Course: Multivariate statistics (AUT23)

Chapter 3: Linear regression

3.18 Time to practice on your own

The exercises 1 and 2 will use the data from the round 10 of the European Social Survey (ESS). You can download the data directly on the ESS website.

The objective is to conduct linear regression to explain the consumption of news about politics and current affairs ('nwspol': watching, reading or listening in minutes). Explanatory variables include a set of political and sociodemographic variables. Political variables can include: political interest ('polintr'), confidence in ability to participate in politics ('cptppola'), and self-placement on the left-right political scale ('lrscale'). Sociodemographic variables can include: gender ('gndr'), age ('agea'), and years of education ('eduyrs'). You can also decide to rely on additional or alternative variables.

Let's start by loading the dataset, selecting the relevant variables, and filtering the data for Switzerland only:

- Show the code

3.18.1 Example 1: stepwise regression

Conduct your own linear regression analysis by following a stepwise logic:

- The variable that shows the highest correlation with the dependent variable is selected (in absolute terms)

- Show the code

- The variable that has the highest semi-partial correlation with the dependent variable from the remaining variables is then selected (in absolute terms). The semi-partial correlation coefficient is the correlation between all of Y and that part of X which is independent of Z.

- Show the code

Each time check whether the variable significantly improves the model.

3.18.2 Example 2: deductive approach

Conduct your own linear regression analysis by following a deductive logic:

- Include the independent variables simultaneously in the regression equation

- Show the code

- Include hierarchically (blockwise) the independent variables into groups and include them in the regression equation group by group (e.g. sociodemographic variables first and political variables then).

- Show the code

Each time check whether the (groups of) variables significantly improve the model.

- Show the code

3.18.3 Example 3: postulates and assumptions

Based on the previous regression model, evaluate the assumptions of linear regression: no multicolinearity, normality of the residuals, no auto-correlation of the residuals (especially for time series and panel data), homoscedasticity. It is also important to check for outliers.

First, we want to check for multicollinearity. In case of multicollinearity issue, regression model is not able to accurately associate variance in the outcome variable with the correct predictor variable, leading to incorrect inferences. Beyond theoretical reflections, there are several steps to test for multicollinearity issues, such as correlation, VIF (and tolerance):

- Show the code

Second, we can test for the normality of the residuals. In order to make valid inferences, the residuals of the regression (differences between the observed value of the dependent variable and the predicted value) should follow a normal distribution. We can examine the normal Predicted Probability (P-P) plot to determine if the residuals are normally distributed (ideally, they they will conform to the diagonal normality line):

- Show the code

Third, we can check for homoscedasticity (variance of the error terms should be constant for all values of the independent variables). In the context of t-tests and ANOVAs, the same concept is referred to as equality (or homogeneity) of variances. We want to assess whether residuals are spread equally along the ranges of predictors (ideally, there should be a horizontal line with equally spread points). We can check this by plotting the predicted values and residuals on a scatterplot:

- Show the code

We can also check for linearity of the residuals, which means that the predictor variables in the regression have a straight-line relationship with the outcome variable (ideally, the plot would not have a pattern where the red line is approximately horizontal at zero):

- Show the code

Nota bene: Using the Durbin-Watson test, we can test the null hypothesis stating that the errors are not auto-correlated with themselves (if p-value > 0.05, we would fail to reject the null hypothesis).

- Show the code

Fourth, we can also verify if we have outliers. A value >2(p+1)/n indicates an observation with high leverage, where p is the number of predictors and n is the number of observations (in our case: 2*(3+1)/1348=0.006).

- Show the code

To extract outliers, you might want to flag observations whose leverage score is more than three times greater than the mean leverage value as a high leverage point.

- Show the code

3.18.4 Example 4: suppressor effect

Suppose a new variable X2 is added to the regression equation in addition to X1. Suppose, X1 explains substantial variance of Y because both variables capture well a certain phenomenon (here: B). Under which circumstances does this increase the model quality (R^2)?

Normally, an increase in R^2 can only be expected if:

- X2 is correlated with Y: when both X2 and Y capture a particular phenomenon (here: C)
- X1 and X2 are only weakly correlated because they predominantly capture different phenomena (here: B and C)

In cases where this ideal scenario is not or only limited valid, R2 will hardly increase, and may even weaken the influence of X1 on Y when X2 is added.

Now, suppose a predictor variable X1 captures 70% of phenomenon A and 30% of another phenomenon B. Y, on the other hand, captures phenomenon B dominantly. Then the variable X1 will correlate only very moderately with Y since the dominance of phenomenon B in Y has virtually prevented a higher correlation. Suppose a new variable X2 is added to the regression equation. Under what circumstances does this increase the model quality?

Chapter 3: Linear regression (*answers)*

3.18 Time to practice on your own

The exercises 1 and 2 will use the data from the round 10 of the European Social Survey (ESS). You can download the data directly on the ESS website.

The objective is to conduct linear regression to explain the consumption of news about politics and current affairs ('nwspol': watching, reading or listening in minutes). Explanatory variables include a set of political and sociodemographic variables. Political variables can include: political interest ('polintr'), confidence in ability to participate in politics ('cptppola'), and self-placement on the left-right political scale ('lrscale'). Sociodemographic variables can include: gender ('gndr'), age ('agea'), and years of education ('eduyrs'). You can also decide to rely on additional or alternative variables.

Let's start by loading the dataset, selecting the relevant variables, and filtering the data for Switzerland only:

➢ Show the code

```
library(foreign)
db <- read.spss(file=paste0(getwd(),
                  "/data/ESS10.sav"),
                use.value.labels = F,
                to.data.frame = T)
sel <- db |>
  dplyr::select(cntry, nwspol, polintr, cptppola, lrscale, gndr, agea, eduyrs) |>
  stats::na.omit() |>
  dplyr::filter(cntry=="CH") # select respondents from Switzerland
# verify the class and range of the variables
sel$nwspol=as.numeric(sel$nwspol)
sel = sel[sel$nwspol<=180,] # maximally 3hours of news consumption
sel$gndr = as.factor(sel$gndr)
sel$gndr = ifelse(sel$gndr=="2", "female", "male")
sel$gndr = as.factor(as.character(sel$gndr))
```

3.18.1 Example 1: stepwise regression

Conduct your own linear regression analysis by following a stepwise logic:

- The variable that shows the highest correlation with the dependent variable is selected (in absolute terms)

- Show the code

```
round(cor(sel[,c("nwspol", "polintr", "cptppola", "lrscale", "agea", "eduyrs")]),2)
##          nwspol polintr cptppola lrscale  agea eduyrs
## nwspol     1.00   -0.32     0.11    0.05  0.31  -0.01
## polintr   -0.32    1.00    -0.47    0.07 -0.19  -0.20
## cptppola   0.11   -0.47     1.00   -0.04 -0.06   0.19
## lrscale    0.05    0.07    -0.04    1.00  0.16  -0.20
## agea       0.31   -0.19    -0.06    0.16  1.00  -0.12
## eduyrs    -0.01   -0.20     0.19   -0.20 -0.12   1.00
reg1 =lm(nwspol ~ polintr, data=sel)
summary(reg1)
##
## Call:
## lm(formula = nwspol ~ polintr, data = sel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73.89 -29.59 -13.89  16.11 134.71
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   88.196      2.826   31.21   <2e-16 ***
## polintr      -14.301      1.162  -12.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.12 on 1346 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.1005
```

```
## F-statistic: 151.4 on 1 and 1346 DF,  p-value: < 2.2e-16
```

- The variable that has the highest semi-partial correlation with the dependent variable from the remaining variables is then selected (in absolute terms). The semi-partial correlation coefficient is the correlation between all of Y and that part of X which is independent of Z.

➢ Show the code

```
res1 = resid(reg1)

round(cor(res1, sel[,c("cptppola", "lrscale", "agea", "eduyrs")]),2)

##      cptppola lrscale agea eduyrs

## [1,]    -0.05    0.07 0.26  -0.08

summary(lm(nwspol ~ polintr + agea, data=sel))

##

## Call:

## lm(formula = nwspol ~ polintr + agea, data = sel)

##

## Residuals:

##     Min      1Q  Median      3Q     Max

## -93.076 -24.208  -8.404  17.573 139.423

##

## Coefficients:

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

## (Intercept)  55.94452    4.16625   13.43   <2e-16 ***

## polintr     -12.05596    1.14124  -10.56   <2e-16 ***

## agea          0.54653    0.05343   10.23   <2e-16 ***

## ---

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

##

## Residual standard error: 35.77 on 1345 degrees of freedom

## Multiple R-squared:  0.166,  Adjusted R-squared:  0.1648
```

```
## F-statistic: 133.9 on 2 and 1345 DF,  p-value: < 2.2e-16
```

Each time check whether the variable significantly improves the model.

3.18.2 Example 2: deductive approach

Conduct your own linear regression analysis by following a deductive logic:

- Include the independent variables simultaneously in the regression equation

➢ Show the code

```
regbt = lm(nwspol ~
             polintr +
             cptppola +
             lrscale +
             agea +
             eduyrs +
             relevel(gndr,"male"),
           data=sel)
summary(regbt)
##
## Call:
## lm(formula = nwspol ~ polintr + cptppola + lrscale + agea + eduyrs +
##     relevel(gndr, "male"), data = sel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.288 -24.295  -7.841  17.618 139.605
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            62.29255    7.80393   7.982 3.06e-15 ***
## polintr               -12.57088    1.33517  -9.415  < 2e-16 ***
```

```
## cptppola                  -0.30173    1.06016  -0.285    0.776
## lrscale                   0.28449    0.49602   0.574    0.566
## agea                      0.52845    0.05564   9.498  < 2e-16 ***
## eduyrs                    -0.32865    0.26416  -1.244    0.214
## relevel(gndr, "male")female  -2.18405    2.00285  -1.090    0.276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.77 on 1341 degrees of freedom
## Multiple R-squared:  0.1685, Adjusted R-squared:  0.1648
## F-statistic:  45.3 on 6 and 1341 DF,  p-value: < 2.2e-16
```

- Include hierarchically (blockwise) the independent variables into groups and include them in the regression equation group by group (e.g. sociodemographic variables first and political variables then).

➢ Show the code

```
regbs = lm(nwspol ~
          agea +
          eduyrs +
          relevel(gndr,"male"),
         data=sel)
summary(regbs)
##
## Call:
## lm(formula = nwspol ~ agea + eduyrs + relevel(gndr, "male"),
##     data = sel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -81.489 -27.006  -8.368  19.546 149.590
##
## Coefficients:
##                          Estimate Std. Error t value   Pr(>|t|)
## (Intercept)               21.32329    4.49269   4.746 0.00000229 ***
## agea                       0.66480    0.05493  12.103    < 2e-16 ***
## eduyrs                     0.29096    0.26232   1.109     0.2676
## relevel(gndr, "male")female -4.16850    2.03222  -2.051     0.0404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.17 on 1344 degrees of freedom
## Multiple R-squared:  0.1002, Adjusted R-squared:  0.09822
## F-statistic:  49.9 on 3 and 1344 DF,  p-value: < 2.2e-16
```

Each time check whether the (groups of) variables significantly improve the model.

- Show the code

```
cat(paste0("R2 from the model with sociodemo variables is: ",
           round(summary(regbs)$adj.r.squared,3),
           "\n",
           "R2 from the model including all the variables is: ",
           round(summary(regbt)$adj.r.squared,3)))
## R2 from the model with sociodemo variables is: 0.098
## R2 from the model including all the variables is: 0.165
```

3.18.3 Example 3: postulates and assumptions

Based on the previous regression model, evaluate the assumptions of linear regression: no multicolinearity, normality of the residuals, no auto-correlation of the residuals (especially for time series and panel data), homoscedasticity. It is also important to check for outliers.

First, we want to check for multicollinearity. In case of multicollinearity issue, regression model is not able to accurately associate variance in the outcome variable with the correct predictor variable, leading

to incorrect inferences. Beyond theoretical reflections, there are several steps to test for multicollinearity issues, such as correlation, VIF (and tolerance):

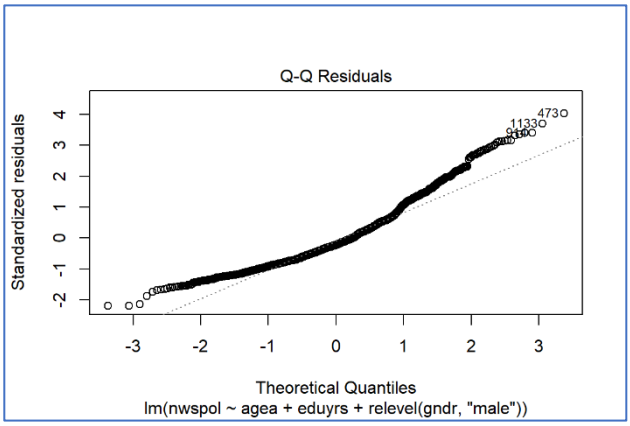
➢ Show the code

```
c <- sel[,-c(1,6)] # removes cntry and gndr
round(cor(c),3)
##         nwspol polintr cptppola lrscale   agea eduyrs
## nwspol    1.000  -0.318    0.105   0.046  0.311 -0.014
## polintr  -0.318   1.000   -0.471   0.072 -0.192 -0.202
## cptppola  0.105  -0.471    1.000  -0.037 -0.057  0.187
## lrscale   0.046   0.072   -0.037   1.000  0.158 -0.203
## agea      0.311  -0.192   -0.057   0.158  1.000 -0.124
## eduyrs   -0.014  -0.202    0.187  -0.203 -0.124  1.000
olsrr::ols_vif_tol(regbs)
##                 Variables Tolerance      VIF
## 1                    agea 0.9838041 1.016463
## 2                  eduyrs 0.9789747 1.021477
## 3 relevel(gndr, "male")female 0.9939485 1.006088
```

Second, we can test for the normality of the residuals. In order to make valid inferences, the residuals of the regression (differences between the observed value of the dependent variable and the predicted value) should follow a normal distribution. We can examine the normal Predicted Probability (P-P) plot to determine if the residuals are normally distributed (ideally, they they will conform to the diagonal normality line):

➢ Show the code

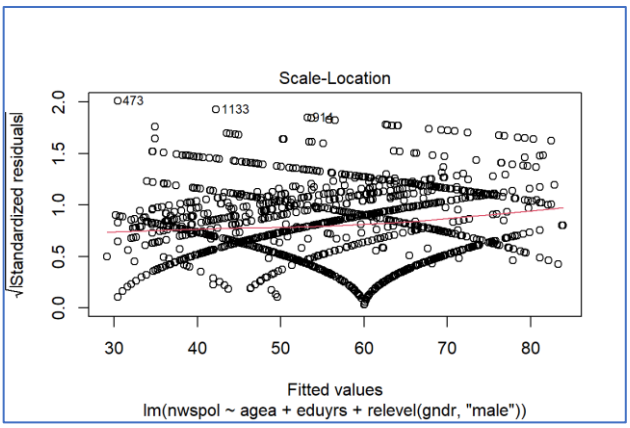
```
plot(regbs, 2)
```



Third, we can check for homoscedasticity (variance of the error terms should be constant for all values of the independent variables). In the context of t-tests and ANOVAs, the same concept is referred to as equality (or homogeneity) of variances. We want to assess whether residuals are spread equally along the ranges of predictors (ideally, there should be a horizontal line with equally spread points). We can check this by plotting the predicted values and residuals on a scatterplot:

➢ Show the code

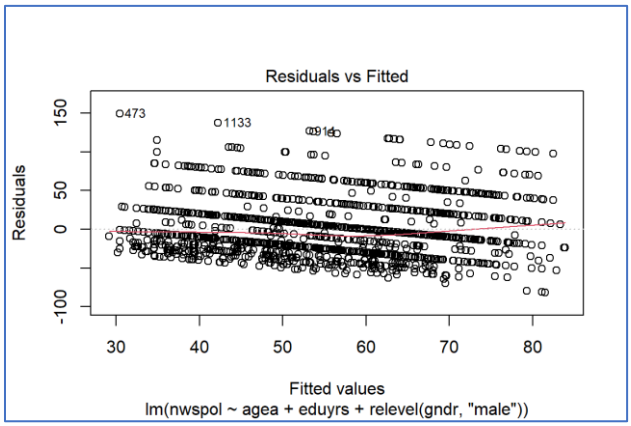
```
plot(regbs, 3)
```



We can also check for linearity of the residuals, which means that the predictor variables in the regression have a straight-line relationship with the outcome variable (ideally, the plot would not have a pattern where the red line is approximately horizontal at zero):

➢ Show the code

```
plot(regbs, 1)
```



Nota bene: Using the Durbin-Watson test, we can test the null hypothesis stating that the errors are not auto-correlated with themselves (if p-value > 0.05, we would fail to reject the null hypothesis).

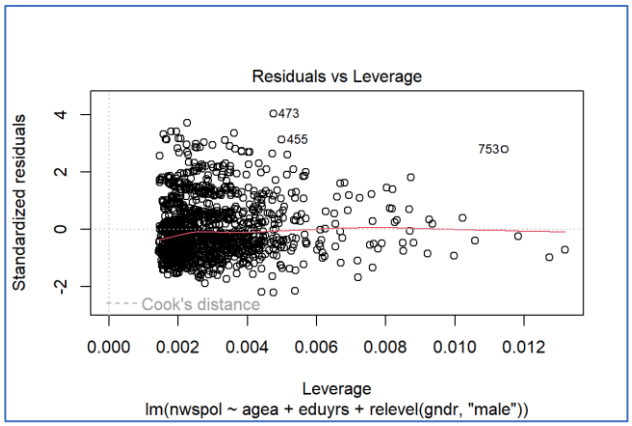
➢ Show the code

```
# car::durbinWatsonTest(regbs)
```

Fourth, we can also verify if we have outliers. A value >2(p+1)/n indicates an observation with high leverage, where p is the number of predictors and n is the number of observations (in our case: 2*(3+1)/1348=0.006).

➢ Show the code

```
plot(regbs, 5)
```



To extract outliers, you might want to flag observations whose leverage score is more than three times greater than the mean leverage value as a high leverage point.

➢ Show the code

```
model_data <- broom::augment(regbs)

high_lev <- dplyr::filter(model_data,.hat>3*mean(model_data$.hat))

high_lev

## # A tibble: 10 × 11

##   .rownames nwspol  agea eduyrs `relevel(gndr, "male")` .fitted .resid    .hat .sigma

##   <chr>     <dbl> <dbl>  <dbl> <fct>                      <dbl>  <dbl>   <dbl>  <dbl>

## 1 40           20    45     25 female                     54.3 -34.3  0.00998   37.2

## 2 140          60    31     25 female                     45.0  15.0  0.0102    37.2

## 3 345          15    39     27 female                     50.9 -35.9  0.0127    37.2

## 4 425          20    20      0 male                       34.6 -14.6  0.0106    37.2

## 5 536          60    37     24 male                       52.9   7.10 0.00935   37.2

## 6 725          80    76      0 female                     67.7  12.3  0.00925   37.2

## # 4 more rows
```

2 more variables: .cooksd <dbl>, .std.resid <dbl>

3.18.4 Example 4: suppressor effect

Suppose a new variable X2 is added to the regression equation in addition to X1. Suppose, X1 explains substantial variance of Y because both variables capture well a certain phenomenon (here: B). Under which circumstances does this increase the model quality (R^2)?

Normally, an increase in R^2 can only be expected if:

- X2 is correlated with Y: when both X2 and Y capture a particular phenomenon (here: C)
- X1 and X2 are only weakly correlated because they predominantly capture different phenomena (here: B and C)

In cases where this ideal scenario is not or only limited valid, R2 will hardly increase, and may even weaken the influence of X1 on Y when X2 is added.

Now, suppose a predictor variable X1 captures 70% of phenomenon A and 30% of another phenomenon B. Y, on the other hand, captures phenomenon B dominantly. Then the variable X1 will correlate only very moderately with Y since the dominance of phenomenon B in Y has virtually prevented a higher correlation. Suppose a new variable X2 is added to the regression equation. Under what circumstances does this increase the model quality?

> Assume that a second predictor variable X2 also dominantly captures phenomenon A. Phenomenon B is only weakly or not at all captured in X2. Since the predictors are controlled simultaneously and reciprocally, the influence of the first predictor variable X1 is “freed” from the dominant influence of phenomenon A and suddenly phenomenon B dominates, which in turn is also dominant in Y. Suddenly there is a strong influence of the predictor variable X1 on Y! In this case the second variable X2 is suppressor for the influence of X1 on Y. Why? Only the residual variance of X1 remains for correlations with Y, but this has very large common variance components with Y.