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.

SUM OF SQUARES DUE TO ERROR

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \textbf{(7.5)}$$

Assessing the Fit of the Simple Linear Regression Model

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$\hat{y}_i = b_0 + b_1x_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	100	9.3	8.0565	1.2435	1.5463
2	50	4.8	4.6652	0.1348	0.0182
3	100	8.9	8.0565	0.8435	0.7115
4	100	6.5	8.0565	−1.5565	2.4227
5	50	4.2	4.6652	−0.4652	0.2164
6	80	6.2	6.7000	−0.5000	0.2500
7	75	7.4	6.3609	1.0391	1.0797
8	65	6.0	5.6826	0.3174	0.1007
9	90	7.6	7.3783	0.2217	0.0492
10	90	6.1	7.3783	−1.2783	1.6341
	Totals	67.0	67.0000	0.0000	8.0288

$$SSE = \sum_{i=1}^{n} e_i^2 = 8.0288$$

- A value closer to 0 indicates a better fit!

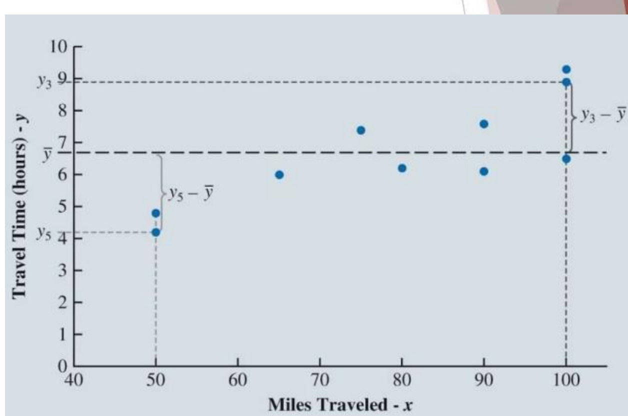
Assessing the Fit of the Simple Linear Regression Model

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:** $y_i - \bar{y}$
  - 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).

TOTAL SUM OF SQUARES, SST

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \textbf{(7.6)}$$

Assessing the Fit of the Simple Linear Regression Model

Sum of Squares Total for the Butler Trucking Simple Linear Regression

SST = 23.9 $hours^2$

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	100	9.3	2.6	6.76
2	50	4.8	−1.9	3.61
3	100	8.9	2.2	4.84
4	100	6.5	−0.2	0.04
5	50	4.2	−2.5	6.25
6	80	6.2	−0.5	0.25
7	75	7.4	0.7	0.49
8	65	6.0	−0.7	0.49
9	90	7.6	0.9	0.81
10	90	6.1	−0.6	0.36
	Totals	67.0	0	23.9

Assessing the Fit of the Simple Linear Regression Model

SUM OF SQUARES DUE TO REGRESSION, SSR

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \textbf{(7.7)}$$

- Measures how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$.

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

$$r^2 \text{ or } R^2$$

  - 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.

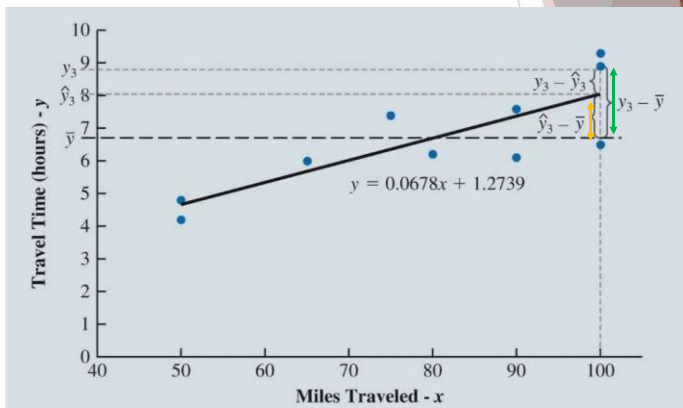
COEFFICIENT OF DETERMINATION

$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \qquad \textbf{(7.9)}$$

Assessing the Fit of the Simple Linear Regression Model

- SST
  - How well the observations (y's) cluster around $\bar{y} = 6.7$
- SSR
  - How well the estimates ($\hat{y}_i$'s) cluster around $\bar{y}$.

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$



Assessing the Fit of the Simple Linear Regression Model

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

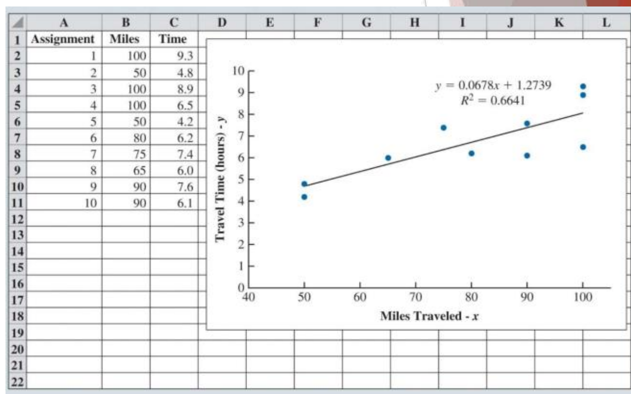
$$r^2 = \frac{SSR}{SST} = \frac{15.8712}{23.9} = 0.6641$$

- 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.



Assignment	Miles	Time
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

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

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon \qquad (7.10)$$

y = dependent variable.

$x_1, x_2, \ldots, x_q$ = independent variables.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ = parameters.

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

The Multiple Regression Model

- Regression Model (cont.):

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon \qquad (7.10)$$

- Interpretation of parameter, β_j:
  - 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 x_j increases by one unit, y increases or decreases by β_j

$$E\left(y \middle| x_2, x_2, \ldots, x_q\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

The Multiple Regression Model

Estimated Multiple Regression Equation:

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_qx_q \qquad \textbf{(7.11)}$$

where

$b_0, b_1, b_2, \ldots, b_q$ = the point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$

$\hat{y}$ = estimated mean value of y given values for $x_1, \ldots, x_q$

The Multiple Regression Model

Least Squares Method and Multiple Regression:

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

Finding:

$b_0, b_1, b_2, \ldots, b_q$ that satisfy $\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} e_i^2$.

- Use sample data to get values of $b_0, b_1, b_2, \ldots, b_q$
- That minimize the sum of squared residuals

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - b_0 - b_1x_1 - \cdots - b_qx_q)^2 = \min \sum_{i=1}^{n} e_i^2 \qquad \textbf{(7.12)}$$

The Multiple Regression Model

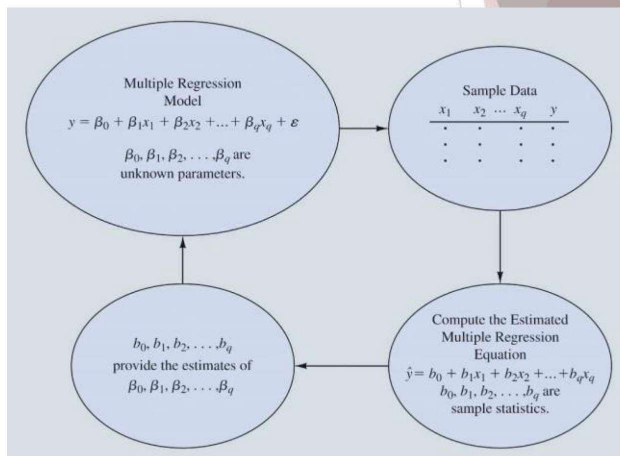
Butler Trucking Company and Multiple Regression:

- The estimated simple linear regression equation,

$$\hat{y}_i = 1.2739 + 0.0678x_i.$$

- 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 Estimation Process for Multiple Regression



The Multiple Regression Model

Butler Trucking Company and Multiple Regression (cont.):

Estimated multiple linear regression with two independent variables:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

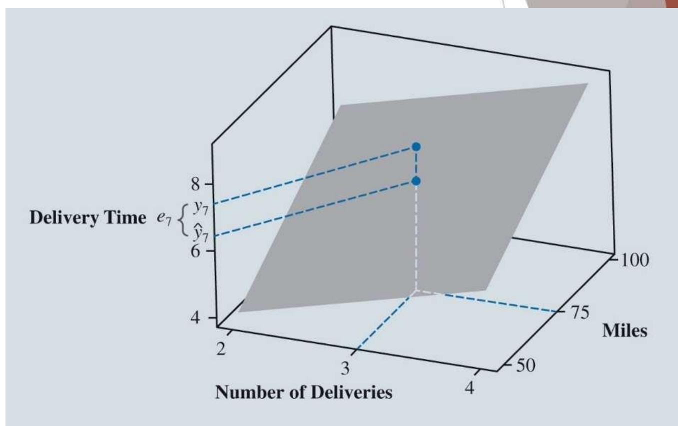
$\hat{y}$ = Estimated mean travel time.
x_1 = Distance traveled.
x_2 = Number of deliveries.

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

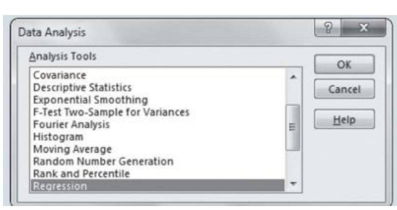
- Instead of a regression line
  - We created a regression plane
- **Notice:**
  - 1. Plane is sloped upward as # of deliveries AND miles increase
- Recall Data:
  - Driver 7: Miles = 75 and Deliveries = 3, Time = 7.4 hours
  - Estimation: Miles = 75 and Deliveries = 3, Time = 7.2 hours
    - Close but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

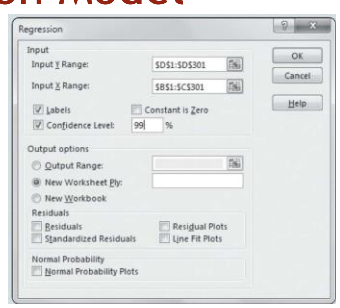


The Multiple Regression Model

Data Analysis Tools Box

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





	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.90407397							
5	R Square	0.817349743							
6	Adjusted R Square	0.816119775							
7	Standard Error	0.829967216							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	915.5160626	457.7580313	664.5292419	2.2419E-110			
13	Residual	297	204.5871374	0.68884558					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	0.127337137	0.20520348	0.620540826	0.53537766	−0.276499931	0.531174204	-0.404649592	0.659323866
18	Miles	0.067181742	0.002454979	27.36551071	3.5398E-83	0.062350385	0.072013099	0.06081725	0.073546235
19	Deliveries	0.68999828	0.029521057	23.37308852	2.84826E-69	0.631901326	0.748095234	0.613465414	0.766531147

Example

$$LFPR_{it} = \beta_o + \beta_1 UR_{it} + \beta_2 IND_{it} + \beta_3 POP_{it} + \beta_4 URBAN_{it} + \beta_5 DEMOG_{it}$$
$$+ \beta_6 EDUC_{it} + \beta_7 HEALTH_{it} + \beta_8 CULTURE_{it} + \beta_9 AMENITY_{it}$$
$$+ \beta_{10} SPATIAL\ SPILLOVER_{it} + \theta_s + \gamma_t + \varepsilon_{it}$$

- 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
- θ_s = State Fixed Effect (Dummy Variable)
- γ_t = Year Fixed Effect (Dummy Variable)

How do you know which variables to include/use?

Example

- What "variables" affect your health?

Health = $\beta_0 + \beta_0 age + \beta_0 weight + \beta_0 heart_{\text{rate}} + \beta_0 gender + \varepsilon$

Health = $87.83 - .165 age\ -\ .385 weight\ -\ .118 heart_{\text{rate}} + 13.208 gender$

- 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.

$R^2 = .577 = 58\%$