CS555, Data Analysis and Visualization Homework 3

Jefferson Parker, [japarker@bu.edu](mailto:japarker@bu.edu)

See the accompanying R code at the end of this document.

1. The data were saved to a tab delimited text file and read into R using the read.table() function.
2. The data was plotted as below, with the number of meals per week containing fish as the explanatory (x) variable and mercury measured in the hair as the dependent variable (y). This arrangement was chosen as it is more likely that eating fish (contaminated with mercury) leads to increased levels of mercury in the blood, which would be deposited in the hair where it is easily measured.



As can be seen, the data is generally linear with a positive association as values of both variables are increasing together. However, the dependent measure of mercury in hair is rather scattered, especially at values of 0 to 10 fish meals per week, so the association is not particularly strong.

Note: there is a potential outlier value at ~21 fish meals per week and 2 mg/g of hair.

1. The correlation coefficient, r, is 0.7 (0.6991094). This confirms our visual evaluation of positive association between the variables and supports the existence of a linear relationship, however we would need to perform a more rigorous test to assert the existence of an actual linear relationship.
2. The regression line for the data is y = 1.688 + 0.276x. Overlaying the line on the plot, we have:



1. The estimate of the slope, 1, is 0.276. This means that, on average, the amount of mercury measured in the head hair of a fisherman increases 0.276 mg/g for an increase of one meal weekly fish containing meal. The estimate of the intercept, 0, is 1.688. This value indicates the average amount of mercury, in mg/g, in fishermen’s hair without eating any meals containing fish. Put another way, 1.688 mg/g is the baseline level of mercury in fishermen’s head hair.
2. The ANOVA table for this regression line is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SS (Sum of Squares) | df (degrees of freedom) | MS (Mean Square) | F-statistic | p-value |
| Regression | 309.24 | 1 | 309.239 | 93.689 | 6.01E-16 |
| Residual | 323.47 | 98 | 3.301 |  |  |
| Total | 632.71 |  |  |  |  |

The table to calculate the Standard Error of 1(hat) is at the end of this document, before the supporting R code.

1. Our hypothesis is that there is no linear association between eating fish and mercury in fishermen’s hair.

H0: 1 = 0

H1: 1 ≠ 0

1. Our test statistic is the F statistic with 1 and 98 degrees of freedom and confidence level  = 0.10 (F1,98,0.1)
2. Our critical value of F is calculated as F1,98,0.1 = 2.75743. Our decision rule is
   1. Reject H0 if F ≥ 2.75743
   2. Otherwise do not reject H0
3. The F-statistic (computed in the ANOVA table above) is 93.689.
4. Given the result that our calculated F statistic is greater than our critical F statistic value, we can reject the null hypothesis and state that there is significant evidence of a linear relationship between the number of meals containing fish, eaten per week, by fishermen and the amount of mercury measured in their head hair.

While the test value supports a linear relationship, the R2 value is only 0.49 which indicates only about half of the observed mercury in measured in fishermen’s hair is associated with eating meals containing fish. This is reflective of the broad scatter observed in the original scatter plots mentioned in question 1. The middling value of R2 = 0.49 would suggest that there are other sources of mercury entering fishermen beyond fish in their diet.

The 90% confidence interval of 1hat is 0.24 – 0.31, which means that the regression line of any given sample relating mercury in fishermen’s hair per weekly meal containing fish will be within this range 90% of the time.

**Table to calculate SE(1hat)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| x | y | yhat | y.min.yhat | y.min.yhat2 | x.min.xbar | x.min.xbar2 |
| 14 | 4.484 | 5.550947 | -1.06694661 | 1.14E+00 | 5.7 | 32.49 |
| 7 | 4.789 | 3.619295 | 1.16970537 | 1.37E+00 | -1.3 | 1.69 |
| 5 | 3.856 | 3.067394 | 0.78860593 | 6.22E-01 | -3.3 | 10.89 |
| 8 | 4.888 | 3.895245 | 0.99275508 | 9.86E-01 | -0.3 | 0.09 |
| 21 | 10.849 | 7.482599 | 3.3664014 | 1.13E+01 | 12.7 | 161.29 |
| 18 | 6.457 | 6.654748 | -0.19774775 | 3.91E-02 | 9.7 | 94.09 |
| 22 | 11.222 | 7.758549 | 3.46345112 | 1.20E+01 | 13.7 | 187.69 |
| 6 | 4.908 | 3.343344 | 1.56465565 | 2.45E+00 | -2.3 | 5.29 |
| 19 | 10.116 | 6.930698 | 3.18530197 | 1.01E+01 | 10.7 | 114.49 |
| 7 | 3.567 | 3.619295 | -0.05229463 | 2.73E-03 | -1.3 | 1.69 |
| 16 | 6.092 | 6.102847 | -0.01084718 | 1.18E-04 | 7.7 | 59.29 |
| 17 | 3.799 | 6.378797 | -2.57979746 | 6.66E+00 | 8.7 | 75.69 |
| 20 | 6.781 | 7.206648 | -0.42564831 | 1.81E-01 | 11.7 | 136.89 |
| 5 | 5.995 | 3.067394 | 2.92760593 | 8.57E+00 | -3.3 | 10.89 |
| 7 | 1.717 | 3.619295 | -1.90229463 | 3.62E+00 | -1.3 | 1.69 |
| 14 | 4.615 | 5.550947 | -0.93594661 | 8.76E-01 | 5.7 | 32.49 |
| 1 | 3.362 | 1.963593 | 1.39840707 | 1.96E+00 | -7.3 | 53.29 |
| 6 | 3.928 | 3.343344 | 0.58465565 | 3.42E-01 | -2.3 | 5.29 |
| 9 | 1.833 | 4.171195 | -2.3381952 | 5.47E+00 | 0.7 | 0.49 |
| 10 | 5.668 | 4.447145 | 1.22085452 | 1.49E+00 | 1.7 | 2.89 |
| 13 | 4.7 | 5.274996 | -0.57499633 | 3.31E-01 | 4.7 | 22.09 |
| 9 | 2.272 | 4.171195 | -1.8991952 | 3.61E+00 | 0.7 | 0.49 |
| 16 | 4.812 | 6.102847 | -1.29084718 | 1.67E+00 | 7.7 | 59.29 |
| 5 | 1.342 | 3.067394 | -1.72539407 | 2.98E+00 | -3.3 | 10.89 |
| 18 | 6.123 | 6.654748 | -0.53174775 | 2.83E-01 | 9.7 | 94.09 |
| 7 | 4.622 | 3.619295 | 1.00270537 | 1.01E+00 | -1.3 | 1.69 |
| 8 | 7.805 | 3.895245 | 3.90975508 | 1.53E+01 | -0.3 | 0.09 |
| 7 | 2.643 | 3.619295 | -0.97629463 | 9.53E-01 | -1.3 | 1.69 |
| 8 | 6.111 | 3.895245 | 2.21575508 | 4.91E+00 | -0.3 | 0.09 |
| 7 | 2.476 | 3.619295 | -1.14329463 | 1.31E+00 | -1.3 | 1.69 |
| 10 | 4.317 | 4.447145 | -0.13014548 | 1.69E-02 | 1.7 | 2.89 |
| 4 | 1.789 | 2.791444 | -1.00244378 | 1.00E+00 | -4.3 | 18.49 |
| 4 | 2.484 | 2.791444 | -0.30744378 | 9.45E-02 | -4.3 | 18.49 |
| 7 | 1.757 | 3.619295 | -1.86229463 | 3.47E+00 | -1.3 | 1.69 |
| 6 | 1.239 | 3.343344 | -2.10434435 | 4.43E+00 | -2.3 | 5.29 |
| 5 | 5.311 | 3.067394 | 2.24360593 | 5.03E+00 | -3.3 | 10.89 |
| 19 | 6.103 | 6.930698 | -0.82769803 | 6.85E-01 | 10.7 | 114.49 |
| 3 | 1.984 | 2.515493 | -0.5314935 | 2.82E-01 | -5.3 | 28.09 |
| 4 | 2.697 | 2.791444 | -0.09444378 | 8.92E-03 | -4.3 | 18.49 |
| 7 | 0.692 | 3.619295 | -2.92729463 | 8.57E+00 | -1.3 | 1.69 |
| 7 | 2.404 | 3.619295 | -1.21529463 | 1.48E+00 | -1.3 | 1.69 |
| 9 | 1.503 | 4.171195 | -2.6681952 | 7.12E+00 | 0.7 | 0.49 |
| 17 | 8.231 | 6.378797 | 1.85220254 | 3.43E+00 | 8.7 | 75.69 |
| 14 | 5.321 | 5.550947 | -0.22994661 | 5.29E-02 | 5.7 | 32.49 |
| 7 | 3.81 | 3.619295 | 0.19070537 | 3.64E-02 | -1.3 | 1.69 |
| 21 | 1.765 | 7.482599 | -5.7175986 | 3.27E+01 | 12.7 | 161.29 |
| 4 | 0.408 | 2.791444 | -2.38344378 | 5.68E+00 | -4.3 | 18.49 |
| 7 | 3.901 | 3.619295 | 0.28170537 | 7.94E-02 | -1.3 | 1.69 |
| 10 | 0.48 | 4.447145 | -3.96714548 | 1.57E+01 | 1.7 | 2.89 |
| 11 | 3.826 | 4.723096 | -0.89709576 | 8.05E-01 | 2.7 | 7.29 |
| 7 | 3.451 | 3.619295 | -0.16829463 | 2.83E-02 | -1.3 | 1.69 |
| 9 | 2.32 | 4.171195 | -1.8511952 | 3.43E+00 | 0.7 | 0.49 |
| 2 | 4.086 | 2.239543 | 1.84645678 | 3.41E+00 | -6.3 | 39.69 |
| 7 | 2.272 | 3.619295 | -1.34729463 | 1.82E+00 | -1.3 | 1.69 |
| 3 | 2.564 | 2.515493 | 0.0485065 | 2.35E-03 | -5.3 | 28.09 |
| 7 | 7.998 | 3.619295 | 4.37870537 | 1.92E+01 | -1.3 | 1.69 |
| 11 | 5.081 | 4.723096 | 0.35790424 | 1.28E-01 | 2.7 | 7.29 |
| 8 | 0.366 | 3.895245 | -3.52924492 | 1.25E+01 | -0.3 | 0.09 |
| 7 | 2.477 | 3.619295 | -1.14229463 | 1.30E+00 | -1.3 | 1.69 |
| 4 | 5.288 | 2.791444 | 2.49655622 | 6.23E+00 | -4.3 | 18.49 |
| 7 | 5.676 | 3.619295 | 2.05670537 | 4.23E+00 | -1.3 | 1.69 |
| 7 | 2.296 | 3.619295 | -1.32329463 | 1.75E+00 | -1.3 | 1.69 |
| 21 | 6.11 | 7.482599 | -1.3725986 | 1.88E+00 | 12.7 | 161.29 |
| 4 | 1.502 | 2.791444 | -1.28944378 | 1.66E+00 | -4.3 | 18.49 |
| 7 | 3.71 | 3.619295 | 0.09070537 | 8.23E-03 | -1.3 | 1.69 |
| 3 | 2.752 | 2.515493 | 0.2365065 | 5.59E-02 | -5.3 | 28.09 |
| 3 | 0.987 | 2.515493 | -1.5284935 | 2.34E+00 | -5.3 | 28.09 |
| 19 | 10.14 | 6.930698 | 3.20930197 | 1.03E+01 | 10.7 | 114.49 |
| 7 | 1.616 | 3.619295 | -2.00329463 | 4.01E+00 | -1.3 | 1.69 |
| 12 | 4.65 | 4.999046 | -0.34904605 | 1.22E-01 | 3.7 | 13.69 |
| 13 | 7.241 | 5.274996 | 1.96600367 | 3.87E+00 | 4.7 | 22.09 |
| 18 | 9.36 | 6.654748 | 2.70525225 | 7.32E+00 | 9.7 | 94.09 |
| 7 | 3.753 | 3.619295 | 0.13370537 | 1.79E-02 | -1.3 | 1.69 |
| 13 | 4.008 | 5.274996 | -1.26699633 | 1.61E+00 | 4.7 | 22.09 |
| 21 | 5.345 | 7.482599 | -2.1375986 | 4.57E+00 | 12.7 | 161.29 |
| 1 | 2.455 | 1.963593 | 0.49140707 | 2.41E-01 | -7.3 | 53.29 |
| 0 | 0.941 | 1.687643 | -0.74664265 | 5.57E-01 | -8.3 | 68.89 |
| 1 | 2.478 | 1.963593 | 0.51440707 | 2.65E-01 | -7.3 | 53.29 |
| 1 | 3.212 | 1.963593 | 1.24840707 | 1.56E+00 | -7.3 | 53.29 |
| 10 | 5.214 | 4.447145 | 0.76685452 | 5.88E-01 | 1.7 | 2.89 |
| 0 | 1.12 | 1.687643 | -0.56764265 | 3.22E-01 | -8.3 | 68.89 |
| 0 | 0.745 | 1.687643 | -0.94264265 | 8.89E-01 | -8.3 | 68.89 |
| 2 | 4.645 | 2.239543 | 2.40545678 | 5.79E+00 | -6.3 | 39.69 |
| 2 | 4.981 | 2.239543 | 2.74145678 | 7.52E+00 | -6.3 | 39.69 |
| 1 | 2.812 | 1.963593 | 0.84840707 | 7.20E-01 | -7.3 | 53.29 |
| 0 | 0.846 | 1.687643 | -0.84164265 | 7.08E-01 | -8.3 | 68.89 |
| 2 | 5.142 | 2.239543 | 2.90245678 | 8.42E+00 | -6.3 | 39.69 |
| 0 | 1.111 | 1.687643 | -0.57664265 | 3.33E-01 | -8.3 | 68.89 |
| 0 | 1.094 | 1.687643 | -0.59364265 | 3.52E-01 | -8.3 | 68.89 |
| 2 | 2.978 | 2.239543 | 0.73845678 | 5.45E-01 | -6.3 | 39.69 |
| 2 | 3.942 | 2.239543 | 1.70245678 | 2.90E+00 | -6.3 | 39.69 |
| 0 | 1.131 | 1.687643 | -0.55664265 | 3.10E-01 | -8.3 | 68.89 |
| 0 | 0.979 | 1.687643 | -0.70864265 | 5.02E-01 | -8.3 | 68.89 |
| 0 | 0.844 | 1.687643 | -0.84364265 | 7.12E-01 | -8.3 | 68.89 |
| 1 | 2.411 | 1.963593 | 0.44740707 | 2.00E-01 | -7.3 | 53.29 |
| 1 | 2.497 | 1.963593 | 0.53340707 | 2.85E-01 | -7.3 | 53.29 |
| 10 | 3.764 | 4.447145 | -0.68314548 | 4.67E-01 | 1.7 | 2.89 |
| 20 | 8.178 | 7.206648 | 0.97135169 | 9.44E-01 | 11.7 | 136.89 |
| 19 | 7.664 | 6.930698 | 0.73330197 | 5.38E-01 | 10.7 | 114.49 |
| 22 | 9.716 | 7.758549 | 1.95745112 | 3.83E+00 | 13.7 | 187.69 |

**Supporting R Code:**

# CS555 Data Analysis and Visualization

# Homework3.R

# Jefferson Parker, japarker@bu.edu

# 20180723

# 1. Save the data to a local file and read into R.

inputDir <- "C:/Users/jparker/Code/Input";

setwd(inputDir);

fishmerc <- read.table(file = "fishmeal\_mercury\_data.txt", header = TRUE);

# 2. Genereate a scatterplot (labels and title).

# Use the plot to describe the fomr, direction and strength of any association.

plot(x = fishmerc$Count.FishMeals,

y = fishmerc$Mercury.mgPerGram,

pch = 20,

xlab = "Meals Containing Fish (weekly)",

ylab = "Mercury in Head Hair (mg/g)",

main = "Mercury Present in Fisherman's Hair"

);

# 3. Calculate the correlation coefficient.

r.fish <- cor(fishmerc$Count.FishMeals, fishmerc$Mercury.mgPerGram);

r.fish;

# 4. Find the least square regression line.

# Write the equation for the line and add it to the plot.

regLine <- lm(fishmerc$Mercury.mgPerGram ~ fishmerc$Count.FishMeals);

regLine;

abline(regLine);

# 5. What are the estimates for beta1 and beta0.

regLine$coefficients;

# 6. Calculate the ANOVA table.

# Formally test beta1 = 0 by F-test or t-test at alpha = 0.10.

anova(regLine);

#summary(aov(fishmerc$Mercury.mgPerGram ~ fishmerc$Count.FishMeals));

# While anova(lm(formula)) and (aov(formula)) generate the ANOVA table in different

# ways, the results are within rounding differences identical.

qf(1 - 0.1, 1, nrow(fishmerc)-2);

# Calculate the table for standard error of b1hat.

betaTable <- fishmerc;

colnames(betaTable) <- c("x", "y");

# Calculate the fitted value yhat with the regression coefficients.

betaTable$yhat <- regLine$coefficients[1] + (regLine$coefficients[2] \* betaTable$x);

# Calculate the terms of the SE of beta1(hat)

betaTable$y.min.yhat <- betaTable$y - betaTable$yhat;

betaTable$y.min.yhat2 <- betaTable$y.min.yhat^2;

betaTable$x.min.xbar <- betaTable$x - mean(betaTable$x);

betaTable$x.min.xbar2 <- betaTable$x.min.xbar^2;

betaTable;

# Also calculate R^2.

r2.fish <- r.fish^2;

r2.fish;

# Also calculate the 90% confidence interval for beta1.

# First, calculate the standard error for beta1hat.

sebeta1hat.top <- sqrt(sum(betaTable$y.min.yhat2/(nrow(betaTable)-2)));

sebeta1hat.bot <- sqrt(sum(betaTable$x.min.xbar2));

se.beta1hat <- sebeta1hat.top/sebeta1hat.bot;

# Next, find our 2-sided t-statistic

t.confint <- qt(0.90, df = nrow(betaTable) - 2);

# Then, calculate the bounds of the confidence interval.

regLine$coefficients[2] + (t.confint \* se.beta1hat);

regLine$coefficients[2] - (t.confint \* se.beta1hat);