting variables to improve linearity;
- Identify unusual observations (outliers, leverage points);
- The Poisson distribution isn't appropriate:
  - Overdispersion: the variability of the estimated rates is larger than would be implied by a Poisson model
    - Correlation in the observed responses
    - Heterogeneity in the rates that hasn't been modeled

# Overdispersion

*Overdispersion*: the variability of the response variable is larger than would be implied by the model

Binomial regression model:

- $\hat{y} = \mu$
- Overdispersed Binomial:  $\hat{y} = \mu$

Poisson regression model:

- $\hat{y} = \mu$
- Overdispersed Poisson:  $\hat{y} = \mu$

Overdispersion Parameter:

- Estimate: where  $D$  is the sum of the squared deviances
- If  $D > 0$  then overdispersed model

# Regression Analysis

## Poisson Regression

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Goodness of Fit Assessment:  
Data Examples

# Data Example 1: High School Awards

**Objective:** To model and predict the number of awards earned by students at one high school for multiple high schools.

**Response Variable:** The number of awards earned by students at a high school per year

## **Predicting Variables:**

- The type of program in which the student was enrolled, with three levels: 1 = "General", 2 = "Academic" and 3 = "Vocational"; and
- The score on the final exam in math.

Acknowledgement: *This data example was acquired from the Institute for Digital Research and Education at University of California, Los Angeles.*

# Goodness-Of-Fit

```
## Deviance Test for GOF
```

```
with(m1, cbind(res.deviance = deviance, df = df.residual,
                p = 1 - pchisq(deviance, df.residual)))
```

	res.deviance	df	p
[1,]	189.4496	196	0.6182274

Test for goodness-of-fit:

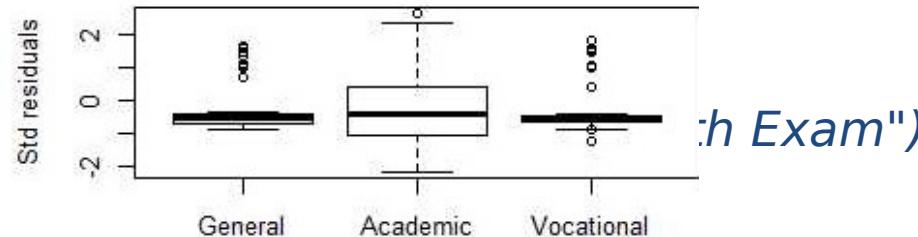
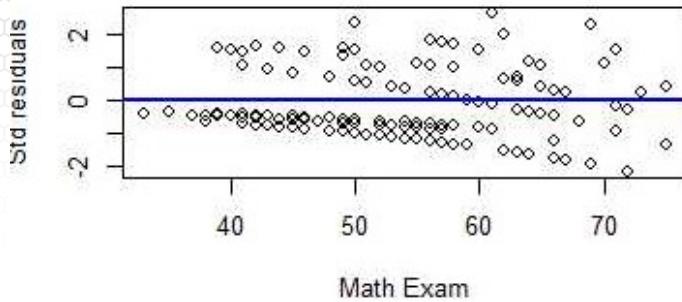
- Using deviance residuals: p-value = 0.61
- Do not reject the null hypothesis of good fit.

# Residual Analysis

```
## Residual Plots
res = resid(m1,type="deviance")
par(mfrow=c(2,2))
plot(awardsdata$math,res,ylab="Std residuals",xlab="Math Exam")
abline(0,0,col="blue",lwd=2)
boxplot(res~prog,ylab = "Std residuals")
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```

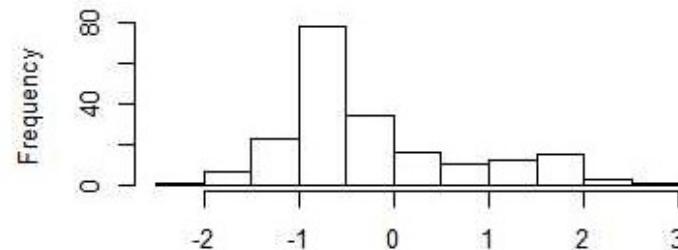
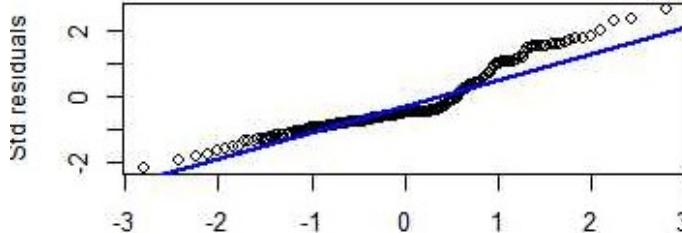
# Residual Analysis

## ## Residual Plots



*(h Exam")*

Normal Q-Q Plot



# Modeling Nonlinear Relationships

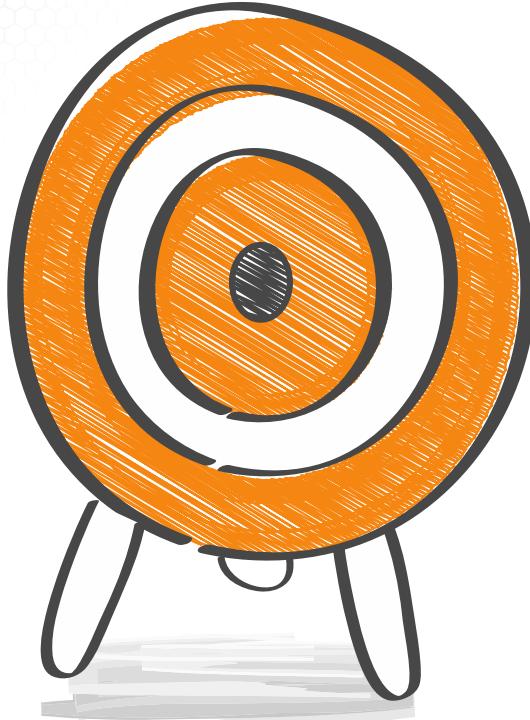
## Fit a logistic regression model with math nonlinearly associated to awards count

```
library(mgcv)
```

```
m2 = gam(num_awards ~ prog + s(math), family="poisson",  
data=awardsdata)
```

- The residuals vs math: downward trend: Consider a **non-parametric** transformation of 'math' predicting variable
- *Nonparametric association*: not specifying the transformation but allowing the data to best identify/fit the transformation
- For this example, we do not see an improvement in the fit.

# Summary



# Regression Analysis

## Model Selection

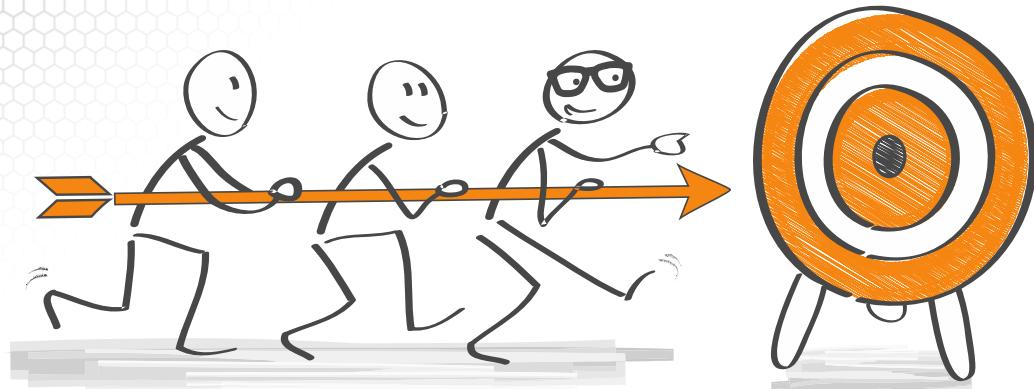
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Introduction

# About this lesson



# Objectives

- High Dimensionality: When we have a very large number of predicting variables to consider, it can be difficult to interpret and work with the fitted model.
  - Multicollinearity: When the predicting variables are correlated, it is important to select variables in such a way that the impact of multicollinearity is minimized.
  - Prediction vs Explanatory Objective: The variables selected for the two objectives will most often be different
-  **Variable Selection** addresses all these objectives.

# Implications and Words of Cautious

- Confounding vs Explanatory Variables: Consider the research hypothesis as well as any potential confounding variables to control for.
- Targeted Predicting Variables: Include in the model the target variable if specified by the research hypothesis
- Over-Interpretation: The selected variables are not necessarily special!
  - Highly influenced by correlations between variables
  - Interpretation of the regression coefficients
  - Causality vs Association

# No Magic Bullet

- Variable selection for large number of predicting variables is an “unsolved” problem in statistics:
- In some sense, model selection is “data mining.”
- Data miners / machine learners often work with many predictors.
- There are no magic procedures to get you the “best model.”

*”All models are wrong, but some are useful.”* George Box

# Notation

Given  $S$  a subset of indices and ( for ) the subset of predicting variables with indices in  $S$ :

- estimated regression coefficients for the submodel with the ( for ) predicting variables
  - fitted values for the submodel with the ( for ) predicting variables (e.g. for regression assuming normality = )
- will refer to this model as the  **$S$  submodel**

# Regression Analysis

## Model Selection

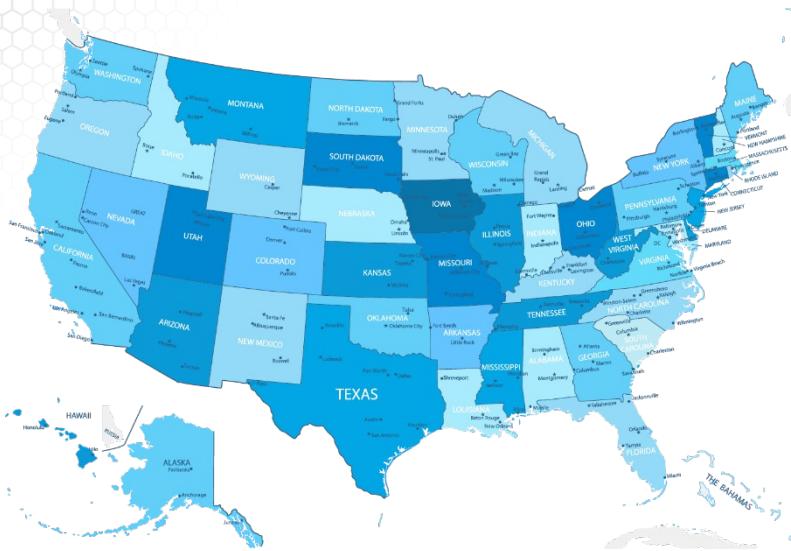
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Data Examples

# Ranking States by SAT Performance



SAT Mean Score by State - Year 1982

790 (South Carolina) -1088  
(Iowa)

*Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?*

# Response & Predicting Variables

**The response variable is:**

**Y** = State average SAT score (verbal and quantitative combined)

**The predicting variables are:**

**“takers”**: % of total eligible students (high school seniors) in the state who took the exam

**“rank”** median percentile of ranking of test takers within their secondary school classes

**“income”**: median income of families of test takers, in hundreds of dollars

**“years”**: average number of years that test takers had in social sciences, natural sciences, and humanities

**“public”**: % of test takers who attended public schools

**“expend”**: state expenditure on secondary schools, in hundreds of dollars per student

# Regression Analysis

`regression.line = lm(sat ~ log(takers) + rank + income + years + public + expend)`  
`summary(regression.line)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	407.53990	282.76325	1.441	0.15675
log(takers)	-38.43758	15.95214	-2.410	0.02032 *
rank	4.11427	2.50166	1.645	0.10734
income	-0.03588	0.13011	-0.276	0.78407
years	17.21811	6.32007	2.724	0.00928 **
public	-0.11301	0.56239	-0.201	0.84168
expend	2.56691	0.80641	3.183	0.00271 **



Model: Model that contains all predictors

- Test for statistical significance:
- : p-value=0.02
  - : p-value>0.1 OR
  - : p-value>0.1
  - : p-value<0.01
  - : p-value>0.1
  - : p-value<0.01

Shall we discard the predicting variables with regression coefficients that are not statistically significant?

NO. Perform variables selection

# Inference on Subset of Coefficients

```
regression.red = lm(sat ~ log(takers) + rank)
```

```
anova(regression.red, regression.line)
```

Model 1: sat ~ log(takers) + rank

Model 2: sat ~ log(takers) + rank + income + years + public  
+ expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	45530				
2	43	26585	4	18945	7.6604	9.42e-05 ***

Testing for a subset of regression coefficients:

Reduced Model (takers and rank only) vs. Full Model

Partial F Test: F-value = 7.6604; P-value  $\approx 0$

# Inference on Subset of Coefficients

```
regression.red = lm(sat ~ log(takers) + rank)
```

```
anova(regression.red, regression.line)
```

Model 1: sat ~ log(takers) + rank

Model 2: sat ~ log(takers) + rank + income + years + public  
+ expend

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	45530				
2	43	26585	4	18945	7.6604	9.42e-05 ***

- **Confounding and explanatory variables:** log(takers) and rank need to be in the model.
  - **Partial F test for explanatory variables:** at least one predicting variable has explanatory power. Which ones?
- 💡 Perform variable selection!!!

# Predicting Bankruptcy

- Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm's prospects.
- Roughly forty years ago Ed Altman showed that publicly available financial indicators can be used to distinguish between firms that are about to go bankrupt and those that are not.

**Which financial indicators are associated with bankruptcy for telecommunications firms?**

*Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University inspired from the honors thesis of Jeffrey Lui.*

# Bankruptcy Data

## Data Sample:

- 25 telecommunication firms, which declared bankruptcy 2000-2002
- 25 telecommunication firms, which did not declare bankruptcy “matched” according to the asset size of the bankrupt firms

## Replicate Experimental Data Setting:

- Matching firms to be comparable with respect to meaningful factors
- Allowing for causal inference

# **Response & Predicting Variables**

**The response variable is:**

**Y** = whether the firm declared bankruptcy

**The predicting variables are:**

**“WC.TA”:** Working capital as a percentage of total assets (in %).

**“RE.TA”:** Retained earnings as a percentage of total assets (in %).

**“EBIT.TA”:** Earnings before interest and taxes as a percentage of total assets (in %).

**“S.TA”:** Sales as a percentage of total assets (in %)

**“BE.TL”:** Book value of equity divided by book value of total

# Exploratory Data Analysis

```
## Read the data from the file
```

```
bankruptcy = read.table("bankruptcy.dat",sep="\t",header=T, row.names=NULL)
attach(bankruptcy)
```

```
## Exploratory analysis
```

```
par(mfrow=c(2,3))
```

```
boxplot(split(WC.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="WC.TA",
main="Boxplot of WC/TA")
```

```
boxplot(split(RE.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="RE.TA",
main="Boxplot of RE/TA")
```

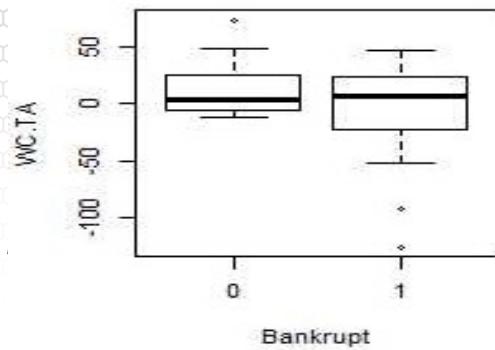
```
boxplot(split(EBIT.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="EBIT.TA",
main="Boxplot of EBIT/TA")
```

```
boxplot(split(S.TA,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="S.TA",
main="Boxplot of S/TA")
```

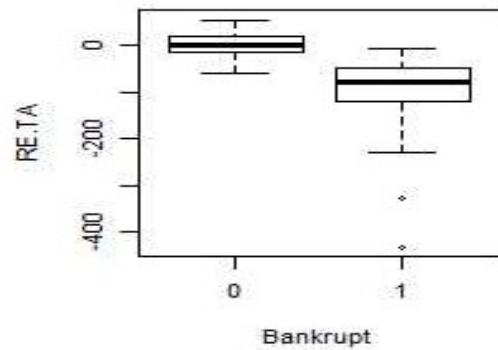
```
boxplot(split(BVE.BVL,Bankrupt),style.bxp="old",xlab="Bankrupt",ylab="BVE.BVL",
main="Boxplot of BVE/BVL")
```

# Exploratory Data Analysis

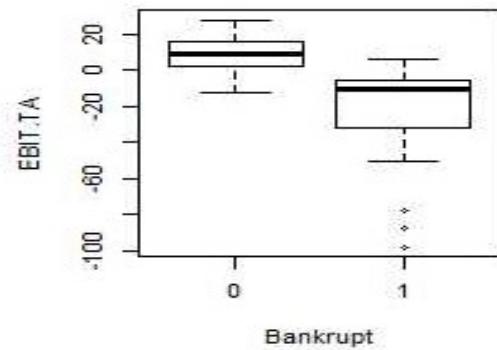
Boxplot of WC/TA



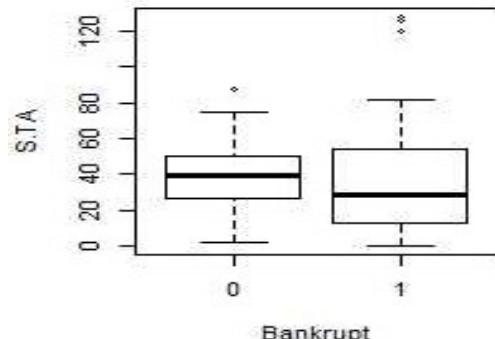
Boxplot of RE/TA



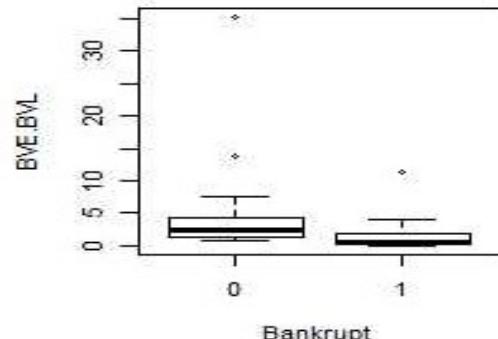
Boxplot of EBIT/TA



Boxplot of S/TA



Boxplot of BVE/BVL



# Regression Analysis

```
bank1 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL, family=binomial)
```

```
summary(bank1)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	7.42646	6.35770	1.168	0.243
WC.TA	-0.15587	0.12208	-1.277	0.202
RE.TA	-0.07605	0.06311	-1.205	0.228
EBIT.TA	-0.49111	0.32260	-1.522	0.128
S.TA	-0.08040	0.09216	-0.872	0.383
BVE.BVL	-2.07764	1.47488	-1.409	0.159



Test for statistical significance:  
All p-values > 0.1: none of the  
coefficients is statistical significant

Null deviance: 69.315 on 49 degrees of freedom

Residual deviance: 11.847 on 44 degrees of freedom

```
gstat = bank1$null.deviance - deviance(bank1)
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-1))
gstat
[1,] 57.46799 4.049594e-11
```



Test for overall regression:  
p-value  $\approx 0$ : The overall  
regression has predictive power

# Regression Analysis

## Model Selection

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Prediction Risk Estimation

# Bias-Variance Tradeoff

## • Variable Selection in Regression with Variance Transformation:

- Including many covariates leads to low bias and high variance
- Including few covariates leads to high bias and

## Poisson Regression

- Prediction Risk: Measure of the Bias-Variance Tradeoff
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR  
for a submodel  $S$ , with the fitted response for model  $S$  and the future observation.

# Bias-Variance Tradeoff

## Variable Selection in Regression with Variance Transformation:

- Including many covariates leads to low bias and high variance
- Including few covariates leads to high bias and

## Poisson Regression

- Prediction Risk: Measure of the Bias-Variance Tradeoff
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR

→ We cannot obtain the prediction risk because we do not have the future observations. *How to estimate?*

# Training Risk

- Replace with actual observations

**Model:** Probability of success given predictor(s)

for a submodel  $S$ , with  $p = p(x_1, \dots, x_p) = P(Y=1|x_1, \dots, x_p)$  the fitted response for model  $S$  and the future observation.

- Use data twice (data snooping): upward bias in the estimate of the risk

- Always prefers larger/more complex model

■ **Correct for the bias**  $g(p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Variable Selection Criteria

## ▪ Correct for the bias:

$$\bullet \widehat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n x_i^2} \equiv \frac{2|S|\hat{\sigma}^2}{n}$$

where  $|S|$  is the model size (number of predictors) and  
is

$$\sum_{i=1}^n y_i(x_i - \bar{x})$$
 the estimated variance based on the full model.  $\frac{2\sqrt{S}\sqrt{\sigma^2}}{n}$

$$\bullet \text{Akaike Information Criterion (AIC)}: = \text{where } n \text{ is the}$$

→ For AIC, we need to replace  $\sigma^2$  with an estimate (from the full model or from the submodel S).

# Variable Selection Criteria (cont'd)

- Bayesian Information Criterion (BIC):

# Where is the true marriage?

- For BIC, we need to replace  $\sigma^2$  with an estimate (from the full model or from the submodel S).
  - $\hat{\sigma}^2 = \hat{\mu}_1 = \hat{\mu}_2 = \dots = \hat{\mu}_k$
  - BIC penalizes complexity more than other approaches and thus preferred in model selection for prediction.

# Variable Selection Criteria (cont'd)

Correct for the bias:  $R_{tr}(S) + \text{Complexity Penalty}$

- Leave-one-out Cross Validation:

- $\hat{R}_{CV}(S) = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_{(i)}(S) - Y_i)^2$
- where  $\hat{Y}_{(i)}(S)$  is the  $i$ -th predicted value from the submodel  $S$  without  $i$ -th observation.
- where  $\hat{Y}_{(i)}(S)$  is the  $i$ -th predicted value from the submodel  $S$  without  $i$ -th observation.

- Leave-one-out Cross Validation Approximation:

- $\hat{R}_{CV}(S) \approx R_{tr}(S) + \frac{2|S|\hat{\sigma}^2(S)}{n}$
- where  $\hat{\sigma}^2(S)$  is the estimated variance based on the  $S$  submodel.

# Variable Selection Criteria (cont'd)

→ ~~Correct for the bias~~  $R_{tr}(S) + \text{Complexity Penalty}$

- Leave-one-out Cross Validation:

- Leave-one-out CV is approximately AIC when the true variance is replaced by the estimate of the variance from the S submodel.  $\text{MSE} \sim \times \frac{2}{n-p-1}$
- Leave-one-out CV penalizes complexity less than Mallow's Cp since (full)

- $\hat{R}_{CV}(S) \approx R_{tr}(S) + \frac{2|S|\hat{\sigma}^2(S)}{n}$
- where  $\hat{\sigma}^2(S)$  is the estimated variance based on the S submodel.

# Generalized Linear Models

- Logistic regression & Poisson regression
- Training Risk

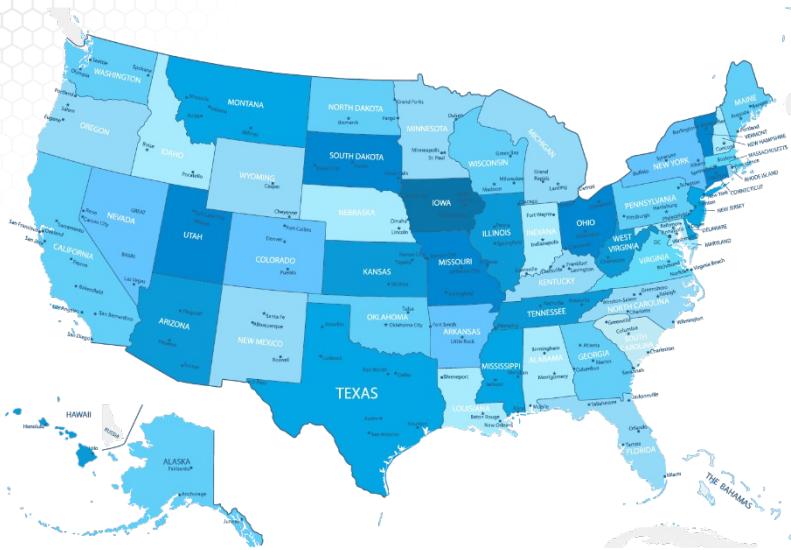
for a submodel  $S$ , with the fitted response for model  $S$  and the future observation.



Correction for the bias

- AIC & BIC are commonly used for model selection for GLMs

# Ranking States by SAT Performance



SAT Mean Score by State - Year 1982

790 (South Carolina) -1088  
(Iowa)

*Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?*

# Model Selection Criteria Using R

```
library(CombMSC)
n = nrow(datasat)
## full model
c(Cp(regression.line,S2=24.86),
AIC(regression.line,k=2),AIC(regression.line,k=log(n)))
[1] 7 472 487
```

## reduced model

```
c(Cp(regression.red,S2=24.86),
AIC(regression.red,k=2),AIC(regression.red,k=log(n)))
[1] 29 499 500
```

**Mallow's Cp: = 24.86 is the estimated standard deviation for the full model**

- BIC: It is similar to AIC except that the AIC complexity is further penalized by  $\log(n)/2$
- The values of the three criteria are different and not comparable
- The full model is better according to all three criteria

# Regression Analysis

## Model Selection

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Search

# Bias-Variance Tradeoff

- ~~Variable Selection: Bias vs Variance~~ The Least Squares estimated coefficients have specific interpretations. Including many covariates leads to low bias and high variance
  - ✓  $\hat{\beta}_0$  is the estimated expected value of the response variable when all predicting variables equal zero;
  - Prediction Risk: Measure of the Bias-Variance Tradeoff
  - ✓  $\hat{\beta}_i$  is the estimated expected change in the response variable associated with one unit of change in the  $i$ -th predicting variable holding fixed all other predictors in the model for a new observation. This is the slope of the fitted regression line for model S and the future observation.

# Bias-Variance Tradeoff

- ~~Variable Selection: Bias vs. Variance~~  
The Least Squares estimated coefficients have specific interpretations. Including many covariates leads to low bias and high variance
- $\hat{\beta}_0$  is the estimated expected value of the response variable when all predicting variables equal zero;
- *Prediction Risk*: Measure of the Bias-Variance Tradeoff

→ Given an estimate of the prediction risk for a submodel  $S$ , choose the submodel with smallest prediction risk. *How to search over all submodels?*

# Model Search

- a. Estimated Regression Coefficients
  - If  $p$  is the number of predicting variables, there are  $2^p$  possible submodels:
- b. Conditional model:
  - $\hat{\beta}_{adv} = 1.192$   
The expected additional gain in sales in thousands for \$100 additional expenditure in advertisement **while holding all other fixed**.
  - If  $p$  small, fit all submodels
  - If  $p$  large, search using heuristics/greedy search
- Stepwise Regression:
  - Forward: Start with no predictor and one at a time
  - $\hat{\beta}_{adv} = 2.772$   
The expected additional gain in sales in thousands for \$100 additional expenditure in advertisement **not accounting for other predicting variables**
  - Backward: Start with all predictors and drop one at a time
  - Forward-Backward: Add and drop one variable at a time iteratively

# Model Search

- ~~if  $p$  is the number of predicting variables, there are  $2^p - 1$  possible regression models~~

- Stepwise regression is a greedy algorithm; it does not guarantee to find the model with the best score
- Forward stepwise regression is preferable over backward stepwise regression
- It does not necessarily select the same model as the one selected using backward stepwise regression

expenditure in advertisement not accounting for other predicting variables

- Forward-Backward: Add and drop one variable at a time iteratively

# Forward Stepwise Regression

- The estimator  $\hat{\beta}$  is unbiased for  $\beta$ .
- The sampling distribution of  $\hat{\beta}$  is normal with the covariance matrix depending on the design matrix and  $\sigma^2$ . But we do not know  $\sigma^2$ !

# Backward Stepwise Regression

- Select criterion for model selection (e.g. AIC)
- Fit *full* model and discard one predictor for all  $j=1, \dots, p$ :
  - the criterion value for the model with the  $j$ -th predictor discarded
  - Select predictor to be discarded with the smallest criterion value if it is smaller than the criterion value for the full model.
- Fit the regressions without predictor and discarding another predictor for all  $j=1, \dots, -1$ :
  - the criterion value for the model with the  $j$ -th predictor discarded
  - Select predictor with the smallest criterion value to discard from the model if smaller than
    - if larger than then stop; the selected model discards only the  $j$ -th predictor;
- Continue discarding predictors until the criterion does not improve

# Backward Stepwise Regression

- Select criterion for model selection (e.g. AIC)
- Fit *full* model and discard one predictor for all  $j=1, \dots, p$ :

→ If the criterion does not improve, then the predictor is discarded

- It cannot be performed if  $p$  larger than  $n$
- More computationally expensive than forward stepwise regression
- It will select larger models if  $p$  large

- If the criterion does not improve, then the predictor is discarded
- if larger than then stop; the selected model discards only the  $-th$  predictor;
  - Continue discarding predictors until the criterion does not improve

# Regression Analysis

## Model Selection

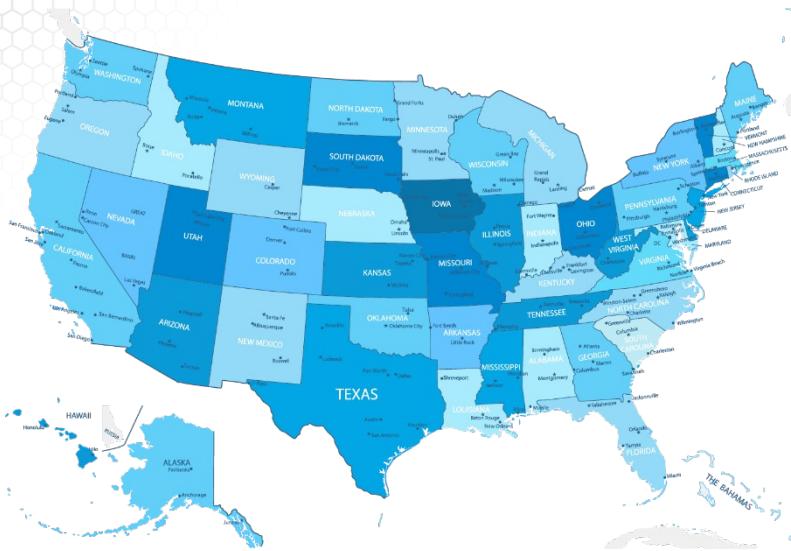
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Model Search: Data Examples

# Ranking States by SAT Performance



SAT Mean Score by State - Year 1982

790 (South Carolina) -1088  
(Iowa)

*Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?*

# Compare All Models

```
library(leaps)
out = leaps(data$at[,-c(1,2)], sat, method =
"Cp")
cbind(as.matrix(out$which),out$Cp)
```

1 2 3 4 5 6  
1 0 0 0 0 0 1 34.026  
1 1 0 0 0 0 0 47.639  
1 0 1 0 0 0 0 187.387  
1 0 0 1 0 0 0 269.647  
1 0 0 0 1 0 0 306.188  
1 0 0 0 0 1 0 307.076  
....

```
best.model = which(out$Cp==min(out$Cp))
cbind(as.matrix(out$which),out$Cp)
[best.model,]
```

1 2 3 4 5 6  
0 0 1 1 1 1 3.581157

The output includes all 64 combinations of predictors with specification of which predictors are in the model and the Cp score value for each model.

The best model with respect to Mallow's Cp criterion: *Years, Public, Expend, Rank* (last four predictors in the input dataset)

**Does not allow for specification of confounding variables!!!**

# Stepwise Regression

## # Forward Stepwise Regression

```
step(lm(sat~log(takers)+rank), scope = list(lower=sat~log(takers)+rank,  
upper = sat~log(takers)+rank+expend+years+income+public), direction =  
"forward")
```

Start: AIC=346.7

sat ~ log(takers) + rank

	Df	Sum of Sq	RSS	AIC
+ expend	1	13149.5	32380	331.66
+ years	1	9827.2	35703	336.55
<none>			45530	346.70
+ income	1	1305.3	44224	347.25
+ public	1	15.9	45514	348.69

Step: AIC=331.66

sat ~ log(takers) + rank + expend

	Df	Sum of Sq	RSS	AIC
		17.857		

Step: AIC=323.9

sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
<none>			26637	323.90
+ income	1	26.6165	26610	325.85
+ public	1	4.5743	26632	325.89

Call:

```
lm(formula = sat ~ log(takers) + rank + expend +  
years)
```

Coefficients:

(Intercept)	log(takers)	rank	expend
388.425	-38.015	4.004	2.423

# Stepwise Regression

## # Forward Stepwise Regression

```
step(lm(sat~log(takers)+rank), scope = list(lower=sat~log(takers)+rank,  
upper = sat~log(takers)+rank+expend+years+income+public), direction =  
"forward")
```

- Stepwise regression in R allows for specification of a reduced model, including confounding variables
- Selected model: expend & years with confounding variables log(takers) & rank

346.70  
+ income 1 1305.3 44224

347.25  
+ public 1 15.9 45514  
348.69

Step: AIC=331.66

sat ~ log(takers) + rank + expend  
Df Sum of Sq RSS AIC

Call:

lm(formula = sat ~ log(takers) + rank + expend +  
years)

Coefficients:

(Intercept)	log(takers)	rank	expend
388.425	-38.015	4.004	2.423
17.857			

# Stepwise Regression (cont'd)

## # Backward Stepwise Regression

```
full = lm(sat~log(takers)+rank+expend+years+income+public)
```

```
minimum = lm(sat~log(takers)+rank)
```

```
step(full, scope = list(lower=minimum, upper = full), direction = "backward")
```

Start: AIC=327.8

sat ~ log(takers) + rank + expend  
+ years + income + public

	Df	Sum of Sq	RSS	AIC
--	----	-----------	-----	-----

- public	1	25.0	26610	325.85
----------	---	------	-------	--------

- income	1	47.0	26632	325.89
----------	---	------	-------	--------

<none>			26585	327.80
--------	--	--	-------	--------

- years	1	4588.8	31174	333.77
---------	---	--------	-------	--------

- expend	1	6264.4	32850	336.38
----------	---	--------	-------	--------

Step: AIC=325.85

sat ~ log(takers) + rank + expend + years +  
income

	Df	Sum of Sq	RSS	AIC
--	----	-----------	-----	-----

- income	1	26.6	26637	323.90
----------	---	------	-------	--------

<none>			26610	325.85
--------	--	--	-------	--------

- years	1	5452.8	32063	333.17
---------	---	--------	-------	--------

- expend	1	7430.3	34040	336.16
----------	---	--------	-------	--------

Step: AIC=323.9

sat ~ log(takers) + rank + expend + years

	Df	Sum of Sq	RSS	AIC
--	----	-----------	-----	-----

<none>			26637	323.90
--------	--	--	-------	--------

- years	1	5743.5	32380	331.66
---------	---	--------	-------	--------

- expend	1	9065.8	35703	336.55
----------	---	--------	-------	--------

# Stepwise Regression (cont'd)

# Backward Stepwise Regression

```
full = lm(sat~log(takers)+rank+expend+years+income+public)
```

```
minimum = lm(sat~log(takers)+rank)
```

```
step(full, scope = list(lower=minimum, upper = full), direction = "backward")
```

- Selected model: expend & years with confounding variables log(takers) & rank
- The same model was selected using forward regression; generally for a large number of predictors the two methods will select different models

Step: AIC=323.9			
sat ~ log(takers) + rank + expend + years			
	Df	Sum of Sq	RSS
<none>			26637 323.90
- years	1	5743.5	32380 331.66
- expend	1	9065.8	35703 336.55

# Predicting Bankruptcy

- Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm's prospects.
- Roughly forty years ago Ed Altman showed that publicly available financial indicators can be used to distinguish between firms that are about to go bankrupt and those that are not.

**Which financial indicators are associated with bankruptcy for telecommunications firms?**

*Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University*

# Compare All Models

```
out = leaps(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL), Bankrupt)
```

```
best.model = which(out$Cp == min(out$Cp))
```

```
as.matrix(out$which)[best.model,]
```

```
1 2 3 4 5
```

```
FALSE TRUE TRUE FALSE TRUE
```

```
bank2 = glm(Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family=binomial, x=T)
```

```
summary(bank2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.29478	1.12317	-0.262	0.7930
RE.TA	-0.05627	0.02745	-2.050	0.0400
EBIT.TA	-0.16763	0.09269	-1.808	0.0705
BVE.BVL	-0.62975	0.39429	-1.597	0.1102

```
gstat = deviance(bank2) - deviance(bank1)
```

```
cbind(gstat, 1-pchisq(gstat,length(coef(bank1))-length(coef(bank2))))
```

```
gstat
```

```
[1,] 4.040336 0.1326332
```

The best model selected with respect to Mallow's Cp: *RE.TA, EBIT.TA, BE.BVL*

- *RE.TA* is now statistically significant at  $\alpha = 0.05$
- Not all coefficients are statistically significant

The null (reduced model) not rejected

# Remove Outlier

```
bankrupt2 = bankrupt[-1,]  
attach(bankrupt2)  
bank3 = glm(Bankrupt ~ WC.TA + RE.TA + EBIT.TA + S.TA + BVE.BVL,  
family=binomial, data=bankrupt2)  
summary(bank3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	265.467	576281.709	0	1
WC.TA	-4.297	12439.717	0	1
RE.TA	-1.516	5131.146	0	1
EBIT.TA	-17.043	35543.170	0	1
S.TA	-2.859	7408.747	0	1
BVE.BVL	-77.540	184903.000	0	1

The model fits perfectly. This is complete separation, and the solution is to simplify the model if that is possible.

# Compare All Models: Without Outlier

```
out = leaps(cbind(WC.TA, RE.TA, EBIT.TA, S.TA, BVE.BVL), Bankrupt)
```

```
best.model = which(out$Cp == min(out$Cp))
```

```
as.matrix(out$which)[best.model,]
```

1	2	3	4	5
---	---	---	---	---

FALSE	TRUE	TRUE	FALSE	TRUE
-------	------	------	-------	------

```
bank4 = glm(Bankrupt ~ RE.TA + EBIT.TA + BVE.BVL, family=binomial, x=T)
```

```
summary(bank4)
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.09166	1.47135	-0.062	0.9503
RE.TA	-0.08229	0.04230	-1.945	0.0517
EBIT.TA	-0.26783	0.15854	-1.689	0.0912
BVE.BVL	-1.21810	0.76536	-1.592	0.1115

```
exp(coef(bank2)[-1])
```

RE.TA	EBIT.TA	BVE.BVL
0.9452862	0.8456655	0.5327273

```
exp(coef(bank4)[-1])
```

RE.TA	EBIT.TA	BVE.BVL
0.9210091	0.7650371	0.2957930

# Stepwise Regression: Without Outlier

```
bank3.select=step(bank3,direction = "backward")  
summary(bank3.select)
```

Start: AIC=12

Bankrupt ~ WC.TA + RE.TA +  
EBIT.TA + S.TA + BVE.BVL

	Df	Deviance	AIC
- S.TA	1	0.000	10.000
<none>		0.000	12.000
- WC.TA	1	9.384	19.384
- RE.TA	1	10.736	20.736
- EBIT.TA	1	14.799	24.799
- BVE.BVL	1	19.027	29.027

Step: AIC=10

Bankrupt ~ WC.TA + RE.TA +  
EBIT.TA + BVE.BVL

	Df	Deviance	AIC
<none>		0.000	10.000
- WC.TA	1	9.384	17.384
- RE.TA	1	12.853	20.853
- EBIT.TA	1	14.867	22.867
- BVE.BVL	1	19.132	27.132

Stepwise regression selects four predictors vs.  
best subset selection selects three predictors.

# Regression Analysis

## Model Selection

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Regularized Regression: Penalties

# Bias-Variance Tradeoff

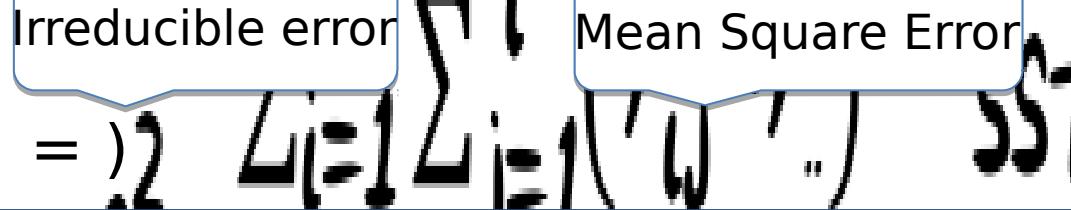
- *Prediction Risk*: Measure of the Bias-Variance Tradeoff

$$\text{Prediction Risk} = \frac{\sum_{i=1}^N \sum_{j=1}^{N-1} (y_{ij} - \bar{y}_{ij})^2}{SST}$$

for a submodel  $S$ , with the fitted response for model  $S$  and the future observation

# Bias-Variance Tradeoff

- *Prediction Risk*: Measure of the Bias-Variance Tradeoff

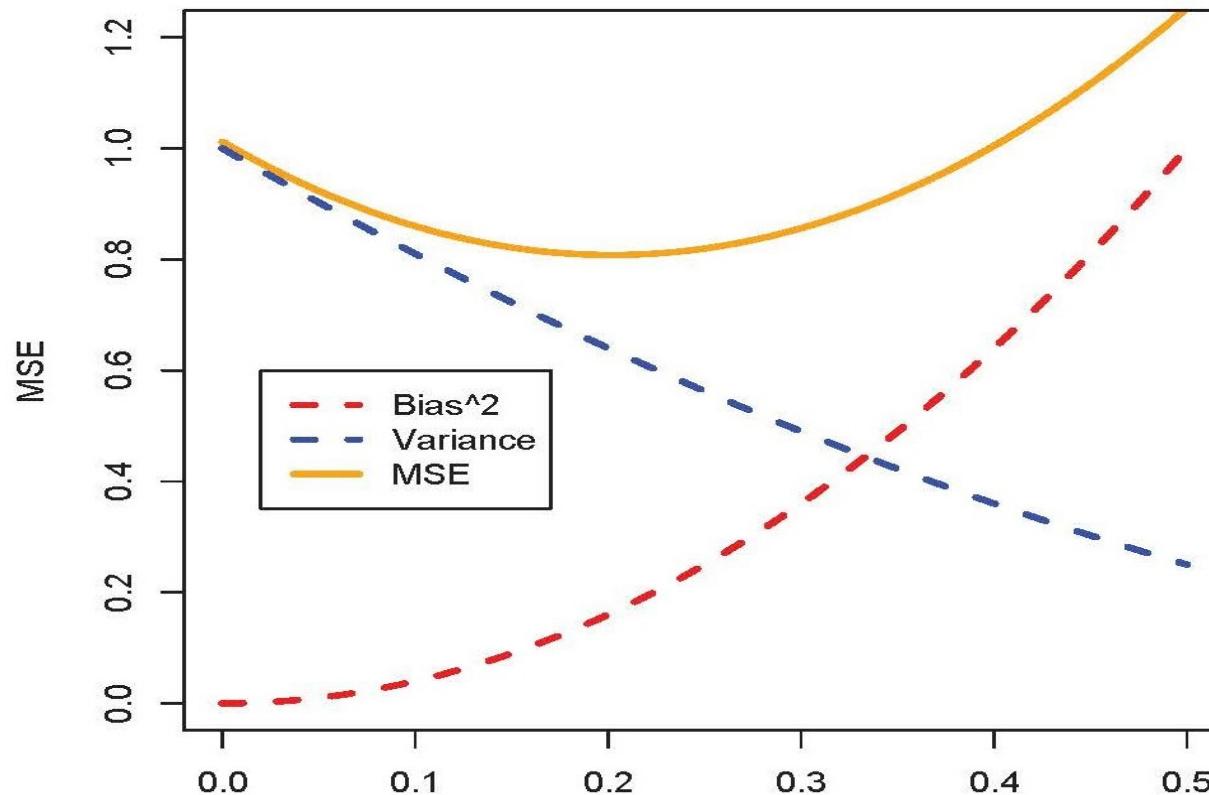


The diagram illustrates the decomposition of Mean Square Error (MSE). It shows a wavy line representing the total error, which is the sum of two components: 'Irreducible error' and 'Mean Square Error'. The 'Irreducible error' is represented by a flat line at the bottom. The 'Mean Square Error' is represented by the vertical distance between the wavy line and the flat line. Arrows point from the labels to their respective parts of the equation.

$$\text{MSE} = \text{Irreducible error} + \text{Mean Square Error}$$

- Sometimes, it is possible to find a model with lower MSE than an unbiased model!
- It is “generic” in statistics: almost always introducing some bias yields a decrease in MSE.

# Bias-Variance Tradeoff



# Biased Regression: Penalties

Not all biased models are better – we need a way to find “good” biased models!

- Penalize large values of  $\beta$ 's jointly. This should lead to “multivariate” shrinkage of the vector  $\beta$ .
- Heuristically, “large” is interpreted as “complex model”.
- Goal is really to penalize “complex” models.
- If truth really is complex, this may not work! (But, it will then be hard to build a good model anyways ...)

# Regularized Regression

**Without Penalization:** Estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  by minimizing the sum of squared errors

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2$$

**With Penalization:** Estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  by minimizing the penalized sum of squared errors

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \text{Penalty}(\beta_1, \dots, \beta_p)$$

The bigger  $\lambda$ , the bigger the penalty for model complexity.

# Regularized Regression

## (cont'd)

$$Q(\beta_1, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \text{Penalty}(\beta_1, \dots, \beta_p)$$

How will the procedure change if we test:

$H_0: \beta_j = b$  vs.  $H_a: \beta_j \neq b$  for some known  $b$ ?

# Comparing Penalties

- penalty: provides the best model given a selection criterion but it requires fitting all submodels
- ~~penalty measures sparsity~~

Example: Consider the following two vectors of length  $p$

~~Test for significance  $\beta_{smoker}$ : p-value=0.0025 thus statistically significant~~

~~Test for overall regression: Null deviance - Residual Deviance = 9.2 with  
u is sparse since it contains many zeros~~

~~& versus &~~

~~• penalty is easy to implement but it does not do~~

~~variable selection~~

# Regression Analysis

## Model Selection

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Regularized Regression:  
Approaches

# Variable Standardization & Notation

For regularized regression,

- Rescale the  $j$ -th predicting variable for  $j=1, \dots, p$  as follows:

**Test for significance  $\beta_{\text{smoker}}$ :** p-value=0.151, not statistically significant

It is recommended to also

- **Test for significance  $\beta_{\text{age}}$ :** p-value < 0, statistically significant

▪ **Use the original scale when fitting the selected model for interpretation of the regression coefficients.**

# Ridge Regression

- Minimize

$$SSE_{\lambda}(\beta) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \sum_{i=1}^n \beta_i^2$$

- The estimated regression coefficients:

where  $\mathbf{I}$  is the identity matrix

When  $\lambda = 0$  we get the least squares estimate (low bias, high variance). When  $\lambda = 0$  we get  $\hat{\beta} = 0$  (high bias, low variance).

- Commonly used to fit a regression model under multicollinearity
- Not used for model selection: it does not “force” any

# Lasso Regression

- Lasso (Least Absolute Shrinkage and Selection)
- Normal Linear Regression: Minimize

$$SSE_{\lambda}(\beta_0, \dots, \beta_p) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda \sum_{j=1}^p |\beta_j|$$

- Generalized Linear Model: Minimize

$$SSE_{\lambda}(\beta_0, \dots, \beta_p) = l(\beta_0, \dots, \beta_p) + \lambda \sum_{j=1}^p |\beta_j|$$

where  $l(\beta)$  is the log-likelihood function.

- The estimated regression coefficients: Use numerical algorithms since there is not a close form expression
- Used for model selection: it can “force” any

# Choosing $\lambda$ : Cross-Validation

- The estimator  $\hat{\beta}$  is unbiased for  $\beta$ .
- The sampling distribution of  $\hat{\beta}$  is normal with the covariance matrix depending on the design matrix and  $\sigma^2$ . But we do not know  $\sigma^2$ !

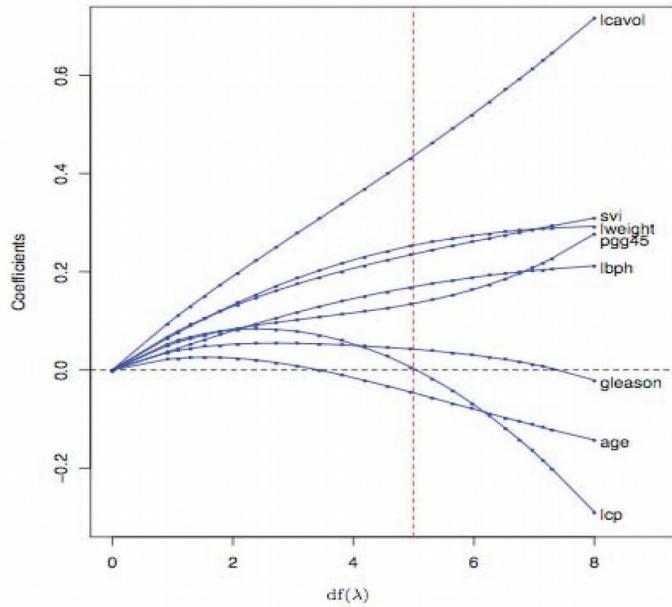
the regression problem

# Cross Validation: How to Split Data?

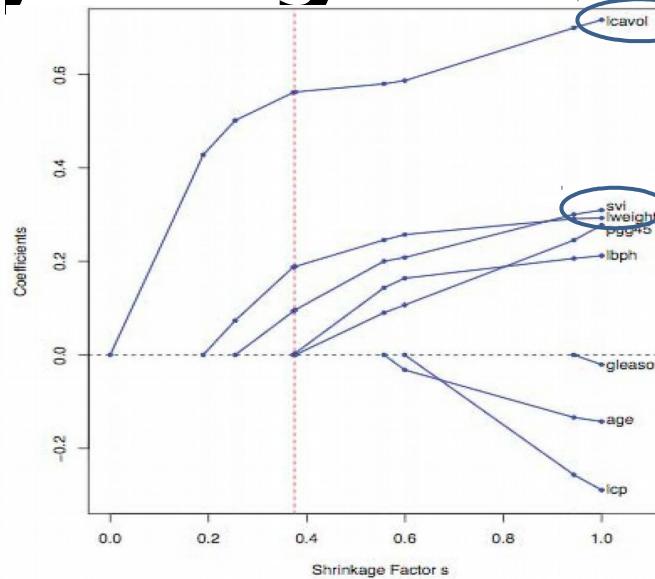
*K-fold cross-validation (KCV)*

- Randomly divide the data into K chunks of approximately equal size.
- For  $k = 1$  to  $K$ ,
  - The training data consist of data without the  $k$ -th fold of data and the testing data consist of the  $k$ -th fold;
  - Compute mean squared error or classification error rate for the  $k$ -th fold testing data.
  - Compute overall error given  $\lambda$
- For a range of  $\lambda$  penalty values, e.g. compute the overall error (e.g. MSE or classification error)
- Select  $\lambda$  penalty providing minimum overall error

# Lasso vs Ridge Regression



**FIGURE 3.8.** Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter  $\lambda$  is varied. Coefficients are plotted versus  $df(\lambda)$ , the effective degrees of freedom. A vertical line is drawn at  $df = 5.0$ , the value chosen by cross-validation.



**FIGURE 3.10.** Profiles of lasso coefficients, as the tuning parameter  $t$  is varied. Coefficients are plotted versus  $s = t / \sum_1^p |\hat{\beta}_j|$ . A vertical line is drawn at  $s = 0.36$ , the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.

Acknowledgement: From Hastie, T., Tibshirani, R., Friedman, J. (2001), *The Elements of Statistical Learning*, Springer Series in Statistics.

# LASSO: Limitations

- $p > n$ : the Lasso can only select up to  $n$  variables;
- $n > p$ : if there exist high correlations among predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression.
- If there is a group of variables with high correlation, the Lasso tends to select only one variable from the group

# Elastic Net

- Minimize

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j|$$

- $L_1$  penalty generates a sparse model.
- $L_2$  penalty
  - Removes the limitation on the number of selected variables;
  - Encourages group effect;
  - Stabilizes the  $L_1$  regularization path

Reference: Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B* 67.2 (2005): 301-320.

# Regression Analysis

## Model Selection

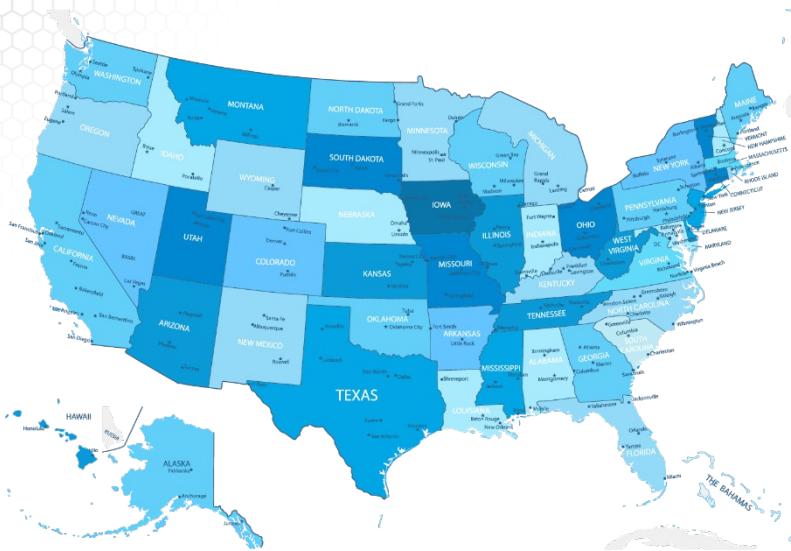
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Regularized Regression: Data  
Examples

# Ranking States by SAT Performance



SAT Mean Score by State - Year 1982

790 (South Carolina) -1088  
(Iowa)

*Which variables are associated with state average SAT scores?  
After accounting for selection biases, how do the states rank?  
Which states perform best for the amount of money they spend?*

# Ridge Regression

*library(MASS)*

```
## Scale the predicting variables and the response variable
ltakers = log(takers)
predictors = cbind(ltakers, income, years, public, expend, ranks)
predictors = scale(predictors)
sat.scaled = scale(sat)

## Apply ridge regression for a range of penalty constants
lambda = seq(0, 10, by=0.25)
out = lm.ridge(sat.scaled~predictors, lambda = lambda)
round(out$GCV, 5)
which(out$GCV == min(out$GCV))
```

2.25

10

```
round(out$coef[,10],4)
```

	predictors	ltakers	predictors	rank	predictors	income	predictors	years
	-0.4771	0.4195	0.0223				0.1796	
	predictors	public	predictors	expend				
	-0.0028		0.1808					

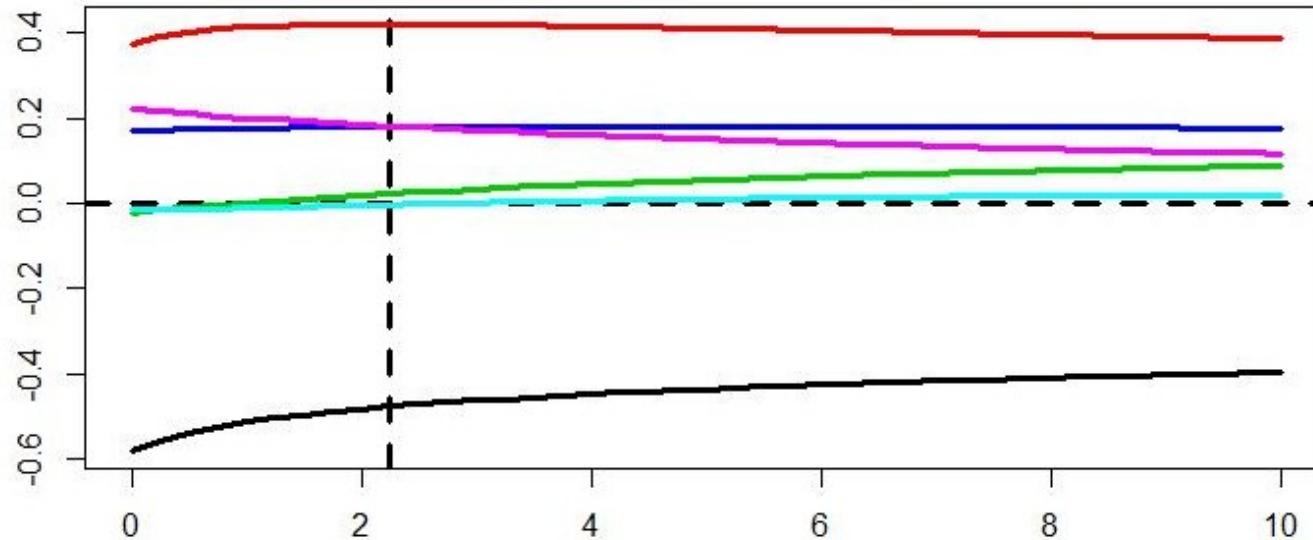


The ridge regression outputs estimates for each lambda in the considered range

The lambda is selected to minimize the (generalized) CV score.

# Ridge Regression

Plot of Regression Coefficients vs. Lambda Penalty Ridge Regression



ession outputs  
ach lambda in the  
ge  
selected to minimize  
d) CV score.

	predictors	itakers	rank	income	years
	predictors	public	expend		
	-0.4771	0.4195	0.0223	0.1796	
	-0.0028	0.1808			

# Lasso Regression

```
library(lars)
object = lars(x = predictors, y = sat.scaled)
object
```

Sequence of LASSO moves:

Itakers rank years expend income public

Var	1	2	4	6	3	5
Step	1	2	3	4	5	6

```
round(object$Cp,2)
```

0	1	2	3	4	5
6					

349.91 103.40 46.89 35.64 3.10 5.09

7.00

```
plot.lars(object)
```

```
plot.lars(object, xvar="df", plottype="Cp")
```



The selected model according to Malow's Cp is at the fourth variable introduced in the model.

# Lasso Regression

```
library(lars)
```

```
object = lars(x = predictors, y = sat.scaled)
```

```
object
```

Sequence of LASSO moves:

Itakers rank years expend income public

Var	1	2	4	6	3	5
Step	1	2	3	4	5	6

```
round(object$Cp)
```

0 1

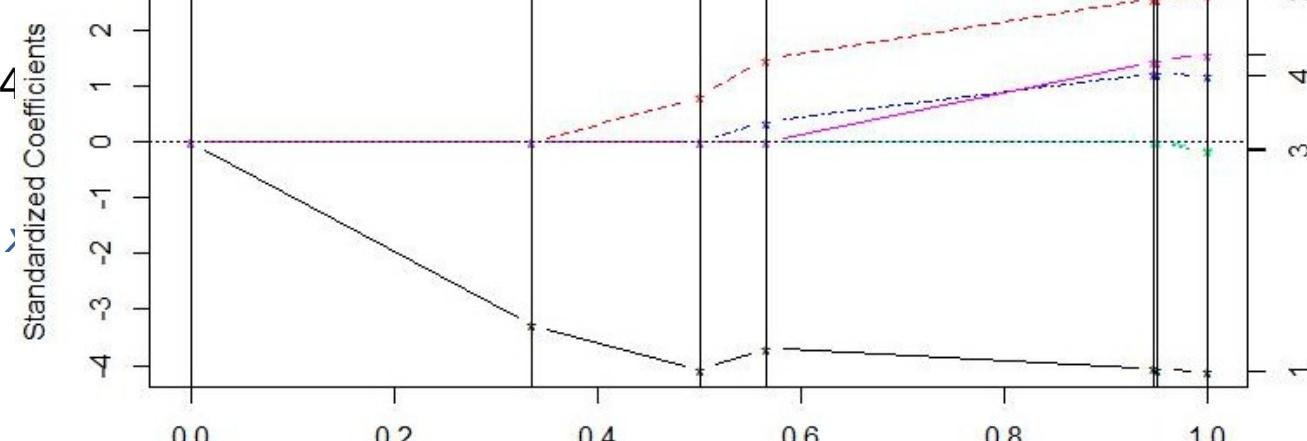
6

349.91 103.40

7.00

```
plot.lars(object)
```

```
plot.lars(object, )
```



The selected model according to Malow's Cp is at the fourth variable introduced in the model.

# Lasso Regression

```
library(lars)
```

```
object = lars(x = predictors, y = sat.scaled)
```

```
object
```

Sequence of LASSO moves:

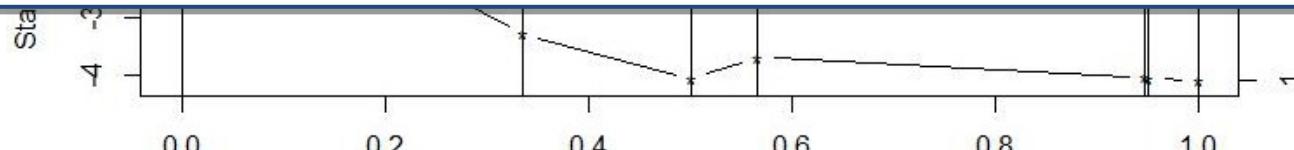
Itakers rank years expend income public

Var	1	2	4	6	3	5
Step	1	2	3	4	5	6



The selected model according to Malow's Cp is at the fourth variable introduced in the model.

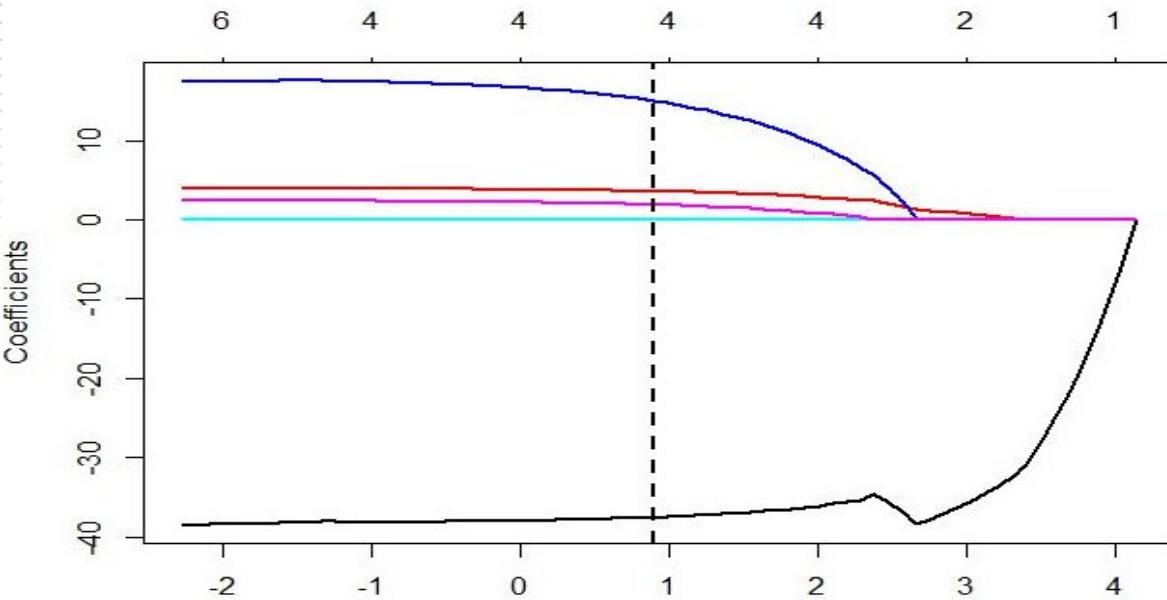
- The order of selected predictors: log(takers), rank, years, expend, income and public -- the first four are selected
- After Lasso variable selection, apply ordinary least squares with the selected predicting variables.



# Lasso & Elastic Net

```
library(glmnet)
Xpred= cbind(ltakers, rank, income, years, public,
expend)
# Find the optimal lambda using 10-fold CV
satmodel.cv=cv.glmnet(Xpred,
sat, alpha=1, nfolds=10)
## Fit lasso model with 100 values for lambda
satmodel = glmnet(Xpred, sat, alpha = 1, nlambda =
100)
## Extract coefficients at optimal lambda
coef(satmodel, s=satmodel.cv$lambda.min)
(Intercept) 478.328624
ltakers     -37.572757
rank        3.587894
income       .
years        15.028032
public       .
expend      1.899913
## Plot coefficient path
```

# Lasso & Elastic Net



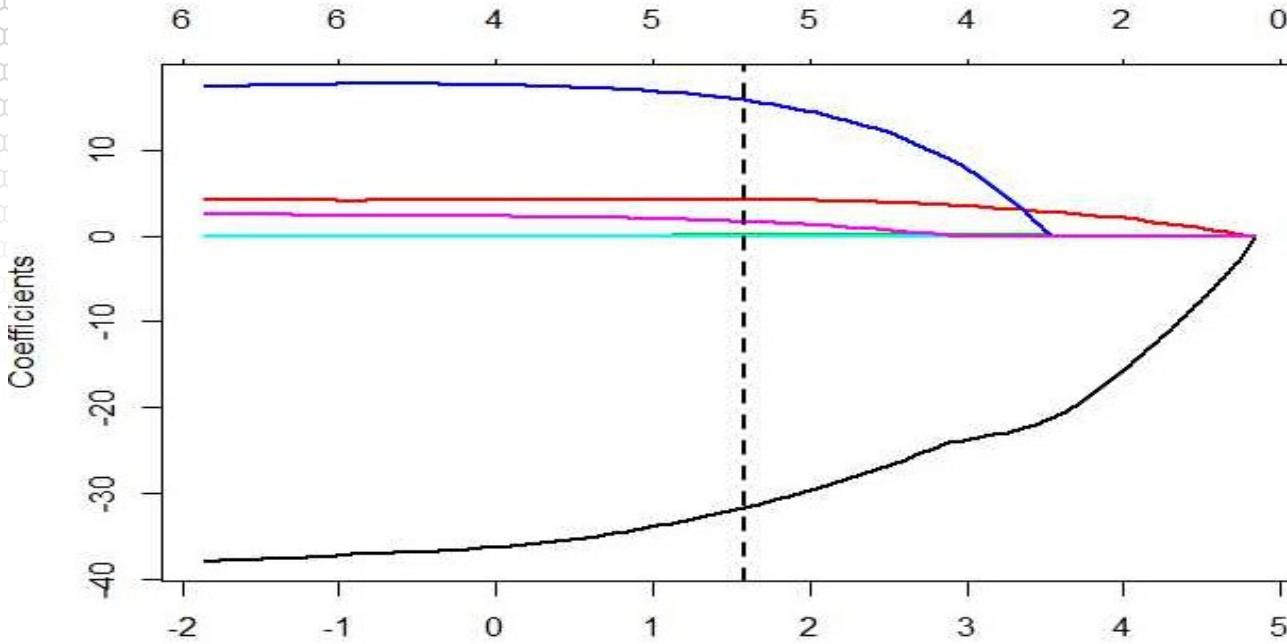
Selected predictors:  $\log(\text{takers})$ , rank, years & expend  
using Lasso & penalty selected using 10-fold CV

public expend 1.899913

# Elastic Net

```
library(glmnet) ## alpha = 1 lasso, alpha=0 ridge
Xpred= cbind(ltakers, rank, income, years, public, expend)
# Find the optimal lambda using 10-fold CV
satmodel.cv=cv.glmnet(Xpred,sat, alpha=0.5,nfolds=10)
## Fit lasso model with 100 values for lambda
satmodel = glmnet(Xpred, sat, alpha = 0.5, nlambda =
100)
## Plot coefficient paths
coef(satmodel,s=satmodel.cv$lambda.min)
ltakers      -31.62400226
rank         4.22409311
income        0.02588644
years        15.81282685
public        .
expend       1.65644751
## Extract coefficients at optimal lambda
plot(satmodel,xvar="lambda", lwd=2)
abline(v=log(satmodel.cv$lambda.min),col='black',lty =
2, lwd=2)
```

# Elastic Net



Selected predictors: Takers, rank, income, years &  
Expend using Elastic Net & penalty selected using 10-fold CV

# Overview of All Selection Approaches

	Log(Takers)	Rank	Income	Years	Public	Expended
Best subset & Mallow's Cp		✗		✗	✗	✗
Stepwise & AIC	✗	✗		✗		✗
Lasso & Mallow's Cp	✗	✗		✗		✗
Lasso & 10-fold CV	✗	✗		✗		✗
Elastic Net & 10-fold	✗	✗	✗	✗		✗

# Overview of All Selection Approaches

	Log(Taker s)	Rank	Income	Years	Public	Expen d
Best subset & Mallow's		✗		✗	✗	✗

- Rank, Years & Expend are selected by all approaches
- Takers is not selected by best subset only
- Income is not selected by any approach

10-fold CV       $\downarrow$        $\downarrow$        $\downarrow$        $\downarrow$        $\downarrow$

Elastic Net  
& 10-fold       $\times$        $\times$        $\times$        $\times$        $\times$

# Predicting Bankruptcy

- Effective bankruptcy prediction is useful for investors and analysts, allowing for accurate evaluation of a firm's prospects.
- Roughly forty years ago Ed Altman showed that publicly available financial indicators can be used to distinguish between firms that are about to go bankrupt and those that are not.

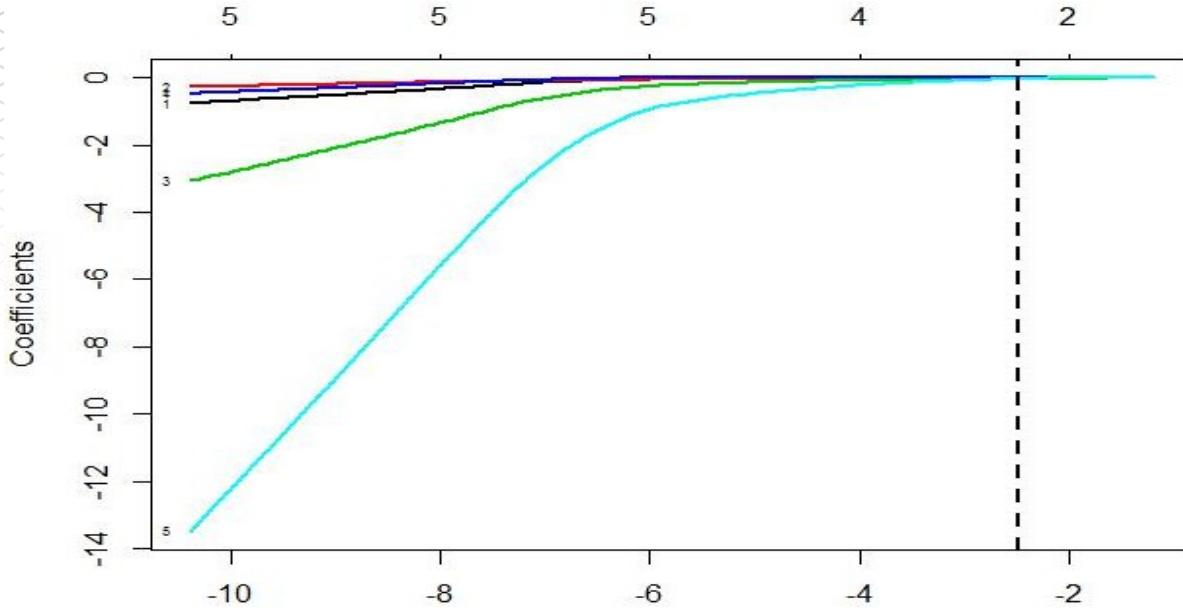
**Which financial indicators are associated with bankruptcy for telecommunications firms?**

*Acknowledgement: This example was provided by Dr. Jeffrey Simonoff from New York University*

# Lasso Regression

```
library(glmnet)
X = cbind(WC.TA,RE.TA,EBIT.TA,S.TA,BVE.BVL)
# 10-fold CV to find the optimal lambda
bank5.cv=cv.glmnet(X,Bankrupt,family=c("binomial"))
# 10-fold CV to find the optimal lambda
bank5=cv.glmnet(X,Bankrupt,family=c("binomial"),alpha=1,type="class",
nfold=10)
## Fit lasso model with 100 values for lambda
bank5 = glmnet(X, Y, family=c("binomial"), alpha = 1, nlambda = 100)
## Extract coefficients at optimal lambda
coef(bank5,s=bank5.cv$lambda.min)
(Intercept) -0.68410870
WC.TA
RE.TA -0.01323255
EBIT.TA -0.03136747
S.TA
BVE.BVL -0.03409259
## Plot coefficient paths
plot(bank5,xvar="lambda",lwd=2)
abline(v=log(bank5.cv$lambda.min),col='black',lty = 2,lwd=2)
```

# Lasso Regression



```
plot(log(bank5.cv$lambda), log(bank5.cv$lambda),  
      xlab="lambda", ylab="log(lambda)", log="xy")  
abline(v=log(bank5.cv$lambda.min), col='black', lty = 2, lwd=2)  
abline(h=0, col='black', lty = 1, lwd=2)  
text(5, 5, "5", col="black", font=2)  
text(5, 4, "5", col="black", font=2)  
text(5, 3, "5", col="black", font=2)  
text(4, 2, "4", col="black", font=2)  
text(2, 2, "2", col="black", font=2)  
text(0, 10, "0", col="black", font=2)  
text(0, -14, "-14", col="black", font=2)  
text(0, -12, "-12", col="black", font=2)  
text(0, -10, "-10", col="black", font=2)  
text(0, -8, "-8", col="black", font=2)  
text(0, -6, "-6", col="black", font=2)  
text(0, -4, "-4", col="black", font=2)  
text(0, -2, "-2", col="black", font=2)  
text(0, 0, "0", col="black", font=2)  
text(0, -3, "-3", col="black", font=2)
```

Selected predictors: RE.TA, EBIT.TA, BVE.BVL  
according to Lasso & penalty selected using 10-fold CV

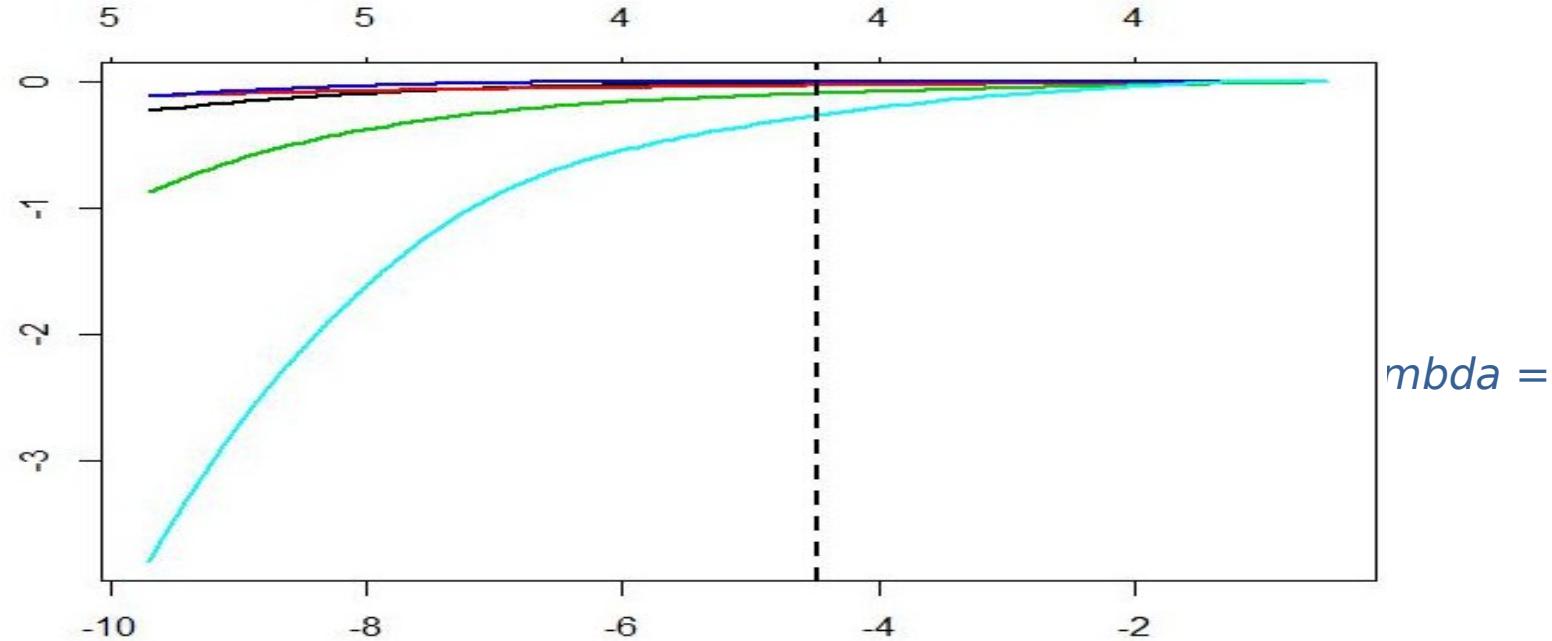
```
abline(v=log(bank5.cv$lambda.min), col='black', lty = 2, lwd=2)
```

# Elastic Net

```
library(glmnet)
## alpha = 1 lasso, alpha=0 ridge
X = cbind(WC.TA,RE.TA,EBIT.TA,S.TA,BVE.BVL)
# 10-fold CV to find the optimal lambda
bank6.cv=cv.glmnet(X,Bankrupt,family=c("binomial"),alpha=
0.5,type="class",nfolds=10)
## Fit lasso model with 100 values for lambda
bank6 = glmnet(X, Bankrupt,family=c("binomial"), alpha = 0.5, nlambda =
100)
## Extract coefficients at optimal lambda
coef(bank6,s=bank6.cv$lambda.min)
(Intercept) -0.57662609
WC.TA      -0.01115371
RE.TA      -0.02826461
EBIT.TA     -0.09143845
S.TA
BVE.BVL    -0.26434840
## Plot coefficient paths
plot(bank6, xvar = "lambda", lwd = 2)
```

# Elastic Net

Coefficients



$mbda =$

Selected predictors: WC.TA, RE.TA, EBIT.TA and , BVE.BVL  
using Elastic Net & penalty selected using 10-fold CV

```
## Plot coefficient paths  
plot(bankE, type = "lambda", lwd = 2)
```

# Overview of All Selection Approaches

	WC.TA	RE.TA	EBIT.TA	S.TA	BVE.BV L
Best subset & Mallow's Cp	✗	✗	✗		✗
Stepwise & AIC		✗	✗		✗
Lasso & 10-fold CV		✗	✗		✗
Elastic Net & 10-fold CV	✗	✗	✗		✗

# Regression Analysis

Regression Analysis in Practice

**Nicoleta Serban, Ph.D.**

*Associate Professor*

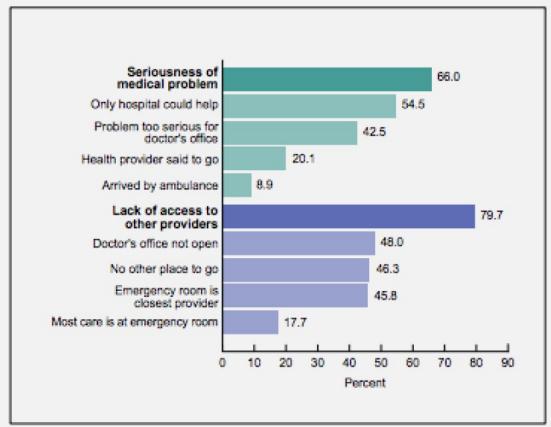
Stewart School of Industrial and Systems Engineering

Emergency Department  
Healthcare Costs

# Emergency Department Healthcare Costs



Figure 1. Percentage who had selected reasons for last emergency room visit, among adults aged 18–64 whose last visit in past 12 months did not result in hospital admission: United States, January–June 2011



## Research Question 1:

What factors impact the healthcare cost due to emergency department encounters?

## Research Question 2:

Is access to primary care providers associated to healthcare costs due to emergency department encounters?

# Emergency Department Healthcare Costs

**Study population:** Medicaid-enrolled adults in four southeast states: Alabama, Arkansas, Louisiana, and North Carolina in 2011

- Medicaid is a health insurance program for the low-income population

**Data Source:** The Medicaid Analytic eXtract (MAX) claims files available from the Centers of Medicare and Medicaid Services (CMS)

- Disclaimer: The research on healthcare cost for the Medicaid population using the MAX claims data has been approved by the Georgia Tech Internal Review Board and by CMS; Do NOT use the data provided for this analysis for other purposes beyond the study in this lecture.
- Additional data sources: US Bureau Census, Health Analytics Group at GT, Robert Wood Johnson Foundation among others.

# Response & Predicting Variables

## Response variable:

- Emergency Department cost aggregated at the census tract level (*EDcost*)
- Number of member months aggregated at the census tract level (*PMPM*)

## Predicting variables are:

**Location:** state (*State*) and census tract identification (*GEOID*)

**Utilization:** three predicting variables measuring the number of claims for the Emergency Department (*ED*), of hospitalizations (*HO*) and physician office (*PO*)

**Population characteristics:** percentages of Medicaid-enrolled adults who are black (*BlackPop*), white (*WhitePop*) or other race/ethnicity (*OtherPop*); percentages of Medicaid-enrolled adults who are health (*HealthyPop*), with chronic conditions (*ChronicPop*) or with complex health problems (*ComplexPop*)

# Controlling Variables

## Selection Bias:

- The utilization of healthcare emergency services is directly driven by the health status of the population utilizing the system. Adults with multiple chronic conditions and/or with complex health problems tend to need emergency healthcare services more than the healthy population.
- Controlling factors: Percentage of population with chronic conditions (*ChronicPop*) or with complex health problems (*ComplexPop*)

## Confounding Variable:

- The number of ED claims is a confounding variable not an explanatory factor for ED cost because it is a measure of utilization of the emergency department which leads to ED healthcare cost; it correlates with both the response and the predicting variables.
- Such confounding variables should not be included in the model.

# Regression Analysis

## Regression Analysis in Practice

**Nicoleta Serban, Ph.D.**

*Associate Professor*

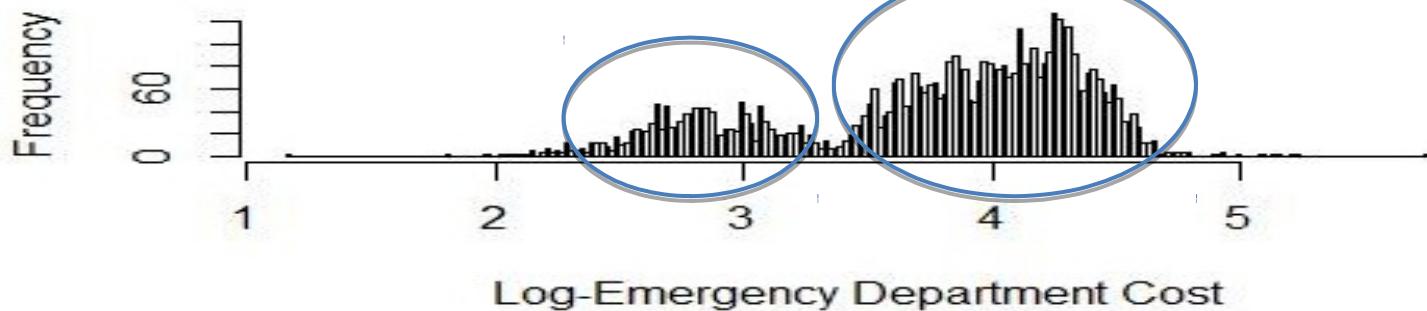
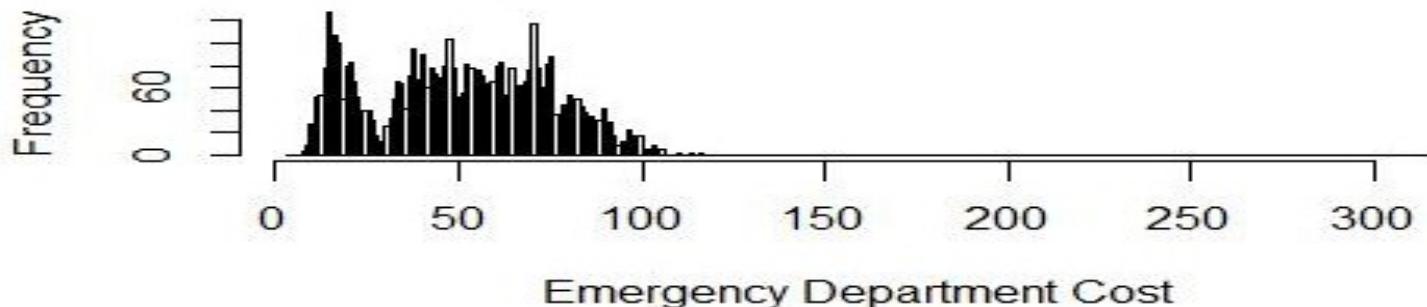
Stewart School of Industrial and Systems Engineering

Exploratory Data Analysis

# Exploratory Data Analysis: Response Variable

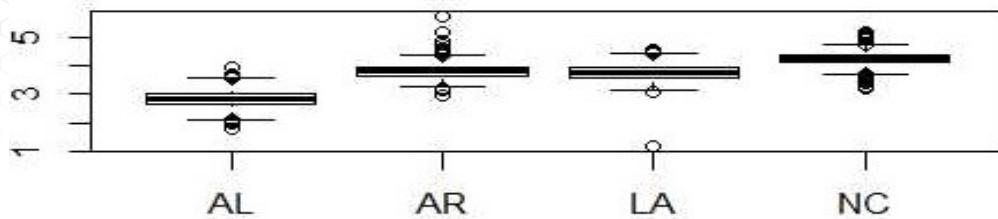
```
## Read the data using the 'read.csv()' R command
dataAdult=read.csv("DataADULT.csv",header=TRUE)
attach(dataAdult)
## Outcome/Response Variable
EDCost.pppm = EDCost/PMPM
## Rescale utilization
dataAdult$PO = PO/PMPM
dataAdult$HO = HO/PMPM
#Histogram of the response variable
par(mfrow=c(2,1))
hist(EDCost.pppm,breaks=300, xlab="Emergency Department Cost", main="")
hist(log(EDCost.pppm),breaks=300, xlab="Log-Emergency Department Cost",
main="")
log.EDCost.pppm = log(EDCost.pppm)
```

# Response Variable



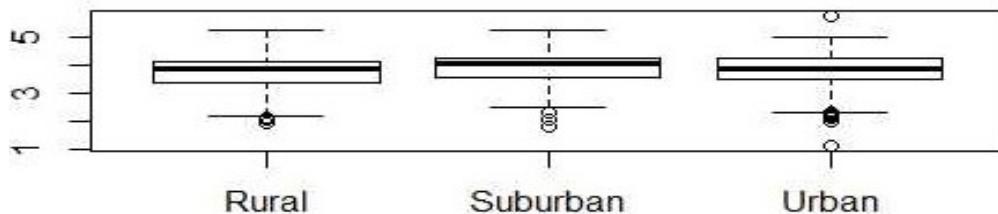
# Exploratory Data Analysis: Response vs Qualitative Predictors

**Variation of log of ED costs by state**



*variation of log of ED costs by*

**Variation of log of ED costs by urbanicity**

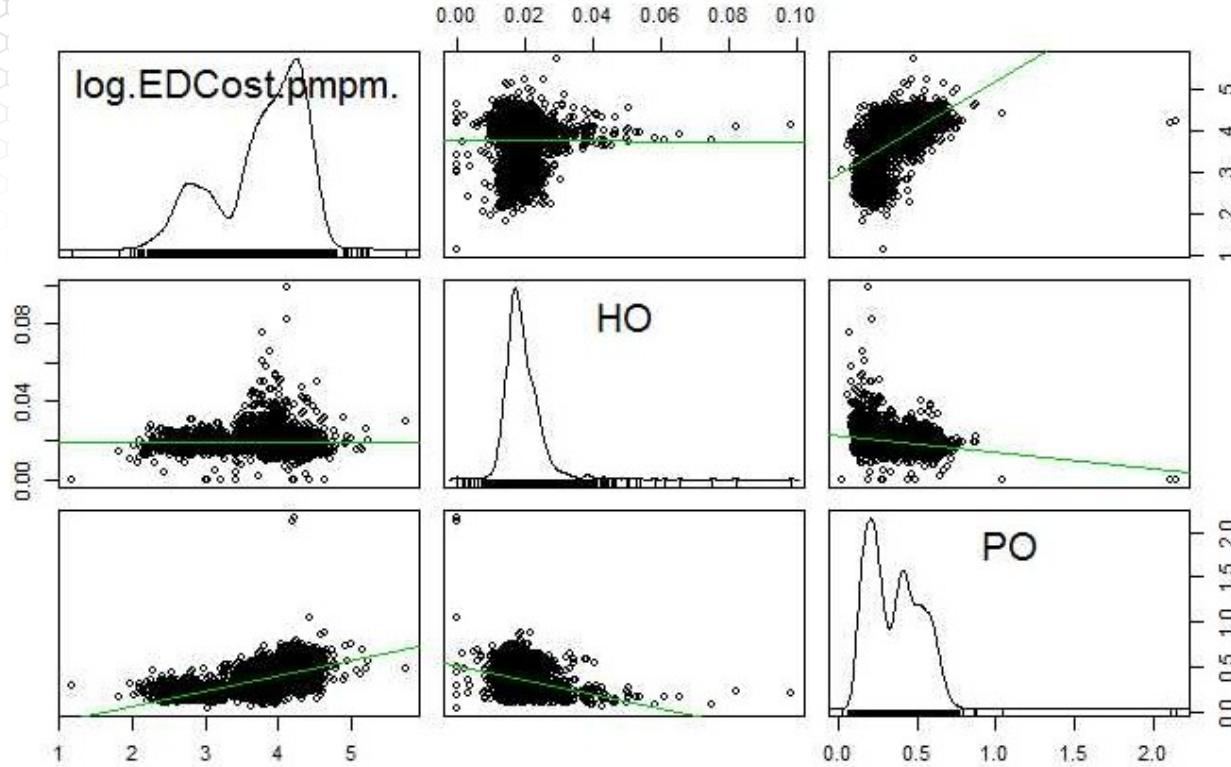


# Exploratory Data Analysis: Response vs Qualitative Predictors

```
## Scatterplot matrix plots
library(car)
## Response vs Utilization
scatterplotMatrix(~log(EDCost.pppm)+HO+PO,smooth=FALSE)
## Response vs Population Characteristics
scatterplotMatrix(~log(EDCost.pppm)
+WhitePop+BlackPop+OtherPop+HealthyPop+
ChronicPop+ComplexPop,smooth=FALSE)
## Response vs Social and Economic Environment Characteristics
scatterplotMatrix(~log(EDCost.pppm)
+Unemployment+Income+Poverty+Education+
Accessibility+Availability+ProvDensity,smooth=FALSE)
## Response vs County Health Rankings
scatterplotMatrix(~log(EDCost.pppm)
+RankingsPCP+RankingsFood+RankingsHousing+
```

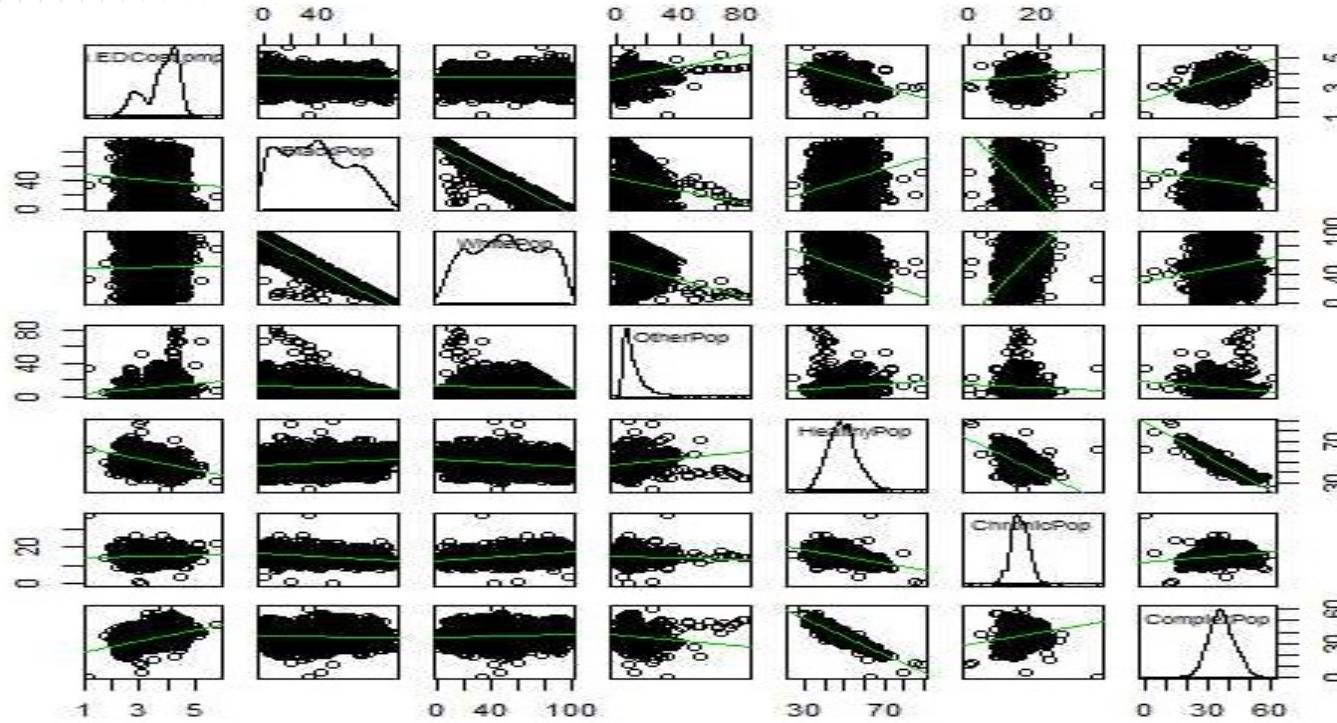
# Response vs Quantitative Predictors

**ED Cost vs Utilization Measures: number of claims for HO and PO**



# Response vs Quantitative Predictors

**ED Cost vs Population Characteristics:** BlackPop, WhitePop, OtherPop, HealthyPop, ChronicPop, ComplexPop

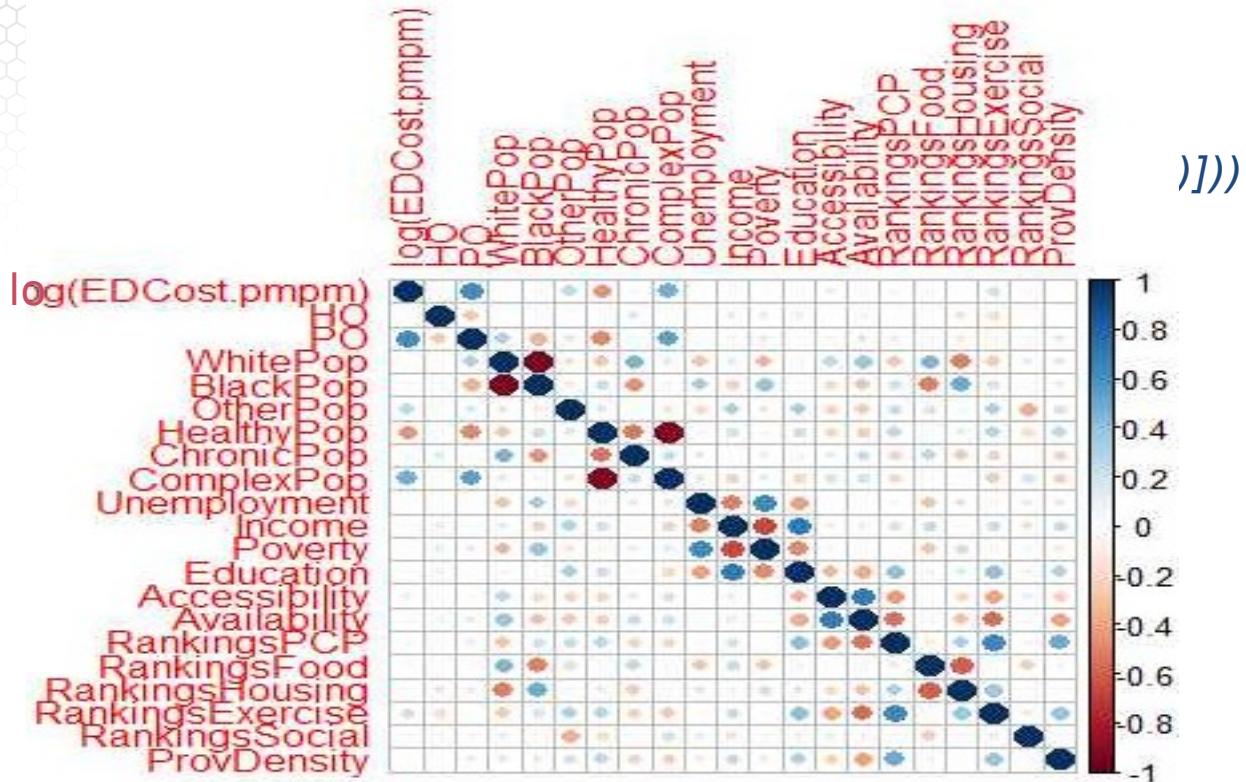


# Response vs. Predicting Variables: Correlation Matrix Plot

```
## Correlation matrix plot
library(corrplot)
corr = cor(cbind(EDCost.pppm,dataAdult[,-c(1,2,3,18)]))
corrplot(corr)
```

lo

# Response vs. Predicting Variables: Correlation Matrix Plot



# Regression Analysis

## Regression Analysis in Practice

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Multiple Regression: Fitted Model  
and Residual Analysis

# Multiple Linear Regression Model

```
## Exclude GEOID, scaling factor (PMPM); confounding factor (ED)
## Exclude OtherPop & ComplexPop because of linear dependence
dataAdult.red = dataAdult[,-c(1,3,4,5,10,13)]
fullmodel = lm(log(EDCost.pppm) ~ ., data = dataAdult.red)
summary(fullmodel)
```

# Multiple Linear Regression Model

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.208e+00	1.175e-01	18.788	< 2e-16	***
StateAR	9.235e-01	1.610e-02	57.353	< 2e-16	***
StateLA	9.081e-01	1.358e-02	66.853	< 2e-16	***
StateNC	1.418e+00	1.650e-02	85.909	< 2e-16	***
HO	1.168e+01	7.587e-01	15.401	< 2e-16	***
PO	1.378e-01	4.114e-02	3.350	0.000815	***
WhitePop	4.416e-03	5.800e-04	7.614	3.16e-14	***
BlackPop	4.894e-03	5.824e-04	8.403	< 2e-16	***
HealthyPop	-9.044e-04	8.160e-04	-1.108	0.267751	.
ChronicPop	-5.949e-03	2.052e-03	-2.899	0.003760	**
Unemployment	4.390e-04	7.377e-04	0.595	0.551797	.
Income	-2.556e-07	2.774e-07	-0.922	0.356769	.
Poverty	-3.306e-04	4.460e-04	-0.741	0.458529	.
Education	-1.447e-03	3.296e-04	-4.390	1.16e-05	***
UrbanicitySuburban	-4.565e-04	1.369e-02	-0.033	0.973406	.
UrbanicityUrban	2.067e-02	1.269e-02	1.629	0.103356	.
Accessibility	-1.965e-03	7.094e-04	-2.770	0.005623	**
Availability	8.037e-02	1.975e-02	4.068	4.81e-05	***
RankingsPCP	7.596e-04	1.819e-04	4.175	3.03e-05	***
RankingsFood	6.586e-03	5.203e-03	1.266	0.205642	.
RankingsHousing	-4.642e-03	1.562e-03	-2.973	0.002967	**
RankingsExercise	3.993e-04	2.332e-04	1.712	0.086907	.
RankingsSocial	-3.895e-04	1.347e-03	-0.289	0.772497	.
ProvDensity	6.042e-02	1.573e-02	3.841	0.000124	***
---					

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

Residual standard error: 0.2321 on 4995 degrees of freedom

Multiple R-squared: 0.8486, Adjusted R-squared: 0.8479

F-statistic: 1218 on 23 and 4995 DF, p-value: < 2.2e-16

ding factor (ED)

**Socio-economic** predicting variables (unemployment, median income, % below the poverty level and rankings with respect social environment) are **not** statistically significant given other predicting variables in the model.

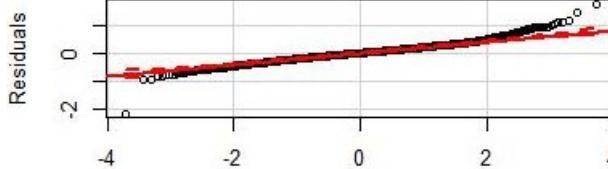
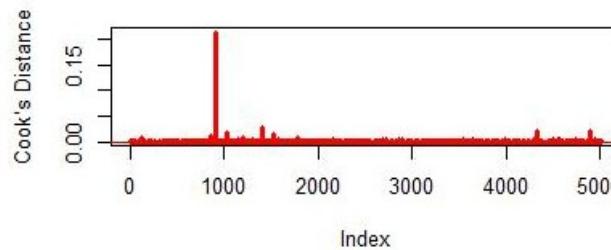
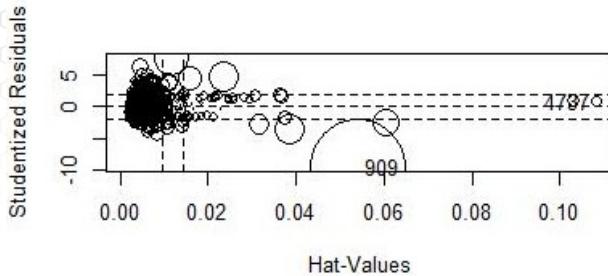
**Access** to primary care (accessibility and availability) is statistically significantly associated to ED cost.

85% of the variability in the ED cost is explained.

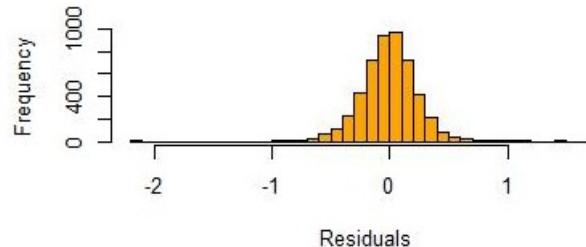
# Residual Analysis: Outliers & Normality

```
## Residuals versus individual predicting variables
full.resid = residuals(fullmodel)
cook = cooks.distance(fullmodel)
par(mfrow=c(2,2))
## Check outliers
influencePlot(fullmodel)
plot(cook,type="h",lwd=3,col="red", ylab = "Cook's
Distance")
## Check Normality
abline(0,0,col="red")
qqPlot(full.resid, ylab="Residuals", main = "")
hist(full.resid, xlab="Residuals", main =
"",nclass=30,col="orange")
```

# Residual Analysis: Outliers & Normality



...  
...,  
nclass=50, col="orange")



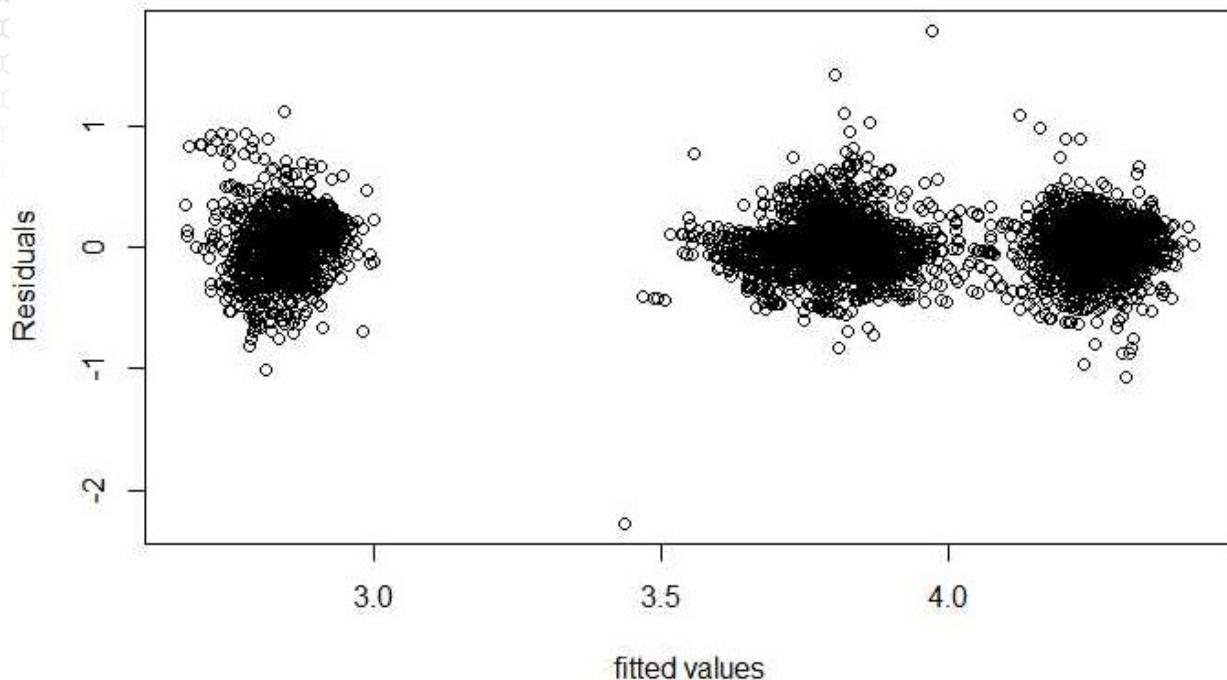
**Outliers:** Observation 909 stands out

**Normality:** Symmetric but heavy tails

# Residual Analysis: Constant Variance & Uncorrelated Errors

```
## Check Constant Variance & Uncorrelated Errors
full.fitted = fitted(fullmodel)
par(mfrow=c(1,1))
plot(full.fitted,full.resid, xlab="fitted values",
ylab="Residuals")
```

# Residual Analysis: Constant Variance & Uncorrelated Errors



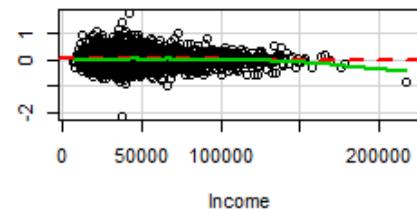
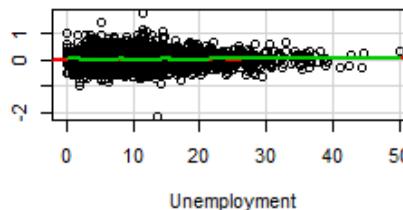
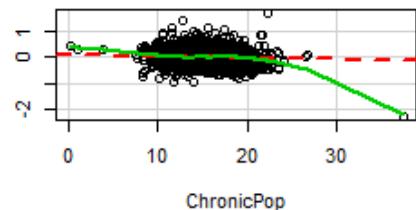
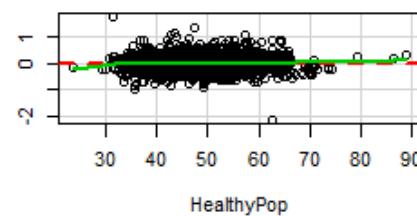
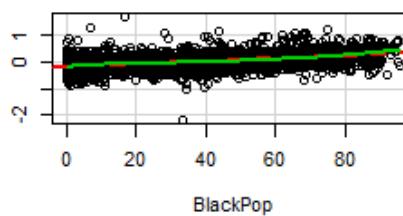
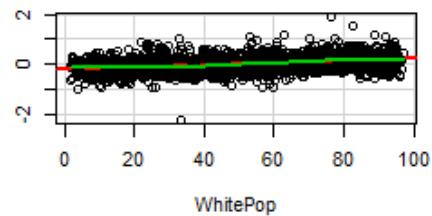
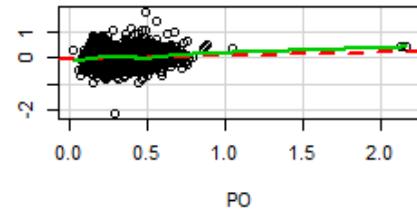
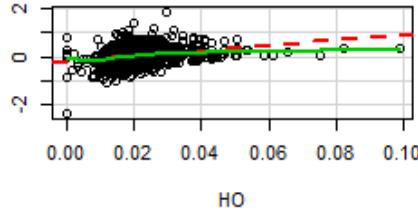
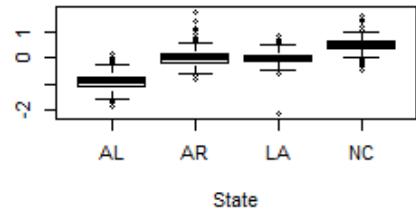
**Constant Variance:**  
No pattern

**Uncorrelated Errors:**  
Three well defined clusters

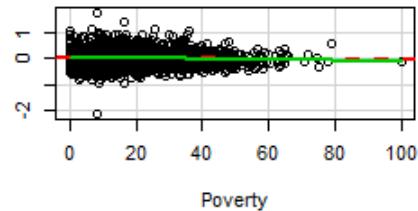
# Residual Analysis: Linearity

```
## Check Linearity  
crPlots(fullmodel, ylab="")
```

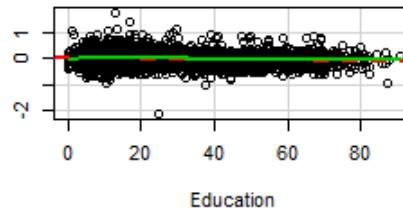
# Residual Analysis: Linearity



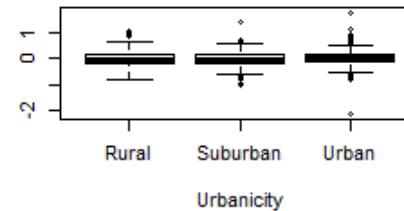
# Residual Analysis: Linearity (cont'd)



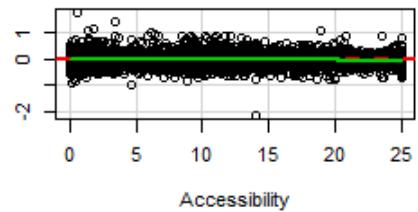
Poverty



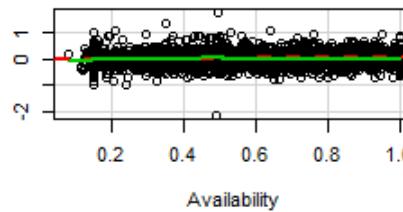
Education



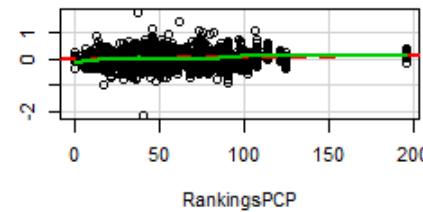
Urbanicity



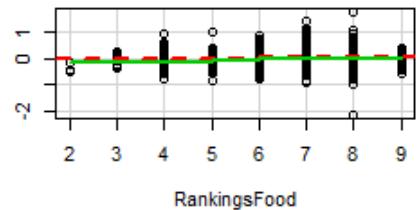
Accessibility



Availability



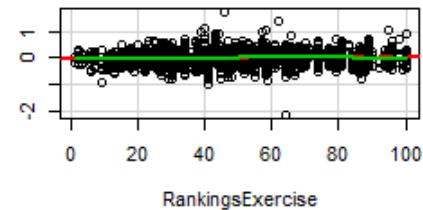
RankingsPCP



RankingsFood



RankingsHousing



RankingsExercise

# Regression Analysis

## Regression Analysis in Practice

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

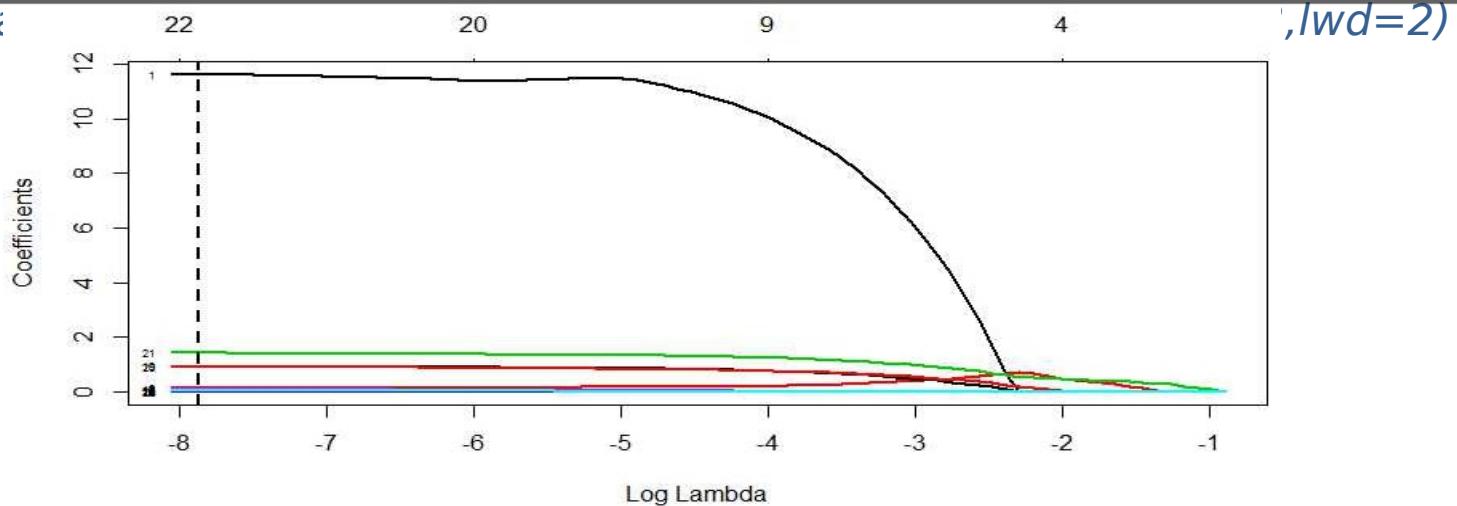
Variable Selection

# Lasso Regression

```
# 10-fold CV to find the optimal lambda
lassomodel.cv=cv.glmnet(predictors,log(EDCost.pmpm),alpha=1,nfolds=10)
## Fit lasso model with 100 values for lambda
lassomodel = glmnet(predictors,log(EDCost.pmpm), alpha = 1, nlambda =
100)
## Plot coefficient paths
plot(lassomodel,xvar="lambda",label=TRUE,lwd=2)
abline(v=log(lassomodel.cv$lambda.min),col='black',lty = 2,lwd=2)
```

# Lasso Regression

- Rankings Social and Suburban dummy variables are not selected according to Lasso & penalty selected using 10-fold CV or Mallow's Cp;
- High-coefficient path corresponds to HO variable;
- Other large-coefficient paths correspond to State dummy variables

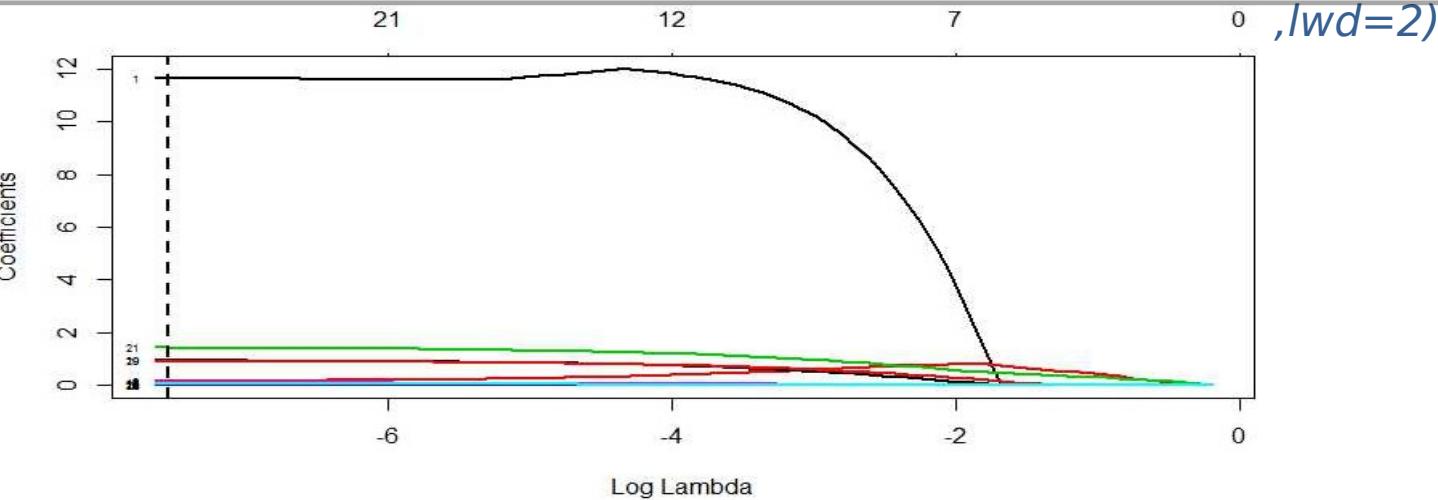


# Elastic Net Regression

```
# 10-fold CV to find the optimal lambda
enetmodel.cv=cv.glmnet(predictors,log(EDCost.pmpm),alpha=0.5,nfolds=10)
## Fit lasso model with 100 values for lambda
enetmodel = glmnet(predictors,log(EDCost.pmpm), alpha = 0.5, nlambd
100)
## Plot coefficient paths
plot(enetmodel,xvar="lambda",label=TRUE, lwd=2)
abline(v=log(enetmodel.cv$lambda.min),col='black',lty = 2,lwd=2)
## Extract coefficients at optimal lambda
coef(enetmodel,s=enetmodel.cv$lambda.min)
```

# Elastic Net Regression

- RankingsSocial is not selected according to Lasso & penalty selected using 10-fold CV or Mallow's Cp;
- High-coefficient path corresponds to HO variable;
- Other large-coefficient paths correspond to State dummy variables



# Stepwise Regression

*full* =

```
lm(log(EDCost.pppm)~HealthyPop+ChronicPop+State+Urbanicity+HO+PO+
  BlackPop+WhitePop+Unemployment+Income+Poverty+Education+
  Accessibility+Availability+ProvDensity+
  RankingsPCP+RankingsFood+RankingsExercise+RankingsSocial)
```

```
minimum = lm(log(EDCost.pppm)~HealthyPop+ChronicPop)
```

# Forward Stepwise Regression

```
forward.model = step(minimum, scope = list(lower=minimum, upper = full),
  direction = "forward")
```

```
summary(forward.model)
```

# Backward Stepwise Regression

```
backward.model = step(full, scope = list(lower=minimum, upper = full),
  direction = "backward")
```

```
summary(backward.model)
```

# Forward- Backward Stepwise Regression

```
both.min.model = step(minimum, scope = list(lower=minimum, upper = full),
  direction = "both")
```

```
summary(both.min.model)
```

# Stepwise Regression

*full* =

```
lm(log(EDCost.pppm) ~ HealthyPop + ChronicPop + State + Urbanicity + HO + PO +  
    BlackPop + WhitePop + Unemployment + Income + Poverty + Education +  
    Accessibility + Availability + ProvDensity +
```

- **Variables not selected by all methods: Unemployment, Income, Poverty, RankingExercise, RankingSocial**
- **State dummy variables followed by number of claims per-member per-month are first selected by forward stepwise regression**

# Backward Stepwise Regression

```
backward.model = step(full, scope = list(lower = minimum, upper = full),  
    direction = "backward")
```

```
summary(backward.model)
```

# Forward- Backward Stepwise Regression

```
both.min.model = step(minimum, scope = list(lower = minimum, upper = full),  
    direction = "both")
```

```
summary(both.min.model)
```

# Stepwise Regression Model

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	2.0271089	0.0995378	20.365	< 2e-16	***
HealthyPop	-0.0005092	0.0007837	-0.650	0.515917	
ChronicPop	-0.0051250	0.0020252	-2.531	0.011418	*
StateAR	0.9324593	0.0155667	59.901	< 2e-16	***
StateLA	0.9003846	0.0118631	75.898	< 2e-16	***
StateNC	1.4268425	0.0157605	90.533	< 2e-16	***
HO	12.0476486	0.7237072	16.647	< 2e-16	***
Education	-0.0016689	0.0002312	-7.218	6.08e-13	***
ProvDensity	0.0605923	0.0156154	3.880	0.000106	***
RankingsPCP	0.0007885	0.0001577	5.000	5.94e-07	***
Availability	0.0756249	0.0191618	3.947	8.03e-05	***
Accessibility	-0.0019930	0.0007001	-2.847	0.004433	**
PO	0.1232428	0.0406869	3.029	0.002466	**
UrbanicitySuburban	-0.0017746	0.0136754	-0.130	0.896758	
UrbanicityUrban	0.0226383	0.0124409	1.820	0.068870	.
BlackPop	0.0050790	0.0005596	9.076	< 2e-16	***
WhitePop	0.0046371	0.0005522	8.398	< 2e-16	***
RankingsFood	0.0158764	0.0040770	3.894	9.98e-05	***
---					
Signif. codes:	0	****	0.001	***	0.01 **
	0.05	**	0.1	**	1

Residual standard error: 0.2322 on 5001 degrees of freedom  
Multiple R-squared: 0.8483, Adjusted R-squared: 0.8478  
F-statistic: 1645 on 17 and 5001 DF, p-value: < 2.2e-16

**Level of urbanicity of the community is not statistically significant at  $\alpha = 0.05$**

**Access to primary care (accessibility and availability) is statistically significantly associated to ED cost.**

**85% of the variability in the ED cost is explained.**

# Stepwise Regression Vs Full Models

## Compare full model to selected model

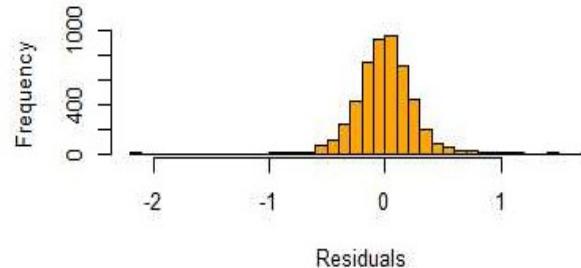
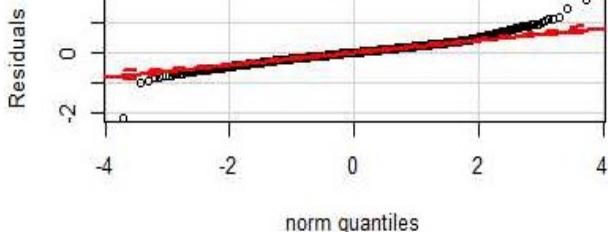
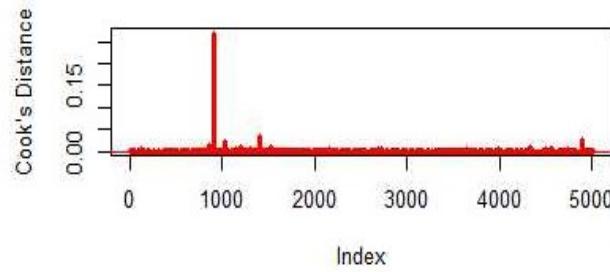
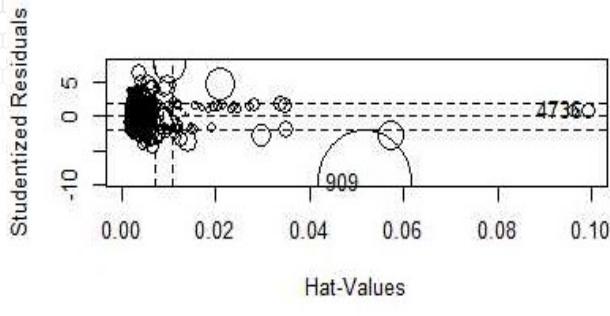
```
reg.step = lm(log(EDCost.pppm) ~ HealthyPop + ChronicPop + State + HO +  
    Education + ProvDensity + RankingsPCP + Accessibility + Availability +  
    PO + Urbanicity + BlackPop + WhitePop + RankingsFood)
```

```
anova(reg.step, full)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	5001	269.56				
2	4996	269.46	5	0.10406	0.3859	0.8588

- **P-value large  $\Rightarrow$  Do not reject the null hypothesis (reduced model)**
- **The reduced model is plausibly as good in terms of explanatory power as the full model.**

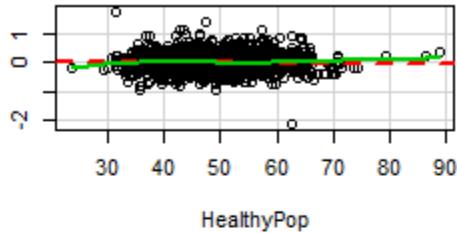
# Residual Analysis: Outliers & Normality



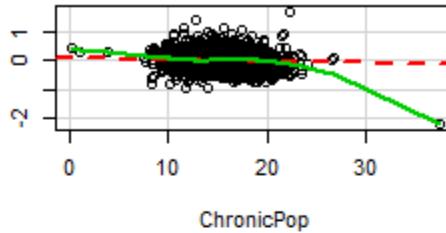
**Outliers:** Observation 909 stands out

**Normality:** Symmetric but heavy tails

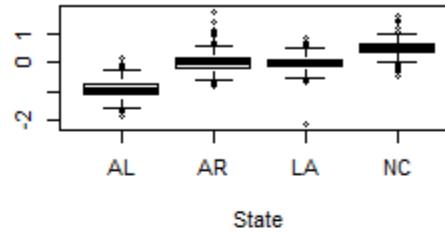
# Residual Analysis: Linearity



HealthyPop

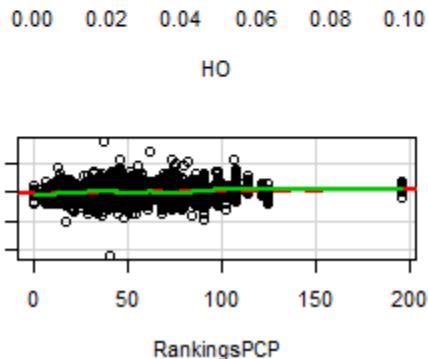


ChronicPop

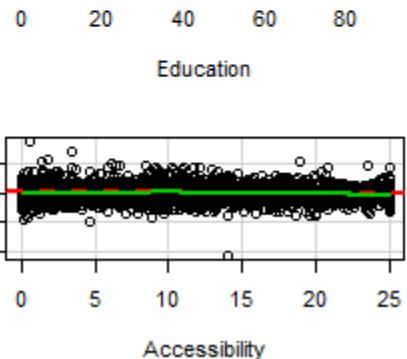


State

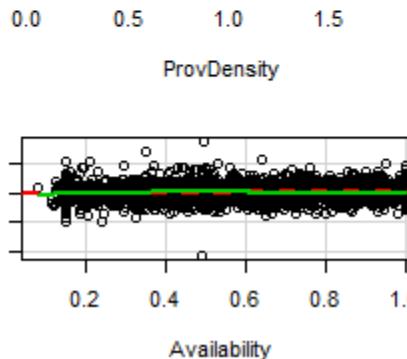
**Linearity:** Some nonlinearity with respect to HO, BlackPop & WhitePop  
**Transformations:** Not an improvement in the fit



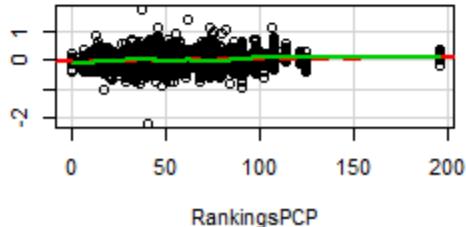
HO



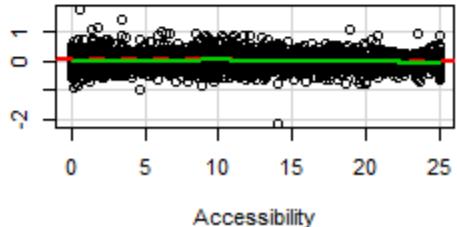
Education



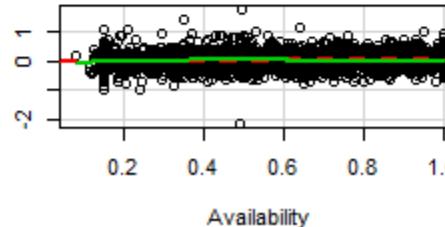
ProvDensity



RankingsPCP



Accessibility



Availability

# Removing Outlier

## With Outlier

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.0271089	0.0995378	20.365	< 2e-16 ***
HealthyPop	-0.0005092	0.0007837	-0.650	0.515917
ChronicPop	-0.0051250	0.0020252	-2.531	0.011418 *
StateAR	0.9324593	0.0155667	59.901	< 2e-16 ***
StateLA	0.9003846	0.0118631	75.898	< 2e-16 ***
StateNC	1.4268425	0.0157605	90.533	< 2e-16 ***
HO	12.0476486	0.7237072	16.647	< 2e-16 ***
Education	-0.0016689	0.0002312	-7.218	6.08e-13 ***
ProvDensity	0.0605923	0.0156154	3.880	0.000106 ***
RankingsPCP	0.0007885	0.0001577	5.000	5.94e-07 ***
Availability	0.0756249	0.0191618	3.947	8.03e-05 ***
Accessibility	-0.0019930	0.0007001	-2.847	0.004433 **
PO	0.1232428	0.0406869	3.029	0.002466 **
UrbanicitySuburban	-0.0017746	0.0136754	-0.130	0.896758
UrbanicityUrban	0.0226383	0.0124409	1.820	0.068870 .
BlackPop	0.0050790	0.0005596	9.076	< 2e-16 ***
WhitePop	0.0046371	0.0005522	8.398	< 2e-16 ***
RankingsFood	0.0158764	0.0040770	3.894	9.98e-05 ***
---				

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

Residual standard error: 0.2322 on 5001 degrees of freedom  
 Multiple R-squared: **0.8483**, Adjusted R-squared: 0.8478  
 F-statistic: 1645 on 17 and 5001 DF, p-value: < 2.2e-16

## Without Outlier

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.9356344	0.0991296	19.526	< 2e-16 ***
HealthyPop	0.0003798	0.0007824	0.485	0.627430
ChronicPop	-0.0010849	0.0020519	-0.529	0.597031
StateAR	0.9379139	0.0154403	60.745	< 2e-16 ***
StateLA	0.8989533	0.0117596	76.444	< 2e-16 ***
StateNC	1.4282364	0.0156224	91.422	< 2e-16 ***
HO	11.5397384	0.7193214	16.043	< 2e-16 ***
Education	-0.0017147	0.0002292	-7.480	8.72e-14 ***
ProvDensity	0.0654339	0.0154862	4.225	2.43e-05 ***
RankingsPCP	0.0007560	0.0001564	4.835	1.37e-06 ***
Accessibility	-0.0018658	0.0006940	-2.688	0.007205 **
Availability	0.0755848	0.0189930	3.980	7.00e-05 ***
PO	0.1338608	0.0403440	3.318	0.000913 ***
UrbanicitySuburban	-0.0006647	0.0135555	-0.049	0.960895
UrbanicityUrban	0.0222961	0.0123314	1.808	0.070654 .
BlackPop	0.0050502	0.0005547	9.105	< 2e-16 ***
WhitePop	0.0044178	0.0005478	8.064	9.14e-16 ***
RankingsFood	0.0162198	0.0040412	4.014	6.07e-05 ***
---				

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

Residual standard error: 0.2301 on 5000 degrees of freedom  
 Multiple R-squared: **0.8504**, Adjusted R-squared: 0.8499  
 F-statistic: 1672 on 17 and 5000 DF, p-value: < 2.2e-16

# Model Interpretation: Access to Care

## Access to primary care:

- Availability – proxy measure of wait times for appointment, measured in level of congestion and takes values between 0 and 1 (the higher the value, the worse the wait time)
- Accessibility – travel distance to primary care providers, measured in miles

## Interpretation:

- An increase of 1% in lack of availability of primary care providers results in 0.0755 unit increase in  $\log(\text{ED cost PMPM})$  given all other predictors fixed
- A reduction of 1 mile in travel distance to primary care providers results in 0.001 unit increase in  $\log(\text{ED cost PMPM})$  given all other predictors fixed
- The correlation between the two measures is 0.695: If Availability is discarded from the model, Accessibility is not statistically significant.

# Model Interpretation: State Differences

## **Location: Comparing ED Costs for AL, AR, LA and NC in 2011**

- ED cost PMPM is  $\exp(0.938)$  higher in AR versus AL, or ED cost is \$30.65 per-member per-year higher in AR versus AL controlling for utilization, access and socio-economics;
- ED cost PMPM is  $\exp(0.899)$  higher in LA versus AL, or ED cost is \$29.48 per-member per-year higher in LA versus AL controlling for utilization, access and socio-economics;
- ED cost PMPM is  $\exp(1.428)$  higher in NC versus AL, or ED cost is \$50.04 per-member per-year higher in NC versus AL controlling for utilization, access and socio-economics;

**Overall interpretation:** Controlling for many potential factors contributing to ED cost, North Carolina pays significantly more while Alabama pays significantly less per-member than other states on emergency care.

# Model Interpretation: Utilization

## Healthcare Utilization:

- PO – number of claims reimbursed for care in the physician office, a proxy of utilization of regular care
- HO – number of claims reimbursed for hospital care, a proxy of utilization of inpatient care

## Interpretation:

- An increase of one claim PMPM for regular care results in 0.133 increase in log of ED cost PMPM given all other predictors fixed
- An increase of one claim PMPM for inpatient care results in 11.54 increase in log of ED cost PMPM given all other predictors fixed

# Regression Analysis

## Regression Analysis in Practice

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Findings

# Access to Care: Intervention

## Access to primary care:

- Availability – proxy measure of wait times for appointment, measured in level of congestion and takes values between 0 and 1 (the higher the value, the worse the wait time)

## Interpretation:

- An increase of 1% in lack of availability of primary care providers results in \$1.078 unit increase in ED cost PMPM given all other predictors fixed

## Policy Research Question:

- Does improvement in availability of primary care providers reduce the cost of ED care?

# Findings: Access Intervention

```
Availability = dataAdult.no.out$Availability
```

```
# Improve Availability to less than 0.5 congestion experienced by all communities
```

```
Availability.interv = Availability
```

```
Availability.interv[Availability>=0.5] = 0.5
```

```
newdata=dataAdult.no.out
```

```
newdata$Availability=Availability.interv
```

```
index = which(Availability>=0.5)
```

```
# Predict by changing availability with all other predictors fixed
```

```
EDCost.predict = predict(reg.step.no.out, newdata,interval="prediction")[,1]
```

```
# Compare predicted to fitted for those communities with intervention
```

```
EDCost.diff.fitted = exp(fitted(reg.step.no.out)) - exp(EDCost.predict)
```

```
hist(EDCost.diff.fitted[index],xlab="Difference in Expected versus Predicted ED Cost")
```

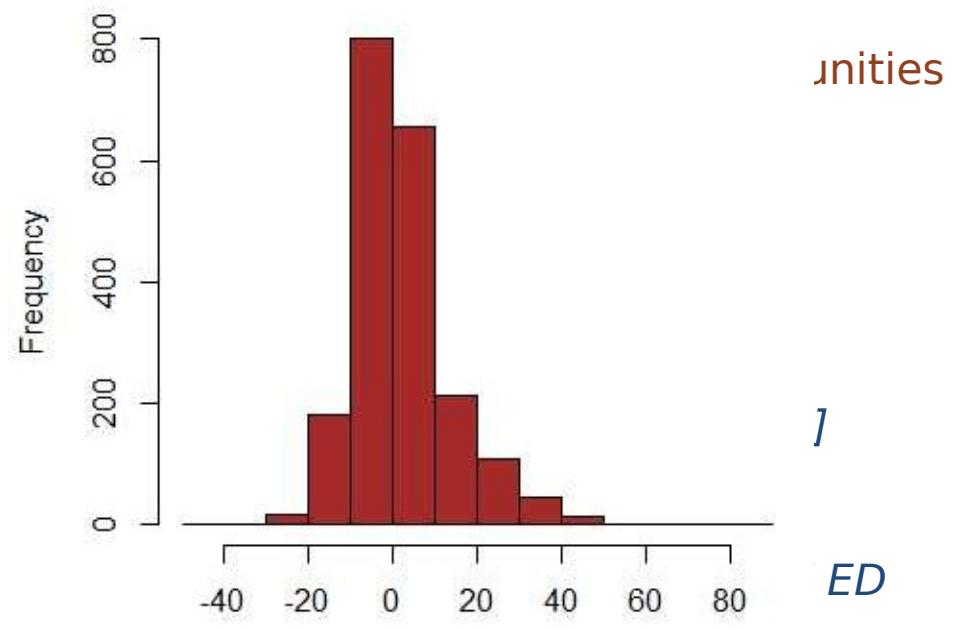
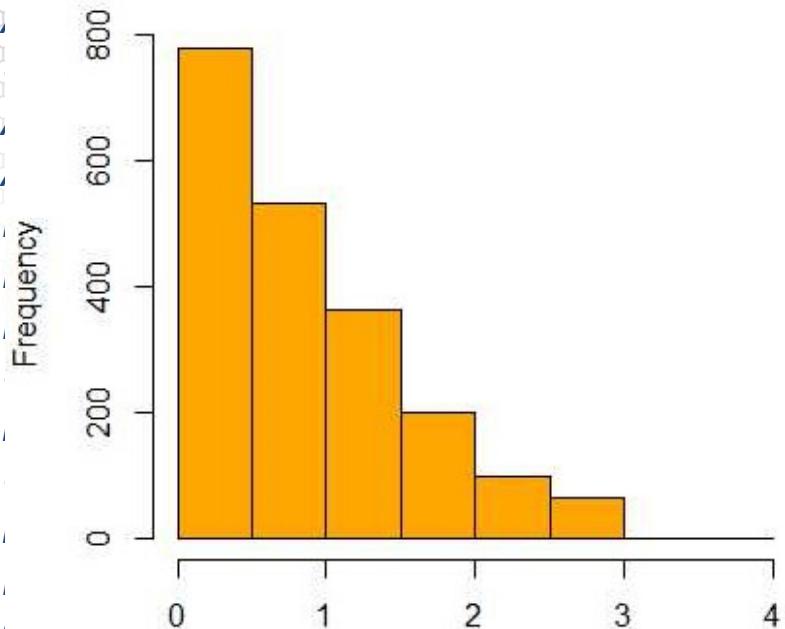
```
# Compare predicted to observed for those communities with intervention
```

```
EDCost.diff.observed = EDCost.pppm[-909] - exp(EDCost.predict)
```

```
summary(EDCost.diff.observed[index])
```

```
hist(EDCost.diff.observed[index],xlab="Difference in Observed versus Predicted ED
```

# Findings: Access Intervention



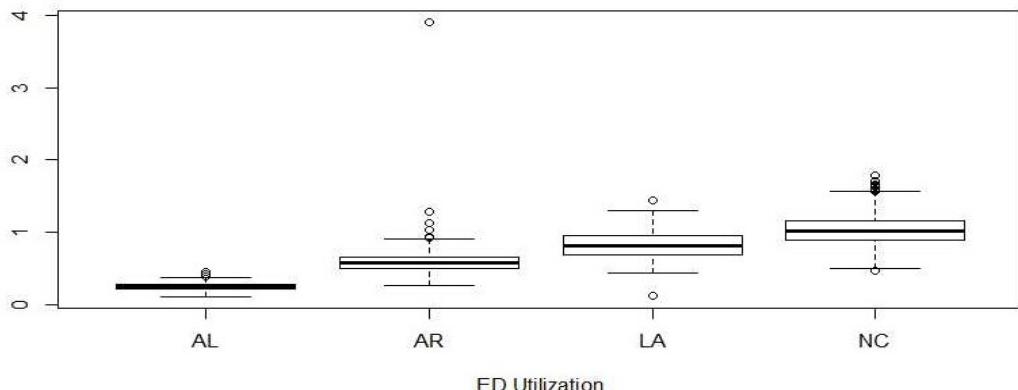
$EDCost.diff.observed = EDCost.pppm[-909] - \exp(EDCost.predict)$

`summary(EDCost.diff.observed[index])`

`hist(EDCost.diff.observed[index], xlab = "Difference in Observed versus Predicted ED`

# Findings: State Variations

- There are large variations in healthcare cost for the ED encounters across the four states, with North Carolina being the leading state and Alabama being the trailing state in cost of ED care; *Why?*
- Medicaid programs vary by states, with different health policies and reimbursements levels.
- ED utilization PMPM is also highest in North Carolina and lowest in Alabama



The correlation between ED cost and ED utilization is 0.899

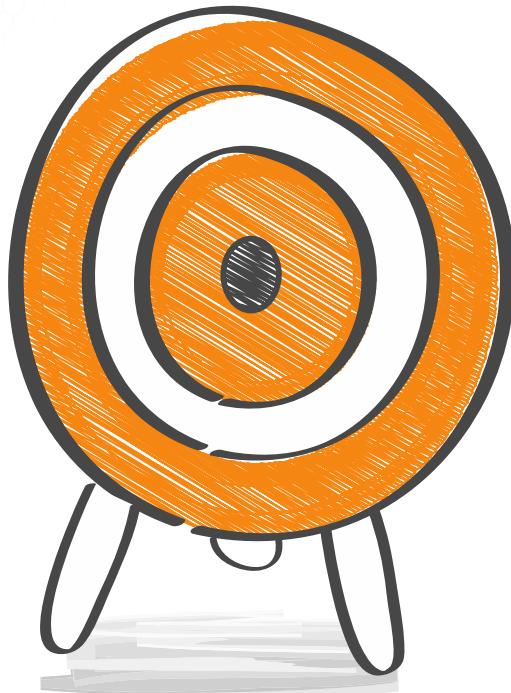
# Findings: Utilization

- Utilization of physician office is positively associated to ED cost of care given the other predicting variables fixed in the model;
- Correlation between utilization of physician office and ED is high (0.54) thus the positive relationship to ED cost may because there are communities with higher utilization of healthcare in general and thus higher ED costs.
- Utilization of inpatient care (hospitalizations) is positively associated to ED cost of care given the other predicting variables fixed in the model;
- There is a very weak correlation between utilization of ED and utilization of inpatient care; further investigation is needed.

# Findings: Other Variables

- Socio-economic variables except for Education are not selected to be included in the reduced model; they do not add additional explanatory power given the other predicting variables in the model;
- Availability of primary care providers is statistically significantly associated to ED cost of care, and intervening to improve availability will show a reduction in the expected ED cost of care according to the fitted model; such analysis however relies on causal inference and thus the regression model not appropriate to address such research question.
- Whether living in urban or rural communities is not statistically significantly associated to ED cost of care given other predicting variables in the model.

# Summary



# Regression Analysis

## Other Regression Methods

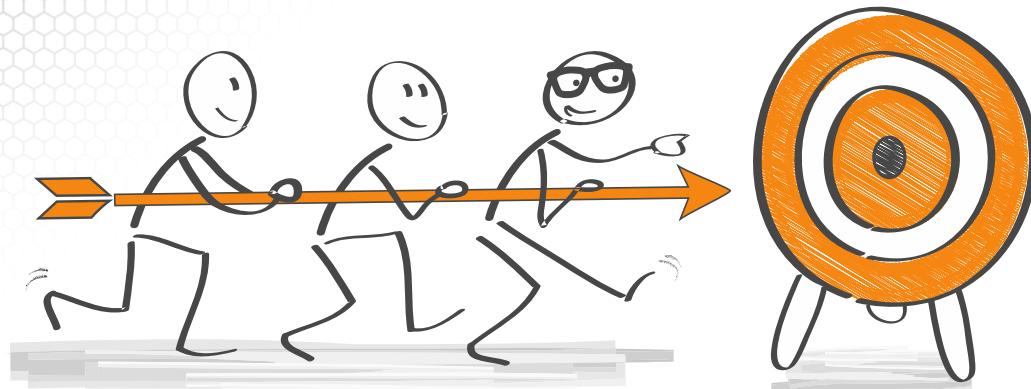
**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Weighted Least Squares  
Regression

# About this lesson



# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1,$

...,<sup>n</sup>  
**Assumptions:**

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**What if the variance is not constant?**

- **Transform the response variable using a variance-stabilizing transformation**
- **Weighted Least Squares Regression**

• *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$

• *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables

• *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Example: Normal Approximation

- Assume the number of diseased individuals in a population of size ; and
- Only observe but generally is large, thus apply the normal approximation (CLT):
  - Use a regression analysis under the normality assumption instead of logistic regression
  - $V(\cdot) =$  thus non-constant variance

# Weighted Least Regression (WLS)

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1,$

$\dots, n$

**Assumptions:** For the vector of errors  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon) = 0$
- *Covariance-Variance Assumption:*  $V(\varepsilon) = \Sigma$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables if  $\Sigma$  is a diagonal matrix
- *Normality Assumption:*  $\varepsilon \sim \text{Normal}$

# Parameter Estimation ( $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ )

To estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ , we find values that minimize squared error:

$$(Y - X\beta)^T \Sigma^{-1} (Y - X\beta)$$


$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

**Statistical Properties**  $E(\hat{\beta}) = \beta$

$$V(\hat{\beta}) = \sigma^2 (X^T \Sigma^{-1} X)^{-1}$$

**Upshot:** The covariance-variance matrix of the error terms is assumed known. However, it is needed for statistical inference.  
*How to get  $\Sigma$ ?*

# Simple WLS

The simplest WLS model:  $V(\cdot)$  with **Standard Linear Regression with log-transformation:**

- Frequency response  $\sim$  Normal Approximation:  $V(\cdot) =$  where  $E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  know.
- Generally  $x_1$  is not known

▪ Use external information: There are some cases where other information on the variance is available (e.g. measurement

**Poisson Regression.**

- $\log(E(Y|x_1, \dots, x_p))$  are  $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- ▪ Use replications. If there are several  $Y$ 's for each  $x$ , estimate  $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR as the sample variance of the replications

▪  $\log(V(Y|x_1, \dots, x_p)) = \alpha \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  a smooth function of using nonparametric regression

# Simple WLS

The simplest WLS model:  $V(s) = 1$

**Standard Linear Regression with log-transformation:**

**Model:** Probability of success given predictor(s)

$$p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$$

- **Im( $y \sim x$ , weights =  $1/w$ ) where  $w$  is a vector of the weights**  
link  $p$  to the predicting variables through a nonlinear *link function*

**How to estimate  $w()$  as a smooth function of  $x$ ?**

- **Several smoothing functions: simplest to use is 'lowess'**

- Use replications. If there are several  $Y$ 's for each  $x$ , estimate  $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR

$\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  as a smooth function of using nonparametric regression

# Regression Analysis

## Other Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Robust Regression

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Multiple Linear Regression

## What if there are outliers?

- If one or two, remove the outliers and fit again ⇒ Compare models with and without outliers
- If many, it is an indication that the normality assumption does not hold ⇒ Use an approach that provides robust estimates to outliers

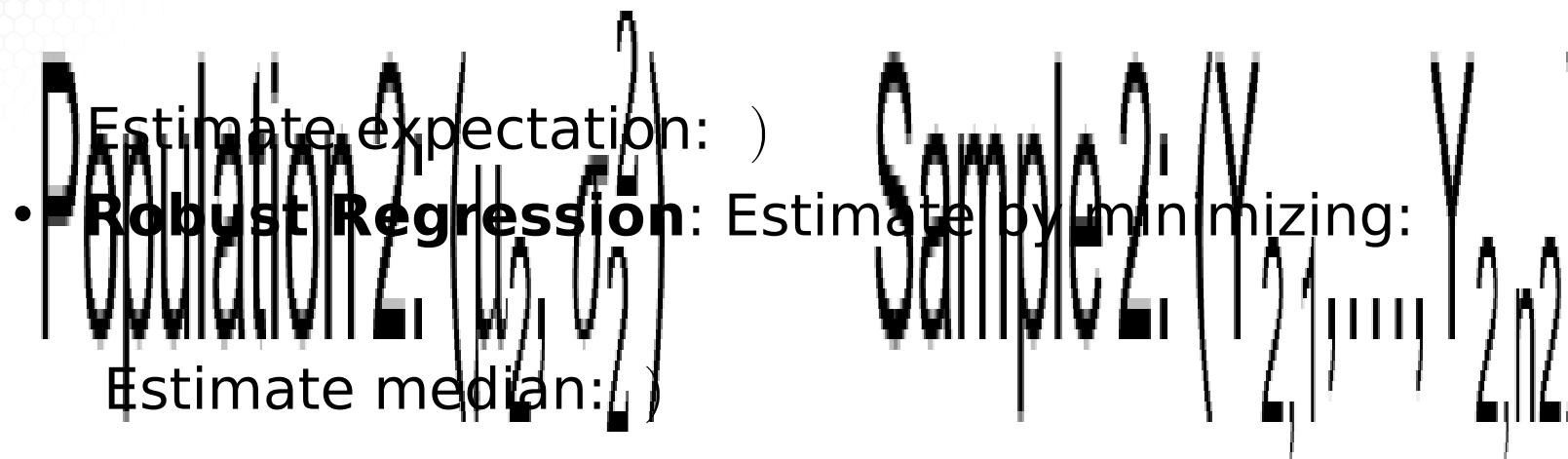
- *Constant Variance Assumption*:  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*:  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption*:  $\varepsilon_i \sim \text{Normal}$

# Example: Departure from Normality

- Assume has a pdf given as
- This has heavier tails than the normal distribution.
- MLE for  $\mu$  to minimize  $\sum (x_i - \mu)^2$ 
  - The estimate of  $\mu$  is the sample median
- Assuming in regression analysis
  - Estimate  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  by minimizing
$$|\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}|$$

# OLS vs Robust Regression

- **Ordinary Least Squares (OLS):** Estimate by minimizing



# Why not always use Robust Regression?

## Estimation Algorithm:

- Not close form expression  $\Rightarrow$  Use numeric algorithm to estimate the regression parameters: Iteratively re-weighted least squares

## Statistical Inference:

- The estimated variance is
- Efficiency comes with a cost: Confidence intervals for Robust Regression are wider than for OLS



# Regression Analysis

## Other Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Nonlinear & Nonparametric  
Regression

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Multiple Linear Regression

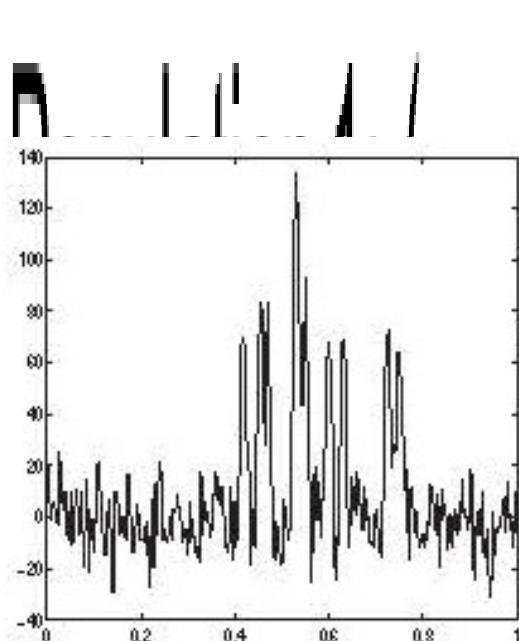
## What if nonlinearity assumption does not hold?

- If it does not hold for few quantitative predicting variables and transformation not known  $\Rightarrow$  Use transformations of the predicting variables to improve fit
- If the relationship between response and the predicting variables is known  $\Rightarrow$  Use Nonlinear Regression
- If it does not hold for many variables  $\Rightarrow$  Use Generalized Additive Regression

- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Example: Nonlinear Regression

- Mass spectroscopy or NMR frequency data:



- Each peak corresponds to one component in the sum, with  $L$  different peaks;
- Assume  $L$  is known
- Parameters:
  - for  $l=1, \dots, L$  which are the centers of the peaks
  - for  $l=1, \dots, L$  which are the amplitudes of the peaks

# Example: Nonlinear Regression

- Mass spectroscopy or NMR frequency data:



## Nonlinear Regression

- **The regression function has a known structure given the predicting variable(s); and**
- **The regression function depends on a series of parameters.**

# Nonlinear Regression

## Model Description:

- Data:  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$
- Model:  $Y_i = f(x_{i1}, \dots, x_{ip} | \theta) + \varepsilon_i, i = 1, \dots, n$
- $f(x_{i1}, \dots, x_{ip} | \theta)$  known up to the parameter Vector:  $\theta$

## Model Estimation:

- Estimate by minimizing with respect to  $\theta$
- Apply numeric algorithms to obtain the estimate for  $\theta$

# Nonlinear Regression

## Model Description:

### Nonlinear vs Linear Regression

- Use least squares approach for estimation;
- Assume same assumptions on the error terms hence goodness of fit can be performed similarly;
- Regression function is nonlinear vs linear in the parameters;
- Estimation of the parameters not in close form expression for nonlinear regression.
- R software: `nls()` vs `lm()` ← Nonlinear regression is more challenging to implement

or  $\theta$

# Nonparametric Regression

## Model Description:

- Data:  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$
- Model:  $Y_i = f(x_{i1}, \dots, x_{ip}) + \varepsilon_i, i = 1, \dots, n$
- $f(x_{i1}, \dots, x_{ip})$  unknown

## Model Estimation:

- *Curse of dimensionality*: To maintain a given degree of accuracy of an estimator, the sample size must increase exponentially with the dimension  $p$ .
  - $n = 30000$  points when  $d = 5$  to get the same accuracy as  $n = 300$  when  $d = 1$ .

# Nonparametric Regression

## Model Description:

### Nonparametric Regression

- The regression function has an unknown structure given the predicting variable(s); and
- The regression function does not depend on parameters.

### Model Estimation.

- *Curse of dimensionality*: To maintain a given degree of accuracy of an estimator, the sample size must increase exponentially with the dimension  $p$ .
  - $n = 30000$  points when  $d = 5$  to get the same accuracy as  $n = 300$  when  $d = 1$ .

# Nonparametric Regression

## (cont'd.)

Generalized Additive Models (GAM) transformation:

**Model:**  $E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  with

- $V(\log(Y)|x_1, \dots, x_p) = (x_1) + \dots + (x_p)$  where  $, \dots,$  are unlabeled smooth functions

Poisson Regression:

**Model Estimation:**

- Backfitting algorithm:  $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR
  - Iterate until convergence: For  $j=1, \dots, p$   
 $\log(Y_i|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  and estimate from regressing

# Nonparametric Regression

(cont'd)

## Nonparametric vs Linear Regression

- The relationship of a predicting variable to the response is assumed unknown
- Estimation using the least squares
- Estimation of the parameters not in close form expression
- R software: **gam()** in **mgcv** or **gam** library
- $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$
- Backfitting algorithm:
  - $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR
    - Iterate until convergence: For  $j=1, \dots, p$
  - $\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  and estimate from regressing  $Y_i$  on  $x_1, \dots, x_p$

# Regression Analysis

## Other Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Time Series Regression

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**What if uncorrelated errors assumption does not hold?**

- **Degrees of freedom are not equal to the sample size**
- **Higher variability or uncertainty than estimated thus less reliable statistical inference**

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Example: Time Series

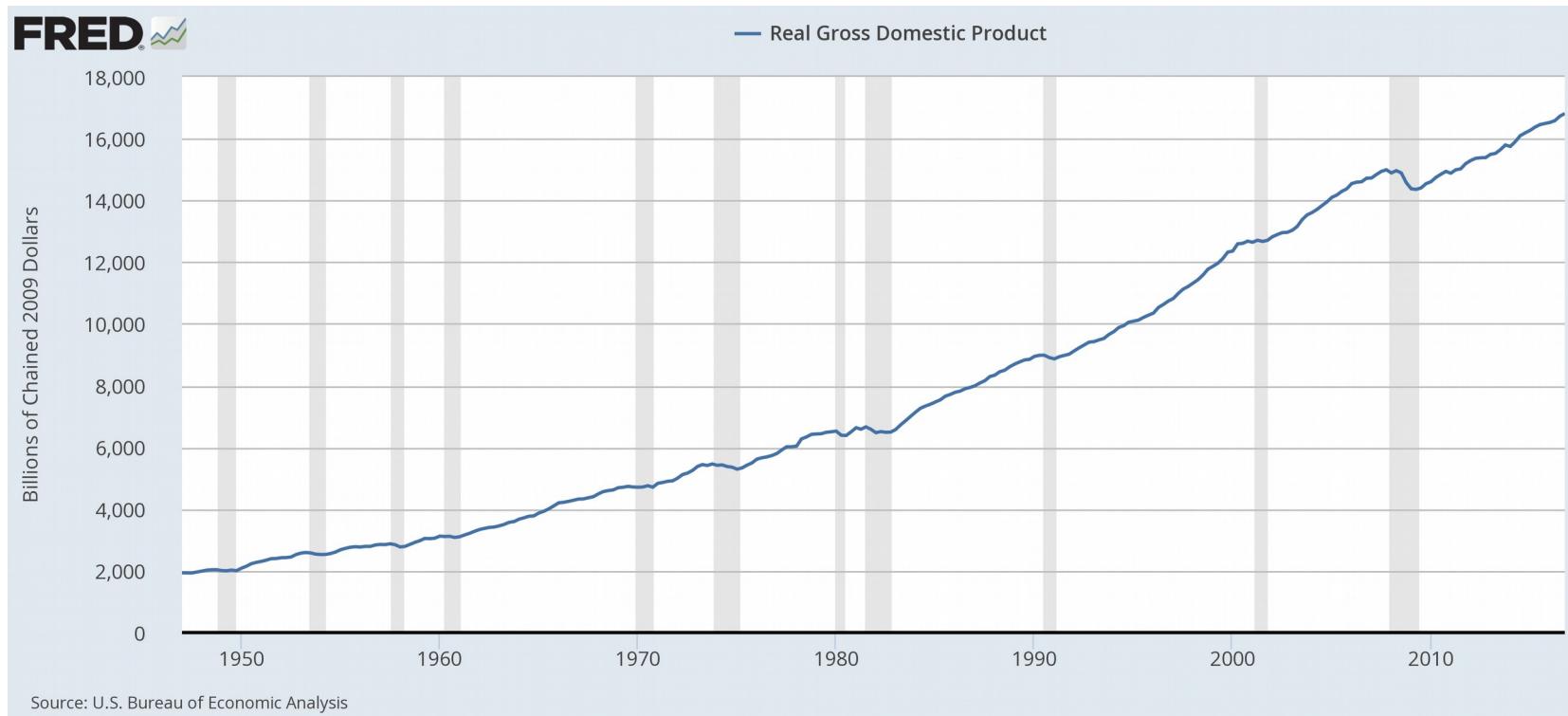
## Correlation in time:

- US yearly GDP
- Monthly sales of Australian red wine
- Monthly accidental deaths in the U.S.
- Monthly interest rates in the U.S.
- Daily Average Temperature from La Harpe station in Hancock County, Illinois
- Daily stock price of IBM stock
- 1-minute intraday S&P500 return

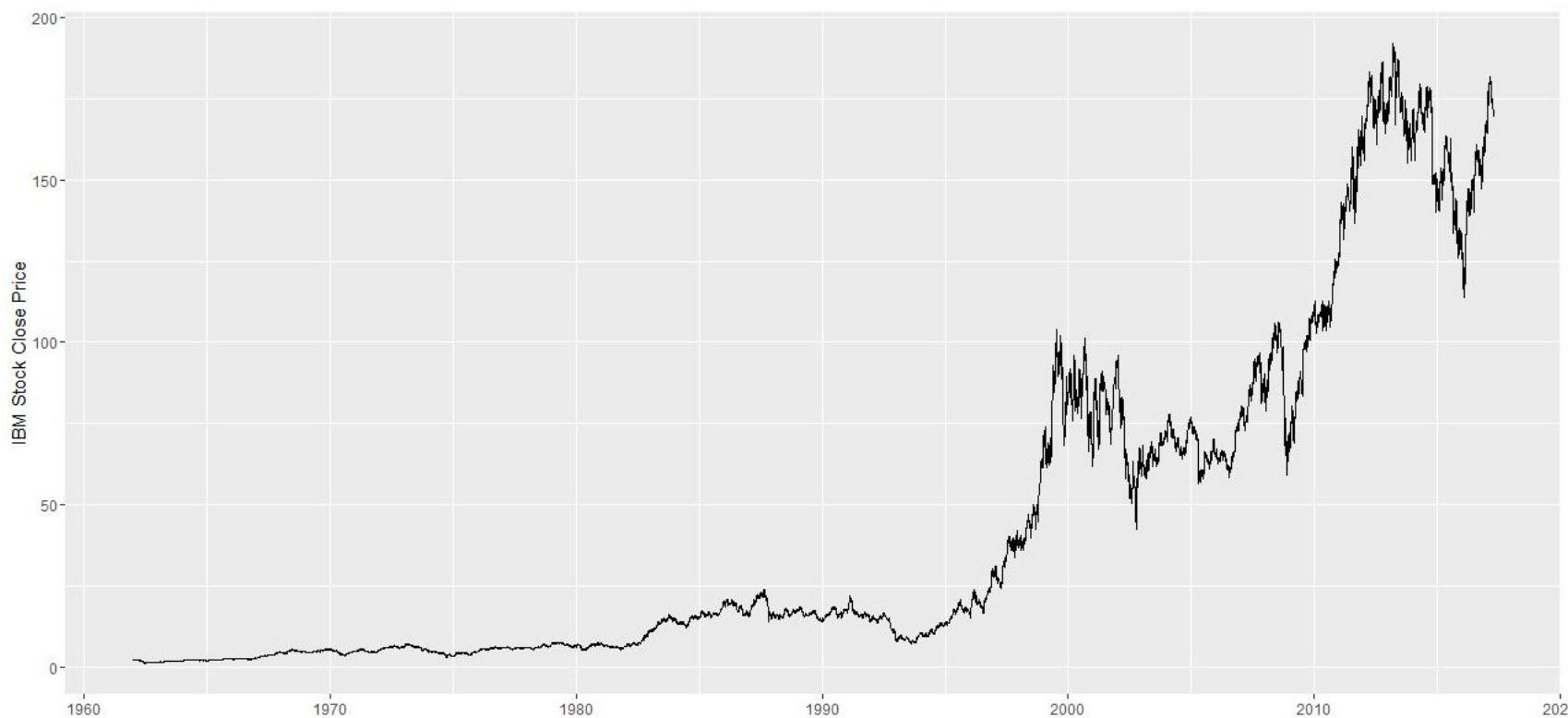
# Time Series: Characteristics

- **Trend**: long-term increase or decrease in the data over time
- **Seasonality**: influenced by seasonal factors (e.g. quarter of the year, month, or day of the week)
- **Periodicity**: exact repetition in regular pattern (seasonal series often called periodic, although they do not exactly repeat themselves)
- **Cyclical trend**: data exhibit rises and falls that are not of a fixed period
- **Heteroskedasticity**: varying variance with time
- **Correlation**: positive (successive observations are similar) or negative (successive observations are dissimilar)

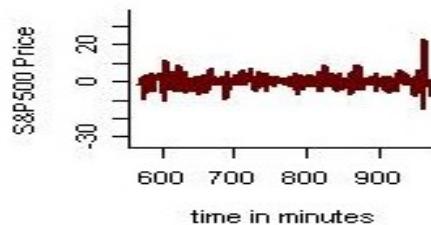
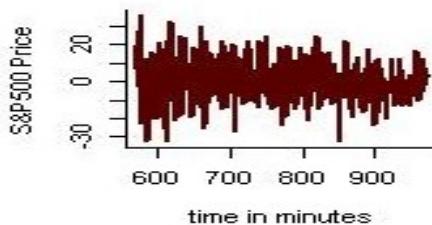
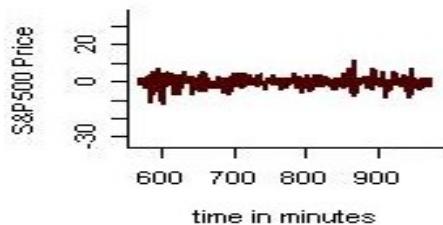
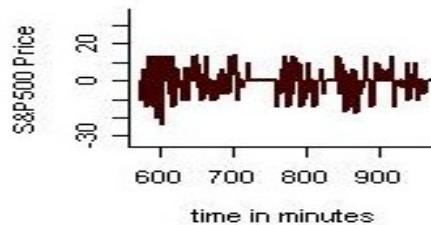
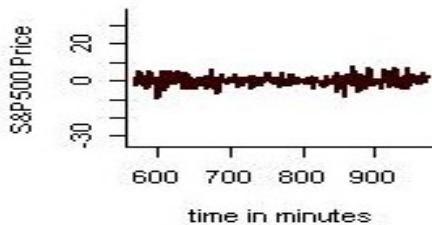
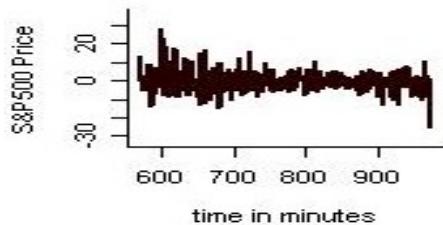
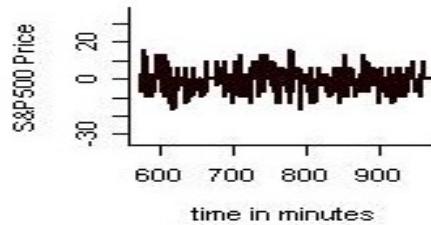
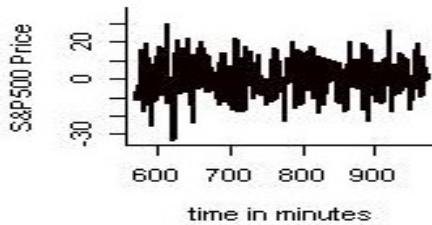
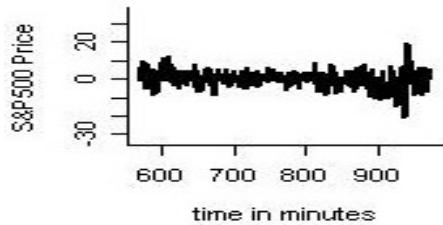
# Example: GDP



# Example: Daily IBM Stock Price



# Example: S&P500 Intraday



# Is Time Series Analysis Necessary?

**Time Series  $\Rightarrow$  Dependence**

- Data redundancy: number of degrees of freedom is smaller than  $T-1$  ( $T$  is the number of observations)
- Data sampling: concentrated about a small part of the probability space

**Ignoring dependence leads to**

- Inefficient estimates of regression parameters
- Poor predictions
- Standard errors unrealistically small (too narrow CI  $\Rightarrow$  improper inferences)

# Time Series: Basics

~~Data with linear trend, seasonal pattern, monthly, quarterly, annual, weekly, daily, hourly, minute, second, ...~~

~~Model:~~

$$E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- is a trend component;
- $V(\log(Y)|x_1, \dots, x_p)$  constant
- is a seasonality component with known periodicity  $d$

~~Poisson Regression:~~

- ~~log(E(Y|X\_1, \dots, X\_p)) component, i.e.  $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  is probability distribution does not change when shifted in time~~
- $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR

~~Estimation: trend and seasonal components are first estimated and subtracted from  $\log(Y|x_1, \dots, x_p)$  to have a left stationary process to be model using time series modeling approaches.~~

# Regression Analysis

## Other Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Spatial Regression

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**What if uncorrelated errors assumption does not hold?**

- **Degrees of freedom: not equal to the sample size**
- **Higher variability or uncertainty than estimated thus less reliable statistical inference**

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Example: Spatial Processes

## Correlation in space:

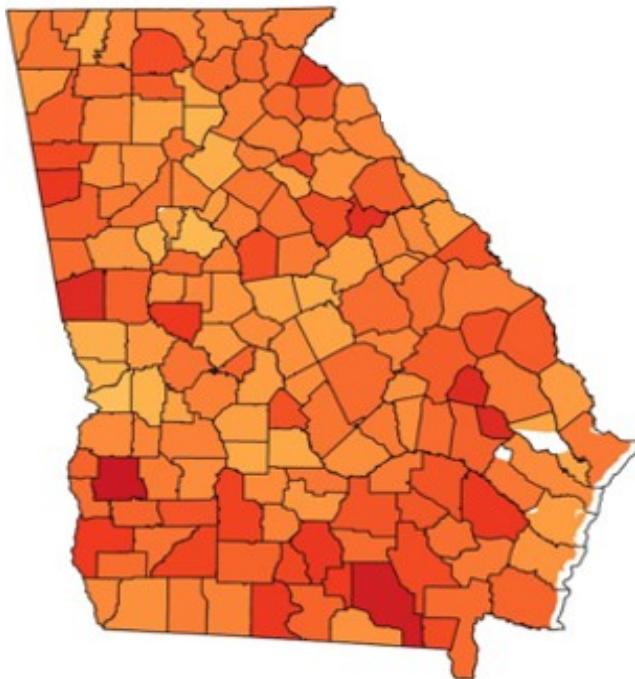
- Travel distance to a primary care provider evaluated at the community level
- The number emergency department visit per member per-month
- Locations of outbreak of a disease, for example, Zika
- Functional Magnetic Resonance Imaging (fMRI)
- Trajectory of the bison in the landscape

# Spatial Process: Characteristics

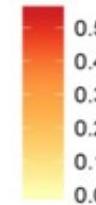
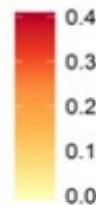
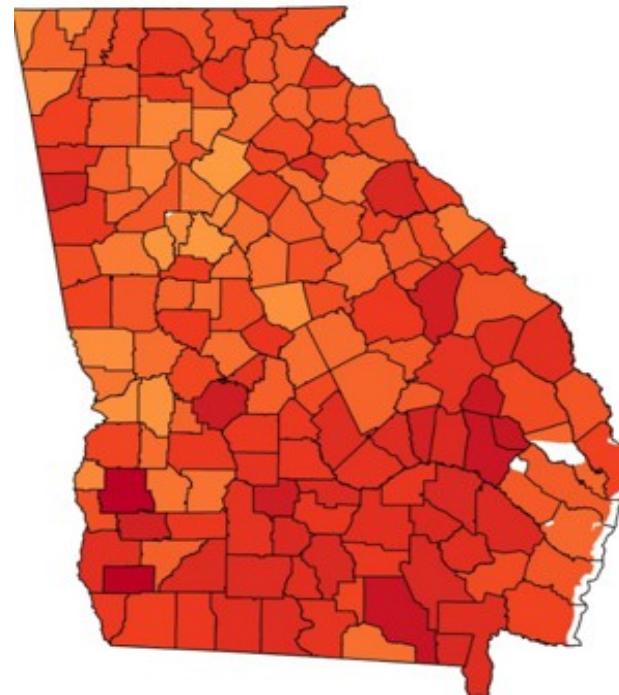
- **Trend**: long-distance increase or decrease in the data over space
- **Smoothness**: observations change slowly as the distance between their location increases
- **Heteroskedasticity**: varying variance with space
- **Continuous vs. Discrete**: at each location observe a numeric value versus a binary response versus a count response
- **Regular vs. irregular design**: the spatial process can be observed within regular or irregular division of the space

# Example 1: ED Care Utilization

*Percentage of **Children** with an ED visit*

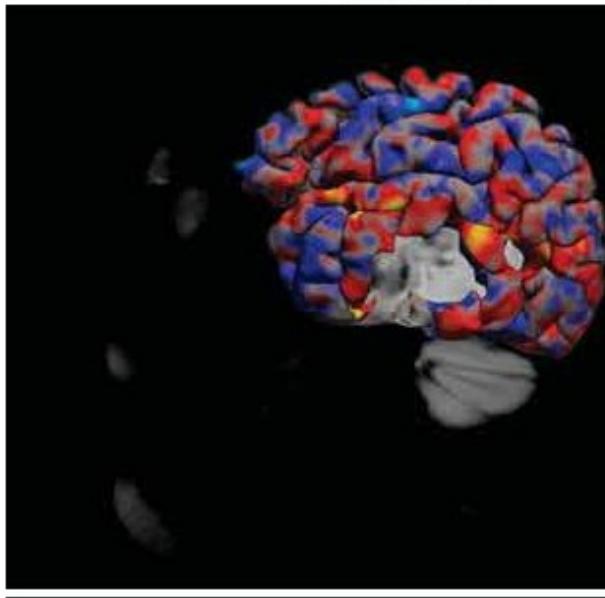


*Percentage of **Adults** with an ED visit*

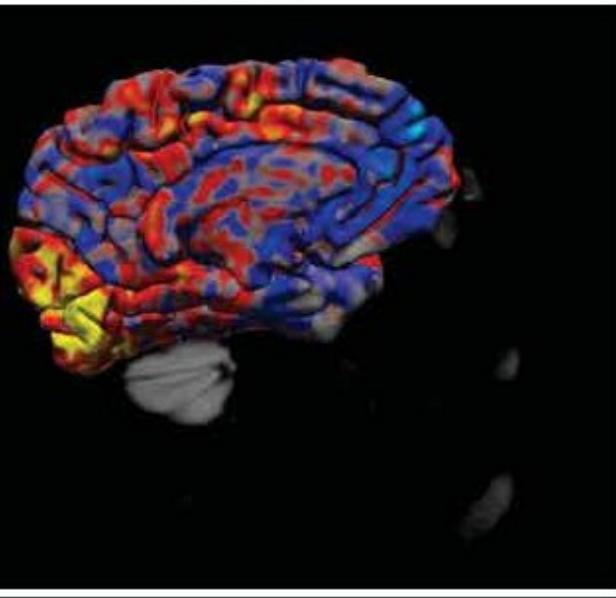


# Example 2: fMRI

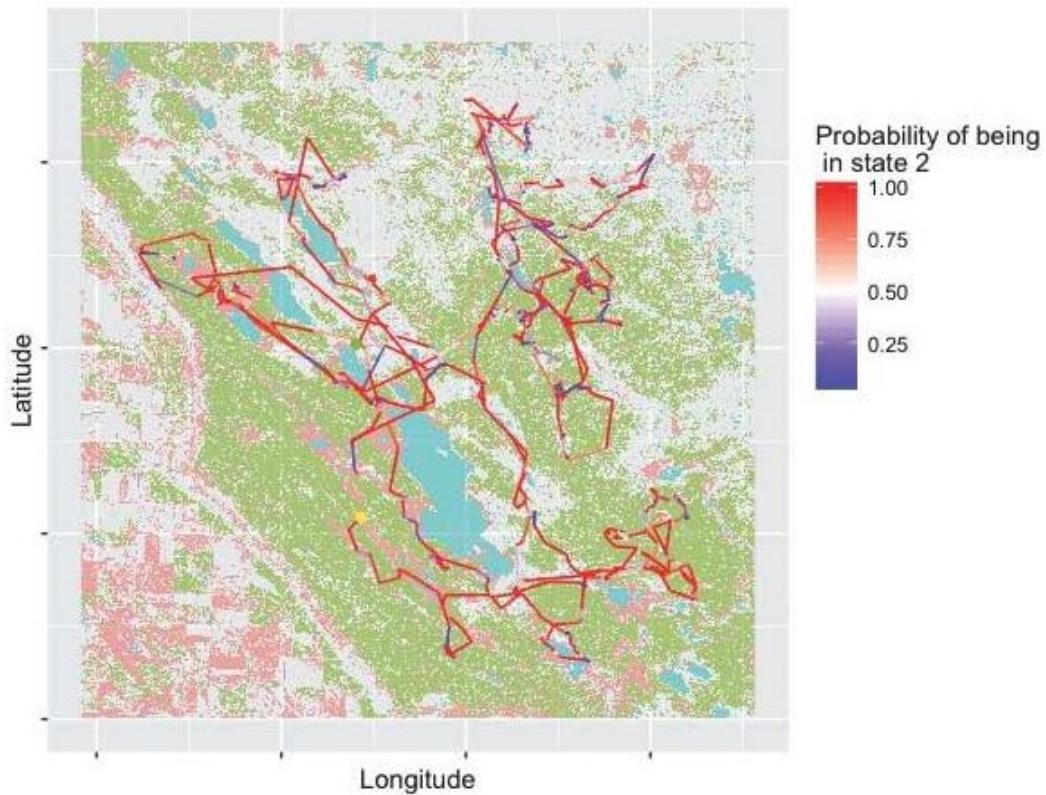
Lateral



Medial



# Example: Bison Trajectory



# Spatial Analysis: Basics

~~Standard linear Regression vs. log-transformations: tract~~

**Model:**

$$E(\log(Y)|x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- numeric response vs. point process
- $V(\log(Y)|x_1, \dots, x_p)$  constant
- stationary vs. non-stationary

**Poisson Regression:**

- isotropic process
- $\log(E(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$  where is the space domain: irregular vs regular
- observation grid  $V(Y|x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR
- Large scale vs. small-scale spatial dependence  
 $\log(V(Y|x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Regression Analysis

## Other Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Mixed Effects Models

# ANOVA Model

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

**Model:**  $Y_{ij} =$  where

•  $\mu$  =  $i$ th group mean decomposed into =

• In some designs, the categorical variable is  
“subject” or experimental setting

- Simplest example: repeated measures, where more than one (identical) measurement is taken on the same setting.

# ANOVA Model: Random Effects

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

**Model:**  $Y_{ij} =$  where

With group mean decomposed into  $=$

- We might be interested in the variability across subjects, i.e. . Is it zero?

# ANOVA Model: Random Effects

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

**Are the assumptions the same as in ANOVA with fixed effects?**

- **In random effects model, the observations are no longer independent (even if the error terms are independent).**
- We might be interested in the variability across subjects, i.e. . Is it zero?

# ANOVA Model: Random Effects

When to use random effects?

- A “group” effect is random if we can think of the responses we observe in the group to be samples from a larger population.
- Example: if collecting data from different medical centers, “center” might be thought of as random.
- Example: if surveying students on different campuses, “campus” may be a random effect.

# Regression Model: Mixed Effects

- In some studies, some factors can be thought of as fixed, others random.
- Suppose we study the effect of a blood pressure meant to lower blood pressure over time and we study  $n$  patients.
- For each patient we record BP at regular intervals over a week (every day, say).
- **Model:**  $Y_{ij} =$
- in this example

# Regression Model: Mixed Effects

- In some studies, some factors can be thought of as fixed, others random.
- **If not all the X's are the same for each subject, or some observations are missing, things are more complicated.**
- **Covariance matrix of Y is more complicated; use a computer to estimate such models!**
- **Model:**  $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_{ij} + \beta_3 W_{ij} + \beta_4 U_{ij} + \epsilon_{ij}$
- in this example

# Regression Analysis

## Regression Methods

**Nicoleta Serban, Ph.D.**

*Associate Professor*

Stewart School of Industrial and Systems Engineering

Regression Analysis: Overview

# Simple Linear Regression

**Data:**  $\{(x_1, Y_1), \dots, (x_n, Y_n)\}$   
**Population 1:**  $(\mu_1, \sigma_1^2)$   
**Model:**  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption* :  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption*:  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption*:  $\varepsilon_i \sim \text{Normal}$

# ANOVA

**Data:**  $Y_{ij}$  for  $j = 1, 2, \dots, n_i$ ;  $i = 1, \dots, k$   
**Population k:**  $(\mu_k, \sigma_k^2)$   
**Model:**  $Y_{ij} = \mu_i + \epsilon_{ij}$  where  $\epsilon_{ij}$  is an error term

## Assumptions:

- **Constant Variance Assumption:**  $\text{Var}(\epsilon_{ij}) = \sigma^2$
- **Independence Assumption:**  $\{Y_{1j}, \dots, Y_{nj}\}$  are independent, random variables
- **Normality Assumption:**  $\sim \text{Normal}(0, \sigma^2)$

# Multiple Linear Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

## Assumptions:

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon_i) = 0$
- *Constant Variance Assumption:*  $\text{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon_i \sim \text{Normal}$

# Logistic Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  binary responses

**Model:** Probability of success given predictor(s)

**Estimate**

link  $p$  to the predicting variables through **logit link function**

**Assumptions:** Standard Linear Regression with log-transformation:

- $E(\log(Y) | x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

- **Linearity Assumption:** constant

**Poisson Regression:**

- **Independence Assumption:**  $Y_1, Y_2, \dots, Y_n$  are independent random variables

- $E(Y | x_1, \dots, x_p) = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$  OR

- **Logit link function:**  $\log(E(Y | x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Poisson Regression

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  count response variable

**Model:** Model the conditional expectation:

**Model:** Probability of success given predictor(s)

**Assumptions:**  $p = p(x_1, \dots, x_p) = P\{Y=1|x_1, \dots, x_p\}$

- *Linearity Assumption:*  $+ \dots +$
- *link p to the predicting variables through a nonlinear link function*  
*Independence Assumption:*  $Y_1, \dots, Y_n$  are independent random variables
- *Variance Assumption:*  $g(p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

# Generalized Linear Model

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$  where  $Y_1, \dots, Y_n$  response variable with **a distribution from the exponential family**

**Model:** Model the conditional expectation:

$$E[\hat{\beta}_1] = \text{OR} \frac{\sigma^2}{S_{xx}}$$

where  $(\cdot)$  is a *link function* and  $(\cdot)$  the *inverse link function* depending on the distribution of  $Y$ .

# Weighted Least Regression (WLS)

**Data:**  $\{(x_{11}, \dots, x_{1p}), Y_1\}, \dots, \{(x_{n1}, \dots, x_{np}), Y_n\}$

**Model:**  $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, i = 1, \dots, n$

**Assumptions:** For the vector of errors  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$

- *Linearity/Mean Zero Assumption:*  $E(\varepsilon) = 0$
- *Covariance-Variance Assumption:*  $V(\varepsilon) = \Sigma$
- *Independence Assumption:*  $\{\varepsilon_1, \dots, \varepsilon_n\}$  are independent random variables
- *Normality Assumption:*  $\varepsilon \sim \text{Normal}$

# Generalized Additive Model (GAM)

**Model:**  $Y_i = f(x_{i1}, x_{ip}) + \epsilon_i$ ,  $i = 1, \dots, n$  with

$f(x_{i1}, x_{ip}) = f_1(x_{i1}) + \dots + f_p(x_{ip})$  where  $f_1, \dots, f_p$  are unknown smooth functions

**Model Estimation:**

- Backfitting algorithm:

- Initialize:  $\hat{\alpha}, \hat{f}_1, \dots, \hat{f}_p$
- Iterate until convergence: For  $j = 1, \dots, p$

$\check{Y}_i = Y_i - \hat{\alpha} - \sum_{k \neq j} f_k(x_{ki})$  and estimate  $f_j$  from regressing  $\check{Y}_i \sim x_{ji}$

# Summary

