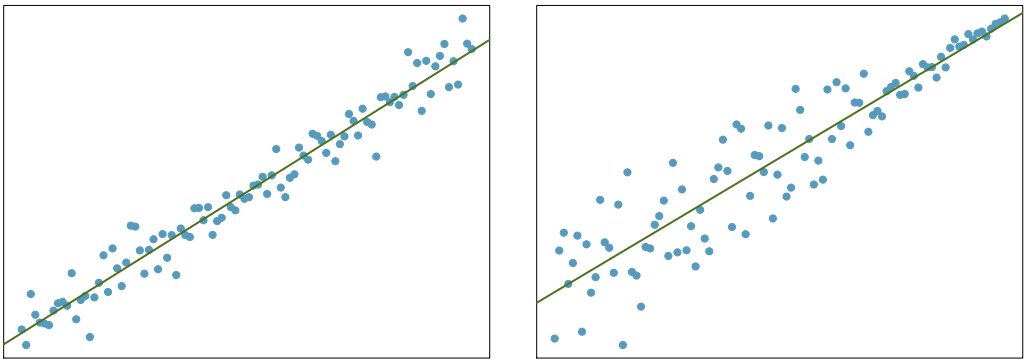
Linear regression

Computational Mathematics and Statistics Camp
University of Chicago
September 2018

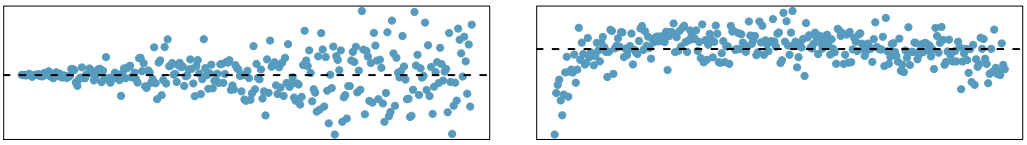
1. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.



(a) (b)

   a. The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant.
   b. The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

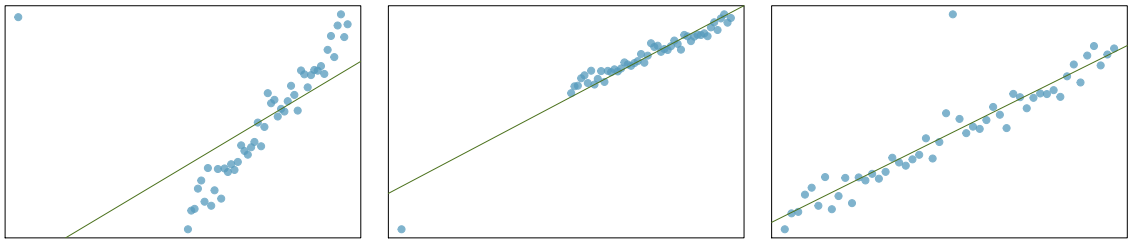
2. Shown below are two plots of residuals remaining after fitting a linear model to two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.



(a) (b)

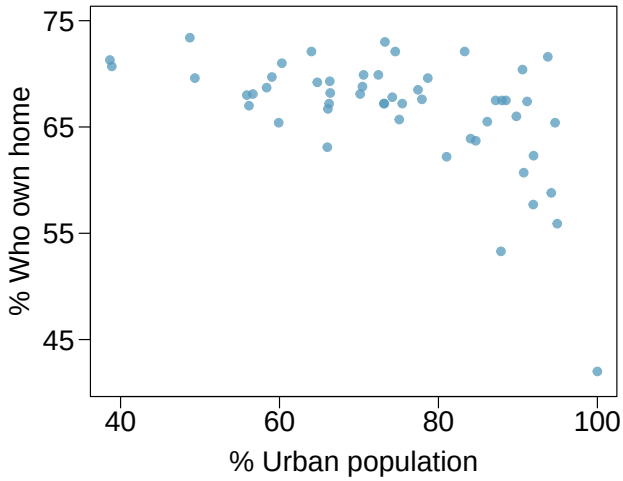
   a. There is a fan shape in the residual plot. Variability around the regression line increases as x increases. Since there is a trend in the residual plot, a linear model would using the methods we have described would not be appropriate for these data.
   b. There is an apparent curvature in the residual plot. A linear model would not be appropriate for these data.

3. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.



a. The outlier is in the upper-left corner. Since it is horizontally far from the center of the data, it is an influential point. Additionally, since the fit of the regression line is greatly influenced by this point, it is a point with high leverage.
b. The outlier is located in the lower-left corner. It is horizontally far from the rest of the data, so it is a high-leverage point. The regression line also would fall relatively far from this point if the fit excluded this point, meaning it the outlier is influential.
c. The outlier is in the upper-middle of the plot. Since it is near the horizontal center of the data, it is not a high-leverage point. This means it also will have little or no influence on the slope of the regression line.

4. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas in 2010. There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.



a. Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas in 2010.

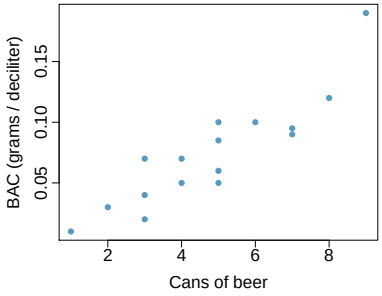
   There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot.

b. The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?

   The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

5. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who

each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatterplot and regression table summarize the findings.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

a. Describe the relationship between the number of cans of beer and BAC.

The relationship appears to be strong, positive and linear. There is one potential outlier, the student who had 9 cans of beer.

b. Write the equation of the regression line. Interpret the slope and intercept in context.

$$\widehat{BAC} = -0.0127 + 0.0180 \times \text{beers}$$

- Slope: For each additional can of beer consumed, the model predicts an additional 0.0180 grams per deciliter BAC.
- Intercept: Students who don't have any beer are expected to have a blood alcohol content of -0.0127.

c. Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.

The hypotheses are as follows:

- H_0: The true slope coefficient of number of beers is zero ($\beta_1 = 0$).
- H_A: The true slope coefficient of number of beers is greater than zero ($\beta_1 > 0$).

The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is approximately 0. (Note that this output doesn't mean the p-value is exactly zero, only that when rounded to four decimal places it is zero.) Therefore the p-value for the one-sided hypothesis will also be very small. With such a small p-value, we reject H_0 and conclude that the data provide convincing evidence that number of cans of beer consumed and blood alcohol content are positively correlated and the true slope parameter is indeed greater than 0.

d. The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

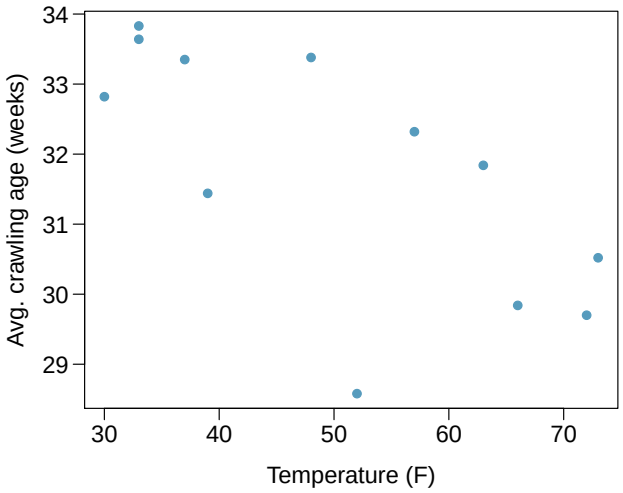
$$R^2 = 0.89^2 = 0.79$$

Approximately 79% of the variability in blood alcohol content can be explained by number of cans of beer consumed.

e. Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?

It would probably be weaker. This study had people of very similar ages, and they also had identical drinks. In bars and elsewhere, drinks vary widely in the amount of alcohol they contain.

6. A previous exercise introduces data on the average monthly temperature during the month babies first try to crawl (about 6 months after birth) and the average first crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53° F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an influential point?



The outlier is located in the bottom center of the plot. The point does not have high leverage since it is near the center of the data on the horizontal axis. Since it is not a point of high leverage, it is also not an influential point.

7. A realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

a. Write the equation of the regression line that includes all of the variables.

$$\widehat{\text{baby weight}} = -80.41 + 0.44 \times \text{gestation} - 3.33 \times \text{parity} - 0.01 \times \text{age} + 1.15 \times \text{height} + 0.05 \times \text{weight} - 8.40 \times \text{smoke}$$

b. Interpret the slopes of `gestation` and `age` in this context.

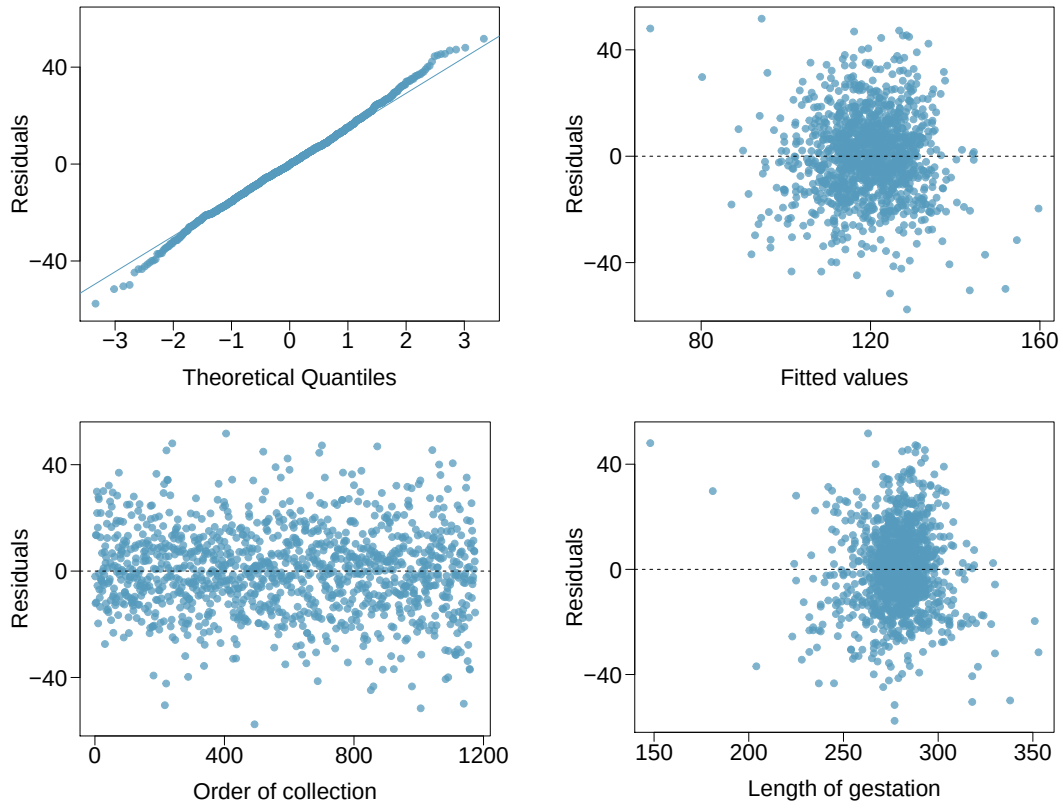
- $\hat{\beta}_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day in length of pregnancy, all else held constant.
- $\hat{\beta}_{age}$: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant.

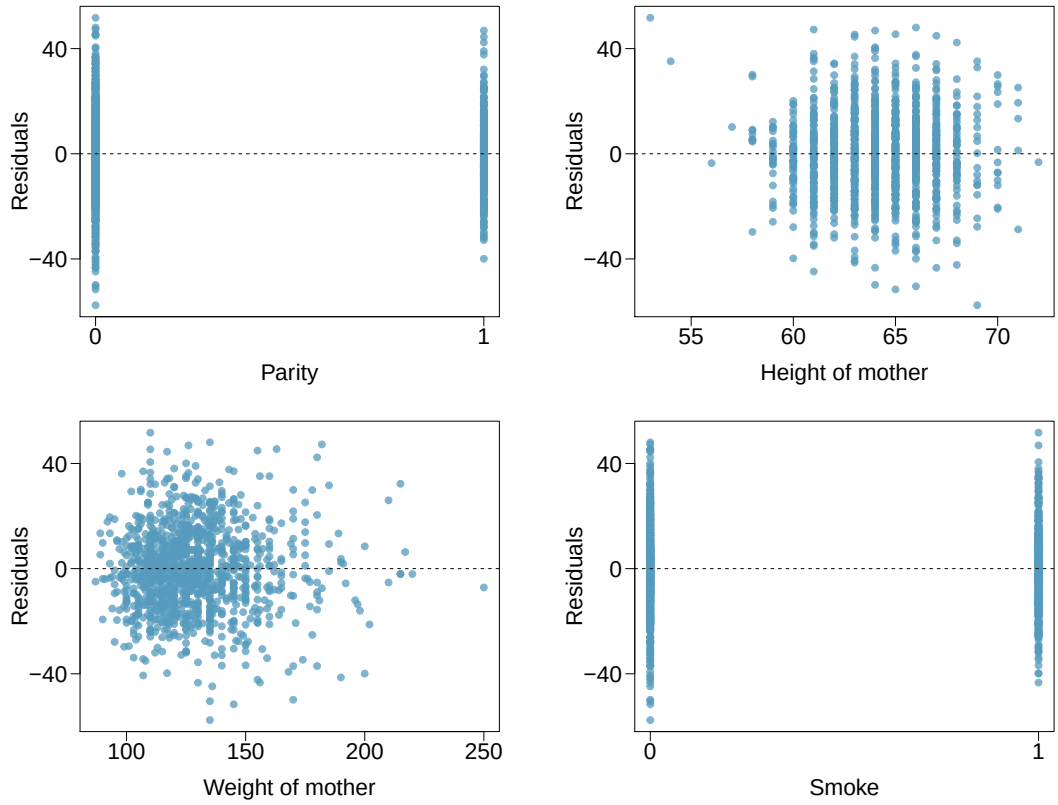
c. The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

The R^2 and the adjusted R^2 can be calculated as follows:

$$R^2 = 1 - \frac{Var(e_i)}{Var(y_i)} = 1 - \frac{249.28}{332.57} = 0.2504$$
$$R^2_{adj} = 1 - \frac{Var(e_i)/(n-p-1)}{Var(y_i)/(n-1)} = 1 - \frac{249.28/(1236-6-1)}{332.57/(1236-1) = 0.2468}$$

8. Determine if the assumptions of the regression model in the previous question are met using the plots below. If not, describe how to proceed with the analysis.





1. Nearly normal residuals: The normal probability plot shows a nearly normal distribution of the residuals, however, there are some minor irregularities at the tails. With a data set so large, these would not be a concern.
2. Constant variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
3. Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
4. Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.

All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

9. A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender. Summary output of the regression model is shown below. Note that male is coded as 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.45	0.35	9.85	0.00
studyweek	0.00	0.00	0.27	0.79
sleepnight	0.01	0.05	0.11	0.91
outnight	0.05	0.05	1.01	0.32
gender	-0.08	0.12	-0.68	0.50

a. Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.

A 95% confidence interval for the slope of gender can be calculated as follows:

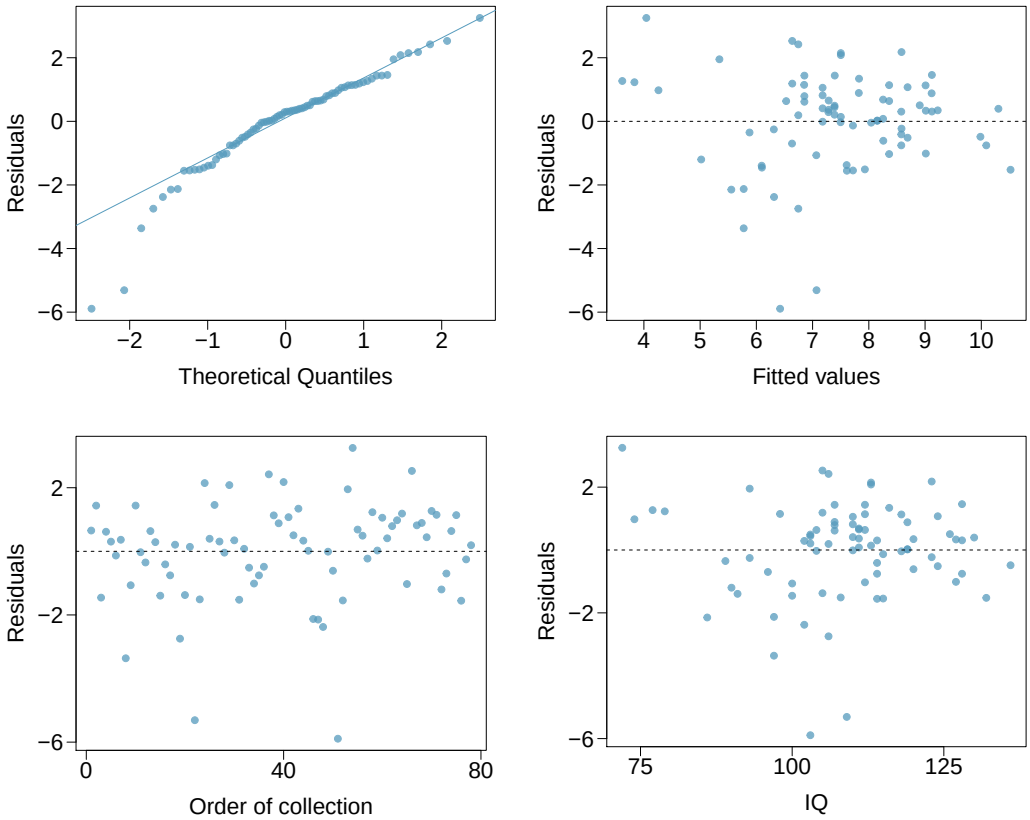
$$\begin{aligned} df &= n - p - 1 = 55 - 4 - 1 = 50 \\ t^*_{50} &= 2.01 \\ \hat{\beta}_{gender} \pm t^*_{df} SE_{gender} &= -0.08 \pm 2.01 \times 0.12 \\ &= -0.08 \pm 0.24 \\ &= (-0.32, 0.16) \end{aligned}$$

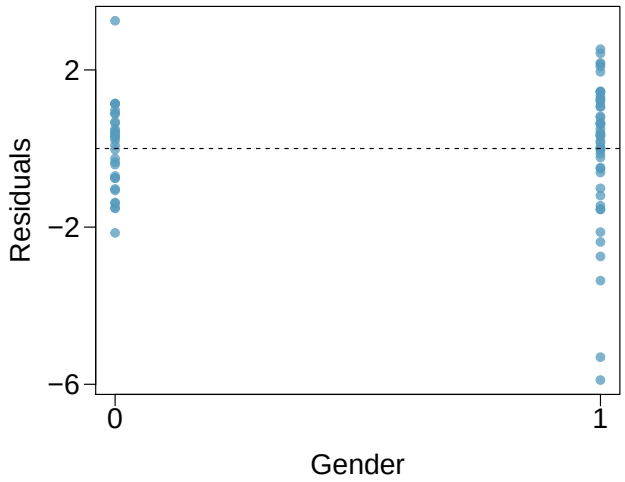
We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model.

b. Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain

Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

10. A regression model for predicting GPA from gender and IQ was fit, and both predictors were found to be statistically significant. Using the plots given below, determine if this regression model is appropriate for these data.





1. Nearly normal residuals: The normal probability plot shows curvature, indicating non-normal residuals.
2. Constant variability of residuals : Couple outliers with high absolute residual values stand out in this plot. In addition, the variability of residuals among the two gender groups is not constant.
3. Independent residuals: While there doesn't seem to be a relationship between the residuals and the order in which the data is collected, this plot reveals couple outliers with large negative residual values.
4. Linear relationships between the response variable and numerical explanatory variables: Residuals vs. IQ and gender plots shows residuals reveals quite a few outliers. There isn't a clear linear relationship between IQ and GPA.

None of the assumptions are met, most likely due to the outliers in the data. Since the model assumptions are not satisfied we cannot rely on this model to conclude a relationship between the explanatory variables and GPA.