p8130 homework 5

Jeffrey Liang

11/15/2020

Problem 1

For this problem youwill use dataset 'Antibodies.csv'from homework 1. In the first assignmentyou generated descriptive statisticsand exploratory plotsfor the Ig-M levels by Normal vs Altered groups. The conclusion was that Ig-M distributions were both right skewedwithhigher levels for the Altered group. Given the non-normal distributions, now you are asked touse an alternative, non-parametrictest to assess and comment on the differencein Ig-M levelsbetween the two groups (please ignore unanswered andmissing values).You can use R to perform the calculations, but make sure you statetest that you used,the hypotheses, test statistic, p-value and provide interpretation in the context of the problem.

a.

H0 : The median of IgM level in Normal smell group is the same as Altered smell groups.

H1 : The median of ImG level is different between groups.

if no tied ranks:

$$T = \frac{|T_1 - n_1(n_1+n_2+1)/2| - \frac{1}{2}}{\sqrt{(n_1n_2/12)(n_1+n_2+1)}}$$

if there's tied ranks:

$$T = \frac{|T_1 - n_1(n_1+n_2+1)/2| - \frac{1}{2}}{\sqrt{(n_1n_2/12)[(n_1+n_2+1) - \sum^{tied\ group} t_i(t_i^2-1)/(n_1+n_2)(n_1+n_2+1)]}}$$

Under the normal-approximation, we reject H0 if $T > Z_{1-\alpha/2}$, with $\text{p}-\text{value} = 2 \times [1 - \Phi(T)]$.

rank	Normal	Altered
1	NA	0.048
2	0.048	NA
3	0.050	NA
4	0.051	NA
6	NA	0.052
6	0.052	NA

A peek at the rank's table, we know that there's tied rank, so we used the adjusted method.

$$t1 = 9157$$

$$T = 2.456 > Z_{1-0.975} = 1.96$$

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  antibody %>% pull(Normal) and antibody %>% pull(Altered)
## W = 9157, p-value = 0.01
## alternative hypothesis: true location shift is not equal to 0
```

So at 95% level, we reject the Null and conclude that the median of IgM level in Normal group differs from the Altered group.■

Problem 2

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

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

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

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

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

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

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

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

Leeing 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)

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

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

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

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

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

with ① $= N\bar{Y} - N(\beta_1 \bar{X} - \beta_1 \bar{X} + \bar{Y})$

$= 0$ $\square$

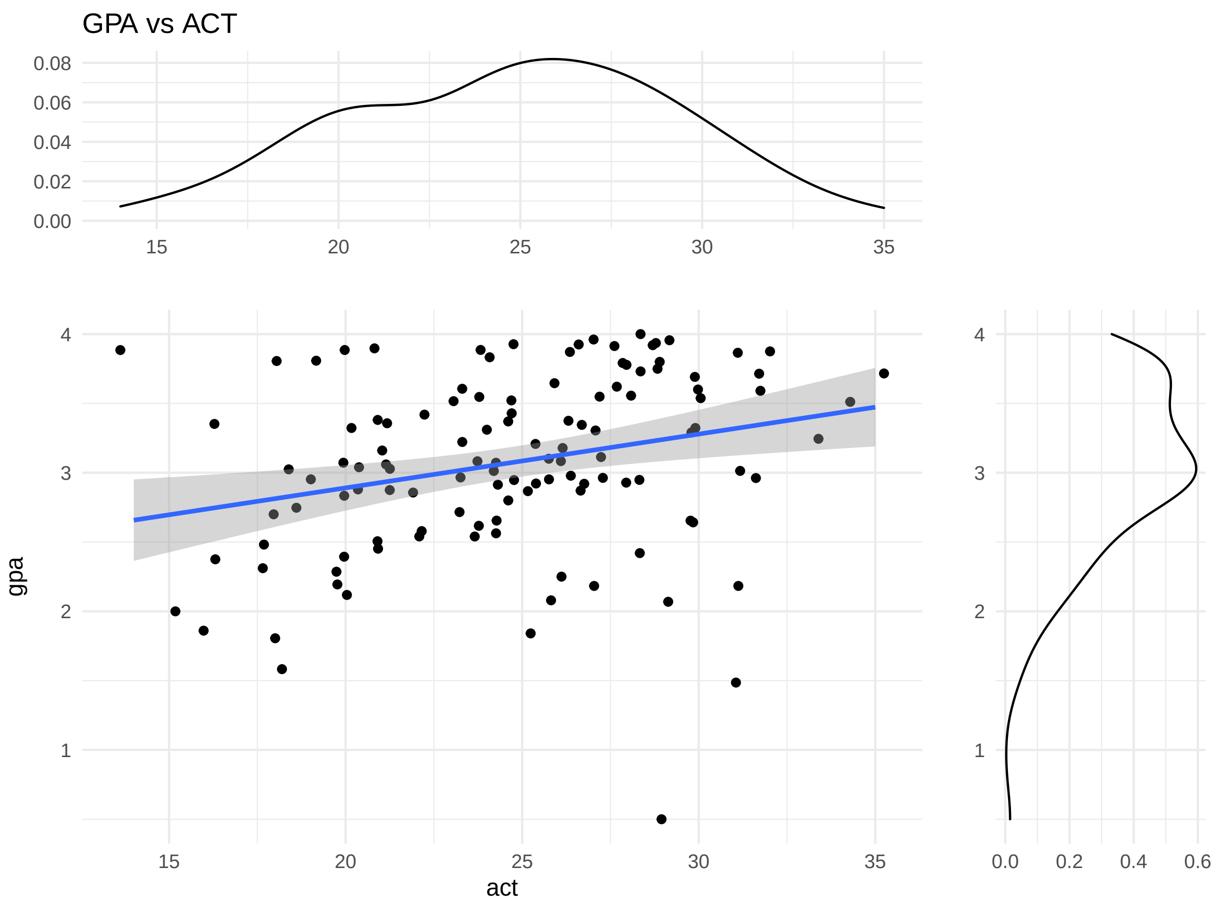
Problem 3

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.




```
## Analysis of Variance Table
```

```
##
## Response: gpa
##           Df Sum Sq Mean Sq F value Pr(>F)
## act         1    3.6    3.59    9.24 0.0029 **
## Residuals 118   45.8    0.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 \ : \ \beta_1 = 0$(there's no association)

$H_1 \ : \ \beta_1 \neq 0$

$error\ sum\ of\ squares(SSE) = \sum_{i=1}^{k}(y_i - \bar{y})^2$

$regression\ sum\ of\ squares(SSR) = \sum_{i=1}^{k}(\hat{y_i} - \bar{y})^2$

$MSR = \frac{\sum_{i=1}^{k}(\hat{y_i}-\bar{y})^2}{k-1}\ (k=2)$

$MSE = \frac{\sum_{i=1}^{k}(y_i-\bar{y})^2}{n-k}$

$F_{statistics} = \frac{MSR}{MSE} \sim F(k-1, n-k)$

$Reject\ H_0\ if\ F > F_{k-1,n-k,1-\alpha}$

$Fail\ reject\ H_0\ if\ F < F_{k-1,n-k,1-\alpha}$

At 95% confidence level, with F = 9.24 and critical value of $F_{0.95,1,n-1}$, we reject the Null hypothesis and conclude that there's linear association between GPA and ACT.

2. $GPA = \hat{\beta}_0 + \hat{\beta}_1 * ACT$

```
##
## 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_1 is

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} * se(\hat{\beta}_1)$$
$$= 0.039 \pm 1.98 * 0.013$$

$$(0.014, 0.064)$$

So at 95% confidence level ,the mean change in GPA per 1 ACT score is somewhere within this interval, and it doesn't contain 0. The AO might take a interest in this association, and include as much as high ACT score candidate for higher average performance in the new enviroment. Or they flag the candidate whom might fall behind and take action to make sure that they catch up.

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.