DUMMY-VARIABLE REGRESSION

Objective:

- Incorporation of qualitative explanatory variables (factors) in conjunction with or without quantitative variables in a linear model.
- The usage of dummy variable repressors coded as dichotomous (two sub groups) and polytomous (more than two sub groups).
- Interpret the interactions between qualitative and quantitative variables in a model which incorporates both.

Linear Model Formulation

Common slope model incorporating additive dummy variables:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \text{ ....1}$$

Where D is the dummy (indicator) variable.

Taking one explanatory variable and one regressor variable:

X is the education and D is modelled by Gender (say D=0 women and D=1 men)

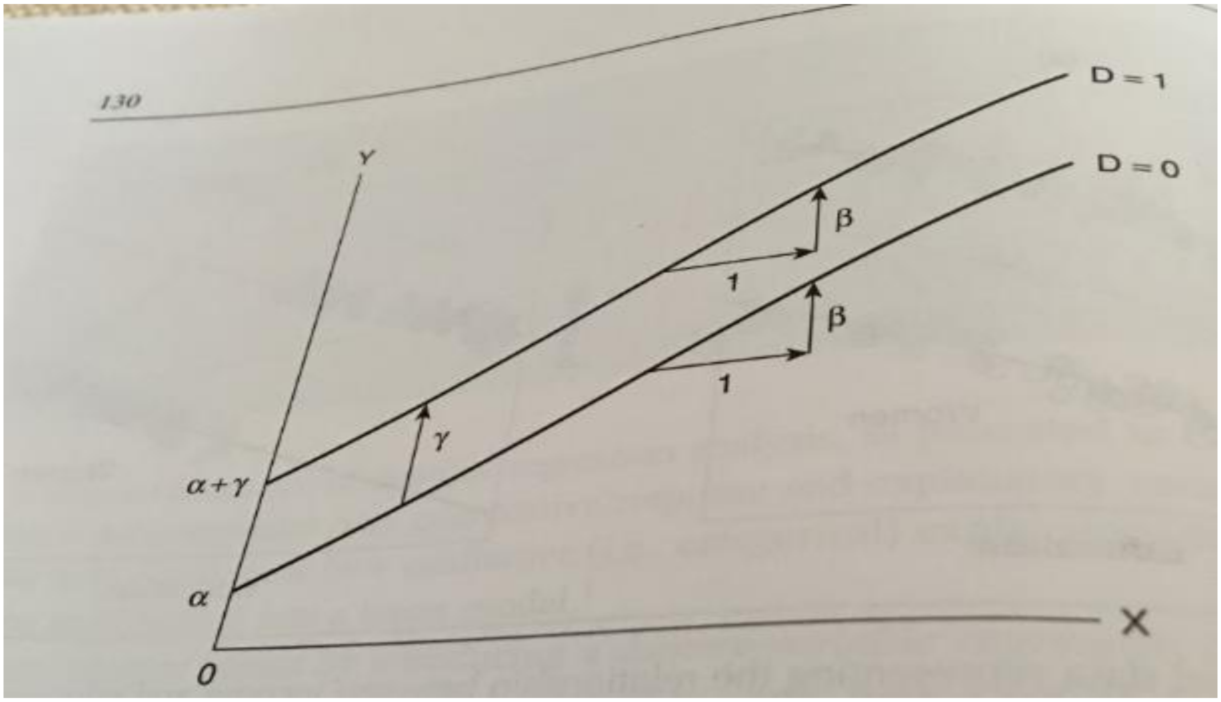
Y is the Income

The dummy variable D (factor with two or more groups or qualitative/categorical) is a regressor variable.

Generally for qualitative variables we use the term regressor instead of explanatory.

The explanatory variable or regressor variable X(education) is a quantitative variable .

If the original explanatory variable is transformed by say squaring or taking a log then the log(X) is no longer a explanatory variable as it is a regressor derived from the explanatory variable.



X is education and Υ is income D=0 Women D=1 Men

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \text{ ....1}$$

For women

$$Y_i = \alpha + \beta X_i + \varepsilon_i \text{ ....2}$$

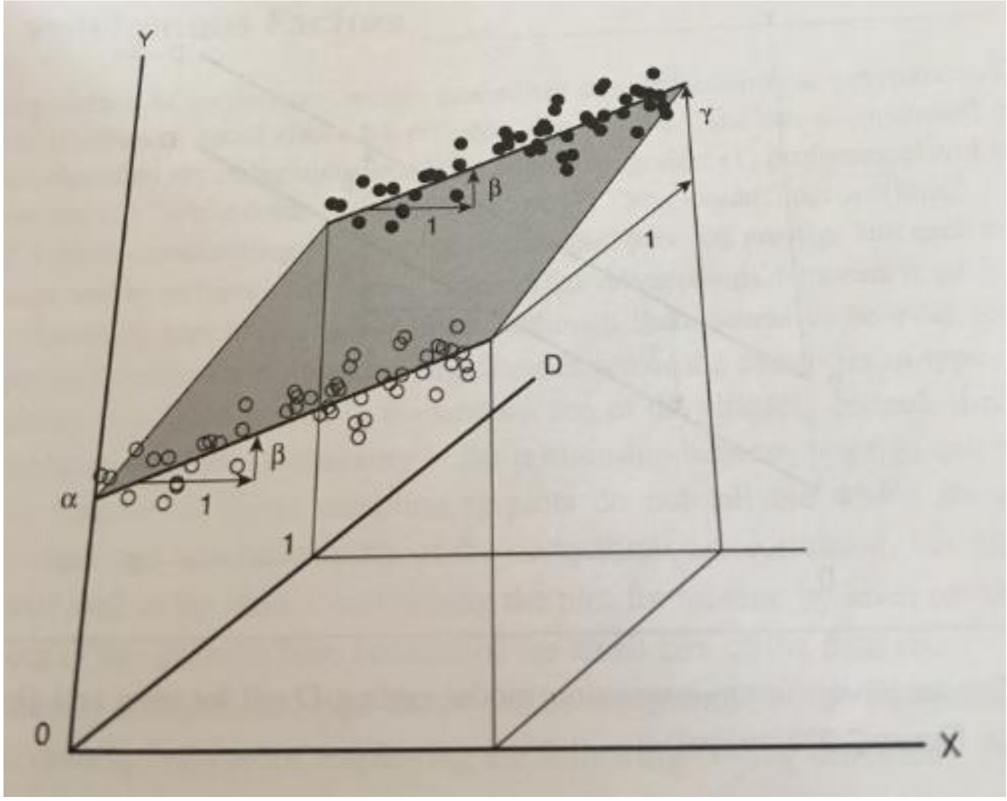
For men

$$Y_i = (\alpha + \gamma) + \beta X_i + \varepsilon_i$$

Here the coefficient Υ gamma represents the difference of income of the two genders. In this example the differences are consistent throughout but it might not be so in other case studies. In this case Υ (a positive quantity) represents the income advantage obtained by men.

The slope model is a good one which incorporates the within gender regressions of income on education (parallel planes) as well as the overall regression (keeping education constant) then effect of gender on the income represented by vertical distance γ .

Here we are incorporating only cases where the two dummy regressors D=0 and D=1 on the edge of the plane.



INCOME

Interpretation about the slopes

To check if the gender effects the income for a particular education level (controlled education) the hypothesis can be formulated as:

$H_0 : \gamma = 0$ and conducting a t test

OR

Conducting an F test by dropping the D.

The Dummy Regressor model can be applied for various quantitative variables. The only assumption required here is that the slopes are the same for the factors with two categories.

If there are more than one quantitative variables the equation can be represented as follows:

$$Y_i = \alpha + \beta_1 X_{i1} + \dots\dots\dots\dots\dots\dots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

D=0 Say for women

$$Y_i = \alpha + \beta_1 X_{i1} + \dots\dots\dots\dots\dots\dots + \beta_k X_{ik} + \varepsilon_i$$

D=1 Say for men

$$Y_i = (\alpha + \gamma) + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \varepsilon_i$$

POLYTOMOUS FACTORS

If there are more than two categories for the qualitative/categorical variable then we need more than one dummy variable.

Suppose we consider the regression of prestige on education and income and type of occupation. The categories of type of education is Professional and managerial, White and Blue Collar.

Category	D1	D2
Professional and Managerial	1	0
White Collar	0	1
Blue Collar(Baseline Category)	0	0

The general model for polytomous factors is:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$$ X_1 Income X_2 Education

The parallel regression planes are as follows:

Blue Collar:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Where α is the Y intercept for Blue collar

White Collar:

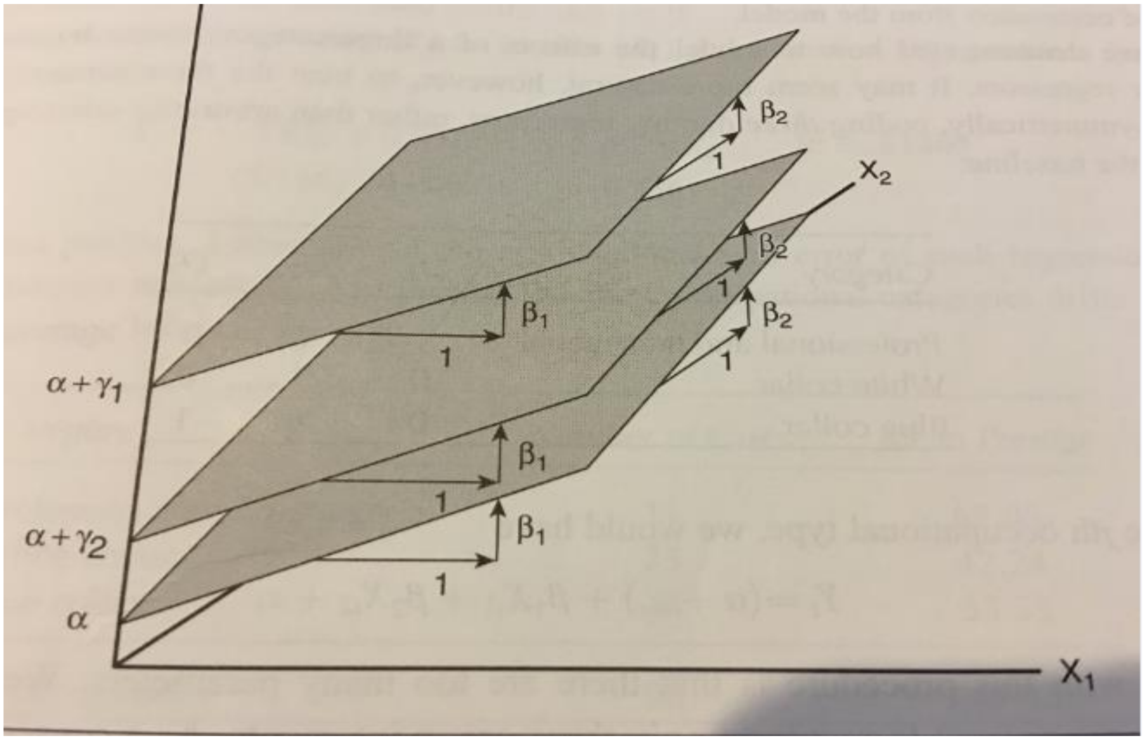
$$Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where γ_2 the vertical distance between white collar and blue collar

Professional and Managerial

$$Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

where γ_1 the vertical distance between professional and blue collar



Blue Collar category is taken as a baseline. Any category could have been taken as a baseline. The blue collar category is coded as D=0 and 0. Here the choice of the baseline is arbitrary.

For a study where the comparison has to be made for experimental and control groups, usually control groups are considered as baseline category

Hypothesis Testing:

To check that there is a difference of Prestige regardless of the occupation (blue collar, white collar or professional) keeping the education and income constant .

$$H_0 : \gamma_1 = \gamma_2 = 0$$

If think about coding the three categories uniformly we can also code it in this format:

Category	D1	D2	D3
Professional and Managerial	1	0	0
White Collar	0	1	0
Blue Collar(Baseline Category)	0	0	1

Any category can be used as a baseline category. Here jth occupational type would be:

$$Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Taking the case above where we have 3 categories

This would gives four parameters and $(\alpha, \gamma_1, \gamma_2, \gamma_3)$ and three regression equations. Therefore we cannot find unique values for these four parameters.

This also causes the issues of collinearity. This means that one of the factor categories can be predicted by the others.

$D_3=1-D_1-D_2$ The dummy variables are perfectly collinear. Therefore the unique least square estimates cannot be calculated.

Therefore we do not use the model of m categorical with m dummy variables. We use the model of m categories m-1 dummy regressor variables.

Category	D_1	D_2	……….	……………	D_{m-1}
1	1	0	0	0	0
2	0	1	0	0	0
3	…	….	…..	…..	…..
……					
……..					
m-1	0	0	0	0	1
m	0	0	0	0	0

The model can be visualized as a m parallel planes with each plane representing each of the m categories. The mth category with all the dummy variables =0 is the baseline category.

```
PrestigeFinalmod<-lm(prestige~income+education,data=Prestige)
PrestigeFinalmod

##
## Call:
## lm(formula = prestige ~ income + education, data = Prestige)
##
## Coefficients:
## (Intercept)       income    education
##   -7.621035     0.001242     4.292108

summary(PrestigeFinalmod)

##
## Call:
## lm(formula = prestige ~ income + education, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9367  -4.8881   0.0116   4.9690  15.9280
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.6210352  3.1162309  -2.446   0.0163 *
## income       0.0012415  0.0002185   5.682 1.45e-07 ***
## education    4.2921076  0.3360645  12.772  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.45 on 95 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.8101
## F-statistic: 207.9 on 2 and 95 DF,  p-value: < 2.2e-16
```

```
> summary(PrestigeFinalmodDum)

Call:
lm(formula = prestige ~ income + education + type, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9529  -4.4486   0.1678   5.0566  18.6320

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6229292  5.2275255  -0.119    0.905
income       0.0010132  0.0002209   4.586 1.40e-05 ***SIGNIFICANT
education    3.6731661  0.6405016   5.735 1.21e-07 ***SIGNIFICANT
typeprof     6.0389707  3.8668551   1.562    0.122difference of mean type pro
                                                  and type bc(baseline) is
                                                  insignificant.

typewc      -2.7372307  2.5139324  -1.089    0.279
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.095 on 93 degrees of freedom
Multiple R-squared:  0.8349,   Adjusted R-squared:  0.8278
F-statistic: 117.5 on 4 and 93 DF,  p-value: < 2.2e-16
```

Here the

$$\hat{Y} = -.62292 + .001013X_1 + 3.673X_2 + 6.039D_1 - 2.73D_2$$

Professional: $\hat{Y} = 5.416 + .001013X_1 + 3.673X_2$ (intercept obtained by-.623+6.039=5.416)

White Collar: $\hat{Y} = -3.360 + .001013X_1 + 3.673X_2$ (intercept obtained by-.623-2.737=-3.360)

Blue Collar: $\hat{Y} = -.623 + .001013X_1 + 3.673X_2$

Notice and compare (Called the Null Model as we have omitted the dummy variable)

Only Income and education

$$\hat{Y} = -.62292 + .0012415X_1 + 4.2921X_2$$

Income and education and Dummy variable (Called the Full Model as we have included the dummy variable)

$$\hat{Y} = -.62292 + .001013X_1 + 3.673X_2 + 6.039D_1 - 2.73D_2$$

See that the coefficients of both income and education decrease when we control for type of occupation.

```
> aggregate(Prestige$prestige, by=list(Prestige$type), FUN=mean)
  Group.1        x
1      bc 35.52727
2    prof 67.84839
3      wc 42.24348
```

Keeping the income and education constant the difference in prestige for bc and wc is 42.24-35.53=6.71

Which points towards the slope -2.73

Keeping the income and education constant the difference in prestige for bc and prof is 67.84-35.53=32.32

Which points towards the slope 6.04

This proves that the greater prestige of professionals as compared to bc and wc are due to differences in education and income.

To test the partial effects of type of occupation ie therefore

Ho: no effect of occupation type on Prestige keeping income and education is kept constant:

$$H_0 : \gamma_1 = \gamma_2 = 0$$

To compare the null model and the full model

If the Full model is significantly different and better than the null model then it makes sense that $R_1 >> R_0$ which means F_0 will be much much higher.

$$F_0 = \frac{n-k-1}{q} * \frac{R_1^2 - R_0^2}{1-R_1^2} = \frac{98-4-1}{2} * \frac{R.83486-.81400}{1-.83486} = 5.874$$

q is the number of regressors omitted from the full model (2 Dummy regressors) to obtain the Null or reduced Model.

With 2 and 93 degrees of freedom. p=.0040 which is statistically ie the Full Model is significantly Different from the Null Model. The difference between R_1 and R_0 is significant.

The effects of income on Prestige are significant keeping type of occupation and education constant 1.40e-05 ***

The effects of education on Prestige are significant keeping type of occupation and income constant

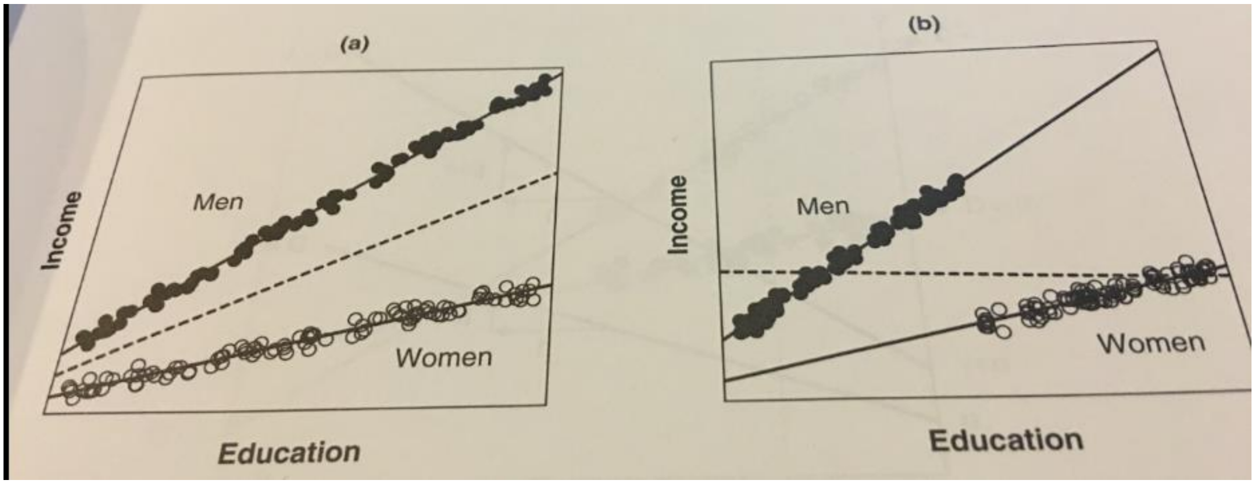
1.21e-07 ***

Modelling Interactions:

A model interactions incorporates the effect on the response variable of two explanatory variables (could be more than two) of which the partial effect of one them on the response variable depends on the value of the other explanatory/regressor variable.

This section explains the interactions between factors (categorical) and explanatory variables.

To identify these interactions the planes depicting the two or more groups of the categorical variable will not be parallel to each other.



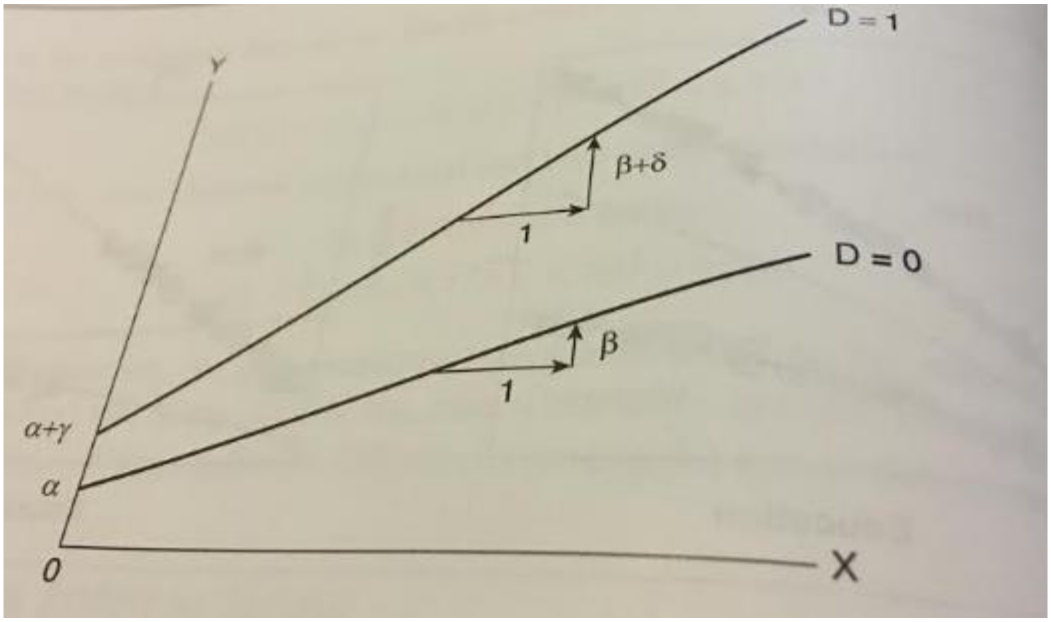
Both the images show that the gender planes are not parallel therefore we can infer that there is interactions of gender and education which affects income. The slope for men is higher in both cases.

The effect of gender at different education levels are different. As the education level increase the income differences of the two genders increases.

Interactions between two explanatory variables does not imply correlation between them.

Combined Regressor Model: Incorporates non parallel slopes

The combined regressor model for both gender together is better than creating two separate model one for each gender. The combined model showcases the gender by education interactions.



$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

Here notice that we have X as the regressor for education, D as the Dummy regressor for gender and an interaction regressor XD to depict the interaction between gender and education.

The interaction regressor XD (product of the two other regressors education and gender) incorporated both the males and females. Since XD is a product of two regressors it is a nonlinear.

But by incorporating the dummy variable we have linearlized the model. This also avoides perfect collinearity ie the coefficient of the interaction do not have a linear relationship.

Is $\delta = \beta + \gamma$? Or any other combination?

For women the equation becomes:

$$Y_i = \alpha + \beta X_i + \gamma(0) + \delta(X_i * 0) + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

For men the equation becomes

$$Y_i = \alpha + \beta X_i + \gamma(1) + \delta(X_i * 1) + \varepsilon_i = (\alpha + \gamma) + (\beta + \delta) X_i + \varepsilon_i$$

α Is the intercept for the gender: women

β Is the slope for the gender: women ie it represents the change in income for women with every unit change in education for them. Remember for the non interaction model β was the common within gender slope.

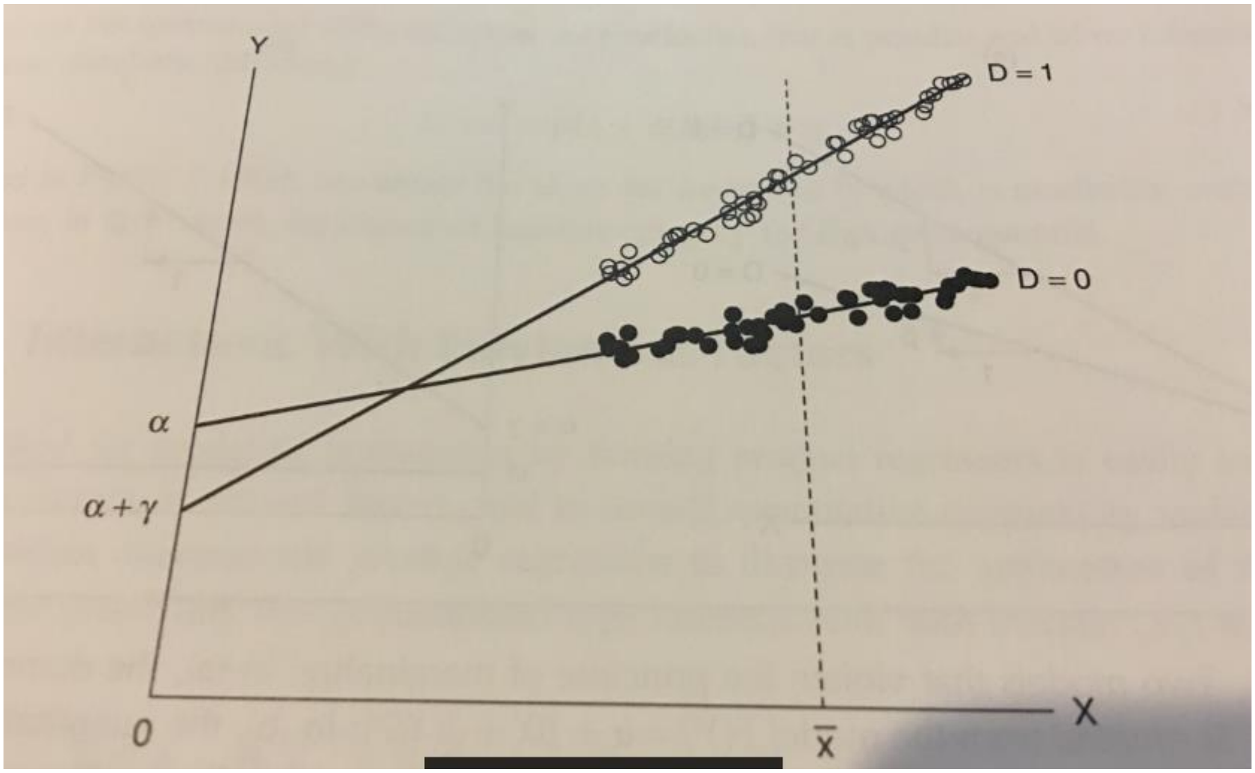
γ Is the difference of intercept for men and women

δ Difference of slopes for men and women. β Is the slope for the gender: women $\beta + \delta$ Is the slope for the gender: men

To check for the presence or absence of interaction we need to test the following Hypothesis:

H$_0$: $\delta = 0$

In this model the γ The education intercept difference between males and females is not consequential because unlike the non interaction model it does not stay the same for all educational level and does not represent the educational advantage of men over women. As we have perceived earlier sometimes for certain data analysis the y intercept where X=0 is not interpretable at all. For example in the diagram given below:



Principle of Marginality

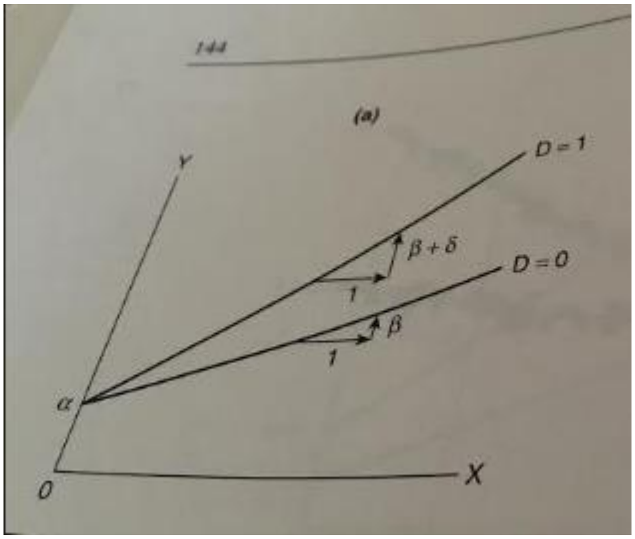
$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$

The model that we create to incorporate interaction should include the main effects (lower order terms : The regressors: X_i and D_i) as well as the higher order terms(The higher order term representing the interaction X_iD_i).

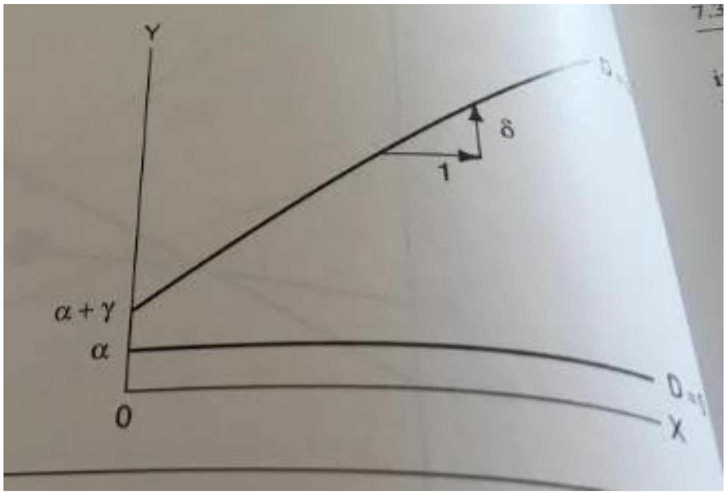
If it does not the model is not appropriate.

For instance if we ignore X or D in the model , the non substantive model would be represented as follows:

$$Y_i = \alpha + \beta X_i + \delta(X_i D_i) + \varepsilon_i$$



$$Y_i = \alpha + \gamma D_i + \delta(X_i D_i) + \varepsilon_i$$



[1]

Both these models are inappropriate

Interactions with Polytomous Factors:

This is the extension of the case where we have not just one factor but many factors (categorical regressors) as well as many quantitative variables.

Example the Prestige case study:

Income (X_i) Education (X_2) are the quantitative explanatory variables

Type of profession Blue Collar, White collar, Professional categorical regressor. Here occupation will interact with both income and education.

The main effects of the explanatory variables X_i and D_i for profession type have to be incorporated as well as the interaction regressors (each explanatory variable with the each factor using the dummy coding schema).

Remember in the factor Profession type we have 3 categories so we will use 2 dummy variables.

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \varepsilon_i$$

Blue collar: Both D_{i1}=0 and D_{i2}=0

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

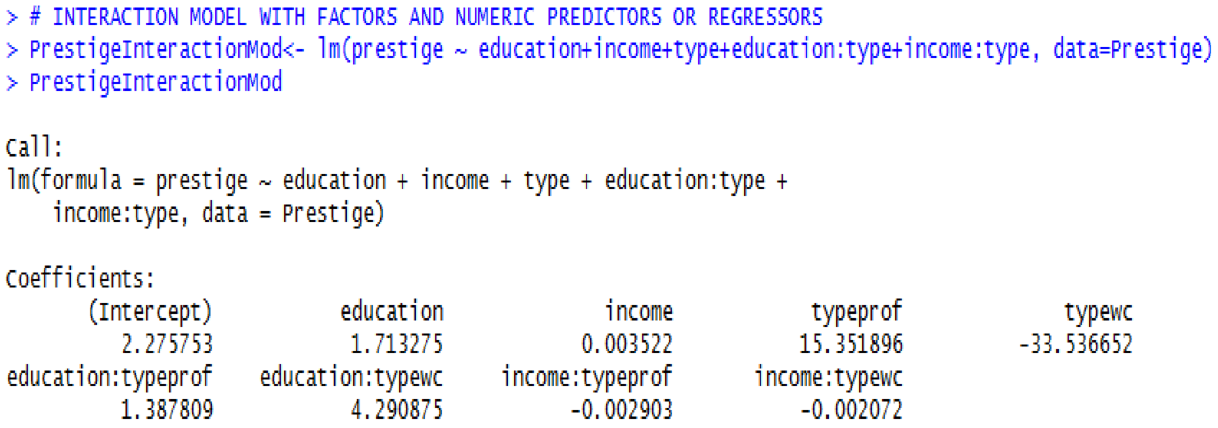
White collar: D_{i1}=0

$$Y_i = (\alpha + \gamma_2) + (\beta_1 + \delta_{12}) X_{i1} + (\beta_2 + \delta_{22}) X_{i2} + \varepsilon_i$$

Professional D_{i2}=0

$$Y_i = (\alpha + \gamma_1) + (\beta_1 + \delta_{11}) X_{i1} + (\beta_2 + \delta_{21}) X_{i2} + \varepsilon_i$$

Baseline category is chosen arbitrarily. Here baseline category is Blue collar.



```
> # INTERACTION MODEL WITH FACTORS AND NUMERIC PREDICTORS OR REGRESSORS
> PrestigeInteractionMod<- lm(prestige ~ education+income+type+education:type+income:type, data=Prestige)
> PrestigeInteractionMod

Call:
lm(formula = prestige ~ education + income + type + education:type + 
    income:type, data = Prestige)

Coefficients:
       (Intercept)          education             income           typeprof             typewc  
          2.275753           1.713275           0.003522          15.351896         -33.536652  
education:typeprof   education:typewc    income:typeprof      income:typewc  
          1.387809           4.290875          -0.002903          -0.002072  
```

$$\hat{Y}_i = 2.276 + .003522 X_1 + 1.713 X_2 + 15.35 D_1 - 33.54 D_2 - 33.54 D_2 - .002903 X_1 D_1 - .002072 X_1 D_2 + 1.388 X_2 D_1 + 4.291 X_2 D_2$$

For Simplistic Interpretation of the dummy regressor the best way is to look at the equations written individually for each profession group:

$$\hat{Y}_i = 2.276 + .003522X_1 + 1.713X_2 + 15.35D_1 - 33.54D_2 - .002903X_1D_1 - .002072X_1D_1 + 1.388X_2D_1 + 4.291X_2D_2$$

X_1 Income X_2 Education

Professional D_1=1 D_2=0

$$\hat{Y}_i = 17.63 + .000619 * INCOME + 3.101 * EDUCATION$$

White Collar D_1=0 D_2=1

$$\hat{Y}_i = -31.26 + .001450 * INCOME + 6.004 * EDUCATION$$

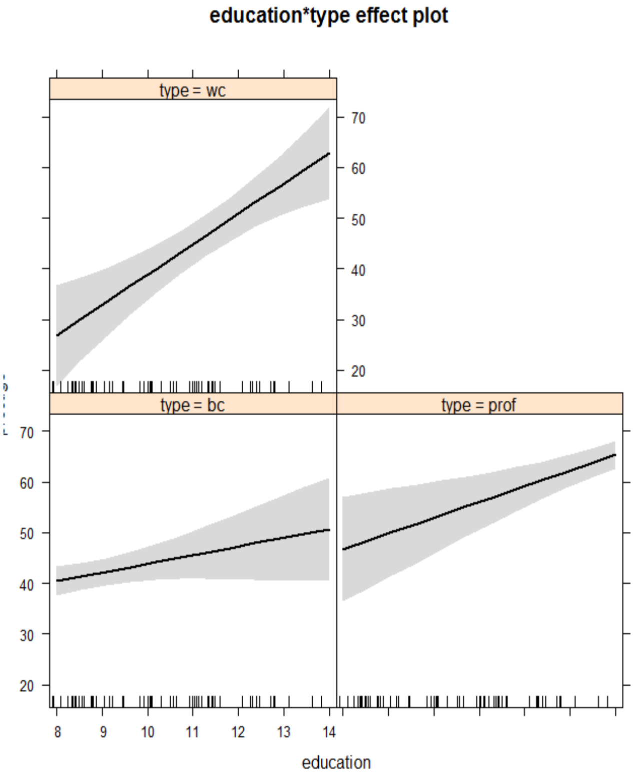
Blue Collar D_1=0 D_2=0

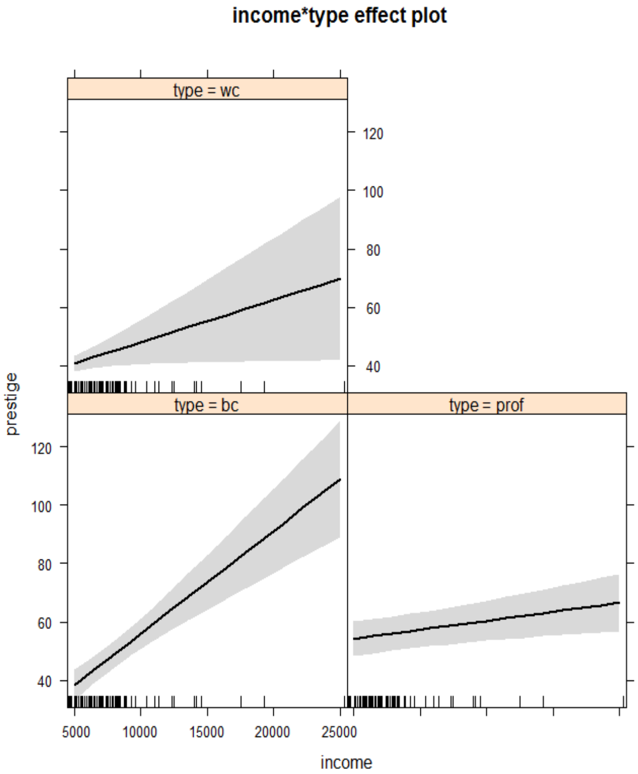
$$\hat{Y}_i = 2.276 + .003522 * INCOME + 1.713 * EDUCATION$$

Please observe that income makes more difference to Prestige for a blue collar as compared to white collar and even more as compared to Professional.

Education makes maximum difference in Prestige for the White collar category followed by Professional category and then Blue collar.

```
> plot(allEffects(PrestigeInteractionMod))
```





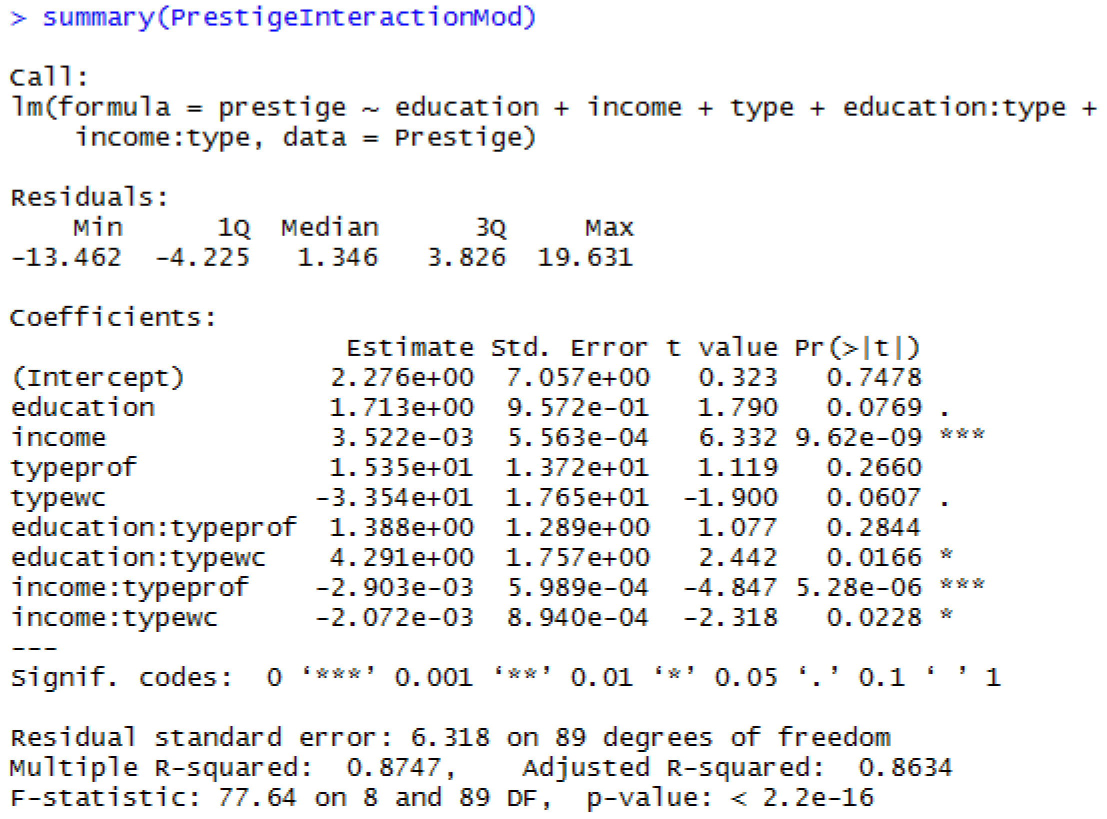
The complex models can also be viewed by the following strategy:

We can observe the factor categories over the ranges of an explanatory variable setting the other explanatory variable to its average value.

For example we can see how prestige changes for wc, bc and prof as the education increases keeping income an average value.

For example we can see how prestige changes for wc, bc and prof as the income increases keeping education an average value of 10.79. The 95% confidence intervals of standard errors can be viewed.

The analysis that we inferred can also be perceived here.



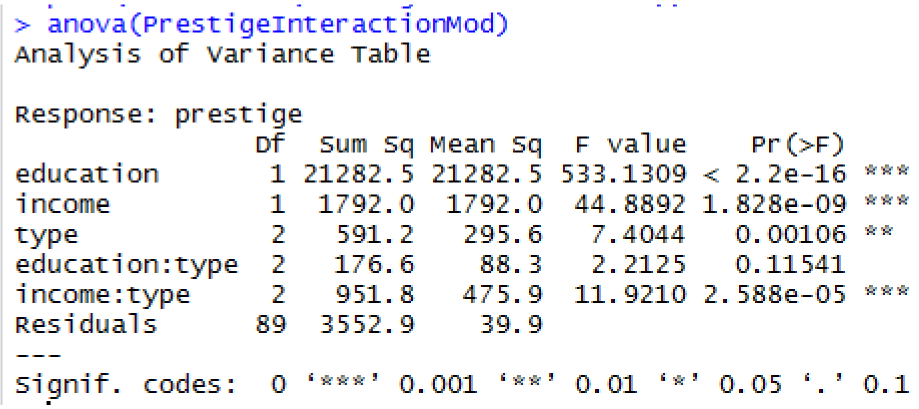
```
> summary(PrestigeInteractionMod)

Call:
lm(formula = prestige ~ education + income + type + education:type +
    income:type, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-13.462  -4.225   1.346   3.826  19.631

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.276e+00  7.057e+00   0.323   0.7478
education           1.713e+00  9.572e-01   1.790   0.0769 .
income              3.522e-03  5.563e-04   6.332 9.62e-09 ***
typeprof            1.535e+01  1.372e+01   1.119   0.2660
typewc             -3.354e+01  1.765e+01  -1.900   0.0607 .
education:typeprof  1.388e+00  1.289e+00   1.077   0.2844
education:typewc    4.291e+00  1.757e+00   2.442   0.0166 *
income:typeprof    -2.903e-03  5.989e-04  -4.847 5.28e-06 ***
income:typewc      -2.072e-03  8.940e-04  -2.318   0.0228 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.318 on 89 degrees of freedom
Multiple R-squared:  0.8747,    Adjusted R-squared:  0.8634
F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16
```



```
> anova(PrestigeInteractionMod)
Analysis of Variance Table

Response: prestige
               Df  Sum Sq Mean Sq  F value    Pr(>F)
education       1 21282.5 21282.5 533.1309 < 2.2e-16 ***
income          1  1792.0  1792.0  44.8892 1.828e-09 ***
type            2   591.2   295.6   7.4044   0.00106 **
education:type  2   176.6    88.3   2.2125   0.11541
income:type     2   951.8   475.9  11.9210 2.588e-05 ***
Residuals      89  3552.9    39.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Source	Null Hypothesis	p-value
Income	$\beta_1 = 0 \mid \delta_{11} = \delta_{12} = 0$	1.828e-09
Education	$\beta_2 = 0 \mid \delta_{21} = \delta_{22} = 0$	<2.26e-16
Type	$\gamma_1 = \gamma_2 = 0 \mid \delta_{11} = \delta_{12} = 0 \ \ \delta_{21} = \delta_{22} = 0$	.00106
Income*Type	$\delta_{11} = \delta_{12} = 0$	2.588e-05
Education*Type	$\delta_{21} = \delta_{22} = 0$	.11541

Interpretation of Hypothesis Testing:

The Incremental F table is created using the Principal of Marginality:

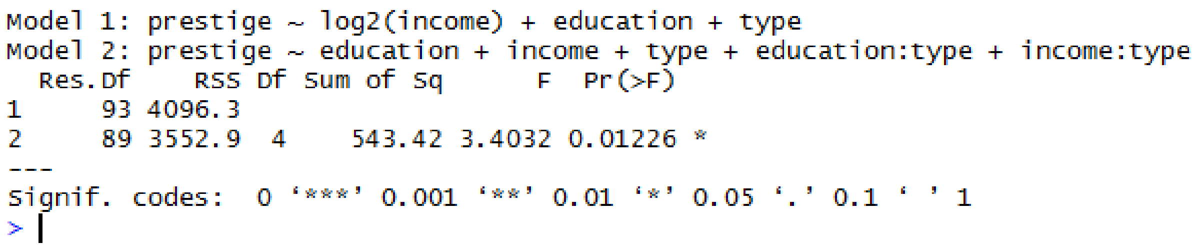
The model that we create to incorporate interaction should include the main effects (lower order terms : The regressors: X_i and D_i) as well as the higher order terms(The higher order term representing the interaction X_iD_i).

The interaction between income and type of profession is significant wheras the interaction between education and type of profession is not significant.

Looking at standard error of the coefficient and the confidence intervals of effects are very broad which informs us that the interactions have not been estimated accurately. This could be due to the small dataset.

The main effect (lower order terms) of Income, Education and Type are significant assuming the higher order terms are insignificant. The Interaction of Income and Type is significant therefore the main effect of Income and Type are of not very consequential.

Comparing Two model



```
Model 1: prestige ~ log2(income) + education + type
Model 2: prestige ~ education + income + type + education:type + income:type
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1     93 4096.3                              
2     89 3552.9  4    543.42 3.4032 0.01226 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

These models are significantly different.

Prediction

```
### Prediction of mean response for cases like this
> predict(PrestigeFinalmod, list(income=2300,education=10), interval="conf")
       fit      lwr      upr
1 38.15558 35.84546 40.46569
> ### Prediction for a single new case
> predict(PrestigeFinalmod, list(income=2300,education=10), interval="pred")
       fit      lwr      upr
1 38.15558 23.18654 53.12461
```

> Confidence intervals of the slopes: estimated regression coefficents: and intercepts

```
> confint(PrestigeFinalmod,conf.level=.95)
                    2.5 %       97.5 %
(Intercept) -1.380754e+01 -1.434534747
income       8.077722e-04  0.001675302
education    3.624935e+00  4.959279979
```

https://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html

https://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/

http://minato.sip21c.org/msb/man/VIF.html https://onlinecourses.science.psu.edu/stat857/node/225