

Assessing the Fit of the Simple Linear Regression Model

The Sums of Squares

The Coefficient of Determination

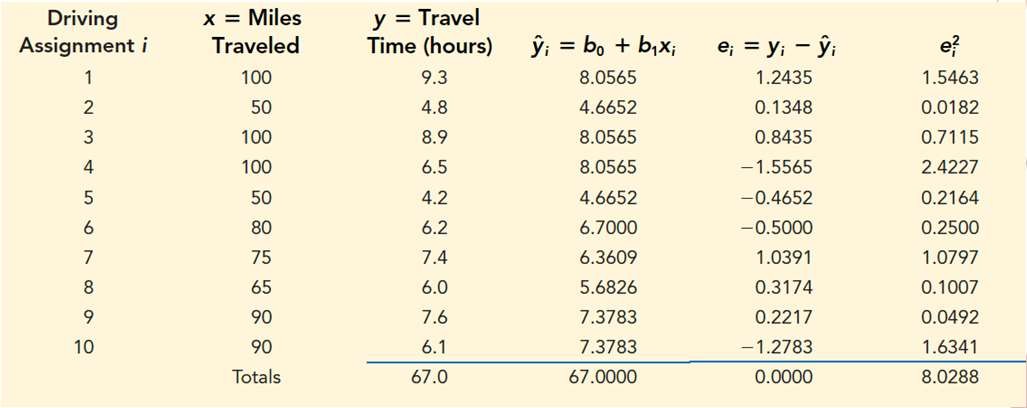
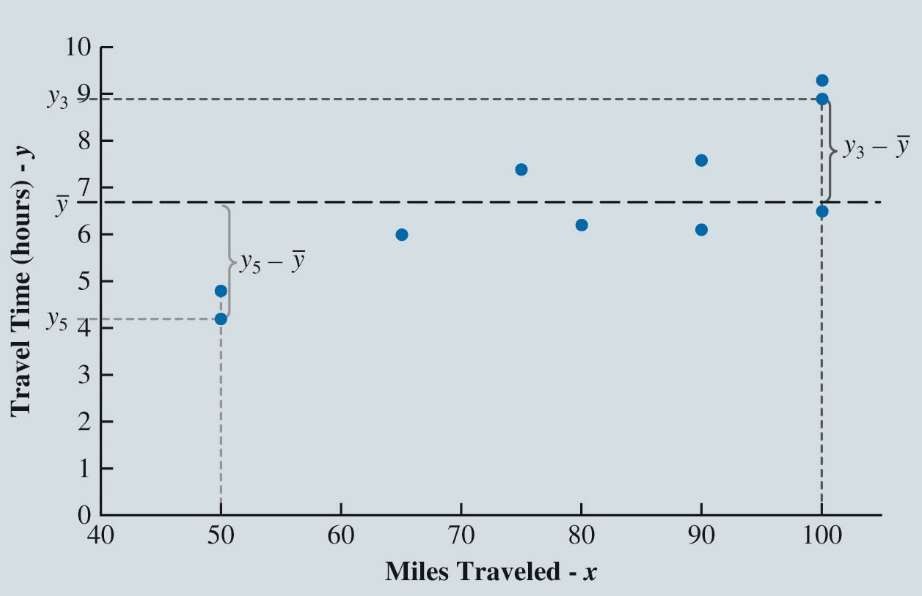
Using Excel’s Chart Tools to Compute the Coefficient of Determination



Assessing the Fit of the Simple Linear Regression Model

**The Sums of Squares:**

* Sum of squares due to error (SSE):
  + is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample.



Assessing the Fit of the Simple Linear Regression Model

*n*

SSE = *e* = 8.0288

 *i*

2

* A value closer to 0 indicates a better fit!

*i* =1



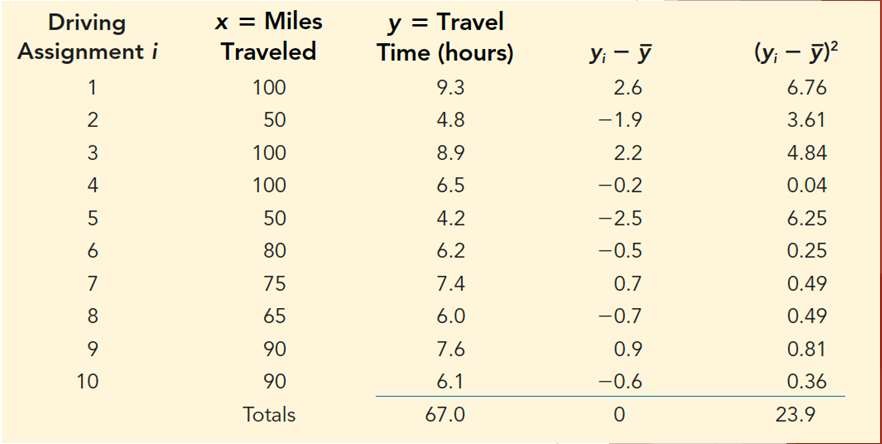
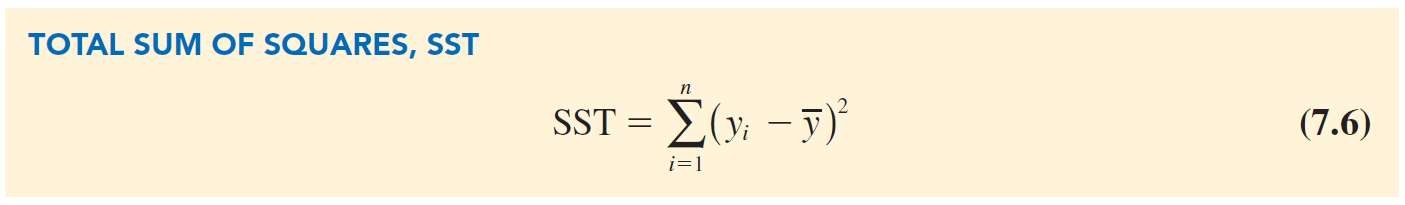
Assessing the Fit of the Simple Linear Regression Model

**What if we wanted to predict the travel time without knowing the miles traveled?**

* Use the sample Mean, 𝑦
  + 𝑦 = 6.7 - Sample mean
    - Over estimates 6 data points
    - Under estimates 4 data points

**How far off is the sample mean?**

* **Difference:** 𝒚𝒊 − 𝒚¯
  + Measure of error involved using 𝑦 to predict travel time.



Assessing the Fit of the Simple Linear Regression Model

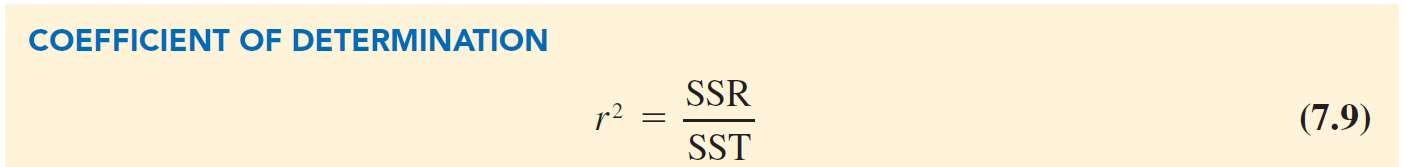
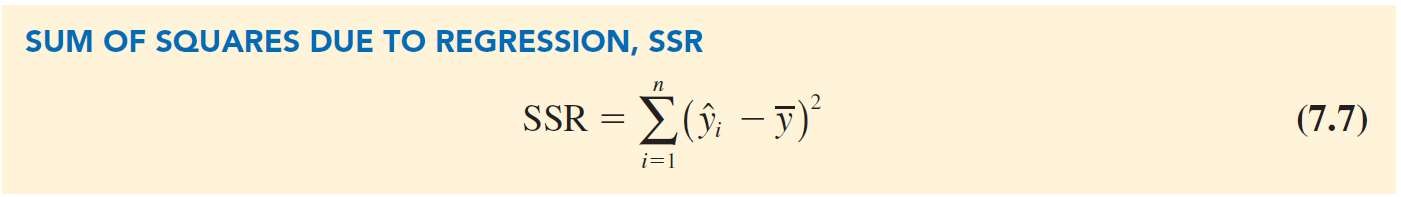
* The corresponding sum of squares is called the total sum of squares (SST).



Assessing the Fit of the Simple Linear Regression Model

Sum of Squares Total for the Butler Trucking Simple Linear Regression

SST = 23.9 ℎ𝑜𝑢𝑟𝑠2



Assessing the Fit of the Simple Linear Regression Model

* Measures how much the 𝑦^ values on the estimated regression line deviate from 𝑦¯.

SST = SSR + SSE

where

* + SST = total sum of squares
  + SSR = sum of squares due to regression
  + SSE = sum of squares due to error.



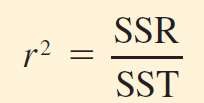
Assessing the Fit of the Simple Linear Regression Model

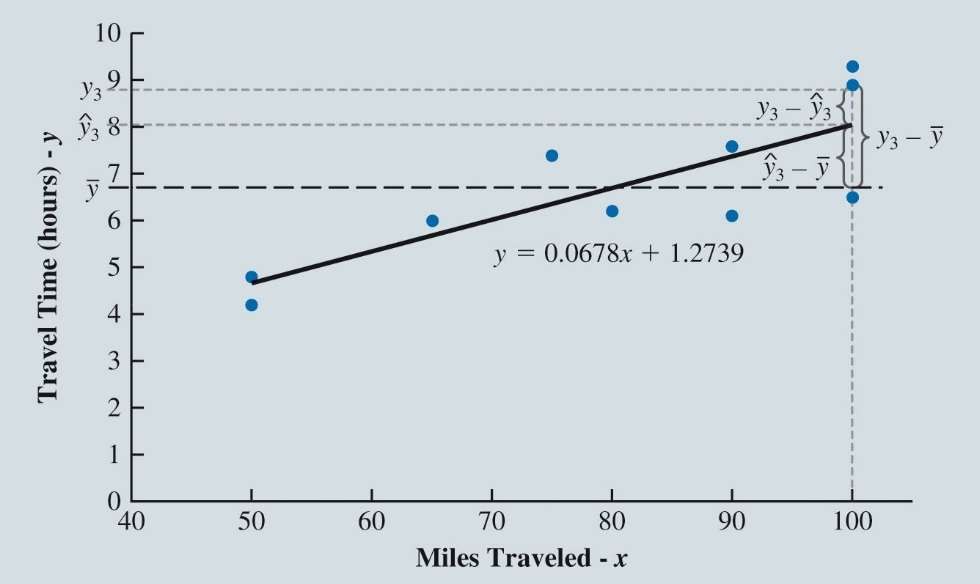
**The Coefficient of Determination:**

* The ratio SSR/SST used to evaluate the goodness of fit for the estimated regression equation;
  + Denoted by

𝑟2 or R2

* + - Take values between zero and one.
    - Interpreted as the percentage of the total sum of squares that can be explained by using the estimated regression equation.





Assessing the Fit of the Simple Linear Regression Model

* **SST**
  + How well the observations (y’s) cluster around 𝑦¯ = 6.7
* **SSR**
  + How well the estimates (𝑦^i ’s) cluster around 𝑦¯.



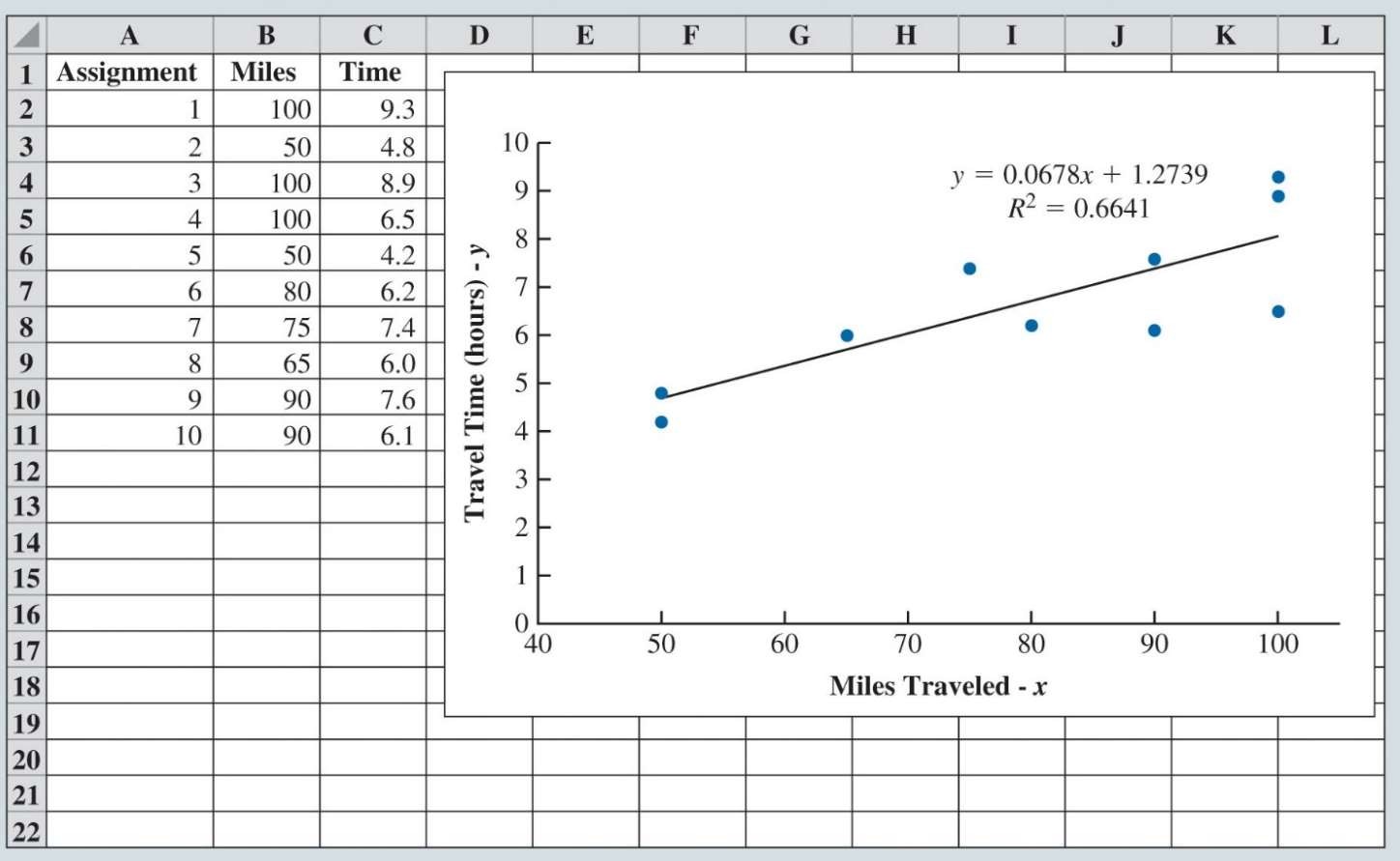
Assessing the Fit of the Simple Linear Regression Model

* For Butler Trucking Company, the value of the coefficient of determination:

𝑟2 = 𝑆𝑆𝑅 = 15.8712 = 0.6641

𝑆𝑆𝑇 23.9

* As a percentage (66.41%)
  + The percentage of the total sum of squares that can be explained by using the estimated regression equation
* In other words:
  + 66.41% of the variability in the values of travel time can be explained by the linear relationship between miles traveled and travel time.



Assessing the Fit of the Simple Linear Regression Model

**Using Excel’s Chart Tools to Compute the Coefficient of Determination:**

* To compute the coefficient of determination :

1. Right-click on any data point in the scatter chart and select **Add Trendline…**

2. When the **Format Trendline** task pane appears:

* + Select **Display R-squared value on chart** in the **Trendline Options** area.



**Only 66.41% of the variation in travel time can be explained by miles traveled?**

**What about the other 33%?**

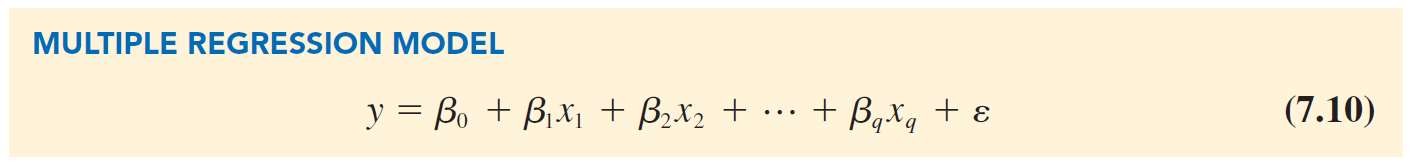
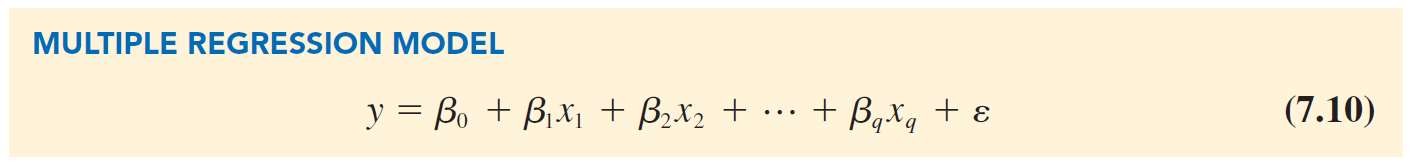
The Multiple Regression Model

Regression Model

Estimated Multiple Regression Equation

Least Squares Method and Multiple Regression Butler Trucking Company and Multiple Regression

Using Excel’s Regression Tool to Develop the Estimated Multiple Regression Equation



The Multiple Regression Model

Multiple Regression Model: How dependent variable y is related to 2 or more independent variables

*y* = dependent variable.

*x*1 , *x*2 , . . ., *xq* = independent variables.

** , ** , ** ,, ** = parameters.

** = error term (accounts for the variability in *y* that cannot be explained by the linear effect of the *q* independent variables).

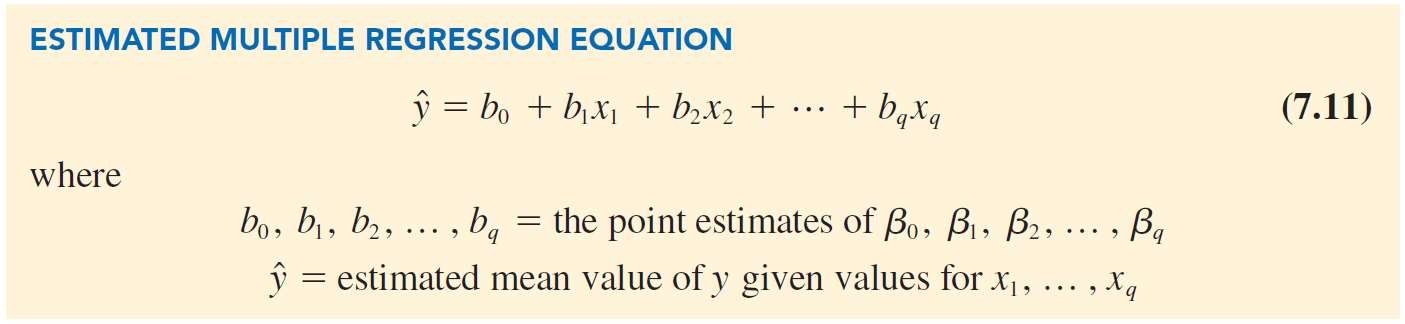
0 1 2 *q*



The Multiple Regression Model

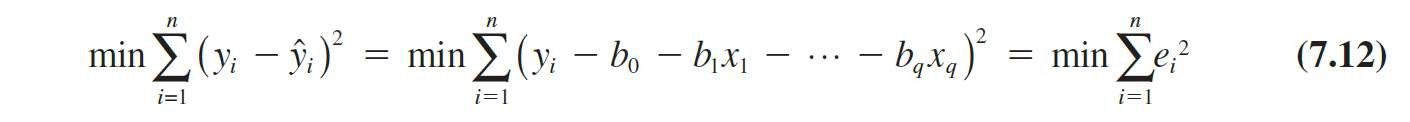
* Regression Model (cont.):
* Interpretation of parameter, 𝛽j:
  + Represents the change in the mean value of the dependent variable *y* that corresponds to a one unit increase in the independent variable
  + In other words, all else constant: as variable 𝑥j increases by one unit, y increases or decreases by 𝛽j

*E* (*y x*2 , *x*2 , . . . , *xq* ) = **0 + **1*x*1 + **2*x*2 ++ *q xq*



The Multiple Regression Model

Estimated Multiple Regression Equation:



The Multiple Regression Model

**Least Squares Method and Multiple Regression:**

* The least squares method is used to develop the estimated multiple regression equation:

**Finding:**

n

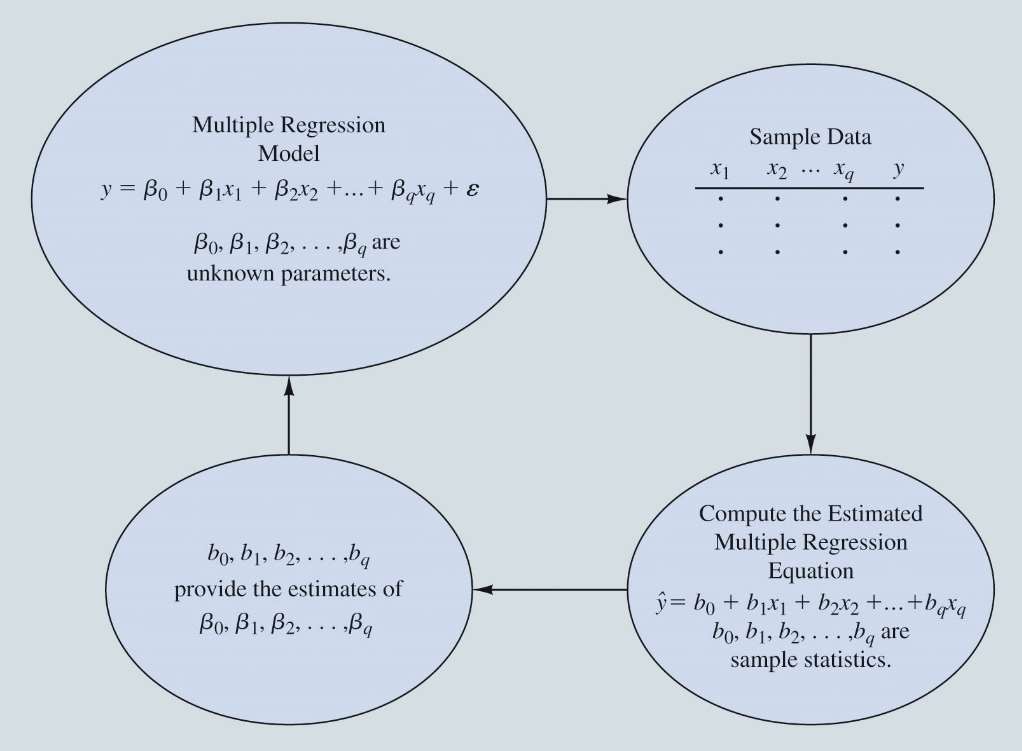
n

𝑏0, 𝑏1, 𝑏2, … , 𝑏q that satisfy min Σ 𝑦i − 𝑦^i 2 = min Σ 𝑒2 .

i

i=1 i=1

* Use sample data to get values of 𝑏0, 𝑏1, 𝑏2, … , 𝑏q
  + That minimize the sum of squared residuals



The Multiple Regression Model

Butler Trucking Company and Multiple The Estimation Process for Multiple Regression

Regression:

* The estimated simple linear regression equation,

*y*ˆ*i* = 1.2739 + 0.0678*xi* .

* The linear effect of the number of miles traveled explains 66.41%
* This implies, 33.59% of the variability in sample travel times remains unexplained
* Other variables?
  + Number of packages/deliveries
  + Weather, traffic, city, rural



The Multiple Regression Model

Butler Trucking Company and Multiple Regression (cont.):

Estimated multiple linear regression with two independent variables:

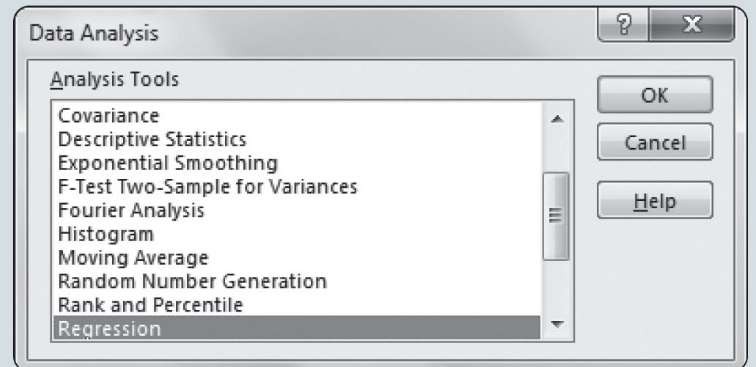
*y*ˆ = *b*0 + *b*1 *x*1 + *b*2 *x*2

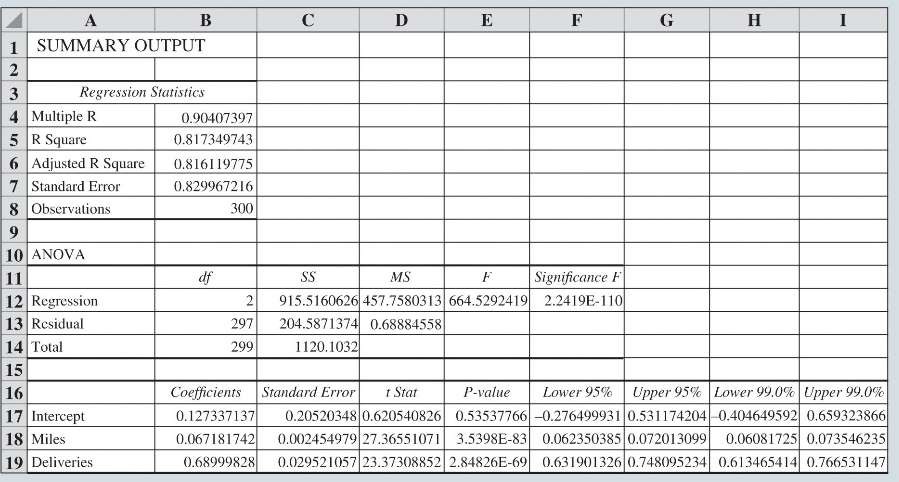
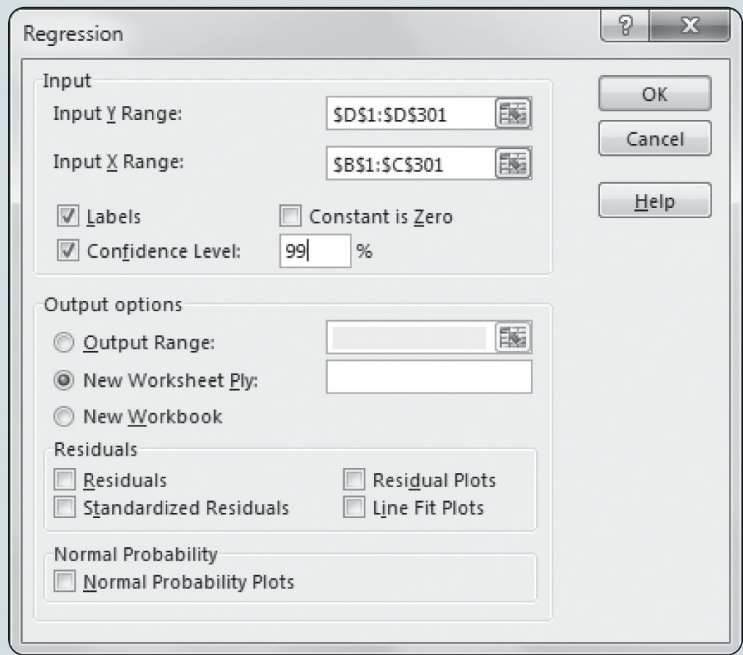
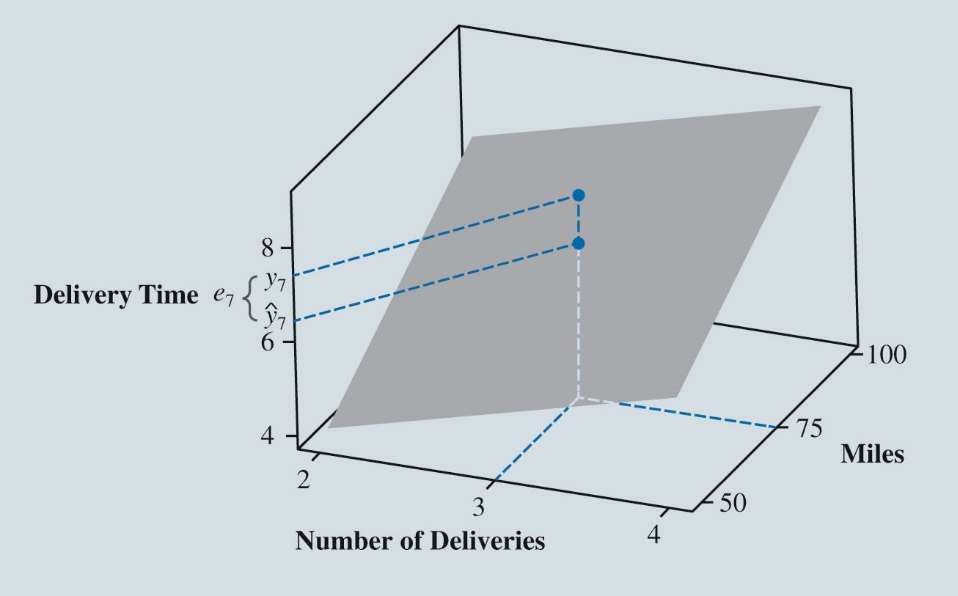
*y*ˆ = Estimated mean travel time.

*x*1 = Distance traveled.

*x*2 = Number of deliveries.

The SST, SSR, SSE and *R*2 are computed using the formulas discussed earlier.





The Multiple Regression Model

Graph of the Regression Equation for Multiple Regression Analysis with Two Independent Variables

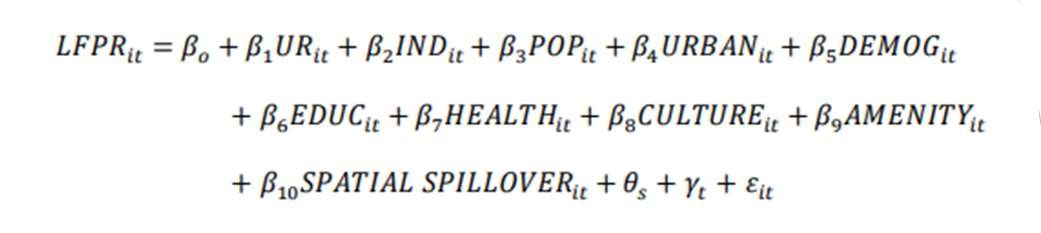
* Instead of a regression line
  + We created a regression plane
* **Notice:**
  + 1. Plane is sloped upward as # of deliveries AND miles increase
* Recall Data:
  + Driver 7: Miles = 75 and Deliveries = 3, Time = 7.4 hours
  + Estimation: Miles = 75 and Deliveries = 3, Time = 7.2 hours
    - Close but a little low



The Multiple Regression Model

**Data Analysis Tools Box**

Using Excel’s Regression Tool to Develop the Estimated Multiple Regression Equation:



Example

* Y = LFPR = County labor force participation rate
* X1 = UR = Unemployment rate
* X2 = IND = Industry composition
* X3 = POP = Population
* X4 = URBAN = Urban or Rural status
* X5 = DEMOG = Demographics
* X6 = EDUC = Education levels













X7 = HEALTH = Life Expectancy

X8 = CULTURE = Social Capital Index

X9 = AMENITY = USDA natural amenity scale

X10 = SPATIAL SPILLOVER = Nearest Neighbor weights

𝜃s = State Fixed Effect (Dummy Variable)

𝛾t = Year Fixed Effect (Dummy Variable)

How do you know which variables to include/use?



Example

* What “variables” affect your health?

Health = 𝛽0 + 𝛽0𝑎𝑔𝑒 + 𝛽0𝑤𝑒𝑖𝑔ℎ𝑡 + 𝛽0ℎ𝑒𝑎𝑟𝑡rate + 𝛽0𝑔𝑒𝑛𝑑𝑒𝑟 + 𝜀

Health = 87.83 − .165𝑎𝑔𝑒 − .385𝑤𝑒𝑖𝑔ℎ𝑡 − .118ℎ𝑒𝑎𝑟𝑡rate + 13.208𝑔𝑒𝑛𝑑𝑒𝑟

* Y intercept (Constant) – Average Health status = 87.83
* Age – As age increase by 1 year, on average, a person’s health status decreases by .165
* Weight – As weight increases by 1 pound, on average, a person’s health status decreases by .385
* Heart rate – As heart rate increases by 1 beat per minute, on average, a person’s health status decreases by .118
* Gender – Average difference between female and male health = 13.208
  + Females have 13.208 point higher health status than males.

𝑅2 = .577 = 58%