Homework 2

PSTAT 120C

Summer 2022 Session B

Reading

The purpose of this portion of the assignment is to guide your reading and help you generate concise reading notes that list the key concepts – generally, terminology, definitions, and theorems. For the submission, treat each bullet point as an exercise and submit your 'answers' as you would a problem set.

- Write the general equation for a **multiple** linear regression model.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{1}$$

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

$$\left(\boldsymbol{X'X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{X'Y} \tag{2}$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'Y} \tag{3}$$

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

The test statistic is:

$$T = \frac{\boldsymbol{a'}\hat{\boldsymbol{\beta}} - \left(\boldsymbol{a'}\boldsymbol{\beta}\right)_{\mathbf{0}}}{S\sqrt{\boldsymbol{a'}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{a}}} \tag{4}$$

And the confidence interval:

$$\boldsymbol{a'}\hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2}} S\sqrt{\boldsymbol{a'}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{a}} \tag{5}$$

- Describe the general process of testing the hypothesis that $\beta_1 = \beta_2 = ... = \beta_k = 0$.
  1. Define the null and alternative hypotheses;
  2. Calculate the value of the test statistic;
  3. Find the critical t-value;
  4. Determine whether the test statistic falls in the rejection region or not;
  5. Calculate the p-value of the test statistic;
  6. Decide whether to reject the null hypothesis or fail to reject, and interpret your decision in the context of the problem.

Practice

The purpose of this portion of the assignment is to help you practice applying concepts in the reading, and in some cases, establish results that will be used later on. Remember that you will be graded on problem attempts, not solutions; do your best and ask questions if you get stuck.

1. Consider the general linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$, where $E[\epsilon] = 0$ and $V(\epsilon) = \sigma^2$. Notice that $\hat{\beta}_1 = \mathbf{a}'\hat{\beta}$, where the vector $\mathbf{a}$ is defined by $a_j = 1$ if $j = i$ and $a_j = 0$ if $j \neq i$.

   Use this to verify that $E[\hat{\beta}_i] = \beta_i$ and $V(\hat{\beta}_i) = c_{ii}\sigma^2$, where c_{ii} is the element in row i and column i of $(\mathbf{X}'\mathbf{X})^{-1}$.

   Here, $\boldsymbol{a}$ is a vector of k zeroes and a single one. Therefore:

$$E\left[\hat{\beta}_i\right] = E\left[\boldsymbol{a'}\hat{\boldsymbol{\beta}}\right] \tag{6}$$
$$= \boldsymbol{a'}E\left[\hat{\boldsymbol{\beta}}\right] \tag{7}$$
$$= \boldsymbol{a'}\boldsymbol{\beta} \tag{8}$$
$$= \beta_i \tag{9}$$

$$V\left(\hat{\beta}_i\right) = V\left(\boldsymbol{a'}\hat{\boldsymbol{\beta}}\right) \tag{10}$$
$$= \boldsymbol{a'}E\left[\hat{\boldsymbol{\beta}}\right]\boldsymbol{a} \tag{11}$$
$$= \boldsymbol{a'}\sigma^2\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{a} \tag{12}$$
$$= \sigma^2\boldsymbol{a'}\left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{a} \tag{13}$$
$$= c_{ii}\sigma^2 \tag{14}$$

2. A real estate agent's computer data listed the selling price Y (in thousands of dollars), the living area x_1 (in 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 $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ was fit to the data obtained by randomly selecting 40 condos currently on the market.

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

The easiest way to answer this question is by recognizing that the F-distributed test statistic is equivalent to $F = \frac{n-(k+1)}{k}\left(\frac{R^2}{1-R^2}\right)$. For this problem, $n = 40$ and $k = 4$, so:

$$F = \frac{35}{4}\left(\frac{.942}{1-.942}\right) \tag{15}$$

$$F = 8.75\,(16.24) \tag{16}$$

$$F = 142.1 \tag{17}$$

This statistic is compared to an F-distribution with 4 numerator or 35 denominator degrees of freedom. It has $p < .0001$, as shown:

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

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

It is statistically significant, so we reject the null hypothesis and 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?

We know that:

$$R^2 = 1 - \frac{SSE}{S_{yy}} \tag{18}$$

$$.942 = 1 - \frac{SSE}{16382.2} \tag{19}$$

$$(.942 - 1) = -\frac{SSE}{16382.2} \tag{20}$$

$$SSE = -0.058\,(-16382.2) \tag{21}$$

$$SSE = 950.168 \tag{22}$$

   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 are testing the null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ versus H_a : at least one $\beta_i \neq 0$. Here, the reduced model is $y = \beta_0 + \beta_1 x_1$, and the complete model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$.

The F-statistic to compare a complete and reduced model is:

$$F = \frac{\frac{(SSE_R - SSE_C)}{(k-g)}}{\frac{SSE_C}{(n-[k+1])}}$$

We know that $SSE_R = 1553$ and $SSE_C = 950.168$. The number of independent variables in the complete model is $k = 4$, the number of independent variables in the reduced model is $g = 1$, and

$n = 40$, so:

$$F = \frac{\frac{1553-950.168}{3}}{\frac{950.168}{40-5}} \tag{23}$$

$$F = \frac{\frac{602.832}{3}}{27.1477} \tag{24}$$

$$F = \frac{200.94}{27.1477} \tag{25}$$

$$F = 7.40 \tag{26}$$

The critical value, F_α with 4 and 35 degrees of freedom is 2.64. Therefore, we reject the null hypothesis and conclude that these variables should be retained in the model.

3. A response Y is a function of three independent variables x_1, x_2, and 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 this model to the $n = 7$ data points shown in the accompanying table.

y	x_1	x_2	x_3
1	-3	5	-1
0	-2	0	1
0	-1	-3	1
1	0	-4	0
2	1	-3	-1
3	2	0	-1
3	3	5	1

```
data <- tibble(y = c(1, 0, 0, 1, 2, 3, 3),
               x_1 = c(-3, -2, -1, 0, 1, 2, 3),
               x_2 = c(5, 0, -3, -4, -3, 0, 5),
               x_3 = c(-1, 1, 1, 0, -1, -1, 1))
```

The easiest way to fit this model is with `lm()`:

```
num_3_mod <- lm(y ~ x_1 + x_2 + x_3, data)
num_3_mod %>%
  summary()
```

```
##
## Call:
## lm(formula = y ~ x_1 + x_2 + x_3, 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 ***
## x_1          0.50000    0.01684   29.70 8.38e-05 ***
## x_2          0.11905    0.00972   12.25 0.001172 **
## x_3         -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
```

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?

One way – again, the easiest way to do this:

```
new_data <- tibble(x_1 = 1, x_2 = -3, x_3 = -1)

predict(num_3_mod, newdata = new_data)
```

```
##        1
```

```
## 2.071429
```

The expected value of Y when $x_1 = 1$, $x_2 = -3$, $x_3 = -1$ is 2.07. This is not equal to the observed value Y_5, which is 2. The reason for the difference is the component of random error, ϵ_5.

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 $\alpha = 0.05$.)

This hypothesis is tested by default with `lm()`; β_3 is significantly different from zero. It produced a test statistic of $t_{.05} = -13.75$ with $df = n - (k + 1) = 7 - 4 = 3$ and $p < .001$.