# *Linear Regression Examples*

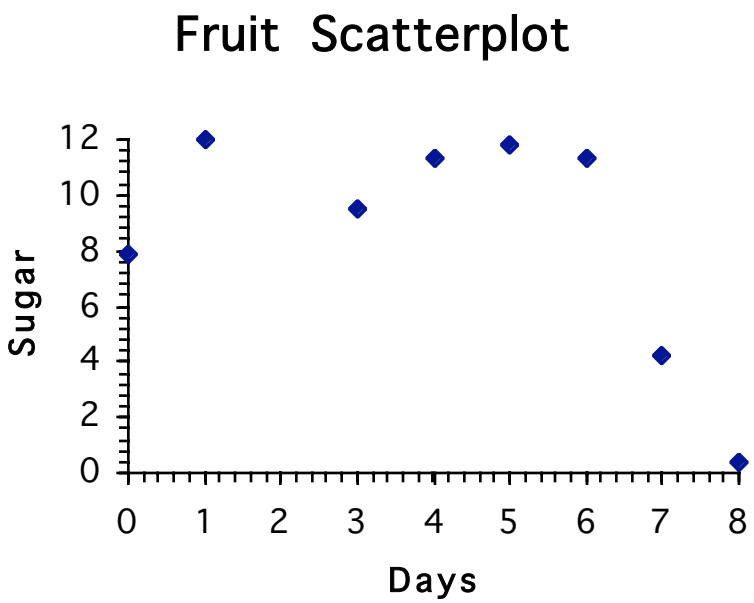
**STAT 314**

1. The data below show the sugar content of a fruit (SUGAR) for different numbers of days after picking (DAYS).

| Days | Sugar |
|---|---|
| 0 | 7.9 |
| 1 | 12.0 |
| 3 | 9.5 |
| 4 | 11.3 |
| 5 | 11.8 |
| 6 | 11.3 |
| 7 | 4.2 |
| 8 | 0.4 |

a. Obtain the estimated regression line to predict sugar content based on the number of days the fruit is left on the tree. Also create the regression ANOVA table.

**Step 1: Scatterplot**




Since we see a slightly linear pattern, linear regression may be appropriate (Assumption 1 is met, but barely).

**Step 2: Compute the Sums of Squares**

Let *x* be DAYS and *y* be SUGAR.

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 200 - \frac{(34)^2}{8} = 200 - 144.5 = 55.5$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 245.1 - \frac{(34)(68.4)}{8} = 245.1 - 290.7 = -45.6$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 709.08 - \frac{(68.4)^2}{8} = 709.08 - 584.82 = 124.26 = SSTo$$

**Step 3: Compute the Least-Squares Linear Regression Equation**

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-45.6}{55.5} = -0.8216$$

$$a = \bar{y} - b\bar{x} = \frac{68.4}{8} - (-0.8216)\left(\frac{34}{8}\right) = 12.0418 \qquad SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = 37.46595$$

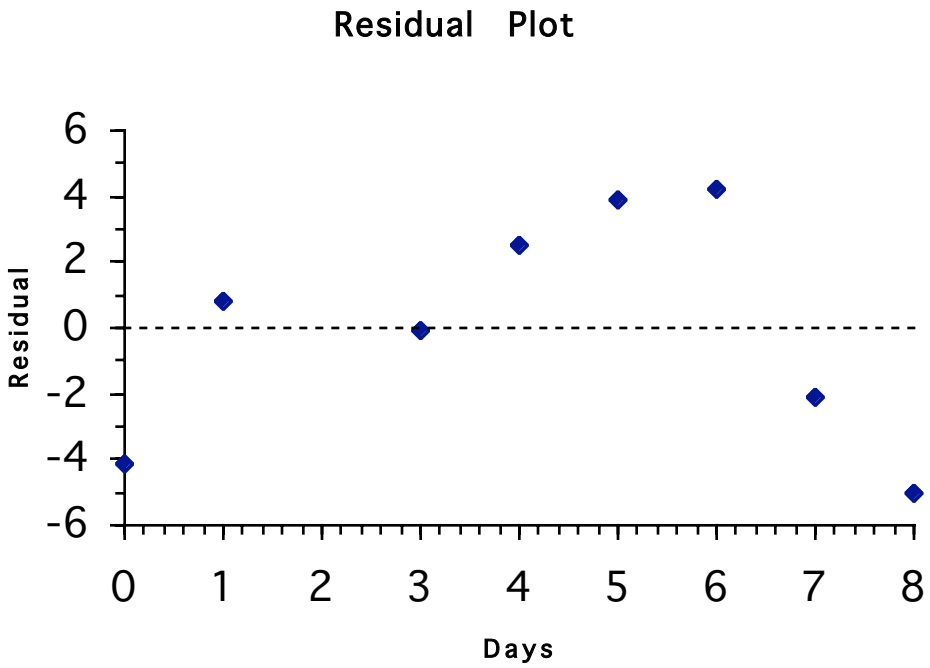
$$\hat{y} = a + bx = 12.0418 + (-0.8219)x = 12.0418 - 0.8219x$$

***ANOVA Table***

| Source | *df* | *SS* | *MS* | *F*-statistic | *p*-value |
|---|---|---|---|---|---|
| Regression | 1 | 37.46595 | 37.46595 | 2.590 | $p > 0.10$ |
| Error | 6 | 86.79405 | 14.46568 | | |
| Total | 7 | 124.26000 | | | |

b. Calculate and plot the residuals against DAYS. Do the residuals suggest a fault in the model?

| Days | Sugar | Predicted $(\hat{y} = 12.0418 - 0.8219x)$ | Residual $(y - \hat{y})$ |
|---|---|---|---|
| 0 | 7.9 | 12.0418 | -4.1418 |
| 1 | 12.0 | 11.2199 | 0.7801 |
| 3 | 9.5 | 9.5761 | -0.0761 |
| 4 | 11.3 | 8.7542 | 2.5458 |
| 5 | 11.8 | 7.9323 | 3.8677 |
| 6 | 11.3 | 7.1104 | 4.1896 |
| 7 | 4.2 | 6.2885 | -2.0885 |
| 8 | 0.4 | 5.4666 | -5.0666 |



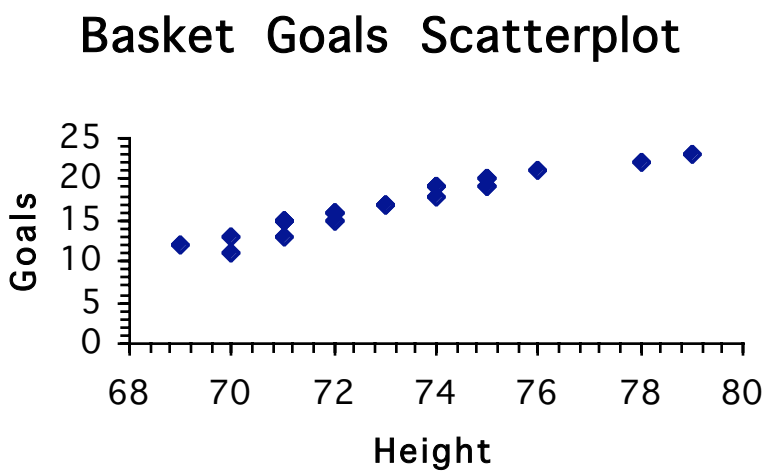

The residuals seem to be randomly scattered with an even variability (a slight increase in variance—Assumption 3 is not met). Therefore, the residual plot seems to indicate that the relationship may be nonlinear (fault in model). This fault is illustrated in the ANOVA test for the slope which indicates that DAYS is not useful as a predictor of SUGAR (fail to reject $H_0$—*p*-value > 0.10).

2. It is generally believed that taller persons make better basketball players because they are better able to put the ball into the basket. The table below lists the heights of a sample of 25 non-basketball athletes and the number of successful baskets made in a 60-second time period.

| Obs. | Height | Goals | Obs. | Height | Goals | Obs. | Height | Goals |
|---|---|---|---|---|---|---|---|---|
| 1 | 71 | 15 | 10 | 74 | 18 | 19 | 78 | 22 |
| 2 | 74 | 19 | 11 | 71 | 13 | 20 | 79 | 23 |
| 3 | 70 | 11 | 12 | 72 | 15 | 21 | 72 | 16 |
| 4 | 71 | 15 | 13 | 73 | 17 | 22 | 75 | 20 |
| 5 | 69 | 12 | 14 | 72 | 16 | 23 | 76 | 21 |
| 6 | 73 | 17 | 15 | 71 | 15 | 24 | 74 | 19 |
| 7 | 72 | 15 | 16 | 75 | 20 | 25 | 70 | 13 |
| 8 | 75 | 19 | 17 | 71 | 15 | | | |
| 9 | 72 | 16 | 18 | 75 | 19 | | | |

a. Perform a regression relating GOALS to HEIGHT to ascertain if there is such a relationship and, if there is, estimate the nature of that relationship. Use the regression ANOVA table to assess the usefulness of HEIGHT as a predictor of GOALS.

**Step 1: Scatterplot**




Since we see a linear pattern, linear regression may be appropriate (Assumption 1 is met).

**Step 2: Compute the Sums of Squares**

Let *x* be HEIGHT and *y* be GOALS.

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 133{,}373 - \frac{(1825)^2}{25} = 148$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 30{,}912 - \frac{(1825)(421)}{25} = 179$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 7321 - \frac{(421)^2}{25} = 231.36 = SSTo$$

**Step 3: Compute the Least-Squares Linear Regression Equation**

$$b = \frac{S_{xy}}{S_{xx}} = \frac{179}{148} = 1.2095$$

$$a = \bar{y} - b\bar{x} = \frac{421}{25} - (1.2095)\left(\frac{1825}{25}\right) = -71.4535 \qquad SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = 216.49324$$

$$\hat{y} = a + bx = -71.4535 + 1.2095x$$

***ANOVA Table***

| Source | *df* | *SS* | *MS* | *F*-statistic | *p*-value |
|---|---|---|---|---|---|
| Regression | 1 | 216.49324 | 216.49324 | 334.931 | < 0.001 |
| Error | 23 | 14.86676 | 0.64638 | | |
| Total | 25 | 231.36 | | | |

Since the *p*-value is extremely small, HEIGHT is useful as a predictor of GOALS.
The relationship the regression suggests is an increase of 1.2 goals for every extra inch of height.

b. Estimate the number of goals to be made by an athlete who is 60 inches tall. How much confidence can be assigned to that estimate?

$$\hat{y} = -71.4535 + 1.2095(60) = 1.1165$$

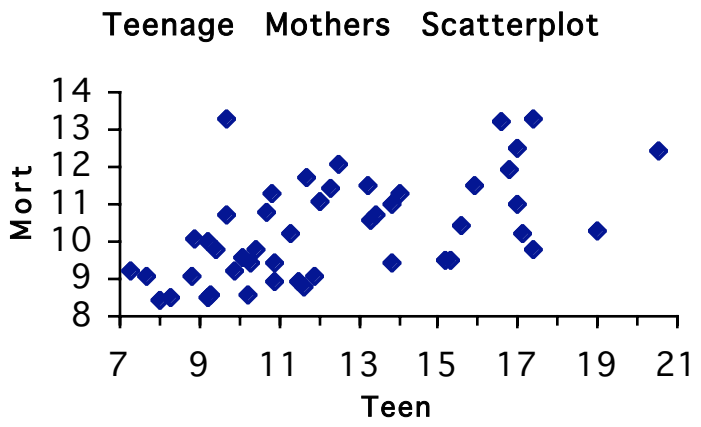
An athlete who is 60 inches (5 feet) tall will make only 1.1165 goals on average in 60 seconds. Very little confidence can be assigned to this estimate since it seems foolish…short people will almost definitely make more than 1 goal in 60 seconds. This is an example of why we should not ***extrapolate*** (predict for *x*-values that are outside of the range of our original sample).

3. It has been argued that many cases of infant mortality rates are caused by teenage mothers who, for various reasons, do not receive proper prenatal care. From the *Statistical Abstract of the United States* we have statistics on the teenage birth rate (per 1000) and the infant mortality rate (per 1000 live births) for the 48 contiguous states. The data are given below, where TEEN denotes the birthrate for teenage mothers and MORT denotes the infant mortality rate.

| State | Teen | Mort | State | Teen | Mort | State | Teen | Mort |
|---|---|---|---|---|---|---|---|---|
| AL | 17.4 | 13.3 | MA | 8.3 | 8.5 | OH | 13.3 | 10.6 |
| AR | 19.0 | 10.3 | MD | 11.7 | 11.7 | OK | 15.6 | 10.4 |
| AZ | 13.8 | 9.4 | ME | 11.6 | 8.8 | OR | 10.9 | 9.4 |
| CA | 10.9 | 8.9 | MI | 12.3 | 11.4 | PA | 11.3 | 10.2 |
| CO | 10.2 | 8.6 | MN | 7.3 | 9.2 | RI | 10.3 | 9.4 |
| CT | 8.8 | 9.1 | MO | 13.4 | 10.7 | SC | 16.6 | 13.2 |
| DE | 13.2 | 11.5 | MS | 20.5 | 12.4 | SD | 9.7 | 13.3 |
| FL | 13.8 | 11.0 | MT | 10.1 | 9.6 | TN | 17.0 | 11.0 |
| GA | 17.0 | 12.5 | NB | 8.9 | 10.1 | TX | 15.2 | 9.5 |
| IA | 9.2 | 8.5 | NC | 15.9 | 11.5 | UT | 9.3 | 8.6 |
| ID | 10.8 | 11.3 | ND | 8.0 | 8.4 | VA | 12.0 | 11.1 |
| IL | 12.5 | 12.1 | NH | 7.7 | 9.1 | VT | 9.2 | 10.0 |
| IN | 14.0 | 11.3 | NJ | 9.4 | 9.8 | WA | 10.4 | 9.8 |
| KS | 11.5 | 8.9 | NM | 15.3 | 9.5 | WI | 9.9 | 9.2 |
| KY | 17.4 | 9.8 | NV | 11.9 | 9.1 | WV | 17.1 | 10.2 |
| LA | 16.8 | 11.9 | NY | 9.7 | 10.7 | WY | 10.7 | 10.8 |

a. Perform a regression to estimate MORT using TEEN as the independent variable. Do the results confirm the stated hypothesis? Interpret the results. Use a regression ANOVA table.

**Step 1: Scatterplot**




Since we see a somewhat linear pattern, linear regression may be appropriate (Assumption 1 is probably met).

**Step 2:** **Compute the Sums of Squares**

Let *x* be TEEN and *y* be MORT.

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 7929.88 - \frac{(596.8)^2}{48} = 509.6667$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 6276.68 - \frac{(596.8)(495.6)}{48} = 114.72$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 5202.72 - \frac{(495.6)^2}{48} = 85.65 = SSTo$$

**Step 3:** **Compute the Least-Squares Linear Regression Equation**

$$b = \frac{S_{xy}}{S_{xx}} = \frac{114.72}{509.6667} = 0.2251$$

$$a = \bar{y} - b\bar{x} = \frac{495.6}{48} - (0.2251)\left(\frac{596.8}{48}\right) = 7.5263 \qquad SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = 25.82213$$

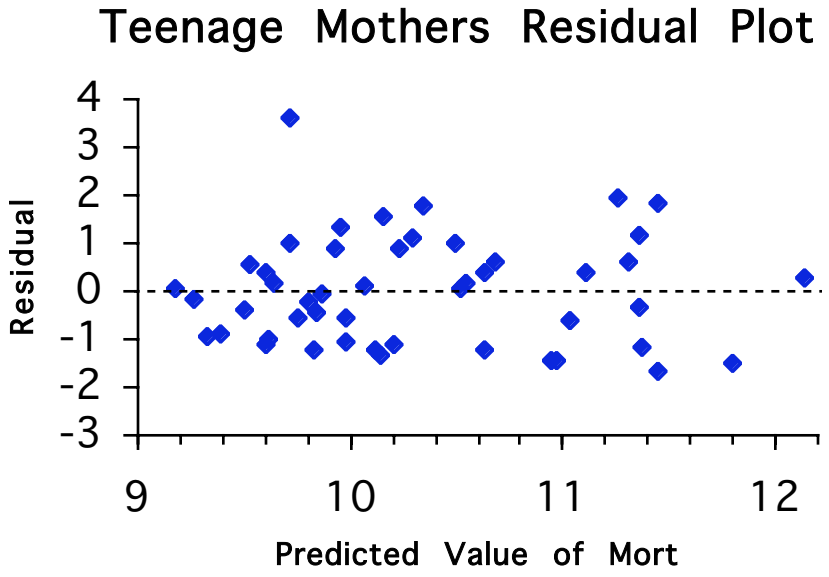
$$\hat{y} = a + bx = 7.5263 + 0.2251x$$

***ANOVA Table***

| Source | *df* | *SS* | *MS* | *F*-statistic | *p*-value |
|---|---|---|---|---|---|
| Regression | 1 | 25.82213 | 25.82213 | 19.854 | < 0.001 |
| Error | 46 | 59.82787 | 1.30061 | | |
| Total | 47 | 85.65000 | | | |

There seems to be an increase in infant mortality with increased teenage birth rate (TEEN is a useful predictor of MORT with a positive slope—small *p*-value). These results seem to confirm the stated hypothesis.

b. Construct a residual plot. Comment on the results.




The residuals seem to be randomly scattered with an even variability (Assumption 3 is met). There may be one outlier since there is a large residual (South Dakota—high mortality rate but low teen birth rate).

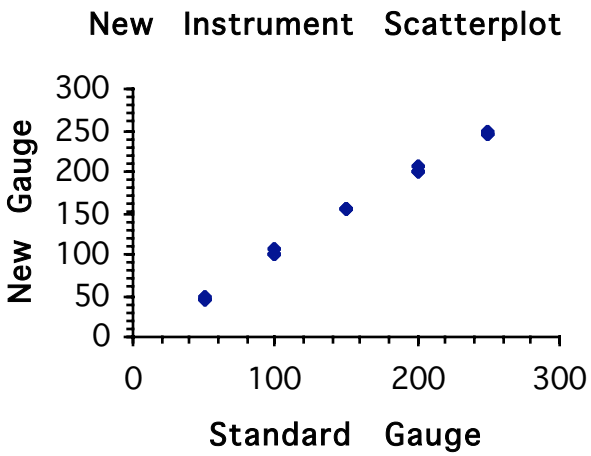
[Note that sometimes a residual plot uses the predicted values on the *x*-axis instead of the predictor values. Interpretations of this type of plot are similar to those of the usual residual plot method.]

4. An experimenter is testing a new pressure gauge against a standard (a gauge known to be accurate) by taking three readings each at 50, 100, 150, 200, and 250 pounds per square inch. The purpose of the experiment is to ascertain the precision and accuracy of the new gauge. The data are shown below.

| Standard Gauge | 50 | 100 | 150 | 200 | 250 |
|---|---|---|---|---|---|
| | 48 | 100 | 154 | 201 | 247 |
| New Gauge | 44 | 100 | 154 | 200 | 245 |
| | 46 | 106 | 154 | 205 | 246 |

As we saw in Example 7.3 both precision and accuracy are important factors in determining the effectiveness of a measuring instrument. Perform the appropriate analysis to determine the effectiveness of this instrument. However, this device has a shortcoming that is of a slightly different nature. Perform the appropriate ANOVA table analyses to find the shortcoming.

**Step 1: Scatterplot**




Since we see a linear pattern, linear regression may be appropriate (Assumption 1 is met).

**Step 2: Compute the Sums of Squares**

Let *x* be STANDARD and *y* be NEW GAUGE.

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 412{,}500 - \frac{(2250)^2}{15} = 75{,}000$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 412{,}500 - \frac{(2250)(2250)}{15} = 75{,}000$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 412{,}716 - \frac{(2250)^2}{15} = 75{,}216 = SSTo$$

**Step 3: Compute the Least-Squares Linear Regression Equation**

$$b = \frac{S_{xy}}{S_{xx}} = \frac{75{,}000}{75{,}000} = 1$$

$$a = \bar{y} - b\bar{x} = \frac{2250}{15} - (1)\left(\frac{2250}{15}\right) = 0$$

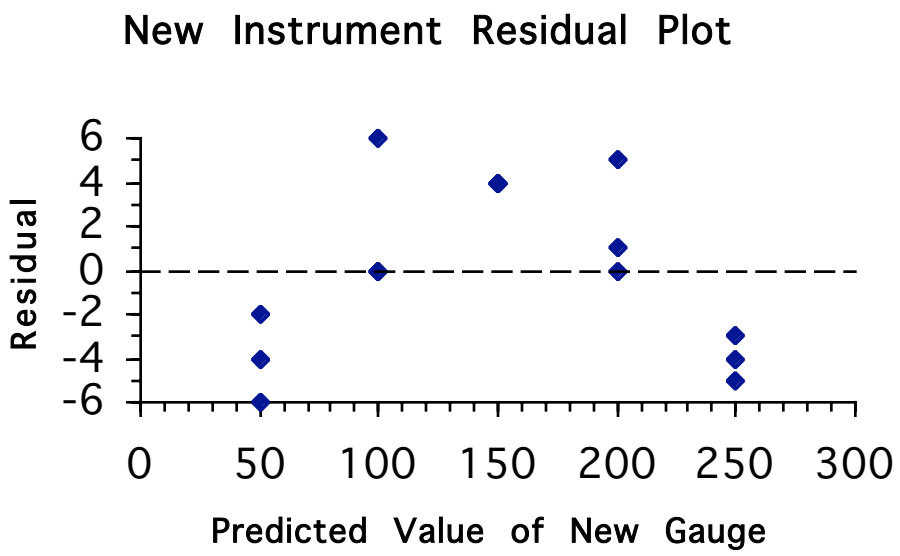
$$SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = 75000$$

$$\hat{y} = a + bx = 0 + (1)x = x$$

***ANOVA Table***

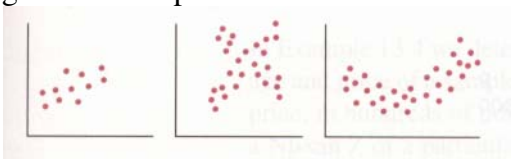
| Source | *df* | *SS* | *MS* | *F*-statistic | *p*-value |
|---|---|---|---|---|---|
| Regression | | 75000 | 75000 | 4513.889 | < 0.001 |
| Error | 13 | 216 | 16.61538 | | |
| Total | 14 | 75216 | | | |

The new gauge is a useful (*p*-value is very small) and accurate instrument (slope of line is 1).




The residuals seem to have an uneven variability (Assumption 3 is not met), and the residuals seem to have a definite concave-down parabolic pattern (Not linear—Assumption 1 is not met). Even though the scatterplot looks nearly linear, the residuals show that a quadratic component should be added to the model. This curvature illustrates the shortcoming of this measuring instrument—the new gauge is precise for the middle values but imprecise for the smaller and larger values. Therefore, the new gauge is accurate (since the slope of the regression line is 1), but it is not precise (uneven variability).

5. For which of the following sets of data points is it reasonable to determine a regression line?



The idea behind finding a regression line is based on the assumption that the data points are actually scattered about a straight line. Only the left data set appears to be scattered about a straight line. Thus, it is reasonable to determine a regression line only for the left set of data.

6. Suppose $r^2 = 1$ for a data set.
   a. What can you say about SSE?

$$r^2 = \frac{SSRegr}{SSTo} = \frac{SSTo - SSE}{SSTo} = 1$$ Therefore, SSE must equal 0.

   b. What can you say about SSRegr?

$$r^2 = \frac{SSRegr}{SSTo} = \frac{SSTo - SSE}{SSTo} = 1$$ Therefore, SSRegr must equal SSTo.

   c. What can you say about the utility of the regression equation for making predictions?

The regression is extremely useful for making predictions since there is a perfect linear relationship between the explanatory and response variables.

7. Suppose $r^2 = 0$ for a data set.
   a. What can you say about SSE?

$$r^2 = \frac{SSRegr}{SSTo} = \frac{SSTo - SSE}{SSTo} = 0$$ Therefore, SSE must equal SSTo.

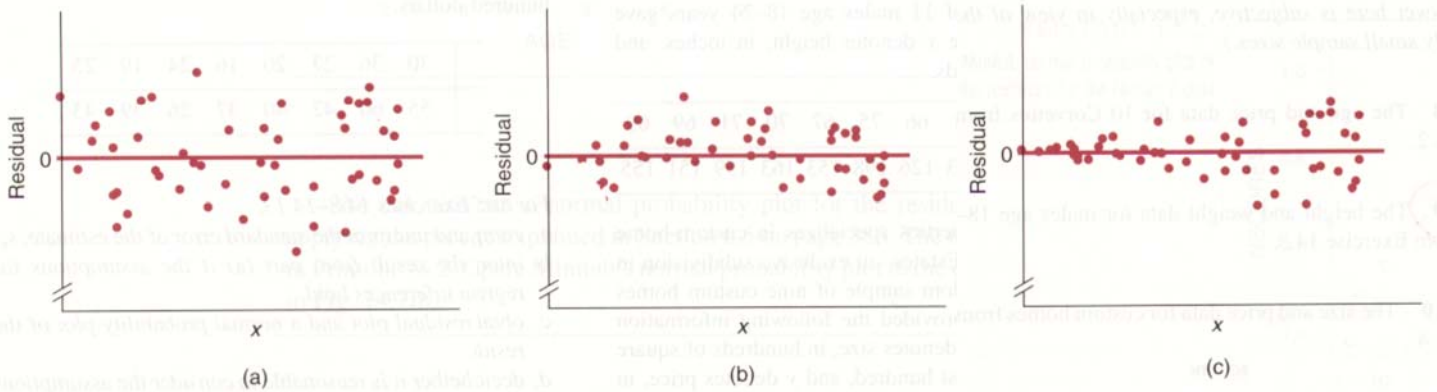
   b. What can you say about SSRegr?

$$r^2 = \frac{SSRegr}{SSTo} = \frac{SSTo - SSE}{SSTo} = 0$$ Therefore, SSRegr must equal 0.

   c. What can you say about the utility of the regression equation for making predictions?

   The regression is totally useless for making predictions since there is absolutely no linear relationship between the explanatory and response variables.

8. The figures below show three residual plots. For each plot, decide whether the graph suggests a violation of one or more of the assumptions for regression analysis. Provide a detailed explanation for your answers.




   a. The graph does not suggest a violation of one or more of the assumptions for regression inferences; all points are randomly scattered in a horizontal band.

   b. Assumption (1) appears to be violated since the points seem to form a (slight) curve indicating that the data do not follow a straight-line pattern.

   c. Assumption (3) appears to be violated since the points form a funnel shape indicating non-constant variability.

# *Extensive Linear Regression Examples*

1. Ten Corvettes between 1 and 6 years old were randomly selected from the classified ads of *The Arizona Republic*. The following data were obtained, where $x$ denotes age, in years, and $y$ denotes price, in hundreds of dollars.

| $x$ | 6 | 6 | 6 | 2 | 2 | 5 | 4 | 5 | 1 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 125 | 115 | 130 | 260 | 219 | 150 | 190 | 163 | 260 | 160 |

a. Discuss what it would mean for the assumptions of regression analysis to be satisfied by the variables under consideration.

If the assumptions for regression inferences are satisfied for a model relating a Corvette's age to its price, this means that there are constants $\alpha$, $\beta$, and $\sigma$ such that, for each age $x$, the prices for Corvettes of that age are normally distributed with mea $\alpha + \beta x$ and standard deviation $\sigma$.

b. Determine the regression equation for the data.

| $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 6 | 125 | 750 | 36 | 15.625 |
| 6 | 115 | 690 | 36 | 13,225 |
| 6 | 130 | 780 | 36 | 16,900 |
| 2 | 260 | 520 | 4 | 67,600 |
| 2 | 219 | 438 | 4 | 47,961 |
| 5 | 150 | 750 | 25 | 22,500 |
| 4 | 190 | 760 | 16 | 36,100 |
| 5 | 163 | 815 | 25 | 26,569 |
| 1 | 260 | 260 | 1 | 67,600 |
| 4 | 160 | 640 | 16 | 25,600 |
| 41 | 1772 | 6403 | 199 | 339,680 |

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 6403 - \frac{41(1772)}{10} = -862.2$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 199 - \frac{(41)^2}{10} = 30.9$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 339{,}680 - \frac{(1772)^2}{10} = 25{,}681.6$$

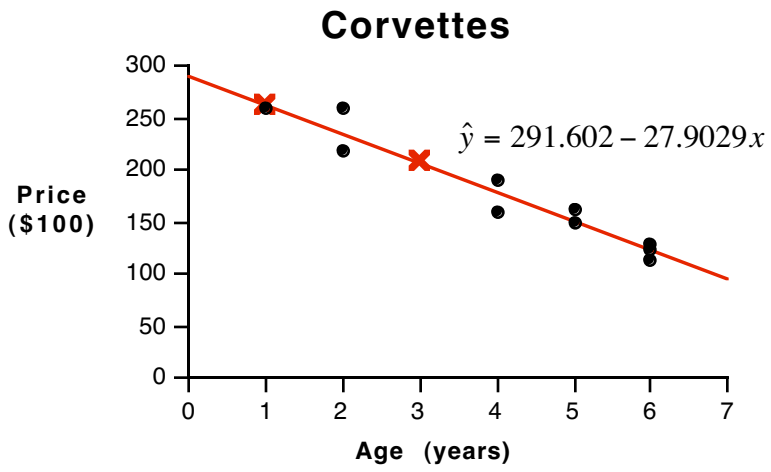
$$\overline{x} = \frac{\sum x}{n} = \frac{41}{10} = 4.1 \qquad \overline{y} = \frac{\sum y}{n} = \frac{1772}{10} = 177.2$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-862.2}{30.9} = -27.9029 \qquad a = \overline{y} - b\overline{x} = 177.2 - (-27.9029)(4.1) = 291.602$$

Therefore, the regression equation is: $\hat{y} = 291.602 - 27.9029x$.

c. Graph the regression equation and the data points.

For $x = 0$, $\hat{y}$ = 291.602 – 27.9029(1) = 263.6991.
For $x = 3$, $\hat{y}$ = 291.602 – 27.9029(3) = 207.8933.




d. Describe the apparent relationship between age and price for Corvettes.

The price for Corvettes tends to decrease as they get older (as age increases).

e. What does the slope of the regression line represent in terms of Corvette prices?

The slope indicates that Corvettes depreciate an estimated $2,790.29 per year.

f. Use the regression equation obtained in part (b) to predict the price of a 2-year-old Corvette; a 3-year-old Corvette.

For a 2-year-old Corvette, $\hat{y}$ = 291.602 – 27.9029(2) = 235.7962 or $23,579.62.
For a 3-year-old Corvette, $\hat{y}$ = 291.602 – 27.9029(3) = 207.8933 or $20,789.33.

g. Identify the predictor and response variables.

The predictor variable is age. The response variable is price.

h. Identify outliers and potential influential observations.

There do not appear to be any outliers or potential influential observations.

i. Compute SSTo, SSRegr, and SSE.

$$SSTo = S_{yy} = 25{,}681.6 \qquad SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = \frac{(-862.2)^2}{30.9} = 24{,}057.9$$

$$SSE = SSTo - SSRegr = 25{,}681.6 - 24{,}057.9 = 1{,}623.7$$

j. Compute the coefficient of determination, $r^2$.

$$r^2 = \frac{SSRegr}{SSTo} = \frac{24{,}057.9}{25{,}681.6} = 0.9367$$

k. Determine the percentage of the total variation in the observed $y$-values that is explained by the regression, and interpret your result.

About 93.67% of the variation in the price data is explained by age.

l. State how useful the regression equation appears to be for making predictions.

The regression equation appears to be very useful for making predictions since the value of $r^2$ is close to 1.

m. Compute the linear correlation coefficient, $r$.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-862.2}{\sqrt{(30.9)(25{,}681.6)}} = -0.967872$$

n. Interpret the value of $r$ in terms of the linear relationship between the two variables in question.

The above value of $r$ suggests a strong negative linear correlation since the value is negative and close to -1.

o. Discuss the graphical interpretation of the value of $r$ and check that it is consistent with the graph you obtained above.

Since the above value of $r$ suggests a strong negative linear correlation, the data points should be clustered closely about a negatively sloping regression line. This is consistent with the graph obtained above.

p. Square $r$ and compare the result with the value of the coefficient of determination ($r^2$) you obtained above.

$(r)^2 = (-0.967872)^2 = 0.9637 = r^2$

This value matches the coefficient of determination that was calculated above.

q. At the 10% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that age is useful as a predictor of price for Corvettes?

**Step 1: Hypotheses**

$H_0$: $\beta = 0$ (*Age is not a useful predictor of price.*)

$H_a$: $\beta \neq 0$ (*Age is a useful predictor of price.*)

**Step 2: Significance Level**

$\alpha = 0.10$

**Step 3: Critical Value(s) and Rejection Region(s)**

$\pm t_{\alpha/2, df = n-2} = \pm t_{0.05, df = 8} = \pm 1.86$

Reject the null hypothesis if $T \leq -1.86$ or if $T \geq 1.86$ ($p$-value $\leq 0.10$).

**Step 4: Test Statistic**

$$T = \frac{b - 0}{s_\varepsilon \Big/ \sqrt{S_{xx}}} = \frac{-27.9029 - 0}{14.2465 \Big/ \sqrt{30.9}} = -10.8873 \qquad (p\text{-value} < 2(0.001) = 0.002)$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1{,}623.7}{8}} = 14.2465$$

**Step 5: Conclusion**

Since -10.8873 ≤ -1.860 ($p$-value < 0.002 ≤ 0.10), we shall reject the null hypothesis.

**Step 6: State conclusion in words**

At the $\alpha = 0.10$ level of significance, there exists enough evidence to conclude that the slope of the population regression line is not zero and, hence, that age is useful as a predictor of price for Corvettes.

r. Obtain a point estimate for the mean price of all 4-year-old Corvettes.

$$\hat{y}^* = 291.602 - 27.9029(4) = 179.9904 = \$17{,}999.04$$

s. Determine a 90% confidence interval for the mean price of all 4-year-old Corvettes.

$$\hat{y}^* \pm t_{\alpha/2, df=n-2} \cdot s_\varepsilon \sqrt{\frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{xx}}}$$

$$179.9904 \pm 1.86 \cdot 14.2464 \sqrt{\frac{1}{10} + \frac{(4 - 4.1)^2}{30.9}}$$

$$[171.5974 \ \ to \ \ 188.3834]$$

We can be 90% confident that the mean price of all four-year-old Corvettes is somewhere between $17,159.74 and $18,838.34.

t. Find the predicted price of a randomly selected 4-year-old Corvette.

$$\hat{y}^* = 291.602 - 27.9029(4) = 179.9904 = \$17{,}999.04 \quad \text{[same as in part (r)]}$$

u. Determine a 90% prediction interval for the price of a randomly selected 4-year-old Corvette.

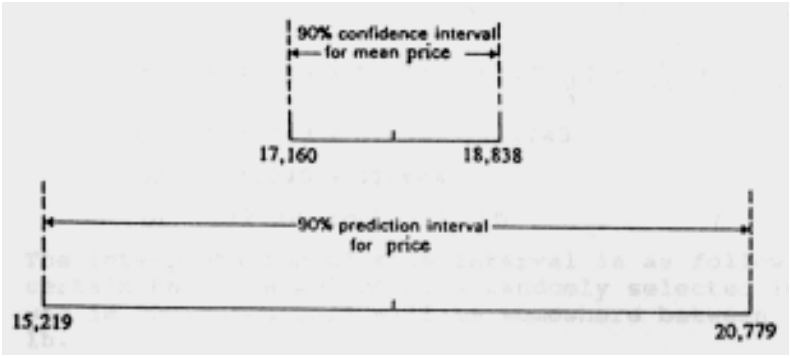
$$\hat{y}^* \pm t_{\alpha/2, df=n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{S_{xx}}}$$

$$179.9904 \pm 1.860 \cdot 14.2464 \sqrt{1 + \frac{1}{10} + \frac{(4 - 4.1)^2}{30.9}}$$

$$[152.1947 \ \ to \ \ 207.7861]$$

We can be 90% certain that the price of a randomly selected four-year-old Corvette is somewhere between $15,219.47 and $20,778.61.

v. Draw a graph showing both the 90% confidence interval from part (s) and the 90% prediction interval from part (u).




w. Why is the prediction interval wider than the confidence interval?

The error in the estimate of the mean price of four-year-old Corvettes is due only to the fact that the population regression line is being estimated by a sample regression line; whereas, the error in the prediction of the price of a randomly selected four-year-old Corvette is due to that fact plus the variation in prices for four-year-old Corvettes.

x. At the 5% significance level, do the data provide sufficient evidence to conclude that age and price of Corvettes are negatively linearly correlated?

**Step 1: Hypotheses**

$H_0$: $\rho = 0$ (*Age and price are not linearly correlated.*)

$H_a$: $\rho < 0$ (*Age and price are negatively linearly correlated.*)

**Step 2: Significance Level**

$\alpha = 0.05$

**Step 3: Critical Value(s) and Rejection Region(s)**

$-t_{\alpha, df=n-2} = -t_{0.05, df=8} = -1.86$

Reject the null hypothesis if $T \le -1.86$ ($p$-value $\le 0.05$).

**Step 4: Test Statistic**

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.967872}{\sqrt{\frac{1-(-0.967872)^2}{8}}} = -10.8874 \qquad (p\text{-value} < 0.001)$$

**Step 5: Conclusion**

Since -10.8874 ≤ -1.86 ($p$-value < 0.001 ≤ 0.05), we shall reject the null hypothesis.

**Step 6: State conclusion in words**

At the $\alpha = 0.05$ level of significance, there exists enough evidence to conclude that age and price for Corvettes are negatively linearly correlated.

2. The National Center for Health Statistics publishes data on heights and weights in *Vital and Health Statistics*. A random sample of 11 males age 18–24 years gave the following data, where $x$ denotes height, in inches, and $y$ denotes weight, in pounds.

| $x$ | 65 | 67 | 71 | 71 | 66 | 75 | 67 | 70 | 71 | 69 | 69 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 175 | 133 | 185 | 163 | 126 | 198 | 153 | 163 | 159 | 151 | 155 |

a. Discuss what it would mean for the assumptions of regression analysis to be satisfied by the variables under consideration.

If the assumptions for regression inferences are satisfied for a model relating an 18–24-year-old male's height to his weight, this means that there are constants $\alpha$, $\beta$, and $\sigma$ such that, for each height $x$, the weights of 18–24-year-old males of that height are normally distributed with mean $\alpha + \beta x$ and standard deviation $\sigma$.

b. Determine the regression equation for the data.

| $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 65 | 175 | 11,375 | 4,225 | 30,625 |
| 67 | 133 | 8,911 | 4,489 | 17,689 |
| 71 | 185 | 13,135 | 5,041 | 34,225 |
| 71 | 163 | 11,573 | 5,041 | 26,569 |
| 66 | 126 | 8,316 | 4,356 | 15,876 |
| 75 | 198 | 14,850 | 5,625 | 39,204 |
| 67 | 153 | 10,251 | 4,489 | 23,409 |
| 70 | 163 | 11,410 | 4,900 | 26,569 |
| 71 | 159 | 11,289 | 5,041 | 25,281 |
| 69 | 151 | 10,419 | 4,761 | 22,801 |
| 69 | 155 | 10,695 | 4,761 | 24,025 |
| 761 | 1761 | 122,224 | 52,729 | 286,273 |

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} = 122{,}224 - \frac{761(1761)}{11} = 394.8182$$

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 52{,}729 - \frac{(761)^2}{11} = 81.6364$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 286{,}273 - \frac{(1761)^2}{11} = 4{,}352.91$$

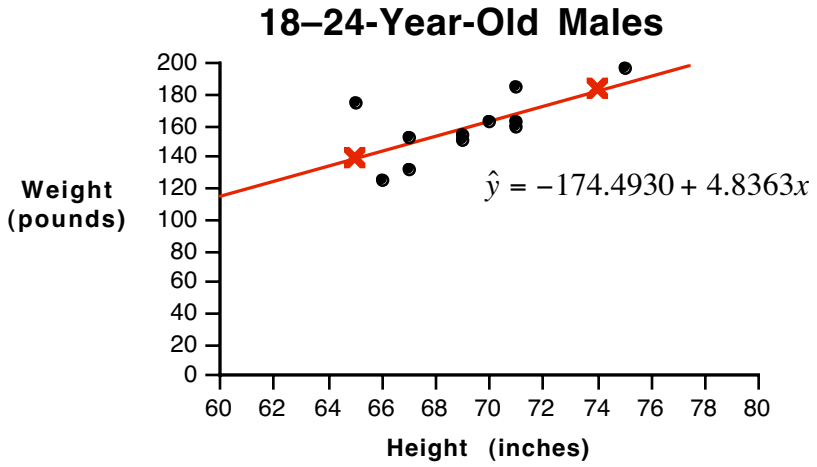
$$\bar{x} = \frac{\sum x}{n} = \frac{761}{11} = 69.1818 \qquad \bar{y} = \frac{\sum y}{n} = \frac{1761}{11} = 160.0909$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{394.8182}{81.6364} = 4.8363 \qquad a = \bar{y} - b\bar{x} = 160.0909 - (4.8363)(69.1818) = -174.4930$$

Therefore, the regression equation is: $\hat{y} = -174.4930 + 4.8363x$.

c. Graph the regression equation and the data points.

For $x = 65$, $\hat{y}$ = -174.4930 + 4.8363(65) = 139.8665.
For $x = 74$, $\hat{y}$ = -174.4930 + 4.8363(74) = 183.3932.




d. Describe the apparent relationship between height and weight for 18–24-year-old males.

Taller 18–24-year-old males tend to weigh more than smaller ones (or weight tends to increase as height increases).

e. What does the slope of the regression line represent in terms of heights and weights for 18–24-year-old males?

The weights of 18–24-year-old males increase an estimated 4.8363 pounds for each increase in height of one inch.

f. Use the regression equation obtained in part (b) to predict the weight of an 18–24-year-old male who is 67 inches tall; 73 inches tall..

For a 67 inches tall male, $\hat{y}$ = -174.4930 + 4.8363(67) = 149.5391 pounds.
For a 73 inches tall male, $\hat{y}$ = -174.4930 + 4.8363(73) = 178.5569 pounds.

g. Identify the predictor and response variables.

The predictor variable is height. The response variable is weight.

h. Identify outliers and potential influential observations.

The observation (65, 175) appears to be an outlier since it is far away from the regression line. The observation (75, 198) seems to be a potential influential observation since it is to the left of the cluster of the rest of the points.

i. Should the above regression equation be used to predict the weight of an 18–24-year-old male who is 68 inches tall? 60 inches tall? Explain your answers.

It is acceptable to use the regression equation to predict the weight of an 18–24-year-old male who is 68 inches tall since that height lies within the range of the heights in the sample data. It is not acceptable (and would be extrapolation) to use the regression equation to predict the weight of an 18–24-year-old male who is 60 inches tall since that height lies outside the range of the heights in the sample data (the range of heights upon which the regression equation is based).

j. For which heights is it reasonable to use the regression equation to predict weight?

It is reasonable to use the regression equation to predict weight for heights between 65 and 75 inches, inclusive.

k. Compute SSTo, SSRegr, and SSResid.

$$SSTo = S_{yy} = 4{,}352.91 \qquad SSRegr = \frac{\left(S_{xy}\right)^2}{S_{xx}} = \frac{(394.8182)^2}{81.6364} = 1{,}909.46$$

$$SSResid = SSTo - SSRegr = 4{,}352.91 - 1{,}909.46 = 2{,}443.44$$

l. Compute the coefficient of determination, $r^2$.

$$r^2 = \frac{SSRegr}{SSTo} = \frac{1{,}909.46}{4{,}352.91} = 0.4387$$

m. Determine the percentage of the total variation in the observed $y$-values that is explained by the regression, and interpret your result.

About 43.87% of the variation in the weight data is explained by height.

n. State how useful the regression equation appears to be for making predictions.

The regression equation appears to be moderately useful for making predictions since the value of $r^2$ is close to 0.5.

o. Compute the linear correlation coefficient, $r$.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{394.8182}{\sqrt{(81.6364)(4{,}352.91)}} = 0.662319$$

p. Interpret the value of $r$ in terms of the linear relationship between the two variables in question.

The above value of $r$ suggests a moderate positive linear correlation since the value is positive and close to 0.5.

q. Discuss the graphical interpretation of the value of $r$ and check that it is consistent with the graph you obtained above.

Since the above value of $r$ suggests a moderate positive linear correlation, the data points should be clustered moderately closely about a positively sloping regression line. This is consistent with the graph obtained above.

r. Square $r$ and compare the result with the value of the coefficient of determination ($r^2$) you obtained above.

$(r)^2 = (0.662319)^2 = 0.4387 = r^2$ This value matches the coefficient of determination that was calculated above.

s. Compute and interpret the standard error of the estimate, $s_\varepsilon$.

$$s_\varepsilon = \sqrt{\frac{SSResid}{n-2}} = \sqrt{\frac{2{,}443.44}{9}} = 16.4771$$

Roughly speaking, on the average, the predicted weight of an 18–24-year-old male in the sample differs from the observed weight by about 16.4771 pounds.

t. Interpret the result from part (s) if the assumptions for regression analysis hold.

Presuming that the variables height ($x$) and weight ($y$) for 18–24-year-old males satisfy Assumptions (1)–(3) for regression analysis, the standard error of the estimate $s_\varepsilon$ = 16.4771 pounds provides an estimate for the common population standard deviation $\sigma$ of weights for all 18–24-year-old males of any particular height.

u. Obtain the residuals and create a residual plot.

| Height $x$ | Residual $e$ |
|---|---|
| 65 | 35.13 |
| 67 | -16.54 |
| 71 | 16.12 |
| 71 | -5.88 |
| 66 | -18.70 |
| 75 | 9.77 |
| 67 | 3.46 |
| 70 | -1.05 |
| 71 | -9.88 |
| 69 | -8.21 |
| 69 | -4.21 |


v. Decide whether it is reasonable to consider that the assumptions for regression analysis are met by the variables in questions. (The answer here is subjective, especially in view of the extremely small sample sizes.)

It appears reasonable to consider the assumptions for regression inferences met for the variables height and weight since we see a random scatter in a horizontal band in the residual plot. However, there is a potential outlier (65, 175) ($e$ = 35.13) which could cast some doubt on the assumptions.

w. Do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and, hence, that height is useful as a predictor of weight for 18–24-year-old males? Use $\alpha = 0.10$.

**Step 1: Hypotheses**

$H_0$: $\beta = 0$ (*Height is not a useful predictor of weight.*)

$H_a$: $\beta \neq 0$ (*Height is a useful predictor of weight.*)

**Step 2: Significance Level**

$\alpha = 0.10$

**Step 3: Critical Value(s) and Rejection Region(s)**

$\pm t_{\alpha/2, df=n-2} = \pm t_{0.05, df=9} = \pm 1.83$

Reject the null hypothesis if $T \leq -1.83$ or if $T \geq 1.83$ ($p$-value $\leq 0.10$).

**Step 4: Test Statistic**

$$T = \frac{b}{s_\varepsilon / \sqrt{S_{xx}}} = \frac{4.8363}{16.4771 / \sqrt{81.6364}} = 2.6520$$

$(0.02 = 2(0.01) < p\text{-value} < 2(0.025) = 0.050)$

**Step 5: Conclusion**

Since $2.6520 \geq 1.83$ $(0.02 < p\text{-value} < 0.05 \leq 0.10)$, we shall reject the null hypothesis.

**Step 6: State conclusion in words**

At the $\alpha = 0.10$ level of significance, there exists enough evidence to conclude that the slope of the population regression line is not zero and, hence, that height is useful as a predictor of weight for 18–24-year-old males.

x. Obtain a 90% confidence interval for the slope, $\beta$, of the population regression line that relates weight to height for males age 18–24. Be sure to interpret your result.

$$b - t_{\alpha/2, df=n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}} \quad to \quad b + t_{\alpha/2, df=n-2} \cdot \frac{s_\varepsilon}{\sqrt{S_{xx}}}$$

$$4.8363 - 1.83 \cdot \frac{16.4771}{\sqrt{81.6364}} \quad to \quad 4.8363 + 1.83 \cdot \frac{16.4771}{\sqrt{81.6364}}$$

$$\left[1.4990 \ \ pounds \ \ to \ \ 8.1736 \ \ pounds\right]$$

We can be 90% confident that, for 18–24-year-old males, the increase in mean weight per one inch increase in height is somewhere between 1.4990 pounds and 8.1736 pounds.

y. Do the data provide sufficient evidence to conclude that the variables height and weight are positively linearly correlated for 18–24-year-old males? Perform the required hypothesis test at the 5% significance level.

**Step 1: Hypotheses**

$H_0$: $\rho = 0$ (*Height and weight are not linearly correlated.*)

$H_a$: $\rho > 0$ (*Height and weight are positively linearly correlated.*)

**Step 2: Significance Level**

$\alpha = 0.05$

**Step 3: Critical Value(s) and Rejection Region(s)**

$t_{\alpha, df = n-2} = t_{0.05, df = 9} = 1.83$

Reject the null hypothesis if $T \geq 1.83$ ($p$-value $\leq 0.05$).

**Step 4: Test Statistic**

$$T = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.662319}{\sqrt{\frac{1 - (0.662319)^2}{9}}} = 2.6520 \qquad (0.01 < p\text{-value} < 0.025)$$

**Step 5: Conclusion**

Since $2.6520 \geq 1.83$ ($0.01 < p$-value $< 0.025 \leq 0.05$), we shall reject the null hypothesis.

**Step 6: State conclusion in words**

At the $\alpha = 0.05$ level of significance, there exists enough evidence to conclude that height and weight are positively linearly correlated for 18–24-year-old males.