Linear Regression 2

Iris Kim (6757827) and Ted Tinker (3223468)

April 13, 2017

Question 1, Part A

$$\mathbb{E}[\hat{\beta}_0|X = x_i] = \mathbb{E}[\overline{y} - \hat{\beta}_1\overline{x}|X = x_i] = \mathbb{E}[\overline{y}] - x_i\mathbb{E}[\hat{\beta}_1]$$

In class, we showed $\mathbb{E}[\hat{\beta}_1] = \beta_1$, so

$$\mathbb{E}[\overline{y}] - x_i\mathbb{E}[\hat{\beta}_1] = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}[y_i]\right) - x_i\beta_1$$

$y_i = \beta_0 + \beta_1 x_i + e_i$, so

$$\frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}[y_i]\right) - x_i\beta_1 = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}[\beta_0 + \beta_1 x_i + e_i]\right) - x_i\beta_1 = \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}[\beta_0] + x_i\mathbb{E}[\beta_1] + \mathbb{E}[e_i]\right) - x_i\beta_1$$

$$= (\frac{1}{n}\dot{n}\beta_0) + (\frac{1}{n}\dot{n}x_i\beta_1) - x_i\beta_1 = \beta_0$$

Q1, Part B

$$Var(\hat{\beta}_1) = Var(\overline{y} - \hat{\beta}_1\overline{x})$$

Using the hint,

$$Var(\overline{y} - \hat{\beta}_1\overline{x}) = Var(\overline{y}) - \overline{x}^2 Var(\hat{\beta}_1) - 2\overline{x}Cov(\overline{y}, \hat{\beta}_1)$$

$$Var(\overline{y}) = Var(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i) = \frac{1}{n^2} \times n \times \sigma^2 = \frac{\sigma^2}{n}$$

$$Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n}(x_i - \overline{x})y_i) = Var(\sum_{i=1}^{n} c_i y_i), \text{ with } c_i = \frac{x_i - \overline{x}}{S_{XX}}$$

$$= \sum_{i=1}^{n} c_i^2 Var(y_i) = \sigma^2\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{S_{XX}}\right)^2\sigma^2 = \frac{\sigma^2}{S_{XX}}$$

Using these values in $Var(\overline{y}) - \overline{x}^2 Var(\hat{\beta}_1) - 2\overline{x}Cov(\overline{y}, \hat{\beta}_1)$, with $Cov(\overline{y}, \hat{\beta}_1) = 0$ and $c_i = \frac{x_i - \overline{x}}{S_{XX}}$, we see

$$\frac{\sigma^2}{n} - \overline{x}^2\frac{\sigma^2}{S_{XX}} = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)$$

Q1, Part C

As a linear combination of $y_1, \ldots, y_n$, with X given, and knowing e_i is distributed with mean 0 and variance σ^2, we may combine the results of Part A and B to understand $\hat{\beta}_0|X$ is distributed as

$$(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right))$$

Q1, Part D

In this hypothesis test, we first declare the null hypothesis to be $H_0 : \beta_0 = 0$, and the alternative hypothesis to be $H_a : \beta_0 \neq 0$. The following code find the 95% confidence interval for β_0 given a dataset of the Nile's flow rate for each year:

```
m1 <- lm(Nile$Nile~Nile$time) # Makes m1 the regression line for the Nile
est.intercept = m1$coefficients[1] # Value of beta_0 hat
x.obs = Nile$time # List of all x-values
SXX = sum((x.obs-mean(x.obs))^2) # Finds SXX
y.obs = Nile$Nile # These two lines prepare to find S
y.hat = m1$fitted.values
S = sqrt(sum(y.obs-y.hat)^2)(100-2) # Divide by degrees of freedom - 2
se.intercept = S*sqrt(1/100 + mean(x.obs)^2/SXX) # Standard error of our estimate
t.percent = qt(.95/2,100-2) # Find t-value

est.intercept - t.percent * se.intercept # Lower bound

est.intercept + t.percent * se.intercept # Upper bound
```

This code gives us the 95% confidence interval for β_0 $(4144, 8120)$. This means that given a linear interpretation of the data (which is not necessarily a solid claim, given the nature of the dataset we are using as an example), we expect that in the year 0, we are 95% confidence that the Nile's true flow rate is in that range. Because 0 is not in that range, we have evidence to reject the null hypothesis $\beta_0 = 0$.

Question 2, Part A

We know

$$\mathbb{E}(Y^* - \hat{y}^*) = \mathbb{E}(Y - \hat{y}|X = x^*) = \mathbb{E}(\beta_0 + \beta_1 x^* + e^* - \hat{y}^*|X = x^*).$$

The expected value of e^* is 0, and we defined $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_2 x^*$, so

$$\mathbb{E}(Y - \hat{y}|X = x^*) = \beta_0 + \beta_1 x^* - \mathbb{E}(\hat{\beta}_0 x^* + \hat{\beta}_1 x^*|X = x^*).$$

We showed in Question 1 that $\mathbb{E}(\hat{\beta}_0|X) = \beta_0$, and we learned in class that $\mathbb{E}(\hat{\beta}_1|X)0 = \beta_1$. Therefore,

$$\mathbb{E}(Y - \hat{y}|X = x^*) = \beta_0 + \beta_1 x^* - \beta_0 - \beta_1 x^* = 0.$$

Q2, Part B

From class, we know that $Var(\hat{y}|X = x^*) = \sigma^2\left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\right)$. The covariance of Y and $\hat{y}$ should be 0, so using the hint from Question 1, Part B,

$$Var(Y^* - \hat{y}^*) = Var(Y - \hat{y}|X = x^*) = Var(Y) + Var(\hat{y}|X = x^*) - 0 = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}\right)$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}\right)$$

Q2, Part C

In Part A, we showed $\mathbb{E}(Y^* - \hat{y}^*) = 0$. In Part B, we showed $Var(Y^* - \hat{y}^*) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}\right)$. As the variance in error e^* is normally distributed, we understand $Y^* - \hat{y}^*$ is distributed as

$$N(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{XX}}\right))$$

Question 3, Part A

After uploading the CSV to R-Studio, we ran the following code:

```
salesLine <- lm(playbill$CurrentWeek~playbill$LastWeek)
confint(salesLine)
```

Which output:

```
                          2.5 %       97.5 %
(Intercept)       -1.424433e+04 27854.099443
playbill$LastWeek  9.514971e-01     1.012666
```

Then, the 95% confidence interval for β_1 ranges from 9.514971e-01 to 1.012666. As 1 is inside that range, it is a plausible value for β_1 using this test.

Q3, Part B

Because 1,000 is in the confidence interval for β_0 found in Part A, -1.424433e+04 to 27854.099443, we cannot reject the null hypothesis $H_0 : \beta_0 = 1,000$. This would imply that for if a play earned $0 in the last week, it would be expected to earn $1,000 this week.

Q3, Part C

Having attached playbill.csv and made the linear regression SalesLine, run the code

```
newdata=data.frame(LastWeek=400000)
ystar <- predict(salesLine,newdata,interval='predict')
```

to outout the best fit and a 95% confidence interval for the sales in the current week given $400,000 in sales last week. The best fit is $399637.5.

Q3, Part D

In Part C, we also find lower bound $359832.8 and upper bound $439442.2. Then, $450,000 does not seem like a reasonable estimate for the expected sales of a broadway show given it sold $400,000 the previous week.

Q3, Part E

Because the 1 is within the 95% confidence interval for β_1, this could be an appropriate rule of thumb.

Question 4, Part A

Using

```
houseLine <- lm(indicators$PriceChange~indicators$‘LoanPaymentsOverdue‘)
confint(houseLine)
```

we produce confidence intervals

```
                                          2.5 %     97.5 %
(Intercept)                           -2.532112 11.5611000
indicators$‘\\tLoanPaymentsOverdue‘ -4.163454 -0.3335853
```

Because the 95% confidence interval for β_1 ranges from -4.163454 to -0.3335853, it contains no non-negative values. Hence, there is evidence to say that the slope is negative.

Q4, Part B

With code similar to that used in Q3, Part C,

```
newdata=data.frame(LoanPaymentsOverdue=4)
ystar <- predict(houseLine,newdata,interval='predict')
```

we find the 95% confidence interval for the percentage change in average price given 4% of loans are is $(-13.13784, 4.178667)$. This interval contains 0, so a 0% change would be a feasible value for the expectation.

Question 5, Part A

$$y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - (\overline{y} - \hat{\beta}_1 \overline{x}) - \hat{\beta}_1 x_i$$

$$= y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x})$$

Q5, Part B

$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$, so

$$y_i - \overline{y} = y_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{x}$$

$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, so $y_i - \hat{\beta}_0 = \hat{\beta}_1 x_i - e_i$. Then, because $\mathbb{E}[e_i|X] = 0$,

$$y_i - \overline{y} = \hat{\beta}_1 x_i - \hat{\beta}_1 \overline{x} = \hat{\beta}_1 (x_i - \overline{x})$$

Q5, Part C

Using Part A,

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} (y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x}))(\hat{y}_i - \overline{y})$$

And, using Part B,

$$\sum_{i=1}^{n} (y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x}))(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} (\hat{\beta}_1 (x_i - \overline{x}) - \hat{\beta}_1 (x_i - \overline{x}))(\hat{y}_i - \overline{y}) = 0$$