

Chapter 6 Homework Answers

John Fox

Soc. 740, Winter 2012

Exercise D6.3

You can construct confidence intervals “manually” from the coefficient estimates and standard errors, or “automatically” using the `confint` function.

Reading the data and filtering out the missing data:

```
> UN <- read.table(
+ "http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/UnitedNations.txt",
+ header=TRUE)
> dim(UN)
[1] 207 13
> UN <- na.omit(UN[,c("tfr", "GDPperCapita", "illiteracyFemale", "contraception")])
> dim(UN)
[1] 118 4
```

Regression of TFR on GDP Per Capita:

```
> mod.gdp <- lm(tfr ~GDPperCapita, data=UN)
> summary(mod.gdp)
```

Call:

```
lm(formula = tfr ~GDPperCapita, data = UN)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.6758 -1.1183 -0.1419  1.0434  3.5492
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.147e+00  1.663e-01  24.939 < 2e-16 ***
GDPperCapita -1.320e-04  2.793e-05  -4.725 6.51e-06 ***
---
```

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

```
Residual standard error: 1.539 on 116 degrees of freedom
Multiple R-squared:  0.1614,    Adjusted R-squared:  0.1542
F-statistic: 22.33 on 1 and 116 DF,  p-value: 6.51e-06
```

```
> confint(mod.gdp)
              2.5 %       97.5 %
(Intercept)  3.8180087865  4.476779e+00
GDPperCapita -0.0001872730 -7.664479e-05
```

Here, the degrees of freedom for error are $n - 2 = 118 - 2 = 116$, and so the critical value of t is $t_{.025,116} = 1.9806$:

```
> qt(.025, df=116, lower.tail=FALSE)
[1] 1.980626
```

Thus, for example, to construct the 95-percent confidence interval for the slope manually, we have

$$\begin{aligned}\beta &= B \pm t_{\alpha/2}SE(B) \\ &= -0.0001320 \pm 1.9806 \times 0.00002793 \\ &= -0.0001320 \pm 0.00005532 \\ &= (-0.0001873, -0.0000767)\end{aligned}$$

which is, within rounding error, the same result as reported by `confint`.

Regression of TFR on Female Illiteracy:

```
> mod.illit <- lm(tfr ~illiteracyFemale, data=UN)
> summary(mod.illit)
```

Call:

```
lm(formula = tfr ~illiteracyFemale, data = UN)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.72565	-0.80639	-0.07995	0.74146	2.40063

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.28938	0.14569	15.71	<2e-16 ***
illiteracyFemale	0.04839	0.00363	13.33	<2e-16 ***

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

Residual standard error: 1.056 on 116 degrees of freedom
Multiple R-squared: 0.6051, Adjusted R-squared: 0.6017
F-statistic: 177.7 on 1 and 116 DF, p-value: < 2.2e-16

```
> confint(mod.illit)
```

	2.5 %	97.5 %
(Intercept)	2.0008375	2.57793113
illiteracyFemale	0.0412037	0.05558304

Regression of TFR on Contraception:

```
> mod.con <- lm(tfr ~contraception, data=UN)
> summary(mod.con)
```

Call:

```
lm(formula = tfr ~contraception, data = UN)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.26855	-0.52558	0.07253	0.71491	2.41757

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.905947	0.212777	27.76	<2e-16 ***
contraception	-0.054669	0.004666	-11.71	<2e-16 ***

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

```
Residual standard error: 1.137 on 116 degrees of freedom
Multiple R-squared:  0.542,    Adjusted R-squared:  0.538
F-statistic: 137.3 on 1 and 116 DF,  p-value: < 2.2e-16
```

```
> confint(mod.con)
                2.5 %    97.5 %
(Intercept)    5.48451554 6.327378
contraception -0.06391189 -0.045427
```

Exercise D6.5

Again, one could construct the confidence intervals manually or via `confint`:

```
> mod.un <- lm(tfr ~GDPperCapita + illiteracyFemale + contraception, data=UN)
> summary(mod.un)
```

Call:

```
lm(formula = tfr ~GDPperCapita + illiteracyFemale + contraception,
    data = UN)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-2.0259 -0.5864  0.0830  0.5195  2.1703
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.021e+00  2.837e-01  14.175 < 2e-16 ***
GDPperCapita  -2.998e-05  1.771e-05  -1.693  0.0933 .
illiteracyFemale 3.188e-02  3.905e-03   8.165 4.83e-13 ***
contraception  -2.884e-02  4.800e-03  -6.008 2.30e-08 ***
```

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

```
Residual standard error: 0.8952 on 114 degrees of freedom
Multiple R-squared:  0.7212,    Adjusted R-squared:  0.7138
F-statistic: 98.28 on 3 and 114 DF,  p-value: < 2.2e-16
```

```
> confint(mod.un)
                2.5 %    97.5 %
(Intercept)    3.459165e+00 4.583120e+00
GDPperCapita  -6.507083e-05 5.107412e-06
illiteracyFemale 2.414574e-02 3.961558e-02
contraception  -3.834571e-02 -1.932883e-02
```

The degrees of freedom for error in the multiple regression are $n - k - 1 = 118 - 3 - 1 = 114$, and the critical t for 95-percent confidence intervals is $t_{.025,114} = 1.9810$:

```
> qt(.025, df=114, lower.tail=FALSE)
[1] 1.980992
```

The 95-percent confidence interval for the GDP coefficient, for example, is therefore

$$\begin{aligned}\beta_1 &= B_1 \pm t_{\alpha/2} \text{SE}(B_1) \\ &= -0.00002998 \pm 1.9810 \times 0.00001771 \\ &= -0.00002998 \pm .00003508 \\ &= (-0.0006506, 0.00000510)\end{aligned}$$

which, notice, includes 0 (and agrees, within rounding error, with the interval reported by `confint`).

R reports the omnibus F -test for the regression, $F_{3,114} = 98.28$, $p \approx 0$, but it doesn't report the ANOVA table. There are many different ways to find the sums of squares. Here are three:

(1) Add up the sums of squares from the sequential analysis of variance:

```
> anova(mod.un)
Analysis of Variance Table

Response: tfr
      Df Sum Sq Mean Sq F value    Pr(>F)
GDPperCapita      1  52.875   52.875   65.987 6.012e-13 ***
illiteracyFemale  1 154.454  154.454  192.755 < 2.2e-16 ***
contraception     1  28.924   28.924   36.096 2.299e-08 ***
Residuals       114  91.348    0.801
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, $\text{RSS} = 91.348$, $\text{RegSS} = 52.875 + 154.454 + 28.294 = 236.253$, and $\text{TSS} = \text{RSS} + \text{RegSS} = 91.348 + 236.253 = 327.601$.

(2) Compute the sums of squares directly:

```
> with(UN, sum((tfr - mean(tfr))^2)) # TSS
[1] 327.6015
>
> sum(residuals(mod.un)^2) # RSS
[1] 91.34823
>
> sum((fitted.values(mod.un) - mean(UN$tfr))^2) # RegSS
[1] 236.2532
```

(3) Formulate the test as an incremental F -test against a null model that contains only the intercept:

```
> anova(mod.0, mod.un)
Analysis of Variance Table

Model 1: tfr ~ 1
Model 2: tfr ~ GDPperCapita + illiteracyFemale + contraception
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     117 327.60
2     114  91.35  3    236.25 98.28 < 2.2e-16 ***
```

The first method suffers slightly from rounding error. In any event, the Anova table is

<i>Source</i>	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>p</i>
Regression	236.25	3	78.75	98.28	$\ll .0001$
Residuals	91.35	114	0.8013		
Total	327.60	117			

The incremental F -test for the GDP coefficient, $H_0: \beta_1 = 0$:

```

> mod.gdp.0 <- update(mod.un, . ~. - GDPperCapita)
> anova(mod.gdp.0, mod.un)
Analysis of Variance Table

Model 1: tfr ~illiteracyFemale + contraception
Model 2: tfr ~GDPperCapita + illiteracyFemale + contraception
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     115 93.644
2     114 91.348  1     2.2958 2.8651 0.09325 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Check: $t^2 = (-1.693)^2 = 2.866$.

One could also compute the incremental F “manually” from the regression or residual sums of squares or R^2 's for the two models: From the output above, we have $R_1^2 = .7212$, while from the following output, $R_0^2 = .7142$.

```

> summary(mod.gdp.0)

Call:
lm(formula = tfr ~illiteracyFemale + contraception, data = UN)

Residuals:
    Min       1Q   Median       3Q      Max
-1.969174 -0.629662  0.005065  0.536075  2.303221

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.990654   0.285399   13.983 < 2e-16 ***
illiteracyFemale 0.032578   0.003914    8.323 2.01e-13 ***
contraception  -0.030950   0.004672   -6.624 1.17e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 0.9024 on 115 degrees of freedom
Multiple R-squared: 0.7142, Adjusted R-squared: 0.7092
F-statistic: 143.7 on 2 and 115 DF, p-value: < 2.2e-16

Thus,

$$\begin{aligned}
F_0 &= \frac{n - k - 1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2} \\
&= \frac{118 - 3 - 1}{1} \times \frac{.7212 - .7142}{1 - .7212} \\
&= 2.862 \text{ on } 1 \text{ and } 114 \text{ } df
\end{aligned}$$

which is the same result within rounding error.