Stat 615/415 (Regression) HW #3 – Solutions

Regression Inference (sec. 2.1-2.5). Regression slope and intercept. Estimation, testing, and prediction

1. (**2.10**) For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

   (a) What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31^oC?

   (b) How much do families whose disposable income is \$23,500 spend, on the average, for meals away from home?

   (c) How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

   SOLUTION.

   (a) Prediction interval. There is only one particular day tomorrow.

   (b) Confidence interval. We are estimating the mean spending for all families with the given income.

   (c) Prediction interval. As in (a), we are predicting power consumption for one particular month.

2. (**2.11**) A person asks if there is a difference between the "mean response at $X = x$" and the "mean of m new observations at $X = x$". Reply.

   SOLUTION. *Yes, there is a difference. The "mean response at $X = x$" is the population parameter, the average of all responses Y in the population whose $X = x$. The "mean of m new observations at $X = x$" is a sample average of these m responses. The latter is a random variable but the former is constant.*

3. (Continued from HW-2) The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

   In the previous homework, we fit a regression model that predicted the time it will take to transmit a 400 Kbyte file. According to this model, the standard deviation of responses is estimated by $s = s_y\sqrt{\frac{n-1}{n-2}(1-r^2)} = 0.0052$.

   (a) Construct a 95% confidence interval for the regression slope.

   (b) Based on this interval, is the slope significant at the 5% level?

   (c) State the null and alternative hypotheses in (b). Calculate the test statistic and the p-value.

   (d) When you answered questions (b) and (c), it was correct to conduct a two-sided a two-sided test. However, in this given example, why does it make more sense to consider a one-sided, right-tail alternative?

   SOLUTION. *We are given: $n = 30$, $\bar{X} = 126$, $s_x = 35$, $\bar{Y} = 0.04$, $s_y = 0.01$, $r = 0.86$ and $s = 0.0052$. In the previous homework, we calculated the sample regression slope,*

$$b_1 = r\left(\frac{s_y}{s_x}\right) = (0.86)\left(\frac{0.01}{35}\right) = 0.000246.$$

Now, estimate its standard deviation $\sigma(b_1)$ by

$$s_{b_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{s}{s_x\sqrt{n-1}} = \frac{0.0052}{35\sqrt{29}} = 0.0000276.$$

(a) *The 95% confidence interval for the regression slope β_1 is*

$$b_1 \pm t_{0.025} s_{b_1} = 0.000246 \pm (2.045)(0.0000276) = 0.000246 \pm 0.000057 = \boxed{[0.000189,\ 0.000303]}\ ,$$

using the critical value $t_{\alpha/2} = t_{0.025} = 2.048$ from the t-distribution with $n-2 = 28$ degrees of freedom.

(b) *Yes, the slope is significant at the 5% level because the 95% confidence interval does not contain 0. So, $H_0 : \beta_1 = 0$ is rejected in favor of $H_1 : \beta_1 \neq 0$.*

(c) *For testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, the test statistic is*

$$t = \frac{b_1}{s_{b_1}} = \frac{0.000246}{0.0000276} = \boxed{8.913}$$

The corresponding p-value is

$$P = 2\boldsymbol{P}\{t > 8.913\} \boxed{\approx 0}\ ,$$

hence, there is a strong evidence that the slope is not zero, and that the transmission time does depend on the size of the file.

(d) *Clearly, the time it takes to transmit a file cannot be a decreasing function of the size. It may only take more time to transmit more data. Thus, the case of a negative slope $\beta_1 < 0$ leading to a decreasing regression function can safely be eliminated. It is only reasonable to test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$.*

4. (Continued from HW-2) At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$. In the previous homework, we fit a regression model that described how the number of miles per gallon depends on the mileage. According to this model, the standard deviation of responses is estimated by $s = s_y\sqrt{\frac{n-1}{n-2}(1-r^2)} = 3.36$.

(a) You purchase a used car with 35,000 miles on it. Predict the number of miles per gallon. Give a 95% prediction interval for your car and a 95% confidence interval for the average number of miles per gallon of all cars with such a mileage.

(b) Do the given data present a significant evidence that cars with higher mileage are less economic? Formulate appropriate null hypothesis and alternative and conduct the test.

Solution.

(a) *In the previous homework, we calculated the regression line $y = 24.77 - 0.0000395x$ and predicted the number of miles per gallon as*

$$\hat{Y} = 24.77 - (0.0000395)(35,000) = \boxed{23.39 \text{ miles per gallon}}$$

Now, the 95% prediction interval for the miles per gallon of this particular car is

$$\begin{aligned}\hat{Y} \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{S_{xx}}} &= 23.39 \pm (1.97)(3.36)\sqrt{1 + \frac{1}{180} + \frac{(35000-24598)^2}{38333558124}} \\ &= 23.39 \pm 6.65 = \boxed{[16.74,\ 30.04]}\ ,\end{aligned}$$

and the 95% confidence interval for the average number of miles per gallon of all cars with 35,000 miles on them is

$$\begin{aligned}\hat{Y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{S_{xx}}} &= 23.39 \pm (1.97)(3.36)\sqrt{\frac{1}{180} + \frac{(35000-24598)^2}{38333558124}} \\ &= 23.39 \pm 0.61 = \textit{[22.78, 24.00]}\ ,\end{aligned}$$

where $S_{xx} = (n-1)s_x^2 = (179)(14634^2) = 38,333,558,124$*, and* $t_{\alpha/2} = t_{0.025} = 1.97$ *is a critical value from the t-distribution with* $n-2 = 178$ *degrees of freedom.*

(b) *To answer this question, we need to test* $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 < 0$*. The test statistic is*

$$t = \frac{b_1}{s_{b_1}} = \frac{b_1}{s/\sqrt{S_{xx}}} = \frac{-0.0000395}{3.36/\sqrt{38333558124}} = -2.30.$$

The (one-sided) P-value for this test is

$$P = \boldsymbol{P}\{t < -2.30\} = 0.011.$$

The null hypothesis is rejected at any level $\alpha > 0.011$*. Although the sample slope is so small, the given data present a significant evidence,* at any significance level above 0.011, *that cars with higher mileage are less economic.*

5. **(2.52+, Stat-615 only)** In this problem, we develop statistical methods for the regression *intercept*. You can review our inference about the regression slope and follow similar steps.

(a) Derive the expression for the variance $\mathrm{Var}(b_0)$.

(b) Derive the t-statistic for testing $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$. Suppose that the standard regression assumptions are satisfied.

(c) Test significance of the intercept for the Miles Per Gallon data in the previous exercise.

(d) Write the general expression for a $(1-\alpha)100\%$ confidence interval for the population intercept β_0.

(e) Construct a 95% confidence interval for the intercept for the Miles Per Gallon data in the previous exercise.

Hint: In (a), notice again that $b_0 = \sum a_i Y_i$ *is a linear combination of responses* $Y_1, \ldots, Y_n$*. In (b,d), use our general methods for t-tests and confidence intervals, applying them to the regression intercept* β_0 *and its estimator* b_0*.*

SOLUTION.

(a) *Since*

$$b_0 = \overline{Y} - b_1\overline{X} = \frac{1}{n}\sum Y_i - \frac{\sum(X_i-\overline{X})Y_i}{S_{xx}}\overline{X} = \sum\left(\frac{1}{n} - \frac{X_i - \overline{X}}{S_{xx}}\right)Y_i$$

is a linear function of $Y_1, \ldots, Y_n$*, we can find its variance as*

$$\begin{aligned}\mathrm{Var}(b_0) &= \sigma^2 \sum\left(\frac{1}{n} - \frac{X_i-\overline{X}}{S_{xx}}\overline{X}\right)^2 \\ &= \sigma^2\left(\sum\left(\frac{1}{n}\right)^2 + \frac{\sum(X_i-\overline{X})^2}{S_{xx}^2}\overline{X}^2 - 2\frac{\overline{X}\sum(X_i-\overline{X})}{nS_{xx}}\right) \\ &= \sigma^2\left(n\left(\frac{1}{n}\right)^2 + \frac{S_{xx}}{S_{xx}^2}\overline{X}^2 - 0\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right)\end{aligned}$$

(b) $t = \dfrac{\textit{Estimator}}{\textit{SD of this estimator}} = \dfrac{b_0}{s_{b_0}} = \dfrac{b_0}{s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}}}$

(c) *For the Miles Per Gallon data in the previous exercise,* $b_0 = 24.77$, $s = 3.36$, $n = 180$, $\overline{X} = 24598$, *and* $S_{xx} = 38333558124$. *So,*

$$s_{b_0} = 3.36\sqrt{\frac{1}{180}+\frac{24598^2}{38333558124}} = 0.491$$

and

$$t = \frac{b_0}{s_{b_0}} = \frac{24.77}{0.491} = 50.47$$

The p-value for this test is practically 0, so the intercept is significant.

(d) *A* $(1-\alpha)100\%$ *confidence interval for the population intercept is*

$$b_0 \pm t_{\alpha/2}s_{b_0} = b_0 \pm t_{\alpha/2}s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{xx}}}$$

(e) *For the Miles Per Gallon data, the 95% confidence interval for the intercept is*

$$b_0 \pm t_{\alpha/2}s_{b_0} = 24.77 \pm (1.97)(0.491) = 24.77 \pm 0.97 = [23.80,\ 25.74]$$

where the critical value $t_{0.025} = 1.97$ *is already found in the previous exercise.*

6. Computer project (**2.4, 2.13ab, 2.67**).
Grade point average (this data set was already used in Homework-2).

(a) Obtain a 99% confidence interval for β_1. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

(b) Test whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of 0.01. State the alternatives, decision rule, and conclusion.

(c) What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

(d) Obtain a 95 percent confidence interval for the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

(e) Mary Jones obtained a score of 28 on the ACT. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

(f) On the same graph, plot

- the data
- the least squares regression line for ACT scores
- the 95 percent confidence band for the true regression line for ACT scores between 20 and 30.

Does the confidence band suggest that the true regression relation has been precisely estimated? Discuss.

SOLUTION. *Results are based on the R output below.*

(a) *[1.274, 2.954] . It does not include 0, and therefore, at the 1% level of significance, the slope* β_1 *is found significant. So, the director of admissions will conclude that the ACT score is an important variable predicting success of students (their GPA) during their freshman year.*

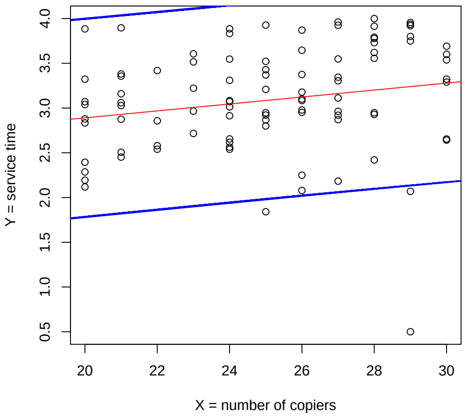
(b) *Test* $H_0 : \beta_1 = 0$ *(no linear association) vs* $H_0 : \beta_1 \neq 0$ *(presence of a linear association). As noted above,* H_0 *is rejected at the 1% level. We conclude that there is significant evidence of a linear association between the ACT score and the freshman GPA.*

(c) *The P-value is* *0.00292* < 0.01 *leading to the same conclusion, rejection of* H_0. *This is consistent with the answers to (a-c).*

(d) *[3.061, 3.341]* . *In a long run of samples, 95% of confidence intervals constructed this way will contain the actual population mean response* $\mu(28) = \mathbf{E}\{Y \mid X = 28\}$.

(e) *[1.959, 4.443]* . *In a long run of samples and students with ACT=28, 95% of prediction intervals constructed this way will contain the actual response* Y.

(f)



R code

```
    # Reading data from the text file, renaming variables, fitting regression
> GPA = read.table("C:\\Data\\Book data\\Chapter  1 Data Sets\\CH01PR19.txt")
> attach(GPA)
> X = V2; Y = V1;
> reg = lm(Y~X)

    # A 99% confidence interval for the slope
> confint(reg,level=0.99)
                  0.5 %     99.5 %
(Intercept) 1.273902675 2.95419590
X           0.005385614 0.07226864

    # Obtaining the p-value for testing the slope
> summary(reg)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
X            0.03883    0.01277   3.040  0.00292 **

    # Confidence interval for the mean response
> predict(reg, data.frame(X=28), interval="confidence")
       fit      lwr      upr
1 3.201209 3.061384 3.341033

    # Prediction interval for the individual response
> predict(reg, data.frame(X=28), interval="prediction")
       fit      lwr      upr
1 3.201209 1.959355 4.443063
```

```
    # Confidence band for the entire population regression line
> n = length(X)
> W = sqrt(qf(0.95,2,n-2))
> Sxx = (n-1)*var(X)
> e = reg$residuals
> s = sqrt( sum(e^2)/(n-2) )
> margin = W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)
> upper.band = predict(reg) + margin
> lower.band = predict(reg) - margin

    # Plots
> plot(X,Y,xlab="X = number of copiers",ylab="Y = service time",xlim=c(20,30))
> abline(reg,col="red")
> lines(X,upper.band,col="blue")
> lines(X,lower.band,col="blue")
```