Homework 5 Solutions
401 Fall 2001
Brian Junker and Kimberly Sellers

Tuesday October 25, 2001

1. RwG problem 15 chapter 3 page 104.

   First, a model needs to be found to help see if the oxygen uptake can be explained by both the relative head width and the type of salamender (either lunged or lungless). To begin, a simple regression analysis is preformed with the salamander type as a dummy variable.

```
> lm.sal1_lm(OxyUp~Type*Ratio, data=sal)
> summary(lm.sal1)

Call: lm(formula = OxyUp ~ Type * Ratio, data = sal)
Residuals:
 Min     1Q Median    3Q Max
  -6 -1.495 0.4031 2.023   3

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)   76.6667   10.6913     7.1709    0.0004
       Type   37.4981   12.9821     2.8884    0.0278
      Ratio   33.3333   69.1924     0.4817    0.6471
 Type:Ratio -325.9837   79.8187    -4.0841    0.0065

Residual standard error: 3.442 on 6 degrees of freedom
Multiple R-Squared: 0.9593
F-statistic: 47.18 on 3 and 6 degrees of freedom, the p-value is 0.0001449
```

   From the regression summary, I see that all of the coefficients are significant except for the Ratio coefficient. However, because the interation term between Type and Ratio is included in the regression, the Ratio term will remain in the model. (Recall that if x*y is included in the regression, then x and y should be included in the regression).

   Overall, the regression statistics look good, with a high R^2 and a pretty small residual standard error. As always, when performing a regression analysis, I next check the residuals to ensure that there are no violations of the regression assumptions.

   Figure 1 contains the residual plots. From the fitted vs. residuals plot, it appears that there may be some non-constant variance. Techniques will be addressed later in the course on how to address this

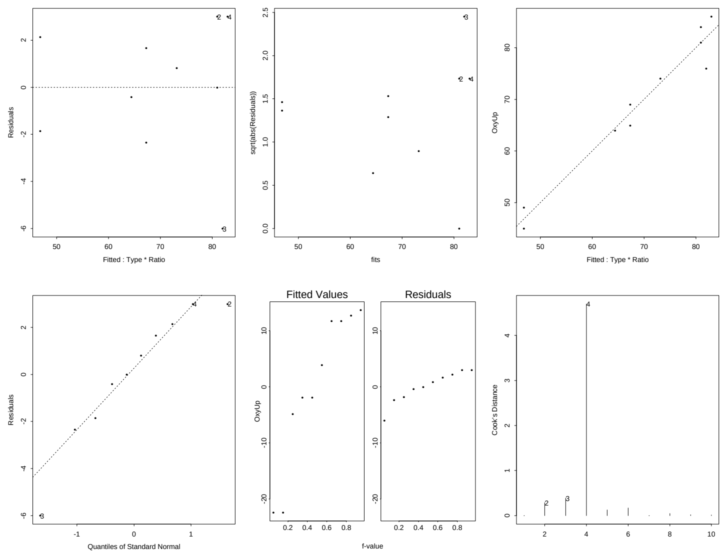

Figure 1: Residual Plot for the Salamander Data

non-conformance, however for now it will suffice to mention that this may be a problem. The normal probability plot looks marginal. It appears as if there may be a negative skew to the residuals. For now, continue on assuming that the residuals are sufficient for the analysis.

Returning to the regression summary, recall that Type is a dummy variable. It takes on values zero for lungless salamanders and one for lunged salamanders. Overall the regression equation is

OxyUp $= 76.6667 + 37.4981 *$ Type $+ 33.3333 *$ Ratio $- 325.9837 *$ Type * Ratio

Of course, this breakes into two regression lines based on the type of salamander. For **LUNGLESS** salamenders, the regression line is obtained by placing zero in for the type variable. Thus,

OxyUp.lungless $= 76.6667 + 33.3333 *$ Ratio

and for **LUNGED** salamanders, place a 1 into the Type variable to obtain the regression line.

OxyUp.lunged $= (76.6667 + 37.4981) + (33.3333 *$ Ratio $- 325.9837 *$ Ratio$) = 114.1648 - 292.6504 *$ Ratio

These results are demonstrated in Figure 2. In the graph, the 'Y' marks are for the lunged salamanders and the 'N' marks are for the lungless salamanders. In addition, both regression lines (one for the lunged and one for the lungless) are plotted on the graph.

```
Note the graph was created by the following commands
> sal
                  Type OxyUp Ratio
  G.porphyriticus    0    81  0.13
          P.ruber    0    84  0.13
     P.glutinosus    0    76  0.16
D.quadramaculatus    0    86  0.19
        A.tiginum    1    45  0.23
     A.talpoideum    1    49  0.23
         A.opacum    1    64  0.17
```

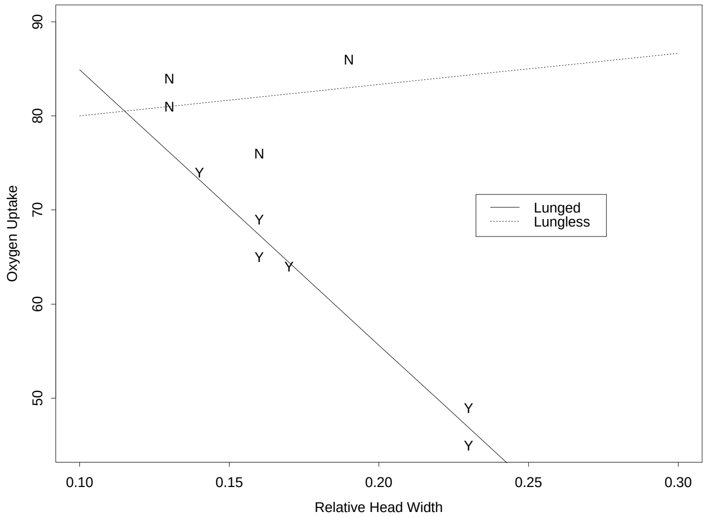

Figure 2: Demonstrative Plot for the Salamander Data

```
  A.macrodactylum    1    65  0.16
      A.maculatum    1    69  0.16
      R.olympicus    1    74  0.14
> sal$Type==0
 [1] T T T T F F F F F F

> lungless_sal[sal$Type==0,]
> lunged_sal[sal$Type==1,]
> lunged
                Type OxyUp Ratio
      A.tiginum    1    45  0.23
   A.talpoideum    1    49  0.23
       A.opacum    1    64  0.17
A.macrodactylum    1    65  0.16
    A.maculatum    1    69  0.16
    R.olympicus    1    74  0.14
> lungless
                  Type OxyUp Ratio
  G.porphyriticus    0    81  0.13
          P.ruber    0    84  0.13
     P.glutinosus    0    76  0.16
D.quadramaculatus    0    86  0.19

> seq(0.1, 0.3, 0.02)
 [1] 0.10 0.12 0.14 0.16 0.18 0.20 0.22 0.24 0.26 0.28 0.30
> x_seq(0.1, 0.3, 0.02)

> plot(lungless$Ratio, lungless$OxyUp, pch="N", xlim=c(0.1, 0.3),
```

```
      ylim=c(45,90), ylab="Oxygen Uptake", xlab="Relative Head Width")
> points(lunged$Ratio, lunged$OxyUp, pch="Y")
> lines(x, 114.1648-292.6504*x,lty=1 )
> lines(x, 76.667+33.333*x, lty=2)
> legend(locator(1), c("Lunged", "Lungless"), lty=1:2)
```

The plot clearly shows how different the relationship is between relative head width and oxygen uptake is for the lunged versus the lungless salamanders.

2. Return to the Prestige Data Set

   (a) The Diary

   To begin to determine what model to choose, the bullets outlined on the homework are followed. First, the scatterplot matrix and the correlation matrix are created. Figure 3 contains the scatterplot matrix.

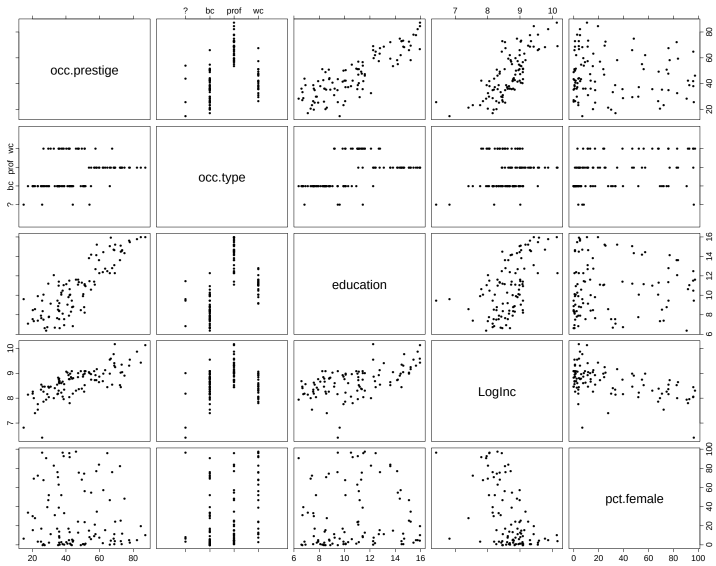

Figure 3: Scatterplot Matrix for New Prestige Data

First, examine the variables which may be correlated with the response, occ.prestige. It appears that occ.type may be correlated with occ.prestige. Realize that because occ.type is categorical, there is no real ordering to the plot. Thus, any ordering to the variables is valid, so we can construct this in a way such that it appears that occ.prestige rises with the occupation types ordered as (?, bc, wc, prof). There also appear to be strong linear components in the education and loginc variables with relation to the occ.prestige. It does not appear that pct.female has much of an effect on occ.prestige.

Next, examine any collinearities between the explanatory variables. From the scatterplot matrix, it appears that (education and loginc), and (loginc and pct.female) may be collinear. To quantify this, the correlation matrix is examined.

```
> cor(newprestige[,1:5])
             occ.prestige  occ.type  education     LogInc  pct.female
occ.prestige    1.0000000 0.3131787 0.85017689  0.7410561 -0.11833419
    occ.type    0.3131787 1.0000000 0.50827437  0.1807217  0.36846592
   education    0.8501769 0.5082744 1.00000000  0.5481051  0.06185286
      LogInc    0.7410561 0.1807217 0.54810506  1.0000000 -0.52254399
  pct.female   -0.1183342 0.3684659 0.06185286 -0.5225440  1.00000000
```

From the correlation matrix, it appears that education and loginc may be the strongest single predictors, with occ.type as a moderate predictor and pct.female a weak predictor. The matrix of correlations indicates that there are no pairwise correlations greater than 0.55, which indicates that there are no strong pairwise correlations.

To examine if there are multiple correlations present, the `collin` function is used.

```
> collin(newprestige[,2:5])
No R^2 >= 0.7
```

Note that only the variables occ.type, education, loginc and pct.female were included as parameters to the collin function. The variable occ.prestige was not included because it is the response and the hope is that it is collinear with the variables.

Thus, it appears that there is no collinearity in this data set (YEAH!).

Next, the regressions of `occ.prestige` on the other continuous variables is examined. Start with `occ.prestige` and `education`.

```
> summary(lm(occ.prestige~education, data=newprestige))

Call: lm(formula = occ.prestige ~ education, data = newprestige)
Residuals:
    Min      1Q Median    3Q   Max
 -26.04 -6.523 0.6611 6.743 18.16

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -10.7320   3.6771    -2.9186   0.0043
  education   5.3609   0.3320    16.1478   0.0000

Residual standard error: 9.103 on 100 degrees of freedom
Multiple R-Squared: 0.7228
F-statistic: 260.8 on 1 and 100 degrees of freedom, the p-value is 0
```

In this regression, we see that education is significant in the regression, and that the R^2 value is 0.7728. Thus, 72% of the variablility is explained by education. This is a fairly good predictor overall, however not all of the available data is being used, so we can probably do better.

Next evaluate the `occ.prestige` and `pct.female` regression.

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

Call: lm(formula = occ.prestige ~ pct.female, data = newprestige)
Residuals:
```

```
    Min      1Q Median    3Q   Max
 -33.44 -12.39 -4.126 13.03 39.18

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  48.6930   2.3076    21.1011   0.0000
 pct.female  -0.0642   0.0538    -1.1917   0.2362

Residual standard error: 17.17 on 100 degrees of freedom
Multiple R-Squared: 0.014
F-statistic: 1.42 on 1 and 100 degrees of freedom, the p-value is 0.2362
```

In this regression, pct.female is not needed. This confirms what we found previously that pct.female is not highly correlated with occ.prestige. A better regression is needed.

Finally, evaluate the `occ.prestige` and `LogInc` regression.

```
> summary(lm(occ.prestige~LogInc, data=newprestige))

Call: lm(formula = occ.prestige ~ LogInc, data = newprestige)
Residuals:
    Min      1Q Median    3Q   Max
 -19.11 -9.342 -1.218 8.101 30.45

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -139.8557   16.9542    -8.2490    0.0000
     LogInc   21.5564    1.9531    11.0368    0.0000

Residual standard error: 11.61 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
```

In this regression, LogInc is significant, and it explains 54% of the variability in the occ.prestige variable. It appears that the LogInc and the education variables will be needed in the final regression.

As suggested, next review the boxplots and regression of occ.prestige on the four dummy variables `Prof, WC, BC` and `Unknown`.

The boxplots are in figure 4. From these, it appears that the Prof occupational type has a different pattern from the other occupational types. The bc and wc appear to be roughly the same, with the medians similar and a lot of overlap in their IQR's. The unknown type appears to be a mixture of the bc and wc types.

Next, examine the regression of occ.prestige on the four dummy variables representing the occupational types.

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

Call: lm(formula = occ.prestige ~ prof + BC + WC + Unknown - 1, data =
newprestige)
```

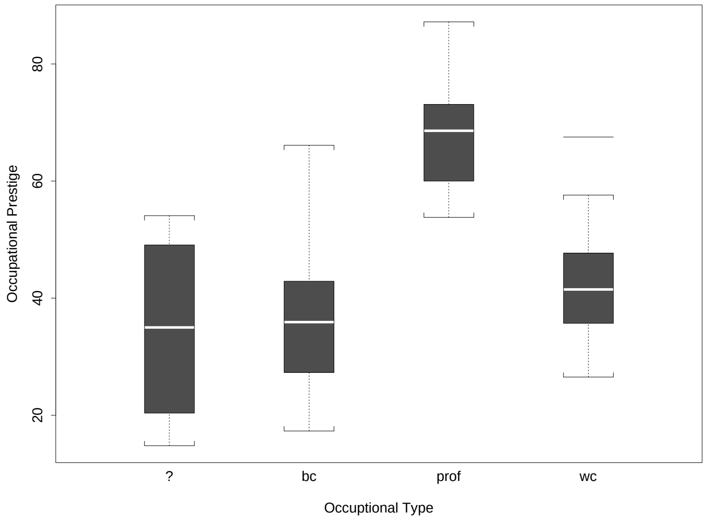

Figure 4: Boxplots comparing Occupational Type to Occupational Prestige

```
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
     BC 36.2067  1.5372    23.5543  0.0000
     WC 42.2435  2.1501    19.6471  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
```

From this, it appears that the profession variables are all significant in the regression. Recall that the R^2 reported by Splus is not accurate for the regression case where there is no intercept. Thus, I need to compute it by hand. I'll show how to do it in detail here, for the rest of the solutions I'll just state it.
The R^2 is 1-(RSS/TSS). To find RSS I use the anova command.

```
> anova(lm(occ.prestige~prof+BC+WC+Unknown-1, data=newprestige))
Analysis of Variance Table

Response: occ.prestige

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
```

```
     prof  1  138339.5 138339.5 1301.055 0.000000e+00
       BC  1   58991.5  58991.5  554.803 0.000000e+00
       WC  1   41043.8  41043.8  386.008 0.000000e+00
  Unknown  1    4823.3   4823.3   45.362 1.124443e-09
Residuals 98   10420.2    106.3
```

From this, I can see that RSS=10420.2. Recall that $TSS = \Sigma_i(Y_i - \bar{Y})^2$. I can compute this from Splus.

```
> sum((newprestige$occ.prestige-mean(newprestige$occ.prestige))^2)
[1] 29895.43
```

Thus, my R^2 should be 1-(10420.2/29895.43) = 0.6514451. Note this is significantly lower than the 0.9589 reported!. My adjusted R^2 is now 0.65144-(3/98)*(1-0.65144) = 0.6407698.
Now, to find the best model, first consider the best model with no interaction terms. First check the full model.

```
>summary(lm(occ.prestige~LogInc+education+pct.female+prof+BC+WC+Unknown-1,
  data=newprestige))

Call: lm(formula = occ.prestige ~ LogInc + education + pct.female +
prof + BC + WC + Unknown - 1, data = newprestige)
Residuals:
    Min      1Q Median    3Q   Max
 -13.74 -3.871 0.3553 4.265 16.75

Coefficients:
               Value Std. Error   t value  Pr(>|t|)
    LogInc   15.5466    2.0107     7.7321    0.0000
 education    2.7516    0.5712     4.8169    0.0000
pct.female    0.0936    0.0306     3.0590    0.0029
      prof -115.6447   17.3898    -6.6501    0.0000
        BC -121.4667   16.2479    -7.4759    0.0000
        WC -124.4908   16.7139    -7.4484    0.0000
   Unknown -112.0209   14.9748    -7.4806    0.0000

Residual standard error: 6.515 on 95 degrees of freedom
Multiple R-Squared: 0.9841
F-statistic: 840.1 on 7 and 95 degrees of freedom, the p-value is 0
```

This regression confirms the intuition developed above, that the variables LogInc, education, prof, BC, WC, Unknown are needed in the model. However, it also has that pct.female is needed in the model, which is counter to the intuition developed above. For now, I will keep it in this model since it appears significant here. The residual plots are in figure 5. The residual analysis appears good, with no evidence of non-constant variance and a good normal probability plot. The R^2 for this model is $1 - (4031.8/29895.43) = 0.865136$. The adjusted R^2 for this model is $0.8651 - (6/95)(1 - 0.8651) = 0.85658$.
This model indicates that there are four parallel lines, one for each of the occupational types. The coefficients on each of the occupation types represent the intercept term for the model with that coefficient.

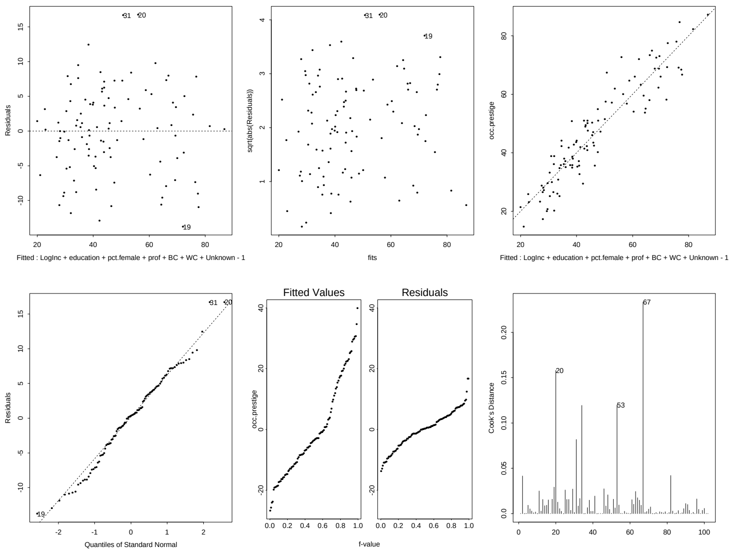

Figure 5: Residual Analysis for full model with no interactions

Note, that running a step wise procedure confirms that this is a good model. To run the stepwise procedure, one of the categorical variables needs to be removed because of the collinearity. In general, if one of the categorical variables needs to be included, then I would include all of the categorical variables. This is for ease of interpretation. The results of the step wise procedure are below. (Note that I had to use a stepwise with an intercept, as stepwise without an intercept gives errors. Thus the R^2 adjused values in the stepwise are correct as reported).

```
> sum.step(stepwise(newprestige[,3:8], newprestige$occ.prestige,
          method="exhaustive", nbest=2), newprestige,
          "occ.prestige",includeAIC=T,includeR2=T, sort=0)
          BIC    AIC    R2   R2a education LogInc pct.female prof BC WC
1(#1)  -373.6 -371.0 0.723 0.720         1      0          0    0  0  0
1(#2)  -388.2 -385.6 0.631 0.628         0      0          0    1  0  0
2(#1)  -350.7 -346.8 0.831 0.828         1      1          0    0  0  0
2(#2)  -365.7 -361.8 0.773 0.769         0      1          0    1  0  0
3(#1)  -348.2 -342.9 0.846 0.842         1      1          0    1  0  0
3(#2)  -349.2 -344.0 0.843 0.838         1      1          0    0  0  1
4(#1)  -347.8 -341.2 0.854 0.848         1      1          1    0  0  1
4(#2)  -348.6 -342.1 0.852 0.846         1      1          1    1  0  0
5(#1)  -346.5 -338.7 0.864 0.857         1      1          1    0  1  1
5(#2)  -349.3 -341.4 0.857 0.849         1      1          1    1  0  1
6(#1)  -348.6 -339.4 0.865 0.857         1      1          1    1  1  1
0(#2)  -436.8 -435.4 0.000 0.000         0      0          0    0  0  0
```

The models with the best AIC, BIC and R^2 are below.

```
          BIC    AIC    R2   R2a education LogInc pct.female prof BC WC
5(#1)  -346.5 -338.7 0.864 0.857         1      1          1    0  1  1
6(#1)  -348.6 -339.4 0.865 0.857         1      1          1    1  1  1
```

The first model and the second model are very close in their BIC and AIC values, and they both have the highest R^2 adjusted value. I prefer the use of model 6(#1) because it includes all of the categorical variable. Thus, the stepwise supports the above analysis.

Next, I consider any interactions. First, I created a series of graphs to see if the slopes may be different between occuptional types for the continuous variable vs. occuptional prestige. For example, in figure 6, I compared the educational level vs the occupational prestige for professionals (Y) vs. non-professionals. In this figure, it appears that there may be different intercepts, but approximately the same slope. I created these graphs for all the possible interaction terms to see if a graph could be found to indicate that an interaction may be needed.

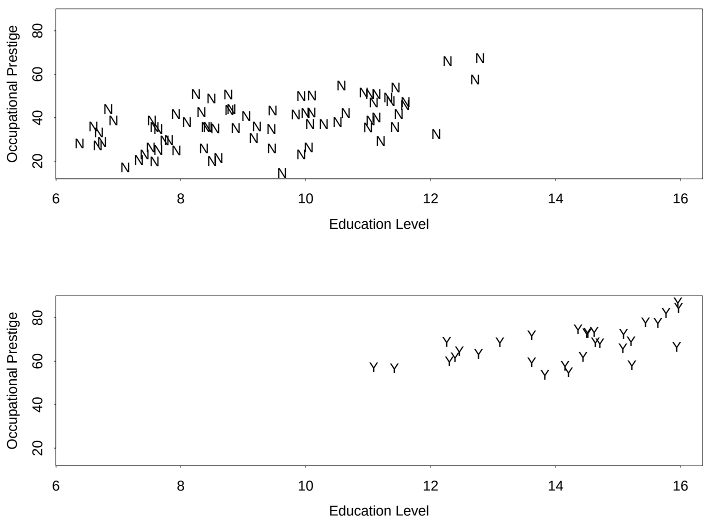

Figure 6: Professional vs. Education Interaction

Looking at all the graphs, there were no graphs which had a strong indication that an interaction term was needed. (though I must admit that this type of visual checking is not the most accurate). If I were to pick the three which were looked most as if they were needed, they would be the interactions between education and the four profession variables. Let me examine a regression including these interactions to see if they are significant.

The regression summary is below.

```
> summary(lm(occ.prestige~LogInc+education+pct.female+prof+BC+WC+Unknown
  + education:prof + education:BC + education:Unknown  -1,
  data=newprestige))

Call: lm(formula = occ.prestige ~ LogInc + education + pct.female +
prof + BC + WC +  Unknown + education:prof + education:BC +
education:Unknown - 1, data = newprestige)
Residuals:
    Min      1Q Median     3Q   Max
 -13.27 -4.437 0.6989 3.896 16.85

Coefficients:
```

```
                      Value Std. Error   t value  Pr(>|t|)
           LogInc   15.4415    1.9831     7.7864    0.0000
        education    4.6823    1.5275     3.0653    0.0029
       pct.female    0.0911    0.0307     2.9659    0.0038
             prof -105.6148   19.9626    -5.2906    0.0000
               BC -124.9149   16.8431    -7.4164    0.0000
               WC -144.7548   21.8876    -6.6136    0.0000
          Unknown  -78.1354   22.8846    -3.4143    0.0010
   education:prof   -2.5674    1.7398    -1.4757    0.1434
     education:BC   -1.4110    1.7011    -0.8295    0.4090
education:Unknown   -5.4656    2.4602    -2.2216    0.0288

Residual standard error: 6.415 on 92 degrees of freedom
Multiple R-Squared: 0.9851
F-statistic: 607.1 on 10 and 92 degrees of freedom, the p-value is 0
```

From the regression summary, it appears that the education:prof and education:BC variables are not needed. Removing these variables and re-running the regression yields the summary below.

```
> summary(lm(occ.prestige~LogInc+education+pct.female+prof+BC+WC+Unknown
  + education:Unknown  -1, data=newprestige), cor=F)

Call: lm(formula = occ.prestige ~ LogInc + education + pct.female +
prof + BC + WC +  Unknown + education:Unknown - 1, data = newprestige)
Residuals:
    Min      1Q Median    3Q   Max
 -14.05 -4.136 0.3466 4.014 16.74

Coefficients:
                      Value Std. Error   t value  Pr(>|t|)
           LogInc   15.5786    1.9845     7.8502    0.0000
        education    3.0148    0.5809     5.1896    0.0000
       pct.female    0.0962    0.0302     3.1841    0.0020
             prof -119.7302   17.3000    -6.9208    0.0000
               BC -124.0124   16.0929    -7.7061    0.0000
               WC -127.8029   16.5896    -7.7038    0.0000
          Unknown  -79.0985   22.9230    -3.4506    0.0008
education:Unknown   -3.8224    2.0344    -1.8789    0.0634

Residual standard error: 6.43 on 94 degrees of freedom
Multiple R-Squared: 0.9847
F-statistic: 755.1 on 8 and 94 degrees of freedom, the p-value is 0
```

This regression appears acceptable, with the education:Unknown on the borderline of being significant. However, the R^2 is recalculated as 1-(3885/29895.43) =0.870047. The adjusted R^2 is now $0.870-(7/94)(1-0.870) = 0.860$. Note that this adjusted R^2 is lower than the adjusted R^2 for the model without any interactions. Thus, I will keep the model with no interaction terms as my model as it has a better fit and it is a simpler model.

Finally, let me graph the residuals as suggested in the homework sheet for the full model with no interaction terms. The graphed residuals based on occuptaional type is in figure 7.

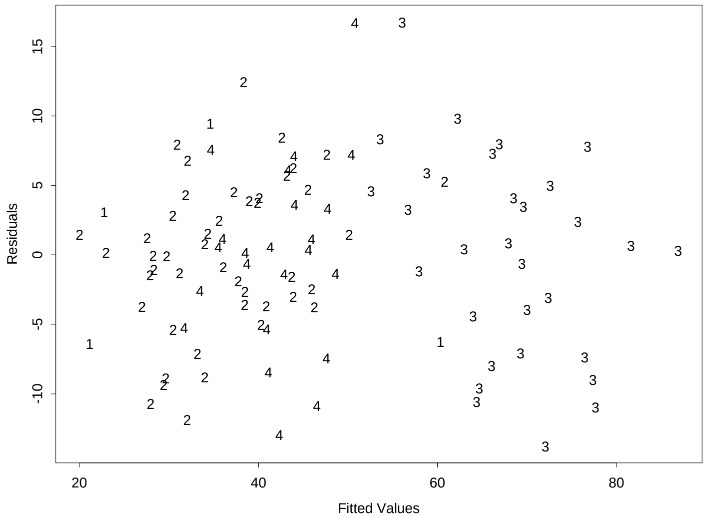

Figure 7: Residuals based on Occupational Type

This figure shows an interesting pattern in the residuals. The 1 (?), 2 (bc) and 4 (wc) residuals all appear to be scattered together with no pattern amoung them. However the 3 (prof) residuals appear to have a separate pattern, distinct from the other residuals. This could be a consequence of what was seen in the box plots, with the professionals having a different occupational prestige than WC, BC or Unknown. Out of curiosity, let's see if it may indicate that an interaction between prof and the other continuous variables may be necessary. The regression summary is below.

```
>  summary(lm(occ.prestige~LogInc+education+pct.female+prof+BC+WC+Unknown
   + education:prof + LogInc:prof + pct.female:prof  -1,
   data=newprestige), cor=F)

Call: lm(formula = occ.prestige ~ LogInc + education + pct.female + prof +
Unknown + education:prof + LogInc:prof + pct.female:prof - 1, data =
newprestige)
Residuals:
    Min      1Q Median     3Q   Max
 -12.92 -3.466  0.512 4.041 15.13

Coefficients:
                    Value Std. Error   t value  Pr(>|t|)
         LogInc   20.3611     2.2952    8.8710    0.0000
      education    2.5609     0.6658    3.8461    0.0002
     pct.female    0.1565     0.0352    4.4453    0.0000
           prof  -25.2398    28.9895   -0.8707    0.3862
             BC -162.1177    18.5420   -8.7433    0.0000
```

```
              WC -166.4262   19.1048    -8.7112    0.0000
         Unknown -148.7041   16.9859    -8.7545    0.0000
  education:prof    0.6052    1.1324     0.5345    0.5943
     LogInc:prof  -14.9978    4.0860    -3.6705    0.0004
pct.female:prof   -0.1824    0.0614    -2.9717    0.0038

Residual standard error: 6.135 on 92 degrees of freedom
Multiple R-Squared: 0.9863
F-statistic: 664.7 on 10 and 92 degrees of freedom, the p-value is 0
```

Here, it is clear that the education:prof variable is not needed. The regression without the education:prof interaction was examined.

```
>  summary(lm(occ.prestige~LogInc+education+pct.female+prof+BC+WC+Unknown+

Call: lm(formula = occ.prestige ~ LogInc + education + pct.female + prof +
Unknown + LogInc:prof + pct.female:prof - 1, data = newprestige)
Residuals:
    Min     1Q Median    3Q   Max
 -12.85 -3.568 0.1753 4.049 14.85

Coefficients:
                    Value Std. Error   t value  Pr(>|t|)
          LogInc   20.0690    2.2207     9.0374    0.0000
       education    2.7701    0.5365     5.1634    0.0000
      pct.female    0.1542    0.0348     4.4302    0.0000
            prof  -25.2266   28.8779    -0.8736    0.3846
              BC -161.3492   18.4150    -8.7618    0.0000
              WC -166.1402   19.0238    -8.7333    0.0000
         Unknown -148.3684   16.9090    -8.7745    0.0000
     LogInc:prof  -14.1043    3.7140    -3.7976    0.0003
pct.female:prof   -0.1767    0.0602    -2.9346    0.0042

Residual standard error: 6.111 on 93 degrees of freedom
Multiple R-Squared: 0.9863
F-statistic: 744.2 on 9 and 93 degrees of freedom, the p-value is 0
```

I next compute the R^2 as $1-(3462/29895.43) = 0.8841963$, and the adjusted R^2 for this model as $0.884-(8/94)(1-0.884) = 0.8741$. Figure 8 includes the residual plot for the above model. From the figure, the pattern is still evident with the prof residuals appearing separate from the other residuals. It is because the fitted values for occuptional prestige are larger for professionals than for WC, BC and Unknown. Thus, the professionals have their fitted values further to the right of the graph than the others.

At this point, I still believe I'll pick the model with no interaction terms. The difference in R^2 between the no interaction term and the model above isn't too big (only two percentage points). In addition, he model with no interactions is more easily interpretable. The residual analysis for both models is comparable, so that neither model violates the regression assumptions. Now is a good time to point out that there is no one correct model, only useful ones. Someone else may

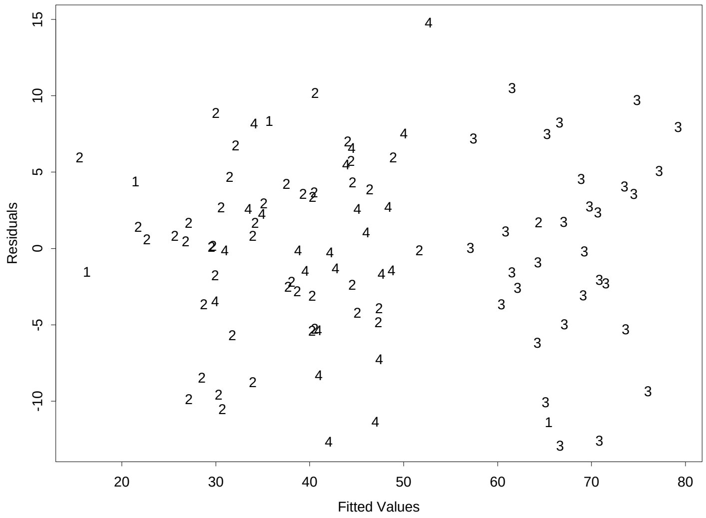

Figure 8: Residuals based on Occupational Type for Larger Model

make a different decision

I would like to do a stepwise procedure with the interaction terms to compare those results. However, it is not so clear how to do this, I can create new variables which contains the interactions. Examples of how to do this are below.

```
> newprestige$prof.inc_newprestige$LogInc*newprestige$prof
> newprestige$bc.inc_newprestige$LogInc*newprestige$BC
> newprestige$wc.inc_newprestige$LogInc*newprestige$WC
```

However, similar to the first stepwise procedure, I can not include all of the variables because of the multicolinearity. To run this stepwise, I'd need to run it on all variables except Unknown, and all interactions except for the interactions including the Unknown interactions. The problem, is that there is no easy way to extrapolate this to a full model. As above, I can include Unknown if any of the other categorical variables are needed. However there is no easy way to determine how to include the Unknown interactions. Thus, I will not do a stepwise procedure for this larger model.

Thus, the final model I have choosen is the model with no interaction terms and all of the non-interaction terms.

(b) A description of the model.

The final model chosen was occupational.prestige $= 15.54 *$ LogInc $+ 0.275 *$ education $+ 0.0936 *$ pct.female $- 115.64 *$ prof $- 121.46 *$ BC $- 124.49 *$ WC $- 112.02 *$ Unknown. The adjusted R^2 for this model was 0.856, with a residual standard error of 6.515 on 95 degrees of freedom. The F-statistic was 840.1 on 7 and 95 degrees of freedom. Recall that without an intercept, the F-statistc tests whether the regression is significantly better than a line at zero. All of these indicate that there was a good fit to the model. The residual are in figure 5. There are no signs of non-constant variance or non-independence. The normal probability plot looks good. Thus the regression assumptions appear to be met.

This model identifies four parallel lines for each of the occupational types. The four regression lines are below.

occupational.prestige.prof $= -115.64 + 15.54 * \text{LogInc} + 0.275 * \text{education} + 0.0936 * \text{pct.female}$

occupational.prestige.BC $= -121.46 + 15.54 * \text{LogInc} + 0.275 * \text{education} + 0.0936 * \text{pct.female}$

occupational.prestige.WC $= -124.49 + 15.54 * \text{LogInc} + 0.275 * \text{education} + 0.0936 * \text{pct.female}$

occupational.prestige.Unknown $= -112.02 + 15.54 * \text{LogInc} + 0.275 * \text{education} + 0.0936 * \text{pct.female}$

Thus, we see that the occupational prestige can be explained by the logarithm of income, the average education level, the average percent female in a profession and whether the profession is characterized by white collar, blue collar, professional or unknown. Everything else constant, a unit increase in income increases the occupational prestige by 15.54. Similearily, a unit increase in education increases the occupational prestige by 0.275, a unit increase in pct.female increases the occupational prestige by 0.0936. For each occupation type, the slope of the line is the same, however the intercepts vary. For professionals, the intercept is -115.64, it is -121.46 for blue collar workers, -124.49 for white collar workers and -112.02 for unknown profession types.

(c) Plots for residual checking.

The residual plots are in figure 5 and were discussed in the model description above. It is also interesting to note the interesting residual patterns, as seen in figure 7. As mentioned with the figure, it appears that the 1 (?), 2 (bc) and 4 (wc) residuals all appear to be scattered together with no pattern amoung them. However the 3 (prof) residuals appear to be to the right of the other residuals. This is because the prof occupation has higher occupational prestige than the other occupation types as seen in figure 4. Because the occupational prestige is higher, the fitted values are higher.