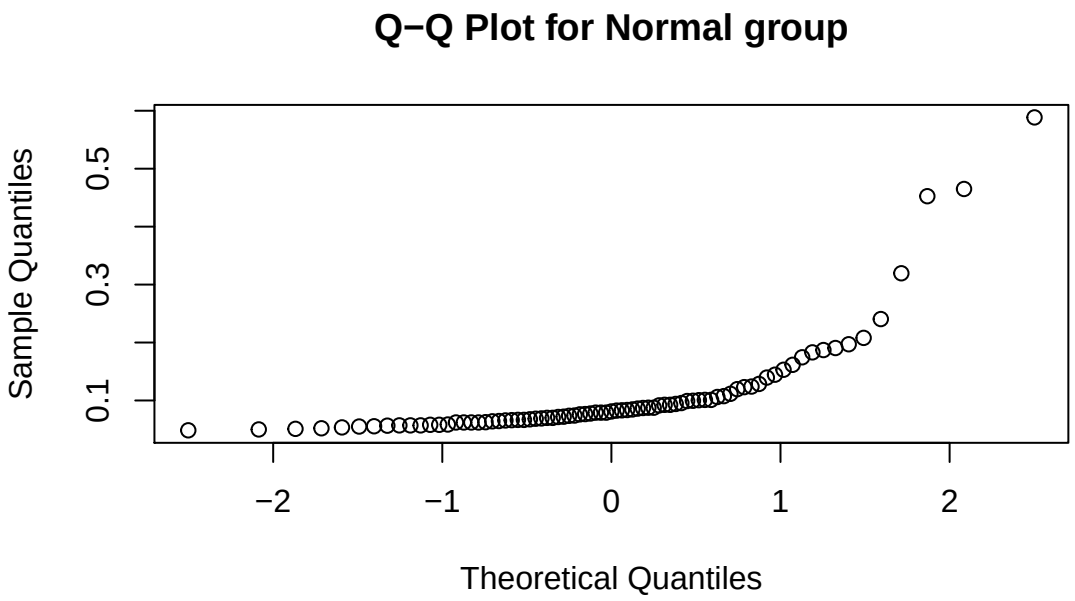
p8130_hw5_rw2844

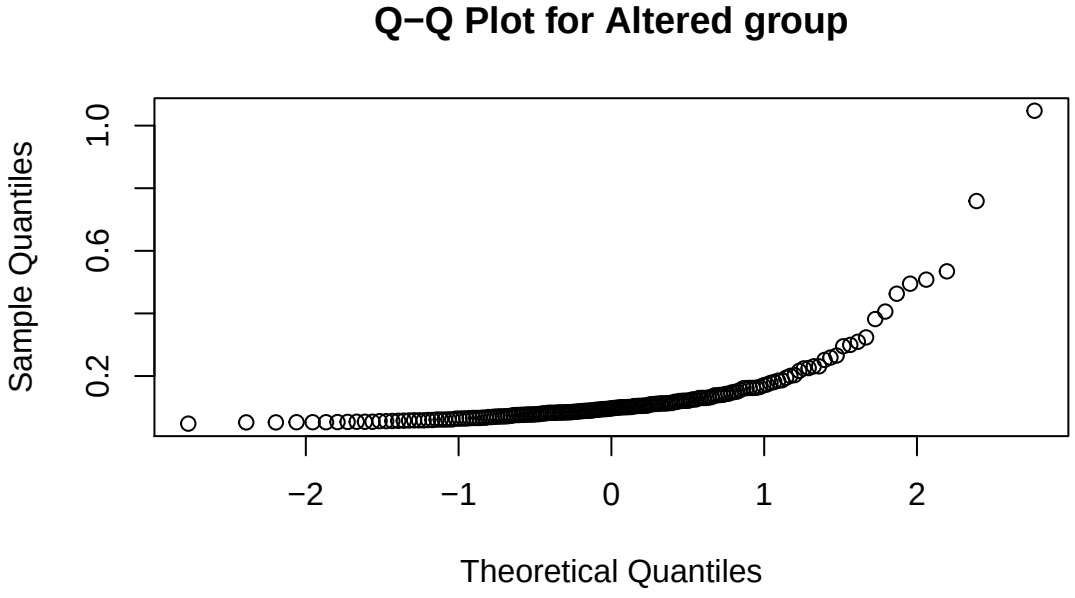
Renjie Wei

11/14/2020

Problem 1

	Altered (N=1047)	Normal (N=410)	Unanswered/Others (N=34)
IgM			
- Mean (SD)	0.130 (0.119)	0.110 (0.092)	0.116 (0.068)
- Median/IQR	0.097 (0.072, 0.137)	0.081 (0.064, 0.107)	0.094 (0.083, 0.119)
- Min - Max	0.048 - 1.048	0.048 - 0.589	0.064 - 0.275
- Missing	869	329	26





From the descriptive statistic and the qq-plots above, we can see that the IgM between two groups are skewed and un-paired. So I decided to use non-parametric method —- Wilcoxon Rank-Sum test.

Let T_1 be the sum of the ranks of the IgM levels of the Normal group, and let T_2 be the levels of the Altered group.

And the hypotheses to be tested are:

H_0 : the medians of the two groups are equal. H_1 : the medians of the two groups are not equal.

And the T_{stat}:

$$T_{stat} = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)\,(n_1+n_2+1)}}$$

And if ties:

$$T_{stat} = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)\left[(n_1+n_2+1) - \sum_{i=1}^{g} t_i\,(t_i^2-1)\,/\,(n_1+n_2)\,(n_1+n_2-1)\right]}}$$

where t_i refers to the number of observations with the same absolute value in the i^{th} group and g is the number of tied groups.

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
6	NA	0.052
9	NA	0.052
9	NA	0.052
9	NA	0.052

As we can see, there are ties, so we need to adjust the T_{stat}. Where $T_1 = 9157$, $g = 153$.

Decision Rules:

Under normal-approximation:n_1 *and* $n_2 \geq 10$, where $n_1 = 81$, and $n_2 = 178$.

Reject H_0 if $T_{stat} > z_{1-\alpha/2}$.

P-value $= 2 \times [1 - \Phi(T_{stat})]$

In our situation, $T_{stat} = 2.456 > z_{0.975} = 1.96$, and p-value is 0.014. We reject the null hypothesis and conclude that the medians of the Normal and Altered smell categories are not equal (using a 0.05 significant level).

Problem 2

a)

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

therefore, $Y_i \sim N(\beta_0 + \beta_1 X_1, \sigma^2)$

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

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

Let $\frac{\partial}{\partial \beta_0} \log L = 0$, $\frac{\partial}{\partial \beta_1} \log L = 0$, $\frac{\partial}{\partial \sigma^2} \log L = 0$

$$\frac{\partial}{\partial \beta_0} \log L = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \Leftrightarrow \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\frac{\partial}{\partial \beta_1} \log L = -\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\Leftrightarrow \quad 0 = \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

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

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

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

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

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

$$\therefore \quad \beta_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$\therefore \quad \beta_0 = \bar{Y} - \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} \bar{X}.$$

b)

$$\sum e_i = \sum (Y_i - \hat{Y}_i)$$

$$= \sum (Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0)$$

$$= n\bar{Y} - (\sum \hat{\beta}_1 X_i + \hat{\beta}_0)$$

$$= n\bar{Y} - (n\hat{\beta}_1 \bar{X} + n\hat{\beta}_0)$$

$\because \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

$\therefore \sum e_i = n\bar{Y} - (n\hat{\beta}_1 \bar{X} + n\hat{\beta}_0)$

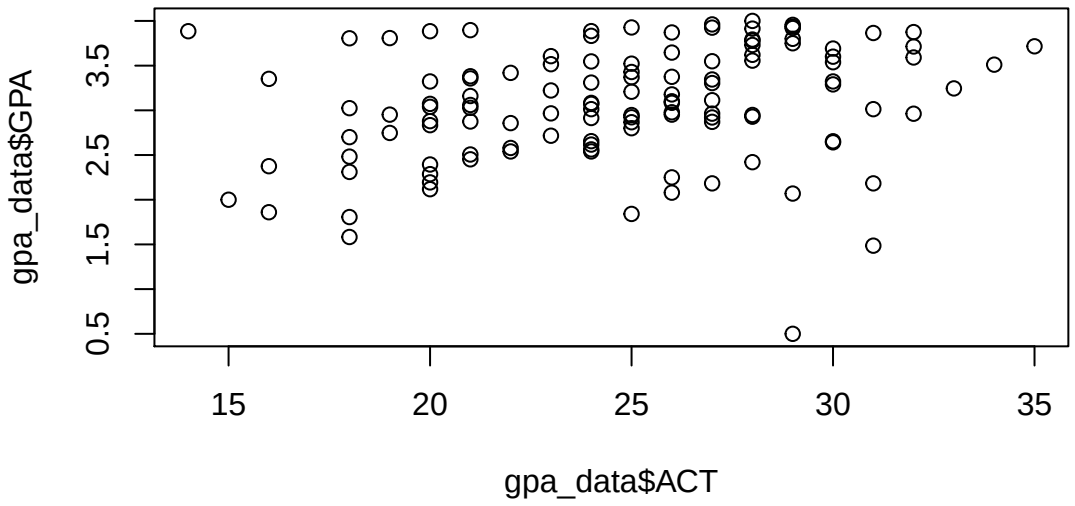
$$= n\bar{Y} - [n\hat{\beta}_1 \bar{X} + n(\bar{Y} - \hat{\beta}_1 \bar{X})]$$

$$= n\bar{Y} - n\bar{Y} = 0.$$ □.

Problem 3

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.



$H_0 \ : \ \beta_1 = 0$

$H_1 \ : \ \beta_1 \neq 0$

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

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

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

$MSE = \frac{\sum_{i=1}^{k}(y_i - \bar{y})^2}{n-k}\ (n = 120)$

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

Decision Rules:

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

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

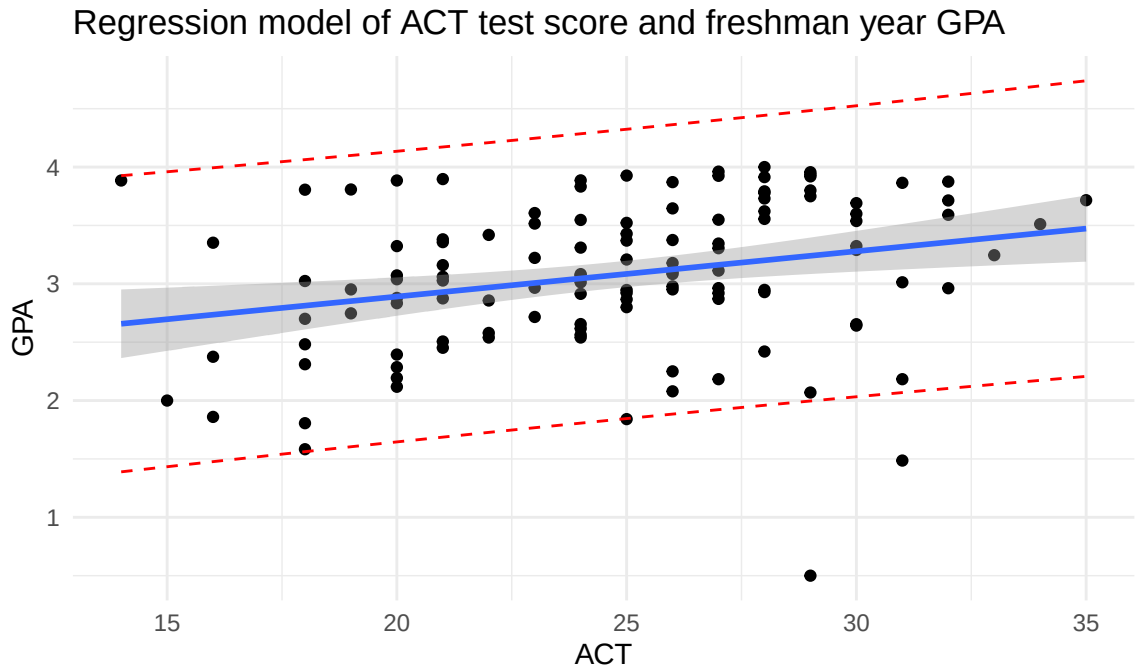
In our situation, $F_{stat} = 9.24 > F_{0.95,1,118} = 3.921$, we reject the Null hypothesis at 0.95 confidence level and conclude that there's linear association between GPA and ACT.

2.

Write the estimated regression line equation:

$$GPA = \hat{\beta}_0 + \hat{\beta}_1 \times ACT$$

$$GPA = 2.114 + 0.039 \times ACT$$



3.

	2.5 %	97.5 %
(Intercept)	1.479	2.750
ACT	0.014	0.064

Confidence Interval for β_1 is :

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

$$(0.014, 0.064)$$

So at 95% confidence level doesn't contain 0. The director of admissions might be interested in whether the confidence interval includes zero this association because that means how much the mean GPA changes when there is one unit change in ACT score. If the 95% CI contains 0, it's hard to say whether the better ACT score brings to better performance in GPA or not. ## 4.

The 95% CI for $ACT = 28$ is $(3.061, 3.341)$. That means the true mean of estimator GPA lies between in this range, with 95% confidence.

5.

The 95% CI for $ACT = 28$ is $(1.959, 4.443)$. That means the true estimator GPA lies between in this range, with 95% confidence.

6.

PI is wider than CI, because CI is the interval of the mean of the estimator while PI is the interval of the estimator itself, the standard error of the mean of the estimator is less than that of the estimator, so we got a narrower interval in CI.