Homework 2 Solutions

Manu Navjeevan

August 24, 2021

Single Linear Regression Theory Review

1. Recall that we define our parameters of interest β_0 and β_1 as the parameters governing the "line of best fit" between Y and X:

$$\beta_0, \beta_1 = \arg\min_{b_0, b_1} \mathbb{E}[(Y - b_0 - b_1 X)^2]. \tag{1}$$

Once we define these parameters we define the regression error term $\epsilon = Y - \beta_0 - \beta_1 X$ which then generates the linear model

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

(a) Using the first order conditions for β_0 and β_1 (set the derivatives of the right hand side of (1) with respect to b_0 and b_1 equal to zero at) show why $\mathbb{E}[\epsilon] = \mathbb{E}[\epsilon X] = 0$.

Answer: Taking derivatives of (1) with respect to b_0 and b_1 gives:

$$\frac{\partial}{\partial b_0} = -2\mathbb{E}[(Y - b_0 - b_1 X)]$$
$$\frac{\partial}{\partial b_1} = -2\mathbb{E}[(Y - b_0 - b_1 X)X]$$

At the true optimizers β_0 and β_1, these derivatives are equal to zero, giving

$$-2\mathbb{E}[(Y - \beta_0 - \beta_1 X)] = 0$$
$$-2\mathbb{E}[(Y - \beta_0 - \beta_1 X)X] = 0$$

Dividing the above equations by -2 and recalling that $\epsilon = Y - \beta_0 - \beta_1 X$ gives the result.

(b) Using the definition of β_0 and β_1 as line of best fit parameters, give an intuitive explanation for why $\mathbb{E}[\epsilon] = 0$.

Answer: If $\mathbb{E}[\epsilon] \neq 0$, this means that our line is either consistently above or below our data, that is we are either consistently "overshooting" or "undershooting" Y. This is clearly not optimal and our line would be closer to Y on average if we adjusted β_0 down (decreased β_0) to fix an "overshoot" or adjusted β_0 up (increased β_0) to fix an "undershoot".

Hypothesis Testing and Confidence Intervals

In the following questions, whenver running a hypothesis test, please state the null and alternative hypotheses, show some work, and state the conclusion of the test.

1. In an estimated simple regression model based on $n = 64$, the estimated slope parameter, $\hat{\beta}_1$, is 0.310 and the standard error of $\hat{\beta}_1$ is 0.082.

(a) What is $\hat{\sigma}^2_{\beta_1}$? Recall σ_{β_1} is the terms such that, approximately for large n,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \sim N(0, \sigma_{\beta_1}).$$

Answer: Recall from lecture that $\text{se}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2_{\beta_1}/n}$. Using this we find that

$$\text{se}(\hat{\beta}_1) = \hat{\sigma}_{\beta_1}/\sqrt{n} \implies \hat{\sigma}^2_{\beta_1} = (\text{se}(\hat{\beta}_1))^2 \cdot n = 0.082^2 \cdot 64 \approx 0.430$$

(b) Test the hypothesis that the slope is zero against the alternative that it is not at the 1% level of significance ($\alpha = 0.01$).

Answer: We are testing, at level $\alpha = 0.01$,the null and alternate hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

We compute our test statistic

$$t^* = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{0.310}{0.082} = 3.78.$$

There are two ways to conduct this test. This first way would be to compute a p-value. Since this is a two sided test, we want to compute $\Pr(|Z| > |t^*|)$ where $Z \sim N(0,1)$. By symmetry of the normal distribution and this is equal to

$$2\Pr(Z > 3.78) = 2(1 - \text{pnorm}(3.78)) = 0.000156 < 0.01 = \alpha$$

Since the p-value is less than α, we reject this null hypothesis.

Another way of running this test would be to reject if $|t^*| > z_{1-\alpha/2} = z_{0.995}$. Using qnorm(0.995) we find $z_{0.995} = 2.57$. Since $|t^*| = 3.78 > 2.57 = z_{0.995}$ we reject this null hypothesis and conclude in favor of our alternative hypothesis that $\beta_1 \neq 0$.

(c) Test the hypothesis that the slope is negative against the alternative that it is positive at the 1% level of significance ($\alpha = 0.01$).

Answer: Formally the hypotheses that we are testing are given:

$$H_0 : \beta_1 \leq 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0.$$

We should already know that we will reject this null hypothesis given our answer in part (b) and since $\hat{\beta} \geq 0$. However, let's conduct this test formally in two ways. We can use the same value $t^* = 3.78$ from part (b). First, let us construct a p-value. Since this is a one sided test with a ">" sign in the alternate hypothesis, the p-value is given $\Pr(Z > t^*) = 1 - \text{pnorm}(3.78) \approx 0.00001$. This is clearly less than $\alpha = 0.01$ so we reject the null hypothesis.

Alternatively, we can compare t^* to $z_{1-\alpha}$ and reject if $t^* > z_{1-\alpha}$. Using qnorm(0.99) we find that $z_{1-\alpha} = 2.326$. Since $t^* = 3.78 > 2.326 = z_{0.99}$ we reject this null hypothesis and conclude in favor of the alternative hypothesis that $\beta_1 > 0$.

(d) Test the hypothesis that the slope is positive against the alternative that it is negative at the 5% level of significance. What is the p-value?

Answer: The hypotheses that we are testing are given

$$H_0 : \beta_1 \geq 0 \quad \text{vs.} \quad H_1 : \beta_1 < 0.$$

Again, we should already know that we will fail to reject this null hypotheses since $\hat{\beta}_1 \geq 0$. But, just to be sure, let's compute the p-value. Since this alternative hypothesis contains a "<" sign the p-value is computed $\Pr(Z < t^*) = \text{pnorm}(3.78) = 0.9999$. This p-value is much larger than $\alpha = 0.05$ so we fail to reject the null hypothesis that $\beta_1 \geq 0$.

(e) Generate a 99% confidence interval for β_1. How can we use this interval to run the hypothesis test in part (b)?

Answer: A 99% confidence interval for β_1 would be given

$$\hat{\beta}_1 \pm z_{0.995} \cdot \text{se}(\hat{\beta}_1) = 0.310 \pm 2.57 \cdot 0.082 = [0.09926, 0.52074].$$

Since zero is not contained in this interval, we know that we would reject the null hypothesis in part (b) in favor of the alternative that $\beta_1 \neq 0$.

2. Consider a simple regression of log-income (income is measured thousands of dollars), Y, against years of education, X. After collecting a sample of size $n = 50$ we estimate the following regression equation.

$$\widehat{Y} = \hat{\beta}_0 + 0.0180X.$$

(a) Using the following information to solve for $\hat{\beta}_0$ as well as the estimated variance $\widehat{\text{Var}}(\hat{\beta}_0)$, which is the square of the standard error.

- The standard error of $\hat{\beta}_0$ is 2.174
- The test statistic, t^*, associated with the hypothesis test for

$$H_0 : \beta_0 = 0 \quad \text{vs.} \quad H_1 : \beta_0 \neq 0,$$

is equal to 1.257.

Answer: The estimated variance is the square of the standard error so

$$\text{Var}(\hat{\beta}_0) = 2.174^2 = 4.72.$$

To find $\hat{\beta}_0$ we invert the test statistic

$$1.257 = t^* = \frac{\hat{\beta}_0 - 0}{2.174} \implies \hat{\beta}_0 = 1.257 \cdot 2.174 = 2.732.$$

(b) Use the following information to solve for the standard error $\hat{\beta}_1$ as well as the estimated variance $\widehat{\text{Var}}(\hat{\beta}_1)$, which is the square of the standard error.

- The test statistic, t^*, associated with the hypothesis test for

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0,$$

is equal to 5.754

Answer: Again, we invert the test statistic. However, in this case we know $\hat{\beta}_1 = 0.0180$.

$$5.754 = t^* = \frac{0.0180 - 0}{\text{se}(\hat{\beta}_1)} \implies \text{se}(\hat{\beta}_1) = \frac{0.0180}{5.754} = 0.003.$$

The Variance of $\hat{\beta}_1$ is the square of the standard error so that $\text{Var}(\hat{\beta}_1) = 0.000009$.

(c) Given that Y is a logged variable, $Y = \log(\text{income})$, how do we interpret $\hat{\beta}_1$?

Answer: Since a one unit increase in $\log(Y)$ corresponds to a $\approx 100\%$ increase in Y and a one unit increase in X is associated with an estimated 0.180 unit increase in $\log(Y)$ we can conclude that we estimate that a one unit increase in years of education is associated with a 1.8% increase in income.

(d) Suppose that we are interested in the average value of log-income for someone with 16 years of education. We want to use the model above to test the hypothesis that the average value of log-income for someone with 16 years of education is less than or equal to 1.85. That is we want to test

$$H_0 : \lambda = \beta_0 + 16\beta_1 \leq 1.85 \quad \text{vs.} \quad H_1 : \lambda = \beta_0 + 16\beta_1 > 1.85.$$

Use the fact that $\widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) = 2.84$ to test this hypothesis at level $\alpha = 0.1$.

Answer: We want to compute the standard error of the linear combination $\hat{\lambda} = \hat{\beta}_0 + 16\hat{\beta}_1$. To do so, we will use the formula $\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$.

$$\begin{aligned} \text{Var}(\hat{\lambda}) &= \text{Var}(\hat{\beta}_0) + 16^2 \text{Var}(\hat{\beta}_1) + 2 \cdot 16 \cdot \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= 4.72 + 256 \cdot 0.000009 + 32 \cdot 2.84 \\ &= 95.6023. \end{aligned}$$

This means that the standard error of $\hat{\lambda}$ is given $\text{se}(\hat{\lambda}) = \sqrt{\text{Var}(\hat{\lambda})} \approx 9.777$. The value of $\hat{\lambda}$ is given $\hat{\lambda} = 2.732 + 16 \cdot 0.0180 = 3.02$. Given this, we can compute our test statistic via the general formulation

$$t^* = \frac{\text{Estimated Value} - \text{Null Hypothesis}}{\text{Standard Error of Estimator}} = \frac{3.02 - 1.85}{9.777} \approx 0.11.$$

Since $0.11 < z_{0.9} = \text{qnorm}(0.9) = 1.281$ we fail to reject this null hypothesis and cannot conclude that the average value of log-income for a person with 16 years of education is greater than 1.85. We can also compute the p-value $\Pr(Z > t^*) = 1 - \text{pnorm}(0.11) = 0.456 > 0.1$ and reach the same conclusion.

(e) Use the above to generate a 90% confidence interval for λ.

Answer: Using $\hat{\lambda} = 3.05$, $\text{se}(\hat{\lambda}) = 9.777$ and $z_{0.95} = 1.65$ we construct a 90% confidence interval

$$\hat{\lambda} \pm z_{0.95} \cdot 9.77 = 3.05 \pm 1.65 \cdot 9.777 = [-13.0825, 19.18205].$$

We are 90% confident that the true value of λ lies in this interval.

3. (Challenge) Suppose we find that $\hat{\beta}_1 > 0$. If we reject the null hypothesis that $\beta_1 = 0$ in favor of an alternative hypothesis that $\beta_1 \neq 0$ at level α, up to what level can we be sure that would we reject the null hypothesis that $\beta_1 \leq 0$ against an alternative that $\beta_1 > 0$? (Please give some explanation here as well as your answer, which will be some multiple of α).

Answer: Recall that we reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the two sided alternative $H_1 : \beta_1 \neq 0$ at level α if $|t^*| > z_{1-\alpha/2}$ where

$$t^* = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)}.$$

Because $\hat{\beta}_1 > 0$ we know that $t^* = |t^*|$, that is we know that t^* is positive. So, we have that $t^* > z_{1-\alpha/2}$. Another way of putting this is that $\Pr(Z > t^*) < \alpha/2$.

We reject the null hypothesis $H_0 : \beta_1 \leq 0$ in favor of the one sided alternative $H_1 : \beta_1 > 0$ at level $\tilde{\alpha}$ if $t^* > z_{1-\tilde{\alpha}}$. Since we know that $t^* > z_{1-\alpha/2}$ we can take $\tilde{\alpha} = \alpha/2$ and still reject this null hypothesis. Alternatively, we know we reject this null hypothesis if the p-value, $p = \Pr(Z > t^*)$ is less than $\tilde{\alpha}$. Since we know that $\Pr(Z > t^*) < \alpha/2$ we can take $\tilde{\alpha} = \alpha/2$ and still reject the null against a one sided alternative.

R^2 and Goodness of Fit

1. Consider the following estimated regression equation.

$$\widehat{Y} = 6.83 + 0.869X.$$

Write the estimated regression equation that would result if

(a) All values of X were divided by 20 before estimation.

Answer: Recall from class that $\hat{\beta}_0$ does not change and $\hat{\beta}_1$ get's scaled by $\frac{1}{c}$ where $c = \frac{1}{20}$. If $\tilde{X} = \frac{1}{20}X$ the new estimated regression line would be

$$Y = 6.83 + \frac{1}{\frac{1}{20}} \cdot 0.869\tilde{X} = 6.83 + 17.38\tilde{X}.$$

(b) All values of Y were divided by 20 before estimation.

Answer: Recall from lecture that scaling Y by c scales the estimated parameters by c as well. The new regression line with $\tilde{Y}$ is given

$$\tilde{Y} = \frac{1}{20}\hat{\beta}_0 + \frac{1}{20}\hat{\beta}_1 X = 0.3415 + 0.04345X.$$

(c) All values of X and Y were divided by 20 before estimation.

Answer: Let's construct new variables $\tilde{Y} = \frac{1}{20}Y$, $\tilde{X} = \frac{1}{20}X$ and consider estimating the new regression line

$$\tilde{Y} = \beta_0^\circ + \beta_1^\circ X + \epsilon^\circ.$$

$$\begin{aligned}
\hat{\beta}_1^\circ &= \frac{\sum_{i=1}^n (\tilde{Y}_i - \bar{\tilde{Y}})(\tilde{X}_i - \bar{\tilde{X}})}{\sum_{i=1}^n (\tilde{X}_i - \bar{\tilde{X}})^2} \\
&= \frac{\sum_{i=1}^n (\frac{1}{20}Y_i - \frac{1}{20}\bar{Y})(\frac{1}{20}X_i - \frac{1}{20}\bar{X})}{\sum_{i=1}^n (\frac{1}{20}X - \frac{1}{20}\bar{X})^2} \\
&= \frac{(\frac{1}{20})^2 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{(\frac{1}{20})^2 \sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \hat{\beta}_1 = 0.869
\end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}_0^\circ &= \bar{\tilde{Y}} - \hat{\beta}_1^\circ \bar{\tilde{X}} \\
&= \frac{1}{20}\bar{Y} - \hat{\beta}_1 \frac{1}{20}\bar{X} \\
&= \frac{1}{20}(\underbrace{\bar{Y} - \hat{\beta}_1 \bar{X}}_{=\hat{\beta}_0}) = \frac{1}{20}6.83 = 0.3415
\end{aligned}$$

So the final regression estimated regression line is

$$\tilde{Y} = 0.3415 + 0.869\tilde{X}.$$

2. Given the quantities in the questions below, calculate and interpret R^2:

(a) $\sum_{i=1}^n (Y_i - \bar{Y})^2 = 631.63$ and $\sum_{i=1}^n \hat{\epsilon}_i^2 = 182.85$.

Answer: We calculate

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{182.85}{631.63} = 0.7105.$$

About 71% of the variance in Y is explained by the linear model with X.

(b) $\sum_{i=1}^{n} Y_i^2 = 5930.94$, $\bar{Y} = 16.035$, $n = 20$, and SSR $= 666.72$.

Answer: We know SSR so we need to calculate SST. Using the useful equality from Homework 1

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\bar{Y})^2 \implies SST = \sum_{i=1}^{n} Y_i^2 - n(\bar{Y})^2 = 5930.94 - 20(16.035)^2 = 788.5155.$$

and so

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{666.72}{788.5155} = 0.8455382.$$

About 85% of the variance in Y is explained by the linear model with X.

3. Suppose $R^2 = 0.7911$, SST $= 552.36$, and $n = 20$. Find $\hat{\sigma}_\epsilon^2$.

Answer: Recall that $\hat{\sigma}_\epsilon^2 = \frac{1}{n} \cdot \text{SSE}$. To get SSE use

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \implies 0.7911 = 1 - \frac{\text{SSE}}{552.36} \implies \text{SSE} = 115.388 \implies \hat{\sigma}_\epsilon^2 = \frac{1}{20}115.388 = 5.7694.$$