p8130 homework 5

Jeffrey Liang

11/15/2020

Problem 1

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  antibody %>% pull(Altered) and antibody %>% pull(Normal)
## W = 8582, p-value = 0.01
## alternative hypothesis: true location shift is not equal to 0
```

Problem 2

a) $Y_i = \beta_0 + \beta_1 X + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

so $Y_i \sim N(\beta_0 + \beta_1 X, \sigma^2)$

$$L(\beta_0, \beta_1, \sigma^2) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}}$$

$$\log L(\beta_0, \beta_1, \sigma^2) = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{n}{2}\log(\sigma^2) - \sum \left(\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial \beta_0} \log L = -\frac{1}{\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_i) \quad \text{——— (1)}$$

$$\frac{\partial}{\partial \beta_1} \log L = -\frac{1}{\sigma^2} \sum X_i (Y_i - \beta_0 - \beta_1 X_i) \quad \text{——— (2)}$$

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{n}{2\sigma^2} - \left(\frac{1}{2\sigma^2}\right)^1 \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_i)^2 \quad \leftarrow (3)$$

Leaving (1), (2), (3) equal to 0.

$$(1) = 0 = \sum (Y_i - \beta_0 - \beta_1 X_i)$$

$$= \sum Y_i - N\beta_0 - \sum X_i \beta_1$$

$$\Rightarrow \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$(2) = 0 = \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

$$= \sum Y_i X_i - \beta_0 X_i - \beta_1 X_i^2$$

$$= \sum Y_i X_i - (\bar{Y} - \beta_1 \bar{X}) X_i - \beta_1 X_i^2$$

$$= \sum Y_i X_i - \bar{Y} X_i + \beta_1 X_i \bar{X} - \beta_1 X_i^2$$

$$= \sum Y_i X_i - \bar{Y} X_i + \beta_1 X_i (\bar{X} - X_i)$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum Y_i X_i - N\bar{Y}\bar{X}}{\sum X_i^2 - N\bar{X}^2}$$

$$\Rightarrow \hat{\beta}_0 = \bar{Y} - \bar{X} \cdot \frac{\sum Y_i X_i - N\bar{Y}\bar{X}}{\sum X_i^2 - N\bar{X}^2}$$

b)

$$\sum o_i = \sum Y - \hat{Y}_i$$

$$= \sum Y - \hat{\beta}_1 X_i - \hat{\beta}_0$$

$$= N\bar{Y} - \sum \hat{\beta}_1 X_i - \hat{\beta}_0$$

$$= N\bar{Y} - N E[\hat{\beta}_1 X - \hat{\beta}_0]$$

$$= N\bar{Y} - N(\beta_1 \bar{X} - \beta_0)$$

with Φ $= N\bar{Y} - N(\beta_1 \bar{X} - \beta_1 \bar{X} + \bar{Y})$

$= 0$ □

Problem 3

The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's GPA at the end of the freshman year (Y) can be predicted from the ACT test score (X). Use data 'GPA.csv' to answer the following questions: You can use R to perform/check the calculations, but you need to show the formulae where asked to do so.

1. Generate a scatter plot and test whether 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 0.05. Write the hypotheses, test statistics, critical value and decision rule with interpretation in the context of the problem. (7.5p)
2. Write the estimated regression line equation in the context of this problem. (2.5p)
3. Obtain a 95% confidence interval for 1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero? (2.5p)
4. Obtain a 95% interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval. Hint: Use R function predict(). (2.5p)
5. Anne obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95% prediction interval. Interpret your prediction interval. Hint: Use R function predict(). (2.5p)
6. Is the prediction interval in part 5) wider than the confidence interval in part 4)? Explain. (2.5p)

PROOF

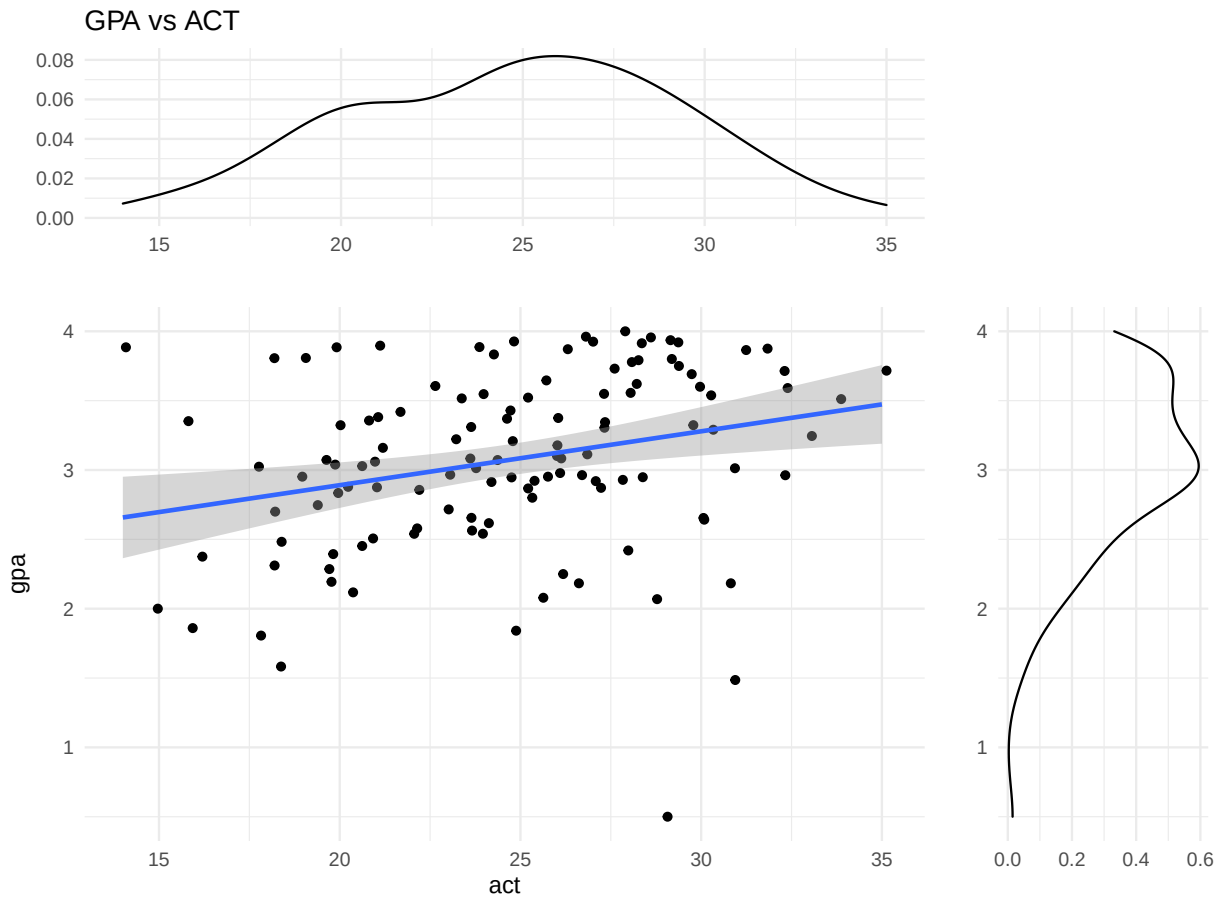
1.

Table 1: Data summary

Name	gpa
Number of rows	120
Number of columns	2
Column type frequency:	
numeric	2
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
gpa	0	1	3.07	0.64	0.5	2.69	3.08	3.59	4
act	0	1	24.73	4.47	14.0	21.00	25.00	28.00	35



2. $GPA = \beta_0 + \beta_1 * ACT$

```
##
## Call:
## lm(formula = gpa ~ act, data = gpa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7400 -0.3383  0.0406  0.4406  1.2274
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1140     0.3209    6.59  1.3e-09 ***
## act           0.0388     0.0128    3.04   0.0029 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.623 on 118 degrees of freedom
## Multiple R-squared:  0.0726, Adjusted R-squared:  0.0648
## F-statistic: 9.24 on 1 and 118 DF,  p-value: 0.00292
```

3. Confidence Interval for β_0 is

$$\begin{aligned} &\hat{\beta_1} \pm t_{n-2,1-\alpha/2} * se(\hat{\beta_1}) \\ &= 0.039 \pm 1.98 * 0.013 \end{aligned}$$

$$(0.014, 0.064)$$

4. The 95% CI for $X_h = 28$ is $(3.061, 3.341)$, with 95% confidence, the true mean of estimator Yh lies between somewhere in this range.
5. The 95% PI for $X_h = 28$ is $(1.959, 4.443)$, with 95% confidence, the true estimator Yh lies between somewhere within this interval.
6. The PI is wider than the CI because CI is interval of the mean of $\hat{Y} = \beta_0 + \beta_1 * X$, while the PI is interval of the actual estimation of $\hat{Y} = \beta_0 + \beta_1 * X + \varepsilon$, the standard error for PI is larger than the CI, so the PI is always larger than the CI.