STA303 - Assignment 2

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA303H1S-L0101

Feb 19th, 2017

Initializing data and libraries once, before starting the questions.

```
# Loading ggplot2, lsmeans & estimability library
library(ggplot2)
library(lsmeans)
library(estimability)

# Loading in the data 1 dataset
data1 = read.table("A3Q1data.txt", sep = " ", header = T)

# Loading in the data 2 dataset
data2 = read.table("donner.txt", sep = " ", header = T)

# Checking type of variables in this data
str(data1)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ pretest : int  24 23 27 28 33 27 38 31 44 42 ...
##  $ posttest: int  45 28 34 50 39 31 59 36 55 60 ...
##  $ trmt    : int  1 2 3 1 2 3 1 2 3 1 ...
```

```
str(data2)
```

```
## 'data.frame':    45 obs. of  3 variables:
##  $ age         : int  23 40 40 30 28 40 45 62 65 45 ...
##  $ sex         : int  1 0 1 1 1 1 0 1 1 0 ...
##  $ survivorship: int  0 1 1 0 0 0 0 0 0 0 ...
```

```
# Converting treatment field in the dataset as a factor variable
data1$trmt = as.factor(data1$trmt)

# Converting sex field in the dataset as a factor variable
data2$sex = as.factor(data2$sex)

# Converting survivorship field in the dataset as a factor variable
data2$survivorship = as.factor(data2$survivorship)
```

Q1 (a-c) - Data 1: mental treatment

(a) Construct the one-way ANOVA analysis for comparing the three treatment means when pretest is ignored. (Show your code, ANOVA output and give your analysis conclusion).

```r
summary(aov(data1$posttest ~ data1$trmt, data = data1))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## data1$trmt   2   1752   876.1   6.706 0.00432 **
## Residuals   27   3528   130.7                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion : Analyzing the one-way ANOVA output, the p-value we calculated was 0.00432. This is a small number indicating that we can reject the null hypothesis with a high degree of significance indicating that atleast one pair of treatment means are unequal.

(b) From the one-way ANOVA in (1a), it involves a F-test for equality of means. Specify a model and a null hypothesis for no therapy effect, then give the formula of the F-ratio and its observed value from data. Use R code to find the critical value of this F-test using $\alpha = 0.05$. Compare the observed F value and the the critical value, what conclusion do you have ? Does it agree with your conclusion based on p-value in (1a) ?

```r
# Applying one-way ANOVA
summary(aov(data1$pretest ~ data1$trmt, data = data1))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## data1$trmt   2   18.5    9.23   0.101  0.904
## Residuals   27 2471.0   91.52               
```

Solution: The model is one-way ANOVA, with the following hypothesis:

The null and alternative hypothesis for the no therapy effect:

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ... = \alpha_i = 0 \\ H_a : \exists i \neq j, s.t. \alpha_i \neq 0 \end{cases}$$

Observed F-ratio: $F^* = \frac{SS_T/(r-1)}{SS_E/(n-r)}$ where:

Sum of Squares Between Groups: $SS_T = \sum_{i=1}^{r} n_i(\overline{Y_{i.}} - \overline{Y})^2$

Sum of Squares Error: $SS_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2 = \sum_{i=1}^{r}(n_i - 1)s_i^2$

Number of Treatments: $r = 3$

Number of Observations: $n = 30$

Observed F-value using R:

```r
# Calculating total means
total_mean = mean(data1$posttest)

# Calculating the treatment means
trmt_means = with(data1, tapply(posttest, trmt, mean))

# Calculating the treatment sizes
trmt_sizes <- with(data1, tapply(posttest, trmt, length))
```

```r
# Treatments we achieve are all of size 10
trmt_sizes
```

```
##  1  2  3
## 10 10 10
```

```r
sum_of_squares_between_groups = sum(10 * (trmt_means - total_mean)^2)

# Calculating the treatment sd
trmt_sds = with(data1, tapply(posttest, trmt, sd))
sum_of_squares_error = sum((10 - 1) * trmt_sds^2)
(sum_of_squares_between_groups/2)
```

```
## [1] 876.1333
```

```r
(sum_of_squares_error/27)
```

```
## [1] 130.6519
```

```r
observed_F_statistic = (sum_of_squares_between_groups/2)/(sum_of_squares_error/27)
observed_F_statistic
```

```
## [1] 6.705862
```

With the null hypothesis, the distribution of this observed F statistic is $F^* \sim F_{r-1,n-r}$. In our case it is $F^* \sim F_{2,27}$

For a significance level of 0.05, our critical value is the 95^{th} percentile of that distribution:

```r
qf(0.95, df1 = 2, df2 = 27)
```

```
## [1] 3.354131
```

We note that the critical values is significantly less than our observed F statistic, which shows high statistic significance and we can conclude to reject the null hypothesis stating that there was no therapy effect and conclude that there indeed was a therapy effect.

This is in accordance with our results in question 1 (a). Both tests are equivalent, in question 1a we rejected the null hypothesis and took on the alternative, stating that atleast one pair of treatment means are unequal which would mean that there was a therapy effect, which is what we concluded in this question. Adding on to this, the test in question 1 a was highly statistically signifant, and in this question, as mentioned before, our critical value was significantly smaller than our observed F statistic.

(c) What is the homogeneity of slopes assumption of ANCOVA ? Why is it important ?

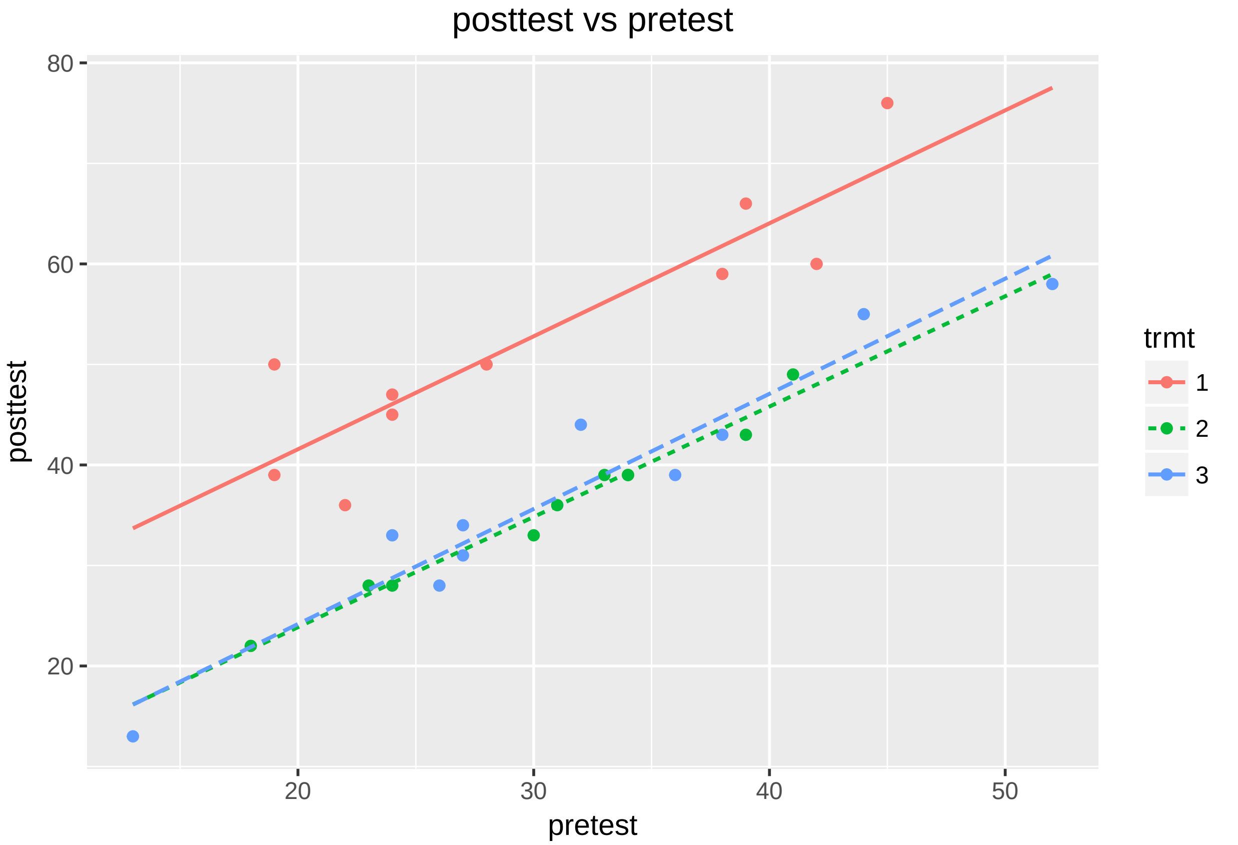
Solution: Whenever an ANCOVA is performed we look at the overall relationship between the outcome (dependent variable) and the covariate: we fit a regression line to the entire data set, ignoring to which group a person belongs. In fitting this overall model we, therefore, assume that this overall relationship is true for all groups of participants. The assumption that regression lines should be parallel among groups is what we call the homogeneity of slopes assumption for ANCOVA.

This assumption is required due to the fact it is important that treatments do not affect the value of the covariates, otherwise if relationships between the outcome (dependent variable) and covariates were to differ across the groups, then the overall regression model would be inaccurate (since it does not represent all of the groups).

Q2 (a-d) - Data 1: mental treatment

(a) Plot posttest versus pretest with a different symbol or color for each treatment. From this plot, does the assumption of homogenous slope look reasonable ?

```
# Setting up an empty plot (posttest versus pretest)
plot_setup = ggplot(data1, aes(x = pretest, y = posttest, group = trmt, colour = trmt,
    linetype = trmt))
# Plotting the points
plot_points = geom_point()

# Plotting the regression lines
plot_lm = geom_smooth(method = "lm", size = 0.7, se = FALSE, fullrange = TRUE)

# Entering a title
plot_title = ggtitle("posttest vs pretest")

# Positioning the title
plot_theme = theme(plot.title = element_text(hjust = 0.5))

# Putting everything together
plot_setup + plot_points + plot_lm + plot_title + plot_theme
```



Solution: The three regression lines are for the three treatments given. Although you could argue that the regression line of treatment 2 and treatment 3 are not exactly parallel, their slopes (of all treatments i.e. treatment 1, 2 and 3) are quite similar, indicating that the homogeneity of slopes assumption is met (looks reasonable).

(b) Specify a model that can be used to access the homogenous regression slope assumption. Evaluate the assumption for this data. Is the homogenous slopes assumption met ?

```
# Homogeneity assumption test
anova(lm(data1$posttest ~ data1$pretest * data1$trmt, data = data1))
```

```
## Analysis of Variance Table
##
## Response: data1$posttest
##                             Df  Sum Sq Mean Sq  F value    Pr(>F)
## data1$pretest                1 2847.06 2847.06 179.9097 1.207e-12 ***
## data1$trmt                   2 2052.23 1026.11  64.8415 2.104e-10 ***
## data1$pretest:data1$trmt     2    0.78    0.39   0.0246    0.9757
## Residuals                   24  379.80   15.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution: We will need to use the ANCOVA model in order to access the homogenous regression slop assumption. The assumptions are (1) linearity of regression (posttest and pretest) must be linear. (2) Homogeneity of variance: equal variances for different treatment classes and observations. (3) Independence of error terms. (4) Normality of error terms. (5) Homogeneity of regression slopes. (which is what we will test.)

ANCOVA model is:

$$posttest_{ik} = \mu + \alpha_i + \sum_{j=1}^{p=1} \beta_j (pretest_{ijk}) + \epsilon_{ik}$$

After running this model, we received a p-value of 0.9757 which is greater than 0.05, therefore we do not have enough evidence to reject the null hypothesis which states the same slope assumption. We conclude in saying that it's okay to assume that the assumption is met.

(c) Fit an ANCOVA model to this data. Report the F-test for a treatment effect, after controlling for the effect of the pretest measurement. (Also show your R code and ANOVA output)

```
# Applying ANOVA on ANCOVA model
anova(lm(data1$posttest ~ data1$pretest + data1$trmt), data = data1)
```

```
## Analysis of Variance Table
##
## Response: data1$posttest
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## data1$pretest  1 2847.06 2847.06 194.503 1.399e-13 ***
## data1$trmt     2 2052.23 1026.11  70.101 3.360e-11 ***
## Residuals     26  380.58   14.64
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: ANCOVA model is:

$$posttest_{ik} = \mu + \alpha_i + \sum_{j=1}^{p=1} \beta_j (pretest_{ijk}) + \epsilon_{ik}$$

Performing ANOVA on this model, the effect for the pretest measurement is accounted for.

Observing the output, we achieved a value of 1.399e-13 (1.399×10^{-13}) for pretest and a value of 3.360e-11 (3.360×10^{-11}) for a treatment effect, after controlling for the effect of the pretest measurement. Both of these values are very small and are significant. We can conclude that the F-test shows pretest being a significant predictor of posttest score and that treatment treatment has a significant effect on posttest score.

(d) Find unadjusted and adjusted post-test score for 3 treatments.

```r
# Finding the unadjusted post-test score for 3 treatments
aggregate(data1$posttest, list(data1$trmt), mean)
```

```
##   Group.1    x
## 1       1 52.8
## 2       2 35.6
## 3       3 37.8
```

```r
# Finding the adjusted post-test score for 3 treatments
lsmeans(lm(data1$posttest ~ data1$pretest + data1$trmt), "data1$trmt")
```

```
##  data1$trmt   lsmean       SE df lower.CL upper.CL
##  1          46.02881 1.294998 26 43.36691 48.69072
##  2          26.91031 1.347213 26 24.14107 29.67955
##  3          32.27020 1.267278 26 29.66527 34.87513
##
## Confidence level used: 0.95
```

Q3 (a-b) - Data 1: mental treatment

(a) If we define Y=posttest-protest as our new dependent variable, fit the oneway ANOVA model to it. How significant of the treatment effect ?

```r
# Definining Y
data1$Y = data1$posttest - data1$pretest

# Fitting one-way ANOVA
anova(lm(data1$Y ~ data1$trmt, data = data1))
```

```
## Analysis of Variance Table
##
## Response: data1$Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## data1$trmt  2 2023.4 1011.70  64.822 4.921e-11 ***
## Residuals  27  421.4   15.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution: The value we received is 4.921e-11. This is a very small number indicating that the treatment effect was highly significant.

(b) Compare the one-way ANOVA with the ANCOVA model in Q-2c, which model do you prefer and why ?

Solution: Comparing ANOVA with ANCOVA, we see that ANOVA measures whether the mean change in the outcome from pre to post differs in the three groups. ANCOVA measures whether the post-test means, adjusted for pre-test scores, differs between groups.

We would prefer to use the ANCOVA model. ANCOVA helps to account for differences created by a covariate, so if the differences between a certain measure don't aid your effect size, but they do account for some of the variation in the model, an ANCOVA would basically take those out of the SSE, reducing the overall noise of your model.

Q4 (a-f) - Data 2: donner party

(a) Fit a logistic regression model to the data with covariates sex and age. Provide the summary output. Give the formula for the estimated curve. What is the fitted male model ? What is the fitted female model ?

```r
# Fitting a logistic regression model
summary(glm(formula = data2$survivorship ~ data2$age + data2$sex, family = "binomial",
    data = data2))
```

```
##
## Call:
## glm(formula = data2$survivorship ~ data2$age + data2$sex, family = "binomial",
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.23041    1.38686   2.329   0.0198 *
## data2$age   -0.07820    0.03728  -2.097   0.0359 *
## data2$sex1  -1.59729    0.75547  -2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Solution: From summary output, we have $b_0 = 3.23041$, $b_1 = -0.07820$, $b_2 = -1.59729$.

The estimated curve is: $(\hat{\pi} = \frac{e^\eta}{1+e^\eta})$

Where: $\eta = 3.23041 - 0.07820 age - 1.59729 I_{male}$

Fitted male model:

$$logit(\hat{\pi}) = 3.23041 - 0.07820\text{age} - 1.59729 = 1.63312 - 0.07820\text{age}$$

Fitted female model:

$$logit(\hat{\pi}) = 3.23041 - 0.07820\text{age}$$

(b) Interpret the intercept and all the slopes from the full fitted model.

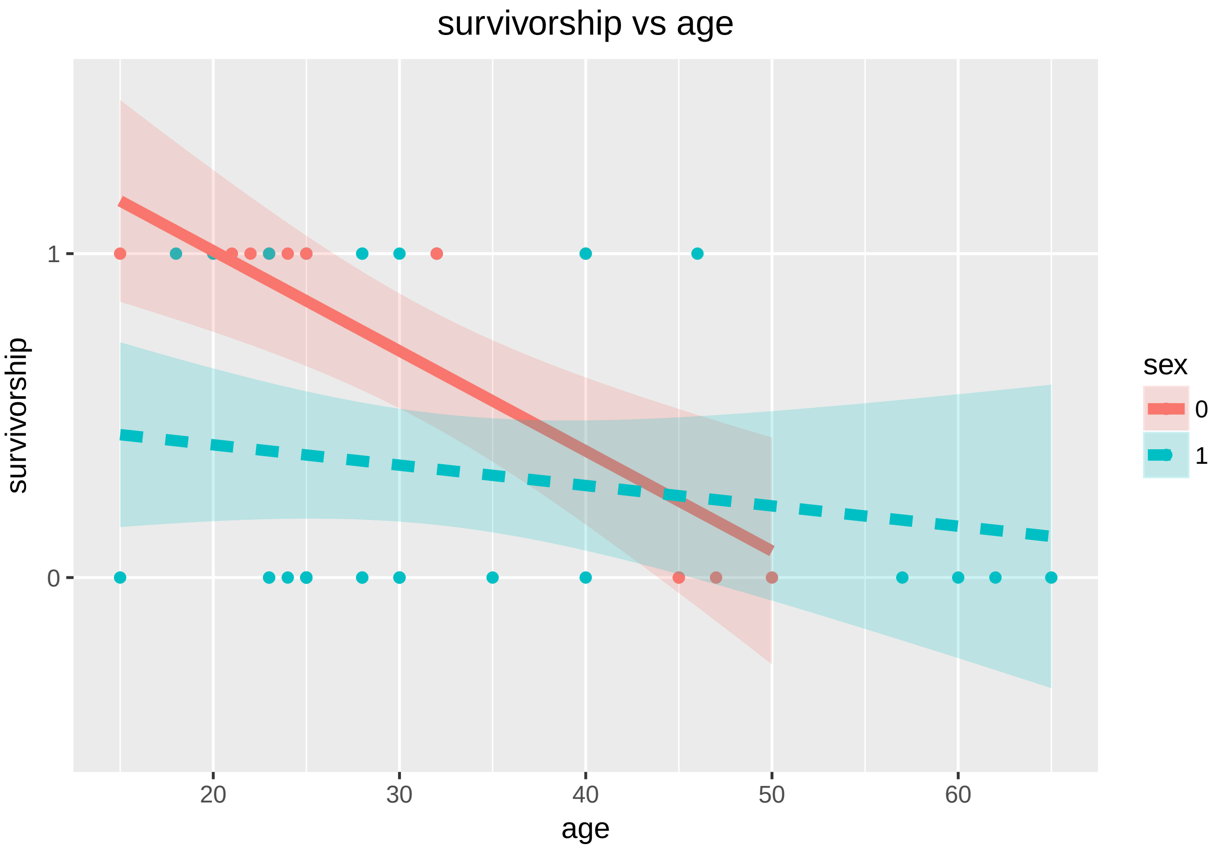
Solution:
$b_0 = 3.23041$: In our case, b_0 is the the log-Odds of survival for females. We see that females have a log-Odds survival of 3.23041.

$b_1 = -0.07820$: In logistic regression, b_1 is the increase in the log-Odds of survival, for a unit change in age, keeping other variables constant. In our case, the log odds of survival decreases by 0.07820 for each unit increase in age.

$b_2 = -1.59729$: In logistic regression, b_2 is the increase in the log-Odds of survival for males compared to females, keeping age constant. In our case, being male compared to female, decreases the log odds of survival by 1.59729.

(c) Plot the logistic regression curve versus age that has both male and female curves on it. Look at this plot, what conclusion do you have comparing the estimated survival probability for a male and a female given age =30 ?

```
# q4 = glm(formula = survivorship ~ age, family = binomial(link='logit'),
# data = data2) plot(data2$age, data2$survivorship) xv = seq(min(data2$age),
# max(data2$age), 0.01) yv <- predict(q4, list(age=xv), type = 'response')
# lines(xv, yv)

# xweight <- seq(0, 6, 0.01) yweight <- predict(q4, list(age =
# xweight),type='response') plot(data2$, mtcars$vs, pch = 16, xlab = 'WEIGHT
# (g)', ylab = 'VS')

# Setting up an empty plot (survivorship versus age)
plot_setup = ggplot(data2, aes(x = age, y = survivorship, group = sex, colour = sex,
    linetype = sex))

# Plotting the points
plot_points = geom_point()

# Plotting the logistic regression curve
plot_lm = stat_smooth(method = "glm", family = binomial, formula = y ~ x, alpha = 0.2,
    size = 2, aes(fill = sex))

# Entering a title
plot_title = ggtitle("survivorship vs age")

# Positioning the title
plot_theme = theme(plot.title = element_text(hjust = 0.5))

# Putting everything together
plot_setup + plot_points + plot_lm + plot_title + plot_theme
```



Comments: Looking at the graph we notice that at around age 30 the estimated survival probability is higher for females compared to males.

(d) What are the estimated probabilities of survival for men and women of ages 25 and 50 ?

Solution: Estimated probabilities,

$$Men : \hat{\pi} = \frac{e^{1.63312-0.07820\text{age}}}{1 + e^{1.63312-0.07820\text{age}}}$$

$$Women : \hat{\pi} = \frac{e^{3.23041-0.07820\text{age}}}{1 + e^{3.23041-0.07820\text{age}}}$$

Plugging in for 25 and 50 for Men and Women we get the following survival probabilities:

- Age 25 (Men) = 0.42021
- Age 50 (Men) = 0.09306
- Age 25 (Female) = 0.78167
- Age 50 (Female) = 0.33635

(e) What is the age at which the estimated probability of survival is 50 percent for women and for men ?

Solution: For females, the age of 50% survival is 41.31 years; for males it is 20.88 years.

(f) Why should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50 ?

Solution: Looking at the data, we notice that there were no females over 50. So if we were to do any comparisons for Donner Party members over 50, it must be with the *assumption* that the model would still hold/be valid and we can not really verify this with the given data. Which is the main reason why we would be reluctant to draw any conclusions about members over 50.