# Exercises Chapter 6

1. An alternative, equivalent representation of the $F$-test statistic is the following:

$$F = \frac{\left(\sum_{i=1}^n \hat{\varepsilon}_{iR}^2 - \sum_{i=1}^n \hat{\varepsilon}_{iU}^2\right)/q}{\left(\sum_{i=1}^n \hat{\varepsilon}_{iU}^2\right)/(n-K)} = \frac{\left(SS_R - SS_U\right)/q}{SS_U/(n-K)},$$

where $\hat{\varepsilon}_{iU}$ are the residuals from the *un*restricted (i.e., the usual) regression of $Y$ on $X$, and where $\hat{\varepsilon}_{iR}$ are the residuals from the *re*stricted ordinary least squares regression which minimizes the following *restricted* version of the OLS-objective function

$$\min_{\tilde{\beta}} S_n(\tilde{\beta}) = (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \quad \text{such that} \quad R\tilde{\beta} - r = 0$$

where the restriction is just the null hypothesis.

**The standard $F$-test.** The standard $F$-test for a linear regression tests the hypothesis that all coefficients except the intercept are equal to zero. In this case, $\hat{\varepsilon}_{iR}$ are simply the residuals from regressing $Y$ on only the intercept. In this standard case we have,

$$F_1 = \frac{\left(\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n \hat{\varepsilon}_{iU}^2\right)/(K-1)}{\left(\sum_{i=1}^n \hat{\varepsilon}_{iU}^2\right)/(n-K)}$$

since here $\hat{\varepsilon}_{iR}^2 = (Y_i - \bar{Y})^2$.

Show that $F_1$ is equal to $F_2$ with

$$F_2 = \frac{R_U^2/(K-1)}{(1-R_U^2)/(n-K)},$$

where $R_U^2$ denotes the coefficient of determination of the unrestricted regression model.

2. Install the R package `AER` and load the package. The `ARE`-package contains the data set `Journals`. Check `?Journals` to learn more about the data. Create the variables `citeprice` (journal price per citations) and age (journal age) as following:

```
> # install.packages("AER")
> suppressMessages(library("AER"))
> ## attach the data-set Journals to the current R-session
> data("Journals", package = "AER")
> ## ?Journals # Check the help file
> ##
> ## Select variables "subs" and "price"
> journals           <- Journals[, c("subs", "price")]
> ## Define variable 'journal-price per citation'
> journals$citeprice <- Journals$price/Journals$citations
> ## Define variable 'journal-age'
> journals$age       <- 2020 - Journals$foundingyear
> ## Check variable names in 'journals'
> names(journals)
```

```
[1] "subs"      "price"     "citeprice" "age"
```

Estimate the coefficients $\beta_1$ and $\beta_2$ of the following linear regression model

$$\log(Y_i) = \beta_1 + \beta_2 \log(X_i) + \varepsilon_i, \quad i = 1, \dots, n$$

with $\log(Y) =$`log(subs)` (i.e., logarithm of the number of library subscriptions) and $\log(X) =$ `log(citeprice)` (i.e., logarithm of the journal price per citations).

(a) Do you have heteroscedastic error-term variances? Explain your answer by discussing a diagnostic plot showing the residuals against the fitted values.

(b) Estimate the standard error of the OLS estimator $\hat{\beta}_2$ using an appropriate variance estimator.

3. Consider the following multiple linear regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i, \quad i = 1, \dots, n$$
$$\text{(in matrix notation)} \quad Y = X\beta + \varepsilon$$

where $\beta = (1, -5, 5)'$, $\varepsilon_i$ is a heteroscedastic error term

$$\varepsilon_i \sim N(0, \sigma_i^2) \quad \text{with} \quad \sigma_i = |X_{3i}|,$$

and where for all $i = 1, \dots, n = 100$:

- $X_{2i} \sim N(10, 1.5^2)$
- $X_{3i} \sim U[0.2, 8]$

You're given the following data generated from this regression model:

```
> set.seed(109) # Sets the "seed" of the random number generators:
> n   <- 100     # Number of observations
> ## Generate two explanatory variables plus an intercept-variable:
> X_1  <- rep(1, n)                      # Intercept
> X_2  <- rnorm(n, mean=10, sd=1.5)     # Draw realizations form a normal distr.
> X_3  <- runif(n, min = 0.2, max = 8) # Draw realizations form a t-distr.
> X    <- cbind(X_1, X_2, X_3)          # Save as a Nx3-dimensional data matrix.
> beta <- c(1, -5, 5)
> ## Generate realizations from the heteroscadastic error term
> eps <- rnorm(n, mean=0, sd=abs(X_3))
> ## Dependent variable:
> Y   <- X %*% beta + eps
```

(a) Compute the theoretical covariance matrix variance $\mathrm{Var}(\hat{\beta})$ of the OLS estimator $\hat{\beta}$ for the given data generating process and the given data.

(b) Use a Monte-Carlo simulation to generate 10000 variance estimates

$$\widehat{\mathrm{Var}}_{\text{HC3},1}(\hat{\beta}_2), \dots, \widehat{\mathrm{Var}}_{\text{HC3},10000}(\hat{\beta}_2)$$

and 10000 variance estimates

$$\widehat{\mathrm{Var}}_{\text{HC3},1}(\hat{\beta}_3), \dots, \widehat{\mathrm{Var}}_{\text{HC3},10000}(\hat{\beta}_3).$$

These estimates represent typical estimation results. (Of course, in practice you observe only one variance estimation result $\widehat{\mathrm{Var}}_{\text{HC3}}(\hat{\beta}_2)$ for $\mathrm{Var}(\hat{\beta}_2)$ and one $\widehat{\mathrm{Var}}_{\text{HC3}}(\hat{\beta}_3)$ for $\mathrm{Var}(\hat{\beta}_3)$.)

(i) Visualize the Monte Carlo realizations for the variance estimates. Add points displaying the sample mean of the Monte Carlo realizations and points displaying the true variance values.

(ii) Do the Monte Carlo realizations $\widehat{\mathrm{Var}}_{\text{HC3},r}(\hat{\beta}_2)$ and $\widehat{\mathrm{Var}}_{\text{HC3},r}(\hat{\beta}_3)$, $r = 1, \dots, 10000$ estimate the true variances $\mathrm{Var}(\hat{\beta}_2)$ and $\mathrm{Var}(\hat{\beta}_3)$ well on average?

(iii) Are there large estimation uncertainties?

4. The `Boston` housing data set (contained in the R package `MASS`) contains observations on housing values in suburbs of Boston. Let's consider the following regression model

$$\text{medv}_i = \beta_1 + \beta_2\text{ptratio}_i + \beta_3\text{lstat}_i + \beta_4\text{age}_i + \beta_5\text{crim}_i + \beta_6\text{nox}_i + \varepsilon_i$$

where $i = 1, \dots, n$ indexes the suburbs. Check `?Boston` in R to get an overview about the variables. You can assume that the assumptions of Chapter 6 hold. The following R code computes the regression estimates:

```
> library("lmtest")  # for coeftest()
> library("sandwich")# for robust se
> library("MASS")    # for Boston housing data
> data("Boston")     # Check: ?Boston; names(Boston)
> lm_obj      <- lm(medv ~ ptratio + lstat + age + crim + nox, data = Boston)
> vcovHC3_mat <- vcovHC(lm_obj, type = "HC3")
> round(coeftest(lm_obj, vcov = vcovHC3_mat), 3)

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   55.746      3.013  18.504   <2e-16 ***
ptratio       -1.181      0.147  -8.046   <2e-16 ***
lstat         -0.868      0.081 -10.662   <2e-16 ***
age            0.060      0.017   3.527   <2e-16 ***
crim          -0.024      0.036  -0.674    0.501
nox           -8.059      3.318  -2.429    0.016 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Use R to test $H_0 : \beta_6 = 0$ versus $H_A : \beta_6 < 0$ by means of an $t$-test. What is the correct $p$-value and what is the test decision when using a significance level of $\alpha = 0.01$?

(b) Use R to test $H_0 : \beta_5 = \beta_6 = 0$ versus $H_A : \beta_5 \neq 0$ and/or $\beta_6 \neq 0$ by means of an $F$-test. What is the marginal significance value in this case?

(c) What is the maximal probability of a type I error if you test the null hypothesis in (b) by means of two separate $t$-tests instead of one $F$-test? How does this compare to the probability of a type I error for the $F$ test in (b).