Exercises (with Solutions) · Chapter 6

1. **Problem**

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

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

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.
Solution
By the definition of the R_U^2 we have that

$$\begin{aligned} R_U^2 &= 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\ 1 - R_U^2 &= \frac{\sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \end{aligned}$$

Inserting R_U^2 and $1 - R_U^2$ into F_2 yields:

$$\begin{aligned} F_2 &= \frac{R_U^2/(K-1)}{(1-R_U^2)/(n-K)} \\ &= \frac{\left(\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\right)/(K-1)}{\left(\frac{\sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\right)/(n-k)} \\ &= \frac{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2\right)/(K-1)}{\left(\sum_{i=1}^{n}\hat{\varepsilon}_{iU}^2\right)/(n-k)} = F_1 \end{aligned}$$

2. **Problem**

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 β_1 and β_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.

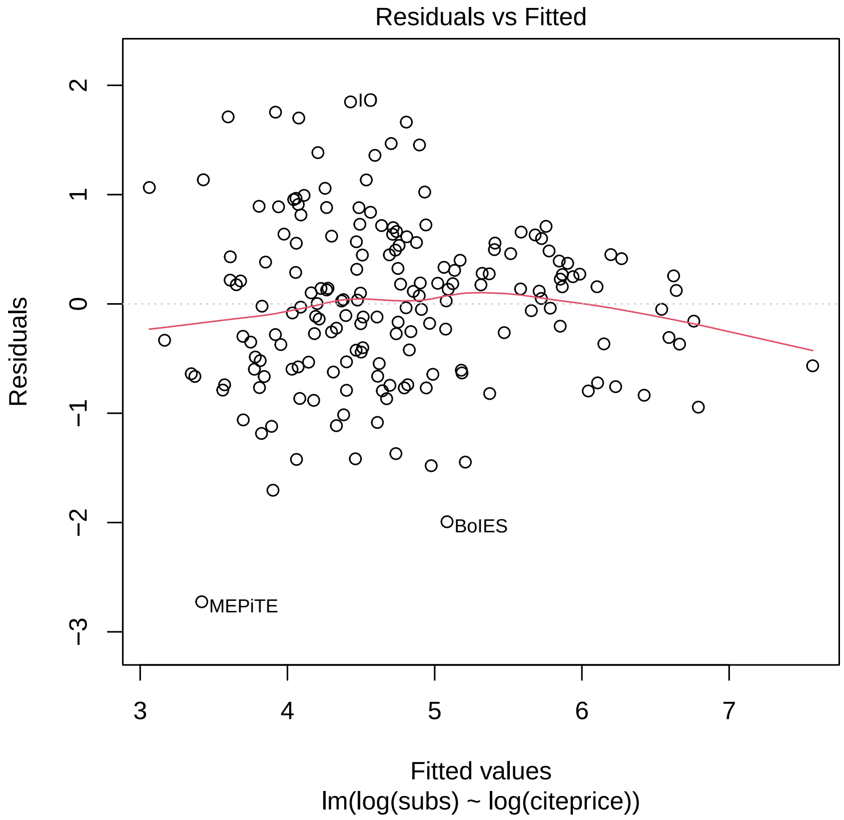
Solution

(a) The error terms seem to have heteroscedastic variances. This can be seen, for instance, by plotting the residuals $\hat{\varepsilon}_i$ against the fitted values $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$.

Note. One usually plots the residuals $\hat{\varepsilon}_i$ against the *fitted values* $\hat{Y}_i$ (not on the explanatory variable X_i), since plotting against the fitted values works also in the case of multiple regressors ($K > 2$).

The following code computes the OLS estimation and shows a typical diagonstic plot for checking heteroscedasticitiy in the residuals:

```
> jour_lm <- lm(log(subs) ~ log(citeprice), data = journals)
> ## Diagnostic plot residuals against fitted values
> ## plot(y=resid(jour_lm), x=fitted(jour_lm))
> ## Or slightly more fancy
> plot(jour_lm, which=1)
```



(b) In case of heteroscedastic error term variances, we need to consider a robust heteroscedasticity consistent variance estimator such as the following one:

$$\widehat{\text{Var}}_{\text{HC3}}(\hat{\beta}) = \left(\frac{1}{n}X'X\right)^{-1} \widehat{E}(\varepsilon_i X_i X_i') \left(\frac{1}{n}X'X\right)^{-1}$$
$$= S_{X'X}^{-1}\widehat{E}(\varepsilon_i^2 X_i X_i')S_{X'X}^{-1}$$
$$\text{with} \quad \widehat{E}(\varepsilon_i X_i X_i') = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}X_i X_i',$$

and where $h_i = [P_X]_{ii}$ is the leverage statistic of X_i.

Here is the estimation result using R:

```
> library("sandwich") # HC robust variance estimation
> ## Robust estimation of the variance of \hat\beta:
> Var_hat_beta_HC3 <- sandwich::vcovHC(jour_lm, type="HC3")
> ## Robust standard error of \hat\beta_2
> sqrt(diag(Var_hat_beta_HC3)[2])

log(citeprice)
    0.03447364

> ## Comparison with the classic standard error estimation
> sqrt(diag(vcov(jour_lm))[2])

log(citeprice)
     0.0356132
```

3. **Problem**

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 $\beta = (1, -5, 5)'$, ε_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

$$\widehat{\mathrm{Var}}_{\mathsf{HC3},1}(\hat\beta_2), \dots, \widehat{\mathrm{Var}}_{\mathsf{HC3},10000}(\hat\beta_2)$$

and 10000 variance estimates

$$\widehat{\mathrm{Var}}_{\mathsf{HC3},1}(\hat\beta_3), \dots, \widehat{\mathrm{Var}}_{\mathsf{HC3},10000}(\hat\beta_3).$$

These estimates represent typical estimation results. (Of course, in practice you observe only one variance estimation result $\widehat{\mathrm{Var}}_{\mathsf{HC3}}(\hat\beta_2)$ for $\mathrm{Var}(\hat\beta_2)$ and one $\widehat{\mathrm{Var}}_{\mathsf{HC3}}(\hat\beta_3)$ 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 $\widehat{\mathrm{Var}}_{\mathsf{HC3},r}(\hat\beta_2)$ and $\widehat{\mathrm{Var}}_{\mathsf{HC3},r}(\hat\beta_3)$, $r = 1, \dots, 10000$ estimate the true variances $\mathrm{Var}(\hat\beta_2)$ and $\mathrm{Var}(\hat\beta_3)$ well on average?

(iii) Are there large estimation uncertainties?

Solution

(a) The theoretical covariance matrix $\mathrm{Var}(\hat\beta)$ of the OLS estimator $\hat\beta$ is given by

$$\mathrm{Var}(\hat\beta) = (X'X)^{-1}\,X'\,\mathrm{Var}(\varepsilon)X\,(X'X)^{-1},$$

where $\mathrm{Var}(\varepsilon) = \mathrm{diag}(X_{31}^2, \dots, X_{3n}^2)$. To compute the values of the variance-covariance matrix $\mathrm{Var}(\hat\beta)$ we can use R as following:

```
> Var_theo           <- solve(t(X) %*% X) %*% t(X) %*% diag(X_3^2) %*%
+                          X %*% solve(t(X) %*% X)
> rownames(Var_theo) <- c("", "", "") # remove row-names
> colnames(Var_theo) <- c("", "", "") # remove col-names
> round(Var_theo, 3)

  7.229 -0.683 -0.073
 -0.683  0.069 -0.008
 -0.073 -0.008  0.057
```

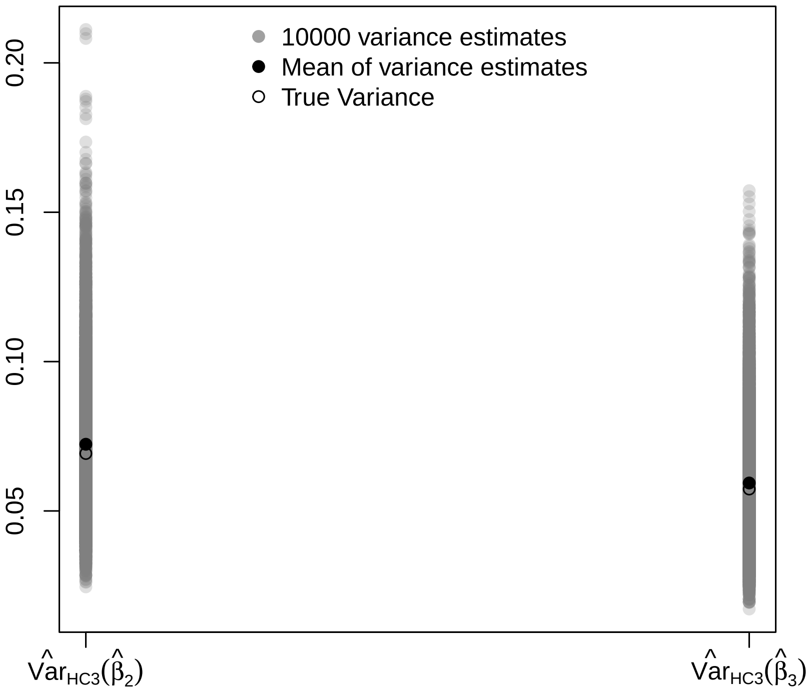
(b i) R code for the Monte-Carlo simulations and the plot:

```
> library("sandwich") # HC robust variance estimation
> MC_reps       <- 10000  # Number of Monte Carlo replications
> VarHC3_estims <- matrix(NA, 3, MC_reps) # Container to collect the results
> for(r in 1:MC_reps){
+ ## Generate new realizations from the heteroscedastic error term
+ eps <- rnorm(n, mean=0, sd=abs(X_3))
+
+ ## Generate new realizations from the dependent variable:
+ Y      <- X %*% beta + eps
+
+ ## Now OLS estimation
+ lm_fit <- lm(Y ~ X - 1) # '-1' since X contains an intercept
+
+ ## Now robust estimation of the variance of \hat\beta:
+ VarHC3_estims[,r] <- diag(sandwich::vcovHC(lm_fit, type="HC3"))
+ }
> VarHC3_estims_means <- rowMeans(VarHC3_estims)
> ## Compare the theoretical variances Var(\hat\beta_2) and Var(\hat\beta_3)
> ## with the means of the 10000 variance estimations
> ## \hat{Var}(\hat\beta_2) and \hat{Var}(\hat\beta_3)
> cbind(diag(Var_theo)[c(2,3)], VarHC3_estims_means[c(2,3)])

       [,1]       [,2]
 0.06923203 0.07232364
 0.05734246 0.05935843

> plot(x=c(1,2), y=c(0,0), ylim=range(VarHC3_estims[c(2,3),]), type="n", axes = FALSE,
+ xlab = "", ylab = "")
> box()
> axis(1, c(1,2), labels=c(expression(hat({Var})[{HC3}](hat(beta)[2])),
+                           expression(hat({Var})[{HC3}](hat(beta)[3]))))
> axis(2)
> points(x=rep(1,MC_reps), y=VarHC3_estims[2,], pch=21, col=gray(.5,.25), bg=gray(.5,.25))
> points(x=1, y=VarHC3_estims_means[2], pch=21, col="black", bg="black")
> points(x=1, y=diag(Var_theo)[2], pch=1)
> points(x=rep(2,MC_reps), y=VarHC3_estims[3,], pch=21, col=gray(.5,.25), bg=gray(.5,.25))
```

```
> points(x=2, y=VarHC3_estims_means[3], pch=21, col="black", bg="black")
> points(x=2, y=diag(Var_theo)[3], pch=1)
> legend("top",
+        legend = c("10000 variance estimates", "Mean of variance estimates",
+                   "True Variance"), bty = "n", pt.bg = c(gray(.5,.75),"black","black"),
+        pch = c(21,21,1), col=c(gray(.5,.75),"black","black"))
```



(ii) On average, the 10000 estimates $\widehat{\mathrm{Var}}_{\text{HC3}}(\hat{\beta}_2)$ and $\widehat{\mathrm{Var}}_{\text{HC3}}(\hat{\beta}_3)$ approximate well the true variances $\mathrm{Var}(\hat{\beta}_2) = 0.069$ and $\mathrm{Var}(\hat{\beta}_2) = 0.057$.

(iii) There are considerable estimation uncertainties (large variances of the estimators) with estimates ranging from 0.025 to 0.211 and from 0.017 to 0.157.

4. **Problem**

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 $H_0 : \beta_6 \geq 0$ 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).

Solution

(a) The regression output reports two-sided t-tests. The observed value of the t-test statistic for $\hat{\beta}_6$ is $t_{\text{obs}} = -2.429$ with a two-sided p-value $p_{\text{two-sided}} = 0.016$, where

$$p_{\text{two-sided}} = 2\min\{\underbrace{P_{H_0}(t \leq t_{\text{obs}})}_{p_{\text{one-sided, lower}}}, \underbrace{P_{H_0}(t \geq t_{\text{obs}})}_{p_{\text{one-sided, upper}}}\}.$$

The correct p-value for $H_A : \beta_6 < 0$, however, is $p_{\text{one-sided, lower}}$.
Since $t_{\text{obs}} = -2.429$ is negative we know that $P_{H_0}(t \leq t_{\text{obs}}) < P_{H_0}(t \geq t_{\text{obs}})$. Therefore,

$$p_{\text{two-sided}} = 2p_{\text{one-sided, lower}}$$

which allows us to compute the correct $p_{\text{one-sided, lower}}$ by

$$\begin{aligned} p_{\text{one-sided, lower}} &= p_{\text{two-sided}}/2 \\ &= 0.016/2 = 0.008. \end{aligned}$$

That is, we can reject the null-hypothesis $H_0 : \beta_6 = 0$ against the alternative $H_A : \beta_6 < 0$ at the significance level of $\alpha = 0.01$.

(b) A test of $H_0 : \beta_5 = \beta_6 = 0$ versus $H_A : \beta_5 \neq 0$ and/or $\beta_6 \neq 0$ by means of an F-test can be conducted as following:

```
> library("car")
> linearHypothesis(lm_obj, c("crim=0", "nox=0"),
+                  vcov = vcovHC3_mat)
```

```
Linear hypothesis test

Hypothesis:
crim = 0
nox = 0

Model 1: restricted model
Model 2: medv ~ ptratio + lstat + age + crim + nox

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F  Pr(>F)
1    502
2    500  2 3.7413 0.02439 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

So, the marginal significance value (p-value) in this case is 0.024. That is, we can reject the null hypothesis at any significance level $\alpha > 0.024$.

(c) When we test $H_0 : \beta_5 = \beta_6 = 0$ versus $H_A : \beta_5 \neq 0$ and/or $\beta_6 \neq 0$ by means of two separate t-tests, we may conduct a type I error in either of the two test-decisions. Therefore, the joint probability of a type I error is

$$P_{H_0}(|t^{(5)}| > c_{1-\alpha/2} \cup |t^{(6)}| > c_{1-\alpha/2}),$$

where $t^{(5)}$ and $t^{(6)}$ denote here the t-tests based on $\hat{\beta}_5$ and $\hat{\beta}_6$ respectively.
Using that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (see Script, Chapter 1.1.2) we can derive the following upper threshold for this joint probability of type I errors (assuming Assumptions 1-4* hold):

$$\begin{aligned} P_{H_0}(|t^{(5)}| > c_{1-\alpha/2} \cup |t^{(6)}| > c_{1-\alpha/2}) &= P_{H_0}(|t^{(5)}| > c_{1-\alpha/2}) + P_{H_0}(|t^{(6)}| > c_{1-\alpha/2}) \\ &\quad - P_{H_0}(|t^{(5)}| > c_{1-\alpha/2} \cap |t^{(6)}| > c_{1-\alpha/2}) \\ &\leq P_{H_0}(|t^{(5)}| > c_{1-\alpha/2}) + P_{H_0}(|t^{(6)}| > c_{1-\alpha/2}) = 2\alpha \end{aligned}$$

So, the maximal probability of a type I error if you test the null hypothesis in (b) by means of two separate t-tests is two times the significance level α of the two separate t-tests. That is, in order to do inference at a chosen α-level, we can conduct each of the two separate t-tests at an $\alpha/2$ significance level. (This is then called a Bonferroni-adjustment.) Such an adjustment makes it generally harder to detect a violation of the null-hypothesis then when using the F test in (b); particularly, in the case of hypothesis involving more than two parameters.