

Lab03: Matrix Operations & Model Diagnostics

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 $\mathbf{y} = (3, 6, 3, 6, 3, 9)^T$ and the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 3 & 1 & 5 \end{pmatrix}^T$$

[a] Calculate **manually** the vector regression coefficients $\mathbf{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$.

The analytical equation for a 2×2 inverse matrix is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{a \cdot d - b \cdot c} \cdot \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \text{ with the denominator } a \cdot d - b \cdot c \text{ being the determinant.}$$



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

$$(\mathbf{X}^T \cdot \mathbf{X}) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 3 & 1 & 5 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 3 & 1 & 5 \end{pmatrix}^T = \begin{pmatrix} 6 & 14 \\ 14 & 46 \end{pmatrix}$$

$$(\mathbf{X}^T \cdot \mathbf{X})^{-1} = \begin{pmatrix} 6 & 20 \\ 20 & 80 \end{pmatrix}^{-1} = \frac{1}{\underbrace{6 \cdot 46 - 14 \cdot 14}_{1/80}} \cdot \begin{pmatrix} 46 & -14 \\ -14 & 6 \end{pmatrix} = \begin{pmatrix} 0.575 & -0.175 \\ -0.175 & 0.075 \end{pmatrix}$$

$$\begin{aligned} (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T &= \begin{pmatrix} 0.575 & -0.175 \\ -0.175 & 0.075 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 1 & 3 & 1 & 5 \end{pmatrix} \\ &= \begin{pmatrix} 0.4 & 0.05 & 0.4 & 0.05 & 0.4 & -0.3 \\ -0.1 & 0.05 & -0.1 & 0.05 & -0.1 & 0.2 \end{pmatrix} \end{aligned}$$

$$(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} = \begin{pmatrix} 0.4 & 0.05 & 0.4 & 0.05 & 0.4 & -0.3 \\ -0.1 & 0.05 & -0.1 & 0.05 & -0.1 & 0.2 \end{pmatrix} \cdot (3, 6, 3, 6, 3, 9)^T = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$$

[b] 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. 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)

```
> X <- matrix(c(1,1,1,1,1,1,1,3,1,3,1,5), nrow=6)
> y <- matrix(c(3,6,3,6,3,9), nrow=6)
> solve(crossprod(X), crossprod(X,y))
      [,1]
[1,]  1.5
[2,]  1.5
```

[c] Use R's matrix operations to calculate for a dependent variable $\mathbf{y} = (3, 6, 9)^T$, the design matrix $\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 5 \end{pmatrix}^T$ 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(1,1,1,1,3,5), nrow=3)
> y <- matrix(c(3,6,9), nrow=3)
> W <- diag(c(3,2,1))
> solve(t(X)%*%W%*%X, t(X)%*%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 \mathbf{W} . [0.5 points]

There are only three different sets of observations:

- (1) $y_1 = 3$ with $X_1 = (1, 1)^T$ with 3 observations
- (2) $y_2 = 6$ with $X_1 = (1, 3)^T$ with 2 observations
- (3) $y_3 = 9$ with $X_1 = (1, 5)^T$ 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 \mathbf{W} provides information how frequent each unique observation is in the full regression system. The weights matrix in-between $\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X}$ and between $\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{y}$ 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 (3.5 points)

Provide the R 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 \mathbf{y} and the 9×3 design matrices \mathbf{X}_1 to \mathbf{X}_4 separate matrix objects into R and show these object in your answer (0.5 points):

$$\mathbf{y} = \begin{bmatrix} 8 \\ 6 \\ 4 \\ 1 \\ 3 \\ 2 \\ 9 \\ 5 \\ 7 \end{bmatrix} \quad \mathbf{X}_1 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \mathbf{X}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \mathbf{X}_4 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

X_1 and X_2 are given in the **indicator coding** scheme (R codes it as `contrasts(factor) <- "contr.treatment"`) whereas X_3 and X_4 are given in the **centered coding** scheme (R codes it as `contrasts(factor) <- "contr.sum"` and Hamilton p 99 calls it **effect coding**). In X_1 and X_3 the last category is suppressed, whereas in X_2 and X_4 the second category is suppressed.

```
> (Y <- matrix(c(8, 6, 4, 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 group means of the observations $\{y_1, y_2, y_3\}$, $\{y_4, y_5, y_6\}$ and $\{y_7, y_8, y_9\}$ as well as the global mean for all observations $\{y_1, \dots, y_9\}$. (0.5 points)

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

[c] Find the four sets of estimated regression parameters by regressing y on the four design matrices X_1 , X_2 , X_3 and X_4 with your linear regression function from task 1 [b] 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 parameters, i.e., $b_{g_1} = -(b_{g_2} + b_{g_3})$. (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}
$y \sim X_1$	cornered	7	-1	-5	—	—	$b_0 + b_1$	$b_0 + b_2$	b_0
$y \sim X_2$	cornered	2	4	—	5	—	$b_0 + b_1$	b_0	$b_0 + b_3$
$y \sim X_3$	centered	5	1	-3	2	b_0	$b_0 + b_1$	$b_0 + b_2$	$b_0 + b_3$
$y \sim X_4$	centered	5	1	-3	2	b_0	$b_0 + b_1$	$b_0 + b_2$	$b_0 + b_3$

Notes:

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

- For the centered coding scheme, the global mean \bar{y}_{global} is only equal to the intercept term b_0 if all groups have the same number of observations.
- 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 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...”.

[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.

[f] Argue, based on the four different coding schemes which, however, give identical predictions \hat{y} , 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)

Comment: All four model specifications lead to the same predicted values for the group means. The size, and therefore the deviation from zero, of a parameter estimate for a particular factor level depends on the coding scheme and the suppressed reference category. Therefore, depending on its size, it may either be either significantly different from zero or not. Due to this arbitrariness in the specification an individual t -test becomes meaningless. For dummy variables, it's appropriate to test the significance of all coefficients simultaneously with a partial F -test, which would tell us that the mean levels of the categories differ significantly from the global mean.

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 the data-frame. **Study the description** of the variable **experience** in the associated online help.

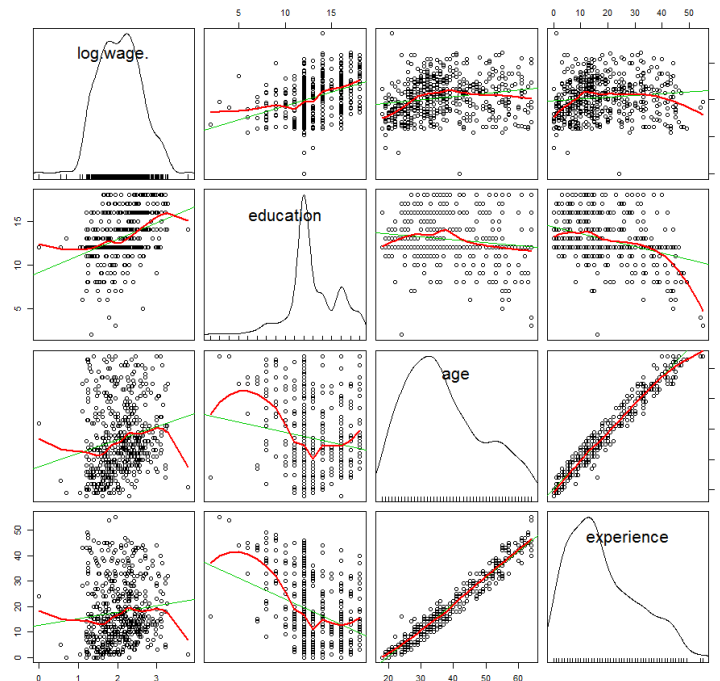
Task 3: Multicollinearity diagnostics (2 points)

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

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

```
> library(AER); library(car)
> data(CPS1985)
> rownames(CPS1985) <- 1:nrow(CPS1985)
```

```
> scatterplotMatrix(~log(wage)+education+age+experience, data=CPS1985, spread=F, 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 relationships.

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

```
> 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 $\log(\text{wage})$. Only 21% of the variation within $\log(\text{wage})$ 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 $\log(\text{wage}) \sim \text{education} + \text{experience} + \text{age}$ and show the output. (1 point)

Address the following points:

- What do the **VIF** tell you?
- What** happened to the significances of the t -tests for the estimated regression parameters of the augmented model and **why**?
- 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 (1 point)

[a] Estimate the model: $\log(\text{wage}) \sim \text{education} + \text{experience} + \text{gender} + \text{occupation} + \text{union}$ and fully interpret the estimated regression model. (0.5 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 sales persons 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. (0.25 points)

```

> model.test <- lm(log(wage)~education+experience+gender+union, data=CPS1985)
> anova(model.test, model.full)
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 ***

```

```

---
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 if advisable refine the model. (0.25 point)

```

> car::residualPlots(model.full, main="Full model")
              Test stat Pr(>|t|)
education     -0.337    0.737
experience   -4.212    0.000
gender         NA      NA

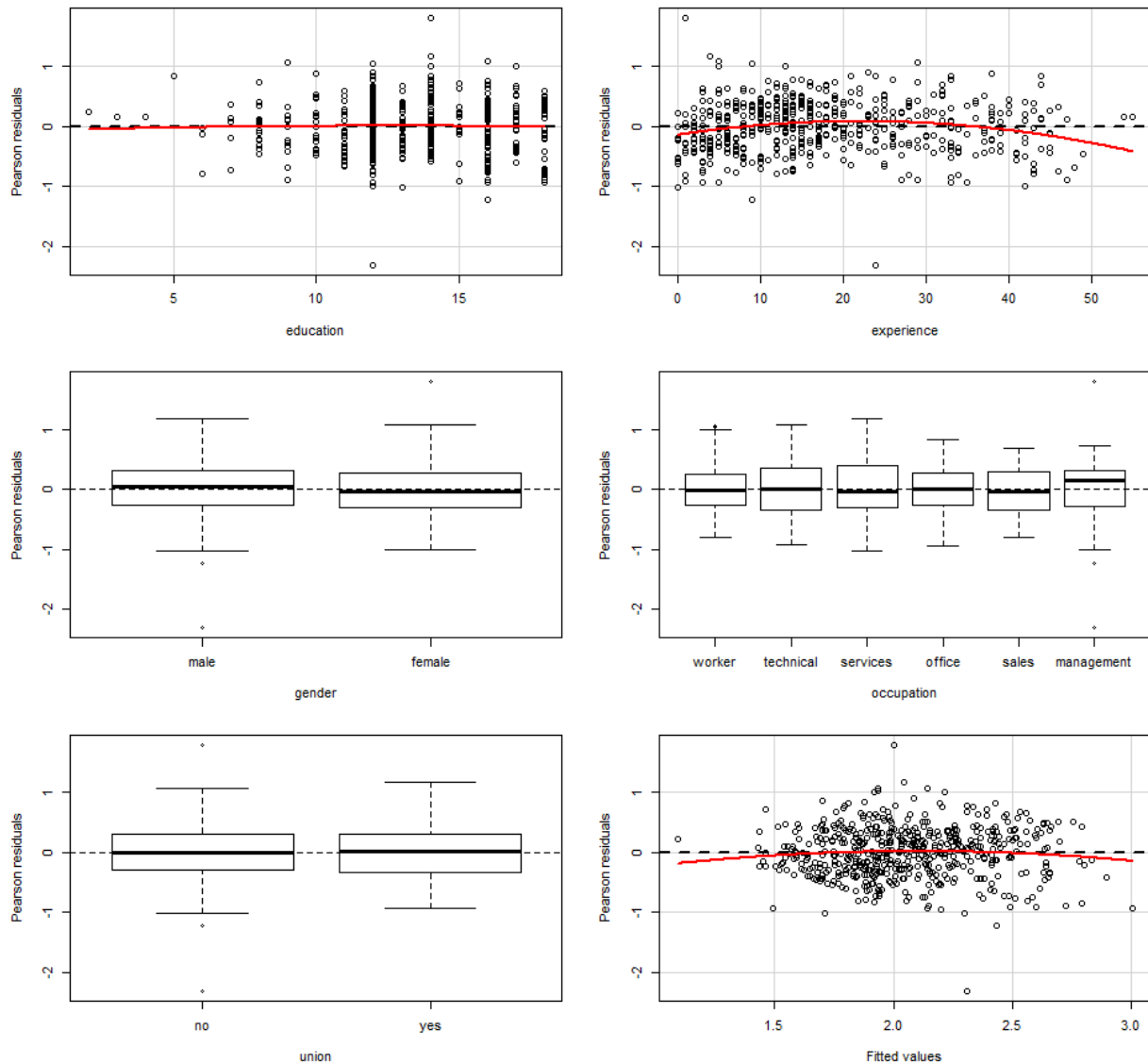
```

```

occupation      NA      NA
union           NA      NA
Tukey test     -1.375   0.169

```

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

`log(wage) ~ education + experience + I(experience^2) + gender + occupation + union`

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

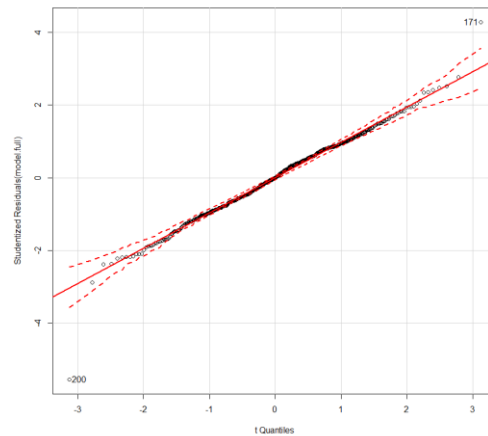
[a] Generate the following plots and *interpret* them for your final model. (0.75 points)

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)
200 171
1 534

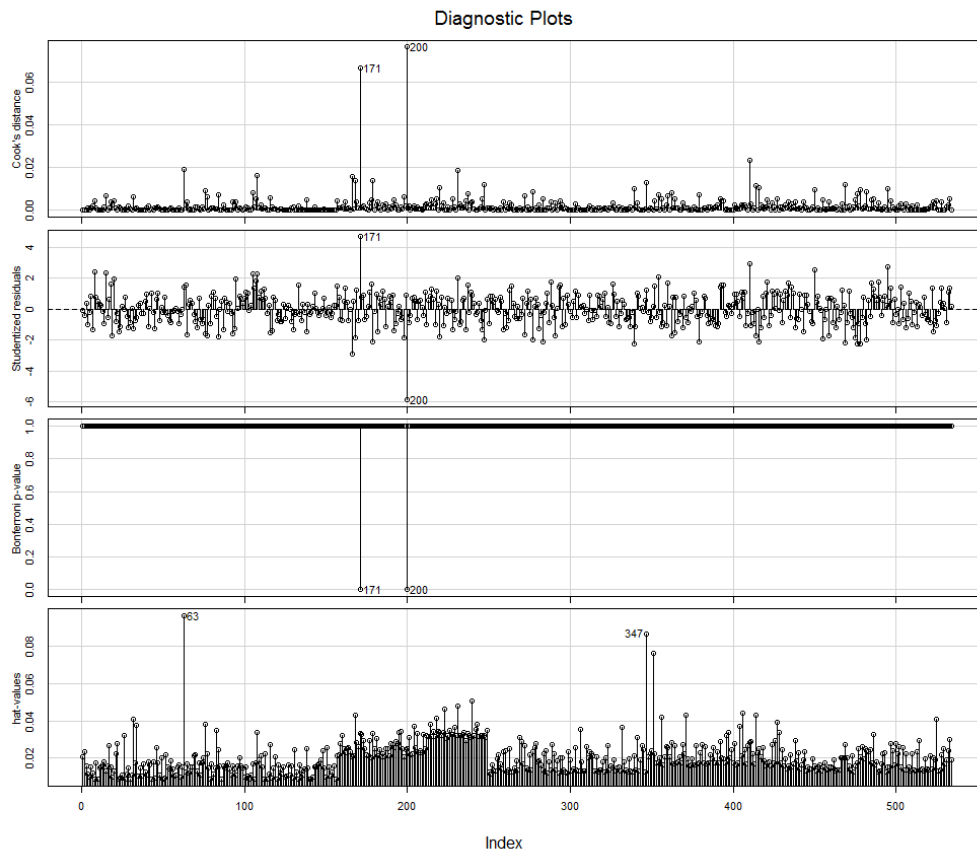
```

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 171th 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)
```



Comments:

Studentized residuals: Two extreme cases 171 and 200 were detected

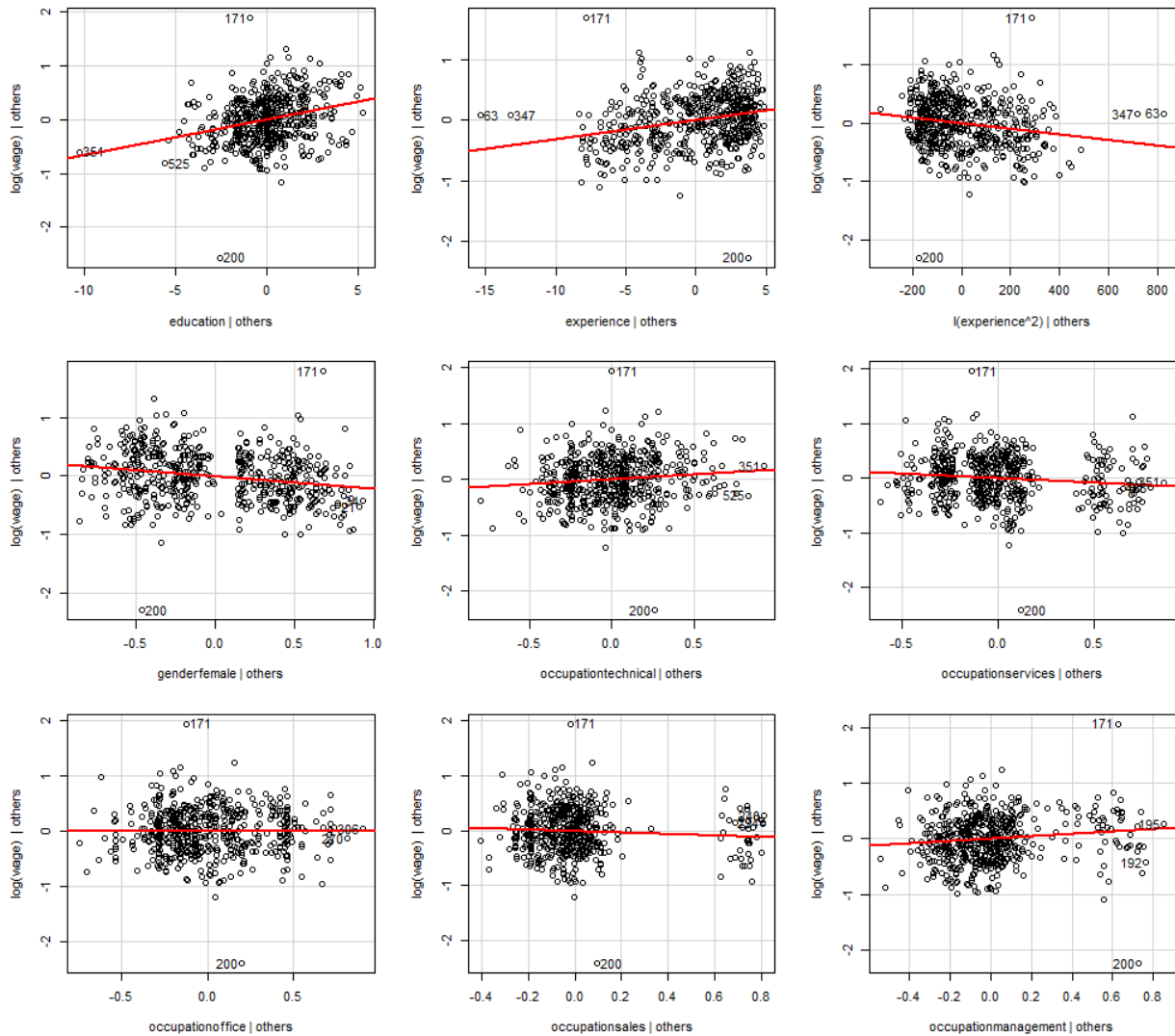
Cook's distance: cases 171 and 200

Bonferoni adjusted p-values: Significant outliers for cases 171 and 200

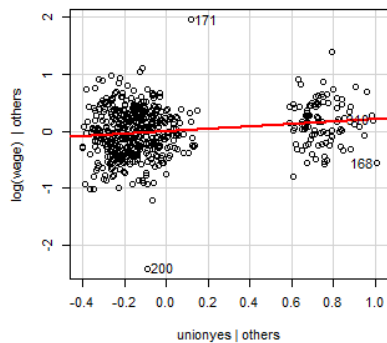
Leverage value: Special combinations of independent variables are prominent for cases 63 and 347

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 171th 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. (0.25 points)

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

Comment: Inspection of data for intensified special case:

```
> CPS1985[id, ]
      wage education experience age ethnicity region gender
63      7.0         3          55 64  hispanic  south  male
171  44.5        14           1  21    cauc   other female
200   1.0        12          24  42    cauc   other  male
347   6.0         4          54  64    cauc   other  male

      occupation      sector union married
63      worker manufacturing    no     yes
171 management          other    no     no
200 management          other    no     yes
347  services            other    no     yes
```

For both cases 171 and 200 the wages were at the extreme end of the hourly wage scale. The high wage in case 171 *does not make sense* for a 21 years old person with 1 year of experience. The low wage for a manager in case 200 does not make sense either. It looks as if these extreme wages are due to *a typo* during data entry.

If the errors in these records cannot be fixed it is highly suggested to drop case 171 and case 200.

The two leverage cases 63 and 347 both exhibit an extremely low education of less than 5 years. Both cases are associated with low-skilled workers. The question ultimately is whether these cases are representative for the underlying study population.

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

```
> CPS1985 <- 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 a correlation between an endogenous regressor and the disturbances lead to an associated regression coefficient estimate, which must be biased. See also Figure 9.1. (0.75 points)

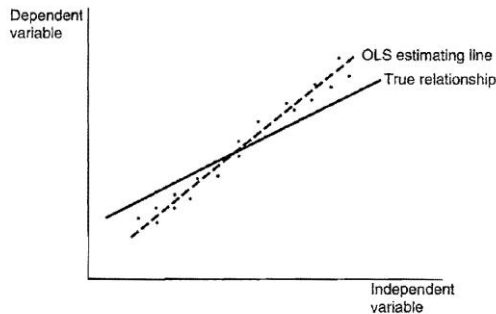


Figure 9.1 Positive contemporaneous correlation.

The dependent variable can be expressed by its two components: The expected value on the true regression line and the disturbance. Thus $Y_i = E[Y_i] + \varepsilon_i$. Without loss of generality let us assume that endogenous regressor x is positively correlated with the disturbance, thus, $Corr(x, \varepsilon) > 0$. For positive deviations from the true regression line, i.e., $\varepsilon_i > 0$, the endogenous regressor will be on average larger and vice versa, for negative deviations. Since the endogenous regressor follows

the deviations of ε_i , the fit will be unrealistically better. This leads to biased estimates of the regression parameters. See Figure 9.1 on the left.

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)

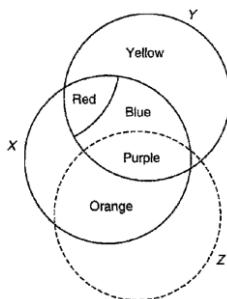


Figure 9.2 Using an instrumental variable Z.

See Figure 9.2. The “red” area denotes the correlation between the disturbance of Y and endogenous regressor X . The objective is to not allow this correlated area to become part of the estimation. This can be achieved by just using that intersection between Y and X (blue and purple), which is free of the “red” area. The instrumental variable Z achieves this by covering this subset (purple area) of X that intersects with Y . Since the purple area explains less variation in Y than the blue and purple area the standard errors will increase. Note, the instrumental variable Z should not provide additional explanation of Y . In the Ballentine’s plot there is, however, a small intersection which is unique to Y and Z .

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:
<code>lwage</code>	Dependent variable: logarithm of wage
<code>educ</code>	Endogenous regressor: education in years
<code>age</code>	Exogenous regressor: age in years
<code>nearc2</code>	First instrument: dummy variable indicating that person lived near a 2 years college
<code>nearc4</code>	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)

```
> summary(ivreg(lwage~educ+age|nearc2+nearc4+age, data=card), diagnostics=T)
```

Call:

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

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 .

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

Residual standard error: 0.5327 on 3007 degrees of freedom

Multiple R-Squared: -0.4398, Adjusted R-squared: -0.4408

Wald test: 112.5 on 2 and 3007 DF, p-value: < 2.2e-16

The **weak instruments test** indicates that the instruments are sufficiently strong in explaining the endogenous regressor education. Thus, they are good instruments for the education variable.

The **Wu-Hausman test** indicates that the regression coefficients for the structural equation at the second stage differ significantly from those obtained by standard ordinary least squares. Thus the ordinary least squares estimates were biased and the regressor education is correlated with the disturbances.

We fail to reject the **Sargan test** at a 5% error probability and thus the instruments are exogenous with regards to the dependent wage variable.