Assignment-8

Oluwanifemi

2024-11-22

```r
library(faraway)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```r
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':
##
##     logit, vif

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

##
## Attaching package: 'alr4'

## The following objects are masked from 'package:faraway':
##
##     cathedral, pipeline, twins
```

```r
data("UN11")
head(UN11)
```

```
##                 region  group fertility   ppgdp lifeExpF pctUrban
## Afghanistan       Asia  other     5.968   499.0    49.49       23
## Albania         Europe  other     1.525  3677.2    80.40       53
## Algeria         Africa africa     2.142  4473.0    75.00       67
## Angola          Africa africa     5.135  4321.9    53.17       59
## Anguilla     Caribbean  other     2.000 13750.1    81.10      100
## Argentina    Latin Amer  other     2.172  9162.1    79.89       93
```

```r
model_1 <- lm(formula = lifeExpF ~ I(log(ppgdp)) + pctUrban + fertility,
data = UN11)

summary(model_1)
```

```
##
## Call:
## lm(formula = lifeExpF ~ I(log(ppgdp)) + pctUrban + fertility,
##     data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6165  -1.6683   0.5406   2.6425  11.2263
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   64.21764    3.84281  16.711  < 2e-16 ***
## I(log(ppgdp))  2.17842    0.42888   5.079 8.83e-07 ***
## pctUrban       0.02116    0.02353   0.900    0.369
## fertility     -4.19652    0.39396 -10.652  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.145 on 195 degrees of freedom
## Multiple R-squared:  0.7456, Adjusted R-squared:  0.7417
## F-statistic: 190.5 on 3 and 195 DF,  p-value: < 2.2e-16
```

```r
View(UN11)
```

Question 1(b)

Conclusion for the Overall F-Test

From the R output, the **overall F-test** evaluates whether at least one of the regression coefficients (excluding the intercept) is significantly different from zero.

- **F-statistic**: 190.5
- **Degrees of Freedom (DF)**: 195
- **p-value**: $< 2.2 \times 10\text{-}16$
- **Significance Level** (α): 0.01

Conclusion: The p-value for the F-test ($< 2.2 \times 10\text{-}16$) is much smaller than the significance level of 0.01. **We reject the null hypothesis**, concluding that the model as a whole is statistically significant.

Reasoning: At least one predictor in the model significantly contributes to explaining the variability in life expectancy.

Question 1(c)

```r
reduced_model <- lm(lifeExpF ~ fertility, data = UN11)

# Perform the nested F-test
anova_result <- anova(reduced_model, model_1)
anova_result
```

```
## Analysis of Variance Table
##
## Model 1: lifeExpF ~ fertility
## Model 2: lifeExpF ~ I(log(ppgdp)) + pctUrban + fertility
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 6528.4
## 2    195 5161.7  2    1366.6 25.814 1.133e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing Whether the Parameters for `log(ppgdp)` and `pctUrban` Are Both Zero

1. Hypotheses

- **Null Hypothesis** (H_0): The parameters for `log(ppgdp)` and `pctUrban` are both 0. These variables do not contribute to explaining the variability in `lifeExpF`.
- **Alternative Hypothesis** (H_a): At least one of the parameters for `log(ppgdp)` or `pctUrban` is not 0. Including these variables significantly improves the model.

2. Test Statistic The test statistic is the F-statistic from the ANOVA output:

$$F = 25.814$$

- **Degrees of Freedom**: - Numerator (df_1): 2 (number of predictors added: `log(ppgdp)` and `pctUrban`). - Denominator (df_2): 195 (residual degrees of freedom from the full model).

3. Significance Level The significance level (α) is 0.05.

4. p-value From the ANOVA output:

$$p = 1.133 \times 10^{-10}$$

5. Decision Compare the p-value to the significance level ($\alpha = 0.05$):

$$p = 1.133 \times 10^{-10} < 0.05$$

Since the p-value is much smaller than 0.05, we **reject the null hypothesis**.

6. Conclusion At the 0.05 significance level, there is sufficient evidence to conclude that the parameters for `log(ppgdp)` and/or `pctUrban` are not 0. Adding these predictors significantly improves the model's ability to explain `lifeExpF`.

Question 1(d)

Problem: ANOVA Table for Testing $H_0 : \beta_1 = \beta_2$

ANOVA Table

The ANOVA table is constructed using the provided and computed values:

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Squares (MS)	F-Statistic
Error (SSE_Ω)	195	5161.65	26.470	-
Error (SSE_ω)	196	5787.45	-	-
Difference	1	625.8	625.8	23.64

(a) Compute Degrees of Freedom

The difference in degrees of freedom is calculated as:

$$df_{\text{difference}} = 196 - 195 = 1.$$

I got 195 from the df from the model_1, I ran earlier, and 196 because if (n-p-1) is 195, the 196 is n-p

(b) Compute Sum of Squares

To get Error (SSE_Ω), we know df is 195 from first anova table, therefore it will be 26.470 * 195 = 5161.65 . To get,(SSE_ω, it will be, 5267.78 + 625.8 = 5787.45

$$SS_{\text{difference}} = 625.8$$

(c) Compute Mean Squares

The mean square difference is calculated using the formula:

$$MS_{\text{difference}} = \frac{SS_{\text{difference}}}{df_{\text{difference}}} = \frac{625.8}{1} = 625.8$$

The mean square error (MSE) is given directly as:

$$MSE = 26.470$$

(d) Compute the F-Statistic

The F-statistic is computed as:

$$F = \frac{MS_{\text{difference}}}{MSE} = \frac{625.8}{26.470} \approx 23.64$$

Conclusion

The F-statistic value of 23.64 can be used to test the null hypothesis H_0, comparing the full and reduced models. Using the F-statistic of 23.64 and the corresponding p-value (as calculated earlier or obtained from statistical software), we reject the null hypothesis H_0 at the 0.05 significance level. This indicates that at least one of the predictors, log(ppgdp) or pctUrban, is significant in predicting life expectancy.

Question 2:

Problem 2: True/False with Explanations

a) Consider a model with 5 predictors. We can use an ANOVA test to test whether $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ holds, and the "difference" degrees of freedom would be 4. False.
- In this hypothesis test, you are testing whether three specific predictors $(\beta_1, \beta_2, \beta_3)$ are simultaneously equal to 0. The "difference" degrees of freedom would equal the number of predictors being tested, which is $df_{\text{difference}} = 3$, not 4. The total number of predictors in the model (5) does not determine the degrees of freedom for the test unless all predictors are included in the hypothesis.

b) It's possible to get a negative ANOVA F test-statistic. False.
- The F-statistic is the ratio of two variances ($MS_{\text{difference}}/MSE$), which are always non-negative. Since variances cannot be negative, the F-statistic is also always non-negative. A negative F-statistic would indicate an error in calculation.

c) I have a linear regression model predicting house prices in NJ from the number of bedrooms and bathrooms. I build a 95% confidence interval for the average house price for houses with three bedrooms and two bathrooms and get the interval (250,000, 650,000). I can say that there is a 95% probability that the average three-bedroom and two-bathroom house price in NJ is between $250,000 and $650,000. False.
- Confidence intervals do not express probabilities about the population parameter after the data have been collected. The correct interpretation is: *We are 95% confident that the true average house price for three-bedroom and two-bathroom houses in NJ lies between $250,000 and $650,000.* The confidence level reflects the long-run proportion of confidence intervals that will contain the true value if repeated sampling were done.

d) If we increase sample size, we should expect our confidence intervals to get narrower. True.
- As sample size increases, the standard error of the estimate decreases because the variance is divided by a larger sample size. Narrower confidence intervals reflect this reduction in uncertainty about the estimate, leading to more precise estimates of the population parameter.