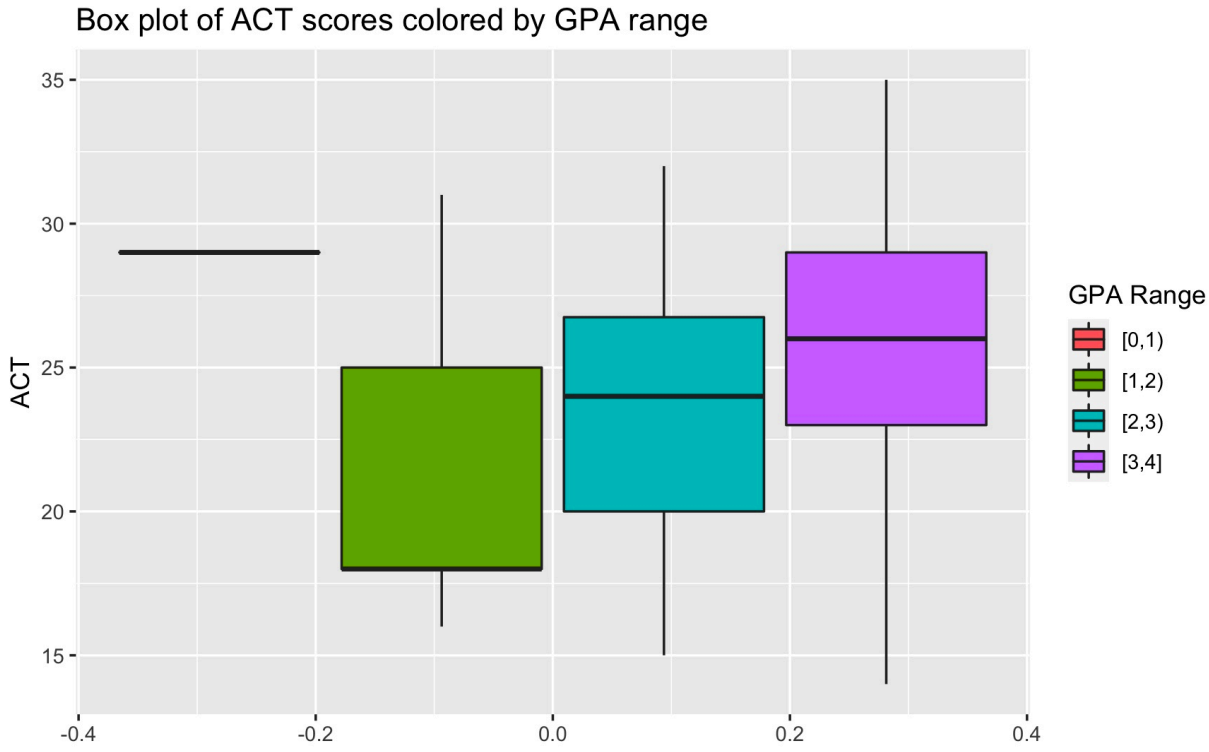
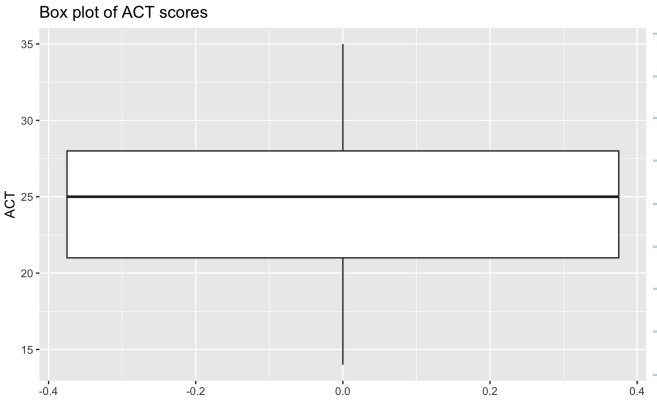
Yuhan Zhu

STT863 Homework 4

3.3. Refer to **Grade point average** Problem 1.19.

a. Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?





Most student get around 24~25 for their ACT score.

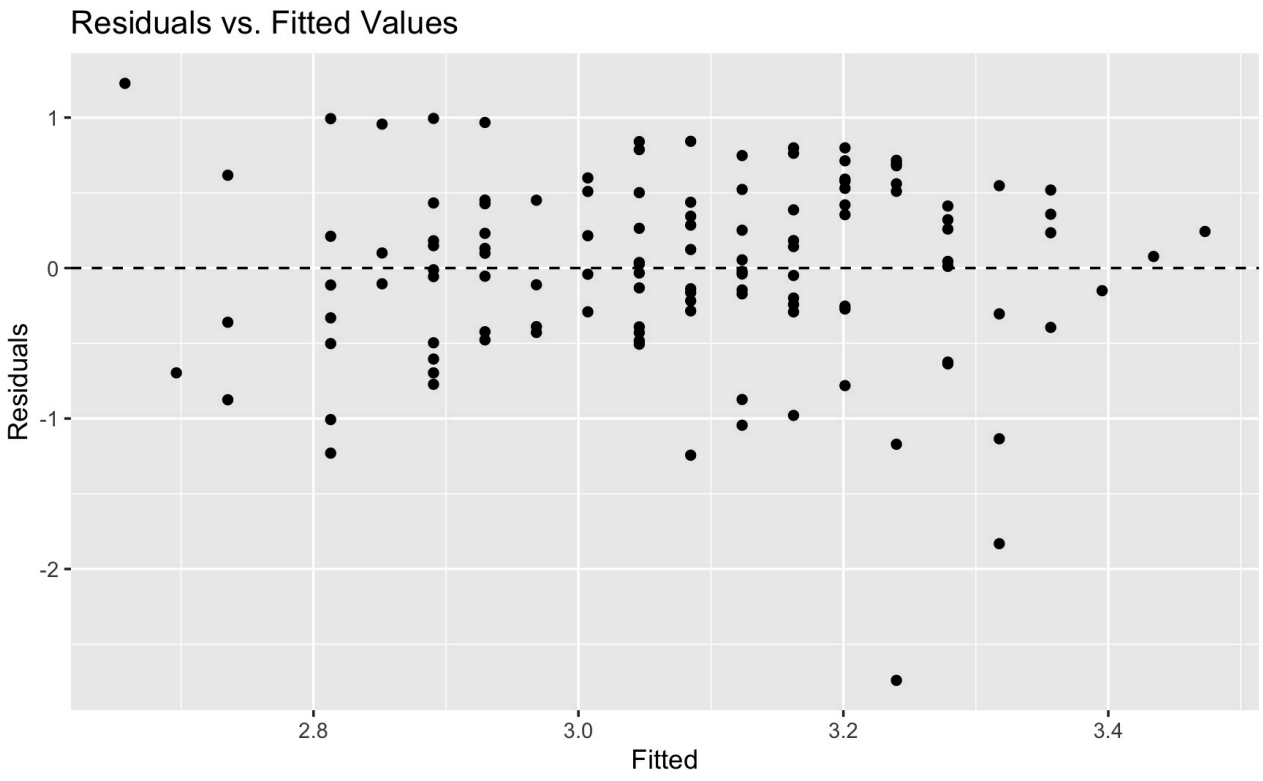
Student can have low GPA but high ACT scores.

Generally, it seems like there are some correlation higher ACT score seems like will likely have a higher GPA.

c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

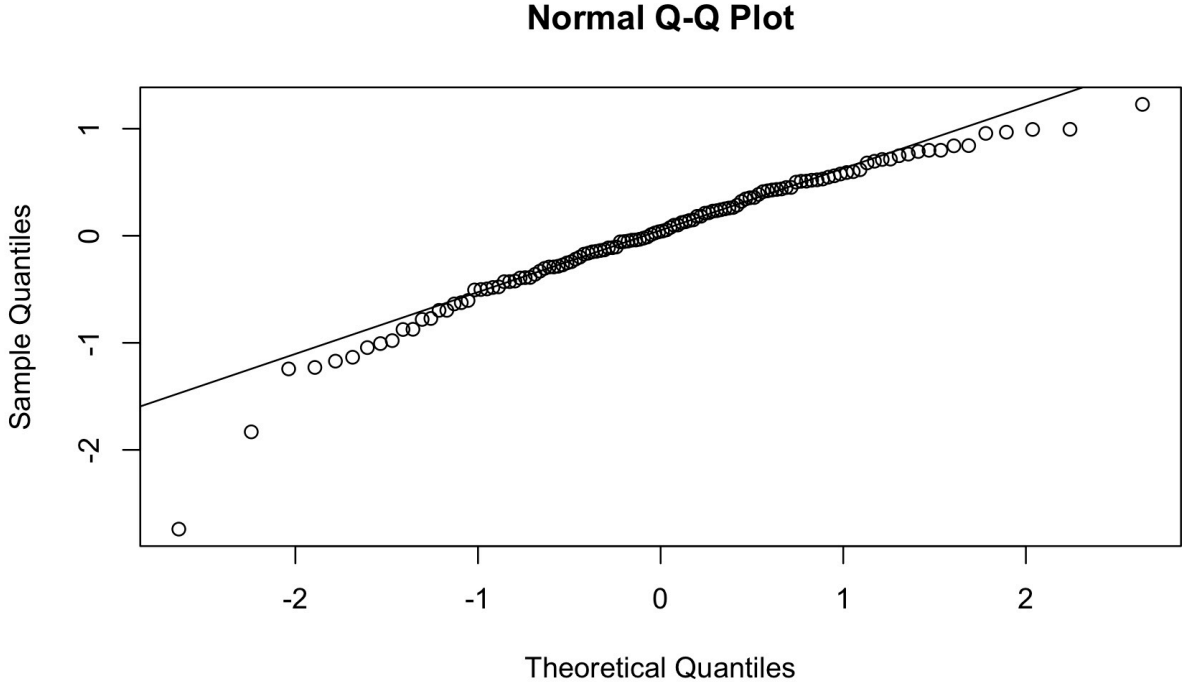
The residual e_i vs fitted values seem to randomly scattered around $e_i = 0$.

There are some outliers



d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

Used 'qqnorm( )' and 'qqline( )' functions



The correlation between ordered residuals and their expected values are $r \approx 0.97336$

From Table B.6 we can see when $n=100$ and $\alpha=0.05$ critical value is 0.987, when $n=40$ $\alpha=0.05$ critical value = 0.972
The residuals didn't fit with the line perfectly. especially the head and tails. Residuals is likely be normal distributed but slight off.

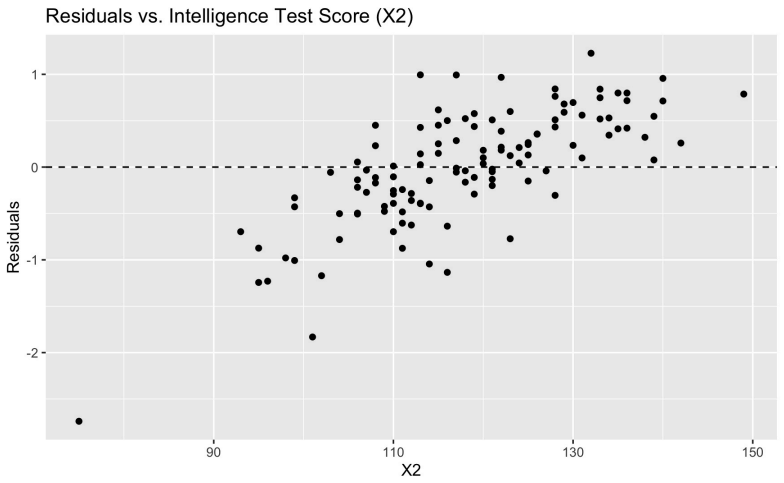
e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

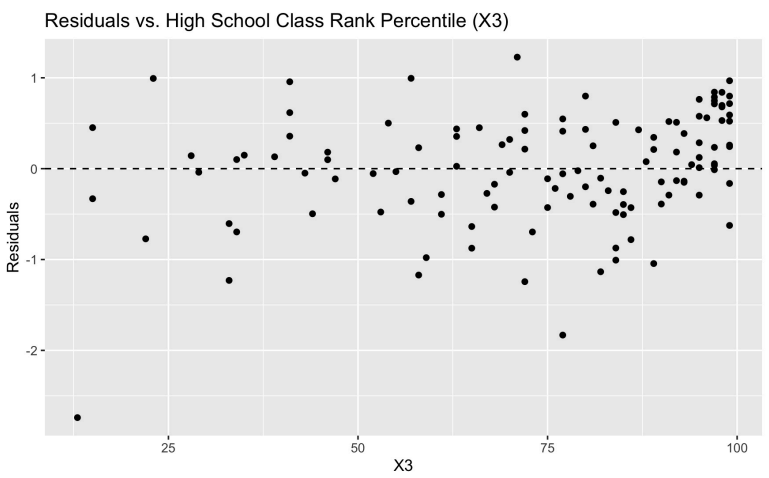
H_0: Variances across two groups $X<26$ and $X\geq 26$ are the same.
H_a $Var(X<26) \neq Var(X\geq 26)$
$\alpha = 0.01$

If P value $< \alpha$, reject H_0
If P value $\geq \alpha$, fail to reject H_0
P-value = 0.3717 $> \alpha$ fail to reject H_0. The variance are pretty constant across all X.
Finding in (e) dose support preliminary findings in part (c).

f. Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i:	1	2	3	...	118	119	120
X_2:	122	132	119	...	140	111	110
X_3:	99	71	75	...	97	65	85





For intelligence score is showing some increasing trend. It may indeciate potential non-linear relationship.

For high school class rank percentile. It indeciate there are linear relationship between GPA and high school rank percentile. It can be another variable we take into consideration in our linear model.