Yu Tian_Mid_Pstat120C

Yu Tian

2022-08-19

```r
# package
library(tidyverse)
library(ggplot2)

# set seed
set.seed(0623)

# Read (and import) the full data set into R using read.csv()
data <- read.csv(file = 'data.csv')

# view the data example in R
data
```

```
##    manufacturer   model_year      mpg   weight displacement class horsepower
## 1           Kia         2015 21.54716 4124.129     178.5575 Truck   237.6715
## 2          Ford         2007 17.02911 4736.041     236.0139 Truck   236.3425
## 3         Mazda         2003 19.33781 3777.898     179.4107 Truck   186.9290
## 4            GM         1986 23.02399 3174.024     190.2972   Car   115.2296
## 5           BMW         2017 22.54566 4650.112     164.4554 Truck   284.4615
## 6           Kia Prelim. 2021 32.38923 3194.868     114.4701   Car   161.8138
## 7           All         2000 22.51440 3400.909     168.2990   Car   168.2936
## 8        Nissan         2014 22.18444 4458.683     208.4433 Truck   245.0790
## 9      Mercedes         1999 21.50476 3879.585     197.3525   Car   222.9132
## 10       Subaru         2012 27.21958 3450.740     137.7964   Car   170.3608
## 11           VW         1990 23.73371 2929.358     122.0215   Car   118.9145
## 12       Toyota         1998 24.57349 3304.248     142.4937   Car   151.3054
## 13       Toyota         2007 19.09633 4461.215     218.8619 Truck   229.8358
## 14           GM         2002 15.44052 4987.675     302.1571 Truck   257.2493
## 15         Ford         1988 16.42429 4357.654     239.6896 Truck   149.6339
```

Question 1

1. Answer the following based on a simple linear regression, predicting $mpg(y)$ with $weight(x_1)$.

(a) Fit the specified model. Write the model equation, including your estimates.

Answer Fit the simple linear regression model $Y = \beta_0 + \beta_1 x_1 + \epsilon$ to the data.

```r
fit = lm(mpg~weight, data=data)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4600 -2.1210 -0.6158  1.6716  7.0659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.267655   5.038457   7.992 2.26e-06 ***
## weight      -0.004678   0.001267  -3.692  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.131 on 13 degrees of freedom
## Multiple R-squared:  0.5119, Adjusted R-squared:  0.4744
## F-statistic: 13.63 on 1 and 13 DF,  p-value: 0.002709
```

From the table above, we can get

$$\hat{\beta}_0 = 40.2677$$

$$\hat{\beta}_1 = -0.004678$$

Thus, the model equations is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \epsilon = 40.2677 - 0.004678 x_1 + \epsilon$$

Using the method of least squares,

```r
mean_y = mean(data$mpg)
mean_y
```

```
## [1] 21.9043
```

```r
mean_x1 = mean(data$weight)
mean_x1
```

```
## [1] 3925.809
```

```r
# n = 15
S_xy = sum(data$weight * data$mpg) - (sum(data$weight) * sum(data$mpg)/15)
S_xy
```

```
## [1] -28575.53
```

```r
S_xx = sum((data$weight)^2) - (sum(data$weight))^2 / 15
S_xx
```

```
## [1] 6109018
```

```r
S_yy = sum((data$mpg)^2) - (sum(data$mpg))^2 / 15
S_yy
```

```
## [1] 261.1102
```

```r
B_1 = S_xy / S_xx
B_1
```

```
## [1] -0.004677598
```

```r
B_0 = mean_y - B_1 * mean_x1
B_0
```

```
## [1] 40.26766
```

So,

$$\bar{y} = 21.9034$$

$$\bar{x_1} = 3925.809$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = -28575.53$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 610908$$

Then,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{8575.53}{610908} = -0.004678$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 40.2677$$

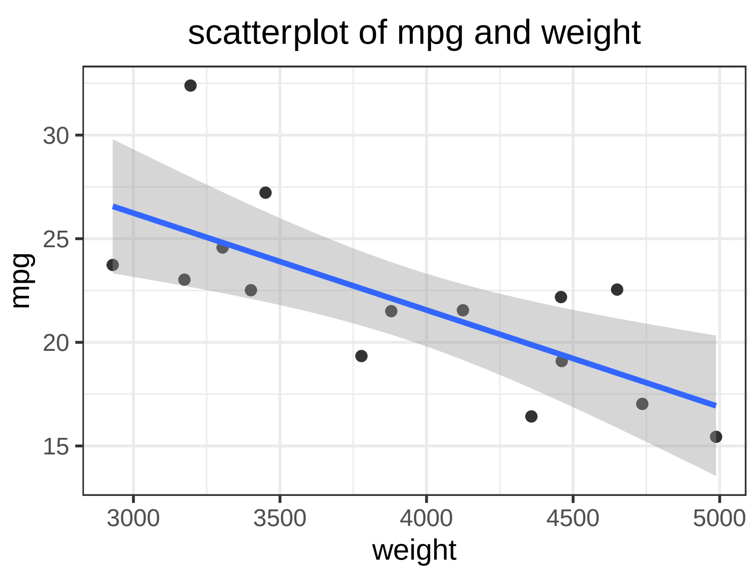
Thus, the model equations is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \epsilon = 40.2677 - 0.004678x_1 + \epsilon$$

(b) Create a scatterplot of mpg and weight. Add a line representing the model, with 95% confidence bands. Does the model appear to fit the data?

```
data %>%
  ggplot(aes(x=weight, y=mpg)) +
  geom_point(alpha = 0.8) +
  labs(title = 'scatterplot of mpg and weight',
       x='weight', y='mpg') +
  geom_smooth(method = 'lm', formula = 'y ~ x', lty = 1, level=0.95) +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

Answer



From the graph above, we can find that the 95% confidence bands covered most data points, but NOT all points are covered. In conclusion, the model appears to fit the data with most data in confidence bands, but it does not fit all data very well.

(c) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Do not interpret the conclusion of this test.

Answer Test of Hypothesis for β_1:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

From the question above, we get $\hat{\beta}_1 = -0.0047$, $S_{xx} = 610908$, $S_{xy} = -28575.53$, $S_{yy} = 261.1102$.

Then,

```
SSE = S_yy - B_1 * S_xy
SSE
```

```
## [1] 127.4454
```

```
S_2 = (1/(15-2)) * SSE
S_2
```

```
## [1] 9.80349
```

```
S_s = sqrt(S_2)
S_s
```

```
## [1] 3.131053
```

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy} = 127.4454$$

$$S^2 = (\frac{1}{n-2})SSE = 9.8035$$

$$s = \sqrt{S^2} = 3.1311$$

Because we interested in the parameter β_1, we need the value

$$c_{11} = \frac{1}{S_{xx}} = \frac{1}{610908}$$

Then,

```
t_value = B_1/(S_s*sqrt(1/S_xx))
t_value
```

```
## [1] -3.692481
```

$$t = \frac{\hat{\beta}_1 - \beta_1}{s\sqrt{c_{11}}} = \frac{-0.0047 - 0}{3.1311 \cdot \sqrt{\frac{1}{610908}}} = -3.6925$$

(Besides, we could get the value of $t = -3.692$ from the table above.)

SSE is based on $df = 15 - 2 = 13$. If we take $\alpha = 0.05$, the value of $t_{\alpha/2} = t_{0.025}$ for 13 df is $t_{0.025} = 2.16$, and the rejection region is

$$\text{reject if } \; |t| >= 2.16.$$

Since 3.6925 is larger than 2.16, we reject the null hypothesis that $\beta_1 = 0$.

Then,

```
p_value = 2*pt(q=-3.6925, df=13, lower.tail=TRUE)
p_value
```

```
## [1] 0.002708503
```

(Beside, we can get the p-value = 0.00271 from the table above.)

With $t = -3.6925$, the p-value is 0.002709, which is obviously less than $\alpha = 0.05$, so we reject the null hypothesis that $\beta_1 = 0$.

Question 2

2. Answer the following based on a multiple linear regression, predicting mpg with weight ($x1$) and engine displacement ($x2$).

(a) Fit the specified model. Write the model equation, including your estimates.

Answer Fit the multiple linear regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to the data.

```
fit2 = lm(mpg~weight+displacement, data=data)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ weight + displacement, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1342 -0.9828 -0.6934  1.4039  5.0779
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.5095516  3.8852963   9.397 6.98e-07 ***
## weight       -0.0003083  0.0015820  -0.195    0.849
## displacement -0.0717513  0.0209294  -3.428    0.005 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 12 degrees of freedom
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7123
## F-statistic: 18.33 on 2 and 12 DF,  p-value: 0.0002248
```

From the table above, we can get

$$\hat{\beta_0} = 36.5096$$
$$\hat{\beta_1} = -0.0003083$$
$$\hat{\beta_2} = -0.07175$$

Thus, the model equations is

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = 36.5095 - 0.0003083 x_1 - 0.07175 x_2 + \epsilon$$

(b) Test the null hypothesis that the slope of x_1, β_1, is equal to zero. State the hypotheses, test statistic, rejection region(s), and p-value. Interpret the conclusion of this test at $\alpha = 0.05$.

Answer Test of Hypothesis for β_1:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

From the table above, we get $\hat{\beta_1} = -0.0003083$ and the value of $t = -0.195$.

Then, since n = 15, k (the number of independent variables in the complete model) = 2, so df (the degrees of freedom) $= n - k - 1 = 15 - 2 - 1 = 12$.

Thus, based on $df = 12$. If we take $\alpha = 0.05$, the value of $t_{\alpha/2} = t_{0.025}$ for 12 df is $t_{0.025} = 2.179$.

The rejection region is

$$\text{reject if } \; |t| >= 2.179$$

Since 0.195 is smaller than 2.179, we fail to reject the null hypothesis that $\beta_1 = 0$.

Then,

```
p_value = 2*pt(q=-0.195, df=12, lower.tail=TRUE)
p_value
```

```
## [1] 0.8486555
```

(Beside, we can get the p-value = 0.849 from the table above.)

With $t = -0.195$, the p-value is 0.8487, which is obviously larger than $\alpha = 0.05$, so we fail to reject the null hypothesis and conclude that the slope is of x_1, β_1 is equal to zero. This means there is not a statistically significant relationship between weight and mpg.

(c) Consider $x_1^* = 3000$ and $x_2^* = 150$. Calculate a 95% confidence interval for $E[Y|x_1 = x_1^*, x_2 = x_2^*]$. Calculate a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$. Interpret both of these intervals in context.

Answer

- Confidence interval:

```
new_data = data.frame(weight=3000,displacement=150)
confidence_interval = predict(fit2,newdata = new_data, interval='confidence', level=0.95)
confidence_interval
```

```
##        fit      lwr      upr
## 1 24.82209 22.35674 27.28745
```

By calculation,

```
# Y
Y = data$mpg
# Y'
Y_transpose = t(Y)
# X
x0 = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
X = cbind(x0,data$weight,data$displacement)
# X'
X_transpose = t(X)
# the inverse of X'X
XX_inverse = solve(X_transpose%*%X)
# beta_hat = (X'X)^-(1)X'Y
beta_hat = XX_inverse %*% X_transpose %*% Y
# SSE = Y'Y-(beta_hat)'X'Y
SSE2 = Y_transpose%*%Y - t(beta_hat)%*%X_transpose%*%Y
# S^2=SSE/[n-(k+1)]
n = 15
k = 2
S2 = SSE2/(n-(k+1))
S = sqrt(S2)
```

```
# a and a'
a = c(1,3000,150)
a_transpose = t(a)
# t_(alpha/2)
t = 2.179
# confidence interval
CI1 = a_transpose%*%beta_hat + t*S*sqrt(a_transpose%*%XX_inverse%*%a)
CI2 = a_transpose%*%beta_hat - t*S*sqrt(a_transpose%*%XX_inverse%*%a)
CI1
```

```
##          [,1]
## [1,] 27.28766
```

```
CI2
```

```
##          [,1]
## [1,] 22.35653
```

Thus, a 95% confidence interval

$$a'\hat{\beta} + t_{\frac{\alpha}{2}} S\sqrt{a'(X'X)^{-1}a} = 27.2876$$

$$a'\hat{\beta} - t_{\frac{\alpha}{2}} S\sqrt{a'(X'X)^{-1}a} = 22.3565$$

From the output (from function) above, a 95% confidence interval for $E[Y|x_1 = x_1^*, x_2 = x_2^*]$ the fitted mpg with (weight) $x_1^* = 3000$ and (distance)$x_2^* = 150$ is about 24.8221. The confidence interval of (22.3567, 27.2875) signifies the range in which the true mpg lies at a 95% level of confidence.

- Prediction Interval:

```
prediction_interval = predict(fit2,newdata = new_data, interval='prediction', level=0.95)
prediction_interval
```

```
##        fit      lwr      upr
## 1 24.82209 19.20524 30.43895
```

```
# prediction interval
PI1 = a_transpose%*%beta_hat + t*S*sqrt(1+a_transpose%*%XX_inverse%*%a)
PI2 = a_transpose%*%beta_hat - t*S*sqrt(1+a_transpose%*%XX_inverse%*%a)
PI1
```

```
##          [,1]
## [1,] 30.43943
```

```
PI2
```

```
##          [,1]
## [1,] 19.20476
```

Thus, a 95% prediction interval

$$a'\hat{\beta} + t_{\frac{\alpha}{2}} S\sqrt{1 + a'(X'X)^{-1}a} = 30.4394$$

$$a'\hat{\beta} - t_{\frac{\alpha}{2}} S\sqrt{1 + a'(X'X)^{-1}a} = 19.2048$$

From the output (of function) above, a 95% prediction interval for y_i, given $x_1 = x_1^*$ and $x_2 = x_2^*$, the fitted mpg with (weight) $x_1^* = 3000$ and (distance)$x_2^* = 150$ is about 24.8221. The prediction interval of (19.2052, 30.4389) signifies the range in which the next mpg lies at a 95% level of prediction. Notice the fitted value is the same as before, but the interval is wider. Since the prediction interval must take into account the variability of the estimators for μ and σ, the interval will be wider.

(d) Which model constitutes the "complete" model and which the "reduced" model? Can x_2 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

Answer The "complete" model is (question 2(a))

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon = 36.5095 - 0.0003083 x_1 - 0.07175 x_2 + \epsilon$$

and the "reduced" model is (question 1(a))

$$Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \epsilon = 40.2677 - 0.004678 x_1 + \epsilon$$

.

Test of Hypothesis for β_2:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

```
F_test = anova(fit, fit2)
F_test
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ weight
## Model 2: mpg ~ weight + displacement
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     13 127.445
## 2     12  64.386  1     63.06 11.753 0.005002 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table above, we get $SSE_R = 127.445$, $SSE_C = 64.386$, and $F = 11.753$.

Thus, the F-statistic is

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{(SSE_C)/(n-(k+1))} = \frac{(127.445 - 64.386)/(2-1)}{(64.386)/(15-(2+1))} = 11.753$$

The tabulated F-value for $\alpha = 0.05$ with $v = 2 - 1 = 1$ numerator *df* and $v2 = 15 - (2+1) = 12$ denominator *df* is 4.7472.

Since $11.753 > 4.7472$, so $F > F_\alpha$, which is the appropriate rejection region, so we reject the null hypothesis $H_0 : \beta_2 = 0$.

From the table above, we can get the p-value is 0.005002, which is obviously smaller than $\alpha = 0.05$, so we reject the null hypothesis and conclude that the slope is of x_2, β_2 is not equal to zero. Thus, x_2 can NOT be dropped from the model without losing predictive information. This means there is a statistically significant relationship between displacement and mpg.

Question 3

3. Consider your answers to the previous questions, then answer the following. Suppose that the true population relationship is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Further suppose that there is a relationship between x_1 and x_2, given by:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

where γ_1 and β_2 are non-zero.

(a) Find the expected values of β_0 and β_1 if the independent variable x_2 is omitted from the regression.

Answer Given

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Since

$$x_2 = \gamma_0 + \gamma_1 x_1 + \delta$$

so

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\gamma_0 + \gamma_1 x_1 + \delta) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2\gamma_0 + \beta_2\gamma_1 x_1 + \beta_2\delta + \epsilon \\ &= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_1 + \beta_2\delta + \epsilon \end{aligned}$$

Thus, the expected values of β_0 is

$$E[\beta_0] = \beta_0 + \beta_2\gamma_0$$

and the expected values of β_1 is

$$E[\beta_1] = \beta_1 + \beta_2\gamma_1$$

(b) Calculate the bias (if any) of β_0 and β_1 when x_2 is omitted.

Answer According to the definition of the bias,

bias of $\beta_0 = E[\beta_0] - \beta_0 = \beta_0 + \beta_2\gamma_0 - \beta_0 = \beta_2\gamma_0$

bias of $\beta_1 = E[\beta_1] - \beta_1 = \beta_1 + \beta_2\gamma_1 - \beta_1 = \beta_2\gamma_1$

(c) What values of γ_1 and β_2 would result in β_0 and β_1 remaining unbiased?

Answer From 3(b), we get bias of $\beta_0 = \beta_2\gamma_0$ and bias of $\beta_1 = \beta_2\gamma_1$.

To make β_0 and β_1 remain unbiased, which means the bias of β_0 and the bias of β_1 are all equal to zero.

Thus,

$$\beta_2 = 0$$

and

$$\gamma_1 = 0$$

(d) In light of the above:

i. What assumption of linear regression is being violated in Question 1? Is this assumption met in Question 2?

ii. In Question 1, are the estimates of β_0 and β_1 BLUE? Why or why not?

Answer

i. The assumption of $E[\epsilon] = 0$ is being violated in Question 1. This assumption is met in Question 2.

ii. No, the estimates of β_0 and β_1 are NOT BLUE, since NOT all assumptions are satisfied.