Regression Analysis

Simple Linear Regression

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

Regression Concepts: Basics

1

About This Lesson

2

Example 1

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

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

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

Georgia Tech

3

Data Example 1

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

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

Georgia Tech

4

Example 2

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

Georgia Tech

5

Data Example 2

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

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

Georgia Tech

6

Example 3

- In 2000 **Bush** and **Gore** were the main candidates for President in the U.S. Buchanan, a strongly conservative candidate, was also on the ballot. In the **state of Florida**, **Bush** and **Gore** essentially tied, hence the counts were examined carefully county by county.

- **Palm Beach County** exhibited strange results. Even though the people in this county are not conservative, many votes were cast for **Buchanan**. Examination of the voting ballot revealed that it was easy to mistakenly vote for **Buchanan (a conservative candidate)** when intending to vote for **Gore**. We will thus predict whether those who voted for **Buchanan** were indeed going for a conservative candidate.

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

Georgia Tech

7

Variables in Regression

The regression framework is characterized by the following:

1. We have one particular variable that we are interested in understanding or modelling, such as sales of a particular product, or the stock price of a publicly traded firm. This variable is called the **response (dependent) variable**, and is usually represented by Y.

2. We have a set of other variables that we think might be useful in predicting or modelling the response variable (say the price of the product, the competitors' price, and so on; or the profits, revenues, financial position of the firm, and so on). These are called the **predicting *or* explanatory (independent) variables**, and are usually represented by x1, x2, etc.

Georgia Tech

8

Variables in Regression

RESPONSE VARIABLE versus PREDICTING VARIABLE?

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

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

Georgia Tech

9

Response vs Predicting Variable

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

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

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

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

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

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

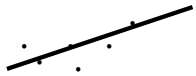
Georgia Tech

10

Linear Regression: General Model

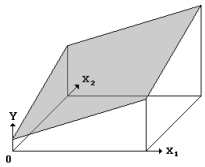
Simple linear regression

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



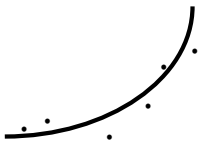
Multiple linear regression

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$



Polynomial Regression

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$



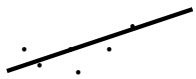
Georgia Tech

11

Linear Regression: General Model

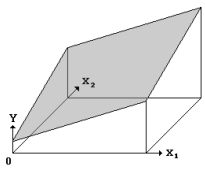
Simple linear regression

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



Multiple linear regression

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$



Polynomial Regression

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

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

Georgia Tech

12

Regression: Basics

A regression analysis is used for:

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

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

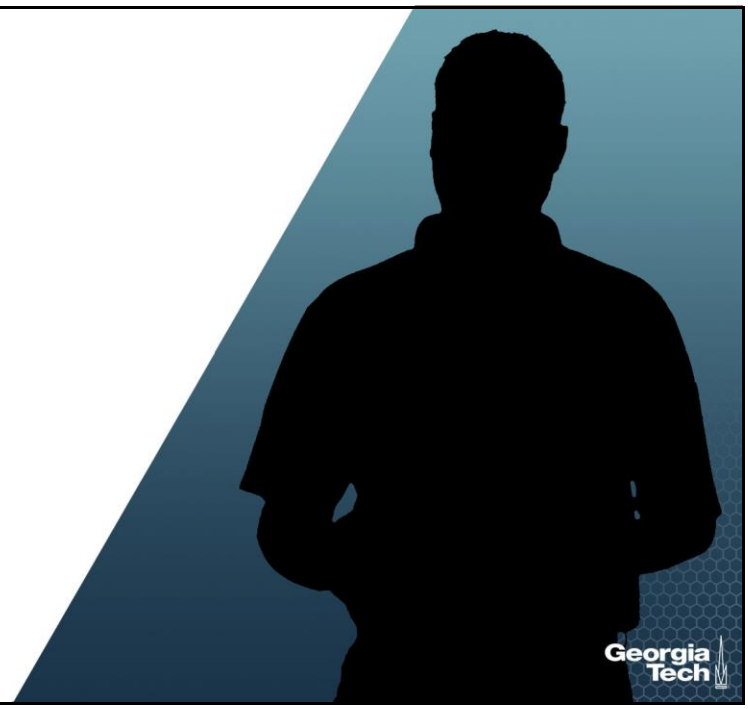
"*All models are wrong, but some are useful*." George Box

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

Georgia Tech

13

Summary



14

Regression Analysis

Simple Linear Regression

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

Regression Concepts: Estimation

1

About This Lesson

2

Simple Linear Regression: Model

Our goal is to find the best line that describes a linear relationship; that is, find (β_0, β_1) where

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

Equivalently, estimating:

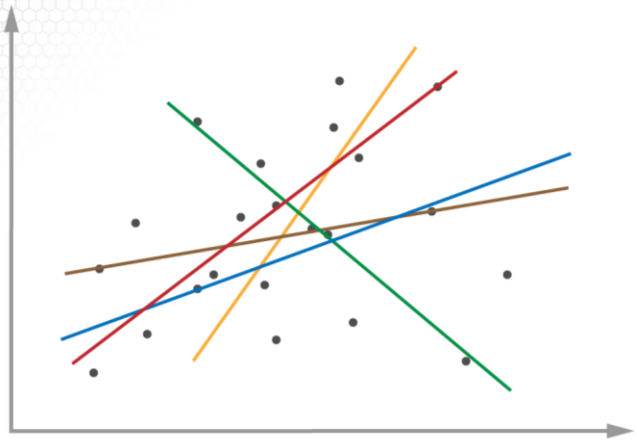
1. β_0 *Intercept*
2. β_1 *Slope*

ε is the deviance of the data from the linear model

Georgia Tech

3

Simple Linear Regression: Model



How to find the best line?

Our goal is to find the line that describes a linear relationship; that is, find (β_0, β_1) where

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

Georgia Tech

4

Simple Linear Regression: Model

Data: $\{(x_1,y_1),...,(x_n,y_n)\}$
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; i = 1,...,n$

Assumptions:

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

Georgia Tech

5

Simple Linear Regression: Model

Data: $\{(x_1,y_1),...,(x_n,y_n)\}$
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; i = 1,...,n$

The model parameters are:
β_0, β_1, σ^2
- ***Unknown regardless how much data are observed***
- ***Estimated given the model assumptions***
- ***Estimated based on data***

Assumptions:

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

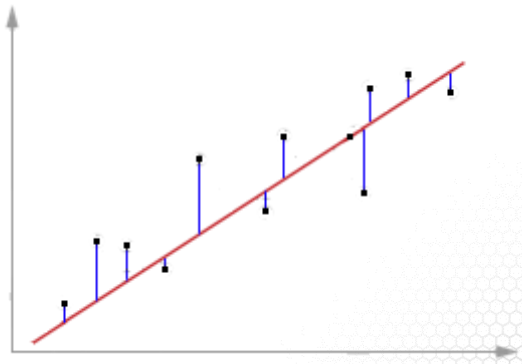
Georgia Tech

6

Model Estimation: Approach

To estimate (β_0, β_1) , we find values that minimize sum of squared errors:

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



Georgia Tech

7

Model Estimation: Approach

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

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

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

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

Georgia Tech

8

Model Estimation: Approach

Begin with the minimization problem:

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

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

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

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

- Result into a system of linear equation in β_0 and β_1
- Solve using linear algebra
- Solutions to the system are $\hat{\beta}_0$ and $\hat{\beta}_1$

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

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

Georgia Tech

9

Fitted Values and Residuals

Given the estimates of b_0 and b_1, we define:

- *Fitted values*: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- *Residuals*: $r_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$
- *Mean squared error*: Estimator for σ^2

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

Georgia Tech

10

Variance Sampling Distribution

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

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

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

↑

Estimating σ^2 ⟵ Sample variance

Variance Sampling Distribution (cont'd)

What is the sample variance estimation?

Basic statistic concept:

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

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

Why n-1?

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

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

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

Why?

Variance Sampling Distribution (cont'd)

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

↑ Replaced by $\hat{\epsilon}_i = \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)$

We lose two degrees of freedom because

$\beta_0 \leftarrow \hat{\beta}_0$

$\beta_1 \leftarrow \hat{\beta}_1$

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

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

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

Georgia Tech

13

Model Parameter Interpretation

Commonly interested in the behavior of β_1

- A positive value of β_1 is consistent with a direct relationship between x and y; **e.g.**, higher values of height are associated with higher values of weight, or lower values of revenue are associated with lower values of profit;
- A negative value of β_1 is consistent with an inverse relationship between x and y; **e.g.**, higher price of a product is associated with lower demand, or a lower inflation rate is associated with a higher savings rate;
- A close-to-zero value of β_1 means that there is not a significant association between x and y.

Georgia Tech

14

Model Estimate Interpretation

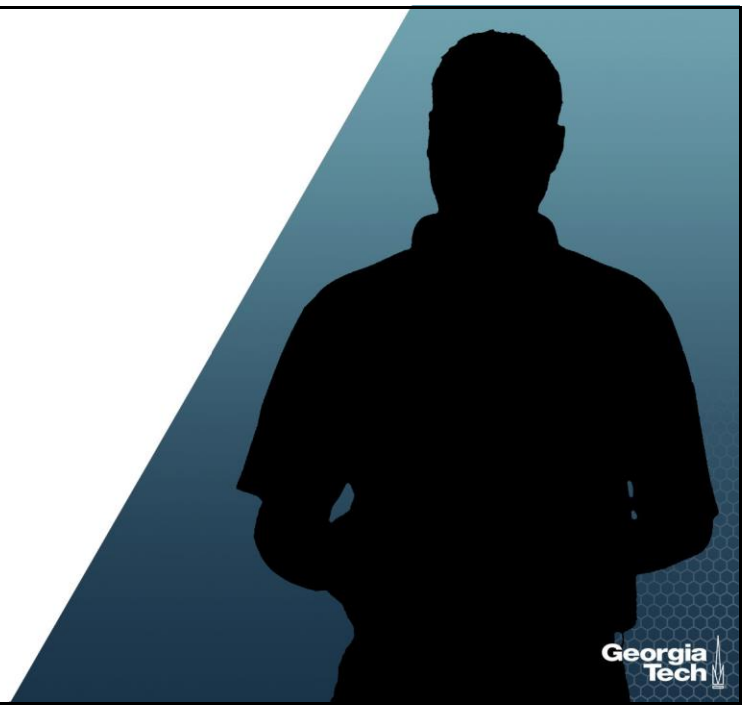
The Least Squares estimated coefficients have specific interpretations:

- $\hat{\beta}_1$ is the estimated expected change in the response variable associated with one unit of change in the predicting variable;
- $\hat{\beta}_0$ is the estimated expected value of the response variable when the predicting variable equals zero.

Georgia Tech

15

Summary



16

Regression Analysis

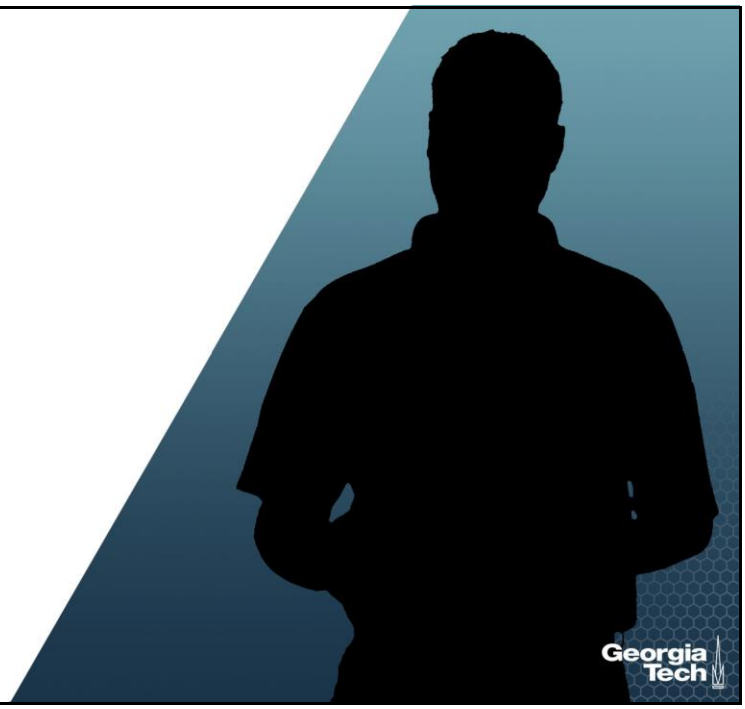
Simple Linear Regression

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

Regression Concepts: Estimation Example

1

About This Lesson



2

Example in R

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

Which are the response and the predicting variables?
Y = **Sales** and X = **Advertising Expenditure**

Georgia Tech

3

Example in R: Estimation

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

B. Interpret the coefficients.

C. What does the model predict sales as the advertising expenditure increases for an additional **$1,000**?

D. What sales would you predict for an advertising expenditure of **$30,000**?

E. What is the estimate of the error variance?

F. What could you say about the sales for an advertising expenditure of **$100,000**?

Georgia Tech

4

Example in R (cont'd)

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

Coefficients:

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

Residual standard error: 101.4 on 23 degrees of freedom
Multiple R-squared: 0.8106, Adjusted R-squared: 0.8024
F-statistic: 98.43 on 1 and 23 DF, p-value: 8.873e-10

Estimated Model Parameters:

$\widehat{\beta}_0$ = -157.3301
$\widehat{\beta}_1$ = 2.7721
$\widehat{\sigma}$ = 101.4

Georgia Tech

5

Example in R (cont'd)

A. Fit a linear regression. What are the estimated regression coefficients and the estimated regression line?
Solution: Estimates (β_0, β_1) are (-157.33, 2.77) and the *regression equation is:*

Sales = -157.33 + 2.77 Adv Expenditure

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

C. What does the model predict as the advertising expenditure increases for an additional \$1,000?
Solution: The increase in sales is 10 × 2.77 = 27.7 thousands.

Georgia Tech

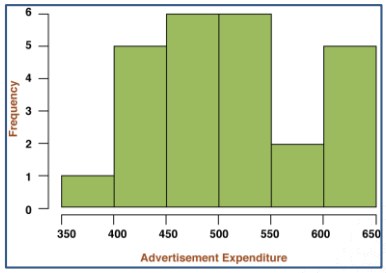
6

Example in R (cont'd)

D. What sales would you predict for an advertisement expenditure of $30,000?
Solution: The predicted sales is
-157.33 + 300 × 2.77 = 673.67 thousands

E. What is the estimate of the error variance?
Solution: Estimate σ^2 with MSE = 10,281.96

F. What could you say about the sales for an advertising expenditure of $100,000?
Solution: An advertisement expenditure of $100,000 or 1000 units is outside of the observed range and thus we cannot predict the sales since this is extrapolation.



Georgia Tech

7

Summary

Georgia Tech

8

Regression Analysis

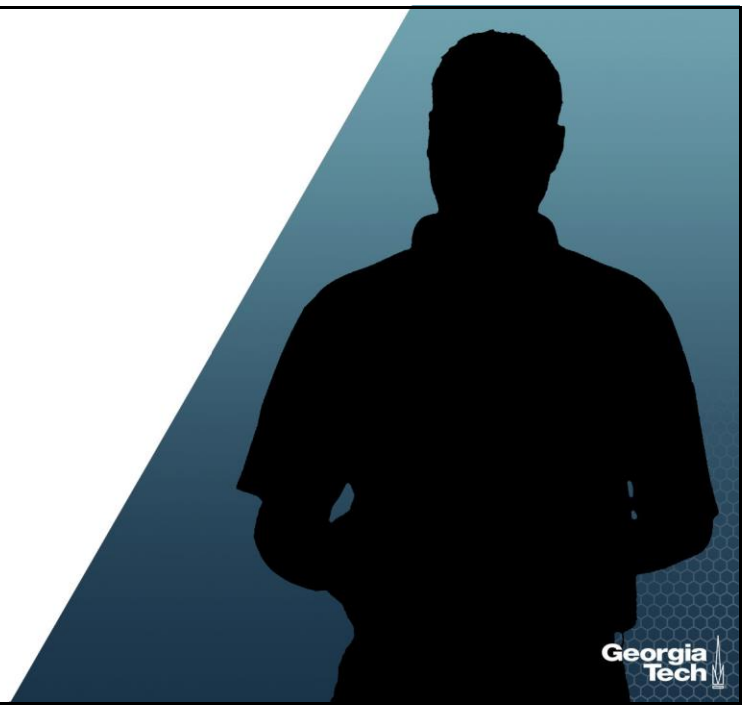
Simple Linear Regression

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

Regression Concepts:
Statistical Inference

1

About This Lesson



2

Regression Estimators: Properties

For the slope parameter β_1, we can show

$$E(\hat{\beta}_1) = \beta_1$$
$$Var(\hat{\beta}_1) = \sigma^2 / S_{xx}$$

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

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

$$= \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{=1}$$

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

Georgia Tech

3

Regression Estimators: Properties

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

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

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

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

Georgia Tech

4

Regression Estimators: Properties

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

We do not know σ^2. We can replace it by MSE, but then the sampling distribution becomes the t-distribution with **n-2** df.

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

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

$$\rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{MSE}}{\text{S}_{\text{XX}}}}} \sim \text{t}_{n-2}$$

Georgia Tech

5

Inference for Slope Parameter

Given the sampling distribution of $\hat{\beta}_1$, we can derive confidence intervals and perform hypothesis testing for β_1:

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

Georgia Tech

6

Confidence Interval Derivation

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

1 - α Confidence interval

$$\rightarrow \hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\frac{\text{MSE}}{S_{xx}}}$$

- $\hat{\beta}_1$: Estimate of β_1
- $t_{\frac{\alpha}{2}, n-2}$: t-critical point
- $\sqrt{\frac{\text{MSE}}{S_{xx}}}$: Standard Deviation/Error of $\hat{\beta}_1$

Sampling distribution of $\hat{\beta}_1$ is t_{n-2}

$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$, $\sigma^2 \leftarrow \text{MSE}$

Georgia Tech

7

Testing the Overall Regression

One way we can test statistical significance is to use the t-test for

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

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

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

Georgia Tech

8

Testing Regression at Different Levels

How will the procedure change if we test:

$$H_0: \beta_1 = c \quad \text{vs.} \quad H_A: \beta_1 \neq c$$

for some known c?

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

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

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

If P-value small **(p-value < 0.01)** ⟶ **Reject**

9

Testing Regression at Different Levels (cont'd)

How will the procedure change if we test:

$H_o: \beta_1 = 0$ **versus** $H_A: \beta_1 > 0$

OR

$H_o: \beta_1 = 0$ **versus** $H_A: \beta_1 < 0$?

What if we want to test for positive relationship

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

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

What if we want to test for negative relationship

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

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

10

Inference for Intercept Parameter

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

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

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

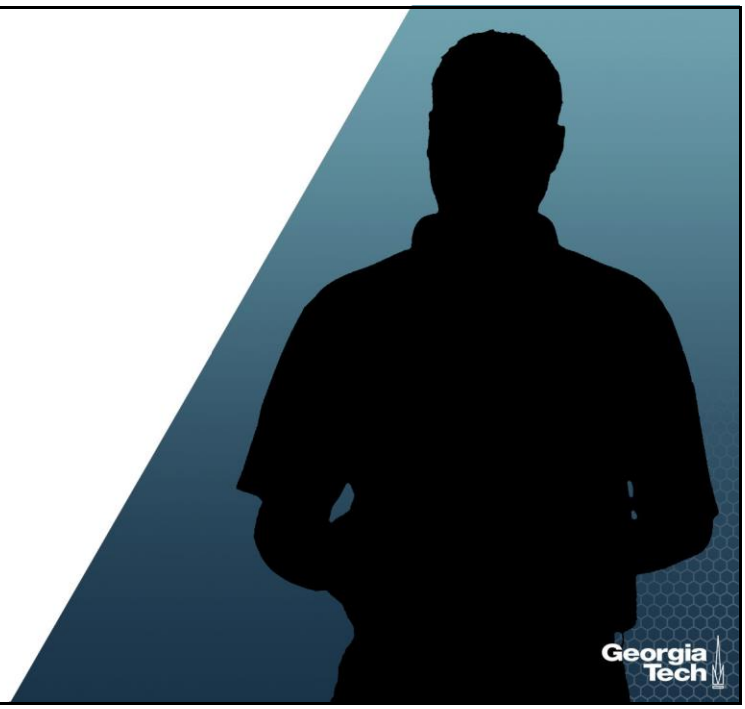
Confidence interval:

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

Georgia Tech

11

Summary



Georgia Tech

12

Regression Analysis

Simple Linear Regression

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

Regression Concepts:

Statistical Inference Examples

About This Lesson

Linear Regression: Example in R

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

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

What inferences can be made on the regression coefficients?

Example in R: Inference

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

b. What is the estimate of the coefficient β_0 and its variance?

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

d. Is the coefficient β_1 statistically positive? What is the p-value of the test? Interpret.

e. Obtain the 99% confidence interval for β_1.

f. What is the p-value of a hypothesis testing procedure?

Example in R (cont'd)

summary(model)

Coefficients:

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

Residual standard error: 101.4 on 23 degrees of freedom

a. The estimate for b_1 is 2.7721. The variance estimate is 0.2794^2. The sampling distribution is a t-distribution with 23 degrees of freedom.
b. The estimate for b_0 is -157.3301. The variance estimate is 145.1912^2.
c. The estimate for b_1 is statistically significant, as evidenced by a p-value of 8.87×10^{-10}

Georgia Tech

Example in R (cont'd)

d. β_1 statistically positive: H_A: $\beta_1 > 0$
We accept the alternative hypothesis because p-value is 4.43×10-10. (The test statistic is 9.921.)

e. The the 99% confidence interval for β_1 is (1.988, 3.557)

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

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

Please read the P-value Statement by the American Statistical Association at: *https://doi.org/10.1080/00031305.2016.1154108*

Georgia Tech

Summary

Regression Analysis

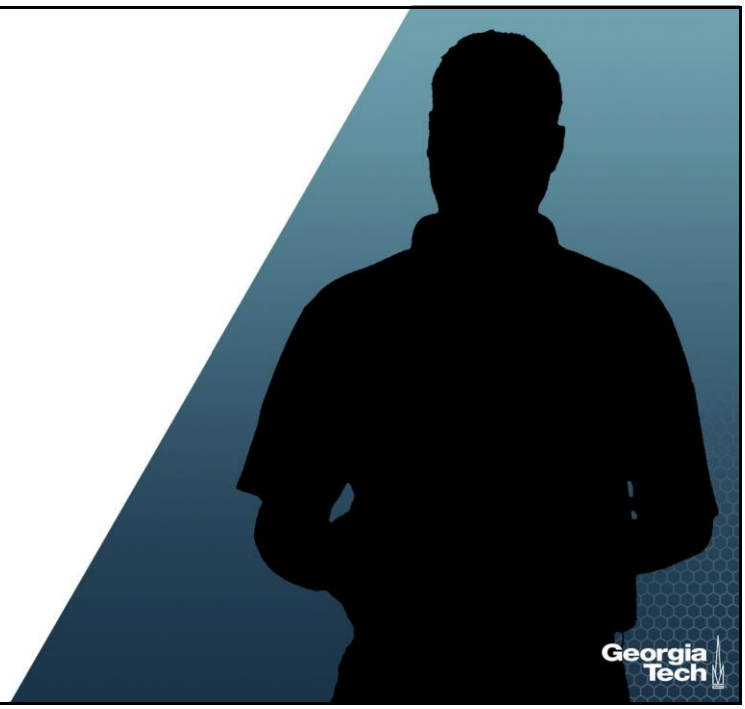
Simple Linear Regression

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

Regression Concepts:
Regression Line and Prediction

1

About This Lesson



2

Estimation vs. Prediction

Interpretation of estimated mean response:

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

Georgia Tech

3

Estimating the Regression Line

At some selected value of x (say x*), we estimate the "mean response" of y (or the regression line) via

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

Because the estimators of β_0 and β_1 are normally distributed, so is $\hat{y}$. That means we can draw inference using $\hat{y}$ if we know expected value and variance.

Georgia Tech

4

Estimating the Regression Line

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

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

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

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

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

Georgia Tech

5

Confidence Interval for Mean Response

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

- Interval length depends on x*
- As x* changes, we can construct a confidence band for $\hat{y}$
- Confidence bands show why extrapolation fails

Georgia Tech

6

Predicting a New Response

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

But the prediction contains two sources of uncertainty:

1. Due to the new (n+1)th observation
2. Due to parameter estimates (of β_0 and β_1)

Georgia Tech

7

Predicting a New Response

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

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

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

Georgia Tech

8

Predicting a New Response

A 100(1 - α)% ***prediction*** interval for a future y* (at x*) is

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

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

Georgia Tech

9

Summary

Georgia Tech

10

Regression Analysis

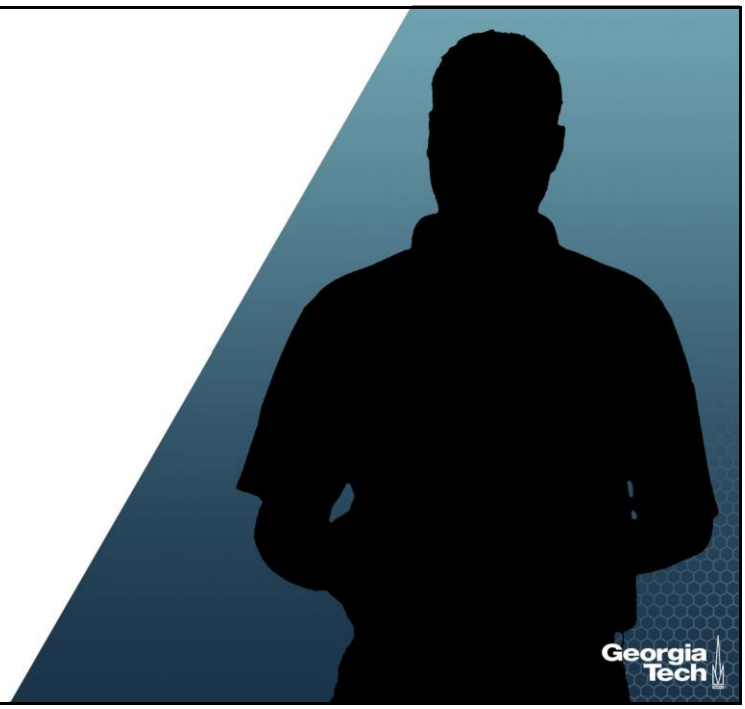
Simple Linear Regression

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

Regression Concepts: Regression Line and Prediction Examples

1

About This Lesson



2

Linear Regression: Example in R

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

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

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

Georgia Tech

3

Example in R: Estimating Regression Line & Prediction

a. What sales would you predict for an advertisement expenditure of **$30,000**?

b. What is the variance estimate of the estimated predicted sales for an advertisement expenditure of **$30,000**?

c. What are the lower and upper limits of predicted sales for an advertisement expenditure of **$30,000** at **99%** confidence level? How will the limits change if we lower the confidence level to **95%**?

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

Georgia Tech

4

Example in R

summary(model)
Coefficients:

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

Residual standard error: 101.4 on 23 degrees of freedom
xbar = mean(ADV)
n = 23+2
mse =101.4^2
var.beta1 = 0.2794^2
sxx = mse/var.beta1
pred.var = mse(1+1/n+(xbar-300)^2/sxx)*
pred.var
[1] 14286.16

Georgia Tech

5

Example in R

summary(model)
Coefficients:

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

Residual standard error: 101.4 on 23 degrees of freedom
xbar = mean(ADV)
n = 23+2
mse =101.4^2
var.beta1 = 0.2794^2
sxx = mse/var.beta1
pred.var = mse(1+1/n+(xbar-300)^2/sxx)*
pred.var
[1] 14286.16

a. For advertising expenditure of \$30,000, the predicted sales is:
-157.33 + 300 × 2.77
= 673.67 thousand

b. The variance of the predicted sales is

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

Georgia Tech

6

Example in R (cont'd)

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

Georgia Tech

7

Example in R (cont'd)

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

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

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

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

Georgia Tech

8

Summary

9

Regression Analysis

Simple Linear Regression

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

Regression Concepts:
Assumptions and Diagnostics

About This Lesson

Simple Linear Regression: Model

Data: $\{(x_1,y_1),\ldots,(x_n,y_n)\}$
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \; i = 1,\ldots,n$

Assumptions:

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

Georgia Tech

Residual Analysis

Residual Values: $\varepsilon_i \rightarrow \hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Graphical display: **Plot of the residuals $\boldsymbol{\varepsilon_i}$**

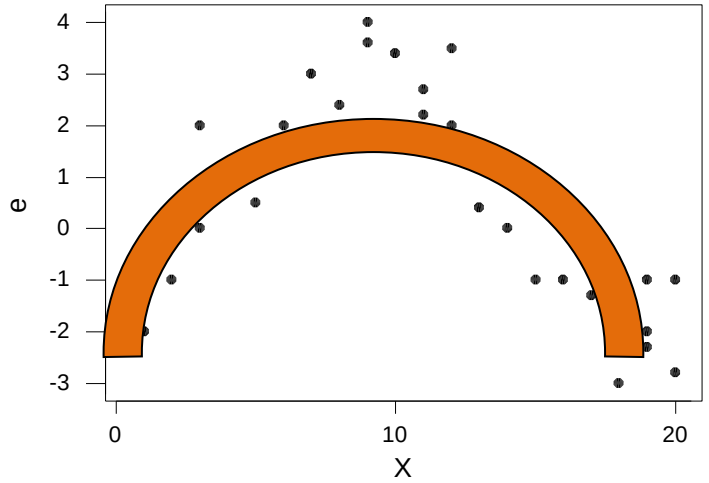
If the scatter of ε_i is **not random around zero line**, it could be that

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

Checking Assumptions: Residual Analysis

Linearity Assumption:

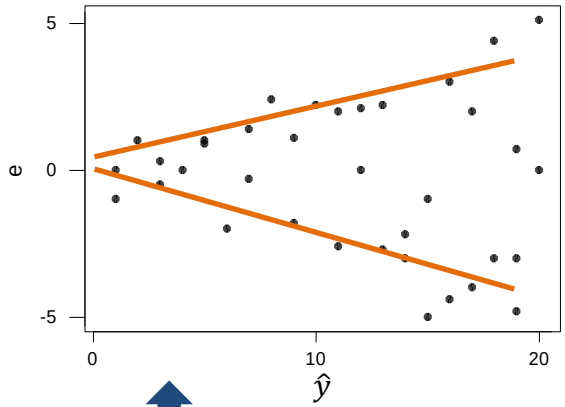
This shows that there may be a non-linear relationship between X and Y.



Checking Assumptions: Residual Analysis

Constant Variance Assumption:

The residuals show larger variance as the fitted values increases.

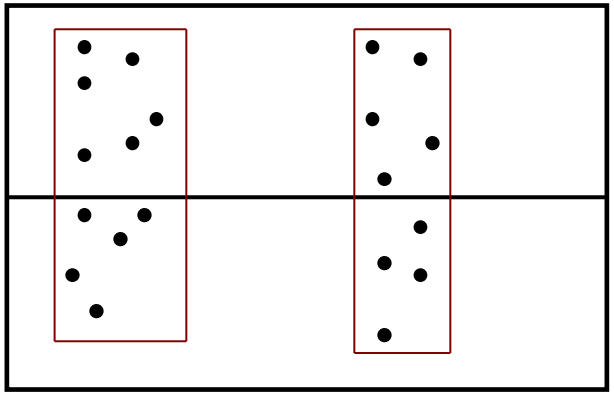


Here, it could be that σ^2 is not constant.

Checking Assumptions: Residual Analysis

Independence Assumption:

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



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

Georgia Tech

Checking the Assumption of Normality

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

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

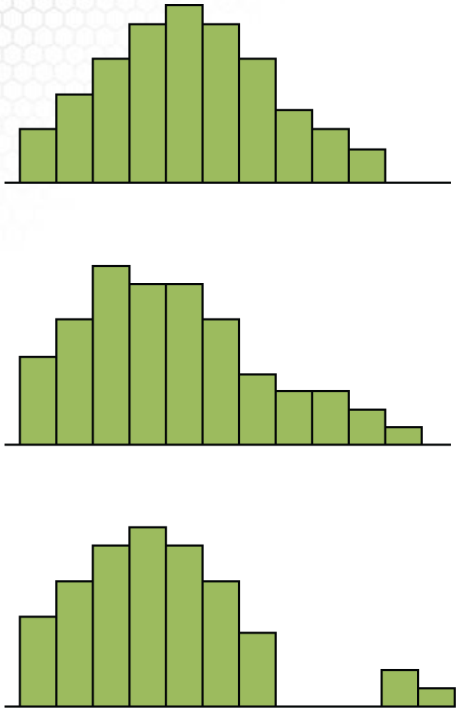
r_i = rank of e_i (between 1, n)

Φ = CDF of Normal Distribution

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

Georgia Tech

Checking the Assumption of Normality



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

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

Georgia Tech

Variable Transformation

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

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

Normality Transformations

Problem: Normality or constant variance assumption does not hold.
Solution: Transform the response variable from y to y* via

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

where the value of λ depends on how Var(Y) changes as X changes.

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

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

Georgia Tech

Summary

Regression Analysis

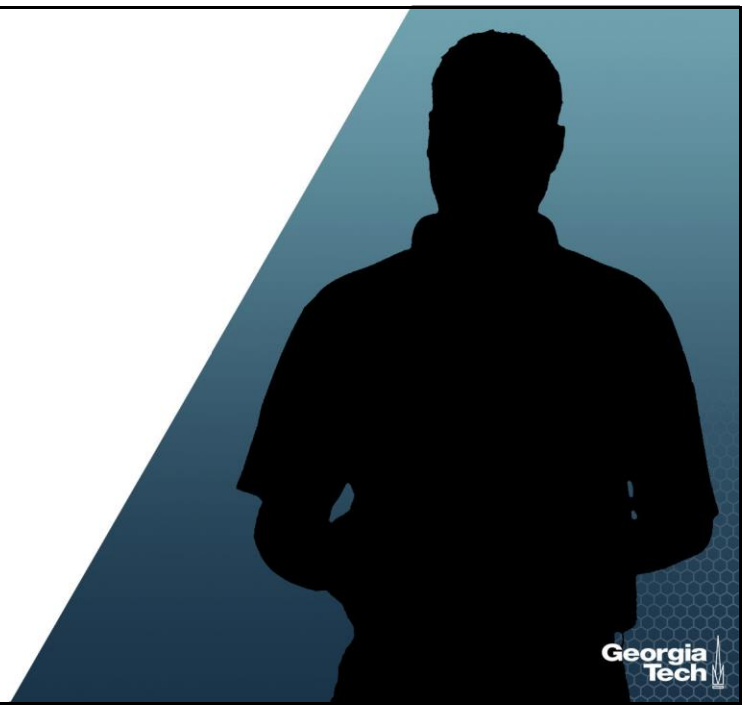
Simple Linear Regression

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

Regression Concepts:
Assumptions and Diagnostics

1

About This Lesson



2

Outliers in Regression

A data point far from the majority of the data (in *y* and/or *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 mean of the x's are called *leverage points*.
- A data point that is far from the mean of either or both the x's and/or the y's are *influential points* if they influence the fit of the regression.
- An outlier may or may not impact the regression fit significantly, thus it may or may not be an influential point.

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

Georgia Tech

3

Checking for Outliers

Look at the **standardized residuals**:

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

Compare the standardized residuals to the -2 to +2 band (or -1 to + 1).

- Standardized residuals bigger than 1 are large.
- Standardized residuals bigger than 2 extremely large.

Most statistics packages will calculate these automatically.

Georgia Tech

4

Coefficient of Determination

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

$R^2 = 1 - SSE / SST$

which is interpreted as:

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

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

R^2 = Proportion of total variability in Y that can be explained by the regression (that uses X)

Georgia Tech

5

Correlation Coefficient

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

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

Correlation coefficient and coefficient of variation:

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

Georgia Tech

6

Summary

7

Regression Analysis

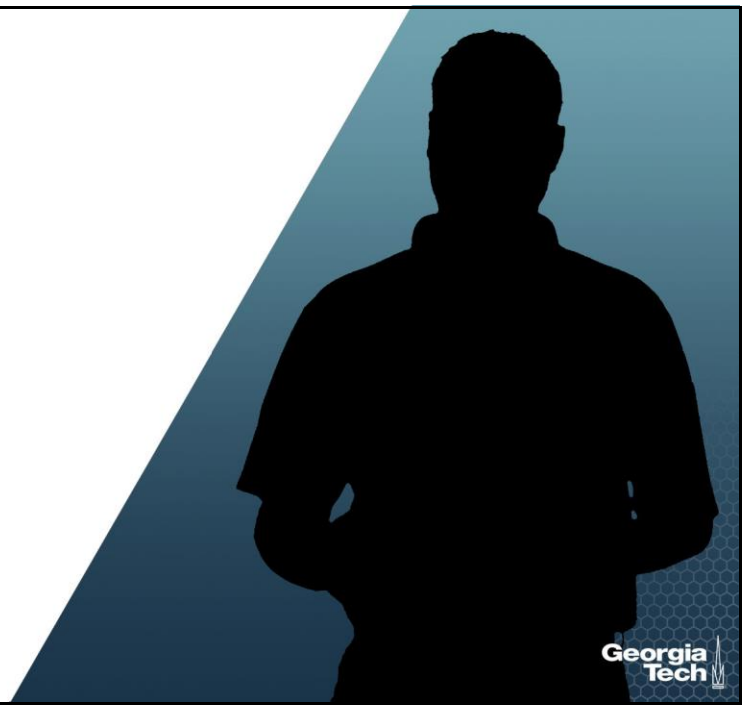
Simple Linear Regression

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

Regression Concepts:
Model Diagnostic Example

1

About This Lesson



2

Linear Regression: Example in R

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

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

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

Georgia Tech

3

Example in R: Residual Analysis

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

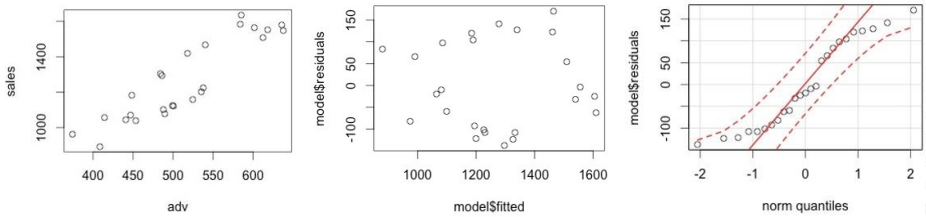
Georgia Tech

4

Example in R (cont'd)

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

b. plot(adv, sales)
plot(model$fitted, model$residuals)
library(car); qqPlot(model$residuals)



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

Example in R (cont'd)

c. *Do you identify any outliers?*

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

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

```
summary(model)$r.squared
[1] 0.8105919
```

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

Summary

7

Regression Analysis

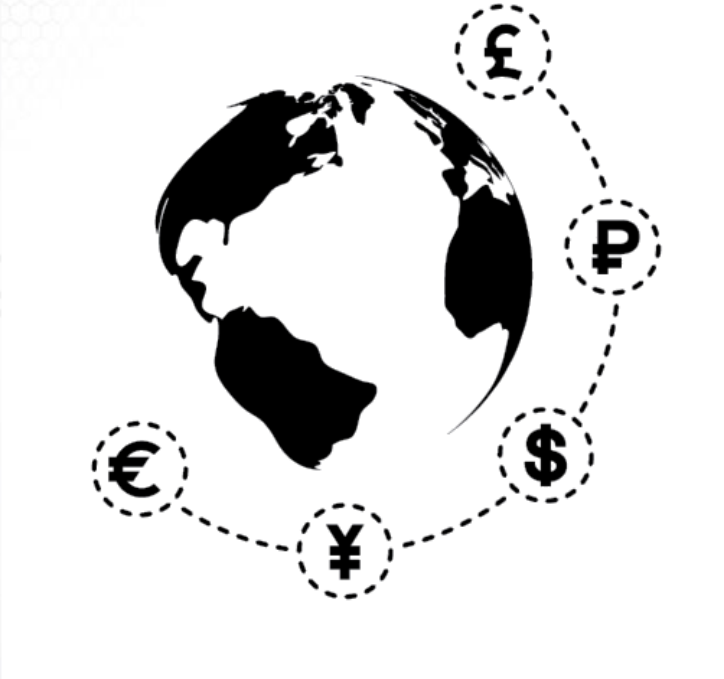
Simple Linear Regression

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

Testing the Theory of Purchasing Power Parity (Part 1)

About This Lesson

Testing the Theory of Purchasing Power Parity



<u>Acknowledgment</u>: This example was made available by Dr. Jeffrey Simonoff from the New York University.

Georgia Tech

Regression Variables

Response Variable: Average annual change in the exchange rate

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

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

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

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

Georgia Tech

Read the Data in R

```
# Use read.table R command: pay attention to the file type to use the correct read file!#
    ppp = read.table("ppp.dat",sep="\t", header=T, row.names=NULL)
## How many countries?
    dim(ppp)
    [1] 40  4
##  Brazil is an outlier and it was not included in the data set initially; I am adding it back as follows
    Addp = data.frame("Brazil",-76,-73,0)
    names(addp) = names(ppp)
## Save the data variables to be recognized by R as separate variables
    ppp = data.frame(rbind(ppp,addp))
    attach(ppp)
## Re-label the 'Developed' column to differentiate between Developed and Developing countries
    Developed[Developed==1] = "Developed"
    Developed[Developed==0] = "Developing"
```

Georgia Tech

Exploratory Data Analysis in R

 # Evaluate the Linear Relationship: Perform a scatter plot of the two variables

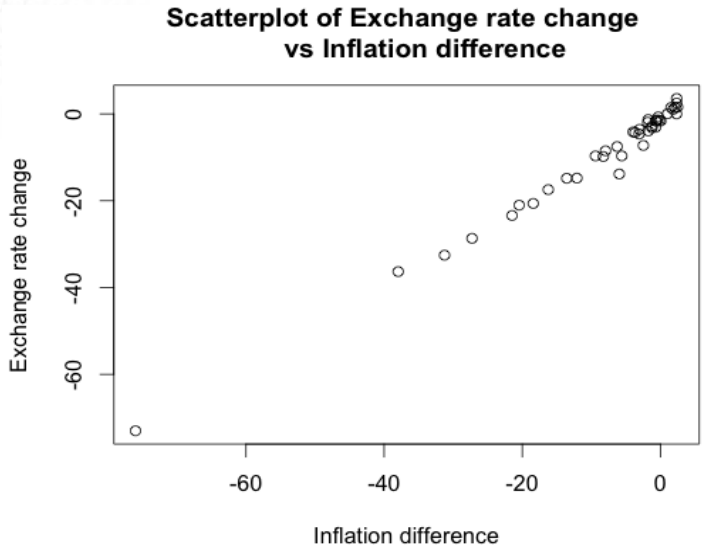
plot(Inflation.difference,Exchange.rate.change, main="Scatterplot of Exchange rate change vs Inflation difference", xlab="Inflation difference",ylab="Exchange rate change")

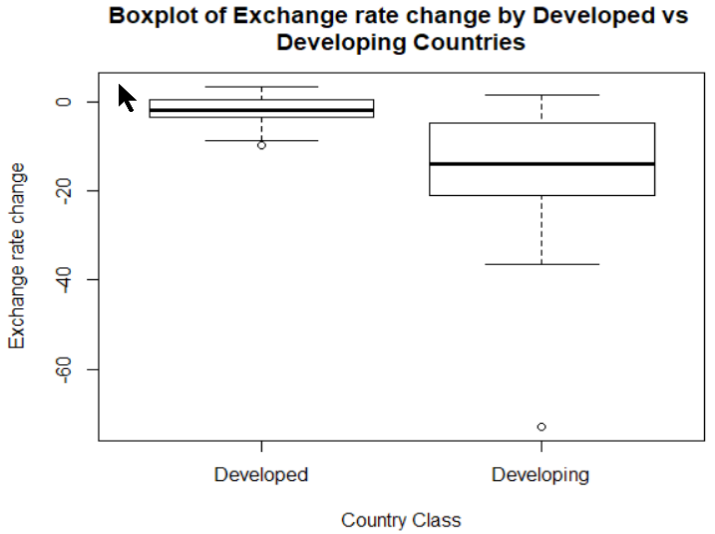
 # Evaluate differences between developed and developing countries

boxplot(Exchange.rate.change~as.factor(Developed), main="Boxplot of Exchange rate change by Developed vs Developing Countries",xlab="Country Class",ylab="Exchange rate change")

Georgia Tech

Exploratory Data Analysis in R





Fitting Linear Regression in R

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

Coefficients:

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

$\hat{\beta}_0$ = -1.5193, se($\hat{\beta}_0$) = 0.2941
$\hat{\beta}_1$ = 0.9618, se($\hat{\beta}_1$) = 0.0178
Test for statistical significance:
β_0: t-value= -5.165, p-value ≈ 0
β_1: t-value= 53.991, p-value ≈ 0

Residual standard error: 1.646 on 39 degrees of freedom
Multiple R-squared: 0.9868, Adjusted R-squared: 0.9865
F-statistic: 2915 on 1 and 39 DF, p-value: < 2.2e-16

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

Georgia Tech

Does the Theory Hold?

The principle of purchasing power parity (PPP) states:

Average annual change in the exchange rate = Difference in average annual inflation rates + Random error

The economic theory says that $\beta_0 = 0,\ \beta_1 = 1$.
The estimates for these coefficients are: $\hat{\beta}_0 = -1.519, \hat{\beta}_1 = 0.961$

Violations of PPP theory with respect to both the intercept and the slope.

Testing the theory:

$\beta_0 = 0$: Based on the t-test of statistical significance we find that β_0 is statistically different from zero.

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

$$\text{T-value} = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = \frac{0.9618 - 1}{0.0178} = -2.1448$$

p-value = $2(1\text{-}P(T_{39} < |-2.1448|)) = 0.038$

Georgia Tech

Hypothesis Testing in R

```
# Perform the hypothesis test for slope coefficient
H0: slope=1
# use the library 'car' available in R (you need to
install this library first then download it)
install.packages("car")
library(car)
linearHypothesis(pppa,c(0,1),rhs=1)
## Alternatively, you can compute the t-value and p-
value as follows:
tvalue = (0.9618−1)/0.01781
pvalue = 2*(1-pt(abs(tvalue),39))
## Use the help menu to learn more about the
functions used above:
help(pt)
help(linearHypothesis)
```

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

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

Georgia Tech

Summary

Regression Analysis

Simple Linear Regression

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

Example 1: Testing the Theory of Purchasing Power Parity

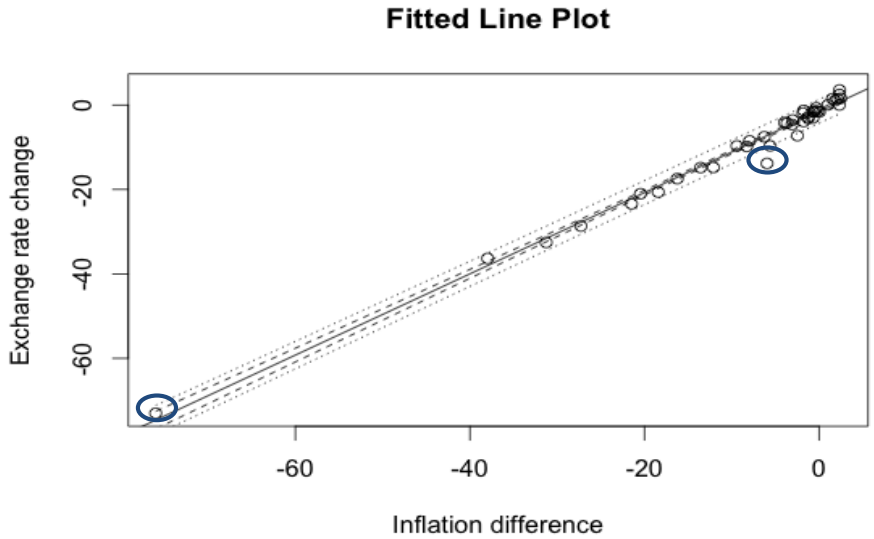
(Part 1)

About This Lesson

Confidence Bands in R

```
# Function for fitted line plot: See ppp-revised.R for this function
#regplot.confbands.fun = function(x, y, confidencelevel=.95, CImean=T, PI=T,
CIregline=F, legend=F){
    #### Modified from a function written by Sandra McBride, Duke University
….}
```

regplot.confbands.fun(Inflation.difference,Exchange.rate.change)



The fitted line plot shows several lines:

- The continuous line is the fitted regression line.
- The wider interrupted line band is the prediction confidence band.
- The narrower interrupted line band is the confidence band.
- The circles correspond to outliers.

Georgia Tech

Confidence and Prediction Intervals

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

Interpretation of the two intervals:

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

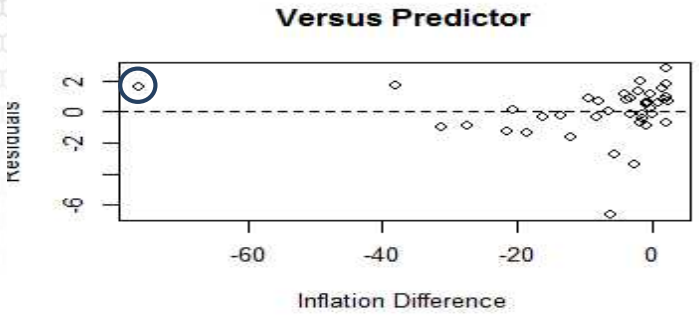
- The 95% confidence limits for the exchange rate change for one country with inflation difference equal to -0.68 are (-5.554,1.207).

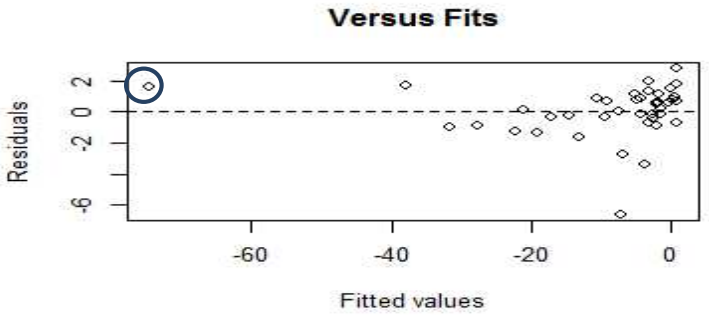
Georgia Tech

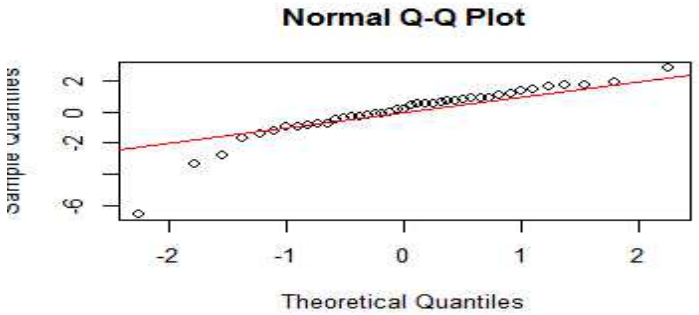
Residual Analysis in R

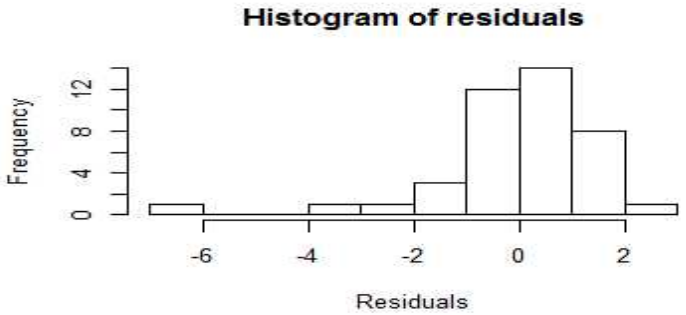
```
par(mfrow=c(2,2))
plot(Inflation.difference, residuals(pppa),xlab="Inflation
Difference",ylab="Residuals",main="Versus Predictor")
abline(h=0,lty=2)
plot(fitted(pppa),residuals(pppa),xlab="Fitted values",ylab="Residuals“, main="Versus Fits")
abline(h=0,lty=2)
qqnorm(residuals(pppa))
abline(0,1,lty=1,col="red")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

Residual Analysis in R









Georgia Tech

Residual Analysis in R

***Leverage Points*:** The isolated point in residual plots is Brazil. Why is Brazil a leverage point?

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

Why do we care about leverage points?

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

Influential Points in Regression Analysis

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

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.37222	0.30517	-4.497	6.31e-05
Inflation.difference	0.99152	0.02626	37.757	< 2e-16

Residual standard error: 1.62 on 38 degrees of freedom
Multiple R-squared: 0.974, Adjusted R-squared: 0.9734

```
## Test whether the slope is equal to 1 (PPP
theory)
    tvalue = (0.9915−1)/ 0.02626
    pvalue = 2*(1-pt(abs(tvalue),38))
```

$\hat{\beta}_0$ = -1.372, se($\hat{\beta}_0$) = 0.305
Statistical significance for β_0:
t-value= -4.497, p-value ≈ 0

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

We are seeing violations of PPP with respect to intercept only.

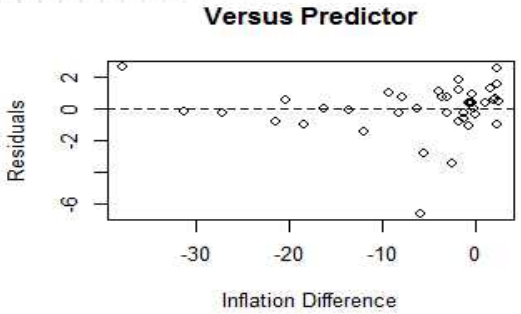
Georgia Tech

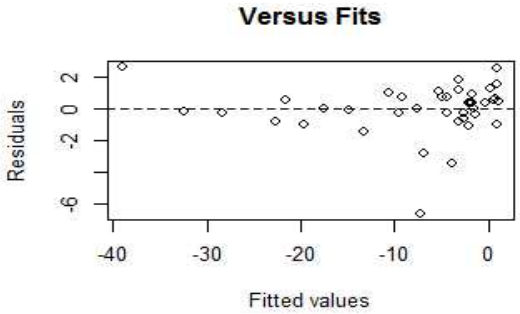
Residual Analysis: Model without Brazil

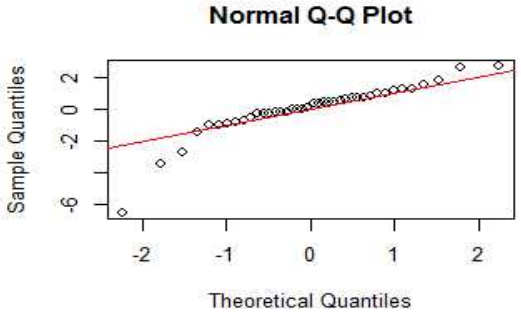
```
par(mfrow=c(2,2))
plot(Inflation.difference, residuals(pppn),xlab="Inflation Difference",ylab="Residuals",
main="Versus Predictor")
abline(h=0,lty=2)
plot(fitted(pppn),residuals(pppn),xlab="Fitted values",ylab="Residuals",main="Versus Fits")
abline(h=0,lty=2)
qqnorm(residuals(pppn))
abline(0,1,lty=1,col="red")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

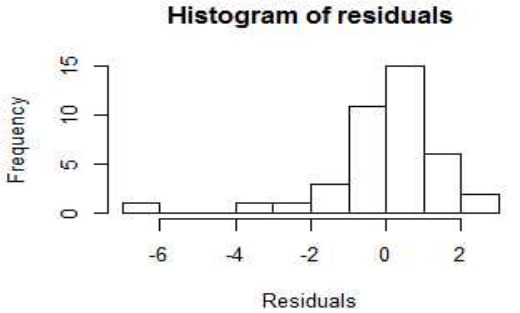
Georgia Tech

Residual Analysis: Model without Brazil









Assumptions:
Linearity: No pattern in the residuals with respect to the predicting variable.
Constant Variance: The variance is higher for higher fitted values. Does not hold.
Uncorrelated Errors: No grouping of the residuals
Normality: Except for the presence of an outlier, it is reasonably symmetric.

Outliers *(observations for which the residual value is away from the range)***:**
The isolated point in the residual plots is Indonesia. Would omitting Indonesia change anything? The strength of the relationship would increase, but so the rejection of PPP.

Testing the Theory of Purchasing Power Parity

Findings:

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

Summary

Regression Analysis

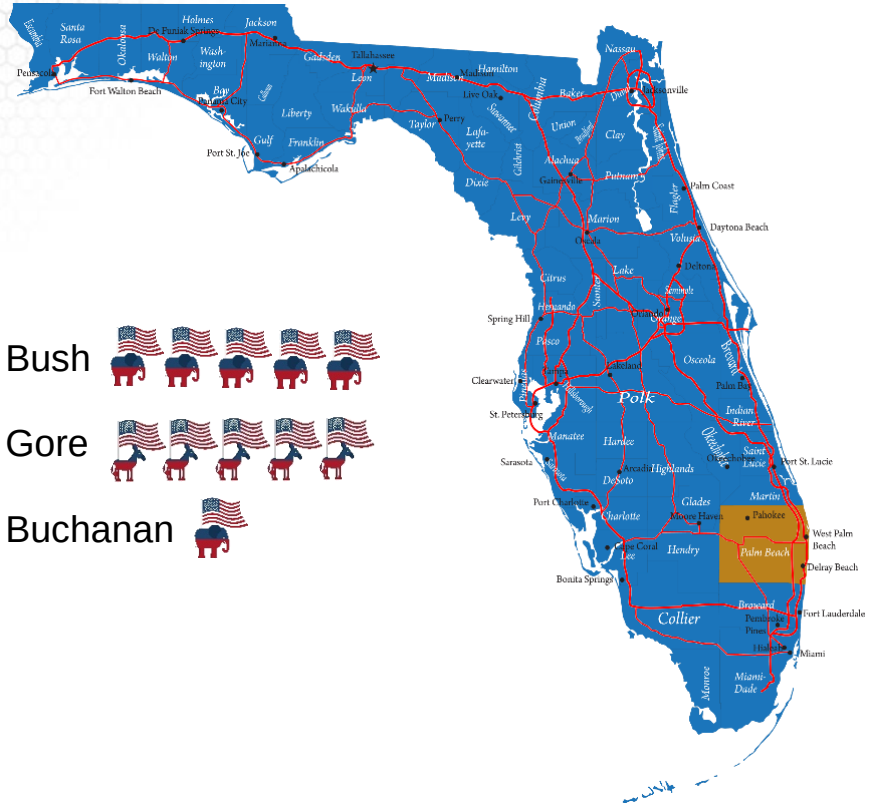
Simple Linear Regression

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

Example 2: 2000 Presidential Elections in Florida

About This Lesson

Elections in 2000: Florida



Data Example in R

Read data with read.table R command which is used for reading ASCII files

elections = read.table("elections.txt",header=TRUE)

Check the data content elections[1:4,]

	co	lat	lon	npop	whit	blac	hisp	o65	hsed	coll	inco	bush	gore	brow
1	1	29.7	82.4	198326	74.4	21.8	4.7	9.4	82.7	34.6	19412	34124	47365	658
2	2	30.3	82.3	20761	82.4	16.8	1.5	7.7	64.1	5.7	14859	5610	2392	17
3	3	30.2	85.6	146223	84.2	12.4	2.4	11.9	74.7	15.7	17838	38637	18850	171
4	4	29.9	82.2	24646	76.1	22.9	2.6	11.8	65.0	8.1	13681	5414	3075	28

	nade	harr	hage	buch	mcre	phil	moor
1	3226	6	42	263	4	20	21
2	53	0	3	73	0	3	3
3	828	5	18	248	3	18	27
4	84	0	2	65	0	2	3

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

Georgia Tech

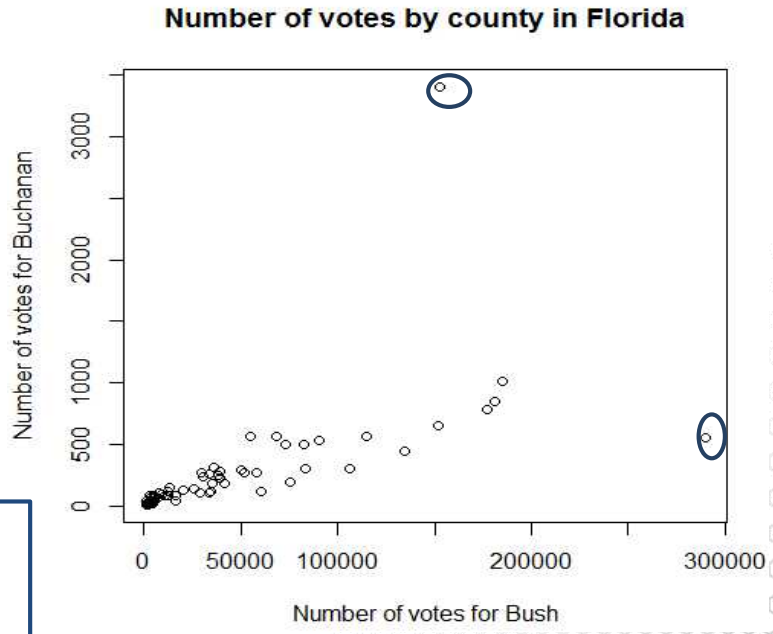
Exploratory Data Analysis in R

Extract number of votes for each candidates

buch = elections$buch
bush = elections$bush

Visualize the relationship between number of votes between Buchanan and Bush

plot(bush,buch,xlab="Number of votes for Bush",ylab="Number of votes for Buchanan", main="Number of votes by county in Florida")
cor(buch,bush)



Linearity Assumption:

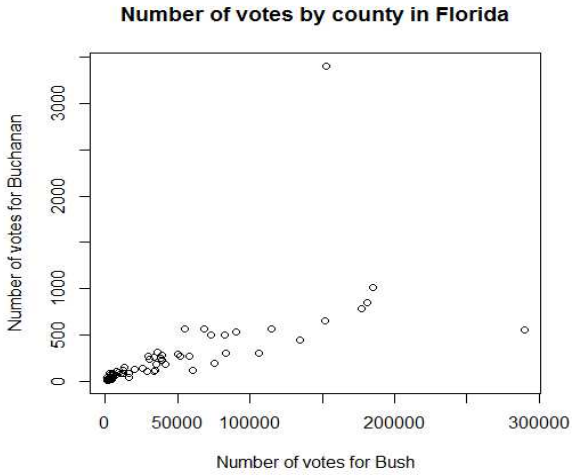
- The scatterplot shows a strong positive relationship between the number of votes for the two candidates except for two outliers, one corresponding to the Palm Beach county. The correlation is high also (0.625).
- Curvature in the relationship – consider transformations

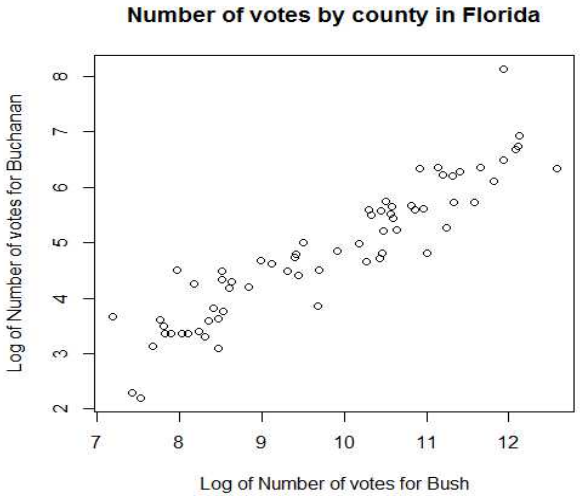
Georgia Tech

Linearity using Transformation

Transform both variables using the log-transformation

plot(log(bush),log(buch),xlab="Log of Number of votes for Bush",ylab="Log of Number of votes for Buchanan", main="Number of votes by county in Florida")
cor(log(bush),log(buch))





Georgia Tech

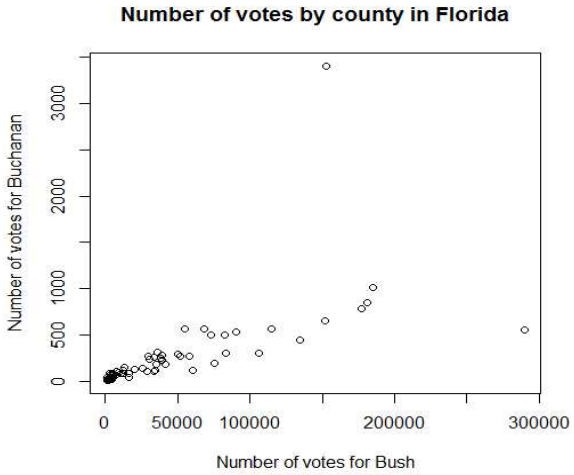
Linearity using Transformation

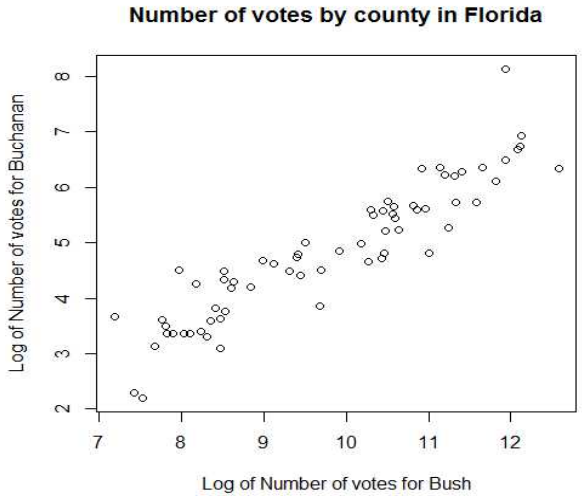
Transform both variables using the log-transformation

plot(log(bush),log(buch),xlab="Log of Number of votes for Bush",ylab="Log of Number of votes for Buchanan", main="Number of votes by county in Florida")
cor(log(bush),log(buch))

Linearity Assumption:

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





Georgia Tech

Linear Regression Analysis

model = lm(log(buch)~log(bush))

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	*-2.55079*	*0.38903*	*-6.557*	*1.04e-08 ****
log(bush)	*0.75620*	*0.03934*	*19.222*	*< 2e-16 ****

---

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

$\hat{\beta}_0 = -2.55$, se($\hat{\beta}_0$) = 0.389
$\hat{\beta}_1 = 0.756$, se($\hat{\beta}_1$) = 0.039
Test for statistical significance:
$\hat{\beta}_0$: t-value= -6.557, p-value ≈ 0
$\hat{\beta}_1$: t-value= 19.22, p-value ≈ 0

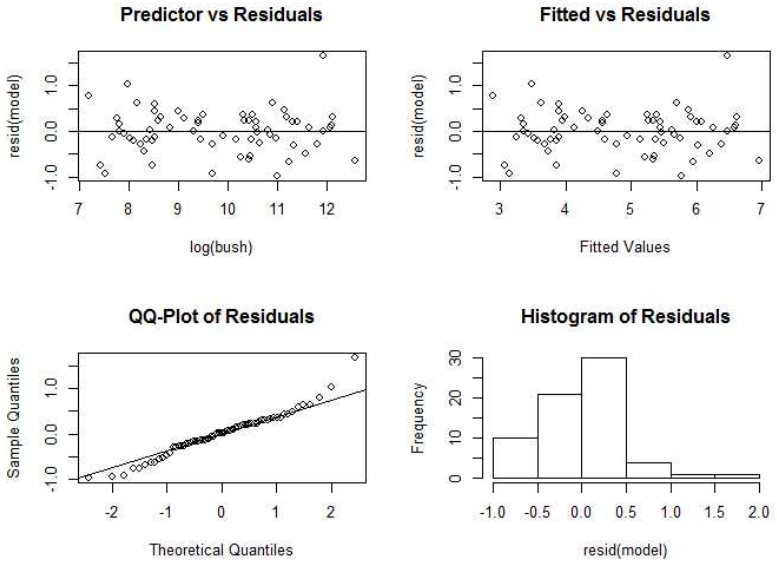
Residual standard error: 0.4672 on 65 degrees of freedom
Multiple R-squared: 0.8504, Adjusted R-squared: 0.8481
F-statistic: 369.5 on 1 and 65 DF, p-value: < 2.2e-16

$\hat{\sigma}$= 0.4672, n-2 = 65
$R^2 \sim 85\%$ *variability explained*

Georgia Tech

Residual Analysis

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



Model Interpretation

Estimated Regression Coefficients

betas = coef(model)

Betas

```
  (Intercept)   log(bush)
 -2.5507857    0.7561963
```

Confidence intervals for the coefficients

confint(model)

```
                2.5 %      97.5 %
(Intercept) -3.3277351  -1.7738363
log(bush)    0.6776289   0.8347638
```

Interpretation:

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

Georgia Tech

Is Palm Beach an Outlier?

```
## Omit Palm Beach
    model.red = lm(log(buch[-50])~log(bush[-50]))
    summary(model.red)
    Coefficients:
                        Estimate    Std. Error  t value  Pr(>|t|)
    (Intercept)       -2.31657    0.35470    -6.531   1.23e-08 ***
    log(bush[-50])  0.72960    0.03599   20.271  < 2e-16 ***
## Obtain the predicted vote count for Palm Beach given the fitted model without
    new = data.frame(bush = bush[50])
    ## The difference between predicted on the original scale and the observed vote count
    buch[50 ]-exp(predict(model.red,new))
    [1] 2809
## Prediction Confidence Interval for log(vote count)
    predict(model.red,new,interval='prediction',level=.95)
## Prediction Confidence Interval on the original scale
    exp(predict(model.red,new,interval='prediction',level=.95))
      fit          lwr          upr
    597.5019 252.738 1412.564
## Is the observed vote count in the prediction interval?
    buch[50]
    [1] 3407
```

Is Palm Beach an Outlier?

```
## Omit Palm Beach
    model.red = lm(log(buch[-50])~log(bush[-50]))
    summary(model.red)
    Coefficients:
                          Estimate    Std. Error  t value  Pr(>|t|)
    (Intercept)       -2.31657    0.35470    -6.531   1.23e-08 ***
    log(bush[-50])  0.72960    0.03599   20.271  < 2e-16 ***
## Obtain the predicted vote count for Palm Beach given the fitted model without
    new = data.frame(bush = bush[50])
    ## The difference between predicted on the original scale and the observed vote count
    buch[50 ]-exp(predict(model.red,new))
    [1] 2809
## Prediction Confidence Interval for log(vote count)
    predict(model.red,new,interval='prediction',level=.95)
## Prediction Confidence Interval on the original scale
    exp(predict(model.red,new,interval='prediction',level=.95))
      fit          lwr          upr
    597.5019 252.738 1412.564
## Is the observed vote count in the prediction interval?
    buch[50]
    [1] 3407
```

Interpretation:

- The difference between predicted and observed vote count for Bush in the Palm Beach county is 2809.
- The upper bound of the prediction confidence interval for the vote count is 1412 which is much lower than the observed vote count, 3407.
- While a difference of 2809 votes is not large given the total U.S. votes, this was particularly decisive for the 2000 elections.
- Recall that George W. Bush won Florida by a margin of 537 votes.

Georgia Tech

Summary