Homework 4 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Due: Tuesday October 9, 2001

1. Prestige Data set Problem 1

   (a) Make a scatter plot matrix of the first four variables in `newprestige` and also a correlation matrix for these four variables. Write a few sentences describing which variables may be related to `occ.prestige`. Are there any (approximate) collinearities evident among `education, income`, and `pct.female`?

   The scatterplot matrix in Figure 1 identifies a few variables which may be related to `occ.prestige`.

```
> pairs(newprestige[,1:4])
```

In particular, the education variable appears to have a positive association with the occ.prestige variable (i.e. when education increases, so does occupational prestige). The scatterplot between the percent female and the occupational prestige is pretty random, indicating that there is not much association. Finally, the scatterplot between the logarithm of the income and the occupational prestige appears to have a positive association (i.e. when the logarithm of the income increases so does the occupational prestige).

To determine if there are any collinearities among the explanatory variables, the scatterplots between education, percent female and the logarithm of income are examined. The scatter plot between education and percent female is mostly random, indicating no collinearity. The scatterplot between education and the logarithm of the income appears to have a positive association, though the association is not very strong. In addition, there also appears to be slight negative association between the percent female and the logarithm of the income. None of these associations appear strong enough to be concerned with collinearity. In addition, it is good to check the correlation matrix between the variables.

```
> cor(newprestige[,1:4])
             occ.prestige  education  pct.female     Loginc
occ.prestige    1.0000000 0.85017689 -0.11833419  0.7410561
   education    0.8501769 1.00000000  0.06185286  0.5481051
  pct.female   -0.1183342 0.06185286  1.00000000 -0.5225440
      Loginc    0.7410561 0.54810506 -0.52254399  1.0000000
```

In the correlation matrix, the largest correlation between the explanatory variables is 0.54, which again, is not large enough to be concerned with collinearity.

   (b) Summarize the result of regressing `occ.prestige` on the three variables `education, pct.female` and `LogInc`. Turn in (i) the fitted regression equation, with standard errors for each regression coefficient; (ii) R^2, R^2_{Adj} and F; (iii) a small set of appropriate diagnostic plots; and (iv) one sentence summarizing the evidence you have provided for goodness of fit of the multiple regression model.

   The regression results are below:

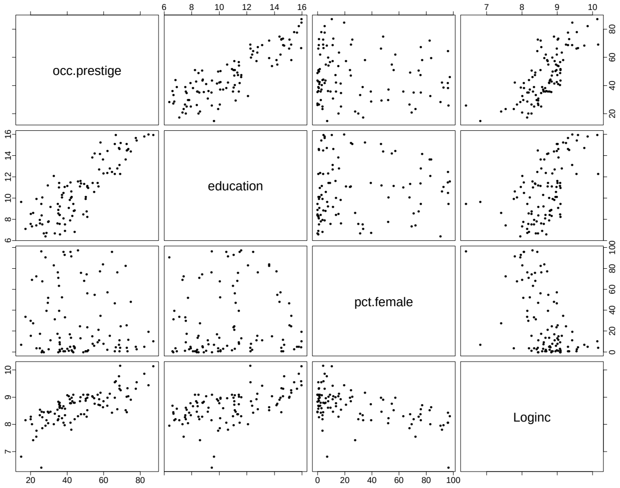

Figure 1: Pairs Plot for the Prestige Dataset

```
> summary(lm(occ.prestige~education+pct.female+Loginc, data=newprestige))

Call: lm(formula = occ.prestige ~ education + pct.female + Loginc,
        data = newprestige )
Residuals:
    Min     1Q Median    3Q   Max
 -17.36 -4.429 -0.101 4.316 19.18

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -110.9658   14.8429   -7.4760    0.0000
  education    3.7305    0.3544   10.5268    0.0000
 pct.female    0.0469    0.0299    1.5685    0.1200
     Loginc   13.4382    1.9138    7.0219    0.0000

Residual standard error: 7.093 on 98 degrees of freedom
Multiple R-Squared: 0.8351
F-statistic: 165.4 on 3 and 98 degrees of freedom, the p-value is 0

Correlation of Coefficients:
           (Intercept) education pct.female
 education  0.5338
pct.female -0.6775     -0.4884
    Loginc -0.9807     -0.6821    0.6666
```

This indicates that the fitted regression line is occ.prestige=-110.96 + 3.73*education + 0.0469*pct.female + 13.43*Loginc.

The standard errors for the regression coefficients are 14.84 for the intercept, 0.35 for education, 0.0299 for pct.female and 1.91 for Loginc. These, together with their respective p-values, indicate that all the variables are needed in the regression except for pct.female.

The R^2 value is 0.8351. Recall that $R^2_{adj} = R^2 - \frac{K-1}{n-K}(1-R^2)$. Thus in this case, $R^2_{adj} = 0.8351 - \frac{4-1}{102-4}(1-0.8351) = 0.830$. This indicates that 83% of the variation in the data is explained by the regression line.

The F Statistic for the regression is 165.4 on 3 and 98 degrees of freedom. This indicates that this regression is significantly better than explaining the data with the mean of occupational prestige.

Some appropriate diagnostic plots are in Figure 2.

```
>  occ.prest.lm_lm(occ.prestige~education+pct.female+Loginc,
   data=newprestige)
> par(mfrow=c(1,2))
> plot(fitted(occ.prest.lm), resid(occ.prest.lm), xlab="Fitted
   Values", ylab="Residuals")
> abline(h=0)
> qqn(resid(occ.prest.lm))
```

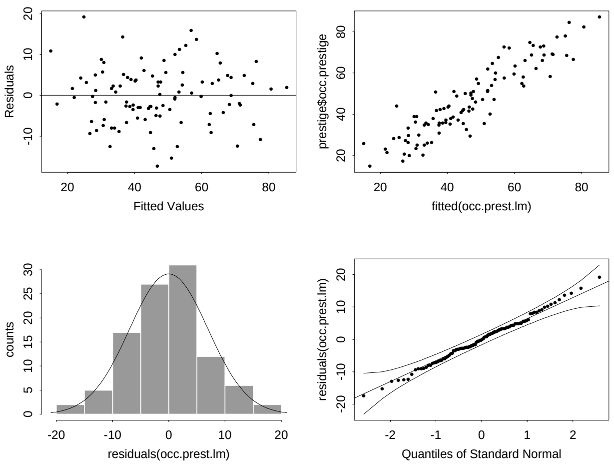

Figure 2: Diagnostics for the occ.prestige Multiple Regression

The fitted vs. residual plot indicates that the residuals are randomly scattered about the 0 line with no apparent trend. The fitted values versus the occupational prestige scores shows the fit of the regression line with the data points. The histogram indicates that the residuals are fairly normal. The normal probability plot indicates that the residuals are normally distributed. Thus the tests from the regression are valid.

Thus, the fit of this model is good, since the residuals appear to be independent normal random variables, and the regression test statistics all indicate that the regression is adequate.

(c) Turn in the output from (see HW sheet). Does the best model listed here agree with your intuition, based on the t statistics in part (b)?

The output is included below.

```
> sum.step(stepwise(newprestige[,2:4],prestige\$occ.prestige,
         method="exhaustive",nbest=2),newprestige, "occ.prestige",
         includeAIC=T, includeR2=T, sort=0)
         BIC    AIC    R2   R2a education pct.female Loginc
1(#1) -373.6 -371.0 0.723 0.720         1          0      0
1(#2) -398.4 -395.8 0.549 0.545         0          0      1
2(#1) -350.7 -346.8 0.831 0.828         1          0      1
2(#2) -370.3 -366.3 0.752 0.747         1          1      0
3(#1) -351.8 -346.6 0.835 0.830         1          1      1
0(#2) -436.8 -435.4 0.000 0.000         0          0      0
```

According to the above output, there are two models which appear roughly equivalent. They are the models

```
         BIC    AIC    R2   R2a education pct.female Loginc
2(#1) -350.7 -346.8 0.831 0.828         1          0      1
3(#1) -351.8 -346.6 0.835 0.830         1          1      1
```

Their R^2 adjusted are almost the same, as are their AIC, BIC values. If I were to pick between these two models, I would choose the simpler model, i.e. the one with fewer parameters. Thus, I would pick the model with the education and the Loginc variables, but not include the pct.female. This agrees with the output from part (b), because the p-values for the education (0.0000) and the Loginc (0.000) are small and the p-value for pct.female (0.1200) is large.

2. [Continuation.] Consider the general regression model. An alternative procedure for calculating the least-squares coefficient $\hat{\beta}_1$ is as follows:

- Regress y on $X_2, ..., X_k$
- Regress X_1 on $X_2, ..., X_k$
- Regress the residuals from the first regression on the residuals from the second regression.

(a) Apply this procedure to the `newprestige` data set and the regression `occ.prestige ~ education + pct.female+LogInc` to obtain the coefficient for `education`. Is it the same as in the multiple regression in part(a)?

Following the above list, I generated the following regressions.

```
> hw4.2.a.lm_lm(occ.prestige~pct.female+Loginc, data=newprestige)
> hw4.2.a2.lm_lm(education~pct.female+Loginc, data=newprestige)
> hw4.2.aresid.lm_lm(resid(hw4.2.a.lm)~resid(hw4.2.a2.lm))
> summary(hw4.2.aresid.lm)

Call: lm(formula = resid(hw4.2.a.lm) ~ resid(hw4.2.a2.lm))
Residuals:
```

```
    Min     1Q Median    3Q   Max
 -17.36 -4.429 -0.101 4.316 19.18

Coefficients:
                      Value Std. Error  t value Pr(>|t|)
       (Intercept)   0.0000   0.6952     0.0000   1.0000
resid(hw4.2.a2.lm)   3.7305   0.3508    10.6336   0.0000

Residual standard error: 7.021 on 100 degrees of freedom
Multiple R-Squared: 0.5307
F-statistic: 113.1 on 1 and 100 degrees of freedom, the p-value is 0

Correlation of Coefficients:
                   (Intercept)
resid(hw4.2.a2.lm)  0
```

The coefficient on the `resid(hw4.2.a2.lm)` variable is 3.71, which matches the coefficient for education obtained above.

(b) Illustrate the regression in part(a) with an appropriate scatter plot, and comment on the usefulness of X_1 in the model, based on this scatter plot (this is the "partial regression leverage plot").

The partial regression leverage plot is seen in Figure 3. This was obtained using the following S+ commands.

```
> plot(resid(hw4.2.a2.lm), resid(hw4.2.a.lm), xlab="Residuals from
lm(education~pct.female+Loginc, data=newprestige)", ylab="Residuals from
lm(occ.prestige~pct.female+Loginc, data=newprestige)" )
> abline(hw4.2.aresid.lm)
> title("Partial Regression Leverage Plot")
```

This plot expands on what was seen in the summary for the β_1 term in the regression summary. That is, the slope of the regression line is 3.73. The scatter of the points around the line shows variations in the data independent of the percent female and logarithm of income variables. Thus, it appears that X_1 is useful in the model because there is clearly an increasing trend which appears larger than the spread of points about the line.

(c) Notice that the intercept in your regression in part (a) is 0. Why is that?

Note that for a simple regression model, $b_0 = \bar{Y} - b_1\bar{X}$. In this case,

$$\bar{Y} = \frac{\sum_{i=1}^{n} e_{i,Y|\mathbf{X}}}{n} = 0 \text{ and } \bar{X} = \frac{\sum_{i=1}^{n} e_{i,X_1|\mathbf{X}}}{n} = 0.$$

Therefore, $b_0 = 0 - b_1(0) = 0$.

A similar approach:
Consider the model $e_{i,Y|\mathbf{X}} = b_0^* + b_1^* e_{i,X_1|\mathbf{X}}$ for two parameters, b_0^* and b_1^*, where $e_{i,Y|\mathbf{X}}$ denotes the residuals formed by regressing Y on $X_2, ..., X_k$ and $e_{i,X_1|\mathbf{X}}$ denotes the residuals formed by regressing X_1 on $X_2, \ldots, X_k$. We want to solve for b_0^*:

$$\begin{aligned} b_0^* &= e_{i,Y|\mathbf{X}} - b_1^* e_{i,X_1|\mathbf{X}}, \\ \underbrace{\sum_{i=1}^{n} b_0^*}_{nb_0^*} &= \underbrace{\sum_{i=1}^{n} e_{i,Y|\mathbf{X}}}_{=0} - \underbrace{b_1^* \sum_{i=1}^{n} e_{i,X_1|\mathbf{X}}}_{=0}, \end{aligned}$$

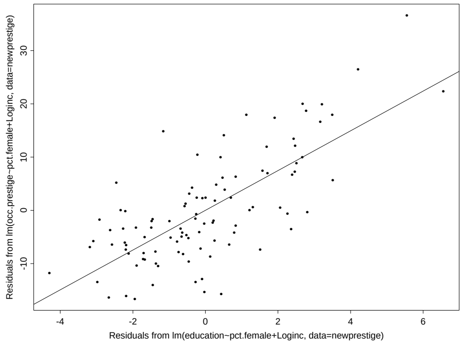

Figure 3: Partial Regression Leverage Plot

therefore, $b_0^* = \frac{0}{n} = 0$.

3. [Continuation.]

(a) Make parallel boxplots of `occ.prestige`, split by `occ.type`, using the `prestige` data set.

Figure 4 displays the boxplots split by occupational type. As a result, we see that professionals have the highest median Pineo-Porter prestige score, with white collar workers, blue collar workers and unknown occupations all having median prestige scores less than or approximately 40. There is at least one high outlier among white collar workers, but no apparent outliers within the other three occupational types.

(b) Summarize the multiple regression `occ.prestige ~ Prof + WC + BC + Unknown - 1` , using the `newprestige` data set.

```
> occ.prestige.lm_lm(occ.prestige ~ Prof + WC + BC + Unknown - 1,
data=newprestige)
> summary(occ.prestige.lm)

Call: lm(formula = occ.prestige \~{} Prof + WC + BC + Unknown - 1, data =
newprestige)
Residuals:
    Min     1Q  Median    3Q   Max
 -19.93 -7.982 -0.3067 6.218 29.89

Coefficients:
          Value Std. Error t value Pr(>|t|)
   Prof 67.9067  1.8826    36.0701  0.0000
```

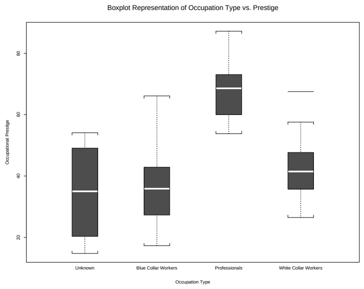

Figure 4: Parallel Boxplots of Occupational Prestige, split by Occupational Type

```
     WC 42.2435  2.1501    19.6471  0.0000
     BC 36.2067  1.5372    23.5543  0.0000
Unknown 34.7250  5.1558     6.7351  0.0000

Residual standard error: 10.31 on 98 degrees of freedom
Multiple R-Squared: 0.9589
F-statistic: 571.8 on 4 and 98 degrees of freedom, the p-value is 0

Correlation of Coefficients:
        Prof WC BC
     WC 0
     BC 0    0
Unknown 0    0  0
```

The command `occ.prestige.lm_lm(occ.prestige ~ Prof + WC + BC + Unknown - 1, data=newprestige)` considers the hypothesis test H_0 : $\texttt{occ.prestige} = \epsilon$ vs. H_1 : $\texttt{occ.prestige} = \beta_1\texttt{Prof} + \beta_2\texttt{WC} + \beta_3\texttt{BC} + \beta_4\texttt{Unknown} + \epsilon$. As a result of the output provided above, we reject H_0 and conclude that each coefficient is significantly different from zero. This is also seen through the F-statistic (571.8 on 4 and 98 degrees of freedom) whose p-value is zero; thus we reject H_0 in favor of the H_1. The multiple regression model explains almost 96% of the variation in the data and there exists no correlation between any of the coefficients, which implies that there exists no collinearity within this model.

The adjusted R^2 was computed with the following command:


```
> occ.R2_summary(occ.prestige.lm)$r.squared
> occ.R2
[1] 0.9589138         # this is the r-squared value listed above
> adj.occ.R2_occ.R2 - ((3/98)*(1-occ.R2)) # here is the formula for
                                            adjusted r-squared
>  round(adj.occ.R2,4)
[1] 0.9576
```

Thus, $R_a^2 = 0.958$ is slightly lower than R^2 but still large at approximately 96 percent. Furthermore, the F-statistic is highly significant with a p-value of zero, rejecting H_0 and verifying that the multiple regression model is the better model to use.

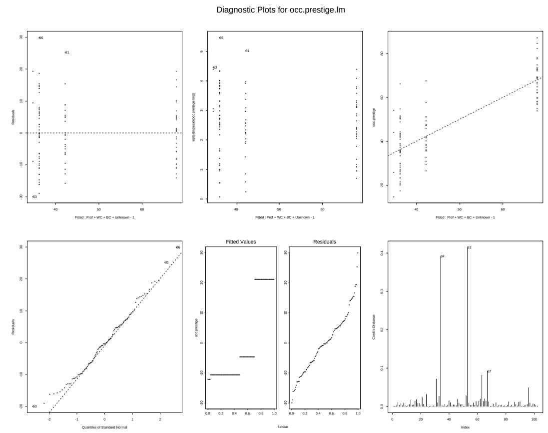

Figure 5: Splus's standard diagnostic plots for occ.prestige.lm

Figure 5 displays the diagnostic plots associated with the multiple regression model named `occ.prestige.lm`. As a result, we see that the residuals are approximately normal with a slightly wider tail, and three outliers (points 31, 96, and

53).

```
> newprestige[c(31,96,53),]
                     occ.prestige education pct.female   LogInc Prof WC BC Unknown
MEDICAL_TECHNICIANS          67.5     12.79      76.04 8.552560    0  1  0       0
             PILOTS          66.1     12.27       0.58 9.549096    0  0  1       0
           NEWSBOYS          14.8      9.62       7.00 6.822197    0  0  0       1
```

From this output, it makes sense that newsboys receive such a low prestige score. Furthermore, within white-collar workers, we see that medical technicians represent the high outlier, and pilots have the maximum prestige score among blue collar workers.

(c) In part (b), give the fitted values and 95% confidence intervals for the occupational prestiges of each of the four groups: Professionals, White collar workers, Blue collar workers, and workers of unknown type.

From the output [provided in part (b)],

```
          Value Std. Error t value Pr(>|t|)
   Prof 67.9067  1.8826    36.0701  0.0000
     WC 42.2435  2.1501    19.6471  0.0000
     BC 36.2067  1.5372    23.5543  0.0000
Unknown 34.7250  5.1558     6.7351  0.0000
```

and $|\ t_{n-K,.05/2}\ |=|\ t_{98,.05/2}\ |= 1.984467$, the 95% confidence intervals are:

Prof: $67.9067 \pm (1.98 * 1.8826) = (64.17915, 71.63425)$

WC: $42.2435 \pm (1.98 * 2.1501) = (37.9863, 46.5007)$

BC: $36.2067 \pm (1.98 * 1.5372) = (33.16304, 39.25036)$

Unknown: $34.7250 \pm (1.98 * 5.1558) = (24.51652, 44.93348)$

Comparing these results to part (a), we note that there is a strong overlap between white-collar, blue-collar, and unknown occupation types, while the prestige scores for professionals is much greater than and does not overlap with the other occupation types. This is consistent with the boxplots of part (a). We note that none of the above confidence intervales contain 0, therefore all of the associated coefficients are significantly different from zero, verifying our claim in part (b).

(d) In part (b), what does the "`- 1`" do? Try the multiple regression `occ.prestige ~ Prof + WC + BC + Unknown`, and explain why this doesn't work (give the specific relationship among the four variables `Prof`, `WC`, `BC` and `Unknown` that illustrates your explanation).

When the "`-1`" is removed from the model, we get the following error,

```
> occ.prestige.lm_lm(occ.prestige ~ Prof + WC + BC + Unknown, data=newprestige)
Error in lm.fit.qr(x, y): computed fit is singular, rank 4
Dumped
```

This is because we have "exhausted" the options with reference to occupation type. Knowing the dummy values associated with three of the four occupation types allows you to determine the value of the fourth occupation type dummy value. Therefore, knowing three of the variables fully explains the fourth variable, hence a severe case of collinearity.

4. The fuel consumption data set described in class is available as `fuel-alr.dat` in the data area of the 401 web pages; create a data frame from it. The variables are

```
(State) = Two-Letter State Code
    Pop = Population in each state
```



```
   Tax = 1971 Gas tax, cents per gallon
  Nlic = 1971 thousands of drivers
   Inc = 1972 per capita income
  Road = 1971 miles of primary highways
 Fuelc = fuel consumption, millions of gallons
  Dlic = 100*Nlic/Pop = proportion of licensed drivers
  Fuel = 1000*Fuelc/Pop = per-capita fuel consumption
```

(a) Make a scatterplot matrix of all of the variables in `fuel`, and also a correlation matrix for these variables (use `round(cor(fuel),4)` to make it fit well in 80 column output). Write a few sentences describing which variables may be related to `fuel`. Are there any (approximate) collinearities evident among the variables?

```
> round(cor(fuel),4)
          Pop     Tax    Nlic     Inc    Road   Fuelc    Dlic    Fuel
  Pop  1.0000 -0.0982  0.9916  0.3394  0.6554  0.9707 -0.3222 -0.4012
  Tax -0.0982  1.0000 -0.1292 -0.0451 -0.4012 -0.1727 -0.3368 -0.2940
 Nlic  0.9916 -0.1292  1.0000  0.3328  0.6568  0.9878 -0.2592 -0.3642
  Inc  0.3394 -0.0451  0.3328  1.0000 -0.0090  0.2591  0.1117 -0.2826
 Road  0.6554 -0.4012  0.6568 -0.0090  1.0000  0.7110 -0.0704  0.0775
Fuelc  0.9707 -0.1727  0.9878  0.2591  0.7110  1.0000 -0.2487 -0.2997
 Dlic -0.3222 -0.3368 -0.2592  0.1117 -0.0704 -0.2487  1.0000  0.5808
 Fuel -0.4012 -0.2940 -0.3642 -0.2826  0.0775 -0.2997  0.5808  1.0000
```

The variable that appears most related to fuel is the proportion of licensed drivers (correlation equals .5808). Further, we see from the correlation matrix and the scatterplot that there are strong collinearities between NLIC and POP, FUELC and POP, and NLIC and FUELC, with a weaker but nonetheless present collinearity between FUELC and ROAD.

(b) Summarize the multiple regression of `Fuel` on all other variables in the `fuel` data frame.

```
> fuel.lm_lm(Fuel ~ Pop + Tax + Nlic + Inc + Road + Fuelc +
Dlic, data=fuel)
> summary(fuel.lm)

Call: lm(formula = Fuel ~ Pop + Tax + Nlic + Inc + Road + Fuelc + Dlic,
data = fuel)
Residuals:
    Min     1Q Median    3Q   Max
 -261.3 -37.77 -12.15 18.03 267.2

Coefficients:
               Value Std. Error  t value  Pr(>|t|)
(Intercept)  59.1185  238.6249    0.2477    0.8055
        Pop   0.0162    0.0324    0.5010    0.6190
        Tax  -1.7319   14.0095   -0.1236    0.9022
       Nlic  -0.1203    0.0954   -1.2606    0.2144
        Inc -25.5418   25.3137   -1.0090    0.3187
       Road   7.3159    6.5322    1.1200    0.2691
      Fuelc   0.0844    0.0579    1.4590    0.1520
       Dlic  10.8673    2.9539    3.6789    0.0007
```

```
Residual standard error: 81.37 on 42 degrees of freedom
Multiple R-Squared: 0.5665
F-statistic: 7.841 on 7 and 42 degrees of freedom, the p-value is 4.656e-06

Correlation of Coefficients:
      (Intercept)     Pop     Tax    Nlic     Inc    Road   Fuelc
  Pop -0.3620
  Tax -0.6996     -0.0541
 Nlic  0.4270     -0.9085 -0.0184
  Inc -0.3514     -0.0686  0.1193 -0.1388
 Road -0.0296     -0.5223  0.3001  0.5539  0.1745
Fuelc -0.3966      0.6155  0.0789 -0.8820  0.2706 -0.5627
 Dlic -0.7058      0.6184  0.2259 -0.5303 -0.2252 -0.3502  0.3592
```

The result of the summary is provided above. As a result, we see the results of the multicollinearity in that none of the variables (with the exception of DLIC) have coefficients that are considered significant; however, the overall F-statistic remains significant with a p-value of 4.66×10^{-6}. Furthermore, when we investigate the correlation of coefficients, we again see the strong relationships between POP and NLIC, and NLIC and FUELC, as well as relationships between POP and FUELC, POP and ROAD, and NLIC and ROAD. These are evidenced because, for each of the above mentioned pairs, $|r| > 0.50$.

(c) Get `collin.q` from the data area of the 401 website, and do `source("collin.q")`. Then run `collin` on `fuel[,-8]` (the `fuel` data frame without the response variable `Fuel`).

```
> collin(fuel[,-8])
Pop     0.993
Nlic    0.997
Road    0.742
Fuelc   0.991
```

From the above output, approximately 99% of the variation from POP is explained by the remaining explanatory variables; similarly, 99% of the variation is explained for NLIC and FUELC. As well, approximately 74% of the variation in ROAD is explained by the other variables. These variables were detected through `collin.q` because the cutoff value within this program is 70%; therefore, if the variation explained is higher than 70% for a particular variable, the program output will inform the investigator.

(d) Use `sum.step` and `stepwise` (using `method="exhaustive"`) to find the best model according to BIC and the best model according to AIC. Examine the `summary()` output from each model, and decide which one you like better.

```
> source("sum.step.q")
> fuel.result_stepwise(fuel[,names(fuel)!="Fuel"],fuel$Fuel,
method="exhaustive",nbest=2)
> sum.step(fuel.result, fuel, "Fuel", includeAIC=T, includeR2=T, sort=0)
         BIC    AIC    R2   R2a Pop Tax Nlic Inc Road Fuelc Dlic
1(#1) -301.1 -299.2 0.337 0.324  0   0    0   0    0     0    1
1(#2) -307.0 -305.1 0.161 0.143  1   0    0   0    0     0    0
2(#1) -298.0 -295.1 0.460 0.437  0   0    0   1    0     0    1
2(#2) -301.1 -298.2 0.388 0.362  1   0    0   0    0     0    1
3(#1) -296.9 -293.1 0.521 0.490  1   0    0   0    1     0    1
3(#2) -297.1 -293.3 0.517 0.486  0   0    1   0    0     1    1
```



```
4(#1) -297.2 -292.4 0.554 0.514  0   0    1   0    1     1    1
4(#2) -297.7 -292.9 0.544 0.503  1   0    0   1    1     0    1
5(#1) -298.6 -292.9 0.564 0.514  0   0    1   1    1     1    1
5(#2) -299.1 -293.3 0.556 0.506  1   0    1   0    1     1    1
6(#1) -300.5 -293.8 0.566 0.506  1   0    1   1    1     1    1
6(#2) -300.6 -294.0 0.564 0.503  0   1    1   1    1     1    1
7(#1) -302.5 -294.9 0.567 0.494  1   1    1   1    1     1    1
```

From the above stepwise procedure, we conclude that the best model is

- by BIC:

```
         BIC    AIC    R2   R2a Pop Tax Nlic Inc Road Fuelc Dlic
3(#1) -296.9 -293.1 0.521 0.490  1   0    0   0    1     0    1
```

- by AIC:

```
         BIC    AIC    R2   R2a Pop Tax Nlic Inc Road Fuelc Dlic
4(#1) -297.2 -292.4 0.554 0.514  0   0    1   0    1     1    1
```

Because the two model choices are not identical, we further consider the adjusted R^2 as a point of reference. According to adjusted R^2,

```
         BIC    AIC    R2   R2a Pop Tax Nlic Inc Road Fuelc Dlic
4(#1) -297.2 -292.4 0.554 0.514  0   0    1   0    1     1    1
5(#1) -298.6 -292.9 0.564 0.514  0   0    1   1    1     1    1
```

either model 4(#1) or 5(#1) should be chosen. Thus, the adjusted R^2 model can be chosen to agree with the model chosen by AIC. However, due to our inability to choose the best model solely based upon BIC or AIC (or even adjusted R^2), we investigate further and consider the summary information for the models:

- Model 1: $\texttt{Fuel} = \beta_0 + \beta_1\texttt{Pop} + \beta_2\texttt{Road} + \beta_3\texttt{Dlic} + \epsilon$

  The associated summary and diagnostics for this model are provided below and in Figure 7, respectively.

```
Call: lm(formula = Fuel ~ Pop + Road + Dlic, data = fuel)
Residuals:
    Min     1Q Median    3Q   Max
 -266.7 -47.77 -6.795 42.36 276.6

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  68.0424  126.3727    0.5384   0.5929
        Pop  -0.0152    0.0038   -4.0339   0.0002
       Road  16.0431    4.4991    3.5659   0.0009
       Dlic   8.4212    2.1705    3.8798   0.0003

Residual standard error: 81.74 on 46 degrees of freedom
Multiple R-Squared: 0.5209
F-statistic: 16.67 on 3 and 46 degrees of freedom, the p-value is 1.807e-07

Correlation of Coefficients:
     (Intercept)     Pop    Road
 Pop -0.3527
Road  0.0824     -0.6700
```

```
Dlic -0.9855      0.3664 -0.1969
```

- Model 2: $\texttt{Fuel} = \beta_0 + \beta_1\texttt{Nlic} + \beta_2\texttt{Road} + \beta_3\texttt{Fuelc} + \beta_4\texttt{Dlic} + \epsilon$

  The associated summary and diagnostics for this model are provided below and in Figure 8, respectively.

```
Call: lm(formula = Fuel ~ Nlic + Road + Fuelc + Dlic, data = fuel)
Residuals:
    Min     1Q Median     3Q   Max
 -262.5 -39.43 -5.527 14.98 273.9

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)   17.2158  119.8788    0.1436    0.8864
       Nlic   -0.0956    0.0340   -2.8126    0.0073
       Road    9.8839    5.1260    1.9282    0.0602
      Fuelc    0.0842    0.0408    2.0661    0.0446
       Dlic    9.4094    2.0548    4.5793    0.0000

Residual standard error: 79.73 on 45 degrees of freedom
Multiple R-Squared: 0.5541
F-statistic: 13.98 on 4 and 45 degrees of freedom, the p-value is 1.73e-07

Correlation of Coefficients:
      (Intercept)    Nlic    Road   Fuelc
 Nlic -0.0440
 Road  0.0227      0.4086
Fuelc -0.0094     -0.9807 -0.5288
 Dlic -0.9845      0.0278 -0.1312  0.0281
```

- Model 3: $\texttt{Fuel} = \beta_0 + \beta_1\texttt{Nlic} + \beta_2\texttt{Inc} + \beta_3\texttt{Road} + \beta_4\texttt{Fuelc} + \beta_5\texttt{Dlic} + \epsilon$

  The associated summary and diagnostics for this model are provided below and in Figure 9, respectively.

```
Call: lm(formula = Fuel ~ Nlic + Inc + Road + Fuelc + Dlic, data = fuel)
Residuals:
    Min    1Q Median    3Q Max
 -253.3 -39.8 -13.62 20.84 274

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)   85.7786  138.3903    0.6198    0.5386
       Nlic   -0.0775    0.0386   -2.0088    0.0507
        Inc  -24.3885   24.5825   -0.9921    0.3266
       Road    9.1970    5.1735    1.7777    0.0824
      Fuelc    0.0672    0.0442    1.5193    0.1358
       Dlic   10.0263    2.1471    4.6696    0.0000

Residual standard error: 79.74 on 44 degrees of freedom
Multiple R-Squared: 0.5638
F-statistic: 11.38 on 5 and 44 degrees of freedom, the p-value is 4.436e-07
```



```
Correlation of Coefficients:
      (Intercept)    Nlic     Inc    Road   Fuelc
 Nlic  0.2025
  Inc -0.4994     -0.4728
 Road -0.0473      0.2935  0.1338
Fuelc -0.2012     -0.9799  0.3879 -0.4311
 Dlic -0.6718      0.1603 -0.2896 -0.1632 -0.0875
```

Therefore, from the above output, we see that the residual error is minimized for Model 2: $\texttt{Fuel} = \beta_0 + \beta_1\texttt{Nlic} + \beta_2\texttt{Road} + \beta_3\texttt{Fuelc} + \beta_4\texttt{Dlic} + \epsilon$. However, due to the strong collinearity within this model between NLIC and FUELC, we should not feel comfortable accepting this model. Therefore, we choose the BIC model choice: $\texttt{Fuel} = \beta_0 + \beta_1\texttt{Pop} + \beta_2\texttt{Road} + \beta_3\texttt{Dlic} + \epsilon$ because this model contains variables that do not appear to be strongly collinear.

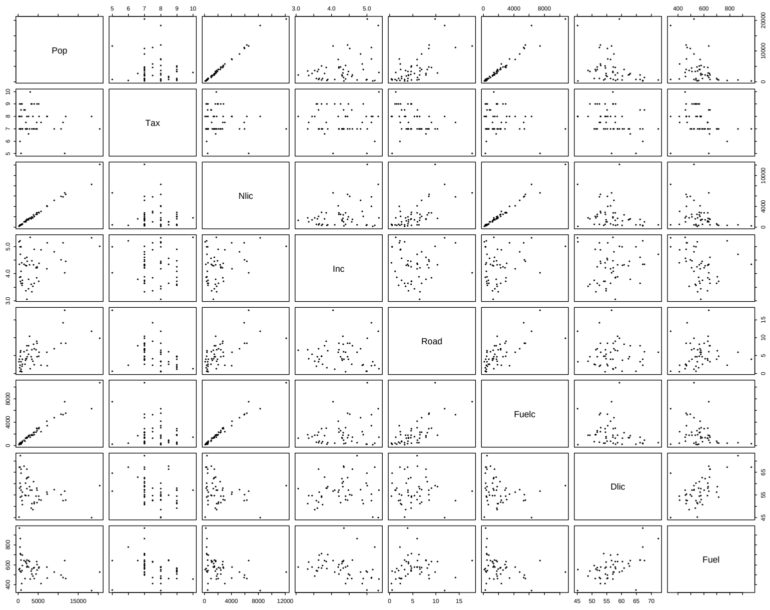

Figure 6: Pairwise scatterplot of variables in fuel-alr.dat dataset

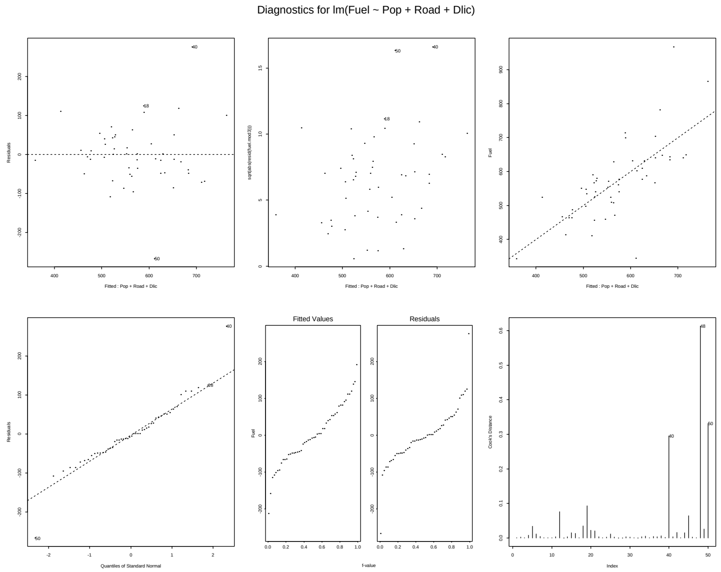

Figure 7: Diagnostics for `lm(Fuel   Pop + Road + Dlic)`

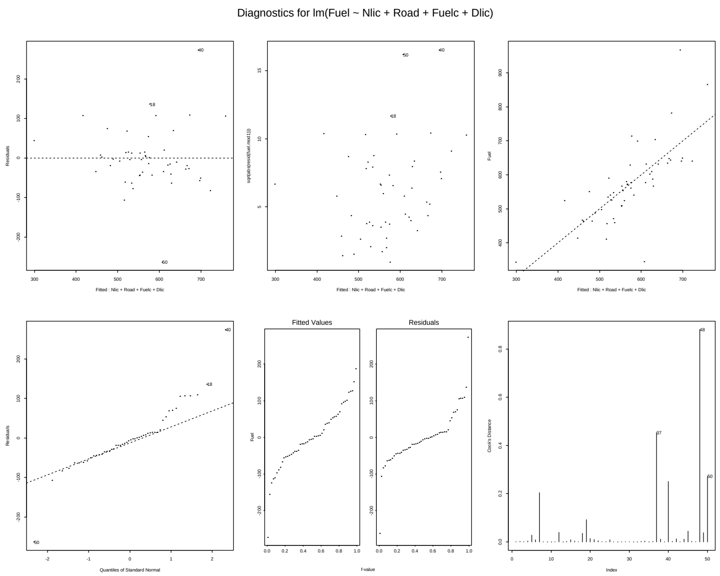

Figure 8: Diagnostics for `lm(Fuel    Nlic + Road + Fuelc + Dlic)`

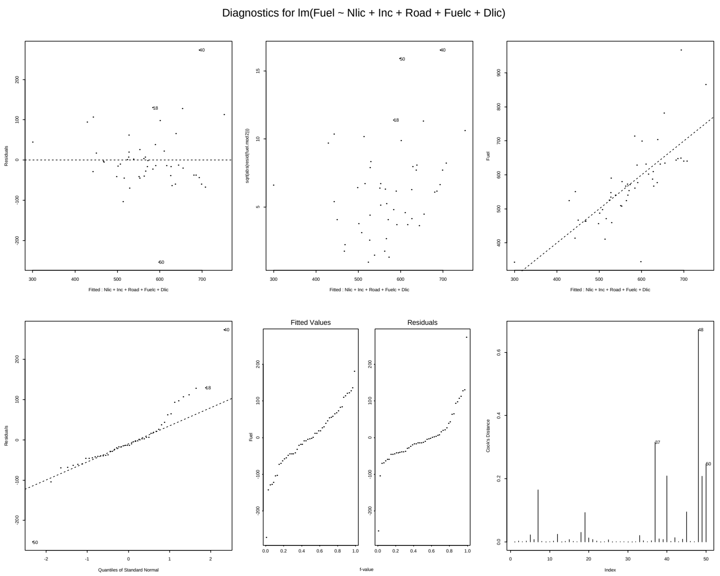

Figure 9: Diagnostics for `lm(Fuel    Nlic + Inc + Road + Fuelc + Dlic)`