Stat Methods I - Homework 3

Justin Reising

October 11, 2018

1.5) When asked to state the simple linear regression model, a student wrote it as follows:(Do you agree?)

$$E[Y_i] = \beta_0 + \beta_1 X_i + \epsilon_i$$

I disagree with the student's response. Note that the simple linear regression model has a slight difference and is as follows: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Note that the response variable Y_i has a probability distribution whose mean is $E[Y_i] = \beta_0 + \beta_1 X_i$ which is known as the deterministic component. Also note that ϵ_i falls off for the $E[Y_i]$ since $E[\epsilon_i] = 0$.

1.7) In a simulation excercise, regression model (1.1) applies with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for $X = 5$.

(a) Can you state the exact probability that Y will fall between 195 and 205. Explain.

Well, based on the the following question, No. But we need the assumption to be satisfied that each ϵ_i $N(0, \sigma^2)$, which we can formally test with the Shapiro-Wilk Test.

(b) If the normal error regression (1.24) is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

With the assumption that each ϵ_i $N(0, \sigma^2)$ as with the normal error regression, then we can find the probability $195 \leq Y \leq 205$ such that with $E[Y] = 100 + 20(5) = 200$ and $\sigma^2 = 25$ by calculating normal probabilities:

```r
pnorm(205,mean = 200,sd = 5) - pnorm(195, mean = 200, sd = 5)
```

```
## [1] 0.6826895
```

1.11) The regression function relating production output by an employee after taking a training program (Y) to the production output before the training program (X) is $E[Y] = 20 + .95X$, where X ranges from 40 to 100. An observer concludes that the training program does not raise production output on the average because β_1 is not greater than 1. Comment.

Well, there is no context for the unit of measure for "production output" as to whether it is some number of units of product being produced or a percentage, etc. To me, a simple linear regression model is not ideal for this type of investigation. The reason being, take two employees, A adn B, where their prior production levels are 40 and 100 respectively. If this unit of measure is percentage, then employee B has no room for improvement and the model will indicate the training had no effect. For employee B however, there is 60 percent of improvement possible and any type of additional training is likely to increase production. We also must consider variation of production between time frames (days,weeks,etc.). How do we know if the observed difference in production is not just a result from likely variations? This type of investigation requires an experimental design in my opinion. All of that aside, after introducing a training program, I would expect those with lower prior production to have higher increases in production after the training than those with higher prior production, meaning that β_1 should have a negative slope. Personally, I think skill and ability, when measured appropriately, is nonlinear in that the better you are at something, the harder you have to work to see improvement.

1.16) Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

I'm not sure what it means to be "fully" valid as oppose to "partially" or "mostly" valid but it does not seem to me that the requirement of the distribution of Y to be normal needs to be satisfied, just as long as the Predictor variables are linearly correlated and it is unbiased. The consequence would be that we would obtain non-normal parameters.

1.17) A person states that b_0 and b_1 in the fitted regression function (1.13) can be estimated by the method of least squares. Comment.

Note that the primary objective of the least squares method is to find values of b_0 and b_1 that minimize $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$. These values are calculated from the sampled data and are unbiased estimators of β_0 and β_1.

1.18) According to (1.17) $\sum e_i = 0$ when regression model (1.1) is fitted to a set of n cases by the method of least squares. Is it also true that $\sum \epsilon_i = 0$? Comment.

Note that each e_i is an unbiased estimator of ϵ_i and the assumption that that each ϵ_i $N(0, \sigma^2)$ so we have the $\sum e_i = 0 \implies \sum \epsilon_i = 0$.

1.19) Grade Point Average. The director of admission of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the from the ACT score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

Table 1: Grade Point Average (Header)

GPA	ACT
3.897	21
3.885	14
3.778	28
2.540	22
3.028	21
3.865	31

(a) Obtain the least squares estimates for β_0 and β_1, and state the estimated regression function.

```
b1<- cor(GPA,ACT)*(sd(GPA)/sd(ACT))
b0<- mean(GPA)-(b1*mean(ACT))
b0
```

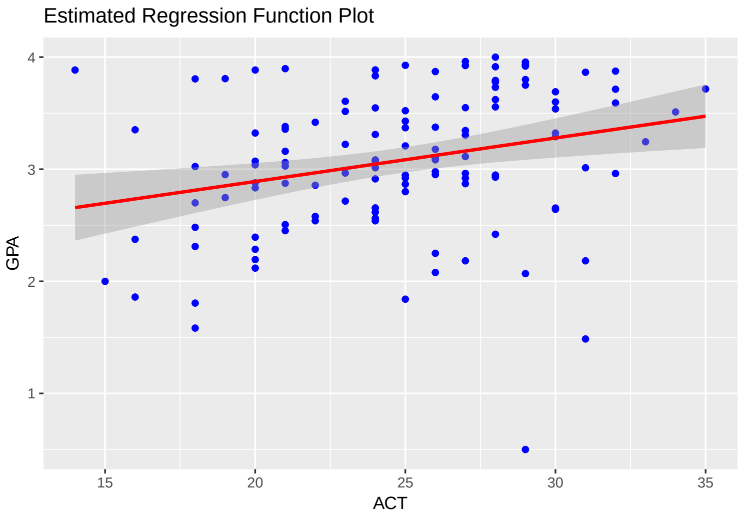
```
## [1] 2.114049
```

```
b1
```

```
## [1] 0.03882713
```

$$E[Y_i] = 2.114 + 0.039X_i$$

(b) Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?



Note that there is some wide variation in the distribution of the residuals from the regression line. One notable observation is a student that scored what looks to be a 29 on the ACT ended up with 0.5 GPA which may be a considerable case for removal of an outlier. Overall it seems to be a decent fit for the data although it does seem quite scattered.

(c) Obtain a point estimate of the mean freshman GPA for the students with ACT scores $X = 30$.

$$\hat{Y} = 2.114 + 0.039(30) = 3.284$$

(d) What is the point estimate of the change in the mean response when entrance test score increases by one point?

$$0.039$$

1.22) Plastic Hardness: Refer to problem 1.3 and 1.14. Sixteen batches of plastic were made and from each batch one test item was molded. Each item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are as shown below, X is the elapsed time in hours, and Y is hardness in Brinell units. Assume that first-orderregression model (1.1) is appropriate.

Table 2: Plastic Hardness (Header)

Hardness	Time
199	16
205	16
196	16
200	16
218	24
220	24

(a) Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here?

```r
b1<- cor(Hardness,Time)*(sd(Time)/sd(Hardness))
b0<- mean(Hardness)-(b1*mean(Time))
b0
```

```
## [1] 212.1693
```

```r
b1
```

```
## [1] 0.4783303
```

$$E[Y_i] = 212.1693 + 0.4783X_i$$

(b) Obtain a point estimate of the mean hardnes when $X = 40$.

$$\hat{Y} = 212.1693 + 0.4783(40) = 231.3$$

(c) Obtain a point estimate of the change in the mean hardnes when X increases by 1 hour.

1.23) Refer to Grade Point Average. Problem 1.19.

(b) Estimate σ^2 and σ. In what units is σ expressed?

```
grade.reg<-lm(GPA~ACT)
SSE_grade<- sum((grade.reg$residuals)^2)
MSE_grade<- SSE_grade/(length(ACT)-2)
MSE_grade
```

```
## [1] 0.3882848
```

```
sqrt(MSE_grade)
```

```
## [1] 0.623125
```

$$\sigma^2 = 0.388 \quad \sigma = 0.623$$

Sigma is measured in uits of GPA.

1.26) Refer to Plastic Hardness Problem 1.22

(b) Estimate σ^2 and
$sigma$. In what units is σ expressed?

```
plastic.reg<-lm(Hardness~Time)
SSE_plastic<- sum((plastic.reg$residuals)^2)
MSE_plastic<- SSE_plastic/(length(Time)-2)
MSE_plastic
```

```
## [1] 10.45893
```

```
sqrt(MSE_plastic)
```

```
## [1] 3.234027
```

$$\sigma^2 = 10.459 \quad \sigma = 3.234$$

Sigma is measured in Brinell units.

1.30) Refer to regression model (1.1). What is the implication for the regression function if $\beta_1 = 0$ so that the model is $Y_i = \beta_0 + \epsilon_i$ How would the regression function plot on a graph?

The regression function would be constant at β_0

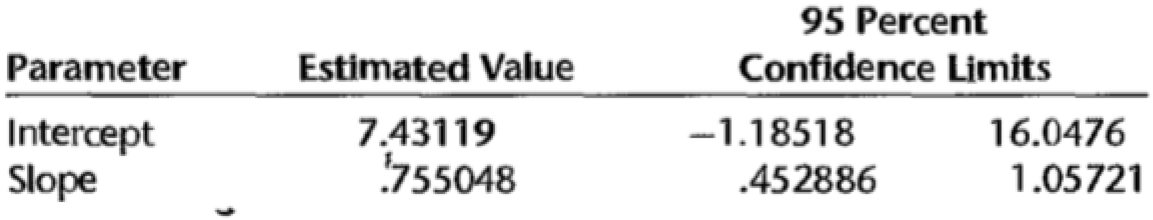
1.33) (Calculus needed) Refer to the regression model $Y_i = \beta_0 + \epsilon_i$ in Exercise 1.30. Derive the least squares estimator of β_0 for this model.

$$Q = \sum_{i=1}^{n}(Y_i - \beta_0)^2$$

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0) &= 0 \\
-2(\sum_{i=1}^{n} Y_i - n\beta_0) &= 0 \\
\sum_{i=1}^{n} Y_i - n\beta_0 &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} Y_i &= \beta_0 \\
\bar{Y} &= \beta_0
\end{aligned}$$

1.40) In fitting regression model (1.1), it was found that observation Y_i fell directly on the fitted regression line (i.e $Y_i = \hat{Y}_i$). If this case were deleted, would the least squares regression line fitted to the remaining $n - 1$ cases be changed? [*HINT* What is the contribution of case i to the least squares criterion Q in (1.8)?]

This would be the equivalent of adding a value equal to the average of a sampling distribution to the sample over and over again. Although n increases, the shape, center, and spread of the distribution remains the same.



		95 Percent	
Parameter	Estimated Value	Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

Figure 1: Regression Coefficients 2.1

2.1) A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in millions of dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the information in Figure 1 above on the regression coefficients:

(a) The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

> The implied level of significance is 95% based on the confidence limits in the figure above. Since the estimated value of of the Slope (β_0) $\neq 0$, then I would have to agree the there is a linear association between Y and X. Whiele I cannot formally assess the claim, it appears to be a warranted conclusion.

2.2) In a test of the alternatives $H_0 : \beta_1 \leq 0$ versus $H_1 : \beta_1 > 0$, an analyst concluded H_0. Does this conclusion imply that there is no linear association between X and Y? Explain.

> As the analyst fails to reject the null hypothesis H_0, we haven't really answered the question of whether or not X and Y are linearly associated. We have however confirmed the claim that X and Y are either negatively correlated or not at all. IF we look at the confidence limits for a 95% or 99% confidence levels the nwe may be able to ascertain if 0 is in the intervals or not, which may give us an indication as to whether or not X and Y are linearly associated.

2.3) A member of a student team is playing an interactive marketing game recieved the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one output of the team's products:

$$\text{Estimated regression equation: } \hat{Y} = 350.7 - 0.18X$$

$$\text{Two-Sided p-value for est. slope: } 0.91$$

The student states: "The message I get here is that the more we spend on advertising this product, the fewer units we sell!!" Comment.

> At first glance of the estimated regression equation, it is true that there is a negative slope, however upon further investigation of the p-value below, we can see that there is not a significance level worth testing that would allow for a p-value that high to conclude the estimated slope in the estimated regression equation.

2.4) Refer to the GPA problem 1.19.

(a) Obtain a 99 percent confidence interval for β_1. Interpret you confidence interval. Does it include 0? Why might the director admissions be interested in whether the confidence interval includes 0?

```
confint(grade.reg,"ACT",level=.99)
```

```
##             0.5 %     99.5 %
## ACT 0.005385614 0.07226864
```

> As you can see above, with 99% confidence the true change in GPA per ACT score point is between 0.0054 and 0.0723. Note that while the lower limit for the 99% Confidence Interval is close to 0, 0 is not in the confidence interval. The director of the admissions would be interested in this because if 0 were in this interval, then it would indicate that there could be no change in GPA per ACT unit and these two variables may not be correlated.

(b) Test, using a test statstic t^*, whether or not the linear association exists between student's ACT score (X) and GPA (Y). Use a significance level of 0.01. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

```
alpha <- 0.01
t_star<- (grade.reg$coefficients[2])/(sqrt(MSE_grade/sum((ACT-mean(ACT))^2)))
t_star
```

```
##      ACT
## 3.039777
```

```
t_star > qt(1-alpha/2, length(ACT)-2)
```

```
##  ACT
## TRUE
```

> Note that since $t^* > t_{(1-\frac{\alpha}{2}, n-2)}$ the rejection rule for the critical value holds as we reject the null hypothesis H_0. As we saw in the 99% confidence interval, we conclude that there is a linear association between ACT scores and GPA.

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

```
pval_grade<- 2*pt(-t_star, length(ACT)-2)
pval_grade
```

```
##         ACT 
## 0.002916604
```

> We obtain a p-value of 0.00292 which is less than our significance level of 0.01 and confirms our conclusion in part (b). This p value indicates that due to random variation, we expect to see approximately 3/1000 samples in which no correlation between X and Y would be evident. So within our predetermined accepted significance level, this result strongly supports the conlcusion in part (b).

2.7) Refer to Plastic Hardness Problem 1.22.

(a) Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99% confidence interval. Interpet your interval estimate.

(b) The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = 0.1$. State tha alternatives, decision rule, and conclusion. What is the P-value of the test?

$$H_0 : \beta_1 = 2 \quad H_1 : \beta_1 \neq 2$$

```
alpha <- 0.01
t_star<- (plastic.reg$coefficients[2]-2)/(sqrt(MSE_plastic/sum((Time-mean(Time))^2)))
t_star > qt(1-alpha/2, length(Time)-2)
```

```
##  Time 
## FALSE
```

```
pval_plastic_2<- 2*pt(-t_star, length(Time)-2)
pval_plastic_2
```

```
##      Time 
## 0.7094445
```

> Note that since $t^* \ngtr t_{(1-\frac{\alpha}{2}, n-2)}$ the rejection rule for the critical value fails so we fail to reject the null hypothesis H_0. With 99% confidence, we conclude that the manufacturer's claim is consistent. The P-value, 0.0709 for this hypothesis test is well over $\alpha = 0.01$, our predetermined level of acceptance.

2.9) Refer to figure 2.2. A student noting that $s\{b_1\}$ is furnished in the printout, asks why $s\left\{\hat{Y_h}\right\}$ is not also given. Discuss.

> It is furnished indirectly as the square root of the MSE for the Error since the standard deviation of a predicted value on the regression line is the standard deviation of the error.

2.13) Refer ot Grade Point Average Problem 1.19

(a) Obtain a 95% interval estimate of the mean Freshman GPA for students whose ACT test score is 28. Interpret you confidence interval.

```r
pred_act<-data.frame(ACT = 28)
predict(grade.reg, pred_act, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

> With 95% confidence, the estimated true mean Freshan GPA for students whose ACT score is 28 is approximately between 3.06 and 3.34.

(b) Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95% prediction interval. Interpret your prediction interval.

```r
predict(grade.reg, pred_act, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

> With 95% probability, the estimated true mean Freshman GPA for students whose observed ACT score is 28 is approximately between 1.96 and 4.44. (Assuming the max GPA is 4.0, then the prediction interval would be 1.96 to 4.0)

(c) Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

> The prediction interval is wider than the confidence interval because the predicted value accounts for added variability from a new sample estimate.

2.16) Refer to Plastic Hardness Problem 1.22.

(a) Obtain a 98% confidence interval for the mean hardness molded items with an elapsed time of 30 hourse. Interpret your confidence interval.

```r
pred_time<-data.frame(Time = 30)
predict(plastic.reg, pred_time, interval = "confidence", level = 0.98)
```

```
##        fit      lwr      upr
## 1 229.6312 227.4569 231.8056
```

> With 98% confidence the true mean hardness for 30 hours time elapsed is between 227.46 and 231.81 Brinell units.

(b) Obtain a 98% prediction interval for the hardness of a newly molded test item with an elapsed time of 30 hourse. Interpret your confidence interval.

```r
predict(plastic.reg, pred_time, interval = "prediction", level = 0.98)
```

```
##        fit      lwr     upr
## 1 229.6312 220.8695 238.393
```

2.17) An analyst fitted a normal error regression model (2.1) and conducted an F test of $\beta_1 = 0$ vs $\beta_1 \neq 0$. The P-Value of the test was 0.033, and the analyst concluded $H_0 : \beta_1 \neq 0$. Was the α level used by the analyst greater than or smaller than 0.033? If the α level had been 0.01, what would have been the appropriate conclusion?

> As the analyst failed to reject the null hypothesis, the α level used had to be greater than 0.033 (presumably 0.05). If the analyst had ised an α level of 0.01, they would have rejected the null hypothesis.

2.18) For conducting statistical tests concerning the parameter β_1, why is the t-test more versatile than the F test?

> The t-test is more versatile since it can be used to test for one-sided and two-sided alternatives where the F-test is a one sided test since the F distribution is just the square of a t distribution.

2.26 Refer to Plastic Hardness Problem 1.22

(a) Set up the ANOVA table.

```
anova_plastic<- anova(plastic.reg)
anova_plastic
```

```
## Analysis of Variance Table
##
## Response: Hardness
##           Df Sum Sq Mean Sq F value            Pr(>F)
## Time       1 5297.5  5297.5  506.51 0.000000000002159 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Test by means of an F test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = 0.01$. State the alternatives, decision rule, and conclusion.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

```
f_star<- anova_plastic$`F value`[1]
f_star > qf(.99,1,length(Time)-2)
```

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

> Note that using the critical value approach, since $F^* > F_{(1-\alpha,1,n-2)}$, we reject the null hypothesis, H_0 and conclude that there is a linear association between the hardness of the plastic and the elapsed time.

(d) Calculate R^2 and r and interpret the values.

```
r<- cor(Time, Hardness)
r
```

```
## [1] 0.9864599
```

```
r^2
```

```
## [1] 0.9731031
```

> 'r' is the sample correlation between 'Time' and 'Hardness' which we can see is highly correlated with a value of 0.986. With $R^2 = 0.9731$, we interpret this as 97.31% of the variation in Hardness is explained by the linear relationship with time elapsed.