**Chapter 9 Exercises**

**Concepts**

1. If a researcher is reading the work of a peer who has run a regression on two variables that the researcher doubts are linearly related, what measure might he inspect for evidence that the true relationship between two variables x and y is not in fact linear?
2. If a researcher takes a new sample of data from a population, does the true regression line change? Why or why not?
3. What sampling factors might introduce error into the relationship between two variables? Is the existence of an error a sign that the data has been collected poorly?
4. What are the fundamental assumptions in linear regression?

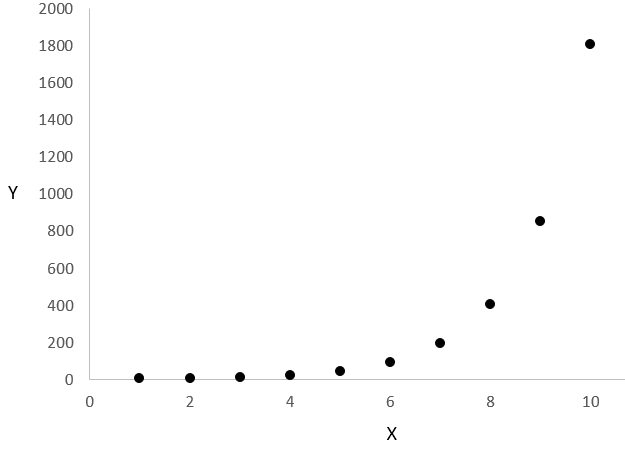
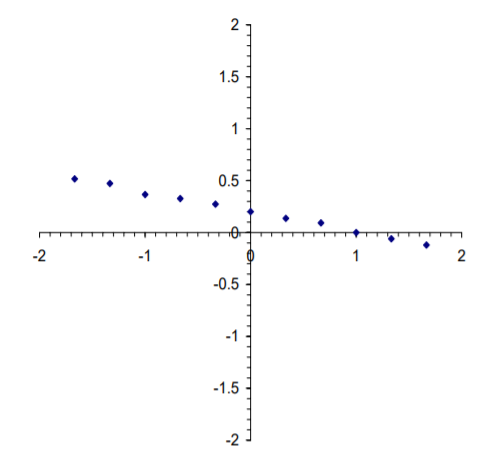
**Exercises**

1. A dataset consists of the following eight (x,y) pairs of values:

|  |  |
| --- | --- |
| **X** | **Y** |
| 0 | 12 |
| 4 | 16 |
| 8 | 22 |
| 15 | 28 |
| 2 | 15 |
| 5 | 14 |
| 13 | 24 |
| 20 | 30 |

* 1. Plot the data in a scatter plot.
  2. Visually inspect and explain whether the relationship between x and y appears to be deterministic or involving randomness.
  3. Based on the plot, explain whether the relationship between x and y appears to be linear or non linear.

1. In the scatter plots below, visually inspect and describe the relationships.

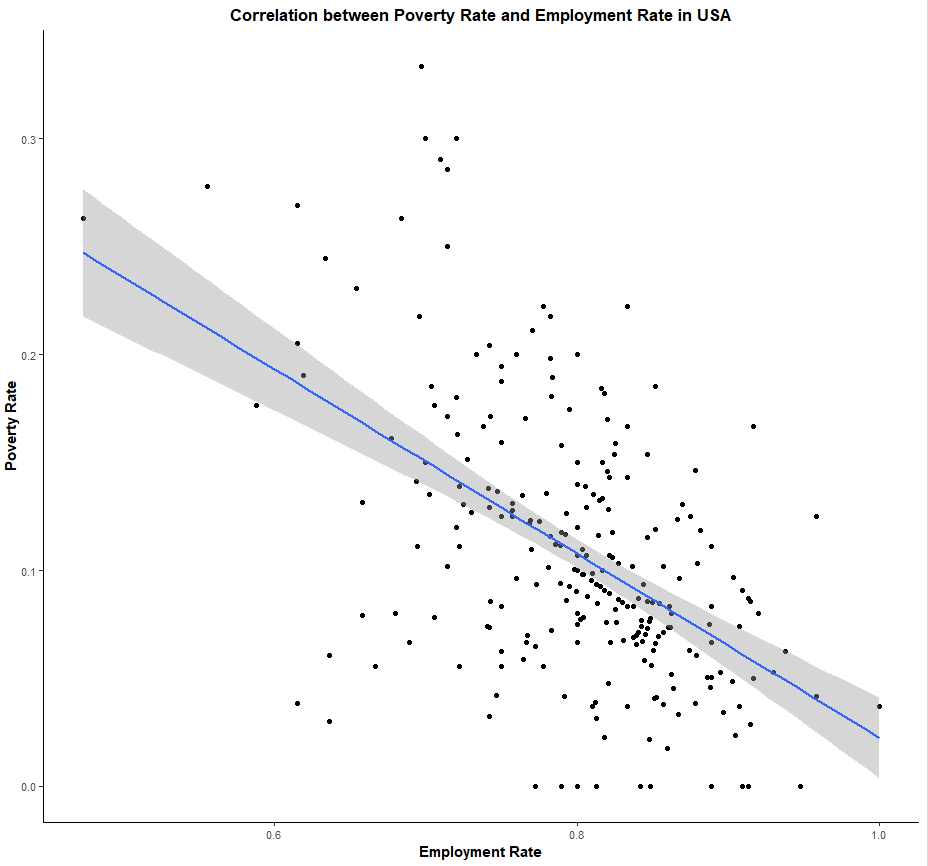


1. Use the following data to answer the questions below:

|  |  |
| --- | --- |
| **X** | **Y** |
| 286 | 77 |
| 252 | 62 |
| 241 | 73 |
| 264 | 74 |
| 260 | 79 |
| 227 | 68 |
| 233 | 74 |
| 259 | 74 |
| 208 | 61 |
| 287 | 76 |

* 1. Fit a regression line to the data.
  2. Test the null hypothesis that the proportion of the variability in y explained by x is zero.

1. What is value is minimized when fitting the regression line during in linear regression analysis? In a plot below, which illustrates the relationship between Poverty rate (dependent variable) and Unemployment rate (independent variable) and includes a fitted regression draw these distances/values.



5. Use the following data to answer the questions below:

|  |  |
| --- | --- |
| **X** | **Y** |
| 7 | 321 |
| 8 | 91 |
| 11 | 59 |
| 8 | 1104 |
| 13 | 138 |
| 10 | 82 |
| 8 | 102 |
| 9 | 325 |
| 7 | 10 |
| 14 | 834 |

* 1. Fit a regression line to the data.
  2. Test the null hypothesis that the proportion of the variability in y explained by x is zero.

6. Use the following data to answer the questions below:

|  |  |
| --- | --- |
| **X** | **Y** |
| 128 | 21 |
| 90 | 90 |
| 92 | 52 |
| 156 | 263 |
| 67 | 8 |
| 140 | 159 |
| 151 | 174 |
| 127 | 135 |
| 110 | 130 |
| 119 | 89 |

* 1. Fit a regression line to the data.
  2. Test the null hypothesis that the proportion of the variability in y explained by x is zero.

1. A regression of per capita GDP on university graduate attainment rates (the number of university-educated persons per thousand adults) is performed on data collected from 42 countries (data entered in US Dollars). The results are summarized below:

Intercept 17032.8

Slope 60.2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** | 5,000 |  |  |  |
| **Residual** |  |  |  |  |
| **Total** | 35,000 |  |  |  |

* 1. Write the regression equation.
  2. Fill in the blanks of the ANOVA table.
  3. What is the predicted per capita GDP of a country where the education completion rate is 598?
  4. In words, what is the meaning of the coefficient 60.2? What does this mean, and how can it be interpreted in the context of the population being studied?
  5. What is the value of the correlation coefficient?
  6. What is the standard deviation of the residuals?
  7. What is the intercept and how would you interpret it?

1. A regression of standardized test scores on family income is performed on data collected from 57 students (data entered in dollars per year). The results are summarized below:

Intercept 250

Slope .015

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** |  |  |  |  |
| **Residual** | 3,300 |  |  |  |
| **Total** | 4,000 |  |  |  |

* 1. Fill in the blanks of the ANOVA table.
  2. What is the predicted test score of a student from a family whose income is 18,000?
  3. In words, what is the meaning of the coefficient .015? What does this mean, and how can it be interpreted in the context of the population being studied?
  4. What is the value of the correlation coefficient?
  5. What is the standard deviation of the residuals?
  6. What is the intercept value and what does this mean in the context of the study?

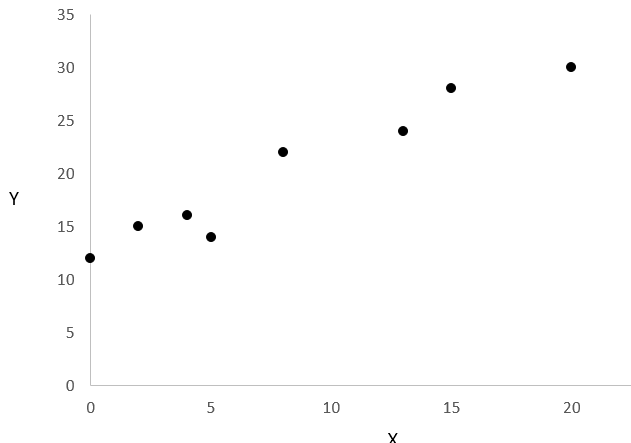
1. Given a linear regression with slope b = 10, sy= 3, sx = 20, and n=52, find the standard error of the estimate (i.e., the standard deviation of the residuals).
2. Calculate the following by hand, without using a computer.

|  |  |
| --- | --- |
| **X** | **Y** |
| 51 | 37 |
| 47 | 28 |
| 50 | 50 |
| 52 | 36 |
| 44 | 31 |
| 51 | 67 |
| 49 | 38 |
| 50 | 46 |
| 48 | 45 |
| 50 | 42 |
| 52 | 54 |
| 47 | 23 |

* 1. Find the regression coefficients (the intercept and slope coefficients).
  2. Estimate the standard error of the residuals about the regression line.
  3. Test the hypothesis that the regression coefficient associated with the independent variable is zero.
  4. Find the value of *R*2

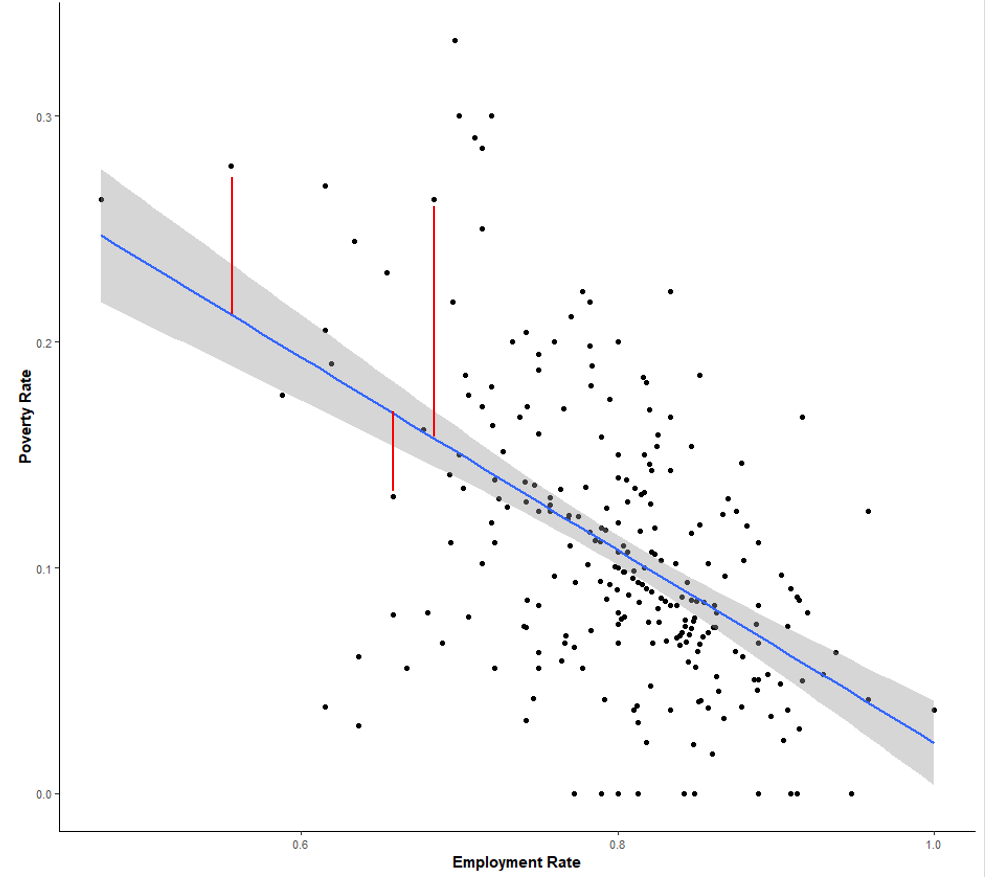
1. Use the SPSS Housing Dataset and SPSS or Excel to answer the following:
   1. Carry out a regression using the average price as the dependent variable and the unemployment rate as the independent variable.
   2. Repeat step a, with the floor area as the independent variable. How do the R2 values compare between the two tests? What does this suggest?
2. Use the Milwaukee Dataset and SPSS or Excel to answer the following:
   1. Carry out a regression using the sale price as the dependent variable and the square feet as the independent variable.
   2. Repeat step a, with the lot size as the independent variable. How do the R2 values compare between the two tests? What does this suggest?

**Chapter 9 Solutions**



* 1. The data show a clear, visual trend with the values of Y rising with the value of X. There does appear to be some variation surrounding the trend, but overall a deterministic relationships appears to exist.
  2. The relationship does appear to be linear and positive.
  3. There is a negative linear relationship between the X and Y variables
  4. There is a positive relationship between the variables. However, that relationship does not appear to be linear.
  5. The regression line can be expressed as y = .2x + 30.2
  6. The observed *F*-ratio is 6.72, which exceeds the critical value 5.32. Thus we reject the null hypothesis.

1. Linear regression minimizes the sum of squared residuals, or the vertical distance between each data point and the estimated regression line. These distances can be visualized as follows on the provided plot. Here the red lines indicate the distances minimized by regression.



* 1. The regression line can be expressed as y = 26.2x + 57.3
  2. The observed F ratio is .251, which is less than the critical value 5.32. Thus we fail to reject the null hypothesis.
  3. The regression line can be expressed as y = 2.1x - 136.5
  4. The observed F ratio is 12.429, which exceeds the critical value 5.32. Thus we reject the null hypothesis.
  5. y = 17,032.8 + 60.2x

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** | 5,000 | 1 | 5,000 | 6.667 |
| **Residual** | 30,000 | 40 | 750 |  |
| **Total** | 35,000 | 41 |  |  |

* 1. The predicted per capita GDP is $53032.4
  2. The slope 60.2 implies that every university-educated person per 1000 people increases the per capita GDP by $60.2
  3. r = 0.378
  4. Standard deviation of the residuals = 27.39
  5. 17,032. This is the expected mean of Y when the value of x = 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sum of squares** | **df** | **Mean square** | ***F*** |
| **Regression** | 700 | 1 | 700 | 11.667 |
| **Residual** | 3300 | 55 | 60 |  |
| **Total** | 4000 | 56 |  |  |

* 1. The predicted test score for such a student is 520.
  2. The coefficient implies that for every dollar per year the student’s family earns, they will score approximately .015 points.
  3. r = 0.418
  4. Standard deviation of the residuals = 7.75
  5. 250. In the context of this study this would be the average test score when a student’s family had zero income. While possible, this level of income is unlikely to occur, which complicates the interpretation of the intercept in this case.

1. 14.81
   1. intercept: -108.7. Slope: 3.0
   2. 10.151
   3. The *F*-value is 5.613, which is greater than the critical value of 4.96. We reject the null hypothesis.
   4. .359
   5. The intercept is 49868.965, and the slope is -1515.454. The significance value associated with the relationship is less than .001, so that we reject the null hypothesis.
   6. The intercept is -1311.770, and the slope is 447.399. The significance value associated with the relationship is less than .001, so that we reject the null hypothesis. However, the *R*2 value associated with floor area as the independent variable is .468, compared with the R2 associated with unemployment, which was .071. The explanation power differs substantially between the two.
   7. The intercept is 1841.9, and the slope is 85.9. The significance value associated with the relationship is less than .001, so that we reject the null hypothesis.
   8. The intercept is 92984.2, and the slope is 6.4. The significance value associated with the relationship is less than .001, so that we reject the null hypothesis. However, the R2 value associated with the floor area as the independent variable is .397, compared with the R2 associated with lot size, which was .045. The explanation power differs substantially between the two, so that price depends more on the size of the housing than on the lot.