STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 4

1. The file `house.csv` contains data on 27 houses sold in Pennsylvania in 1977. We have the option of treating the number of bedrooms as numerical or categorical. For this analysis, we will fit an ANCOVA model to predict house prices, treating the number of bedrooms as categorical (which may be more appropriate - why?)

   (a) Fit a parallel regression ANCOVA model using price as the response variable, the number of bedrooms as the categorical predictor and the continuous variables Taxes, LotSize, LivingSpace and Age. Test whether the number of bedrooms is related to the house price. Create the appropriate plots to check diagnostics.

   Solution:

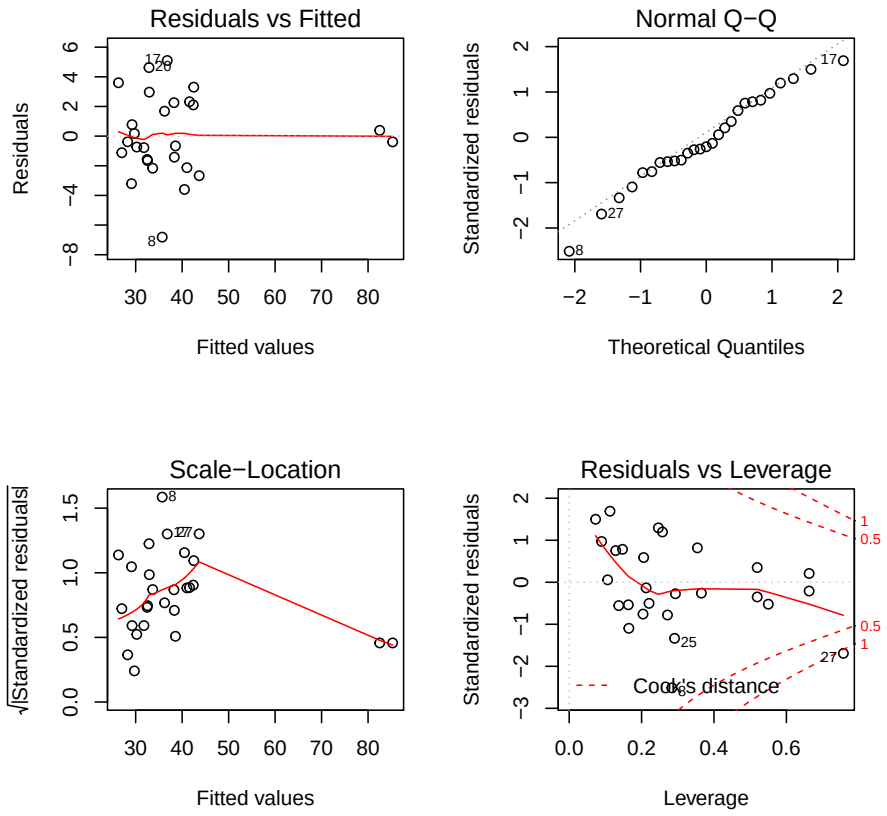
```
> house <- read.table("House.txt", header = TRUE)
> house.lm <- lm(Price ~ Age+Taxes+LivingSpace+LotSize+factor(Bedrooms),
                 data=house)
> anova(house.lm)

Analysis of Variance Table

Response: Price
                 Df Sum Sq Mean Sq  F value    Pr(>F)
Age               1  512.4   512.4  49.9544 1.003e-06 ***
Taxes             1 3951.4  3951.4 385.2033 4.485e-14 ***
LivingSpace       1  489.3   489.3  47.7029 1.382e-06 ***
LotSize           1    3.5     3.5   0.3409  0.566165
factor(Bedrooms)  3  170.7    56.9   5.5481  0.006588 **
Residuals        19  194.9    10.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   So, the number of bedrooms is clearly a significant factor in the price of a house.

```
> par(mfrow=c(2, 2))
> plot(house.lm)
```



Based on these plots there does not seem to be a problem with normality or heteroscedasticity. The residual plot does show that there are two somewhat odd points, but these correspond to the two five-bedroom houses in the dataset and thus we can understand why they appear out where they do. The residuals versus leverage plot point to the last data point (data point 27) as a potentially influential one which might require further investigation. It turns out that this data point has a rather large taxation value. Investigating further removing this data point has a dramatic effect on the parameter estimates.

```
> coefficients(house.lm)
```

```
      (Intercept)               Age             Taxes       LivingSpace
      11.79611402       -0.02378062        1.85661387        6.92485450
          LotSize factor(Bedrooms)3 factor(Bedrooms)4 factor(Bedrooms)5
       0.41643700       -0.08447372       -1.26386892       18.09741094
```

```
> house.lm1 <- lm(Price ~ Age+Taxes+LivingSpace+LotSize+factor(Bedrooms),
+                 data=house[-27, ])
> coefficients(house.lm1)
```

```
      (Intercept)               Age             Taxes       LivingSpace
       8.21024135        0.02463474        3.13084827        2.67364598
```

```
      LotSize factor(Bedrooms)3 factor(Bedrooms)4 factor(Bedrooms)5
0.18330737        1.36074948       -1.11686966       17.57133570
```

(b) What is the expected difference in price between a three and a four-bedroom house? Find a 95% confidence interval for this value. Repeat this exercise for the difference in price between a four and a five bedroom house. Comment on the results.

Solution:

```
> attach(house)
> bed3 <- ifelse(Bedrooms==3, 1, 0)
> bed4 <- ifelse(Bedrooms==4, 1, 0)
> bed5 <- ifelse(Bedrooms==5, 1, 0)
> beds <- cbind(bed3, bed4, bed5)
> house.reg<-lm(Price ~ Age+Taxes+LivingSpace+LotSize+beds, data=house)
> cc <- c(0, 0, 0, 0, 0, -1, 1, 0)
> diff34 <- as.vector(t(cc)%*%house.reg$coef)
> Xmat <- cbind(1, Age, Taxes, LivingSpace, LotSize, beds)
> XtXi <- solve(t(Xmat)%*%Xmat)
> rtMSE <- summary(house.reg)$sigma
> sd <- rtMSE*sqrt(as.vector(t(cc)%*%XtXi%*%cc))
> upper <- diff34 + (qt(0.975, house.reg$df.residual)*sd)
> lower <- diff34 - (qt(0.975, house.reg$df.residual)*sd)
> cbind(lower, diff34, upper)

         lower    diff34    upper
[1,] -5.319775 -1.179395 2.960984

> cc <- c(0, 0, 0, 0, 0, 0, -1, 1)
> diff45 <- as.vector(t(cc)%*%house.reg$coef)
> sd <- rtMSE*sqrt(as.vector(t(cc)%*%XtXi%*%cc))
> upper <- diff45 + (qt(0.975, house.reg$df.residual)*sd)
> lower <- diff45 - (qt(0.975, house.reg$df.residual)*sd)
> cbind(lower, diff45, upper)

        lower   diff45    upper
[1,] 8.084931 19.36128 30.63763
```

The estimate for the difference between a three and four bedroom house is negative, indicating that three bedroom houses are more expensive. However, the confidence intervals contain the value 0. This means that there is no apparent difference in price between a three and four bedroom house. On the other hand, there is a large difference (nearly $20,000) between a four and a five bedroom house.

(c) Calculate the correlation matrix of the predictors (including the indicator variables associated with the number of bedrooms) and comment on its structure, particularly

in relation to your results in part (b).

Solution:

```
> options(digits=4)
> Xmat <- cbind(Age, Taxes, LivingSpace, LotSize, beds)
> cor(Xmat)
```

```
                Age    Taxes LivingSpace  LotSize    bed3     bed4    bed5
Age          1.0000 -0.37113    -0.17806 -0.38027 -0.2791  0.41416 -0.1740
Taxes       -0.3711  1.00000     0.83238  0.68673 -0.3589  0.03193  0.7696
LivingSpace -0.1781  0.83238     1.00000  0.70327 -0.4038  0.08416  0.8766
LotSize     -0.3803  0.68673     0.70327  1.00000 -0.2654  0.01639  0.5939
bed3        -0.2791 -0.35893    -0.40380 -0.26545  1.0000 -0.69693 -0.3688
bed4         0.4142  0.03193     0.08416  0.01639 -0.6969  1.00000 -0.1512
bed5        -0.1740  0.76957     0.87663  0.59386 -0.3688 -0.15119  1.0000
```

Notice that there are some large correlations between the continuous predictors and the indicators for the number of bedrooms. This means that multicollinerity may be a problem. In addition, it may partly explain the reason for the strange result in part (b) whereby three and four bedroom houses seemed to have the same price. The difference in parameter estimates for the indicators of three and four bedroom houses must be interpreted in the light of the restriction that all the other predictors are held constant. However, there is an indication that the number of bedrooms in a house is correlated with the other predictors, and thus such an estimate may not be very meaningful. In other words, from the perspective of the population from which the data was gathered, it may well be that four bedroom houses are more expensive than those with only three bedrooms because the three and four bedroom houses in the area differ not only in their number of bedrooms but also in all their other predictor values. Nonetheless, if the prices are adjusted for the other variables, then there may be no difference between a three and a four bedroom house. The trick, of course, is finding a four bedroom house which has similar values for its predictors to those of the three bedroom houses!

(d) A stepwise variable selection procedure identifies the model including the predictors Taxes, LivingSpace and Bedrooms as the best model. Perform a cross-validation to assess the predictive capability of the model by using the last 7 data points as the validation set and the first 20 data points as the modelling set. Compare these predictions to the actual values using the formula:

$$\sum_{i=21}^{27} (Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i$ is the predicted value from the regression fit on the modelling set. Do the same for the model fit in part (a). Which model seems best from this perspective?

Solution:

```
> Xmat1 <- cbind(Taxes,LivingSpace,beds)
> reg1<-lm(Price[1:20]~Taxes[1:20]+LivingSpace[1:20]+beds[1:20,])
> sum((Price[21:27]-(coef(reg1)[1]+Xmat1[21:27,]%*%coef(reg1)[-1]))^2)

[1] 233.8

> Xmat2 <- cbind(Age,Taxes,LivingSpace,LotSize,beds)
> reg2<-lm(Price[1:20]~Age[1:20]+Taxes[1:20]+LivingSpace[1:20]+
+ LotSize[1:20]+beds[1:20,])
> sum((Price[21:27]-(coef(reg2)[1]+Xmat2[21:27,]%*%coef(reg2)[-1]))^2)

[1] 260.1
```

So, from this perspective the smaller model is better, though not substantially.

2. As an example of an Analysis of Covariance (ANCOVA) model, the lecture notes presents an analysis of some data on Teacher Effectiveness (Example 2 on pages 13 to 18). The data for this example are available on Wattle. Use R to repeat the analyses described in the lecture notes (the original analysis used S-Plus rather than R). Can you still fit the models described in the lecture notes using R and still get the same output shown in the lecture notes?

Solution:

Please refer to R file for this question on Wattle.