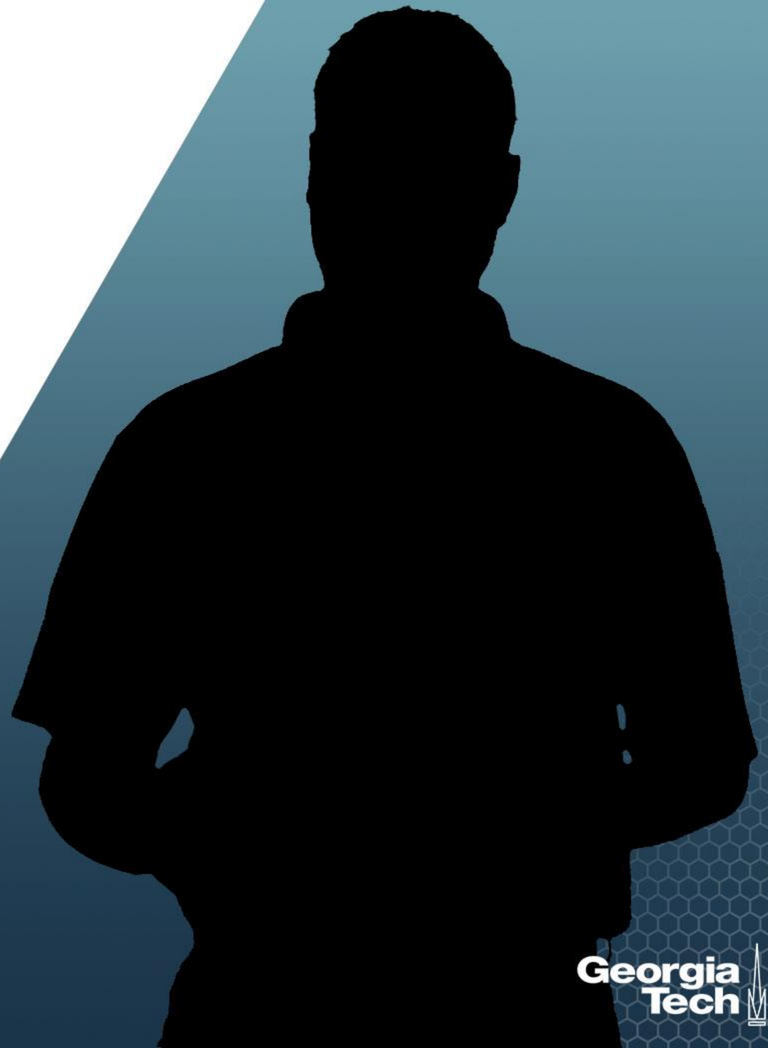
Regression Analysis

Multiple Linear Regression

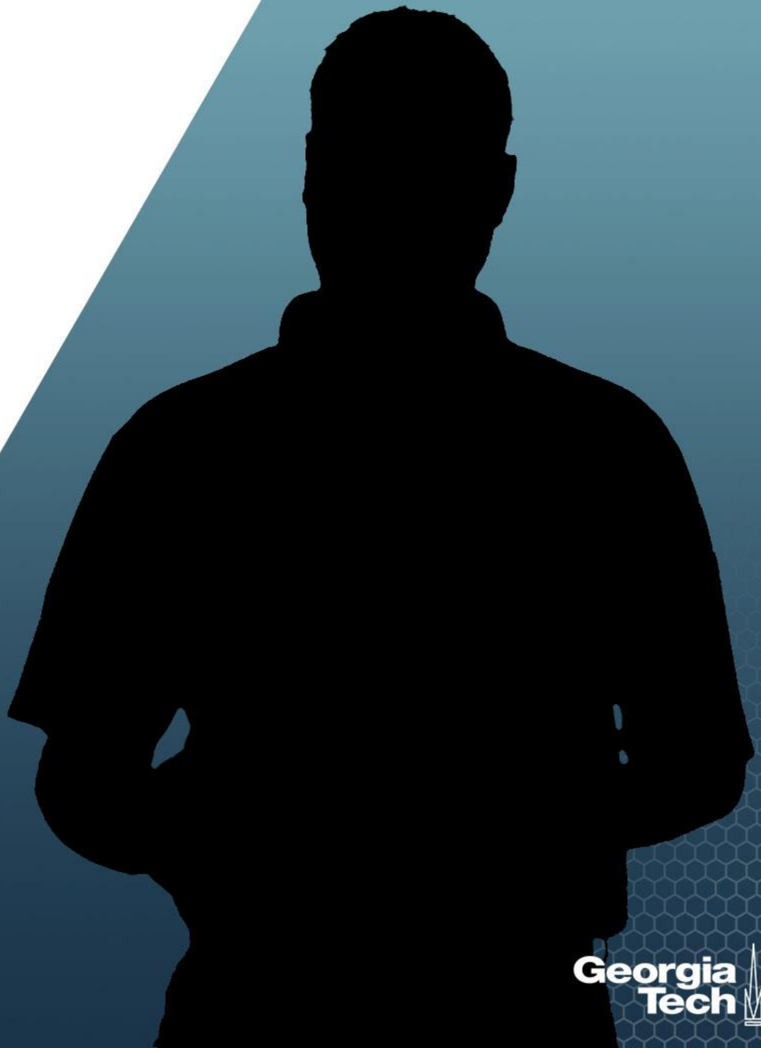
Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Objectives and Examples

Georgia Tech

About This Lesson



Linear Regression: Example 1

Linear Regression: Data Example 1

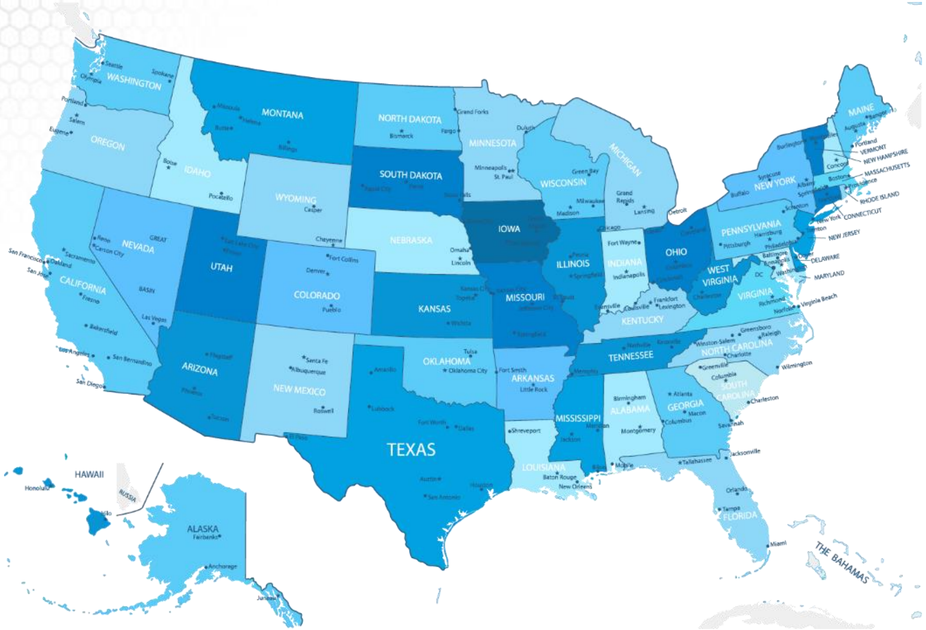
The response variable is:

Y = Sales (in thousands of dollars)

The predicting variables are:

X_1 = Amount (in hundreds of dollars) spent on advertising
X_2 = Total amount of bonuses paid
X_3 = Market share in each territory
X_4 = Largest competitor's sales
X_5 = 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:

$\boldsymbol{Y}$ = State average SAT score (verbal and quantitative combined)

The predicting variables are:

$\boldsymbol{X}_1$ = % of total eligible high school seniors in the state who took the exam
$\boldsymbol{X}_2$ = Median income of families of test takers, in hundreds of dollars
$\boldsymbol{X}_3$ = Average number of years that test takers had in social sciences, natural sciences, and humanities
$\boldsymbol{X}_4$ = % of test takers who attended public schools
$\boldsymbol{X}_5$ = State expenditure on secondary schools, in hundreds of dollars per student
$\boldsymbol{X}_6$ = Median percentile of ranking of test takers within their secondary school classes

Linear Regression: Example 3

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Data Source: UCI Machine Learning Repository

Linear Regression: Data Example 2

The response variable is:

$\boldsymbol{Y}$ = Hourly count rentals of bikes

The predicting variables are:

$\boldsymbol{X}_1$ = Day of the week
$\boldsymbol{X}_2$ = Month of the year
$\boldsymbol{X}_3$ = Hour of the day (ranging 0-23)
$\boldsymbol{X}_4$ = Year (2011, 2012)
$\boldsymbol{X}_5$ = Holiday Indicator
$\boldsymbol{X}_6$ = Weather condition (with four levels from good weather for level 1 to severe condition for level 4)
$\boldsymbol{X}_7$ = Normalized temperature
$\boldsymbol{X}_8$ = Normalized humidity
$\boldsymbol{X}_9$ = Wind speed

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 explanatory variables
3. **Testing** hypotheses of association relationships

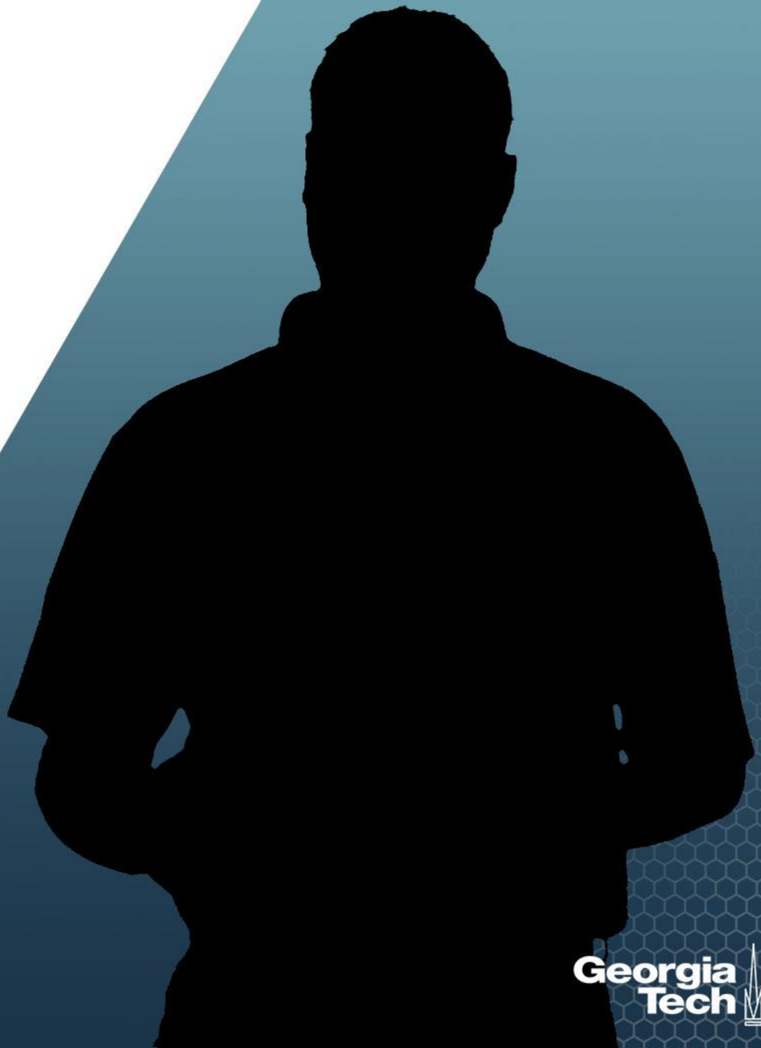
Linear Regression: The basis of what we will discussing in most of this course is the linear model. Virtually all other methods for studying dependence among variables are variations on the idea of linear regression.

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

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

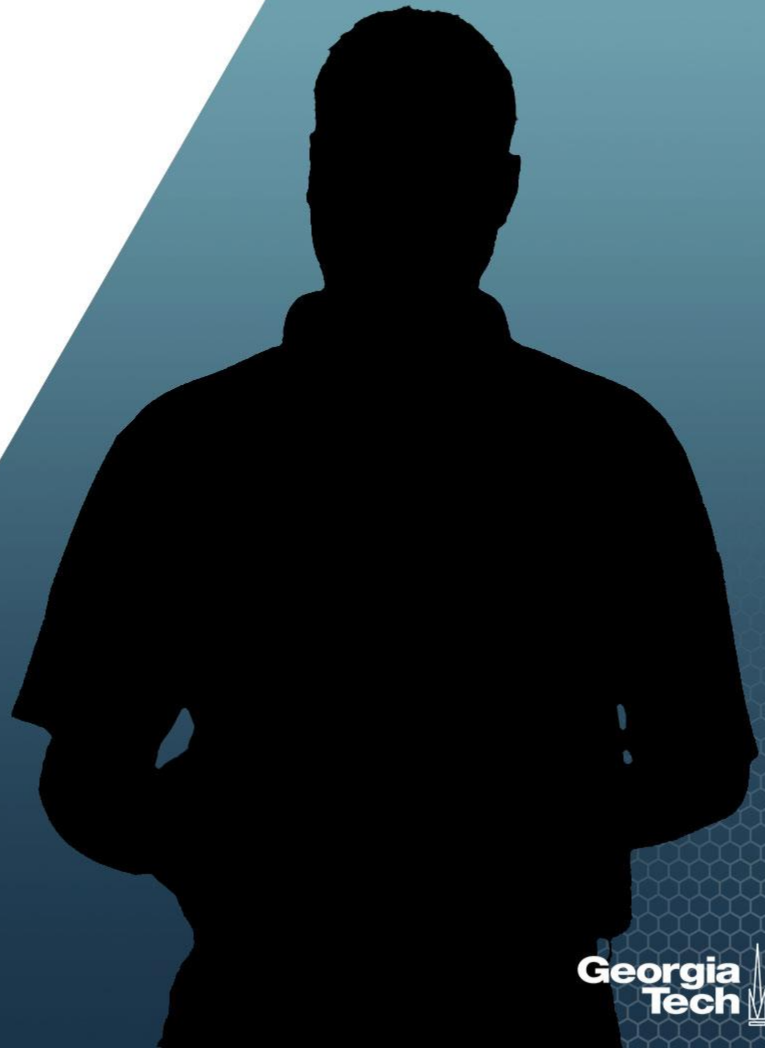
Georgia Tech

Summary



Regression Analysis

Multiple Linear Regression



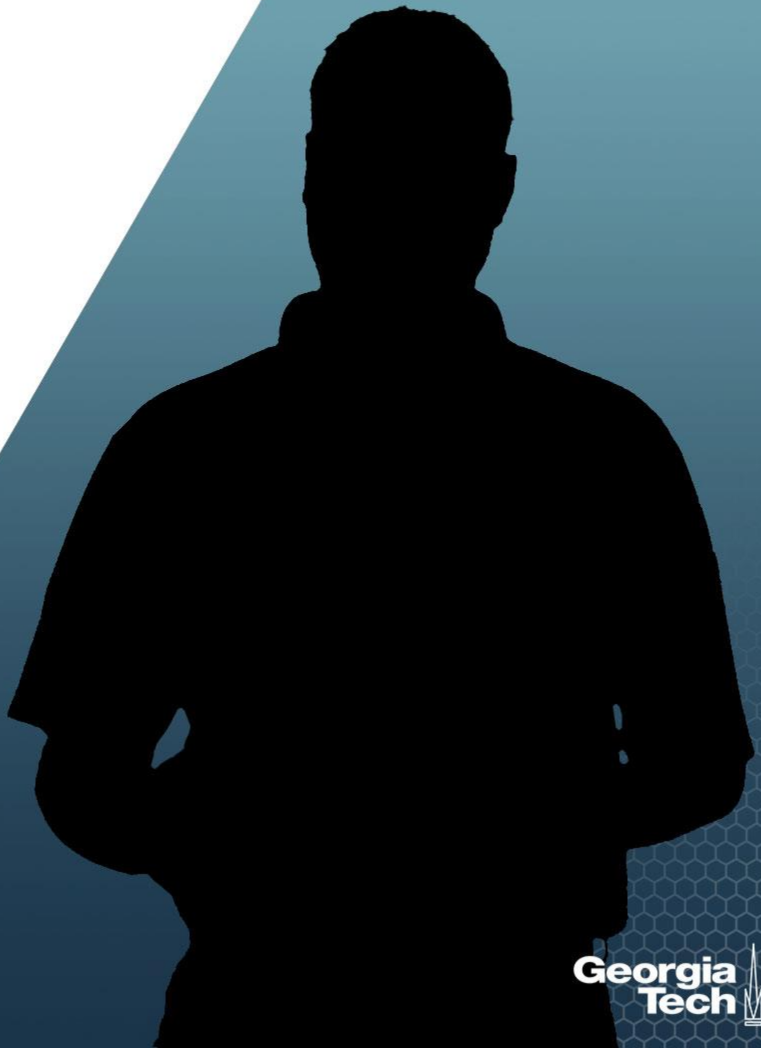
Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering

Basics of Multiple Regression

About This Lesson



Multiple Linear Regression: Model

Data: $\{(x_{1,1},\dots,x_{1,p}),y_1\},\dots,\{(x_{n,1},\dots,x_{n,p}),y_n\}$
Model: $Y_i=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\cdots+\beta_p x_{i,p}+\varepsilon_i,\ \ i=1,\dots,n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i)=0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i)=\sigma^2$
- *Independence Assumption*: $\{\varepsilon_1,\dots,\varepsilon_n\}$ are independent random variables
- ε_i ~ Normally distributed *for confidence/prediction intervals, hypothesis testing*

Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \dots, x_{1,p}), y_1\}, \dots, \{(x_{n,1}, \dots, x_{n,p}), y_n\}$
Model: $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \;\; i = 1, \dots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \dots, \varepsilon_n\}$ are independent random variables
- $\varepsilon_i \sim$ Normally distributed *for confidence/prediction intervals, hypothesis testing*

The model parameters are: $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1, \dots, \beta_p, \sigma^2}$

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

Multiple Linear Regression: Model

Data: $\{(x_{1,1},\dots,x_{1,p}),y_1\},\dots,\{(x_{n,1},\dots,x_{n,p}),y_n\}$
Model: $Y_i=\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\cdots+\beta_p x_{i,p}+\varepsilon_i,\ \ i=1,\dots,n$

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

Response

$$Y=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix}$$

Design Matrix

$$\mathbf{X}=\begin{bmatrix}1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p}\\1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p}\\\vdots & \vdots & \vdots & \vdots & \vdots\\1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p}\end{bmatrix}$$

Coefficients

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{pmatrix}$$

Error

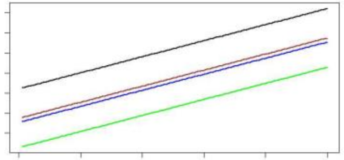
$$\boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix}$$

Georgia Tech

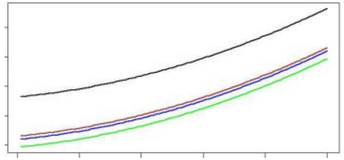
Model Flexibility: Main Effects & Interactions

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

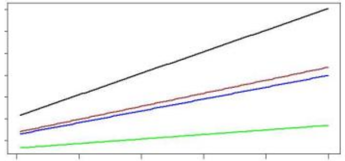
- ***1st* Order Model:**
  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$



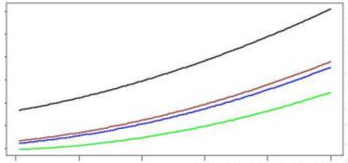
- ***2nd* Order Model:**
  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$



- ***1st* Order Interaction Model:**
  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$



- ***2nd* Order Interaction Model:**
  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$



Georgia Tech

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

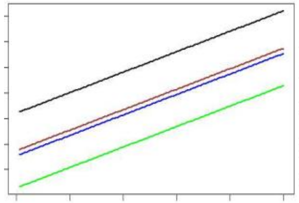
Quantitative and Qualitative Variables

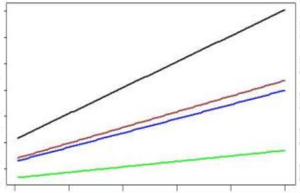
Multiple Linear Regression: Multiple quantitative/qualitative predicting variables

$\boldsymbol{x_1}$ quantitative
$\boldsymbol{x_2}$ qualitative with three levels: $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{D_3}$ dummy variables

Model: $\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 d_1} + \boldsymbol{\beta_3 d_2} + \boldsymbol{\varepsilon}$ ➡ **Intercept varies**

$$\left.\begin{array}{l} \text{If } d_1{=}0, d_2{=}0\text{: } \boldsymbol{\beta_0 + \beta_1 x_1} \\ \text{If } d_1{=}1, d_2{=}0\text{: } (\boldsymbol{\beta_0{+}\beta_2}) + \boldsymbol{\beta_1 x_1} \\ \text{If } d_1{=}0, d_2{=}1\text{: } (\boldsymbol{\beta_0{+}\beta_3}) + \boldsymbol{\beta_1 x_1} \end{array}\right\} \textbf{Parallel regression lines}$$





If $\boldsymbol{x_1}$ $\boldsymbol{x_2}$ interaction: Nonparallel regression lines

Linear Regression: Example 1

Linear Regression: Example 1

Quantitative Predicting Variables:

$\boldsymbol{X}_1$ = The amount (in hundreds of dollars) spent on advertising in 1999
$\boldsymbol{X}_2$ = The total amount of bonuses paid in 1999
$\boldsymbol{X}_3$ = The market share in each territory
$\boldsymbol{X}_4$ = The largest competitor's sales

Qualitative Predicting Variable:

$\boldsymbol{X}_5$ = Indicates the region of the office (1 = south, 2 = west, 3 = midwest)

Linear Regression: Example 3

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Linear Regression: Example 3

Qualitative predicting variables:

$\boldsymbol{X}_1$ = Day of the week
$\boldsymbol{X}_2$ = Month of the year
$\boldsymbol{X}_3$ = Hour of the day (ranging 0-23)
$\boldsymbol{X}_4$ = Year (2011, 2012)
$\boldsymbol{X}_5$ = Holiday Indicator
$\boldsymbol{X}_6$ = Weather condition (with four levels from good weather for level 1 to severe condition for level 4)

Quantitative predicting variables:

$\boldsymbol{X}_7$ = Normalized temperature
$\boldsymbol{X}_8$ = Normalized humidity
$\boldsymbol{X}_9$ = Wind speed

Linear Regression: Example 3

Qualitative predicting variables:

X_1 = Day of the week
X_2 = Month of the year
X_3 = Hour of the day (ranging 0-23)
X_4 = Year (2011, 2012)
X_5 = Holiday Indicator
X_6 = Weather condition (with four levels from good weather for level 1 to severe condition for level 4)

Quantitative predicting variables:

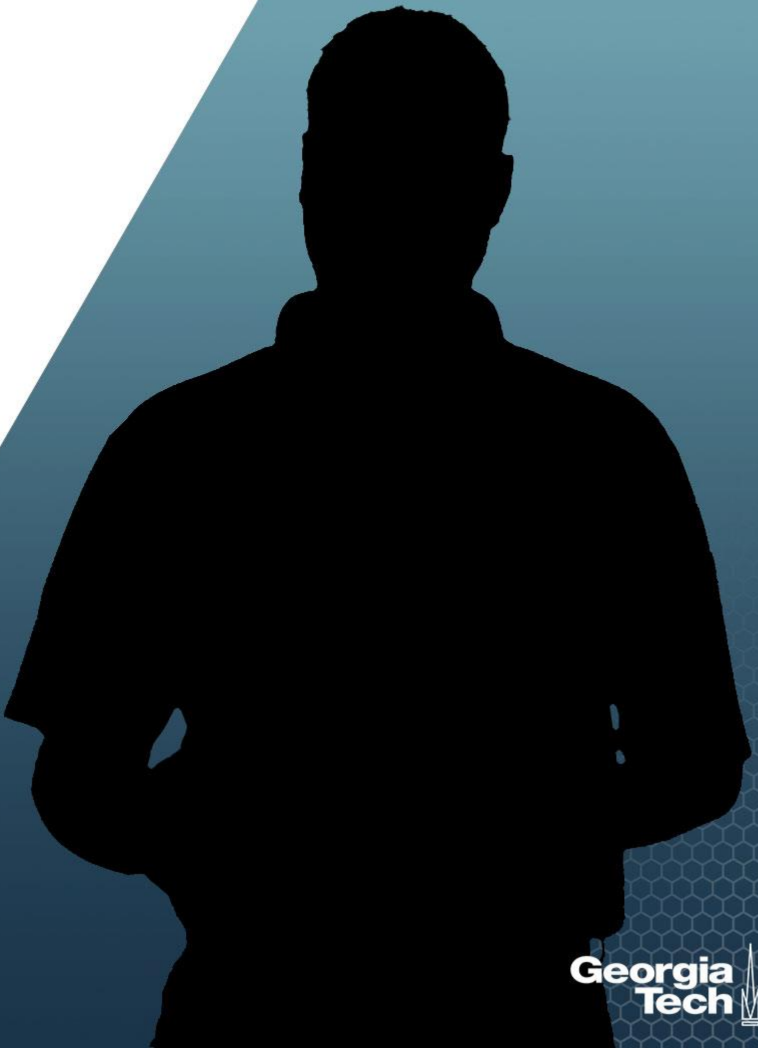
X_7 = Normalized temperature
X_8 = Normalized humidity
X_9 = Wind speed

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

- If observations are made over many years, then consider it to be *quantitative*
- If observations are made over only a few years, then consider it to be *qualitative*

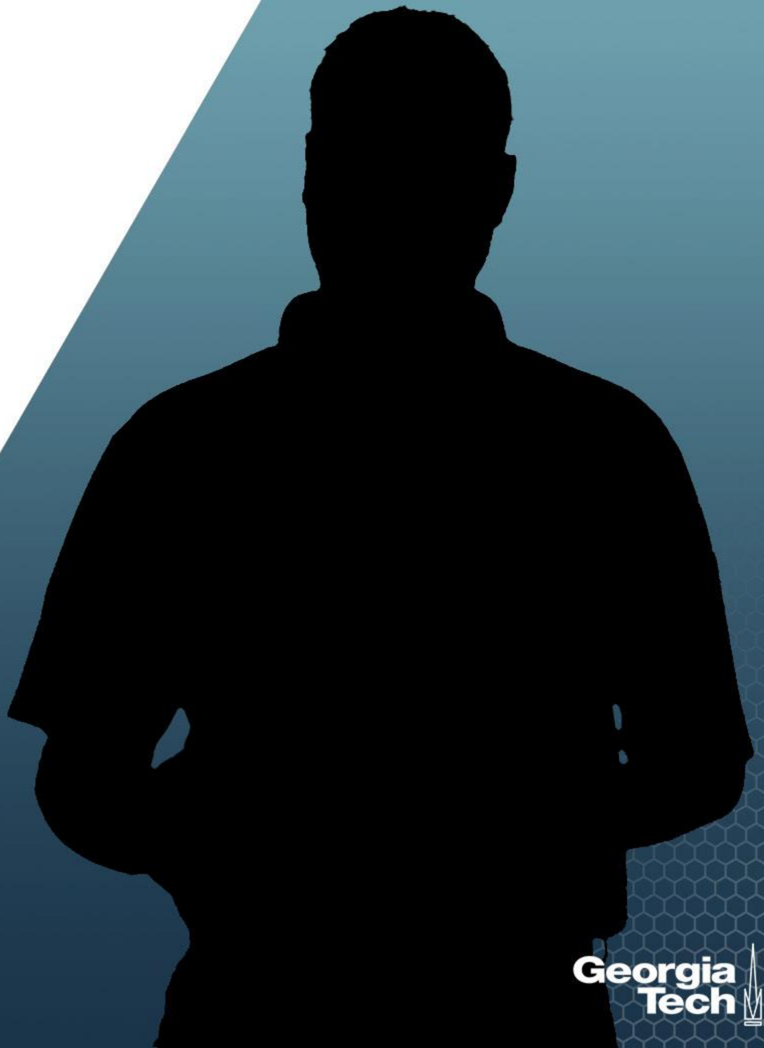
Georgia Tech

Summary



Regression Analysis

Multiple Linear Regression



Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Regression Parameter Estimation

About This Lesson

Parameter Estimation $\left(\beta_0, \beta_1, \ldots, \beta_p\right), \sigma^2$

To estimate $\left(\beta_0, \beta_1, \ldots, \beta_p\right)$, find values $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ that minimize squared error:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_p x_{i,p})\right)\right)^2 = \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

By linear algebra (Orthogonal Decomposition Theorem) or differentiation:

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = 0$$

So

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

If $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible,

$$\boxed{\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}}$$

Parameter Estimation $\left(\beta_0, \beta_1, \ldots, \beta_p\right), \sigma^2$

The fitted values are $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, so

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

where $\mathbf{H} \equiv \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ is called the **hat matrix** because multiplying $\boldsymbol{y}$ by $\mathbf{H}$ gives $\widehat{\boldsymbol{y}}$.

The residuals are:

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

To estimate σ^2,

$$\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}/(n - p - 1)$$

Parameter Estimation $(\beta_0, \beta_1, \dots, \beta_p), \sigma^2$

The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, so

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

where $\mathbf{H} \equiv \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}$ is called the **hat matrix** because multiplying $\boldsymbol{y}$ by $\mathbf{H}$ gives $\hat{\boldsymbol{y}}$.

The residuals are:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{y}$$

To estimate σ^2,

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}/(n - p - 1)$$

The estimator of $\boldsymbol{\sigma}^2$ is MSE

Assuming $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ are normally distributed

- **MSE ~ χ^2 with *n-p-1* degrees of freedom**

Parameter Estimation

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}}}{n-p-1} = \frac{\sum \hat{\boldsymbol{\varepsilon}}_i^2}{n-p-1} \sim \boldsymbol{\chi}^2_{n-p-1}$$

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

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

Estimating σ^2

Sample variance

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

Parameter Estimation

Recall that $\varepsilon_i = \left(y_i - \left(\beta_0 + \beta_1 \mathrm{x}_{i,1} + \cdots + \beta_p \mathrm{x}_{i,p} \right) \right)$

Replaced by $\hat{\varepsilon}_i = \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_{i,1} + \cdots + \hat{\beta}_p \mathrm{x}_{i,p} \right) \right)$

Use *p*+1 degrees of freedom because

$$\beta_0 \leftarrow \hat{\beta}_0$$
$$\beta_1 \leftarrow \hat{\beta}_1$$
$$\vdots$$
$$\beta_p \leftarrow \hat{\beta}_p$$

Thus, assuming that

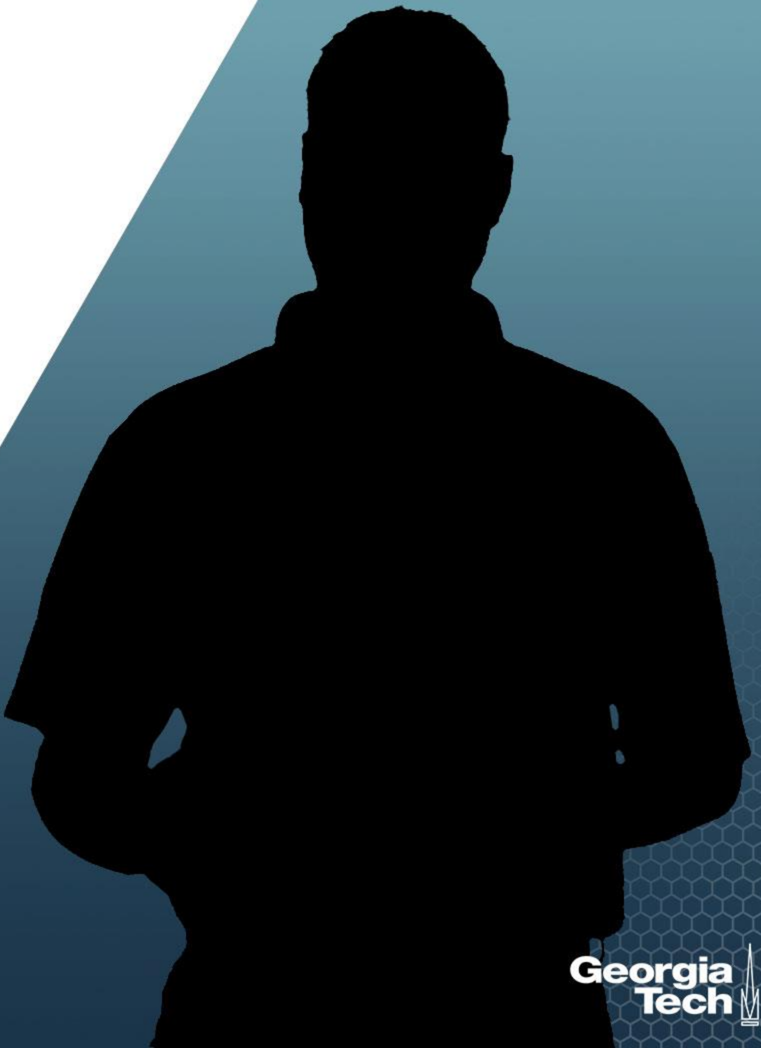
$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$$

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

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

Georgia Tech

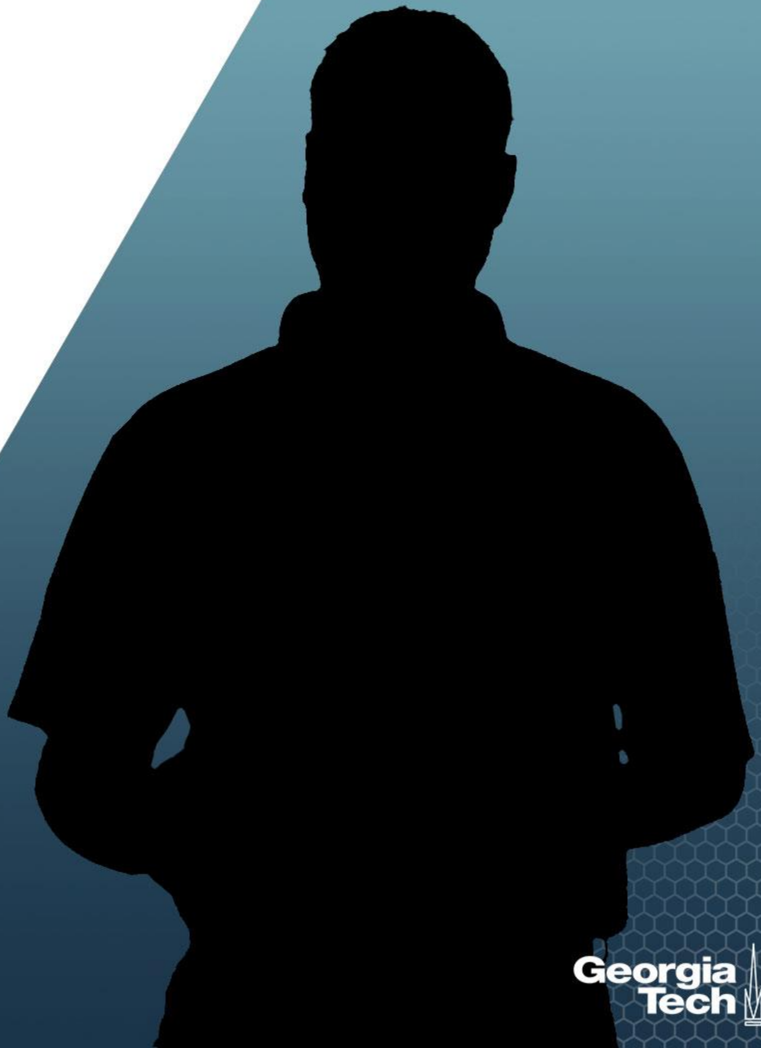
Summary



Regression Analysis

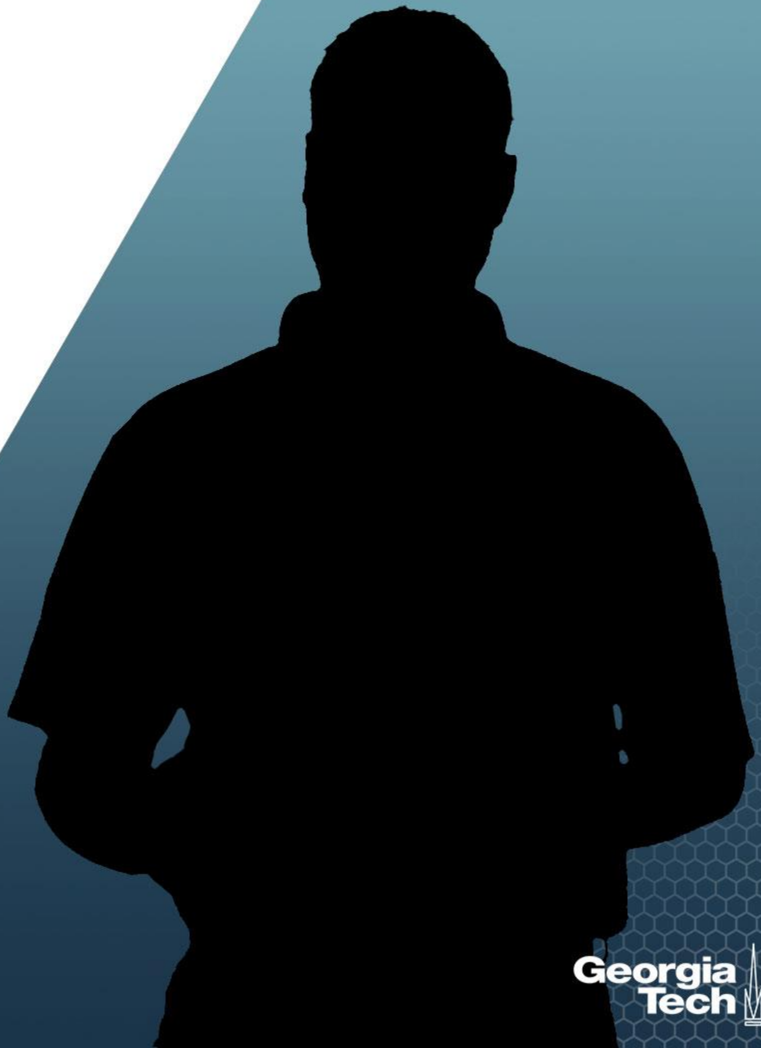
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Model Interpretation

About This Lesson



Model Interpretation: Parameters

The Least Squares estimated coefficients have specific interpretations:

$\widehat{\boldsymbol{\beta}}_0$ The estimated expected value of the response variable when all predicting variables equal zero.

$\widehat{\boldsymbol{\beta}}_i$ The estimated expected change in the value of the response variable associated with one unit of change in the value of the i^{th} predicting variable (i.e., associated with a one-unit change in $\boldsymbol{x}_i$, where i is any of $1, \ldots, p$), holding all other predictors in the model fixed (i.e., holding fixed $\boldsymbol{x}_j$ for $j = 1, \ldots, p$ where $j \neq i$).

Model 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 conditionally, *i.e., conditional of all other predicting variables in the model*.

The estimated regression coefficients for conditional and marginal relationships can differ not only in magnitude but also in sign or direction of the relationship.

Georgia Tech

Model 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: $COLGPA = 1.3 + 0.7(HSGPA) - 0.0003(SAT)$.

- **Incorrect Interpretation:** Higher values of SAT are associated with lower values of College GPA.
- **Correct Interpretation:** Higher values of SAT are associated with lower values of college GPA, ***holding high school GPA fixed***.

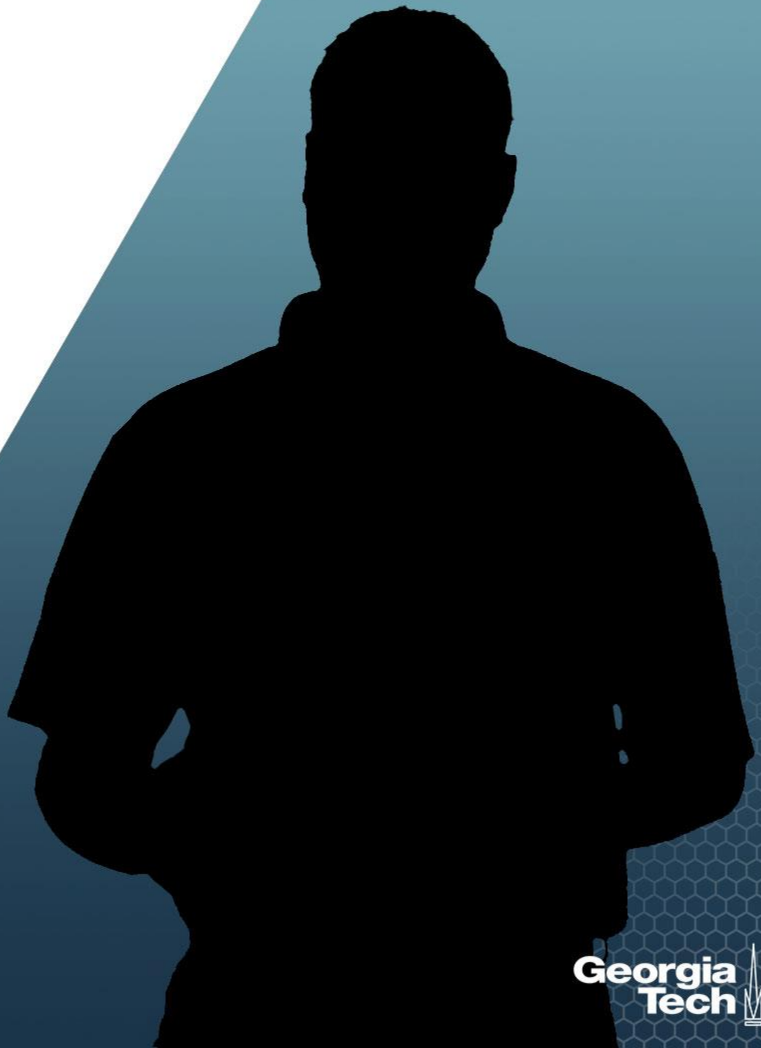
The coefficients of a multiple regression <u>must not</u> be interpreted marginally!

Different Roles of Predicting Variables

Predicting variables can be distinguished as:

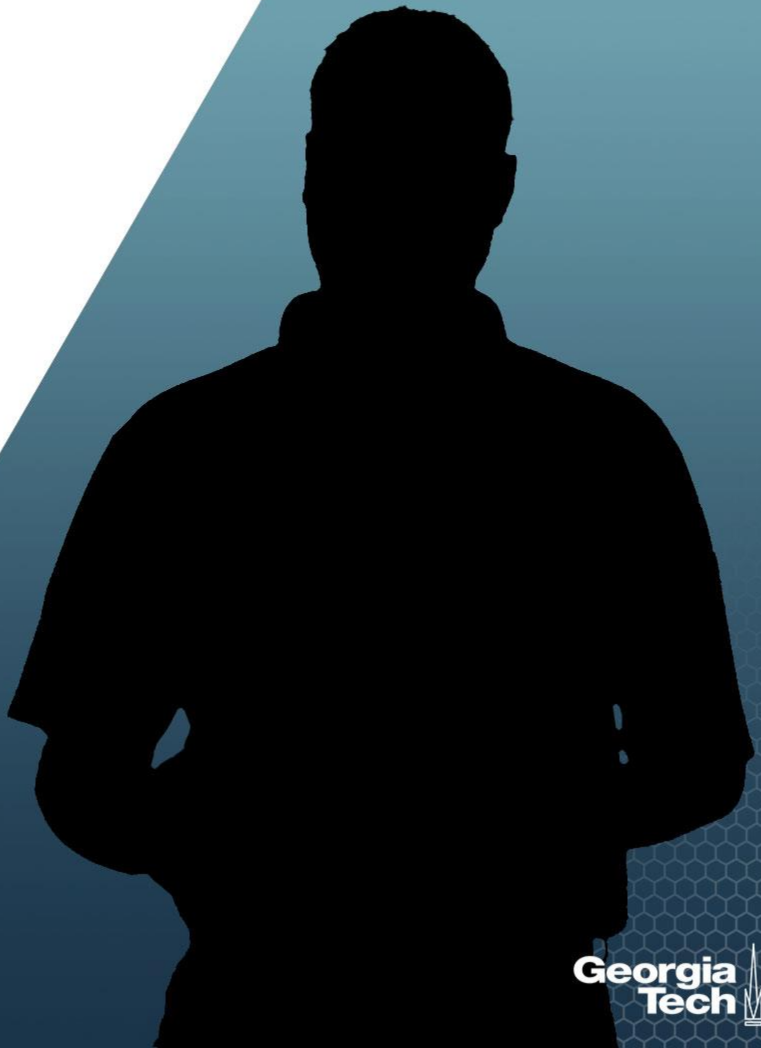
- ***Controlling*** – to control for bias selection in the sample. They are used as 'default' variables in order to capture more meaningful relationships.
- ***Explanatory*** – to explain variability in the response variable. They may be included in the model even if other "similar" variables are in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

About This Lesson



Regression Analysis

Multiple Linear Regression



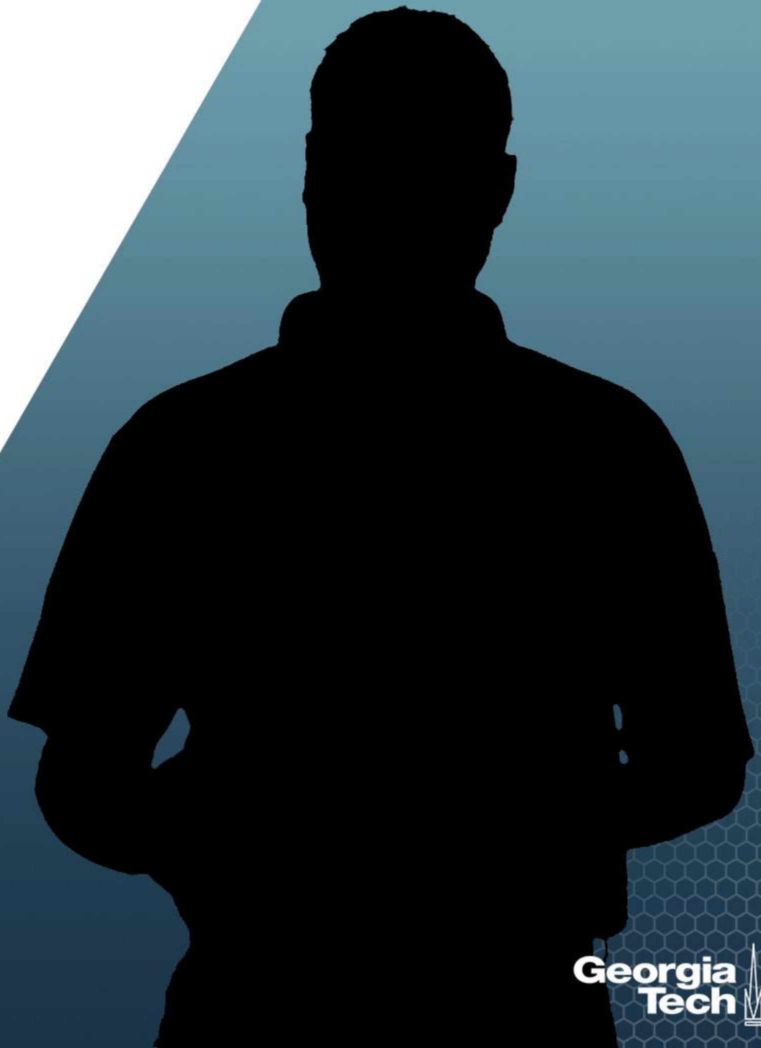
Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering

Regression Parameter Estimation: Data Example

About This Lesson



Georgia Tech

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 office is located (1 = south, 2 = west, 3 = midwest)

Response Variable:

Y = yearly sales (in thousands of dollars)

Georgia Tech

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. the marginal (one predictor) model.

c. What change does the full regression model predict for yearly sales as the advertisement expenditure increases by an additional $1,000? Is this prediction different when compared to that from the simple linear model with the advertisement expenditure variable only?

d. What is the estimate of the error variance under the full model? Is it different from that under 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
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

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
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

a. **Estimated Regression Coefficients**

b. **Conditional model**:

$\hat{\beta}_{adv} = \mathbf{1.4092}$

The expected additional gain in sales in thousands for $100 additional expenditure in advertisement **while holding all other fixed.**

Marginal model:

$\hat{\beta}_{adv} = \mathbf{2.772}$

The expected additional gain in sales in thousands for $100 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
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9407
F-statistic: 64.42 on 6 and 18 DF, p-value: 3.466e-11

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

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

d. Under the full model, the variance estimate is $(55.57)^2$. Under the simple linear model, the variance estimate was $(101.4)^2$.

Why? More variability in the response is explained when including multiple predicting variables versus only one.

Georgia Tech

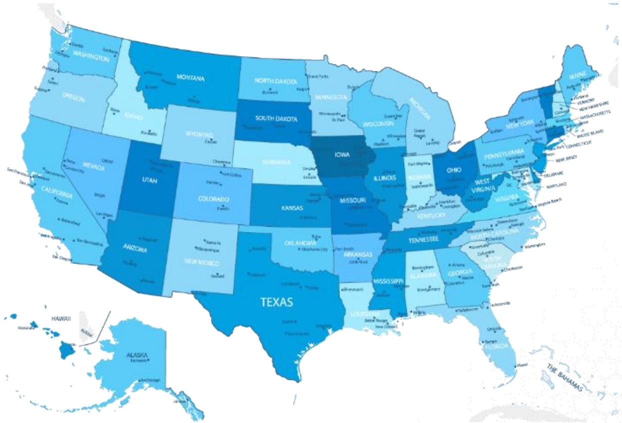
Linear Regression: Example 2

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

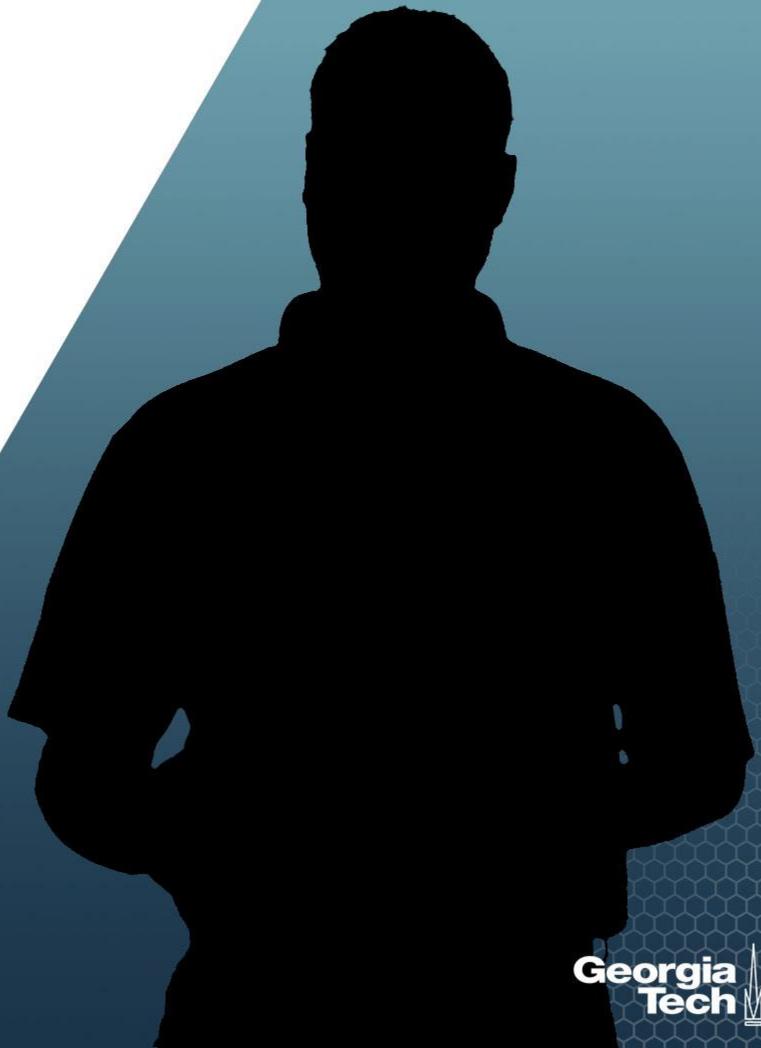
Controlling Factors:

- 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



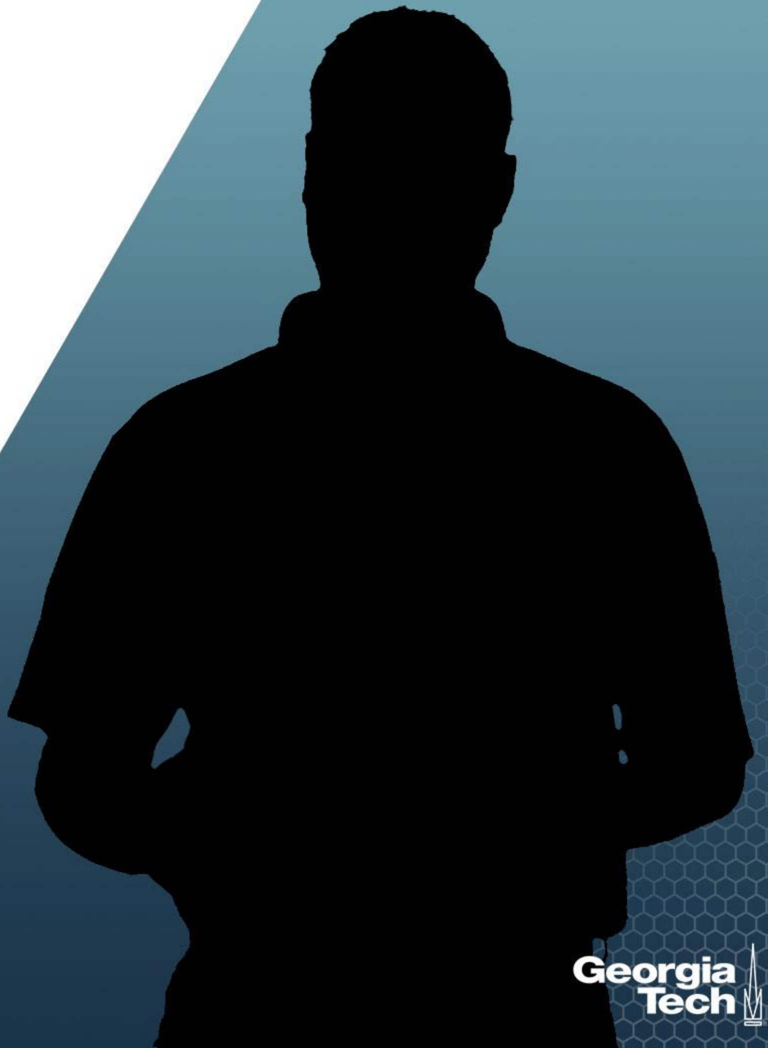
SAT Mean Score by State — Year 1982
790 (South Carolina) – 1088 (Iowa)

Summary



Regression Analysis

Multiple Linear Regression



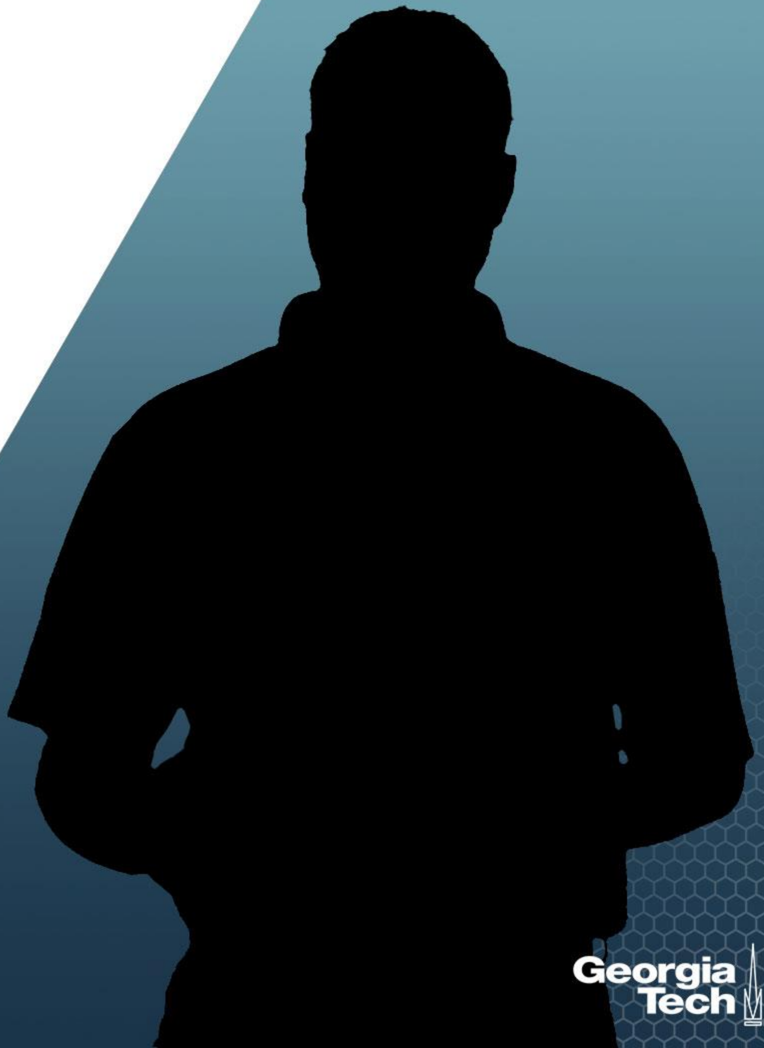
Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering

Inference for Regression Parameters

About This Lesson



Sampling Distribution

$$\mathrm{E}(\hat{\beta}) = \beta$$
$$\mathrm{V}(\hat{\beta}) = \sigma^2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} = \Sigma$$

Furthermore, $\widehat{\boldsymbol{\beta}}$ is a linear combination of $\{y_1,\ldots,y_n\}$. If we assume that $\varepsilon_i \sim$ Normal $(0, \sigma^2)$, then $\widehat{\boldsymbol{\beta}}$ is also distributed as

$$\widehat{\boldsymbol{\beta}} \sim \mathbf{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Properties of Regression Estimators

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

σ^2 is unknown!

Replace σ^2 with $\hat{\sigma}^{\mathbf{2}} = \mathrm{MSE}$

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-p-1} \sim \chi^2_{n-p-1}$$

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

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

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

Confidence Interval Estimation

We can derive confidence intervals for β_j using this *t* sampling distribution:

$$\boxed{\hat{\beta}_j \pm \left(t_{\alpha/2,\, n-p-1}\right)\left(\mathrm{SE}(\hat{\beta}_j)\right)}$$

Is $\boldsymbol{\beta_j}$ statistically significant?

- Check whether zero is in the confidence interval

Why is this a *t*-interval?

Georgia Tech

Confidence Interval Estimation

Why is this a *t*-interval?

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{V}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)} \sim \mathrm{T}_{n-p-1} \longrightarrow$$

t-interval for β_j

1-α Confidence Interval for β_j $\longrightarrow$

$$\underbrace{\hat{\beta}_j}_{\text{Estimate of } \beta_j} \pm \underbrace{\left(t_{\alpha/2,\, n-p-1}\right)}_{t\text{-critical point}} \underbrace{\left(\mathrm{SE}(\hat{\beta}_j)\right)}_{\text{Standard Deviation/Error of } \hat{\beta}_j}$$

Testing Statistical Significance

To test for statistical significance of β_j given all other predicting variables in the model, use a *t*-test for $\mathrm{H_0}$ and $\mathrm{H_a}$:

$$\boxed{\begin{gathered}\mathrm{H_0}\text{: } \beta_j = 0 \\ \text{vs.} \\ \mathrm{H_a}\text{: } \beta_j \neq 0\end{gathered}}$$

$$\boxed{t - \mathrm{value} = \frac{\hat{\beta}_j - 0}{\mathrm{SE}(\hat{\beta}_j)}}$$

- Reject $\mathrm{H_0}$ if |*t*-value| gets too large
- Interpret rejecting the null hypothesis as β_j being statistically significant

Georgia Tech

Testing Statistical Significance

How will the procedure change if

$$\mathrm{H_0}: \beta_j = b \quad \text{vs.} \quad \mathrm{H_a}: \beta_j \neq b$$

for some known null value b?

we test

$$\boxed{t - \text{value} = \frac{\hat{\beta}_j - b}{\mathrm{SE}(\hat{\beta}_j)}}$$

- Reject $\mathrm{H_0}$ if |*t*-value| is large
  - For significance level α, if $|t - \text{value}| > t_{\alpha/2,\, n-p-1}$ → reject $\mathrm{H_0}$
- Alternatively, compute a p-value based on the probability that the *t* distribution is greater than the absolute value of the *t*-value:

$$\text{p-value} = 2\text{Prob}(T_{n\text{-}p\text{-}1} > |t\text{-value}|)$$

- If p-value is small (e.g., < 0.01) → reject $\mathrm{H_0}$

Testing Statistical Significance

How will the procedure change if we test whether a coefficient is statistically positive or negative?

Test for Statistically Positive

$$\boxed{\begin{array}{c}\mathrm{H_0}:\ \beta_j \leq 0 \\ \text{vs.} \\ \mathrm{H_a}:\ \beta_j > 0\end{array}}$$

$$\text{p-value} = \mathrm{Prob}(\mathrm{T}_{n\text{-}p\text{-}1} > t\text{-value})$$

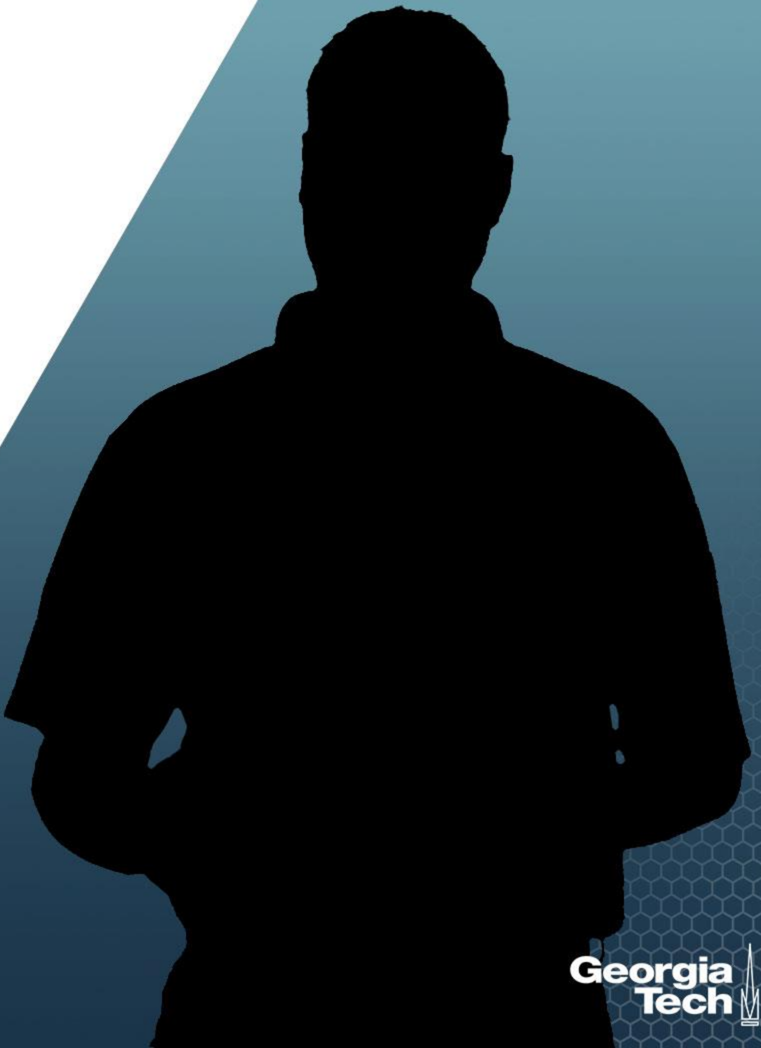
Test for Statistically Negative

$$\boxed{\begin{array}{c}\mathrm{H_0}:\ \beta_j \geq 0 \\ \text{vs.} \\ \mathrm{H_a}:\ \beta_j < 0\end{array}}$$

$$\text{p-value} = \mathrm{Prob}(\mathrm{T}_{n\text{-}p\text{-}1} < t\text{-value})$$

Georgia Tech

Summary

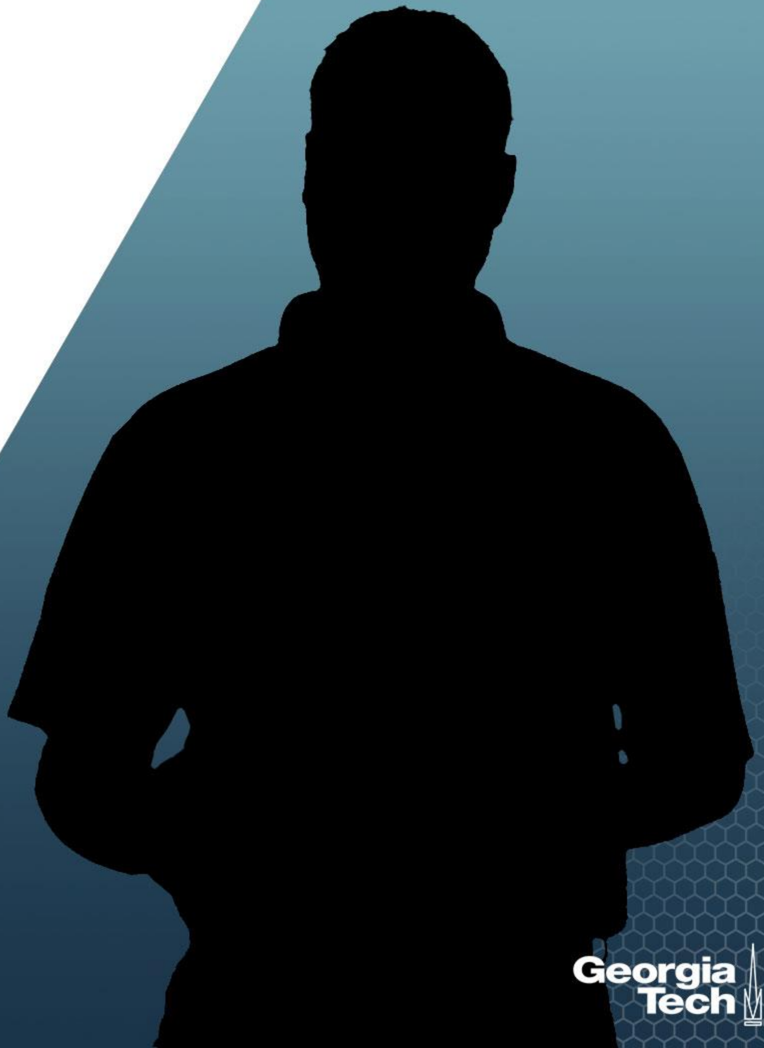


Regression Analysis

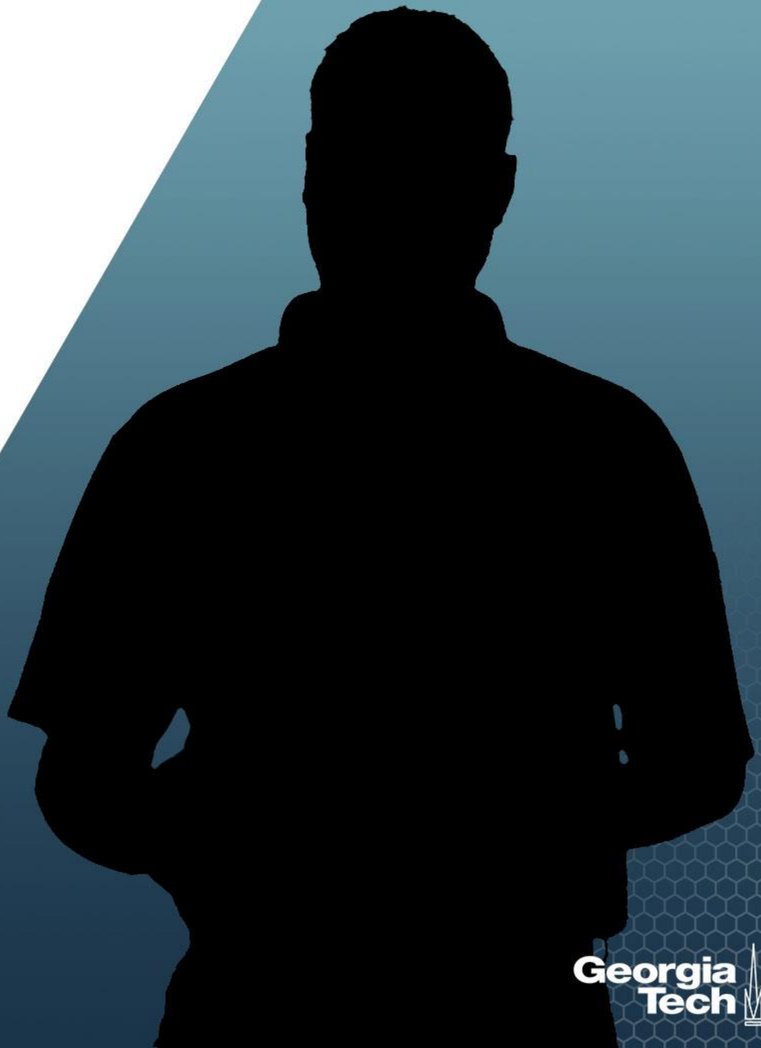
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Testing for Subsets of Regression Parameters



About This Lesson



Testing Overall Regression

Analysis of Variance (ANOVA) for multiple regression:

Variability Source	DF	Sum of Squares	Mean SS	F-Statistic
Regression	p	SSReg	SSReg / p	MSSReg / MSE
Residual	n-p-1	SSE	SSE / (n-p-1)	
Total	n-1	SST		

$$\mathrm{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Null hypothesis: All predictor coefficients are 0, i.e., $\mathbf{H_0}\text{:}\ \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Reject $\mathbf{H_0}$ if F-statistic is large (> $F_{\alpha,\, p,\, n\text{-}p\text{-}1}$ for α significance level, *p* and *n*-*p*-1 df).

- At least one of the coefficients is different from zero at the α significance level.

p-value = Prob($F_{p,\, n\text{-}p\text{-}1}$ > F-statistic) for F-distribution with *p* and *n*-*p*-1 df.

- Reject $\mathbf{H_0}$ if p-value is small.

Testing Subsets of Coefficients

Analysis of Variance (ANOVA):

$$\mathrm{SST}(X_1, \cdots, X_p) = \mathrm{SSReg}(X_1, \cdots, X_p) + \mathrm{SSE}(X_1, \cdots, X_p)$$

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

$\boldsymbol{SSReg(X_1)}$: Sum of squares (SS) explained using only X_1

$\boldsymbol{SSReg(X_2|X_1)}$: ***Extra*** SS explained using X_2 in addition to X_1

$\boldsymbol{SSReg(X_3|X_1, X_2)}$: ***Extra*** SS explained using X_3 in addition to X_1 and X_2

$\boldsymbol{SSReg(X_p|X_1, \cdots, X_{p-1})}$: ***Extra*** SS explained using X_p in addition to X_1 , X_2 … X_{p-1}

Georgia Tech

Testing Subsets of Coefficients

- Does X_1 alone significantly aid in predicting Y?
  - SSReg(X_1) vs. SSE(X_1)
- Does the addition of X_2 significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1?
  - SSReg(X_2 | X_1) vs. SSE(X_1, X_2)
- Does the addition of X_3 significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1 and X_2?
  - SSReg(X_3 | X_1, X_2) vs. SSE(X_1, X_2, X_3)
- Does the addition of X_p significantly contribute to the prediction of Y after accounting (controlling) for the contribution of X_1,..., X_{p-1}?
  - SSReg(X_p | X_1, ..., X_{p-1}) vs. SSE(X_1, X_2, ..., X_p)

Georgia Tech

Testing Subsets of Coefficients

Partial F-test:

- Consider a full model with two sets of predictors, X_1, ..., X_p (perhaps controlling factors) and $(Z_1$, ..., $Z_q)$ (perhaps additional explanatory factors):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha_1 Z_1 + \cdots + \alpha_q Z_q + \varepsilon$$

- Test whether any of the Z factors add explanatory power to the model:

$$\mathbf{H_0}\text{: } \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \quad \textbf{vs.} \quad \mathbf{H_a}\text{: } \alpha_{\mathrm{i}} \neq 0 \text{ for at least one } \alpha_{\mathrm{i}}, \mathrm{i} = 1, \cdots, q$$

$$\mathrm{F} - \text{statistic} = \mathrm{F}_{partial}$$
$$= \frac{\mathrm{SSReg}(Z_1, \ldots, Z_q | X_1, \ldots, X_p)/q}{\left(\mathrm{SSE}(Z_1, \ldots, Z_q, X_1, \ldots, X_p)/(n - p - q - 1\right)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large ($\text{F-statistic} > \mathrm{F}_{\alpha,\, q,\, n\text{-}p\text{-}q\text{-}1}$)
  - At least one coefficient is different from zero at the α significance level

Testing for Statistical Significance

- Consider a full model with the set of predictors, X_1, ..., X_p and an additional predicting variable Z:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha Z + \varepsilon$$

- Test whether Z has explanatory or predictive power:

$$\mathbf{H_0}:\ \alpha = 0 \text{ vs } \mathbf{H_a}:\ \alpha \neq 0$$

$$\mathrm{F-statistic} = \mathrm{F}_{partial} = \frac{\mathrm{SSReg}(\mathrm{Z}|X_1, \ldots, X_p)/1}{\left(\mathrm{SSE}(Z, X_1, \ldots, X_p)/(n-p-2\right)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large ($\text{F-statistic} > \mathrm{F}_{\alpha,\,1,\,n\text{-}p\text{-}2}$)

This is equivalent to testing for statistical significance using the t-test

Testing for Statistical Significance

- Consider a full model with the set of predictors, X_1, ..., X_p and an additional predicting variable Z:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \alpha Z + \varepsilon$$

- Test whether Z has explanatory or predictive power:

$$\mathbf{H_0}\colon \alpha = 0 \text{ vs } \mathbf{H_a}\colon \alpha \neq 0$$

$$\mathrm{F-statistic} = \mathrm{F}_{partial} = \frac{\mathrm{SSReg}(\mathrm{Z}|X_1, \dots, X_p)/1}{\left(\mathrm{SSE}(Z, X_1, \dots, X_p)/(n-p-2)\right)}$$

- Reject $\mathbf{H_0}$ if F-statistic is large ($\text{F-statistic} > \mathrm{F}_{\alpha,\,1,\,n\text{-}p\text{-}2}$)

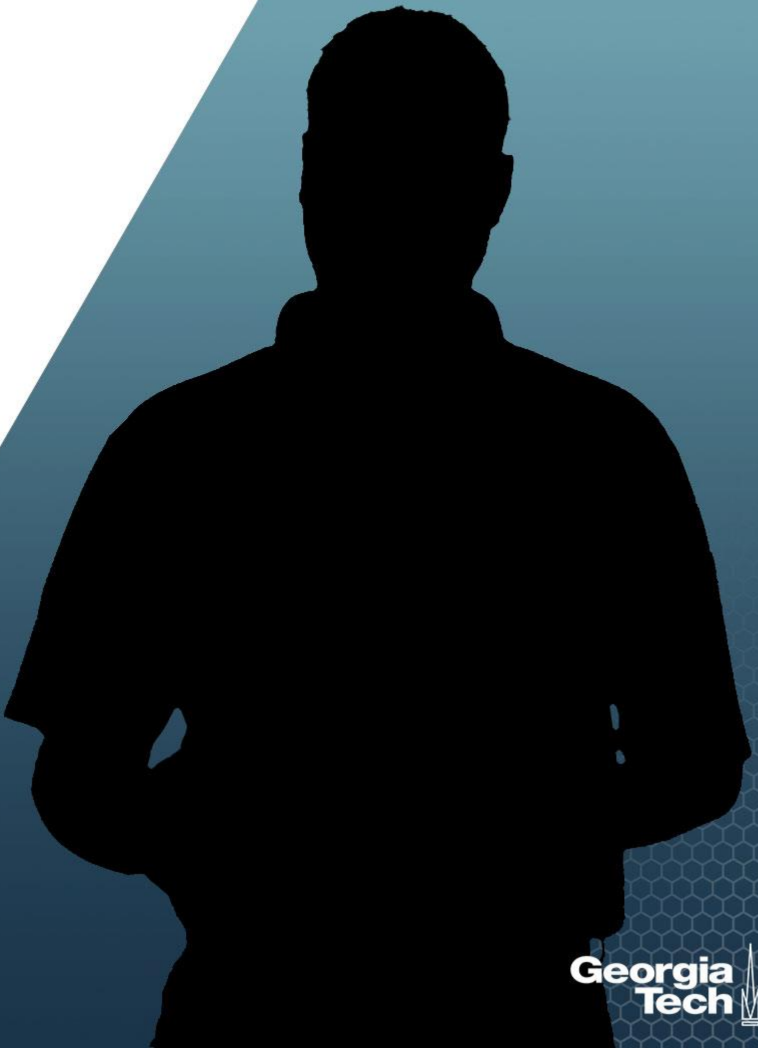
This is equivalent to testing for statistical significance using the t-test

- Interpretation of the t-test for statistical significance is conditional on other predicting variables being in the model.
- The relationship between Y and X is statistically significant given all other predicting variables being in the model.

Do not perform variable selection based on the p-values of the t-tests!

Georgia Tech

Summary



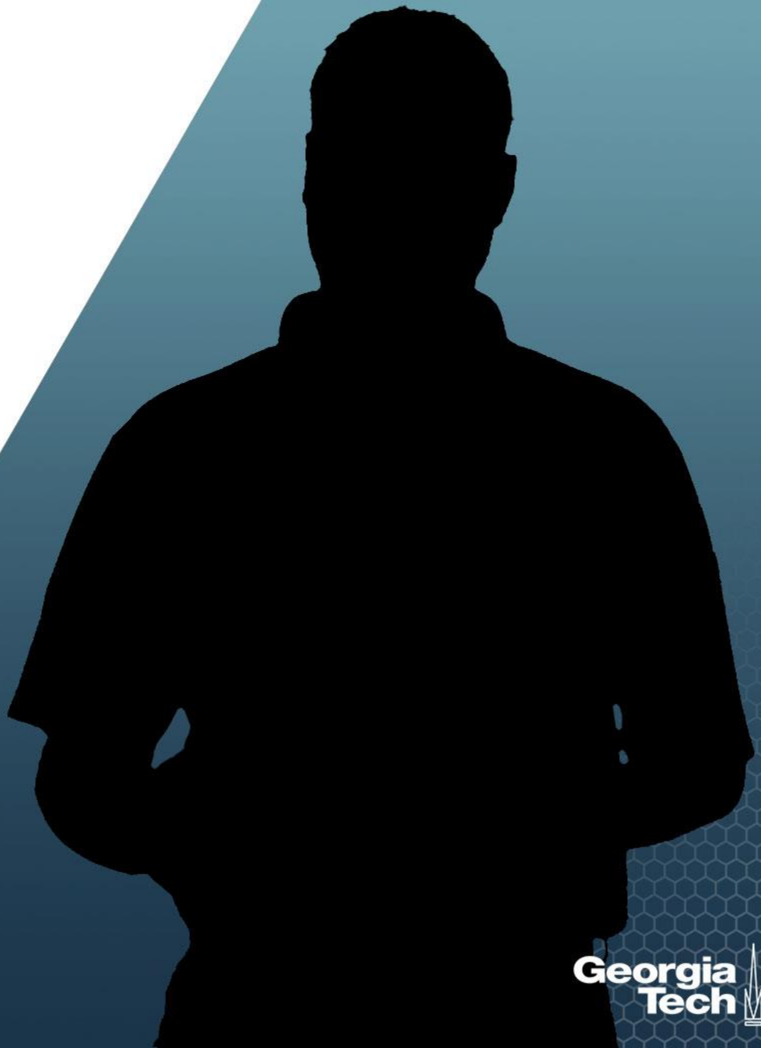
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.

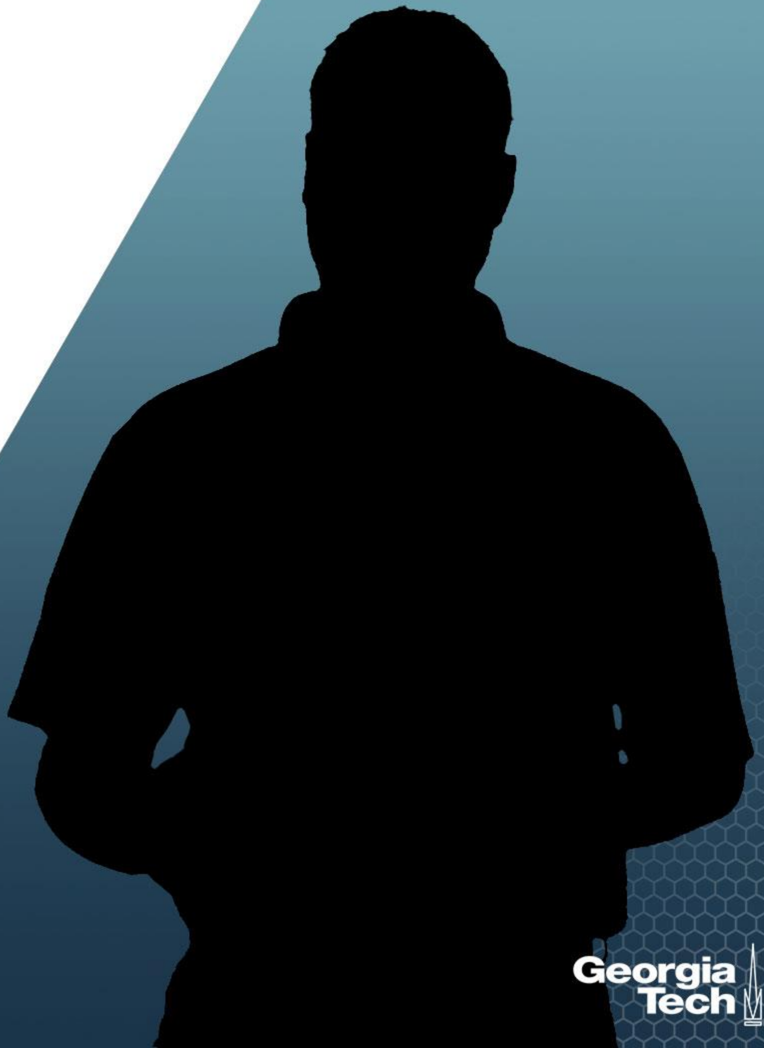
Professor

School of Industrial and Systems Engineering



Statistical Inference:
Data Example

About This Lesson



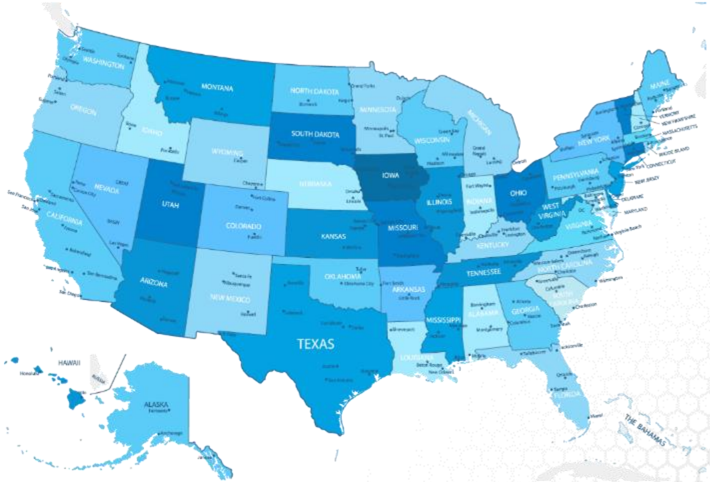
Linear Regression: Example 2

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



Example 2: Inference on Coefficients

a. What is the estimate of the coefficient β_1 and its variance? Interpret. What is its sampling distribution?

b. Is the coefficient β_1 statistically significant? What is the p-value of the test. Interpret.

c. What is the F-statistic for overall regression? Do we reject the null hypothesis that all regression coefficients are zero?

d. Obtain the 99% confidence interval for β_1.

e. Given the controlling factors, test the null hypothesis that the coefficients of the other variables 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	0.957353
years	22.610082	6.314577	3.581	0.000866 ***
public	-0.464152	0.579104	-0.802	0.427249
expend	2.212005	0.845972	2.615	0.012263 *

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

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

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	0.957353
years	22.610082	6.314577	3.581	0.000866 ***
public	-0.464152	0.579104	-0.802	0.427249
expend	2.212005	0.845972	2.615	0.012263 *

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

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

a. **Estimation and distribution**:
$\hat{\beta}_{takers} = \boxed{-0.480}$
$se(\hat{\beta}_{takers}) = \boxed{0.693}$
t-dist. with $\boxed{43}$ degrees of freedom

b. **Test for statistical significance**:
$\hat{\beta}_{takers}$: t-value = **-0.692**
p-value **> 0.1**

c. **Test for overall regression:**
F-value = **51.91**
p-value **≈ 0**

Georgia Tech

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:

β_{takers}: **[-2.349701, 1.389541]**

Interpretation: The interval includes zero, thus it is plausible that the regression coefficient to be zero given all 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 ***

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 coefficients**:

H_0: Reduced Model (*takers* and *rank* only) vs. H_A: Full Model

Partial F Test:

F-value = **8.6221**

P-value **≈ 0**

Georgia Tech

Example 2: Inference on Coefficients

e. **Testing for a subset of regression coefficients *(continued)***:

Test $\mathrm{H}_0: \beta_{income} = \beta_{years} = \beta_{public} = \beta_{expend} = 0$

How was the F-statistic computed?

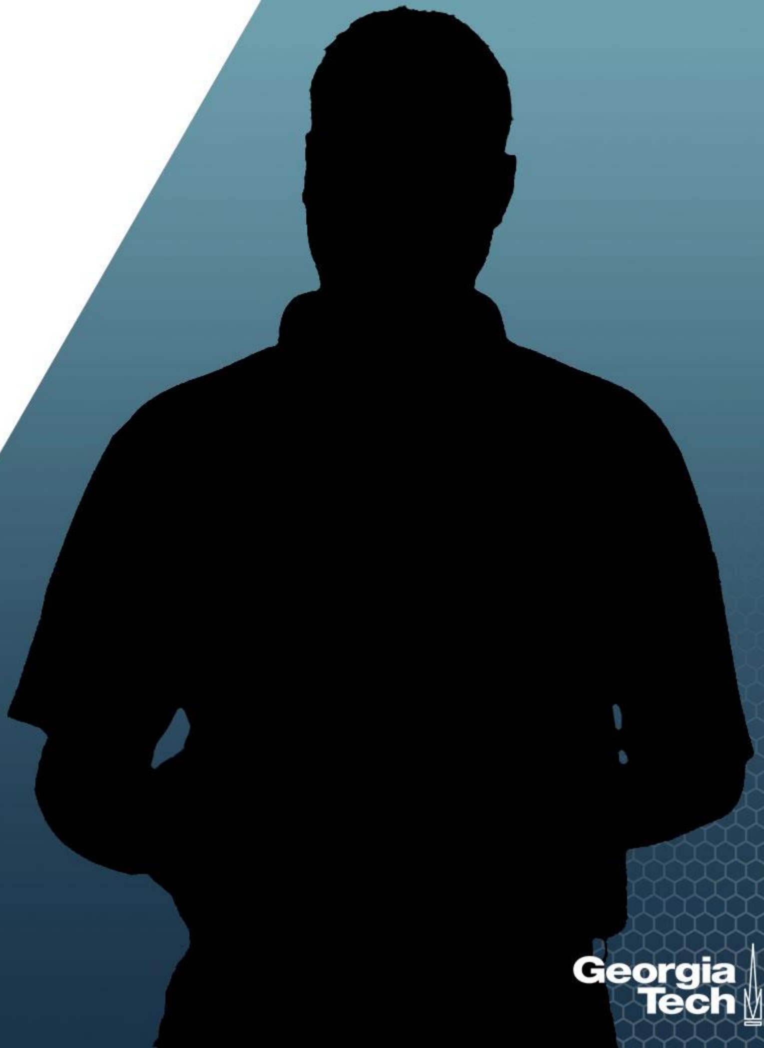
$$\text{F-statistic} = \frac{\mathrm{SSReg}(income, years, public, expend \mid takers, rank)/4}{\mathrm{SSE}/(50-6-1)}$$

The p-value is computed as

$$\mathrm{Prob}\left(\mathrm{F}_{4,43} > \mathrm{F}-\mathrm{statistic}\right) = 1 - \mathrm{Prob}\left(\mathrm{F}_{4,43} < \mathrm{F}-\mathrm{statistic}\right)$$

Interpretation: The p-value is approximately 0, so reject the null hypothesis. We conclude that at least one predictor among *income*, *years*, *public* and *expend* will be significantly associated with states' average SAT scores.

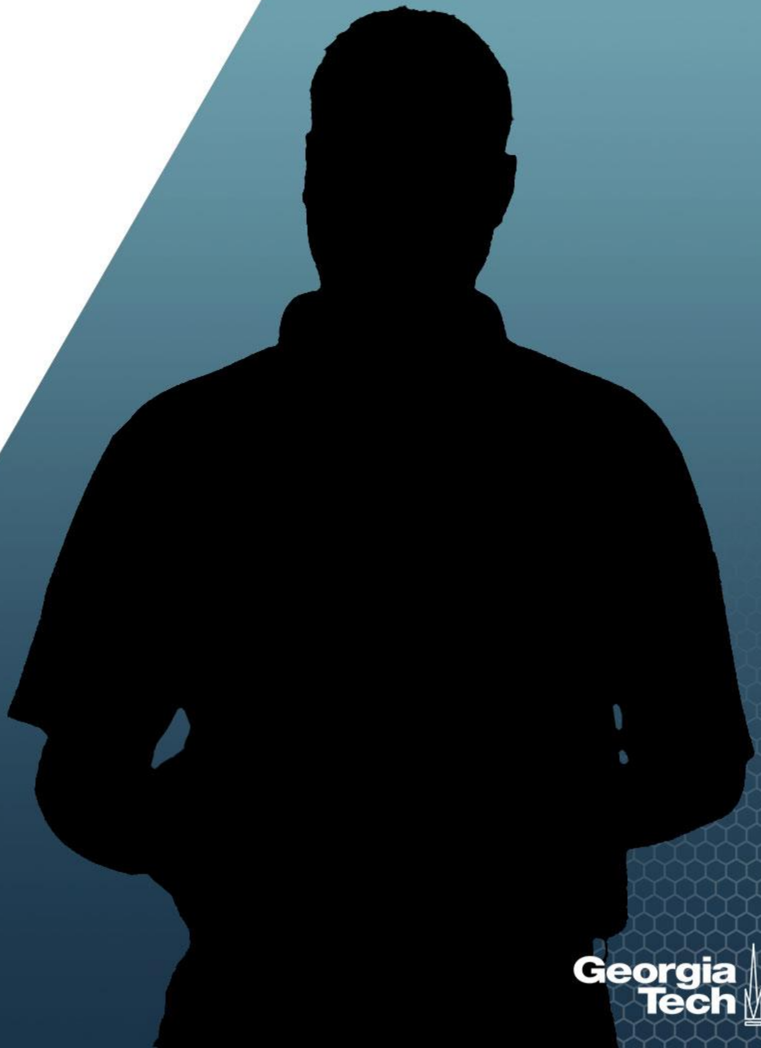
Summary



Regression Analysis

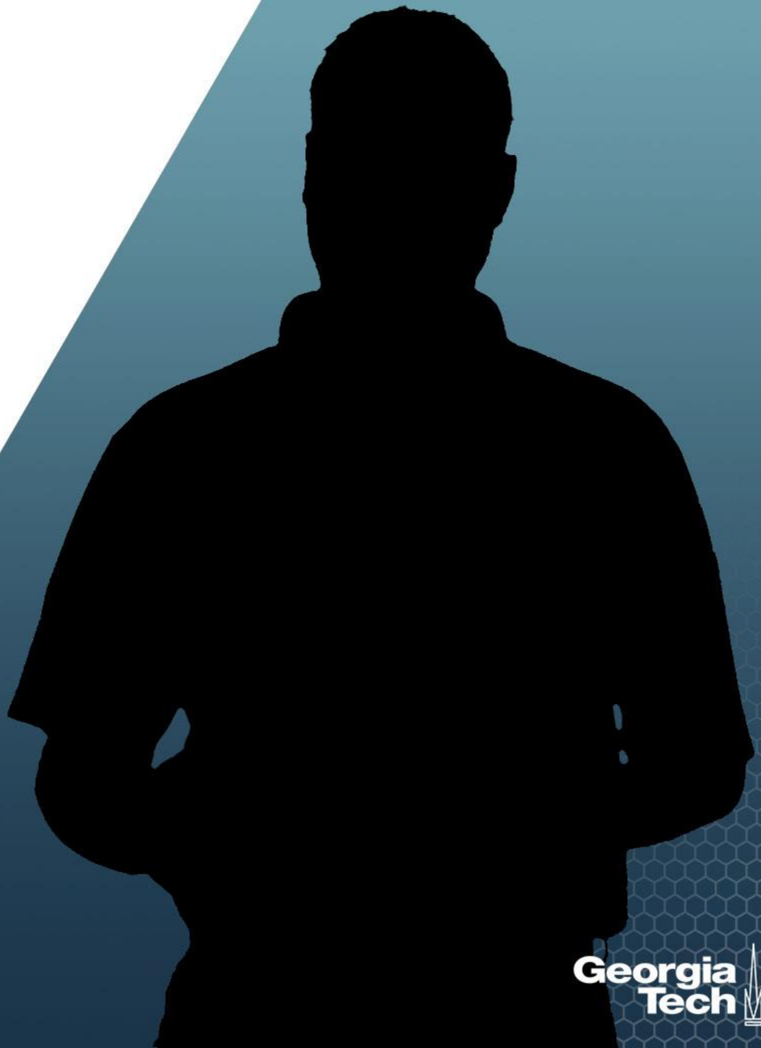
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Estimating the Regression Line and Predicting a New Response

About This Lesson



Estimating the Regression Line

At some selected value of $\boldsymbol{x}$, say $\boldsymbol{x}^*$, estimate the "mean response" of y (the regression line) via

$$\boxed{\hat{Y}|x^* = \hat{\beta}_0 + \hat{\beta}_1 {x^*}_1 + \hat{\beta}_2 {x^*}_2 + \cdots + \hat{\beta}_p {x^*}_p = {\boldsymbol{x}^*}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}$$

- Because the estimators of $\boldsymbol{\beta}$ are normally distributed, so is $\hat{Y}$.
- If we know the expected value and variance, we can use the normal distribution of $\hat{Y}$ to draw inferences on the regression line.

Estimating the Regression Line

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

$$\mathrm{E}\left(\hat{Y}|\boldsymbol{x}^*\right) = \boldsymbol{x}^{*\boldsymbol{T}}\boldsymbol{\beta} = \beta_0 + \beta_1 x^*{}_1 + \beta_2 x^*{}_2 + \cdots + \beta_p x^*{}_p$$

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

If we replace the unknown variance with its estimator, $\hat{\sigma}^2$ = MSE, the sampling distribution becomes a *t*-distribution with *n*-*p*-1 degrees of freedom.

Confidence Interval for Regression Line

The (1 - α) **Confidence Interval** for the *mean response* (or regression line) for <u>one</u> instance of predicting variables $\boldsymbol{x}$ is:

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

The (1 - α) **Confidence Surface** for all possible instances of the predicting variables is:

$$\boxed{\hat{y}|\boldsymbol{x} \pm \sqrt{(p+1)\mathrm{F}_{\alpha,\, p+1,\, n-p-1}} \sqrt{\hat{\sigma}^2 \boldsymbol{x}^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{x}}}$$

Predicting a New Response

- One of the primary motivations for regression is to use the regression equation to predict future responses.
- The predicted regression line is the same as the estimated regression line.
- But a prediction is not the same as the regression line estimation. The prediction contains *two* sources of uncertainty:
  - From the parameter estimates (of βs)
  - From the new observation(s)

Predicting a New Response *(cont'd)*

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

The new observation is independent of the regression data, so the total variation in predicting $\boldsymbol{y}|\boldsymbol{x}$ is

$$\mathrm{Var}\left(\hat{Y}|\boldsymbol{x}^*\right) = \sigma^2\, \boldsymbol{x}^{*\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}^* + \sigma^2 = \hat{\sigma}^2\left(1 + \boldsymbol{x}^{*\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*\right)$$

Predicting a New Response *(cont'd)*

The (1 - α) **Prediction Interval** for one new (future) y (at $\boldsymbol{x}$) is

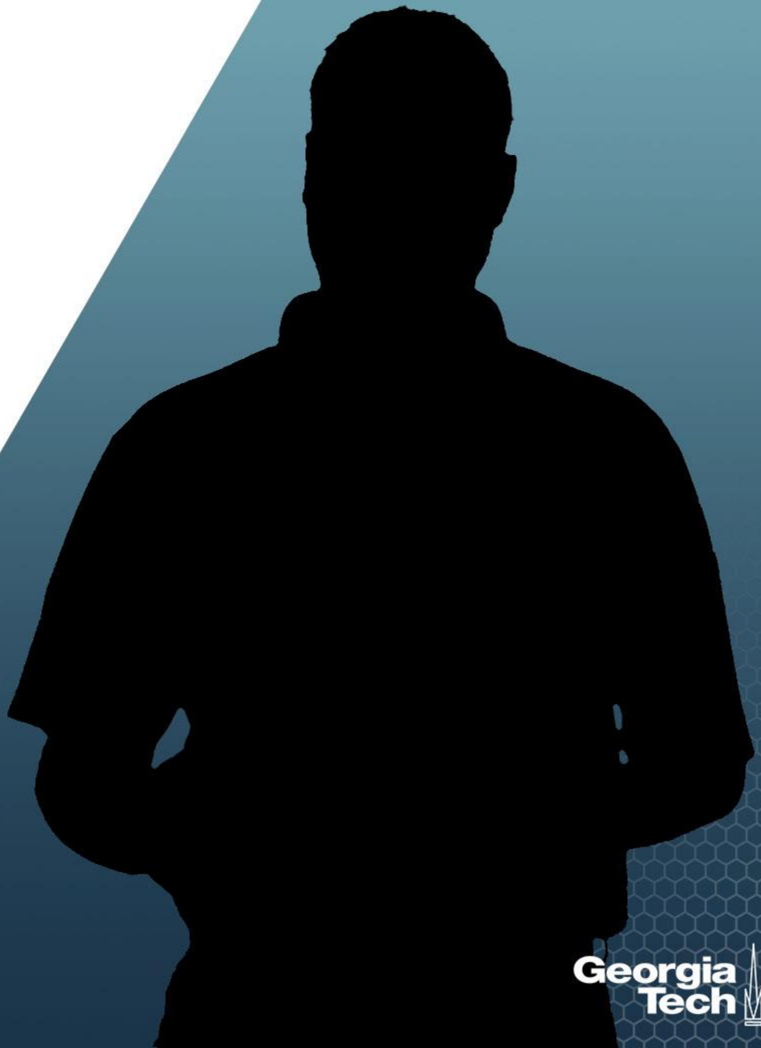
$$\boldsymbol{x}^{*\mathrm{T}}\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2,\, n-p-1} \sqrt{\hat{\sigma}^2\left(1 + \boldsymbol{x}^{*\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}^*\right)}$$

$\hat{y} = \boldsymbol{x}^{*\mathrm{T}}\widehat{\boldsymbol{\beta}}$ is the same as the line estimate, but the *Prediction Interval* is wider than the *Confidence Interval* for the mean response.

The (1 - α) **Prediction Interval** for *m* new (future) ys (at $\boldsymbol{x}^*$) is

$$\hat{y}|\boldsymbol{x}^* \pm \sqrt{m\mathrm{F}_{\alpha,\, m,\, n-p-1}} \sqrt{\hat{\sigma}^2\left(1 + \boldsymbol{x}^{*\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}^*\right)}$$

Summary



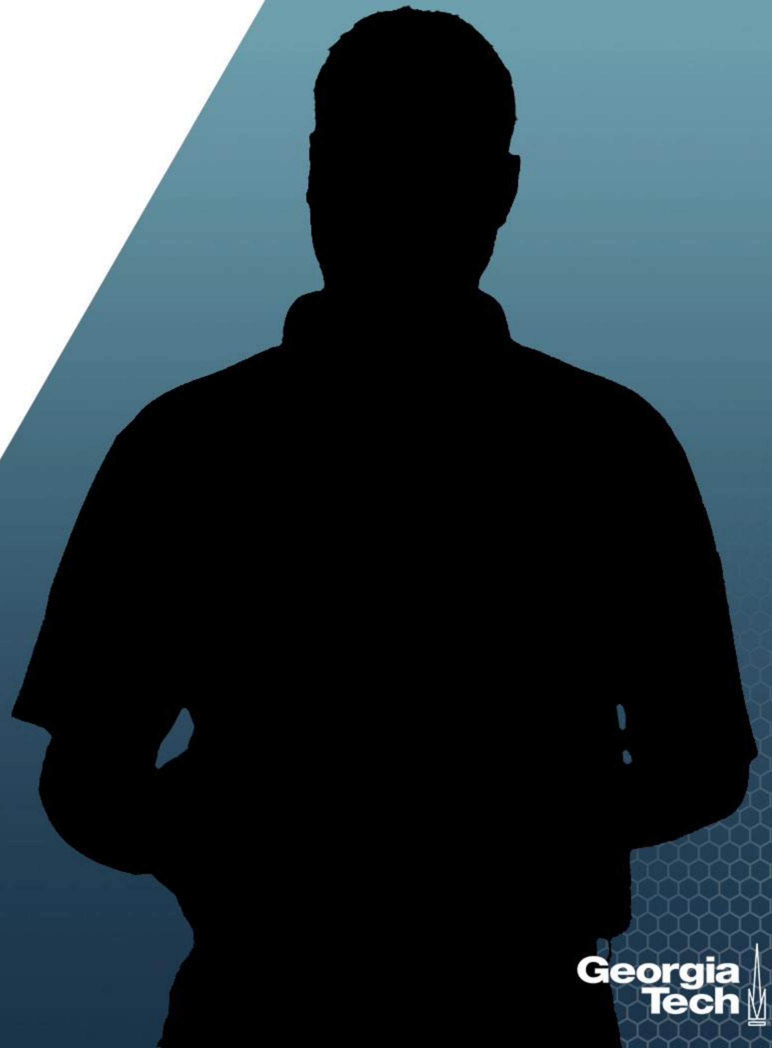
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.

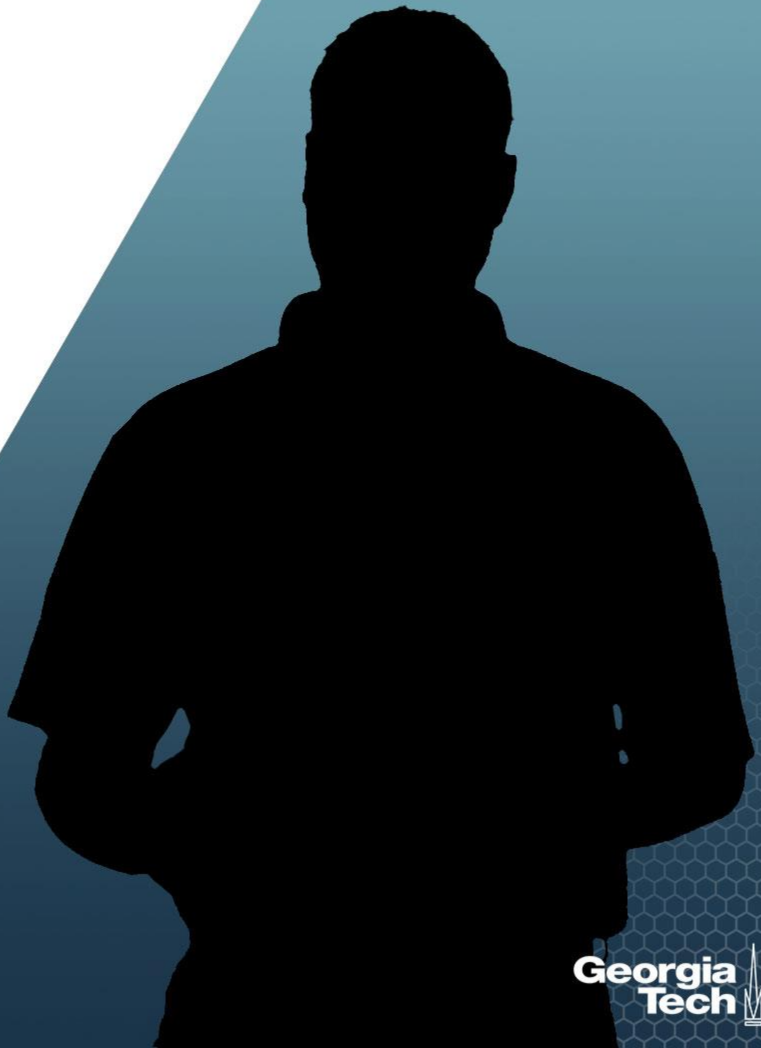
Professor

School of Industrial and Systems Engineering



Estimating Regression Line & Predicting a New Response: Data Example

About This Lesson



Linear Regression: Example 1

Quantitative Predicting Variables:

- X_1 = amount (in hundreds of dollars) spent on advertising in 1999
- X_2 = total amount of bonuses paid in 1999
- X_3 = market share in each territory
- X_4 = largest competitor's sales (thousands)

Qualitative Predicting Variable:

- X_5 = indicates region where office is located (1 = south, 2 = west, 3 = midwest)

Response Variable:

- Y = yearly sales (in thousands of dollars)

Example 1: Mean Response & Prediction

a. For <u>all</u> offices with the characteristics such as those of the first office:
   - What is the <u>average estimated sales</u>?
   - What is the standard deviation?
   - What is the 95% confidence interval for this mean response?

b. If the first office's largest competitor's sales increase to $303,000 (assuming everything else fixed):
   - What sales would you predict for the first office?
   - What is its standard deviation?
   - What is the 95% prediction interval for this prediction?

Example 1: Mean Response Estimation

```
s2 = summary(model)$sigma^2 # Variance estimate
X = model.matrix(model) # Design Matrix
xstar = X[1,] # First office data for formula
resp.var = s2*(xstar%*%solve(t(X)%*%X)%*%xstar) # Variance formula
sqrt(resp.var)

         [,1]
[1,] 33.19118

newdata = meddcor[1,-1] # First office data for confidence interval
predict(model, newdata, interval="confidence") # Confidence Interval

       fit      lwr      upr
1 934.7767 865.0446 1004.509
```

Example 1: Mean Response Estimation

```
s2 = summary(model)$sigma^2 # Variance estimate
X = model.matrix(model) # Design Matrix
xstar = X[1,] # First office data for formula
resp.var = s2*(xstar%*%solve(t(X)%*%X)%*%xstar) # Variance formula
sqrt(resp.var)

         [,1]
[1,] 33.19118

newdata = meddcor[1,-1] # First office data for confidence interval
predict(model, newdata, interval="confidence") # Confidence Interval

       fit      lwr      upr
1 934.7767 865.0446 1004.509
```

a. **Average estimated sales (mean response for sales)**:

$$\hat{y} = \mathbf{934.777}$$

Estimated standard deviation:

$$\mathrm{se}(\hat{y}) = \mathbf{33.191}$$

95% Confidence Interval:

(865.045, 1004.509)

Interpretation: For offices with the same characteristics as the first, the average estimated sales are \$934,777, with a lower bound of \$865,045 and an upper bound of \$1,004,509.

Georgia Tech

Example 1: Mean Response Estimation

```
## Change the competitor's sales for prediction of future observation
xstar.new = xstar
xstar.new[5] = 303

# Variance formula
pred.var = s2*(1+xstar.new%*%solve(t(X)%*%X)%*%xstar.new)
sqrt(pred.var)

         [,1]
[1,] 64.31099

# Prediction Interval
predict(model, xstar.new[-1], interval="prediction")

       fit      lwr      upr
1 911.0569 775.9446 1046.169
```

Example 1: Prediction

Change the competitor's sales for prediction of future observation

```
xstar.new = xstar
xstar.new[5] = 303
```

 # Variance formula

```
pred.var = s2*(1+xstar.new%*%solve(t(X)%*%X)%*%xstar.new)
sqrt(pred.var)
```

```
      [,1]
[1,] 64.31099
```

 # Prediction Interval

```
predict(model, xstar.new[-1], interval="prediction")
```

```
       fit      lwr      upr
1 911.0569 775.9446 1046.169
```

b. Predicted sales of the first office given the higher competitor's sales:

$$\hat{y} = \mathbf{911.057}$$

Estimated standard deviation:

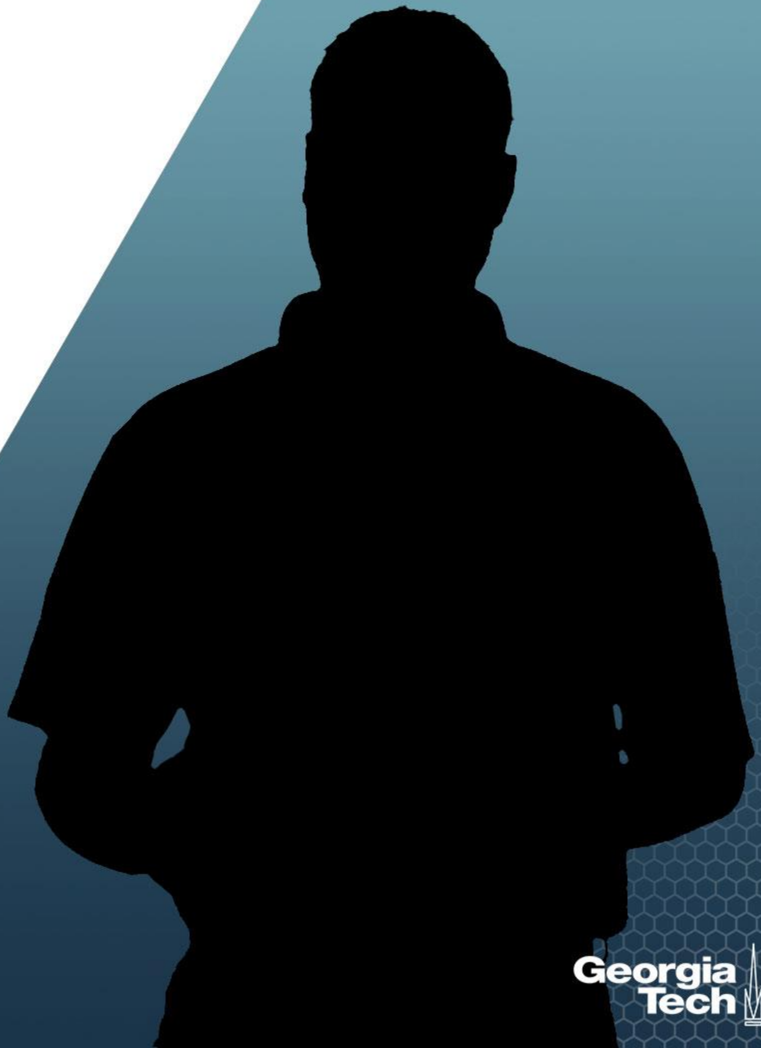
$$\text{se}(\hat{y}) = \mathbf{64.311}$$

95% Confidence Interval:

$$\boxed{\mathbf{(775.945,\ 1046.169)}}$$

Interpretation: If the competitor's sales increase to \$303,000 (from \$202,220), the predicted sales reduce by \$23,720 (from \$934,777 to \$911,057). Since this is prediction, the standard deviation increases.

Summary



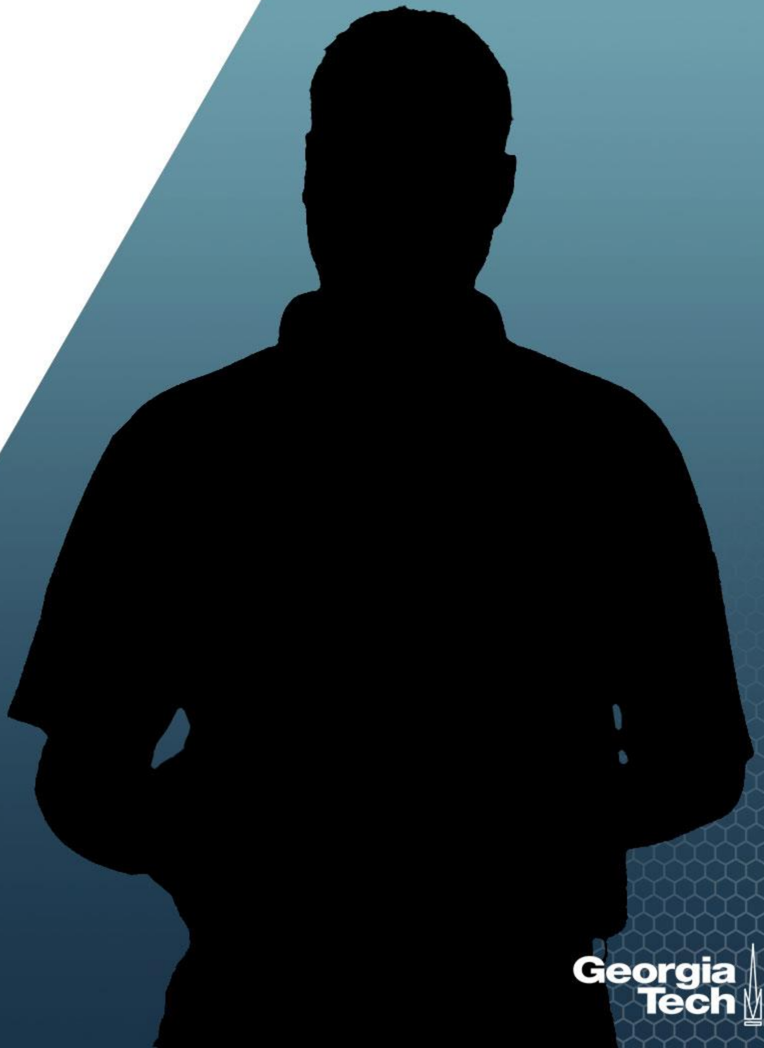
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Assumptions and Diagnostics

About This Lesson



Multiple Linear Regression: Model

Data: $\{(x_{1,1}, \dots, x_{1,p}), y_1\}, \dots, \{(x_{n,1}, \dots, x_{n,p}), y_n\}$

Model: $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,\ \ i = 1, \dots, n$

Assumptions:

- *Linearity/Mean Zero Assumption*: The relationship between the response variable and each predicting variable is linear. (For each j, $j = 1, \dots, p$, y_i and x_{ij} are linearly related, $i = 1, \dots, n$.) $\mathrm{E}(\varepsilon_i) = 0$
- *Constant Variance Assumption*: $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- *Independence Assumption*: $\{\varepsilon_1, \dots, \varepsilon_n\}$ are independent random variables
- Assumption that ε_i ~ Normal *for confidence/prediction intervals, hypothesis testing*

Properties of the Errors & Residuals

Properties of (true) errors:

- $\mathrm{E}(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathrm{I}$

Properties of the (estimated) residuals:

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
- $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ (or $\mathrm{E}(\hat{\varepsilon}_i) = 0$)
- $\mathbf{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (or $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2\left(1 - h_{i,i}\right)$
  - Where $\mathbf{H}$ is the hat matrix, and $h_{i,i}$ is the *i*-th element on its diagonal

Properties of the Errors & Residuals

Properties of (true) errors:

- $\mathrm{E}(\varepsilon_i) = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$
- $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{I}$

Properties of the (estimated) residuals:

- $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
- $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ (or $\mathrm{E}(\hat{\varepsilon}_i) = 0$)
- $\mathbf{V}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$ (or $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2\left(1 - h_{i,i}\right)$
  - Where $\mathbf{H}$ is the hat matrix, and $h_{i,i}$ 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:**

$$r_i = \hat{\varepsilon}_i / \left(\widehat{\sigma}\sqrt{1 - h_{i,i}}\right)$$

Residuals Analysis

Standardized Residual Values: $r_i = \hat{\varepsilon}_i / \left(\widehat{\sigma}\sqrt{1 - h_{i,i}}\right)$

Graphical assessment of MLR assumptions:

- Plot standardized residuals r_i against each predictor
  - *Linearity*
- Plot standardized residuals r_i against fitted values
  - *Constant Variance & Independence*
- QQ normal plot & histogram
  - *Normality*

Residuals Analysis

Standardized Residual Values: $r_i = \hat{\varepsilon}_i / \left(\widehat{\sigma}\sqrt{1 - h_{i,i}}\right)$

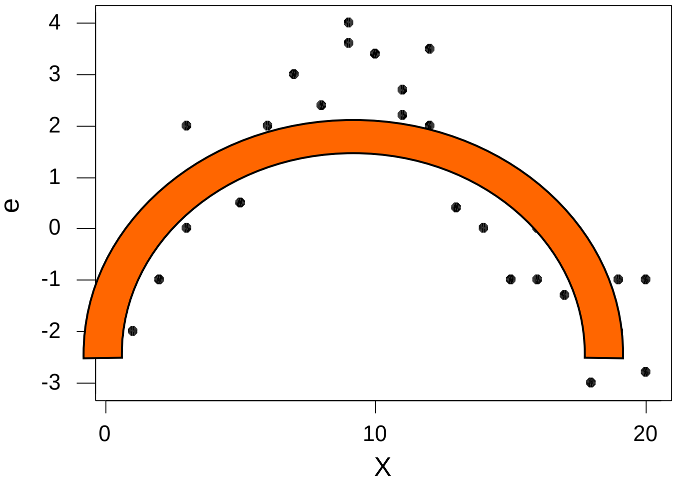
Graphical assessment of MLR assumptions:

- Plot standardized residuals r_i against each predictor
  - *Linearity*
- Plot standardized residuals r_i against 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

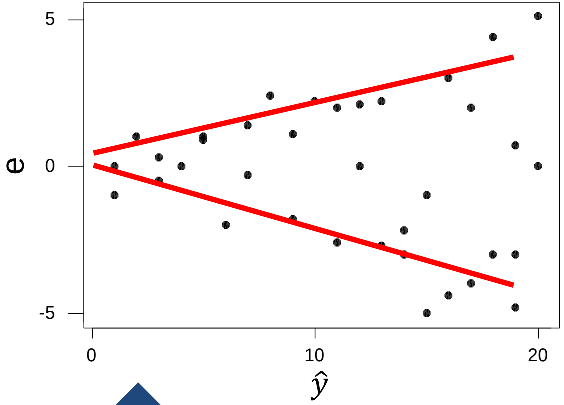
Linearity: Plot the residuals against each predicting variable.



This shows that there may be a non-linear relationship between *X* and *Y*.

Residual Analysis: Constant Variance Assumption

Constant Variance: Plot the residuals against fitted values.



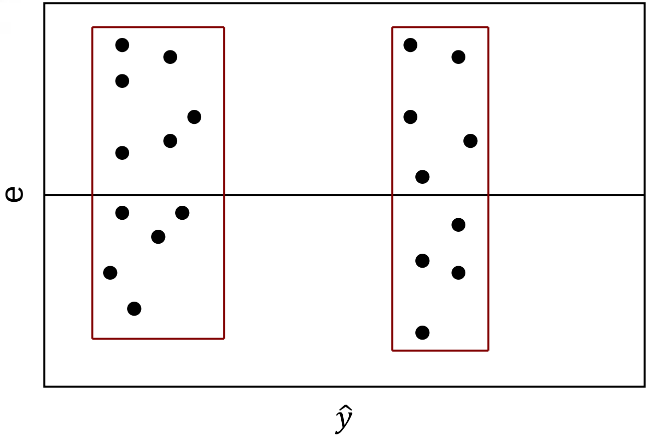
The residuals show larger variance as the fitted values increase.

Here, it is an example for which σ^2 is not constant.

Georgia Tech

Residual Analysis: Independence Assumption

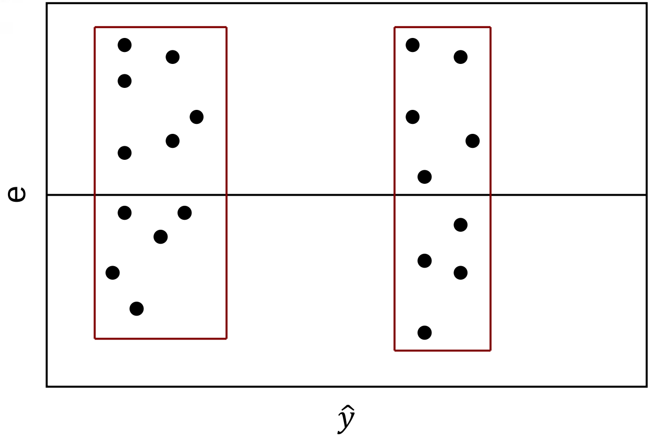
Independence *(uncorrelated errors)***:** Plot the residuals against fitted values.



- There are clusters of residuals.
- The independence assumption does not hold.

Residual Analysis: Independence Assumption

Independence *(uncorrelated errors)*: Plot the residuals against fitted values.



- There are clusters of residuals.
- The independence assumption does not hold.

- **Using residual analysis, we are actually checking for uncorrelated errors, not independence.**
- **Independence is a more complicated matter. If the data are from a randomized experiment, then independence holds, but most data are from 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 (Q-Q) Plot**

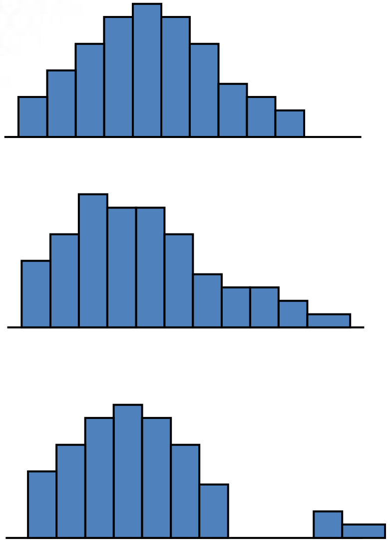
y-axis:	e_i
x-axis:	$\Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$

r_i = rank of e_i (between 1, n)
Φ = CDF of Normal Distribution

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

Residual Analysis: Normality Assumption

A complementary approach for checking for the normality assumption is by plotting the histogram of the residuals.



Normality Assumption:
The residuals should have an approximately symmetric, unimodal distribution, 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 one or more predicting variable *X* might not have a linear relationship with the response variable *Y*.

- To model the nonlinear relationship, we transform X by some nonlinear function such as

$$\boxed{f(x) = x^a}$$

or

$$\boxed{f(x) = \log(x)}$$

Normality Transformation

Problem: Constant variance or/and normality assumption

Solution: Transform the response variable from $\boldsymbol{y}$ to $\boldsymbol{y}^*$ via

$$\boldsymbol{y}^* = \boldsymbol{y}^{\lambda}$$

where the value of λ depends on how $\mathrm{Var}(\boldsymbol{y})$ changes as x changes.

$\sigma_y(x) \propto const$	$\lambda = 1$	(don’t transform)
$\sigma_y(x) \propto \sqrt{\mu_x}$	$\lambda = 1/2$	$\boldsymbol{y}^* = \sqrt{\boldsymbol{y}}$
$\sigma_y(x) \propto \mu_x$	$\lambda = 0$	$\boldsymbol{y}^* = \ln(\boldsymbol{y})$
$\sigma_y(x) \propto \mu_x^2$	$\lambda = -1$	$\boldsymbol{y}^* = \frac{1}{\boldsymbol{y}}$

Georgia Tech

Outliers in Regression

A data point far from the majority of the data (in *y* and/or any *x*) may be called an *outlier*, especially if it does not follow the general trend of the rest of the data.

- Data points that are far from the means of the *X*s or near the edge of the observation space are called *leverage points*.
- A data point that is far from the means of *y* and/or an *x* is called an *influential point* if it influences the fit of the regression.
- Excluding a leverage point may or may not the regression fit significantly, thus a leverage point may or may not be an influential point.

The upshot: Sometimes there are good reasons to exclude subsets of data (e.g., errors in data entry or experimental errors). Sometimes an outlier belongs in the data. Outliers should always be examined.

Georgia Tech

Checking for Outliers

Cook's Distance: $$\mathrm{D}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^{\mathrm{T}}(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{(p+1)\widehat{\sigma}^2}$$

where $\widehat{\boldsymbol{Y}}_{(i)}$ are the fitted values from the model fitted without the *i*th observation (i.e., excluding the *i*th observation from the data) and $\widehat{\boldsymbol{Y}}$ are the fitted values from the model fitted with the *i*th observation (i.e., including all observations).

Cook's Distance measures how much the estimated parameter values in the regression model change when the *i*th observation is removed.

Rule of Thumb: $\mathrm{D}_i > 4/n$, $\mathrm{D}_i > 1$, OR any "large" D_i should be investigated.

Checking for Outliers

Cook's Distance: $$\mathrm{D}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^{\mathrm{T}}(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{(p+1)\widehat{\sigma}^2}$$

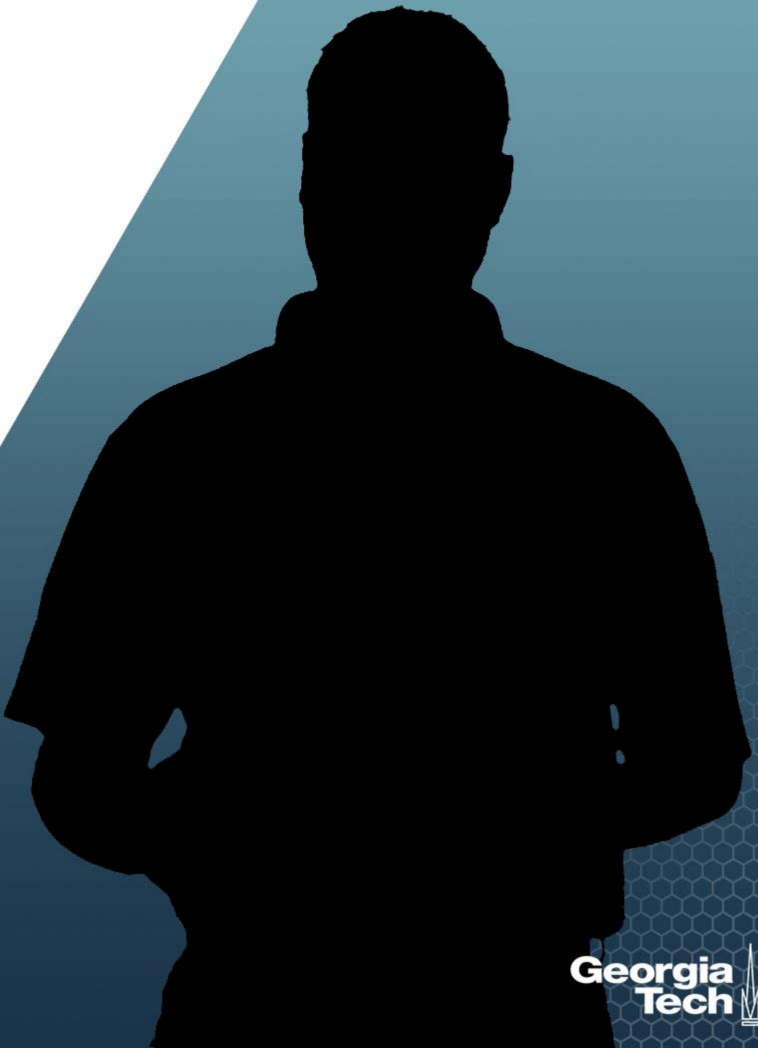
where $\widehat{\boldsymbol{Y}}_{(i)}$ are the fitted values from the model fitted without the *i*th observation (i.e., excluding the *i*th observation from the data) and $\widehat{\boldsymbol{Y}}$ are the fitted values from the model fitted with the *i*th observation (i.e., including all observations).

Cook's Distance measures how much the estimated parameter values in the regression model change when the *i*th observation is removed.

Rule of Thumb: $\mathrm{D}_i > 4/n$, $\mathrm{D}_i > 1$, OR any "large" D_i should be investigated.

- Outliers: are those few observations with much larger Cook's distance than the rest of observations;
- If a large number of outliers, then they probably point to a heavy tailed distribution rather than truly extreme values.

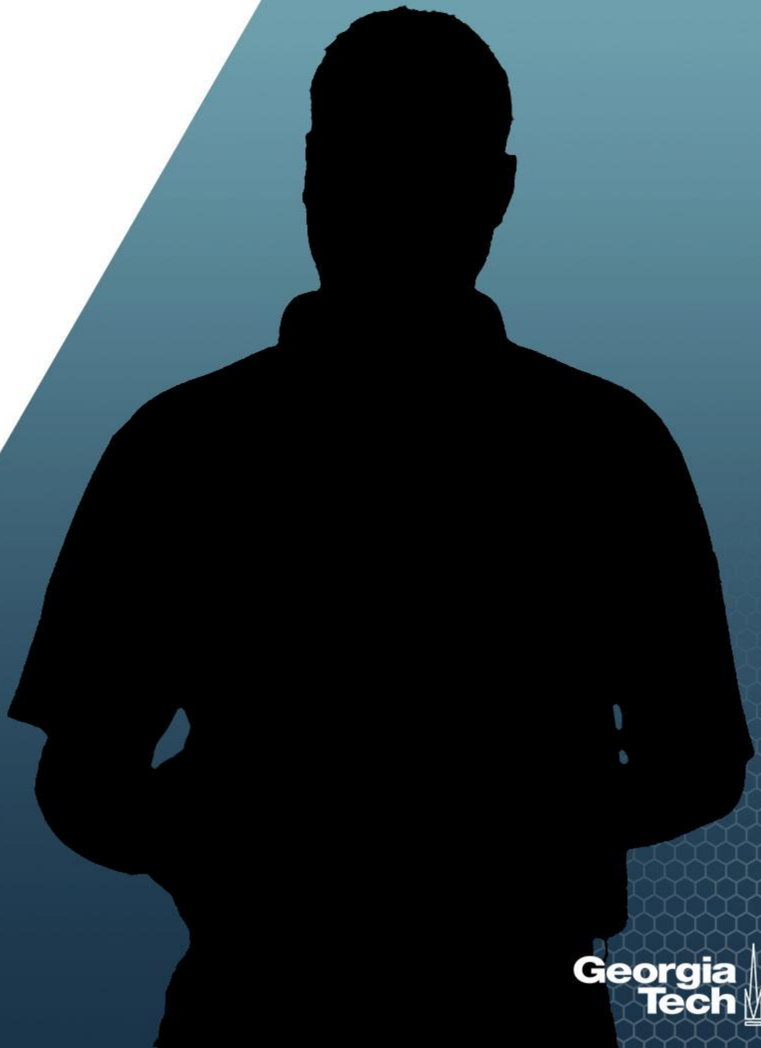
Summary



Regression Analysis

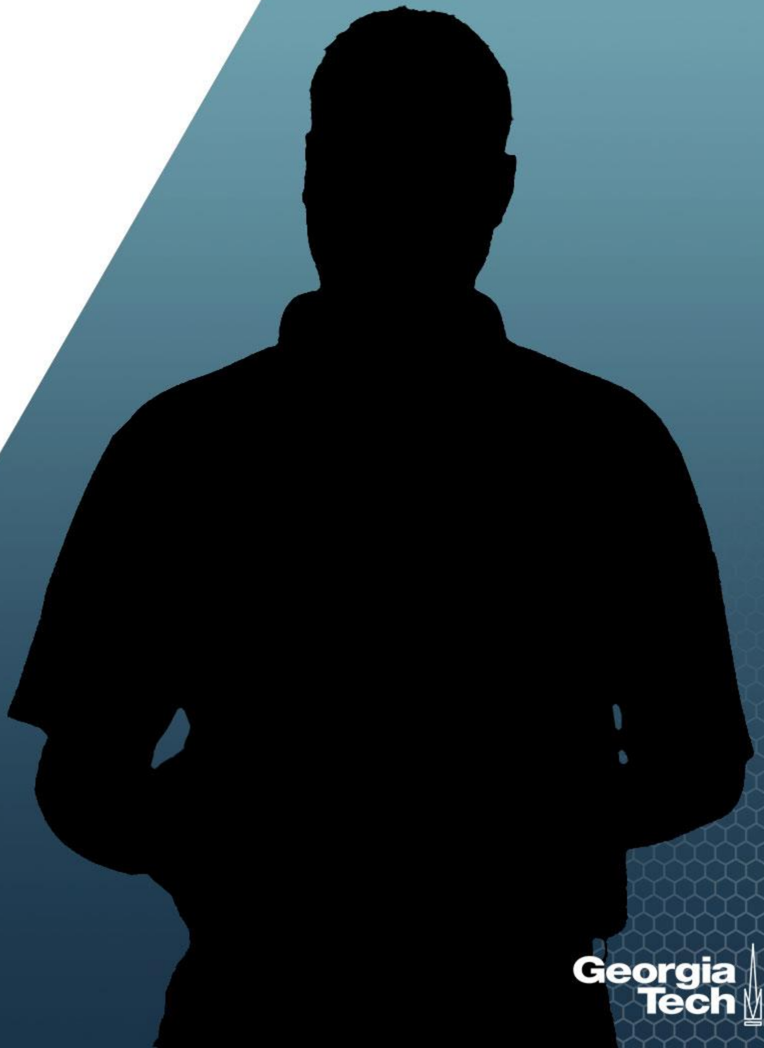
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Assumptions and Diagnostics: Data Example

About This Lesson



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 = Indicates the region of the office (1 = south, 2 = west, 3 = midwest)

Response Variable:

Y = Sales (in thousands of dollars) in 1999

Georgia Tech

Residual Analysis: Example 1

a. Do the assumptions hold? Provide the graphical displays needed to support the diagnostics. Interpret.

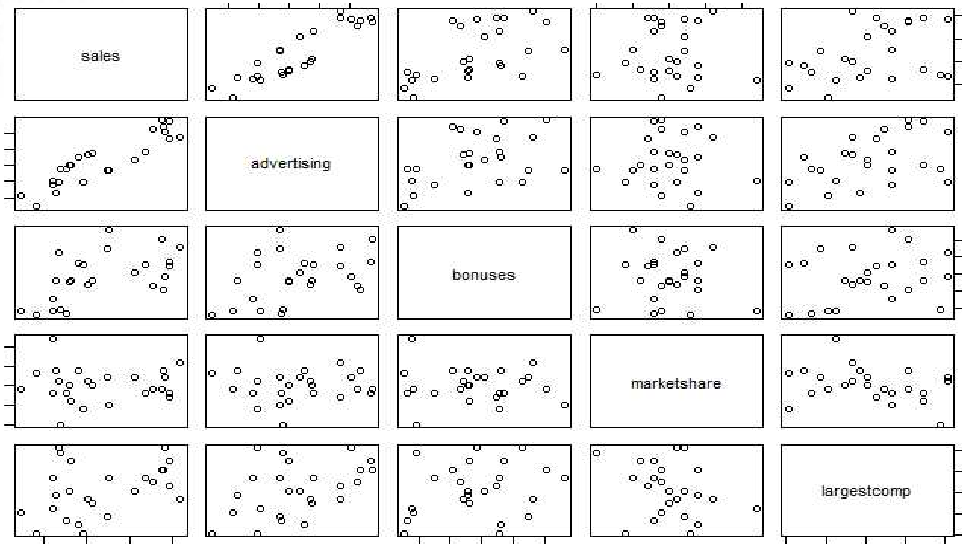
b. If one or more assumptions do not hold, what transformations do you suggest? Did the residual diagnoses improve with the suggested transformations?

c. Do you identify any outliers?

Linearity Assumption

Scatter plot matrix of sales and numeric predicting variables

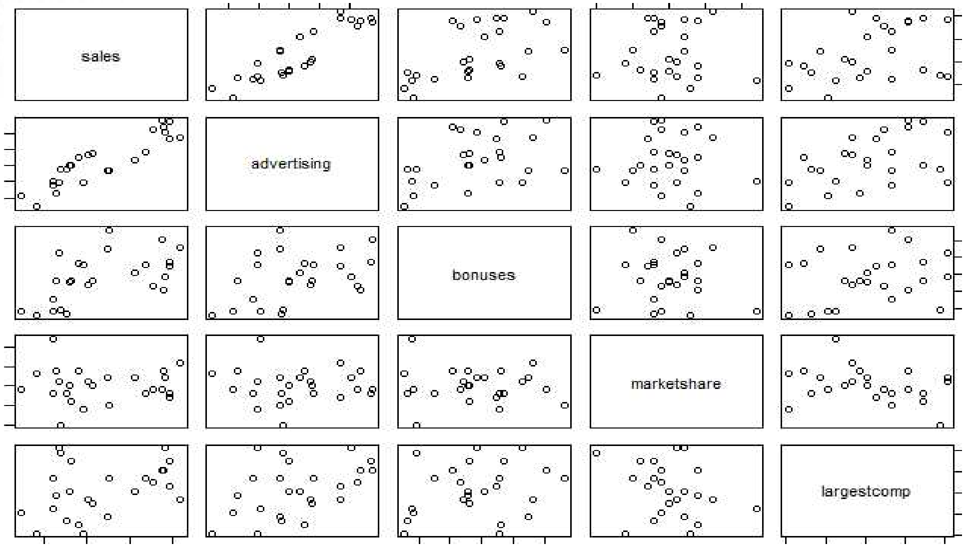
plot(meddcor[,1:5])



Linearity Assumption

Scatter plot matrix of sales and numeric predicting variables

plot(meddcor[,1:5])



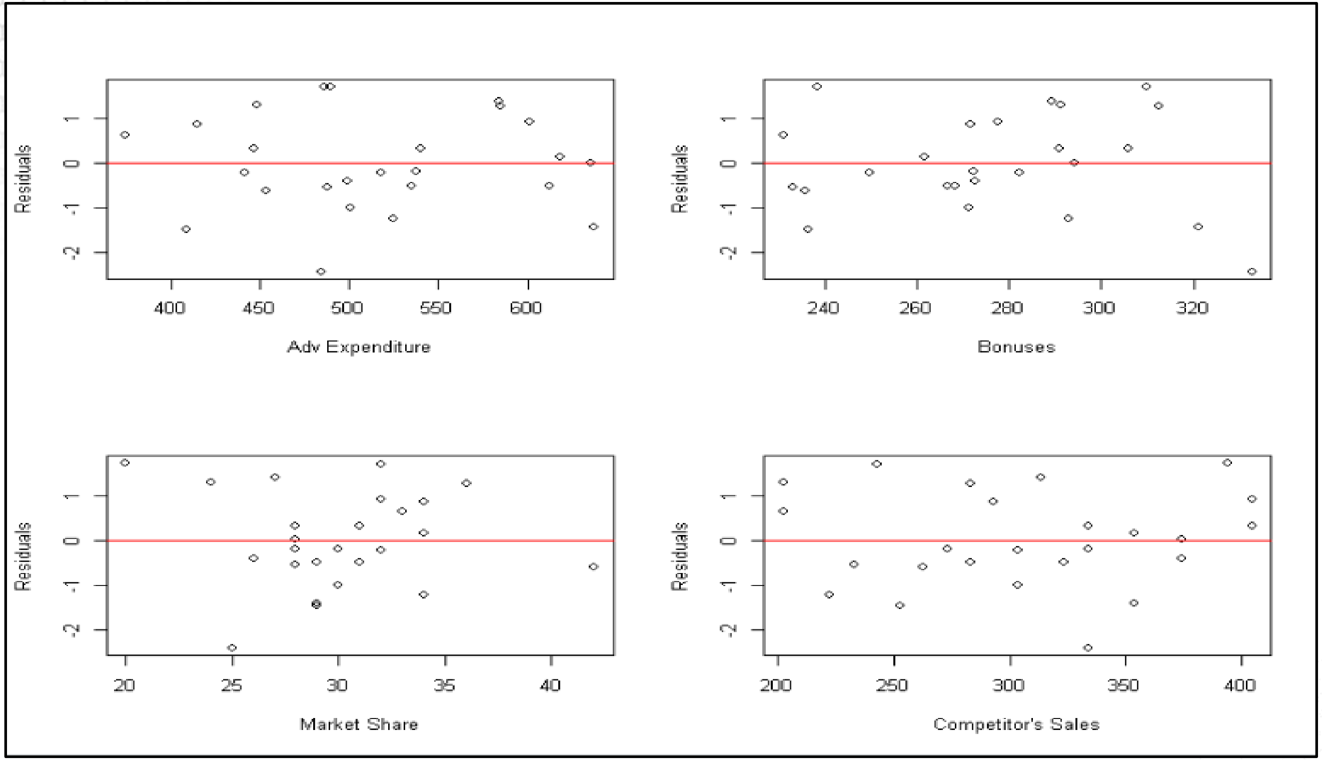
- Linearity assumption holds for all predicting variables.
- For advertisement expenditure, bonus amount, and competitor's sales, the relationship with sales is strongly linear.

Georgia Tech

Linearity Assumption

```
## Standardized Residuals versus individual predicting variables
resids = stdres(model)
par(mfrow =c(2,2))
plot(meddcor[,2],resids,xlab="Adv Expenditure",ylab="Residuals")
abline(0,0,col="red")
plot(meddcor[,3],resids,xlab="Bonuses",ylab="Residuals")
abline(0,0,col="red")
plot(meddcor[,4],resids,xlab="Market Share",ylab="Residuals")
abline(0,0,col="red")
plot(meddcor[,5],resids,xlab="Competitor's Sales",ylab="Residuals")
abline(0,0,col="red")
```

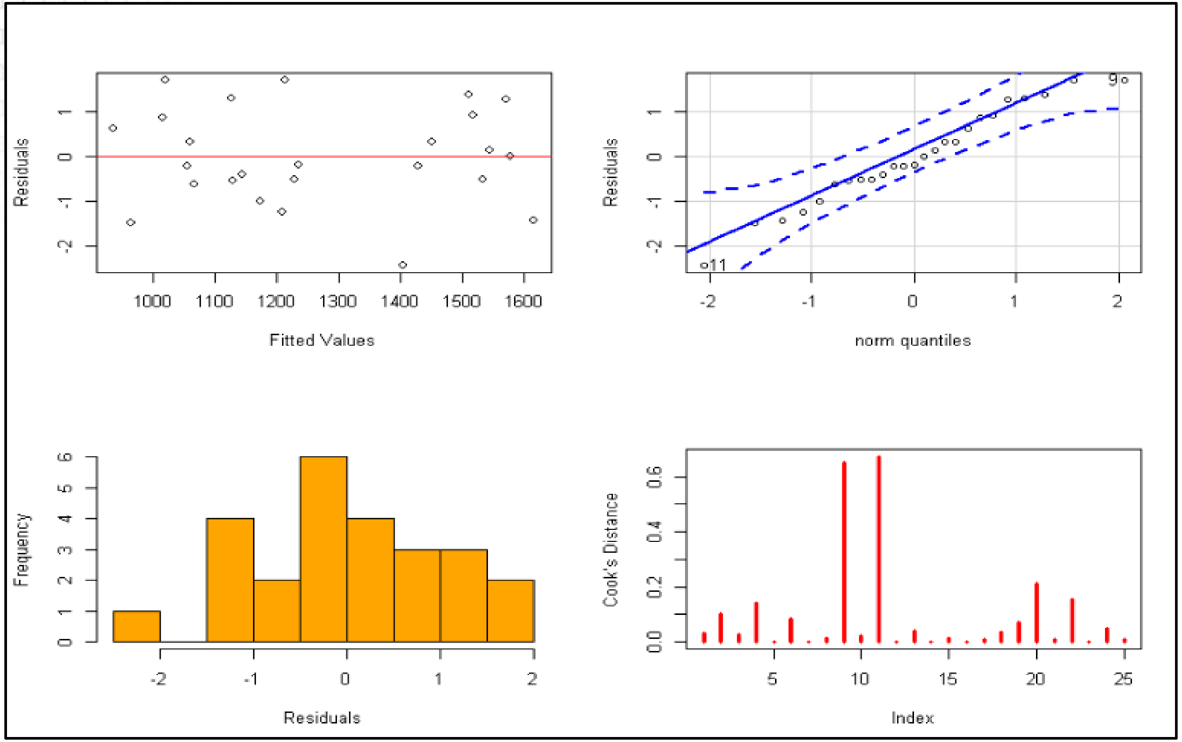
Linearity Assumption



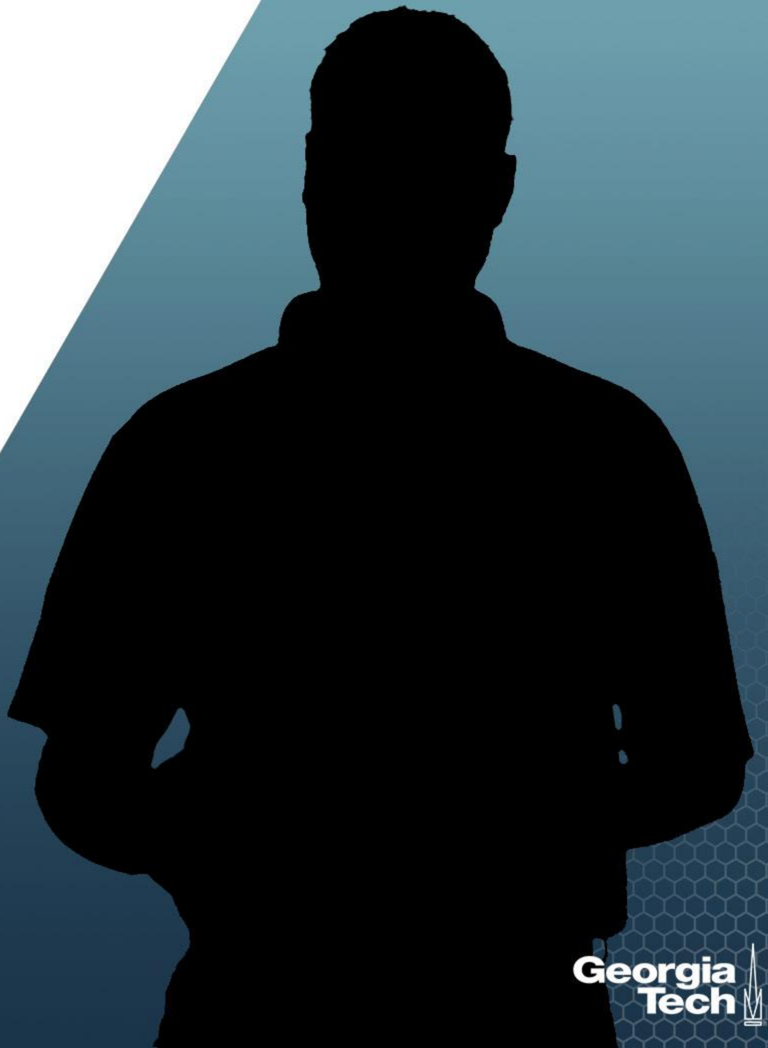
Residual Analysis: Other Assumptions

```
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



Summary

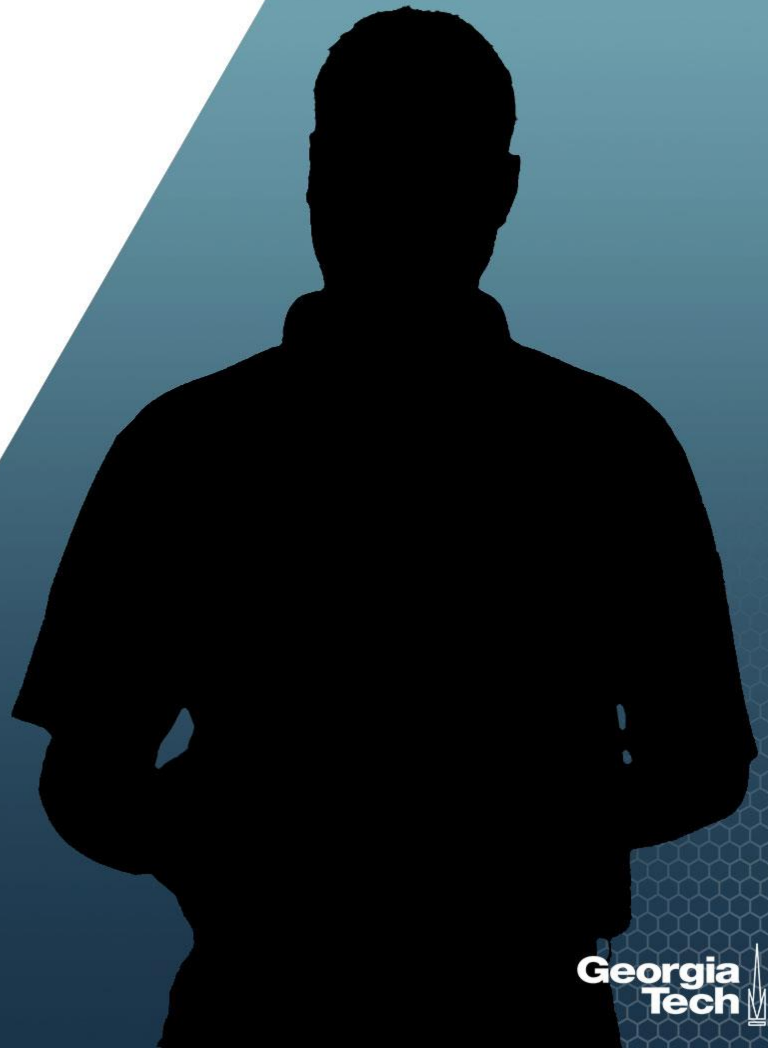


Regression Analysis

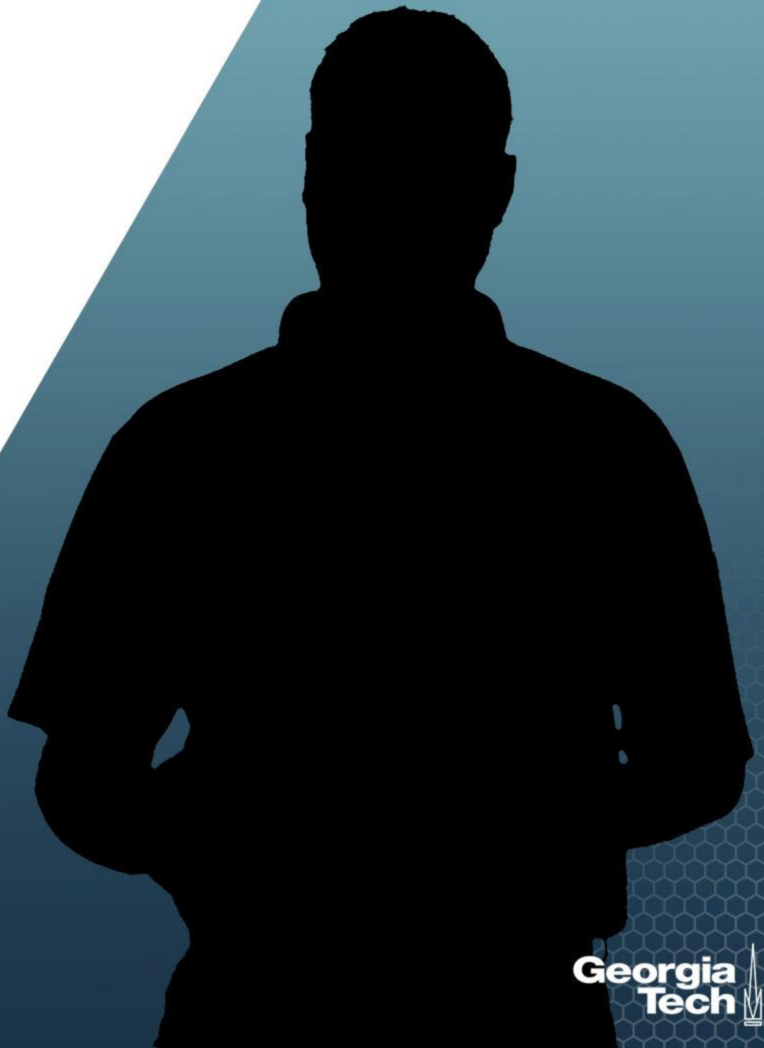
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Model Evaluation and Multicollinearity



About This Lesson



R^2: Coefficient of Determination

A measure that efficiently summarizes how well the *X*s can be used to predict *Y* is **R^2** (called *R-squared* or the *coefficient of determination*):

$$\mathrm{R}^2 = 1 - \mathrm{SSE}/\mathrm{SST}$$

where

$$\mathrm{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

R^2 is interpreted as the proportion of total variability in *Y* that can be explained by the linear regression model.

R^2: Notation and Terminology

SSE, **SST**, and **SSR** refer to ***sum of squared errors***, ***sum of squares total***, and ***sum of squares for regression***. Unfortunately, the field of statistics abounds in inconsistent terminology and notation.

- **SSE:** sometimes denoted $\mathbf{SS_{error}}$ or $\mathbf{SS_{err}}$, is also known as **RSS** (residual sum of squares) and $\mathbf{SS_{res}}$ (sum of squared residuals, sometimes **SSR**).
- **SST:** sometimes written as $\mathbf{SS_{total}}$ or $\mathbf{SS_{tot}}$. It is also called total sum of squares and written as **TSS**.
- **SSR:** also called the sum of squares due to regression, and it is sometimes written as $\mathbf{SS_{reg}}$. It's also called explained sum of squares (**ESS**). Don't confuse ESS with SSE, and, for R^2, remember that SSR is SS regression, not SS residuals!

Model Evaluation

- **F-test for overall regression**
  - $\mathrm{H}_0\colon \beta_1 = \beta_1 = \cdots = \beta_p = 0$
  - $\mathrm{F}_0 = \mathrm{MSR}/\mathrm{MSE} \sim \mathrm{F}(p, n\text{-}p\text{-}1)$
    - $\mathrm{MSR} = \mathrm{SSR}/p$
      - $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
    - $\mathrm{MSE} = \mathrm{SSE}/(n\text{-}p\text{-}1)$

- **Coefficient of determination**
  - $\mathrm{R}^2 = 1 - \mathrm{SSE}/\mathrm{SST} = \mathrm{SSR}/\mathrm{SST}$
  - R^2 increases when additional predictors are added to a model
    - Such increase might not indicate increased explanatory power

- **Adjusted coefficient of determination**
  - Penalizes for more predictors
  - adjusted $\mathrm{R}^2 = 1 - (n\text{-}1)(1\text{-}\mathrm{R}^2)/(n\text{-}p\text{-}1)$

Correlation Coefficient

A statistic that efficiently summarizes how well one of the *X*s is *linearly* related to *Y* (or to another *X*) is ρ, the (Pearson) correlation coefficient:

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

- Can be used to evaluate the *linear* relationship between the response variable and any of the predicting variables, X_j
  - Useful when looking for transformations of predicting variables
- Can also be used to evaluate correlation between predicting variables
  - Can help detect near linear dependence (multicollinearity)

Multicollinearity

Recall that finding the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$$

depends on $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ being invertible (nonsingular or nondegenerate). From linear algebra, a square matrix is invertible if and only if its columns are linearly independent (i.e., no column is a linear combination of the others).

If that doesn't hold, the ordinary least squares estimator of $\widehat{\boldsymbol{\beta}}$ doesn't exist. That's probably due to a specification error where one or more predictors should be eliminated as redundant (e.g., if years and number of rings were included in a model for trees).

Even if the columns of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ are linearly independent, some problems might arise if the value of one predictor can be closely estimated from the other predictors. We call this condition *multicollinearity* or *near collinearity*.

Georgia Tech

Multicollinearity

- Indications that near collinearity is present:
  - The estimated coefficients $\hat{\boldsymbol{\beta}}$ are unstable: When the value of one predictor changes slightly, the fitted regression coefficients change dramatically
  - The standard error of $\hat{\boldsymbol{\beta}}$ is artificially large
  - The overall F statistic is significant, but individual *t*-statistics are not
- Prediction may be affected
  - The relationship to the response may change widely
- Some computational algorithms are sensitive to multicollinearity
- <u>But no</u> inflation or deflation in $\mathbf{R}^2$

Multicollinearity Diagnosis

Compute the variance inflation factor (VIF_j) for each predicting variable X_j

$$\boxed{\mathrm{VIF}_j = \frac{1}{1 - \mathrm{R}_j^2}}$$

where R_j^2 is the coefficient of determination for the regression of X_j against all other predicting variables.

What is an acceptable VIF (i.e., multicollinearity is not problematic)?

- $\mathrm{VIF} < \max\left(10,\ 1/\left(1 - \mathrm{R}^2_{\mathrm{model}}\right)\right)$
  - $\mathrm{R}^2_{\mathrm{model}}$ is the coefficient of determination for the original model
  - Rule of thumb only

Multicollinearity Diagnosis

Steps:

1. For $j = 1,...,p$, regress X_j against all other X_i, $i = 1, \ldots, p, i \neq j$ (i.e., $\mathrm{X}_j \sim \mathrm{X}_1, \ldots, \mathrm{X}_{j-1}, \mathrm{X}_{j+1}, \ldots, \mathrm{X}_p$).
2. For each regression run, compute R^2 for that regression (i.e., compute R_j^2 for $j = 1, ..., p$).
3. For each regression run, compute VIF_j based on the computed R_j^2 for that regression.
4. If any $\mathrm{VIF}_j \geq 10$ and it is also $\geq 1/\left(1 - \mathrm{R}_{\text{model}}^2\right)$, where $\mathrm{R}_{\text{model}}^2$ is the coefficient of determination for the original model, the test is positive for multicollinearity.

High multicollinearity is not detected if each $\mathrm{VIF}_j < \max(10, \frac{1}{1-\mathrm{R}_{\text{model}}^2})$.

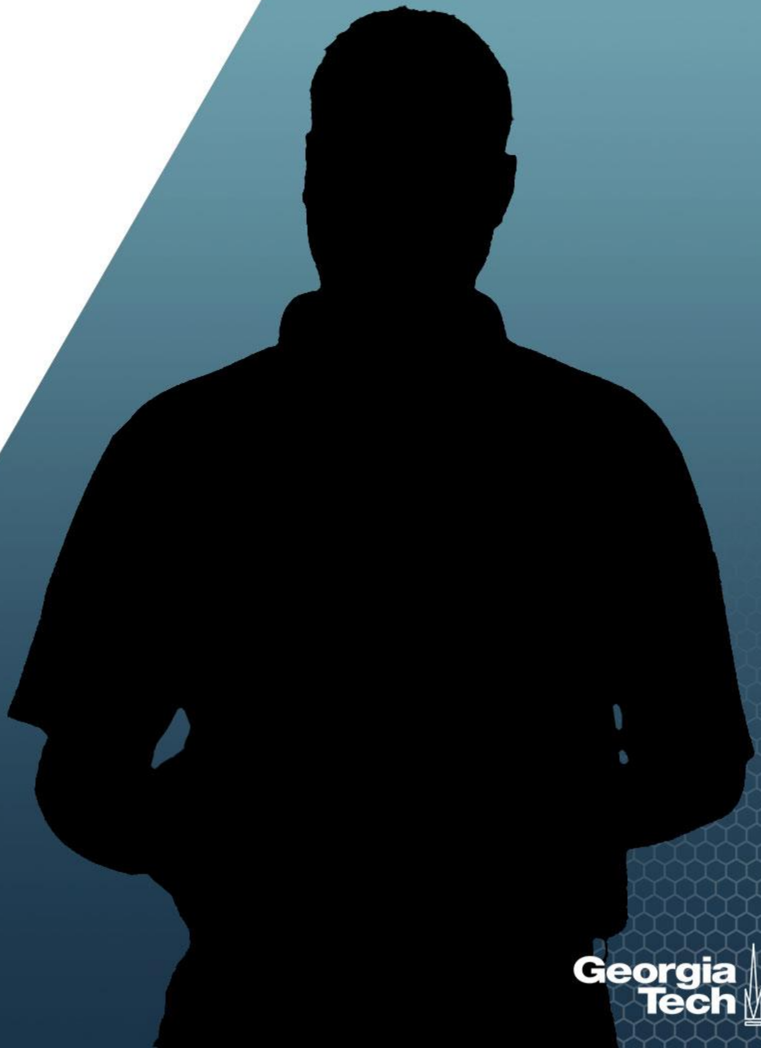
Georgia Tech

Multicollinearity Interpretation

VIF measures the proportional increase in the variance of $\hat{\beta}_i$ compared to what it would have been if the predicting variables had been completely uncorrelated.

- VIF of 1 (the minimum possible VIF) means the tested predictor is not correlated with the other predictors
- The higher the VIF:
  - The more correlated a predictor is with the other predictors
  - The more the standard error is inflated
  - The larger the confidence interval
  - The less likely it is that a coefficient will be evaluated as statistically significant

Summary



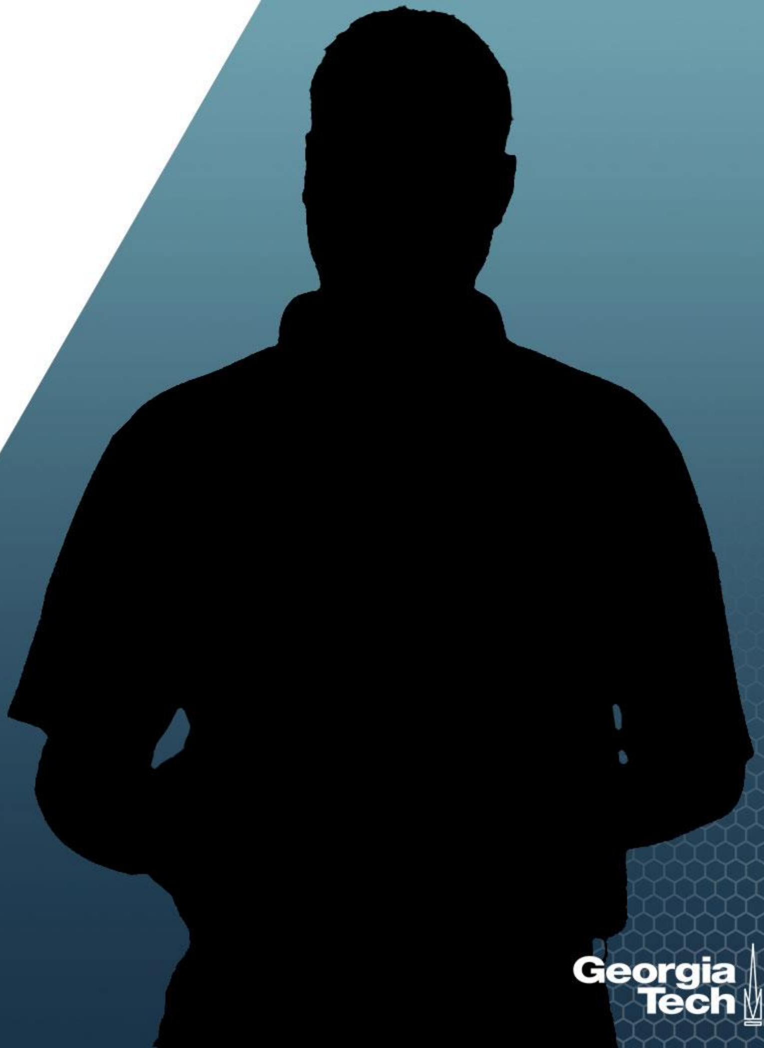
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Multicollinearity: Data Example

About This Lesson



Model Evaluation: 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 office is located (1 = south, 2 = west, 3 = midwest)

Response Variable:

Y = yearly sales (in thousands of dollars)

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

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

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

b. The R function vif() outputs the generalized VIF (GVIF), which specializes to the usual VIF in the case of a single coefficient.

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

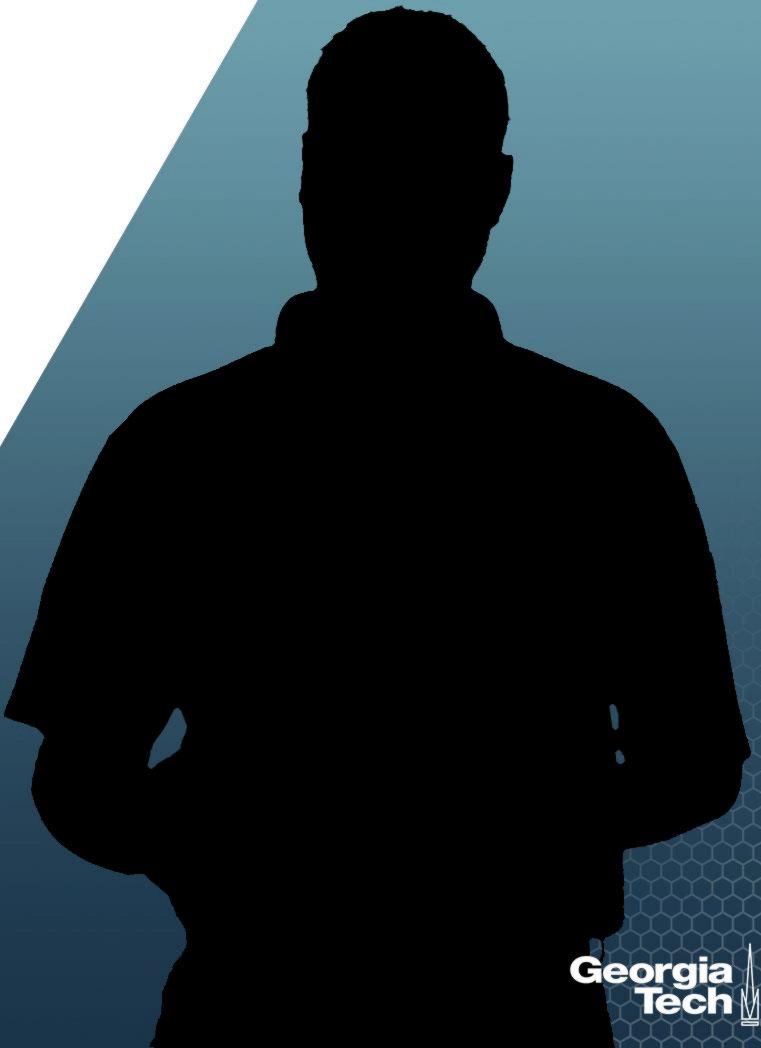
summary(model)$r.squared

0.9555032

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

Georgia Tech

Summary



Regression Analysis

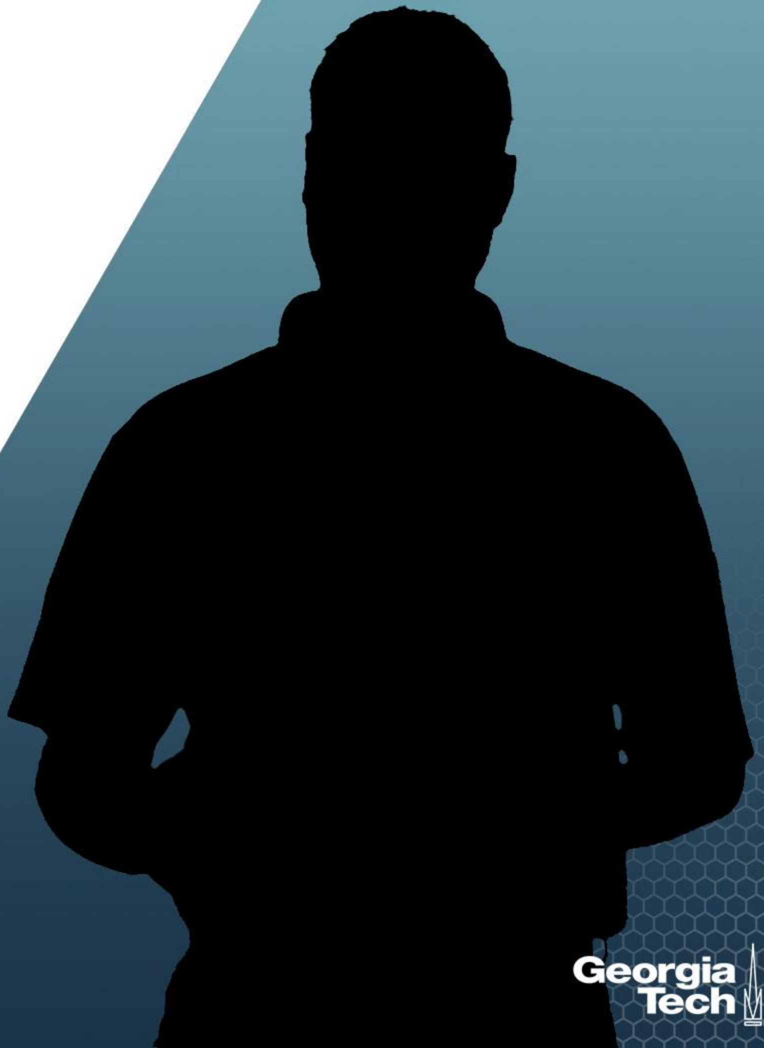
Multiple Linear Regression

Nicoleta Serban, Ph.D.

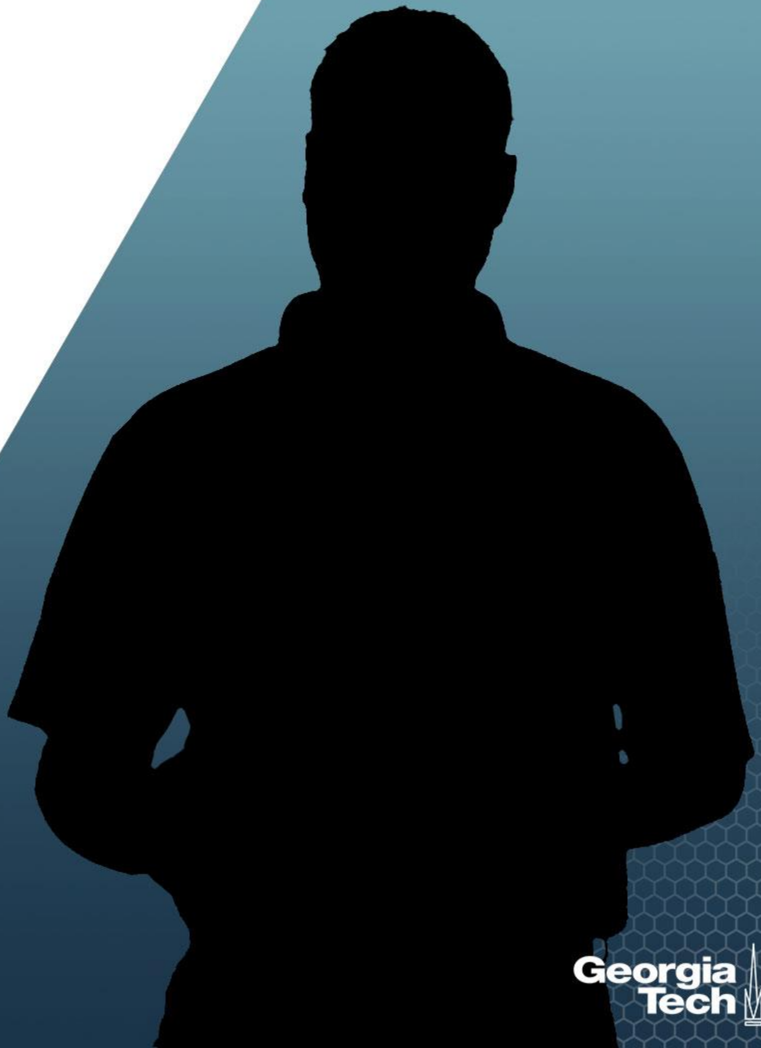
Professor

School of Industrial and Systems Engineering

Ranking States by SAT Performance: Exploratory Analysis



About This Lesson



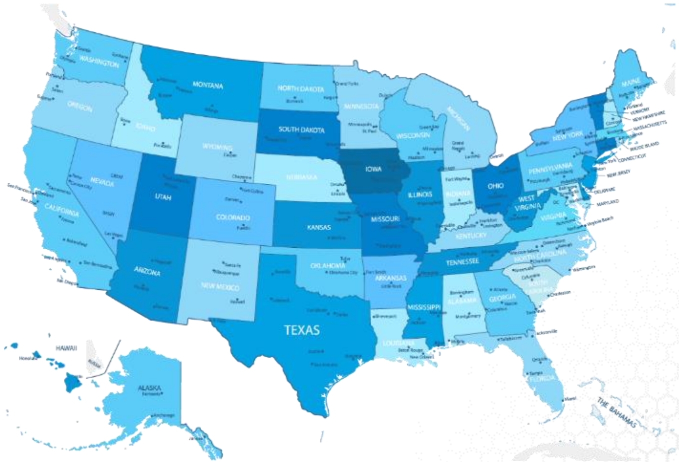
Linear Regression: Example 2

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

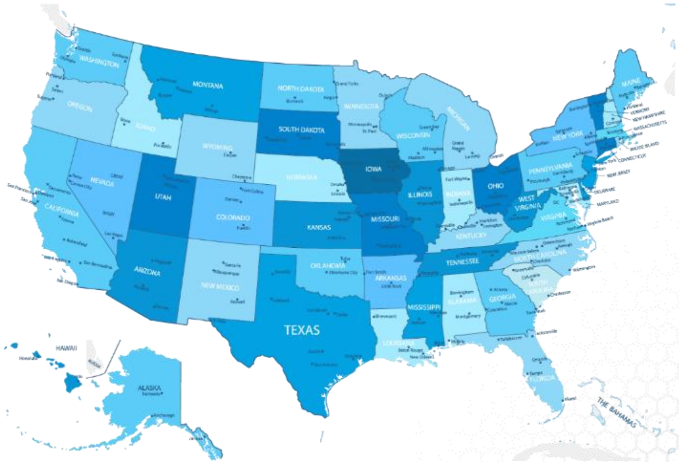
Explanatory Factors:

- 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



Ranking States by SAT Performance

- ***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

Response Variable:

sat	State average SAT score (verbal and quantitative combined)

Predicting Variables:

takers	% of eligible students (high school seniors) in 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 scores had low percentages of takers.
- Those taking the test tend to be in the higher median percentiles of rankings of test takers within their secondary school classes.

Controlling Factors:

takers	% of eligible students (high school seniors) in 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)
[1] 50 8

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

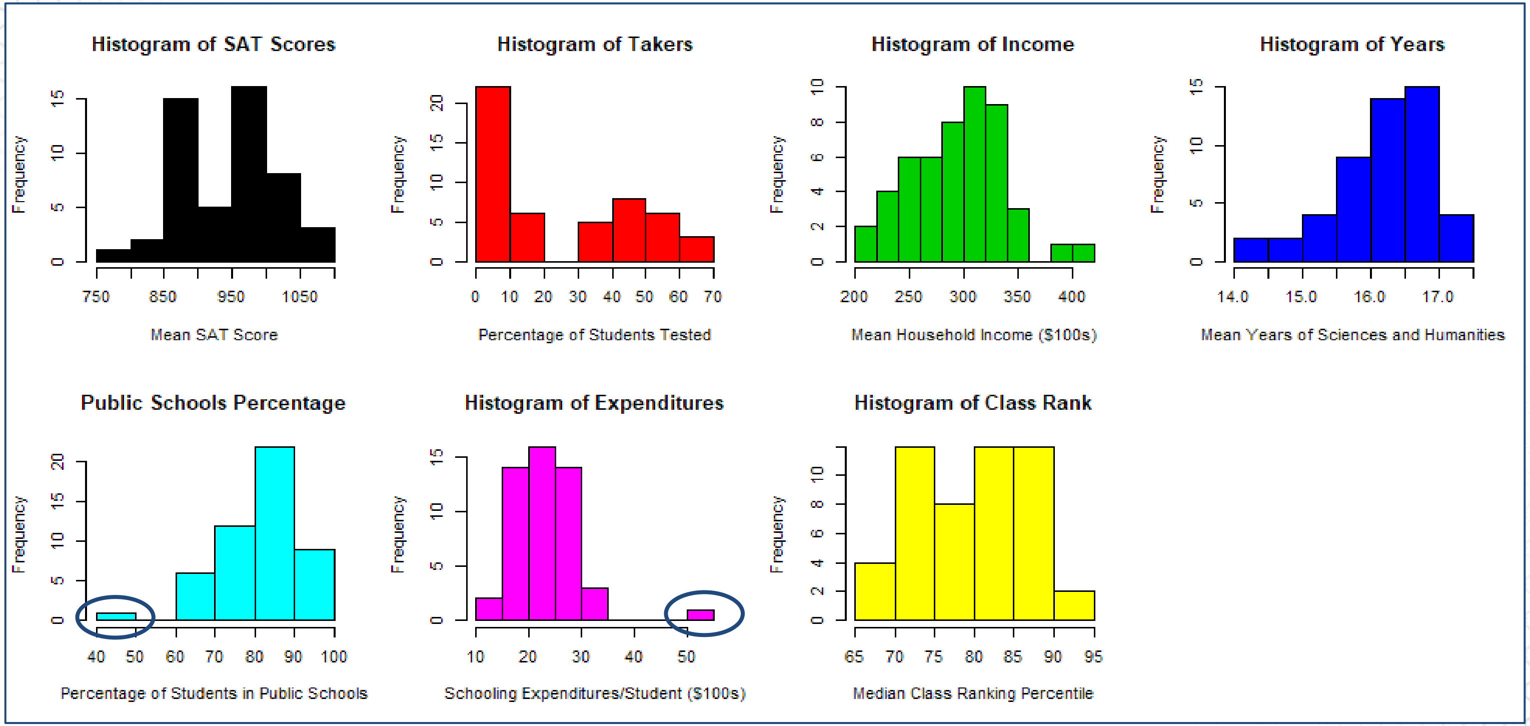
The data consist of 50 rows, each corresponding to a U.S. state.

Georgia Tech

Exploratory Data Analysis in R

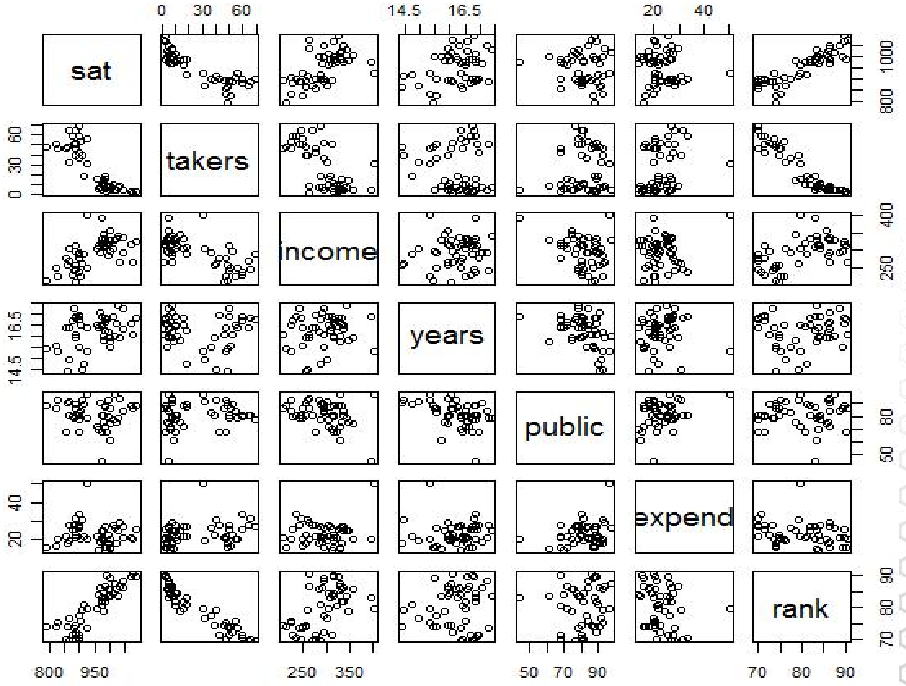
```
## Evaluate the shape of the distribution of each predicting variable and of the response variable
par(mfrow = c(2, 4))
hist(sat, main = "Histogram of SAT Scores", xlab = "Mean SAT Score", col = 1)
hist(takers, main = "Histogram of Takers", xlab = "Percentage of Students Tested", col = 2)
hist(income, main = "Histogram of Income", xlab = "Mean Household Income ($100s)", col = 3)
hist(years, main = "Histogram of Years", xlab = "Mean Years of Sciences and Humanities", col = 4)
hist(public, main = "Public Schools Percentage", xlab = "Percentage of Students in Public Schools", col = 5)
hist(expend, main = "Histogram of Expenditures", xlab = "Schooling Expenditures/Student ($100s)", col = 6)
hist(rank, main = "Histogram of Class Rank", xlab = "Median Class Ranking Percentile", col = 7)

## Evaluate the scatter plot matrix of the data, ignoring the first column
par(mfrow = c(1, 1))
plot(data[,-1])

## Explore the correlation coefficients
round(cor(data[,-1]), 2)
```

Exploratory Data Analysis in R

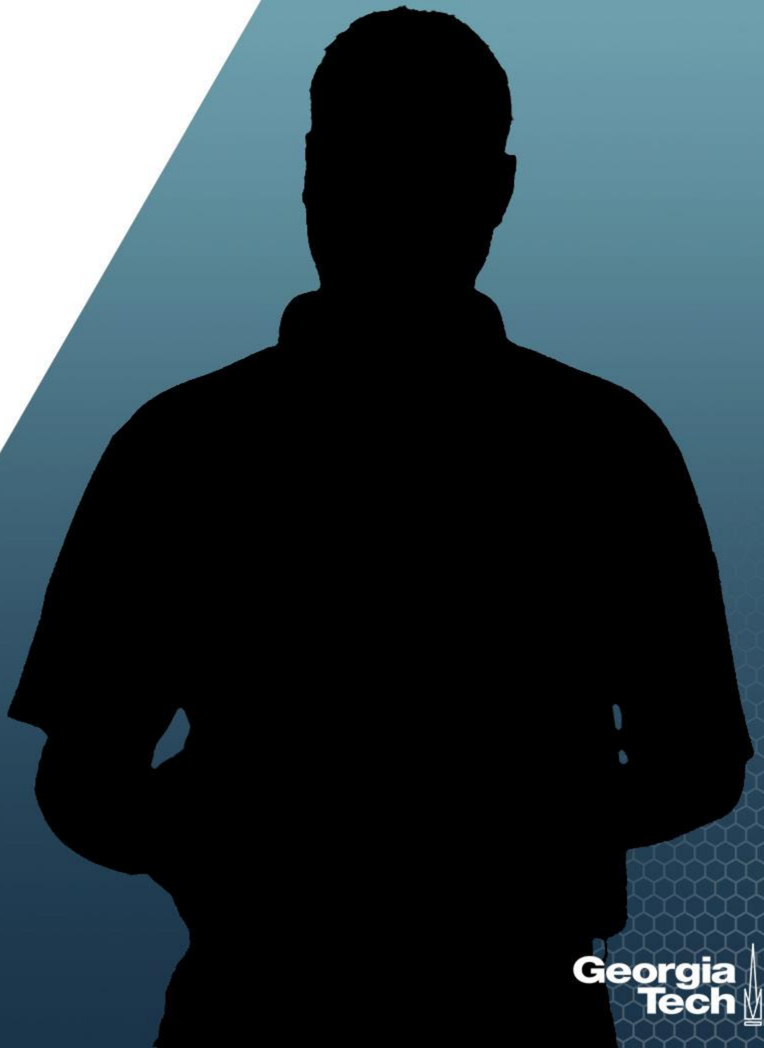


Exploratory Data Analysis in R (Cont'd)

	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



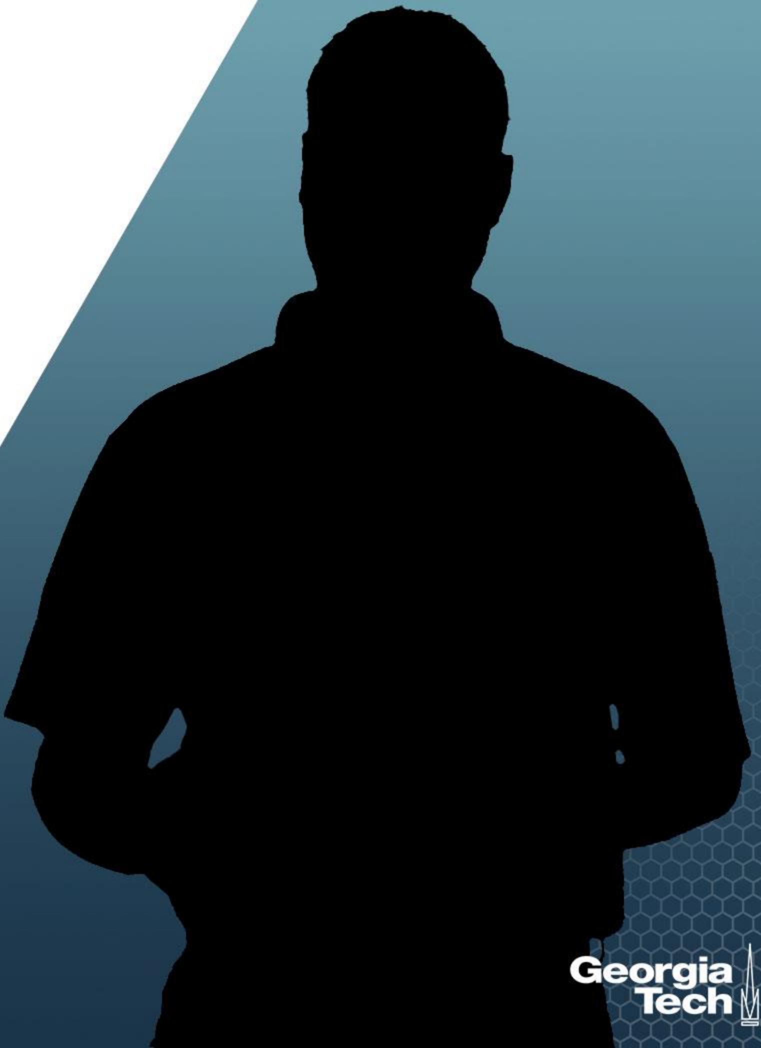
Summary



Regression Analysis

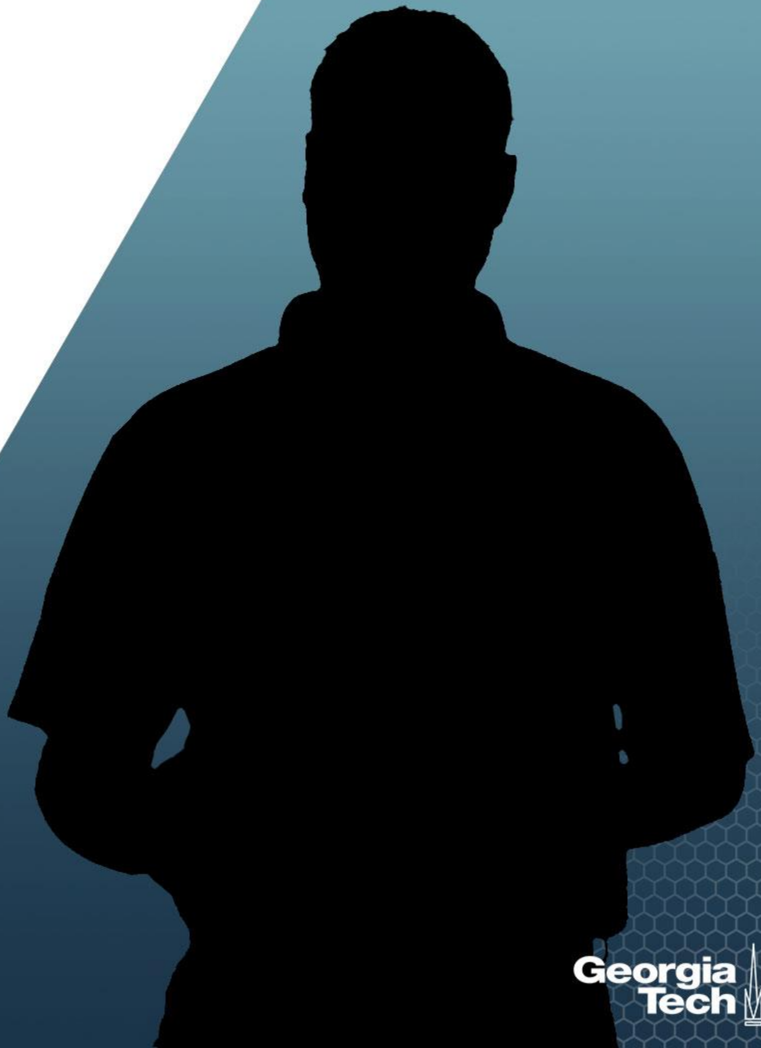
Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Ranking States by SAT Performance: Regression Analysis

About This Lesson



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   0.693711  -0.692 0.492628
rank          8.476217   2.107807   4.021 0.000230 ***
income       -0.008195   0.152358  -0.054 0.957353
years        22.610082   6.314577   3.581 0.000866 ***
public       -0.464152   0.579104  -0.802 0.427249
expend        2.212005   0.845972   2.615 0.012263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

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	0.693711	-0.692	0.492628
rank	8.476217	2.107807	4.021	0.000230 ***
income	-0.008195	0.152358	-0.054	0.957353
years	22.610082	6.314577	3.581	0.000866 ***
public	-0.464152	0.579104	-0.802	0.427249
expend	2.212005	0.845972	2.615	0.012263 *

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

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

Test for statistical significance:

$\hat{\beta}_{takers}$ Pr(>|t|) ≈ 0.4926 > 0.01
$\hat{\beta}_{rank}$ Pr(>|t|) ≈ 0.0002 < 0.01
$\hat{\beta}_{income}$ Pr(>|t|) ≈ 0.9574 > 0.01
$\hat{\beta}_{years}$ Pr(>|t|) ≈ 0.0009 < 0.01
$\hat{\beta}_{public}$ Pr(>|t|) ≈ 0.4272 > 0.01
$\hat{\beta}_{expend}$ Pr(>|t|) ≈ 0.0123 ≈ 0.01

$\hat{\sigma}$ = 26.34, df = n-p-1 = 43
R^2 ≈ 0.879 ⇒ 87.9% of variability explained

Georgia Tech

Testing for Subsets of Coefficients

```
## Compare models: reduced with controlling variables only vs. full with all variables
anova(regression.line)
Analysis of Variance Table

Response: sat
          Df  Sum Sq  Mean Sq  F value    Pr(>F)
takers     1  181024   181024 260.8380  < 2.2e-16 ***
rank       1   11209    11209  16.1512  0.0002313 ***
income     1    2858     2858   4.1182  0.0486431 *
years      1   16080    16080  23.1701   1.86e-05 ***
public     1     252      252   0.3631  0.5499447
expend     1    4745     4745   6.8369  0.0122629 *
Residuals 43   29842      694

## compute partial-F statistic
fstat = ((2858+16080+252+4745)/4)/(29842/43)
pvalue = 1-pf(fstat,4,43)
pvalue
[1] 3.349778e-05
```

Testing for Subsets of Coefficients

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

How were the F-statistic and the p-value computed?

$$\mathrm{F-statistic} = \frac{\mathrm{SS_{Reg}}(income, public, years, expend \mid takers, rank)/4}{\mathrm{SSE}/(50-6-1)}$$

$$\mathrm{Pr}\left(\mathrm{F}_{4,43} > \mathrm{F-statistic}\right) = 1 - \mathrm{Pr}\left(\mathrm{F}_{4,43} < \mathrm{F-statistic}\right)$$

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.

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 model with the two controlling factors 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$res, decreasing = TRUE)
## Reorder states. Create table including state name, new and old order.
states = factor(data[order.vec, 1])
newtable = data.frame(State = states, Residual = as.numeric(round(reduced.line$res[order.vec], 1)), oldrank = (1:50)[order.vec])
newtable
```

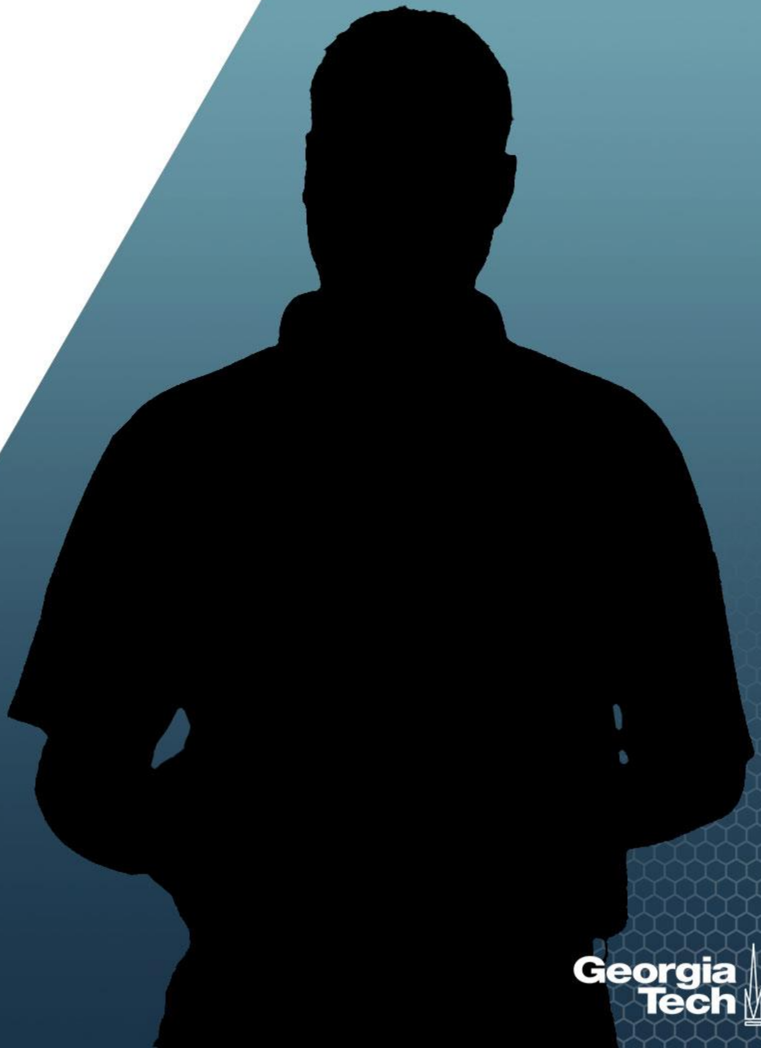
Using Residuals to Create Better Rankings

	State	Residual	oldrank
1	Connecticut	53.9	35
2	Iowa	53.5	1
3	NewHampshire	45.8	28
4	Massachusetts	41.9	41
5	NewYork	40.9	36
6	Minnesota	40.6	7
7	Kansas	35.8	4
8	SouthDakota	33.4	2
⋮			
43	Arkansas	-31.2	12
44	WestVirginia	-38.9	25
45	Nevada	-45.4	30
46	Mississippi	-49.3	16
47	Texas	-50.3	45
48	Georgia	-63.0	49
49	NorthCarolina	-71.3	48
50	SouthCarolina	-98.5	50

After controlling for selection bias, Connecticut moved from 35^{th} to 1^{st}.

After controlling for selection bias, Mississippi moved from 16^{th} to 46^{th}.

Summary

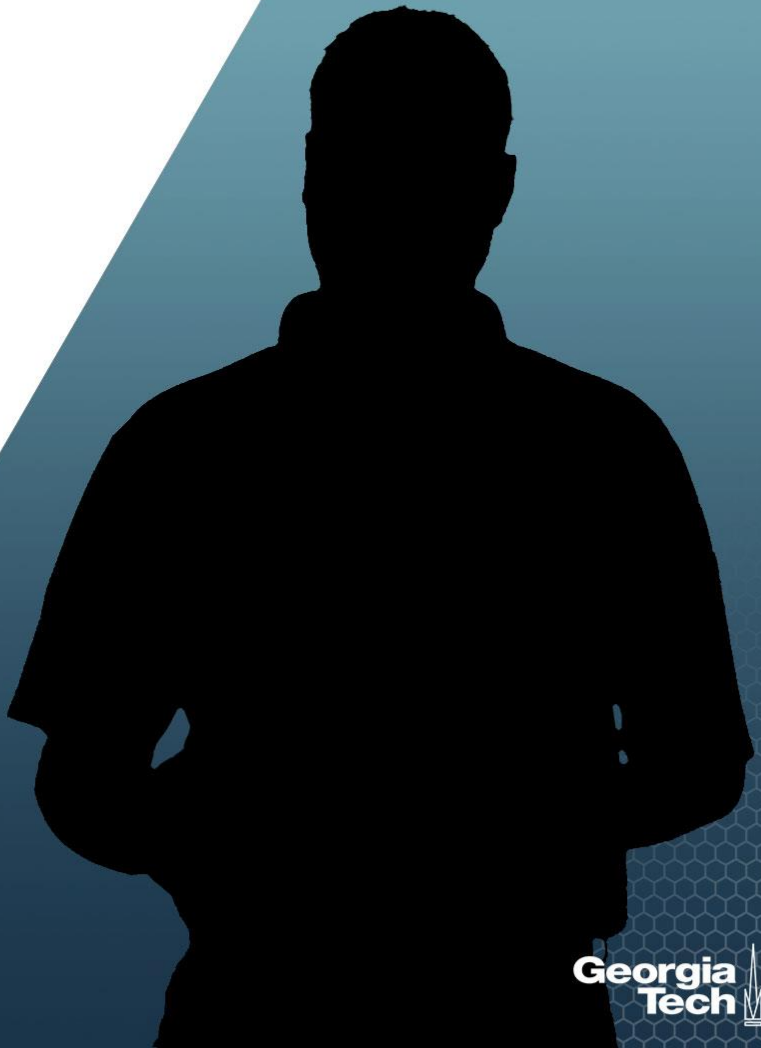


Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

Ranking States by SAT Performance: Model Fit



About This Lesson

Residual Analysis

To evaluate assumptions:

- ***Constant variance & uncorrelated errors***
  - Response variable or fitted values vs residuals
- ***Linearity***
  - Predicting variables vs residuals
- ***Normality***
  - Histogram and QQ normal plot

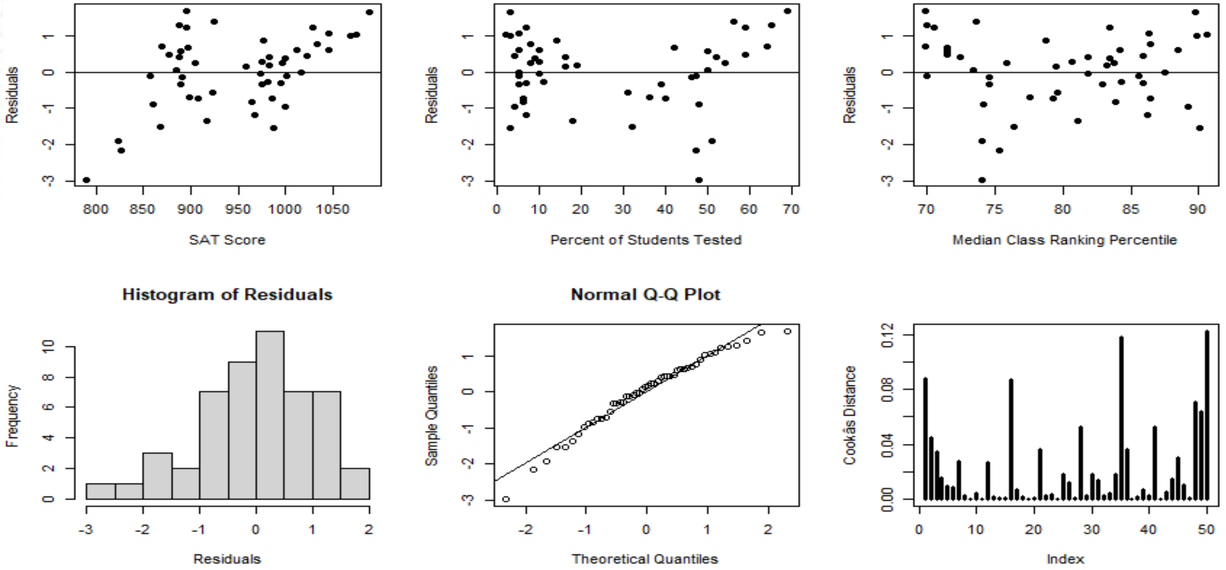
To evaluate outliers:

- Cook's distance plots

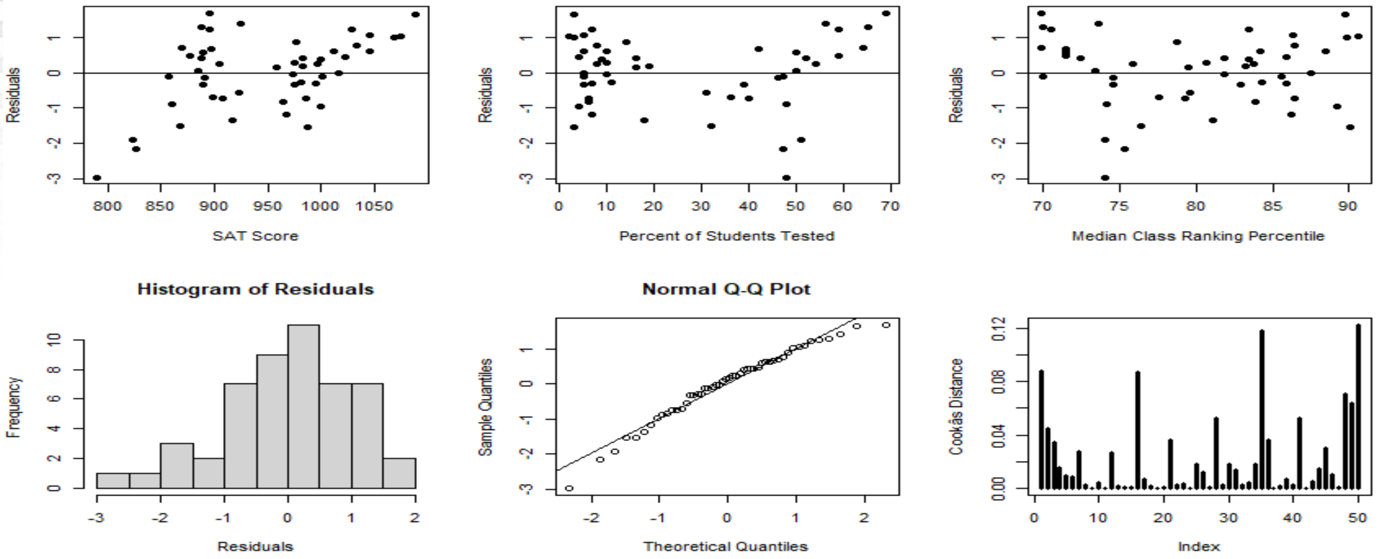
Residual Analysis

```
## Residual analysis for the reduced model
res = stdres(reduced.line)
cook = cooks.distance(reduced.line)
par(mfrow = c(2,3))
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")
qqnrom(res)
qqline(res)
plot(cook,type="h",lwd=3, ylab = "Cook’s Distance")
```

Residual Analysis



Residual Analysis



- Transform the predicting variable Percent of Students Tested (*takers*)
- Reanalyze heavy tailed residuals and outliers after transformation

Georgia Tech

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
```

Linear Regression Analysis in R

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

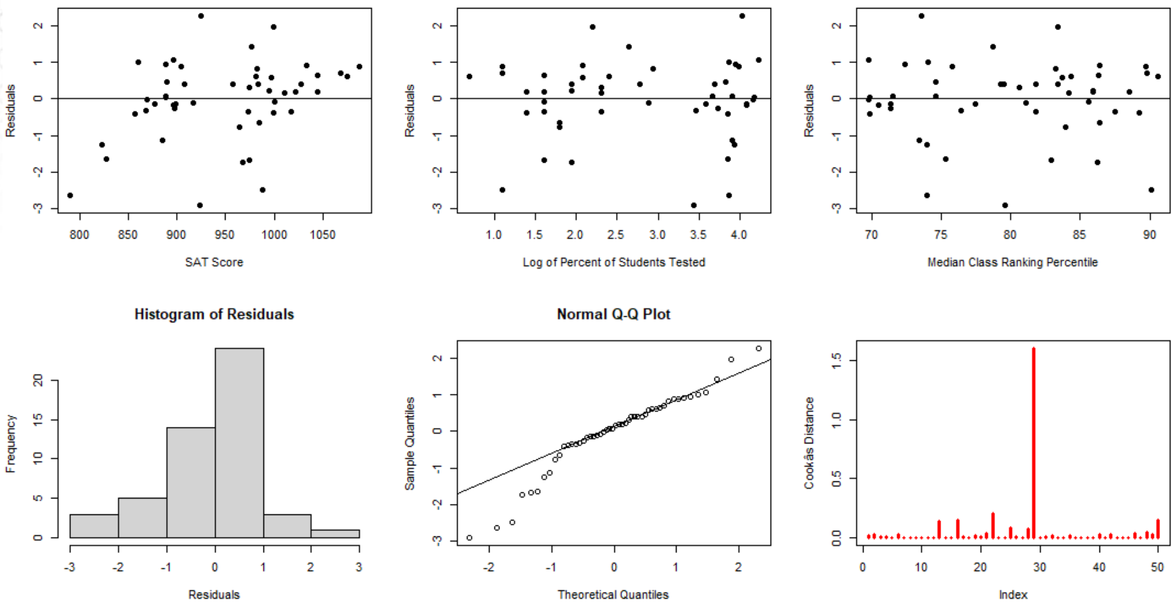
```
Coefficients:
              Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) 407.53990  282.76325   1.441  0.15675
log(takers) -38.43758   15.95214  -2.410  0.02032 *
rank          4.11427    2.50166   1.645  0.10734
income       -0.03588    0.13011  -0.276  0.78407
years        17.21811    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:

$\hat{\beta}_{\log(takers)}$ Pr(>|t|) ≈ 0.0203 < 0.01
$\hat{\beta}_{rank}$ Pr(>|t|) ≈ 0.1073 > 0.01
$\hat{\beta}_{income}$ Pr(>|t|) ≈ 0.7840 > 0.01
$\hat{\beta}_{years}$ Pr(>|t|) ≈ 0.0093 < 0.01
$\hat{\beta}_{public}$ Pr(>|t|) ≈ 0.8417 > 0.01
$\hat{\beta}_{expend}$ Pr(>|t|) ≈ 0.0027 < 0.01

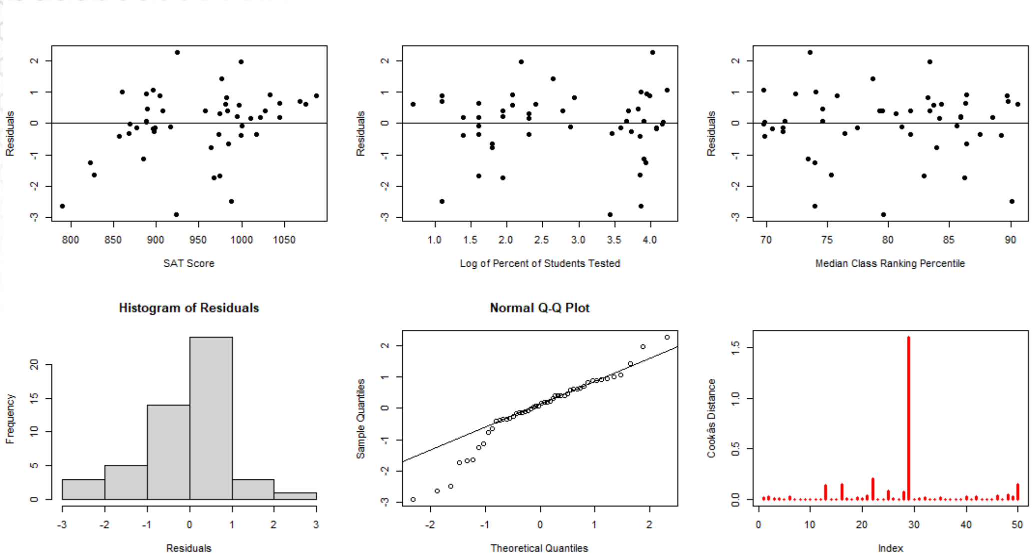
$\hat{\sigma} = 24.86$, df = n-p-1 = 43
$R^2 \approx 0.892 \Rightarrow$ 89.2% of variability explained

Residual Analysis



Georgia Tech

Residual Analysis



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

State SAT Performance: Findings

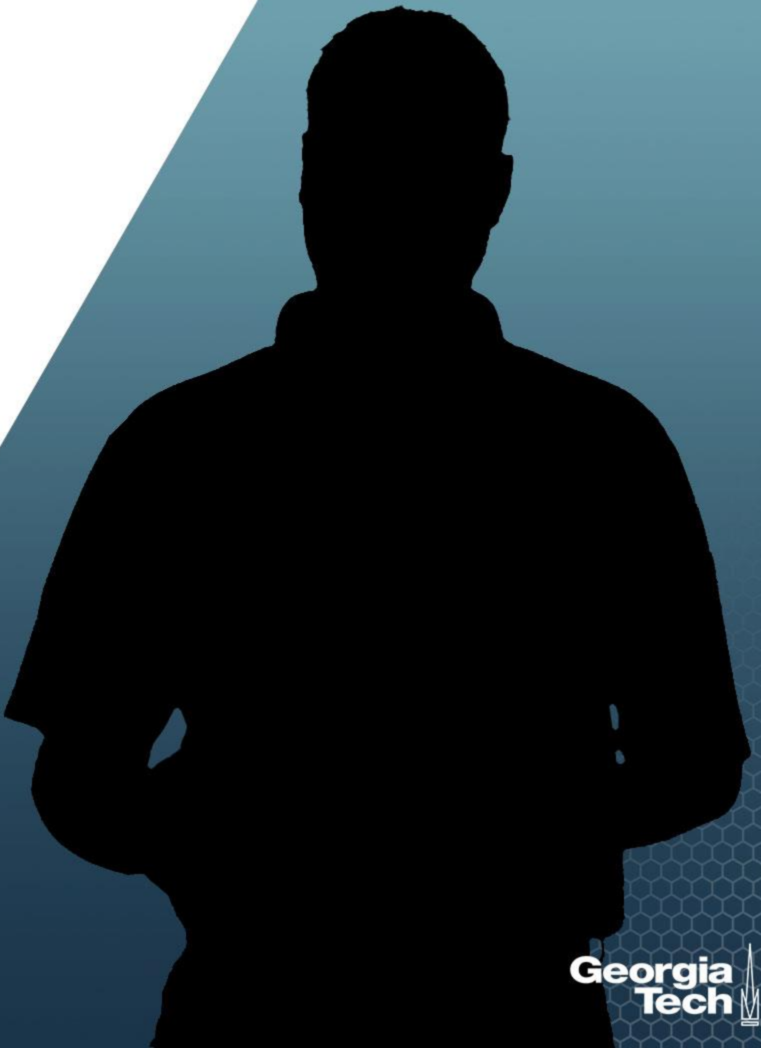
- Given all other predictors in the model:
  - Percent of students taking SAT from a public school and family income of test takers are not statistically significantly associated to SAT score
  - A $100 increase in the expenditure on secondary schools results in a 2.56-point increase in the SAT score
  - One additional year that test takers had in social sciences, natural sciences, and humanities leads to a 17.2-point 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
  - For example, Connecticut moves from 35th to 1st, Massachusetts from 41st to 4th, and New York from 36th to 5th

Summary

Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Predicting Demand for Rental Bikes: Exploratory Data Analysis

About This Lesson

Predicting Demand for Rental Bikes

Bike sharing systems are of great interest due to their important role in traffic management.

Dataset: Historical data for years 2011-2012 for the bike sharing system in Washington D.C.

Data Source: UCI Machine Learning Repository

***Acknowledgement**: This example was prepared with support from students in the Masters of Analytics program, including Naman Arora, Puneeth Banisetti, Mani Chandana Chalasani, Joseph (Mike) Tritchler and Kevin West*

Georgia Tech

Response & Predicting Variables

The response variable is:

Y (*Cnt*): Total bikes rented by both casual & registered users together

The qualitative predicting variables are:

Season: Season which the observation is made (1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall)
Yr: Year on which the observation is made
Mnth: Month on which the observation is made
Hr: Day on which the observation is made (0 through 23)
Holiday: Indictor of a public holiday or not (1 = public holiday, 0 = not a public holiday)
Weekday: Day of week (0 through 6)
Weathersit: Weather condition (1 = Clear, Few clouds, Partly cloudy, Partly cloudy, 2 = Mist & Cloudy, Mist & Broken clouds, Mist & Few clouds, Mist, 3 = Snow, Rain, Thunderstorm & Scattered clouds, Ice Pallets & Fog)

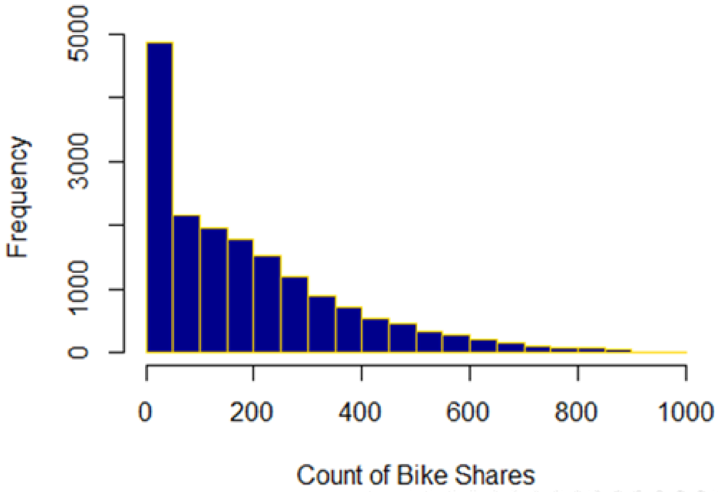
The quantitative predicting variables are:

Temp: Normalized temperature in Celsius
Atemp: Normalized feeling temperature in Celsius
Hum: Normalized humidity
Windspeed: Normalized wind speed

Exploratory Data Analysis in R

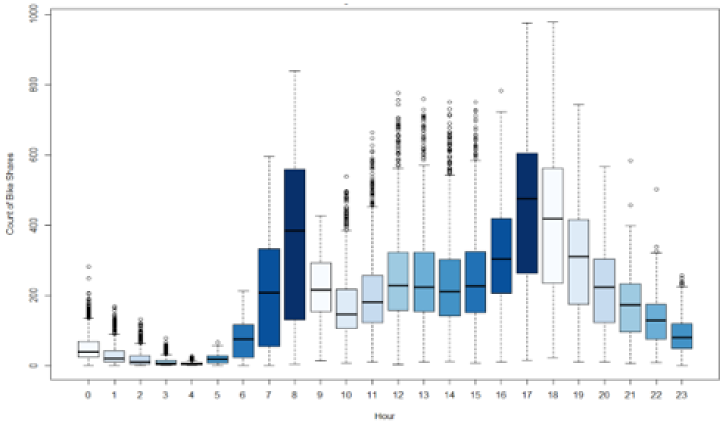
```
## Read data using read.csv
data<-read.csv("Bikes.csv")
dim(data)[1] # how many observations?
[1] 17379
## Test initial intuitions/assumptions on the behavior of the data
hist(data$cnt,
     main="",
     xlab="Count of Bike Shares",
     border="gold",
     col="darkblue")
```

The frequency of zero bike shares is high, which skews the demand data.



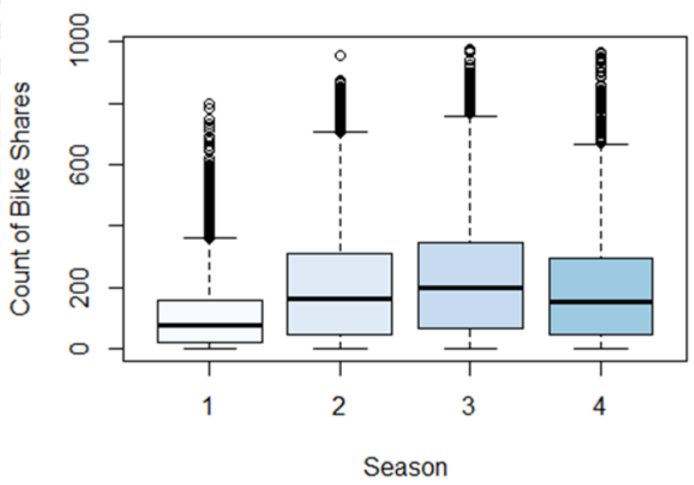
Exploratory Data Analysis in R (cont’d)

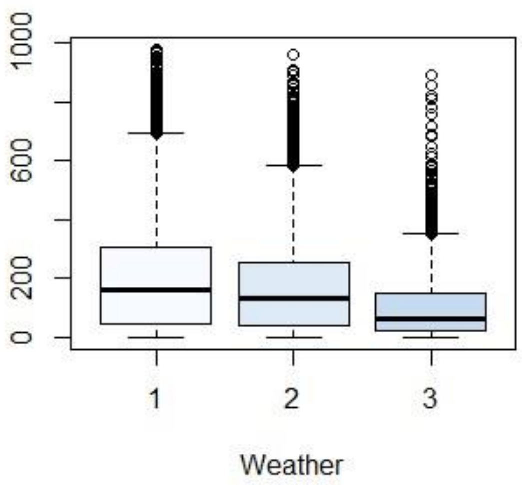
```
## Evaluate intuitions/assumptions on the behavior of the data and understand patterns
boxplot(cnt~hr,
        main="",
        xlab="Hour",
        ylab="Count of Bike Shares",
        col=blues9,
        data=data)
```



The number of bike shares between hour 0 and hour 6 is low. The majority activity as expected is focused between hour 7 and hour 23, peaking at hour 8 and hour 17.

Exploratory Data Analysis in R (cont’d)



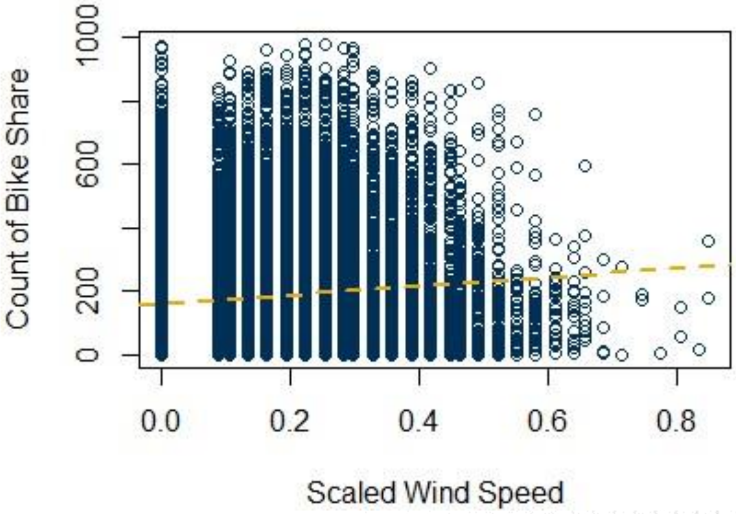


The number of bikes rented during winter are the lowest.

The number of bikes decreases as the weather becomes unfavorable.

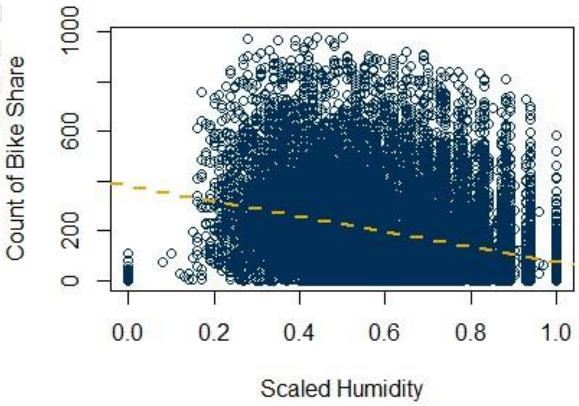
Exploratory Data Analysis in R (cont'd)

```
plot(data$windspeed,
      data$cnt,
      xlab='Scaled Wind Speed',
      ylab='Count of Bike Share',
      main='',  col="darkblue")

abline(lm(cnt~windspeed, data=data),
      col=buzzgold,
      lty=2, lwd=2)
```

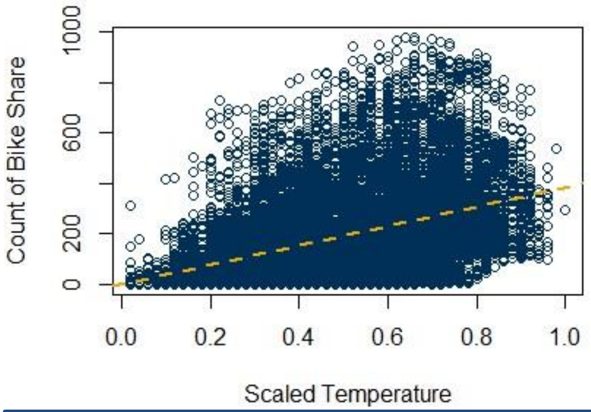


The count of rental bikes seems to decrease as windspeed increases.

Exploratory Data Analysis in R (cont'd)



The count of rental bikes seems to decrease as humidity increases although the demand varies within similar ranges at varying humidity levels.



The count of rental bikes seems to increase as temperature increases however with much wider variability at larger temperature levels.

Preparing the Data

```
## Divide data into train and test data
# Set seed for reproducibility
set.seed(9)
# Test Train split
sample_size = floor(0.8*nrow(data))
picked = sample(seq_len(nrow(data)),size = sample_size)

# Remove irrelevant columns from training data
train = data[picked,]
train <- train[-c(1,2,9,15,16)]

## Converting the numerical cateogrical variables to predictors
train$season = as.factor(train$season)
train$yr = as.factor(train$yr)
train$mnth = as.factor(train$mnth)
train$hr = as.factor(train$hr)
train$holiday = as.factor(train$holiday)
train$weekday = as.factor(train$weekday)
train$weathersit = as.factor(train$weathersit)
```

Fitting the Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-79.4356	7.4390	-10.678	< 2e-16 ***
season2	34.9268	5.4110	6.455	1.12e-10 ***
season3	27.0055	6.4438	4.191	2.80e-05 ***
season4	65.3435	5.4690	11.948	< 2e-16 ***
yr1	85.3415	1.7487	48.804	< 2e-16 ***
mnth2	4.1666	4.3853	0.950	0.342060
mnth3	16.4733	4.9267	3.344	0.000829***
mnth4	12.5834	7.3038	1.723	0.084936 .
mnth5	26.4616	7.8357	3.377	0.000735 ***
mnth6	11.5056	8.0535	1.429	0.153131
mnth7	-7.8872	9.0547	-0.871	0.383736

⋮

 # Applying multiple linear regression model

model1 = lm(cnt ~ .,data=train)
summary(model1)

In the full output there are 51 predictor rows in addition to the intercept.

Statistical Significance

```
# Applying multiple linear regression model
model1 = lm(cnt ~ .,data=train)
summary(model1)

# Find insignificant values
which(summary(model1)$coeff[,4]>0.05)
```

mnth2	mnth4	mnth6	mnth7	mnth8	mnth11	mnth12
6	8	10	11	12	15	16

Statistically insignificant variables at 0.05 significance level:

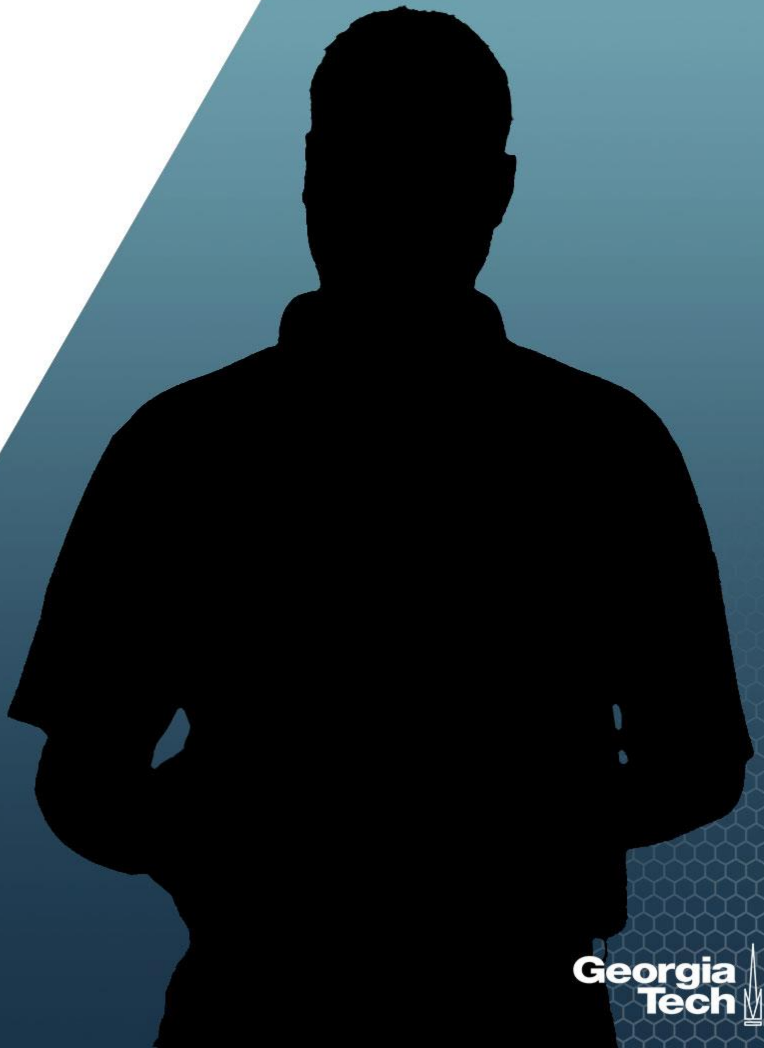
- Month-2, month-4, month-6, month-7, month-8, month-11, month-12 are not statistically different from month-1 (baseline)

Summary

Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering



Predicting Demand for Rental Bikes: Regression Analysis

Georgia Tech

About This Lesson

Linear Regression Analysis in R

 # Applying multiple linear regression model
*model1 = **lm**(cnt ~ .,data=train)*
***summary**(model1)*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-79.4356	7.4390	-10.678	< 2e-16 ***
season2	34.9268	5.4110	6.455	1.12e-10 ***
season3	27.0055	6.4438	4.191	2.80e-05 ***
season4	65.3435	5.4690	11.948	< 2e-16 ***
yr1	85.3415	1.7487	48.804	< 2e-16 ***
mnth2	4.1666	4.3853	0.950	0.342060
mnth3	16.4733	4.9267	3.344	0.000829***
mnth4	12.5834	7.3038	1.723	0.084936 .
mnth5	26.4616	7.8357	3.377	0.000735 ***
mnth6	11.5056	8.0535	1.429	0.153131

⋮

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 101.8 on 13851 degrees of freedom
Multiple R-squared: 0.6852,   Adjusted R-squared: 0.684
F-statistic: 591 on 51 and 13851 DF, p-value: < 2.2e-16

In the full output there are 51 predictor rows in addition to the intercept.

$\hat{\sigma} = \boxed{101.8}$
df = n-p-1 = 13,903 - 51 - 1 = $\boxed{13,851}$
$R^2 \approx \boxed{0.6852} \approx$ 68.5% variability explained

Coding Dummy Variables in R

```
## Create Dummy Variables
weathersit  = data$weathersit
weathersit.1 = rep(0,length(weathersit))
weathersit.1[weathersit==1] = 1
weathersit.2 = rep(0,length(weathersit))
weathersit.2[weathersit==2] = 1
weathersit.3 = rep(0,length(weathersit))
weathersit.3[weathersit==3] = 1

## Include all dummy vars without intercept
fit.1 = lm(cnt ~ weathersit.1 + weathersit.2 +weathersit.3 - 1)

## Include 3 dummy variables with intercept
fit.2 = lm(cnt ~ weathersit.1 + weathersit.2)

## Use categorical variable
weathersit  = as.factor(data$weathersit)
fit.3 = lm(cnt ~ weathersit)
```

summary(fit.1)

	Estimate	Std. Error	t value	Pr(>|t|)
weathersit.1	204.869	1.680	121.97	<2e-16 ***
weathersit.2	175.165	2.662	65.80	<2e-16 ***
weathersit.3	111.501	4.758	23.43	<2e-16 ***

summary(fit.2)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	111.501	4.758	23.43	<2e-16 ***
weathersit.1	93.369	5.046	18.50	<2e-16 ***
weathersit.2	63.665	5.452	11.68	<2e-16 ***

summary(fit.3)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	204.869	1.680	121.972	<2e-16 ***
weathersit2	-29.704	3.148	-9.437	<2e-16 ***
weathersit3	-93.369	5.046	-18.503	<2e-16 ***

Georgia Tech

Coding Dummy Variables in R

```
## Create Dummy Variables
weathersit  = data$weathersit
weathersit.1 = rep(0,length(weathersit))
weathersit.1[weathersit==1] = 1
weathersit.2 = rep(0,length(weathersit))
weathersit.2[weathersit==2] = 1
weathersit.3 = rep(0,length(weathersit))
weathersit.3[weathersit==3] = 1

## Include all dummy vars without intercept
fit.1 = lm(cnt ~ weathersit.1 + weathersit.2 +weathersit.3 - 1)

## Include 3 dummy variables with intercept
fit.2 = lm(cnt ~ weathersit.1 + weathersit.2)

## Use categorical variable
weathersit  = as.factor(data$weathersit)
fit.3 = lm(cnt ~ weathersit)
```

summary(fit.1)

	Estimate	Std. Error	t value	Pr(>|t|)
weathersit.1	204.869	1.680	121.97	<2e-16 ***
weathersit.2	175.165	2.662	65.80	<2e-16 ***
weathersit.3	111.501	4.758	23.43	<2e-16 ***

summary(fit.2)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	111.501	4.758	23.43	<2e-16 ***
weathersit.1	93.369	5.046	18.50	<2e-16 ***
weathersit.2	63.665	5.452	11.68	<2e-16 ***

summary(fit.3)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	204.869	1.680	121.972	<2e-16 ***
weathersit2	-29.704	3.148	-9.437	<2e-16 ***
weathersit3	-93.369	5.046	-18.503	<2e-16 ***

Codding Dummy Variables

R Sets the “first” class as being the baseline

- If a different class is the baseline, either use dummy variables or specify with ‘contr.treatment’

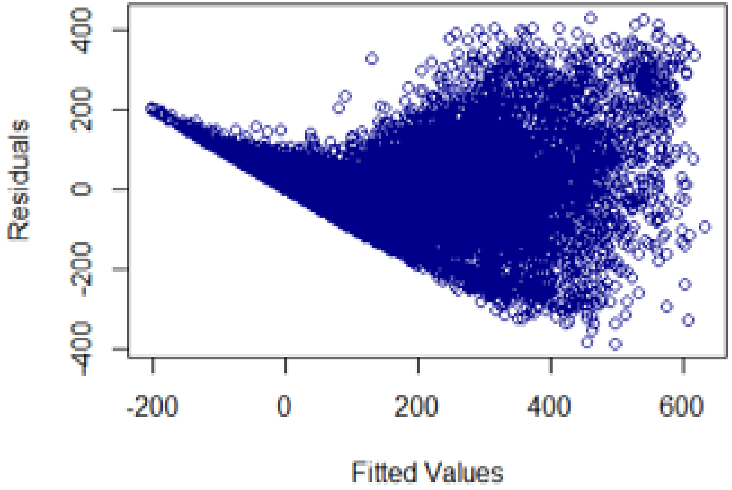
Be careful when using a model without intercept in R!

- No baseline comparison

Goodness of Fit: Constant Variance Assumption

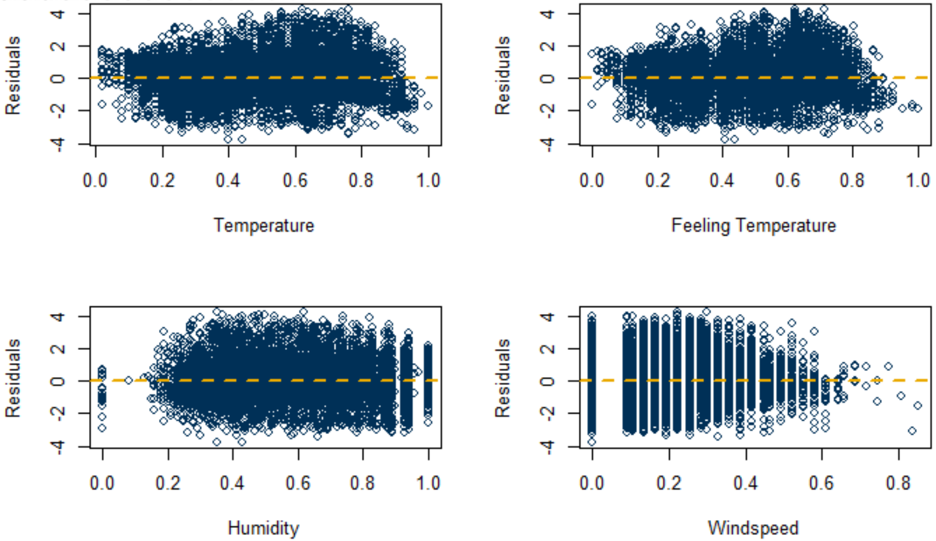
Fitting the model
 # Creating scatterplot of residuals vs fitted values

```
resids = rstandard(model1)
fits = model1$fitted
plot(fits,
     resids,
     xlab="Fitted Values",
     ylab="Residuals",
     main="Scatterplot",
     col="darkblue")
```



- The constant variance assumption does not hold -- the variance increases when moving from lower to higher fitted values.
- The residuals, at low y values, seem to follow a straight-line pattern. The linear pattern in the beginning suggests that the response variable stays constant for a range of predictor values.

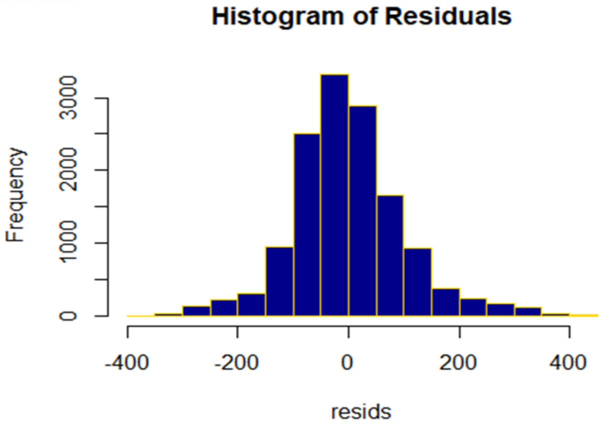
Goodness of Fit: Linearity Assumption



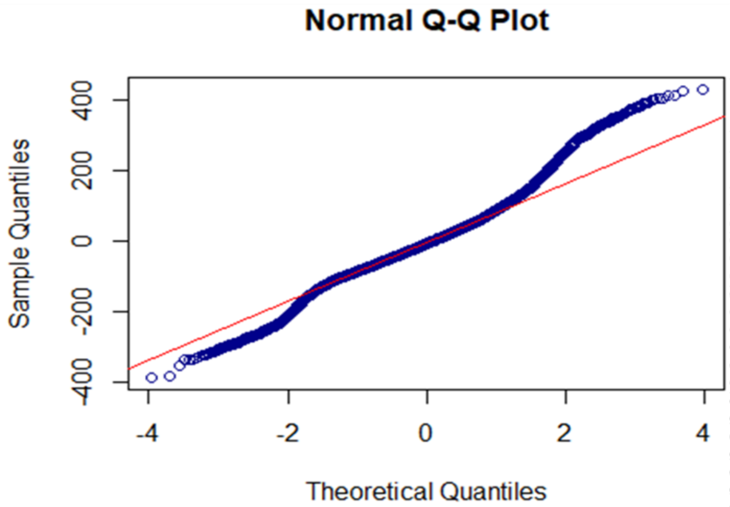
The residuals do not vary with the any of the numeric predicting variables. No transformation of the predicting variable is needed.

Goodness of Fit: Normality Assumption

```
## Checking normality
# histogram
hist(resids,
     nclass=20,
     col="darkblue",
     border="gold",
     main="Histogram of residuals")
```



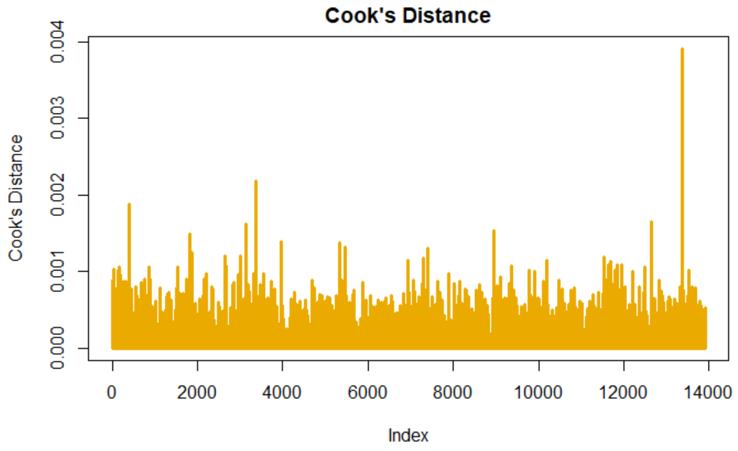

```
# q-q plot
qqnorm(resids,
     col="darkblue")
qqline(resids,
     col="red")
```



Goodness of Fit: Outliers

 # Cook's Distance

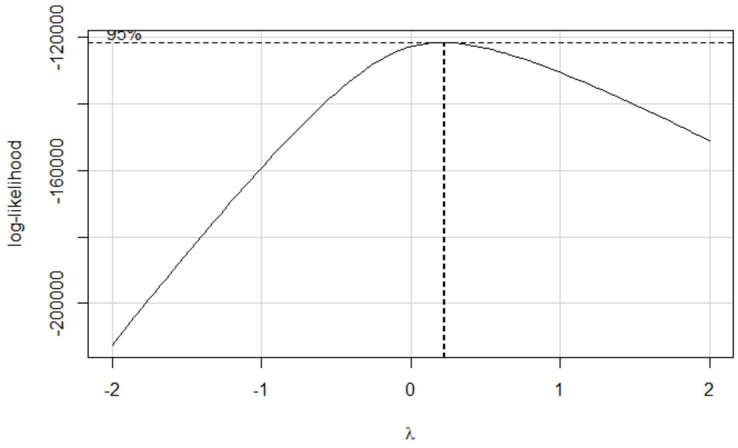
```
cook = cooks.distance(model1)

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



There is one observation with a Cook's Distance noticeably higher than the other observations. However, its Cook's distance is close to 0.004, suggesting that there are likely no outliers.

Transformation of the Response Variable

```
## Box Cox transformation
bc <- boxcox(model1)
lambda <- bc$x[which(bc$y==max(bc$y))]

## Fitting the model with square root transformation
model2<-lm(sqrt(cnt)~.,data=train)
summary(model2)
```



- The optimal value of lambda or the power provided by the Box Cox transformation is 0.22.
- Generally, when the response data consist of count data, a theoretically recommended transformation is the square root, corresponding to a 0.5 power transformation.

Georgia Tech

Regression Analysis after Transformation

Fitting the model with square root transformation

*model2<-**lm**(**sqrt**(cnt)~.,data=train)*
***summary**(model2)*

Find Insignificant Values

***which**(**summary**(model2)$coeff[,4]>0.05)*

mnth2	mnth4	mnth6	mnth7	mnth8	mnth10	mnth11	weekday1
6	8	10	11	12	14	15	41

Multicollinearity

***vif**(model2)*

	GVIF	Df	GVIF^(1/(2*Df))
season	165.308	3	2.343
yr	1.025	1	1.012
mnth	323.778	11	1.300
hr	1.771	23	1.012
holiday	1.121	1	1.059
weekday	1.137	6	1.011
weathersit	1.386	2	1.085
temp	51.283	1	7.161
atemp	43.748	1	6.614
hum	1.921	1	1.386
windspeed	1.251	1	1.118

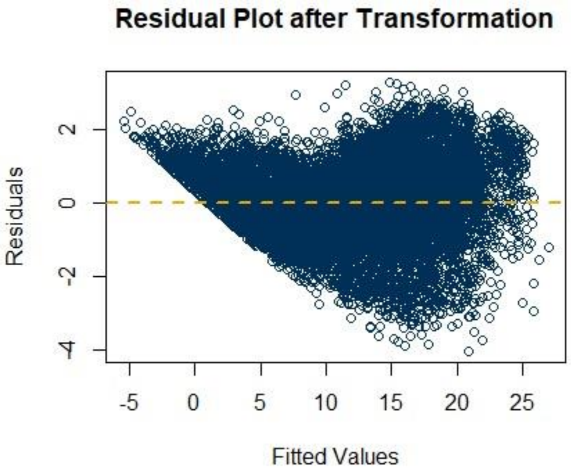
Model Performance

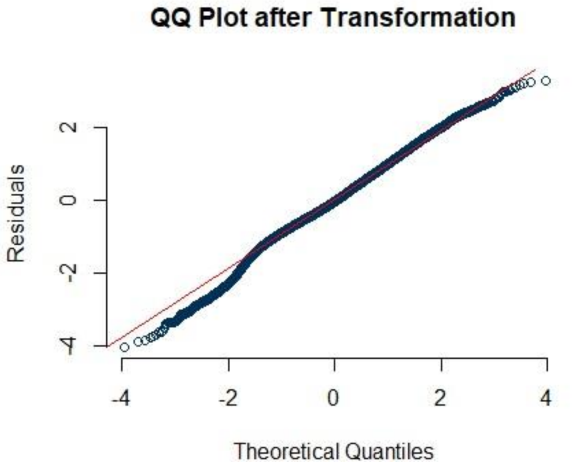
***summary**(model2)**$**r.squared*

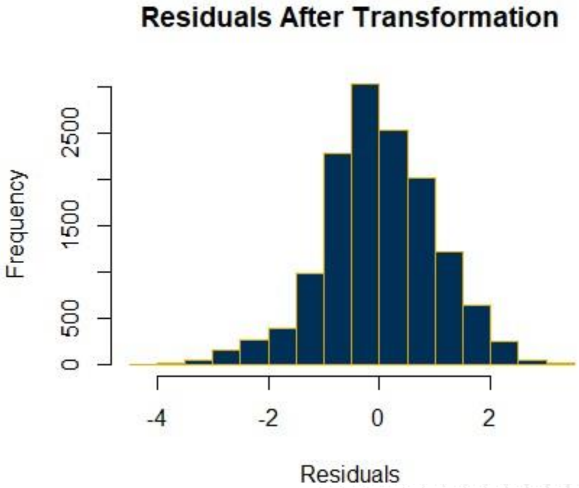
[1] 0.786535

As VIFs of the season, mnth, temp, atemp factors are greater than max(10, $1/(1-R^2)$), it indicates there is a problem of multicollinearity in the linear model. So, we should not use all the predictors in the model.

Goodness of Fit after Transformation





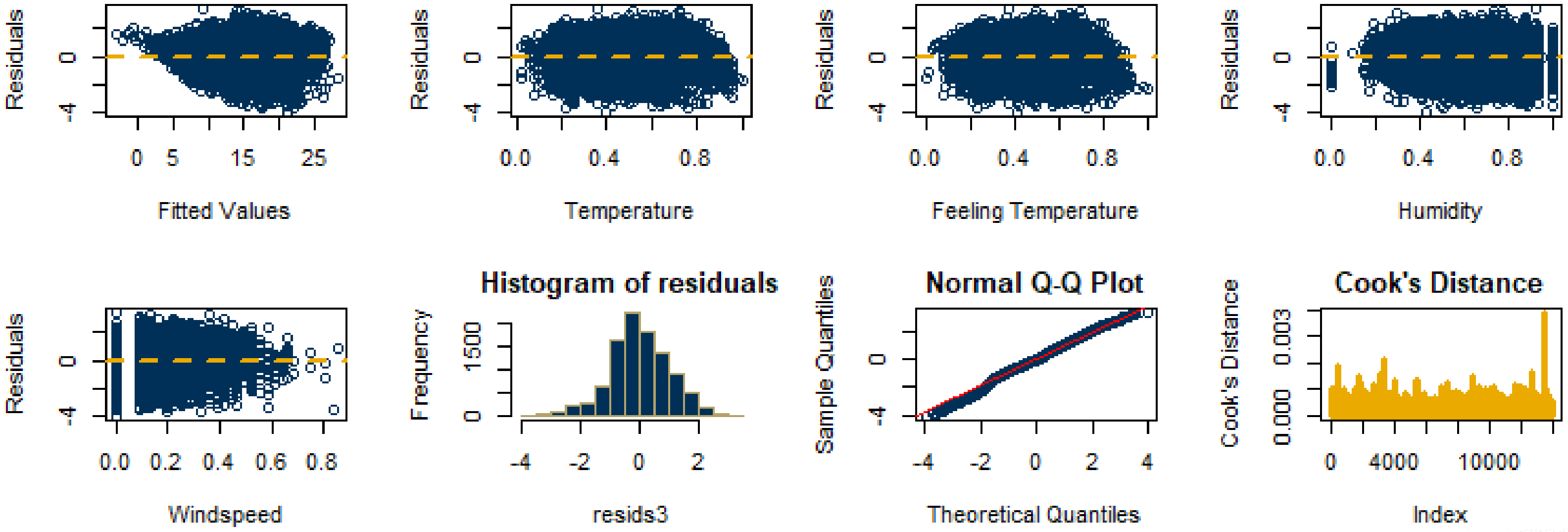


The constant variance assumption is still violated. The transformation has not improved the goodness of fit even though the model performance is better with respect to the coefficient of determination.

Removing Low Demand Data

```
## Remove data for hours 0-6
hrs <- as.numeric(data$hr)
data_red <- data[which(hrs>=7),]
## Test/Train Data
set.seed(9) # for uniformity
sample_size  <- floor(0.8*nrow(data_red))
picked  <- sample(seq_len(nrow(data_red)),size = sample_size)
train_red  <- data_red[picked, -c(1,2,9,15,16)]
test_red  <- data_red[-picked, -c(1,2,9,15,16)]

## Fitting the model with square root transformation
model3<-lm(sqrt(cnt)~.,data=train_red)
summary(model3)$r.squared
[1] 0.6579021
df<-which(summary(model3)$coeff[,4]>0.05)

   mnth7 mnth11 mnth12   hr14   hr15   hr20
      11     15     16     23     24     29
```

Goodness of Fit without Low Demand Data



- The constant variance assumption is still violated even for the model without the low demand data and with the transformed response.
- The implication of the constant variation assumption violation is that the uncertainty in predicting bike demand when in high demand will be higher than estimated using the multiple regression models in this lesson.

Summary

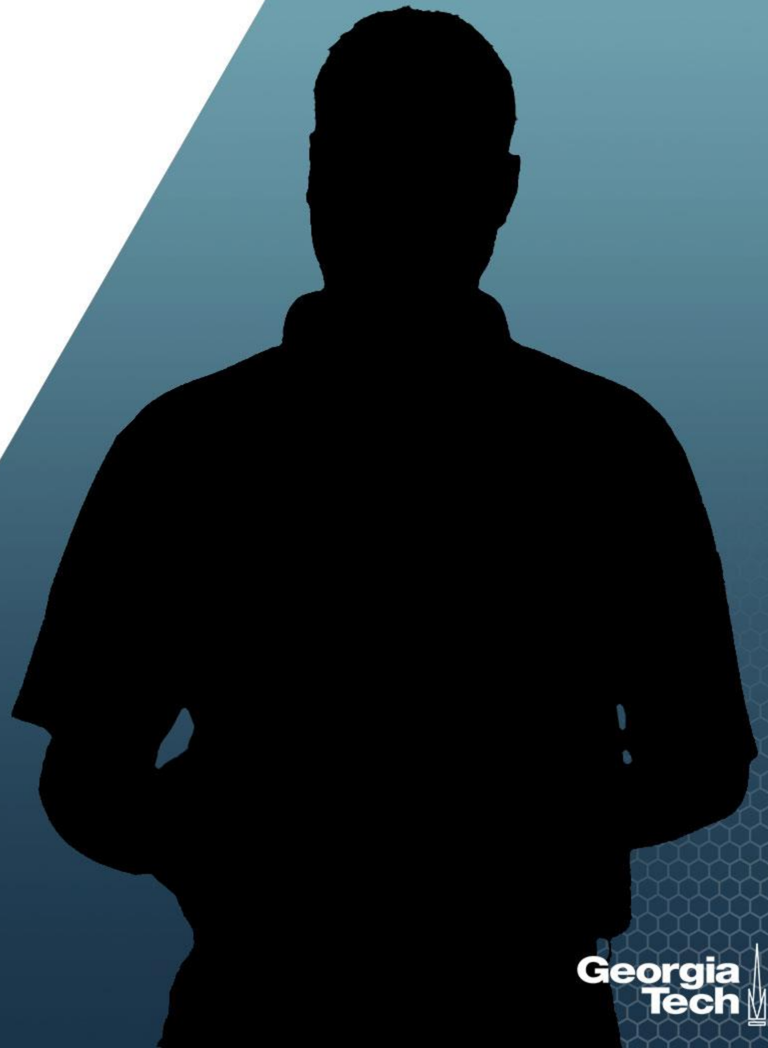
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering



Predicting Demand for Rental Bikes: Prediction, Interpretation

Georgia Tech

About This Lesson

Prediction

```
## Read New Data (Test Data)
test=data[-picked,]
test <- test[-c(1,2,9,15,16)]

## Prepare the test data the same as the training data
## Convert the numerical categorical variables to
predictors in the test data
test$season = as.factor(test$season)
test$yr = as.factor(test$yr)
test$mnth = as.factor(test$mnth)
test$hr = as.factor(test$hr)
test$holiday = as.factor(test$holiday)
test$weekday = as.factor(test$weekday)
test$weathersit = as.factor(test$weathersit)

## Build a prediction for model1 with the test data
# Specify whether a confidence or prediction interval
pred = predict(model1, test, interval = 'prediction')
```

Apply similar R code for the other two models.

Prediction (cont'd)

Read New Data (Test Data)

```
test=data[-picked,]
test <- test[-c(1,2,9,15,16)]
```

Prepare the test data the same as the training data
Convert the numerical categorical variables to predictors in the test data

```
test$season = as.factor(test$season)
test$yr = as.factor(test$yr)
test$mnth = as.factor(test$mnth)
test$hr = as.factor(test$hr)
test$holiday = as.factor(test$holiday)
test$weekday = as.factor(test$weekday)
test$weathersit = as.factor(test$weathersit)
```

Build a prediction for model1 with the test data
 # Specify whether a confidence or prediction interval

```
pred = predict(model1, test, interval = 'prediction')
```

Apply similar R code for the other two models.

Prediction Output

	Fit	lwr	upr
6	-104.3303581	-3.038988e+02	95.238132
9	239.0013629	3.941481e+01	438.587917
30	-82.5358710	-2.822639e+02	117.192193
35	58.5579012	-1.410152e+02	258.130976
38	22.5421861	-1.770914e+02	222.175777
44	102.8402463	-9.671724e+01	302.397729
47	-40.1522581	-2.396963e+02	159.391774
48	-69.0241889	-2.685984e+02	130.549996
63	334.4570824	1.349013e+02	534.012852
65	176.2306906	-2.336174e+01	375.823119
68	-31.2412576	-2.308027e+02	168.320195
69	-45.1215422	-2.446761e+02	154.433034
78	69.0246421	-1.305309e+02	268.580201
82	99.6552263	-9.989334e+01	299.203794
85	176.4458539	-2.309072e+01	375.982429
87	289.1456026	8.960119e+01	488.690014

Prediction Accuracy

Prediction Error Measures

- Compare observed response Y_i to the predicted Y_i*
- Mean squared prediction error (MSPE) = $\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i*)^2$
- Mean absolute prediction errors (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|Y_i - Y_i*|$
- Mean absolute percentage error (MAPE) = $\frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - Y_i*|}{|Y_i|}$
- Precision error (PM) = $\frac{\sum_{i=1}^{n}(Y_i - Y_i*)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- Confidence Interval error (CIM) = $\frac{1}{n}\sum_{i=1}^{n} I(Y_i \notin CI)$

Prediction Error Measure Insights

Mean squared prediction error (MSPE)

- Appropriate for linear regression model prediction but depends on scale and it is sensitive to outliers

Mean absolute prediction errors (MAE)

- Not appropriate for linear regression model prediction and depends on scale but robust to outliers

Mean absolute percentage error (MAPE)

- Not appropriate for linear regression model prediction but it does not depend on scale and robust to outliers

Precision error (PM)

- Appropriate for linear regression model and does not depend on scale

Confidence Interval error (CIM)

Prediction Error Measure Insights

Mean squared prediction error (MSPE)

- Appropriate for linear regression model prediction but depends on scale and it is sensitive to outliers

Mean absolute prediction errors (MAE)

- Not appropriate for linear regression model prediction and depends on scale but robust to outliers

Mean absolute percentage error (MAPE)

- Not appropriate for linear regression model prediction but it does not depend on scale and robust to outliers

Precision error (PM)

- Appropriate for linear regression model and does not depend on scale

Confidence Interval error (CIM)

While MAE and MAPE are commonly used to evaluate prediction error, I recommend using the precision measure.

-- Regression models are estimated using by minimizing sum of least squares hence the accuracy error shall be best of squared differences not absolute differences

Prediction Accuracy: Model 1

Save Predictions to compare with observed data

```
pred1 <- predict(model1, test, interval = 'prediction')
test.pred1 <- pred1[,1]
test.lwr1 <- pred1[,2]
test.upr1 <- pred1[,3]

# Mean Squared Prediction Error (MSPE)
mean((test.pred1-test$cnt)^2)
[1] 10304.95
# Mean Absolute Prediction Error (MAE)
mean(abs(test.pred1-test$cnt))
[1] 74.52024
# Mean Absolute Percentage Error (MAPE)
mean(abs(test.pred1-test$cnt)/test$cnt)
[1] 2.724609
# Precision Measure (PM)
sum((test.pred1-test$cnt)^2)/sum((test$cnt-mean(test$cnt))^2)
[1] 0.3101164
# CI Measure (CIM)
(sum(test$cnt<test.lwr1)+sum(test$cnt>test.upr1))/nrow(test)
[1] 0.06904488
```

Accuracy Measures

MSPE = $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overset{*}{Y}_i\right)^2$

MAE = $\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \overset{*}{Y}_i\right|$

MAPE = $\frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i-\overset{*}{Y}_i\right|}{|Y_i|}$

PM = $\frac{\sum_{i=1}^{n}\left(Y_i-\overset{*}{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$

Prediction Accuracy
MSPE = 10304
MAE = 74.52
MAPE = 2.72
PM = 0.31
CIM = 0.069

Prediction Accuracy: Model 3

```
## Save Predictions to compare with observed data
pred3 <- predict(model3, test_red, interval = 'prediction')
test.pred3 <- pred3[,1]^2
test.lwr3 <- pred3[,2]^2
test.upr3 <- pred3[,3]^2

 # Mean Squared Prediction Error (MSPE)
mean((test.pred3-test_red$cnt)^2)
 [1] 11271.78
# Mean Absolute Prediction Error (MAE)
mean(abs(test.pred3-test_red$cnt))
[1] 78.67701
# Mean Absolute Percentage Error (MAPE)
mean(abs(test.pred3-test_red$cnt)/test_red$cnt)
[1] 0.5172032
# Precision Measure (PM)
sum((test.pred3-test_red$cnt)^2)/sum((test_red$cnt-mean(test_red$cnt))^2)
[1] 0.316168
# CI Measure (CIM)
(sum(test_red$cnt<test.lwr3)+sum(test_red$cnt>test.upr3))/nrow(test_red)
[1] 0.060984
```

Prediction Accuracy
MSPE = 11271
MAE = 78.67
MAPE = 0.517
PM = 0.361
CIM = 0.061

Model Comparison

Model	MSPE	Precision.Measure	Adjusted.R.Squared	R squared
Full MLR	10304.95	0.310	0.684	0.685
MLR (sqrt transformation)	8955.41	0.271	0.784	0.785
MLR (sqrt transformation-no low demand data)	11271.78	0.362	0.656	0.658

- **The model with the square-root transformation outperforms the other models in terms of predictive power as reflected in the Precision Measure and R squared.**
- **The constant variance assumption is violated across all models.**

Georgia Tech

Summary

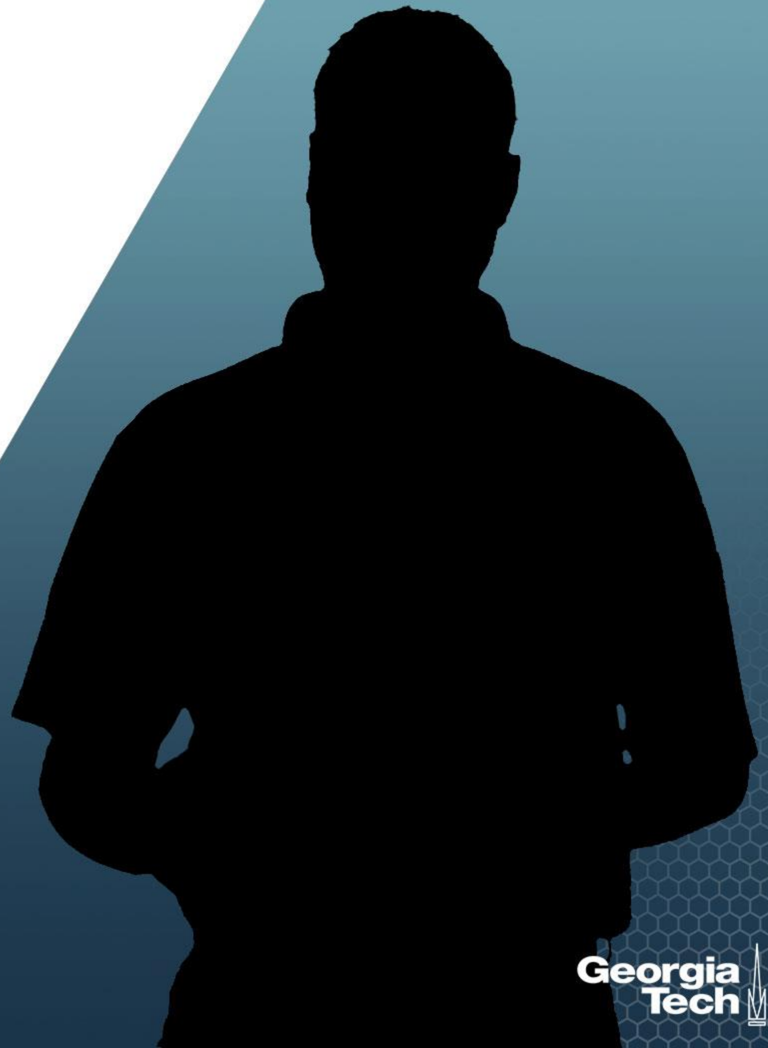
Regression Analysis

Multiple Linear Regression

Nicoleta Serban, Ph.D.

Professor

School of Industrial and Systems Engineering



Predicting Demand for Rental Bikes: P-values and Large Sample Size

About This Lesson

The P-value Problem: Basis Statistics

- Basic statistics under large sample size:

$$Z_1, \ldots, Z_n \sim N(\mu, \sigma^2) \Rightarrow \bar{Z} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Hypothesis testing for the mean:

$H_0\!: \mu = 0$ vs. $H_A\!: \mu \neq 0$

- P-value and sample size:

$\boldsymbol{p - value = 2P(Z > \sqrt{n}|\frac{\bar{Z}-0}{\sigma}|)}$ **is approximately *0* with *n* very large**

"Inflated" Significance: Conclusions based on small-sample statistical inferences using large samples can be misleading.

Samples Can Make the Insignificant...Significant!

The P-value Problem: Regression Analysis

- Hypothesis testing for the statistical significance of the regression coefficients:

  $H_0\colon \beta_i = 0 \text{ vs. } H_A\colon \beta_i \neq 0$

- P-value and sample size:

  $p - value = 2P(T_{n-p-1} > |t - value|)$
  is approximately *0* with *n* very large

- Misleadingly, reject the null hypothesis of zero coefficient – all or most relationship are statistically significant.

"Inflated" Statistical Significance: Conclusions based on small-sample statistical inferences on the regression coefficients using large samples can be misleading.

The P-value Problem: Approach

- Sub-sampling: Sample the observed data, e.g. 10-20% of the sample size
- Apply the regression model to each sub-sampled data
- Repeat for B times, e.g. B=100
- **Output**:

  Sub-sample 1: $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}, \dots, \hat{\beta}_{p,1}$ & corresponding p-values $pv_{0,1}, pv_{1,1}, \dots, pv_{p,1}$

  Sub-sample 2: $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{p,2}$ & corresponding p-values $pv_{0,,2}, pv_{1,2}, \dots, pv_{p,2}$

  …..

  Sub-sample B: $\hat{\beta}_{0,B}, \hat{\beta}_{1,B}, \dots, \hat{\beta}_{p,B}$ & corresponding p-values $pv_{0,B}, pv_{1,B}, \dots, pv_{p,B}$
- Empirical distributions of the regression coefficients and the p-values

The P-value Problem: Approach

- Sub-sampling: Sample the observed data, e.g. 10-20% of the sample size
- Apply the regression model to each sub-sampled data
- Repeat for B times, e.g. B=100
- **Output**:

  Sub-sample 1: $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}, \dots, \hat{\beta}_{p,1}$ & corresponding p-values $pv_{0,1}, pv_{1,1}, \dots, pv_{p,1}$

  Sub-sample 2: $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{p,2}$ & corresponding p-values $pv_{0,,2}, pv_{1,2}, \dots, pv_{p,2}$

  …..

  Sub-sample B: $\hat{\beta}_{0,B}, \hat{\beta}_{1,B}, \dots, \hat{\beta}_{p,B}$ & corresponding p-values $pv_{0,B}, pv_{1,B}, \dots, pv_{p,B}$
- Empirical distributions of the regression coefficients and the p-values

Theoretical Underpinning:
- Statistical significance (or lack of it) can be identified based on the distribution of the p-values; specifically, if the empirical distribution is approximately uniform between 0 and 1, then we don't have statistical significance.
- Statistical significance (or lack of it) can be identified based on the confidence interval of the regression coefficient derived from the empirical distribution.

The P-value Problem: Approach (cont’d)

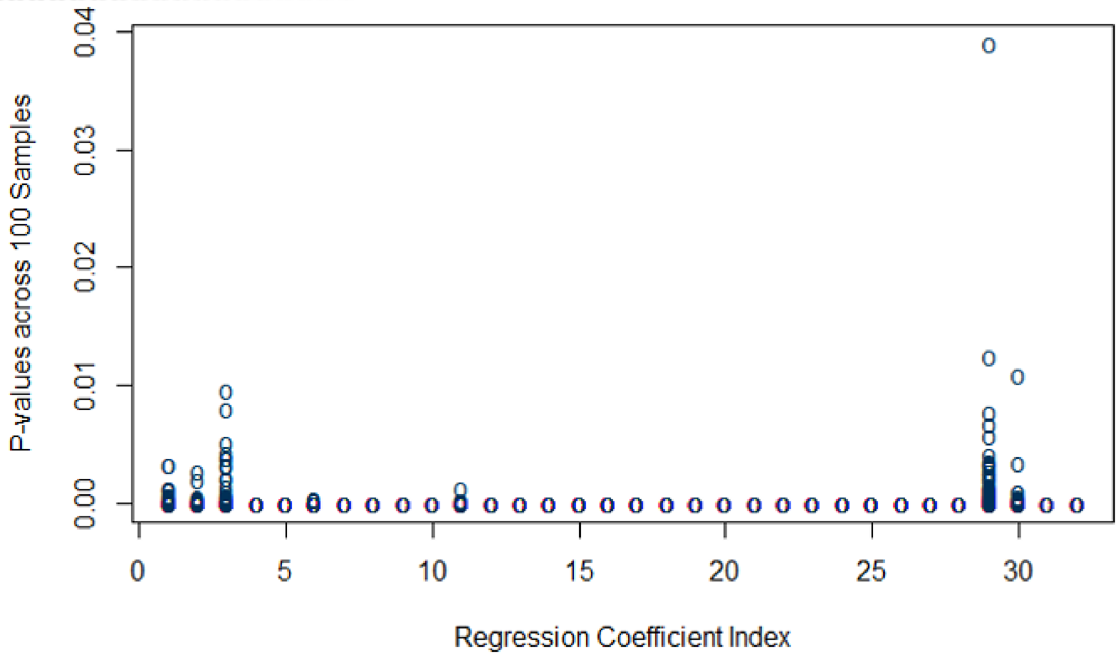
```
## Approach: Subsample 40% of the initial data sample & repeat 100 times
count = 1
n = nrow(train)
B = 100
ncoef = dim(summary(model1)$coeff)[1]
pv_matrix = matrix(0,nrow = ncoef,ncol = B)
while (count <= B) {
        # 40% random sample of indices
        subsample = sample(n, floor(n*0.4), replace=FALSE)
        # Extract the random subsample data
        subdata = train[subsample,]
        # Fit the regression for each subsample
        submod = lm(sqrt(cnt)~.,data=subdata)
        # Save the p-values
        pv_matrix[,count] = summary(submod)$coeff[,4]
        # Increment to the next subsample
        count = count + 1
}
# Count pvalues smaller than 0.01 across the 100 (sub)models
alpha = 0.01
pv_significant = rowSums(pv_matrix < alpha)
```

Statistical Significance

```
## Which regression coefficients are statistically significant?
idx_scoef = which(pv_significant>=95)
## Show the p-values of the significant coefficients in model2
cbind(summary(model2)$coeff[idx_scoef,c(1,4)],
      Freq=pv_significant[idx_scoef])
## Plot the 100 p-values of the significant coefficients
matplot(pv_matrix[idx_scoef,],
     xlab="Regression Coefficient Index",
     ylab="P-values across 100 Samples",
     type="p",
     pch="o",
     col=gtblue)
```

	Estimate	Pr(>|t|)	Freq
(Intercept)	1.670	0	100
season2	1.370	0	100
season3	1.380	0	100
season4	2.720	0	100
yr1	2.800	0	100
hr1	-1.630	0	100
hr2	-2.570	0	100
hr3	-3.770	0	100
hr4	-4.190	0	100
hr5	-2.360	0	100
hr6	1.480	0	100
hr7	6.820	0	100
hr8	10.700	0	100
hr9	7.500	0	100
hr10	5.440	0	100
hr11	6.210	0	100
hr12	7.450	0	100
hr13	7.310	0	100
hr14	6.770	0	100
hr15	7.090	0	100
hr16	9.020	0	100
hr17	12.700	0	100
hr18	12.100	0	100
hr19	9.440	0	100
hr20	7.020	0	100
hr21	5.380	0	100
hr22	3.860	0	100
hr23	2.000	0	100
holiday1	-0.986	0	98
weekday5	0.723	0	99
weathersit3	-2.650	0	100
hum	-2.580	0	100

Statistical Significance (cont'd)



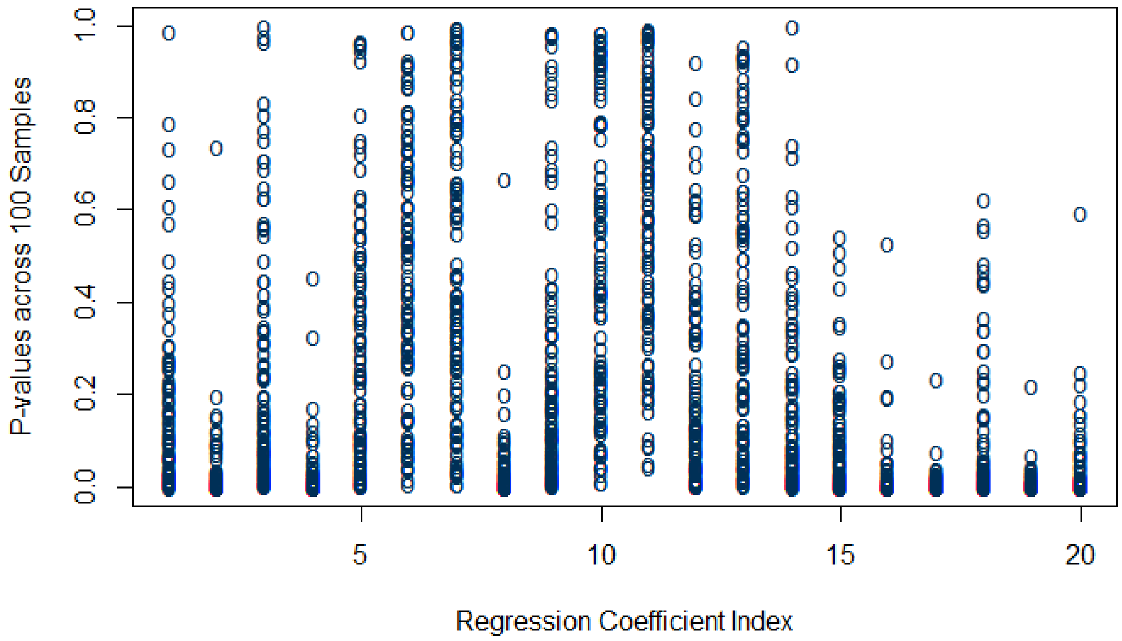
Statistical significance: Most P-values are small across all sub-samples

Lack Statistical Significance

```
# Which regression coefficients are not statistically significant?
idx_icoef = which(pv_significant<85)
# Show the p-values of the non-significant coefficients in model2
cbind(summary(model2)$coeff[idx_icoef,c(1,4)],
      Freq=pv_significant[idx_icoef])
# Plot the 100 p-values of the non-significant coefficients
matplot(pv_matrix[idx_icoef,],
     xlab="Regression Coefficient Index",
     ylab="P-values across 100 Samples",
     type="p",
     pch="o",
     col=gtblue)
```

	Estimate	Pr(>|t|)	Freq
mnth2	0.379	0.005	12
mnth3	0.676	0.000	68
mnth4	0.516	0.021	11
mnth5	1.108	0.000	66
mnth6	0.499	0.043	7
mnth7	-0.326	0.240	1
mnth8	0.300	0.267	2
mnth9	1.052	0.000	64
mnth10	0.516	0.020	7
mnth11	-0.241	0.260	1
mnth12	-0.038	0.826	0
weekday1	0.229	0.024	9
weekday2	0.174	0.080	4
weekday3	0.283	0.004	16
weekday4	0.344	0.001	35
weekday6	0.530	0.000	79
weathersit2	-0.346	0.000	74
temp	3.847	0.000	38
atemp	4.879	0.000	84
windspeed	-1.101	0.000	59

Lack Statistical Significance



Lack of statistical significance: Uniform Distribution of P-values

Statistical Significance Summary

- Most regression coefficients remain statistically significant for 95% of the sub-samples, supporting statistical significance for these factors

- Statistical significance is not supported for most of months and weekdays as well as for temperature and windspeed factors given that other relevant factors, such as season and weather situation are in the model.

- While the 85% cutoff was used for the frequency of p-values being smaller than the significance level 0.01, other lower cut-offs, such as 50%, can be used.

Summary