Linear Models

Lab Week 10 Solutions

Question 1

(a)

```
cars=read.table("carsales.txt",header = TRUE)
cars
```

```
##    RetailSales    GNP Increase GNP.Increase
## 1          978 1112.5        0          0.0
## 2         1123 1143.0        0          0.0
## 3         1125 1169.3        0          0.0
## 4         1260 1204.7        0          0.0
## 5         1121 1248.9        0          0.0
## 6         1275 1277.9        0          0.0
## 7         1257 1308.9        0          0.0
## 8         1381 1344.0        1       1344.0
## 9         1172 1358.8        1       1358.8
## 10        1368 1383.8        1       1383.8
## 11        1382 1416.3        1       1416.3
## 12        1454 1430.9        1       1430.9
## 13        1260 1416.6        1       1416.6
## 14        1462 1440.9        1       1440.9
```

```
car_mod=lm(RetailSales~GNP+Increase+GNP:Increase,data=cars)
summary(car_mod)
```

```
##
## Call:
## lm(formula = RetailSales ~ GNP + Increase + GNP:Increase, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.26  -59.04   20.54   43.82  107.55
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -254.7117   585.6561  -0.435   0.6729
## GNP             1.1721     0.4835   2.424   0.0358 *
## Increase     -452.5303  1447.2017  -0.313   0.7609
## GNP:Increase    0.3016     1.0623   0.284   0.7822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.93 on 10 degrees of freedom
## Multiple R-squared:  0.7197, Adjusted R-squared:  0.6356
## F-statistic: 8.558 on 3 and 10 DF,  p-value: 0.004101
```

The t-test statistic for the interaction term is t=0.284, the p-value is 0.7822, so we have no evidence of an interaction between GNP and increase.

Method 2 - use `*` (most convenient)

```
car_mod=lm(RetailSales~GNP*Increase,data=cars)
summary(car_mod)
```

```
##
## Call:
## lm(formula = RetailSales ~ GNP * Increase, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -123.26  -59.04   20.54   43.82  107.55
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -254.7117   585.6561  -0.435   0.6729
## GNP             1.1721     0.4835   2.424   0.0358 *
## Increase     -452.5303  1447.2017  -0.313   0.7609
## GNP:Increase    0.3016     1.0623   0.284   0.7822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.93 on 10 degrees of freedom
## Multiple R-squared:  0.7197, Adjusted R-squared:  0.6356
## F-statistic: 8.558 on 3 and 10 DF,  p-value: 0.004101
```

(b)

```
car_mod=lm(RetailSales~GNP+Increase,data=cars)
summary(car_mod)
```

```
##
## Call:
## lm(formula = RetailSales ~ GNP + Increase, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -132.81  -56.06   19.65   52.41  102.98
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -330.2931   499.3971  -0.661   0.5220
## GNP            1.2346     0.4122   2.995   0.0122 *
## Increase     -42.4557    89.6101  -0.474   0.6449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.26 on 11 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.666
## F-statistic: 13.96 on 2 and 11 DF,  p-value: 0.0009579
```

The test statistic t=-0.474, p-value p=0.6449, we have no evidence of an effect of Increase.

Question 2

(a)

```
pill=read.table("pill.txt",header = TRUE)
pill
```

```
##     y Dummy1 Dummy2
## 1   9      0      0
## 2  12      0      0
## 3  17      0      0
## 4  13      0      0
## 5   5      0      0
## 6  19      0      0
## 7   6      0      0
## 8   8      0      0
## 9  10      0      0
## 10 11      0      0
## 11  2      0      0
## 12 14      0      0
## 13 15      0      0
## 14  1      0      0
## 15 12      0      0
## 16  4      1      0
## 17 15      1      0
## 18  8      1      0
## 19  6      1      0
## 20  9      1      0
## 21  8      1      0
## 22 18      1      0
## 23  0      1      0
## 24 12      1      0
## 25  6      1      0
## 26  7      1      0
## 27 10      1      0
## 28 11      1      0
## 29  2      1      0
## 30  6      1      0
## 31 13      0      1
## 32  7      0      1
## 33  2      0      1
## 34  0      0      1
## 35 11      0      1
## 36  8      0      1
## 37  5      0      1
## 38  3      0      1
## 39 10      0      1
## 40  9      0      1
## 41  8      0      1
## 42  8      0      1
## 43  4      0      1
## 44  7      0      1
## 45  1      0      1
```

Note, that

Response y <- number of days with colds

Reference group <- Placebo

Dummy1 <- New Pill

Dummy2 <- Vitamin C

```
pill_mod=lm(y~Dummy1+Dummy2,data=pill)
summary(pill_mod)
```

```
##
## Call:
## lm(formula = y ~ Dummy1 + Dummy2, data = pill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.267 -2.400  0.600  2.733  9.867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.267      1.187   8.646 7.15e-11 ***
## Dummy1        -2.133      1.679  -1.270   0.2109
## Dummy2        -3.867      1.679  -2.303   0.0263 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.599 on 42 degrees of freedom
## Multiple R-squared:  0.1124, Adjusted R-squared:  0.07018
## F-statistic:  2.66 on 2 and 42 DF,  p-value: 0.08168
```

Fit a linear model, with Dummy1 and Dummy2 as predictors,

$$y_i = \beta_0 + \beta_1 \text{Dummy1}_\text{i} + \beta_2 \text{Dummy2}_\text{i} + \varepsilon_i$$

The group means are β_0, $\beta_0 + \beta_1$ and $\beta_0 + \beta_2$.

From the summary output, the estimates are:

mean(Placebo) = 10.267

mean(New Pill) = 10.267-2.133 = 8.134

mean(Vitamin C) = 10.267- 3.867 = 6.4

Alternatively, the group sample means can be calculated directly as:

```
c(mean(pill$y[pill$Dummy1+pill$Dummy2==0]),
mean(pill$y[pill$Dummy1==1]),
mean(pill$y[pill$Dummy2==1]))
```

```
## [1] 10.266667  8.133333  6.400000
```

or

```
c(mean(pill$y[1:15]),
mean(pill$y[16:30]),
mean(pill$y[31:45]))
```

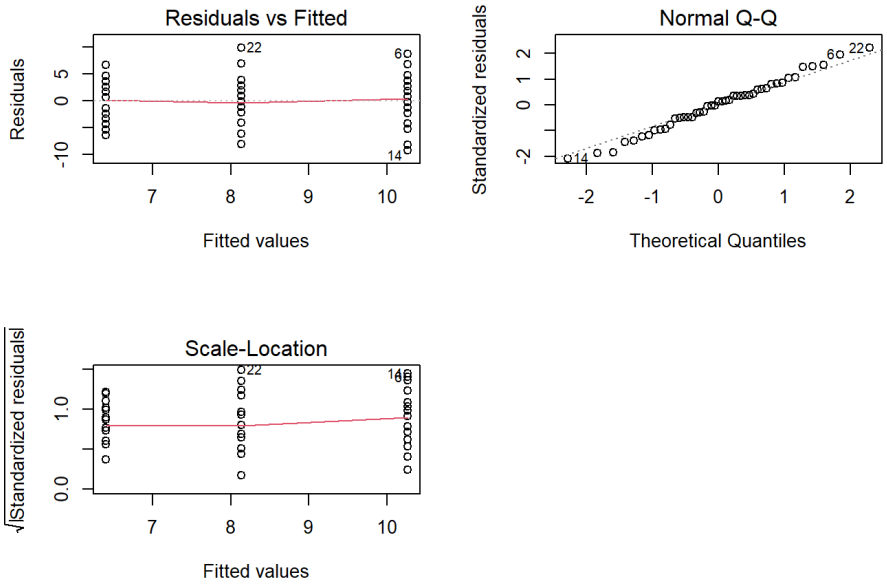
```
## [1] 10.266667  8.133333  6.400000
```

(b)

The groups have the same mean if the $\beta_1 = 0$ and $\beta_2 = 0$. This is the hypothesis tested by the F test in the summary function (see output in part (a)). F=2.66, p=0.08168, so we have weak evidence that the group mean number of colds differ. At a 5% significance level we would accept the null hypothesis that the groups have the same mean number of colds.

```
pill_mod=lm(y~Dummy1+Dummy2,data=pill)
par(mfrow=c(2,2))
plot(pill_mod)
```

```
## hat values (leverages) are all = 0.06666667
##  and there are no factor predictors; no plot no. 5
```



The residual plots look reasonable, no evidence of violations of assumptions.

(c)

Looking at the summery table, we know that the placebo group is the reference group (both Dummy variables are 0), so $b_1 = -2.133$ is the mean of the New pill group minus the mean of the Placebo group. Hence patients in the New Pill group had fewer colds than patients in the placebo group, however the p-value tells us this difference is not significant (p=0.2109). Likewise $b_2 = -3.867$, so vitamin C is also better (less colds) than the Placebo, however this time we have evidence (p=0.0263) that vitamin C is better (less colds) than the Placebo.

Question 3

```
sleep=read.table("sleep.txt",header = TRUE)
sleep
```

```
##    Score Group2 Group3
## 1   8.95      0      0
## 2   6.48      0      0
## 3   8.04      0      0
## 4   7.81      0      0
## 5   7.72      0      0
## 6   7.50      0      0
## 7   6.21      0      0
## 8   6.90      0      0
## 9   7.70      1      0
## 10  8.04      1      0
## 11  5.81      1      0
## 12  5.96      1      0
## 13  6.61      1      0
## 14  7.30      1      0
## 15  6.07      1      0
## 16  7.46      1      0
## 17  5.99      0      1
## 18  5.78      0      1
## 19  6.79      0      1
## 20  7.60      0      1
## 21  6.43      0      1
## 22  5.78      0      1
## 23  5.85      0      1
## 24  6.00      0      1
```

```
sleep_mod=lm(Score~Group2+Group3,data=sleep)
summary(sleep_mod)
```

```
##
## Call:
## lm(formula = Score ~ Group2 + Group3, data = sleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2412 -0.5109 -0.1050  0.5316  1.4988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4513     0.2856  26.087  < 2e-16 ***
## Group2       -0.5825     0.4039  -1.442  0.16404
## Group3       -1.1737     0.4039  -2.906  0.00845 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8079 on 21 degrees of freedom
## Multiple R-squared:  0.2868, Adjusted R-squared:  0.2188
## F-statistic: 4.222 on 2 and 21 DF,  p-value: 0.02877
```

If level of sleep is not related to performance, than the coefficients for both dummy variables would be 0, i.e. $\beta_1 = \beta_2 = 0$. To test this we can use the F-test above. The test statistic F=4.222, p=0.02877, so we have evidence, at 5% significance level, that sleep is related to performance.

Testing whether the mean performance score for group 2 and group 3 differs from the group 1 is equivalent to testing whether the relevant coefficient is zero or not. Hence, based on partial t-tests, with p-values of 0.16404 (for group 2) and 0.00845 (for group 3),
the performance of group 1 is significantly different from group 3, but not from group 2.

Question 4

(a)

We fit the following model with Light Blond as the reference group:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$\mathbb{E}(y_i) = \begin{cases} \beta_0 + \beta_1 & \text{DarkBlond} \\ \beta_0 + \beta_2 & \text{LightBrunette} \\ \beta_0 + \beta_3 & \text{DarkBrunette} \\ \beta_0 & \text{LightBlond} \end{cases}$$

```
hair <- read.table('blonds.txt',header = T)
model<- lm(Pain ~ DarkBlond + LightBrunette + DarkBrunette, data = hair)
summary(model)
```

```
##
## Call:
## lm(formula = Pain ~ DarkBlond + LightBrunette + DarkBrunette,
##     data = hair)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.20  -5.45  -0.50   4.30  13.60
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     59.200      3.655  16.198 6.53e-11 ***
## DarkBlond       -8.000      5.169  -1.548 0.142507
## LightBrunette  -16.700      5.482  -3.046 0.008166 **
## DarkBrunette   -21.800      5.169  -4.218 0.000746 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.172 on 15 degrees of freedom
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.4912
## F-statistic: 6.791 on 3 and 15 DF,  p-value: 0.004114
```

The estimated regression equation is

Pain = 59.2 - 8.00 DarkBlond - 16.7 LightBrunette - 21.8 DarkBrunette

The estimated mean pain scores are:

for Dark Blond, $b_0 + b_1 = 59.2 - 8.0 = 51.2$

for Light Brunette, $b_0 + b_2 = 59.2 - 16.7 = 42.5$

for Dark Brunette, $b_0 + b_3 = 59.2 - 21.8 = 37.4$

for Light Blond, $b_0 = 59.2$

(b)

The F-test here tests whether the mean pain score across all hair color groups is the same. This is because testing whether hair colour is associated with pain threshold is equivalent to testing whether all three predictor coefficients are zero or not (so that all 4 groups have the same mean). With a p-value of 0.004114, there does seem to be strong evidence to suggest that at least one group mean differs from the others.

Question 5

(a)

Method 1 -

X has 4 columns:

- Intercept: a column of 1's
- `FormatA` : 1 when `Format=="A"` , 0 otherwise
- `FormatB` : 1 when `Format=="B"` , 0 otherwise
- `FormatC` : 1 when `Format=="C"` , 0 otherwise

Check

```
time<-read.table('Timeformat.txt',header = T, stringsAsFactors = T)
levels(time$Format)
```

```
## [1] "a" "A" "B" "C"
```

```
model1 <-lm(Time ~ Format, time)
data.frame(time$Format,model.matrix(model1))
```

```
##    time.Format X.Intercept. FormatA FormatB FormatC
## 1            a            1       0       0       0
## 2            a            1       0       0       0
## 3            a            1       0       0       0
## 4            a            1       0       0       0
## 5            a            1       0       0       0
## 6            A            1       1       0       0
## 7            A            1       1       0       0
## 8            A            1       1       0       0
## 9            A            1       1       0       0
## 10           A            1       1       0       0
## 11           B            1       0       1       0
## 12           B            1       0       1       0
## 13           B            1       0       1       0
## 14           B            1       0       1       0
## 15           B            1       0       1       0
## 16           C            1       0       0       1
## 17           C            1       0       0       1
## 18           C            1       0       0       1
## 19           C            1       0       0       1
## 20           C            1       0       0       1
```

Method 2 -

X has 4 columns:

- `Formata` : 1 when `Format=="a"` , 0 otherwise
- `FormatA` : 1 when `Format=="A"` , 0 otherwise
- `FormatB` : 1 when `Format=="B"` , 0 otherwise
- `FormatC` : 1 when `Format=="C"` , 0 otherwise

Check

```
model2 <-lm(Time ~ -1 + Format, data=time)
data.frame(time$Format,model.matrix(model2))
```

```
##    time.Format Formata FormatA FormatB FormatC
## 1            a       1       0       0       0
## 2            a       1       0       0       0
## 3            a       1       0       0       0
## 4            a       1       0       0       0
## 5            a       1       0       0       0
## 6            A       0       1       0       0
## 7            A       0       1       0       0
## 8            A       0       1       0       0
## 9            A       0       1       0       0
## 10           A       0       1       0       0
## 11           B       0       0       1       0
## 12           B       0       0       1       0
## 13           B       0       0       1       0
## 14           B       0       0       1       0
## 15           B       0       0       1       0
## 16           C       0       0       0       1
## 17           C       0       0       0       1
## 18           C       0       0       0       1
## 19           C       0       0       0       1
## 20           C       0       0       0       1
```

(b)

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
C=matrix(c(0,1,-1,0,
           0,0,1,-1),2,4,byrow = TRUE)
linearHypothesis(model1,C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## FormatA - FormatB = 0
## FormatB - FormatC = 0
##
## Model 1: restricted model
## Model 2: Time ~ Format
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     18 42.416
## 2     16 23.404  2    19.012 6.4987 0.008592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic F=6.4987, p=0.008592

```
linearHypothesis(model2,C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## FormatA - FormatB = 0
## FormatB - FormatC = 0
##
## Model 1: restricted model
## Model 2: Time ~ -1 + Format
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     18 42.416
## 2     16 23.404  2    19.012 6.4987 0.008592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 6

(a)

```
data(iris)
iris2<-iris
iris2$versicolor <- (iris2$Species == "versicolor")*1
head(iris2)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species versicolor
## 1          5.1         3.5          1.4         0.2  setosa          0
## 2          4.9         3.0          1.4         0.2  setosa          0
## 3          4.7         3.2          1.3         0.2  setosa          0
## 4          4.6         3.1          1.5         0.2  setosa          0
## 5          5.0         3.6          1.4         0.2  setosa          0
## 6          5.4         3.9          1.7         0.4  setosa          0
```

```
m<-glm(versicolor~.-Species, data = iris2, family = binomial)
summary(m)
```

```
##
## Call:
## glm(formula = versicolor ~ . - Species, family = binomial, data = iris2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1280  -0.7668  -0.3818   0.7866   2.1202
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    7.3785     2.4993   2.952 0.003155 **
## Sepal.Length  -0.2454     0.6496  -0.378 0.705634
## Sepal.Width   -2.7966     0.7835  -3.569 0.000358 ***
## Petal.Length   1.3136     0.6838   1.921 0.054713 .
## Petal.Width   -2.7783     1.1731  -2.368 0.017868 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.95  on 149  degrees of freedom
## Residual deviance: 145.07  on 145  degrees of freedom
## AIC: 155.07
##
## Number of Fisher Scoring iterations: 5
```

We use the values of the estimated coefficinets of the logistic regression model from the R output.

```
Xb = 7.3785-0.2454*2.2 -2.7966*0.5 + 1.3136*1.2 -2.7783*2.2
p =exp(Xb)/(1+exp(Xb))
p
```

```
## [1] 0.7118488
```

So, the estimate the probability that an iris with Petal.Length = 1.2cm and Petal.Width=2.2cm, Sepal.Length=2.2cm and Sepal.Width=0.5cm belongs to the versicolor specie is

$$\hat{p}_i = \frac{e^{7.3785-0.2454\times 2.2-2.7966\times 0.5+1.3136\times 1.2-2.7783\times 2.2}}{1+e^{7.3785-0.2454\times 2.2-2.7966\times 0.5+1.3136\times 1.2-2.7783\times 2.2}} = 0.7118$$

(b)

According to the logistic regression model, the log odds is given by

$$\ln\left(\frac{p_i}{1-p_i}\right) = (X\beta)_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and the odds is calculated as

$$\frac{p_i}{1-p_i} = e^{(X\beta)_i} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4}$$

```
Xb = 7.3785-0.2454*2.2 -2.7966*0.5 + 1.3136*1.2 -2.7783*2.2
Xb
```

```
## [1] 0.90438
```

```
exp(Xb)
```

```
## [1] 2.4704
```

Therefore, the estimated odds that an iris with Petal.Length = 1.2cm, Petal.Width=2.2cm, Sepal.Length=2.2cm and Sepal.Width=0.5cm belongs to the versicolor specie is

$$\frac{\hat{p}_i}{1-\hat{p}_i} = e^{0.90438} = 2.47$$

(c)

There is no evidence at 5% level (p=0.705634 > 0.05) that the predictor Sepal.Length is significant.

(d)

We look at the output of the likelihood ratio test, which compares the two models:

```
m1 <- glm(versicolor ~ Sepal.Width + Petal.Width, data = iris2, family = binomial)
m <- glm(versicolor ~ Sepal.Length + Sepal.Width + Petal.Length+ Petal.Width , data = iris2,
family = binomial)
anova(m1,m,test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: versicolor ~ Sepal.Width + Petal.Width
## Model 2: versicolor ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       147     151.67
## 2       145     145.07  2   6.5987  0.03691 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The test statistic is Diviance= 151.67 - 145.07 = 6.5987 with p-value = 0.03691 < 0.05. We reject the null hypothesis at 5% level that a model with only Sepal.Width and Petal.Width is better than a full model with all the predictors Sepal.Length, Sepal.Width, Petal.Length and Petal.Width.

Instead of reading the p-value from the output, we can compare the observed value of the deviance D = 151.67 - 145.07 = 6.5987 to the upper 5% point of the χ^2_2 distribution, which can be computed by

```
qchisq(0.95,2)
```

```
## [1] 5.991465
```

Therefore, we reject H_0 under a 5% level (6.5987 > 5.991465) and conclude that the model with all the predictors is preferred.