CS555, Data Analysis and Visualization Homework 4

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 comma separated value (.csv) file and read into R using the read.csv() function. Column names were changed for convenience.
2. The scatterplot of Prestige Score and Years of education is shown below (Figure 1).



Figure 1: Scatterplot of Prestige Score and Years of Education

In the plot, we can observe that the variables have a positive linear association; however it could be argued that the association is curved with two segments from 6 – 12 years and the second from 12 – 16 years.

The strength of the association could be described as increasing with Years of Education, with less scatter among the data points above 12 years, but this is likely just an artifact of there being fewer observations above 12 years.

The correlation of the two variables is r = 0.85 (0.8501769). Given this value, the coefficient of determination (R2 ~ 0.72) tells us that the model explains just over 70% of the variation in the data, which is consistent with our observation with fewer data points at the higher end of the education range and more scatter at the lower end.

1. A simple linear regression was performed according to the plot presented in Question 2. The plot of residuals against the Years of Education is below (Figure 2).

To assess the fit of the model with the necessary assumptions for valid linear regression I evaluated the following:

* The data appear to have a linear association, as detailed in question 2 (Figure 1).
* We can only assume that the observations are independent, as we do not have any detail of the data collection methods.

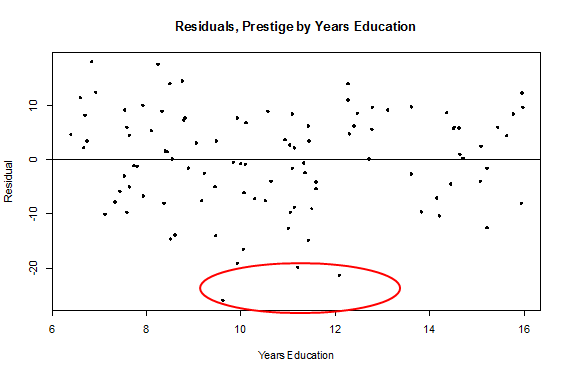


Figure 2: Residuals Plot of linear model relating Prestige Score and Years of Education



Figure 3: Histogram of Residuals of the linear model relating Prestige Score and Years of Education.

* A histogram of the residual values (Figure 3) was generated to determine normality of the residuals. The graph indicates that the residual values are somewhat left skewed, with a mean closer to five than to zero.
* Finally, we tested for the presence of outlier values or influence points. Visually, there are three points on the residual plot (circled in red, Figure 2) that are outside of the expected range of residuals. The residual values have an IQR of -6.5 to 6.7, and a maximum value of just over 18. Ideally the residuals should be evenly distributed in the vertical along the regression line, but the lowest residual value is -26.

Using the R influence.measures() function, I identified three influential observations in this regression model (Table 1). The R2 value for the full data model is 0.72. By removing observation any one of these observations we get a 1% improvement in fit (R2 = 0.73) between the model and the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Row Number | Job Title | Years Education | Income | Percentage Women | Prestige Score |
| 41 | FILE\_CLERKS | 12.09 | 3016 | 83.19 | 32.7 |
| 46 | COLLECTORS | 11.2 | 4741 | 47.06 | 29.4 |
| 53 | NEWSBOYS | 9.62 | 918 | 7 | 14.8 |

Table 1: Influencer observations in the Prestige Score vs Years Education regression model.

1. Calculating the least squares regression line to predict prestige from education, income and percentage of women gives us the formula:

To test the significance of this model:

* Our hypothesis is that all of the slope variables are zero:

H0: education = income = percentwomen = 0

H1: education and/or income and/or percentwomen ≠ 0

At confidence  = 0.05

* Our test statistic is F = RegMS/ResMS
* Our decision rule is if F < Fk,n-k-1, we reject the null hypothesis, otherwise we fail to reject. Alternatively, if P() <  we reject the null hypothesis.
* Using the anova() function to extract the F statistic and p-values for significance of these variables in predicting Prestige we find:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| Years Education | 1 | 21608.4 | 21608.4 | 350.9741 | < 2.2e-16 |
| Income | 1 | 2248.1 | 2248.1 | 36.5153 | 2.74E-08 |
| Percentage Women | 1 | 5.3 | 5.3 | 0.0858 | 0.7702 |
| Residuals | 98 | 6033.6 | 61.6 |  |  |

Table 2: ANOVA table for prediction of Prestige from Education, Income and Percentage women in an occupation.

* Given the high F values and low p-values, both the Years of Education and Income terms are found to be significant, so we can reject the null hypothesis.

1. The model was found to have significant  terms (Table 2). The strongest contributor to the Prestige score to the linear regression model in Question 4 is Years of Education (p < 2.2-16). The R2 value for this variable is 0.91, meaning that 91% of the predicted Prestige Score is based on Years of Education. The second strongest contributor to Prestige score is income. While the  coefficient is only 0.001, meaning Prestige increases 1/10th of 1 unit per dollar income, the R2 value for income is 0.09 indicating that a full 9% of the regression model is explained by that variable.

The 95% confidence interval for the slope term () for Years of Education is 0 to 8.3, while the 95% confidence interval the slope term () for income is 0 to 0.003 dollars.

1. The residuals plot for fitted values against the residuals values for the regression model of question 4 is shown below (Figure 4).



Figure 4: Plot of Residuals vs Fitted Values for the regression model of Prestige vs Education, Income and Percentage Women.

The plot *appears* to have a curvilinear association, with downward slope from 0 to 45, upward slope from 45 to 55, and downward slope from 55 to 90. However, it is kind of late and this could just be my brain making things up. However, the residuals do taper with a larger vertical range at the lowest fitted values, to a more narrow vertical range at higher fitted values, suggesting there is something not proper within the data for a regression analysis.

1. There are several influencer points in the latter regression model, including two of the influencers in the regression from question 1 ().

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Row Number | Job Title | Years Education | Income | Percentage Women | Prestige Score |
| 2 | GENERAL\_MANAGERS | 12.26 | 25879 | 4.02 | 69.1 |
| 17 | LAWYERS | 15.77 | 19263 | 5.13 | 82.3 |
| 24 | PHYSICIANS | 15.96 | 25308 | 10.56 | 87.2 |
| 46 | COLLECTORS | 11.2 | 4741 | 47.06 | 29.4 |
| 53 | NEWSBOYS | 9.62 | 918 | 7 | 14.8 |
| 67 | FARMERS | 6.84 | 3643 | 3.6 | 44.1 |
| 84 | SEWING\_MACH\_OPERATORS | 6.38 | 2847 | 90.67 | 28.2 |

Table 3: Influencer points in the regression model of Prestige Score, Years Education, Income and Percentage Women.

**Supporting R Code:**

# CS555 Data Analysis and Visualization

# Homework4.R

# Jefferson Parker, japarker.bu.edu

# 20180728

# 1. Save the data to .csv file and read into R for analysis.

setwd('C:/Users/jparker/Code/Input/');

canjobdata <- read.csv(file = "Canada\_JobData.csv", header = TRUE);

# Remove extra periods from colname(s)

colnames(canjobdata) <- c("Job.Title", "Years.Education", "Income", "Perc.Women", "Prestige.Score");

# 2. Plot the assocation between Years of Education and Prestige Score

# Calculate the Correlation

plot(x = canjobdata$Years.Education,

y = canjobdata$Prestige.Score,

main = "Prestige by Years Education",

xlab = "Years of Education",

ylab = "Prestige Score",

pch = 20

);

cor(canjobdata$Years.Education, canjobdata$Prestige.Score);

# 3. Perform a simple linear regression.

# Generate a residual plot.

#

# Determine if the model assumptions were met.

# 1. The relationship is linear without major outliers.

# 2. The observations are INDEPENDENT.

# 3. The variation of the response around the regression line is CONSTANT.

# 4. The residuals are NORMALLY DISTRIBUTED.

job.model <- lm(Prestige.Score ~ Years.Education, data = canjobdata);

job.resid <- resid(job.model);

plot(x = canjobdata$Years.Education,

y = job.resid,

xlab = "Years Education",

ylab = "Residual",

main = "Residuals, Prestige by Years Education",

pch = 20

);

abline(h = 0);

hist(job.resid);

# Are there any outliers or influence points?

# If so, identify them by ID and comment on their effect.

#

# The influence.measures() function returns three lists.

# The second ($is.inf) contains six logical vectors of values,

# within this, the cov.r column marks the influencer values.

job.influence <- data.frame(influence.measures(job.model)$is.inf);

canjobdata[which(job.influence$cov.r),];

summary(job.model);

summary(lm(Prestige.Score ~ Years.Education, data = canjobdata[-41, ]));

summary(lm(Prestige.Score ~ Years.Education, data = canjobdata[-46, ]));

summary(lm(Prestige.Score ~ Years.Education, data = canjobdata[-53, ]));

# 4. Calculate the Least Squares Regression model to predicte Prestige

# from Education, Income and Percentage of Women.

prestige.model <- lm(Prestige.Score ~ Years.Education + Income + Perc.Women, data = canjobdata);

summary(prestige.model);

# Formally test whether these predictors are associated with Prestige

# at alpha = 0.05.

anova(prestige.model);

# 5. Calculate the 95% Confidence intervals for significant variables.

# Reference: https://stat.ethz.ch/pipermail/r-help/2008-April/160538.html

# How to extract coefficient standard errors from a linear model.

prestige.coef <- data.frame(coef(summary(prestige.model)));

rownames(prestige.coef);

prestige.coef$upper.95 <- prestige.coef$Estimate + (prestige.coef$t.value \* prestige.coef$Std..Error);

prestige.coef$lower.95 <- prestige.coef$Estimate - (prestige.coef$t.value \* prestige.coef$Std..Error);

prestige.coef;

# 6. Generate a residuals plot showing the fitted values from teh regression against the residuals.

plot(x = prestige.model$fitted.values,

y = prestige.model$residuals,

xlab = "Fitted values",

ylab = "Residuals",

main = "Prestige Regression, Residuals vs Fits",

pch = 20

);

abline(h = 0);

# 7. Are there any outliers or influence points?

prestige.influence <- data.frame(influence.measures(prestige.model)$is.inf);

canjobdata[which(prestige.influence$cov.r),];