uwa6xv_M07_HW

Alanna Hazlett

2024-03-20

Problem 1

Use the dataset swiss. The goal of the data set was to assess how fertility rates in the Swiss (French-speaking) provinces relate to a number of demographic variables.

```
library(datasets)
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
Data<-swiss
```

(a)
Previously you fit a model with the fertility measure as the response variable and used all the other variables as predictors. Now, consider a simpler model, using only the last three variables as predictors: Education, Catholic, and Infant.Mortality. Carry out an appropriate hypothesis test to assess which of these two models should be used. State the null and alternative hypotheses, find the relevant test statistic, p-value, and state a conclusion in context. (For practice, try to calculate the test statistic by hand.)

```
result.full<-lm(Fertility~.,data=Data)
result.reduced<-lm(Fertility ~ Education + Catholic + Infant.Mortality, data = Data)
```

The appropriate hypothesis test is a general liner F test.

$$H_0 : \hat{\beta}_{Agriculture} = \hat{\beta}_{Examination} = 0$$
$$H_a : at\ least\ one\ \neq 0$$

```
anova(result.reduced,result.full)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Education + Catholic + Infant.Mortality
## Model 2: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 2422.2
## 2     41 2105.0  2     317.2 3.0891 0.05628 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F_0 = \frac{(SS_R(F) - SS_R(R)) \ / \ r}{(SS_{res}(F)) \ / \ (n-p)} = \frac{(SS_{res}(R) - SS_{res}(F)) \ / \ r}{(SS_{res}(F)) \ / \ (n-p)}$$

$$F_0 = \frac{(2422.2 - 2105) \ / \ 2}{2105 / \ (47-7)} = 3.0138$$

Finding p-value and critical value to compare t statistic to:

```
(1-pf(3.0138,2,(47-7)))
```

```
## [1] 0.06037167
```

```
qf(1-(0.05/2),2,(47-7))
```

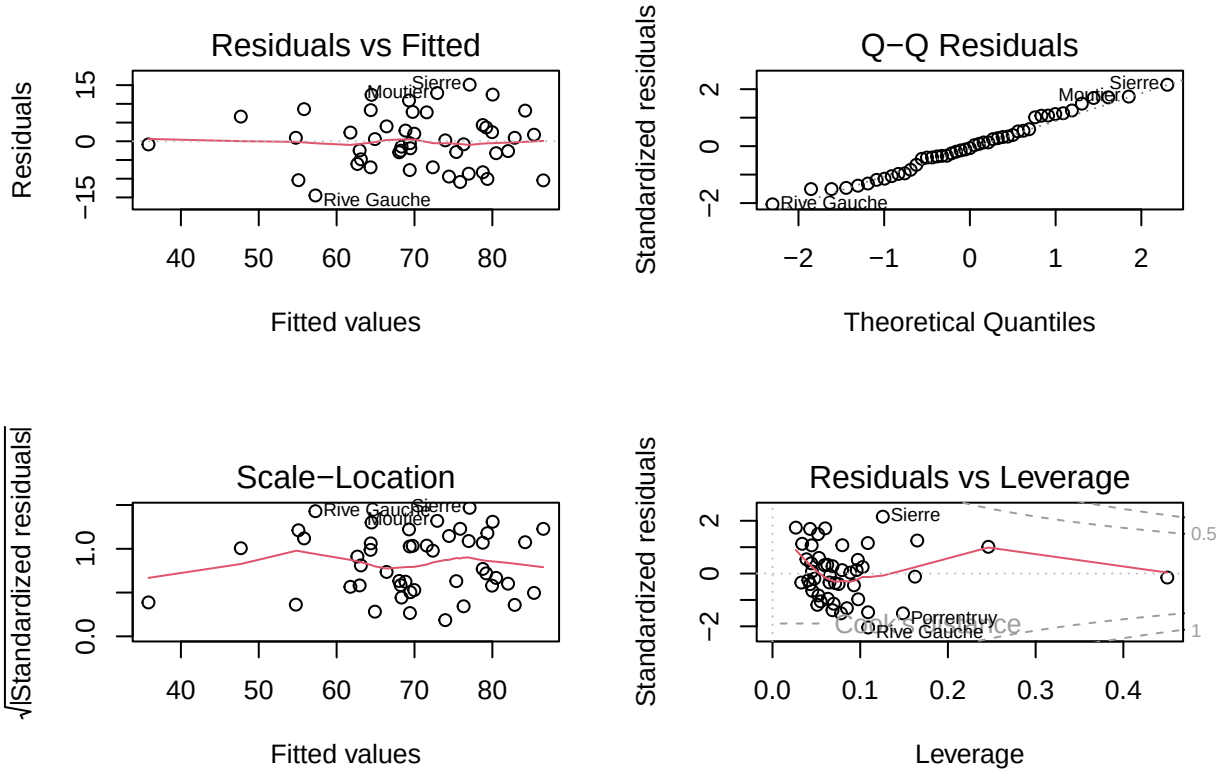
```
## [1] 4.050992
```

Our p-value is larger than our significance level of 0.05, so we fail to reject the null hypothesis. This supports that we should utilize the reduced model, drop the predictors Agriculture and Examination.

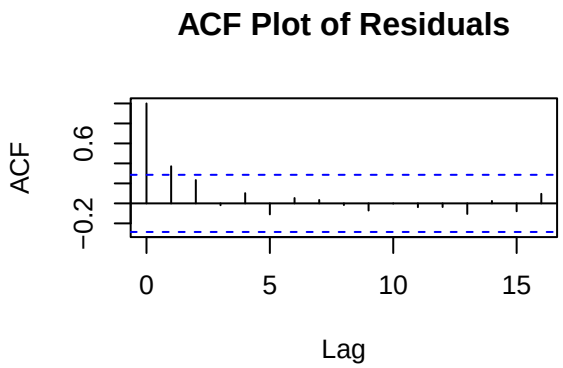
(b)

For the model you decide to use from part 1a, assess if the regression assumptions are met.

```
par(mfrow=c(2,2))
plot(result.reduced)
```



```
acf(result.reduced$residuals, main="ACF Plot of Residuals")
```



In Residuals vs Fitted we can see that the mean of the errors is very near zero, so assumption one is met. We can see that the variance of the errors is relatively constant from left to right, so assumption two is met. In the Q-Q Residuals plot we can see that the errors are normally distributed, so assumption four is met. In the ACF of Residuals we do see one value that surpasses our dashed line, however the majority of the points are within, so assumption 3 indicating that the errors are independent is met.

Problem 2

Hospital Infection Risk

(a)

Based on the t statistics, which predictors appear to be insignificant? Finding critical value to compare t statistic to:

```
qt(1-(0.05/2),113-6)
```

```
## [1] 1.982383
```

The predictors that appear to be insignificant are Age, Census, and Bed.

(b)

Based on your answer in part 2a, carry out the appropriate hypothesis test to see if those predictors can be dropped from the multiple regression model. Show all steps, including your null and alternative hypotheses, the corresponding test statistic, p-value, critical value, and your conclusion in context.

$$H_0 : \hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 0$$

$$H_a : at\ least\ one\ coefficient \neq 0$$

Our test statistic is:

$$F_0 = \frac{(0.136 + 5.101 + 0.028)\ /\ 3}{105.413/\ (113 - 6)} = 1.7814$$

Finding critical value to compare F0 statistic to and finding p-value:

```
qf(1-(0.05/2),3,(113-6))
```

```
## [1] 3.240674
```

```
(1-pf(1.7814,3,(113-6)))
```

```
## [1] 0.1550967
```

We fail to reject the null hypothesis, our data support the alternative hypothesis. This means we can not drop all 3 of these terms, Age, Census, and Bed. We should use the full model, comprised of Stay, Cultures, Age, Census, and Bed.

(c)

Suppose we want to decide between two potential models:

* Model 1: using x1, x2, x3, x4 as the predictors for InfctRsk
* Model 2: using x1, x2 as the predictors for InfctRsk

Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

$$H_0 : \hat{\beta}_3 = \hat{\beta}_4 = 0$$

$$H_a : at\ least\ one\ coefficient \neq 0$$

Our test statistic is:

$$F_0 = \frac{(0.136 + 5.101)\ /\ 2}{(105.413 + 0.028)/\ (113 - 5)} = 1.7814$$

Finding critical value to compare F0 statistic to and finding p-value:

```
qf(0.975,2,(113-5))
```

```
## [1] 3.817797
```

```
1-pf(2.6820,2,(113-5))
```

```
## [1] 0.0729834
```

We fail to reject the null hypothesis, our data support the alternative hypothesis. This means we can not drop both of these terms, Age and Census, we should use the full model (Model 1) that is comprised of Stay, Cultures, Age, and Census.

Problem 3

Explain how this output indicates the presence of multicollinearity in this regression model.

For both of our predictors LeftFoot and RtFoot, the t statistics are insignificant. We can determine this, because the p-values for these predictor variables are larger than our significance level. This is problematic, as these are important/the only predictor variables we have for this model.