SPRING 2014 : <u>STAT - 1</u> , Section 10.

CHAPTER 13: DESCRIPTIVE METHODS IN REGRESSION AND CORRELATION

- Recall that a linear equation can be written as $y = b_0 + b_1 x$ and recall the definition of *slope* b_1 and the y $-$ *intercept* b_0. TRY 13.5 a, b, 13.7 a, b.

  In Section 13.2 carefully read the Table 13.2 which gives a bivariate data set on the *Age* and *Price* for a sample of 11 *Orion* cars. Follow this example in the text as we define *scatterplot* (see p.603). The line that best fits the data set ("closest" to the scatterplot) is called the *Regression Line* and its equation represented by $\hat{y} = b_0 + b_1 x$ is called the *Regression Equation*. We use ŷ as the *predicted value* for an input value of x. Accordingly, for the data point (x_i, y_i) , y_i is the *observed value* of y . We call x the *independent variable* or the *predictor variable* and y the *dependent variable* or the *response variable*. Study THE DEFINITION 13.3 AND FORMULA 13.3 (p.605). As in Chapter 3, we will use the COMPUTING FORMULA TO DETERMINE S_{xx}, S_{xy}, S_{yy} for the bivariate data sets we encounter. If you do not pay attention to this instruction and use the Defining Formula for computations you are on your own. Follow the computations for *Orion data* in Example 13.4, of S_{xx}, S_{xy}, b_1, b_0, the regression equation, its interpretation, interpretatin of the slope, prediction using the regression equation, pitfalls of *extrapolation*, outliers and influential obsrvations.

  After understanding the *Orion* example, TRY 13.51, 13.53, 13.55. Also TRY 13.35, 13.37, 13.39, 13.57 and 13.59.

- In 13.3, we learn Definition 13.5 (p.617), SST, SSR, SSE and the *Coefficient of Determination* r^2 $(= \frac{SSR}{SST})$. The Regression idenitity states $SST = SSR + SSE$. Again USE THE COMPUTING FORMULAS (p.620) for SST, SSR, SSE as used in Example 13.8 (p.621). DO NOT USE the defining formula computations in Example 13.7 (p. 618$-$620);

they will lead to the correct answers but by lengthy and tedious calcultions.

TRY 13.79, 13.80, 13.81, 13.89, 13.91, 13.93, 13.106, 13.107.

- We introduce the *Linear Correlation Coefficient* $- r$ (what is another name for it based on its developer?) and its computing formula. Study its properties (p. 624−625). Example 13.10 follows through with the *Orion* data. Read pages 627−628.

  TRY 13.108, 13.109, 13.110, 13.111, 13.112, 13.113, 13.114, 13.123, 13.125, 13.127, 13.145.

- Read about Legendre, Method of Least Squares and Gauss in the biography section.

- For the *Orion* data we have
  $n = 11, \ \sum x_i \ = \ 58 \ , \sum x_i^2 \ = \ 326 \ \ \sum y_i \ = \ 975 \ ,$
  $\sum y_i^2 \ = \ 96,129 \ \ \sum x_i y_i \ = \ 4,732$ (see p.637 and p.653).

- For Exercise 13.53, we have
  $n = 11, \ \sum x_i \ = \ 723 \ , \sum x_i^2 \ = \ 48,747 \ \ \sum y_i \ = \ 156.5 \ ,$
  $\sum y_i^2 \ = \ 2,523.251 \ \ \sum x_i y_i \ = \ 10,486.0$.

- For Exercise 13.51 we have
  $n = 10, \ \sum x_i \ = \ 41 \ , \sum x_i^2 \ = \ 199 \ \ \sum y_i \ = \ 3422 \ ,$
  $\sum y_i^2 \ = \ 1,196,690 \ \ \sum x_i y_i \ = \ 13,168.$

  For each of the above three examples calculate:

  S_{xx} , S_{xy} , S_{yy} , b_1 , b_0 , ŷ , SST , SSR , SSE , r^2 , r , .

  Then interpret ŷ , r^2 , r , and the coefficient of determination.

Recall that $S_{xx} = \left[(\sum x_i^2) - \frac{(\sum x_i)^2}{n} \right]$

$$S_{xy} = \left[(\sum x_i y_i) - \frac{(\sum x_i)(\sum y_i)}{n} \right]$$

$$S_{yy} = \left[(\sum y_i^2) - \frac{(\sum y_i)^2}{n} \right]$$

$$b_1 = \frac{S_{xy}}{S_{xx}} \ , \ b_0 = \frac{1}{n} \left[(\sum y_i) - b_1 (\sum x_i) \right]$$

$\hat{y} = b_0 + b_1 x$ is the regression line

$$SST = S_{yy} \ , SSR = \frac{S_{xy}^2}{S_{xx}} \ , \ SSE = SST - SSR$$

$$r^2 = \text{Coefficient of Determination} = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Go over the chapter for the interpretation of $\hat{y}$, b_1 , b_0 , r^2 , and r.

For the *Orion* data we get

$$S_{xx} = \left[(326) - \frac{(58)^2}{11} \right] = 20.1818$$

$$S_{xy} = \left[(4,732) - \frac{(58)(975)}{11} \right] = -408.9091$$

$$S_{yy} = \left[(96,129) - \frac{(975)^2}{11} \right] = 9,708.5455$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{-408.9091}{20.1818} = -20.2613$$

$$b_0 = \frac{1}{n} \left[(\sum y_i) - b_1 (\sum x_i) \right] = \left(\frac{1}{11}\right) [975 - (-20.2613)(58)] = 195.4687$$

$$\hat{y} = 195.47 - 20.2613\, x \text{ (in } 100s).$$

(See p.630 above Formula 13.2)

SST = 9708.5455 , SSR = 8285.0218 , SSE = SST − SSR = 1423.5237.

r^2 = Coefficient of Determination $= \frac{(-408.9091)^2}{(20.1818)(9708.5455)} =$ 0.8534

$$r = \frac{(-408.9091)}{\sqrt{(20.1818)(9{,}708.5455)}} = -0.9238$$

Verify that for 13.51

$$S_{xx} = \left[(199) - \frac{(41)^2}{10}\right] = 30.9$$

$$S_{xy} = \left[(13,168) - \frac{(41)(3{,}422)}{10}\right] = -862.20$$

$$S_{yy} = \left[(1,196,690) - \frac{(3422)^2}{10}\right] = 25,681.60$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = -27.9029$$

$$b_0 = \frac{1}{n}\left[(\sum y_i) - b_1(\sum x_i)\right] = (\tfrac{1}{10})[3,422 - (-27.9029)(41)] = 456.6019$$

$\hat{y} = 456.6019 - 27.9029\, x$, $SST = 25,681.60$, $SSR = 24,057.8913$, $SSE = 1,623.7087 = r^2 = 0.9368 = r = -0.9679.$

Verify that for 13.53,

$$S_{xx} = \left[(148,747) - \frac{(723)^2}{11}\right] = 1226.1818$$

$$S_{xy} = \left[(10,486.0) - \frac{(723)(156.5)}{11}\right] = 199.6818$$

$S_{yy} \ = \ \left[(2,523.251) - \frac{(156.5)^2}{11}\ \right] \ = \ 296.6828$

$b_1 \ = \ \frac{S_{xy}}{S_{xx}} = 0.1628$

$b_0 \ = \ \frac{1}{n}\left[(\sum y_i) - b_1(\sum x_i)\right] = 3.5269$

ŷ $= \ 3.5269 + 0.1628\ x$, $SST = 296.6828$, $SSR \ = \ 32.5179$, $SSE \ = \ 264.1649 \ = \ r^2 \ = \ 0.1096 \ = \ r \ = \ 0.331.$