HW1

Robert Chandler

2024-06-16

Setup

Begin by reading the dataset for this set of problems:

```
le <- read.csv("./life_expectancy.csv")
```

Problem 1

Estimate the parameters (β_0, and β_1) for a linear regression to predict Y based on X. Complete the following with details.

First, set X, Y:

```
X <- le$Medical.doctors
Y <- le$X2015Life.expectancy
```

1.a

```
Xbar <- mean(X)
```

$\bar{X} = \frac{1}{n}\sum X_i = 19.4401124$

1.b

```
Ybar <- mean(Y)
```

$\bar{Y} = \frac{1}{n}\sum Y_i = 71.641573$

1.c

```
ss_xy = sum((X - Xbar) * (Y - Ybar))
```

$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 1.8444757 \times 10^4$

1.d

```
ss_x = sum((X - Xbar)^2)
```

$\sum(X_i - \bar{X})^2 = 5.2922519 \times 10^4$

1.e

Start with b_1 calculation since b_0 can be easily calculated from there:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

```
b1 = ss_xy / ss_x
```

Now calculate b_0 using b_1:

$$b_0 = \bar{Y} - b_1\bar{X}$$

```
b0 = Ybar - b1 * Xbar
```

$b_0 = 64.8662313$

1.f

Using previously calculated value for b_1:

$b_1 = 0.3485238$

1.g

First, calculate $\hat{Y}$:

$$\hat{\mathbf{Y}} = b_0 + b_1\mathbf{X}$$

```
Yhat <- b0 + b1 * X
```

then proceed with SSE:

```
sse = sum((Y - Yhat)^2)
```

$SSE = \sum(Y_i - \hat{Y})^2 = 5219.1755668$

1.h

```
n = length(X)
dfe = n - 2
```

We have 176 degrees of freedom of error, and we already know SSE, so

```
mse = sse / dfe
```

$MSE = \frac{SSE}{DFE} = 29.6544066$

1.i

```
sst = sum((Y - Ybar)^2)
```

$SST = \sum(Y_i - \bar{Y})^2 = 1.1647612 \times 10^4$

1.j

```
ssr = sum((Ybar - Yhat)^2)
```

$SSR = \sum(\bar{Y} - \hat{Y}_i)^2 = 6428.4367928$

To determine whether the two sides of the equation $SST = SSR + SSE$ are equal, we will compare their absolute difference against a very small threshold:

```
thresh = 1e-10

abs(sst - (ssr + sse)) < thresh
```

```
## [1] TRUE
```

Since the condition above is true, the values are close enough for us to consider them equal.

Problem 2

Calculations

Obtain t-values at $\alpha = 0.1, 0.05$ using the $n - 2 = 176$ degrees of freedom for this problem:

```
t_0.1 = qt(1 - 0.1 / 2, df = dfe)
t_0.05 = qt(1 - 0.05 / 2, df = dfe)
```

Next, obtain the standard error of the estimation $s\{b_1\}$ and use it to obtain the margin of error for each t value:

```
stderr_b1 = sqrt(mse / ss_x)
me_0.1 = t_0.1 * stderr_b1
me_0.05 = t_0.05 * stderr_b1
```

Answers

In order to estimate the linear impact of X on Y, at a confidence of $100(1 - \alpha)\%$, you should use the critical value, or the t value denoted as $t(1 - \alpha/2, n - 2)$, which has a value of 1.6535574 at $\alpha = 0.1$, and 1.9735344 at $\alpha = 0.05$. The standard error of the estimation is:

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}} = 0.0236714$$

The margin error, or $t \cdot SE$, of the confidence interval is 0.0391421 at $\alpha = 0.1$, and 0.0467164 at $\alpha = 0.05$.

Problem 3

Perform a hypothesis test on the linear impact of X on Y, with a t test with a significant value of 0.1.

The "linear impact" of X on Y refers to the change in Y attributed to a change in X, which is just the slope of the regression line, β_1. We assume the use of a two-sided test against $\beta_1^* = 0$ since we lack any other information to form our hypotheses, so:

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

and we define our test statistic as:

$$t^* = \frac{(b_1 - 0)}{s\{b_1\}}$$

We will reject H_0 if the p value of the 2-sided test is less than $\alpha = 0.1$, which is equivalent to the condition $|t^*| > t(1 - \frac{\alpha}{2}, n - 2)$

```
t_star = b1 / stderr_b1
```

$t^* = 14.7233986$, so we need to compare this to the t value for $\alpha = 0.1$, which was calculated in the previous problem as $t(1 - \frac{0.1}{2}, 176) = 1.6535574$.

Since $|t^*| = 14.7233986 > 1.6535574 = t(1 - \frac{0.1}{2}, 176)$, this t value falls in the rejection region.

To double-check, we can calculate the p value for this:

```
p = 2 * (1 - pt(t_star, dfe))
```

$p \approx 0 \ll 0.1 = \alpha$, which confirms that we should reject H_0.

In conclusion, there is an extremely low chance of observing the data in this dataset under the assumption that there is no linear impact of X on Y, so we can say that a change in X value brings about a statistically significant impact on the linear change in Y.

Problem 4

Use R to obtain a summary of this SLR model. Highlight the following concepts on the output, the notation, the values, and finally an interpretation.

We can summarize the model using `lm()`:

```
model <- lm(Y ~ X)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1102  -3.5062   0.4287   4.0203  11.7057
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.86623    0.61511  105.45   <2e-16 ***
## X            0.34852    0.02367   14.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.446 on 176 degrees of freedom
## Multiple R-squared:  0.5519, Adjusted R-squared:  0.5494
## F-statistic: 216.8 on 1 and 176 DF,  p-value: < 2.2e-16
```

We will use these results to describe the following:

4.a

The standard error of the point estimate of the linear impact of X on Y:

The point estimate of the linear impact of X on Y is b_1. The standard error of b_1 estimates the standard deviation of the sampling distribution of b_1. This is the "spread" of the values we would get when repeatedly sampling b_1 while holding the level of X constant, and its value is shown under the `Coefficients` section of the summary above, at the intersection of the `Std. Error` column and the row labeled `X`. It has a value of 0.0236714, which agrees with the value calculated manually in Problem 2.

4.b

The residual standard error:

The residuals are the differences between the point predictions $\hat{Y}_i$ and the observed Y_i values. The residual standard error estimates the standard deviation of these residuals, or how much spread we expect to see on the residuals. It is notated as s and can be found by taking the square root of s^2, also known as the MSE. It can be found under the heading `Residual standard error` in the summary above and it has a value of 5.445586.

4.c

The degree of freedom of the residual:

The degrees of freedom of the residual, sometimes notated DFE (degrees of freedom of error), can be found in the summary above on the same line as the residual standard error, under the `Residual standard error` heading, where it notes on how many degrees of freedom that value is. In this case, it has a value of 176, which is equal to $n - 2$.

4.d

The mean square of the standard error:

The mean square error, notated MSE or s^2, is closely related to the residual standard error discussed above. It is another measure of the spread of the residual values, but it is equal to the residual standard error squared, and can be calculated by taking the sum of each residual value squared and dividing by the degrees of freedom. It is minimized by the least squares solution. It is not explicitly shown in the summary above, but can easily be calculated by squaring the residual standard error from 4.b. Alternatively, it can be found by calling the `anova()` function on the model returned by `lm()`:

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## X           1 6428.4  6428.4  216.78 < 2.2e-16 ***
## Residuals 176 5219.2    29.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

where it is found at the intersection of the `Residuals` row and the `Mean Sq` column. It has a value of 29.6544066.

4.e

The standard deviation of the dependent variable Y, denoted by s_y, and briefly explain how it is related to the total sum of variance, $SST = \sum(Y_i - \bar{Y})^2$:

The (sample) standard deviation of Y is denoted s_y and it describes the variation in the observed Y_i values about their mean. It is related to SST by the following equation:

$$s_y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{SST}{n-1}}$$

so as SST increases/decreases, so does s_y. It is not readily available in the model summary, but it can be derived using the information in the `anova()` table above by summing the rows in the `Sum Sq` column, which yields SST, and then dividing by the sum of the rows in the `Df` column and taking the square root of the result. Thus, it has a value of $s_y = 8.1120725$.

Problem 5

5.a

The tendency, or the form by which of the response variable, Y, varies with X can be estimated with a linear function (**TRUE**). The linear function has a true form of $\beta_0 + \beta_1 X$ in the population domain.

5.b

At a general $X = X_h$ level, the predicted value is estimated by $\beta_0 + \beta_1 X_h$. Both β_0 and β_1 are variables and can be estimated by b_0 and b_1 on a sample (**FALSE**).

5.c

The deviation between the actual response variable Y and the predicted Y, or $\hat{Y}$ at a given $X = X_h$ level is called the random error and is denoted by ε, which can be estimated with a value denoted by e in a sample.

5.d

This random error is assumed to have a distribution of $N(0, \sigma^2)$ (**TRUE**), where the standard deviation, σ, can be estimated by the standard error term denoted by s computed from a sample.

5.e

The actual response variable, $Y = \beta_0 + \beta_1 X + \varepsilon$, represents the linear relationship between X and Y. The two "ingredients" in this relationship can be identified as $\beta_0 + \beta_1 X$ and ε.

Update X

For the remaining problems, we consider X to be the number of nurses and midwives rather than doctors.

```
X = le$Nurses
```

Problem 6

Use a significance level of 0.05, or confidence level of 0.95, and suppose the prediction is made at $X_h =$ the mean of X, or $\bar{X}$. Complete the confidence interval questions.

First, compute necessary calculations for the confidence interval:

```
alpha = 0.05
Xbar = mean(X)
model <- lm(Y ~ X)
b0 = coef(model)[[1]]
b1 = coef(model)[[2]]
```

```
Yhat_h = b0 + b1 * Xbar
mse = anova(model)["Residuals", "Mean Sq"]
std_err_mean = sqrt(mse * (1/n + 0))
t_star = qt(1 - alpha / 2, n - 2)
std_err_pred = sqrt(mse + std_err_mean^2)
m = 3
std_err_predmean = sqrt(mse / m + std_err_mean^2)
```

6.a

To estimate the mean response value of Y, the point estimate can be estimated as

$$\hat{Y}_h = b_0 + b_1 X_h = 66.0672276 + 0.1183351 \cdot 47.1064607 = 71.641573$$

The standard error of this estimation is denoted as the formula

$$s\{\hat{Y}_h\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]} = \sqrt{37.2419684\left[\frac{1}{178} + \frac{(47.1064607 - 47.1064607)^2}{\sum(X_i - 47.1064607)^2}\right]}$$

and computed as 0.4574107.

The t-value is denoted by

$$t(1 - \alpha/2, n - 2) = t(1 - 0.05/2, 178 - 2)$$

and computed as 1.9735344.

Therefore, the confidence interval for this mean response is

$$\begin{aligned}\hat{Y}_h \pm t(1 - \alpha/2, n - 2)s\{\hat{Y}_h\} &= 71.641573 \pm 1.9735344 \cdot 0.4574107 \\ &= 71.641573 \pm 0.9027157 \\ &= (70.7388573, 72.5442888)\end{aligned}$$

6.b

To predict the single response (the next observation value), the point estimate can be estimated as

$$\hat{Y}_h = b_0 + b_1 X_h = 66.0672276 + 0.1183351 \cdot 47.1064607 = 71.641573$$

The standard error of this estimation is denoted

$$s\{\text{pred}\} = \sqrt{MSE + s^2\{\hat{Y}_h\}} = \sqrt{37.2419684 + 0.4574107^2}$$

and computed as 6.119738.

The t-value is denoted by

$$t(1 - \alpha/2, n - 2) = t(1 - 0.05/2, 178 - 2)$$

and computed as 1.9735344.

Therefore, the confidence interval for this single new prediction is

$$\begin{aligned}\hat{Y}_h \pm t(1 - \alpha/2, n - 2)s\{\text{pred}\} &= 71.641573 \pm 1.9735344 \cdot 6.119738 \\ &= 71.641573 \pm 12.0775133 \\ &= (59.5640597, 83.7190864)\end{aligned}$$

6.c

To predict the mean of 3 responses (the average of the next m observation values, where X_h is the same for all m observations, the point estimate can be estimated as

$$\hat{Y}_h = b_0 + b_1 X_h = 66.0672276 + 0.1183351 \cdot 47.1064607 = 71.641573$$

The standard error of this estimation is denoted

$$s\{\text{predmean}\} = \sqrt{\frac{MSE}{m} + s^2\{\hat{Y}_h\}} = \sqrt{12.4139895 + 0.4574107^2}$$

and computed as 3.5529163.

The t-value is denoted by

$$t(1 - \alpha/2, n - 2) = t(1 - 0.05/2, 178 - 2)$$

and computed as 1.9735344.

Therefore, the confidence interval for this single new prediction is

$$\begin{aligned} \hat{Y}_h \pm t(1 - \alpha/2, n - 2)s\{\text{predmean}\} &= 71.641573 \pm 1.9735344 \cdot 3.5529163 \\ &= 71.641573 \pm 7.0118024 \\ &= (64.6297706, 78.6533755) \end{aligned}$$

6.d

When estimating the mean response value given the X_h = median of X, the corresponding standard error is **bigger than** when X_h = mean of X, because **the variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$ (due to the $(X_h - \bar{X})^2$ term in the numerator), so when $X_h = \bar{X}$, which is itself the mean, it must be the case that the error is smaller than when X_h is the median since the variability must be smaller**.

6.e

When estimating the mean of 10 responses, the corresponding standard error is **smaller** than when estimating the mean of 3 responses at the same X level, because **this is essentially an extension of the central limit theorem, and the more responses we add to our calculation, the tighter our distribution will get and the closer we will be to approaching the true mean response. Mathematically, as m increases, the denominator of the variance/error term increases, driving error down**.

Problem 7

We know that both ANOVA F test and T-test can be used to address the significance of the linear impact of X on Y, $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$. We have completed the T-test in the previous questions, now complete an ANOVA F-test.

Start by calculating the various sums:

```
sst = sum((Y - Ybar)^2)
Yhat = b0 + b1 * X
sse = sum((Y - Yhat)^2)
ssr = sum((Yhat - Ybar)^2)
```

No new α was given, so we assume it is the same $\alpha = 0.05$ as in Problem 6.

7.a

SST can be computed with the formula $SST = \sum(Y_i - \bar{Y})^2$ and has a value of 1.1647612×10^4, the degree of freedom is computed with the formula: $n - 1$ and has a value of 177.

SSE can be computed with the formula $SSE = \sum(Y_i - \hat{Y}_i)^2$ and has a value of 6554.5864322, the degree of freedom is computed with the formula: $n - 2$ and has a value of 176.

SSR can be computed with the formula $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ and has a value of 5093.0259274, the degree of freedom is computed with the formula: 1 and has a value of 1.

7.b

Compute the test statistic for the F test, F^*. The formula is

$$F^* = \frac{MSR}{MSE} = \frac{SSR}{SSE/(n-2)} = \frac{5093.0259274}{6554.5864322/(178-2)}$$

```
f_star = ssr / (sse / (n - 2))
f_crit = qf(1 - alpha, 1, n - 2)
```

and has a value of 136.7550146.

If we consider an identical set of input data being used, the F^* obtained in the one-tailed F test is related to the t^* obtained from the two-tailed t test in Problem 3 by the following relation: $F^* = (t^*)^2$. However, we changed our X input between the two problems, so the actual values between the two problems are not related in this way.

7.c

The critical value of the F-test can be denoted by $F(1-\alpha, 1, n-2)$ and has a value of $F(1-0.05, 1, 178-2) = F(0.95, 1, 176) = 3.894838$.

7.d

Compute the p-value for the F-test with the formula

$$p = P\{F(1, n-2) > F^*\} = P\{F(1, 176) > 136.7550146\}$$

```
p_f_star = 1 - pf(136.75, 1, n - 2)
```

and a value of 0.

Had these two tests been using the same X and α values, they would theoretically have the same p-value, and although they appear the same since they are both so close to zero here, they are **not** actually the same value theoretically.

Problem 8

8.a

General linear test (GLT) constructs and then compares two models that establish under H_0 and H_a. Specifically, the full model is established under H_a, and the reduced model is established under H_0.

8.b

The total error in the full model, SSE(Full) or SSE(F) has a value of $SSE(F) = SSE = 6554.5864322$ and a degree of freedom of $n - 2 = 176$. The total error in the reduced model, SSE(Reduced) or SSE(R) has a value of $SSE(R) = SST = 1.1647612 \times 10^4$ and a degree of freedom of $n - 1 = 177$.

8.c

Discuss the connection between the Global F test and the GLT F test in the following perspectives:

8.c.i

The null and alternative hypothesis:

The null and alternative hypotheses can be defined identically for the Global F test and the GLT F test:

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

In both cases, this choice of hypotheses allows us to draw conclusions about whether or not the additional parameter β_1 helps to reduce the variation in Y_i about the fitted regression function.

8.c.ii

The test statistic:

When in simple linear regression, the test statistic of the GLT is identical to the one from the ANOVA/global F test, since it is defined by a difference in the SSE of the reduced and full models which ultimately simplifies to MSR/MSE, which is the same ratio that the statistic in the global F test simplifies to under SLR.

8.c.iii

Situations when the two methods are equivalent, and situations when only GLT test is appropriate:

While the two methods are equivalent under simple linear regression, GLT can be extended to highly complex tests of linear models with multiple parameters as in the case of Multiple Linear Regression, while the Global F test is not appropriate in these cases.