**Quantitative Methods**

**List of Exercises N. 7**

**Selected Exercises from McClave (2014) – Chapter 11**

**11.1 Probabilistic Models**

1. (5, 6, 7). Plot the following lines and give their slopes and intercepts. Why do we generally prefer a probabilistic model to a deterministic model? Give examples for when the 2 types of models might be appropriate.

a) y = 4 + x

b) y = 5 - 2x

c) y = - 4 + 3x

d) y = - 2x

e) y = x

f) y = 0.50 + 1.5x

2. (9). If a straight-line probabilistic relationship related the mean E(y) to an independent variable x, does it imply that every value of the variable y will always fall exactly on the line of means? Why or why not?

**11.2 Fitting the Model: The Least Squares Approach**

3. (10, 11). The following table is used for making the preliminary computations for finding the least squares line for the given pairs of x and y values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Xi | Yi | Xi2 | Xi Yi |
|  | 7 | 2 |  |  |
|  | 4 | 4 |  |  |
|  | 6 | 2 |  |  |
|  | 2 | 5 |  |  |
|  | 1 | 7 |  |  |
|  | 1 | 6 |  |  |
|  | 3 | 5 |  |  |
| Totals | ΣXi= | ΣYi = | ΣXi2 = | Σ Xi Yi = |

1. Complete the table.
2. Find SSxx.
3. Find and .
4. Find the least squares line.
5. Find SSxy.
6. Find 1.
7. Find 0.

After the least squares line has been obtained, the table below (which is similar to the prior table) can be used for (1) comparing the observed and the predicted values of y and (2) computing SSE.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Xi | Yi |  | Yi - |  | (Yi - )2 |
|  | 7 | 2 |  |  |  |  |
|  | 4 | 4 |  |  |  |  |
|  | 6 | 2 |  |  |  |  |
|  | 2 | 5 |  |  |  |  |
|  | 1 | 7 |  |  |  |  |
|  | 1 | 6 |  |  |  |  |
|  | 3 | 5 |  |  |  |  |
|  |  |  |  | Σ(Yi - ) = |  | SSE = Σ(Yi - )2 = |

1. Complete the table.
2. Plot the least suares line on a scatterplot of the data. Plot the following line on the same graph:

14 – 2.5 x.

1. Show that SSE is larger for the line in part i than it is for the least squares line.

4. (20, FEMA). ***Public corruption and bad weather***. The Federal Emergency Management Agency (FEMA) provides disaster relief for states impacted by natural disasters (e.g. hurricanes, tornados, floods). DO these bad wheather windfalls lead to public corruption? This was the research question of interest in an article published in the (Journal of Law and Economics (Nov. 2008). Data on y = average annual number of public corruption convictions (per 100,000 residents) and x = average annual FEMA relief (in dollars) per capita for each of the 50 states were used in the investigation.

1. Access the data saved in the file and construct a scatterplot. Do you observe a trend?
2. Fit the simple linear regression model E(y) = β0 + β1 x to the data and obtain estimates of the intercept and slope.
3. Practically interpret the estimated intercept and estimated slope.

5. (25, TOPBUS). ***Survey on the top business schools***. Each year, the Wall Street Journal and Harris Interactive track the opinions and summarize the results in the Business School Survey. One of these surveys included rankings of 76 business schools. Survey data for the top 10 business schools are given in the table above. All data are saved in the file.

1. Select one of the variables as the dependent variable, y, and another as the independent variable, x. Use your knowledge of the subject area and common sense to help you select the variables.
2. Fit the simple linear model, E(y) = β0 + β1 x, to the data for all 76 business schools. Interpret the estimates of the slope and y-intercept.

**11.3 Model Assumptions**

6. (36, FEMA). ***Public corruption and bad weather***. Refer to the Journal of Law and Economics (No. 2008) study of the link between Federal Emergency Management Agency (FEMA) disaster relief and public corruption. In exercise 4, you used the data in the file to fit a straight-line model relating a state’s average annual number of public corruption convictions (y) to the state’s average annual FEMA relief (x).

1. Estimate σ2, the variance of the random error term in the model.
2. Estimate σ, the standard deviation of the random error term in the model.
3. Which of the estimates, part a or part b, can be interpreted practically? Why?
4. Make a statement about how accurate the model is in predicting a state’s average annual number of public corruption convictions.

7. (37, OJUICE). ***Sweetness of orange juice***. Refer to the quality of orange juice produced at a juice manufacturing plant. Use a simple linear regression to predict the sweetness index (y) from the amount of pectin (x) in the orange juice.

1. Find the values of SSE, s2 and s for this regression.
2. Explain why it is difficult to give a practical interpretation of the value s2.
3. Give a practical interpretation of the value f s.

**11.4 Assessing the Utility of the Model: Making Inferences about the Slope β1**

8. (51, BDAYS). ***Software millionaires and birthdays***. Refer to the study of whether birth decade can predict the number of software millionaires born in the decade. The data are reproduced in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Decade | Total US births (millions) | N. of Software Millionaires Birthdays | N. of CEO Birthdays (in a random sample of 70 companies from the Fortune 500 list) |
| 1920 | 28582 | 3 | 2 |
| 1930 | 24374 | 1 | 2 |
| 1940 | 31666 | 10 | 23 |
| 1950 | 40530 | 14 | 38 |
| 1960 | 38808 | 7 | 9 |
| 1970 | 33309 | 4 | 0 |

1. Construct a 95% confidence interval for the slope of the model, E(y) = β0 + β1 x, where x = total number of US births and y = number of software millionaire birthdays. Give a practical interpretation of the interval assuming that .
2. Construct a 95% confidence interval for the slope of the model, E(y) = β0 + β1 x, where x = number of CEO birthdays and y = number of software millionaire birthdays. Give a practical interpretation of the interval assuming that .
3. Can you conclude that number of software millionaires born in a decade is linearly related to total number of people born in the USA? Number of CEOs born in the decade?

9. (52). ***Beauty and electoral success***. Are good looks an advantage when running for political office? This was the question of interest in an article published in the Journal of Public Economics (Feb. 2010). The researchers focused on a sample of 641 nonincumbent candidates for political office in Finland. Photos of each candidate were evaluated by non-Finnish subjects; each evaluator assigned a beauty rating – measured on a scale of 1 (lowest rating) to 5 (highest rating) – to each candidate. The beauty ratings for each candidate were averaged, then the average was divided by the standard deviation for all candidates to yield a beauty index for each candidate. (Note: A 1-unit increase in the index represents a 1-standard-deviation increase in the beauty rating.) The relative success (measured as a percentage of votes obtained) of each candidate was used as the dependent variable (y) in a regression analysis. One of the independent variables in the model was beauty index (x).

1. Write the equation of a simple linear regression relating y to x.
2. Does the intercept of the equation (part a) have a practical interpretation? Explain.
3. The article reported the estimated slope of the equation (part a) as 22.91. Give a practical interpretation of this value.
4. The standard error of the slope estimate was reported as 3.73. Use this information and the estimate from part c to conduct a test for a positive slope α = 0.01. Give the appropriate conclusion in the words of the problem.

**11.5 The Coefficients of Correlation and Determination**

10. (69, BDAYS). ***Software milionaires and birthdays***. Refer to the study of the seemingly disproportionate number of software millionaires born around the year of 1955. The data are reproduced in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Decade | Total US births (millions) | N. of Software Millionaires Birthdays | N. of CEO Birthdays (in a random sample of 70 companies from the Fortune 500 list) |
| 1920 | 28582 | 3 | 2 |
| 1930 | 24374 | 1 | 2 |
| 1940 | 31666 | 10 | 23 |
| 1950 | 40530 | 14 | 38 |
| 1960 | 38808 | 7 | 9 |
| 1970 | 33309 | 4 | 0 |

1. Find the coefficient of determination for the simple linear regression model relating number (y) of software millionaire birthdays in a decade to total number (x) of US births. Interpret the result.
2. Find the coefficient of determination for the simple linear regression model relating number (y) of software millionaires birthdays in a decade to number (x) of CEO birthdays. Interpret the result.
3. The consulting statistician argued that the software industry appears to be no different from any other industry with respect to producing millionaires in a decade. Do you agree? Explain.

11. (73, TOPBUS). ***Survey on the top business schools***. Refer to the Wall Street Journal (Sep. 25, 2005) Business School Survey and the data saved in the file. Find and interpret r and r2 for the simple linear regression relating y = percentage of graduates with job offers and x = tuition cost and then fit the simple linear model.

**11.6 Using the model for estimation and prediction**

12. (86, PGA). ***Ranking driving performance of professional golfers***. Refer to The Sport Journal (Winter 2007) study of a new method for ranking the total driving performance of golfers on the PGA tour, Exercise 21. You fit a straight-line model relating driving accuracy (*y*) to driving distance (*x*) to the data saved in the file. Of interest is predicting *y* and estimating *E(y)* when *x* = 300 yards*.*

a) Find and interpret a 95% prediction interval for *y.*

b) Find and interpret a 95% confidence interval for *E(y).*

c) If you are interested in knowing the average driving accuracy of all PGA golfers who have driving distance of 300 yards, which of the intervals is relevant? Explain.

**11.7 A complete example**

13. (100, EBI). ***Buying income and households***. Sales and Marketing Management determined the “effective buying income” (EBI) of the average household in a state. Can the EBI be used to predict retail sales per household in the store-group category “eating and drinking places”?

a) Use the data for 13 states given in the table below to find the least squares line relating retail sales per household (*y*) to average household EBI (*x*).

b) Plot the least squares line, as well as the actual data points, on a scatterplot.

c) Based on the graph, part **b**, give your opinion regarding the predictive ability of the least squares line.

d) Find a 95% confidence interval for the slope of the line.

e) Use the results, part **d**, to assess the adequacy of the straight-line model.

|  |  |  |
| --- | --- | --- |
| State | Average household buying income ($) | Retail sales: Eating and drinking places ($ per household) |
| Connecticut | 60,998 | 2,553.8 |
| New Jersey | 63,853 | 2,154.8 |
| Michigan | 46,915 | 2,523.3 |
| Minnesota | 44,717 | 2,278.6 |
| Florida | 42,442 | 2,475.8 |
| South Carolina | 37,848 | 2,358.4 |
| Mississippi | 34,490 | 1,538.4 |
| Oklahoma | 34,830 | 2,063.1 |
| Texas | 44,729 | 2,363.5 |
| Colorado | 44,571 | 3,214.9 |
| Utah | 43,421 | 2,653.8 |
| California | 50,713 | 2,215.0 |
| Oregon | 40,597 | 2,144.0 |