HW5

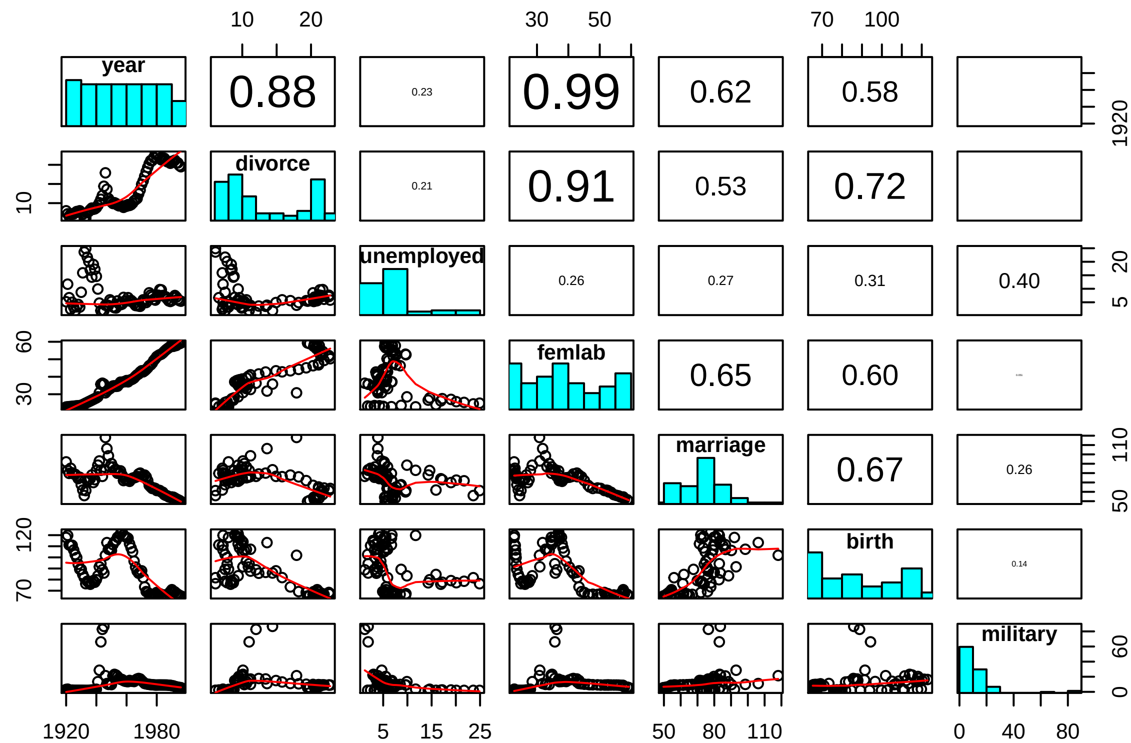
Jordan Garrett

12/4/2019

1)

Exploratory Data Analysis:

```
data(divusa)

panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(divusa, diag.panel = panel.hist, lower.panel=panel.smooth,
      upper.panel = panel.cor, cex.labels=1,font.labels = 2)
```



Step-wise Analysis:

```
step.model1 <- step(lm(divorce ~ ., data = divusa), direction = 'both')
```

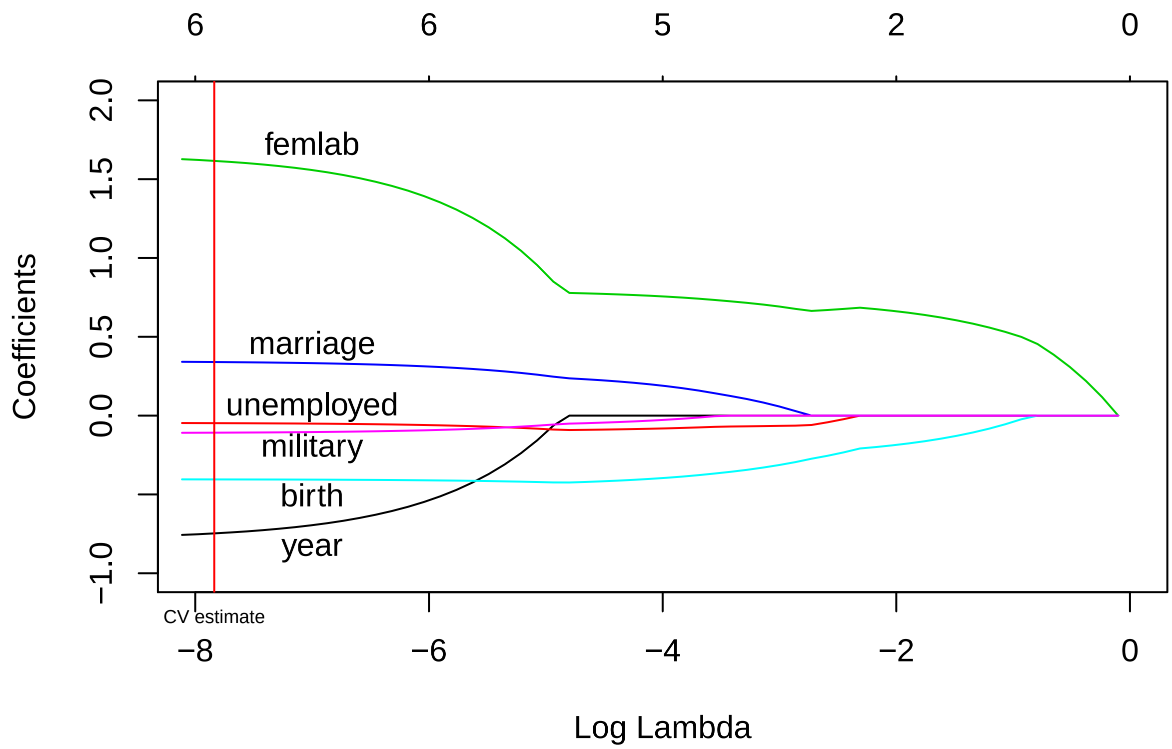
```
## Start:  AIC=70.41
## divorce ~ year + unemployed + femlab + marriage + birth + military
##
##              Df Sum of Sq    RSS     AIC
## - unemployed  1     1.925 162.12  69.330
## <none>                    160.20  70.410
## - military    1    22.231 182.43  78.417
## - year        1    33.199 193.40  82.912
## - marriage    1    90.468 250.66 102.884
## - femlab      1   113.214 273.41 109.572
## - birth       1   144.897 305.10 118.015
##
## Step:  AIC=69.33
## divorce ~ year + femlab + marriage + birth + military
##
##              Df Sum of Sq    RSS     AIC
## <none>                    162.12  69.330
## + unemployed  1     1.925 160.20  70.410
## - military    1    20.957 183.08  76.691
## - year        1    42.054 204.18  85.089
## - marriage    1   126.643 288.77 111.779
## - femlab      1   158.003 320.13 119.718
## - birth       1   172.826 334.95 123.203
```

LASSO:

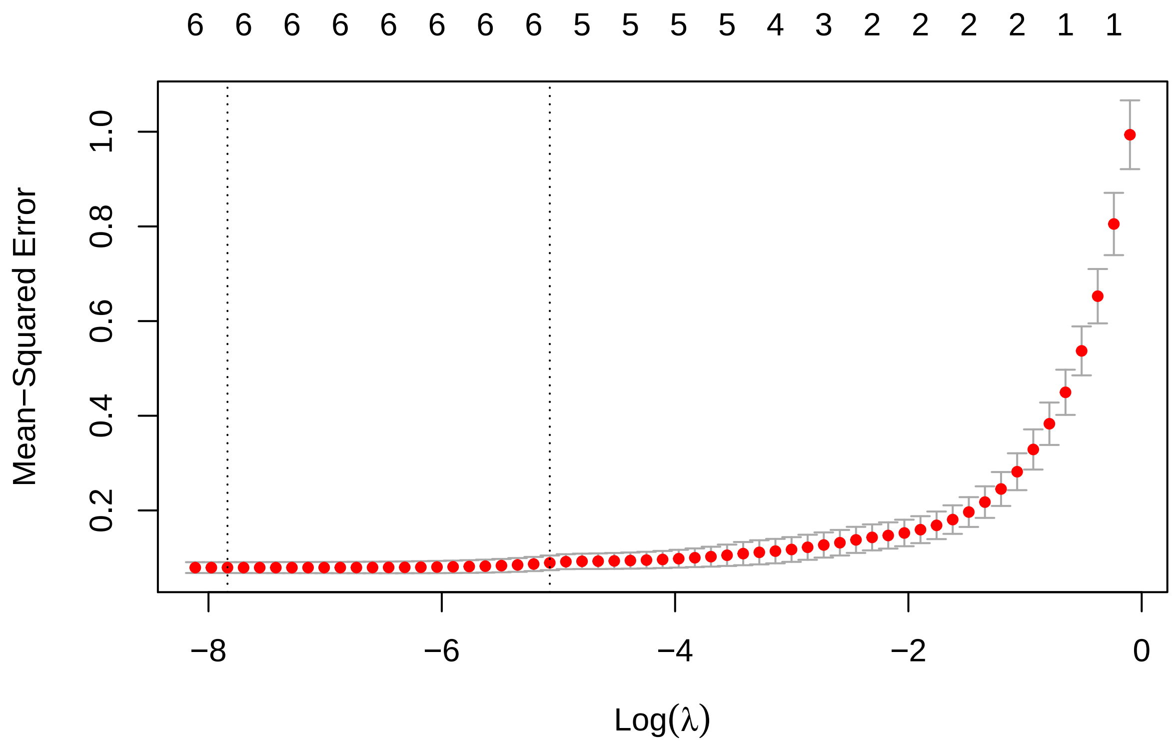
```r
#center and scale to mean of 0 and sd of 1
scaled.df1 <- as.data.frame(scale(divusa))
summary(scaled.fit <- lm(divorce ~ . , data = scaled.df1))
```

```
##
## Call:
## lm(formula = divorce ~ ., data = scaled.df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51309 -0.16250 -0.01649  0.13136  0.61190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.924e-16  3.041e-02   0.000 1.000000
## year        -8.016e-01  2.104e-01  -3.809 0.000297 ***
## unemployed  -4.421e-02  4.820e-02  -0.917 0.362171
## femlab       1.677e+00  2.384e-01   7.033 1.09e-09 ***
## marriage     3.468e-01  5.515e-02   6.287 2.42e-08 ***
## birth       -4.027e-01  5.061e-02  -7.957 2.19e-11 ***
## military    -1.120e-01  3.595e-02  -3.117 0.002652 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2668 on 70 degrees of freedom
## Multiple R-squared:  0.9344, Adjusted R-squared:  0.9288
## F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

```r
fit.1 <- lm(divorce ~ ., data = divusa)

set.seed(800)
X <- model.matrix(scaled.fit)[,-1]
fit.lasso1 <- glmnet(X, scaled.df1$divorce, lambda.min=0, nlambda=101, alpha=1)
plot(fit.lasso1, xvar="lambda", xlim=c(-8,0), ylim = c(-1,2))
text(-7,coef(fit.lasso1)[c(3:5),length(fit.lasso1$lambda)]+0.1,labels=colnames(X)[2:4])
text(-7,coef(fit.lasso1)[c(2,6,7),length(fit.lasso1$lambda)]-0.1,
     labels=colnames(X)[c(1,5,6)])
fit.lasso.cv1 <- cv.glmnet(X, scaled.df1$divorce, lambda.min=0, nlambda=101)
abline(v=log(fit.lasso.cv1$lambda.min), col="red")
mtext("CV estimate", side=1, at=log(fit.lasso.cv1$lambda.min), cex=.6)
```




```r
plot(fit.lasso.cv1)
```



Mallows Cp:

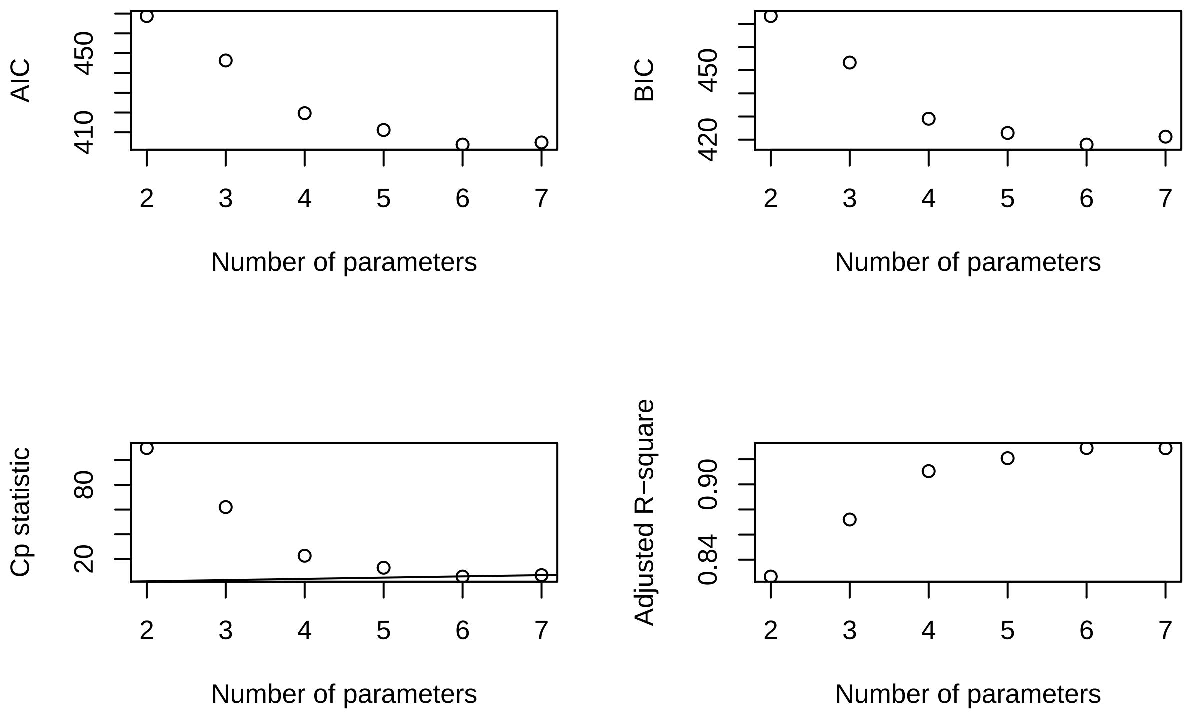
```
a <- regsubsets(divorce ~ ., data=divusa,method="exhaustive")
rs <- summary(a)
```

Cross Validation:

```
CV_error <- cv.glm(divusa, glm(divorce ~ unemployed + femlab +
                                marriage + birth, data=divusa), K = 10)$delta
```

Plotting measures of model quality for seleciton:

```
n <- nrow(divusa)
AIC <- n*log(rs$rss) + 2*(2:7)
BIC <- n*log(rs$rss) + log(n)*(2:7)
par(mfrow=c(2,2))
plot(2:7,AIC,xlab="Number of parameters",ylab="AIC")
plot(2:7,BIC,xlab="Number of parameters",ylab="BIC")
plot(2:7,rs$cp,xlab="Number of parameters",ylab="Cp statistic")
abline(0,1)
plot(2:7, rs$adjr2, xlab="Number of parameters", ylab="Adjusted R-square")
```



The following model was selected based off of the quesiton's selection criterion:

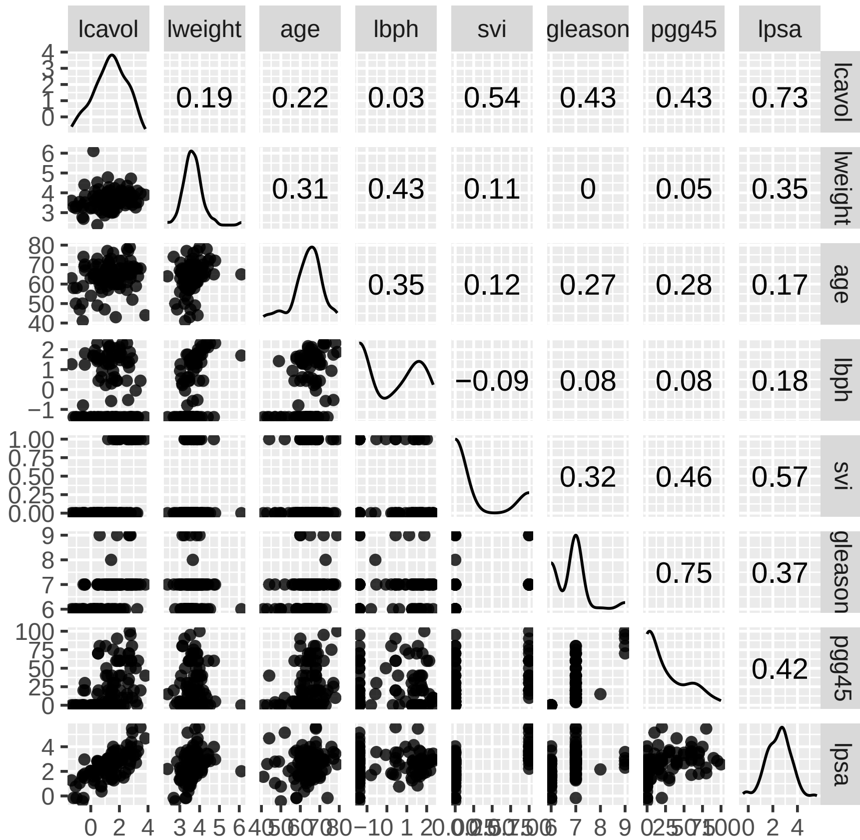
$$divorce_i = -0.22(year_i) + 0.85(femlab_i) + 0.16(marriage_i) - 0.11(birth_i) - 0.04(military_i)$$

This model had the lowest AIC score (289.85), yeilded the smallest adjusted cross-validation error (3.08), explained the highest amount of variation $Adj.R^2 = 0.93, F_{5,71} = 199.7, p < 2.2e-16$, had the smallest Mallow's cp statistic (5.84), and each of its regression coefficients were signifcant for a $\alpha = 0.001$ threshold (except for miliarty, which was significant at $p < 0.01$). It is worth mentioning that if the response variable was tranformed to $divorce^{-1}$ that a much better model is converged upon, but it was not chosen since the quality of model diagnostics was not included in the selection criteria.

2)

```
data("prostate")
```

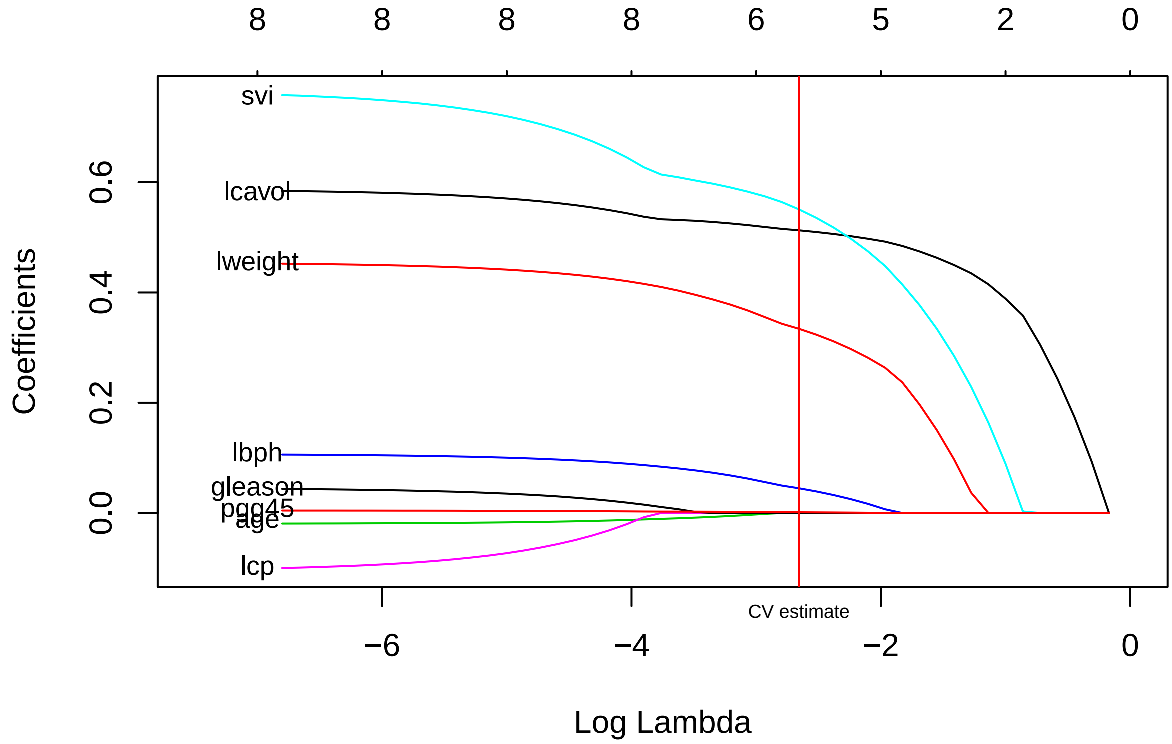
```r
ggscatmat(prostate, columns = -6, alpha=0.8)
```



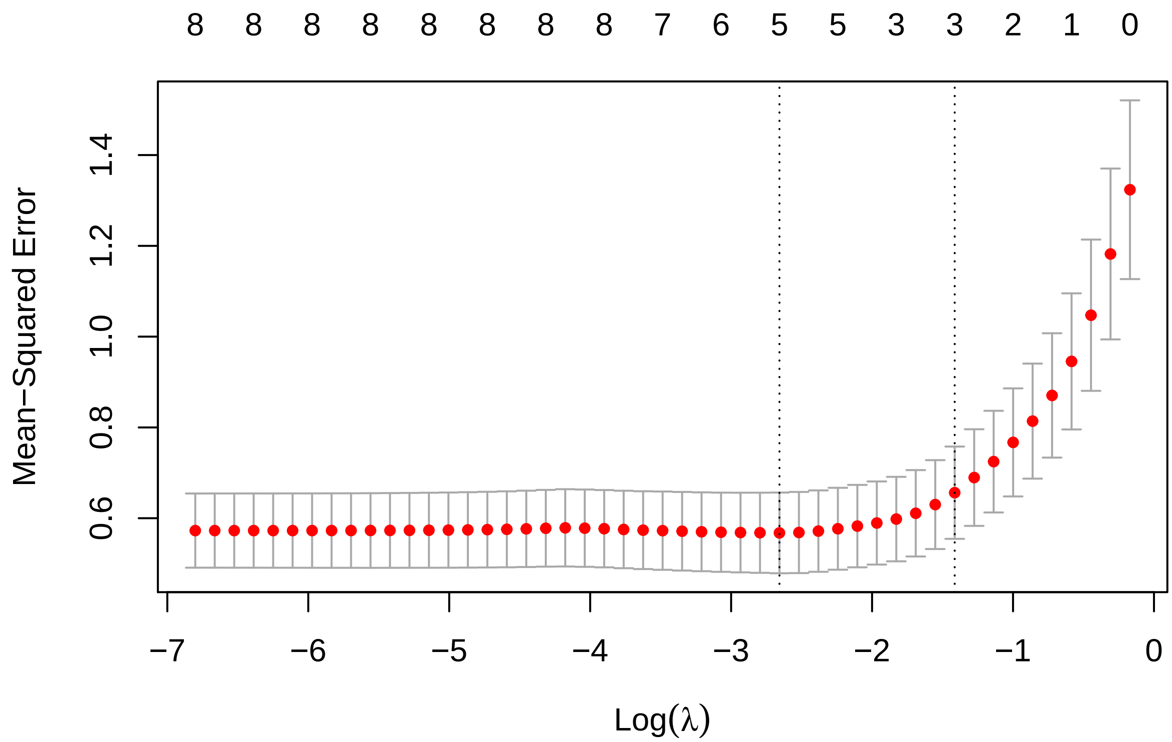

```r
prostate_fit <- lm(lpsa ~ ., data = prostate)
```

Some covariates are fairly correlated with one another, so LASSO regression was implemented.

```r
set.seed(800)
X <- model.matrix(prostate_fit)[,-1]
fit.lasso <- glmnet(X, prostate$lpsa, lambda.min=0, nlambda=101, alpha=1)
plot(fit.lasso, xvar="lambda", xlim=c(-7.5,0))
text(-7,coef(fit.lasso)[-1,length(fit.lasso$lambda)],labels=colnames(X),cex=0.8)
fit.lasso.cv <- cv.glmnet(X, prostate$lpsa, lambda.min=0, nlambda=101)
abline(v=log(fit.lasso.cv$lambda.min), col="red")
mtext("CV estimate", side=1, at=log(fit.lasso.cv$lambda.min), cex=.6)
```




```r
plot(fit.lasso.cv)
```




```r
step.model_prostate <- step(prostate_fit, direction = "both")
```

```
## Start:  AIC=-58.32
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -60.231
## - pgg45    1    0.5258 44.689 -59.174
## - lcp      1    0.6740 44.837 -58.853
## <none>                 44.163 -58.322
## - age      1    1.5503 45.713 -56.975
## - lbph     1    1.6835 45.847 -56.693
## - lweight  1    3.5861 47.749 -52.749
## - svi      1    4.9355 49.099 -50.046
## - lcavol   1   22.3721 66.535 -20.567
##
## Step:  AIC=-60.23
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -60.789
## <none>                 44.204 -60.231
## - pgg45    1    1.1920 45.396 -59.650
## - age      1    1.5166 45.721 -58.959
```

```
## - lbph     1    1.7053 45.910 -58.560
## + gleason  1    0.0412 44.163 -58.322
## - lweight  1    3.5462 47.750 -54.746
## - svi      1    4.8984 49.103 -52.037
## - lcavol   1   23.5039 67.708 -20.872
##
## Step:  AIC=-60.79
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -61.374
## <none>                 44.867 -60.789
## + lcp      1    0.6623 44.204 -60.231
## - age      1    1.2649 46.131 -60.092
## - lbph     1    1.6465 46.513 -59.293
## + gleason  1    0.0296 44.837 -58.853
## - lweight  1    3.5647 48.431 -55.373
## - svi      1    4.2503 49.117 -54.009
## - lcavol   1   25.4189 70.285 -19.248
##
## Step:  AIC=-61.37
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                 45.526 -61.374
## - age      1    0.9592 46.485 -61.352
## + pgg45    1    0.6590 44.867 -60.789
## + gleason  1    0.4560 45.070 -60.351
## + lcp      1    0.1293 45.396 -59.650
## - lbph     1    1.8568 47.382 -59.497
## - lweight  1    3.2251 48.751 -56.735
## - svi      1    5.9517 51.477 -51.456
## - lcavol   1   28.7665 74.292 -15.871
```

```r
summary(step.model_prostate)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

3)

```r
murder <- read.csv("/Users/owner/Desktop/PSTAT_220A/HW5_3.txt",sep=" ", header=T)
murder <- murder[-5]
```

```r
summary(murder.fit <- lm(y ~ x2 + x3, data = murder))
```

```
##
## Call:
## lm(formula = y ~ x2 + x3, data = murder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9019 -2.8101  0.1569  1.7788 10.2709
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.0725     6.7265  -5.065 9.56e-05 ***
## x2            1.2239     0.5682   2.154   0.0459 *
## x3            4.3989     1.5262   2.882   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.648 on 17 degrees of freedom
## Multiple R-squared:  0.802,  Adjusted R-squared:  0.7787
## F-statistic: 34.43 on 2 and 17 DF,  p-value: 1.051e-06
```

```r
new_point <- data.frame("x1" = 150,000, "x3"=10, "x2"=9)
p2 <- predict(murder.fit, newdata=new_point, se=T,interval="prediction")
p2
```

```
## $fit
##       fit       lwr      upr
## 1 20.9322 -2.950571 44.81497
##
## $se.fit
## [1] 10.32135
##
## $df
## [1] 17
##
## $residual.scale
## [1] 4.648482
```

4)

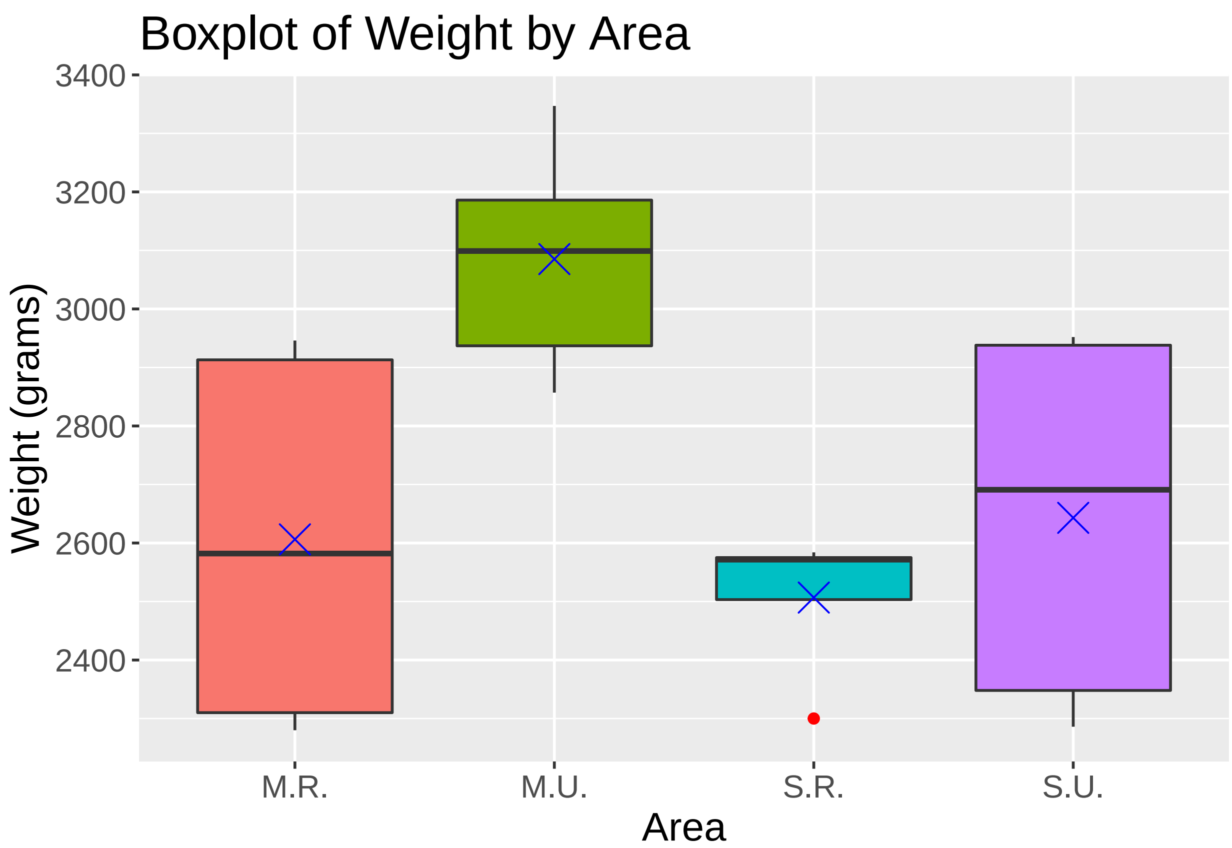
```
birth.weight <- data.frame("M.R."=c(2582,2946,2280,2913,2310),
                           "M.U."=c(3347,3099,3186,2857,2937),
                           "S.R."=c(2572,2571,2300,2584, NA),
                           "S.U."=c(2952,2348,2286,2691,2938))

birth.weight.an <- na.omit(melt(birth.weight))
```

```
## Using  as id variables
```

```
colnames(birth.weight.an) <- c("Area","Weight")
birth.weight.an$Area <- factor(birth.weight.an$Area,levels = c("M.R.","M.U.",
                                                                "S.R.","S.U."))
birth.weight.an$ID <- c(1:nrow(birth.weight.an))
```

a)

```
ggplot(birth.weight.an, aes(x = Area, y = Weight, fill = Area)) +
  geom_boxplot(outlier.color = 'red')  + labs(title="Boxplot of Weight by Area",
        x ="Area", y = "Weight (grams)") +
  stat_summary(fun.y=mean, fun.args = c(trim = 0, na.rm=F),
               geom="point", shape=4, size=5, color="blue") +
  theme(legend.position = "none", text = element_text(size = 15))
```



I anticipate that there will be some differences in the means (represented by a blue "X") between the birth weight of babies in each of the areas shown in the box plot. The mean and spread of the distribution for babies born in the Urban Midwest area seems to be the most likely to be significantly different from all other locations.

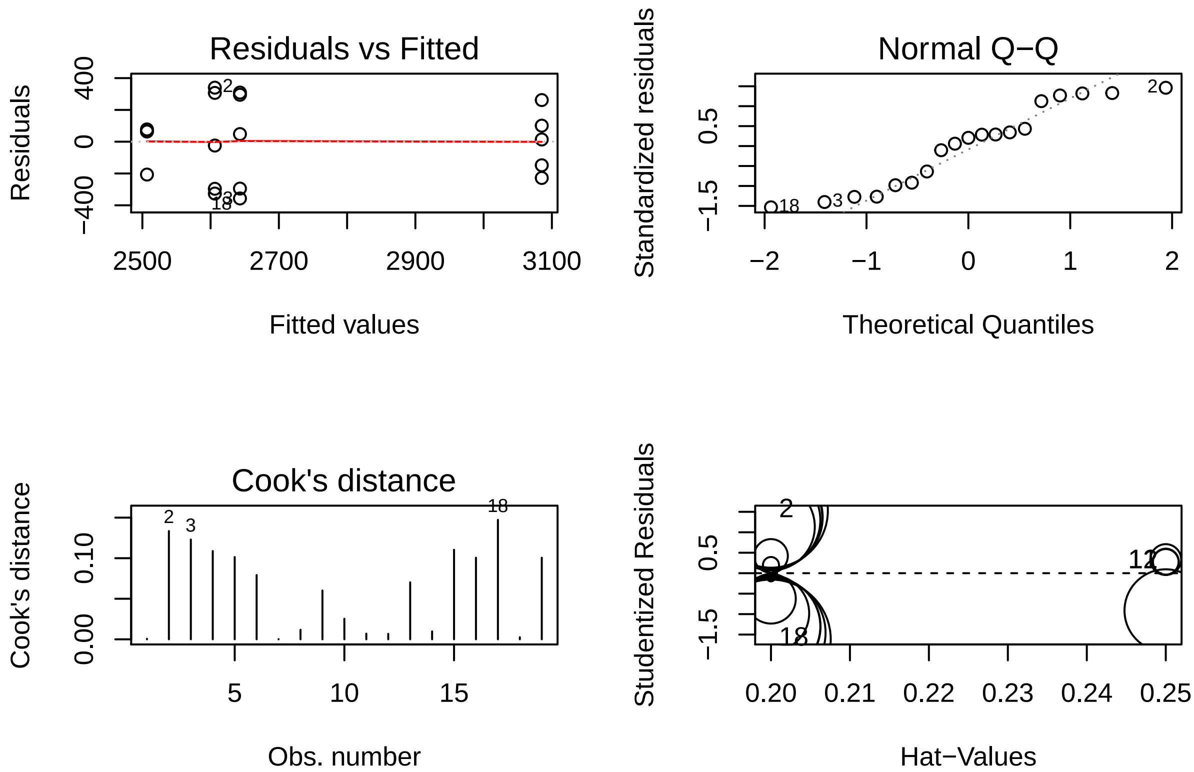
b)

```r
summary(birth_fit1 <- aov(Weight ~ Area, data = birth.weight.an))
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## Area         3  943136  314379   4.652 0.0172 *
## Residuals   15 1013628   67575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant main effect of area.

```r
par(mfrow=c(2,2))
plot(birth_fit1, which = c(1,2,4))
influencePlot(birth_fit1)
```




```
##         StudRes  Hat        CookD
## 2   1.5246073 0.20 0.133490171
## 11  0.2807980 0.25 0.007000532
## 12  0.2764709 0.25 0.006787600
## 18 -1.6157631 0.20 0.147346211
```

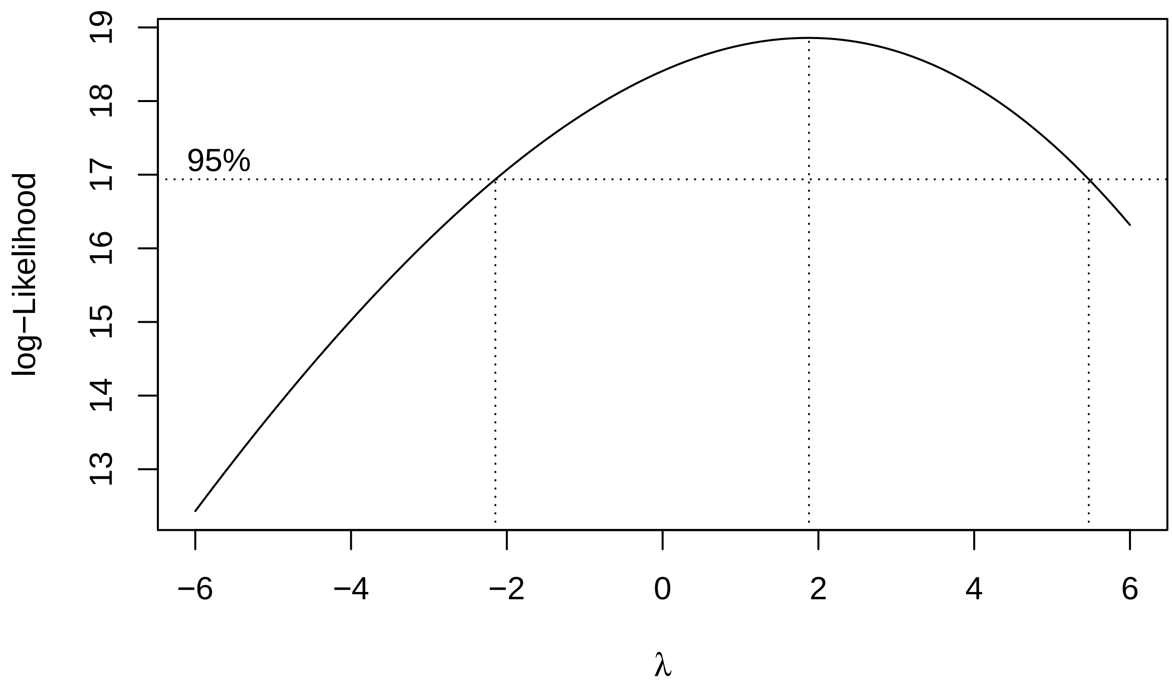
Distribution of standardized residuals appears as though it might be non-normal. Further, some of the residuals appear to have a high Cook's distance and leverage. A Box-Cox test was conducted to dtermine if any transformations were needed, and an outlier test was performed to detect highly influential points.

```r
outlierTest(birth_fit1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 18 -1.615763            0.12845           NA
```

No outliers detected in the model's residualsd under a Bonferroi corrected threshold of $\alpha = 0.05$.

```r
boxcox(birth_fit1, lambda = seq(-6,6,.2))
```



Box Cox 95% CI includes 1, so no tranformation necessary.

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: TH.data

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

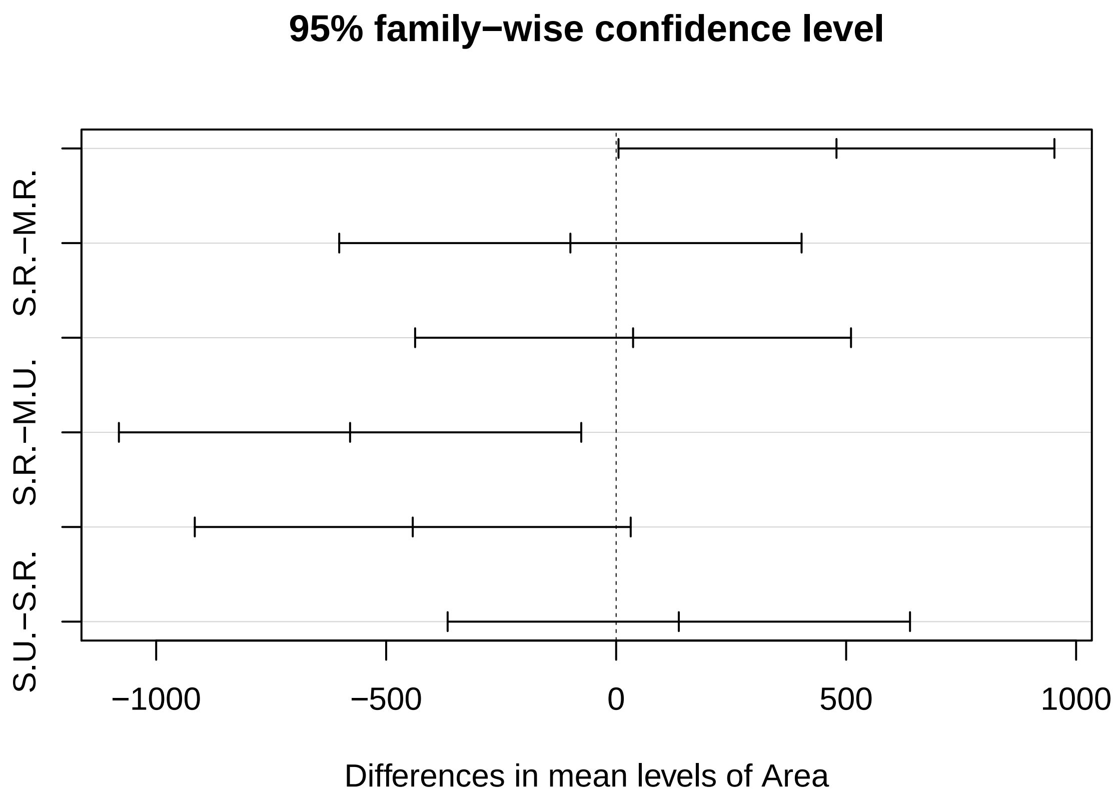
```
birth.bonCI <- PostHocTest(birth_fit1,method = "bonferroni")
birth.bonExtrac <- unite(as.data.frame(round(cbind(birth.bonCI[["Area"]][,2],
                                                    birth.bonCI[["Area"]][,3]),2)),
                         "CI",sep = " , ")$CI
birth.scheffeCI <- PostHocTest(birth_fit1,method = "scheffe")
birth.scheffeExtrac <- unite(as.data.frame(round(cbind(birth.scheffeCI[["Area"]][,2],
                                                    birth.scheffeCI[["Area"]][,3]),2)),
                             "CI",sep = " , ")$CI

birth.tukCI <-TukeyHSD(birth_fit1)
```

```r
birth.tukExtrac <- unite(as.data.frame(round(cbind(birth.tukCI[["Area"]][,2],
                                                    birth.tukCI[["Area"]][,3]),2)),
                              "CI",sep = " , ")$CI

birth.CIs <-  data.frame("Mean Diff" = birth.bonCI[['Area']][,1],
                         "Bonferonni" = birth.bonExtrac,
                         "Tukey" = birth.tukExtrac,
                         "Scheffe" = birth.scheffeExtrac,
                         "CI Level" = rep(0.95,length(birth.bonCI[['Area']][,1])))
birth.CIs
```

```
##             Mean.Diff         Bonferonni             Tukey           Scheffe
## M.U.-M.R.    479.00   -20.19 , 978.19     5.15 , 952.85   -37.31 , 995.31
## S.R.-M.R.    -99.45  -628.92 , 430.02  -602.04 , 403.14  -647.08 , 448.18
## S.U.-M.R.     36.80  -462.39 , 535.99  -437.05 , 510.65  -479.51 , 553.11
## S.R.-M.U.   -578.45 -1107.92 , -48.98 -1081.04 , -75.86 -1126.08 , -30.82
## S.U.-M.U.   -442.20   -941.39 , 56.99   -916.05 , 31.65   -958.51 , 74.11
## S.U.-S.R.    136.25  -393.22 , 665.72  -366.34 , 638.84  -411.38 , 683.88
##             CI.Level
## M.U.-M.R.       0.95
## S.R.-M.R.       0.95
## S.U.-M.R.       0.95
## S.R.-M.U.       0.95
## S.U.-M.U.       0.95
## S.U.-S.R.       0.95
```

Since we have unequal n_i and we did not have any planned comparisons, then we have to compare Bonferonni, Tukey and Scheffe confidence intervals. The method that yields the smallest confidence intervals is the one we used for pairwise comparisons. For all of the possible pairwise comparisons, the Tukey method yeilded the smallest confidence intervals.

```r
plot(TukeyHSD(birth_fit1))
```




```
se <- function(x) { sd(na.omit(x))/sqrt(length(na.omit(x))) }
sum.stats <- group_by(birth.weight.an, Area) %>% summarise(mean.weight = mean(Weight),
                                                          se.weight = se(Weight))
```

Tukey pairwise comparisons indicate that the average birth weight (grams) between babies in Midwest Urban ($M = 3085.2 \pm 87.44$) areas are significantly different from those born in either Southern Rural ($M = 2506.75 \pm 68.98$) or Midwest Rural areas ($M = 2606.2 \pm 142.18$). All other pairwise comparisons with Southern Urban areas ($M = 2643 \pm 141.29$) were nonsignificant.

c)

```
UvR_comp <- contrast(lm(Weight ~ Area, data=birth.weight.an),
         list(Area=as.factor(c("M.U.","S.U."))),
         list(Area=as.factor(c("M.R.","S.R."))),
         type="average")

MvS_comp <- contrast(lm(Weight ~ Area, data=birth.weight.an),
         list(Area=as.factor(c("M.U.","M.R."))),
         list(Area=as.factor(c("S.U.","S.R."))),
         type="average")

quick.CI <- function(con,k){
```

```r
  crit.bon <- qt(df=con$df,0.05/2/k,lower.tail=F)
  con.bon.upper <- con$Contrast + crit.bon*con$SE
  con.bon.lower <- con$Contrast - crit.bon*con$SE
  con.bon.band <- paste(round(con.bon.lower,2),round(con.bon.upper,2), sep = " , ")

  g <- length(con$X) - 1
  crit_scheffe <- qf(0.05,g,con$df,lower.tail = F)
  upper_scheffe <- con$Contrast + sqrt(g*crit_scheffe)*con$SE
  lower_scheffe <- con$Contrast - sqrt(g*crit_scheffe)*con$SE
  scheffe.ci <- paste(round(lower_scheffe,2),round(upper_scheffe,2), sep = " , ")

  return(c(con.bon.band,scheffe.ci))
}

prior_comps <- data.frame("Pair"=c("Urban-Rural","Midwest-Southern"),
                          "Mean Diff"= c(UvR_comp$Contrast,MvS_comp$Contrast),
                          "Bonferonni" = c(quick.CI(UvR_comp,2)[1],quick.CI(MvS_comp,2)[1]),
                          "Scheffe"= c(quick.CI(UvR_comp,2)[2],quick.CI(MvS_comp,2)[2]))
prior_comps
```

```
##               Pair Mean.Diff      Bonferonni         Scheffe
## 1      Urban-Rural   307.625   9.26 , 605.99  -68.7 , 683.95
## 2 Midwest-Southern   270.825 -27.54 , 569.19 -105.5 , 647.15
```

I decided to use the bonferonni correction method for multiple comparisons since it resulted in a smaller confidence range. There was a significant difference between the mean weight of babies in Urban areas ($M = 2864.1 \pm 107.55$) when compared to Rural areas ($M = 2562 \pm 81.94$), whose bonferonni 95% CI was [9.26,605.99]. In contrast, there were no significant differences between the mean weight of babies in Midwestern locations ($M = 2845.7 \pm 112.09$) when compared to Southern locations ($M = 2582.44, \pm 83.14$), whose bonferonni 95% CI was [-27.54,569.19].

d)

I would use the Scheffe's method to further test the relationship between birth weights of Urban and Rural areas, since it is more conservative, thus it attenuates the potential of committing a Type I error more so than the Bonferonni method. According to Scheffe's method, the difference between the means of Urban and Rural birth weigths is nonsignificant (95% CI [-68.7,683.95]).

5)

An appropriate model for this problem would be:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where α represents the main effect of poison, β represents the main effect of treatment, i is the level of poisons ($i = 1, 2, 3$), j is the levels of treatments ($j = 1, 2, 3, 4$), $(\alpha\beta)_{ij}$ is the interaction between poison i and treament j, k is a specific observation ($k = 1, 2, ...N$) , and ϵ are random errors. The levels of poison and treatment are both fixed, thus this is a fixed-effects model. The model assumes the following:

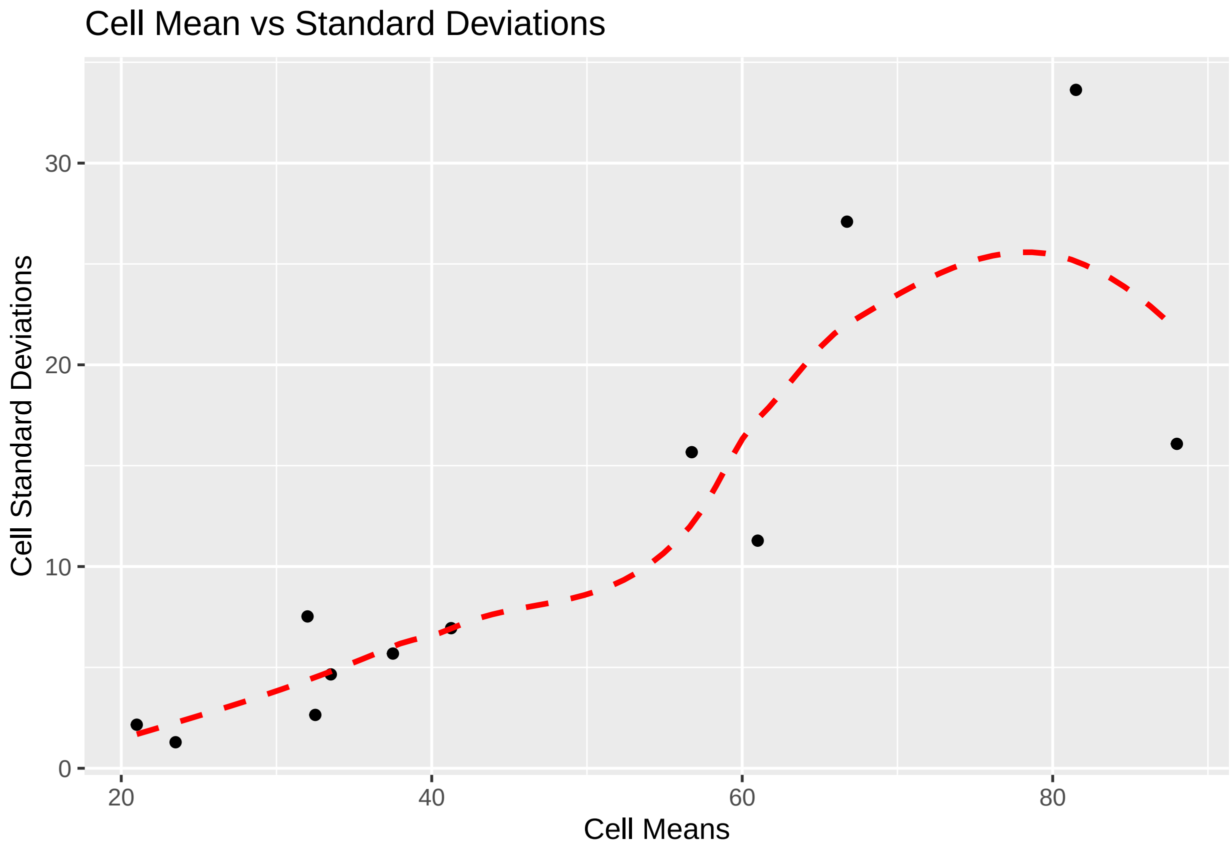
- The populations from which the samples were acquired are normal.
- Samples are independent from one another
- Constant variance of the sampled populations

b)

```
poison.df <- data.frame("SurvivalT" = c(31,45,46,43,
                                        36,29,40,23,
                                        22,21,18,23,
                                        82,110,88,72,
                                        92,61,49,124,
                                        30,37,38,29,
                                        43,45,63,76,
                                        44,35,31,40,
                                        23,25,24,22,
                                        45,71,66,62,
                                        56,102,71,38,
                                        30,36,31,33),
                        "Treatment" = c(rep(1,12),rep(2,12),rep(3,12),rep(4,12)),
                        "Poison"=rep(c(rep(1,4),rep(2,4),rep(3,4)),4))

poison.df$Treatment <- factor(poison.df$Treatment, labels= c("B1","B2","B3","B4"))
poison.df$Poison <- factor(poison.df$Poison, labels = c("A1","A2","A3"))
```

```
cell_summary <- group_by(poison.df,Poison,Treatment) %>%
  summarise(mean.SurvT = mean(SurvivalT),
            sd.SurvT = sd(SurvivalT))
```

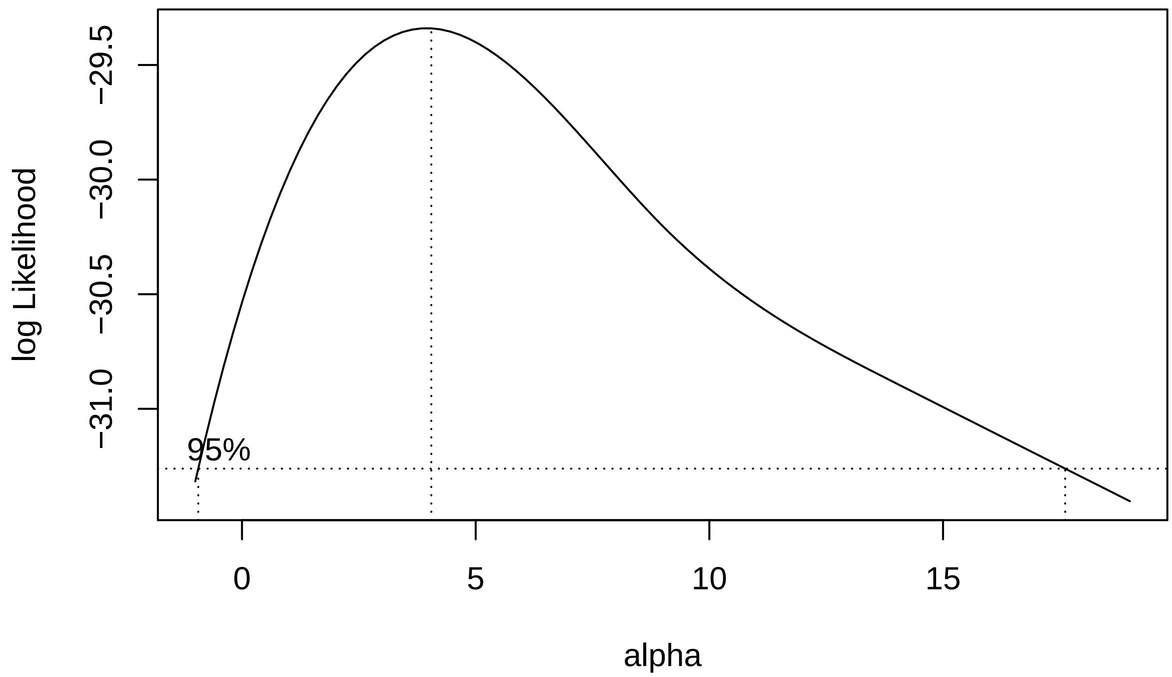
```
ggplot(cell_summary, aes(x=mean.SurvT,y=sd.SurvT)) +
  geom_point() + labs(x="Cell Means",y="Cell Standard Deviations",
                      title="Cell Mean vs Standard Deviations")+
  geom_smooth(method="loess",se=F,color="red",linetype="dashed")
```



The plot suggests that the assumption of constant variance across sampled populations is violated. Increases in cell mean correpond to increases in cell standard deviation.

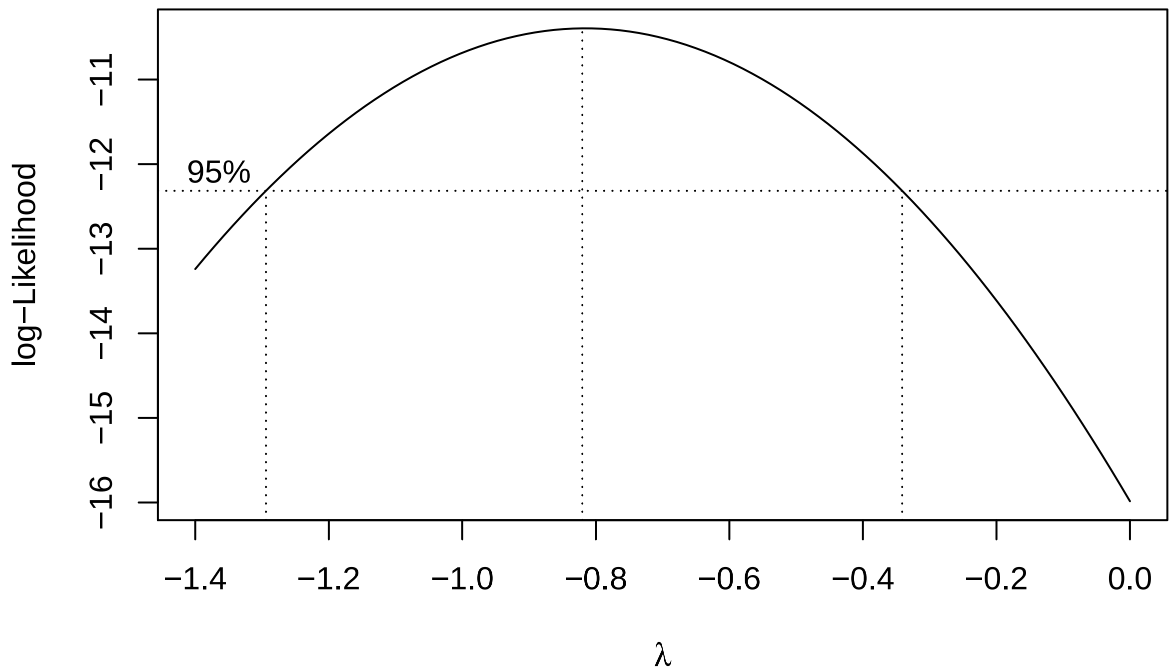
c)

```
logtrans(sd.SurvT ~ mean.SurvT, alpha = seq(-1,19,5), data = cell_summary)
```



This plot does corroborate the previously stated conclusion of non-constant variance being violated. Since 1 is within the 95% CI of alpha, then no transformation is required for the logarithmic model.

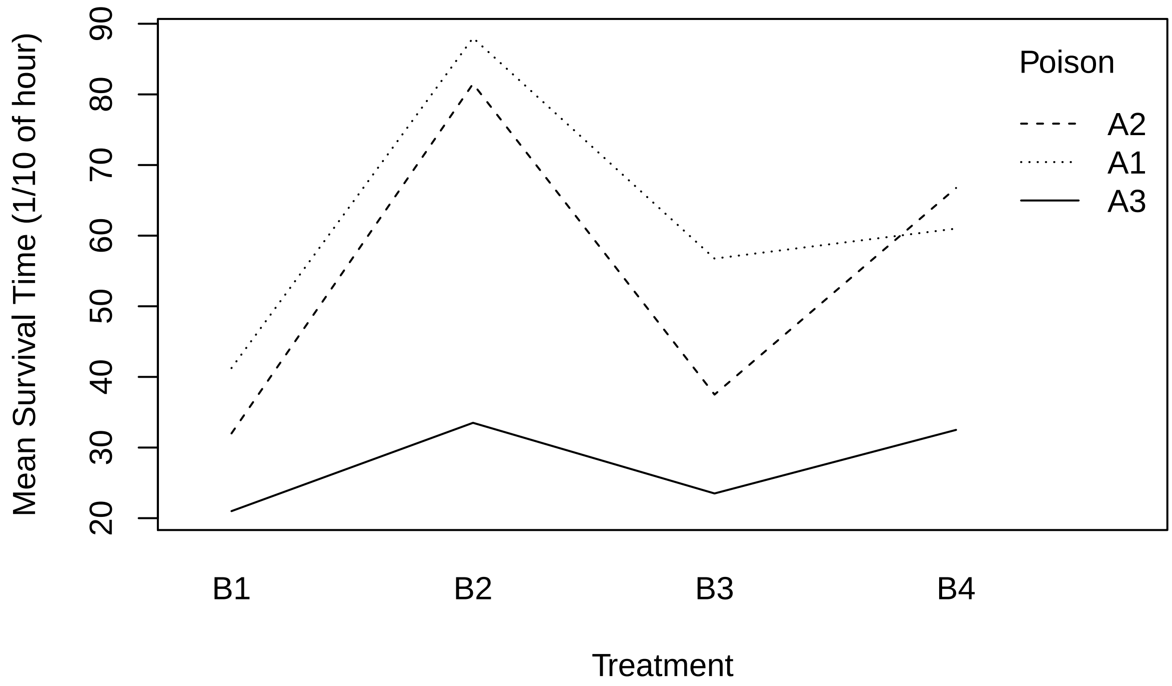
```
boxcox(SurvivalT ~ Treatment * Poison, lambda=seq(-1.4,0,0.1), data=poison.df)
```



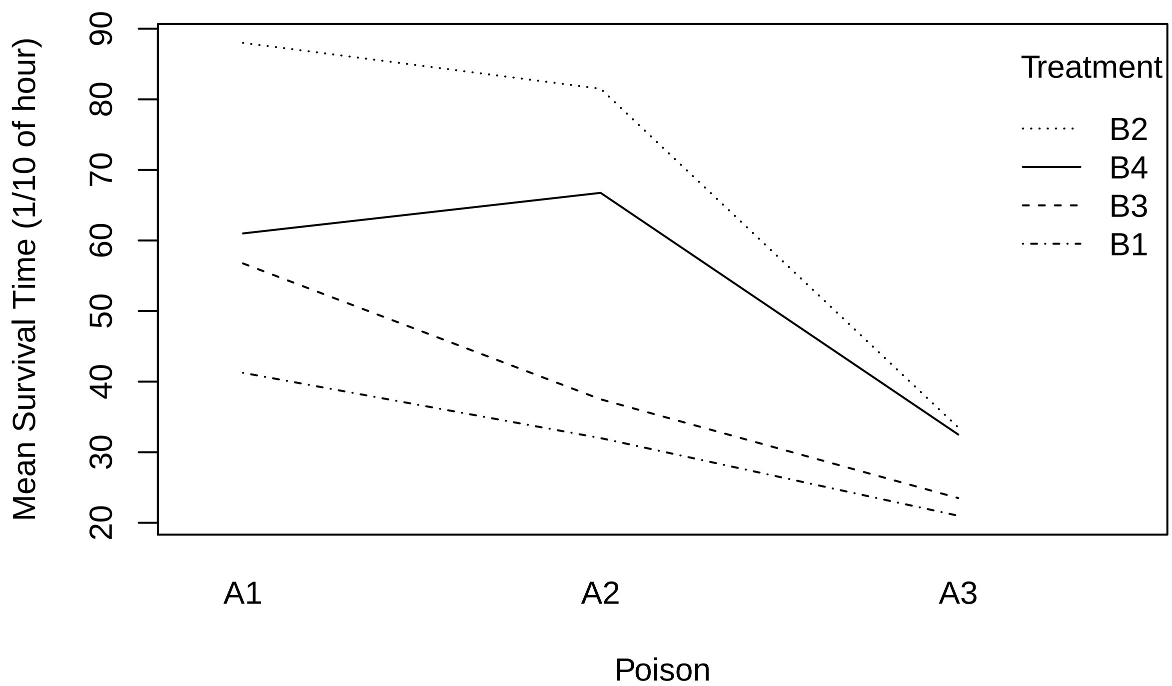
Box-Cox suggests that a transformation of *survival time*$^{-1}$ is required. This transformation is what will be used for subsequent analysis.

d)

```
interaction.plot(poison.df$Treatment,poison.df$Poison,poison.df$SurvivalT,
                 xlab="Treatment",ylab="Mean Survival Time (1/10 of hour)",
                 trace.label = "Poison",cex=0.5,xpd=1)
```




```
interaction.plot(poison.df$Poison,poison.df$Treatment,poison.df$SurvivalT,
                 xlab="Poison",ylab="Mean Survival Time (1/10 of hour)",
                 trace.label = "Treatment")
```



Looking at the first figure, it is clear that there is a main effect of poison, as indicated by the large difference between between treatment means for poison A3 compared to A1 & A2. Figure 2 suggests a main effect of treatement, with differences between treatment B2 & B4 compared to B1 & B3 being the greatest. Both plots also indicate that an interaction is present, since the lines are not perfectly parrallel.

e)

```r
summary(poison_fit <- aov(I(SurvivalT^-1) ~ Treatment * Poison, data = poison.df))
```

```
##                  Df   Sum Sq   Mean Sq F value   Pr(>F)    
## Treatment         3 0.002041 0.0006805   28.34 1.38e-09 ***
## Poison            2 0.003488 0.0017439   72.64 2.31e-13 ***
## Treatment:Poison  6 0.000157 0.0000262    1.09    0.387    
## Residuals        36 0.000864 0.0000240                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

f)

$$H_0 : (\alpha\beta)_{ij} = 0$$

where α is the effect of poison, β is the effect of treatment, $\alpha\beta$ is their interaction. (i = 1,2,3) for the levels of poison, while (j = 1,2,3,4) for the levels of treatment. Differences in survival time between types of poison was not dependent on the type of treatment received($F_{6,564} = 1.09, p = 0.387$). It is appropriate to interpret the main effects of poison and treatment. If the interaction term was significant, then it would only be appropirate to interpret the simple main effects.

g)

$$H_0: \quad poison_i = 0 \quad for \quad (i = 1, 2, 3) \qquad treatment_j = 0 \quad for \quad (j = 1, 2, 3, 4)$$

There was both a significant main effect of poison type on survival time ($F_{2,56} = 72.64, p = 2.31e-13$) and a significant main effect of treatment type on survival time ($F_{3,56} = 28.34, p = 1.38e-09$).

6)

a)

$$earning_{ijk} = subject_i + degree_j + (subject * degree)_{ij} + \epsilon_{ijk}$$

```r
pay <- c(1.7,1.9,2.5,2.3,2.6,2.4,2.7,2.8,2.5,2.6,
              1.8,2.1,2.7,2.4,2.6,2.4,2.5,2.9,3.0,2.8,2.7,2.3,2.8,
              2.5,2.7,2.9,2.5,2.6,2.8,2.7,2.9,3.5,3.3,3.6,3.4,3.7,3.6,
              3.7,3.8,3.9,3.3,3.4,3.3,3.5,3.6)
subject <- c(1,1,2,2,2,2,3,3,4,4,
             1,1,2,2,2,2,2,3,3,3,3,4,4,
             rep(1,8),2,2,2,2,rep(3,5),rep(4,5))
degree <- c(rep(1,10),rep(2,13),rep(3,22))

earnings.df <- data.frame(cbind(pay*1000,subject,degree))
colnames(earnings.df) <- c("Earnings","Subject","Degree")

earnings.df$Subject <- factor(earnings.df$Subject,labels = c("Hum",
                                                       "Soc",
                                                       "Engin",
                                                       "Manage"))
earnings.df$Degree <- factor(earnings.df$Degree,labels = c("Bach",
                                                       "Mast",
                                                       "Doc"))
```

Set to zero condition

```r
options(contrasts=c("contr.treatment","contr.poly"))
model.matrix(lm(Earnings ~ Subject * Degree, data = earnings.df))
```

```
##    (Intercept) SubjectSoc SubjectEngin SubjectManage DegreeMast DegreeDoc
## 1            1          0            0             0          0         0
## 2            1          0            0             0          0         0
## 3            1          1            0             0          0         0
## 4            1          1            0             0          0         0
## 5            1          1            0             0          0         0
## 6            1          1            0             0          0         0
## 7            1          0            1             0          0         0
## 8            1          0            1             0          0         0
## 9            1          0            0             1          0         0
```

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

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

```
## 26                         0                       0                        0
## 27                         0                       0                        0
## 28                         0                       0                        0
## 29                         0                       0                        0
## 30                         0                       0                        0
## 31                         0                       0                        0
## 32                         1                       0                        0
## 33                         1                       0                        0
## 34                         1                       0                        0
## 35                         1                       0                        0
## 36                         0                       1                        0
## 37                         0                       1                        0
## 38                         0                       1                        0
## 39                         0                       1                        0
## 40                         0                       1                        0
## 41                         0                       0                        1
## 42                         0                       0                        1
## 43                         0                       0                        1
## 44                         0                       0                        1
## 45                         0                       0                        1
## attr(,"assign")
##  [1] 0 1 1 1 2 2 3 3 3 3 3 3
## attr(,"contrasts")
## attr(,"contrasts")$Subject
## [1] "contr.treatment"
##
## attr(,"contrasts")$Degree
## [1] "contr.treatment"
```

Sum to zero condition

```
options(contrasts=c("contr.sum","contr.poly"))
model.matrix(lm(Earnings ~ Subject * Degree, data = earnings.df))
```

```
##    (Intercept) Subject1 Subject2 Subject3 Degree1 Degree2 Subject1:Degree1
## 1            1        1        0        0       1       0                1
## 2            1        1        0        0       1       0                1
## 3            1        0        1        0       1       0                0
## 4            1        0        1        0       1       0                0
## 5            1        0        1        0       1       0                0
## 6            1        0        1        0       1       0                0
## 7            1        0        0        1       1       0                0
## 8            1        0        0        1       1       0                0
## 9            1       -1       -1       -1       1       0               -1
## 10           1       -1       -1       -1       1       0               -1
## 11           1        1        0        0       0       1                0
## 12           1        1        0        0       0       1                0
## 13           1        0        1        0       0       1                0
## 14           1        0        1        0       0       1                0
## 15           1        0        1        0       0       1                0
## 16           1        0        1        0       0       1                0
## 17           1        0        1        0       0       1                0
## 18           1        0        0        1       0       1                0
```

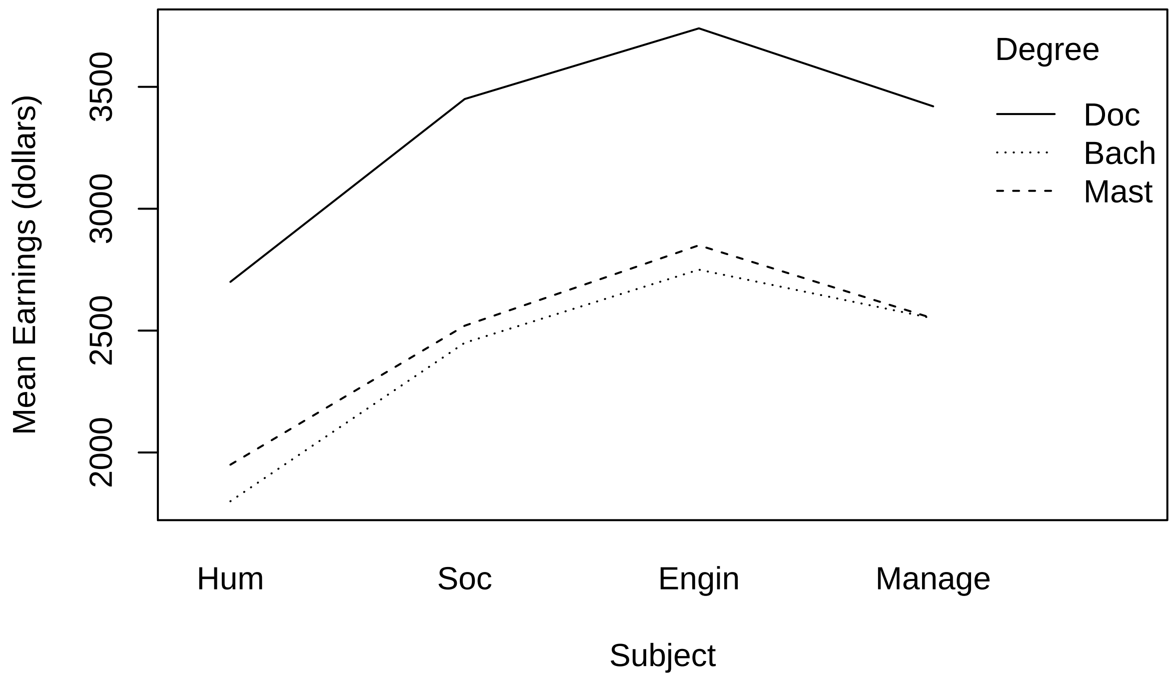
```
## 19           1        0        0        1      0      1               0
## 20           1        0        0        1      0      1               0
## 21           1        0        0        1      0      1               0
## 22           1       -1       -1       -1      0      1               0
## 23           1       -1       -1       -1      0      1               0
## 24           1        1        0        0     -1     -1              -1
## 25           1        1        0        0     -1     -1              -1
## 26           1        1        0        0     -1     -1              -1
## 27           1        1        0        0     -1     -1              -1
## 28           1        1        0        0     -1     -1              -1
## 29           1        1        0        0     -1     -1              -1
## 30           1        1        0        0     -1     -1              -1
## 31           1        1        0        0     -1     -1              -1
## 32           1        0        1        0     -1     -1               0
## 33           1        0        1        0     -1     -1               0
## 34           1        0        1        0     -1     -1               0
## 35           1        0        1        0     -1     -1               0
## 36           1        0        0        1     -1     -1               0
## 37           1        0        0        1     -1     -1               0
## 38           1        0        0        1     -1     -1               0
## 39           1        0        0        1     -1     -1               0
## 40           1        0        0        1     -1     -1               0
## 41           1       -1       -1       -1     -1     -1               1
## 42           1       -1       -1       -1     -1     -1               1
## 43           1       -1       -1       -1     -1     -1               1
## 44           1       -1       -1       -1     -1     -1               1
## 45           1       -1       -1       -1     -1     -1               1
##    Subject2:Degree1 Subject3:Degree1 Subject1:Degree2 Subject2:Degree2
## 1                 0                0                0                0
## 2                 0                0                0                0
## 3                 1                0                0                0
## 4                 1                0                0                0
## 5                 1                0                0                0
## 6                 1                0                0                0
## 7                 0                1                0                0
## 8                 0                1                0                0
## 9                -1               -1                0                0
## 10               -1               -1                0                0
## 11                0                0                1                0
## 12                0                0                1                0
## 13                0                0                0                1
## 14                0                0                0                1
## 15                0                0                0                1
## 16                0                0                0                1
## 17                0                0                0                1
## 18                0                0                0                0
## 19                0                0                0                0
## 20                0                0                0                0
## 21                0                0                0                0
## 22                0                0               -1               -1
## 23                0                0               -1               -1
## 24                0                0               -1                0
## 25                0                0               -1                0
## 26                0                0               -1                0
```

```
## 27                 0                 0                -1                 0
## 28                 0                 0                -1                 0
## 29                 0                 0                -1                 0
## 30                 0                 0                -1                 0
## 31                 0                 0                -1                 0
## 32                -1                 0                 0                -1
## 33                -1                 0                 0                -1
## 34                -1                 0                 0                -1
## 35                -1                 0                 0                -1
## 36                 0                -1                 0                 0
## 37                 0                -1                 0                 0
## 38                 0                -1                 0                 0
## 39                 0                -1                 0                 0
## 40                 0                -1                 0                 0
## 41                 1                 1                 1                 1
## 42                 1                 1                 1                 1
## 43                 1                 1                 1                 1
## 44                 1                 1                 1                 1
## 45                 1                 1                 1                 1
##    Subject3:Degree2
## 1                 0
## 2                 0
## 3                 0
## 4                 0
## 5                 0
## 6                 0
## 7                 0
## 8                 0
## 9                 0
## 10                0
## 11                0
## 12                0
## 13                0
## 14                0
## 15                0
## 16                0
## 17                0
## 18                1
## 19                1
## 20                1
## 21                1
## 22               -1
## 23               -1
## 24                0
## 25                0
## 26                0
## 27                0
## 28                0
## 29                0
## 30                0
## 31                0
## 32                0
## 33                0
## 34                0
```

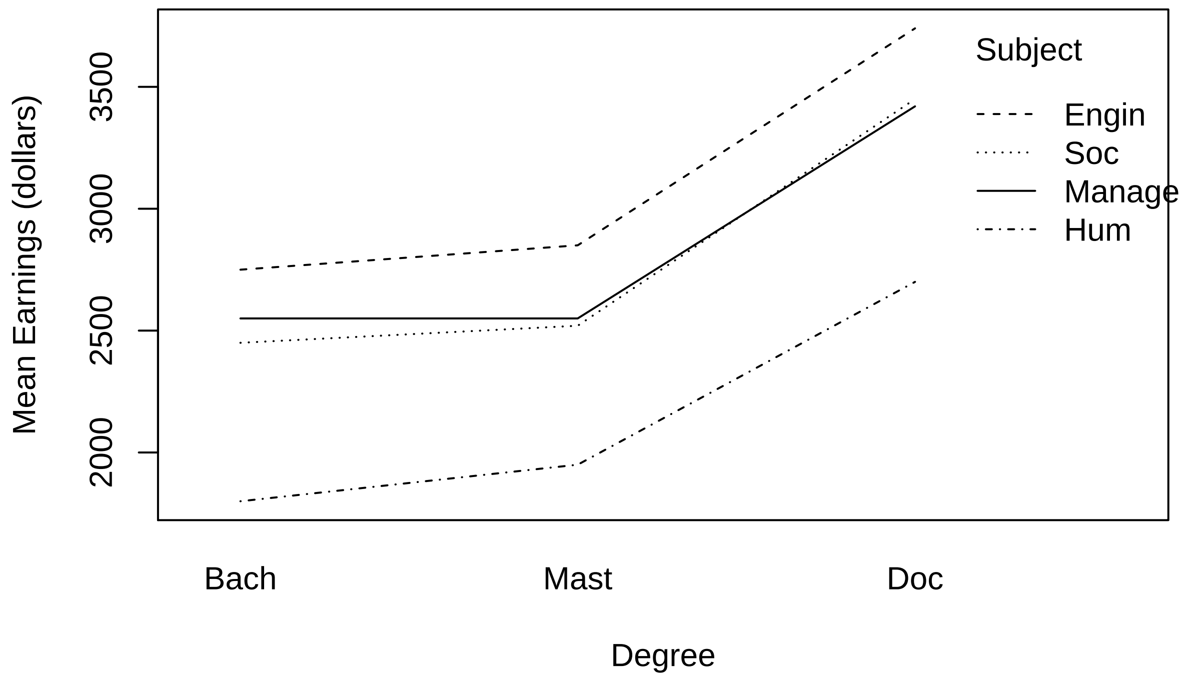
```
## 35                 0
## 36                -1
## 37                -1
## 38                -1
## 39                -1
## 40                -1
## 41                 1
## 42                 1
## 43                 1
## 44                 1
## 45                 1
## attr(,"assign")
##  [1] 0 1 1 1 2 2 3 3 3 3 3 3
## attr(,"contrasts")
## attr(,"contrasts")$Subject
## [1] "contr.sum"
##
## attr(,"contrasts")$Degree
## [1] "contr.sum"
```

b)

```r
options(contrasts=c("contr.treatment","contr.poly"))
interaction.plot(earnings.df$Subject,earnings.df$Degree,earnings.df$Earnings,
                 xlab="Subject",ylab="Mean Earnings (dollars)",
                 trace.label = "Degree",cex=0.5,xpd=1)
```




```r
interaction.plot(earnings.df$Degree,earnings.df$Subject,earnings.df$Earnings,
                 xlab="Degree",ylab="Mean Earnings (dollars)",
                 trace.label = "Subject",cex=0.5,xpd=1)
```



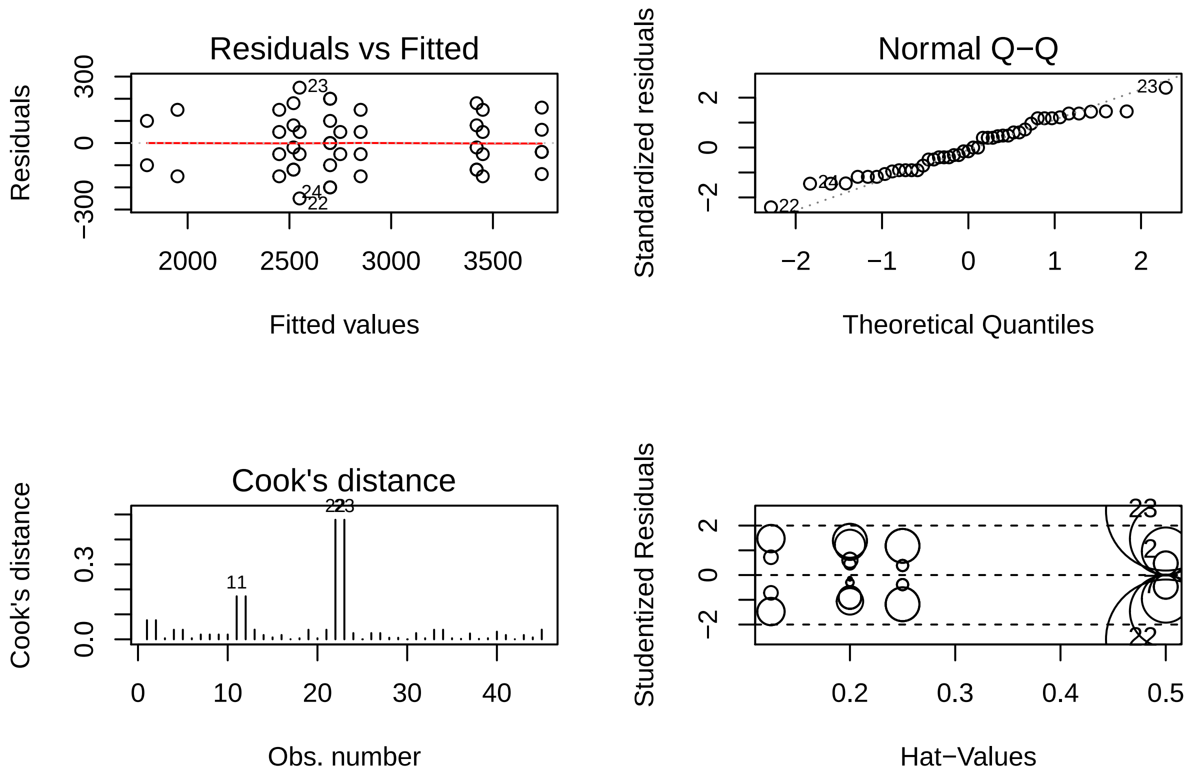

```
summary(aov(Earnings ~ Subject * Degree, data = earnings.df))
```

```
##                 Df  Sum Sq Mean Sq F value  Pr(>F)
## Subject          3 4167567 1389189   63.85 7.9e-14 ***
## Degree           2 8382452 4191226  192.63 < 2e-16 ***
## Subject:Degree   6   44425    7404    0.34    0.91
## Residuals       33  718000   21758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction plots between degree level and subject matter indicate that there is no significant interaction between these variables in determining earnings per course. This conclusion is corroborated by the anova table, where the interaction term is non-significant.

c

```
par(mfrow=c(2,2))
plot(earning.fit <- aov(Earnings ~ Subject * Degree, data = earnings.df),
     which = c(1,2,4))
influencePlot(earning.fit)
```




```
##        StudRes Hat      CookD
## 2   0.9575518 0.5 0.07660167
## 7  -0.4737129 0.5 0.01915042
## 22 -2.5971836 0.5 0.47876045
## 23  2.5971836 0.5 0.47876045
```

Assumptions of normality and constant variance appear to be satisified. There do appear to be some highly influential points that may be outliers, though.

```
outlierTest(earning.fit)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 23 2.597184           0.014085      0.63383
```

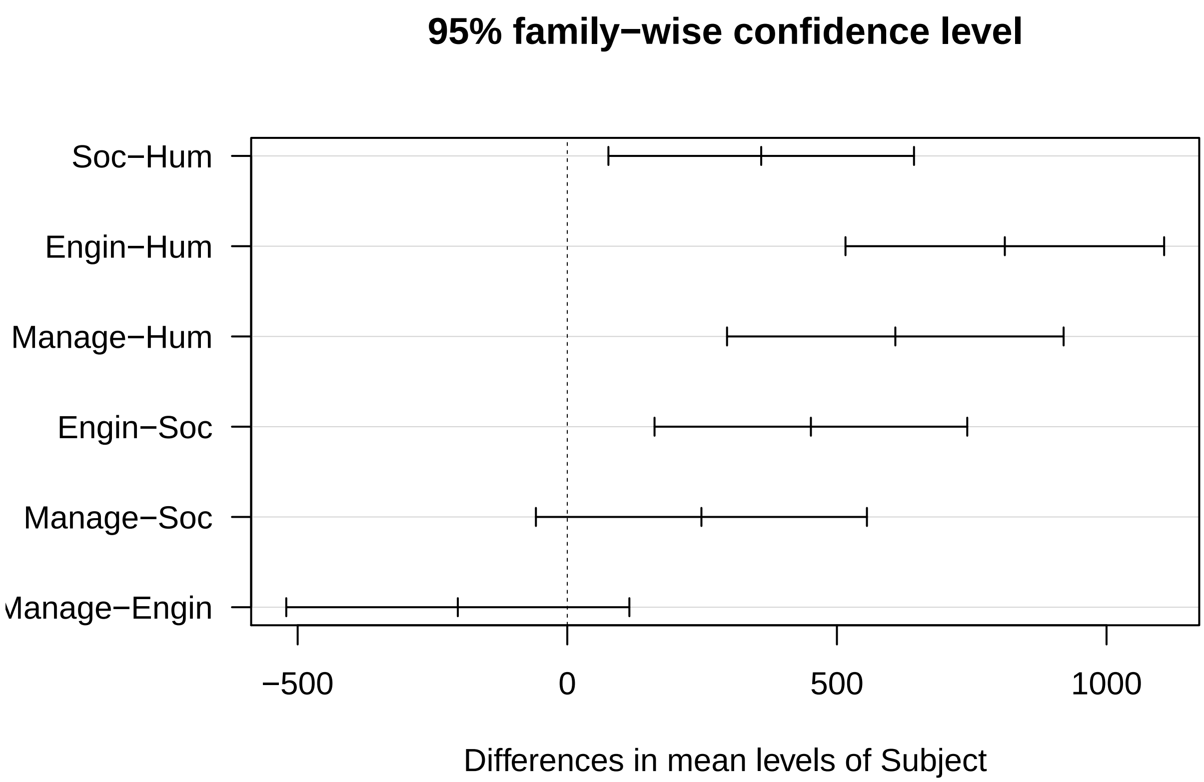
No outliers detected using a bonferonni corrected threshold.

d)

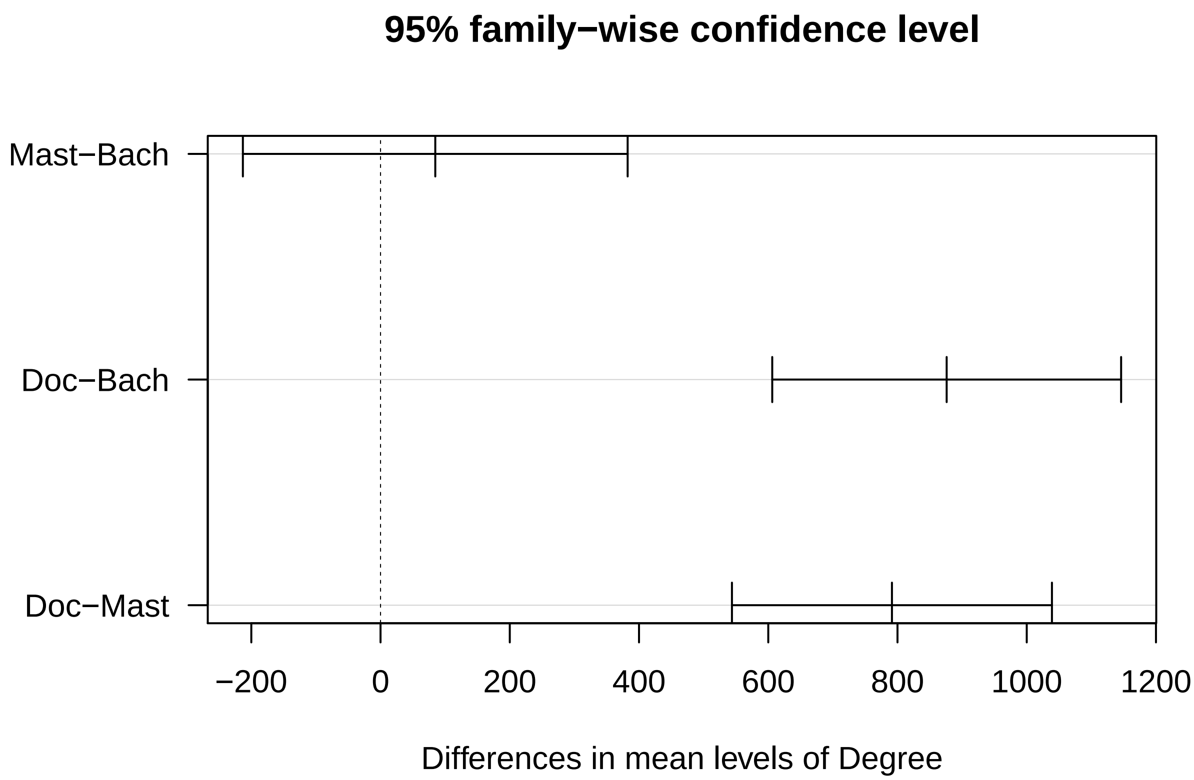
Post-Hoc pairwise comparisons are appropriate since there was a significant main effect for both degree level and subject matter.

```r
earnings.postSub <- ScheffeTest(earning.fit,which=c("Subject"))
earnings.postDeg <- ScheffeTest(earning.fit,which=c("Degree"))
```

```r
par(mar=c(5,6,4,1)+.4)
plot(earnings.postSub, las = 1)
```




```r
plot(earnings.postDeg, las = 1)
```



Post-hoc comparisons of the earnings of professors for different subjects were conducted using the scheffe's method. All comparisons of earnings between Social ($M = \$2785 \pm 132$), Humanities ($M = \$2425 \pm 126$), Engineering ($M = \$3236 \pm 149$), and Management ($M = \$3033 \pm 162$) professors were significant (\$ p < 0.01\$) except for the following: Management vs Social or Engineering. 95% scheffe's simultaneous CI for each comparison are presented above. If 0 is within the confidence band, then it is a nonsignificant difference.

A similar analysis was done for comparisons of earnings between professors who had obtained different levels of degrees. Doctrates ($M = \$3236 \pm 96.1$) earned significantly more money compared to Masters ($M = \$2538 \pm 93.7$) and Bachelors ($M = \$2400 \pm 111$). There was no significant different between earnings of Masters and Bachelors professors. 95% scheffe's simultaneous CI for these comparisons are also presented above.

e)

The highest paid adjunct professors were those who had doctorates and taught engineering ($M = \$3740 \pm 51$), while the lowest paid were those who had Bachelors and taught humanities ($M = \$1800 \pm 100$). Since a comparison between these groups of teachers is unplanned and there was no significant interaction between the factors Subject-Degree, I used the scheffe's method to correct for multiple comparisons (most conservative).

```
earning.cell_summary <- group_by(earnings.df,Subject,Degree) %>%
  summarise(mean.Earn = mean(Earnings), se.Earn = se(Earnings))

mse <- summary(earning.fit)[[1]][[3]][4]
con_sum <- (1/2) + (1/5)
```

```r
SE.c <- sqrt(mse * con_sum)
g <- length(levels(earnings.df$Subject)) * length(levels(earnings.df$Degree))

crit_f <- qf(0.01, g-1, nrow(earnings.df) - g, lower.tail = F)

doc.eng <- max(earning.cell_summary$mean.Earn)
bach.hum <- min(earning.cell_summary$mean.Earn)

diff <- doc.eng-bach.hum

scheffe.lower.earn <- diff - sqrt((g-1)*crit_f) * SE.c
scheffe.upper.earn <- diff + sqrt((g-1)*crit_f) * SE.c

paste(round(scheffe.lower.earn,2), round(scheffe.upper.earn,2), sep = " , ")
```

```
## [1] "1250.26 , 2629.74"
```

The 99% CI for the mean difference between the highest paid adjunct professor vs the lowest paid adjunct professor (M_{Diff} = \$1940) was [\$1250.26,\$2629.74], indicating that this difference was indeed significant.

7)

```r
data("teengamb")
```

```r
teengamb$sex <- factor(teengamb$sex, labels = c("M","F"))
teen.fit <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
```

a)

```r
summary(teen.fit)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sexF        -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

The coeficients for income and females are significant.

b)

The sex coeffiienct in these results represents the change in gamblind expenditure from male to females when holding all other covariates constant. That change is a decrease in gambling expdenditure of 22.12 pounds per year.

c)

```r
avg.data <- as.data.frame(rbind(colMeans(teengamb[-c(1,5)])))
avg.data$sex <- factor(0, label = "M")
colnames(avg.data) <- c("status","income","verbal","sex")
male.average <- predict(teen.fit,newdata = avg.data,
                        se=T,interval="prediction")
male.average
```

```
## $fit
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
##
## $se.fit
## [1] 4.687496
##
## $df
## [1] 42
##
## $residual.scale
## [1] 22.69034
```

```r
max.data <- as.data.frame(lapply(teengamb[-c(1,5)],max))
max.data$sex <- factor(0, label = "M")
colnames(max.data) <- c("status","income","verbal","sex")
male.max <- predict(teen.fit,newdata = max.data,
                        se=T,interval="prediction")
male.max
```

```
## $fit
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
##
## $se.fit
## [1] 14.40753
##
```

```
## $df
## [1] 42
##
## $residual.scale
## [1] 22.69034
```

The confidence interval for the prediction of a male with maximum values of the covariates status, income, and verbal was larger since such an observation is farther from the mean of the model fit.

d)

```r
summary(teen.fit2 <- lm(gamble ~ income, data = teengamb))
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```r
AIC(teen.fit)
```

```
## [1] 433.5561
```

```r
AIC(teen.fit2)
```

```
## [1] 439.7158
```

The first model, which included the covariates sex, status, income and verbal, performed better than the model that only included income. It explained a greater proportion of the variation in gambling ($R^2 = 0.53, Adj.R^2 = 0.48, F_{4,42} = 11.69, p = 1.815e-06$), and yeilded a smaller AIC value of 433.57. The added variation explained is likely due to the inclusion of sex as a covariate, which was significant in the first model ($\beta_{sex} = -22.11, p = 0.01$)