Exercises

3.3.1 The data below contain the weights (in lbs) and highway miles per gallon (MPG) of several cars. Calculate SS_{Reg}, SS_{Res}, SS_{Tot}, and R^2 for the linear regression equation fit to this data.

Weight (x)	3250	3425	2400	2250
MPG (y)	26	28	37	38

3.3.2 The data below contain the amounts of nitrogen applied to different corn fields (in lbs/acre) and the resulting yields (in bushels/acre). Calculate SS_{Reg}, SS_{Res}, SS_{Tot}, and R^2 for the linear regression equation fit to this data.

Nitrogen (x)	0	60	120	180	240
Yield (y)	78	90	140	162	210

3.3.3 The data below contain the diameter of the trunk at chest height and volume of wood in several pine trees. Use the trendline function in Excel to model Volume in terms of Diameter with several different linearizable models and select the "best" one.

Diameter	32	29	24	45	20	30	26	40	24	18
Volume	185	109	95	300	30	125	55	246	60	15

3.3.4 The data below contain the number of manatee deaths in Florida believed to be caused by watercraft for the years 1983 – 2005 where Year 1 corresponds to 1983 (data from the Florida Fish and Wildlife Conservation Commission, Marine Mammal Pathobiology Laboratory). Use different linearizable models to model Deaths in terms of Year and select the "best" one. Is comparing R^2 values sufficient when selecting a "best" model? Explain.

Year	1	2	3	4	5	6	7	8	9	10	11	12
Deaths	15	24	44	44	39	43	50	47	53	38	35	49
Year	13	14	15	16	17	18	19	20	21	22	23	24
Deaths	42	60	54	67	82	78	81	95	73	69	79	92

3.3.5 Another measure of how well a linear regression line fits a set of data is the **standard error of estimate**, denote by s_e. It is defined by

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

Equivalently, we could define it by

$$s_e = \sqrt{\frac{SS_{Res}}{n-2}}$$

The smaller s_e is, the better the fit. This quantity is also an estimate of the standard deviation of the "noise" ε in the relationship $y = \beta_0 + \beta_1 x + \varepsilon$.

Modify the worksheet **R2** to calculate s_e for the linear model fit to the data with some noise. How close is s_e to the true standard deviation of the "noise?" Try different values of the standard deviation of the "noise."

3.3.6 Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the least–squares regression equation fit to a set of data $\{(x_i,\, y_i) \,:\, i = 1, \ldots, n\}$ and let $\hat{y}_i$ be an estimate of y_i based off this equation.

a. Modify the worksheet **R2** to illustrate these two properties:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0 \;\text{ and }\; \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$$

b. Also modify the worksheet **R2** to illustrate this property:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

In other words, show that

$$SS_{Tot} = SS_{Reg} + SS_{Res}$$

c. Illustrate that none of these three properties hold if we use a slope other than $\hat{\beta}_1$ or a y–intercept other than $\hat{\beta}_0$ to calculate $\hat{y}_i$. In other words, show that these properties do not hold if we use a slope other than that given by the Excel formula **SLOPE** or a y–intercept other that that given by **INTERCEPT** to calculate the predicted value of y.

d. Use the properties in part 1 to prove the property in part 2. **Hint**: Start with

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n} ((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i))^2$$

expand the right–hand side and rewrite so that you get the terms $\sum\limits_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, $\sum\limits_{i=1}^{n} (y_i - \hat{y})^2$, $\sum\limits_{i=1}^{n} (y_i - \hat{y}_i)$, and $\sum\limits_{i=1}^{n} x_i\,(y_i - \hat{y}_i)$.

3.3.7 Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the least–squares regression equation fit to a set of data $\{(x_i,\, y_i) \,:\, i = 1, \ldots, n\}$. Modify the worksheet **R2** to illustrate this property:

$$R^2 = \hat{\beta}_1^2 \cdot \frac{n \sum x_i^2 - \left(\sum x_i\right)^2}{n \sum y_i^2 - \left(\sum y_i\right)^2}.$$