KENT STATE UNIVERSITY

MIS-64036: Business Analytics

Lecture VI

Rouzbeh Razavi, PhD

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Agenda

- **Quick Recap of Correlation**
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Sum of Squares, SS

Before we start our discussion, let us introduce the Sum of Squares notation. For variables X and Y :

$$SS_X = \sum \left(X - \bar{X}\right)^2 \text{ can also be represened as } SS_{XX}$$

$$SS_Y = \sum \left(Y - \bar{Y}\right)^2$$

$$SS_{XY} = \sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)$$

Example: Sum of Squares, SS

Console ~/

```
> X=c(1,7,8,9,10)
> Y=c(1,8,5,-2,0)
> SSX=sum((X-mean(X))^2)
> SSX
[1] 50
> SSY=sum((Y-mean(Y))^2)
> SSY
[1] 65.2
> SSXY=sum((X-mean(X))*(Y-mean(Y)))
> SSXY
[1] -5
```

Recall Correlation

- Correlation is a measure of the degree of relatedness of variables.

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$$

$$= \frac{SS_{XY}}{\sqrt{(SS_{XX})(SS_{YY})}}$$

```
> SSXY/sqrt(SSX*SSY)
[1] -0.08757118
> cor(X,Y)
[1] -0.08757118
>
```

Degrees of Correlation

- The term (r) is a measure of the linear correlation of two variables
  - The number ranges from -1 to 0 to +1
    - Positive correlation: as one variable increases, the other variable increases
    - Negative correlation: as one variable increases, the other one decreases
    - No correlation: the value of r is close to 0
  - Closer to +1 or -1, the higher the correlation between the dependent and the independent variables

KENT STATE UNIVERSITY

Computation of r for the Economics Example

Day	Interest X	Futures Index Y	X^2	Y^2	XY
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00
Summations	92.93	2,725	720.220	619,207	21,115.07

Computation of r Economics Example

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

$$= \frac{(21{,}115.07) - \frac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \frac{(92.93)^2}{12}\right]\left[(619{,}207) - \frac{(2725)^2}{12}\right]}}$$

$$= .815$$

Agenda

- Quick Recap of Correlation
- **Predictive Modeling: Simple Linear Regression**
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Predictive Models: Simple Linear Regression

- Prediction: If you know something about X, this knowledge helps you predict something about Y.

- When considering correlation, both X and Y are treated equally. In regression, however, one variable X, is considered as the independent (predictor) variable which tries to predict the dependent (target) variable, Y.

Simple Regression Analysis

- Bivariate (two variables) linear regression -- the most elementary regression model
  - dependent variable, the variable to be predicted, usually called Y
  - independent variable, the predictor or explanatory variable, usually called X
  - Usually the first step in this analysis is to construct a scatter plot of the data
- Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models

Equation of the Simple Regression Line

$$\hat{y} = b_0 + b_1 x$$

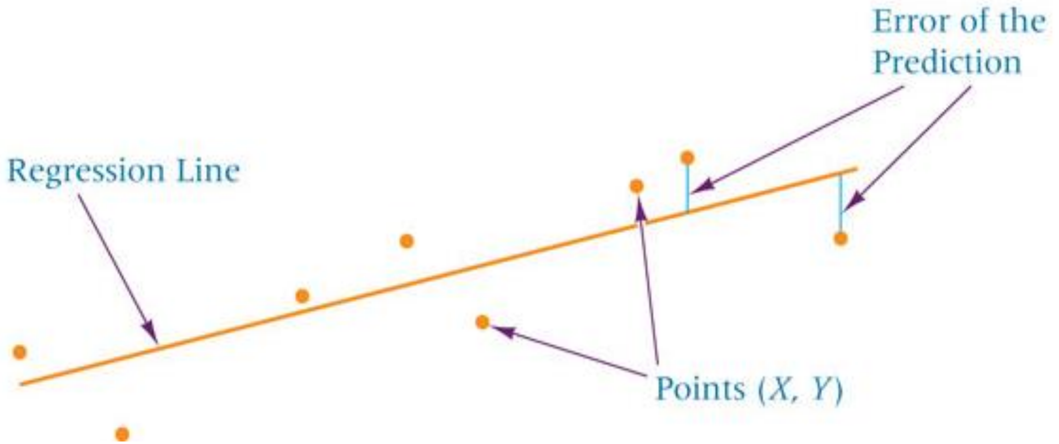
$$where: \quad b_0 = \text{the sample intercept}$$

$$b_1 = \text{the sample slope}$$

$$\hat{y} = \text{the predicted value of } y$$

Least Squares Analysis

- Least squares analysis is a process whereby a regression model is developed by producing the minimum sum of the squared error values
- The vertical distance from each point to the line is the error of the prediction.



Least Squares Analysis

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{\sum X^2}{n}}$$

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{\sum Y}{n} - b_1\frac{\sum X}{n}$$

Least Squares Analysis

$$SS_{XY} = \sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$$

$$SS_{XX} = \sum \left(X - \bar{X}\right)^2 = \sum X^2 - \frac{\sum X^2}{n}$$

$$b_1 = \frac{SS_{XY}}{SS_{XX}}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$$

Solving for b1 and b0 of the Regression Line: Airline Cost Example

Number of Passengers X	Cost ($1,000) Y	X^2	XY
61	4.28	3,721	261.08
63	4.08	3,969	257.04
67	4.42	4,489	296.14
69	4.17	4,761	287.73
70	4.48	4,900	313.60
74	4.30	5,476	318.20
76	4.82	5,776	366.32
81	4.70	6,561	380.70
86	5.11	7,396	439.46
91	5.13	8,281	466.83
95	5.64	9,025	535.80
97	5.56	9,409	539.32
$\sum X = 930$	$\sum Y = 56.69$	$\sum X^2 = 73,764$	$\sum XY = 4,462.22$

Solving for b1 and b0 of the Regression Line: Airline Cost Example

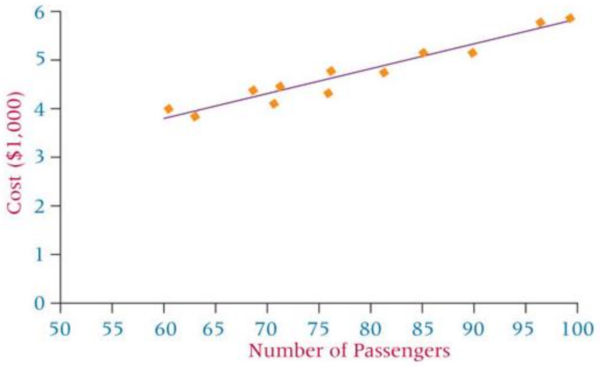
$$SS_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 4{,}462.22 - \frac{(930)(56.69)}{12} = 68.745$$

$$SS_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 73{,}764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \frac{56.69}{12} - (.0407)\frac{930}{12} = 1.57$$

$$\hat{Y} = 1.57 + .0407\,X$$



Least Squares Analysis: R Code

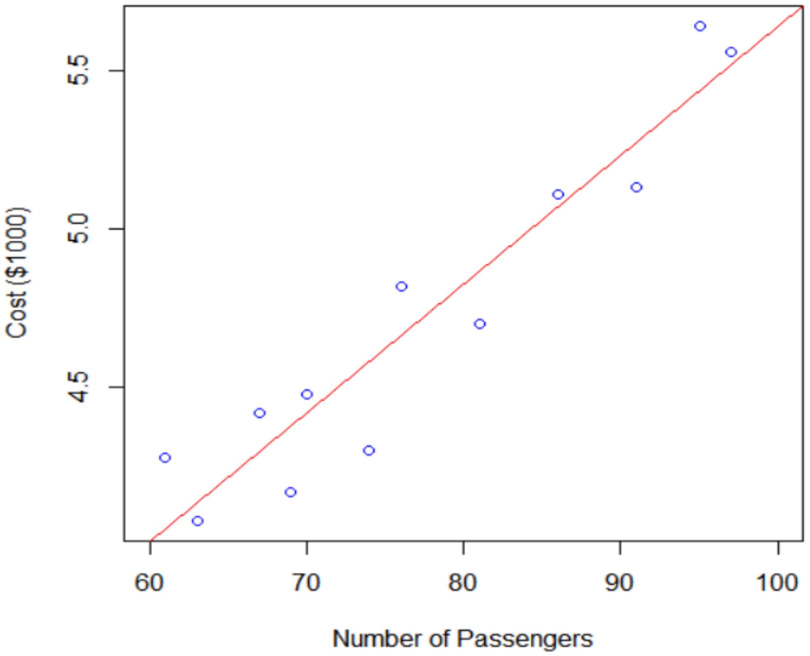
```
X=c(61,63,67,69,70,74,76,81,86,91,95,97)
Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)

SSXY=sum((X-mean(X))*(Y-mean(Y)))
SSX=sum((X-mean(X))^2)
b1=  SSXY/SSX
b1
b0=mean(Y)-b1*mean(X)
b0
```

Least Squares Analysis: R Code

```
> SSXY=sum((X-mean(X))*(Y-mean(Y)))
> SSX=sum((X-mean(X))^2)
> b1=  SSXY/SSX
> b1
[1] 0.0407016
> b0=mean(Y)-b1*mean(X)
> b0
[1] 1.569793
```

Least Squares Analysis: R Code

```
plot(X,Y,xlim=c(60, 100),xlab="Number of Passengers", ylab="Cost ($1000)", col="blue")
abline(lsfit(X, Y),col = "red")
```



Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- **Simple Linear Regression: Residual Analysis**
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Residual Analysis

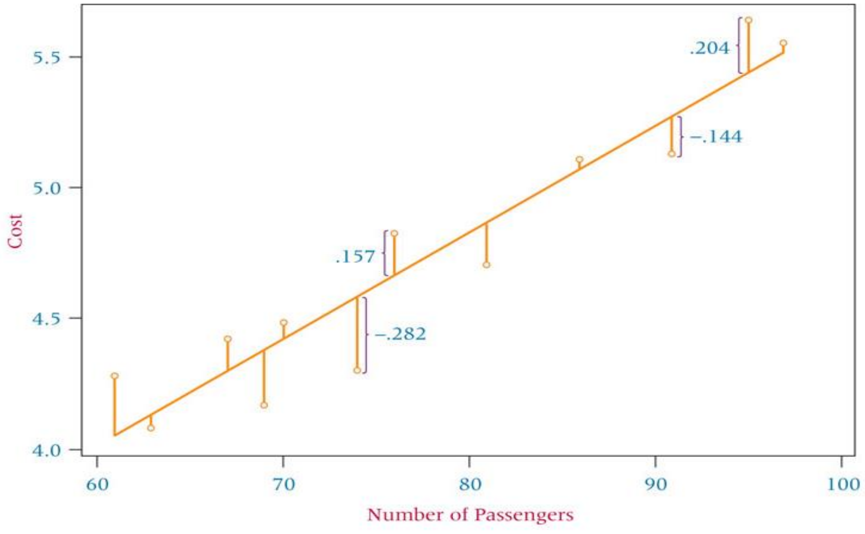
- Residual is the difference between the actual y values and the predicted $\hat{y}$ values.
- Reflects the error of the regression line at any given point.

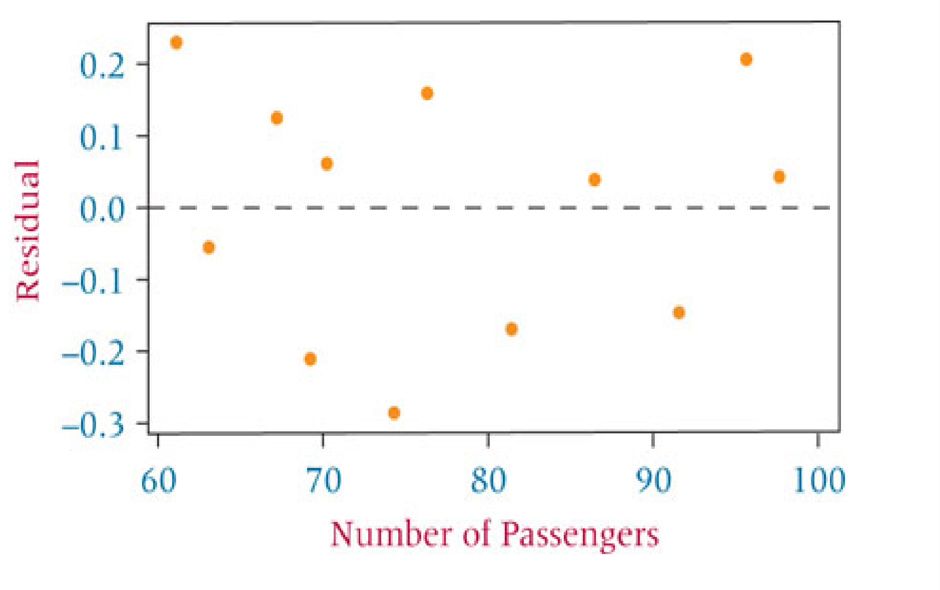
Residual Analysis: Airline Cost Example

Number of Passengers X	Cost ($1,000) Y	Predicted Value $\hat{Y}$	Residual $Y - \hat{Y}$
61	4.28	4.053	.227
63	4.08	4.134	-.054
67	4.42	4.297	.123
69	4.17	4.378	-.208
70	4.48	4.419	.061
74	4.30	4.582	-.282
76	4.82	4.663	.157
81	4.70	4.867	-.167
86	5.11	5.070	.040
91	5.13	5.274	-.144
95	5.64	5.436	.204
97	5.56	5.518	.042

$$\sum (Y - \hat{Y}) = -.001$$

Residual Analysis for Number of Passengers



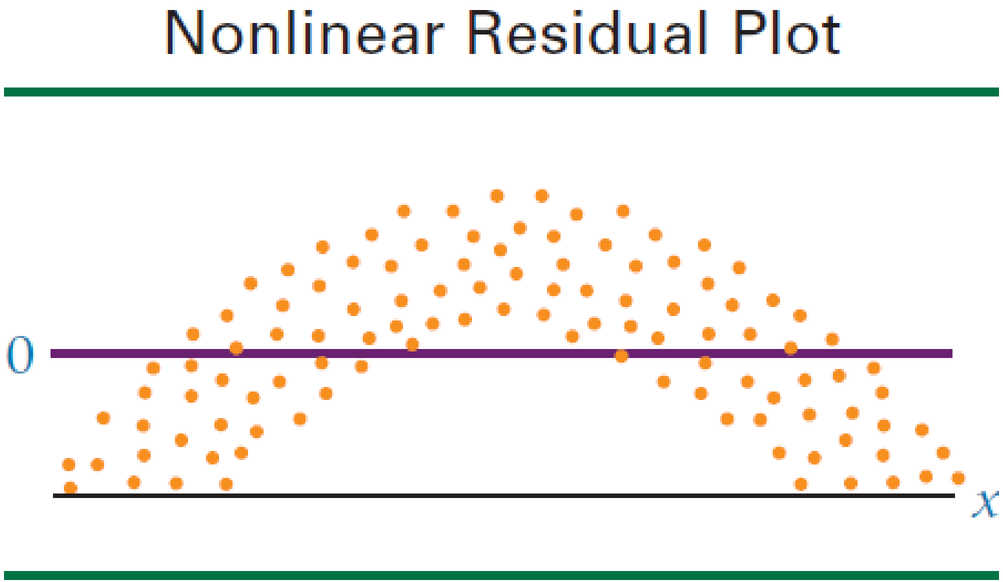


Outliers: data points that lie apart from the rest of the points. They can produce large residuals and affect the regression line.

Using Residuals to Test the Assumptions of the Regression Model

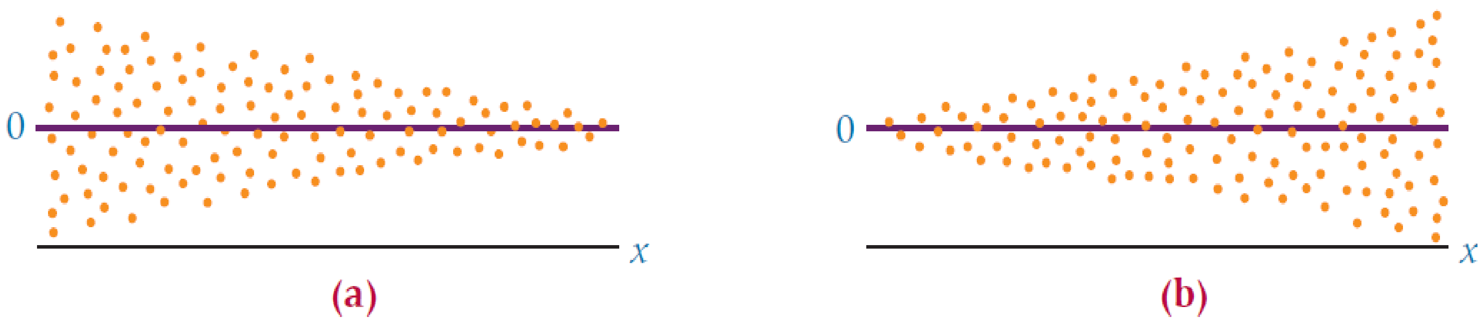
- The assumptions of the regression model
  - The model is linear
  - The error terms have constant variances
  - The error terms are independent
  - The error terms are normally distributed

Using Residuals to Test the Assumptions of the Regression Model

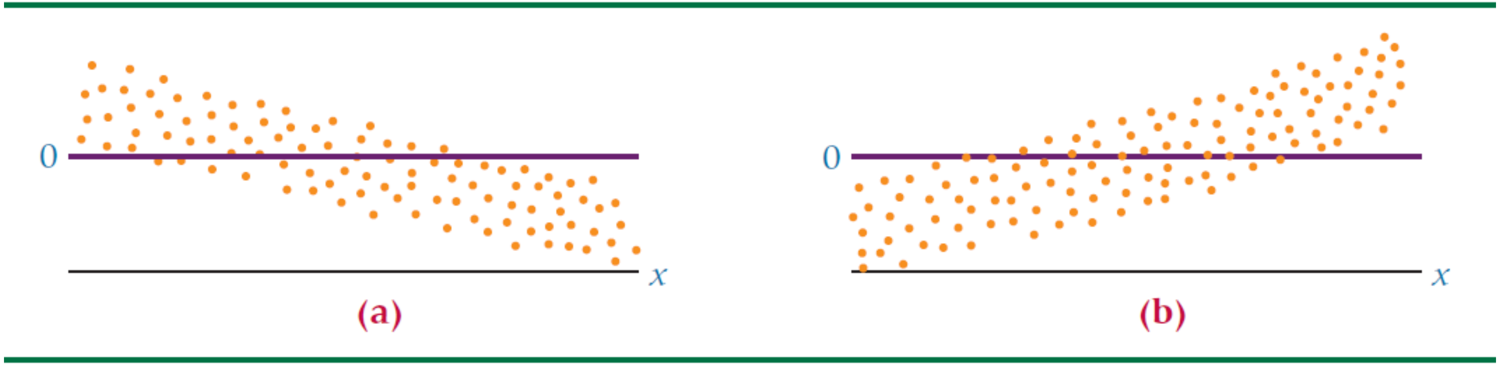


Using Residuals to Test the Assumptions of the Regression Model

Nonconstant Error Variance



Using Residuals to Test the Assumptions of the Regression Model



Graphs of Nonindependent Error Terms

Standard Error of the Estimate

- Residuals represent errors of estimation for individual points.
- A more useful measurement of error is the standard error of the estimate.
- The standard error of the estimate, denoted **se**, is a standard deviation of the error of the regression model.

Standard Error of the Estimate

Sum of Squares Error

$$SSE = \sum \left(Y - \hat{Y}\right)^2$$

$$= \sum Y^2 - b_0 \sum Y - b_1 \sum XY$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

Determining SSE for the Airline Cost Example

Number of Passengers X	Cost ($1,000) Y	Residual $Y - \hat{Y}$	$(Y - \hat{Y})^2$
61	4.28	.227	.05153
63	4.08	-.054	.00292
67	4.42	.123	.01513
69	4.17	-.208	.04326
70	4.48	.061	.00372
74	4.30	-.282	.07952
76	4.82	.157	.02465
81	4 .70	-.167	.02789
86	5.11	.040	.00160
91	5.13	-.144	.02074
95	5.64	.204	.04162
97	5.56	.042	.00176

$$\sum (Y - \hat{Y}) = -.001 \qquad \sum (Y - \hat{Y})^2 = .31434$$

Sum of squares of error = SSE = .31434

Standard Error of the Estimate for the Airline Cost Example

Sum of Squares Error

Standard Error of the Estimate

$$SSE = \sum \left(Y - \hat{Y}\right)^2$$

$$= 0.31434$$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

$$= \sqrt{\frac{0.31434}{10}}$$

$$= 0.1773$$

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- **Simple Linear Regression: Coefficient of Determination**
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Coefficient of Determination, r^2

- The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x)
- The coefficient of determination ranges from 0 to 1.
- An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x.
- An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x.

Coefficient of Determination, r^2

Surprise, Surprise!

$$r^2 = (r)^2$$

Coefficient of Determination= (Correlation Coefficient)^2

Coefficient of Determination

$$SS_{YY} = \sum\left(Y-\bar{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

$$SS_{YY} = \exp lained \ \text{var}\, iation + un\exp lained \ \text{var}\, iation$$

$$SS_{YY} = SSR + SSE$$

$$1 = \frac{SSR}{SS_{YY}} + \frac{SSE}{SS_{YY}}$$

$$r^2 = \frac{SSR}{SS_Y}$$

$$= 1 - \frac{SSE}{SS_Y}$$

$$= 1 - \frac{SSE}{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}$$

$$\boxed{0 \le r^2 \le 1}$$

Coefficient of Determination for the Airline Cost Example

$$SSE = 0.31434$$

$$SS_{YY} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 270.9251 - \frac{(56.69)^2}{12} = 3.11209$$

$$r^2 = 1 - \frac{SSE}{SS_Y}$$

$$= 1 - \frac{.31434}{3.11209}$$

$$= .899$$

89.9% of the variability of the cost of flying a Boeing 737 is accounted for by the number of passengers.

Coefficient of Determination : R Code

```
X=c(61,63,67,69,70,74,76,81,86,91,95,97)
Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)
SSYY=sum((Y-mean(Y))^2)
SSXY=sum((X-mean(X))*(Y-mean(Y)))
SSX=sum((X-mean(X))^2)
b1=  SSXY/SSX
b0=mean(Y)-b1*mean(X)
Y_Estimated=X*b1+b0
Residuals= Y-Y_Estimated
SSE=sum((Residuals -mean(Residuals))^2)
r2=1-SSE/SSYY
r2
cor(X,Y)^2
```

```
> X=c(61,63,67,69,70,74,76,81,86,91,95,97)
> Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)
> SSYY=sum((Y-mean(Y))^2)
> SSXY=sum((X-mean(X))*(Y-mean(Y)))
> SSX=sum((X-mean(X))^2)
> b1=  SSXY/SSX
> b0=mean(Y)-b1*mean(X)
> Y_Estimated=X*b1+b0
> Residuals= Y-Y_Estimated
> SSE=sum((Residuals -mean(Residuals))^2)
> r2=1-SSE/SSYY
> r2
[1] 0.8990839
> cor(X,Y)^2
[1] 0.8990839
```

Hypothesis Tests for the Slope of the Regression Model

- A hypothesis test can be conducted on the sample slope of the regression model to determine whether the population slope is significantly different from zero.
- Testing the slope of the regression line to determine whether the slope is different from zero is important.
- If the slope is not different from zero, the regression line is doing nothing more than the average line of y predicting y.

R Makes Life Easy: Airline Cost Example

```
X=c(61,63,67,69,70,74,76,81,86,91,95,97)
Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)

Model=lm(Y ~X) # lm() is the function to create linear model of Y from X
Model$coefficients
Model$residuals
Model$fitted.values
summary(Model)
```

Hypothesis Test: Airline Cost Example

```
> X=c(61,63,67,69,70,74,76,81,86,91,95,97)
> Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)
>
> Model=lm(Y ~X)
> Model$coefficients
(Intercept)           X
  1.5697928   0.0407016
> Model$residuals
          1           2           3           4           5           6           7           8
 0.22740971 -0.05399349  0.12320012 -0.20820308  0.06109532 -0.28171107  0.15688573 -0.1662226
          9          10          11          12
 0.03986975 -0.14363825  0.20355536  0.04215216
> Model$fitted.values
       1        2        3        4        5        6        7        8        9       10       11
4.052590 4.133993 4.296800 4.378203 4.418905 4.581711 4.663114 4.866622 5.070130 5.273638 5.436445
      12
5.517848
```

Hypothesis Test: Airline Cost Example

```
> summary(Model)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.28171 -0.14938  0.04101  0.13162  0.22741 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.569793   0.338083   4.643 0.000917 ***
X           0.040702   0.004312   9.439 2.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1772 on 10 degrees of freedom
Multiple R-squared:  0.8991, Adjusted R-squared:  0.889 
F-statistic: 89.09 on 1 and 10 DF,  p-value: 2.692e-06
```

Example II

Executives of a video rental chain want to predict the success of a potential new store. The company's researcher begins by gathering information on number of rentals and average family income from several of the chain's present outlets.

Rentals=c(710, 529,314,504,619,428,317,205,468,545,607,694)

Average_Family_Income_k=c(65,43,29,47,52,50,46,29,31,43,49,64)

Develop a regression model to predict the number of rentals per day by the average family income. Comment on the output.

Example II

model=lm(Rentals~Average_Family_Income_k)

summary(model)

```
> summary(model)

Call:
lm(formula = Rentals ~ Average_Family_Income_k)

Residuals:
    Min      1Q  Median      3Q     Max 
-181.54  -32.10    6.87   65.90  128.85 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)   
(Intercept)                9.729    115.515   0.084  0.93455   
Average_Family_Income_k   10.626      2.454   4.330  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.13 on 10 degrees of freedom
Multiple R-squared:  0.6522, Adjusted R-squared:  0.6174 
F-statistic: 18.75 on 1 and 10 DF,  p-value: 0.001489
```

Example II

Rentals=10.626* Average_Family_Income_k+ 9.729

The model says, there will be a 10.62 unit increase in average number of rentals for every $1000 increase in the average households income of the local neighborhood. ➔ more income, more spending on video rentals

So based on this model, what would be the average rental for a neighborhood where the average household income is $86k ?

Rentals=10.626* 86+ 9.729=923.565

How good is this model? Answer: the R-squared (R^2) is 0.652 that says the model explains 65% of the variability of the target variable (i.e. Rental) .

Example II

Use the following data to build model that predicts the flight duration (in hours) given the distance between the source and destination.

Origin	Destination	Distance in km	Flight duration	Flight duration in hours
London	Amsterdam	365	1h 10m	1.167
London	Budapest	1462	2h 20m	2.333
London	Bratislava	1285	2h 15m	2.250
Bratislava	Paris	1096	2h 5m	2.083
Bratislava	Berlin	517	1h 15m	2.250
Vienna	Dublin	1686	2h 50m	2.833
Vienna	Amsterdam	932	1h 55m	1.917
Amsterdam	Budapest	1160	2h 10m	2.167

Example II

Use the following data to build model that predicts the flight duration (in hours) given the distance between the source and destination.

```
Distance=c(365,1462,1285,1096,517,1686,932,1160)

Duration=c(1.167,2.333,2.25,2.083,2.25,2.833,1.917,2.167)


Model=lm(Duration~Distance)

Model$coefficients

summary(Model)
```

Example II

```
> Distance=c(365,1462,1285,1096,517,1686,932,1160)
> Duration=c(1.167,2.333,2.25,2.083,2.25,2.833,1.917,2.167)
>
> Model=lm(Duration~Distance)
> Model$coefficients
(Intercept)    Distance
1.233589015 0.000838679
> summary(Model)

Call:
lm(formula = Duration ~ Distance)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37271 -0.10536 -0.06554  0.01676  0.58281

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2335890  0.2911120   4.238  0.00545 **
Distance    0.0008387  0.0002547   3.292  0.01657 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3024 on 6 degrees of freedom
Multiple R-squared:  0.6437, Adjusted R-squared:  0.5843
F-statistic: 10.84 on 1 and 6 DF,  p-value: 0.01657
```

Example II

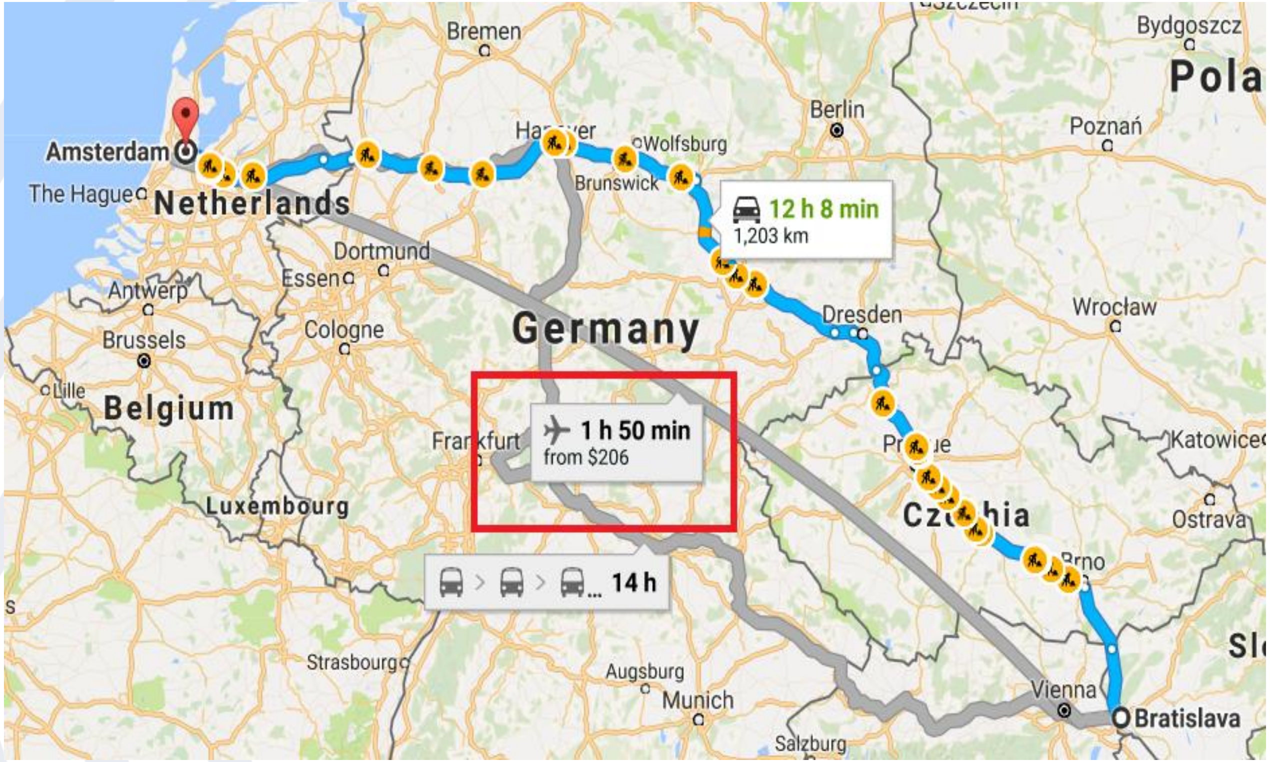
Duration= 0.0008387* Distance+ 1.2335890

We can reason that the flight duration time consists of two times - the first is the time to take off and the landing time; the second is the time that the airplane moves at a certain speed in the air. The first time is some constant. The second time depends linearly on the distance.

With this model, what how long will be the flight from Bratislava to Amsterdam if the distance is 978km?

Duration= 0.0008387* 978+1.2335890=2.053838 => 2:03’ (2 hours and 3 minutes)

Example II


Bremen
Bydgoszcz
Pola
Berlin
Poznań
Amsterdam
Wolfsburg
Brunswick
The Hague
Netherlands
12 h 8 min
1,203 km
Dortmund
Essen
Antwerp
Wrocław
Dresden
Brussels
Cologne
Germany
Lille
Belgium
1 h 50 min
from $206
Katowice
Luxembourg
Ostrava
14 h
Strasbourg
Augsburg
Munich
Vienna
Bratislava
Salzburg


NOT BAD

Example II: Even Lazier With R. predict() function

Duration= 0.0008387* Distance+ 1.2335890

Predicted_value <- predict(Model, data.frame(Distance=c(978)))

```
> Predicted_value  <- predict(Model,  data.frame(Distance=c(978)))
> Predicted_value
       1
2.053817
```

Example II: Even Lazier With R. predict() function

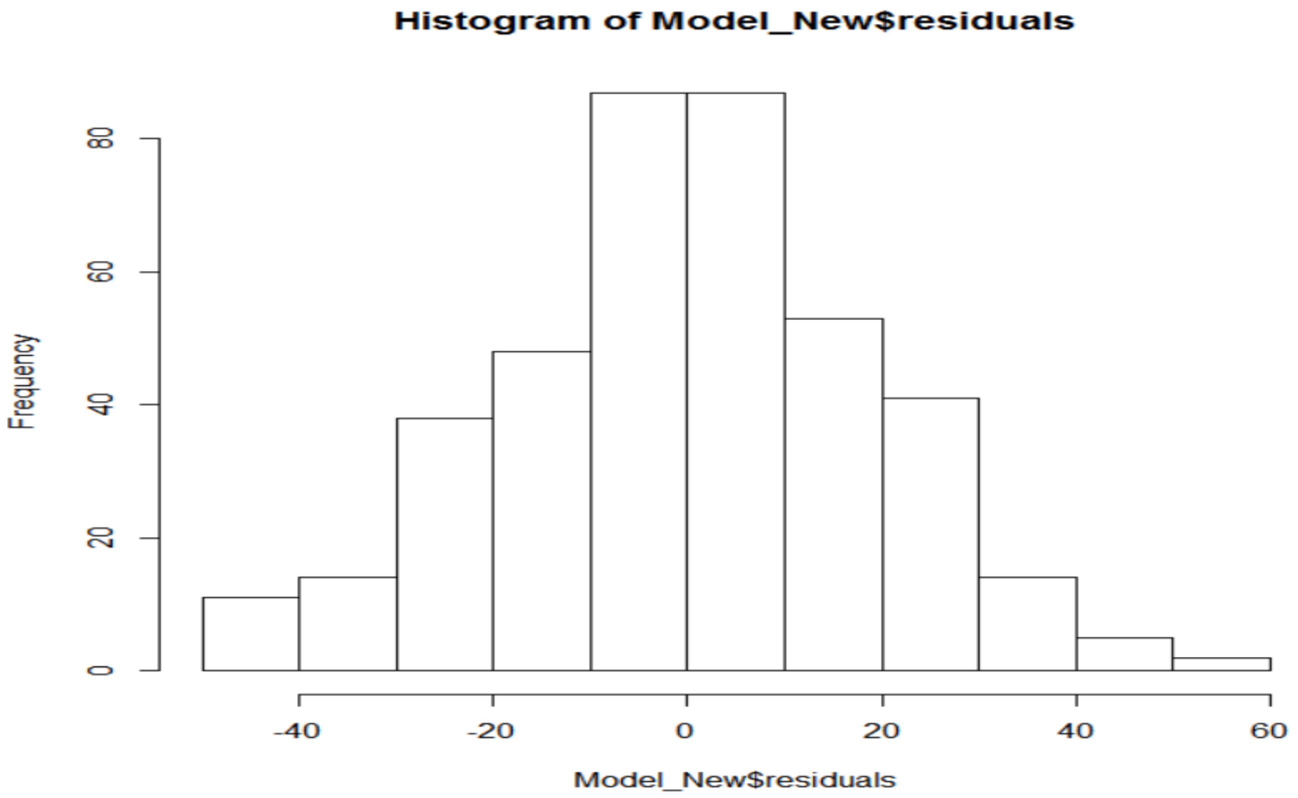
```
library(ISLR) #install if you get error i.e. install.packages('ISLR')
Model_New=lm(Carseats$Price~Carseats$CompPrice)
Model_New=lm(Price~CompPrice,data=Carseats)
summary(Model_New)
hist(Model_New$residuals)
```

Example III

```
> library(ISLR) #install if you get error i.e. install.packages('ISLR')
> Model_New=lm(Carseats$Price~Carseats$CompPrice)
> Model_New=lm(Price~CompPrice,data=Carseats)
> summary(Model_New)

Call:
lm(formula = Price ~ CompPrice, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.473 -12.183   0.197  12.925  56.540 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.94110    7.90436   0.372     0.71    
CompPrice    0.90301    0.06278  14.384   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.23 on 398 degrees of freedom
Multiple R-squared:  0.342, Adjusted R-squared:  0.3404 
F-statistic: 206.9 on 1 and 398 DF,  p-value: < 2.2e-16

> hist(Model_New$residuals)
```

Same commands

Example III: Residuals Has Normal Distribution

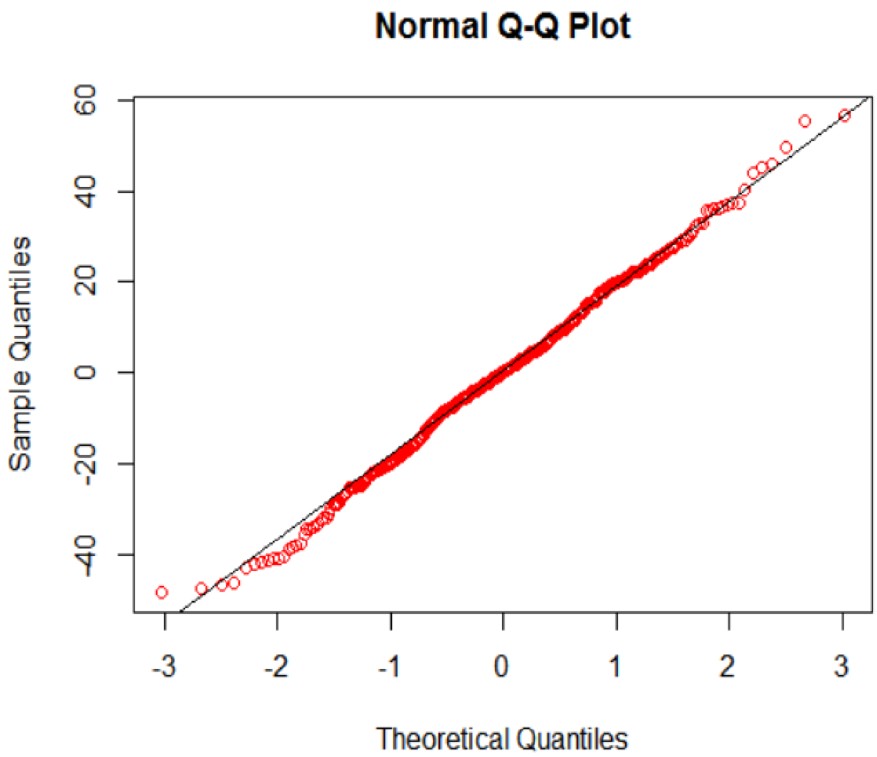


Example II: qqnorm() function

```
qqnorm(Model_New$residuals,col="red")
qqline(Model_New$residuals)
```

More effective way to check normality of residuals!

We need samples to follow the Line.



Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- **Simple Linear Regression: Prediction Interval**
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

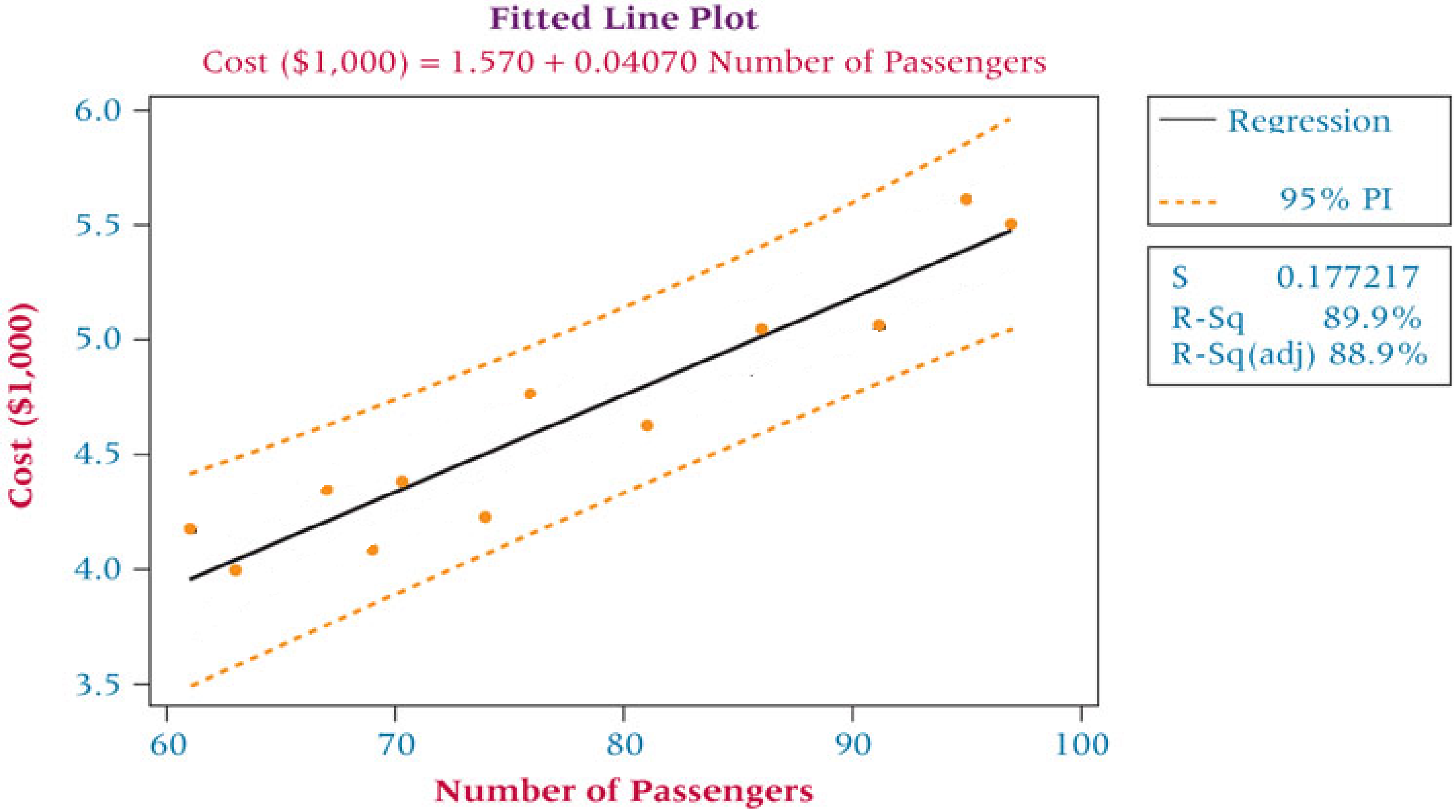
Prediction Interval to Estimate Y for a given value of X

$$\hat{Y} \pm t_{\frac{\alpha}{2}, n-2} S_e \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\mathrm{SSxx}}}$$

$where:\ x_0 =$ a particular value of x

$$\mathrm{SSxx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

Cost ~ Number of Passengers Example



Example IV

No_Pass_X=c(61,63,67,69,70,74,76,81,86,91,95,97)

Cost_Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)

Build a linear regression model of cost based on the number of passengers. Then calculate the 90% confidence interval for estimating the cost of the flight with 84 passengers.

Example IV

```
No_Pass_X=c(61,63,67,69,70,74,76,81,86,91,95,97)

Cost_Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)

Model=lm(Cost_Y~No_Pass_X)

summary(Model)

SSXX=sum((No_Pass_X-mean(No_Pass_X))^2)
SSXX
```

CI=95% => α=0.05 => we need

$t_{0.025,df=n-2=10}$ = qt(0.025,10) = -2.22

$t_{0.075,df=n-2=10}$ = qt(0.975,10) = 2.22

Example IV

```
> No_Pass_X=c(61,63,67,69,70,74,76,81,86,91,95,97)
> Cost_Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)
> Model=lm(Cost_Y~No_Pass_X)
> summary(Model)

Call:
lm(formula = Cost_Y ~ No_Pass_X)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.28171 -0.14938  0.04101  0.13162  0.22741 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.569793   0.338083   4.643 0.000917 ***
No_Pass_X   0.040702   0.004312   9.439 2.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1772 on 10 degrees of freedom
Multiple R-squared:  0.8991, Adjusted R-squared:  0.889 
F-statistic: 89.09 on 1 and 10 DF,  p-value: 2.692e-06

> SSXX=sum((No_Pass_X-mean(No_Pass_X))^2)
> SSXX
[1] 1689
> 
```

Example IV

```
Y_hat=predict(Model,  data.frame(No_Pass_X=c(84)))
Y_hat
##Alternatively you could write as
Model$coefficients
Y_hat=Model$coefficients[1]+84*Model$coefficients[2]
Y_hat
mean(No_Pass_X)
```

Example IV

```
> Y_hat=predict(Model,  data.frame(No_Pass_X=c(84)))
> Y_hat
       1
4.988727
> ##Alternatively you could write as
> Model$coefficients
(Intercept)   No_Pass_X
  1.5697928   0.0407016
> Y_hat=Model$coefficients[1]+84*Model$coefficients[2]
> Y_hat
(Intercept)
   4.988727
> mean(No_Pass_X)
[1] 77.5
```

Prediction Interval to Estimate *Y* for a given value of *X*

$$4.98 - 2.22 \times 0.177\sqrt{1 + \frac{1}{12} + \frac{(84 - 77.5)^\wedge 2}{1689}} \le Y_{84} \le 4.98 + 2.22 \times 0.177\sqrt{1 + \frac{1}{12} + \frac{(84 - 77.5)^\wedge 2}{1689}}$$

$$4.57 \le Y_{84} \le 5.40$$

So, we probability of 95%, the cost of a flight with 84 passengers is between \$4,570 and \$5,400

Everything in ONE Line with R

```
> predict(Model,  data.frame(No_Pass_X=c(84)),interval = "prediction",level = 0.95)
       fit      lwr      upr
1 4.988727 4.573021 5.404434
```

You are Welcome !

Intervals for Coefficient Estimates

```
No_Pass_X=c(61,63,67,69,70,74,76,81,86,91,95,97)

Cost_Y=c(4.28,4.08,4.42,4.17,4.48,4.3,4.82,4.7,5.11,5.13,5.64,5.56)

Model=lm(Cost_Y~No_Pass_X)

confint(Model,  level = 0.9)
```

```
> confint(Model,  level = 0.9)
                   5 %       95 %
(Intercept) 0.95703055 2.18255500
No_Pass_X   0.03288603 0.04851716
```

Example V

Is the amount of money spent by companies on advertising a function of the total sales of the company? Show are sales income and advertising cost data for seven companies published by Advertising Age.

Company	Advertising ($ millions)	Sales ($ billions)
Wal-Mart	1,073	351.1
Procter & Gamble	4,898	68.2
AT&T	3,345	63.1
General Motors	3,296	207.3
Verizon	2,822	93.2
Ford Motor	2,577	160.1
Hewlett-Packard	829	91.7

Example V

Use the data to develop a regression line to predict the amount of advertising by sales. Compute s_e and r^2. Assuming $\alpha = .05$, test the slope of the regression line. Comment on the strength of the regression model.

```
Advertising_Million=c(1073,4898,3345,3296,2822,2577,829)
Sales_Billion=c(351.1,68.2,63.1,207.3,93.2,160.1,91.7)
```

```
Model=lm(Sales_Billion~Advertising_Million)
summary(Model)
```

Example V

```
> Model=lm(Sales_Billion~Advertising_Million)
> summary(Model)

Call:
lm(formula = Sales_Billion ~ Advertising_Million)

Residuals:
       1        2        3        4        5        6        7 
 144.467    0.579  -60.962   81.458  -49.869    8.127 -123.800 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)  
(Intercept)         245.62874   86.26639   2.847   0.0359 *
Advertising_Million  -0.03634    0.02887  -1.259   0.2637  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.1 on 5 degrees of freedom
Multiple R-squared:  0.2406,	Adjusted R-squared:  0.08874 
F-statistic: 1.584 on 1 and 5 DF,  p-value: 0.2637
```

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- **Multiple Linear Regression**
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Multiple Regression Models

- Regression analysis with two or more independent variables or with at least one nonlinear predictor is called multiple regression analysis.

Multiple Regression Model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

$where:$ $\hat{Y} =$ predicted value of Y

$b_0 =$ estimate of regression constant

$b_1 =$ estimate of regression coefficient 1

$b_2 =$ estimate of regression coefficient 2

$b_3 =$ estimate of regression coefficient 3

$b_k =$ estimate of regression coefficient k

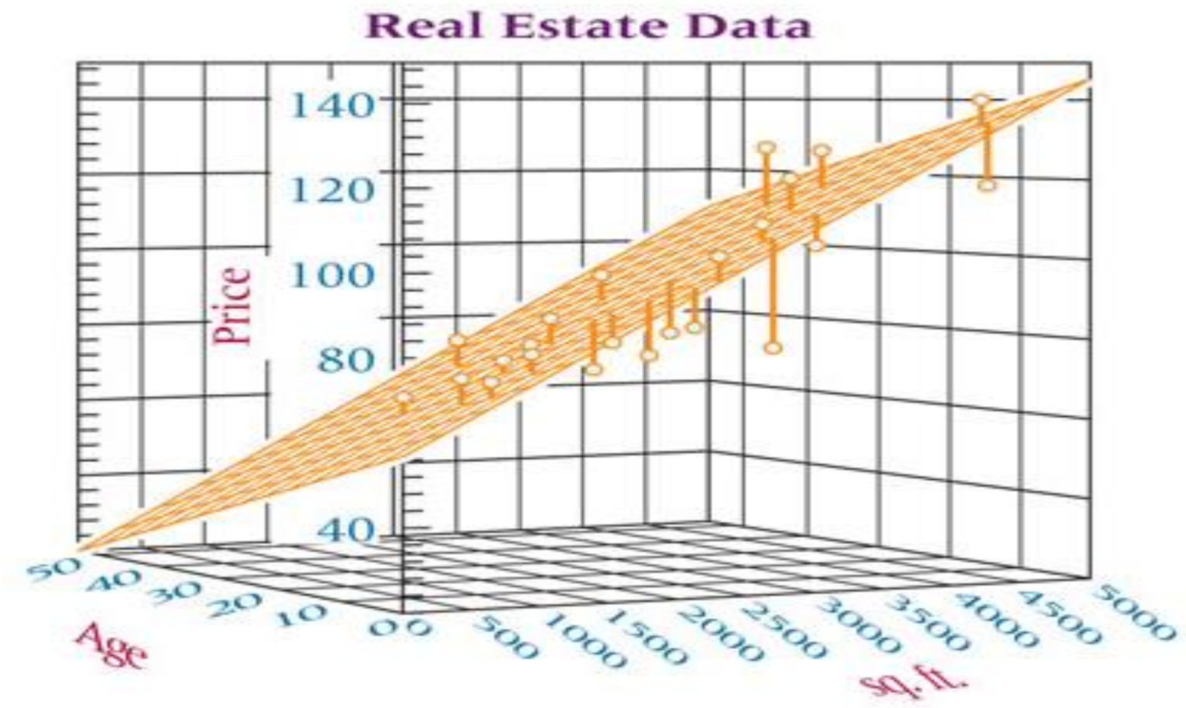
$k =$ number of independent variables

Multiple Regression Model with Two Independent Variables (First-Order)

- The simplest multiple regression model is one constructed with two independent variables.

- In such multiple regression analysis, the resulting model produces a response surface.

Response Plane for First-Order Two-Predictor Multiple Regression Model



Determining the Multiple Regression Equation

- The simple regression equations for determining the sample slope and intercept given in earlier material are the result of using methods of calculus to minimize the sum of squares of error for the regression model.

- The formulas are established to meet an objective of minimizing the sum of squares of error for the model.

- The regression analysis shown here is referred to as least squares analysis.

Example VI

- A real estate study was conducted in a small Louisiana city to determine what variables, if any, are related to the market price of a home. Suppose the researcher wants to develop a regression model to predict the market price of a home by two variables, "total number of square feet in the house" and "the age of the house."

Real Estate Data

Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2	Observation	Market Price ($1,000) Y	Square Feet X_1	Age (Years) X_2
1	63.0	1,605	35	13	79.7	2,121	14
2	65.1	2,489	45	14	84.5	2,485	9
3	69.9	1,553	20	15	96.0	2,300	19
4	76.8	2,404	32	16	109.5	2,714	4
5	73.9	1,884	25	17	102.5	2,463	5
6	77.9	1,558	14	18	121.0	3,076	7
7	74.9	1,748	8	19	104.9	3,048	3
8	78.0	3,105	10	20	128.0	3,267	6
9	79.0	1,682	28	21	129.0	3,069	10
10	63.4	2,470	30	22	117.9	4,765	11
11	79.5	1,820	2	23	140.0	4,540	8
12	83.9	2,143	6				

Example Model Output in Minitab (not R)

The regression equation is

Price = 57.4 + 0.0177 Sq.Feet - 0.666 Age

Predictor	Coef	StDev	T	P
Constant	57.35	10.01	5.73	0.000
Sq.Feet	0.017718	0.003146	5.63	0.000
Age	-0.6663	0.2280	-2.92	0.008

S = 11.96   R-Sq = 74.1%   R-Sq(adj) = 71.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	0.000
Residual Error	20	2861.0	143.1		
Total	22	11050.7			

Predicting the Price of Home

$$\hat{Y} = 57.4 + 0.0177\,X_1 - 0.666\,X_2$$

$$For\ \ X_1 = 2500 \text{ and } X_2 = 12,$$

$$\hat{Y} = 57.4 + 0.0177(2500) - 0.666(12)$$

$$= 93.658 \text{ thousand dollars}$$

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- **Multiple Linear Regression: Evaluating Multiple Regression Models**
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Evaluating the Multiple Regression Model

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

H_a: At least one of the regression coefficients is $\neq 0$

Testing the Overall Model

F-test

$$
\begin{array}{ll}
H_0: \beta_1 = 0 & H_0: \beta_3 = 0 \\
H_a: \beta_1 \neq 0 & H_a: \beta_3 \neq 0 \\
& \vdots \\
H_0: \beta_2 = 0 & H_0: \beta_k = 0 \\
H_a: \beta_2 \neq 0 & H_a: \beta_k \neq 0
\end{array}
$$

Significance Tests for Individual Regression Coefficients

t-test

Testing the Overall Model for the Real Estate Example

- F-test: A rejection of the null hypothesis indicates that at least one of the independent variables is adding significant predictability for y.

- The *F* value is 28.63=> because p = 0.000.

- The null hypothesis is rejected, and there is at least one significant predictor of house price in this analysis.

KENT STATE UNIVERSITY

Testing the Overall Model for the Real Estate Example

$H_0: \beta_1 = \beta_2 = 0$

H_a : At least one of the regression coefficients is $\neq 0$

$F_{.01,2,20} = 5.85$

$F_{Cal} = 28.63 > 5.85$, reject H_0.

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1} \qquad F = \frac{MSR}{MSE}$$

ANOVA

	df	SS	MS	F	p
Regression	2	8189.72[illegible]	4094.86	28.63	.000
Residual (Error)	20	2861.017	143.1		
Total	22	11050.74			

Testing the Individual Coefficients of Model for the Real Estate Example

- With simple regression, a *t* test of the slope of the regression line is used to determine whether the population slope of the regression line is different from zero. This is done for <u>each coefficient individually</u>.

- Fail to reject the null hypothesis - the regression model has no significant predictability for the dependent variable.

Significance Test of the Regression Coefficients for the Real Estate Example

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0$$

$$H_a: \beta_2 \neq 0$$

$t_{.025,20} = 2.086$

$t_{Cal} = 5.63 > 2.086$, **reject** H_0.

	Coefficients	Std Dev	t Stat	p
x_1 (Sq.Feet)	0.0177	0.003146	5.63	.000
x_2 (Age)	-0.666	**0.2280**	-2.92	.008

R Implementation of Real Estate Example

```
Market_Price_K=c(63,65.1,69.9,76.8,73.9,77.9,74.9,78,79,83.4,79.5,83.9,79.7,84.5,96,109.5,102.5,121,104.9,128,129,117.9,140)

Square_Feet=c(1605,2489,1553,2404,1884,1558,1748,3105,1682,2470,1820,2143,2121,2485,2300,2714,2463,3076,3048,3267,3069,4765,4540)

House_Age=c(35,45,20,32,25,14,8,10,28,30,2,6,14,9,19,4,5,7,3,6,10,11,8)

Model=lm(Market_Price_K~Square_Feet+House_Age)
summary(Model)
anova(Model)
```

R Implementation of Real Estate Example

```
> summary(Model)

Call:
lm(formula = Market_Price_K ~ Square_Feet + House_Age)

Residuals:
     Min       1Q   Median       3Q      Max 
-27.7018  -6.8938  -0.1728   7.1340  23.9361 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 57.350746  10.007152   5.731 1.31e-05 ***
Square_Feet  0.017718   0.003146   5.633 1.64e-05 ***
House_Age   -0.666348   0.227997  -2.923  0.00842 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.96 on 20 degrees of freedom
Multiple R-squared:  0.7411,	Adjusted R-squared:  0.7152 
F-statistic: 28.63 on 2 and 20 DF,  p-value: 1.353e-06
```

R Implementation of Real Estate Example

```
> anova(Model)
Analysis of Variance Table

Response: Market_Price_K
            Df Sum Sq Mean Sq F value    Pr(>F)    
Square_Feet  1 6967.8  6967.8 48.7087 8.976e-07 ***
House_Age    1 1221.9  1221.9  8.5417  0.008418 ** 
Residuals   20 2861.0   143.1                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Coefficient of Multiple Determination (R^2)

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	2	8189.7	4094.9	28.63	.000
Error	20	2861.0	143.1		
Total	22	11050.7			

SSR, SSE, SS_{yy}

$$R^2 = \frac{\text{SSR}}{\text{SS}_{yy}} = \frac{8189.7}{11050.7} = .741$$

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{2861.0}{11050.7} = .741$$

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- **Multiple Linear Regression: Indicator (Dummy) Variables**
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects

Indicator (Dummy) Variables

- Some variables are referred to as qualitative variables
  - Qualitative variables do not yield quantifiable outcomes
  - Qualitative variables yield nominal- or ordinal-level information; used more to categorize items.
- Qualitative variables are referred to as indicator, or dummy variables

Monthly Salary Example

As an example, consider the issue of sex discrimination in the salary earnings of workers in some industries.

In examining this issue, suppose a random sample of 15 workers is drawn from a pool of employed laborers in a particular industry and the workers' average monthly salaries are determined, along with their age and gender. The data are shown in the following table. As sex can be only male or female, this variable is coded as a dummy variable with 0 = female, 1 = male.

Data for the Monthly Salary Example

Monthly Salary ($1,000)	Age (10 years)	Sex (1 = male, 0 = female)
2.548	3.2	1
2.629	3.8	1
2.011	2.7	0
2.229	3.4	0
2.746	3.6	1
2.528	4.1	1
2.018	3.8	0
2.190	3.4	0
2.551	3.3	1
1.985	3.2	0
2.610	3.5	1
2.432	2.9	1
2.215	3.3	0
1.990	2.8	0
2.585	3.5	1

Regression Output for the Monthly Salary Example

The regression equation is
Salary = 1.732 + 0.111 Age + 0.459 Gender

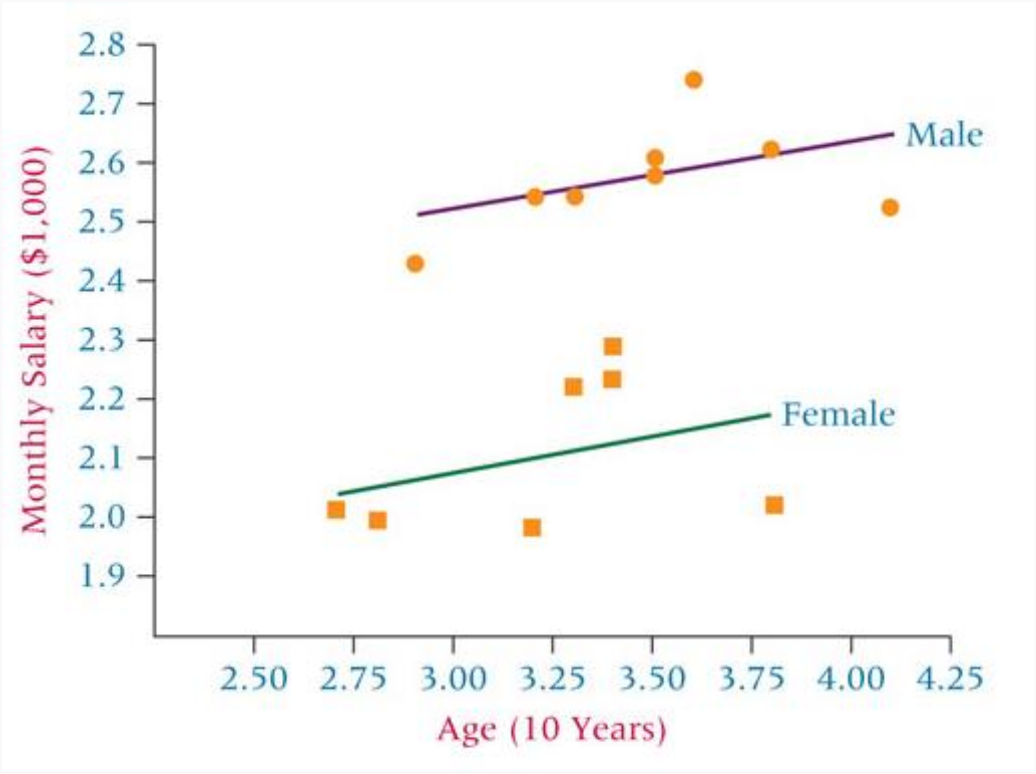
Predictor	Coef	StDev	T	P
Constant	1.7321	0.2356	7.35	0.000
Age	0.11122	0.07208	1.54	0.149
Gender	0.45868	0.05346	8.58	0.000

S = 0.09679 R-Sq = 89.0% R-Sq(adj) = 87.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0.90949	0.45474	48.54	0.000
Error	12	0.11242	0.00937		
Total	14	1.02191			

Regression Output
for the Monthly Salary Example



Dummy Variables with More than Two Levels

- You can use binary coding 1 and 0, only if you have no more than two levels (e.g. Sex: only male and female levels)

- For example, if the there are 3 levels and you code the categorical data as 1,2 and 3, the resulting model is going to be WRONG (because you are assuming that the distance between the first and the second category levels is the same as the distance between the second and the third levels).

- In such cases, you should explicitly defined the variable as a factor.

R Example: Predicting Sales of Baby Car Seats

```
library(ISLR)  # install.packages('ISLR') if you had errors
MyData<-Carseats[,1:8]
str(MyData) # shows which variables are factor or numerical
Model=lm(Sales~.,data=MyData) #Use all other columns to predict Sales
 summary(Model)
```

R Example: Predicting Sales of Baby Car Seats

```
> library(ISLR)  # install.packages('ISLR') if you had errors
> MyData<-Carseats[,1:8]
> str(MyData) # shows which variables are factor or numerical
'data.frame': 400 obs. of  8 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 .
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
> Model=lm(Sales~.,data=MyData) #Use all other columns to predict Sales
```

Location of the Shelve in the Store

R Example: Predicting Sales of Baby Car Seats

```
> summary(Model)

Call:
lm(formula = Sales ~ ., data = MyData)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7634 -0.6869  0.0231  0.6564  3.3245 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.3592591  0.5241924   10.22   <2e-16 ***
CompPrice        0.0929101  0.0041451   22.41   <2e-16 ***
Income           0.0158393  0.0018395    8.61   <2e-16 ***
Advertising      0.1141444  0.0080124   14.25   <2e-16 ***
Population       0.0003004  0.0003622    0.83    0.407    
Price           -0.0953926  0.0026726  -35.69   <2e-16 ***
ShelveLocGood    4.8399824  0.1526478   31.71   <2e-16 ***
ShelveLocMedium  1.9570591  0.1255736   15.59   <2e-16 ***
Age             -0.0459990  0.0031817  -14.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 391 degrees of freedom
Multiple R-squared:  0.8722, Adjusted R-squared:  0.8696 
F-statistic: 333.6 on 8 and 391 DF,  p-value: < 2.2e-16
```

R Example: Predicting Sales of Baby Car Seats

R considers one of the levels (by default the first level alphabetically, in this case “bad”) as a default value and then suggests how the model can be modified for cases when the two other level values (i.e. “good” and “medium”) are true.

In the previous examples, the coefficient for ShelvLocGood is 4.83, which means that if the Shelve Location was “Good” the sales would have been 4.83 units (in thousands) more compared to when the Shelve Location was “Bad” when all other variables are the same. Similarly, the sales would be 1.95 units more if the Shelve Location was “Medium” when compared to the base where the Shelve Location is “Bad”

R Example: Predicting Sales of Baby Car Seats

Lets see what would happen if "ShelvLoc" was considered as a numeric variable:

```
library(ISLR)  # install.packages('ISLR') if you had errors
MyData<-Carseats[,1:8]
MyData$ShelveLoc=as.numeric(MyData$ShelveLoc)
str(MyData)
Model=lm(Sales~.,data=MyData)
summary(Model)
```

R Example: Predicting Sales of Baby Car Seats

```
> MyData$ShelveLoc=as.numeric(MyData$ShelveLoc)
> str(MyData)
'data.frame':	400 obs. of  8 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ..
 $ CompPrice  : num  138 111 113 117 141 124 115
 $ Income     : num  73 48 35 100 64 113 105 81
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ..
 $ Population : num  276 260 269 466 340 501 45
 $ Price      : num  120 83 80 97 128 72 108 120
 $ ShelveLoc  : num  1 2 3 3 1 1 3 2 3 3 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76
```

R Example: Predicting Sales of Baby Car Seats

```
> Model=lm(Sales~.,data=MyData)
> summary(Model)

Call:
lm(formula = Sales ~ ., data = MyData)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0869 -1.2875 -0.4220  0.9601  4.8720 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.7852660  0.9782459   5.914 7.28e-09 ***
CompPrice    0.0932986  0.0075747  12.317  < 2e-16 ***
Income       0.0143093  0.0033603   4.258 2.58e-05 ***
Advertising  0.1289357  0.0146146   8.822  < 2e-16 ***
Population   0.0001162  0.0006618   0.176    0.861    
Price       -0.0926404  0.0048811 -18.979  < 2e-16 ***
ShelveLoc    0.6079784  0.1125391   5.402 1.14e-07 ***
Age         -0.0467581  0.0058141  -8.042 1.06e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 1.864 on 392 degrees of freedom
Multiple R-squared:  0.5722,	Adjusted R-squared:  0.5645 
F-statistic: 74.89 on 7 and 392 DF,  p-value: < 2.2e-16
```

Significant Loss of Model’s Accuracy

R^2 was 0.87 Previously!

R Example: Predicting Sales of Baby Car Seats

One reason for the very poor performance of the model was the order of the mapping levels to number, the default is alphabetical order resulted in

“Bad” -> 1, “Good” ->2 and “Medium” -> 3 , let’s reorder Good and Medium and see if the performance improves:

```
MyData$ShelveLoc[MyData$ShelveLoc==3]=4
MyData$ShelveLoc[MyData$ShelveLoc==2]=3
MyData$ShelveLoc[MyData$ShelveLoc==4]=2
```

We changed all the 3s to 4 (could be any other unused number), then all the 2s to 3 and now all the 4s (i.e. old 3s) to 2. So we have:

R Example: Predicting Sales of Baby Car Seats

Let's try this code now:

```
library(ISLR)  # install.packages('ISLR') if you had errors
MyData<-Carseats[,1:8]
MyData$ShelveLoc=as.numeric(MyData$ShelveLoc)
MyData$ShelveLoc[MyData$ShelveLoc==3]=4
MyData$ShelveLoc[MyData$ShelveLoc==2]=3
MyData$ShelveLoc[MyData$ShelveLoc==4]=2
Model=lm(Sales~.,data=MyData)
summary(Model)
```

R Example: Predicting Sales of Baby Car Seats

```
> summary(Model)

Call:
lm(formula = Sales ~ ., data = MyData)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.65838 -0.72204 -0.02325  0.67879  3.10035 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.7207965  0.5492733   4.953 1.09e-06 ***
CompPrice    0.0927108  0.0042453  21.839  < 2e-16 ***
Income       0.0162403  0.0018819   8.630  < 2e-16 ***
Advertising  0.1143102  0.0082065  13.929  < 2e-16 ***
Population   0.0003560  0.0003707   0.960    0.338    
Price       -0.0952571  0.0027372 -34.801  < 2e-16 ***
ShelveLoc    2.4058070  0.0781070  30.801  < 2e-16 ***
Age         -0.0467704  0.0032541 -14.373  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 392 degrees of freedom
Multiple R-squared:  0.8656, Adjusted R-squared:  0.8632 
F-statistic: 360.7 on 7 and 392 DF,  p-value: < 2.2e-16
```

Much better but the first model is still better

Agenda

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- **Multiple Linear Regression: Variable Importance**

Variable Importance

Which variables are more important in the model? i.e. the loss of which variables impact the model accuracy most?

The coefficient values doesn't tell you anything. i.e. higher coefficient does not mean higher importance of a variable. For example if the "income" is presented in ($1000) the coefficient is going to be 1000 times smaller while the importance of the variable is the same in both cases.

The importance of the variable is defined by how much of the total variability is explained by that variable. We use **ANOVA** (Analysis of Variance) for this.

Variable Importance

```
library(ISLR)  # install.packages('ISLR') if you had errors
MyData<-Carseats[,1:8]
Model=lm(Sales~.,data=MyData)
anova(Model)
T=anova(Model)
T$Variable_Importance_Perentage=T[,2]/sum(T[,2])
T
```

Variable Importance

```
> T$Variable_Importance_Perentage=T[,2]/sum(T[,2])
> T
Analysis of Variance Table

Response: Sales
             Df  Sum Sq Mean Sq   F value   Pr(>F) Variable_Importance_Perentage
CompPrice     1   13.07   13.07   12.5632 0.00044                       0.00411
Income        1   79.07   79.07   76.0262 0.00000                       0.02485
Advertising   1  219.35  219.35  210.8985 0.00000                       0.06893
Population    1    0.38    0.38    0.3677 0.54463                       0.00012
Price         1 1198.87 1198.87 1152.6682 0.00000                       0.37673
ShelveLoc     2 1047.47  523.74  503.5551 0.00000                       0.32916
Age           1  217.39  217.39  209.0108 0.00000                       0.06831
Residuals   391  406.67    1.04                                         0.12779
```

Nearly 70% of variability is explained by these two variables

Remember R2 was .87 which means 13% of the variability was not explained (residuals)

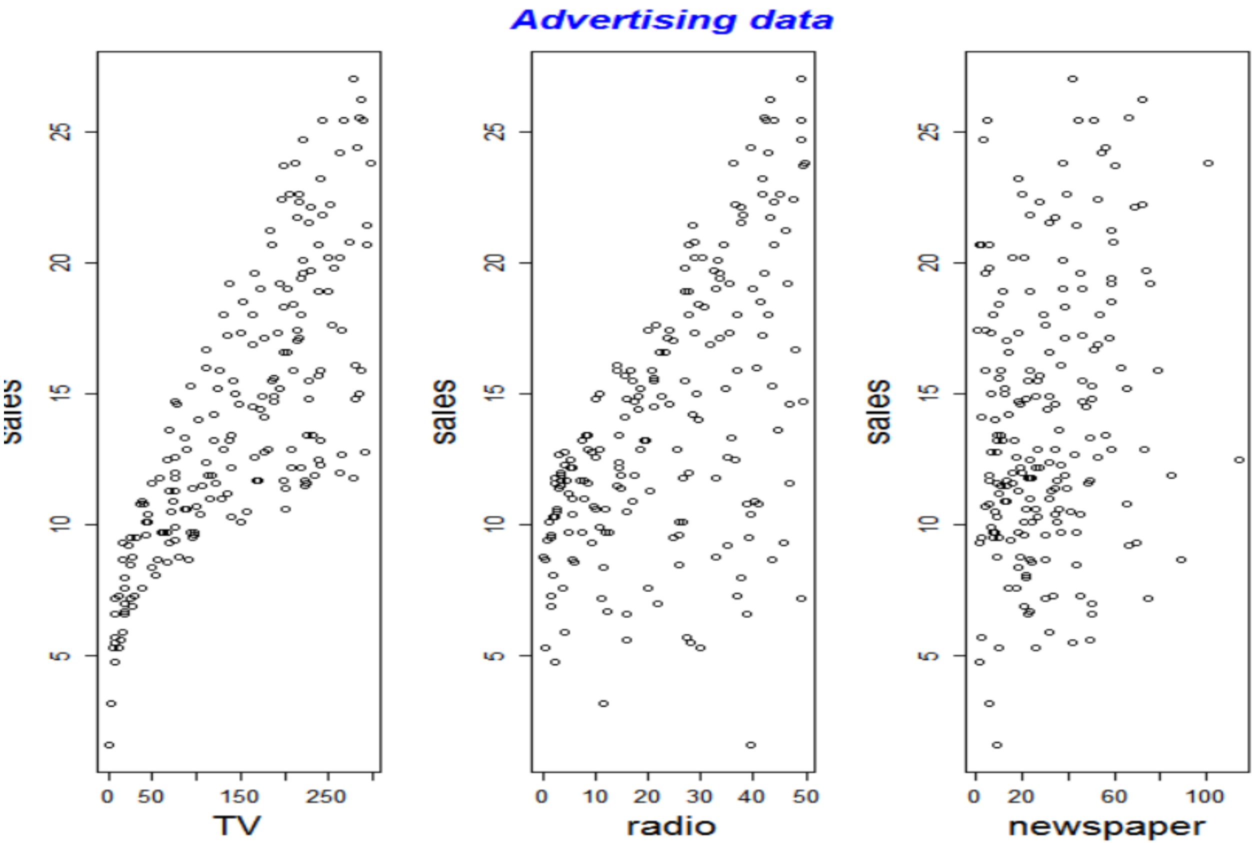
What We Have Covered!

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- **Simple Vs. Multiple Regression: Correlation Effects**

Simple Vs. Multiple Regression: Correlation Effects

- Simple regression models may not tell you the whole story, specially predictors are correlated
- For example, a single predictor may show some predictive power in describing the target variable which could be merely due to correlation effects
- Example consider y the response (target) and x1 and x2, two correlated variables, as predictors. x1 has high descriptive power, x2 does not.
- If you build a multiple regression model, y~x1+x2, the effect of x2 would be small. But, in a simple model y~x2, x2 seems to be predictive of y because in absence of x1, x2 receives the credit due to the correlation to x1. When x1 presents, it can speak for itself so x2 can be ignored!

Simple Vs. Multiple Regression: Example



Simple Vs. Multiple Regression: Example

Question we would like to answer:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy (interaction) among the advertising media?

Simple Vs. Multiple Regression: Example

```
Advertising = read.csv("http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv",
                row.names=1)

summary(lm(sales ~ TV, data= Advertising )) #simple regression
summary(lm(sales ~ radio, data= Advertising )) #simple regression
summary(lm(sales ~ newspaper, data= Advertising )) #simple regression
cor(Advertising[,1:3] )
summary(lm(sales ~ ., data= Advertising )) #multiple regression
```

Simple Regression: Example

```
> summary(lm(sales ~ TV, data= Advertising ))

Call:
lm(formula = sales ~ TV, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3860 -1.9545 -0.1913  2.0671  7.2124 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Simple Regression: Radio

```
> summary(lm(sales ~ radio, data= Advertising ))

Call:
lm(formula = sales ~ radio, data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7305  -2.1324   0.7707   2.7775   8.1810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
radio        0.20250    0.02041   9.921   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,	Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

Simple Regression: Newspaper

```
> summary(lm(sales ~ newspaper, data= Advertising ))

Call:
lm(formula = sales ~ newspaper, data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2272  -3.3873  -0.8392   3.5059  12.7751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212, Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Correlations

```
> cor(Advertising )
                  TV      radio  newspaper     sales
TV        1.00000000 0.05480866 0.05664787 0.7822244
radio     0.05480866 1.00000000 0.35410375 0.5762226
newspaper 0.05664787 0.35410375 1.00000000 0.2282990
sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Multiple Regression

```
> summary(lm(sales ~ ., data= Advertising )) #multiple regression

Call:
lm(formula = sales ~ ., data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

86% change the coefficient for newspaper is zero i.e. no predictive power.

Together, we are all stronger! R^2 is higher than individual Simple Regression Models.

Multiple Regression: ANOVA

Newspaper doesn't explains almost any of the variability .

```
> anova(lm(sales ~ ., data= Advertising ))
Analysis of Variance Table

Response: sales
           Df Sum Sq Mean Sq   F value Pr(>F)
TV          1 3314.6  3314.6 1166.7308 <2e-16 ***
radio       1 1545.6  1545.6  544.0501 <2e-16 ***
newspaper   1    0.1     0.1    0.0312 0.8599
Residuals 196  556.8     2.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

86% change the coefficient for newspaper is zero i.e. no predictive power.

What We Have Covered!

- Quick Recap of Correlation
- Predictive Modeling: Simple Linear Regression
- Simple Linear Regression: Residual Analysis
- Simple Linear Regression: Coefficient of Determination
- Simple Linear Regression: Prediction Interval
- Multiple Linear Regression
- Multiple Linear Regression: Evaluating Multiple Regression Models
- Multiple Linear Regression: Indicator (Dummy) Variables
- Multiple Linear Regression: Variable Importance
- Simple Vs. Multiple Regression: Correlation Effects


ANY QUESTIONS
?