Lab03: Matrix Operations, Model Diagnostics & IV

## Part 1: Matrix Operations (5 points)

## Task 1: Manual matrix operations and regression analysis with matrices [1.5 points]

You are given a vector of the dependent variable and the design matrix

[a] Calculate ***manually*** the vector regression coefficients .   
The analytical equation for a inverse matrix is

Type your calculations with Word’s equation editor or ***clearly*** by hand. Use 3 decimal points precision. (0.5 points)

[b] Write your own OLS  function using the dependent vector and the associated design matrix as input. Your function should return the vector of the estimated regression coefficients. Repeat the analysis from task 1 [a] using your  function and compare the estimated regression coefficients with those in task 1 [a]. (0.5 point)

y <- matrix(c(3,6,3,6,3,9),ncol=1)

x0 <- rep(1,6)

x1 <- c(1,3,1,3,1,5)

X <- cbind(x0,x1)

tXX<- t(X)%\*%X

tXXInv <- solve(tXX)

(b <- tXXInv%\*%t(X)%\*%y)

[,1]

x0 1.5

x1 1.5

The result is same

[c] Use 's matrix operations to calculate for a dependent variable , the design matrix and the diagonal weights matrix the weighted regression coefficients with the formula . (0.5 points)

X <- matrix(c(1,1,1,1,3,5),nrow = 3)

W <- matrix(c(3,0,0,0,2,0,0,0,1),nrow = 3,ncol = 3)

y <- matrix(c(3,6,9),ncol=1)

(b <- (X %>% t() %\*% W %\*% X) %>% solve() %\*% (X %>% t()) %\*% W %\*% y)

[,1]

[1,] 1.5

[2,] 1.5

[d] Compare the estimated regression coefficients from task 1 [b] with those from task 1 [c]. Explain why they are identically. Hint: what is the effect of the weights matrix . (0.5 points)

The first row of matrix X in [c] is c(1,1), the weight is 3. In task[a], c(1,1) repeat 3 times, since no weight matrix exists in task[a], so we could use the frequency to represents their weights. In this situation, matrices in task[c] and task[a] indicates the same information, that’s why we would get identical results from those two tasks.

### Task 2: Coding schemes of categorical variables (3 points)

Provide the  syntax code of your answers. You can either use the **lm(…)** or your coded ordinary least squares function for this task

[a] Enter the matrix and the design matrices to separate matrix objects into  and show these object in your answer (0.5 points):

and are given in the ***indicator coding*** scheme ( codes it as **contrasts(factor) <- "contr.treatment"**) whereas and are given in the ***centered coding*** scheme ( codes it as **contrasts(factor) <- "contr.sum"** and Hamilton p 99 calls it ***effect*** coding). In and the last category is suppressed, whereas in and the second category is suppressed due to the redundancy among a full set of indicator variables.

row1 <- rep(1,9)

y <- matrix(c(8,6,4,1,3,2,9,5,7),ncol=1)

row2 <- c(rep(1,3),rep(0,6))

row3 <- c(rep(0,3),rep(1,3),rep(0,3))

row4 <- c(rep(0,6),rep(1,3))

row5 <- c(rep(1,3),rep(0,3),rep(-1,3))

row6 <- c(rep(0,3),rep(1,3),rep(-1,3))

row7 <- c(rep(1,3),rep(-1,3),rep(0,3))

row8 <- c(rep(0,3),rep(-1,3),rep(1,3))

(x1 <- cbind(row1,row2,row3))

row1 row2 row3

[1,] 1 1 0

[2,] 1 1 0

[3,] 1 1 0

[4,] 1 0 1

[5,] 1 0 1

[6,] 1 0 1

[7,] 1 0 0

[8,] 1 0 0

[9,] 1 0 0

(x2 <- cbind(row1,row2,row4))

row1 row2 row4

[1,] 1 1 0

[2,] 1 1 0

[3,] 1 1 0

[4,] 1 0 0

[5,] 1 0 0

[6,] 1 0 0

[7,] 1 0 1

[8,] 1 0 1

[9,] 1 0 1

(x3 <- cbind(row1,row5,row6))

row1 row5 row6

[1,] 1 1 0

[2,] 1 1 0

[3,] 1 1 0

[4,] 1 0 1

[5,] 1 0 1

[6,] 1 0 1

[7,] 1 -1 -1

[8,] 1 -1 -1

[9,] 1 -1 -1

(x4 <- cbind(row1,row7,row8))

row1 row7 row8

[1,] 1 1 0

[2,] 1 1 0

[3,] 1 1 0

[4,] 1 -1 -1

[5,] 1 -1 -1

[6,] 1 -1 -1

[7,] 1 0 1

[8,] 1 0 1

[9,] 1 0 1

[b] Calculate the three group means of the observations , and as well as the global mean for all observations . (0.5 points)

mean = 6, mean = 2, mean = 7

mean = 5

[c] Find the four sets of estimated regression parameters for the intercept and group coefficients by regressing on the four design matrices , , and with your linear regression function that you have developed in task 1 [b] and enter these estimates into the table below (see columns *Assign Estimated Regression Coefficients*). (0.5 points)

Hints: (i) in the ***centered*** coding scheme the coefficient for the missing category can be calculated as the ***negative sum*** of the two other estimated parameters, i.e., . (ii) For the cornered coding scheme the values for the ***dashed*** cells cannot be calculated from the regression results.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Assign Estimated Regression Coefficients | | | | Give Expressions for the Means in Terms of the Estimate Regression Coefficients | | | |
| Model | **Coding** |  |  |  |  |  |  |  |  |
| y~X1 | ***cornered*** | 7 | -1 | -5 | ─ | ─ | 6 | 2 | 7 |
| y~X2 | ***cornered*** | 2 | 4 | ─ | 5 | ─ | 6 | 2 | 7 |
| y~X3 | ***centered*** | 5 | 1 | -3 | 2 | 5 | 6 | 2 | 7 |
| y~X4 | ***centered*** | 5 | 1 | -3 | 2 | 5 | 6 | 2 | 7 |

[d] For each design matrix the global mean and group means , and can be expressed as a function of the estimated regression coefficients. (0.5 point)  
Find the expressions for the means and write them into columns labels by “*Give Expressions…*” using the parameter symbols, e.g., , note that this is an invalid expression.

**For**  ,

**For**  ,

**For**

**For**

[e] Which coding scheme has a more *intuitive interpretation*? Justify your answer. (0.5 points)

I personally prefer effect coding. The b weights are now such that they specify the deviation of the identified group from the grand mean. This makes more sense to me. Because I could figure out the impact of each group compared with the global mean. And the significance of the b weights tells whether the group differs significantly from the grand mean rather than from a chosen cell mean.

[f] *Argue*, based on the four different coding schemes which, however, give identical predictions , whether it make more sense to test individual regression coefficients with a single *t*-tests or whether a simultaneous partial *F*-test of all coefficients associated with the factor is more appropriate? Think in terms of change coefficient values in dependence of the employed coding scheme. (0.5 points)

For the center encoding method (models 1 and 2), the regression coefficient varies from the different target category selection, So I think it is meaningless to apply a single t-test on those regression coefficients. In contrast, the effect encoding always gives a set of identical regression coefficients, in this situation, a single t-test is acceptable.

## Part 2: Model Building and Diagnostics (4 points)

Open the **CPS1985** data-frame with **data("CPS1985",package="AER")**. Assign new row-names with the statement **rownames(CPS1985) <- 1:nrow(CPS1985)** to your data-frame. ***Study the description*** of the variable **experience** in the associated online help.

### Task 3: Multicollinearity diagnostics (2 points)

[a] For the variables **~log(wage)+education+age+experience** generate a scatterplot matrix. (0.5 points)

scatterplotMatrix(~log(wage)+education+age+experience,data =CPS1985,

main="Relationship between log(wage) and a set of independent variables",

pch=1, smooth=list(span = 0.35,lty.smooth=1, col.smooth="red", col.var="red"),

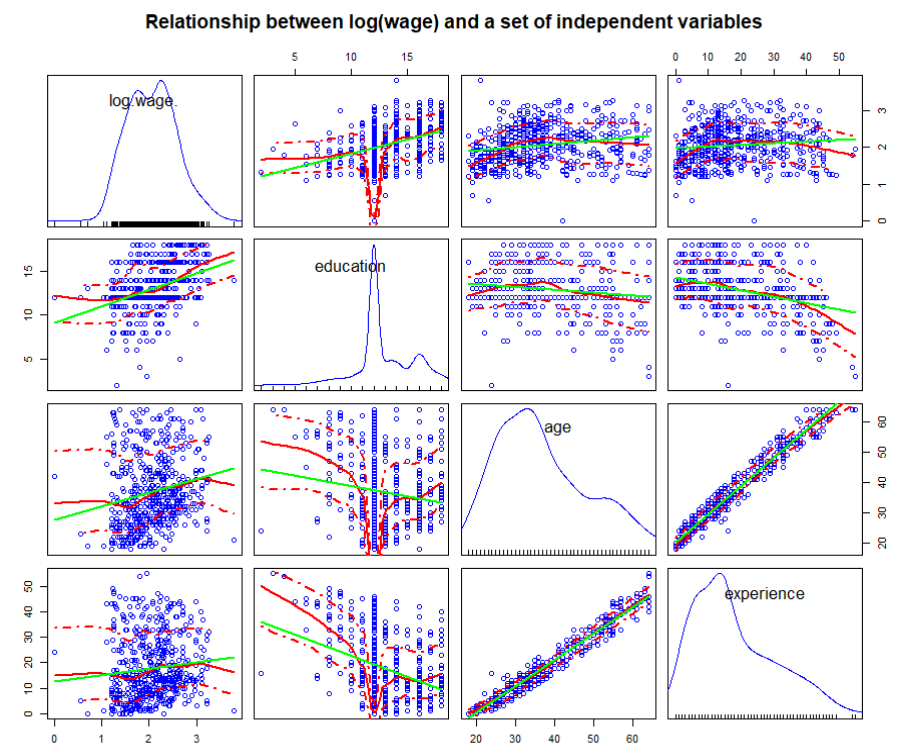
regLine=list(col="green"))

Based on the definition of the variables and the scatterplot matrix, which variables do you expect to be multicollinear? Justify your decisions.

Age and Experience

cor(CPS1985$age,CPS1985$experience)

[1] 0.9779612



[b] Estimate the model **log(wage)~education+experience** and calculate the ***variance inflation factors***. Fully interpret the estimated model and the ***VIF***s. (0.5 points)

model.2 <- lm(log(wage)~education+experience+age,data =CPS1985)

summary(model.2)

Residuals:

Min 1Q Median 3Q Max

-2.03371 -0.33057 0.04223 0.31897 1.83976

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.594169 0.124428 4.775 2.33e-06 \*\*\*

education 0.096414 0.008310 11.603 < 2e-16 \*\*\*

experience 0.011774 0.001756 6.707 5.10e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4695 on 531 degrees of freedom

Multiple R-squared: 0.2115, Adjusted R-squared: 0.2085

F-statistic: 71.21 on 2 and 531 DF, p-value: < 2.2e-16

vif(model.2)

education experience

1.142049 1.142049

From this linear model, both education and experience have a significant impact on wages. When education goes up 1 year, the wage level goes up 9% correspondingly. And experience increase 1 year, the wage level goes up 1% correspondingly. These two dependent variables would explain 21% differences among wages. Both two dependent variables have relatively low VIF score (close to 1), means they do not a strong relationship with each other.

[c] Estimate the augmented model **log(wage)~education+experience+age** and show the output. (1 point)

model.3 <- lm(log(wage)~education+experience+age,data =CPS1985)

summary(model.3)

Residuals:

Min 1Q Median 3Q Max

-2.03367 -0.33094 0.04165 0.31958 1.84066

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.84480 0.71884 1.175 0.240

education 0.13805 0.11791 1.171 0.242

experience 0.05353 0.11796 0.454 0.650

age -0.04173 0.11786 -0.354 0.723

Residual standard error: 0.4699 on 530 degrees of freedom

Multiple R-squared: 0.2117, Adjusted R-squared: 0.2072

F-statistic: 47.44 on 3 and 530 DF, p-value: < 2.2e-16

vif(model.3)

education experience age

229.5738 5147.9190 4611.4008

Address the following questions:

1. What do the ***VIF***s tell you?

Both experience and age have a very large VIF score, which means they have a strong correlation with other variables, which support our previous result, they are related to each other. High VIF score also indicates our model is not reliable.

1. ***What*** happened to the significances of the *t*-tests for the estimated regression parameters of the augmented model and ***why***?

Neither regression coefficient in our model is significant. because the multicollinearity would enlarge the standard error for each parameter, so even the regression coefficient is different from 0, we still fail to reject the none-hypothesis.

1. Why does the global *F*-test still remain significant?

The presence of multicollinearity neither under nor overstates the F-statistic or R-squared, because multicollinearity doesn’t impact model fit and OLS is still unbiased with MC. In plain terms, the F-statistic (and R-squared) is unaffected by multicollinearity.

### Task 4: Refined model specification (1 point)

[a] Estimate the model: **log(wage)~education+experience+gender+occupation+union** and *fully interpret* the estimated regression model. (0.5 point)

lm(formula = log(wage) ~ education + experience + gender + occupation + union, data = CPS1985)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.972050 0.132893 7.315 9.74e-13 \*\*\*

education 0.072296 0.009931 7.280 1.23e-12 \*\*\*

experience 0.010775 0.001670 6.454 2.49e-10 \*\*\*

genderfemale -0.203606 0.041860 -4.864 1.52e-06 \*\*\*

occupationtechnical 0.161965 0.069502 2.330 0.02017 \*

occupationservices -0.198521 0.061204 -3.244 0.00126 \*\*

occupationoffice -0.018791 0.063715 -0.295 0.76817

occupationsales -0.150690 0.082108 -1.835 0.06703 .

occupationmanagement 0.209102 0.076316 2.740 0.00635 \*\*

unionyes 0.216589 0.051117 4.237 2.68e-05 \*\*\*

Residual standard error: 0.4323 on 524 degrees of freedom

Multiple R-squared: 0.3404, Adjusted R-squared: 0.3291

F-statistic: 30.05 on 9 and 524 DF, p-value: < 2.2e-16

Based on this model, when the education year goes up 1 unit, the wage level would increase with almost 7%. And the experience goes up 1 year, the wage level would correspondingly increase by 1%. Compared with males, the salary of females is lower by almost 20%. And if you work in a union, the wage would roughly 21% higher than normal. As for occupation, based on the wage of tradesperson or assembly line worker, the wage of workers from technical and management positions would roughly 16% and 21% than them respectively. With the same standard, the wage of workers from services and sales positions would roughly 19% and 15% lower than them. The wage of workers who work in an office does not have a significant difference from them. All independent variables mentioned above would explain 34% variation in wage.

[b] Test whether the factor **occupation** is significant and if necessary refine the model specification accordingly. (0.25 points)

anova(model.5,model.4)

Analysis of Variance Table

Model 1: log(wage) ~ education + experience + gender + union

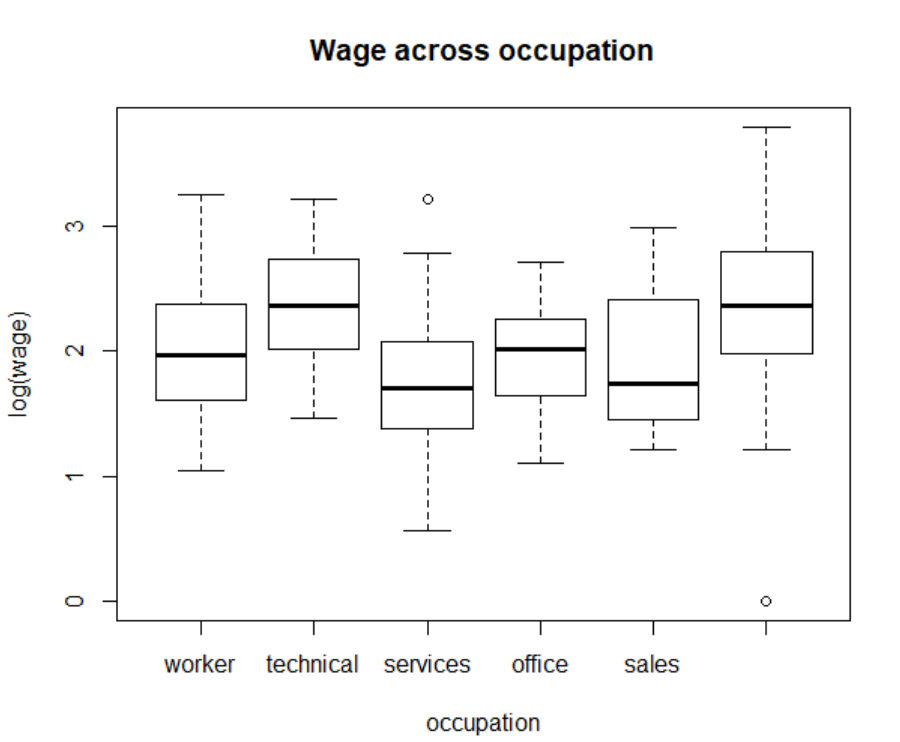
Model 2: log(wage) ~ education + experience + gender + occupation + union

Res.Df RSS Df Sum of Sq F Pr(>F)

1 529 105.092

2 524 97.915 5 7.1769 7.6816 5.535e-07 \*\*\*

boxplot(log(wage)~occupation, data=CPS1985, main="Wage across occupation")



From the figure above, we could discover that different types of the occupation have distinct wage distribution. And according to the result from the partial F test, the factor occupation also improves our model significantly.

[c] Investigate the model with **car::residualPlots( )**. Discuss the output and decide whether it is advisable to refine the model. (0.25 point)

Test stat Pr(>|Test stat|)

education -0.3367 0.7365

experience -4.2117 2.985e-05 \*\*\*

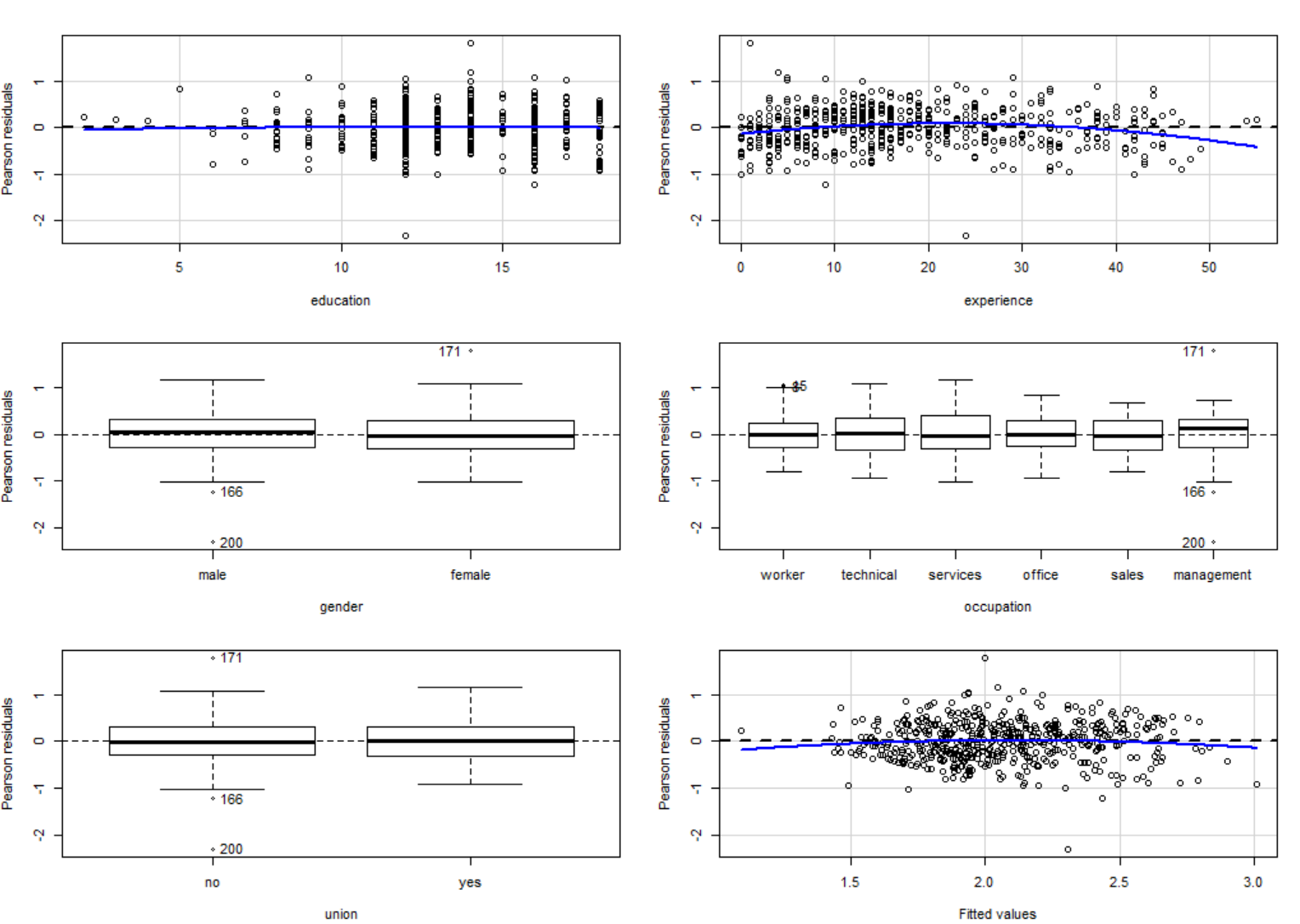
gender

occupation

union

Tukey test -1.3752 0.1691

Yes, from both the figure and statistic result, we could find that the experience may have quadratic relationship with wages. Therefore, we could add squared experience as a new dependent variable.



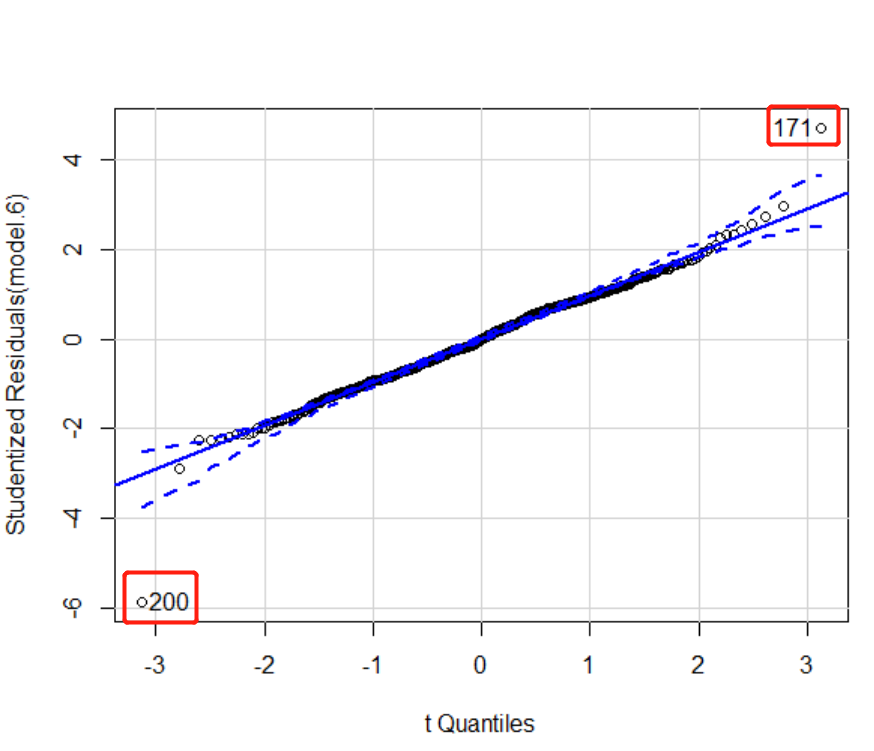
### Task 5: Case statistics of the final model (1 point)

[a] Generate the following plots and *interpret* them for your final model. (0.75 points)

1. Identify the two most extreme observations with a **car::qqPlot( )** and interpret it.

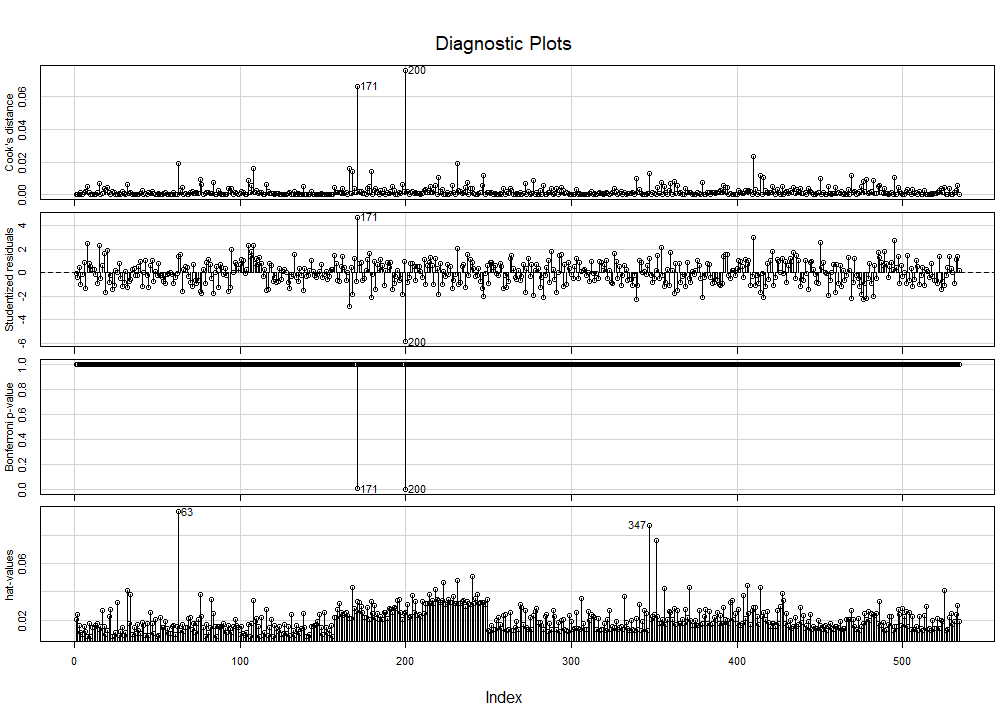
model.6 <- update(model.4, .~.+I(experience^2))

The residual for 171st and 200th observation is far from the center(0) of distribution, which means that two observations could not be predicted well using our model



1. Identify *potential* extreme observations with a **car::influenceIndexPlot( )** and interpret the plots.

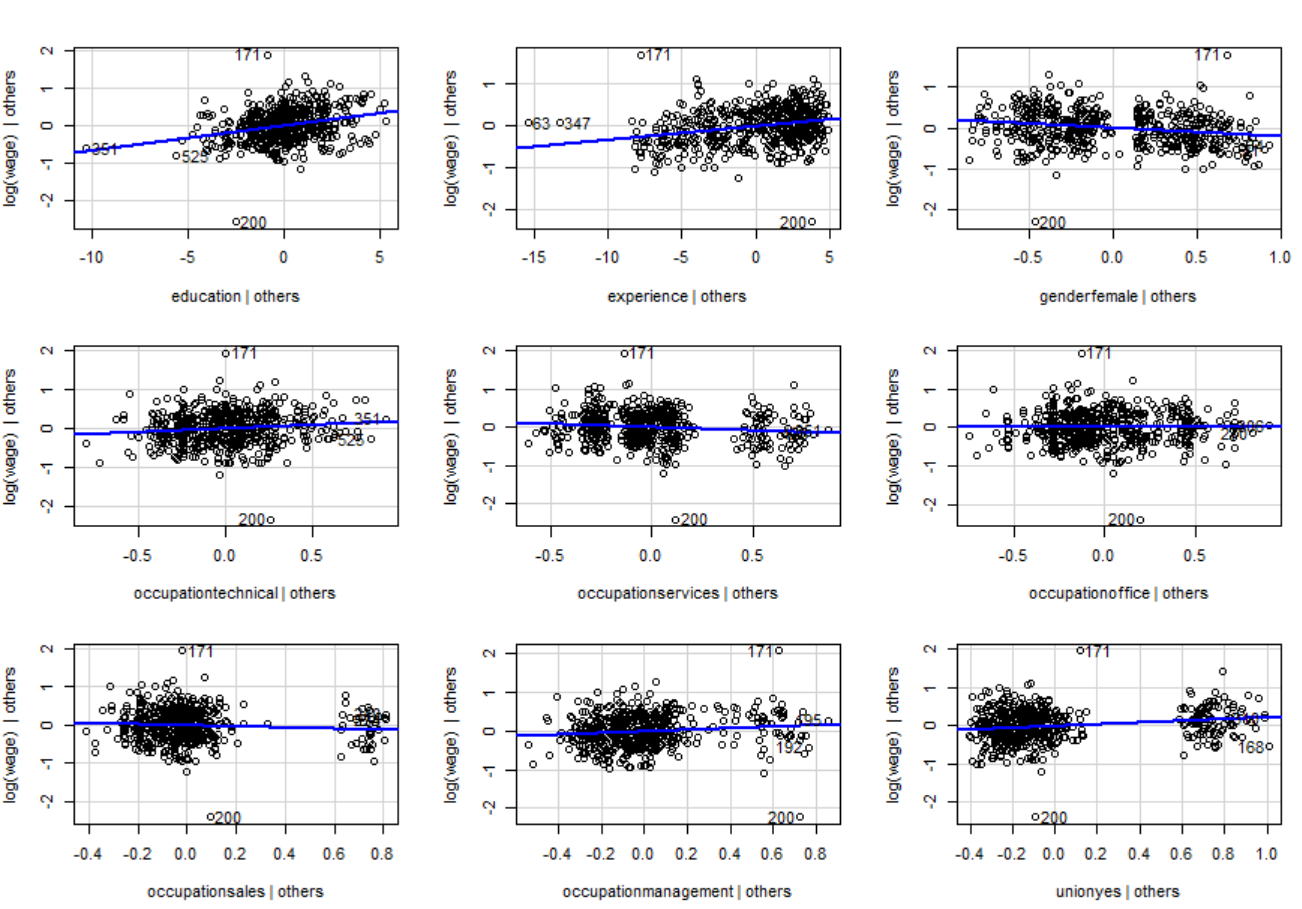
From cook's distance, residual and p-value, 171st and 200th observation could be regarded as the extreme records. Cook's distance refers to how far, on average, predicted y-values will move if the observation in question is dropped from the data set. The first plot tells about that the 171st and 200th observations have biggest impact on the model. The second figure indicates that we could not interpret or predict those two observations well using our model. The third plot also demonstrates those two observations are extreme compare to our assumed distribution. The fourth one is the predicted value plot.



1. Identify the two most extreme observation with a **car::avPlots( )** and interpret the plots.

Same as above, 171st and 200th observations.

Same result as above, 171st and 200th observations. This function explores the partial effects on all dependent variables and all the plots exhibit that those variables could not explain those two observations. Therefore, we could only either input additional variables or delete those two records.



[b] Inspect the ***two*** most extreme observations in the data-frame by examining their records. (0.25 points)

1. Discuss their attributes and argue if they are representative of the underlying population.

CPS1985[c(171,200),]

wage education experience age region gender occupation sector union married

171 44.5 14 1 21 other female management other no no

200 1.0 12 24 42 other male management other no yes

tapply(CPS1985$wage,CPS1985$occupation,summary)

$management

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 7.25 10.62 12.70 16.39 44.50

The woman who gets the highest wage has diminutive work experience, and not work for a union. The man who gets the lowest wage has 24 years of work experience and server as management. The relationship between wages and this information conflicts with our common sense. Therefore, I thought those two observations are not good for being representative of the underlying population.

1. Drop them from the data-frame and show your code for doing so.

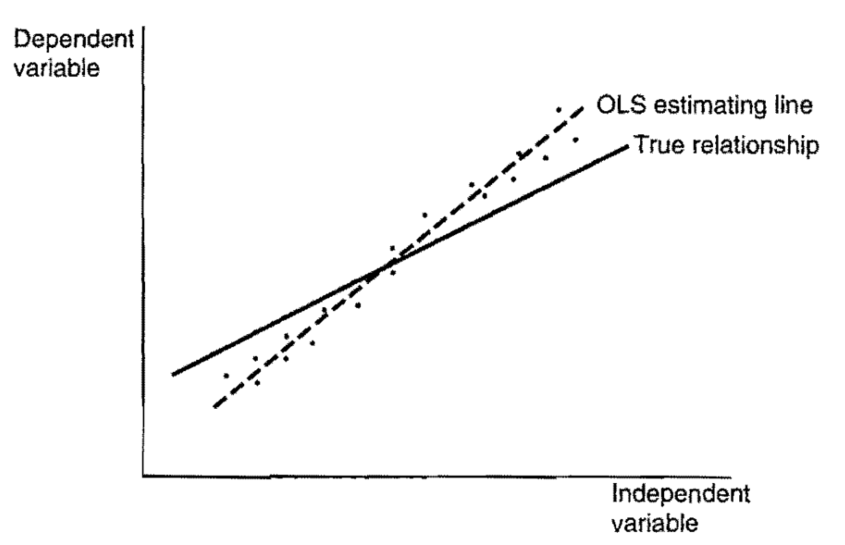
df\_new<- CPS1985[-c(171,200),]

## Part 3: Instrumental Variable Regression (3 points)

Use Kennedy’s Chapter 9 on “Instrumental Variable Estimation” (see IVRegKennedy2008.pdf) to explain the following topics in ***your own words***:

### Task 6: Explain Figure 9.1

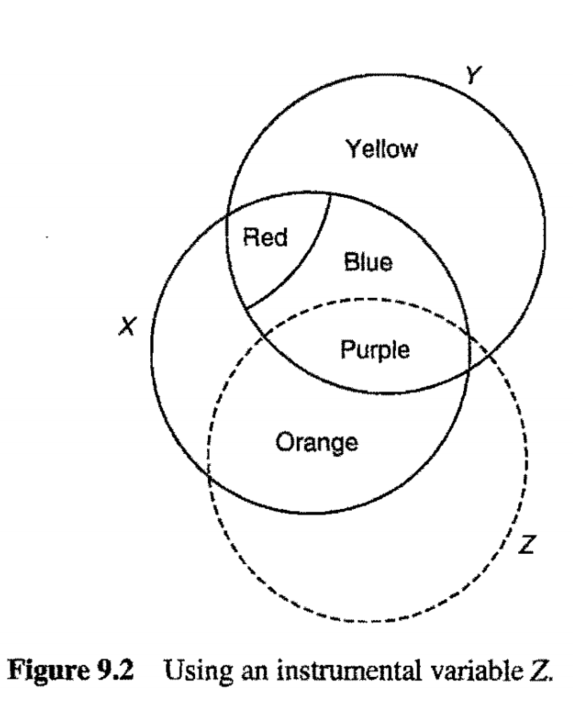
Why will the correlation between an *endogenous regressor* and the *disturbances* lead to a *biased* estimate of the associated regression coefficient. See also Figure 9.1. (0.75 point)



OLS is a method for minimizing the residual sum of the error term. Therefore, the OLS regression line would make observations evenly distributed on both sides. If the disturbance goes larger when the independent variable goes higher, which as shown in the picture above, the slope of the OLS regression line would go more inclined than the true relationship, that is why you would get a biased regressor using OLS under this circumstance, vis versa.

### Task 7: Explain Figure 9.2

Explain in details the components (i.e., the color segments) of Ballentine plot in Figure 9.2 and their relevance for the instrumental variable estimation approach. (1.25 points)



Y represents the dependent variable and X represents the independent variable. Red and Yellow area indicates error term related and unrelated to X respectively. So if we explore the relationship between X and Y, the information from Red, Blue and Purple would be used, then we would get a biased regressor since the Red area is correlated with X. Circle Z represents IV variable, so it does not overlap with Red and Yellow area (error term).it takes variation in X that matches up with a variation in Z and uses only this variation to compute the slope estimate. Therefore, only the purple area would be used for estimating IV slope. Since the variation of Y in the purple area entirely caused by X, the IV slope is unbiased.

### Task 8: Model with IV regression

The Stata dataset **card.dta** (use **foreign::read.dta( )**) with 3010 observations has several relevant variables aiming at explaining the percentage change in wages of respondents:

|  |  |
| --- | --- |
| Variable: | Description: |
| lwage | Dependent variable: logarithm of wage |
| educ | Endogenous regressor: education in years |
| age | Exogenous regressor: age in years |
| nearc2 | First instrument: dummy variable indicating that person lived near a 2 years college |
| nearc4 | Second instrument: dummy variable indicating that person lived near a 4 years college |

Build an instrumental variables model to explain the dependent variable and interpret the outcome of the ***weak instruments***, ***Wu-Hausman*** and ***Sargant*** tests. (1 point)

ivreg(formula = lwage ~ educ + age | age + nearc2 + nearc4, data = df2)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.661632 0.341735 7.789 9.26e-15 \*\*\*

educ 0.183175 0.024517 7.471 1.03e-13 \*\*\*

age 0.041632 0.003101 13.425 < 2e-16 \*\*\*

Diagnostic tests:

df1 df2 statistic p-value

Weak instruments 2 3006 33.647 3.54e-15 \*\*\*

Wu-Hausman 1 3006 51.938 7.21e-13 \*\*\*

Sargan 1 NA 3.118 0.0775 .

**Weak instruments:** *cor(educ,nearc2) = 0.04735098, cor(educ,nearc4) = 0.1442402.* The correlation between endogenous regressor and instruments are significant, so those two instruments would be considered as strong instruments, which means the IV estimates are reliable.

**Wu-Hausman:** The P-value are relatively low, so we could reject the H0 hypothesis and accept that education is significantly correlated with the disturbance, and we should perform the IV regression.

**Sargant:** *cor((df2$lwage - fitted(cig.iv)),df2$nearc2) = 0.0305* . The P-value indicates that we

take a 7% risk to say that at least one instrument variable is correlated with the disturbance (fromthe code we could see that the relationship between nearc2 and residual is 3%). Therefore, this

instrument variable is not ideal.