Sample solutions

Stat 8051 — Homework 3

Problem 1: ALR Exercise 6.7

First create the new variables and build the two models:

6.7.1

```
data(fuel2001)
fuel2001$Dlic <- 1000*fuel2001$Drivers/fuel2001$Pop
fuel2001$Fuel <- 1000*fuel2001$FuelC/fuel2001$Pop
fuel2001$Income <- fuel2001$Income/1000
fuel2001$logMiles <- log(fuel2001$Miles,2)

m0 = lm(Fuel~Tax+Dlic+Income+logMiles, data=fuel2001)
m1 = lm(Fuel~logMiles+Income+Dlic+Tax, data=fuel2001)
```

6.7.1 The type-I ANOVAs are as follows:

```
> anova(m0)
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value    Pr(>F)
Tax        1  26635   26635  6.3254 0.0154602 *
Dlic       1  79378   79378 18.8506 7.692e-05 ***
Income     1  61408   61408 14.5833 0.0003997 ***
logMiles   1  34573   34573  8.2104 0.0062592 **
Residuals 46 193700    4211
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> anova(m1)
Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value    Pr(>F)
logMiles   1  70478   70478 16.7371 0.0001711 ***
Income     1  49996   49996 11.8731 0.0012264 **
Dlic       1  63256   63256 15.0221 0.0003353 ***
```

```
Tax        1  18264   18264  4.3373 0.0428733 *
Residuals 46 193700    4211
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

6.7.2

```
> Anova(m0, type=2)
Anova Table (Type II tests)

Response: Fuel
          Sum Sq Df F value    Pr(>F)
Tax        18264  1  4.3373 0.0428733 *
Dlic       56770  1 13.4819 0.0006256 ***
Income     32940  1  7.8225 0.0075078 **
logMiles   34573  1  8.2104 0.0062592 **
Residuals 193700 46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> Anova(m1, type=2)
Anova Table (Type II tests)

Response: Fuel
          Sum Sq Df F value    Pr(>F)
logMiles   34573  1  8.2104 0.0062592 **
Income     32940  1  7.8225 0.0075078 **
Dlic       56770  1 13.4819 0.0006256 ***
Tax        18264  1  4.3373 0.0428733 *
Residuals 193700 46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Statistics about the last coefficient term in type-II ANOVA is the same as that in type-I ANOVA table of the same model.

Problem 2: ALR Exercise 6.8

Under the setup of multiple linear regression, the total sum of squares can be divided into 3 parts:

$$SS_{Total} = SS_{\bar{y}} + SS_{Reg} + RSS_{AH}$$

where $SS_{\bar{y}}, SS_{Reg}$ are sum of squares due to the intercept term and regression coefficients, respectively. Also, when testing for coefficient of predictors all being 0, we have $SYY = RSS_{NH} = SS_{Reg} + RSS_{AH} \Rightarrow SS_{Reg} = RSS_{NH} - RSS_{AH}$. Finally, $df_{NH} = n - (p -$

$p'), df_{AH} = n - p'$. Thus we have that

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}} = \left(\frac{n-p'}{p}\right)\frac{SS_{Reg}}{RSS_{AH}} = \left(\frac{n-p'}{p}\right)\frac{R^2}{1-R^2}$$

dividing both sides by SYY and putting $R^2 = SS_{Reg}/SYY$.

Problem 3: ALR Exercise 6.9

```
> m1 <- lm(Y ~ X1 + I(X1^2) + X2 + I(X2^2) + X1:X2, cakes)
> m1 = lm(Y~(X1+X2)^2 + I(X1^2) + I(X2^2), cakes)
> m2 <- update(m1, ~ . - X1:X2)
> m3 <- update(m1, ~ . - I(X1^2))
> m4 <- update(m1, ~ . - X1 - I(X1^2) - X1:X2)
```

Checking for $H_0: \boldsymbol{\beta}_5 = 0$ vs. $H_a: \boldsymbol{\beta}_5 \neq 0$:

```
> anova(m2, m1)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + I(X1^2) + I(X2^2)
Model 2: Y ~ (X1 + X2)^2 + I(X1^2) + I(X2^2)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      9 4.2430
2      8 1.4707  1    2.7722 15.079 0.004654 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Checking for $H_0: \boldsymbol{\beta}_2 = 0$ vs. $H_a: \boldsymbol{\beta}_2 \neq 0$:

```
> anova(m3, m1)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + I(X2^2) + X1:X2
Model 2: Y ~ (X1 + X2)^2 + I(X1^2) + I(X2^2)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      9 4.3785
2      8 1.4707  1    2.9077 15.816 0.004079 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Checking for $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_5 = 0$ vs. H_a: not all 0:

```
> anova(m4, m1)
Analysis of Variance Table
```

```
Model 1: Y ~ X2 + I(X2^2)
Model 2: Y ~ (X1 + X2)^2 + I(X1^2) + I(X2^2)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     11 11.4739
2      8  1.4707  3    10.003 18.137 0.0006293 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Thus all the null hypotheses in the question are rejected at 95% confidence level.

Problem 4: ALR Exercise 6.14

6.14.1

```
> A = lm(log(acrePrice)~year, data=MinnLand)
> summary(A)

Call:
lm(formula = log(acrePrice) ~ year, data = MinnLand)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1008 -0.3773  0.1285  0.4365  2.2624

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.939e+02  3.984e+00  -48.67   <2e-16 ***
year         1.005e-01  1.985e-03   50.60   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6808 on 18698 degrees of freedom
Multiple R-squared:  0.1204,Adjusted R-squared:  0.1204
F-statistic:  2560 on 1 and 18698 DF,  p-value: < 2.2e-16
```

The average change in mean log acre-price is 0.1 per year, which means that median change in acre-price is $\exp(0.1) = 1.1$ dollars per year.

6.14.2

```
> MinnLand$fyear = factor(paste(MinnLand$year))
> B = lm(log(acrePrice)~fyear, data=MinnLand)
> summary(B)

Call:
```

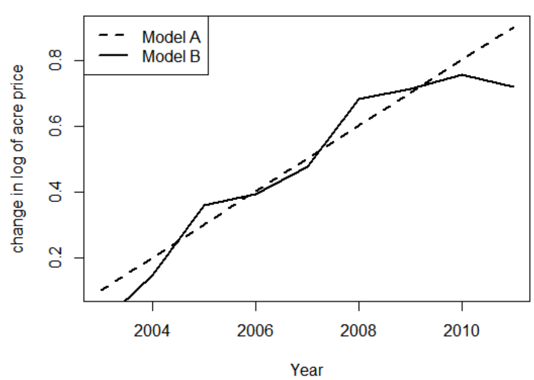
```
lm(formula = log(acrePrice) ~ fyear, data = MinnLand)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9499 -0.3785  0.1301  0.4354  2.3456 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.27175    0.02848 255.345  < 2e-16 ***
fyear2003   -0.00155    0.03207  -0.048    0.961    
fyear2004    0.14794    0.03155   4.689 2.76e-06 ***
fyear2005    0.36026    0.03176  11.343  < 2e-16 ***
fyear2006    0.39392    0.03195  12.329  < 2e-16 ***
fyear2007    0.47682    0.03186  14.965  < 2e-16 ***
fyear2008    0.68364    0.03162  21.620  < 2e-16 ***
fyear2009    0.71407    0.03355  21.284  < 2e-16 ***
fyear2010    0.75733    0.03260  23.231  < 2e-16 ***
fyear2011    0.72071    0.03526  20.437  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6775 on 18690 degrees of freedom
Multiple R-squared:  0.1293,Adjusted R-squared:  0.1289 
F-statistic: 308.5 on 9 and 18690 DF,  p-value: < 2.2e-16
```

The change per year is no longer linear, and results in lack of fit. This is demonestrated in the plot below:



6.14.3 If the year-specific coefficients of B increase linearly, then it becomes nothing but model A.

6.14.4

```
> anova(A,B)
Analysis of Variance Table

Model 1: log(acrePrice) ~ year
Model 2: log(acrePrice) ~ fyear
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  18698 8666.9
2  18690 8579.2  8    87.686 23.878 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

This demonstrates significant lack of fit.

Alternate method (requires `alr3` package)

```
> library(alr3); pureErrorAnova(A)
Analysis of Variance Table

Response: log(acrePrice)
                Df Sum Sq Mean Sq  F value    Pr(>F)
year             1 1186.8 1186.77 2585.395 < 2.2e-16 ***
Residuals    18698 8666.9    0.46
 Lack of fit     8   87.7   10.96   23.878 < 2.2e-16 ***
 Pure Error  18690 8579.2    0.46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Problem 5: ALR Exercise 8.2

8.2.1

```
> library(MASS)
> par(mfrow=c(1,2))
> z = boxcox(lm(Distance~Speed, data=stopping))
> z$x[which.max(z$y)] # power at which log-likelihood is maximized
[1] 0.4242424
>
> invResPlot(lm(Distance~Speed, data=stopping))
      lambda       RSS
1  0.4849737  4463.944
2 -1.0000000 33149.061
3  0.0000000  7890.434
4  1.0000000  7293.835
> par(mfrow=c(1,1))
```

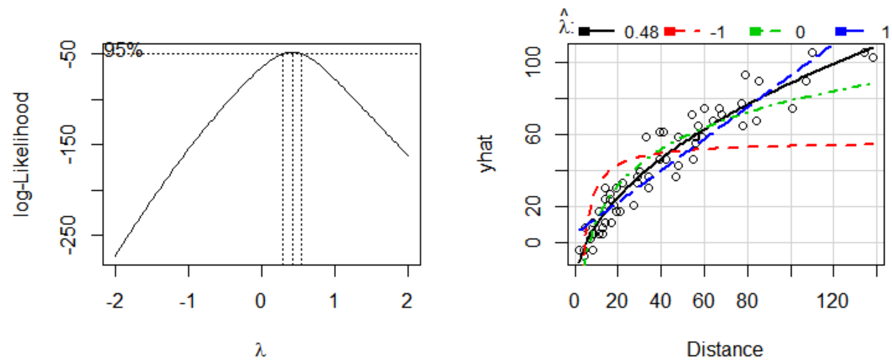


Figure 1: (L) Box-cox plot and (R) inverse response plot for optimal power transformation on `Distance`

Box-cox transformation suggests raising to a power of $\hat{\lambda} = 0.424$, while inverse response plot suggests $\hat{\lambda} = 0.48$. Hence for practical purposes, it makes sense to doa a square-root transformation on the response.

8.2.2 See figure 2.

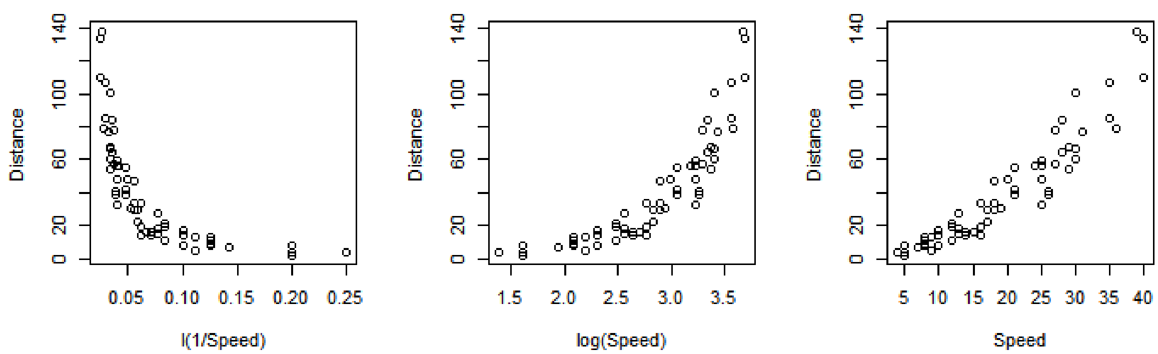


Figure 2: (L) Plot of λ-power transformed predictor vs. response. $\lambda = -1, 0, 1$ left to right

8.2.3 A power transformation of $\lambda = 2$ on the predictor kind of makes the plot more linear, but there is non-constant variance that has to be taken care of (Figure 3).

8.2.4

```
> reg1 = lm(Distance~Speed+I(Speed^2),
+             weight=1/Speed^2, data=stopping)
> reg2 = lm(Distance^.5~Speed, data=stopping)
```

```
>
> plot(Distance~Speed, data=stopping)
> lines(fitted.values(reg1)~stopping$Speed,
+       type = "l",lwd=2)
> lines((fitted.values(reg2))^2~stopping$Speed,
+       type = "l",lwd=2, lty=2)
```

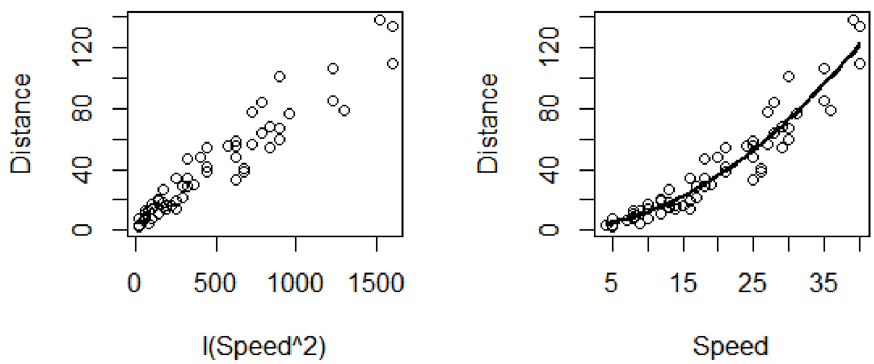


Figure 3: (Left) Plot after transformation with $\lambda = 2$, (Right) Plot of fitted regressions in 8.2.4

As we can see, the plots of the two fitted mean functions are very similar.

Problem 6: ALR Exercise 9.11

We are given the $\hat{e}_i$'s and h_{ii}'s. From there D_i and t_i's are calculated as follows:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \quad t_i = r_i\sqrt{\frac{n-p'-1}{n-p'-r_i^2}}, \quad D_i = \frac{r_i^2}{p'}\frac{h_{ii}}{1-h_{ii}}$$

Here we are testing for one outlier and df of $\hat{\sigma} = 46$, so $n - p' - 1 = 45 \Rightarrow p' = 5$. Thus we finally get the following:

```
> Fuel = c(514.279, 374.164, 426.349, 842.792, 317.492)
> ehat = c(-163.145, -137.599, -102.409, -183.499, -49.452)
> h = c(.256, .162, .206, .084, .415)
> sighat = 64.891
> n = 51; p1 = 5
> r = ehat/(sighat*sqrt(1-h))
> t = r*sqrt((n-p1-1)/(n-p1-r^2))
> D = r^2*h/(p1*(1-h))
> d = (data.frame(cbind(r,t,D),
+                 row.names = c("Alaska", "NY", "Hawaii", "Wyoming", "Dist. Col")))
> d
```

```
                    r          t         D
Alaska    -2.9147602 -3.1927822 0.5846591
NY        -2.3163746 -2.4376317 0.2074525
Hawaii    -1.7711013 -1.8147106 0.1627659
Wyoming   -2.9546191 -3.2465847 0.1601094
Dist. Col -0.9963719 -0.9962917 0.1408527
> pmin(n*2*pt(-abs(t),46),1)
[1] 0.1296661 0.9541017 1.0000000 0.1112947 1.0000000
```

The largest outlier test statistic is 3.2466, and the Bonferroni adjusted p-values mean that none of the five data points can be declared outliers. Since the Cook's distance is largest for Alaska, this is the most influential among the five.