# HW2: OLS and Linear Algebra

Kaori Hirano

2023-06-14

## Packages

```
# load packages here
library(plyr)
suppressPackageStartupMessages(library(tidyverse))
```

## Data

```
# import data here
dcbikeshare <- read.csv('data/bikeshare-day.csv')
```

## Data Wrangling

### Q1

```
# code here
d <- dcbikeshare
# as.factor(d$season)
d$season <- mapvalues(d$season, from = c('1','2','3', '4'),
                      to = c("Winter", "Spring", "Summer", "Fall"))
# assigns strings to number values

d$season <- fct_relevel(d$season, c('Spring', 'Summer', 'Fall', 'Winter'))
#relevels with spring as baseline
```

```
# levels(d$season)
```

## Q2

```
# recodes 0 to no and 1 to yes
d$hbin <- mapvalues(d$holiday, from = c('0', '1'), to = c('no', 'yes'))
# cbind(head(d$hbin), head(d$holiday)) # checks accuracy
d$workbin <- mapvalues(d$workingday, from = c('0', '1'), to = c('no', 'yes'))
# cbind(d$workbin, d$workinday) # checks accuracy
```

## Q3

```
# recodes year to 2011 and 2012
d$years <- mapvalues(d$yr, from = c('0', '1'), to = c('2011', '2012'))
# cbind(d$years, d$yr) # checks accuracy
```

## Q4

```
# changes number values to corresponding weather types as strings
d$weathersit <- as.factor(d$weathersit)

 d <- d %>%
  mutate(weathersit = case_when(
    weathersit == "1"   ~ "clear",
    weathersit == "2"   ~ "mist",
    weathersit == "3"   ~ "light precipitation",
    weathersit == "4"   ~ "heavy precipitation",
    TRUE                ~ weathersit
  ))
# head(d$weathersit) check
```

### Q5

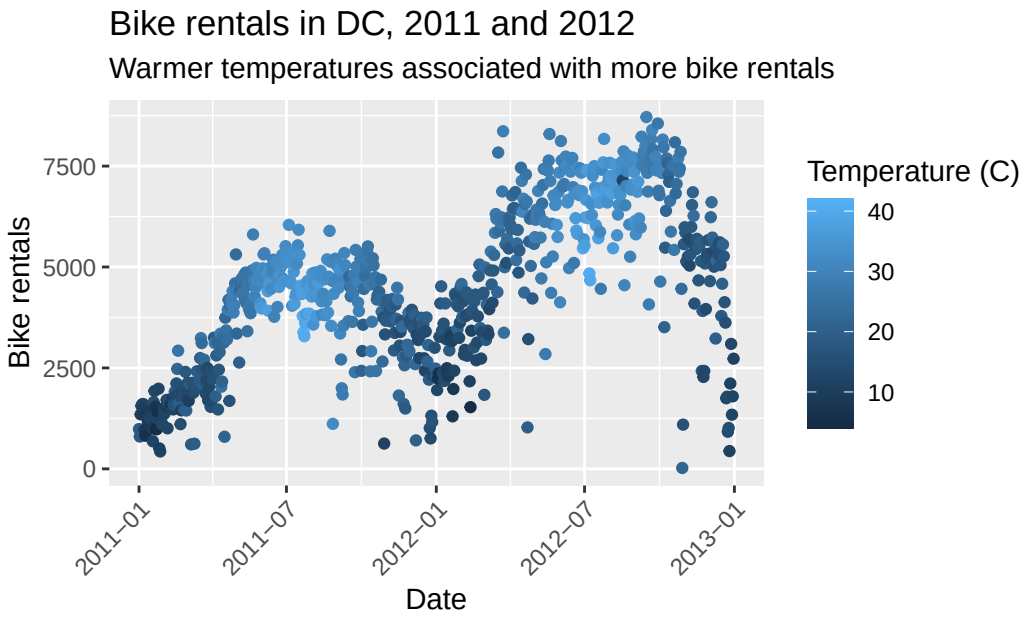
```
# multiplies normalized values by their stated maximum values to get raw values
d$raw_temp <- d$temp * 41
d$raw_ftemp <- d$atemp * 50
d$raw_hum <- d$hum * 100
d$raw_ws <- d$windspeed * 67
```

## Data Visualization

### Q6

```
d$dteday <- as.factor(d$dteday)
d$date <- as.Date(d$dteday)

# recreates image with raw feeling temp
ggplot(d, aes(x = date, y = cnt, color = raw_ftemp)) +
  geom_point() +
  labs(title = "Bike rentals in DC, 2011 and 2012",
       subtitle = "Warmer temperatures associated with more bike rentals",
       x = "Date", y = "Bike rentals",
       color = "Temperature (C)",
       caption = "Source: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




The graph supports that warmer temperatures tend to be associated with higher bike rentals. Bike rentals are lower in winter months, then trend upward in summer months before going back down. Overall, there do appear to be more rentals in 2012 than 2011, even though both years follow the more in warmer temperatures and less in cooler temperatures pattern.

## Modeling

### Q7

```
lm1 <- lm(cnt ~ raw_temp, d) # num rentals by raw temperature
summary(lm1)
```

```
Call:
lm(formula = cnt ~ raw_temp, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4615.3 -1134.9  -104.4  1044.3  3737.8
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1214.642    161.164   7.537 1.43e-13 ***
raw_temp     161.969      7.444  21.759  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1509 on 729 degrees of freedom
Multiple R-squared:  0.3937,    Adjusted R-squared:  0.3929
F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

```r
temp <- (63 - 32) * (5/9) # converts to celcius
new <- data.frame(raw_temp = (temp)) # makes dataframe for predict

predict(lm1, newdata = new, interval = 'confidence') #gets CI interval
```

```
    fit     lwr     upr
1 4004.1 3885.57 4122.63
```

```r
predict(lm1, newdata = new, interval = 'predict') # gets pred interval
```

```
    fit      lwr      upr
1 4004.1 1038.464 6969.737
```

a) There is a relationship between the total daily bike rentals and the daily temperature (b = 161.969, F(1,729) = 473.5, p < .0001)

b) The relationship seems to be moderate as seen by the adjusted R squared of .3929.

c) The relationship is positive as indicated by the positive R squared.

d) The predicted number of bike rentals with a temp of 63 degrees F is 4004. The associated 95% confidence interval is (3885.57, 4122.63) and prediction interval is (1038.464, 6969.737).

## Q8

```r
lm2 <- lm(cnt ~ raw_ftemp, d) # num rentals by raw temperature
summary(lm2)
```

```
Call:
lm(formula = cnt ~ raw_ftemp, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4598.7 -1091.6   -91.8  1072.0  4383.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  945.824    171.291   5.522 4.67e-08 ***
raw_ftemp    150.037      6.831  21.965  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1504 on 729 degrees of freedom
Multiple R-squared:  0.3982,    Adjusted R-squared:  0.3974
F-statistic: 482.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

```r
summary(lm1)
```

```
Call:
lm(formula = cnt ~ raw_temp, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4615.3 -1134.9  -104.4  1044.3  3737.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1214.642    161.164   7.537 1.43e-13 ***
raw_temp     161.969      7.444  21.759  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1509 on 729 degrees of freedom
Multiple R-squared:  0.3937,    Adjusted R-squared:  0.3929
F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

Feeling temperature is a better predictor of bike rentals, although not by large amount. Both have a statistically significant result ($p < .001$), but there is a lower standard error for feeling temperature and more importantly a larger F-statistic and adjusted R2.

## Q9

```r
# fit full model
# names(d)
lm_full <- lm(cnt ~ season + years + hbin + workbin + weathersit +
                raw_temp + raw_ftemp + raw_hum + raw_ws + (raw_ftemp * hbin), d)
summary(lm_full)
```

```
Call:
lm(formula = cnt ~ season + years + hbin + workbin + weathersit +
    raw_temp + raw_ftemp + raw_hum + raw_ws + (raw_ftemp * hbin),
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3675.0  -379.5    72.9   474.1  3341.2

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2715.141    268.534  10.111  < 2e-16 ***
seasonSummer                    -276.949    100.345  -2.760  0.00593 **
seasonFall                       409.792     96.165   4.261 2.30e-05 ***
seasonWinter                   -1130.561    113.544  -9.957  < 2e-16 ***
years2012                       2014.066     61.705  32.640  < 2e-16 ***
hbinyes                        -1384.379    495.409  -2.794  0.00534 **
workbinyes                       119.679     67.867   1.763  0.07826 .
weathersitlight precipitation -1907.149    207.547  -9.189  < 2e-16 ***
weathersitmist                 -420.244     81.286  -5.170 3.04e-07 ***
raw_temp                        102.997     34.015   3.028  0.00255 **
```

```
raw_ftemp                          18.762     30.456   0.616  0.53808
raw_hum                           -13.591      2.957  -4.596 5.09e-06 ***
raw_ws                            -40.639      6.491  -6.261 6.59e-10 ***
hbinyes:raw_ftemp                  34.440     20.625   1.670  0.09539 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 821.5 on 717 degrees of freedom
Multiple R-squared:  0.8234,    Adjusted R-squared:  0.8202
F-statistic: 257.1 on 13 and 717 DF,  p-value: < 2.2e-16
```

a) Yes, there is a relationship between the predictors and the response, as seen by the adjusted R-squared value of .8142 indicating a high level of correlation.

b) the predictors that appear to have a statistically significant relationship to the amount of bike rentals are the season, the year, holiday status, weather type, raw temperature, raw humidity, and raw windspeed.

c) the coefficients for the season suggest that there is a negative relationship for summer and winter with bike rentals and a positive relationship with fall when compared to a baseline level of spring's effect on rentals.

## Using Linear Algebra to Do Regression

### Q10

```r
X <- model.matrix(lm_full) # gets matrix
# head(X)
```

### Q11

```r
y <- model.frame(lm_full)$cnt # saying take this model, then subset to ONLY get count
```

### Q12

```r
(model <- (solve(t(X) %*% X)) %*% t(X) %*% y) # follows formula given in pdf
```

```
                                    [,1]
(Intercept)                     2715.14054
seasonSummer                    -276.94853
seasonFall                       409.79210
seasonWinter                   -1130.56069
years2012                       2014.06598
hbinyes                        -1384.37860
workbinyes                       119.67853
weathersitlight precipitation -1907.14902
weathersitmist                 -420.24376
raw_temp                        102.99712
raw_ftemp                        18.76168
raw_hum                         -13.59065
raw_ws                          -40.63921
hbinyes:raw_ftemp                34.44039
```

```r
# will give us a vector because we're getting the effects of each on y and y is one row
summary(lm_full)
```

```
Call:
lm(formula = cnt ~ season + years + hbin + workbin + weathersit +
    raw_temp + raw_ftemp + raw_hum + raw_ws + (raw_ftemp * hbin),
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-3675.0  -379.5    72.9   474.1  3341.2

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2715.141    268.534  10.111  < 2e-16 ***
seasonSummer                   -276.949    100.345  -2.760  0.00593 **
seasonFall                      409.792     96.165   4.261 2.30e-05 ***
seasonWinter                  -1130.561    113.544  -9.957  < 2e-16 ***
years2012                      2014.066     61.705  32.640  < 2e-16 ***
hbinyes                       -1384.379    495.409  -2.794  0.00534 **
workbinyes                      119.679     67.867   1.763  0.07826 .
weathersitlight precipitation -1907.149    207.547  -9.189  < 2e-16 ***
weathersitmist                 -420.244     81.286  -5.170 3.04e-07 ***
raw_temp                        102.997     34.015   3.028  0.00255 **
```

```
raw_ftemp                         18.762     30.456   0.616  0.53808
raw_hum                          -13.591      2.957  -4.596 5.09e-06 ***
raw_ws                           -40.639      6.491  -6.261 6.59e-10 ***
hbinyes:raw_ftemp                 34.440     20.625   1.670  0.09539 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 821.5 on 717 degrees of freedom
Multiple R-squared:  0.8234,    Adjusted R-squared:  0.8202
F-statistic: 257.1 on 13 and 717 DF,  p-value: < 2.2e-16
```

The intercept column from the summary stats is the same as the output from the matrix.