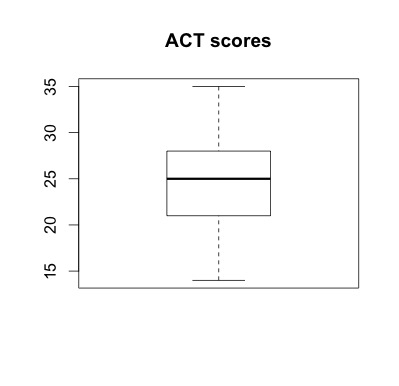
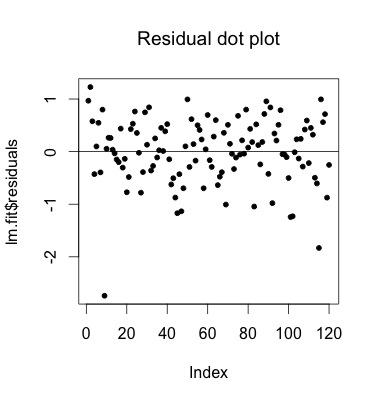
**Question 3**

**a) A boxplot for the ACT scores Xi**

****

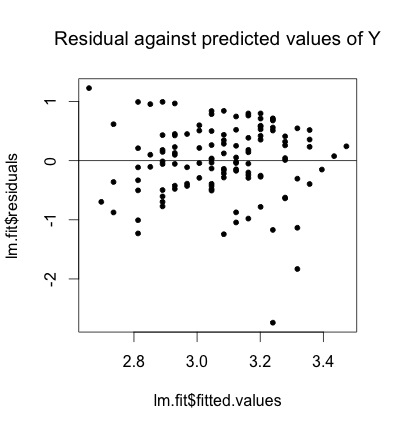
**The distribution seems to be pretty symmetric, and there is no obvious outlier for the ACT scores.**

**b) A dot plot for residuals**

****

**The pattern of the residual seems to be a little asymmetric with respect to the mean residual line. Also, the point at the lower left corner seems to be an outlier.**

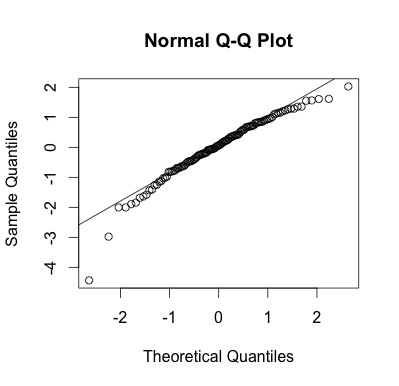
**c) Residual against Y hat**

****

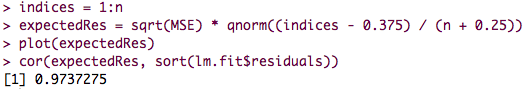
**When plotting residual against fitted value, we will be able to see if there is any relationship between the magnitude of the residual and the magnitude of the fitted value. Ideally, we should not observe any pattern between them.**

**In this plot, the distribution of the residual doesn’t looks very uniform. Also, the point on the lower right might be an outlier.**

**d) Normal probability plot**

****

**The correlation between ordered residual and their expected values under normality is 0.9737275**

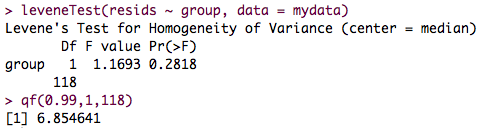
****

**For n = 100 and a = 0.05, the critical value for correlation coefficient between order residual and expected value under normality is 0.979. Since our sample size is n = 120, the critical value is going to be at least 0.979.**

**However, our observed correlation is 0.973 < 0.979, which is going to be smaller than the critical value when n = 120, a = 0.05. Therefore, the correlation between actual residual and expect residual is not significant, which implies that the distribution of the residuals might not be normally distributed.**

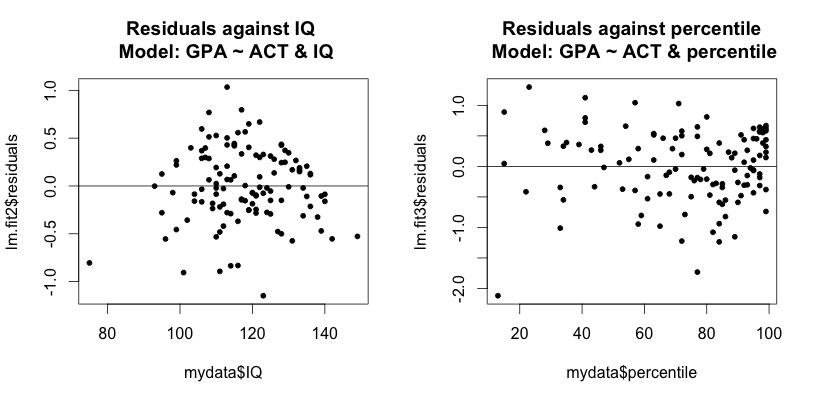
**e) Brown-Forsythe test: whether the error variance varies with the level of X.**

**Divide the X – ACT score by 26.**

****

**The absolute value of the observed F value is 1.1693, which is smaller than F(a = 0.01, 1, 118) = 6.854641. Therefore, we retain the null hypothesis, which indicates the variance is constant. This conclusion supports the observation from c.**

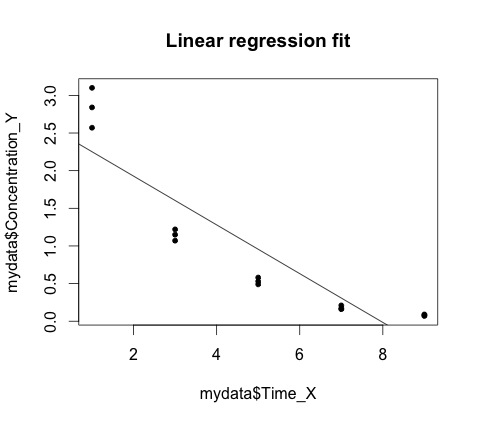
**f)**

****

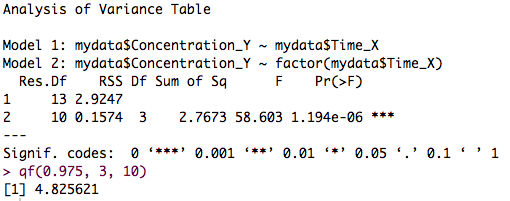
**When adding IQ or percentile rank to the linear model, the pattern of the residuals is not becoming more uniform. So, it might not be very reasonable to adding them as predictors.**

**Question 4**

**a) Fit the simple linear regression model**

****

**b) Perform F test of lack of fit, a = 0.025**

****

**H0: Yij = ß0 + ß1Xj + eij**

**H1: Yij = uij + eij**

**Since Fobs > Fcrit, we reject the null hypothesis, which suggests the relationship is not linear.**

**C) Does the test indicate what regression function is appropriate?**

**The F lack of fit test suggests that the true relation is unlikely to be linear, but it doesn’t provide any information about what this relation might be.**

**From the shape of the data, it seems that the data is curvilinear, and a quadratic regression function might be more appropriate.**

**Question 5**

**Based on the assumption that the intercept term is 0,**

**The full model: Yij = uj + eij**

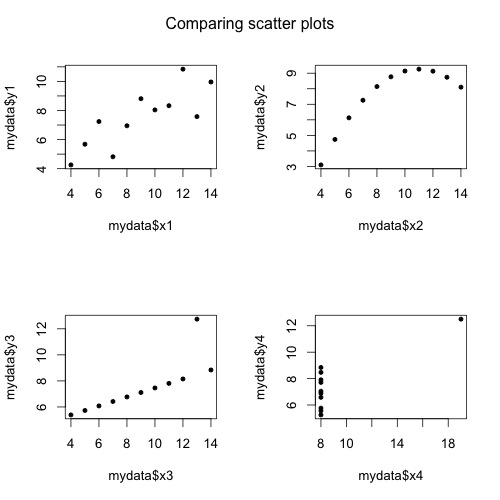
**The reduced model: Yij = ß1Xj + eij**

**The degrees of freedom for SSPE is n – c = 20 – 10 = 10**

**The degrees of freedom for SSE is n – 1 = 20 – 1 = 19 (we lost 1 degree of freedom for estimating beta 1)**

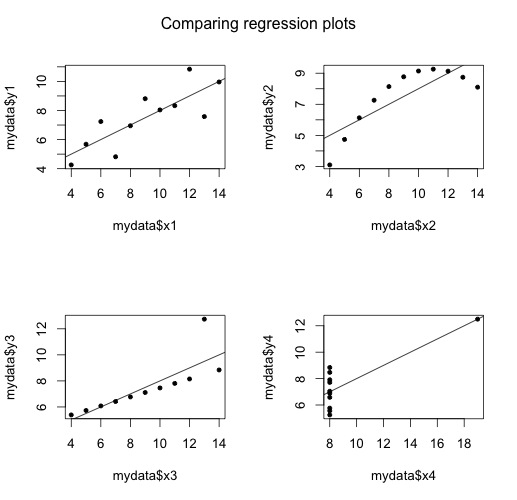
**Question 6**

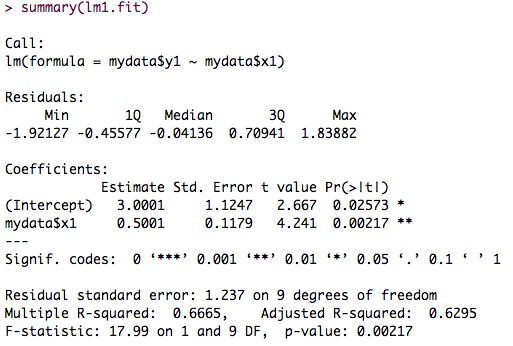
**1) Scatter plot for each data set**

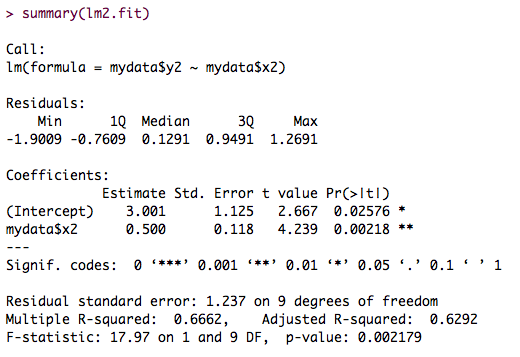


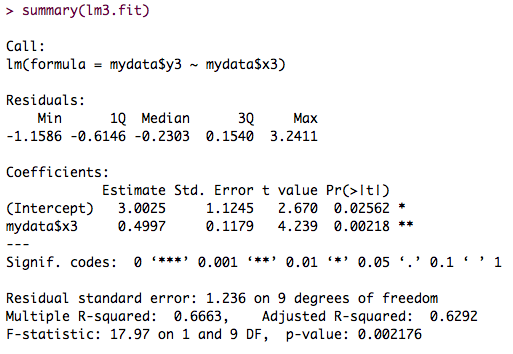
**I think a SLR might fit data1 well. For the third dataset, a SLR will provide a good fit after removing the outlier point. It is not very reasonable to user SLR for other data sets.**

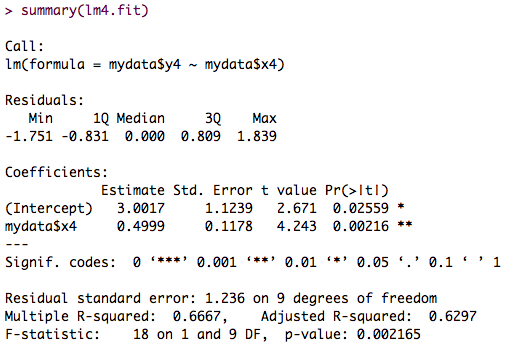
**2) Compare SLR models between these data sets**











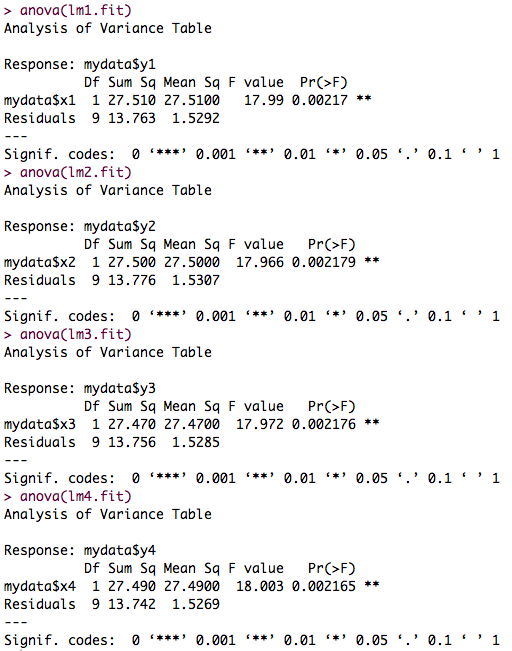
**3) Write out the fitted SLR line, R-square, and MSE**

The regression function for all data set is roughly:

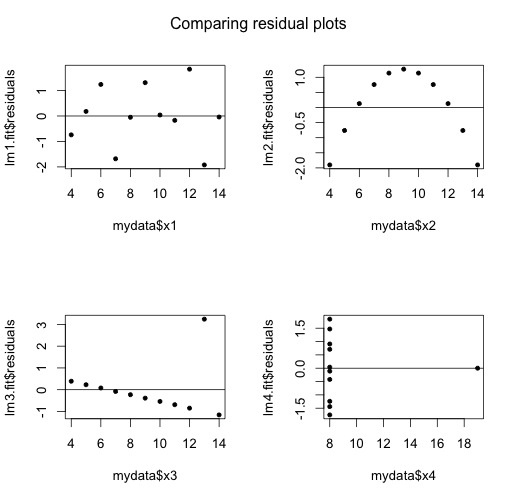
The R2 for all data sets is about: 0.66666

The MSE for all data sets is about: 1.52

**4) The ANOVA tables for all data sets**

****

**5) Residual plots for each data set**



**6) Describe the residual plot you obtained. What features do you notice?**

**Although each data set has completely different pattern, the regression function, ANOVA results, MSE are highly similar. Still, the residual plots indicate that the only reasonable regression line is the first one.**

**The residual plot for the second data shows a cyclical non-independence, which suggests we might need a more complicate regression function to fit the data well.**

**The third and the fourth data set have an outlier, which had a strong influence on the regression fit.**