Homework 2 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Thursday, September 20, 2001

1. Learn about variables in the Prestige.dat data set.

   In 1971, the Census of Canada collected information regarding the occupational prestige of various occupations throughout Canada. The data collected consisted of seven variables which are described below.

   (a) Occupational Title

   This variable includes the occupational title for each of the 102 occupations included in the study. The titles were:

   GOV ADMINISTRATORS, GENERAL MANAGERS, ACCOUNTANTS, PURCHASING OFFICERS, CHEMISTS, PHYSICISTS, BIOLOGISTS, ARCHITECTS, CIVIL ENGINEERS, MINING ENGINEERS, SURVEYORS, DRAUGHTSMEN, COMPUTER PROGRAMERS, ECONOMISTS, PSYCHOLOGISTS, SOCIAL WORKERS, LAWYERS, LIBRARIANS, VOCATIONAL COUNSELLORS, MINISTERS, UNIVERSITY TEACHERS, PRIMARY SCHOOL TEACHERS, SECONDARY SCHOOL TEACHERS, PHYSICIANS, VETERINARIANS, OSTEOPATHS CHIROPRACTOR, NURSES, NURSING AIDES, PHYSIO THERAPSTS, PHARMACISTS, MEDICAL TECHNICIANS, COMMERCIAL ARTISTS, RADIO TV ANNOUNCERS, ATHLETES, SECRETARIES, TYPISTS, BOOKKEEPERS, TELLERS CASHIERS, COMPUTER OPERATORS, SHIPPING CLERKS, FILE CLERKS, RECEPTIONSTS, MAIL CARRIERS, POSTAL CLERKS, TELEPHONE OPERATORS, COLLECTORS, CLAIM ADJUSTORS, TRAVEL CLERKS, OFFICE CLERKS, SALES SUPERVISORS, COMMERCIAL TRAVELLERS, SALES CLERKS, NEWSBOYS, SERVICE STATION ATTENDENTS, INSURANCE AGENTS, REAL ESTATE SALESMEN, BUYERS, FIREFIGHTERS, POLICEMEN, COOKS, BARTENDERS, FUNERAL DIRECTORS, BABYSITTERS, LAUNDERERS, JANITORS, ELEVATOR OPERATORS, FARMERS, FARM WORKERS, ROTARY WELL DRILLERS, BAKERS, SLAUGHTERERS 1, SLAUGHTERERS 2, CANNERS, TEXTILE WEAVERS, TEXTILE LABOURERS, TOOL DIE MAKERS, MACHINISTS, SHEET METAL WORKERS, WELDERS, AUTO WORKERS, AIRCRAFT WORKERS, ELECTRONIC WORKERS, RADIO TV REPAIRMEN, SEWING MACH OPERATORS, AUTO REPAIRMEN, AIRCRAFT REPAIRMEN, RAILWAY SECTIONMEN, ELECTRICAL LINEMEN, ELECTRICIANS, CONSTRUCTION FOREMEN, CARPENTERS, MASONS, HOUSE PAINTERS, PLUMBERS, CONSTRUCTION LABOURERS, PILOTS, TRAIN ENGINEERS, BUS DRIVERS, TAXI DRIVERS, LONGSHOREMEN, TYPESETTERS, BOOKBINDERS.

   (b) Average Education of Incumbents (in years)

This variable contains the average educational level across the different occupations. Table 1 contains the numerical summary of the average years of education. The minimum average education level was 6.38 years for sewing machine operators and the maximum average education level was 15.97 years for University Teachers. The mean of 10.74 years is slightly higher than the median of 10.54 years.

Figure 1 contains a histogram of the average education figures. As seen in the histogram, the data does not appear to follow any standard distribution. The most frequent education level is between 11 and 12 years of education.

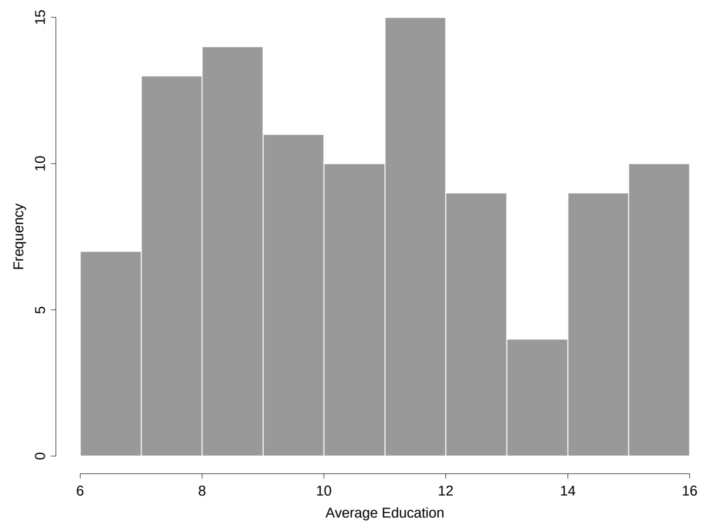

Figure 1: Average Age of Incumbents

(c) Average Income of Incumbents (in dollars)

This variable contains the average income of all the people working in each given profession. Table 1 contains the numerical summary of the incomes. The minimum salary was \$611 for babysitters, and the maximum salary was \$25,880 for General Managers. The mean of \$6,738 is higher than the median of \$5,930 indicating a positive skew.

Figure 2 contains a histogram of the average income of the incumbents. The histogram shows that the distribution is skewed to the right, with a number of high outlying observations. The most frequent income level is about \$5,000.

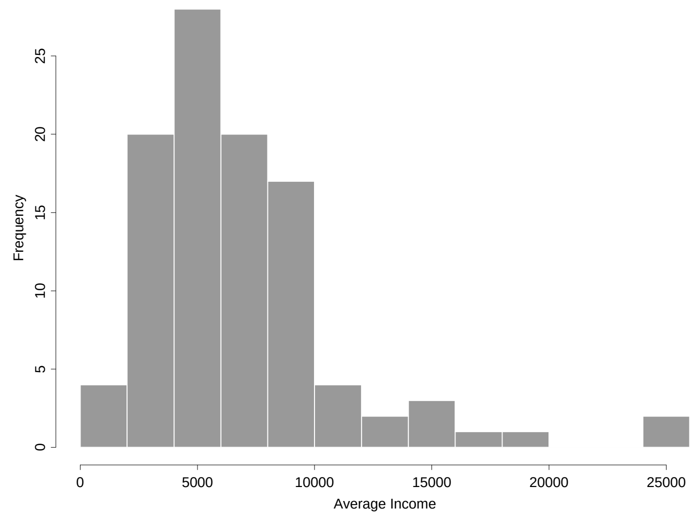

Figure 2: Average Income of Incumbents

(d) Percent of Incumbents who are Women

This variable contains the percent of incumbents who are women in each field. Table 1 contains the numerical summary. The minimum (0%) was obtained by five different occupations: Firefighters, rotary well drillers, railway sectionmen, train engineers and longshoremen. The maximum (97.51%) was in the secretary occupation. The mean of 46.83% is higher than the median of 43.6% indicating a positive skew.

Figure 3 contains a boxplot of the percentage of women across all occupations. The boxplot indicates that the distribution is highly skewed right.

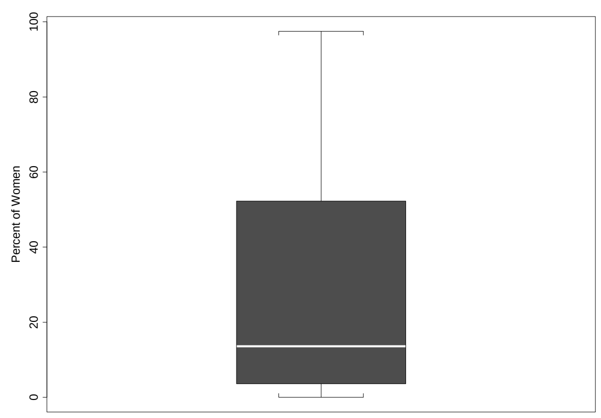

Figure 3: Percent of Incumbents who are Women

(e) Pineo Porter Prestige Score for the Occupation

This variable contains the Pineo Porter Prestige Score for the occupations. Table 1 contains the numerical summary. The minimum of 14.8 was obtained by newsboys, and the maximum of 87.2 was obtained by Physicians. The mean of 46.83 is higher than the median of 43.6 indicating a positive skew.

Figure 4 contains a histogram of the prestige scores. The data is skewed right, with a very short lower tail. The most frequent score is between 30 and 40.

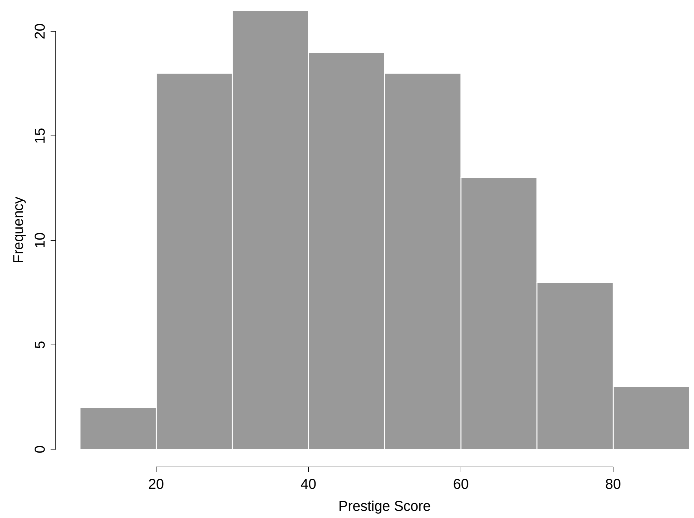

Figure 4: Prestige Scores

(f) Canadian Census Occupational Code

This variable contains the Canadian Census occupational code. Thus there are 102 codes, one for each of occupational titles listed above. Table 1 contains the numerical summary. The minimum code of 1113 corresponds to general administrators and the maximum code of 9517 corresponds to bookbinders.

Figure 5 contains a histogram of the occupational codes. The data follows no distribution.

(g) Type of Occupation

This variable contains the type of occupation, either professional, white collar, or blue collar. This variable contains some missing data, i.e. there were some occupations which were not classified. Specifically, there were 30 professional occupations, 23 white collar occupations, 45 blue collar occupations and 4 were not classified.

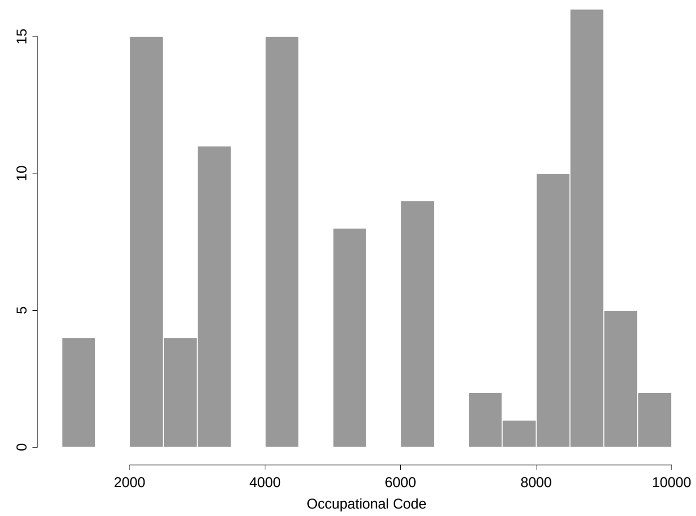

Figure 5: Census Occupational Code

Variable	Minimum Value	First Quartile	Median	Mean	Third Quartile	Maximum
Education	6.38	8.445	10.54	10.74	12.65	15.97
Income	611	4,106	5,930	6,738	8,817	25,880
Percent Women	0	3.59	13.6	28.98	52.2	97.51
Prestige Score	14.8	35.23	43.6	46.83	59.28	87.2
Occupational Code	1113	3120	5135	5402	8312	9517

Table 1: Numerical Summaries for the Occupational Prestige data

2. Simple Regression of Prestige on Income

   During the 1971 Canadian Census, data was collected to learn more about the people who work at 102 different occupations. Two of the measures collected were the average income of the workers in that occupation and the Pineo Porter Prestige Score for the occupations. A natural question regarding the data is whether the higher incomes correspond to higher prestige scores. To better understand this, a linear regression of the average income was performed onto the prestige score. The regression model was:

   Prestige Score = 27.1412 + 0.0029 * (Income).

   A graph of the average income and the prestige score along with the regression results is in Figure 6. The dots on the graph represent the actual data and the solid line represents results of the regression analysis. For example, the regression predicts that an average income of about $5,000 would have a prestige score of approximately 41. This graph shows that there is a positive trend in the data (i.e. as the income rises, so does the occupational prestige). Figure 7 contains a graph of the residuals. These indicate the differences between the regression model's predicted value and the actual value obtained for the 102 data points. The three largest residuals represent the three data points which in which the model performed the worst. These data points represent the occupations of General Managers, Ministers and Physio Therapists. The model under-represented the general managers, and over-represented the ministers and physio-therapists. Because there appears to be a trend in the graph of the residuals (i.e. between fitted values of 30 and 50, the residuals appear to be increasing, and between fitted values of 50 and 70, the residuals appear to be decreasing), this may not be the best model to describe the data.

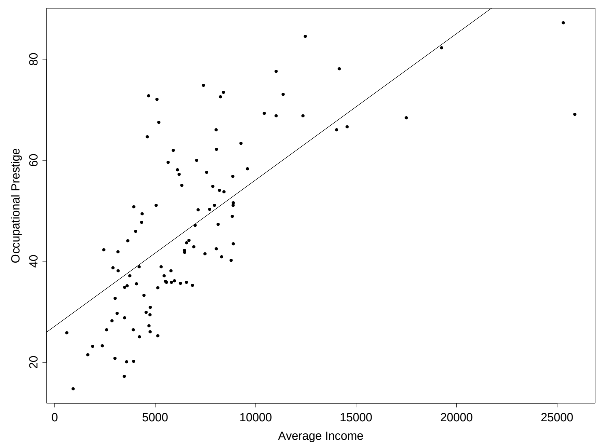

Figure 6: Income Level vs. Prestige Score

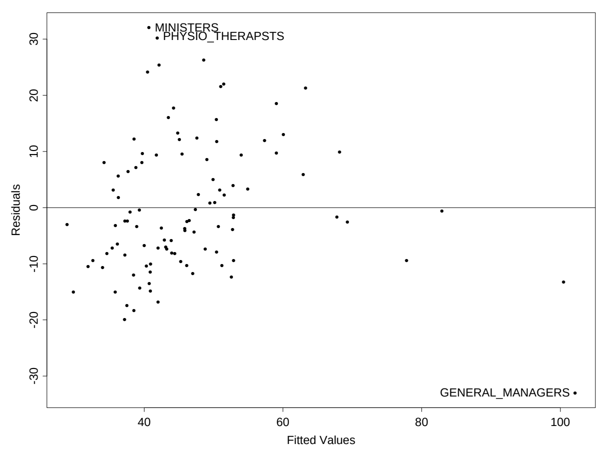

Figure 7: Residuals from the Regression Model

3. Leaded Gasoline and Lead Burden in Newborn Infants.

   With the research indicating that lead poisoning slows the development of a child's intellectual ability, it is important to study how lead from various environmental sources has an impact on a child's development. In 1980-1981 there was a study at a major Boston hospital which measured the mean lead concentration in the umbilical-cord blood of babies and the concentration of lead in room air for different families. To analyze if there is a relationship between these, a the data was graphed and regression analysis was performed. Figure 8 contains the data and the regression line. The figure shows that the data do exhibit a generally increasing trend and that the linear regression line appears appropriate.

   The regression equation is:

(umbilical-cord blood) = 4.96 + 8.97 * (lead concentration in air).

The regression statistics indicate that the intercept is is significant to the equation, with a p-value of 0.0000. The lead concentration in the air is also significant with a p-value of 0.0056. In both cases we reject the null hypothesis that the regression coefficients are zero. The regression intercept indicates that when the lead concentration in the air is zero, that the average lead reading in the umbilical-cord blood is 4.96 units. The coefficient for the lead concentration in the air indicates that for each unit increase in the lead concentration, the average umbilical-cord blood reading will increase by 8.97 units.

The residual standard error 0.5974. This means that, on average, the true umbilical-cord blood reading will be about 0.5974 units away from the regression line.

The R^2 value for the regression is 0.4862. This means that the linear relationship explains about 48.62% of the sample variance of the umbilical cord blood readings.

Thus, the numerical summaries indicate that the regression did fairly well. To complete the analysis, consider the residual analysis in Figure 9. The residuals in the figure appear to be randomly scattered throughout the graph, that is, there does not appear to be any trend in the residuals. This indicates that the residuals appear to be independent.

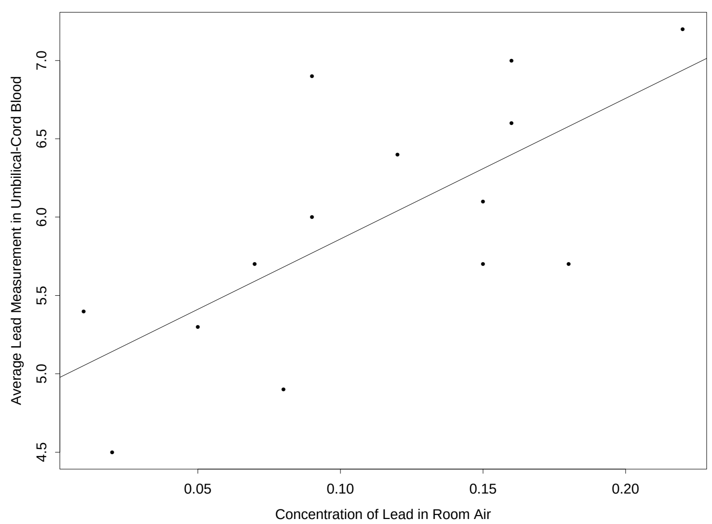

Figure 8: Regression of Umbilical-Cord Blood Concentration on Lead Air Concentration

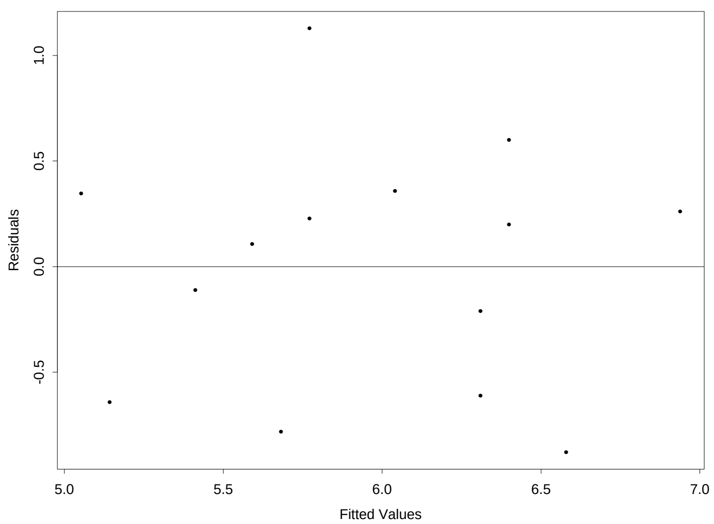

Figure 9: Regression Residuals

4. Transformations of Income

   To examine income, I first examined the histogram of income. See Figure 10.

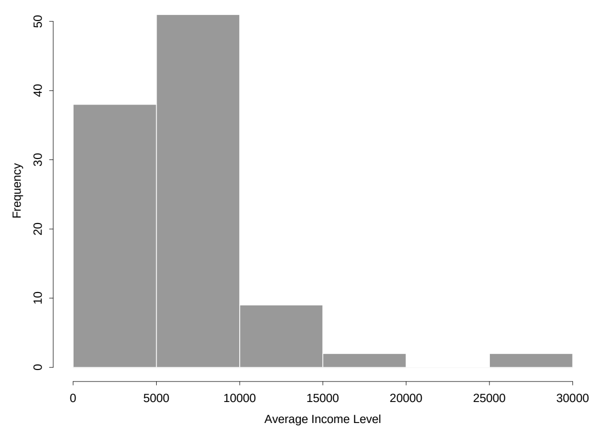

Figure 10: Histogram of Average Income Level

From the histogram, it is clear that the data is skewed right and not very symmetric. To make the distribution more symmetric, examine two transformations, the log and the square root.

Figure 11 shows that the log transformation of the data shifted the histogram from having a very strong positive skew to a histogram which is relative symmetric (note that the mean is 8.661 and the median is 8.686). The Normal Probability Plot shows that the data is roughly normal, however the tails are slightly heavy. The symmetry plot also shows that the data is roughly symmetric, with the two most extreme values being slightly not symmetric.

Figure 12 shows that the square root transformation of the data shifted the data from having very

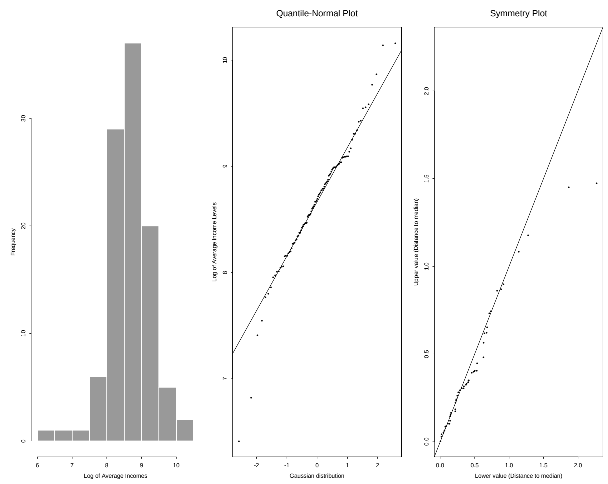

Figure 11: Examination of the Log of the Average Income Levels

strong positive skew to a slight positive skew (note that the mean is 79.2 and the median is 77.01). The quantile normal plot shows that the normality assumption is reasonable, however the right tail is a bit heavy. The symmetry plot supports the fact that the right tail is a bit heavy, with roughly 13 data points falling off of the symmetry line.

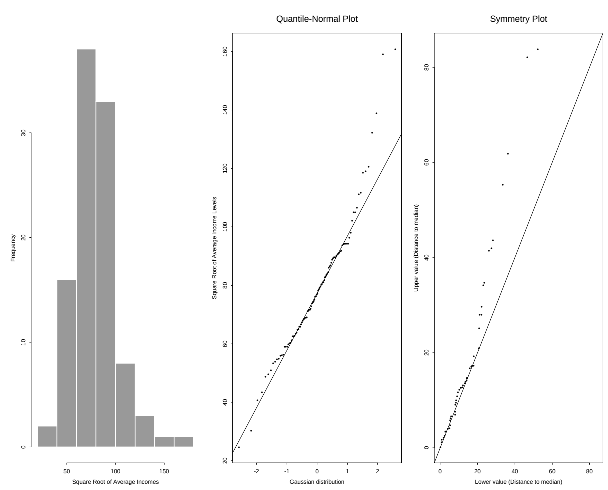

Figure 12: Examination of the Square Root of the Average Income Levels

Reviewing all the data concerning both the log and the square root transformations, I believe that the log transformation produces the better transformation. Though it is difficult to determine which normal probability plot looks better, the histogram of the log transformation appears more symmetric, and the symmetry plot for the log transformation has fewer points off of the symmetry line.

5. Re-do Problem #2 using the best transformation found in #4.

   During the 1971 Canadian Census, data was collected to learn more about the people who work at 102 different occupations. Two of the measures collected were the average income of the workers in that occupation and the Pineo Porter Prestige Score for the occupations. A natural question regarding the data is whether the higher incomes correspond to higher prestige scores. Because the income levels were heavily concentrated toward lower values, a logarithmic transformation was taken to create a model which better explains the data. Thus, to better understand this, a linear regression of the logarithm of the average income was performed onto the prestige score. The regression model was:

   Prestige Score = -139.8557 + 21.55 * Logarithm(Income).

   A graph of the logarithm of the average income and the prestige score along with the regression results using the logarithm scale is in Figure 13. The same graph on the untransformed scale is in figure 15. The dots on the graph represent the actual data and the solid line represents results of the regression analysis. For example, the regression predicts that if the logarithm of the average income is about 8, then the regression model predicts that occupational would have a prestige score of approximately 32. This graph shows that there is a positive trend in the data (i.e. as the logarithm of income rises, so does the occupational prestige). Figure 14 contains a graph of the residuals. These indicate the differences between the regression model's predicted value and the actual value obtained for the 102 data points. The three largest residuals represent the three data points which in which the model performed the worst. These data points represent the occupations of Babysitters, Ministers and Physio Therapists. The model over-represented all three of these groups. Because the data points in the residual graph exhibit no patterns, this model is a good model.

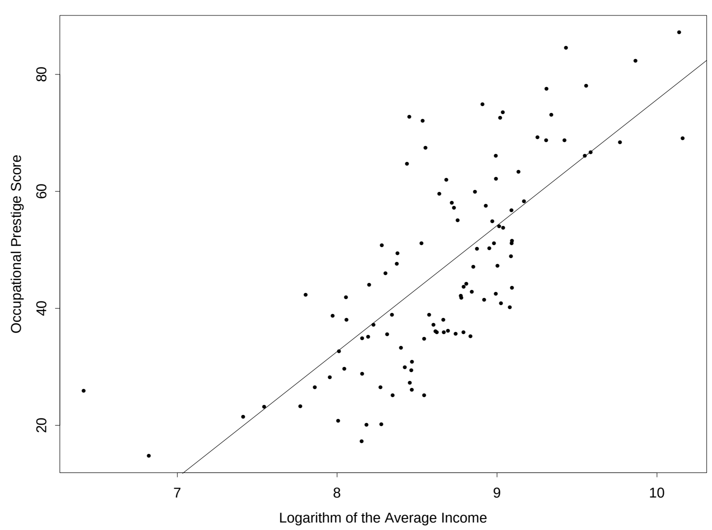

Figure 13: Regression of Logarithm of Income Level vs. Prestige Score, Transformed Scale

For examining the regression model and creating a better model, the regression with the logarithm of the income is a better model. However, it is also more difficult to explain to the layperson. If I were to communicate with a non-statistical collaborator, I may choose the non-optimal model for the sake of ease of interpretation. It all depends on how much "worse" the non-transformed model performs and what type of conclusions I'm trying to draw from the model.

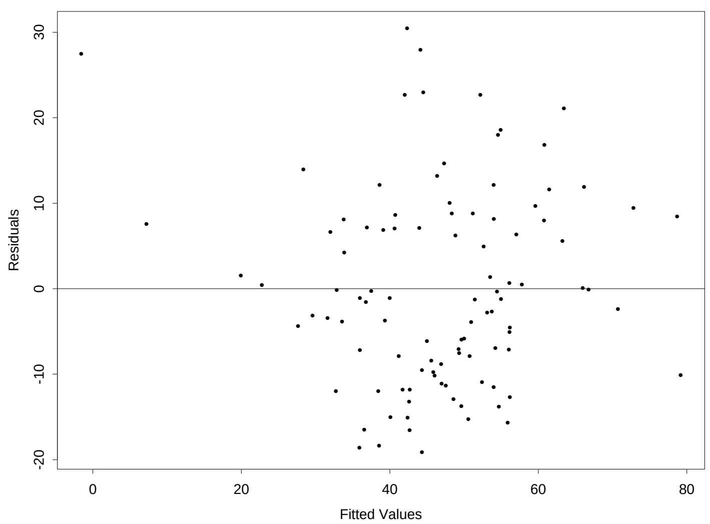

Figure 14: Residuals from the Regression Model

Observation	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	1	5.5	-2	-5	4	10
2	2	7.0	-1	-3.5	1	3.5
3	3	9.5	0	-1	0	0.0
4	4	13.0	1	2.5	1	2.5
5	5	17.5	2	7.0	4	14.0
Total	15	52.5	0	0	10	30

Table 2: Numerical Summaries for the Occupational Prestige data

6. Regression with X and Y

   Consider the error free, non-linear model $Y = 5 + X^2/2$. The Y-values for X=1,2,3,4,5 are 5.5, 7.0, 9.5, 13.0, 17.5. Figure 16 contains a plot fo the data points and the linear regression line y=1.5*3x.

   Table 2 has been created to summarize some of the data.

   Note that the three central points are below the regression line while the largest and smallest points x-values fall above the regression line. This occurs because the data is actually a quadratic and follows a U shape pattern. The line must cross the U where the line minimizes the sum of squared residuals.

   Note that if we set $b_1 = \frac{\Sigma(X-\bar{(X)})(Y-\bar{(Y)})}{\Sigma(X-(X))^2} = \frac{30}{10} = 3$. This matches the slope of the regression line. Further $b_0 = \bar{Y} - b_1\bar{X} = \frac{52.5}{5} - 3 * \frac{15}{5} = 10.5 - 3 * 3 = 1.5$. This matches the intercept of the regression line.

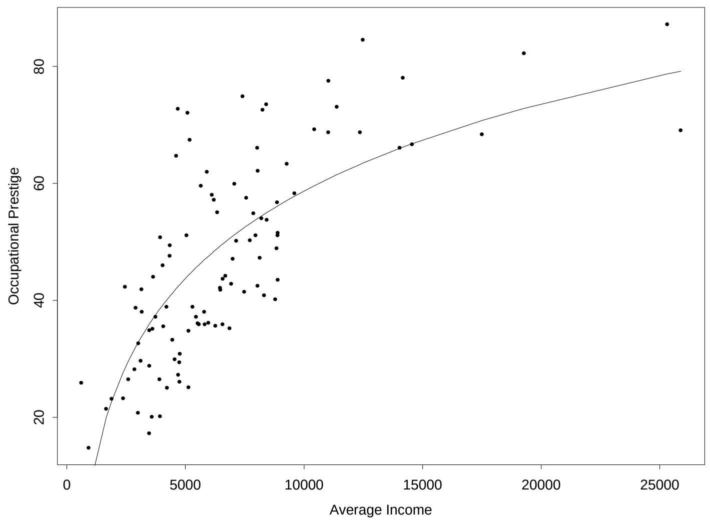

Figure 15: Regression of Logarithm of Income Level vs. Prestige Score, Untransformed Scale

Appendix - S+ Code

1. Problem 1

```
# To read in the data
> prestige_read.table("../HW2/Prestige.dat")

# To get the occupational titles
> dimnames(prestige)

# To get the education data
> summary(education)
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
 6.38   8.445  10.54 10.74   12.65 15.97
> prestige[prestige$education==6.38,]
                      education income pct.female occ.prestige occ.code
SEWING_MACH_OPERATORS      6.38   2847      90.67         28.2     8563
                      occ.type
SEWING_MACH_OPERATOR        bc
> prestige[prestige$education==15.97,]
                    education income pct.female occ.prestige occ.code occ.type
UNIVERSITY_TEACHERS     15.97  12480      19.59         84.6     2711
prof
> hist(prestige$education, xlab="Average Education", ylab="Frequency")

# To get the Income Data
> summary(income)
 Min. 1st Qu. Median Mean 3rd Qu.  Max.
```

```
  611    4106   5930 6798    8187 25880
> prestige[prestige$income==611,]
            education income pct.female occ.prestige occ.code occ.type
BABYSITTERS      9.46    611      96.53         25.9     6147        ?
> max(prestige$income)
[1] 25879
> prestige[prestige$income==25879,]
                 education income pct.female occ.prestige occ.code occ.type
GENERAL_MANAGERS     12.26  25879       4.02         69.1     1130
> hist(prestige\income, nclass=10,xlab="Average Income",ylab="Frequency")

# Percent Women Data
> summary(pct.female)
 Min. 1st Qu. Median  Mean 3rd Qu.  Max.
    0   3.592   13.6 28.98    52.2 97.51
> prestige[prestige$pct.female==0,]
                     education income pct.female occ.prestige occ.code
        FIREFIGHTERS      9.47   8895          0         43.5     6111
ROTARY_WELL_DRILLERS      8.88   6860          0         35.3     7711
  RAILWAY_SECTIONMEN      6.67   4696          0         27.3     8715
     TRAIN_ENGINEERS      8.49   8845          0         48.9     9131
        LONGSHOREMEN      8.37   4753          0         26.1     9313
                     occ.type
        FIREFIGHTERS       bc
ROTARY_WELL_DRILLERS       bc
  RAILWAY_SECTIONMEN       bc
     TRAIN_ENGINEERS       bc
        LONGSHOREMEN       bc
> prestige[prestige$pct.female==97.51,]
            education income pct.female occ.prestige occ.code occ.type
SECRETARIES     11.59   4036      97.51           46     4111       wc
> rwgbox(prestige$pct.female,ylab="Percent of Women")


# Pineo Porter Prestige Scores
> summary(occ.prestige)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
 14.8   35.23   43.6 46.83   59.28 87.2
> prestige[prestige$occ.prestige==14.8,]
         education income pct.female occ.prestige occ.code occ.type
NEWSBOYS      9.62    918          7         14.8     5143        ?
> prestige[prestige$occ.prestige==87.2,]
           education income pct.female occ.prestige occ.code occ.type
PHYSICIANS     15.96  25308      10.56         87.2     3111     prof
> hist(prestige$occ.prestige, xlab="Prestige Score", ylab="Frequency")

# Occupatinal Code
```

```
> summary(occ.code)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 1113    3120   5135 5402    8312 9517
> prestige[prestige$occ.code==1113,]
                   education income pct.female occ.prestige occ.code
occ.type
GOV_ADMINISTRATORS     13.11  12351      11.16         68.8     1113
prof
> prestige[prestige$occ.code==9517,]
            education income pct.female occ.prestige occ.code occ.type
BOOKBINDERS      8.55   3617      70.87         35.2     9517       bc
> hist(prestige$occ.code,nclass=20)

# Type of Occupation
> summary(occ.type)
 ? bc prof wc
 4 45   30 23
```

2. Problem 2

```
> prestige.lm_lm(occ.prestige~income)
> summary(prestige.lm)

Call: lm(formula = occ.prestige ~ income)
Residuals:
    Min     1Q Median    3Q   Max
 -33.01 -8.378 -2.378 8.432 32.08

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 27.1412  2.2677    11.9686  0.0000
     income  0.0029  0.0003    10.2243  0.0000

Residual standard error: 12.09 on 100 degrees of freedom
Multiple R-Squared: 0.5111
F-statistic: 104.5 on 1 and 100 degrees of freedom, the p-value is 0

Correlation of Coefficients:
       (Intercept)
income -0.8493
> plot(income, occ.prestige, xlab="Average Income", ylab="Occupational Prestige")
> abline(prestige.lm)
> plot(prestige.lm$fitted, prestige.lm$resid, xlab="Fitted Values", ylab="Residuals")
>  abline(h=0)
```

3. Problem 3

```
> lead.lm_lm(blood~air, data=lead, na.action=na.omit)
```

```
> summary(lead.lm)

Call: lm(formula = blood ~ air, data = lead, na.action = na.omit)
Residuals:
     Min      1Q Median     3Q   Max
 -0.8791 -0.5098 0.1544 0.3256 1.129

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  4.9634  0.3354    14.8003  0.0000
        air  8.9760  2.6637     3.3697  0.0056

Residual standard error: 0.5974 on 12 degrees of freedom
Multiple R-Squared: 0.4862
F-statistic: 11.35 on 1 and 12 degrees of freedom, the p-value is 0.005573

Correlation of Coefficients:
    (Intercept)
air -0.8794
 > plot(lead$air, lead$blood, xlab="Concentration of Lead in Room Air", ylab="Average Lead Measurement in Umbilical-Cord Blood")
> abline(lead.lm)
> plot(lead.lm$fitted, lead.lm$residuals, xlab="Fitted Values", ylab="Residuals")
> abline(h=0)
> ps("../HW2/bloodresid.ps")
```

4. Problem 4

```
> hist(prestige$income, xlab="Average Income Level", ylab="Frequency")
> par(mfrow=c(1,3))
> hist(log(prestige$income), xlab="Log of Average Incomes", ylab="Frequency")
> qqn(log(prestige$income), ylab="Log of Average Income Levels")
> symplot(log(prestige$income))
> hist(sqrt(prestige$income), xlab="Square Root of Average Incomes", ylab="Frequency")
> qqn(sqrt(prestige$income), ylab="Square Root of Average Income Levels")
> symplot(sqrt(prestige$income))
> summary(log(prestige$income))
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  6.415  8.320   8.688  8.661  9.010  10.160
> summary(sqrt(prestige$income))
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  24.72  64.08   77.01  79.20  90.48  160.90
```

5. Problem 5

```
> prestige.lm2_lm(log(prestige$income)~prestige$occ.prestige)
> summary(prestige.lm2)

Call: lm(formula = log(prestige$income) ~ prestige$occ.prestige)
```

```
Residuals:
    Min      1Q Median     3Q    Max
 -1.712 -0.2392 0.1108 0.2468 0.9334

Coefficients:
                        Value Std. Error t value Pr(>|t|)
          (Intercept)  7.4674  0.1151    64.8775  0.0000
prestige$occ.prestige  0.0255  0.0023    11.0368  0.0000

Residual standard error: 0.3991 on 100 degrees of freedom
Multiple R-Squared: 0.5492
F-statistic: 121.8 on 1 and 100 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                      (Intercept)
prestige$occ.prestige -0.9392
> plot(log(prestige$income), prestige$occ.prestige,xlab="Logarithm of the Average Income", ylab="Occupational Prestige Score")
> abline(prestige.lm2)
> plot(prestige.lm2$fitted, prestige.lm2$resid, xlab="Fitted Values", ylab="Residuals")
> abline(h=0)
```

6. Problem 6

```
> x_1:5
> y_5+x^2/2
> y
[1]  5.5  7.0  9.5 13.0 17.5
> xy.lm_lm(y~x)
> summary(xy.lm)

Call: lm(formula = y ~ x)
Residuals:
 1    2  3    4 5
 1 -0.5 -1 -0.5 1

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 1.5000 1.1328     1.3241  0.2773
          x 3.0000 0.3416     8.7831  0.0031

Residual standard error: 1.08 on 3 degrees of freedom
Multiple R-Squared: 0.9626
F-statistic: 77.14 on 1 and 3 degrees of freedom, the p-value is 0.003109

Correlation of Coefficients:
  (Intercept)
x -0.9045
> plot(x, y, xlab="X Values", ylab="Y Values")
```

```
> abline(xy.lm)
> x
[1] 1 2 3 4 5
> y
[1]  5.5  7.0  9.5 13.0 17.5
> x-mean(x)
[1] -2 -1  0  1  2
> y-mean(y)
[1] -5.0 -3.5 -1.0  2.5  7.0
> (x-mean(x))^2
[1] 4 1 0 1 4
> (x-mean(x))*(y-mean(y))
[1]  10.0   3.5   0.0   2.5  14.0
> sum(x)
[1] 15
> sum(y)
[1] 52.5
> sum(x-mean(x))
[1] 0
> sum(y-mean(y))
[1] 0
> sum((x-mean(x))^2)
[1] 10
> sum((x-mean(x))*(y-mean(y)))
[1] 30
> sum((x-mean(x))*(y-mean(y)))
[1] 30
> 10/30
[1] 0.3333333
> mean(y)-3*mean(x)
[1] 1.5
```

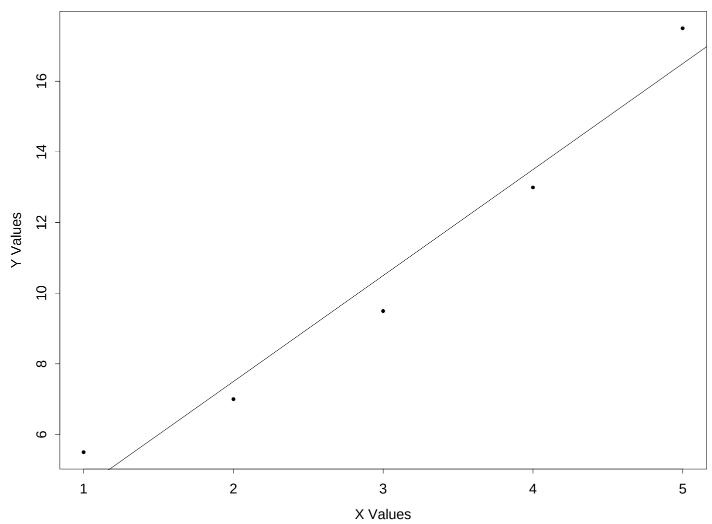

Figure 16: X and Y Datapoints with Regression Line