PSTAT 120C HW 2

TJ Sipin

2022-08-17

Reading

Write the general equation for a multiple linear regression model.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Write the least-squares estimation for a multiple linear regression in matrix form.

$$
(X'X)\hat{\beta} = X'Y
$$

$$
\begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \dots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \dots & \sum x_{i1}x_{ip} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \dots & \sum x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ip} & \sum x_{ip}x_{i1} & \sum x_{ip}x_{i2} & \dots & \sum x_{ip}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \vdots \\ \sum x_{ip} y_i \end{bmatrix}
$$

$$
\hat{\beta} = (X'X)^{-1}X'Y
$$

State the test statistic and confidence interval formulas for a linear function of parameters in multiple linear regression.

Test statistic formula to test the hypothesis $H_0 : \mathbf{a}'\beta = (\mathbf{a}'\beta)_0$:

$$T = \frac{\mathbf{a}'\hat{\beta} - \mathbf{a}'\hat{\beta}}{S\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

Confidence interval formula:

$$\mathbf{a}'\hat{\beta} \pm t_{\alpha/2} S \sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X}^{-1})\mathbf{a}}.$$

Describe the general process of testing the hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.

We define the reduced and complete models below:

$$
\begin{aligned}
&\text{model R:} Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g + \epsilon \\
&\text{model C:} Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \beta_{g+2} x_{g+2} + \cdots + \beta_k x_k + \epsilon.
\end{aligned}
$$

If the terms not included in the reduced model $x_{g+1}, x_{g+2}, \ldots, x_k$ contribute substantial quality of information to predict Y not contained within the terms of the reduced model, then the model C should predict with a smaller error of prediction than model R. In other words, if at least one $\beta_i \neq 0$ for $i = g+1, g+2, \ldots, k$, then $\text{SSE}_C < \text{SSE}_R$; the greater the difference, the stronger the evidence to support the alternative hypthosis that at least one $\beta_i \neq 0$ for $i = g+1, g+2, \ldots, k$ and to reject the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

Practice

1. Consider the general linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$, where $\mathbb{E}(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$. Notice that $\hat{\beta}_i = \mathbf{a}'\hat{\beta}$, where the vector $\mathbf{a}$ is defined by $a_j = 1$ if $j = 1$ and $a_j = 0$ if $j \neq i$. Use this to verify that $\mathbb{E}[\hat{\beta}_i] = \beta_i$ and $V(\hat{\beta}_i) = c_{ii}\sigma^2$, where c_{ii} is the element in row i and column i of $(\mathbf{X}'\mathbf{X})^{-1}$.

Solution.

We have the equation $\hat{\beta} = \mathbf{(X'X)^{-1}X'Y}$ and $\mathbf{Y} = \mathbf{X}\beta$ and the property that $\mathbb{E}[\epsilon] = 0$.

$$

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
\mathbb{E}\left[\hat{\beta}\right] &= \mathbb{E}\left[(X'X)^{-1}X'Y\right] \\
&= \mathbb{E}\left[(X'X)^{-1}X'X\beta\right] \\
&= \mathbb{E}\left[I\beta\right] \\
&= \beta \\
\implies \mathbb{E}\left[\beta_i\right] &= \mathbb{E}\left[\mathbf{a}'\hat{\beta}\right] \\
&= \mathbf{a}'\,\mathbb{E}\left[\hat{\beta}\right] \\
&= \mathbf{a}'\beta \\
&= \beta_i.
\end{aligned}
$$

$$

On the other hand, the variance $V(\hat{\beta}_i)$ is given by

$$
\begin{aligned}
V(\hat{\beta}) &= V\left[(X'X)^{-1}X'Y\right] \\
&= (X'X)^{-1}X'V[Y]\left((X'X)^{-1}X'\right)' \\
&= (X'X)^{-1}X'\sigma^2\left((X'X)^{-1}X'\right)' \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2 I(X'X)^{-1} \\
\implies V(\hat{\beta}_i) &= V(\mathbf{a}'\hat{\beta}) \\
&= \mathbf{a}'\mathbf{a}\sigma^2(X'X)^{-1} \\
&= c_{ii}\sigma^2.
\end{aligned}
$$

2. A real estate agent's computer data listed the selling price Y (in thousands of dollars), the living area x_1 lin hundreds of square feet), the number of floors x_2 , number of bedrooms x_3 , and number of bathrooms x_4 for newly listed condominiums. The multiple regression model $\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ was fit to the data obtained by randomly selecting 40 condos currently on the market.

a. If $R^2 = 0.942$, is there significant evidence to conclude that at least one of the independent variables contributes significant information for the prediction of selling price?

We have the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ and the alternative hypothesis $H_\alpha : \beta_i \neq 0$ for some $i = 1, 2, 3, 4$.

We test this using the formula $F = \frac{n-(k+1)}{k} \frac{R^2}{1-R^2}$, and find that $F = \frac{40-(4+1)}{4} \frac{0.942}{1-0.942} = 142.1$ The test statistic is an F distribution with degrees of freedom $v_1 = k = 4$ and $v_2 = n - (k+1) = 35$. The following function produces the p-value for this statistic:

```
pf(142.1, df1 = 4, df2 = 35, lower.tail = FALSE)
```

```
## [1] 4.011082e-21
```

As 4.01e^{-21} is much less than 0.05, then we have sufficient evidence to conclude that at least one of the independent variables contributes significant information for the prediction of selling price.

b. If $S_{yy} = 16382.2$, what is SSE?

$$
\begin{aligned}
R^2 &= \frac{S_{yy} - SSE}{S_{yy}} \\
0.942 &= \frac{16382.2 - SSE}{16382.2} \\
15432.03 &= 16382.2 - SSE \\
SSE &= 16382.2 - 15432.03 \\
SSE &= 950.17
\end{aligned}
$$

c. The realtor theorizes that square footage, x_1 , is the most important predictor variable, and that the other variables can be left out without losing much prediction information. A simple linear regression of selling price vs. square footage was fit using the same 40 condos, and its SSE was 1553. Can the other independent variables, x_2, x_3, and x_4 be dropped from the model without losing predictive information? Test at the $\alpha = 0.05$ significance level.

We test the null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. The SSE of the complete model is $SSE_C = 950.17$ and the SSE of the reduced model is $SSE_R = 1553$. Our test statistic is given by:

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} \\ = \frac{(1553 - 950.17)/(4-1)}{950.17/(40-4-1)} \\ = 7.40.$$

The test statistic follows an F distribution with $v_1 = 3$ and $v_2 = 35$ degrees of freedom. The p-value is given by:

```
pf(7.4, df1 = 3, df2 = 35, lower.tail = F)
```

```
## [1] 0.0005778566
```

The p-value is 0.000578 which is less than 0.05, so there is sufficient evidence to reject the null hypothesis that the other predictors x_2, x_3, x_4 can be dropped from the model without losing predictive information.

3. A response Y is a function of three independent variables x_1, x_2, x_3 that are related as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

a. Fit the model to the $n = 7$ data points given in the code below.

```
data <- data.frame(y =  c(1, 0, 0, 1, 2, 3, 3),
                   x1 = c(-3, -2, -1, 0, 1, 2, 3),
                   x2 = c(5, 0, -3, -4, -3, 0, 5),
                   x3 = c(-1, 1, 1, 0, -1, -1, 1))
data
```

```
##   y x1 x2 x3
## 1 1 -3  5 -1
## 2 0 -2  0  1
## 3 0 -1 -3  1
## 4 1  0 -4  0
## 5 2  1 -3 -1
## 6 3  2  0 -1
## 7 3  3  5  1
```

```
fit <- lm(y ~ ., data = data); fit
```

```
##
## Call:
## lm(formula = y ~ ., data = data)
##
## Coefficients:
## (Intercept)           x1           x2           x3
##       1.429        0.500        0.119       -0.500
```

b. Predict Y when $x_1 = 1, x_2 = -3, x_3 = -1$. Compare the result with the observed data in row 5 of the table. Why are these values not equal?

```
y = 1.429 + .5*1 + .119*-3 + -1*-0.5; y
```

```
## [1] 2.072
```

The prediction $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = 2.072$ is the expected value $\mathbb{E}(Y)$. The value $y = 2$ is taking into account the error term ϵ_5 . That is, $2 = \mathbb{E}(Y) + \epsilon_5 = 2.072 + \epsilon_5$, where $\epsilon_5 = -0.072$.

c. Do the data present sufficient evidence to indicate that x_3 contributes information for the prediction of Y? Test the hypothesis $H_0 : \beta_3 = 0$ using $\alpha = 0.05$.

```
fit %>% summary()
```

```
##
## Call:
## lm(formula = y ~ ., data = data)
##
## Residuals:
##        1        2        3        4        5        6        7
## -0.02381  0.07143 -0.07143  0.04762 -0.07143  0.07143 -0.02381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.42857    0.03367   42.43 2.88e-05 ***
## x1           0.50000    0.01684   29.70 8.38e-05 ***
## x2           0.11905    0.00972   12.25 0.001172 **
## x3          -0.50000    0.03637  -13.75 0.000833 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08909 on 3 degrees of freedom
## Multiple R-squared:  0.9975, Adjusted R-squared:  0.9951
## F-statistic:   407 on 3 and 3 DF,  p-value: 0.0002058
```

The p-value for the test statistic is 0.000833, which is less than 0.05, which means that we reject the null hypothesis that x_3 does not contribute information for the prediction of y.

```r
knitr::opts_chunk$set(echo = TRUE)

library(dplyr)
pf(142.1, df1 = 4, df2 = 35, lower.tail = FALSE)
pf(7.4, df1 = 3, df2 = 35, lower.tail = F)
data <- data.frame(y =  c(1, 0, 0, 1, 2, 3, 3),
                   x1 = c(-3, -2, -1, 0, 1, 2, 3),
                   x2 = c(5, 0, -3, -4, -3, 0, 5),
                   x3 = c(-1, 1, 1, 0, -1, -1, 1))
data
fit <- lm(y ~ ., data = data); fit
y = 1.429 + .5*1 + .119*-3 + -1*-0.5; y
fit %>% summary()
```