ST 503 Homework 2

Eric Warren

2024-01-22

Contents

1 Problem 1

1.1 Part A

What is the rank of X?

We know that the matrix of $X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$. Now from this, we can solve it through row echelon form (not going to lie does not look fun) or we can look at the columns and find linear dependencies to see if we can reduce the rank. Looking at column 1, this is just adding column 4 + column 5 + column 6 ($a_1 = a_4 + a_5 + a_6$) which means column 1 is linear dependent and we can get rid of it. Now for column 2 this is also linear dependent. This is found from adding columns 4, 5, 6 together and subtracting from column 3 ($a_2 = a_4 + a_5 + a_6 - a_3$). Now we cannot get column 3 from any combinations of columns 4, 5, and 6 so this is linearly independent along with the last 3 columns. Therefore, we have 4 linearly independent columns and the rank(X) = 4. This can also be checked via software and calculators which also shows that rank(X) = 4.

1.2 Part B

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

We know that we are going to have infinite many solutions to the normal equations because we do not have full column rank. Now we can find the normal solutions by writing out what $(X^T X)\hat{\beta} = X^T y$. We know that $\hat{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$, $X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$, $X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{.1} \\ y_{.2} \\ y_{.3} \end{pmatrix}$, and $X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$. So our normal equation is $\begin{pmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{.1} \\ y_{.2} \\ y_{.3} \end{pmatrix}$, which gets us the following system of equations:

- $6\mu + 3\alpha_1 + 3\alpha_2 + 2\beta_1 + 2\beta_2 + 2\beta_3 = y_{..}$
- $3\mu + 3\alpha_1 + \beta_1 + \beta_2 + \beta_3 = y_{1.}$
- $3\mu + 3\alpha_2 + \beta_1 + \beta_2 + \beta_3 = y_{2.}$
- $2\mu + \alpha_1 + \alpha_2 + 2\beta_1 = y_{.1}$
- $2\mu + \alpha_1 + \alpha_2 + 2\beta_2 = y_{.2}$
- $2\mu + \alpha_1 + \alpha_2 + 2\beta_3 = y_{.3}$

Again, we know that we are going to have infinite many solutions to the normal equations because we do not have full column rank (or really full rank of the matrix). Moreover this means we can have a different $\hat{\beta}$ but still the same solution (for example our $\hat{y_{11}}$ value).

1.3 Part C

Show that $\alpha_1 - \alpha_2$ is estimable.

We know that $\alpha_1 - \alpha 2$ is the column vector $\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$. Note that we have that

$$e(X^T) = span\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \}$$

Now to get $\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ we can say that $\begin{pmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ which shows that $\alpha_1 - \alpha_2$ is estimable within the column space.

1.4 Part D

Show that $\beta_1 - 2\beta_2 + \beta_3$ is estimable.

We know that $\beta_1 - 2\beta_2 + \beta_3$ is the column vector $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{pmatrix}$. Note that we have that

$$e(X^T) = span\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \}$$

Now to get $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{pmatrix}$ we can say that $\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} - 2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ which shows that $\beta_1 - 2\beta_2 + \beta_3$ is estimable within the column space.

1.5 Part E

Use `R` to check your answers in part (C) and (D).

First we will show that $\alpha_1 - \alpha_2$ is estimable.

```
# Import library
library(estimability)

# Make matrix columns
col1 <- rep(1, 6)
col2 <- c(rep(1, 3), rep(0, 3))
col3 <- c(rep(0, 3), rep(1, 3))
col4 <- rep(c(1, 0, 0), 2)
col5 <- rep(c(0, 1, 0), 2)
col6 <- rep(c(0, 0, 1), 2)

# Make matrix
matrix <- matrix(c(col1, col2, col3, col4, col5, col6), nrow = 6)

# Is alpha1 - alpha2 estimable? TRUE means yes
is.estble(c(0, 1, -1, 0, 0, 0), nbasis = nonest.basis(matrix))
```

```
## [1] TRUE
```

As we can see, $\alpha_1 - \alpha_2$ is estimable.

Now we will show that $\beta_1 - 2\beta_2 + \beta_3$ is estimable.

```
# Is beta1 - 2 beta2 + beta3 estimable? TRUE means yes
is.estble(c(0, 0, 0, 1, -2, 1), nbasis = nonest.basis(matrix))
```

```
## [1] TRUE
```

As we can see, $\beta_1 - 2\beta_2 + \beta_3$ is estimable.

2 Problem 2

The dataset `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors.

```
# install.packages(c("faraway", "tidyverse")) # Uncomment if you do not have packages
library(faraway)
library(tidyverse)
(
  gambling <- as_tibble(teengamb) %>%
    mutate(sex = ifelse(sex == 0, "Male",
                        ifelse(sex == 1, "Female", "Error")))
)
```

```
## # A tibble: 47 x 5
##    sex    status income verbal gamble
##    <chr>   <int>  <dbl>  <int>  <dbl>
##  1 Female     51   2          8   0
##  2 Female     28   2.5        8   0
##  3 Female     37   2          6   0
##  4 Female     28   7          4   7.3
##  5 Female     65   2          8  19.6
##  6 Female     61   3.47       6   0.1
##  7 Female     28   5.5        7   1.45
##  8 Female     27   6.42       5   6.6
##  9 Female     43   2          6   1.7
## 10 Female     18   6          7   0.1
## # i 37 more rows
```

```
# Check to make sure the number of "Error" values is 0; want it to be TRUE
nrow(gambling %>% filter(sex == "Error")) == 0
```

```
## [1] TRUE
```

```
# Fit the model
gamblingFit <- lm(gamble ~ ., gambling)
```

2.1 Part A

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

```
summary(gamblingFit)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = gambling)
##
## 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)  0.43732   14.75429   0.030    0.9765
## sexMale     22.11833    8.21111   2.694    0.0101 *
## status       0.05223    0.28111   0.186    0.8535
## income       4.96198    1.02539   4.839 0.0000179 ***
## 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: 0.000001815
```

As we can see by looking at the Adjusted R-Squared, we can see this value is 0.4816495, which means that 48.1649452% of variation in the response is explained by these predictors.

2.2 Part B

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

Here we can find the entry with the largest positive residual.

```
# Get observation itself
gambling[which.max(gamblingFit$residuals), ]
```

```
## # A tibble: 1 x 5
##   sex   status income verbal gamble
##   <chr>  <int>  <dbl>  <int>  <dbl>
## 1 Male      27     10      4    156
```

```
# Get row number
which(gambling$gamble == 156)
```

```
## [1] 24
```

As we can see here our case number (or row / observation in the data set) is row number 24.

2.3 Part C

Compute the mean and median of the residuals.

```
mean(gamblingFit$residuals)
```

```
## [1] 0.00000000000000004039414
```

```
median(gamblingFit$residuals)
```

```
## [1] -1.451392
```

Here we can see that the mean of the residuals is 0 which makes sense since the expected value of residuals is 0. The median of the residuals, however, is -1.4513921, which shows that there could be high outliers in gambling that is driving our model to make decisions to include those values. Moreover, this means that the majority of our predicted values are greater than their corresponding actual gambling amounts.

2.4 Part D

Compute the correlation of the residuals with the fitted values.

```
cor(gamblingFit$fitted.values, gamblingFit$residuals)
```

```
## [1] 0.00000000000000003688059
```

Here we can see that our correlation between our fitted values and residuals for this model is about 0 which is basically saying there is no correlation whatsoever between what our fitted values are and what are residuals might be.

2.5 Part E

Compute the correlation of the residuals with the income.

```
cor(gamblingFit$model$income, gamblingFit$residuals)
```

```
## [1] 0.00000000000000006367805
```

Here we can see that our correlation between our income and residuals for this model is about 0 which is basically saying there is no correlation whatsoever between what our income is and what are residuals might be.

2.6 Part F

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

By looking at our output in **Part A**, we saw that our partial slope for the `sex` variable was 22.1183301, meaning that if all other predictors are held constant, the difference in predicted expenditure on gambling for a male would be about 22.1183301 more pounds per year compared to a female.

3 Problem 3

The dataset `uswages` is drawn as a sample from the Current Population Survey in 1988.

3.1 Part A

Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education.

```
# Make linear model
linearModel <- lm(wage ~ educ + exper, uswages)

# Show summary (output)
summary(linearModel)
```

```
##
## 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          0.00000178 ***
## educ          51.1753     3.3419  15.313 < 0.0000000000000002 ***
## exper          9.7748     0.7506  13.023 < 0.0000000000000002 ***
## ---
```

```
## 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: < 0.00000000000000022
```

Based on this model we can say that for every one additional year of education a person receives, we predict that their wage would go up by about 51.1752676 dollars. We can also say that for everyone one additional year of experience a person receives, we predict that their wage would go up by about 9.7747668 dollars.

3.2 Part B

Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education.

```
# Make linear model
logModel <- lm(log(wage) ~ educ + exper, uswages)

# Show summary (output)
summary(logModel)
```

```
##
## 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 <0.0000000000000002 ***
## educ        0.090506   0.005167   17.52 <0.0000000000000002 ***
## exper       0.018079   0.001160   15.58 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Based on this model we can say that for every one additional year of education a person receives, we predict that their wage would go up by about e raised to 0.0905063 or 1.0947284 dollars. We can also say that for everyone one additional year of experience a person receives, we predict that their wage would go up by about e raised to 0.0180786 or 1.018243 dollars.

3.3 Part C

Which interpretation is more natural?

It is more natural to look at a multiple linear regression model without any transformations and just say the coefficients since there is no additional math calculations needed to understand what the coefficient

values are in terms of increasing by one unit for the predictor has on the effect of the response. However, just because it is easier to say does not mean we should experiment with other models. Let us look at the adjusted R-squared values. For the multiple linear regression model we can see from **Part A** the adjusted R-square value is 0.1342524, which means that 13.4252388% of variation in the response is explained by these predictors. However with taking the log of the response with our multiple linear regression model we can see from **Part B** the adjusted R-square value is 0.1740342, which means that 17.4034156% of variation in the response is explained by these predictors. Generally, a higher adjusted R-square value indicates a better fit of the regression model to the data so we might have a hunch our latter model is better for prediction purposes, despite the former model being easier to communicate what the coefficients mean.