Multiple Linear Regression| R programming language

Preliminaries

Script Editor

```
rm(list = ls())
directory <- "C:/Econometrics/DataR/"

# Install packages
PackageNames <- c("tidyverse", "stargazer", "magrittr", "car")
for(i in PackageNames){
  if(!require(i, character.only = T)){
    install.packages(i, dependencies = T)
    require(i, character.only = T)
  }
}
```

Data Exploration

Script Editor

```
wage1 <- read.csv("C:\\Users\\amalz\\OneDrive\\Desktop\\wage1.csv")
wage1 %>%
  select(wage, educ, exper, tenure) %>%
  head(05)
wage1 %>%
  select(wage, educ, exper, tenure) %>%
  str
wage1 %>%
  select(wage, educ, exper, tenure) %>%
  stargazer(type = "text")
```

Console

```
> wage1 %>%
+   select(wage, educ, exper, tenure) %>%
+   head(05)
  wage educ exper tenure
1 3.10   11     2      0
2 3.24   12    22      2
3 3.00   11     2      0
4 6.00    8    44     28
5 5.30   12     7      2
> wage1 %>%
+   select(wage, educ, exper, tenure) %>%
+   str
'data.frame':	526 obs. of  4 variables:
 $ wage  : num  3.1 3.24 3 6 5.3 ...
 $ educ  : int  11 12 11 8 12 16 18 12 12 17 ...
 $ exper : int  2 22 2 44 7 9 15 5 26 22 ...
 $ tenure: int  0 2 0 28 2 8 7 3 4 21 ...
> wage1 %>%
+   select(wage, educ, exper, tenure) %>%
+   stargazer(type = "text")

==========================================
Statistic  N   Mean  St. Dev.  Min    Max
------------------------------------------
wage      526 5.896   3.693   0.530 24.980
educ      526 12.563  2.769     0     18
exper     526 17.017  13.572    1     51
tenure    526 5.105   7.224     0     44
------------------------------------------
```

Multiple Linear Regression| R programming language

Multiple regression (Example 1)

Data 01: wage1.csv

Script Editor

```
model_multiple1 <- lm(wage ~ educ + exper, wage1)
model_multiple2 <- lm(wage ~ educ + exper + tenure, wage1)
summary(model_multiple1)
summary(model_multiple2)
```

Console

```
> model_multiple1 <- lm(wage ~ educ + exper, wage1)
> model_multiple2 <- lm(wage ~ educ + exper + tenure, wage1)
> summary(model_multiple1)

Call:
lm(formula = wage ~ educ + exper, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5532 -1.9801 -0.7071  1.2030 15.8370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.39054    0.76657  -4.423 1.18e-05 ***
educ         0.64427    0.05381  11.974  < 2e-16 ***
exper        0.07010    0.01098   6.385 3.78e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.257 on 523 degrees of freedom
Multiple R-squared:  0.2252,	Adjusted R-squared:  0.2222 
F-statistic: 75.99 on 2 and 523 DF,  p-value: < 2.2e-16
> summary(model_multiple2)

Call:
lm(formula = wage ~ educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6068 -1.7747 -0.6279  1.1969 14.6536 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
educ         0.59897    0.05128  11.679  < 2e-16 ***
exper        0.02234    0.01206   1.853   0.0645 .  
tenure       0.16927    0.02164   7.820 2.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,	Adjusted R-squared:  0.3024 
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

Interpretations and explanations of components of results same as that of simple regression

Multiple Linear Regression| R programming language

Multiple regression: Predicted values and residuals

Script Editor

```
wage1 %<>% mutate(wagehat = fitted(model_multiple2),
                  uhat = residuals(model_multiple2))
wage1 %>%
  select(wage, wagehat, uhat) %>%
  head(10)
wage1 %>%
  select(wage, wagehat, uhat) %>%
  stargazer(type = "text")
```

Console

```
> wage1 %<>% mutate(wagehat = fitted(model_multiple2),
+                   uhat = residuals(model_multiple2))
> wage1 %>%
+   select(wage, wagehat, uhat) %>%
+   head(10)
    wage   wagehat       uhat
1   3.10  3.760560 -0.6605599
2   3.24  5.144853 -1.9048527
3   3.00  3.760560 -0.7605599
4   6.00  7.641447 -1.6414468
5   5.30  4.809760  0.4902401
6   8.75  8.265911  0.4840889
7  11.25  9.428610  1.8213903
8   5.00  4.934350  0.0656505
9   3.60  5.572748 -1.9727481
10 18.18 11.355782  6.8242176
> wage1 %>%
+   select(wage, wagehat, uhat) %>%
+   stargazer(type = "text")

===========================================
Statistic  N   Mean  St. Dev.  Min    Max
-------------------------------------------
wage      526 5.896   3.693   0.530  24.980
wagehat   526 5.896   2.044   -1.934 13.008
uhat      526 -0.000  3.076   -7.607 14.654
-------------------------------------------
```

Interpretations and explanations of components of results same as that of simple regression

Multiple Linear Regression| R programming language

Multiple regression: Partialling out

Partialing out is a statistical technique used in regression analysis to isolate the effect of a specific variable (here, educ*) on the dependent variable (*wage*), removing the influence of other control variables (e.g.,* exper *and* tenure*). The goal is to show that the coefficient of* educ *in the full regression model is the same as the coefficient of the residual (*ehat*) in a simplified regression model i.e.* 0.59897

Script Editor

```
# wage = beta0 + beta1*educ + beta2*exper + beta3*tenure + u
# Same as the model 'model_multiple2'
summary(model_multiple2)

# educ = alpha0 + alpha2*exper + alpha3*tenure + e
model_partial <- lm(educ ~ exper + tenure, wage1)
summary(model_partial)

# predict residuals ehat
wage1 %<>% mutate(ehat = resid(model_partial))

# wage = gamma0 + beta1*ehat + v
lm(wage ~ ehat, wage1) %>% summary
```

Console

```
> summary(model_multiple2)

Call:
lm(formula = wage ~ educ + exper + tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6068 -1.7747 -0.6279  1.1969 14.6536 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.87273    0.72896  -3.941 9.22e-05 ***
educ         0.59897    0.05128  11.679  < 2e-16 ***
exper        0.02234    0.01206   1.853   0.0645 .  
tenure       0.16927    0.02164   7.820 2.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.084 on 522 degrees of freedom
Multiple R-squared:  0.3064,	Adjusted R-squared:  0.3024 
F-statistic: 76.87 on 3 and 522 DF,  p-value: < 2.2e-16
```

The coefficient of educ captures the effect of educ *on wage after accounting for the effects of* exper *and* tenure.

Multiple Linear Regression| R programming language

Multiple regression: Partialling out

Console *ctd...*

```
> # educ = alpha0 + alpha2*exper + alpha3*tenure + e
> model_partial <- lm(educ ~ exper + tenure, wage1)
> summary(model_partial)

Call:
lm(formula = educ ~ exper + tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4285  -1.3536  -0.2055   1.6550   5.9791

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.574964   0.184324  73.647  < 2e-16 ***
exper       -0.073785   0.009761  -7.559 1.83e-13 ***
tenure       0.047680   0.018337   2.600  0.00958 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.63 on 523 degrees of freedom
Multiple R-squared:  0.1013,	Adjusted R-squared:  0.09791
F-statistic: 29.49 on 2 and 523 DF,  p-value: 7.327e-13
```

This regression isolates the portion of educ *that is independent of* exper *and* tenure*. The residual from this regression,* ehat*, represents the variation in* educ *that is not explained by* exper *and* tenure.

```
> # predict residuals ehat
> wage1 %<>% mutate(ehat = resid(model_partial))
> # wage = gamma0 + beta1*ehat + v
> lm(wage ~ ehat, wage1) %>% summary

Call:
lm(formula = wage ~ ehat, data = wage1)

Residuals:
   Min     1Q Median     3Q    Max
-5.302 -2.059 -0.922  1.234 16.306

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.89610    0.14584   40.43   <2e-16 ***
ehat         0.59897    0.05561   10.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.345 on 524 degrees of freedom
Multiple R-squared:  0.1812,	Adjusted R-squared:  0.1797
F-statistic:   116 on 1 and 524 DF,  p-value: < 2.2e-16
```

This regression uses the residual ehat *instead of the original* educ*. Here,* ehat *reflects the "pure" variation in* educ*, excluding its correlation with* exper *and* tenure*.*

The partialing-out process removes the overlap (shared variance) of educ *with other variables (*exper *and* tenure*). This ensures that the isolated effect of* educ *on* wage *remains the same whether it is modeled directly or indirectly through its residual (*ehat*).* ehat *represents the part of* educ *that is uncorrelated with* exper *and* tenure*, which is essentially what* educ *contributes uniquely in the full model.*

Multiple Linear Regression| R programming language

Multiple regression (Example 2)

Data 02: CEOSAL1.csv

Script Editor

```r
# CEO Salary Example
CEOSAL1 <- read.csv(paste0(directory, "CEOSAL1.csv"))
CEOSAL1 %>%
  select(salary, lsalary, roe, sales, lsales) %>%
  head(10)
CEOSAL1 %>%
  select(salary, lsalary, roe, sales, lsales) %>%
  str
CEOSAL1 %>%
  select(salary, lsalary, roe, sales, lsales) %>%
  stargazer(type = "text")

# Linear form
model_linear <- lm(salary ~ roe + sales, CEOSAL1)
summary(model_linear)

# Linear-log model
model_linear_log <- lm(salary ~ roe + lsales, CEOSAL1)
summary(model_linear_log)

# Log-linear form
model_log_linear <- lm(lsalary ~ roe + sales, CEOSAL1)
summary(model_log_linear)

# Log-log form
model_log_log <- lm(lsalary ~ roe + lsales, CEOSAL1)
summary(model_log_log)
```


Console

```
+   select(salary, lsalary, roe, sales, lsales) %>%
+   head(10)
   salary  lsalary  roe   sales    lsales
1    1095 6.998509 14.1 27595.0 10.225389
2    1001 6.908755 10.9  9958.0  9.206132
3    1122 7.022868 23.5  6125.9  8.720281
4     578 6.359574  5.9 16246.0  9.695602
5    1368 7.221105 13.8 21783.2  9.988894
6    1145 7.043160 20.0  6021.4  8.703075
7    1078 6.982863 16.4  2266.7  7.726080
8    1094 6.997596 16.3  2966.8  7.995239
9    1237 7.120444 10.5  4570.2  8.427312
10    833 6.725034 26.3  2830.0  7.948032
> CEOSAL1 %>%
+   select(salary, lsalary, roe, sales, lsales) %>%
+   str
'data.frame':   209 obs. of  5 variables:
 $ salary : int  1095 1001 1122 578 1368 1145 1078 1094 1237 833 ...
 $ lsalary: num  7 6.91 7.02 6.36 7.22 ...
 $ roe    : num  14.1 10.9 23.5 5.9 13.8 ...
 $ sales  : num  27595 9958 6126 16246 21783 ...
 $ lsales : num  10.23 9.21 8.72 9.7 9.99 ...
> CEOSAL1 %>%
+   select(salary, lsalary, roe, sales, lsales) %>%
+   stargazer(type = "text")

=====================================================
Statistic  N    Mean     St. Dev.    Min      Max
-----------------------------------------------------
salary    209 1,281.120 1,372.345    223     14,822
lsalary   209   6.950     0.566     5.407    9.604
roe       209  17.184     8.519     0.500    56.300
sales     209 6,923.793 10,633.270 175.200 97,649.900
lsales    209   8.292     1.013     5.166    11.489
```

Multiple Linear Regression| R programming language

Console *ctd...*

```
Call:
lm(formula = salary ~ roe + sales, data = CEOSAL1)

Residuals:
    Min      1Q  Median      3Q     Max
-1501.8  -492.6  -232.0   123.3 13575.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.306e+02  2.239e+02   3.710 0.000267 ***
roe         1.963e+01  1.108e+01   1.772 0.077823 .
sales       1.634e-02  8.874e-03   1.842 0.066973 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1359 on 206 degrees of freedom
Multiple R-squared:  0.02917,	Adjusted R-squared:  0.01975
F-statistic: 3.095 on 2 and 206 DF,  p-value: 0.04739
Call:
lm(formula = salary ~ roe + lsales, data = CEOSAL1)

Residuals:
    Min      1Q  Median      3Q     Max
-1024.1  -443.2  -223.3    68.8 13666.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1482.29     815.97  -1.817   0.0707 .
roe            22.67      10.98   2.065   0.0402 *
lsales        286.26      92.33   3.100   0.0022 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1339 on 206 degrees of freedom
Multiple R-squared:  0.05718,	Adjusted R-squared:  0.04803
F-statistic: 6.247 on 2 and 206 DF,  p-value: 0.002323
Call:
lm(formula = lsalary ~ roe + sales, data = CEOSAL1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52016 -0.27115 -0.00942  0.25605  2.69491

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.585e+00  8.750e-02  75.258  < 2e-16 ***
roe         1.494e-02  4.329e-03   3.452 0.000674 ***
sales       1.565e-05  3.468e-06   4.512 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.531 on 206 degrees of freedom
Multiple R-squared:  0.1295,	Adjusted R-squared:  0.121
F-statistic: 15.32 on 2 and 206 DF,  p-value: 6.264e-07
all:
m(formula = lsalary ~ roe + lsales, data = CEOSAL1)

esiduals:
    Min      1Q  Median      3Q     Max
0.9464 -0.2888 -0.0322  0.2261  2.7830

oefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept) 4.362167   0.293878  14.843  < 2e-16 ***
oe         0.017872   0.003955   4.519 1.05e-05 ***
sales      0.275087   0.033254   8.272 1.62e-14 ***
--
ignif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

esidual standard error: 0.4822 on 206 degrees of freedom
ultiple R-squared:  0.282,	Adjusted R-squared:  0.275
-statistic: 40.45 on 2 and 206 DF,  p-value: 1.519e-15
```

Interpretations and explanations of components of results same as that of simple regression

Multiple Linear Regression| R programming language

Perfect collinearity

When perfect collinearity exists, R drops one variable to ensure the regression is estimable. This ensures that the matrix of explanatory variables is invertible. Perfect collinearity makes it impossible to estimate the unique contribution of collinear variables. It typically arises when variables are created as linear transformations of one another (e.g., creating male = 1 - female).

Script Editor

```r
# Model for wage with female
model_no_collinearity <- lm(wage ~ educ + female, wage1)
summary(model_no_collinearity)

# Male is an exact linear function of female (perfect collinearity)
wage1 %<>% mutate(male = 1 - female)

# Model for wage with male
model_no_collinearity1 <- lm(wage ~ educ + male, wage1)
summary(model_no_collinearity1)

# Try to run regression with both female and male
model_collinearity <- lm(wage ~ educ + female + male, wage1)
summary(model_collinearity)
# This model cannot be estimated because of perfect collinearity
# R drops one variable, but it chooses which one to drop

# Run regression with "no constant" option
model_no_constant <- lm(wage ~ 0 + educ + female + male, wage1)
summary(model_no_constant)
```

Console

```
Call:
lm(formula = wage ~ educ + female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9890 -1.8702 -0.6651  1.0447 15.4998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.62282    0.67253   0.926    0.355
educ         0.50645    0.05039  10.051  < 2e-16 ***
female      -2.27336    0.27904  -8.147 2.76e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.186 on 523 degrees of freedom
Multiple R-squared:  0.2588,	Adjusted R-squared:  0.256
F-statistic: 91.32 on 2 and 523 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = wage ~ educ + male, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9890 -1.8702 -0.6651  1.0447 15.4998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.65055    0.65232  -2.530   0.0117 *
educ         0.50645    0.05039  10.051  < 2e-16 ***
male         2.27336    0.27904   8.147 2.76e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.186 on 523 degrees of freedom
Multiple R-squared:  0.2588,	Adjusted R-squared:  0.256
F-statistic: 91.32 on 2 and 523 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression| R programming language

Console *ctd...*

```
Call:
lm(formula = wage ~ educ + female + male, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9890 -1.8702 -0.6651  1.0447 15.4998 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.62282    0.67253   0.926    0.355    
educ         0.50645    0.05039  10.051  < 2e-16 ***
female      -2.27336    0.27904  -8.147 2.76e-15 ***
male              NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.186 on 523 degrees of freedom
Multiple R-squared:  0.2588,	Adjusted R-squared:  0.256 
F-statistic: 91.32 on 2 and 523 DF,  p-value: < 2.2e-16
Call:
lm(formula = wage ~ 0 + educ + female + male, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9890 -1.8702 -0.6651  1.0447 15.4998 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
educ    0.50645    0.05039  10.051   <2e-16 ***
female -1.65055    0.65232  -2.530   0.0117 *  
male    0.62282    0.67253   0.926   0.3548    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.186 on 523 degrees of freedom
Multiple R-squared:  0.7914,	Adjusted R-squared:  0.7902 
F-statistic: 661.5 on 3 and 523 DF,  p-value: < 2.2e-16
```


Model Type	Variables Included	Outcome
No Collinearity	`educ`, `female`	Model runs successfully; coefficients estimated.
With Collinearity	`educ`, `female`, `male`	One variable (`female` or `male`) is dropped automatically.
No Constant	`educ`, `female`, `male`	Both variables included; intercept removed.

Multiple Linear Regression| R programming language

Multicollinearity using VIF: Data exploration

Multicollinearity is when regressors are highly correlated with each other.

Script Editor

```
elemapi2 <- read.csv("C:\\Users\\amalz\\OneDrive\\Desktop\\elemapi2.csv")
elemapi2 %<>% select(api00, avg_ed, grad_sch, col_grad)
str(elemapi2)
stargazer(elemapi2, type = "text")
head(elemapi2,05)
```

Console

```
> str(elemapi2)
'data.frame':	400 obs. of  4 variables:
 $ api00   : int  693 570 546 571 478 858 918 831 860 737 ...
 $ avg_ed  : num  NA NA NA 1.91 1.5 ...
 $ grad_sch: int  0 0 0 0 0 31 43 30 33 7 ...
 $ col_grad: int  0 0 0 9 0 36 34 50 42 27 ...
> stargazer(elemapi2, type = "text")

==========================================
Statistic  N   Mean   St. Dev.  Min   Max
------------------------------------------
api00     400 647.622 142.249   369   940
avg_ed    381  2.668   0.764   1.000 4.620
grad_sch  400  8.637   12.131    0     67
col_grad  400 19.698   16.471    0    100
------------------------------------------
> head(elemapi2,05)
  api00 avg_ed grad_sch col_grad
1   693     NA        0        0
2   570     NA        0        0
3   546     NA        0        0
4   571   1.91        0        9
5   478   1.50        0        0
```


Multicollinearity using VIF: Correlation table

Script Editor

```
elemapi2 %>%
  select(-api00) %>%
  na.omit %>% # remove samples with NA
  cor
```

Console

```
> elemapi2 %>%
+   select(-api00) %>%
+   na.omit %>% # remove samples with NA
+   cor
            avg_ed  grad_sch  col_grad
avg_ed   1.0000000 0.7973054 0.8088654
grad_sch 0.7973054 1.0000000 0.4218830
col_grad 0.8088654 0.4218830 1.0000000
```

Multiple Linear Regression| R programming language

Multicollinearity using VIF: Regression and Calculating VIF

Note: If VIF>10, then we must remove that variable.

Script Editor

```
(model_high_vif <- lm(api00 ~ avg_ed + grad_sch + col_grad, elemapi2))
vif(model_high_vif)
```

Console

```
Call:
lm(formula = api00 ~ avg_ed + grad_sch + col_grad, data = elemapi2)

Coefficients:
(Intercept)       avg_ed     grad_sch     col_grad
    162.694      203.903       -1.091       -2.426

> vif(model_high_vif)
   avg_ed  grad_sch  col_grad
10.786031  4.536570  4.780201
```

VIF for avg_ed is >10, hence the model suffers from multicollinearity and must be rectified by removing this variable.

Multicollinearity using VIF: Regression after removing variable with VIF>10

Script Editor

```
(model_low_vif <- lm(api00 ~ grad_sch + col_grad, elemapi2))
vif(model_low_vif)
```

Console

```
Call:
lm(formula = api00 ~ grad_sch + col_grad, data = elemapi2)

Coefficients:
(Intercept)     grad_sch     col_grad
    545.109        5.829        2.648

> vif(model_low_vif)
grad_sch col_grad
1.245412 1.245412
```

This model is free from multicollinearity as all VIF<10. Also note that calculation of VIF and analysis of multicollinearity is not possible for simple linear regression.

Multiple Linear Regression| R programming language

Omitted variable bias: Data exploration

Omitted variable bias is when an omitted variable causes biased coefficients

Script Editor

```
HTV <- read.csv(paste0(directory, "HTV.csv"))
HTV %<>% select(wage, educ, abil)
str(HTV)
stargazer(HTV, type = "text")
head(HTV)
```

Console

```
> HTV <- read.csv("C:\\Users\\amalz\\OneDrive\\Desktop\\HTV.csv")
> HTV %<>% select(wage, educ, abil)
> str(HTV)
'data.frame':   1230 obs. of  3 variables:
 $ wage: num  12.02 8.91 15.51 13.33 11.07 ...
 $ educ: int  15 13 15 15 13 18 13 12 13 12 ...
 $ abil: num  5.03 2.04 2.48 3.61 2.64 ...
> stargazer(HTV, type = "text")

=============================================
Statistic   N    Mean  St. Dev.  Min    Max
---------------------------------------------
wage      1,230 13.288  9.082   1.024  91.309
educ      1,230 13.037  2.354     6      20
abil      1,230 1.797   2.184   -5.631 6.264
---------------------------------------------
> head(HTV)
       wage educ     abil
1 12.019231   15 5.027738
2  8.912656   13 2.037170
3 15.514334   15 2.475895
4 13.333333   15 3.609240
5 11.070110   13 2.636546
6 17.482517   18 3.474334
```


Omitted variable bias: Regression of the true model

 # True model with educ and ability

wage = beta0 + beta1*educ + beta2*abil + u

Script Editor

```
model_true <- lm(wage ~ educ + abil, HTV)
summary(model_true)
beta1 <- coef(model_true)["educ"]
beta1
beta2 <- coef(model_true)["abil"]
beta2
```

Console

```
Call:
lm(formula = wage ~ educ + abil, data = HTV)

Residuals:
    Min      1Q  Median      3Q     Max
-16.696  -4.568  -1.233   2.380  70.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.5226     1.5434  -1.634  0.10242
educ          1.1530     0.1272   9.068  < 2e-16 ***
abil          0.4333     0.1371   3.161  0.00161 **
```

Console *ctd...*

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.443 on 1227 degrees of freedom
Multiple R-squared:  0.1372,	Adjusted R-squared:  0.1358
F-statistic: 97.56 on 2 and 1227 DF,  p-value: < 2.2e-16

> beta1 <- coef(model_true)["educ"]
> beta1
    educ
1.153016
> beta2 <- coef(model_true)["abil"]
> beta2
    abil
0.433261
```

Omitted variable bias: Regression between the independent variables

```
# Model between ability and education
# abil = delta0 + delta1*educ + v
```

Script Editor

```
model_abil <- lm(abil ~ educ, HTV)
summary(model_abil)
delta1 <- coef(model_abil)["educ"]
delta1
```

Console

```
Call:
lm(formula = abil ~ educ, data = HTV)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6935 -1.0553  0.2122  1.1662  4.7699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.3890     0.2822  -19.10   <2e-16 ***
educ          0.5512     0.0213   25.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.758 on 1228 degrees of freedom
Multiple R-squared:  0.3529,	Adjusted R-squared:  0.3523
F-statistic: 669.6 on 1 and 1228 DF,  p-value: < 2.2e-16

> delta1 <- coef(model_abil)["educ"]
> delta1
     educ
0.5511552
```

Multiple Linear Regression| R programming language

Omitted variable bias: Regression of the compromised (biased) model

 # Model where ability is omitted variable, so coefficient on educ is biased

wage = (beta0+beta2*delta0) + (beta1+beta2*delta1)*educ +(beta2*v+u)

Script Editor

```
model_omitted <- lm(wage ~ educ, HTV)
summary(model_omitted)
beta1_biased <- coef(model_omitted)["educ"]
beta1_biased
```

Console

```
Call:
lm(formula = wage ~ educ, data = HTV)

Residuals:
    Min      1Q  Median      3Q     Max
-17.370  -4.558  -1.267   2.579  69.722

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8574     1.3601  -3.571 0.000369 ***
educ          1.3918     0.1027  13.557  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.474 on 1228 degrees of freedom
Multiple R-squared:  0.1302,	Adjusted R-squared:  0.1295
F-statistic: 183.8 on 1 and 1228 DF,  p-value: < 2.2e-16

> beta1_biased <- coef(model_omitted)["educ"]
> beta1_biased
    educ
1.39181
```

Bias =beta2*delta1

Biased coefficient= beta1+beta2*delta1 = 1.3981

Note:- The coefficient on educ is biased but has lower standard error

Omitted variable bias: Regression of the compromised (biased) model

 # Homoscedasticy is when the variance of the error is constant for each x

 # Heteroscedasticy is when the variance of the error is not constant for each x

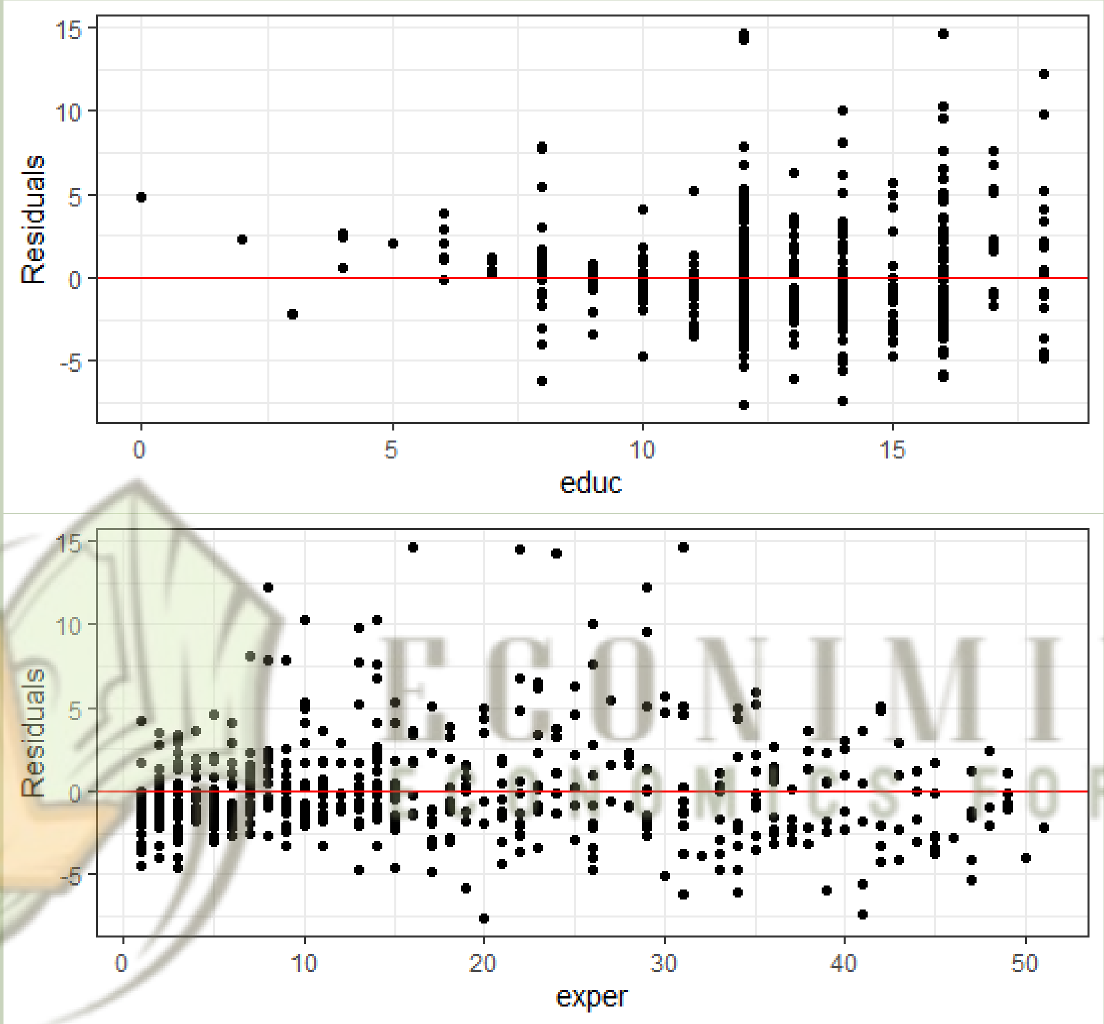
Script Editor

```
# Plotting residuals with "geom_point"
ggplot(data = wage1, mapping = aes(x = educ)) +
  theme_bw() +
  geom_point(mapping = aes(y = uhat)) +
  geom_hline(yintercept = 0, col = 'red') + # add a horizontal line
  ylab(label = "Residuals") # change y-axis label

ggplot(data = wage1, mapping = aes(x = exper)) +
  theme_bw() +
  geom_point(mapping = aes(y = uhat)) +
  geom_hline(yintercept = 0, col = 'red') +
  ylab(label = "Residuals")
```

Multiple Linear Regression| R programming language

Plots pane



Graphs show heteroscedasticity for educ and homoscedasticity for exper