Sample answer Lab03: Matrix Operations & Model Diagnostics

Handed out: Monday, March 22, 2021

Return date: Friday, April 2, 2021 by midnight into the link **LAB03SUBMIT** in eLearning.

Grading: This lab counts 13 % towards your final grade

Part 1: Matrix Operations (4 points)

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

You are given a vector of the dependent variable $\mathbf{y} = (2,5,2,5,2,9)^T$ and the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 2 & 4 & 2 & 6 \end{pmatrix}^T$$

[a] Enter $\mathbf{y}$ and $\mathbf{X}$ into ®. Write your own OLS ® function using the dependent vector $\mathbf{y}$ and the associated design matrix $\mathbf{X}$ as input. Your function should return the vector of the estimated regression coefficients. (0.5 point)

```
X <- matrix(c(1,1,1,1,1,1,2,4,2,4,2,6), nrow=6)
y<- matrix(c(2,5,2,5,2,9), nrow=6)
solve(crossprod(X), crossprod(X,y))
     [,1]
[1,] -1.5
[2,]  1.7
```

[b] Use ®'s matrix operations to calculate for a dependent variable $\mathbf{y} = (2,5,9)^T$, the design matrix $\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \end{pmatrix}^T$ and the diagonal weights matrix $\mathbf{W} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ the weighted regression coefficients with the formula $\mathbf{b}_w = (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{y}$. (0.5 points)

```
X <- matrix(c(1,1,1,2,4,6), nrow=3)
y<- matrix(c(2,5,9), nrow=3)
W <- diag(c(3,2,1))
solve(t(X)%*%W%*%X, t(X)%*%W%*%y)
     [,1]
[1,] -1.5
[2,]  1.7
```

[c] Compare the estimated regression coefficients from task 1 [a] with those from task 1 [b]. Explain why they are identically. Hint: You may see the effect of the weights matrix $\mathbf{W}$ by running the estimation equation step-by-step one term at a time. (0.5 points)

There are only three different sets of observations:

      (1) y_1 = 2 with $X_1 = (1,2)^T$ with 3 observations
      (2) y_2 = 4 with $X_1 = (1,4)^T$ with 2 observations
   (3) y_3 = 6 with $X_1 = (1,6)^T$ with 1 observation

Comment: If we compare the original independent and dependent matrices in question (a) with those of question (c), we see that matrices in question (c) are suppressing duplicate observations and just show the unique observations. The weights matrix **W** provides information how frequent each unique observation is in the full regression system. The weights matrix in-between $\mathbf{X^T \cdot W \cdot X}$ and between $\mathbf{X^T \cdot W \cdot y}$ replicates the observations according to their frequency in the original dataset with 6 observations. Consequently, the same results, that is, of regression coefficients, are obtained by both equations.

Task 2: Coding schemes of categorical variables (2.5 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 9×1 matrix **y** and the 9×3 design matrices $\mathbf{X}_1$ to $\mathbf{X}_4$ separate matrix objects into ® and show these object in your answer (0.25 points):

$$\mathbf{y} = \begin{bmatrix} 7 \\ 5 \\ 3 \\ 1 \\ 3 \\ 2 \\ 9 \\ 5 \\ 7 \end{bmatrix} \quad \mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \mathbf{X}_4 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$\mathbf{X}_1$ and $\mathbf{X}_2$ are given in the ***indicator coding*** scheme (® codes it as **contrasts(factor) <- "contr.treatment"**) whereas $\mathbf{X}_3$ and $\mathbf{X}_4$ are given in the ***centered coding*** scheme (® codes it as **contrasts(factor) <- "contr.sum"** and Hamilton p 99 calls it ***effect*** coding). In $\mathbf{X}_1$ and $\mathbf{X}_3$ the last category is suppressed, whereas in $\mathbf{X}_2$ and $\mathbf{X}_4$ the second category is suppressed due to the redundancy among a full set of indicator variables.

```
(y <- matrix(c(7, 5, 3, 1, 3, 2, 9, 5, 7), ncol = 1))
(X1 <- matrix(c(rep(1, 12), rep(0, 9), rep(1, 3), rep(0, 3)), ncol = 3))
(X2 <- matrix(c(rep(1, 12), rep(0, 12), rep(1, 3)), ncol = 3))
(X3 <- matrix(c(rep(1, 12), rep(0, 3), rep(-1, 3), rep(0, 3), rep(1, 3),rep(-
1, 3)), ncol = 3))
(X4 <- matrix(c(rep(1, 12), rep(-1, 3), rep(0, 6), rep(-1, 3), rep(1, 3)),
ncol = 3))
```

[b] Calculate the three group means $\bar{y}_{g_1}$, $\bar{y}_{g_2}$ and $\bar{y}_{g_3}$ either manually or with ® of the observation groups $g_1 = \{y_1, y_2, y_3\}$, $g_2 = \{y_4, y_5, y_6\}$ and $g_3 = \{y_7, y_8, y_9\}$ as well as the global mean $\bar{y}_{global}$ for all observations $global = \{y_1, …, y_9\}$. (0.25 points)

```
(mean.group1 <- mean(y[1:3]))
5
(mean.group2 <- mean(y[4:6]))
2
(mean.group3 <- mean(y[7:9]))
7
(mean.global <- mean(y))
4.67
```

[c] Find the four sets of estimated regression coefficients for the intercept b_0 and group coefficients b_{g_1} and b_{g_2} or b_{g_3} by regressing **y** on the four design matrices X_1, X_2, X_3 and X_4 with your linear regression function that you have developed in task 1 [a] and enter these estimates into the table below (see columns *Assign Estimated Regression Coefficients*). (0.5 points)

```
Best <- function (x,y){solve(t(x)%*%x)%*%t(x)%*%y}
(parameter1 <- Best(X1,y))
(parameter2 <- Best(X2,y))
(parameter3 <- Best(X3,y))
(parameter4 <- Best(X4,y))
```

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 group parameters, for example, $b_{g_1} = -(b_{g_2} + b_{g_3})$. (ii) For the cornered coding scheme the values for the **dashed** cells cannot be calculated from the regression results.

Model	Coding	Assign Estimated Regression Coefficients				Give Expressions for the Means in Terms of the Estimate Regression Coefficients			
		b_0	b_{g_1}	b_{g_2}	b_{g_3}	$\bar{y}_{global}$	$\bar{y}_{g_1}$	$\bar{y}_{g_2}$	$\bar{y}_{g_3}$
y~X1	*cornered*	7	-2	-5	—	—	$b_0 + b_1$	$b_0 + b_2$	b_0
y~X2	*cornered*	2	3	—	5	—	$b_0 + b_1$	b_0	$b_0 + b_3$
y~X3	*centered*	4.67	0.33	-2.67	2.33	b_0	$b_0 + b_1$	$b_0 + b_2$	$b_0 + b_3$
y~X4	*centered*	4.67	0.33	-2.67	2.33	b_0	$b_0 + b_1$	$b_0 + b_2$	$b_0 + b_3$

Notes:

(1) In the cornered coding scheme, the intercept b_0 is equal to the mean of the suppressed group.

(2) For the centered coding scheme, the global mean $\hat{y}_{global}$ is only equal to the intercept term b_0 if all groups have the same number of observations.

(3) For the centered coding scheme, the negative sum of the regression coefficients is equal to the coefficient of the suppressed group.

[d] For each design matrix the global mean $\bar{y}_{global}$ and group means $\bar{y}_{g_1}$, $\bar{y}_{g_2}$ and $\bar{y}_{g_3}$ can also be expressed as a *function of the estimated regression coefficients* b_0, b_{g_1}, b_{g_2} and/or b_{g_3} in the columns *Assign Estimated Regression Coefficients*. Find the expressions for the means in terms of the regression parameters. (1 point)

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

Comment: the centered coding scheme is more intuitive to be interpreted. No matter which reference category is suppressed, the estimated regression coefficients remain the same. In other words, the interpretation of the relationships between dependent variables and the factor levels is invariant of the suppressed reference category. In contrast, the regression coefficients in the cornered coding scheme depend on the suppressed reference category. Moreover, in the centered coding scheme the regression coefficients measure the variation in the mean levels of the categories around the global mean.

Part 2: Model Building and Diagnostics (9 points)

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

Task 3: Multicollinearity diagnostics (5 points)

[a] For the variables **~log(wage)+education+age+experience** generate a scatterplot matrix. (1 point)

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

```
library(car)
scatterplotMatrix(~log(wage)+education+age+experience, data=CPS1985, spread =
FALSE, id.cex=1.5)
```
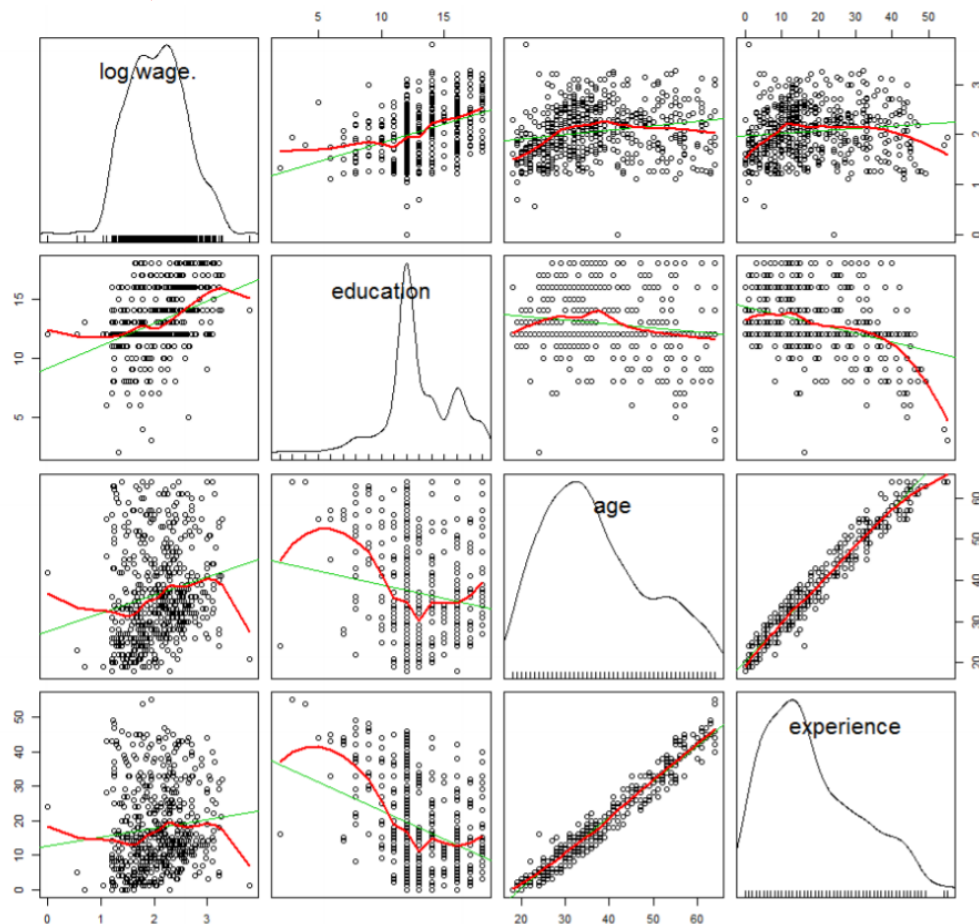

Comment: We expect **age, experience** and **education** to be perfectly multicollinear because the definition of experience is *experience = age − education − 6*. In other words, **experience** is a linear function of **age** and **education**. Therefore, there the variables measure jointly redundant information. Investigating the bivariate plot of **experience-age** displays a strong linear relationship.

[b] Estimate the model **log(wage)~education+experience** and calculate the *variance inflation factors*. Fully interpret the estimated model and the *VIF*s. (1 point)

```
model1 <- lm(log(wage)~education+experience, data=CPS1985)

summary(model1)

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(model1)
education experience
  1.142049   1.142049
```
Comment: Both education and experience are significant and have positive impact on log(wage). Only 21% of the variation within **log(wage)** is explained by **education** and **experience**. The variance inflation factors of **education** and **experience** are smaller than 10. Thus, no multicollinearity exists in this model.

[c] Estimate the augmented model **log(wage)~education+experience+age** and show the output. (3 points)

Address the following points:

   i.   What do the *VIF*s of this model tell you?

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

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

```
model2 <- lm(log(wage)~education+experience+age, data=CPS1985)
summary(model2)
Call:
lm(formula = log(wage) ~ education + experience + age, data = CPS1985)
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(model2)
education experience       age
  229.5738   5147.9190   4611.4008
```

Comment: Due to the substantial degree of collinearity, none of the variables is significant since their standard errors become substantially inflated, but the overall F-statistic remains highly significant. The F-test is significant because jointly the independent variables still influence the dependent variable. The variance inflation factors of **education**, **age**, and **experience** are drastically larger than 10. This example demonstrates that **education**, **age**, and **experience** are highly collinear and at least one of the redundant variables should be dropped.

Task 4: Refined model specification (2 point)

[a] Estimate the model: **log(wage)~education+experience+gender+occupation+union**

and _fully interpret_ the estimated regression model. (1 point)

```
model.full <- lm(log(wage)~education+experience+gender+occupation+union,
data=CPS1985)
summary(model.full)
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 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
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
```

Comment: the regression coefficients of all metric variables are significant. People, who have higher education and experience as well as work in a union job, will have higher wages than other people. However, females seem to be discriminated because on average they earn less money. As the **experience** continues to accumulate over the years wages of workers continue to rise.

The occupation factor indicates that technical and management employees earn more than workers (this is the suppressed reference category) whereas service employees earn less. Office workers and salespersons earn approximately the same as workers, who constitute the reference category, since their coefficients are insignificant.

[b] Test whether the factor **occupation** is significant and if necessary refine the model specification accordingly. (0.5 points)

```
model.test <- lm(log(wage)~education+experience+gender+union, data=CPS1985)
anova(model.test, model.full)
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 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comment: the p-value of the partial F-test is much smaller than 0.05, so the factor **occupation** is significant and should remain in the model. The wages differ by occupation category. It's not necessary to refine the model.
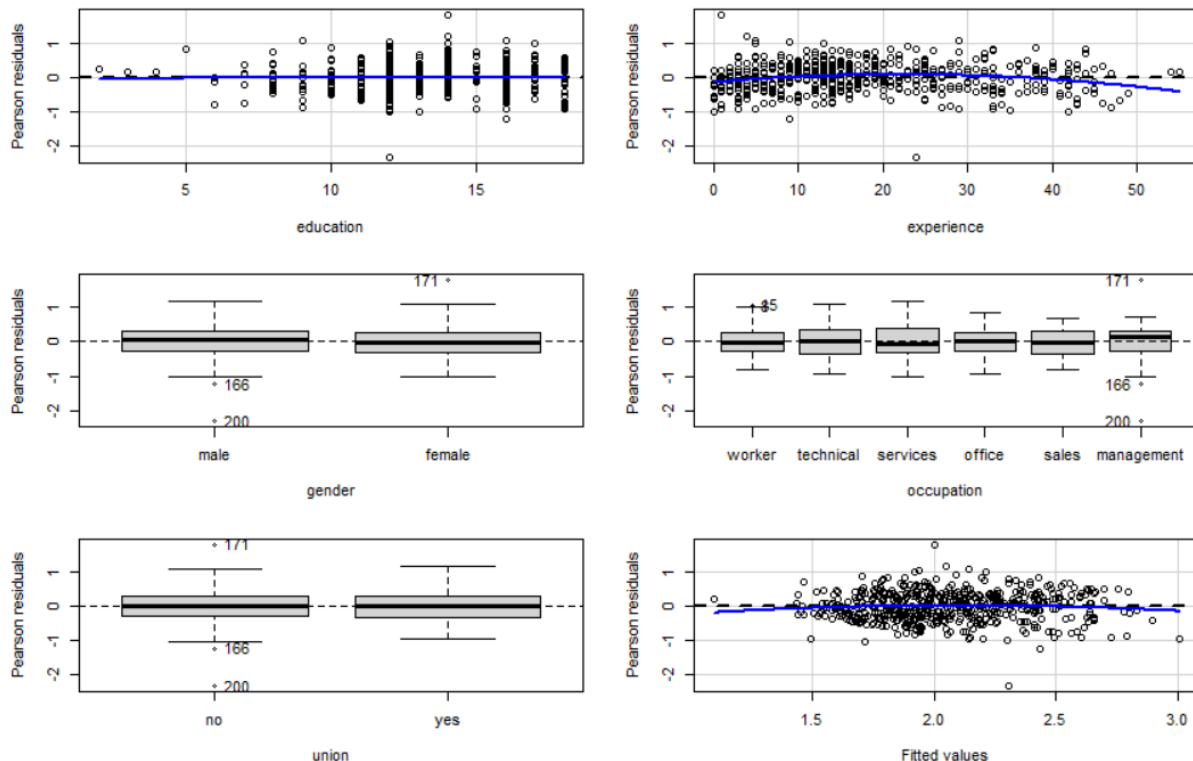
[c] Investigate the model with `car::residualPlots( )`. Discuss the output and decide whether it is advisable to refine the model. (0.5 point)

```
car::residualPlots(model.full, main="Full model")
           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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Full model

Comment: The significant specification test for **experience+experience²** as well as the quadratic lowess line in the residual plot for **experience** clearly indicate the need for a quadratic specification of models as

$$\mathbf{log(wage) \sim education + experience + I(experience^2) + gender + occupation + union}$$

Task 5: Case statistics of the final model (2 points)

[a] Generate the following plots and _interpret_ them for your final model. (1.5 points)

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

```
model.final <-
lm(log(wage)~education+experience+I(experience^2)+gender+occupation+union,
data=CPS1985)
car::qqPlot(model.final, id.n=2)
[1] 171 200
```
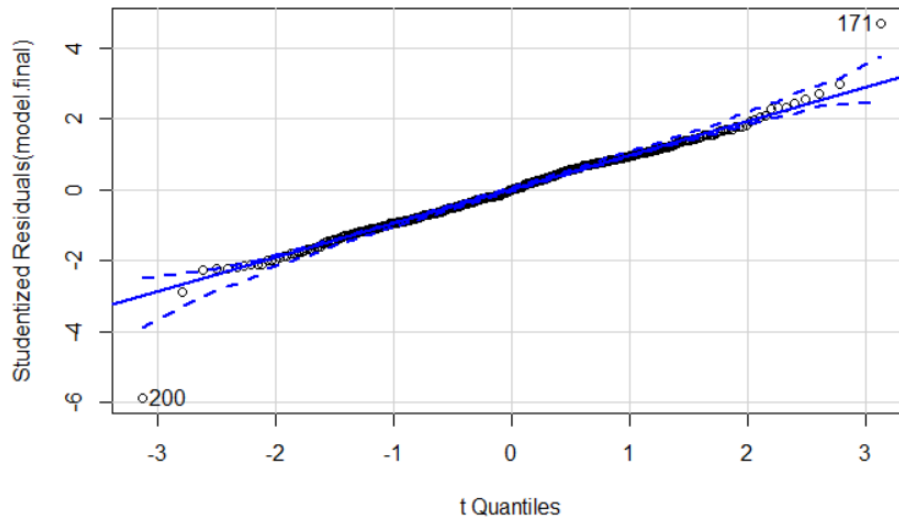

Comment: Most standardized residuals are within the confidence interval around straight line associated with equal quantiles for the observed studentized residuals and their theoretical t-distribution. However, the 171st and 200th observations deviate substantially and are therefore potential outliers.

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

```
car::influenceIndexPlot(model.final, id.n=2)
```


Diagnostic Plots
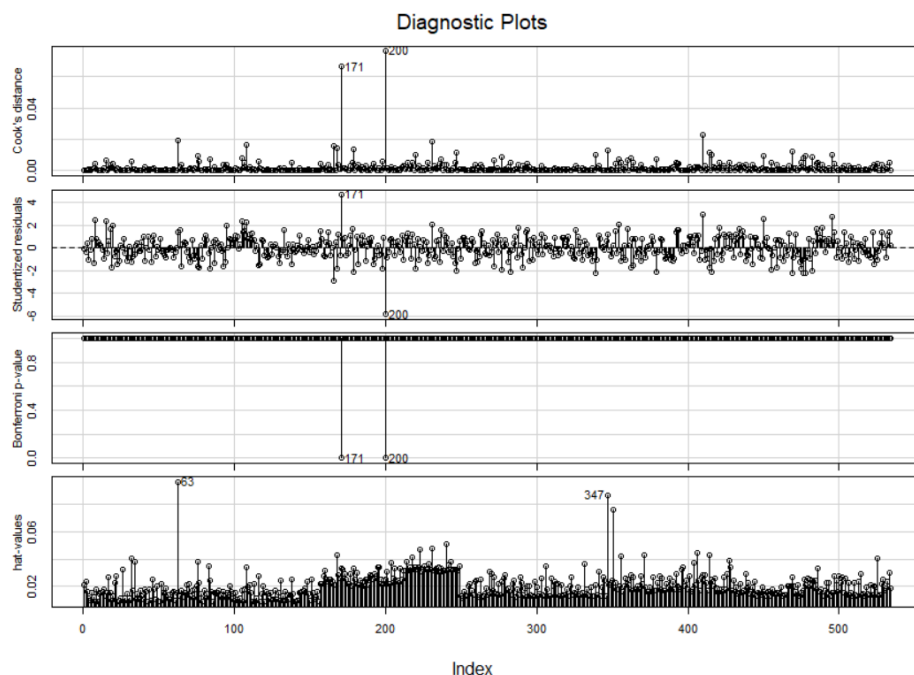
Comments:
Studentized residuals: Two extreme cases 171 and 200 were detected.
Cook's distance: cases 171 and 200
Bonferoni adjusted p-values: Significant outliers for cases 171 and 200
Leverage value: Special combinations of independent variables are prominent for cases 63 and 347

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

```
car::avPlots(model.final, id.n=2)
```
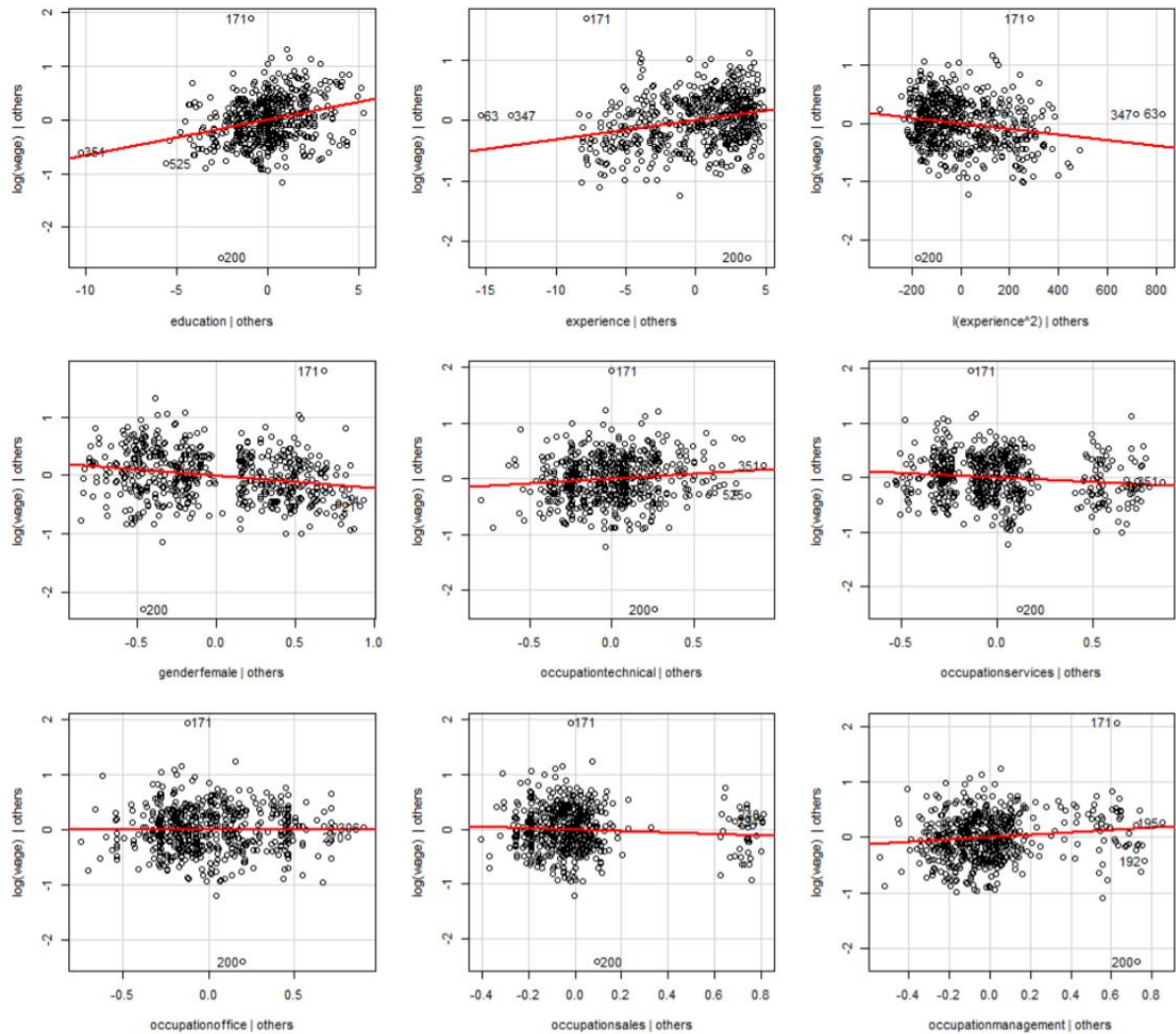

Added-Variable Plots
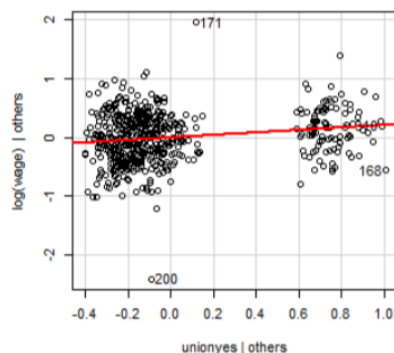
Comment: The 171st and 200th observations are outliers, which affect all added-variable plots most.

[b] Inspect the *two* most extreme observations in the data-frame by examining their records. (0.5 points)

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

```
CPS1985[c(63,171,200,347), ]
```

	WAGE	EDUCATION	EXPERIENCE	AGE	ETHNICITY	REGION	GENDER	OCCUPATION	SECTOR	UNION	MARRIED
63	7	3	55	64	hispanic	south	male	worker	manufacturing	no	yes
171	44.5	14	1	21	cauc	other	female	management	other	no	no
200	1	12	24	42	cauc	other	male	management	other	no	yes
347	6	4	54	64	cauc	other	male	services	other	no	yes

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

```
CPS1985 <- CPS1985[-c(171, 200), ]
```