Homework 2

Greg DePaul

2023-03-28

Problem 6 - Crime Rate and Education Simple Linear Regression Model

We are given the following summary statistics:

```
n <- 84
sum_x <- 6602
sum_y <- 597341
sum_x_squared <- 522098
sum_y_squared <- 4796548849
sum_xy <- 46400230
mean_x <- sum_x / n
mean_y <- sum_y / n
```

Perform analysis under the simple linear regression model.

(a) Based on the scatter plot, comment on the relationship between percentage of high school graduates and crime rate. It appears that there is a negative correlation between percentage of graduates and crime rate. But because of the spread of the data, we expect this correlation to be closer to zero than -1.

(b) Calculate the least-squares estimators: $\hat{\beta}_0, \hat{\beta}_1$. Write down the fitted regression line. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$. From the previous homework, we can use the formulas:

$$\hat{\beta}_1 = \frac{\overline{XY} - \frac{1}{n}\sum_{i=1}^{n} X_i Y_i}{\overline{X}^2 - \frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

```
beta_1 <- (mean_x * mean_y - sum_xy / n) / (mean_x^2 - sum_x_squared / n)
beta_0 <- mean_y - beta_1 * mean_x
```

We can get their values:

```
print(beta_0)
```

```
## [1] 20517.6
```

```
print(beta_1)
```

```
## [1] -170.5752
```

So the fitted regression line is:

$$\hat{Y} = 20517.6 - 170.5752X$$

(c) Calculate error sum of squares (SSE) and mean squared error (MSE). What is the degrees of freedom of SSE? To calulcate SSE, we expand it until we reach our summary statistics:

$$\begin{aligned}
SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \\
&= \sum_{i=1}^{n}(Y_i^2 - 2Y_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) + (\hat{\beta}_0 + \hat{\beta}_1 X_i)^2) \\
&= \sum_{i=1}^{n}(Y_i^2 - 2\hat{\beta}_0 Y_i - 2\hat{\beta}_1 X_i Y_i + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 X_i + \hat{\beta}_1^2 X_i^2) \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - 2\hat{\beta}_0\left(\sum_{i=1}^{n} Y_i\right) - 2\hat{\beta}_1\left(\sum_{i=1}^{n} X_i Y_i\right) + \left(\sum_{i=1}^{n}\hat{\beta}_0^2\right) + 2\hat{\beta}_0\hat{\beta}_1\left(\sum_{i=1}^{n} X_i\right) + \hat{\beta}_1^2\left(\sum_{i=1}^{n} X_i^2\right) \\
&= \left(\sum_{i=1}^{n} Y_i^2\right) - 2\hat{\beta}_0\left(\sum_{i=1}^{n} Y_i\right) - 2\hat{\beta}_1\left(\sum_{i=1}^{n} X_i Y_i\right) + n\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1\left(\sum_{i=1}^{n} X_i\right) + \hat{\beta}_1^2\left(\sum_{i=1}^{n} X_i^2\right)
\end{aligned}$$

```
SSE <- sum_y_squared - 2 * beta_0 * sum_y - 2 * beta_1 * sum_xy + n*beta_0^2 + 2 * beta_0 * beta_1 * su
MSE <- SSE / (n - 2)
```

```
print(SSE)
```

```
## [1] 455273165
```

```
print(MSE)
```

```
## [1] 5552112
```

We recall that

$$\text{degrees of freedom}(SSE) = n - 2 = 82$$

(d) Calculate the standard errors for the LS estimators $\hat{\beta}_0, \hat{\beta}_1$, respectively. Observe, we can rewrite the standard error of our estimators to be:

$$SE\{\hat{\beta}_0\} = \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)} = \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}\right)}$$

```
s_beta_0 <- sqrt(MSE * (1 / n + mean_x^2 / (sum_x_squared - n * mean_x^2)))
print(s_beta_0)
```

```
## [1] 3277.643
```

$$SE\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = \sqrt{\frac{MSE}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}$$

```
s_beta_1 <- sqrt(MSE / (sum_x_squared - n * mean_x^2))
print(s_beta_1)
```

```
## [1] 41.57433
```

(e) Test whether or not there is a linear association between crime rate and percentage of high school graduates at significance level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$
- Test Statistic: $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}}$
- Null distribution of T^* is $t_{n-2} = t_{82}$

We use the two-sided t-test. The rule becomes is we reject H_0 if the following comparison is true:

$$|T^*| > b_{critical}$$

where $b_{critical}$ is defined as the value such that

$$P(|t_{82}| > b_{critical}) < \alpha = 0.01$$

```
crit_val  <- qt(1 - 0.01 / 2, df = 82)
T_star <- beta_1 / s_beta_1
print(abs(T_star) > crit_val)
```

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

From this, we reject the null hypothesis to conclude that there is a significant association between crime rate and percentage of high school graduates.

(f) What is an unbiased estimator for β_0? Construct a 99% confidence interval for β_0. Interpret your confidence interval. On the previous homework, we proved that $\hat{\beta}_0$ is an unbiased estimator for β_0. Further, we know that for a given $1 - \alpha$ confidence interval for β_0, we can estimate it using

$$CI_{99\%}(\beta_0) = \hat{\beta}_0 \pm t(1 - \frac{\alpha}{2}, n - 2)SE(\hat{\beta}_0) = \hat{\beta}_0 \pm t_{82}(0.995)SE(\hat{\beta}_0)$$

```
crit_val <- qt(1 - 0.01 / 2, df = 82)
left_val <- beta_0 - s_beta_0*crit_val
print(left_val)
```

```
## [1] 11874.05
```

```
right_val <- beta_0 + s_beta_0*crit_val
print(right_val)
```

```
## [1] 29161.15
```

So we get the condidence interval to be:

$$CI_{99\%}(\beta_0) = (11874.05, 29161.15)$$

```
X_h <- 85
Y_h <- beta_0 + beta_1*X_h
crit_val <- qt(1 - 0.05 / 2, df = 926)
s_Y_h <- sqrt(MSE * (1/n + (X_h - mean_x)^2 / (sum_x_squared - n * mean_x^2)))
left_val <- Y_h - crit_val*s_Y_h
print(left_val)
```

(g) Construct a 95% confidence interval for the mean crime rate for counties with percentage of high school graduates being 85. Interpret your confidence interval.

```
## [1] 5292.313
```

```
right_val <- Y_h + crit_val*s_Y_h
print(right_val)
```

```
## [1] 6745.105
```

So we get the condidence interval to be:

$$CI_{95\%}(Y_{85}) = (5292.313, 6745.105)$$

```
X_h <- 85
Y_h <- beta_0 + beta_1*X_h
crit_val <- qt(1 - 0.05 / 2, df = 926)
s_Y_h <- sqrt(MSE * (1 + 1/n + (X_h - mean_x)^2 / (sum_x_squared - n * mean_x^2)))
left_val <- Y_h - crit_val*s_Y_h
print(left_val)
```

(h) County A has a high-school graduates percentage being 85. What is the predicted crime rate of county A? Construct a 95% prediction interval for the crime rate. Compare this interval with the one from part (g), what do you observe?

```
## [1] 1337.713
```

```
right_val <- Y_h + crit_val*s_Y_h
print(right_val)
```

```
## [1] 10699.7
```

So we get the condidence interval to be:

$$CI_{95\%}(Y_{85(new)}) = (1337.713, 10699.7)$$

It makes sense that the prediction interval of a new outcome is wider than the confidence interval of the mean response in part (g).

(i) Would additional assumption be needed in order to conduct parts (e)-(h)? If so, please state what it is. We assume the normal error model in order to generate the exact tests for parts (e) through (h).

Problem 7 - Crime Rate and Education ANOVA

(a) Calculate sum of squares: SSTO, SSE and SSR. What are their respective degrees of freedom? We already calculated SSE:

```
print(SSE)
```

```
## [1] 455273165
```

To calculate SSR:

$$SSR = \hat{\beta}_1^2 \left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right)$$

```
SSR <- beta_1^2 * (sum_x_squared - n * mean_x^2)
print(SSR)
```

```
## [1] 93462942
```

Lastly, we calculate SSTO by

$$SSTO = SSE + SSR$$

```
SSTO <- SSE + SSR
print(SSTO)
```

```
## [1] 548736108
```

(b) Calculate the mean squares. We already calculated MSE:

```
print(MSE)
```

```
## [1] 5552112
```

And the regression mean square is simply:

$$MSR = SSR$$

```
MSR <- SSR
print(MSR)
```

```
## [1] 93462942
```

(c) Summarize results from parts (a) and (b) into an ANOVA table.

	SS	df	MS
Regression	SSR = 93462942	1	MSR = 93462942
Error	SSE = 455273165	82	MSE = 5552112
Total	SSTO = 548736108	83	

(d) Assume Normal error model. Use the F test to test whether or not there is a linear association between crime rate and percentage of high school graduates at significance level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. Under the normal error model, we know $SSE \sim \sigma^2 \chi^2_{82}$. Therefore,

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$
- Test Statistic: $F^* = \frac{MSR}{MSE}$
- Null distribution of F^* is $F_{1,82}$

We use the one-side F-test. The rule becomes is we reject H_0 if the following comparison is true:

$$F^* > F(1-\alpha, 1, 82) = F(0.99, 1, 82)$$

```
F_star <- MSR / MSE
crit_val <- qf(0.99, 1, 82, lower.tail = TRUE)
print(F_star > crit_val)
```

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

From this, we reject the null hypothesis to conclude that there is a significant association between crime rate and percentage of high school graduates.

(e) Compare your calculation from part (d) with those from part (e) of Problem 6. What do you observe? In both, we see signficant evidence to reject the idea that crime rate is independent of graduation rat.

```
R_squared <- SSR / SSTO
print(R_squared)
```

(f) Calculate the coefficient of determination R.

```
## [1] 0.170324
```

(g) Calculate the correlation coefficient r between crime rate and percentage of high school graduates. Compare r^2 with R^2. What do you observe?

$$corr(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{n\left(\sum_{i=1}^{n} X_i Y_i\right) - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{n\left(\sum_{i=1}^{n} X_i^2\right) - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} Y_i^2\right) - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

```
corr <- (n * sum_xy - sum_x * sum_y) / (sqrt(n*sum_x_squared - sum_x^2) * sqrt(n*sum_y_squared - sum_y^
print(corr)
```

```
## [1] -0.4127033
```

```
print(corr^2)
```

```
## [1] 0.170324
```

Notice, we see that

$$r^2 = (-0.4127033)^2 = 0.170324 = R^2$$