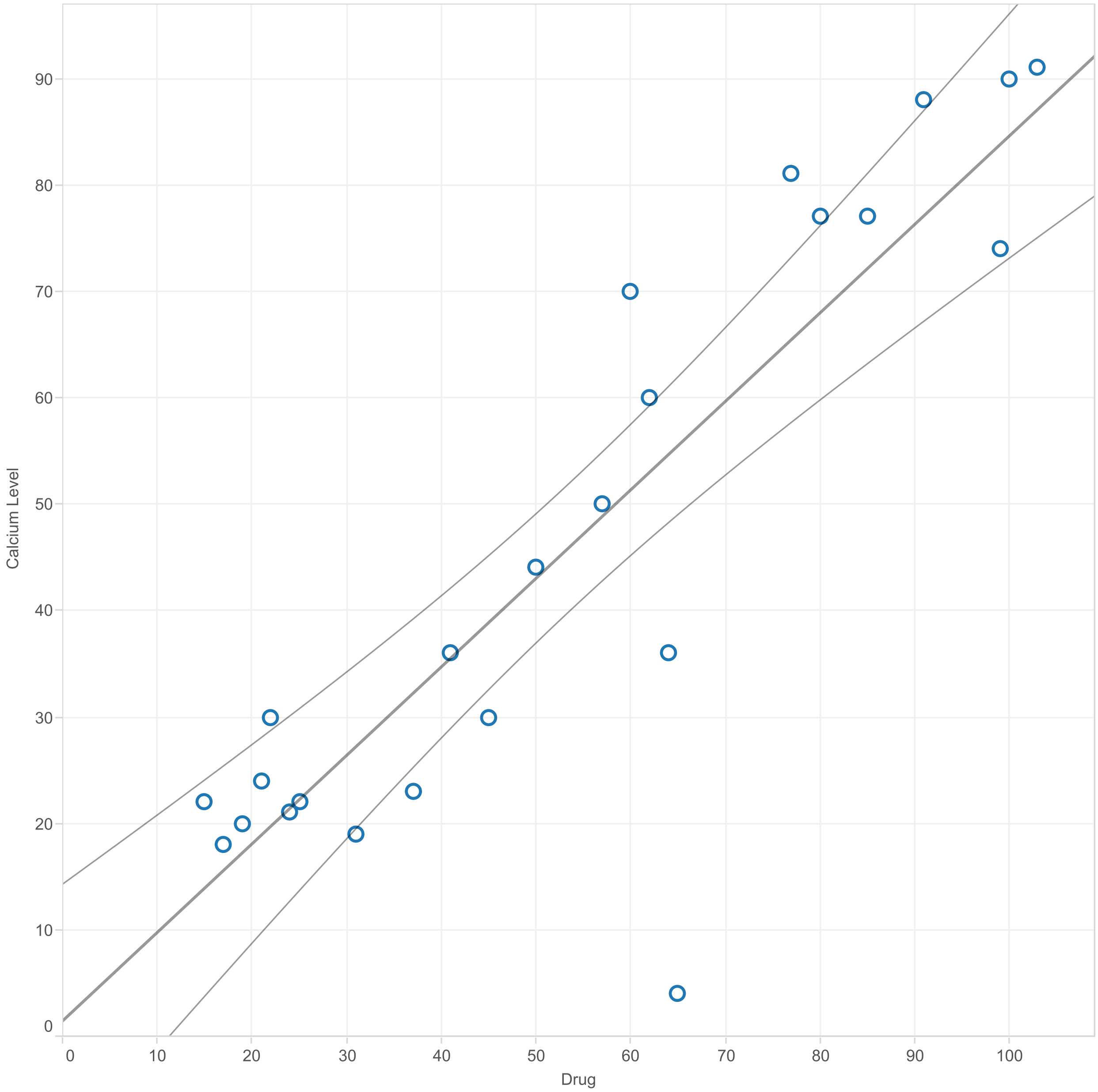
Project 2

Question 1:

Sheet 1



Drug vs. Calcium Level.

Let’s decide if taking more drug will affect Calcium level based upon a specific set of data. Given the data, I prepared a scatter plot (with trend lines and statistics) where we have Drug vs. Calcium Level. Drug is the explanatory variable and Calcium Level is the response variable

The data displayed on the graph are clustered as to resemble a line rising from left to right. Calculating the correlation coefficient and we get r = + 0.8669. Since the slope of the line is positive and r = + 0.8669 > + 0.7, there is a strong linear positive correlation between the two sets of data. This means that according to this set of data, the more drug I take, the higher Calcium level I will get. Note that just this set of data showed a positive correlation does not mean that the relationship is positive for **all** sets of data concerning drug and Calcium level. We should also check outlier condition: a few Calcium levels seem to straggle away from the main pattern. After checking on them by taking them off, the relationship still remains a strong linear positive correlation, so they are not extreme enough to be called outliers.

Question 2:

PLAN A: LINEAR MODEL

ANOVA and Summary:

Trend Lines Model

A linear trend model is computed for Calcium Level given Drug. The model may be significant at p <= 0.05.

Number of modeled observations:	24
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	22
SSE (sum squared error):	4445.83
MSE (mean squared error):	202.083
R-Squared:	0.751666 (This is correlation coefficient r squared)
Standard error:	14.2156
p-value (significance):	< 0.0001

Based on primitive analysis, the correlation says, “The linear association between these two variables is fairly strong”, but it doesn’t tell us what the line is. But why I chose linear model in the first place? It is intuitive to start with simple model. Based the previous calculation, r-squared is + 0.8669 which is very high and entails a linear model fits the data well.

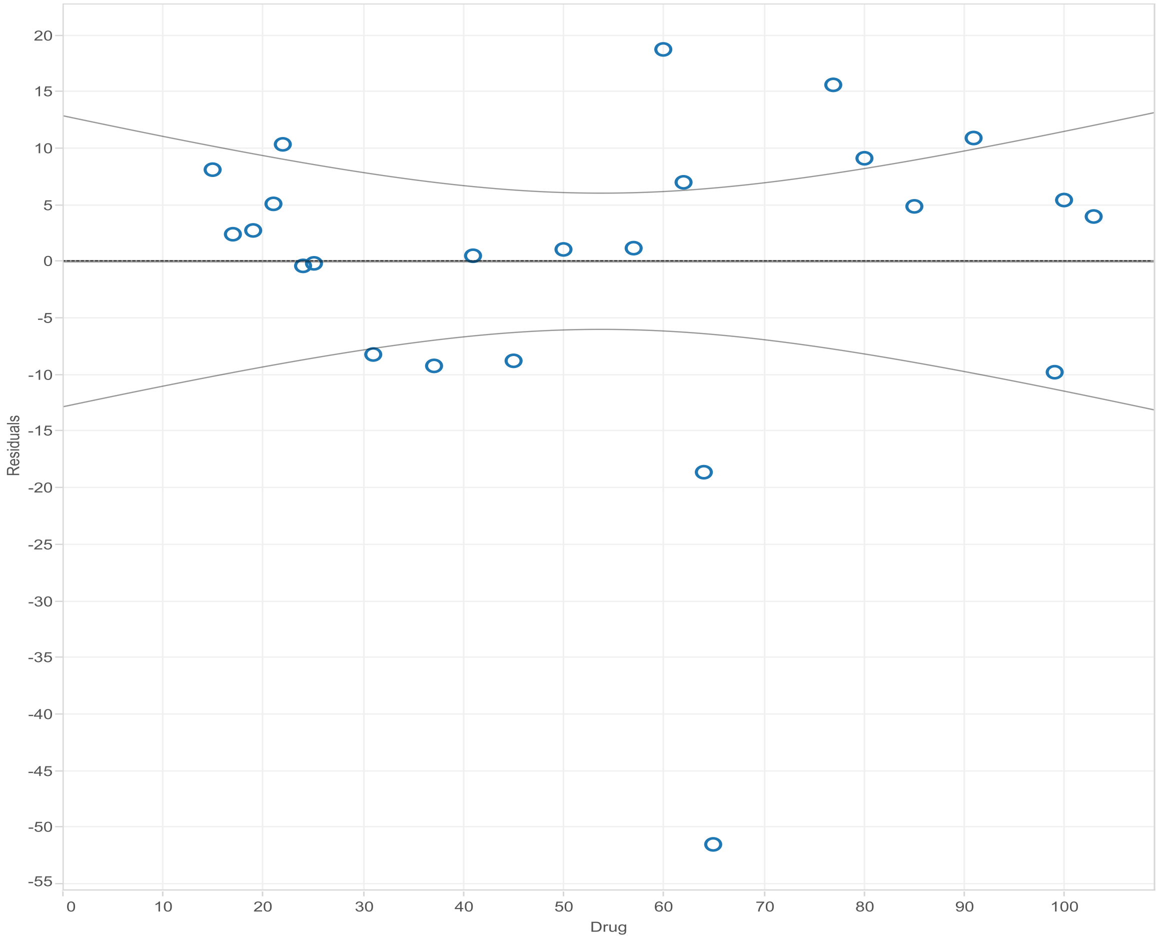
To calculate the line, I will first choose a linear model written as follows: ^ Calcium Level = b_0 + b_1 * Drug. We write b_0 and b_1 for the intercept and slope of the line who are called the coefficients of the the linear model. B_1 is the **slope**, which tells how rapidly ^Calcium Level changes with respect to Drug. B_0 is the **intercept**, which tells where the line hits the y-axis.

Using the formulas for linear regression calculation b_1 = r * (S_Calcium Level/ S_Drug) and b_0 = Calcium Level_bar – b_1 * Drug_bar, I get relevant statistics as follows:

b_0 = 1.42633
b_1 = 0.831603
P-value: < 0.0001
Equation: **Calcium Level = 0.831603*Drug + 1.42633**

To check the fitness of this linear model, we show scatterplot of the residuals versus drug:

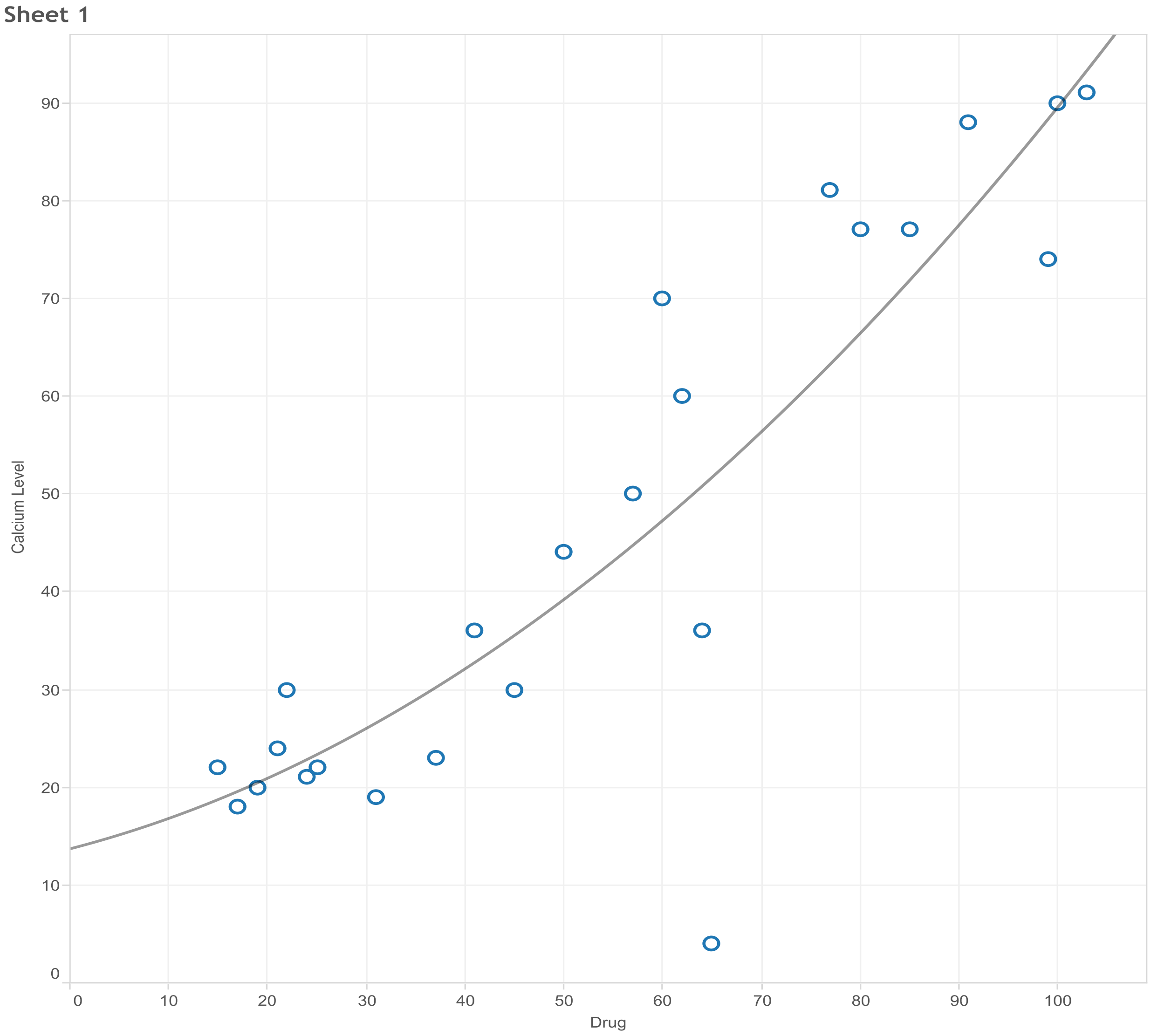
Sheet 1



Drug vs. Residuals . Details are shown for Residuals .

Unfortunately, there seems some pattern and shape from the scatterplot of residuals, which is not what a good model will give. So I proceed to PLAN B: Non linear model, in the hope of a better model for the data. Let’s explore!

PLAN B: POLYNOMIAL DEGREE 2



Drug vs. Calcium Level. Details are shown for Calcium Level.

Compared to the linear model, it seems this non linear model with model equation Calcium Level = 0.0049713*Drug^2 + 0.261175*Drug + 13.6939 fits the data better. Really?

Let's look at the summary

ANOVA and Summary:

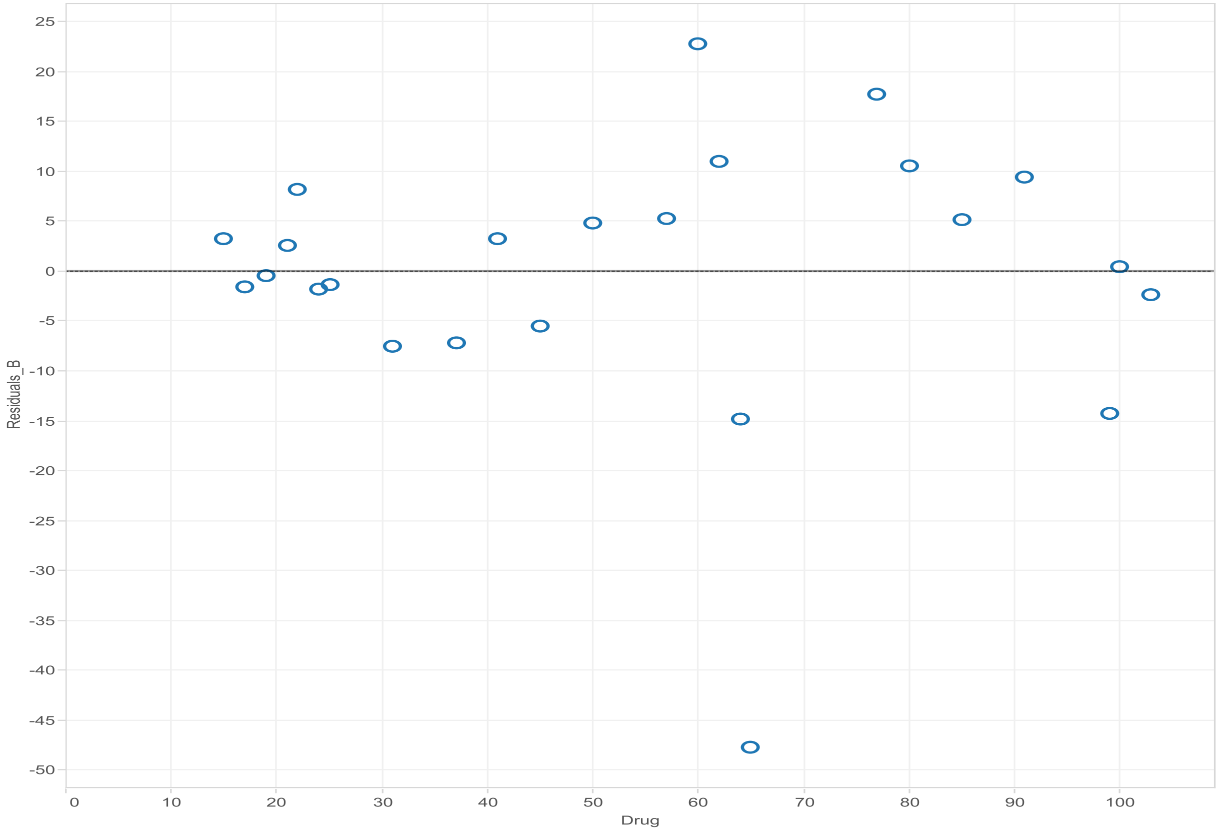
Trend Lines Model

A polynomial trend model of degree 2 is computed for Calcium Level given Drug. The model may be significant at p <= 0.05.

Model formula:	(Drug^2 + Drug + intercept)
Number of modeled observations:	24
Number of filtered observations:	0
Model degrees of freedom:	3
Residual degrees of freedom (DF):	21
SSE (sum squared error):	4171.95
MSE (mean squared error):	198.664
R-Squared:	0.766965
Standard error:	14.0948
p-value (significance):	< 0.0001

Here the polynomial model's R-Squared is 0.766965 which is larger than the R-Squared with linear model 0.751666. And if we present the scatterplot the residuals:

Sheet 1

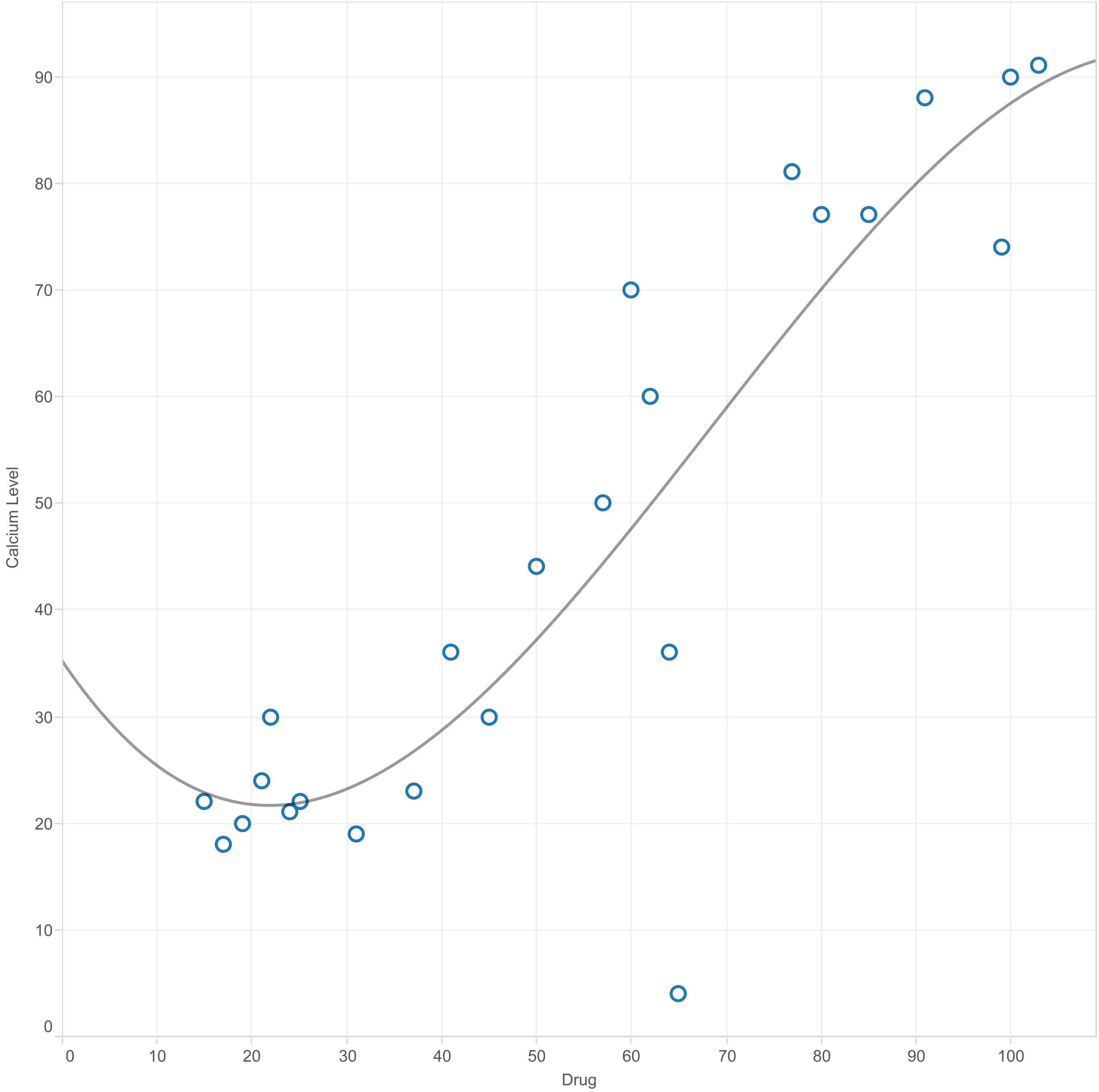


Drug vs. Residuals_B. Details are shown for Residuals_B.

It seems polynomial model has less shape and pattern than the linear model's scatterplot of residuals. So for now Plan B is better than Plan A.

PLAN C: POLYNOIAL DEGREE 3

Sheet 1



Drug vs. Calcium Level. Details are shown for Calcium Level.

I tried to fit the data with polynomial model with degree of 3. Let's look at the summary

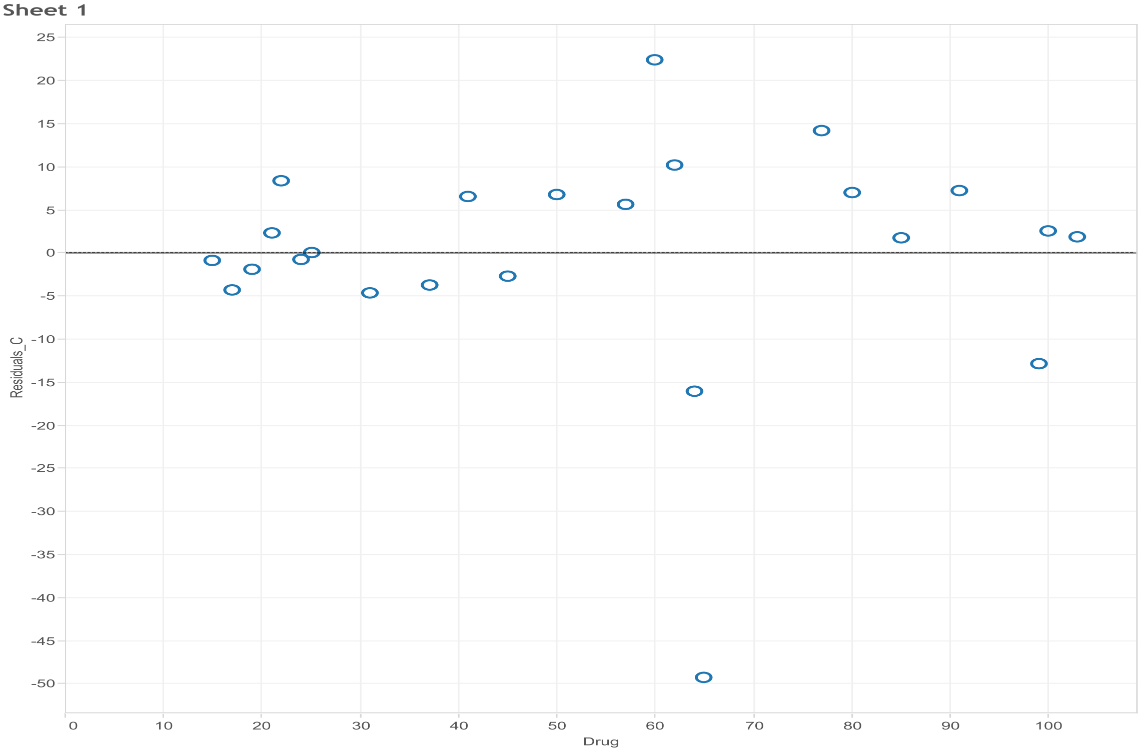
Trend Lines Model

A polynomial trend model of degree 3 is computed for Calcium Level given Drug. The model may be significant at p <= 0.05.

Model formula:	(Drug^3 + Drug^2 + Drug + intercept)
Number of modeled observations:	24
Number of filtered observations:	0
Model degrees of freedom:	4
Residual degrees of freedom (DF):	20
SSE (sum squared error):	4028.13
MSE (mean squared error):	201.407
R-Squared:	0.774998
Standard error:	14.1918
p-value (significance):	< 0.0001

This model Calcium Level = -0.000176086*Drug^3 + 0.036069*Drug^2 + -1.32296*Drug + 35.1837 has R-Squared 0.774998 which is slightly higher than polynomial model with degree of 2.

Let’s look at residual plot:



Drug vs. Residuals_C. Details are shown for Residuals_C.

Indeed, there seems to be less pattern and shape compared to the previous two. But is this the best model for real world application? If we look closely we find that when no drug is taken, the Calcium level is around 35, way more than upper bound of normal range.

CONCLUSION

By exploring linear and polynomial models with degree both 2 and 3, I conclude polynomial model with degree 2 and R-Squared **0.766965, Calcium Level = 0.0049713*Drug^2 + 0.261175*Drug + 13.6939** works better than linear model with R-Squared 0.751666, **Calcium Level = 0.831603*Drug + 1.42633** and polynomial model with degree 3 with R-Squared 0.774998, Calcium Level = -0.000176086*Drug^3 + 0.036069*Drug^2 + -1.32296*Drug + 35.1837. This conclusion is based on the comparisons of scatterplot of residuals and also considerations on real world situations. With the given data and considerations for real world application, there seems to be no better model than the polynomial one with degree of 2.

The model I choose is a polynomial one with degree of two **Calcium Level = 0.0049713*Drug^2 + 0.261175*Drug + 13.6939.** If we look closely at it, we know that when no drug taken (Drug = 0), the Calcium Level is 13.6939 which should represent the normal and natural Calcium level in human body. After doing some research, I find that the normal Calcium level is (range < 10.5), and Classical primary hyperparathyroidism is diagnosed when the calcium level is above the normal range (calcium >**10.5 mg/dL).** This analysis shows that in real world, my polynomial model with those specific coefficients is not very useful as expected. However, 13.6939 is only approximately 3.1 higher than the upper bound of normal range 10.5, compared to 35 which is 20 more higher than the upper bound 10.5, polynomial model with degree 2 is the most acceptable one. If given data size is bigger, I aim to calculate a model that is polynomial with higher degrees and on the other hand has reasonable coefficients that comply with the normal range of Calcium level when Calcium level is 0.