Linear Models HW # 1

Liam Flaherty

Professor Maity

NCSU: ST503-651

May 20, 2024

1) Consider the analysis of covariance (ANCOVA) model: $\boldsymbol{y_{i,j} = \mu + \alpha_i + x_{i,j}\beta + \epsilon_{i,j}}$ for $\boldsymbol{i = 1, 2, 3}$ and $\boldsymbol{j = 1, \dots, n}$.

a. Write the model in matrix form, clearly specifying all model components.

Our equations are:

$$\begin{aligned}
y_{1,1} &= 1 \cdot \mu + 1 \cdot \alpha_1 + 0 \cdot \alpha_2 + 0 \cdot \alpha_3 + x_{1,1}\beta + \epsilon_{1,1} \\
\vdots &= \vdots \\
y_{1,n} &= 1 \cdot \mu + 1 \cdot \alpha_1 + 0 \cdot \alpha_2 + 0 \cdot \alpha_3 + x_{1,n}\beta + \epsilon_{1,n} \\
y_{2,1} &= 1 \cdot \mu + 0 \cdot \alpha_1 + 1 \cdot \alpha_2 + 0 \cdot \alpha_3 + x_{2,1}\beta + \epsilon_{2,1} \\
\vdots &= \vdots \\
y_{2,n} &= 1 \cdot \mu + 0 \cdot \alpha_1 + 1 \cdot \alpha_2 + 0 \cdot \alpha_3 + x_{2,n}\beta + \epsilon_{2,n} \\
y_{3,1} &= 1 \cdot \mu + 0 \cdot \alpha_1 + 0 \cdot \alpha_2 + 1 \cdot \alpha_3 + x_{3,1}\beta + \epsilon_{3,1} \\
\vdots &= \vdots \\
y_{3,n} &= 1 \cdot \mu + 0 \cdot \alpha_1 + 0 \cdot \alpha_2 + 1 \cdot \alpha_3 + x_{3,n}\beta + \epsilon_{3,n}
\end{aligned}$$

This can be neatly placed in the matrix form $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ where $\underset{\sim}{y}$ is the $3n \times 1$ **response vector**, X is the $3n \times 5$ model matrix of **covariates (predictors)**, $\underset{\sim}{\beta}$ is the 5×1 vector of the **regression coefficients**, and $\underset{\sim}{\epsilon}$ is the vector of our **error terms**.

Explicitly, we have:

$$\begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,n} \\ y_{2,1} \\ \vdots \\ y_{2,n} \\ y_{3,1} \\ \vdots \\ y_{3,n} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & x_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1,n} \\ 1 & 0 & 1 & 0 & x_{2,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,n} \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \end{bmatrix} + \begin{bmatrix} \epsilon_{1,1} \\ \vdots \\ \epsilon_{1,n} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{2,n} \\ \epsilon_{3,1} \\ \vdots \\ \epsilon_{3,n} \end{bmatrix}$$

b. Is the model matrix $\boldsymbol{X}$ full column rank?

No, the null space must be non-trivial since the first column is the sum of the middle three columns.

2) Consider the teen gambling data, teengamb, in the R package faraway.

a. Write a brief description of the dataset. Produce some numerical and graphical summaries of the dataset.

According to the documentation for the faraway package, the teengamb dataset consists of 47 observations on 5 variables dealing with teenage gambling in Britain. The variables collected on the observations include sex (0 for male, 1 for female), status (an integer score based on the parents' socioeconomic status), income (in pounds per week), verbal (an integer score giving the number of words correctly defined out of 12 tested), and gamble (in pounds spent on gambling per year). The specifics are given in Figure 0.1 below, while some data visualizations are given in Figure 0.2.

```
> library(faraway)
> summary(teengamb)                          #Get a glimpse of the data#
      sex               status          income           verbal          gamble
 Min.    :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
 1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
 Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
 Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
 3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
 Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
> str(teengamb)                              #get a glimpse of the data#
'data.frame':   47 obs. of  5 variables:
 $ sex   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ status: int  51 28 37 28 65 61 28 27 43 18 ...
 $ income: num  2 2.5 2 7 2 3.47 5.5 6.42 2 6 ...
 $ verbal: int  8 8 6 4 8 6 7 5 6 7 ...
 $ gamble: num  0 0 0 7.3 19.6 0.1 1.45 6.6 1.7 0.1 ...
```

Figure 0.1: Dataset Description

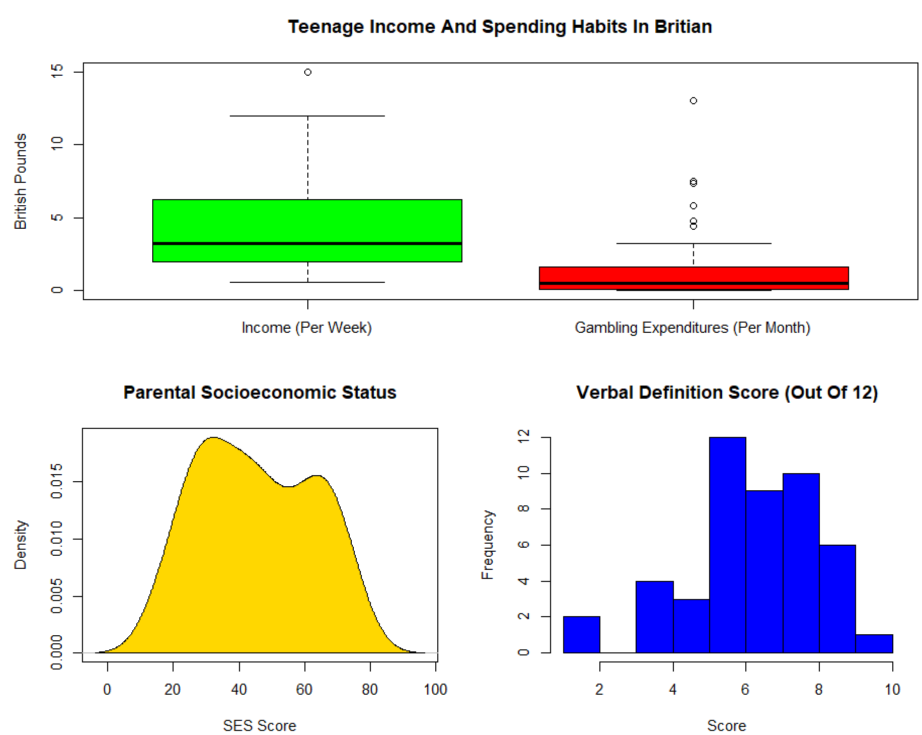

Figure 0.2: Data Visualizations

b. Fit a linear model using the lm() function with gample variable as response, and the income variable as predictors, and report the regression coefficients.

With this simple model, there are only two regression coefficients to report, the slope (which is 5.52), and the intercept (which is -6.325).

```
> #####3. Regression#####
> out=lm(teengamb$gamble ~ teengamb$income)                    #response (y) ~ predictor (x)#
> out

Call:
lm(formula = teengamb$gamble ~ teengamb$income)

Coefficients:
    (Intercept)  teengamb$income
         -6.325            5.520
```

Figure 0.3: R Code For Regression

c. Write the mathematical form of the model you fit in part b. Clearly define each component in your model.

In a general least squares scenario, our model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for a given predictor x_i where the y_i is our response, the β's are our regression coefficients, and the ϵ is our error. Here we have $y_i = -6.352 + 5.52(x_i) + \epsilon_i$. In matrix form:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} -6.352 \\ 5.52 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

d. Compute the mean and standard deviation of gamble and income for males (sex=0) and females (sex=1) separately. Comment on the results.

The mean annual gambling expenditure for males in the dataset was about 29.78 pounds per year versus about 3.87 pounds per year for females (standard deviations of about 37.32 and 5.15 respectively). Such an extreme difference suggests a two-sample problem might be the way to go when fitting a model.

```
> #####4. Sex  Breakdown#####
> number.males=sum(teengamb$sex==0)                                 #count total males#
> number.females=sum(teengamb$sex==1)                               #count total females#
>
> male=data.frame(c("mean", "standard deviation"),                  #dummy column#
+                 rbind(                                            #stack rows on top of each other#
+                 sapply(teengamb[which(teengamb$sex==0),], mean),  #mean of each column, filtered to males#
+                 sapply(teengamb[which(teengamb$sex==0),], sd)     #sd of each column, filtered to males#
+                 ))
> names(male)=c(paste0("MALE", "(n=", number.males, ")"),           #rename first colummn#
+               names(male)[2:ncol(male)])
>
> female=data.frame(c("mean", "standard deviation"),                #dummy column#
+                   rbind(                                          #stack rows on top of each other#
+                   sapply(teengamb[which(teengamb$sex==1),], mean), #mean of each column, filtered to females#
+                   sapply(teengamb[which(teengamb$sex==1),], sd)   #sd of each column, filtered to females#
+                   ))
> names(female)=c(paste0("FEMALE", "(n=", number.females, ")"),     #rename first colummn#
+                 names(female)[2:ncol(female)])
>
> male
          MALE(n=28) sex   status   income   verbal   gamble
1               mean   0 52.00000 4.976071 6.821429 29.77500
2 standard deviation   0 16.43393 4.086625 2.143959 37.32418
> female
        FEMALE(n=19) sex   status   income   verbal   gamble
1               mean   1 35.26316 4.149474 6.421053 3.865789
2 standard deviation   0 13.42817 2.598240 1.346427 5.150730
```

Figure 0.4: Differences In Male And Female

e. Fit the same linear regression as in part b, but separately for males and females. Report the regression coefficients.

The slope and intercepts are 6.518 and -2.66 for males, and 0.1749 and 3.14 for females.

```
> #####5. Sex Differences Income/Spending Regression#####
> maledf=teengamb[which(teengamb$sex==0),]                  #filter to only males#
> femaledf=teengamb[which(teengamb$sex==1),]                #filter to only females#
>
> maleout=lm(maledf$gamble ~ maledf$income)                 #response(y) ~ predictor(x)#
> maleslope=maleout[[1]][[2]]                               #just get numeric output#
> maleintercept=maleout[[1]][[1]]                           #just get numeric output#
>
> femaleout=lm(femaledf$gamble ~ femaledf$income)
> femaleslope=femaleout[[1]][[2]]
> femaleintercept=femaleout[[1]][[1]]
> maleout

Call:
lm(formula = maledf$gamble ~ maledf$income)

Coefficients:
  (Intercept)  maledf$income
       -2.660          6.518

> femaleout

Call:
lm(formula = femaledf$gamble ~ femaledf$income)

Coefficients:
    (Intercept)  femaledf$income
         3.1400           0.1749
```

Figure 0.5: Linear Model For Males And Females

f. Create a scatterplot between gamble (in y axis) and income (x axis), and color the points by sex. Then add two fitted regression lines from part e to the plot.

When filtering results by gender, our model is significantly different than when we report results together (i.e. sex seems to be a moderating variable).

```
plot(teengamb$income, teengamb$gamble,                          #predictor(x), response(y)#
     pch=ifelse(teengamb$sex==0, 17, 16),                       #pch is shape of datapoints#
     col=ifelse(teengamb$sex==0, "blue", "hotpink"),            #col differentiates between male female#
     main="Relationship Between Income And Gambling",
     xlab="Income (Weekly)",
     ylab="Gambling Expenditures (Annually)")
abline(maleintercept, maleslope, col="blue", lwd=2, lty=2)      #add male regression line. lwd is width#
abline(femaleintercept, femaleslope, col="hotpink", lwd=2, lty=2)  #add female reg line. lty is dashed#
legend("topleft", legend=c("Male", "Female"),
       fill=c("blue", "hotpink"))
```

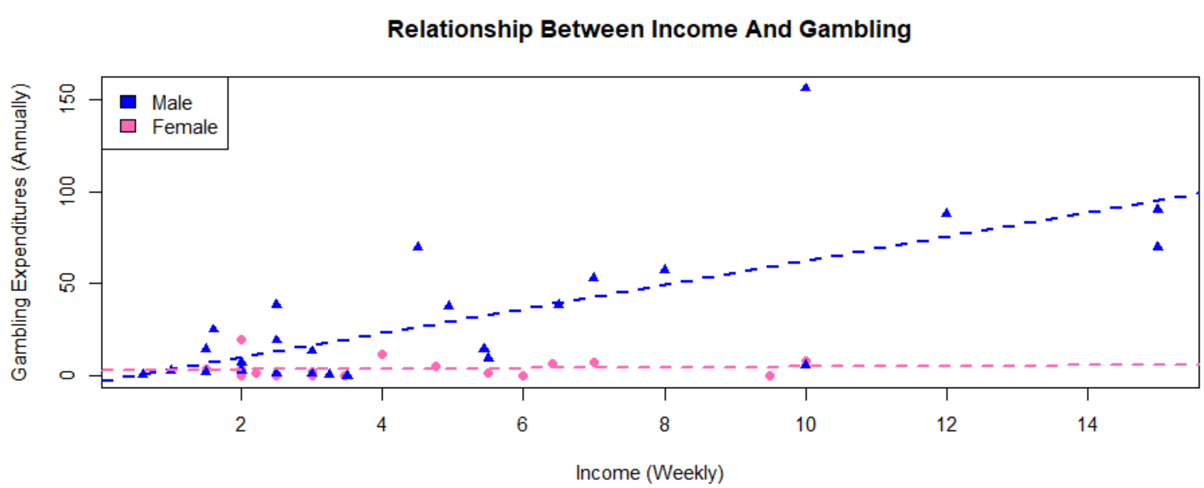

Figure 0.6: Income And Spending Regression By Sex

3) Consider the simple linear regression model $y_i = \beta_0 + x_i\beta_1 + \epsilon_i$ for $i = 1, \ldots, n$ where the x variable has been centered and scaled so that $\sum x_i = 0$ and $\sum x_i^2 = 1$.

a. Write the model matrix, X.

The model matrix is $\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$. This is multiplied by the regression coefficients $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and added to the error terms $\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$ to yield our predicted values.

b. Write the expression for $X^T X$ and solve the normal equations.

By the rules of matrix multiplication, $X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$. By assumption of the problem, this is the identity in $\mathbb{K}^{2\times 2}$, $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

The normal equation is $X^T X \underset{\sim}{\beta} = X^T \underset{\sim}{y}$. Since $X^T X$ was determined to be the identity, we have $\underset{\sim}{\beta} = X^T \underset{\sim}{y} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$.

R Code

```r
##########Written By Liam Flaherty For ST503 HW1##########
#####1. Load Required Packages#####
install.packages("faraway")
library(faraway)
summary(teengamb)                    #Get a glimpse of the data#
str(teengamb)                        #get a glimpse of the data#

#From ?faraway--
#sex is 0 male, 1 female,
#status is socioeconomic status score based on parents' occupation#
#income is in pounds per week#
#verbal is score out of 12 words#
#gamble is expenditure on gambling in pounds per year#

#####2. Simple Data Visualization#####
boxplot(teengamb$income, teengamb$gamble/12,                     #Show income and expenses together#
        names=c("Income (Per Week)",                             #Can't use xlab#
                "Gambling Expenditures (Per Month)"),
        ylab="British Pounds",
        col=c("green", "red"),
        main="Teenage Income And Spending Habits In Britian")

par(mfrow=c(1,2))                                                #Show 4 plots together#

plot(density(teengamb$status),                                   #Just for variety, hist likely better#
     main="Parental Socioeconomic Status",
     xlab="SES Score", ylab="Density")
polygon(density(teengamb$status), col="gold")                    #fill in for effect#

hist(teengamb$verbal, main="Verbal Definition Score (Out Of 12)",
     xlab="Score", ylab="Frequency", col="blue")
par(mfrow=c(1,1))                                                #Put plots back to normal#
```

```r
#####3. Regression#####
out=lm(teengamb$gamble ~ teengamb$income)                        #response (y) ~ predictor (x)#
out

#####4. Sex  Breakdown#####
number.males=sum(teengamb$sex==0)                                #count total males#
number.females=sum(teengamb$sex==1)                              #count total females#

male=data.frame(c("mean", "standard deviation"),                 #dummy column#
                rbind(                                           #stack rows on top of each other#
                sapply(teengamb[which(teengamb$sex==0),], mean), #mean of each column, filtered to males#
                sapply(teengamb[which(teengamb$sex==0),], sd)    #sd of each column, filtered to males#
                ))
names(male)=c(paste0("MALE", "(n=", number.males, ")"),          #rename first colummn#
              names(male)[2:ncol(male)])

female=data.frame(c("mean", "standard deviation"),                 #dummy column#
                  rbind(                                           #stack rows on top of each other#
                  sapply(teengamb[which(teengamb$sex==1),], mean), #mean of each column, filtered to females#
                  sapply(teengamb[which(teengamb$sex==1),], sd)    #sd of each column, filtered to females#
                  ))
names(female)=c(paste0("FEMALE", "(n=", number.females, ")"),      #rename first colummn#
                names(female)[2:ncol(female)])

male
female
```

```
#####5. Sex Differences Income/Spending Regression#####
maledf=teengamb[which(teengamb$sex==0),]                   #filter to only males#
femaledf=teengamb[which(teengamb$sex==1),]                 #filter to only females#

maleout=lm(maledf$gamble ~ maledf$income)                  #response(y) ~ predictor(x)#
maleslope=maleout[[1]][[2]]                                #just get numeric output#
maleintercept=maleout[[1]][[1]]                            #just get numeric output#

femaleout=lm(femaledf$gamble ~ femaledf$income)
femaleslope=femaleout[[1]][[2]]
femaleintercept=femaleout[[1]][[1]]

plot(teengamb$income, teengamb$gamble,                     #predictor(x), response(y)#
     pch=ifelse(teengamb$sex==0, 17, 16),                  #pch is shape of datapoints#
     col=ifelse(teengamb$sex==0, "blue", "hotpink"),       #col differentiates between male female#
     main="Relationship Between Income And Gambling",
     xlab="Income (Weekly)",
     ylab="Gambling Expenditures (Annually)")
abline(maleintercept, maleslope, col="blue", lwd=2, lty=2)          #add male regression line. lwd is width#
abline(femaleintercept, femaleslope, col="hotpink", lwd=2, lty=2)   #add female reg line. lty is dashed#
legend("topleft", legend=c("Male", "Female"),
       fill=c("blue", "hotpink"))
```