Stat 139: Homework 7 Solution

Incorporate the R output in the HW write-up. Choose the output wisely, no need to print-out everything. No more than **one page** of output should be displayed for each problem and the relevant parts should be **highlighted**.

- **A hard copy of your solutions is due at noon on Friday, in a drop-box outside the Science Center 300 suite**. Solutions should not be submitted by email unless arrangements are made with a TF prior to the deadline (or unless late, see below).
- Please print your code and attach to the end of your solution. Use comments to indicate the code associated with each question. Partial credit will be available for R code only.
- See syllabus for details on homework and collaboration policies: your lowest homework score will be dropped; acknowledge your collaborators; solutions submitted electronically within 24 hours after deadline will be graded with a penalty; solutions more than 24 hours late will receive no credit.

Related Reading: Ramsey and Schafer, Chapter 7. You may also find Lecture 15 useful for this homework (slides and video).

Supplementary Theory: Regression Analysis: Theory, Methods, and Applications, A. Sen and M. Srivastava.: Chapter 1: Introduction (you may skip Sec. 1.7 for now).

R code: Ramsey and Schafer, Chapter 7.

1. (35 points) In class we learned that the classical simple linear regression model assumes the following distribution of responses:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where ϵ_i are i.i.d and $\epsilon_i \sim N(0, \sigma^2)$. Estimators of regression parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are derived by minimizing the sum of squared residuals,

$$SSR(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \Big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\Big)^2 \tag{2}$$

(a) (5 points) In Lecture 15 we showed that

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{3}$$

by taking a partial derivative of 2 w.r.t. $\hat{\beta}_0$. Use a partial derivative w.r.t $\hat{\beta}_1$ and 3 to show that the least-squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{4}$$

Solution:

$$\frac{\partial}{\partial \hat{\beta}_1} SSR(\hat{\beta}_0, \hat{\beta}_1) = 0 = -2\sum_{i=1}^{n}(X_i(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)))$$

$$= -2\sum_{i=1}^{n}(X_i(Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i)))$$

Dropping the factor of 2 and splitting the sum:

$$= \sum_{i=1}^{n} X_i(Y_i - \bar{Y}) + \hat{\beta}_1 \sum_{i=1}^{n} X_i(\bar{X} - X_i)$$

Since $\sum(Y_i - \bar{Y})$ *and* $\sum(\bar{X} - X_i)$ *both equal zero, we can subtract constant factors of these:*

$$= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - X_i)$$

Because $(\bar{X} - X_i) = -(X_i - \bar{X})$*, we can write the new equation*

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

And dividing gives the desired result:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Next we will prove that sampling distributions of the estimators are the following:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right),$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right). \tag{5}$$

In class we showed that both estimators are unbiased, i.e., $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$.

(b) (5 points) Next, explain why 4 can be simplified to

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{6}$$

Solution*:*

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \frac{\sum_{i=1}^{n}(X_i - \bar{X})\bar{Y}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

because $\sum_{i=1}^{n}(X_i - \bar{X})\bar{Y} = \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

(c) (10 points)Using 6, show that

$$Var(\hat{\beta}_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Note that the normality of $\hat{\beta}_1$ comes from the fact that it is a linear combination of Y_i's, which are normally distributed.
***Solution**: Because Y_i's are independent,*

$$\begin{aligned} Var(\hat{\beta}_1) =& Var\left(\frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \sum_{i=1}^n Var\left(\frac{(X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\ =& \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)^2 Var(Y_i) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}\sigma^2 \\ =& \sigma^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}. \end{aligned}$$

Next, we will derive the sampling distribution of $\hat{\beta}_0$.

(d) (5 points) Recall the following property of the variance: for two (non-independent) random variables Z_1 and Z_2 and fixed numbers a and b,

$$Var(aZ_1 + bZ_2) = a^2 Var(Z_1) + b^2 Var(Z_2) + 2abCov(Z_1, Z_2).$$

It turns out that $Cov(\bar{Y}, \hat{\beta}_1) = 0$. In other words, the estimate of the slope of a regression line is not correlated with the average response. Intuitively, knowing the average response will not tell us anything about the slope of the regression line. Using this fact, as well as 3, verify that

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right).$$

Again, the normality of $\hat{\beta}_0$ comes from the normality of both components in 3.
***Solution**:*

$$\begin{aligned} Var(\hat{\beta}_0) =& Var(\bar{Y} - \hat{\beta}_1 \bar{X}) = Var(\bar{Y}) + Var(\hat{\beta}_1 \bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X}) \\ =& \frac{\sigma^2}{n} + \bar{X}^2 \sigma^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right). \end{aligned}$$

(e) (10 points) Based on the sampling distributions of the estimators of regression coefficients derived above, describe how the precision of the estimators can be increased.
***Solution**: Considering that*

$$Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right).$$

and

$$Var(\hat{\beta}_1) = \sigma^2 \frac{1}{(n-1)S_X^2},$$

the precision of the estimators can be increased by increasing the sample size, n, and increasing the variance of X (if it is possible to control the selection of X values). Both

precision will also be higher when the standard deviation about the regression line, σ, is smaller (e.g., if the measurement error of the response is small). However, this may be impossible to control. Finally, the slope will be estimated more precisely closer to $\bar{X}$. Note that this is related to the fact that the regression line provides the most precise estimation around the middle point of the data, $(\bar{X}, \bar{Y})$.

2. (15 points) Complete parts (a) through (d) in Exercise 24 (**Decline in Male Births**) in Chapter 7 (3rd ed., or Exercise 26 in the 2nd edition). The data are posted online.
***Solution**: (a) (3 pts) R-code and output:*

```
> RegModel1 <- lm(Denmark ~ Year, data = Problem3_data)
> summary(RegModel1)

Call:
lm(formula = Denmark ~ Year, data = Problem3_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.987e-01  4.080e-02  14.673   <2e-16 ***
Year        -4.289e-05  2.069e-05  -2.073   0.0442 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.001803 on 43 degrees of freedom

> RegModel2 <- lm(Netherlands ~ Year, data = Problem3_data)
> summary(RegModel2)

Call:
lm(formula = Netherlands ~ Year, data = Problem3_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.724e-01  2.792e-02   24.08  < 2e-16 ***
Year        -8.084e-05  1.416e-05   -5.71 9.64e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.001233 on 43 degrees of freedom

> RegModel3 <- lm(Canada ~ Year, data = Problem3_data)
> summary(RegModel3)

Call:
lm(formula = Canada ~ Year, data = Problem3_data)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.338e-01  5.480e-02  13.390 3.98e-11 ***
Year        -1.112e-04  2.768e-05  -4.017 0.000738 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.000768 on 19 degrees of freedom

> RegModel4 <- lm(USA ~ Year, data = Problem3_data)
> summary(RegModel4)

Call:
lm(formula = USA ~ Year, data = Problem3_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.201e-01  1.860e-02  33.340  < 2e-16 ***
Year        -5.429e-05  9.393e-06  -5.779 1.44e-05 ***

Residual standard error: 0.0002607 on 19 degrees of freedom
```

(b) (4 pts) T-statistics are -2.07 (Denmark), -5.72 (Netherlands), -4.02 (Canada), and -5.78 (USA). There is overwhelming evidence that the proportion of male births is declining (in each case we reject the hypothesis that $H_0 : \beta_1 = 0$).
(c) (4 pts) Because $\hat{\sigma}$ is the lowest for the US, the precision of the estimate of β_1 is the highest. Therefore, the denominator of the t-statistic for the US is small, which results in larger test statistic.
(d) (4 pts) Although the X's and the sample sizes are the same, it looks like the observations for the US are concentrated much closer around the regression line. In other words, as mentioned in (c), $\hat{\sigma}$ is lower for the US.