# Lab 6 Tutorial

*Nicholas Nagle*

*April 6, 2015*

## Homework

Your homework assignment is to analyze the prestige data, using "income" as the dependent variable. That is, develop a regression model that helps to explain the income in an occupation, using any of the other data in the the Prestige data set as a potential predictor. Describe your methodology from beginning to end. I have done this using Prestige as the dependent variable, and you should follow along with that before continuing.

## Tutorial

```
library(dplyr)
library(ggplot2)
library(GGally) # for ggpairs
library(car) # for the diagnostic plots and the Prestige data
```

First, load the data

```
data(Prestige)
```

Always look at the data data to see what the variables look like:

```
head(Prestige)
```

```
##                    education income women prestige census type
## gov.administrators      13.11  12351 11.16     68.8   1113 prof
## general.managers        12.26  25879  4.02     69.1   1130 prof
## accountants             12.77   9271 15.70     63.4   1171 prof
## purchasing.officers     11.42   8865  9.11     56.8   1175 prof
## chemists                14.62   8403 11.68     73.5   2111 prof
## physicists              15.64  11030  5.13     77.6   2113 prof
```

```
summary(Prestige)
```

```
##    education         income          women          prestige
##  Min.   : 6.38   Min.   :  611   Min.   : 0.00   Min.   :14.8
##  1st Qu.: 8.45   1st Qu.: 4106   1st Qu.: 3.59   1st Qu.:35.2
##  Median :10.54   Median : 5930   Median :13.60   Median :43.6
##  Mean   :10.74   Mean   : 6798   Mean   :28.98   Mean   :46.8
##  3rd Qu.:12.65   3rd Qu.: 8187   3rd Qu.:52.20   3rd Qu.:59.3
##  Max.   :15.97   Max.   :25879   Max.   :97.51   Max.   :87.2
##      census       type
##  Min.   :1113   bc  :44
```

```
##  1st Qu.:3120   prof:31
##  Median :5135   wc  :23
##  Mean   :5402   NA's: 4
##  3rd Qu.:8312
##  Max.   :9517
```

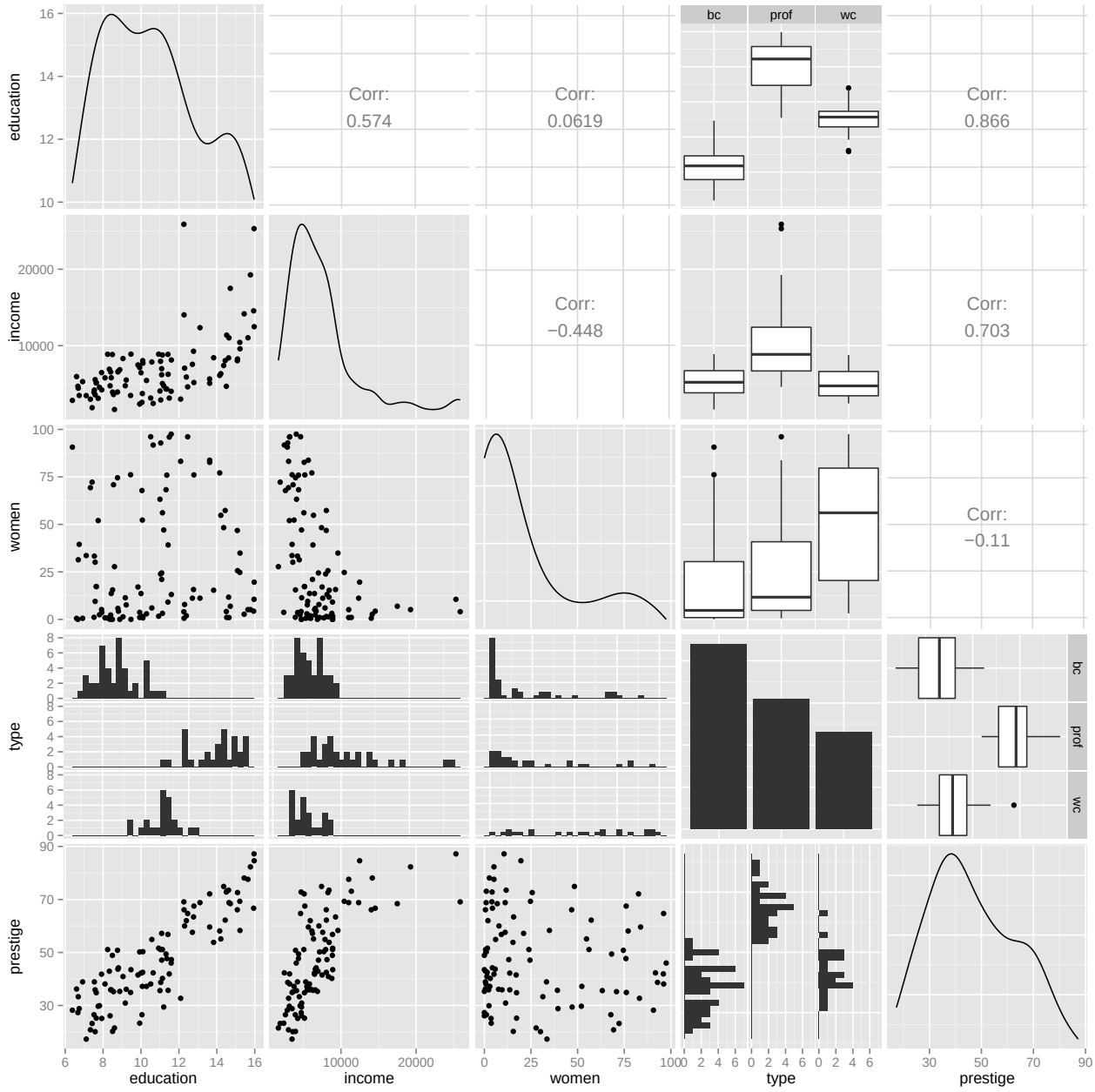
We have data organized by occupation, with the prestige, average income, average eduation, percent women in the occupation, and how it is classified according to "white collar", "blue collar" and "professional."

The first thing I should do is to theorize what the coefficients should be. I expect that income and education should be positively associated with prestige, because I think that our society values income, and I hope that it values education. I'm not certain what the relationship with percent women is. To be honest, I am concerned that it might be negative, i.e. that society devalues "women's work". There's also the type variable, of white collar, blue collar and professional. I expect these to be associated with prestige (blue collar, then white collar, then professional the highest). But I don't know *why* these are important. Perhaps it is related to income and education, in which case the type variable is not significant as it's own effect. Or perhaps there are other differences between the types that are separate from education and income. I don't have any clear expectations for the final relationship between occupational type and prestige.

If this were an academic paper, I would formally state my hypothesis at this point. Example: I hypothesize that there is a significant relationship between gender and prestige, even after controlling for education and income. Moreover, I expect this relationship to be negative.

Now, we can visually plot the data. We've used ggpairs in the past and I like that one. Note that I have adjusted the figure height and width in the .Rmd file from last week. The plot is now much more legible (though the font may be small now)

```
# Note the filtering out of missing values before plotting and
#  1) selecting to drop census, and put the dependent var on the bottom row
ggpairs(Prestige %>% filter(!is.na(type)) %>% select(education, income, women, type, prestige))
```




Looking at this, I would note the following:

1. Prestige, (the dependent variable) has a pretty normalish distribution.
2. There is a relationship between Prestige and education. I might be able to model it with a linear relation.
3. There is a relationship between Prestige and income. But it is nonlinear. It might be downward bending, or perhaps there is a kink at 10000. A log relationship of income might make it more linear.
4. There does not appear to be a relationship between Prestige and % women.
5. There is a strong relationship between prestige and type. In particular, professional jobs carry prestige.
6. There are many strong relationships between the covariates. Apart from women, everything has a correlation above 0.5.

That's a lot from just one graph. The correlations between X variables is concerning. In particular, I am thinking about that fact that just because a bivariate relationship appears linear or nonlinear, this may

change in a multivariate regression. Also, just because women does not appear to be related to prestige, it is (negatively) related to income. Thus, women may be a confounding variable on the true effect of income. I better include it in the regression "just in case"

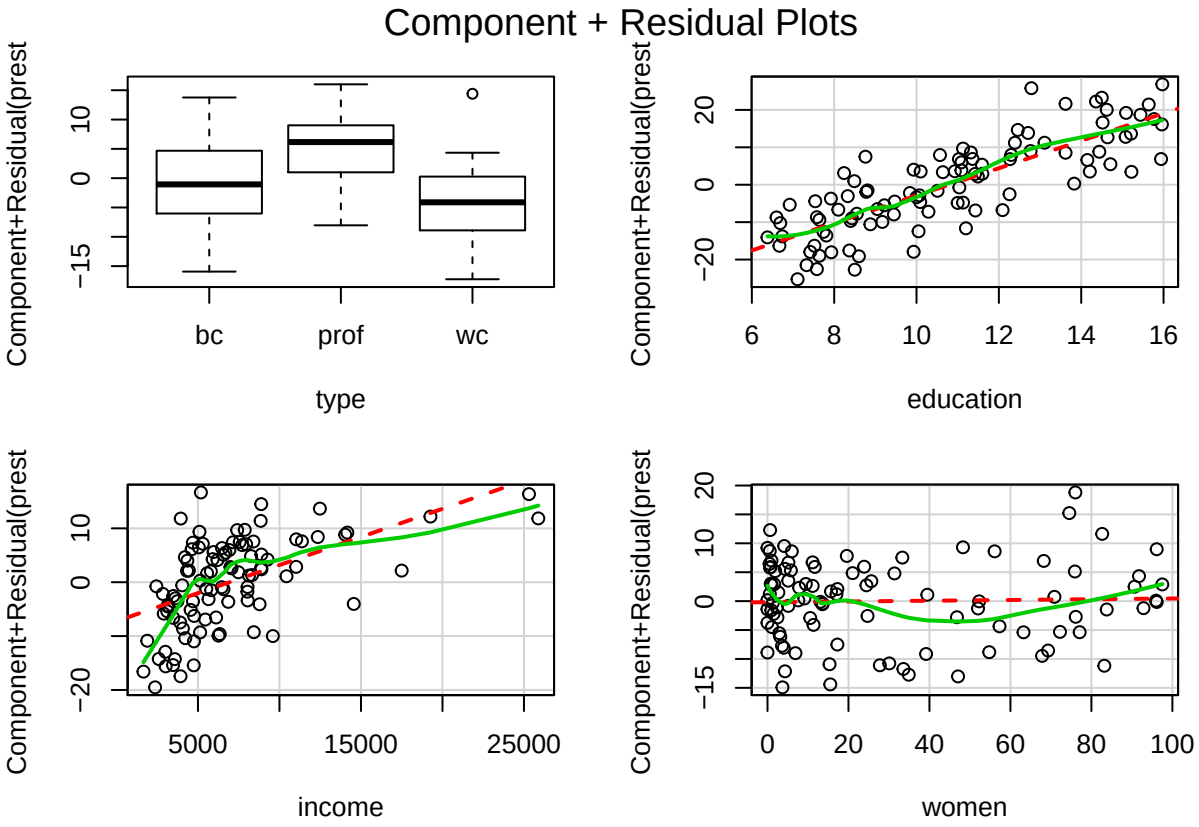
Run a regression with everything in it.

```
reg0 <- lm(prestige ~ type + education + income + women, data=Prestige)
summary(reg0)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + income + women, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.748  -4.482   0.312   5.248  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.813903   5.331156   -0.15  0.87899
## typeprof     5.905197   3.937700    1.50  0.13713
## typewc      -2.917072   2.665396   -1.09  0.27663
## education    3.662356   0.645830    5.67  1.6e-07 ***
## income       0.001043   0.000262    3.98  0.00014 ***
## women        0.006443   0.030378    0.21  0.83249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.13 on 92 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.826
## F-statistic: 93.1 on 5 and 92 DF,  p-value: <2e-16
```

Education and income are significant. But I remember that the relationship with education was potentially nonlinear. I can check that with a component plus residual plot:

```
crPlots(reg0)
```



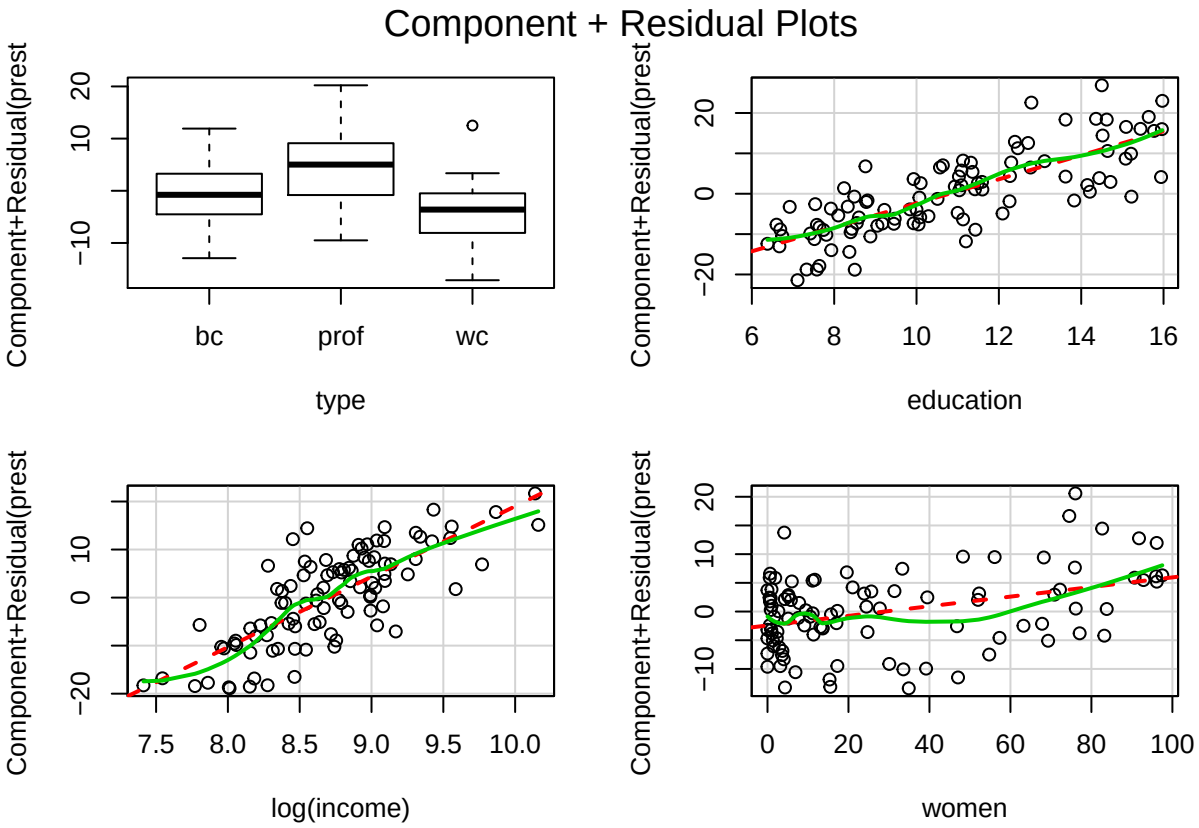

Oh, yes, there is definitely nonlinearity with income in the multivariate regression. Let's try to fix that. I can try to log income, to log prestige, or both. If I log prestige, that will change ALL of the relations, so let's avoide that if I can (unless it makes sense on scientific grounds to do so).

```
reg1 <- lm(prestige ~ type + education + log(income) + women, data=Prestige)
summary(reg1)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + log(income) + women,
##     data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.88  -4.06   0.55   4.21  16.64
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -115.6722    18.8018   -6.15  2.0e-08 ***
## typeprof       5.2919     3.5558    1.49    0.140
## typewc        -3.2160     2.4065   -1.34    0.185
## education      2.9738     0.6021    4.94  3.5e-06 ***
## log(income)   14.6552     2.3115    6.34  8.4e-09 ***
## women          0.0838     0.0322    2.60    0.011 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.44 on 92 degrees of freedom
```

```
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.865,  Adjusted R-squared:  0.858
## F-statistic:  118 on 5 and 92 DF,  p-value: <2e-16
```

```
crPlots(reg1)
```




The plots look much better. But does it make sense? The original regression said that a 1 dollar change in income causes a *beta* change in prestige. The log transformation says a 1% increase in income causes a *beta* change in prestige. 1% income change at $5000 is $250. A 1% change at $20000 is $1000. Does it make sense that you can "buy" relatively cheaply at first, but it gets harder and harder to buy? Yes, I think that makes sense. So I would stick with the log income regression.

Now, back to the regression. Women is now significant. Barely, but significant all the same.
% Women is a trick variable, though. There are many jobs with very few women, and then some jobs with 20-100% women. It appears that these jobs at the high end may have more prestige that we still aren't accounting for. This would suggest that there are increasing returns to prestige at the high end of % women, which is the opposite of what we saw with income.

We could fit a parabola to % women, to get that increasing return. Should that parabola have a bottom at women =0, or somewhere else?

```
# A fit with the parabola centered at women = 0
reg2 <- lm(prestige~type+education+log(income)+I(women^2), data=Prestige)
summary(reg2)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + log(income) + I(women^2),
```

```
##     data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.814  -4.224  -0.006   4.236  16.573
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.17e+02   1.71e+01   -6.83  9.2e-10 ***
## typeprof     5.52e+00   3.47e+00    1.59   0.1150
## typewc      -3.24e+00   2.33e+00   -1.39   0.1685
## education    2.93e+00   5.90e-01    4.96  3.2e-06 ***
## log(income)  1.49e+01   2.13e+00    6.99  4.2e-10 ***
## I(women^2)   1.07e-03   3.30e-04    3.23   0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.32 on 92 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.87,   Adjusted R-squared:  0.863
## F-statistic:  123 on 5 and 92 DF,  p-value: <2e-16
```

```
# A fit with the parabola centered somewhere else...
reg3 <- lm(prestige~type+education+log(income)+ women + I(women^2), data=Prestige)
summary(reg3)
```

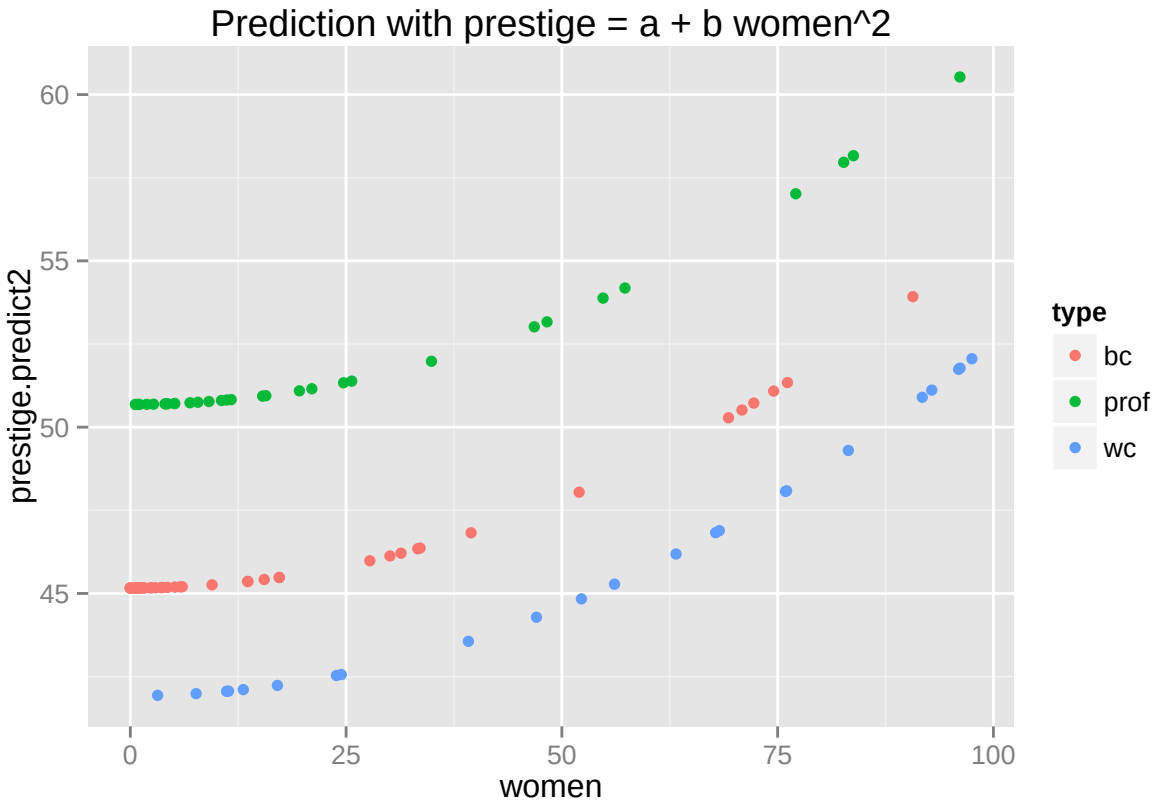
```
##
## Call:
## lm(formula = prestige ~ type + education + log(income) + women +
##     I(women^2), data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.911  -4.343   0.345   3.914  16.944
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.10e+02   1.87e+01   -5.87  7.1e-08 ***
## typeprof     6.03e+00   3.51e+00    1.72    0.089 .
## typewc      -2.84e+00   2.37e+00   -1.20    0.234
## education    2.97e+00   5.91e-01    5.02  2.6e-06 ***
## log(income)  1.41e+01   2.29e+00    6.17  1.9e-08 ***
## women       -8.21e-02   8.56e-02   -0.96    0.340
## I(women^2)   1.86e-03   8.92e-04    2.09    0.040 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.33 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.872,  Adjusted R-squared:  0.863
## F-statistic:  103 on 6 and 91 DF,  p-value: <2e-16
```

It can be helpful to see what these predictions look like.
To do that, we can use the `predict` function. But remember, in a regression, the coefficents are if everything

else is held fixed. So we can create prediction data where we hold education and income fixed. We could hold type fixed, but it is easy to hold fixed in a plot, so I won't.

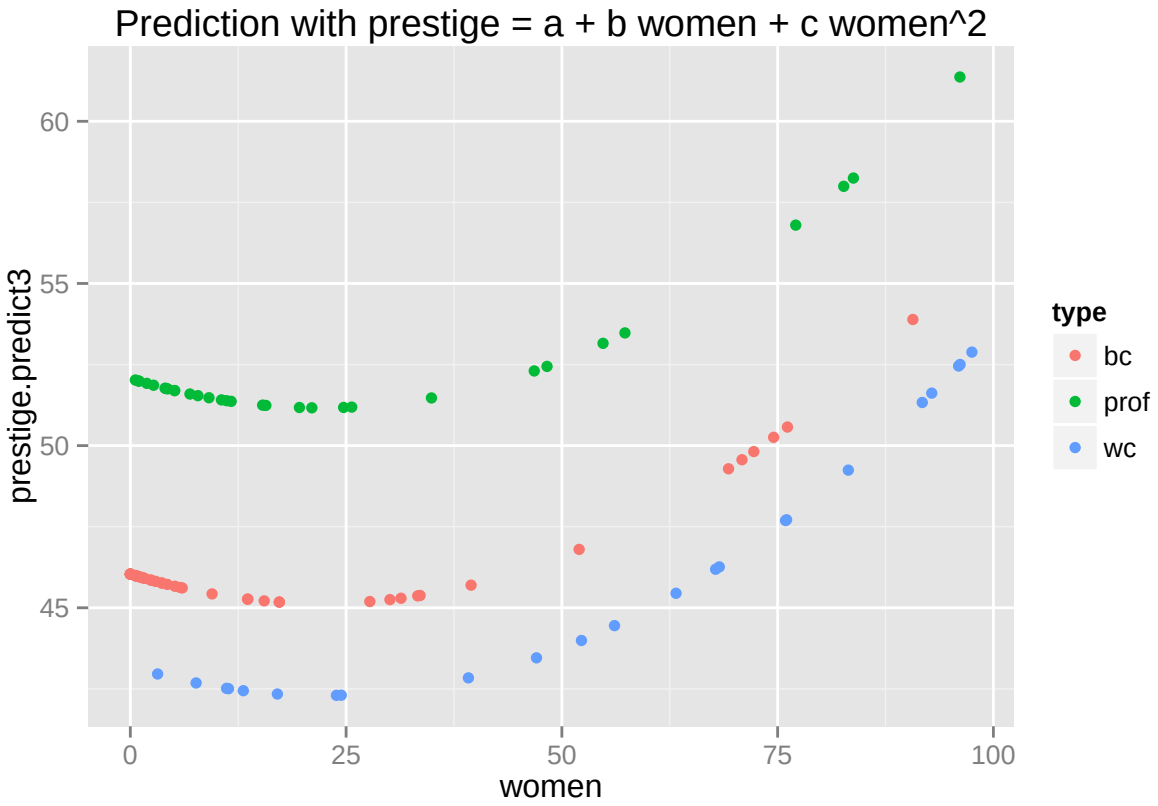
```
# First, copy the data, but then fix eduation and income to set values
pred.data <- Prestige %>% mutate(education=12, income=5000)
# Now, do the predictions
pred.data$prestige.predict2 <- predict(reg2, newdata = pred.data)
pred.data$prestige.predict3 <- predict(reg3, newdata = pred.data)
# And plot
ggplot(pred.data) + geom_point(aes(x=women, y=prestige.predict2, color=type))+
  labs(title = "Prediction with prestige = a + b women^2")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
ggplot(pred.data) + geom_point(aes(x=women, y=prestige.predict3, color=type)) +
  labs(title = "Prediction with prestige = a + b women + c women^2")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```




I am curious what are these professions with many women...

```
Prestige %>% arrange(desc(women))
```

```
##     education income women prestige census type
## 1       11.59   4036 97.51     46.0   4111   wc
## 2        9.46    611 96.53     25.9   6147 <NA>
## 3       10.51   3161 96.14     38.1   4175   wc
## 4       12.46   4614 96.12     64.7   3131 prof
## 5       11.49   3148 95.97     41.9   4113   wc
## 6       11.04   2901 92.86     38.7   4171   wc
## 7       10.64   2448 91.76     42.3   4133   wc
## 8        6.38   2847 90.67     28.2   8563   bc
## 9       13.62   5648 83.78     59.6   2731 prof
## 10      12.09   3016 83.19     32.7   4161   wc
## 11      13.62   5092 82.66     72.1   3137 prof
## 12      14.15   6112 77.10     58.1   2351 prof
## 13       9.45   3485 76.14     34.9   3135   bc
## 14      12.79   5180 76.04     67.5   3156   wc
## 15      11.36   4330 75.92     47.7   4143   wc
## 16       8.76   3942 74.54     50.8   8534   bc
## 17       7.42   1890 72.24     23.2   8221   bc
## 18       8.55   3617 70.87     35.2   9517   bc
## 19       7.33   3000 69.31     20.8   6162   bc
## 20      11.32   4348 68.24     49.4   4131   wc
## 21      10.05   2594 67.82     26.5   5137   wc
## 22      11.00   4075 63.23     35.6   4197   wc
## 23      14.44   8049 57.31     62.2   2311 prof
```

```
## 24      11.13   5052 56.10     51.1   4192   wc
## 25      14.21   6336 54.77     55.1   2331 prof
## 26      10.07   3739 52.27     37.2   4173   wc
## 27       7.74   3116 52.00     29.7   6121   bc
## 28      14.36   7405 48.28     74.9   2315 prof
## 29      11.20   4741 47.06     29.4   4191   wc
## 30      15.08   8034 46.80     66.1   2733 prof
## 31       6.74   3485 39.48     28.8   8278   bc
## 32      11.43   6259 39.17     35.7   4193   wc
## 33      15.22   9593 34.89     58.3   2391 prof
## 34       7.11   3472 33.57     17.3   6191   bc
## 35       7.54   4199 33.30     38.9   8213   bc
## 36       6.69   4443 31.36     33.3   8267   bc
## 37       7.58   3582 30.08     20.1   6193   bc
## 38       8.60   1656 27.75     21.5   7182   bc
## 39      15.09   8258 25.65     72.6   2133 prof
## 40      15.21  10432 24.71     69.3   3151 prof
## 41      11.09   6992 24.44     47.1   5172   wc
## 42      11.03   7956 23.88     51.1   5191   wc
## 43      11.09   6197 21.03     57.2   3314 prof
## 44      15.97  12480 19.59     84.6   2711 prof
## 45       7.64   5134 17.26     25.2   8215   bc
## 46       7.64   5134 17.26     34.8   8215   bc
## 47       9.84   7482 17.04     41.5   5130   wc
## 48      12.77   9271 15.70     63.4   1171 prof
## 49       8.50   3930 15.51     20.2   6123   bc
## 50      13.83   8425 15.33     53.8   2183 prof
## 51       8.43   5811 13.62     35.9   8513   bc
## 52      10.00   6462 13.58     42.2   9511   bc
## 53      11.60   8131 13.09     47.3   5171   wc
## 54      14.62   8403 11.68     73.5   2111 prof
## 55       9.17   4761 11.37     30.9   4153   wc
## 56      13.11  12351 11.16     68.8   1113 prof
## 57      12.71   7562 11.15     57.6   3337   wc
## 58      15.96  25308 10.56     87.2   3111 prof
## 59       7.58   5562  9.47     35.9   9171   bc
## 60      11.42   8865  9.11     56.8   1175 prof
## 61      11.44   8206  8.13     54.1   3373 <NA>
## 62      12.30   7059  7.83     60.0   2163 prof
## 63       9.22   5511  7.62     36.1   4172   wc
## 64       9.62    918  7.00     14.8   5143 <NA>
## 65      14.71  17498  6.91     68.4   3117 prof
## 66      10.57   7869  6.01     54.9   6141   bc
## 67       8.78   6573  5.78     43.7   8515   bc
## 68       7.92   6477  5.17     41.8   8335   bc
## 69      15.64  11030  5.13     77.6   2113 prof
## 70      15.77  19263  5.13     82.3   2343 prof
## 71      15.94  14558  4.32     66.7   3115 prof
## 72       8.81   6686  4.28     44.2   8313   bc
## 73      14.50   4686  4.14     72.8   2511 prof
## 74      12.26  25879  4.02     69.1   1130 prof
## 75       9.93   2370  3.69     23.3   5145   bc
## 76       6.84   3643  3.60     44.1   7112 <NA>
## 77       7.93   4224  3.59     25.1   9173   bc
```

```
## 78      11.13   8780  3.16     40.2   5133   wc
## 79      10.29   5449  2.92     37.2   8537   bc
## 80      15.44  14163  2.69     78.1   2141 prof
## 81       7.81   4549  2.46     29.9   8785   bc
## 82       8.40   6565  2.30     35.9   8333   bc
## 83      12.39   5902  1.91     62.0   2161 prof
## 84      10.93   8891  1.65     51.6   6112   bc
## 85      10.09   8043  1.50     42.5   8311   bc
## 86       9.05   8316  1.34     40.9   8731   bc
## 87       7.52   3910  1.09     26.5   8798   bc
## 88      14.52  11377  1.03     73.1   2143 prof
## 89       9.93   7147  0.99     50.2   8733   bc
## 90      14.64  11023  0.94     68.8   2153 prof
## 91       8.10   5795  0.81     38.1   8581   bc
## 92      10.10   7716  0.78     50.3   8582   bc
## 93       8.24   8880  0.65     51.1   8780   bc
## 94       8.33   6928  0.61     42.9   8791   bc
## 95      12.27  14032  0.58     66.1   9111 prof
## 96       6.92   5299  0.56     38.9   8781   bc
## 97       6.60   5959  0.52     36.2   8782   bc
## 98       9.47   8895  0.00     43.5   6111   bc
## 99       8.88   6860  0.00     35.3   7711   bc
## 100      6.67   4696  0.00     27.3   8715   bc
## 101      8.49   8845  0.00     48.9   9131   bc
## 102      8.37   4753  0.00     26.1   9313   bc
```

Ahhh, where are the row.names?
It turns out that the author of dplyr doesn't like row.names for reasons I don't agree with at all and so he just ignores them. We need to add a column with the row.names.

```
Prestige %>% mutate(occ=row.names(.))  %>% arrange(desc(women))
```

```
##     education income women prestige census type                    occ
## 1       11.59   4036 97.51     46.0   4111   wc              secretaries
## 2        9.46    611 96.53     25.9   6147 <NA>              babysitters
## 3       10.51   3161 96.14     38.1   4175   wc      telephone.operators
## 4       12.46   4614 96.12     64.7   3131 prof                  nurses
## 5       11.49   3148 95.97     41.9   4113   wc                 typists
## 6       11.04   2901 92.86     38.7   4171   wc            receptionsts
## 7       10.64   2448 91.76     42.3   4133   wc         tellers.cashiers
## 8        6.38   2847 90.67     28.2   8563   bc    sewing.mach.operators
## 9       13.62   5648 83.78     59.6   2731 prof  primary.school.teachers
## 10      12.09   3016 83.19     32.7   4161   wc              file.clerks
## 11      13.62   5092 82.66     72.1   3137 prof         physio.therapsts
## 12      14.15   6112 77.10     58.1   2351 prof               librarians
## 13       9.45   3485 76.14     34.9   3135   bc            nursing.aides
## 14      12.79   5180 76.04     67.5   3156   wc      medical.technicians
## 15      11.36   4330 75.92     47.7   4143   wc       computer.operators
## 16       8.76   3942 74.54     50.8   8534   bc       electronic.workers
## 17       7.42   1890 72.24     23.2   8221   bc                  canners
## 18       8.55   3617 70.87     35.2   9517   bc              bookbinders
## 19       7.33   3000 69.31     20.8   6162   bc               launderers
## 20      11.32   4348 68.24     49.4   4131   wc              bookkeepers
```

```
## 21      10.05   2594 67.82     26.5   5137   wc              sales.clerks
## 22      11.00   4075 63.23     35.6   4197   wc             office.clerks
## 23      14.44   8049 57.31     62.2   2311 prof               economists
## 24      11.13   5052 56.10     51.1   4192   wc           claim.adjustors
## 25      14.21   6336 54.77     55.1   2331 prof            social.workers
## 26      10.07   3739 52.27     37.2   4173   wc             postal.clerks
## 27       7.74   3116 52.00     29.7   6121   bc                     cooks
## 28      14.36   7405 48.28     74.9   2315 prof             psychologists
## 29      11.20   4741 47.06     29.4   4191   wc                collectors
## 30      15.08   8034 46.80     66.1   2733 prof secondary.school.teachers
## 31       6.74   3485 39.48     28.8   8278   bc         textile.labourers
## 32      11.43   6259 39.17     35.7   4193   wc             travel.clerks
## 33      15.22   9593 34.89     58.3   2391 prof    vocational.counsellors
## 34       7.11   3472 33.57     17.3   6191   bc                  janitors
## 35       7.54   4199 33.30     38.9   8213   bc                    bakers
## 36       6.69   4443 31.36     33.3   8267   bc           textile.weavers
## 37       7.58   3582 30.08     20.1   6193   bc        elevator.operators
## 38       8.60   1656 27.75     21.5   7182   bc              farm.workers
## 39      15.09   8258 25.65     72.6   2133 prof               biologists
## 40      15.21  10432 24.71     69.3   3151 prof              pharmacists
## 41      11.09   6992 24.44     47.1   5172   wc      real.estate.salesmen
## 42      11.03   7956 23.88     51.1   5191   wc                    buyers
## 43      11.09   6197 21.03     57.2   3314 prof        commercial.artists
## 44      15.97  12480 19.59     84.6   2711 prof       university.teachers
## 45       7.64   5134 17.26     25.2   8215   bc            slaughterers.1
## 46       7.64   5134 17.26     34.8   8215   bc            slaughterers.2
## 47       9.84   7482 17.04     41.5   5130   wc         sales.supervisors
## 48      12.77   9271 15.70     63.4   1171 prof               accountants
## 49       8.50   3930 15.51     20.2   6123   bc                bartenders
## 50      13.83   8425 15.33     53.8   2183 prof       computer.programers
## 51       8.43   5811 13.62     35.9   8513   bc              auto.workers
## 52      10.00   6462 13.58     42.2   9511   bc              typesetters
## 53      11.60   8131 13.09     47.3   5171   wc          insurance.agents
## 54      14.62   8403 11.68     73.5   2111 prof                 chemists
## 55       9.17   4761 11.37     30.9   4153   wc           shipping.clerks
## 56      13.11  12351 11.16     68.8   1113 prof        gov.administrators
## 57      12.71   7562 11.15     57.6   3337   wc       radio.tv.announcers
## 58      15.96  25308 10.56     87.2   3111 prof               physicians
## 59       7.58   5562  9.47     35.9   9171   bc              bus.drivers
## 60      11.42   8865  9.11     56.8   1175 prof       purchasing.officers
## 61      11.44   8206  8.13     54.1   3373 <NA>                 athletes
## 62      12.30   7059  7.83     60.0   2163 prof              draughtsmen
## 63       9.22   5511  7.62     36.1   4172   wc             mail.carriers
## 64       9.62    918  7.00     14.8   5143 <NA>                 newsboys
## 65      14.71  17498  6.91     68.4   3117 prof  osteopaths.chiropractors
## 66      10.57   7869  6.01     54.9   6141   bc         funeral.directors
## 67       8.78   6573  5.78     43.7   8515   bc          aircraft.workers
## 68       7.92   6477  5.17     41.8   8335   bc                   welders
## 69      15.64  11030  5.13     77.6   2113 prof              physicists
## 70      15.77  19263  5.13     82.3   2343 prof                 lawyers
## 71      15.94  14558  4.32     66.7   3115 prof             veterinarians
## 72       8.81   6686  4.28     44.2   8313   bc                machinists
## 73      14.50   4686  4.14     72.8   2511 prof                 ministers
## 74      12.26  25879  4.02     69.1   1130 prof          general.managers
```

```
## 75       9.93   2370  3.69     23.3   5145   bc service.station.attendant
## 76       6.84   3643  3.60     44.1   7112 <NA>                   farmers
## 77       7.93   4224  3.59     25.1   9173   bc              taxi.drivers
## 78      11.13   8780  3.16     40.2   5133   wc     commercial.travellers
## 79      10.29   5449  2.92     37.2   8537   bc        radio.tv.repairmen
## 80      15.44  14163  2.69     78.1   2141 prof               architects
## 81       7.81   4549  2.46     29.9   8785   bc            house.painters
## 82       8.40   6565  2.30     35.9   8333   bc       sheet.metal.workers
## 83      12.39   5902  1.91     62.0   2161 prof                 surveyors
## 84      10.93   8891  1.65     51.6   6112   bc                 policemen
## 85      10.09   8043  1.50     42.5   8311   bc           tool.die.makers
## 86       9.05   8316  1.34     40.9   8731   bc        electrical.linemen
## 87       7.52   3910  1.09     26.5   8798   bc    construction.labourers
## 88      14.52  11377  1.03     73.1   2143 prof            civil.engineers
## 89       9.93   7147  0.99     50.2   8733   bc               electricians
## 90      14.64  11023  0.94     68.8   2153 prof          mining.engineers
## 91       8.10   5795  0.81     38.1   8581   bc             auto.repairmen
## 92      10.10   7716  0.78     50.3   8582   bc        aircraft.repairmen
## 93       8.24   8880  0.65     51.1   8780   bc      construction.foremen
## 94       8.33   6928  0.61     42.9   8791   bc                  plumbers
## 95      12.27  14032  0.58     66.1   9111 prof                    pilots
## 96       6.92   5299  0.56     38.9   8781   bc                carpenters
## 97       6.60   5959  0.52     36.2   8782   bc                    masons
## 98       9.47   8895  0.00     43.5   6111   bc              firefighters
## 99       8.88   6860  0.00     35.3   7711   bc      rotary.well.drillers
## 100      6.67   4696  0.00     27.3   8715   bc        railway.sectionmen
## 101      8.49   8845  0.00     48.9   9131   bc           train.engineers
## 102      8.37   4753  0.00     26.1   9313   bc              longshoremen
```

We see that there are jobs such as secretaries, nurses, teachers, therapists and medical technicians with a relatively high degree of prestige.
The regression results tell us that these jobs have more prestige than should be expected based on income, education and type.

The stories that can be told from this data are facinating.

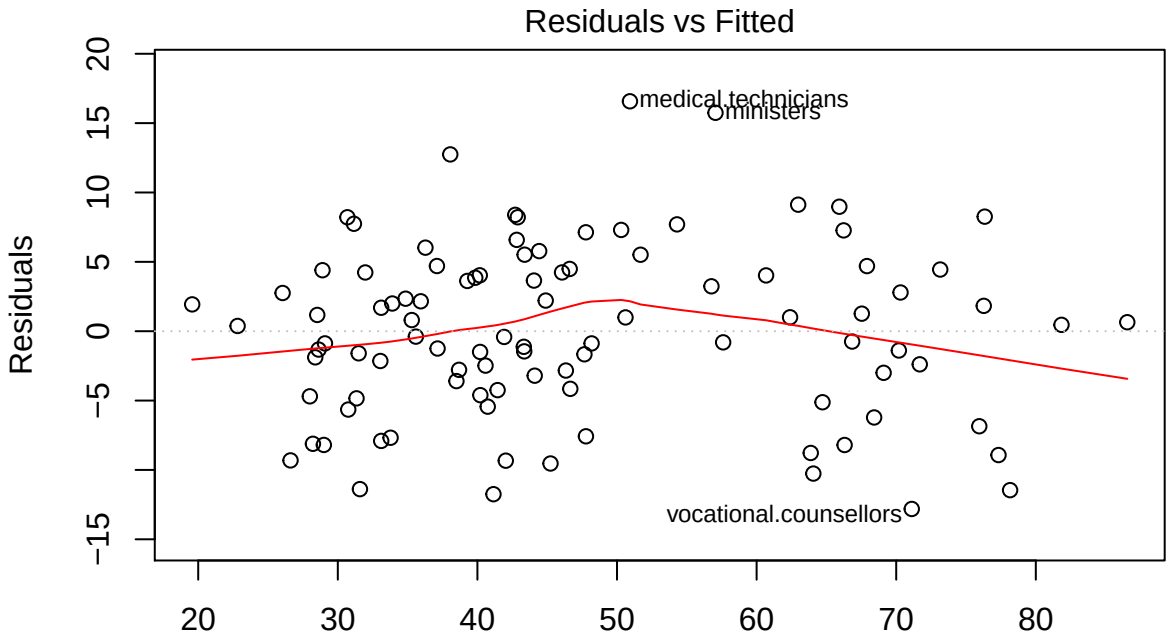
- Do these jobs have high prestige becuase they have many women (is it causal)?
- Perhaps these jobs are underpaid because they have many women (is it causal)?
- Are women self selecting into these jobs?
- Perhaps women value prestige higher than income, when compared to men?
- Or perhaps women don't necessarily value prestige higher, but are tacitly told, "you should expect lower income because there is a lot of prestige in these occupations (part of your"reward" is non-monetary)?

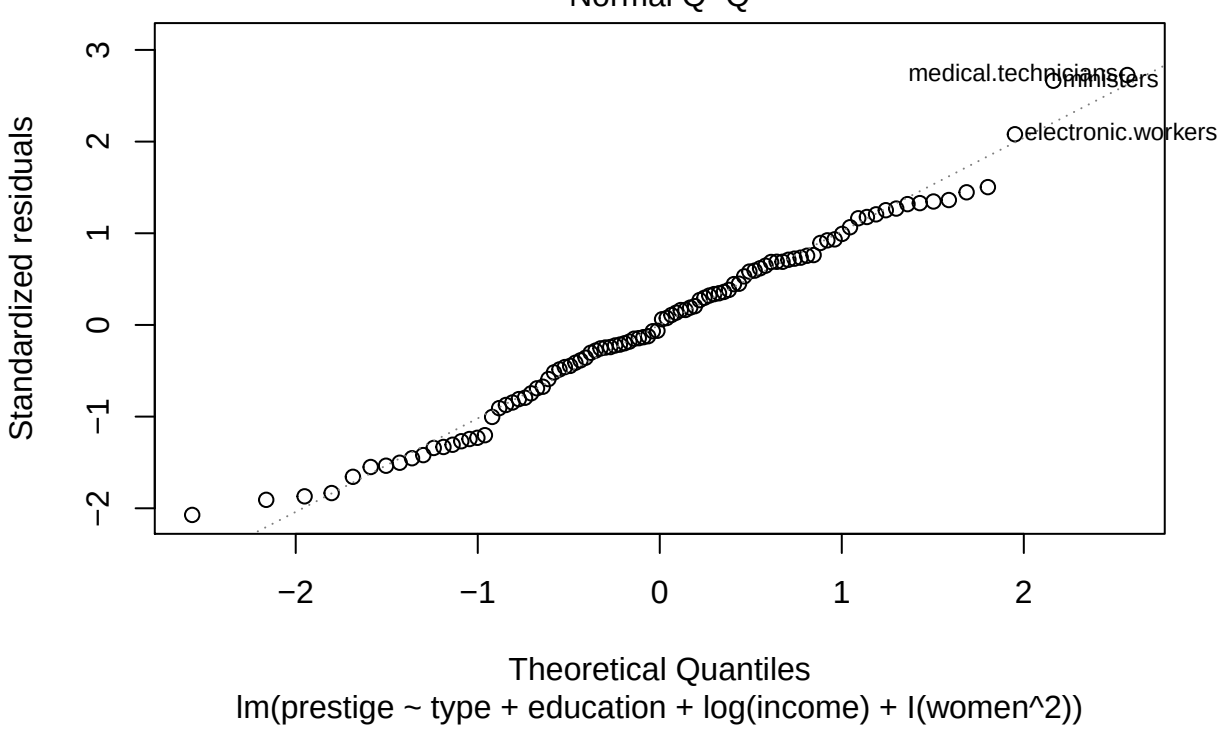
These data have suggestes interesting questions, but can not answer them all. Answering these questions would require different data, and perhaps different methods.

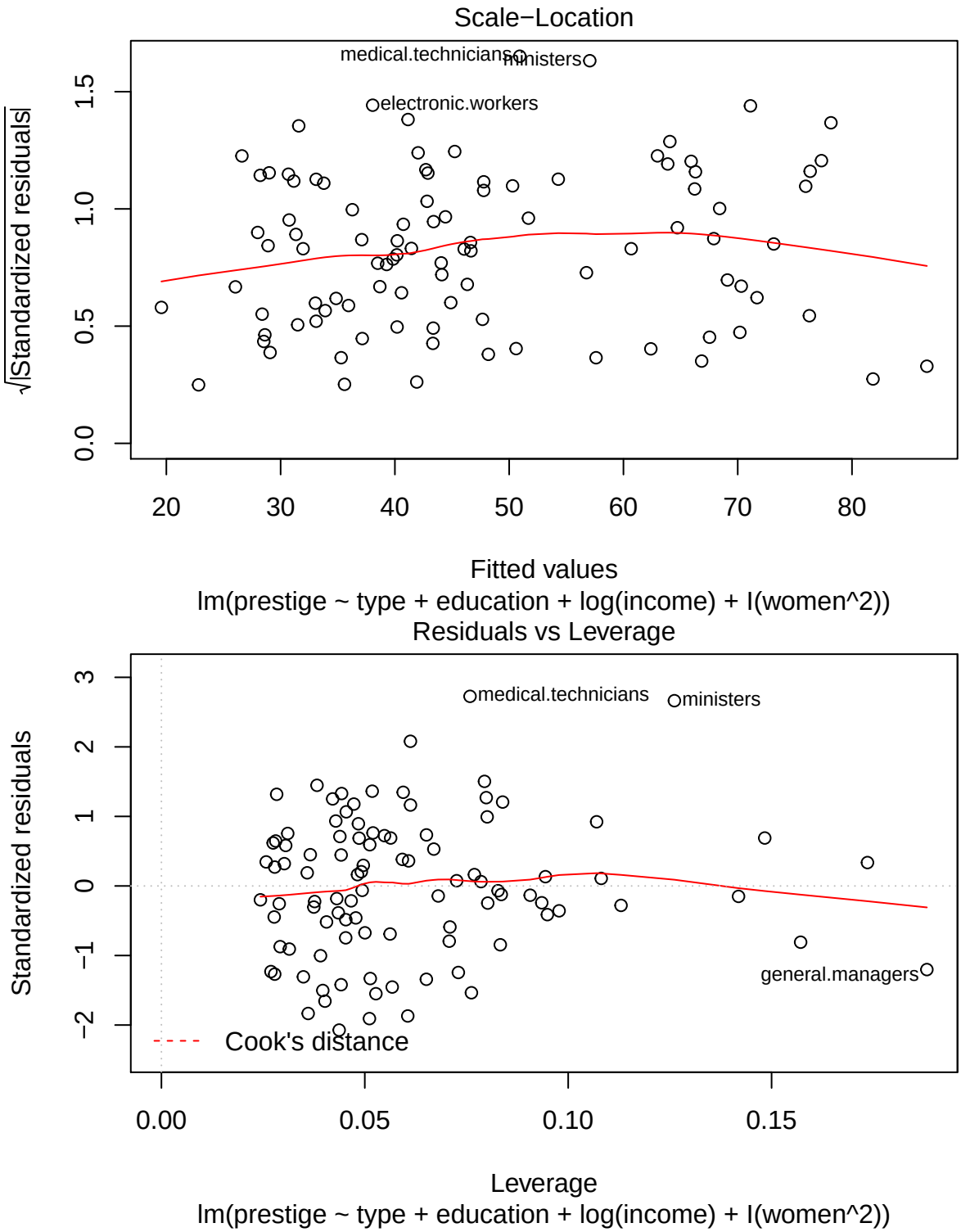
Before reporting these results, we need to do more diagnostic checking.
We should look at the residual plots...

Here are the standard residual plots:

```
plot(reg2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(prestige ~ type + education + log(income) + I(women^2))
medical.technicians
ministers
vocational.counsellors

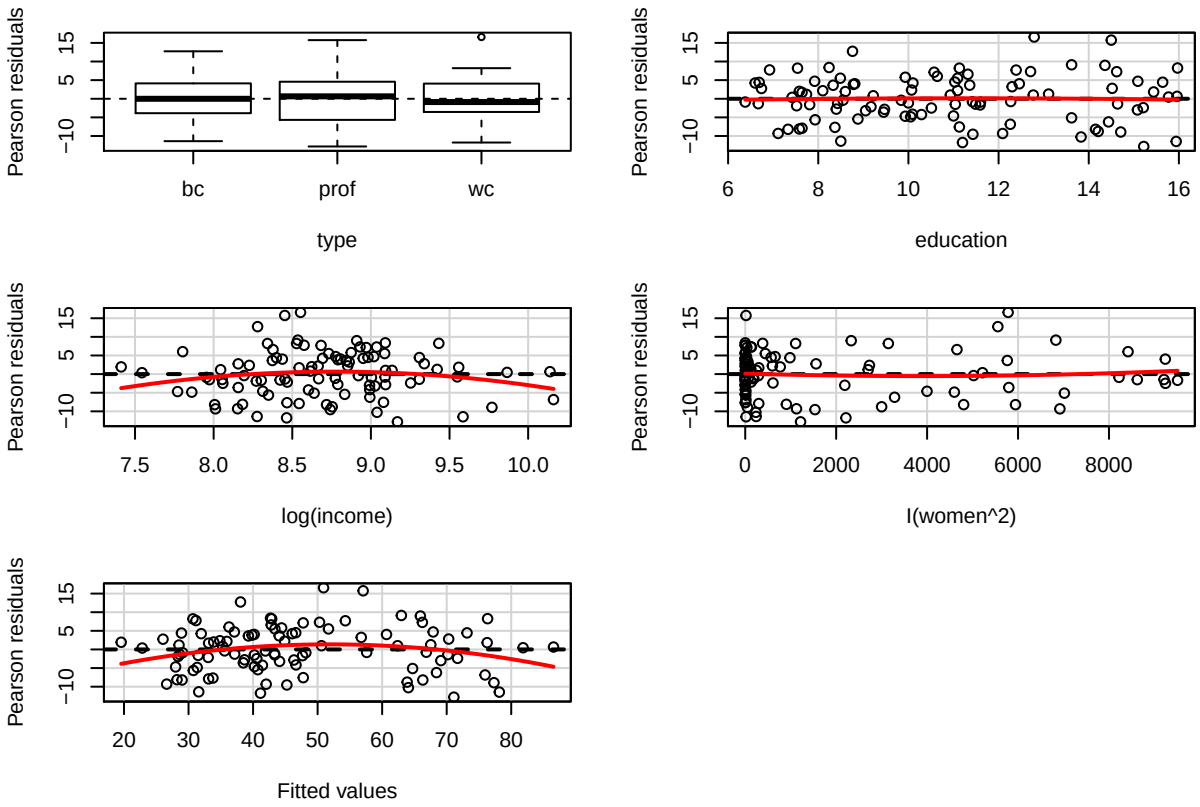

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(prestige ~ type + education + log(income) + I(women^2))
medical.technicians
ministers
electronic.workers




There is still some nonlinearity in the residuals somewhere. We may not be done yet...

There is a helpful function in the car package called `residualPlots`. Instead of just residual vs fitted values, it also gives you residual vs individual values. This is kinda like a residual plus component plot, but without the component. The residualPlots function also does a test for a curve to the residuals. There should be no curve.

```
residualPlots(reg2)
```




```
##             Test stat Pr(>|t|)
## type               NA       NA
## education      -0.202    0.841
## log(income)    -1.577    0.118
## I(women^2)      0.468    0.641
## Tukey test     -2.101    0.036
```

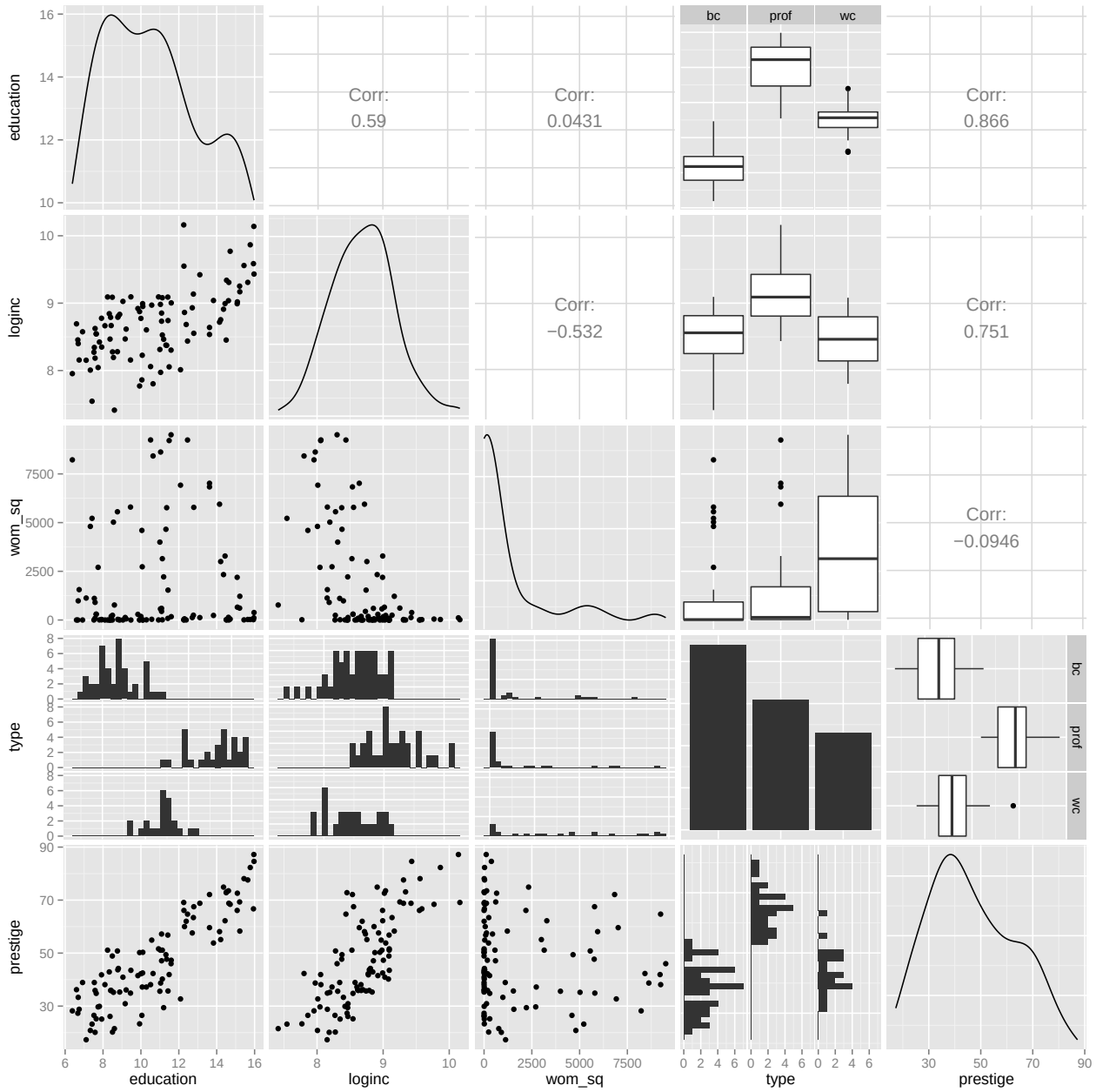
There is a curve, both with regard to the fitted values (we already knew that) and with regard to log income. Other than than. This suggests logincome as the possible source of the nonlinearity. I have two initial thoughts: (1) is that the log transformation of education may not have been sufficient or (2) log income may be interacting with other variables (even though we see no curve with regard to the other variables). We know that there are correlations between all of the X variables, and it's possible that they aren't complely accounted for by the linear regression so far.

At this point, we've done a few variable transformations. We never went back to the pair plot to see what this looked like.

```
ggpairs( Prestige %>% filter(!is.na(type)) %>% mutate(loginc = log(income), wom_sq = women^2) %>% selec
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




I immediately noticed something... there still appears to be zero relationship between women and prestige! We can check that...

```
summary(lm(prestige~I(women^2), data=Prestige))
```

```
##
```

```
## Call:
## lm(formula = prestige ~ I(women^2), data = Prestige)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.22 -12.08  -3.71  13.46  39.22
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.055958   2.027414   23.70   <2e-16 ***
## I(women^2)  -0.000666   0.000600   -1.11     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.2 on 100 degrees of freedom
## Multiple R-squared:  0.0122, Adjusted R-squared:  0.00227
## F-statistic: 1.23 on 1 and 100 DF,  p-value: 0.27
```

What is happening? It turns out that the weak, but probably significant effect of women is probably being masked by the larger effects of education and logincome. Bivariate scatter plots do not always show the effect in a multivariate setting.

Anyway, there are strong relations between loginc and women^2 and between loginc and education.

```
reg4 <- lm(prestige~type + log(income)*education + log(income)*I(women^2), data=Prestige)
summary(reg4)
```

```
##
## Call:
## lm(formula = prestige ~ type + log(income) * education + log(income) *
##     I(women^2), data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.454  -4.483   0.168   4.079  16.850
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.17e+02   5.61e+01   -3.86  0.00021 ***
## typeprof                 5.75e+00   3.45e+00    1.67  0.09843 .
## typewc                  -4.36e+00   2.40e+00   -1.82  0.07195 .
## log(income)              2.63e+01   6.44e+00    4.08  9.7e-05 ***
## education                1.20e+01   4.86e+00    2.47  0.01543 *
## I(women^2)               4.72e-03   6.55e-03    0.72  0.47353
## log(income):education   -1.02e+00   5.42e-01   -1.88  0.06291 .
## log(income):I(women^2) -4.44e-04   7.96e-04   -0.56  0.57832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.27 on 90 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.875,  Adjusted R-squared:  0.866
## F-statistic: 90.2 on 7 and 90 DF,  p-value: <2e-16
```

Whoa. There are significant relations between log income and education. But women has completely disappeared. It would be premature to drop women entirely. Consider this, there are two coefficients for women now. The t-test measures whether these are each zero, individually. But should they both be zero? That's a question for the F-test, which we can do with an anova. First, we run a regression like reg4, but without women entirely. Then we do an anova

```
reg5 <- lm(prestige~type + log(income)*education, data=Prestige)
anova(reg4, reg5)
```

```
## Analysis of Variance Table
##
## Model 1: prestige ~ type + log(income) * education + log(income) * I(women^2)
## Model 2: prestige ~ type + log(income) * education
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     90 3536
## 2     92 3966 -2      -430 5.47 0.0057 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It's significant, so women should be in the model. Having women there does increase our prediction ability.

An alternative to reg 5 is a model without the interaction between loginc and women...

```
reg6 <- lm(prestige~type + log(income)*education + I(women^2), data=Prestige)
summary(reg6)
```

```
##
## Call:
## lm(formula = prestige ~ type + log(income) * education + I(women^2),
##     data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.643  -4.586   0.237   4.032  16.080
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.02e+02   4.98e+01   -4.07   0.0001 ***
## typeprof               5.63e+00   3.43e+00    1.64   0.1037
## typewc                -4.15e+00   2.36e+00   -1.76   0.0816 .
## log(income)            2.47e+01   5.75e+00    4.29  4.4e-05 ***
## education              1.08e+01   4.34e+00    2.49   0.0147 *
## I(women^2)             1.07e-03   3.26e-04    3.27   0.0015 **
## log(income):education -8.89e-01   4.86e-01   -1.83   0.0707 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.24 on 91 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.875,  Adjusted R-squared:  0.867
## F-statistic:  106 on 6 and 91 DF,  p-value: <2e-16
```

```r
anova(reg6, reg4) # Compare with the big model
```

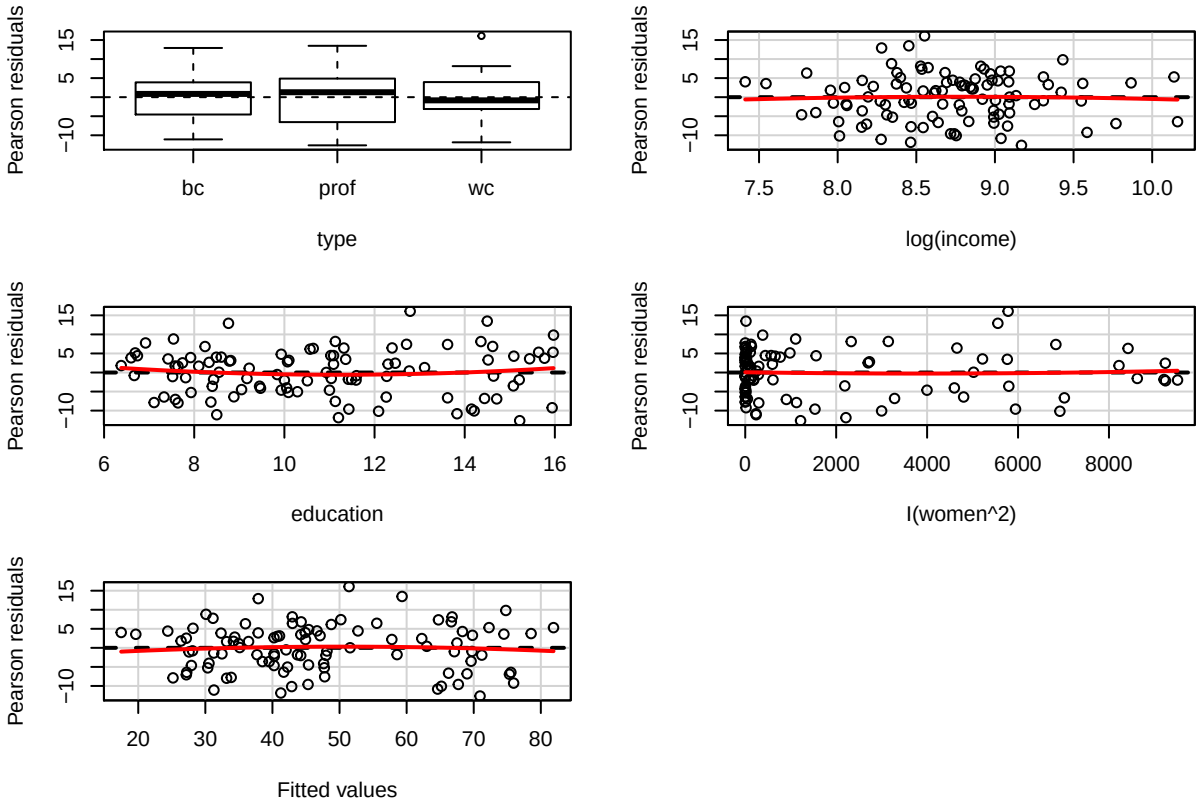
```
## Analysis of Variance Table
##
## Model 1: prestige ~ type + log(income) * education + I(women^2)
## Model 2: prestige ~ type + log(income) * education + log(income) * I(women^2)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     91 3548
## 2     90 3536  1      12.2 0.31   0.58
```

```r
anova(reg6, reg2) # Compare with the smaller model
```

```
## Analysis of Variance Table
##
## Model 1: prestige ~ type + log(income) * education + I(women^2)
## Model 2: prestige ~ type + education + log(income) + I(women^2)
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     91 3548
## 2     92 3679 -1      -130 3.35  0.071 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that there is no difference between this model and the big one, but there is a difference between this one and the smaller, linear model without interactions.

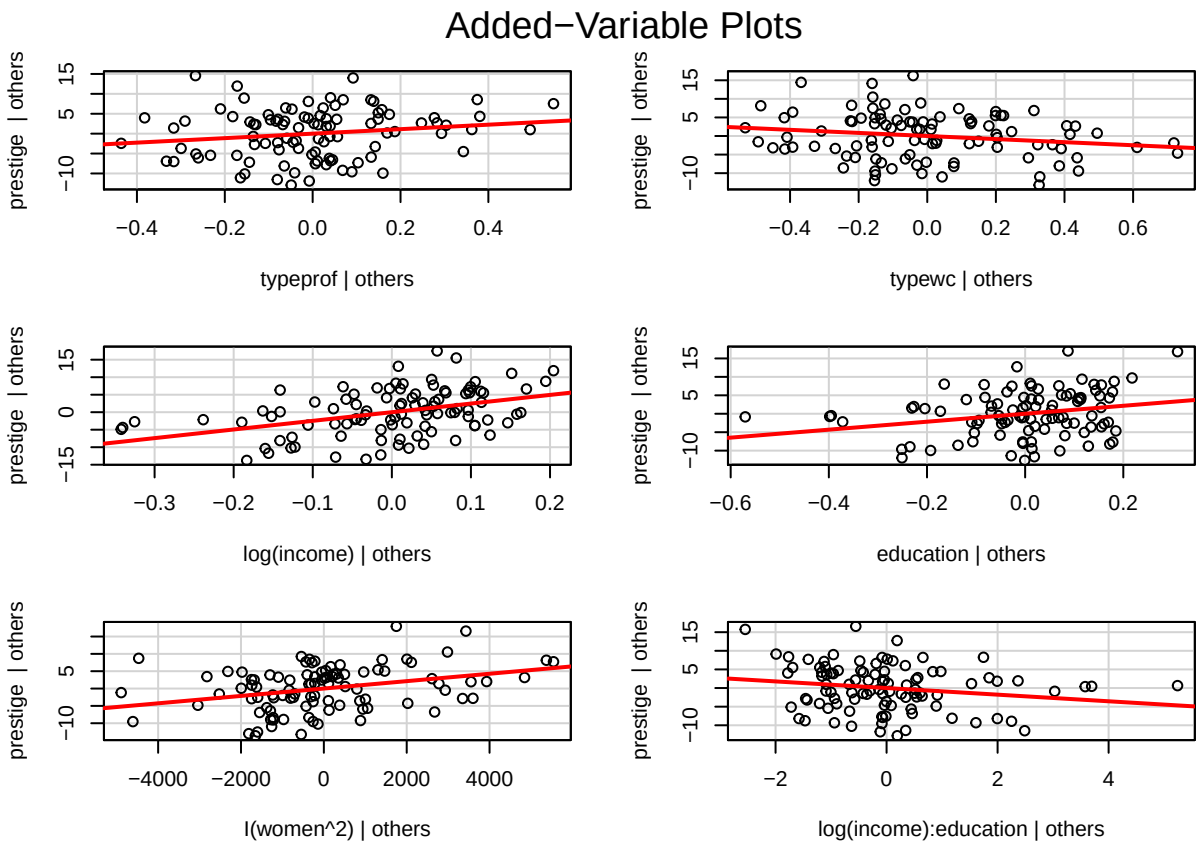
```r
residualPlots(reg6)
```



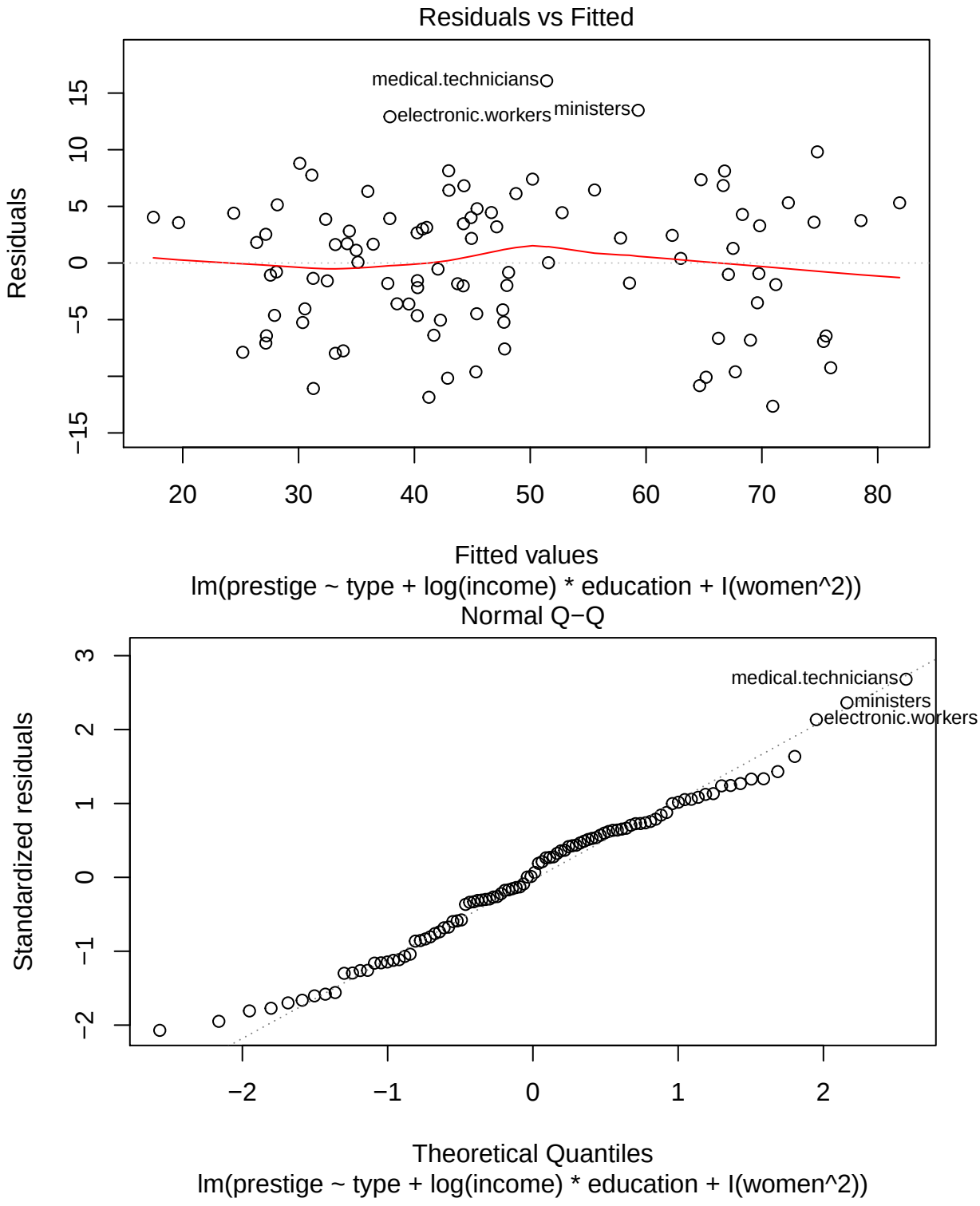
```
##             Test stat Pr(>|t|)
## type               NA       NA
## log(income)    -0.366    0.715
## education       1.343    0.183
## I(women^2)      0.247    0.806
## Tukey test     -1.312    0.190
```

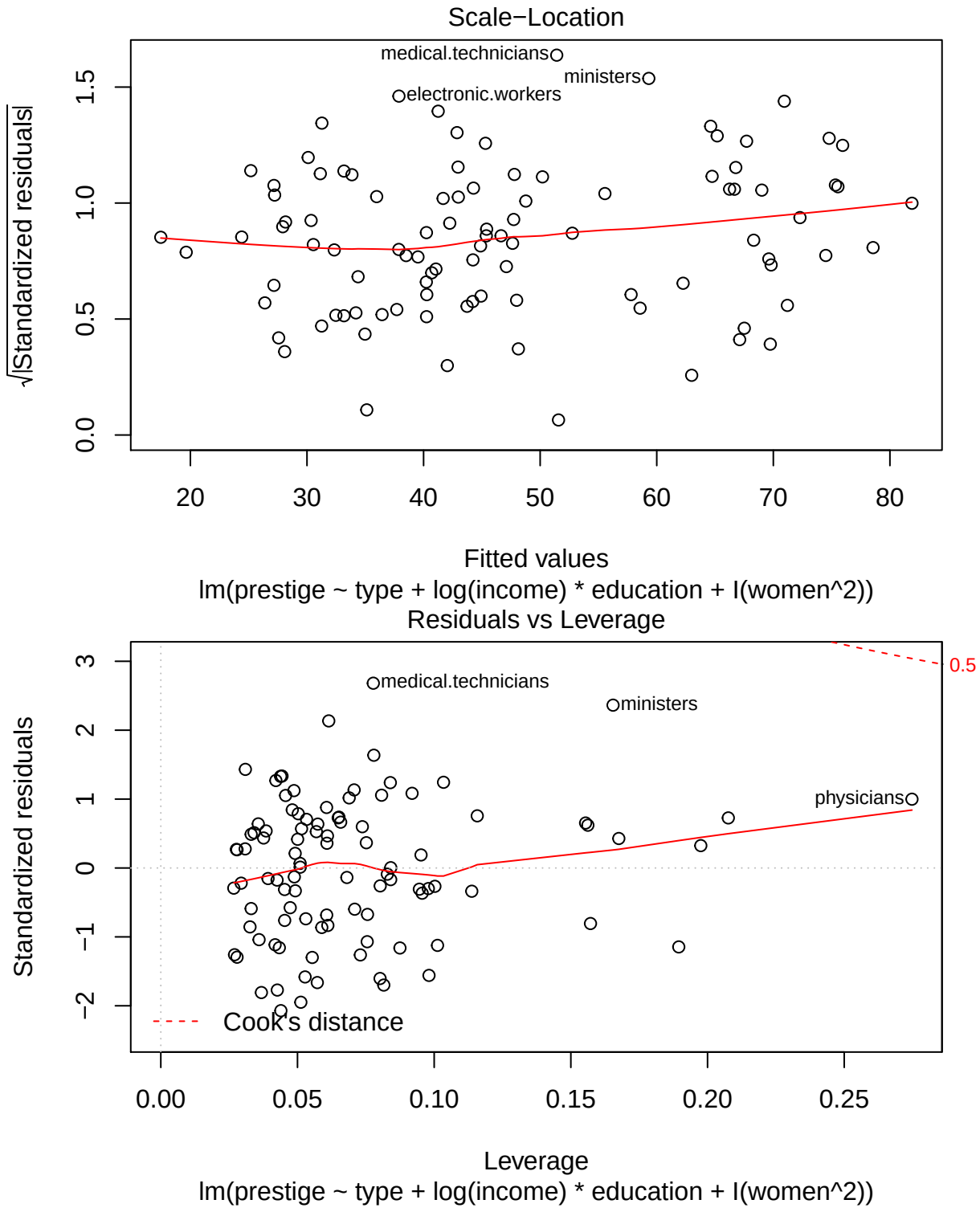
That's really quite beautiful...

```
avPlots(reg6)
```




```
plot(reg6)
```


Residuals vs Fitted
medical.technicians
electronic.workers
ministers
Residuals
15
10
5
0
−5
−15
20
30
40
50
60
70
80
Fitted values
lm(prestige ~ type + log(income) * education + I(women^2))
Normal Q−Q
medical.technicians
ministers
electronic.workers
Standardized residuals
3
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(prestige ~ type + log(income) * education + I(women^2))




All of these plots aren't bad. I think we've got a good model.

One final thing... it would be good to visualize what the effect of the regression is. We'll do that with the predict function again...

```
# First, copy the data, but then fix eduation and income to set values
pred.data <- Prestige
```

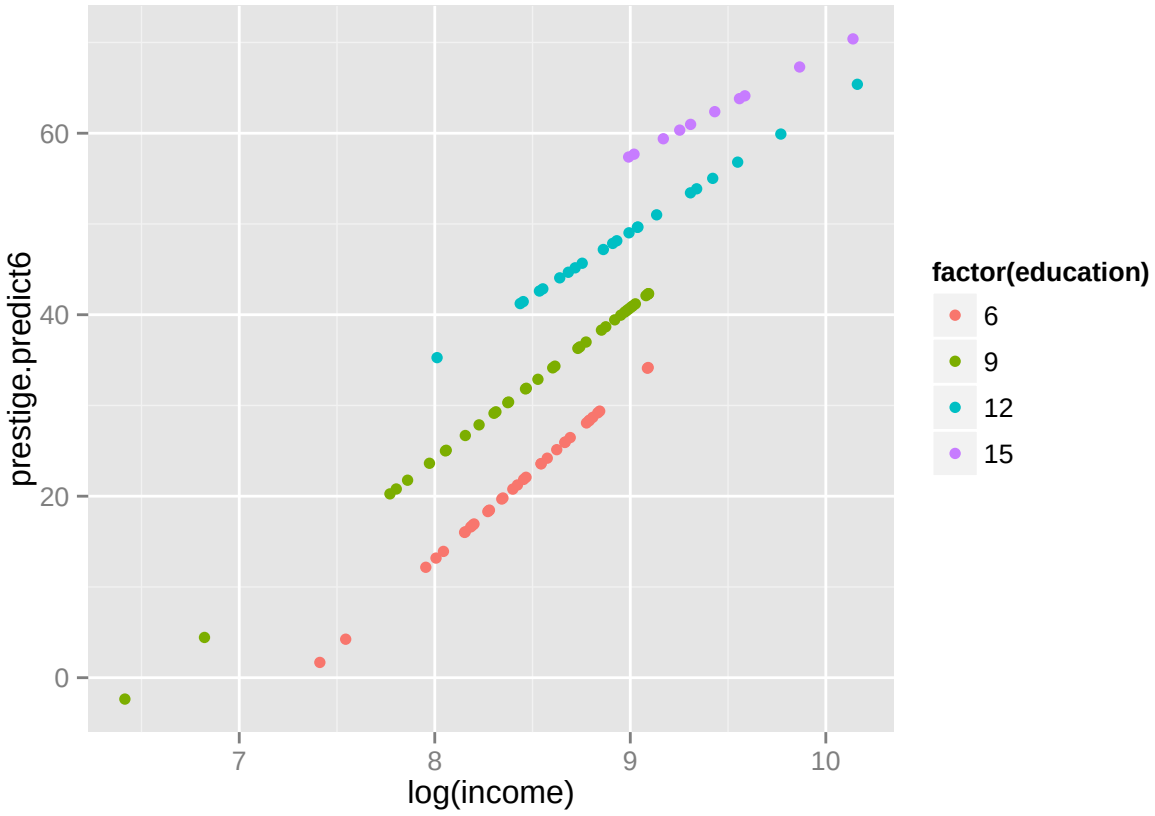
```
# Now, group the income data.  There is a little bit of magic here
# %/% will divide by eliminate the remainder.
#  So look at how I can group these face education numbers...
seq(0,20)
```

```
##  [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

```
(seq(0,20) %/% 4) * 4
```

```
##  [1]  0  0  0  0  4  4  4  4  8  8  8  8 12 12 12 12 16 16 16 16 20
```

```
# Now, do the predictions
pred.data$education <- (pred.data$education %/% 3)*3
pred.data$women <- 10 # Fix women
pred.data$type <- 'wc' # Fix type
pred.data$prestige.predict6 <- predict(reg6, newdata = pred.data)
ggplot(pred.data) + geom_point(aes(x=log(income), y=prestige.predict6, color=factor(education)))
```




You can clearly see the diminishing relationship between income and education.

At this point, I am extremely happy with the regression. I have significant relations, and they all make sense.

In addition to describing the regression coefficients, I would probably write something like this: "A regression including interactions between log(income) and percent women was also performed. That regression did not improve the model fit as measured by an Analysis of Variance ($F(1,91)=.3112$, $p=.5783$) and is not reported here."