# Compulsory exercise 1: Group 13

TMA4315: Generalized linear (mixed) models H2023

Benjamin Sigbjørnsen, Johan Vik Mathisen, Penelope Malaspina and Martinius Theodor Singdahlsen

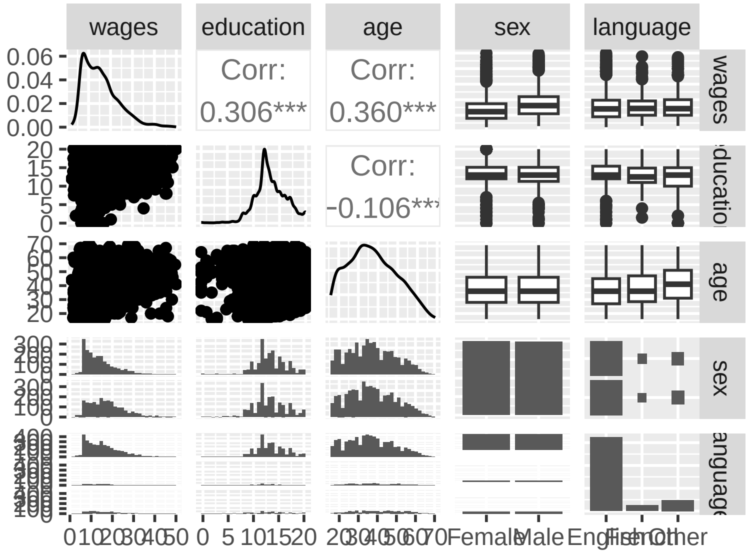
21 September, 2023

Packages:

```
library(car)
library(ggplot2)
library(GGally)
library(mylm)
```

## Part 1)

```
data(SLID, package = "carData")
SLID <- SLID[complete.cases(SLID), ]
ggpairs(SLID)
```



We can see that there are significant correlations between some of the variables. We have three continuous variables and two categorical variables.

In multiple linear regression analysis, we assume that the residuals of the regression are iid Gaussian. We also assume that the relationship between the response and the covariates are linear.

## Part 2)

## a)

The coefficient estimates are found by finding $\hat{\beta}$, the MLE of $\beta$.

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Print of the linear model made with `mylm` follows:

```
model1 <- mylm(wages ~ education, data = SLID)
print.mylm(model1)
```

```
## Info about object
##
## Coefficients
##                [,1]
## (Intercept) 4.972
## education   0.792
```

We can see that the print of a linear model made with the function `lm` gives us the same coefficients.

```
model1b <- lm(wages ~ education, data = SLID)
print(model1b)
```

```
##
## Call:
## lm(formula = wages ~ education, data = SLID)
##
## Coefficients:
## (Intercept)    education
##      4.9717       0.7923
```

## b)

The estimated covariance matrix of the parameter estimates are found by calculating

$$\hat{\sigma}^2 (X^T X)^{-1},$$

where

$$\hat{\sigma}^2 = \frac{SSE}{n-p}.$$

This is the restricted MLE of $\sigma^2$.

We can in the summary outputs below find estimates and standard errors of the intercept and regression coefficient for this model. The estimates are: $\beta_{intercept} = 4.97$ and $\beta_{education} = 0.792$. We also have $SE(\beta_{intercept}) = 0.534$ and $SE(\beta_{education}) = 0.391$.

The standard errors are found by looking at the square root of entries on the diagonal of the covariance matrix. The z-values from the tests $H_0 : \beta = 0$ and $H_1 : \beta \neq 0$ are found from the formula

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

where $\hat{\beta}$ is the parameter estimate. The p-values are the probabilities of obtaining the respective z-values given that $H_0$ is correct.

The intercept describes expected wage with no education, while the education parameter describes how expected wage changes with one more year of education. We can see that both parameters are significant.

The summary of the linear model made with the `mylm` function follows.

```
summary.mylm(model1)
```

```
## Summary of object
##
##              Estimate Std. Error  z-value      p-value
## (Intercept)  4.97169    0.53423  9.30633 1.323215e-20
## education    0.79231    0.03906 20.28670 1.685233e-91
##   F-statistic p-value (chi square)
## 1    411.3438                    0
## R-squared:[1] 0.09358627
```

The table found in the summary output below, where the model is fitted with ´lm´ model, is the same.

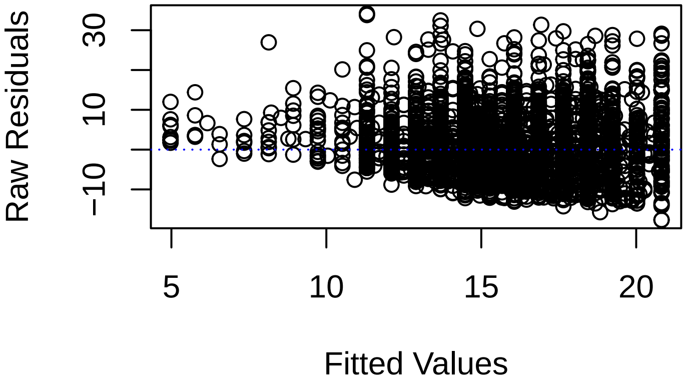
```
summary(model1b)
```

```
##
## Call:
## lm(formula = wages ~ education, data = SLID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.688  -5.822  -1.039   4.148  34.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.97169    0.53429   9.305   <2e-16 ***
## education    0.79231    0.03906  20.284   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.492 on 3985 degrees of freedom
## Multiple R-squared:  0.09359,    Adjusted R-squared:  0.09336
## F-statistic: 411.4 on 1 and 3985 DF,  p-value: < 2.2e-16
```

### c)

We implement a plot function with the fitted values on the $x$-axis and the residuals on the $y$-axis.

```
plot.mylm(model1)
```



From the plot we can see that for the lower values the residual are mostly positive, while that changes the higher the values get. As we can see from the plot, which shows a cone-shaped pattern, we can deduce that the variance seems to increase with higher fitted values. This means that one of the key assumptions of linear regression, homoscedasticity (which means that the residuals have constant variance), is violated. This is called heteroscedasticity and it means that our model assumptions are not entirely reasonable and we should be careful when conducting inference.

## d)

After a scaling, the $\chi^2$-distribution is the limiting distribution of an $F$-distribution as the denominator degrees of freedom goes to infinity. The normalization is $\chi^2$=(numerator degrees of freedom)$\cdot F$.

The residual sum of squares (SSE) is calculated by

$$SSE = \sum_{i=1}^{n} \epsilon_i^2,$$

and in our case is equal to

```
## SSE: 223694.3
```

And the degrees of freedom are the number of observations minus the number of fitted parameters which in our case are equal to

```
## DF: 3984
```

The sum of squares (SST) for this model is calculated by

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

In our case it is equal to

```
## SST: 246790.5
```

After defining a significance level $\alpha = 0.05$ we test the significance using a $\chi^2$-test.

```
## Chi-square test statistic:  411.3438
```

```
## p-val:  0
```

Because the p-value is below any reasonable significance level, we say that the regression is significant.

We know that the $\chi^2$-statistic has $p$-1 degrees of freedom and in simple linear regression ($p$=2) this means that $\chi^2$-statistic has 1 degree of freedom and seeing as the $z$-statistic is standard normal distributed, its square $z^2$ will be $\chi^2$ distributed with 1 degree or freedom.

This means that the relationship between the $\chi^2$-statistic and $z$-statistic is that the square of the $z$-statistic is equal to the $\chi^2$ statistic so they reject at the same critical value the null hypothesis.

```
## z-statistic: 411.5503
```

```
## Chi^2-statistic: 411.3438
```

We confirm that the critical values coincide by listing them for quantiles.

```
interval = c(0.7, 0.8, 0.9, 0.95)
interval2 = c(0.4, 0.6, 0.8, 0.9)
cat(abs(qnorm(interval)), "\n")
```

```
## 0.5244005 0.8416212 1.281552 1.644854
```

```
cat(sqrt(qchisq(interval2, 1)))
```

```
## 0.5244005 0.8416212 1.281552 1.644854
```

### e)

The coefficient of determination,$R^2$, is used to assess the goodness of fit of a regression model. It is calculated:

$$R^2 = 1 - \frac{SSE}{SST}$$

Our $R^2$ is

```
## 0.09358627
```

A low coefficient of determination such as ours indicates that the regression model does not explain much of the variance in the dependent variable.

## Part 3)

### a, b)

Here we have decided to merge (a) and (b) together as there was nothing that had to be changed to get the function `mylm()` to work for multiple linear regression. In this part of the problem set we will show that the linear regression works for multiple predictors and analyse the result of the fit. We will use the predictors `education` and `age` to regress on the response `wage`. Here, all the variables are treated as continuous variables.

```
model2 <- mylm(wages ~ education + age, data = SLID)
summary.mylm(model2)
```

```
## Summary of object
##
##              Estimate Std. Error  z-value       p-value
## (Intercept) -6.02165    0.61885 -9.73045  2.235950e-22
## education    0.90146    0.03576 25.21202 2.957148e-140
## age          0.25709    0.00895 28.72476 1.872433e-181
##   F-statistic p-value (chi square)
## 1    660.5437                    0
## R-squared:[1] 0.2490697
```

Applying our function `mylm()` gives the model,

$$y_{\text{wage}} = -6.022 + 0.901 * x_{\text{education}} + 0.257 * x_{\text{age}}. \tag{1}$$

The estimated coefficients and their corresponding standard errors and p-values from the z-test can be found in the printout from the function `summary.mylm(model2)` above.

However to be clear, the estimated coefficient for education, $\beta_{education}$, is 0.901, the estimated coefficient for age, $\beta_{age}$, is 0.257 and the intercept, $\beta_0$, is $-6.022$. The standard errors of the estimated coefficients, $\beta_0, \beta_{education}, \beta_{age}$, are $0.619, 0.036, 0.009$. Preforming the z-test to test the significance of the parameters, $\beta_0, \beta_{education}, \beta_{age}$, individually we get the p-values $2.236 * 10^{-22}, 2.957 * 10^{-140}, 1.872 * 10^{-181}$.

We can see that both parameters and the intercept contribute significantly to the prediction of the response by looking at the p-values of the z-test. The p-values are all smaller than $10^{-20}$, which is quite significant and smaller than 0.05.

Looking at the interpretation of the parameters, we first need to explain what the predictors and response is. The response, wage is the hourly wage rate from all jobs an individual may have. The predictor education is a continuous variable of how many years an individual has been in school. Thus for each year a person has been in school, this predicts the hourly wage rate of an individual to be 0.901 more. The predictor age is the age of an individual. From the coefficient $\beta_{age}$ we see that the hourly wage of an individual is predicted to increase by 0.257 per year lived. Every individual shares the intercept which is negative. It does not make sense to earn a negative amount. However, when we consider that the predictor age ranges from 16 to 69 and education ranges from 0 to 20, we can see that the predicted value in this range is positive.

### c)

We would like to investigate when and why the parameters differ when we fit nested models of the model fitted in task (a) and (b). We will fit two models where one has the predictor age and the other has the predictor education. Both models has the response wage.

```
# model wages~education
summary.mylm(model1)
```

```
## Summary of object
##
##              Estimate Std. Error  z-value      p-value
## (Intercept)  4.97169    0.53423  9.30633 1.323215e-20
## education    0.79231    0.03906 20.28670 1.685233e-91
##   F-statistic p-value (chi square)
## 1    411.3438                    0
## R-squared:[1] 0.09358627
```

```
# model wages~age
model3 <- mylm(wages ~ age, data = SLID)
summary.mylm(model3)
```

```
## Summary of object
##
##              Estimate Std. Error  z-value       p-value
## (Intercept)  6.89090    0.37400 18.42486  8.299792e-76
## age          0.23311    0.00958 24.32834 9.829575e-131
##   F-statistic p-value (chi square)
## 1    591.5713                    0
## R-squared:[1] 0.1292891
```

This yields the simple linear regression models

$$
\begin{aligned}
y_{\text{wage}} &= 4.972 + 0.792 * x_{\text{education}}, \qquad (2)\\
y_{\text{wage}} &= 6.891 + 0.233 * x_{\text{age}}. \qquad (3)
\end{aligned}
$$

We can see that none of the coefficients in the three different models are equal to each other. However looking at the standard error of the coefficient age in both models age is included, it is approximately 0.01 in both models. The predictors have a difference of 0.0238, so they are relatively similar, but more than two standard deviations apart. This is not the case for the predictor education as the coefficients are quite different.

If we look at the equation for how we obtain the estimates for the parameters,

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \tag{4}$$

we can observe that we get a different answer depending on what $X$ and $Y$ is. Since we always predict wage, $Y$ has stayed constant. However, the predictors we have used has varied. Thus we have had different data matrices, $X$, which has produced different results for our parameters.

However different data matrices can produce the same estimated coefficients. Assume we fit a linear model with the predictor $a$, and an additional nested model with the predictors $a$ and $b$. Say that the data for the simple linear regression model of predictor $a$ is, $X_a : (n \times 1)$, and the data for the model fitted with predictors $a$ and $b$ is, $X_f : (n \times 2)$. Here both data try to predict the same response $y$ and they are nested. Further more $X_a$ is the first column of $X_f$. If the data is orthogonal to each other with unit length, we will get the following estimates for $\beta_a$ corresponding to the model with predictor $a$,

$$\begin{aligned}\hat{\beta}_a &= (X_a^T X_a)^{-1} X_a^T Y &(5)\\ &= (I)^{-1} X_a^T Y &(6)\\ &= X_a^T Y. &(7)\end{aligned}$$

The estimate for $\hat{\beta}_f = (\hat{\beta}_1, \hat{\beta}_2)$ will be,

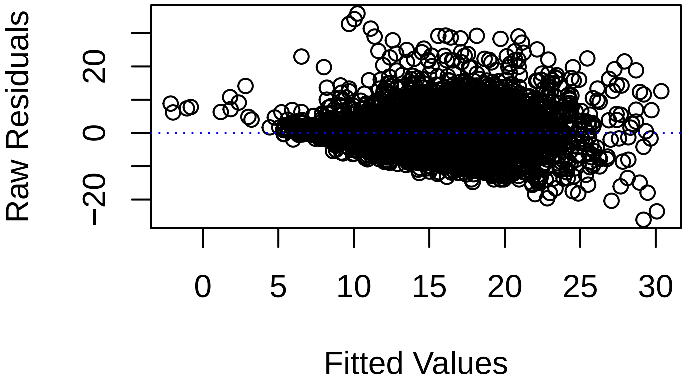
$$\begin{aligned}\hat{\beta}_f &= (X_f^T X_f)^{-1} X_f^T Y &(8)\\ &= (I)^{-1} X_f^T Y &(9)\\ &= X_f^T Y. &(10)\end{aligned}$$

This gives $\hat{\beta}_1 = \hat{\beta}_a$ since the first column of $X_f$ is $X_a$. This can appear if the predictor $b$ is a confounder of $a$ and the response, meaning that $b$ causes $a$ and the response. However this does not appear to be happening here.

## Part 4

We begin by fitting the first model as described in the task.

```r
model4 <- mylm(wages ~ sex + age + language + I(education)^2, data = SLID)
plot.mylm(model4)
```



```r
summary.mylm(model4)
```

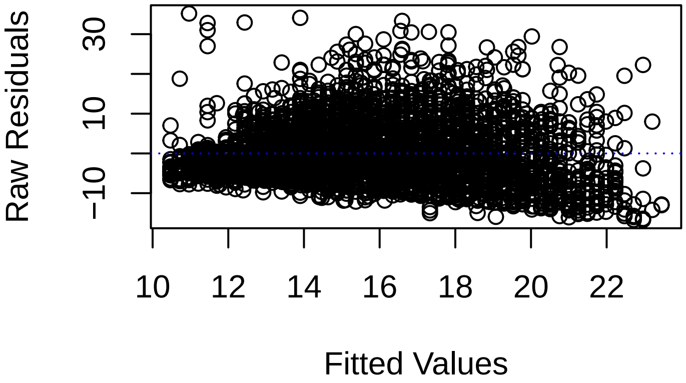
```
## Summary of object
##
##                  Estimate Std. Error   z-value       p-value
## (Intercept)      -7.88878    0.61219 -12.88624  5.380163e-38
## sexMale           3.45541    0.20917  16.51975  2.644739e-61
## age               0.25514    0.00871  29.28215 1.750691e-188
## languageFrench   -0.01522    0.42668  -0.03568  9.715386e-01
## languageOther     0.14260    0.32502   0.43876  6.608349e-01
## I(education)      0.91661    0.03476  26.37157 2.903861e-153
##   F-statistic p-value (chi square)
## 1    336.7157                    0
## R-squared:[1] 0.2972641
```

From the p-values of the regression we observe that `age`, `sex` and `education^2` are highly significant and that `language` is not.

From the regression coefficients we see that the higher wage is associated with longer education, older age and being born male.

We fit the second model as described in the task.

```
model5 <- mylm(wages ~ age + language + age * language, data = SLID)
plot.mylm(model5)
```



```
summary.mylm(model5)
```

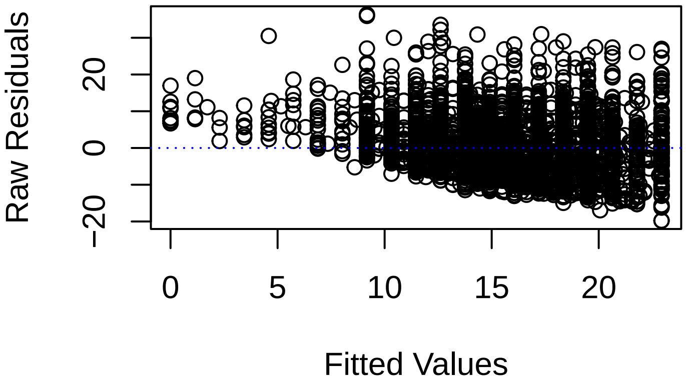
```
## Summary of object
##
##                       Estimate Std. Error  z-value       p-value
## (Intercept)            6.55579    0.41063 15.96525  2.231621e-57
## age                    0.24485    0.01069 22.91293 3.453351e-116
## languageFrench         2.86063    1.59587  1.79252  7.305035e-02
## languageOther          0.84862    1.23503  0.68713  4.920022e-01
## age:languageFrench    -0.08393    0.04045 -2.07480  3.800531e-02
## age:languageOther     -0.03701    0.02934 -1.26170  2.070579e-01
##   F-statistic p-value (chi square)
## 1    120.1751                    0
## R-squared:[1] 0.1311705
```

From the p-values of the regression we observe that `age` and the interaction term `age*French` are the significant features. Again `age` is highly significant, but the interaction term is moderately significant. As the feature `French` is not significant, it is reason to believe that the interaction term is significant as a result of `age` being so.

By removing the interaction term and `language` the model would probably improve.

We fit the third model as described in the task.

```
model6 <- mylm(wages ~ education - 1, data = SLID)
plot.mylm(model6)
```



```
summary.mylm(model6)
```

```
## Summary of object
##
##             Estimate Std. Error  z-value p-value
## education   1.1467    0.00877 130.8104       0
##   F-statistic p-value (chi square)
## 1         Inf                  NaN
## R-squared:[1] 0.07389171
```

From the p-value and z-value of the regression we observe that `education` is highly significant.

By including the intercept the model would be better, as the model seems to underestimate the wages for individuals with lower education levels, as seen by the residual vs fitted plot.

From the scatterplot in task 1, we see that the wage distribution has a long tail. Therefore a transformation of the response could be beneficial. From all three models in part 4 we see from the residual vs fitted plot that the residuals are heteroscedastic. One possible remedy to this issue is to exclude samples where the individual has worked for less than $x$ number of years, as there seems to be less variability in the starting salaries, and that the variability increases before it evens out after a number of years.

## mylm()

```
mylm <- function(formula, data = list(), contrasts = NULL, ...) {
    # Extract model matrix & responses
    mf <- model.frame(formula = formula, data = data)
    X <- model.matrix(attr(mf, "terms"), data = mf, contrasts.arg = contrasts)
    y <- model.response(mf)
    terms <- attr(mf, "terms")

    # a) Calculating regression coefficients
    Bhat = solve(t(X) %*% X) %*% t(X) %*% y

    # b)

    # Calculate fitted values
    fitted_values = X %*% Bhat
    # Calculate the residuals
    residuals = y - X %*% Bhat
    # Estimate for sigma squared
    sigma_Squared = as.numeric(1/(length(y) - ncol(X) + 1) * t(residuals) %*% residuals)
```

```
# Calculating the covariance matrix of coefficients
cov_Bhat = sigma_Squared * solve(t(X) %*% X)
# Standard errors of the regression coefficients
SE_Bhat = as.numeric(sqrt(diag(cov_Bhat)))

# Finding z-values and p-values
z_Bhat = rep(0, length(Bhat))
p_Value = rep(0, length(Bhat))
for (i in 1:length(Bhat)) {
    z_Bhat[i] = Bhat[i]/SE_Bhat[i]
    p_Value[i] = 2 * pnorm(abs(z_Bhat[i]), lower.tail = FALSE)
}

# d) Calculating the SSE and the degrees of freedom
SSE <- sum((t(residuals) %*% residuals))
# Number of responses
n <- length(y)
# Number of predictors including intercept
p <- length(Bhat)
# Degrees of freedom
df <- (n - (p + 1))

# Calculating SST
ressidual_H0 <- y - mean(y)
SST <- sum(t(ressidual_H0) %*% ressidual_H0)

# Testing the significance of the regression with a chi^2-test
F_statistic <- ((df) * (SST - SSE))/(SSE * (p - 1))

F_pvalue <- 1 - pchisq(F_statistic * (p - 1), df = (p - 1))

# Find the critical values for both tests assuming the significance level
# of 0.05
critical_z <- qnorm(1 - (0.05/2))
critical_chi <- qchisq(1 - 0.05, df)

# Calculate R^2
R_suared <- 1 - SSE/SST

# Store the results in the list est
est <- list(terms = terms, model = mf, coefficients = Bhat, fitted_values = fitted_values,
    residuals = residuals, SE = SE_Bhat, z = z_Bhat, p = p_Value, SSE = SSE,
    SST = SST, critical_z = critical_z, critical_chi = critical_chi, F_statistic = F_statistic,
    F_pvalue = F_pvalue, R_suared = R_suared, df = df)

# Store call and formula used
est$call <- match.call()
est$formula <- formula

# Set class name. This is very important!
class(est) <- "mylm"

# Return the object with all results
```

```r
    return(est)
}

print.mylm <- function(object, ...) {
    # Code here is used when print(object) is used on objects of class 'mylm'
    cat("Info about object\n\n")
    cat("Coefficients\n")
    print(round(object$coefficients, 3))
}

summary.mylm <- function(object, ...) {
    # Code here is used when summary(object) is used on objects of class 'mylm'
    cat("Summary of object\n\n")

    # Creating summary table
    summary_Table = data.frame(matrix(c(round(object$coefficients, 5), round(object$SE,
        5), round(object$z, 5), object$p), nrow = length(object$coefficients)))
    colnames(summary_Table) <- c("Estimate", "Std. Error", "z-value", "p-value")
    rownames(summary_Table) <- rownames(object$coefficients)

    summary_F = data.frame(matrix(c(object$F_statistic, object$F_pvalue), nrow = 1,
        byrow = TRUE))
    colnames(summary_F) <- c("F-statistic", "p-value (chi square)")

    print(summary_Table)
    print(summary_F)

    cat("R-squared:")
    print(object$R_suared)


}

plot.mylm <- function(object, ...) {
    # Code here is used when plot(object) is used on objects of class 'mylm'

    # c) Creating a plot function
    plot(object$fitted_values, object$residuals, xlab = "Fitted Values", ylab = "Raw Residuals")
    abline(a = 0, b = 0, col = "blue", lty = 3)
}
```