Home Work 3

Daniel Lupercio

4/13/2021

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	X2	X3
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

When $X_2 = 1$ **and** $X_3 = 0$

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2(1)$$
$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

When $X_3 = 1$ **and** $X_2 = 0$

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_3(1)$$
$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$$

When $X_2 = 0$ **and** $X_3 = 0$

$$E[Y] = \beta_0 + \beta_1 X_1$$

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

(I) With X_1 **fixed, does wearing a bump cap reduce the expected severity of injury as compared with wearing no protection?** We state that H_0 : Wearing a bump cap reduces the expected severity of injury ($\beta_3 \geq 0$). Compared to H_a : Wearing no protection reduces the expected severity of injury ($\beta_3 < 0$).

(II) With X_1 **fixed, is the expected severity of injury the same when wearing a hard hat as when wearing a bump cap?** We state that H_0 : The expected severity of injury when wearing a hard hat is the same as when wearing a bump cap ($\beta_2 = \beta_3$). Compared to H_a : The expected severity of injury when wearing a hard hat is not the same as when wearing a bump cap ($\beta_2 \neq \beta_3$).

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).

```
CDI <- read.csv("/Users/daniel421/Desktop/STAT707/CDI_data2.csv", header = TRUE)
```

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 S and 0 otherwise

```
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)
```

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

```
##
## 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
```

$$\hat{Y} = -207.5 - 0.000515X_1 + 0.107X_2 + 149X_3 + 145.5X_4 + 191.2X_5$$

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

The t-value at $\alpha = 0.10$ and $Df = 434$:

```r
qt(0.95,434)
```

```
## [1] 1.648372
```

$$
\begin{aligned}
t(1-\alpha/2; n-p) &= \\
t(1-0.10/2, 440-6) &= \\
t(0.95, 434) &= \\
1.648372
\end{aligned}
$$

The Standard error for X3 = 86.83, while the standard error for X4 = 85.15

$s(b3 - b4) = 86.83 - 85.15 = 1.68$

So our 90% Confidence Interval is represented by:

$$
\begin{aligned}
\hat{Y} \pm t(1-\alpha/2; n-p) * s(\hat{Y}) \\
3.5 \pm (1.648372) * 1.68
\end{aligned}
$$

```r
3.5 + ((1.648372)*1.68)
```

```
## [1] 6.269265
```

```r
3.5 - ((1.648372)*1.68)
```

```
## [1] 0.730735
```

We state with 90% confidence the number of active physicians in the NC region differs from the NE region is (0.731, 6.269).

c. Test whether any geographic effects are present; use $\alpha = 0.10$. State the alternatives, decision rule and conclusion. What is the P-value of the test?

$$
\begin{aligned}
H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \\
H_a : not\ all\ \beta_3, \beta_4, \beta_5 = 0
\end{aligned}
$$

```r
CDI_fit_2 <- lm(num_physicians ~ total_pop + total_income, data = CDI);
summary(CDI_fit_2)
```

```
##
## Call:
## lm(formula = num_physicians ~ total_pop + total_income, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1849.1  -198.3   -71.4    39.7  3755.3
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.444e+01  3.283e+01  -1.963   0.0503 .
## total_pop     5.310e-04  2.775e-04   1.914   0.0563 .
## total_income  1.072e-01  1.297e-02   8.269 1.64e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 568 on 437 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8993
## F-statistic:  1961 on 2 and 437 DF,  p-value: < 2.2e-16
```

```r
anova(CDI_fit_2)
```

```
## Analysis of Variance Table
##
## Response: num_physicians
##               Df     Sum Sq    Mean Sq F value    Pr(>F)
## total_pop      1 1243181164 1243181164 3853.88 < 2.2e-16 ***
## total_income   1   22058054   22058054   68.38 1.638e-15 ***
## Residuals    437  140967081     322579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

It is worth noting that $SSR = (X3, X4, X5|X1, X2) = 140967081 - 139093455$.

$$F^* = \frac{140967081 - 139093455}{437 - 434} \div \frac{139093455}{434}$$

$$F^* = 1.9487$$

$$F^* \leq F(1-\alpha, df_R - df_F, df_F),\ concude\ H_0$$
$$F^* > F(1-\alpha, df_R - df_F, df_F),\ concude\ H_A$$

$$F(0.90, 3, 434) = 2.096$$

Our F-statistic, F^* is $\leq$ to $F(0.90, 3, 434)$. Thus, we fail to reject the null hypothesis and conclude that $\beta_3 = \beta_4 = \beta_5 = 0$. With a p-value of 0.121

```
## [1] 0.1210319
```

HW3

Saturday, April 24, 2021 2:03 PM

2

8.31 a. $\hat{Y} = b_0 + b_1 x + b_{11} x^2$ (8.11) Centered

$(X^2 - 2X\bar{X} + \bar{X}^2) b_{11}$

$b_{11} X^2 - 2 b_{11} X \bar{X} + b_{11} \bar{X}^2$

$\hat{Y} = b_0' + b_1' X + b_{11}' X^2$ (8.12) original

$= b_0 + b_1 (X - \bar{X}) + b_{11} (X - \bar{X})^2$

$= b_0 + b_1 X - b_1 \bar{X} + b_{11} X^2 - 2 b_{11} X \bar{X} + b_{11} \bar{X}^2$

Let $b_0' = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2$

$b_1' = b_1 - 2 b_{11} \bar{X}$

$b_{11}' = b_{11}$

and substitute these back into 8.12 ✓

b. coefficients of $b_0' = [1, -\bar{X}, \bar{X}^2]$

$b_1' = [0, 1, -2\bar{X}]$

$b_{11}' = [0, 0, 1]$

$\sigma^2\{W\} = \sigma^2\{AY\} = A \sigma^2\{Y\} A'$

$A = \begin{bmatrix} 1 & -\bar{X} & \bar{X}^2 \\ 0 & 1 & -2\bar{X} \\ 0 & 0 & 1 \end{bmatrix}$ $\sigma^2\{b\} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01}^2 & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{bmatrix}$

$b_0' = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2$

$b_1' = b_1 - 2 b_{11} \bar{X}$ $\Rightarrow$

$b_{11}' = b_{11}$

$$\underset{1\times3}{\begin{bmatrix} b_0' \\ b_1' \\ b_1' \end{bmatrix}} = \underset{3\times1}{\begin{bmatrix} b_0 - b_1\bar{X} + b_{11}\bar{X}^2 \\ b_1 - 2 b_{11} \bar{X} \\ b_{11} \end{bmatrix}}$$

B'

$$\underset{3\times3}{\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}}^{A} \cdot \underset{3\times1}{\begin{bmatrix} b_0 \\ b_1 \\ b_{11} \end{bmatrix}}^{B}$$

5.46

$\sigma^2(B') = \sigma(w) = \sigma(A, B) = A \sigma^2(B) A'$ But what is A?

$b_1 - 2 b_{11} \bar{X} = \underbrace{b_1 \left(1 - \frac{2\bar{X} b_{11}}{b_1}\right)}_{a_2}$

$b_0 - b_1 \bar{X} + b_{11} \bar{X}^2 = \underbrace{b_0 \left(1 - \frac{b_1\bar{X} + b_{11}\bar{X}^2}{b_0}\right)}_{a_1}$

$$A = \begin{bmatrix} b_0\left(1 - \frac{b_1\bar{X} + b_{11}\bar{X}^2}{b_0}\right) & 0 & 0 \\ 0 & b_1\left(1 - \frac{2\bar{X}b_{11}}{b_1}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus from 5.46 : $\sigma^2(B') = A\{\sigma^2(B)\}A'$

$$\begin{bmatrix} b_0\left(1 - \frac{b_1\bar{X} + b_{11}\bar{X}^2}{b_0}\right) & 0 & 0 \\ 0 & b_1\left(1 - \frac{2\bar{X}b_{11}}{b_1}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_{11} \end{bmatrix} \begin{bmatrix} b_0\left(1 - \frac{b_1\bar{X} + b_{11}\bar{X}^2}{b_0}\right) & 0 & 0 \\ 0 & b_1\left(1 - \frac{2\bar{X}b_{11}}{b_1}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Homework 3

Wednesday, April 14, 2021 2:33 PM

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).

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

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

If $X_2 = 1$, we have a commericial bank and $X_3 = 0$

$$E[Y] = (\beta_0 + \beta_2) + \beta_1 X_1$$

If $X_3 = 1$, we have a mutual savings bank and $X_2 = 0$

$$E[Y] = (\beta_0 + \beta_3) + \beta_1 X_1$$

If $X_2, X_3 = -1$ we have a savings and loan bank

$$E[Y] = (\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1$$

c. Interpret each of the following quantities

(1) β_2

The amount of profit or loss from the response function when X_2 is coded 1 (commercial bank) and X_3 is coded 0.

(2) β_3

The amount of profit or loss from the response function when X_3 is coded 1 (Mutual savings bank) and X_2 is coded 0.

(3) $-\beta_2 - \beta_3$

The amount of profit or loss from the response function when looking at a "savings and loan bank."