Multivariate Regression Modelling

To get a summary look of the data:

1.education. Average education of occupational incumbents, years, in 1971.

2.income. Average income of incumbents, dollars, in 1971.

3.women. Percentage of incumbents who are women.

4.prestige. Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

5.census .Canadian Census occupational code.

6.type. Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

Summary of the data briefly is as follows:

```
library("car")
```

```
## Warning: package 'car' was built under R version 3.1.3
```

```
str(Prestige)
```

```
## 'data.frame':    102 obs. of  6 variables:
##  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
##  $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

In the regression below we try to regress prestige points(independant variable) against education level attained, income and percentage of women in that field.

```
lm1 <-lm( prestige ~ education + income + women, data= Prestige)
summary(lm1)
```

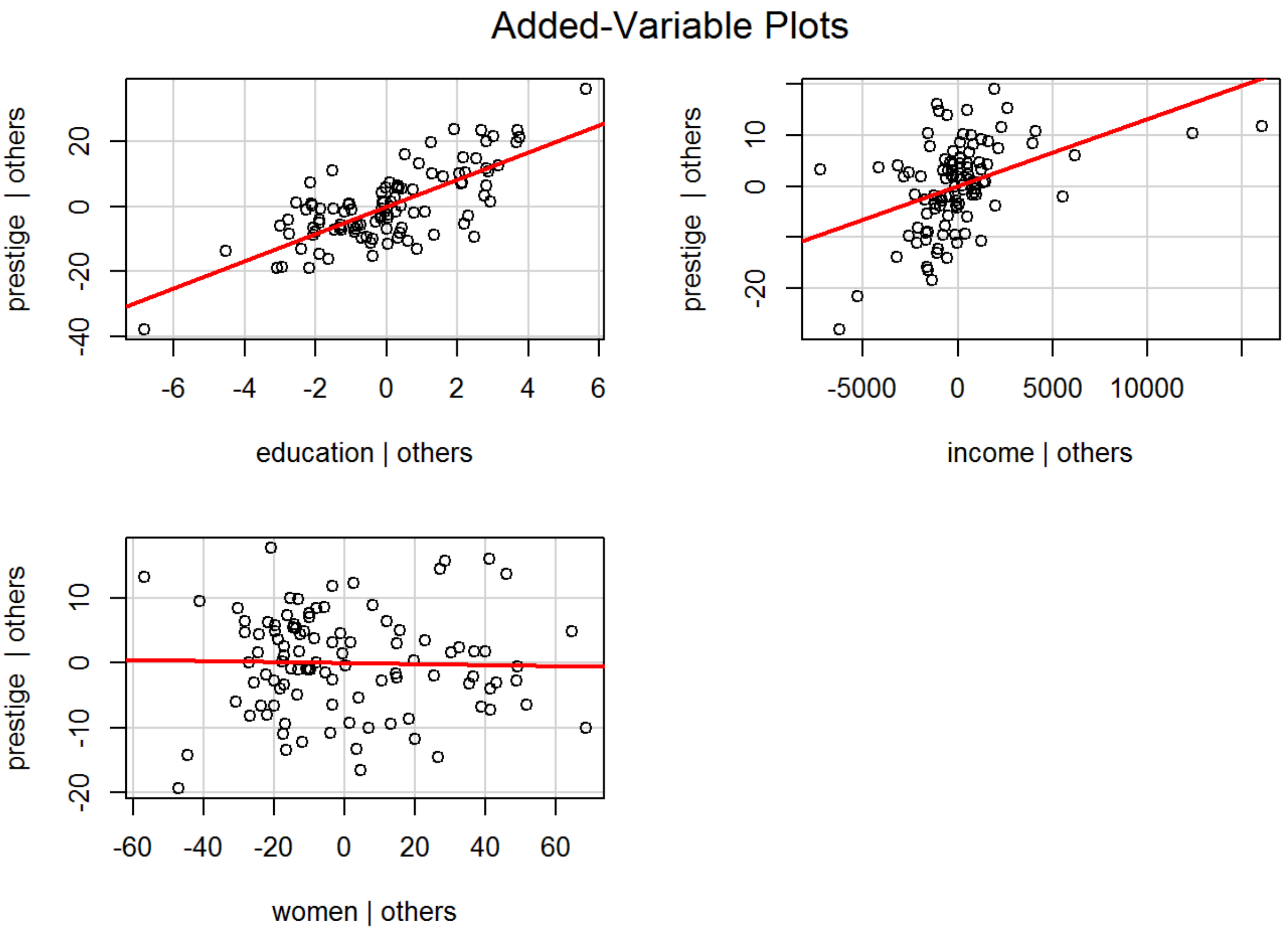
```
##
## Call:
## lm(formula = prestige ~ education + income + women, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.825  -5.333  -0.136   5.159  17.504
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.794334   3.239089   -2.10    0.039 *
## education    4.186637   0.388701   10.77  < 2e-16 ***
## income       0.001314   0.000278    4.73  7.6e-06 ***
## women       -0.008905   0.030407   -0.29    0.770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.85 on 98 degrees of freedom
## Multiple R-squared:  0.798,  Adjusted R-squared:  0.792
## F-statistic:  129 on 3 and 98 DF,  p-value: <2e-16
```

As expected from the above summary, we can see that holding everyother factor constant as education increases the prestige also increases. Similarly holding all other factors contant as income increases prestige also increases. Although I expected to see some sort of a relationship between women and prestige points but this is not observed . No statistically significant results observed.

The plots given below verify our interpretation of the model.

Graphs outcome vs predictor variables holding the rest constant (also called partial-regression plots)

```
avPlots(lm1)
```



Similar regression as above but using the log of education to scale the variable in comparison to other variables.

```
lm2 <-lm( prestige ~ education + log(income) + type, data= Prestige)
summary(lm2)
```

```
##
## Call:
## lm(formula = prestige ~ education + log(income) + type, data = Prestige)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.51  -3.75   1.01   4.36  18.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -81.202     13.743   -5.91  5.6e-08 ***
## education      3.284      0.608    5.40  5.1e-07 ***
## log(income)   10.487      1.717    6.11  2.3e-08 ***
## typeprof       6.751      3.618    1.87    0.065 .
## typewc        -1.439      2.378   -0.61    0.546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.64 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.855,  Adjusted R-squared:  0.849
## F-statistic:  138 on 4 and 93 DF,  p-value: <2e-16
```

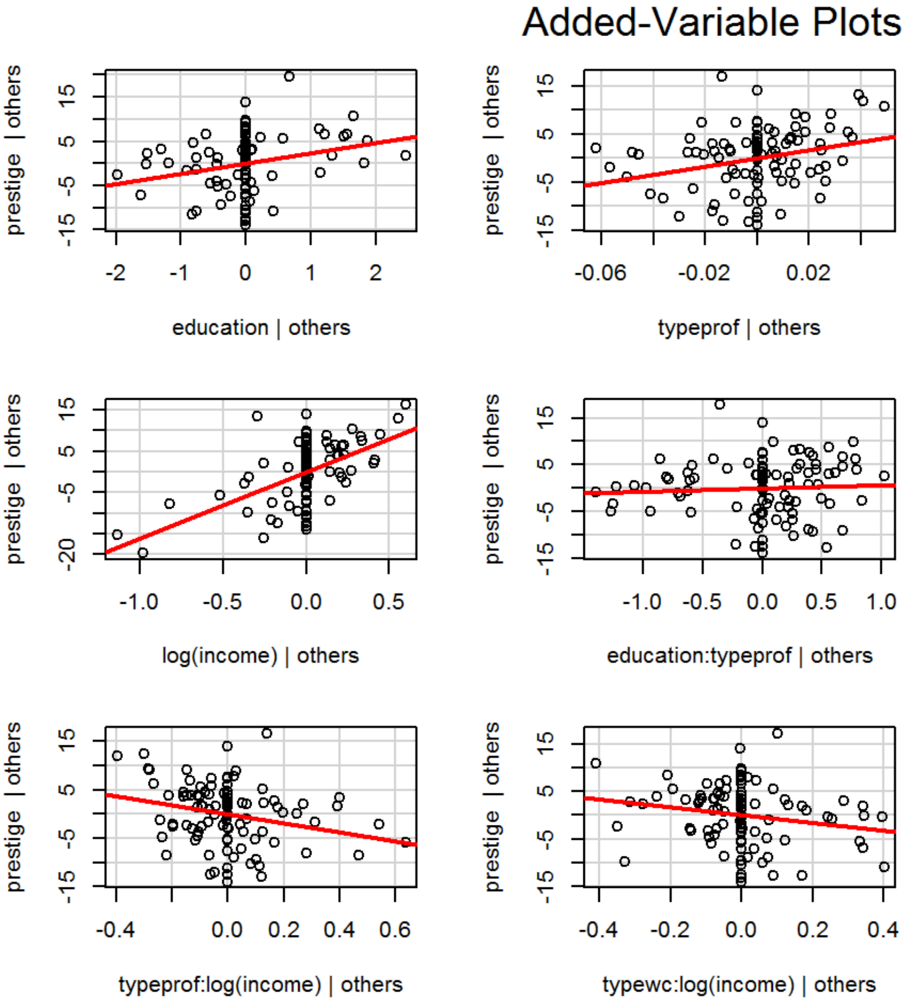
Adding interaction terms to the above analysis- We add interaction terms to the model to see if our model has improved- (We try to observe if there is an interaction between 1. Job type and education 2. Log of income and Job type) We can expect the interaction terms to improve our model

```
lm3 <-lm( prestige ~ education*type + log(income)*type , data= Prestige)
summary(lm3)
```

```
##
## Call:
## lm(formula = prestige ~ education * type + log(income) * type,
##     data = Prestige)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13.97  -4.12   1.21   3.83  18.06
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -120.046     20.158   -5.96  5.1e-08 ***
## education               2.336      0.928    2.52   0.0136 *
## typeprof               85.160     31.181    2.73   0.0076 **
## typewc                 30.241     37.979    0.80   0.4280
## log(income)            15.982      2.606    6.13  2.3e-08 ***
## education:typeprof      0.697      1.290    0.54   0.5900
## education:typewc        3.640      1.759    2.07   0.0414 *
## typeprof:log(income)   -9.429      3.775   -2.50   0.0143 *
## typewc:log(income)     -8.156      4.403   -1.85   0.0673 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.41 on 89 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.859
## F-statistic: 75.1 on 8 and 89 DF,  p-value: <2e-16
```

Similarly as above we use partial regression plots- Help identify the effect (or influence) of an observation on the regression coefficient of the predictor variable.

```
avPlots(lm3)
```



From the anova analysis below - we observe that adding interaction terms to the model improves the model and this result is statistically significant.

```
anova(lm2,lm3)
```

```
## Analysis of Variance Table
##
## Model 1: prestige ~ education + log(income) + type
## Model 2: prestige ~ education * type + log(income) * type
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     93 4096
## 2     89 3655  4       441 2.68  0.036 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Update on project-

I am keen on exploring the intersection between behavioural sciences and policy making primarily focussed on human decision making and risk analysis. A broad area of work in this domain, which I hope to develop further, is “Analysis of risky behaviour in adolescents”.

Questions to be answered:

An underlying motivation behind this study is that adolescents are more prone to risky behaviour which could have long term repercussions on their lives in the long run. A few rudimentary questions that I have in mind are -

1. What are the kinds of risks that adolescents take?
2. Can these risks be categorised in segments ranging from least impact to highest impact?
3. Are their differences in financial conditions which can impact or influence this behaviour?
4. Is there a causal relationship between cultural upbringing and risky behaviour? (A part of this can also include, if belonging to a particular race will have any discernible impact on the individual).
5. If the risks identified above differ between genders? (Exploring if there is a gender specific bias for some types of risky behaviour).
6. Subject to availability of data, if we can compare the risk taking behaviour between adolescents and adults.

The above questions are tentative. I am still in the process of talking to professors and literature review to understand if any prior research has explored these areas and what their findings might be.

Data-sets for this project: There are available data-sets which I could use for my research. One such set is the “The National Longitudinal study of Adolescent and Adult Health”, aka add health dataset asks a lot of relevant questions that could constitute a quantitative study of the risk taking behaviour in adolescents.

Another way this problem is being approached is through the DOSPERT scale. It is a psychometric tool used for measuring domain specific risk taking attitudes and behaviours.

Expected Results and impact: The above research will be useful in understanding risky behaviour. This analysis can be further expanded as a predictive model perhaps, which can help curb or reduce the repercussions of such behaviour.