Homework Assignment 3

Isabella Chittumuri

4/26/2021

```
setwd("~/Documents/Hunter College/Spring 2021/Stat 707/HW")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

8.21

In a regression analysis of on-the-job head injuries of warehouse laborers caused by falling objects, Y is a measure of severity of the injury, X_1 is an index reflecting both the weight of the object and the distance it fell, and X_2 and X_3 are indicator variables for nature of head protection worn at the time of the accident, coded as follows:

Type of Protection	X_2	X_3
Hard hat	1	0
Bump cap	0	1
None	0	0

The response function to be used in the study is $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

a.

Develop the response function for each type of protection category.

Hard hat $(X_2 = 1, X_3 = 0)$:

$$
\begin{aligned}
E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
&= \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) \\
&= (\beta_0 + \beta_2) + \beta_1 X_1
\end{aligned}
$$

Bump hat ($X_2 = 0, X_3 = 1$):

$$
\begin{aligned}
E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
&= \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(1) \\
&= (\beta_0 + \beta_3) + \beta_1 X_1
\end{aligned}
$$

None ($X_2 = 0, X_3 = 0$):

$$
\begin{aligned}
E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\
&= \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) \\
&= \beta_0 + \beta_1 X_1
\end{aligned}
$$

b.

For each of the following questions, specify the alternatives H_0 and H_a for the appropriate test:

(1) With X_1 fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection?

$$
\begin{aligned}
H_0 &: \beta_3 = 0 \\
H_a &: \beta_3 \neq 0
\end{aligned}
$$

In words:

H_0, not wearing bump cap reduces expected severity of injury

H_a, wearing bump cap reduces expected severity of injury

(2) With X_1 fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?

$$
\begin{aligned}
H_0 &: \beta_2 = \beta_3 \\
H_a &: \beta_2 \neq \beta_3
\end{aligned}
$$

In words:

H_0, wearing hard hat and bump cap has same expected severity of injury

H_a, wearing hard hat and bump cap does not have same expected severity of injury

8.31

a.

Derive the expressions for $b_0', b_1', and b_{11}'$ in (8.12)

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X + b_{11} X^2 \\ &= b_0 + b_1(X - \bar{X}) + b_{11}(X - \bar{X})^2 \\ &= b_0 + b_1 X - b_1 \bar{X} + b_{11} X^2 - 2b_{11} X \bar{X} + b_{11} \bar{X}^2 \\ &= (b_0 - b_1 \bar{X} + b_{11} \bar{X}^2) + (b_1 - 2b_{11} \bar{X}) X + b_{11} X^2\end{aligned}$$

$$\begin{aligned}Hence: b_0' &= b_0 - b_1 \bar{X} + b_{11} \bar{X}^2 \\ b_1' &= b_1 - 2b_{11} \hat{X} \\ b_{11}' &= b_{11}\end{aligned}$$

b.

Using (5.46) obtain the variance-covariance matrix for the regression coefficients pertaining to the original X variable in terms of the variance-covariance matrix for the regression coefficients pertaining to the transformed x variable.

(5.46) $\sigma^2[W] = \sigma^2[AY] = A\sigma^2[Y]A'$

coefficients of: $b_0' = [1, -\bar{X}, \bar{X}^2]$ $b_1' = [0, 1, -2\bar{X}]$ $b_{11}' = [0, 0, 1]$

$$A = \begin{bmatrix} 1 & -\bar{X} & \bar{X}^2 \\ 0 & 1 & -2\bar{X} \\ 0 & 0 & 1 \end{bmatrix}$$

$$\sigma^2[b] = \begin{bmatrix} \sigma^2 b_0 & \sigma b_0 b_1 & \sigma b_0 b_2 \\ \sigma b_1 b_0 & \sigma^2 b_1 & \sigma b_1 b_2 \\ \sigma b_2 b_0 & \sigma b_2 b_1 & \sigma^2 b_2 \end{bmatrix}$$

$$A' = \begin{bmatrix} 1 & 0 & 0 \\ -\bar{X} & 1 & 0 \\ \bar{X}^2 & -2\bar{X} & 1 \end{bmatrix}$$

$$A\sigma^2[b]A' = \begin{pmatrix} 1 & -\bar{X} & \bar{X}^2 \\ 0 & 1 & -2\bar{X} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma^2 b_0 & \sigma b_0 b_1 & \sigma b_0 b_2 \\ \sigma b_1 b_0 & \sigma^2 b_1 & \sigma b_1 b_2 \\ \sigma b_2 b_0 & \sigma b_2 b_1 & \sigma^2 b_2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -\bar{X} & 1 & 0 \\ \bar{X}^2 & -2\bar{X} & 1 \end{pmatrix}$$

Matrix multiplication is associative: $ABC = (AB)C = A(BC)$

$$A\sigma^2[b] = \begin{bmatrix} \sigma^2 b_0 - \bar{X}\sigma b_1 b_0 + \bar{X}^2 \sigma b_2 b_0 & \sigma b_0 b_1 - \bar{X}\sigma^2 b_1 + \bar{X}^2 \sigma b_2 b_1 & \sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2 \sigma^2 b_2 \\ \sigma b_1 b_0 - 2\bar{X}\sigma b_2 b_0 & \sigma^2 b_1 - 2\bar{X}\sigma b_2 b_1 & \sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2 \\ \sigma b_2 b_0 & \sigma b_2 b_1 & \sigma^2 b_2 \end{bmatrix}$$

$$A\sigma^2[b]A' = \begin{bmatrix} \sigma^2 b_0 - \bar{X}\sigma b_1 b_0 + \bar{X}^2 \sigma b_2 b_0 & \sigma b_0 b_1 - \bar{X}\sigma^2 b_1 + \bar{X}^2 \sigma b_2 b_1 & \sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2 \sigma^2 b_2 \\ \sigma b_1 b_0 - 2\bar{X}\sigma b_2 b_0 & \sigma^2 b_1 - 2\bar{X}\sigma b_2 b_1 & \sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2 \\ \sigma b_2 b_0 & \sigma b_2 b_1 & \sigma^2 b_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -\bar{X} & 1 & 0 \\ \bar{X}^2 & -2\bar{X} & 1 \end{bmatrix}$$

Whole Matrix:

$$A\sigma^2[b]A' = \begin{bmatrix} \sigma^2 b_0 - \bar{X}\sigma b_1 b_0 + \bar{X}^2\sigma b_2 b_0 - \bar{X}(\sigma b_0 b_1 - \bar{X}\sigma^2 b_1 + \bar{X}^2 \sigma b_2 b_1) + \bar{X}^2(\sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2\sigma^2 b_2) & \sigma b_0 b_1 - \bar{X}\sigma^2 b_1 - \\ \sigma b_1 b_0 - 2\bar{X}\sigma b_2 b_0 - \bar{X}(\sigma^2 b_1 - 2\bar{X}\sigma b_2 b_1) + \bar{X}^2(\sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2) & \sigma^2 b_1 - \\ \sigma b_2 b_0 - \bar{X}(\sigma b_2 b_1) + \bar{X}^2(\sigma^2 b_2) & \end{bmatrix}$$

Separated First Column:

$$(A\sigma^2[b]A')_1 = \begin{pmatrix} \sigma^2 b_0 - \bar{X}\sigma b_1 b_0 + \bar{X}^2\sigma b_2 b_0 - \bar{X}(\sigma b_0 b_1 - \bar{X}\sigma^2 b_1 + \bar{X}^2 \sigma b_2 b_1) + \bar{X}^2(\sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2\sigma^2 b_2) \\ \sigma b_1 b_0 - 2\bar{X}\sigma b_2 b_0 - \bar{X}(\sigma^2 b_1 - 2\bar{X}\sigma b_2 b_1) + \bar{X}^2(\sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2) \\ \sigma b_2 b_0 - \bar{X}(\sigma b_2 b_1) + \bar{X}^2(\sigma^2 b_2) \end{pmatrix}$$

Separated Second Column:

$$(A\sigma^2[b]A')_2 = \begin{bmatrix} \sigma b_0 b_1 - \bar{X}\sigma^2 b_1 + \bar{X}^2\sigma b_2 b_1 - 2\bar{X}(\sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2\sigma^2 b_2) \\ \sigma^2 b_1 - 2\bar{X}\sigma b_2 b_1 - 2\bar{X}(\sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2) \\ \sigma b_2 b_1 - 2\bar{X}(\sigma^2 b_2) \end{bmatrix}$$

Separated Third Column:

$$(A\sigma^2[b]A')_3 = \begin{bmatrix} \sigma b_0 b_2 - \bar{X}\sigma b_1 b_2 + \bar{X}^2\sigma^2 b_2 \\ \sigma b_1 b_2 - 2\bar{X}\sigma^2 b_2 \\ \sigma^2 b_2 \end{bmatrix}$$

After simplifications:

$$\begin{aligned}
\sigma(b_0') &= \sigma_0^2 - 2\bar{X}\sigma_{01} + 2\bar{X}^2\sigma_{02} + \bar{X}^2\sigma_1^2 - 2\bar{X}^3\sigma_{12} + \bar{X}^4\sigma_2^2 \\
\sigma(b_1') &= \sigma_1^2 - 4\bar{X}\sigma_{12} + 4\bar{X}^2\sigma_2^2 \\
\sigma(b_2') &= \sigma_2^2 \\
\sigma(b_0', b_1') &= \sigma_{01} - 2\bar{X}\sigma_{02} + 3\bar{X}^2\sigma_{12} - \bar{X}\sigma_1^2 - 2\bar{X}^3\sigma_2^2 \\
\sigma(b_0', b_2') &= \sigma_{02} - \bar{X}\sigma_{12} + \bar{X}^2\sigma_2^2 \\
\sigma(b_1', b_2') &= \sigma_{12} - 2\bar{X}\sigma_2^2
\end{aligned}$$

where $\sigma_0^2 = \sigma^2 b_0, \sigma_{01} = \sigma b_0, b_1$, etc

8.34

In a regression study, three types of banks were involved, namely, commercial, mutual savings, and savings and loan. Consider the following system of indicator variables for type of bank:

Type of Bank	X_2	X_3
Commercial	1	0
Mutual Savings	0	1
Savings and loan	-1	-1

a.

Develop a first-order linear regression model for relating last year's profit or loss (Y) to size of bank (X_1) and type of bank (X_2, X_3).

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon$

b. State the response functions for the three types of banks.

Commercial ($X_2 = 1, X_3 = 0$):

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 \end{aligned}$$

Mutual savings ($X_2 = 0, X_3 = 1$):

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(1) \\ &= (\beta_0 + \beta_3) + \beta_1 X_1 \end{aligned}$$

Savings and loan ($X_2 = -1, X_3 = -1$):

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2(-1) + \beta_3(-1) \\ &= (\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1 \end{aligned}$$

c.

Interpret each of the following quantities: (1) β_2, (2) β_3, (3) $-\beta_2 - \beta_3$

(1) β_2 shows how much higher profit the commercial banks yield from the response function than for other banks of any size firm

(2) β_3 shows how much higher profit the mutual savings banks yield from the response function than for other banks of any size firm

(3) $-\beta_2 - \beta_3$ shows how much lower profit the savings loan banks yield from the response function than for other two banks of any size firm

8.39

Refer to the CDI data set in Appendix C.2. The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5).

a.

Fit a first-order regression model. Let $X_3 = 1$ if NE and 0 otherwise, $X_4 = 1$ if NC and 0 otherwise, and $X_5 = 1$ if South and 0 otherwise. In the event they are all 0, then it is West.

```r
CDI <- read.csv("CDI_Data.csv", header = F)

names(CDI) <- c("ID", "county", "state", "land_area", "total_pop", "precent_pop_18_34", "percent_pop_65

CDI$X3 <- ifelse(CDI$geographic_region == 1,1,0)
CDI$X4 <- ifelse(CDI$geographic_region == 2,1,0)
CDI$X5 <- ifelse(CDI$geographic_region == 3,1,0)
```

```r
mod <- lm(num_physicians ~ total_pop + total_income + X3 + X4 + X5, data = CDI)
summary(mod)
```

```
##
## Call:
## lm(formula = num_physicians ~ total_pop + total_income + X3 +
##     X4 + X5, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.075e+02  7.028e+01  -2.952  0.00332 **
## total_pop     5.515e-04  2.835e-04   1.945  0.05243 .
## total_income  1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3            1.490e+02  8.683e+01   1.716  0.08685 .
## X4            1.455e+02  8.515e+01   1.709  0.08817 .
## X5            1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

First order regression model: $\hat{Y} = -207.5 + 0.0005X_1 + 0.107X_2 + 149X_3 + 145.5X_4 + 191.2X_5$

b.

Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the northeastern region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

General formula:

$b_k \pm t(\frac{1-\infty}{2}; n-p)s(b_k)$

For comparison:

$$
\begin{aligned}
&(b_3 - b_4) \pm t(\frac{1-\infty}{2}; n-p)s(b_3 - b_4) \\
&149 - 145.5 \pm t(\frac{1-0.1}{2}; 440-6) * (86.83 - 85.15) \\
&3.5 \pm t(0.95; 434) * 1.68 \\
&3.5 \pm 1.645 * 1.68
\end{aligned}
$$

$$
\begin{aligned}
3.5 + 1.645(1.68) = 6.26 \approx 6.3 \\
3.5 - 1.645(1.68) = 0.73
\end{aligned}
$$

c.

Test whether any geographic effects are present;'use $\alpha = .10$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Alternatives:

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ $H_a : not all \beta_k = 0, (where k = 3, 4, 5)$

General Decision Rule

$$
\begin{aligned}
F^* \leq F(1-\alpha, df_R - df_F, df_F), \text{ fail to reject } H_0 \\
F^* > F(1-\alpha, df_R - df_F, df_F), \text{ reject } H_0
\end{aligned}
$$

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##                 Df     Sum Sq    Mean Sq   F value    Pr(>F)
## total_pop        1 1243181164 1243181164 3878.9792 < 2.2e-16 ***
## total_income     1   22058054   22058054   68.8256 1.369e-15 ***
## X3               1      21097      21097    0.0658   0.79764
## X4               1      23046      23046    0.0719   0.78871
## X5               1    1829483    1829483    5.7084   0.01731 *
## Residuals      434  139093455     320492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nrow(CDI)
```

```
## [1] 440
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = num_physicians ~ total_pop + total_income + X3 +
##     X4 + X5, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.075e+02  7.028e+01  -2.952  0.00332 **
## total_pop     5.515e-04  2.835e-04   1.945  0.05243 .
## total_income  1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3            1.490e+02  8.683e+01   1.716  0.08685 .
## X4            1.455e+02  8.515e+01   1.709  0.08817 .
## X5            1.912e+02  8.003e+01   2.389  0.01731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

$$
\begin{aligned}
&SSR(X_3, X_4, X_5|X_1, X_2) = 21097 + 23046 + 1829483 = 1873626 \\
&SSE(X_1, X_2, X_3, X_4, X_5) = 139093455
\end{aligned}
$$

$$
\begin{aligned}
F^* &= \frac{\frac{SSR(X_3,X_4,X_5|X_1,X_2)}{3}}{\frac{SSE(X_1,X_2,X_3,X_4,X_5)}{n-6}} \\
&= \frac{\frac{1873626}{3}}{\frac{139093455}{440-6}} \\
&= 1.9487
\end{aligned}
$$

```r
# 1-a, df, n-p (440-6)
qf(1-0.1, 3, 434)
```

```
## [1] 2.096449
```

$$1.9487 \leq 2.096449, \text{ fail to reject } H_0$$

```r
# another way to test full and reduced model
f_mod <- lm(num_physicians ~ total_pop + total_income + X3 + X4 + X5, data = CDI)
r_mod <- lm(num_physicians ~ total_pop + total_income, data = CDI)
anova(r_mod,f_mod, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: num_physicians ~ total_pop + total_income
## Model 2: num_physicians ~ total_pop + total_income + X3 + X4 + X5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    437 140967081
## 2    434 139093455  3   1873626 1.9487  0.121
```

P-value of F^* test = 0.121