**MULTICOLLINEARITY**

If our regression equation is represented as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$X_1$ provides us with some information about Y and $X_2$ provides us with some more information of Y

For multiple regression we assume that the Xs are independent of each other.

If they are correlated(more then two) with each other then the problem of multicollinearity occurs.

Remember

$\beta_1$= is the change in Y with every unit change in $X_1$ keeping $X_2$ constant. This is for controlling/adjusting for confounding. It is the partial derivative $X_1$ to obtain partial Y keeping $X_2$ constant.

$\beta_2$ =is the change in Y with every unit change in $X_1$ keeping $X_2$ constant. This is for controlling/adjusting for confounding. It is the partial derivative $X_2$ to obtain partial Y keeping constant.

When there is multicollinearity exists then this is not true

$\beta_1 \neq$ is the change in Y with every unit change in $X_1$ keeping $X_2$ constant. This is for controlling/adjusting for confounding.

$\beta_2 \neq$is the change in Y with every unit change in $X_1$ keeping $X_2$ constant. This is for controlling/adjusting for confounding.

The impact of $X_1$ on Y is impacted by $X_2$ and the impact of $X_2$ on Y is impacted by $X_1$.

For multiple regression we are actually trying to identify the combined effect of $X_1$ and $X_2$ on Y but we might also be looking at the unique impact of $X_1$ and $X_2$ on Y.

**Consequences of Multicolinearity:**

1) The variances(square of standard errors) of the estimators of regression coefficients of β that is B will be inflated. We know that we calculate the t statistics by the formula

$$t = \frac{B}{SE(B)}$$

Therefore as SE(B) increases ,the t value decreases and p value increases and this might change a significant explanatory variable into a nonsignificant variable. This might cause us to not reject Null when actually we should.

2. Magnitude of B might be different from what we are expecting.

3. The sign of B might be opposite of what we are expecting.

4. Adding or removing one or more explanatory variable X causes large changes in B

5. Removing some data might change B drastically

6. Some times F is significant but t for the B coefficient is not significant .

Detection of Multicollinearity:

Variance Inflation Factor

Variance Inflation factor is a more rigorous way to check for multicollinearity as compared to Correlation Coefficient.

$$VIF = \frac{1}{1 - R_i^2}$$

VIF for $X_1$ can be obtained by regressing $X_1$ on $X_2$ and $X_3$

$X_1=a_0+a_1X_1+a_2X_2$

If VIF is too high then there is multicollinearity.

If VIF=1 then no collinearity

If VIF>5 then there is multicollinearity. If it is 5 then the variance of B is 5 times what it actually should be if there was no correlation.

**Solutions for multiCollinearity**

1. Drop the variable causing collinearity. If there are many explanatory variables then stepwise regression can be applied. Drop the least significant variable : has highest pvalue amongst the correlated variable.

2. If the number of variables are small then the solution is to not mention the impact of $X_1$ and $X_2$ on Y since the they are correlated.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

We can still conclude the combined impact of the variables on Y but not the individual impacts.

3. Stepwise regression

```
> lm_one = lm(prestige ~ education+income , data=Prestige)
> summary(lm_one)

Call:
lm(formula = prestige ~ education + income, 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 *
education    4.2921076  0.3360645  12.772  < 2e-16 ***
income       0.0012415  0.0002185   5.682 1.45e-07 ***
---
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

> vif(lm_one)
education    income
 1.491621  1.491621
>
```

```
> lm_two = lm(prestige ~ education+income+type , data=Prestige)
> summary(lm_two)

Call:
lm(formula = prestige ~ education + income + 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
education    3.6731661  0.6405016   5.735 1.21e-07 ***
income       0.0010132  0.0002209   4.586 1.40e-05 ***
typeprof     6.0389707  3.8668551   1.562    0.122
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

> vif(lm_two)
             GVIF Df GVIF^(1/(2*Df))
education 5.973932  1        2.444163
income    1.681325  1        1.296659
type      6.102131  2        1.571703
>
```

```
> lm_three<- lm(prestige ~ education+type+education:type, data=Prestige)
> summary(lm_three)

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

Residuals:
     Min       1Q   Median       3Q      Max
-19.7095  -5.3938   0.8125   5.3968  16.1411

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.2936     8.6470  -0.497    0.621
education            4.7637     1.0247   4.649 1.11e-05 ***
typeprof            18.8637    16.8881   1.117    0.267
typewc             -24.3833    21.7777  -1.120    0.266
education:typeprof  -0.9808     1.4495  -0.677    0.500
education:typewc     1.6709     2.0777   0.804    0.423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.827 on 92 degrees of freedom
Multiple R-squared:  0.8012,    Adjusted R-squared:  0.7904
F-statistic: 74.14 on 5 and 92 DF,  p-value: < 2.2e-16

> vif(lm_three)
                     GVIF Df GVIF^(1/(2*Df))
education         12.56332  1        3.544477
type           11059.20462  2       10.254889
education:type 16806.89241  2       11.386018
>
```

```
> lm_four<- lm(prestige ~ education+income+type+education:type+income:type, data=Prestige)
> summary(lm_four)

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

> vif(lm_four)
                     GVIF Df GVIF^(1/(2*Df))
education         16.82500  1        4.101829
income            13.44351  1        3.666539
type           11278.44113  2       10.305339
education:type 21307.65754  2       12.081864
income:type      188.92503  2        3.707425

> step(lm_four, direction="backward")
Start:  AIC=369.87
prestige ~ education + income + type + education:type + income:type

                Df Sum of Sq    RSS    AIC
<none>                       3552.9 369.87
- education:type  2    238.40 3791.3 372.24
- income:type     2    951.77 4504.6 389.13

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

>
```

AIC is to Akaike's Information Criterion: lower pvalues are better.