Linear Models HW # 2

Liam Flaherty

Professor Maity

NCSU: ST503-651

May 27, 2024

1) Consider the linear model with response vector $\boldsymbol{y} = [y_{1_1}\ y_{1_2}\ y_{1_3}\ y_{2_1}\ y_{2_2}\ y_{2_3}]^T$, parameter vector is $\boldsymbol{\beta} = [\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3]^T$, and model matrix $\boldsymbol{X}$ as follows:

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

a. What is the rank of $\boldsymbol{X}$?

Four. This can be seen by inspection, since the first column is the sum of the last three columns (which are all clearly independent), and since only one of the second or third columns can also be independent (the last three columns less the second column yields the third column).

b. Write the normal equations. Explain why the normal equations have infinitely many solutions.

xxxxxxx In general, the normal equation is $(X^TX)\beta = X^Ty$. Since X has less than full column rank (by part a), X^TX is not invertible. We must then use a generalized inverse to arrive at our estimate, $\underset{\sim}{\hat{\beta}} = (X^TX)^- X^Ty$. While $(X^TX)^-$ is not unique, $\underset{\sim}{\hat{\beta}}$ is.

c. Show that $\boldsymbol{\alpha_1 - \alpha_2}$ is estimable. Don't use any software.

A linear function $C^T\beta$ is estimable if and only if C is in the column space of X^T (C^T is in the row space of X). Here, our function is $C^T\beta = \alpha_1 - \alpha_2 = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}\beta$, so $C^T = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$. This is in the row space of X, since it is the first row of X less the fourth row of X; $\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \implies C^T = X_{\cdot,1} - X_{\cdot,4}$.

d. Show that $\boldsymbol{\beta_1 - 2\beta_2 + \beta_3}$ is estimable. Don't use any software.

In the same vein as part c, our function is $C^T\beta = \beta_1 - 2\beta_2 + \beta_3 = \begin{bmatrix} 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}\beta$. This is in the row space of X, since it is the first row of X, less two of the second row of X, plus the third row of X; $\begin{bmatrix} 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} - 2\begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$.

e. Use R to check your answers in part c and d above.

```
> #####2. Estimability######
> mymatrix=cbind(c(1,1,1,1,1,1), c(1,1,1,0,0,0),
+                c(0,0,0,1,1,1), c(1,0,0,1,0,0),
+                c(0,1,0,0,1,0), c(0,0,1,0,0,1))
>
> qr(mymatrix)$rank                               #4, as derived#
[1] 4
>
> cvec1=c(0,1,-1,0,0,0)                           #Question 1c#
> nb1=nonest.basis(mymatrix)                      #use 'estimability' package#
> is.estble(cvec1, nb1)
[1] TRUE
>
> cvec2=c(0,0,0,1,-2,1)                           #Question 1d#
> nb2=nonest.basis(mymatrix)                      #use 'estimability' package#
> is.estble(cvec2, nb2)
[1] TRUE
```

2) The dataset 'teengam' concerns a study of teenage gambling in Britian. Fit a regression model with the expenditure on gambling as the response and the sex, status, income, and verbal score as predictors.

a. Present the output. What percentage of variation in the response is explained by these predictors?

The predictors explain about 52.7% of the variation in the response.

```
> #####3. Gambling Regression#####
> gambling_regression=lm(gamble ~ ., data=teengamb)
> summary(gambling_regression)

Call:
lm(formula = gamble ~ ., 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
sex         -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

> summary(gambling_regression)$r.squared
[1] 0.5267234
```

b. Which observation has the largest (positive) residual? Give the case number.

The twenty-fourth person in the database had the largest positive residual. The model underestimated his predicted annual spending on gambling by over 94 pounds.

```
> ###3c. Residuals###
> gambling_yhat=predict(gambling_regression, teengamb)          #calculated automatically from lm()#
> gambling_yobs=teengamb$gamble
> residuals=data.frame(
+   names=paste0("Subject_", 1:nrow(teengamb)),
+   yhat=gambling_yhat,
+   yobs=gambling_yobs,
+   error=gambling_yobs-gambling_yhat
+ )
> residuals=residuals[order(-residuals$error),]
> residuals
        names       yhat   yobs       error
24 Subject_24 61.7477826 156.00  94.2522174
36 Subject_36 24.3948736  70.00  45.6051264
5   Subject_5 -9.9194692  19.60  29.5194692
37 Subject_37 17.9527471  38.50  20.5472529
```

c. Compute the mean and median of the residuals.

The mean residual was virtually zero. The median residual was about -1.45.

```
> mean(residuals$error)
[1] -1.014968e-14
> median(residuals$error)
[1] -1.451392
```

d. Compute the correlation of the residuals with the fitted values.

The residuals and fitted values were largely uncorrelated.

```
> cor(residuals$yhat, residuals$error)
[1] -1.312094e-17
> cor(teengamb[order(teengamb$gamble),"income"],
+     residuals[order(residuals$yobs),"error"])
[1] 0.02779156
```

e. Compute the correlation of the residuals with the income.

The correlation of the residuals with income was about 0.03.

```
> cor(residuals$yhat, residuals$error)
[1] -1.312094e-17
> cor(teengamb[order(teengamb$gamble),"income"],
+     residuals[order(residuals$yobs),"error"])
[1] 0.02779156
```

f. For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

Since male/female is binary (female is coded as '1'), if all other predictors are held constant, then the difference in prediction would be the estimate for the sex variable. In this case, the predictor was about -22.12, so a male would be predicted to spend about 22.12 pounds more annually compared to females.

```
> coef(gambling_regression)["sex"]
      sex
-22.11833
```

3) The dataset 'uswages' is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the resopnse and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?

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
```

```r
#####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"))
```