YuHW02ST430

Haozhe (Jerry) Yu

2023-09-21

Question 1

For this problem, use the grade point average data described in KNNL Problem #1.19

```
actgpa <- as_tibble(read.delim2("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatas
            sep = "",
            strip.white=TRUE,
            col.name = c("GPA","ACT")
            ))
actgpa$GPA <- as.numeric(actgpa$GPA)
actgpa
```

```
## # A tibble: 119 x 2
##      GPA   ACT
##    <dbl> <int>
##  1  3.88    14
##  2  3.78    28
##  3  2.54    22
##  4  3.03    21
##  5  3.86    31
##  6  2.96    32
##  7  3.96    27
##  8  0.5     29
##  9  3.18    26
## 10  3.31    24
## # i 109 more rows
```

```
actgpam <- lm(GPA~ ACT,data=actgpa)

sactgpam <- summary(actgpam)

# Questions Answered will be with in-line code. I will provide an example of the inline code I use to d
```

a. What is the estimate of sigma from this analysis?

The estimate for σ is 0.619263 with 0.619263 degrees of freedom.

Example inline code: `sactgpam$sigma`

b. Give a point estimate and 95% confidence interval for the slope and interpret each of these in words.

The point estimate for the slope is 0.0403633 and the 95% confidence interval for the slope is 0.0151496 to 0.0655771. This indicates that 95% of intervals from samples of the same size will contain β_1.

c. Give a point estimate and 95% confidence interval for the y-intercept.

The point estimate for the y-intercept is 2.0678879 and the 95% confidence interval for the y-intercept is 1.4336451 to 2.7021307. This indicates that 95% of intervals from samples of the same size will contain β_0.

d. Obtain a 95% interval estimate of the mean GPA for students whose ACT test score is 28. Interpret your confidence interval.

```
oned <- tibble(ACT=28)
```

The 95% confidence interval for students whose ACT score is 28 is 3.0590342 to 3.3370875. This indicates that 95% of intervals from samples of the same size will contain the true mean GPA of students whose ACT is 28.

e. Predict GPA using a 95% prediction interval for students whose ACT test score is 28.

The 95% prediction interval for students whose ACT score is 28 is 1.963788 to 4.4323337. This indicates that 95% of students in the population who scored a 28 on their ACT will have a GPA in this range.

f. Would it be reasonable to consider inference on the intercept for this problem? Please provide justification for your answer.

No. It would not be reasonable to consider inference on the intercept for this problem. The ACT is scored on a scale of 1 to 36, so it is impossible for a student to score a 0 on the intercept, making inference on the intercept unreasonable.

g. For each of the following hypothesis tests, give the value of the test statistic, the degrees of freedom, the p-value (if you cannot obtain the pvalue, give the critical value for the test statistic), and clearly state your conclusion.

```
#get t values for hypothesis testing
t0 <- summary(actgpam)[[4]][2,3]
p20 <- summary(actgpam)[[4]][2,4]
actgpam07 <- lm(GPA~ACT,data = actgpa,offset = 0.07*ACT)
t07 <- summary(actgpam07)[[4]][2,3]
p207 <- summary(actgpam07)[[4]][2,4]

pv <- as.data.frame(do.call(rbind, list(
  c("2 sided, 0.00",
```

```
    p20,
    "as p <0.05, we reject H0 that beta1 =0 at alpha =0.05"),
  c("lower tail, 0.00",
    pt(t0,117,lower.tail = TRUE),
    "as p >0.05, we fail to reject H0 at alpha=0.05, negative slope not statistically supported"),
  c("upper tail, 0.00",
    pt(t0,117,lower.tail = FALSE),
    "as p <0.05, we reject H0 at alpha=0.05, positive slope statistically supported"),
  c("2 sided, 0.07",
    p207,
    "as p <0.05, we reject H0 that beta1 = 0.07 at alpha =0.05"),
  c("lower tail, 0.07",
    pt(t07,117,lower.tail = TRUE),
    "as p <0.05, we reject H0 at alpha=0.05, slope less than 0.07 statistically supported"),
  c("upper tail, 0.07",
    pt(t07,117,lower.tail = FALSE),
    "as p >0.05, we fail to reject H0 at alpha=0.05, slope more than 0.07 not statistically supported")
)))
names(pv) <- c("Hypothesis","p-value", "Conclusion")
as_tibble(pv)
```

```
## # A tibble: 6 x 3
##   Hypothesis       `p-value`            Conclusion
##   <chr>            <chr>                <chr>
## 1 2 sided, 0.00    0.00194362480062543  as p <0.05, we reject H0 that beta1 =0 ~
## 2 lower tail, 0.00 0.999028187599687    as p >0.05, we fail to reject H0 at alp~
## 3 upper tail, 0.00 0.000971812400312717 as p <0.05, we reject H0 at alpha=0.05,~
## 4 2 sided, 0.07    0.0216380964185368   as p <0.05, we reject H0 that beta1 = 0~
## 5 lower tail, 0.07 0.0108190482092684   as p <0.05, we reject H0 at alpha=0.05,~
## 6 upper tail, 0.07 0.989180951790732    as p >0.05, we fail to reject H0 at alp~
```

```
pv
```

```
##         Hypothesis              p-value
## 1    2 sided, 0.00  0.00194362480062543
## 2 lower tail, 0.00    0.999028187599687
## 3 upper tail, 0.00 0.000971812400312717
## 4    2 sided, 0.07   0.0216380964185368
## 5 lower tail, 0.07   0.0108190482092684
## 6 upper tail, 0.07    0.989180951790732
##                                                                                Conclusion
## 1                                          as p <0.05, we reject H0 that beta1 =0 at alpha =0.05
## 2       as p >0.05, we fail to reject H0 at alpha=0.05, negative slope not statistically supported
## 3                  as p <0.05, we reject H0 at alpha=0.05, positive slope statistically supported
## 4                                       as p <0.05, we reject H0 that beta1 = 0.07 at alpha =0.05
## 5             as p <0.05, we reject H0 at alpha=0.05, slope less than 0.07 statistically supported
## 6 as p >0.05, we fail to reject H0 at alpha=0.05, slope more than 0.07 not statistically supported
```

Question 2

For this problem use the "plastic hardness" data described in the text with problem 1.22 on page 36. Make sure you understand which column is X and which is Y and read in the data

accordingly.

```
phard <- as_tibble(read.table("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdataset
            sep = "",
            strip.white=TRUE,
            col.name = c("Hardness_Brinell","Hrs_Elapsed")
            ))

phard
```

```
## # A tibble: 16 x 2
##    Hardness_Brinell Hrs_Elapsed
##               <dbl>       <dbl>
##  1              199          16
##  2              205          16
##  3              196          16
##  4              200          16
##  5              218          24
##  6              220          24
##  7              215          24
##  8              223          24
##  9              237          32
## 10              234          32
## 11              235          32
## 12              230          32
## 13              250          40
## 14              248          40
## 15              253          40
## 16              246          40
```

a) Run the linear regression to predict hardness from time and state the estimated regression equation. Give a 95% confidence interval for the slope. Explain to someone not familiar with statistics what this confidence interval means. Remember that time is measured in hours and hardness is measured in Brinell units.

```
plm <- lm(Hardness_Brinell~Hrs_Elapsed,data=phard)
```

The 95% confidence interval for the slope is 1.8404996 to 2.2282504. What this means is that if you were to take tons more items of plastic from each batch, we would estimate that for 95% of those batches, the actual rate of increase in hardness in Brinell units per hour would be in this interval. So basically, if we were to repeat this experiment an infinite amount of times, 95% of the confidence intervals we get in each experiment will contain the actual rate of the plastic's increase in strength.

example inline code: `confint(plm,level = 0.95)`

b) Describe the results of the significance test for the slope that you get in your software output. State the hypotheses being tested, the test statistic with degrees of freedom, the P-value, and your conclusion in a brief sentence.

We are testing the 2 tailed hypothesis with an H0 that that slope =0 (not linear) and an Ha that the slope =/= 0, as our test statistic is 22.51 with 14 degrees of freedom, we get a p value of 2.163e-12, allowing us to reject our null hypothesis at α =0.05.

c) c) Explain why or why not inference on the intercept is reasonable (i.e., of interest) in this case.

Inference on the intercept is not reasonable as you would not feasibly use plastic before allowing it to cool off.

d) Continue with the same dataset. Give an estimate of the mean hardness that you would expect after 36 and 43 hours, and a 95% confidence interval for each estimate. Which confidence interval is wider and why is it wider?

```
f3 <- tibble(Hrs_Elapsed=c(36,43))

pci <- predict(plm,f3,interval="confidence",level=0.95)
pconf <- tibble(
  Hrs_Elapsed =f3[["Hrs_Elapsed"]],
  Estimated_Mean = pci[, "fit"],
  Lower_Bound = pci[, "lwr"],
  Upper_Bound = pci[, "upr"]
)
pconf
```

```
## # A tibble: 2 x 4
##   Hrs_Elapsed Estimated_Mean Lower_Bound Upper_Bound
##         <dbl>          <dbl>       <dbl>       <dbl>
## 1          36           242.        240.        244.
## 2          43           256.        253.        259.
```

The confidence interval for for 43 hours is wider as 43 is farther from the center of our data. As Variance is a quadratic function it grows as we get farther away from the center.

e) Again, using the same dataset, give a prediction for the hardness that you would expect for an individual piece of plastic after 43 hours; and a 95% prediction interval for this quantity.

```
pp <- predict(plm, f3[2, 1], interval = "predict", level = 0.95)

ppred <- tibble(
  Hrs_Elapsed = f3[[2, 1]],
  Predicted_Hardness = pp[, "fit"],
```

```
  Lower_Bound = pp[, "lwr"],
  Upper_Bound = pp[, "upr"]
)

ppred
```

```
## # A tibble: 1 x 4
##   Hrs_Elapsed Predicted_Hardness Lower_Bound Upper_Bound
##         <dbl>              <dbl>       <dbl>       <dbl>
## 1          43               256.        248.        264.
```

Hardness is in Brinell units.

Question 3

An investigative study collected 40 observations from the Wabash river at random locations near Lafayette. Each observation consisted of a measure of water pH (X) and fish count (Y). The researchers are interested in how the acidity of the water affects the number of fishes. Complete the following ANOVA table for the regression analysis. State the null and alternative hypotheses for the F-test as well as your conclusion in sentence form. You may use the critical F (critical t) approach or the p-value approach.

```
q3 <- tibble(
  "Source" = c("Model","Error","Corrected Total"),
  "degrees of freedom" = c(1,38,39),
  "Sum of Squares" = c(55.30,4.70,60.00),
  "Mean Square" = c(55.30,4.70/38,NA),
  "F-value" = c(55.30/(4.70/38),NA,NA),
  "P-value" = c((1-pf(55.30/(4.70/38),1,38)),NA,NA)
)
q3
```

```
## # A tibble: 3 x 6
##   Source `degrees of freedom` `Sum of Squares` `Mean Square` `F-value` `P-value`
##   <chr>                <dbl>            <dbl>         <dbl>     <dbl>     <dbl>
## 1 Model                    1             55.3        55.3        447.         0
## 2 Error                   38              4.7         0.124       NA         NA
## 3 Corre~                  39             60          NA           NA         NA
```

As the p value is 0, we reject the null hypothesis and conclude that there is evidence for a linear association between water pH and fish count.

Question 4

4. For this problem, use the surgical unit data described in KNNL. Page 350 and table # 9.1. CH09TA01

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation.

- Original data: 108 patients
- Preliminary study: the first 54 patients with the first four variables

Variables: - X1: blood clotting score - X2: prognostic index - X3: enzyme function test score - X4: liver function test score - X5: age, in years - X6: indicator variable for gender (0=male; 1=female) - X7 and X8: indicator variables for history of alcohol use:

The response variable (Y) was the number of weeks the patients survived after the operation.

Read these variables into R

```
surg.data <- read.table(
  "Datasets/Surgical Unit.txt",
  header = FALSE,
  col.names = c(
    "clot",
    "PI",
    "enzy",
    "liver",
    "age",
    "gender",
    "mod_use",
    "heavy_use",
    "sur_time",
    "ln_sur_time"
  )
)

attach(surg.data)
gender = factor(gender)

surg.data1 <- surg.data[, seq(1, 4)] # new data set

surm <- lm(sur_time~clot,surg.data)

ssurm <- summary(surm)

cp <- tibble(clot=7.5)
```

a. What is the estimate of sigma from this analysis?

The estimate for σ is 376.3276211 with 52 degrees of freedom.

b. Give a point estimate and 95% confidence interval for the slope and interpret each of these in words.

The point estimate of for the slope is 85.9083177 and the 95% confidence interval for the slope is 21.200545 to 150.6160904. This indicates that 95% of intervals from samples of the same size will contain β_1.

c. Give a point estimate and 95% confidence interval for the y-intercept.

The point estimate for the slope is 205.2561551 and the 95% confidence interval for the y-intercept is -182.8236829 to 593.335993. This indicates that 95% of intervals from samples of the same size will contain

β_0.

d. Obtain a 95% interval estimate of the mean sur_time for patients whose clot score is 7.5. Interpret your confidence interval.

The 95% confidence interval for patients whose clot score is 7.5 is 698.2426072 to 1000.8944688. This indicates that 95% of intervals from samples of the same size will contain the true mean survival time of patients with a clot score of 7.5.

e. Predict sur_time using a 95% prediction interval for patients whose clot score is 7.5.

The 95% prediction interval for patients whose clot score is 7.5 is 79.3990398 to 1619.7380363.