Linear Models HW # 2

Liam Flaherty

Professor Maity

NCSU: ST503-651

May 27, 2024

1) Consider the linear model with response vector $y = [y_{11}\ y_{12}\ y_{13}\ y_{21}\ y_{22}\ y_{23}]^T$, parameter vector is $\beta = [\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3]^T$, and model matrix 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 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.

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 $(X^TX)^-$ to arrive at our estimate, $\underset{\sim}{\hat{\beta}} = (X^TX)^-X^Ty$. And while $(X^TX)^-$ is not unique, $\underset{\sim}{\hat{\beta}}$ is.

c. Show that $\alpha_1 - \alpha_2$ 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 $C^T\beta = \alpha_1 - \alpha_2 = [0\ 1\ -1\ 0\ 0\ 0]\ \beta$, so $C^T = [0\ 1\ -1\ 0\ 0\ 0]$. This is in the row space of X, since it is the first row of X less the fourth row of X; $[0\ 1\ -1\ 0\ 0\ 0] = [1\ 1\ 0\ 1\ 0\ 0] - [1\ 0\ 1\ 0\ 1\ 0] \implies C^T = X_{\cdot,1} - X_{\cdot,4}$.

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

In the same vein as part c, our function is $C^T\beta = \beta_1 - 2\beta_2 + \beta_3 = [0\ 0\ 0\ 1\ -2\ 1]\ \beta$. 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; $[0\ 0\ 0\ 1\ -2\ 1] = [1\ 1\ 0\ 1\ 0\ 0] - 2\,[1\ 1\ 0\ 0\ 1\ 0] + [1\ 1\ 0\ 0\ 0\ 1]$.

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?

The model for the predicted weekly wages is approximately $50 for every year of education plus $10 for every year of experience, less about $240. More specifically, the linear model predicts that with every additional year of education, one would be expected to earn an additional $51.18 a week (in 1992 dollars, deflated by PCE inflation).

```
> #####4. Wage Regression#####
> wage_regression=lm(wage~ educ+exper, data=uswages)
> summary(wage_regression)

Call:
lm(formula = wage ~ educ + exper, data = uswages)

Residuals:
    Min      1Q  Median      3Q     Max
-1018.2  -237.9   -50.9   149.9  7228.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
educ          51.1753     3.3419  15.313  < 2e-16 ***
exper          9.7748     0.7506  13.023  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 427.9 on 1997 degrees of freedom
Multiple R-squared:  0.1351,    Adjusted R-squared:  0.1343
F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

In contrast, the same model but with logged weekly wages predicts every additional year of education would be expected to net a person about $0.09 log-dollars a week. While there is some argument to be made that the positive intercept in the log model makes an interpretation at the extremes more reasonable, in total, the interpretation makes more since when the prediction isn't logged.

```
> log_wage_regression=lm(log(wage)~ educ+exper, data=uswages)
> summary(log_wage_regression)

Call:
lm(formula = log(wage) ~ educ + exper, data = uswages)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7533 -0.3495  0.1068  0.4381  3.5699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.650319   0.078354   59.35   <2e-16 ***
educ        0.090506   0.005167   17.52   <2e-16 ***
exper       0.018079   0.001160   15.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6615 on 1997 degrees of freedom
Multiple R-squared:  0.1749,    Adjusted R-squared:  0.174
F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

R Code

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

#From ?faraway->teengamb
#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#

#From ?faraway->uswages#
#Wage is real (weekly) wage in dollars deflated by PCE to 1992 base#
#educ and exper is in years#
#race is 1 if black, 0 if white (no other races)#
#smsa is 1 if living in metro area, else 0#
#ne, mw, we, so are location (1 if live in ne, 0 else)#
#pt is 1 if part time, 0 if not#
```

```r
#####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#

cvec1=c(0,1,-1,0,0,0)                            #Question 1c#
nb1=nonest.basis(mymatrix)                       #use 'estimability' package#
is.estble(cvec1, nb1)

cvec2=c(0,0,0,1,-2,1)                            #Question 1d#
nb2=nonest.basis(mymatrix)                       #use 'estimability' package#
is.estble(cvec2, nb2)
```

```r
#####3. Gambling Regression#####
gambling_regression=lm(gamble ~ ., data=teengamb)      #response ~ predictors (all other columns)#
summary(gambling_regression)
summary(gambling_regression)$r.squared                 #variation explained by predictors#



###3a. Coefficients (automatic vs by hand just for practice)###
gambling_coef=gambling_regression$coefficients          #calculated from lm()$

x=model.matrix(gamble ~ ., data=teengamb)               #try to calculate by hand#
y=teengamb$gamble
manual_coef=solve(t(x) %*%x)%*%t(x)%*% y                #bhat is (X^TX)^{-1}X^T Y if X invertible#

compare_coef=round(data.frame(                          #theoretical matches results#
  manual_coef, gambling_coef),2)
compare_coef
```

```
###3b. Predictions (automatic vs by hand just for practice)###
gambling_yhat=predict(gambling_regression, teengamb)          #calculated from lm()#
predict_df=data.frame(
  name=paste0("Subject_", 1:nrow(teengamb)),
  gambling_yhat
)
predict_df=predict_df[order(-predict_df$gambling_yhat),]      #sort large to small#

manual_yhat=vector()                                          #initialize#
for(i in 1:nrow(teengamb)) {                                  #get sumproduct of coeff and obs#
  manual_yhat[i]=
    sum(teengamb[i,-which(names(teengamb)=="gamble")]*
      as.vector(gambling_coef)[-1])+
      as.vector(gambling_coef)[1]
}

compare_yhat=round(data.frame(gambling_yhat, manual_yhat),2)  #theoretical matches results#
compare_yhat
```

```
#3c. Residual SE (auto vs by hand just for practice)#
res=gambling_regression$residual
sse=sum(res^2)                        #or e^Te#
n=nrow(x)
r=qr(x)$rank
sigmasq=sse/(n-r)                     #estimate error variance with \sigma^2=(e^Te) / (n-rank(x))#
manual_sigma=sqrt(sigmasq)

manual_sigma
summary(gambling_regression)$sigma



###3d. Standard Errors (auto vs by hand just for practice; cov(beta)=sigma^2(X^TX)^{-1})###
gambling_se=summary(gambling_regression)$coefficients[, "Std. Error"]

sigma_hat=summary(gambling_regression)$sigma
manual_se=sigma_hat*sqrt(diag(solve(t(x)%*%x)))       #"by hand" calc#
manual_se                                             #SE(\beta_j)=\sigma\sqrt(\diag(X^TX^{-1}))#

compare_se=data.frame(gambling_se, manual_se)
compare_se
```

```
###3e. 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

mean(residuals$error)                                           #virtually 0#
median(residuals$error)                                         #about -1.45#



###3f. Correlations and Coefficients###
cor(residuals$yhat, residuals$error)                            #virtually 0#
cor(teengamb[order(teengamb$gamble),"income"],                  #about 0.03#
    residuals[order(residuals$yobs),"error"])

coef(gambling_regression)["sex"]
```

```
#####4. Wage Regression#####
wage_regression=lm(wage~ educ+exper, data=uswages)
wage_regression$coefficients
summary(wage_regression)

log_wage_regression=lm(log(wage)~ educ+exper, data=uswages)
summary(log_wage_regression)
```