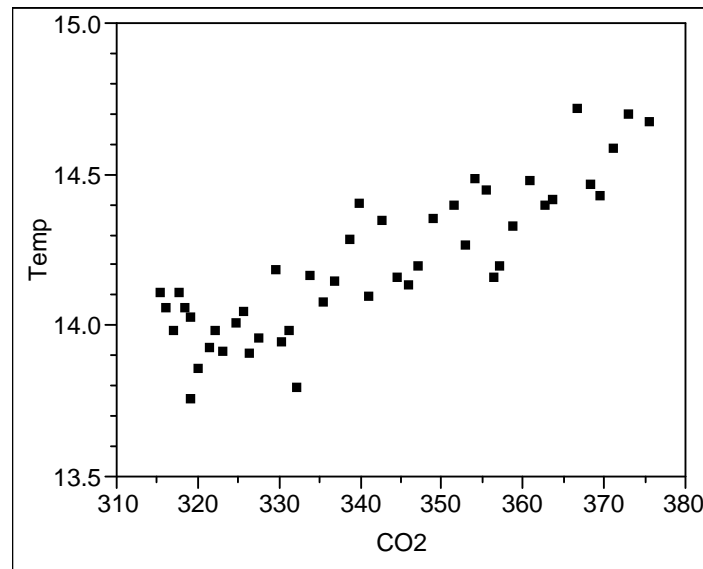


Statistics 101 – Homework 4 - Solution

1. Global warming is a hot topic these days. One of the factors that may explain increases in global temperatures is the amount of carbon dioxide in the atmosphere. Annual atmospheric carbon dioxide (CO₂) concentrations measured as parts per million by volume (ppmv) are derived from air samples collected at Mauna Loa Observatory in Hawaii. Additionally the annual average global temperatures (degrees Celsius) are recorded. The data plotted below are for years from 1958 through 2003.



- a) Answer the questions Who? and What? for this problem.

Who? Years from 1958 through 2003.

**What? Annual atmospheric CO₂ concentrations (ppmv)
Annual average global temperatures (degrees Celsius)**

- b) The value of the correlation for these data is most likely to be?
-0.95 -0.67 -0.31 +0.03 +0.55 +0.86

Explain the reason for your choice.

+0.86. The correlation has to be positive because as CO₂ increases, temperature tends to increase. There is a fairly strong linear relationship because the data points are clustered around the increasing trend.

- c) Write a sentence or two explaining what this correlation means for these data. Write about CO₂ concentrations and global temperatures rather than about correlation coefficients.

In general, above average CO₂ concentrations are associated with above average temperatures and below average CO₂ concentrations are associated with below average temperatures.

- d) If the temperature were recorded in degrees Fahrenheit rather than degrees Celsius, how would the correlation change? Explain your answer.

The value of the correlation would not change. Correlation is not affected by a change of scale.

- e) A news reporter looking at these data writes; “It is clear that increasing CO₂ in the atmosphere causes global temperature to increase.” Explain why this is not an appropriate conclusion from these data alone.

Although this is a tempting conclusion, remember correlation is not causation. We have simply observed the CO₂ concentration and temperature. There could be other explanations for the correlation. We need additional information and subject matter knowledge (about climate change) that is not provided in this data alone.

2. The October 2005 issue of Consumer Reports (CR) has an article on the difference between Environmental Protection Agency (EPA) mileage ratings (miles per gallon) and mileage results (miles per gallon) obtained by Consumer Reports in test-track driving. The EPA mileage ratings are the ones posted on new car sales stickers. Below are mileage values for 10 different vehicle types.

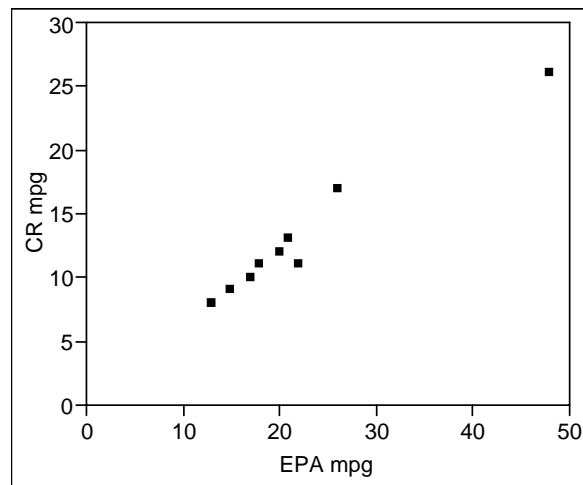
| | | | | | | | | | | |
|--------------|----|----|----|----|----|----|----|----|----|----|
| Vehicle type | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| EPA mileage | 15 | 21 | 13 | 48 | 22 | 17 | 20 | 13 | 26 | 18 |
| CR mileage | 9 | 13 | 8 | 26 | 11 | 10 | 12 | 8 | 17 | 11 |

- a) Answer the questions Who? and What? for this problem.

Who? Vehicle types.

What? EPA mileage (mpg), CR mileage (mpg).

- b) Plot the data.



- c) Compute the mean and standard deviation for the EPA mileage. Round final answers to 2 decimal places.

mean = 21.3 mpg, standard deviation = 10.24 mpg

- d) Compute the mean and standard deviation for the CR mileage. Round final answers to 2 decimal places.

mean = 12.5 mpg, standard deviation = 5.44 mpg

- e) The correlation between the EPA mileage and the CR mileage is $r=0.9839$. Explain in words what this correlation means.

There is a very strong positive linear relationship between the EPA mileage and the CR mileage.

- f) Compute the estimate of the slope for the least squares regression line. Round final answer to 4 decimal places.

$$b_1 = r \frac{s_y}{s_x} = 0.9839 \left(\frac{5.44}{10.24} \right) = 0.5227$$

- g) Give an interpretation of the estimated slope within the context of the problem.

For every one mile per gallon increase in the EPA mileage, the CR mileage increases, on average, only 0.5227 miles per gallon.

- h) Compute the estimate of the intercept for the least squares regression line. Round final answer to 4 decimal places.

$$b_0 = \bar{y} - b_1 \bar{x} = 12.5 - 0.5227(21.3) = 1.366 \text{ mpg}$$

- i) Give the equation of the least squares regression line. Use this equation to predict the Consumer Reports mileage for a vehicle the EPA rates at 30 miles per gallon.

$$\text{Predicted CR mileage} = 1.366 + 0.5227 * (\text{EPA mileage})$$

$$\text{Predicted CR mileage} = 1.366 + 0.5227 * (30) = 17.047 \text{ or } 17 \text{ mpg}$$

- j) How would you describe vehicle 4? Choose all that apply: outlier in regression, high leverage value, influential value. Explain your choice(s) briefly.

Vehicle 4 is not an outlier in regression because it does not have a large residual. Vehicle 4 is a high leverage value because it has a large value for the explanatory variable (EPA mileage). Vehicle 4 is an influential value because if removed the slope and intercept of the regression line would change quite a bit.

- k) Give the value of R^2 for this regression. Give an interpretation of this value within the context of the problem.

$R^2=(r)^2=(0.9839)^2=0.968$. 96.8% of the variation in CR mileage can be explained by the linear relationship with EPA mileage.

3. The December 2003 issue of Kiplinger's Personal Finance published data on the 2004 model year cars and trucks. The weight (pounds) and EPA estimated highway mileage (mpg) for 57 cars are displayed below. The data, Fuel Economy Data, are on the main Stat 101 course page (www.stat.iastate.edu/courses/stat101.html) under Homework 4. Follow the instructions in the JMP Guide to download/open the data set. Use JMP to look at the distribution of Highway mpg and the relationship between Weight and Highway mpg. Use the JMP output to help you answer the questions below. Be sure to attach the JMP output to your assignment.
- a) Describe the distribution of Highway mpg values. Make sure to include in your description the five number summary, the mean and standard deviation, and the shape of the histogram. Are there any outliers?

All summary measures have units of miles per gallon (mpg).

Five number summary:

minimum = 25, $Q_L = 27.5$, median = 29, $Q_U = 31$, maximum = 40

mean = 29.86, standard deviation = 3.36

The shape of the histogram is mound in the high 20's. There is also a skew to the right.

All Highway values of 37 mpg and above could be outliers.

- b) Describe the scatterplot of Highway mpg versus Weight. Give the regression equation for predicting Highway mpg from Weight. Use the regression equation to predict the Highway mpg for a car weighing 3000 pounds. Give an interpretation of the estimated slope. Give the value of R^2 and an interpretation of this value. Finally, describe the plot of residuals versus weight values and make note of any potential problems with the regression.

The scatterplot shows a moderately strong negative linear association. Above average values for Highway mpg correspond to below average values for weight.

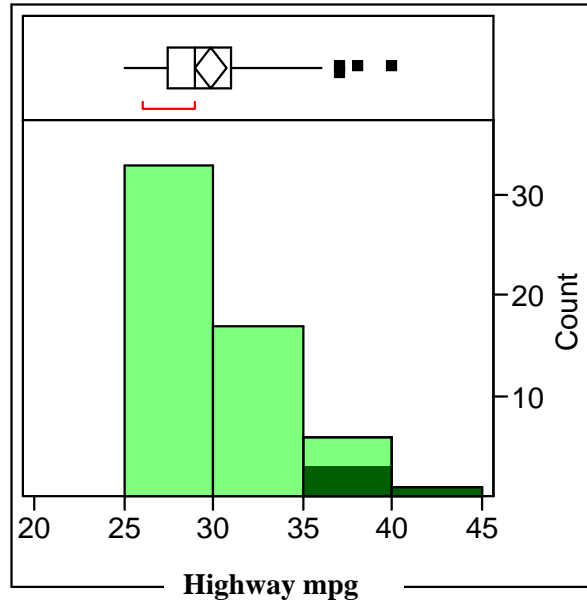
Predicted Highway mpg = $50.91 - 0.00657 * \text{Weight}$

Predicted Highway mpg = $50.91 - 0.00657 * 3000 = 50.91 - 19.71 = 31.2$ mpg

For each additional pound of Weight, the Highway mpg decreases, on average, by 0.00657 miles per gallon. Or for each additional 1000 pounds of weight, the Highway mpg decreases, on average, by 6.57 miles per gallon.

$R^2 = 0.517$. 51.7% of the variation in Highway mpg can be explained by the linear relationship with Weight.

The plot of residuals versus Weight appears to be fairly random. This would mean that a linear model is about the best we can do. Note: some may see the diagonal lines, lower left to upper right, in the residuals. This is a result of cars with different weights having the same Highway mpg. A more precise measurement of Highway mpg may get rid of this pattern.

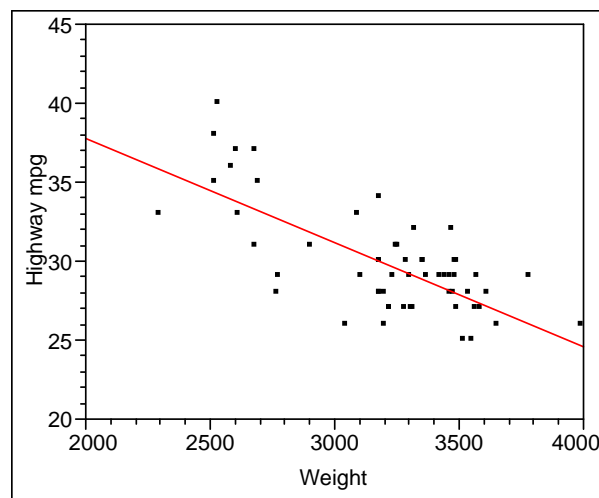


Quantiles

Moments

| | | | | |
|--------|----------|--------|---------|-----------|
| 100.0% | maximum | 40.000 | Mean | 29.859649 |
| 75.0% | quartile | 31.000 | Std Dev | 3.3564362 |
| 50.0% | median | 29.000 | N | 57 |
| 25.0% | quartile | 27.500 | | |
| 0.0% | minimum | 25.000 | | |

Bivariate Fit of Highway mpg By Weight



Linear Fit

$$\text{Highway mpg} = 50.911039 - 0.0065723 \text{ Weight}$$

Summary of Fit

| | |
|----------------------------|----------|
| RSquare | 0.51737 |
| RSquare Adj | 0.508595 |
| Root Mean Square Error | 2.352872 |
| Mean of Response | 29.85965 |
| Observations (or Sum Wgts) | 57 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Ratio |
|----------|----|----------------|-------------|----------|
| Model | 1 | 326.39685 | 326.397 | 58.9589 |
| Error | 55 | 304.48034 | 5.536 | Prob > F |
| C. Total | 56 | 630.87719 | | <.0001 |

Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob> t |
|-----------|-----------|-----------|---------|---------|
| Intercept | 50.911039 | 2.759268 | 18.45 | <.0001 |
| Weight | -0.006572 | 0.000856 | -7.68 | <.0001 |

