Regression and Anova

Take-home Project

Mohamed Salem

Abstract

This paper attempts to identify the relationship between a biomarker value of colorectal cancer and a certain gene expression value. We have been provided with a dataset that includes 189 observations; however, the data does not seem to follow any discernible linear pattern. Therefore, we attempt to apply a number of useful statistical tools to address any issues and accomodations that need be made such that we can carry out proper inference on our data, identify the underlying relationship, if one exists, and make predictive claims about the value of one variable given the other. For those purposes, we identify two models that can be applied in this situation, a log transformed model, and a polynomial model. We take a look at the assumptions underlying the models, whereby assumptions are examined sequentially and are often followed by statistical tests to verify their non-violation in the presented case study. We find that, for the given case, some assumptions of our models are violated. We implement a number of countermeasures to remedy those violations, and in the end we are able, with some degree of caution, to make inferential claims about the relationship between the biomarker value of colorectal cancer and a certain gene expression value.

Introduction:

The relationship between a biomarker value of colorectal cancer and a certain gene expression value is not obviously apparent. The variables present a number of difficulties that make identifying a relationship a non-straightforward process. Nonetheless, establishing such a relationship would be greatly valuable in terms of understanding the disease and predicting its incidence. Throughout this work, we will examine our dataset by constructing descriptive visualisations of its structure and providing the acccompannying deductions. After establishing a satisfactory level of familiarity with our data, we attempt to find some suitable functional form to represent the relationship. The assumptions underlying the fitting methodology are examined and where violations are met, countermeasures will be taken to address them. We will present a number of tests, while providing the mathematical expressions of test statistics and their associated hypotheses for convenience to the reader. Mathematical formulations for the two models utilized in this paper will also be provided, along with any modifications made to the data or within the model structure. This work begins by attempting to identify a suitable model for predicting Gene Expression from Biomarker value. After that is accomplished, we choose to then use a different model for the prediction of the Biomarker value from Gene Expression; while we can use the same functional or form assumptions for both cases, we choose to present two models to allow diversity and create multiple options. The work here is henceforth presented in three sections, one for the log transformed model, followed by a section for the (inverse) polynomial model, and finally a third section where we present our results and conclude.

Model and Methodology:

Log Transformed Model

We begin our work by furnishing descriptive visualisations of our dataset. We can clearly observe from our scatterplot that there seems to be no direct linear relationship between our variables. From our boxplots and histograms, we can also conclude that the data for both the biomarker and gene expression is most likely not normally distrubuted. We observe two outliers in the biomarker data, with one of those two being so far away from the center of the data that we suspect it may be a measurement error. We observe multiple outliers in the Gene Expression data, however the outliers are numerous and within range of each other, which makes us suspect measurement error less than we do for biomarker data, however, we will investigate this throughout the paper. Since the scatterplot does not visually point towards the existence of a linear

relationship between gene expression and biomarker value, then if we are still interested in using a linear model, we will most likely have to carry out some transformation of our data. We begin by going back to the scatterplot and fitting a simple linear regression line, after which, we will compute Cook's distance for our observations to identify any influential points that may be removed.

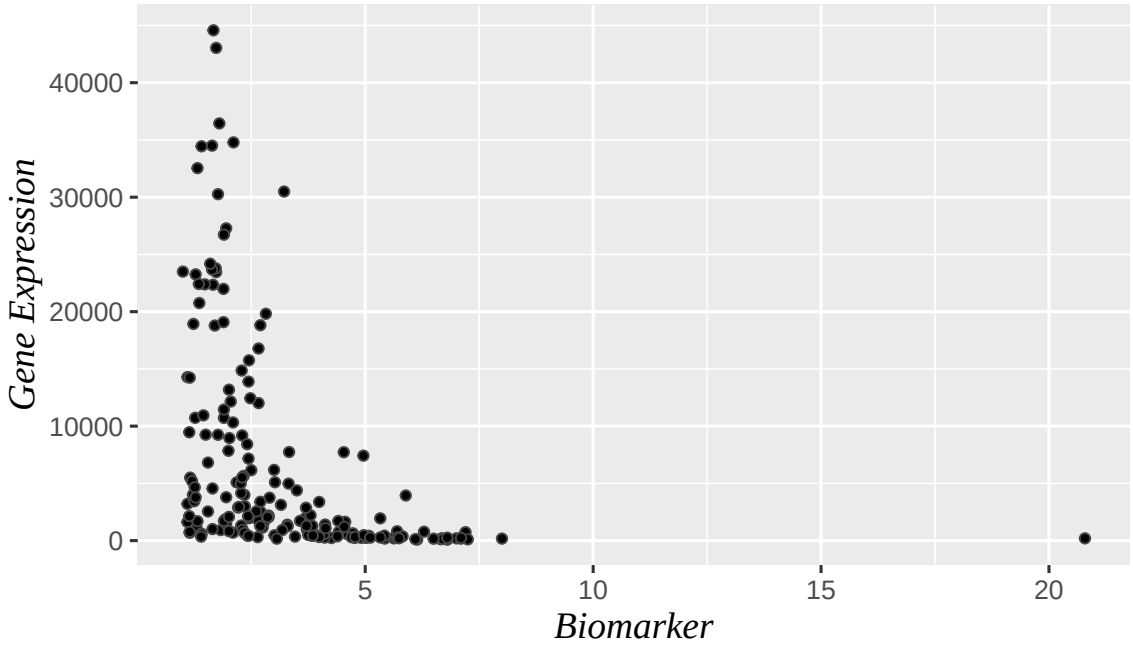

Figure 1: Scatterplot of Biomarker vs. Gene Expression

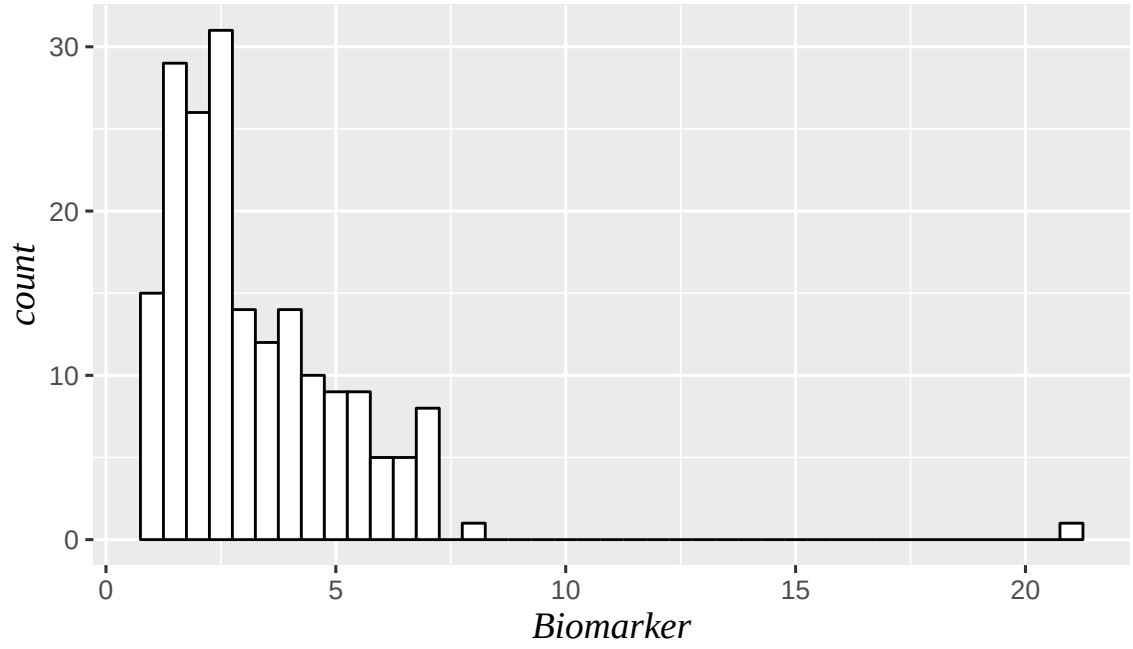

Figure 2: Histogram of Biomarker

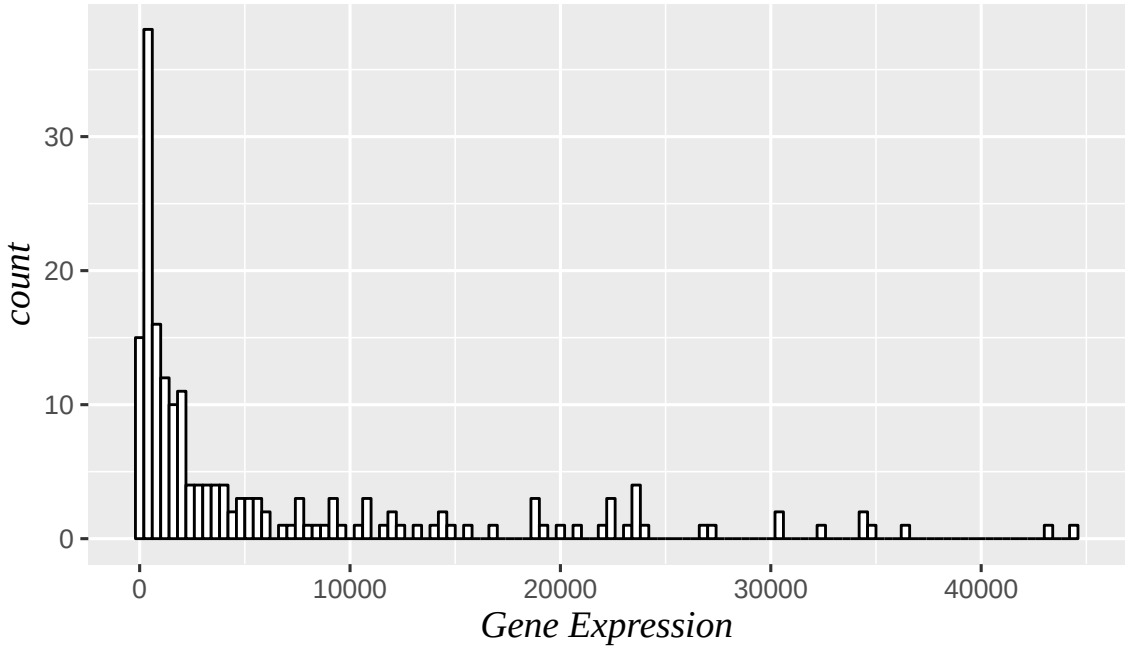

Figure 3: Histogram of Gene Expression

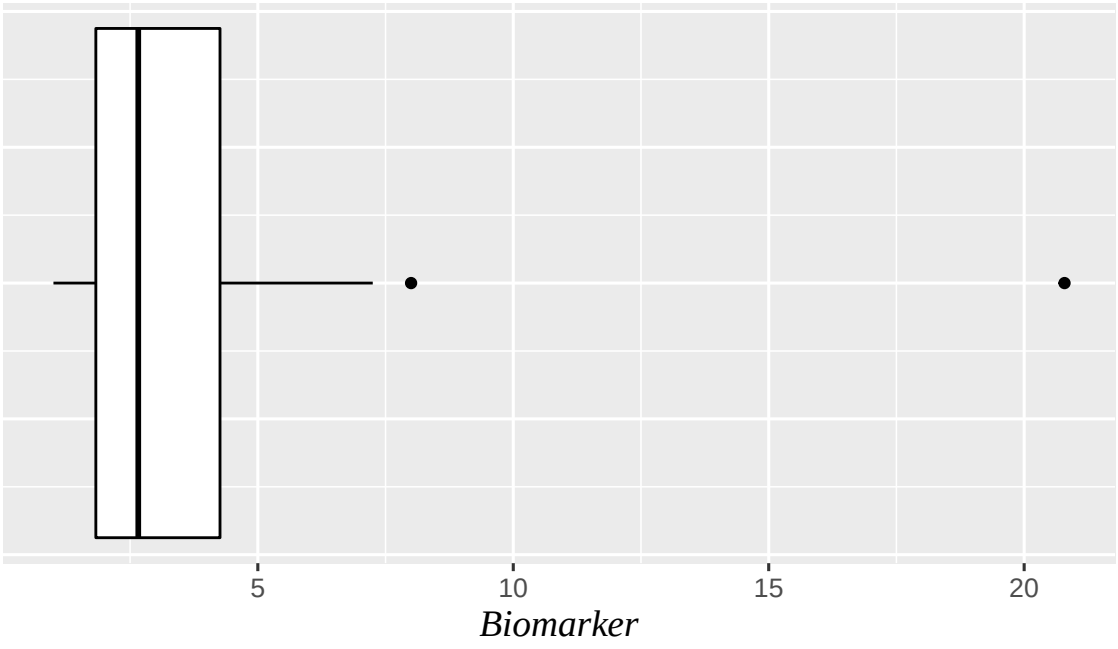

Figure 4: Boxplot of Biomarker

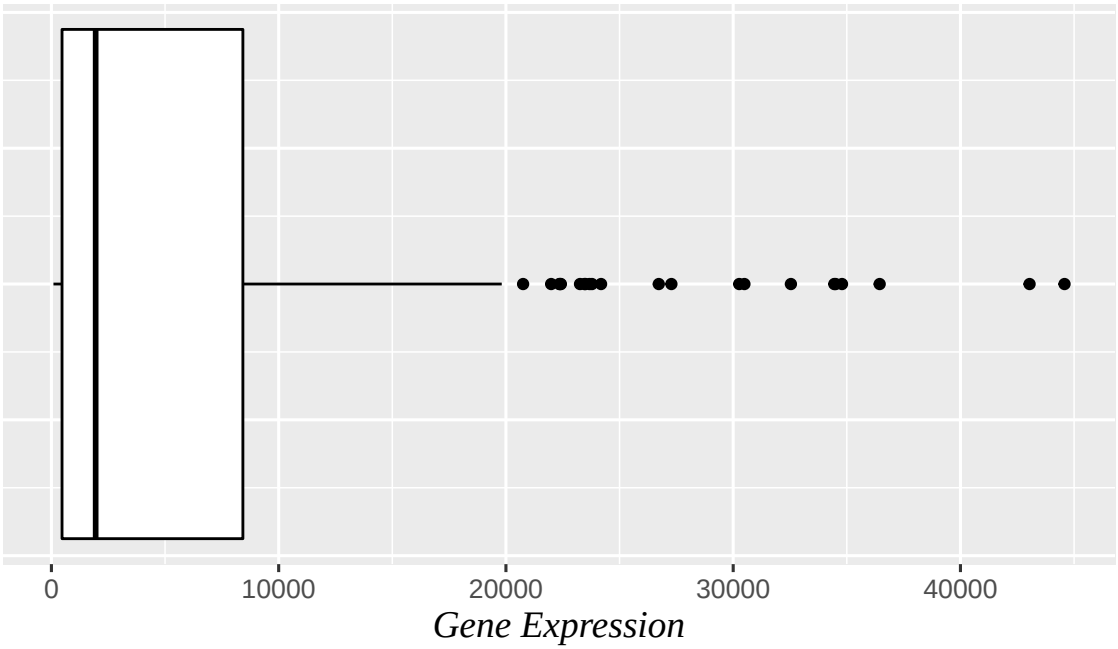

Figure 5: Boxplot of Gene Expression

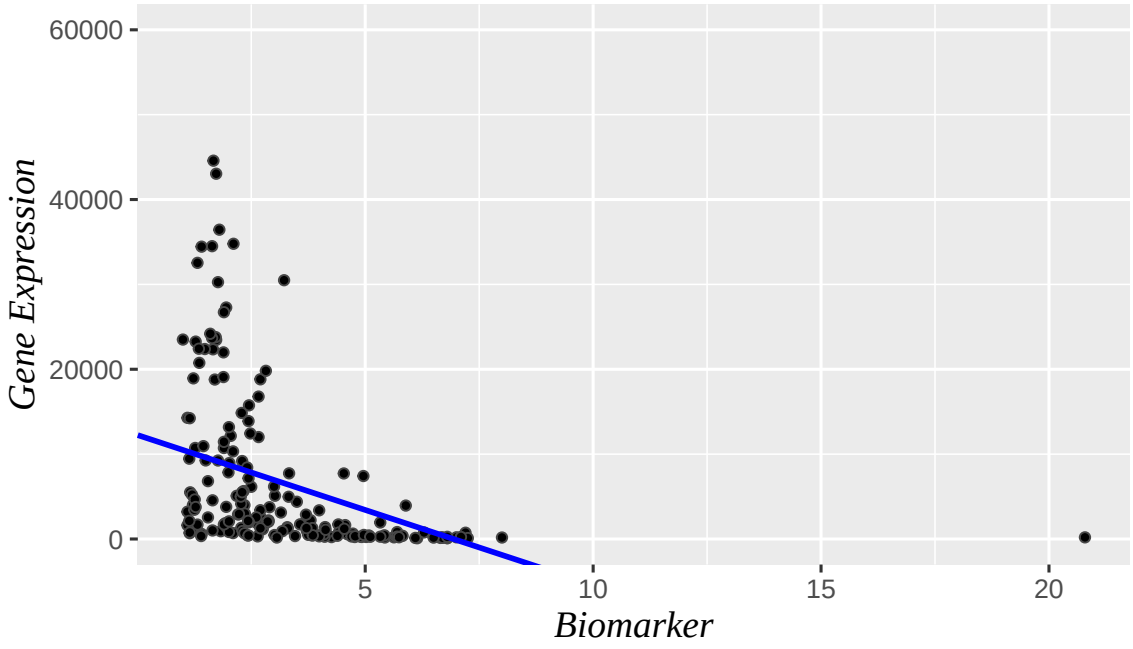

Figure 6: Scatterplot with fitted line

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             Gene Expression
```

```
## ------------------------------------------------
## Biomarker                    -1,767.089***
##                               (297.440)
##
## Constant                     12,256.640***
##                               (1,155.635)
##
## ------------------------------------------------
## Observations                      189
## R2                               0.159
## Adjusted R2                      0.154
## Residual Std. Error    8,688.067 (df = 187)
## F Statistic           35.295*** (df = 1; 187)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Table 1: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	2664182663	2664182663	35.295	0
Residuals	187	14115230170	75482514		

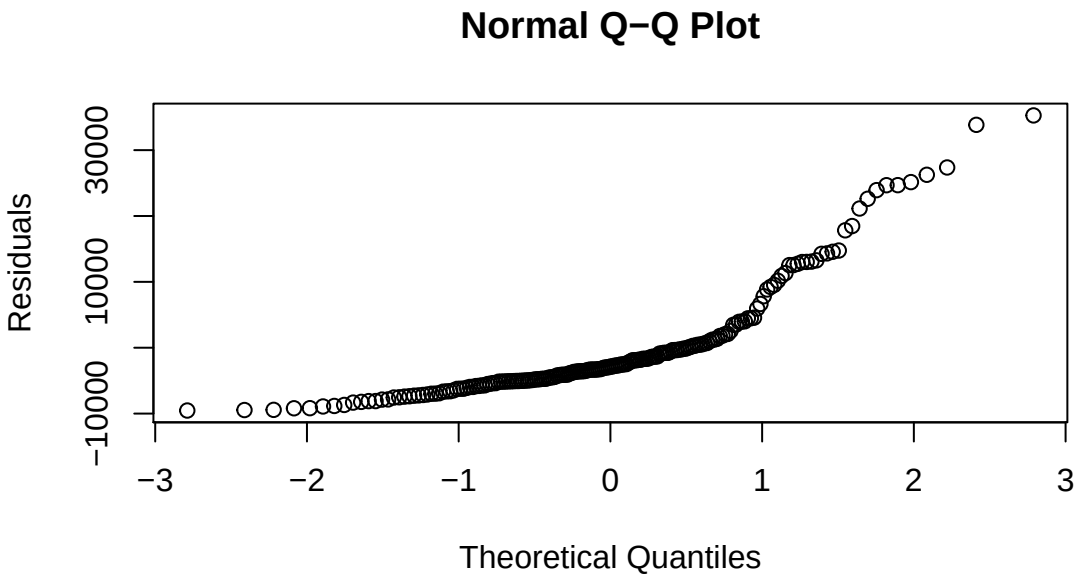

Figure 7: Normal Probability Plot

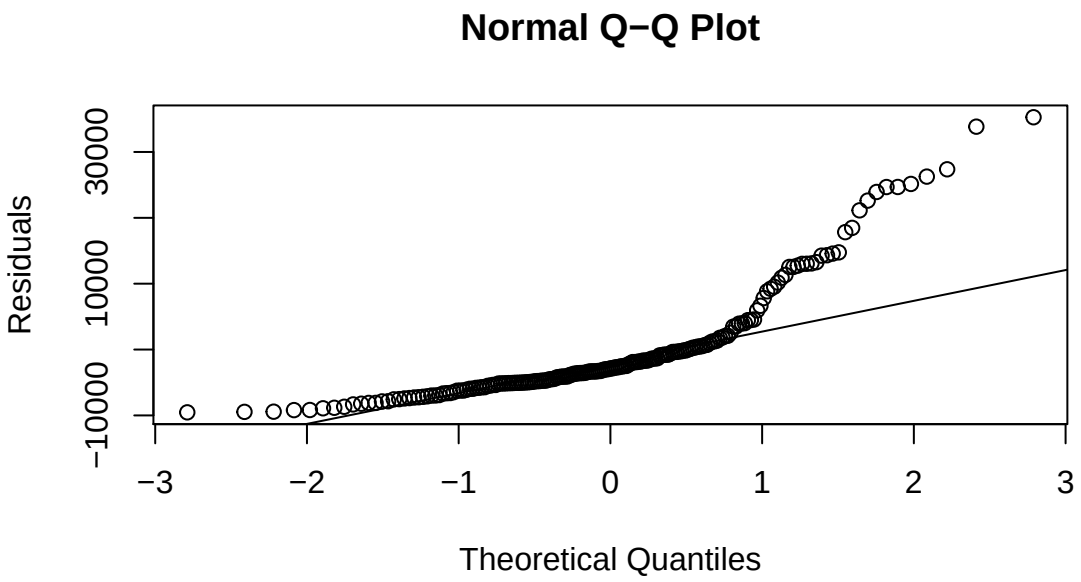

Figure 8: Normal Probability Plot

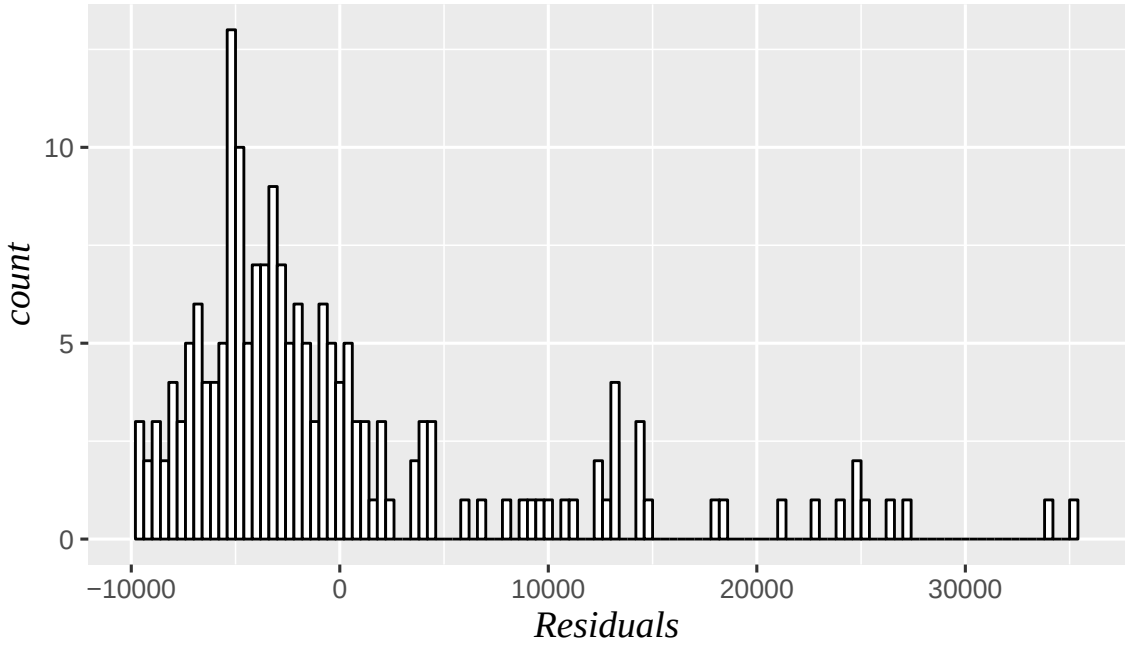

Figure 9: Histogram of Residuals

In the figure above, we fitted the simple linear model, so that we can begin our initial residual analysis, such that:

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$

Before we make any statements about the slope of the fitted simple linear model, we first comment on the normality of the resulting residuals; we can clearly observe from the Normal Probability plot that the residuals do not seem to follow a normal distribution, but rather a distribution with a very heavy right tail. This is confirmed by the histogram of the residuals. Therefore we will not make inferential statements using the t-test on the slope coefficient produced by the simple linear model. We note that this simple linear fit only explained about 15% of the variability in our data set, which was expected given the non-linearity of the relationship. We will now move on to test for the presence of any influential points that may have adversely impacted our fit.

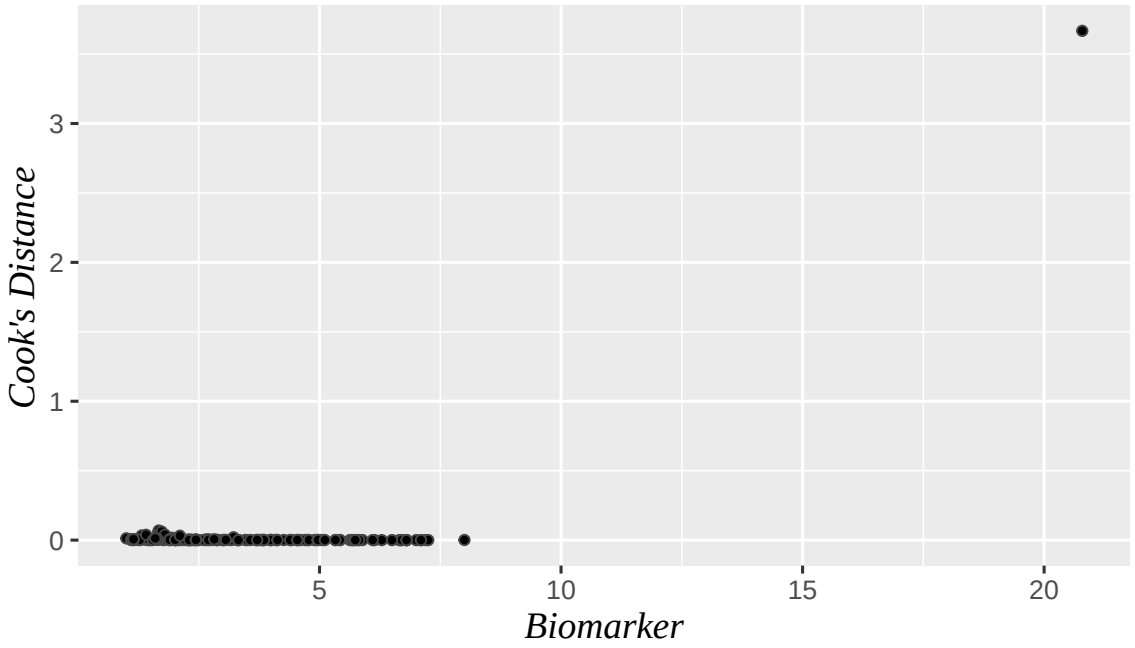

Figure 10: Cook's Distance

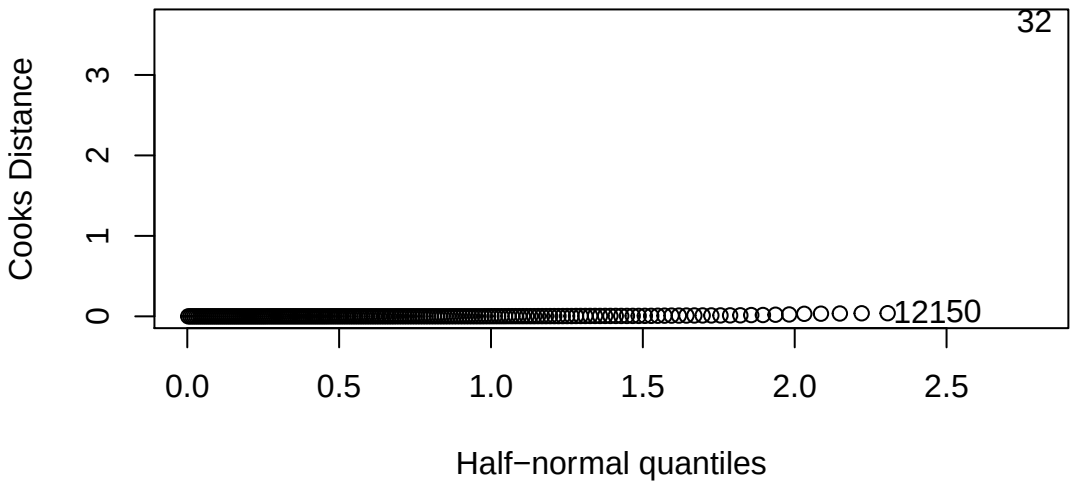

Figure 11: Cook's Distance Standardized Residuals

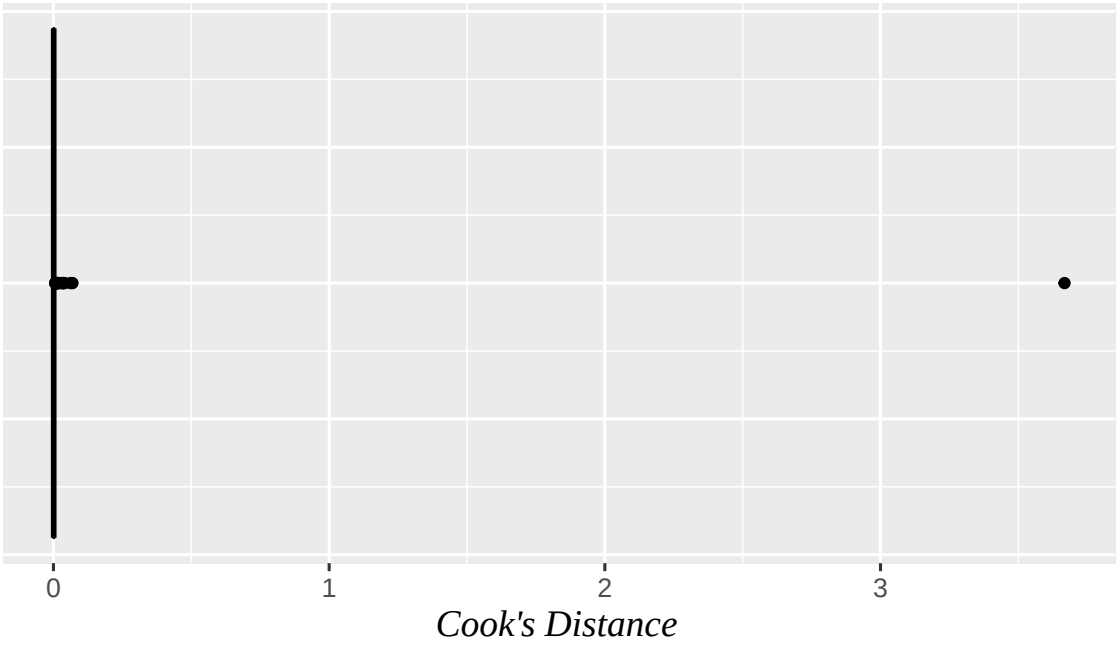

Figure 12: Cook's Distance Boxplot

Above, we present a Cook's distance plot using residuals, another using standardized residuals, and a boxplot of Cook's distances.We can clearly see that there is one point in the data that is much more influential than its peers. we had previously suspected this point of measurement error. Therefore we will exclude this observation from our dataset as a cautionary measure. We provide the statistic used to compute Cook's

distance:

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{\rho.\hat{\sigma}^2} \quad where\ j \neq i\ \ and\ \ \rho = no.\ of\ parameters$$

Now we fit a regression line with the problematic point excluded.

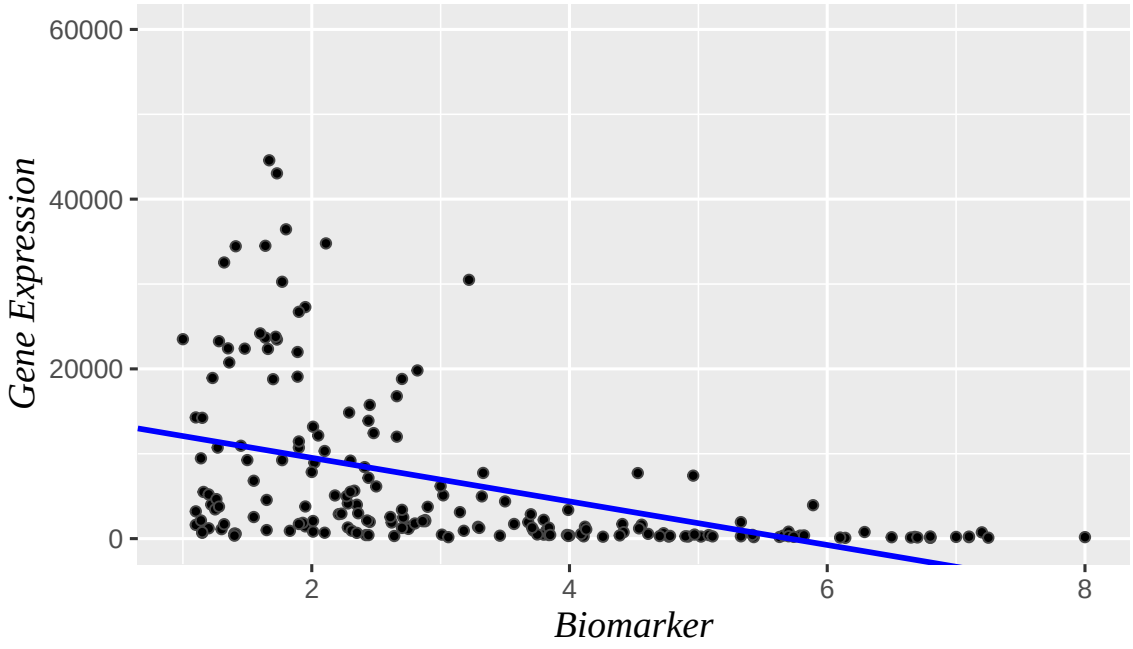

Figure 13: Scatterplot with fitted line

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             Gene Expression
## ------------------------------------------------
## Biomarker                    -2,566.790***
##                               (360.571)
##
## Constant                     14,652.100***
##                               (1,293.870)
##
## ------------------------------------------------
## Observations                      188
## R2                               0.214
## Adjusted R2                      0.210
## Residual Std. Error    8,409.928 (df = 186)
## F Statistic           50.676*** (df = 1; 186)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Table 2: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	3584132830	3584132830	50.676	0
Residuals	186	13155200678	70726885		

Now a linear relationship, seems more reasonable, yet, it is still not entirely obvious, and we are still, despite improvement, only able to explain about 21% of the variation in our data. We note once more, that due to the violation of our normality assumption mentioned above, we are unable to make inferential statements about the fitted models, yet. We will instead, now conduct a transformation of the data. To determine what kind of transformation we wish to apply, we will first examine whether or not we have heteroscedasticity in the data. Following that, we will carry out the box-cox methodology.

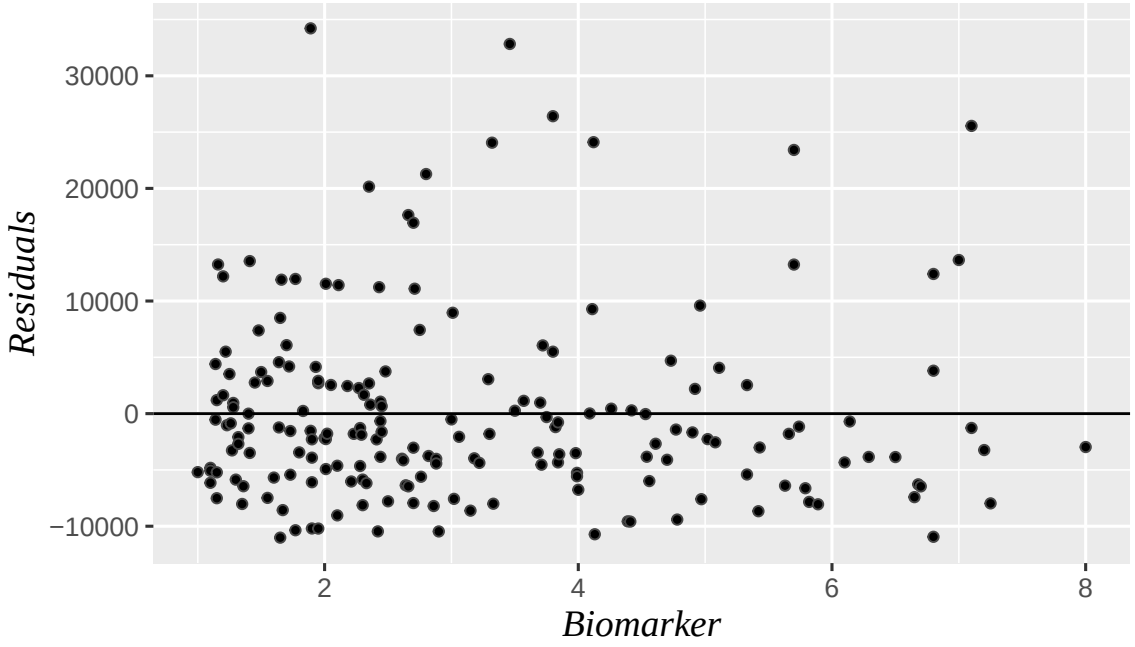

Figure 14: Residual Plot

From the residual plot, it doesn't seem like we have a problem with non-constant variance. To confirm we will perform the Breusch-Pagan and the Brown-Forsythe tests which have test statistics of the following form:

$$BP_F = \frac{SSReg/p}{SSRes/n-p-1} = \frac{\sum_i (\hat{e}_i^2 - \bar{e}^2)^2/p}{\sum_i (e_i^2 - \hat{e}_i^2)^2/n-p-1}$$

Where $\hat{e}_i^2$ is obtained from the below linear model:

$$e_i^2 = \delta_0 + \delta_1.x_i + \xi_i$$

While for the BP_{LM} statistic we have:

$$BP_{LM} = n.\frac{SSReg}{SST} = n.R^2$$

While the BF test statistic is:

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{\sum_i (d_{i1} - \bar{d}_1)^2 + \sum_i (d_{i2} - \bar{d}_2)^2}{n - p - 1}$$

Where the hypotheses in both tests are:

$$H_0 : constant\ variance$$

$$H_a : non - constant\ variance$$

Table 3: Constancy of Variance Tests

BP test	BF test
2.54e-05	1e-07

From the results displayed in the table above and the residual plot, we cannot rule out non-constant variance for the model we are working on. Therefore, we will attempt to apply a transformation that addresses both the non-linearity in our model, and the evidence against constant variance. Three common transformations in this case are: $\sqrt{X}$, $1/X$, and $\log X$ (where log represents the natural logarithm). We will first attempt a log transformation of our independent variable, biomarker, and visually examine the results.

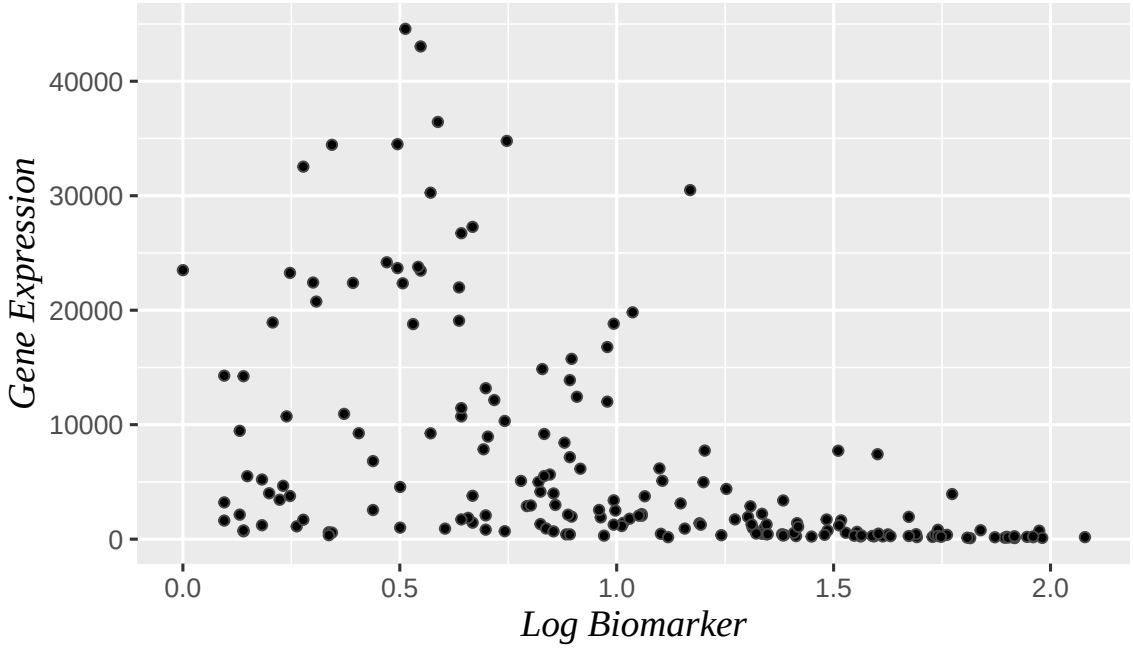

Figure 15: Scatterplot of Log Biomarker vs. Gene Expression

The log transformation does not seem to have been very effective in producing a linearly shaped scatterplot. This is due to the biomedical value variable taking on values in a small range. Therefore, we will now turn instead towards transforming the dependent variable, Gene Expression, via the log transformation.

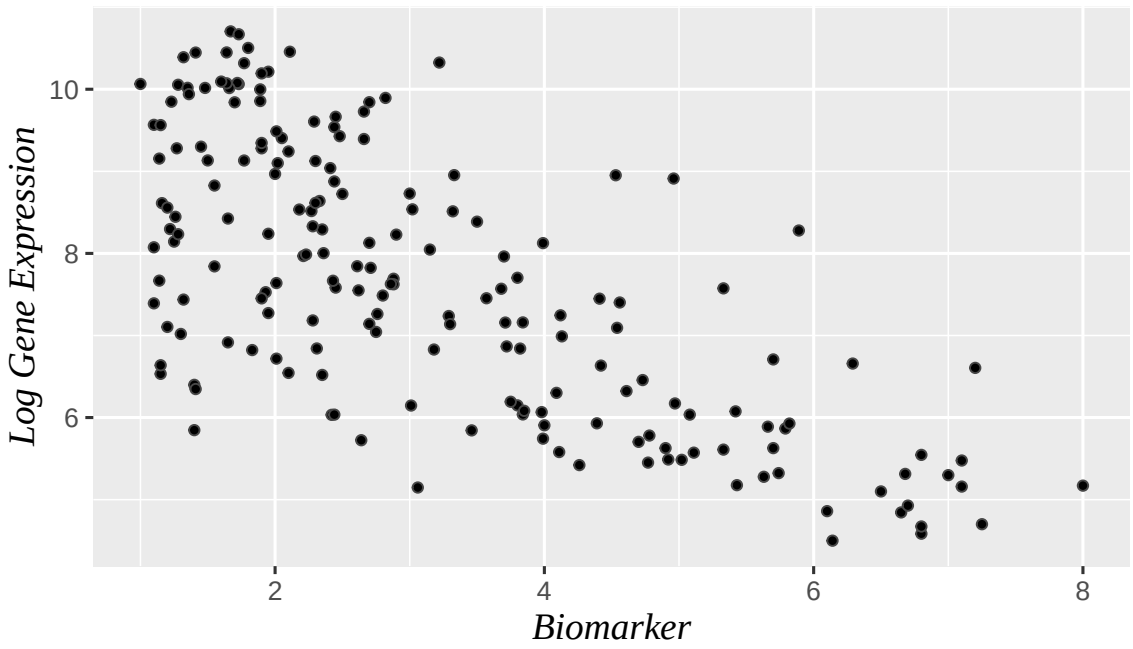

Figure 16: Scatterplot of Log Biomarker vs. Gene Expression

We observe that this transformation produces a much more linear looking scatterplot, however, it still seems to be somewhat curvilinear. Therefore we will now transform both the dependent and the independent variables by the log transformation, and the model we fit is:

$$\begin{aligned} log(Y_i) &= \beta_0 + log(X_i).\beta_1 + \epsilon_i \\ &= Y^* = \beta_0 + X^*.\beta_1 + \epsilon_i \end{aligned}$$

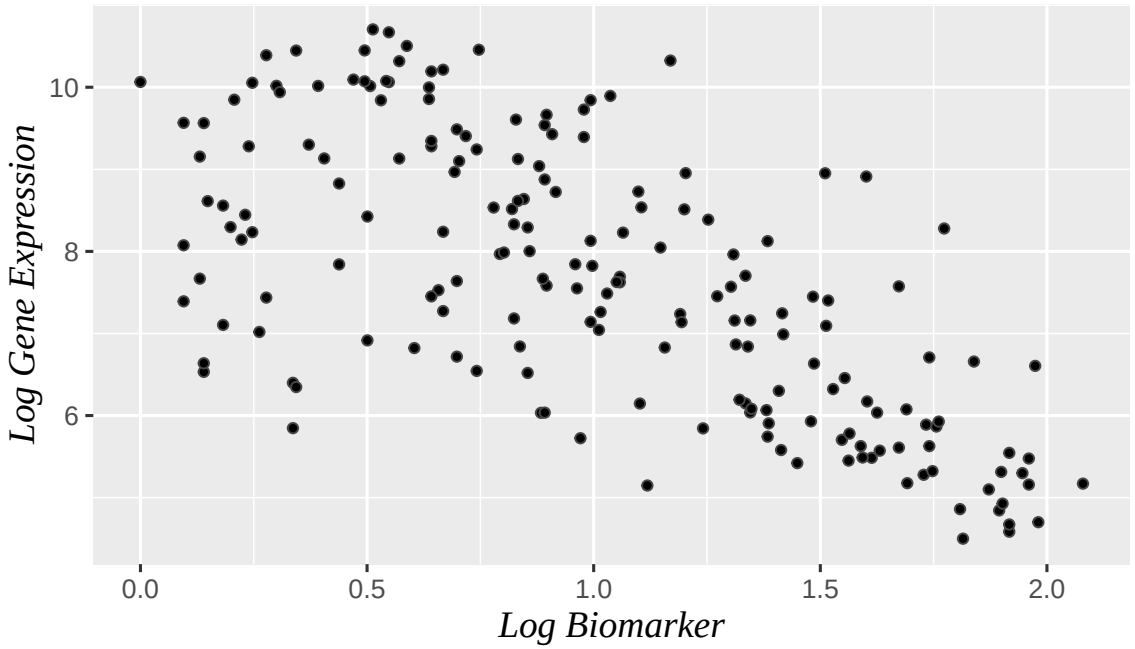

Figure 17: Scatterplot of Log Biomarker vs. Gene Expression

Now, our data seems to much more closely follow a linear pattern than before. For additional certainty we apply the Box-Cox methodologyand find that our transformation is also supported by the box-cox transformation methodology for finding the best lambda:

```
##
## Results of Box-Cox Transformation
## ---------------------------------
##
## Objective Name:                  PPCC
##
## Linear Model:                    lmfit
##
## Sample Size:                     188
##
##  lambda      PPCC
##    -2.0 0.8137339
##    -1.5 0.8787012
##    -1.0 0.9500199
##    -0.5 0.9862248
##     0.0 0.9973994
##     0.5 0.9780660
##     1.0 0.9259406
##     1.5 0.8654906
##     2.0 0.8063981
```

Now we see a relationship that looks more linear than before. However, before we fit our model, we once more need to carry out some descriptive statistics and check our most crucial model assumptions, namely constant variance, normality, and independence of the residuals. We begin by displaying the same visualisations presented earlier, this time for the transformed variables, which we will follow by constructing a residual plot after fitting a simple linear model to the transformed variables:

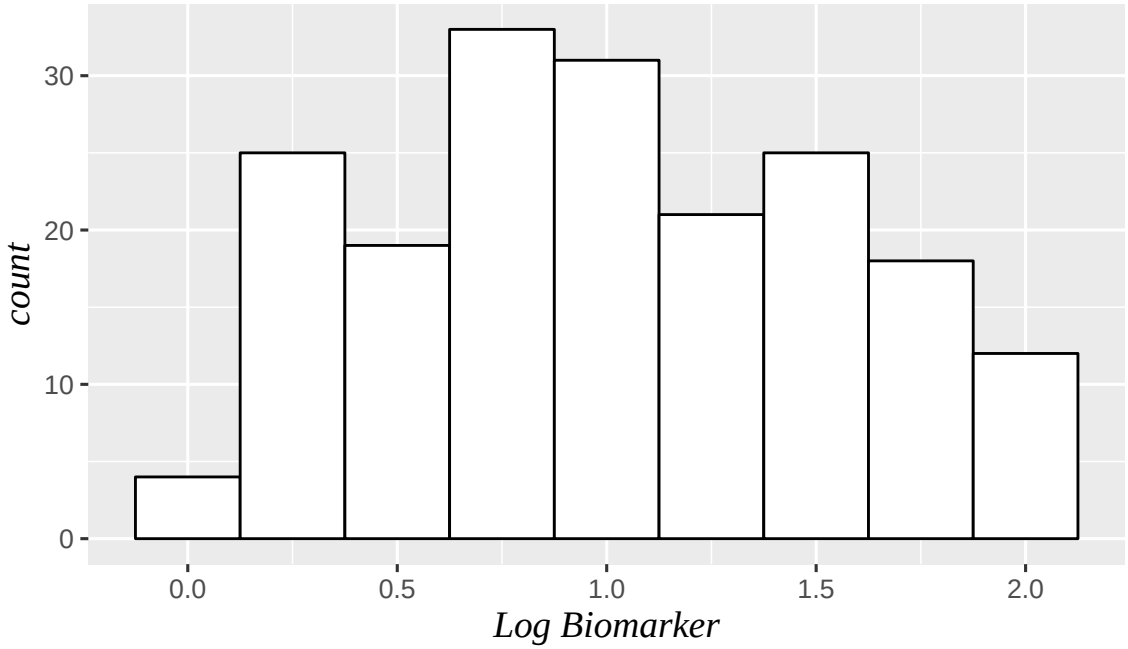

Figure 18: Histogram of Log Biomarker

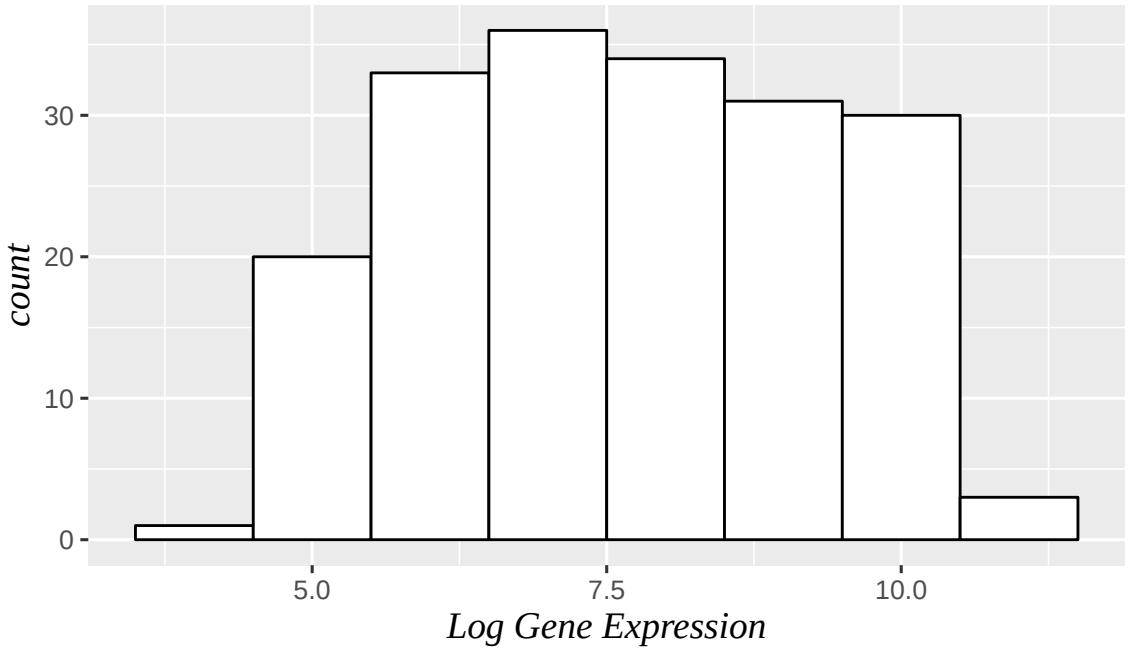

Figure 19: Histogram of Log Gene Expression

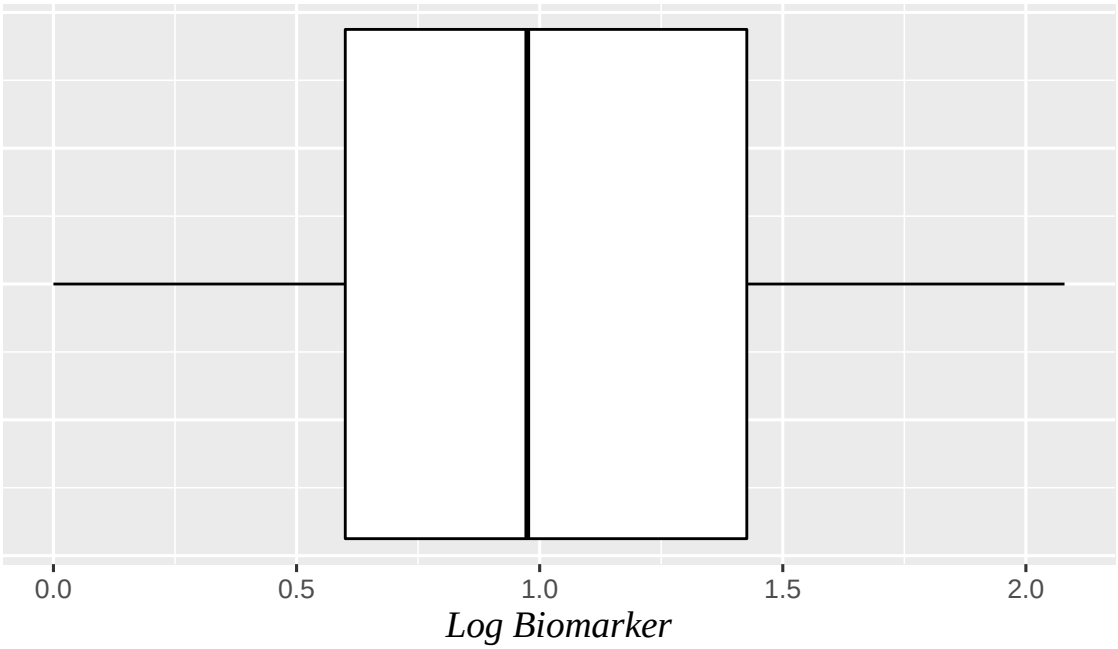

Figure 20: Boxplot of Log Biomarker

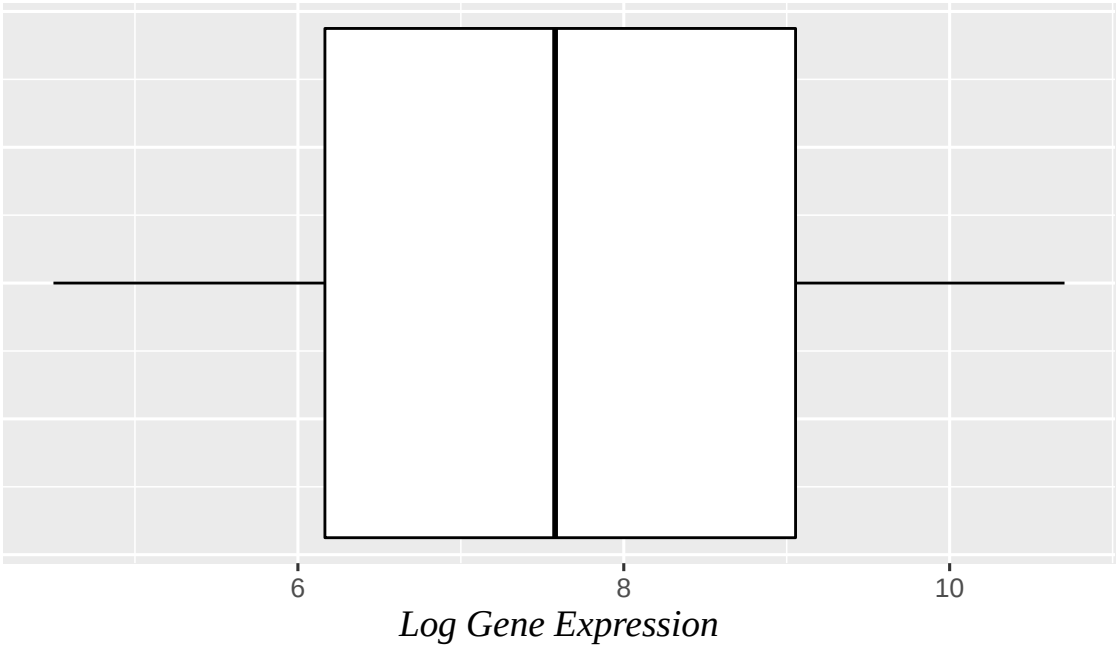

Figure 21: Boxplot of Log Gene Expression

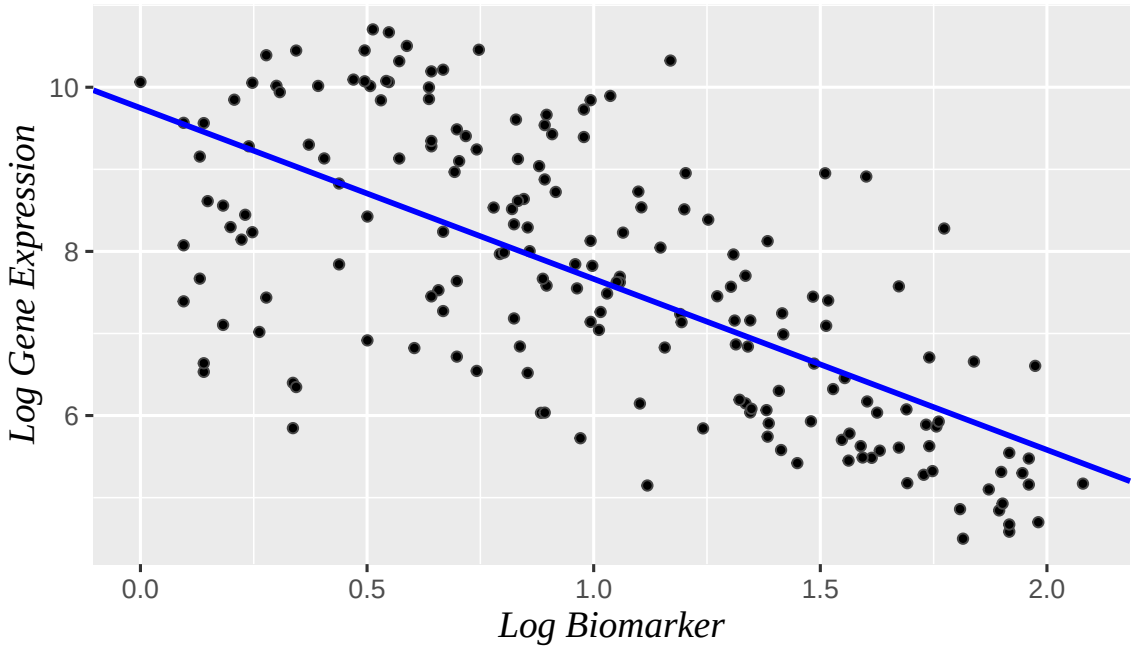

Figure 22: Fitted line on transformed model

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                          Log Gene Expression
## ------------------------------------------------
## Log Biomarker                -2.083***
##                               (0.166)
##
## Constant                     9.747***
##                               (0.190)
##
## ------------------------------------------------
## Observations                    188
## R2                             0.458
## Adjusted R2                    0.455
## Residual Std. Error      1.219 (df = 186)
## F Statistic          157.284*** (df = 1; 186)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Table 4: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	3584132830	3584132830	50.676	0
Residuals	186	13155200678	70726885		

We now observe that our model does a much better job of accounting for the variability in the dependent variable, being able to account for almost 46% of that variability. The linear fit also seems visually reasonable.

We will now check to see if the violation of normality has been resolved by using the same plots as before and applying the Shapiro-Wilk test, which has the following test statistic:

$$SW = \frac{\sum_i (a_i.e_i)^2}{\sum_i e_i^2}$$

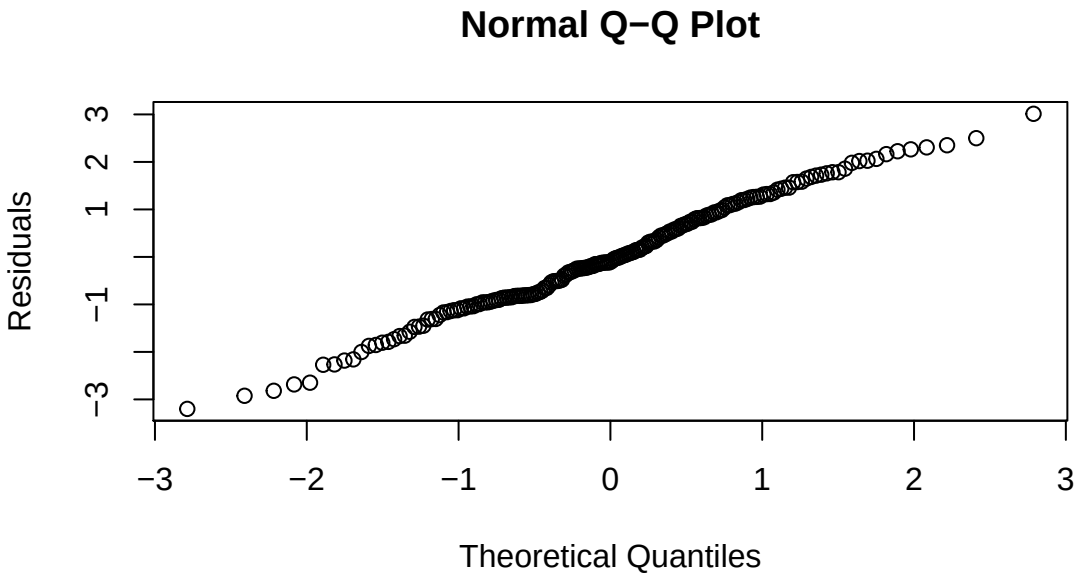

Figure 23: Normal Probability Plot

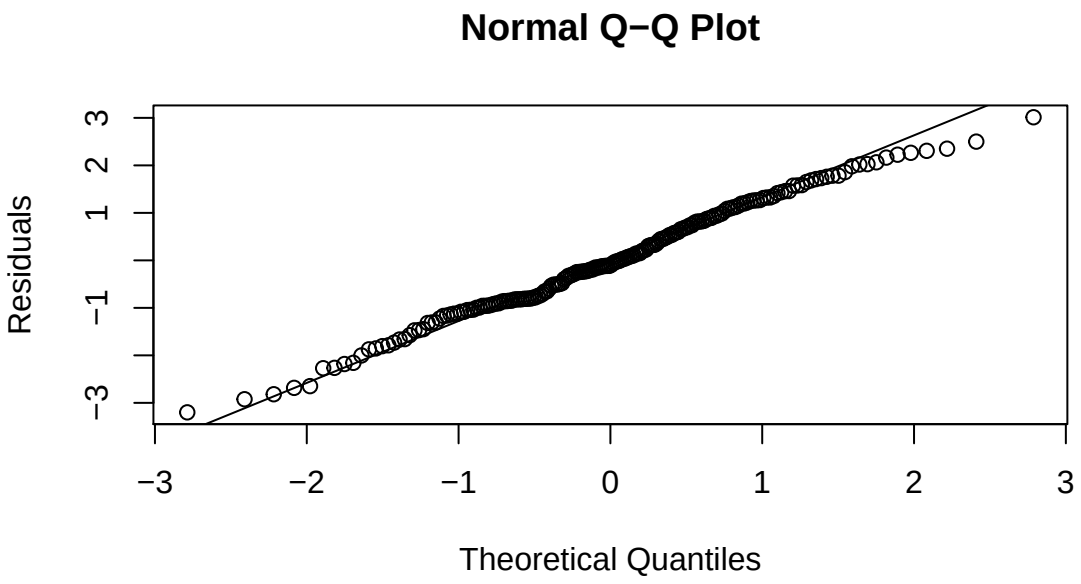

Figure 24: Normal Probability Plot

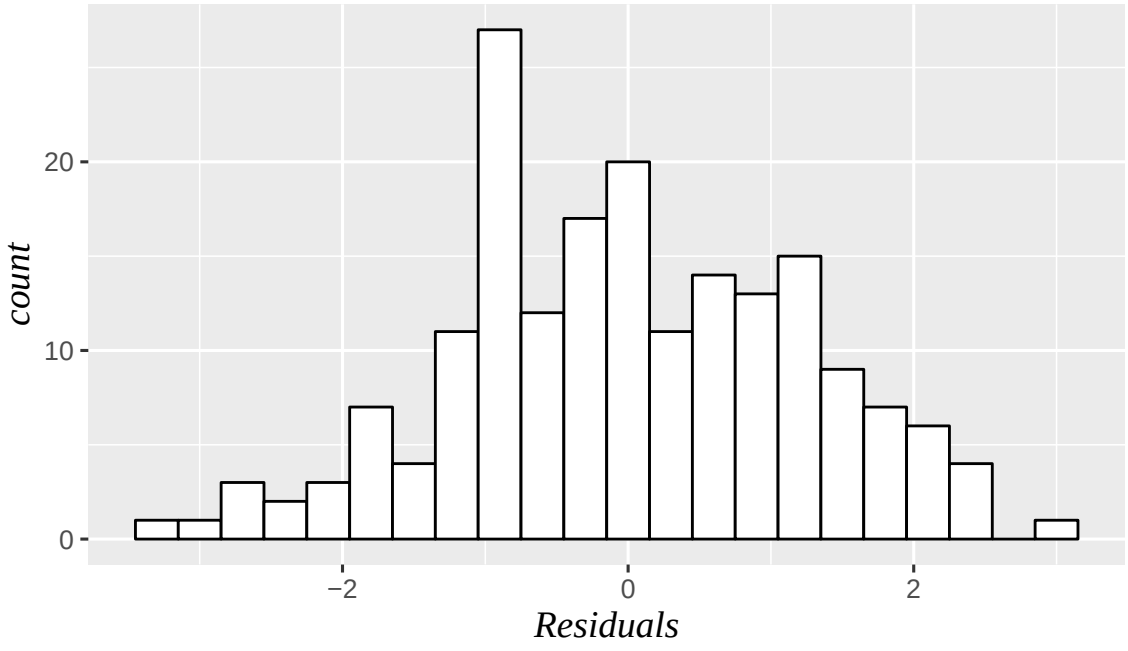

Figure 25: Histogram of Residuals

Our histograms and boxplots of the transformed variables show that the residuals are now approximately normally distributed. This is supported by our Shapiro-Wilk test statistic which has an associated p-value of 0.415538. Next we will examine the residual plot for constancy of variance.

From the residual plot, it seems the constant variance assumption is satisfied. We will once more carry out

our Breusch-Pagan, and Brown-Forsyth test for additional evidence, but first we will once again test for influential points as the presence of these may influence the results of our tests, and our transformation of the model may have caused the appearance of otherwise previously undetected influential points. To test for influential points, we compute and plot Cook's Distance for each value of the residuals.

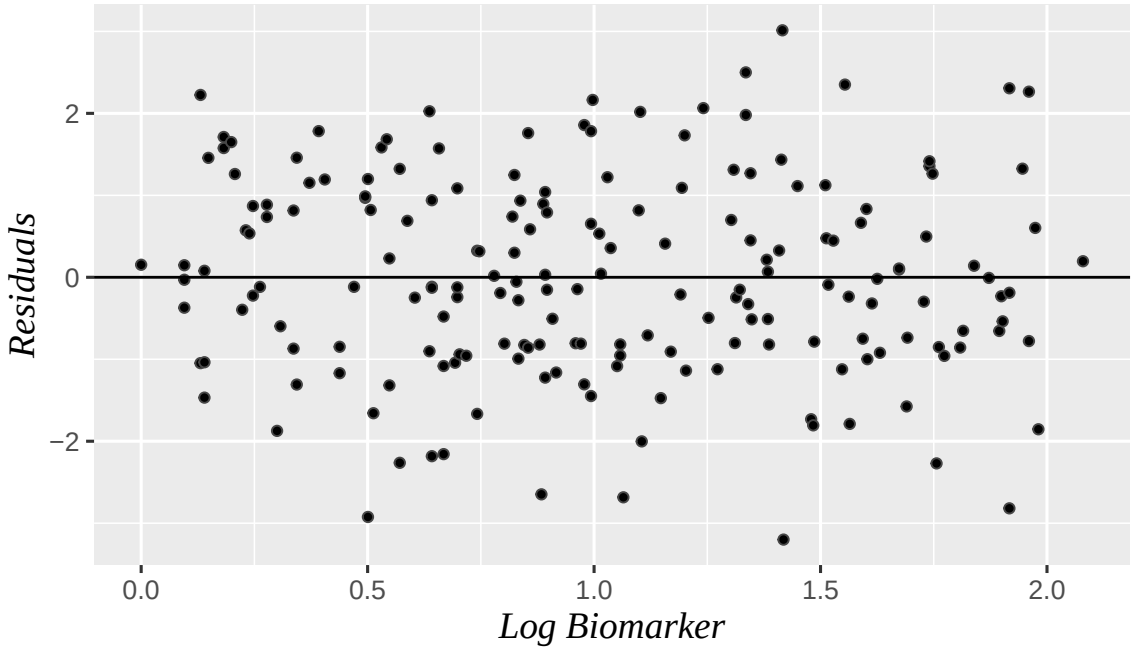

Figure 26: Residual Plot

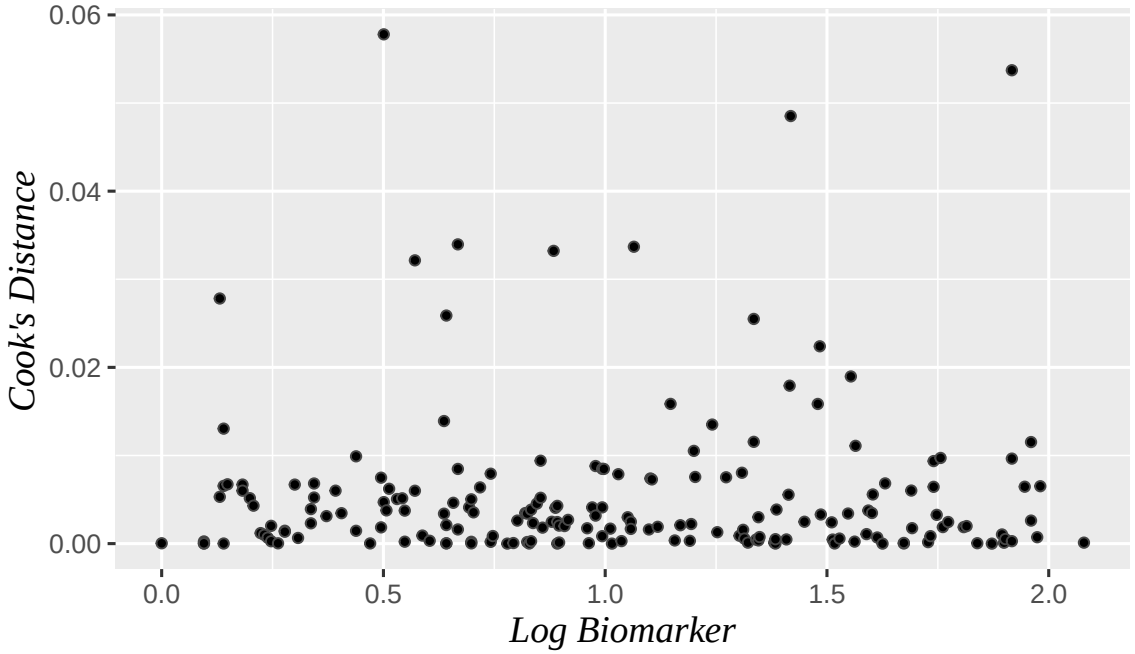

Figure 27: Cook's Distance

We see from the Cook's distance plot, that all values are relatively within range of each other, and therefore we will proceed without further elimination of any other points. But first, we go back to testing for constancy of variance:

Table 5: Constancy of Variance Tests

BP test	BF test
2.54e-05	0.0012711

Our tests provide evidence that is contradictory to what we deduced from the residual plot alone; both the BP and BF test found evidence against the null hypothesis of constant variance. Therefore, our next course of action will be to construct a weighted least squares linear model. We begin by setting the weights as the inverse of the squared residuals resulting from fitting a simple linear model of absolute residuals from the previous transformed model to our transformed independent variables. such that:

$$w_i.Y_i^* = w_i.X_i^*.\beta + w_i\epsilon_i$$

$$Y_i^{*'} = X_i^{*'}.\beta + \epsilon_i$$

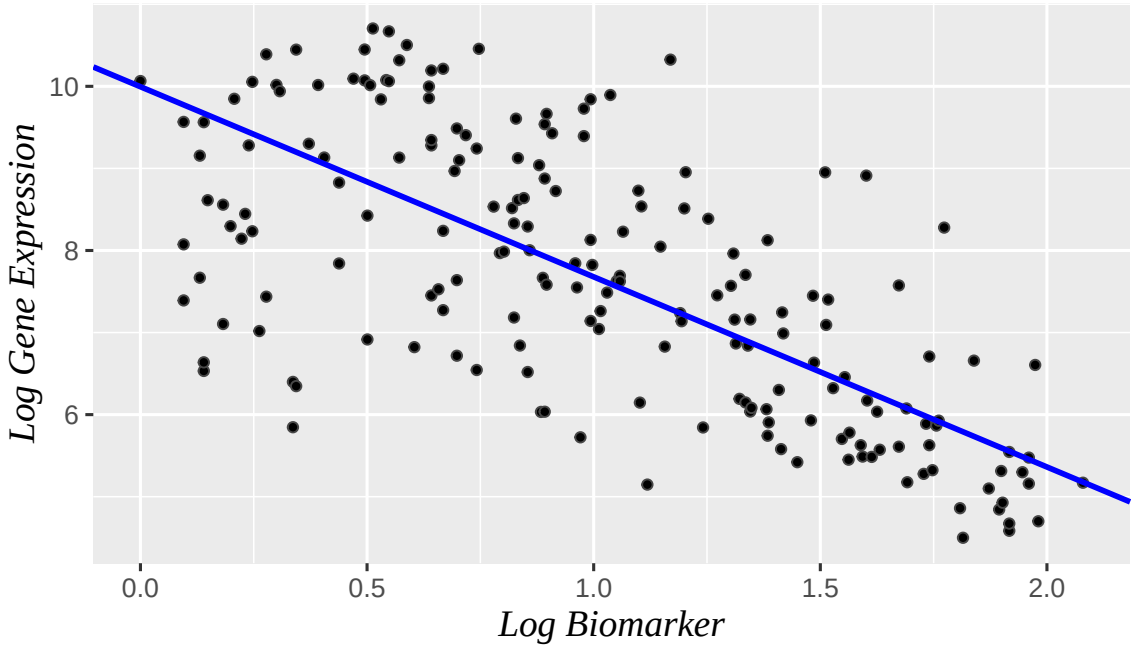

Figure 28: Weighted Least Squares Fitted Line

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                          Log Gene Expression
## ------------------------------------------------
## Log Biomarker                 -2.317***
##                                (0.154)
##
```

```
## Constant                      9.995***
##                                (0.206)
##
## ------------------------------------------------
## Observations                     188
## R2                              0.549
## Adjusted R2                     0.547
## Residual Std. Error       1.206 (df = 186)
## F Statistic           226.733*** (df = 1; 186)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Table 6: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	3584132830	3584132830	50.676	0
Residuals	186	13155200678	70726885		

We have now fitted our weighted least squares model over the log-transformed variables. Doing a quick diagnostic check of the residuals, we find that they are approximately normal (however, the Normal QQ plot shows a thinner right tail, but normality will not be necessary for the purposes of prediction). Our Shapiro-Wilk test also points towards normality not being violated, with an associated p-value of 0.5278906.

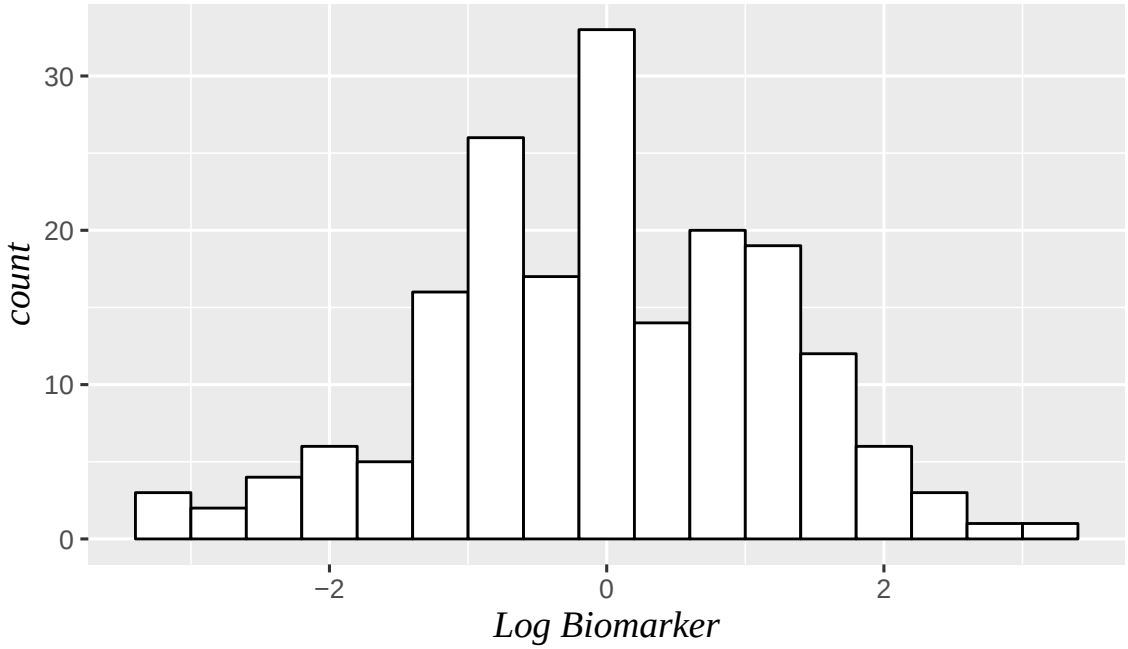

Figure 29: Histogram of Residuals

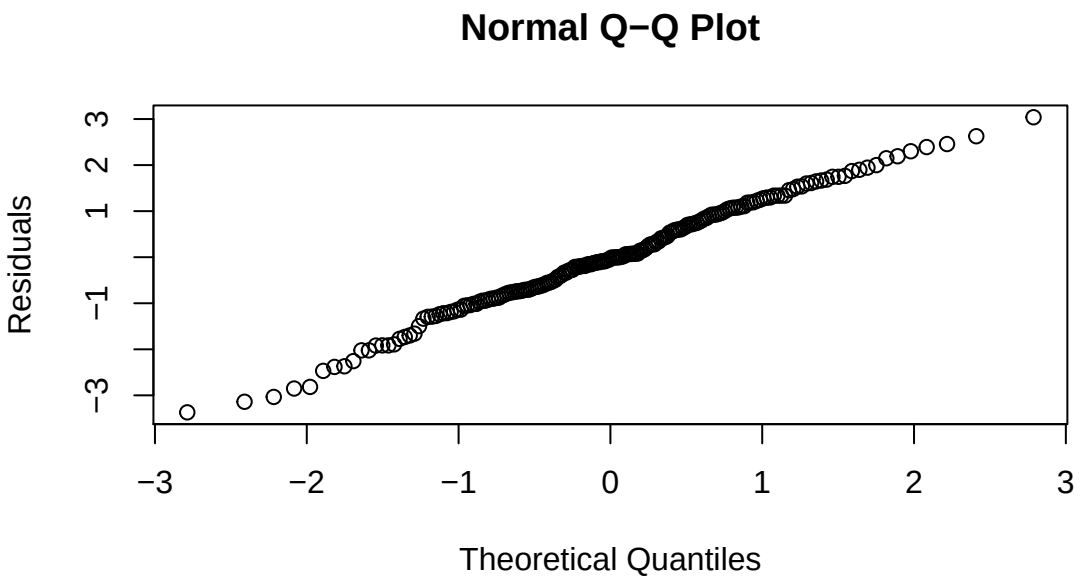

Figure 30: Normal Probability Plot

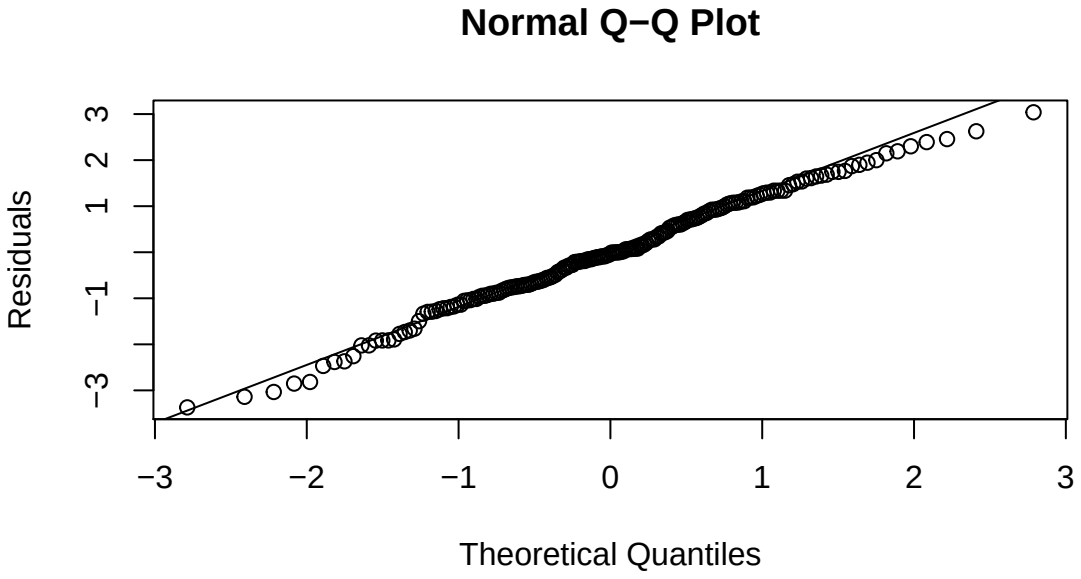

Figure 31: Normal Probability Plot

We also carry out our BP and BF tests once more and find that the problem of non-constant variance has still not been resolved at the type I error level of 5%, as evidenced by the small p-values displayed in the table below, and the scatterplot of the WLS residuals. Therefore, we will adjust our weights by attempting to examine the relationship between the residuals and the independent variable more closely.

Table 7: Constancy of Variance Tests

BP test	BF test
0	0.0003359

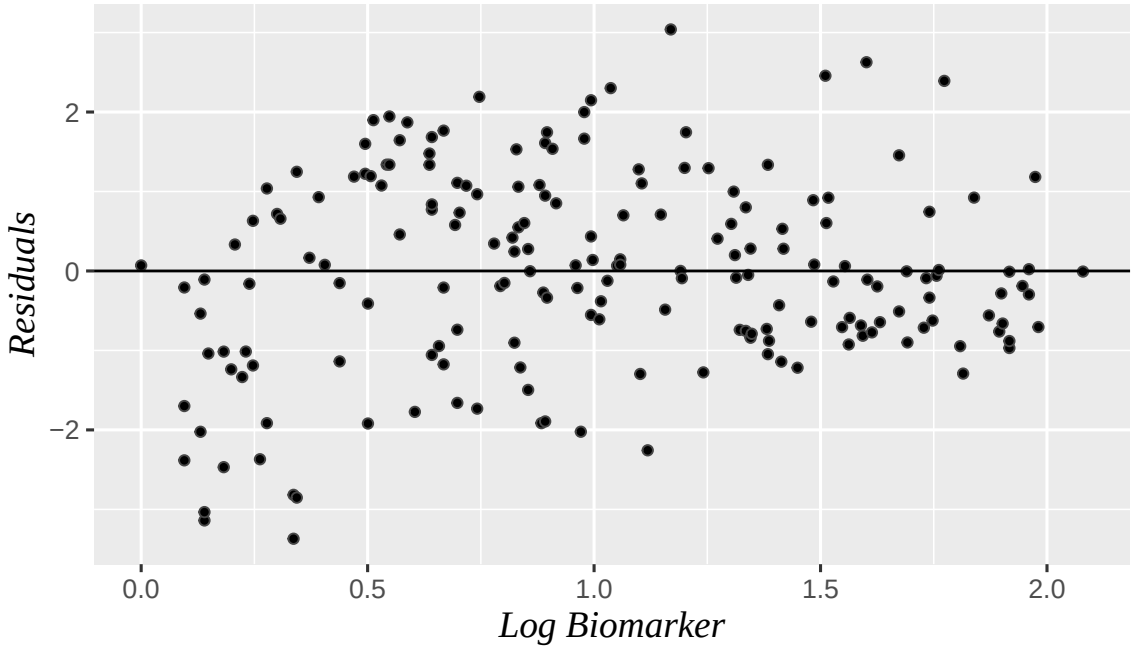

Figure 32: Residual Plot

By visually examining the resulting residual plot from our weighted least squares model, we observe, what appears to us to be a curvilinear relationship between the residuals and the independent variable. The relationship might fit well with a square root transformation of the independent variable when regressing absolute residuals on the Log Biomarker. Therefore, we will construct our weights as follows:

$$w_i = \frac{1}{\hat{u}_i^2}$$

where

$$u_i = \sqrt{X^*}.\delta_i + \epsilon_i$$

and

$$\hat{u}_i = \sqrt{X^*}.\hat{\delta}_i$$

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                          Log Gene Expression
## ------------------------------------------------
## Log Biomarker                 -2.324***
```

```
##                                (0.159)
##
## Constant                      10.023***
##                                (0.208)
##
## ------------------------------------------------
## Observations                     188
## R2                              0.535
## Adjusted R2                     0.532
## Residual Std. Error       1.204 (df = 186)
## F Statistic           213.803*** (df = 1; 186)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Table 8: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	3584132830	3584132830	50.676	0
Residuals	186	13155200678	70726885		

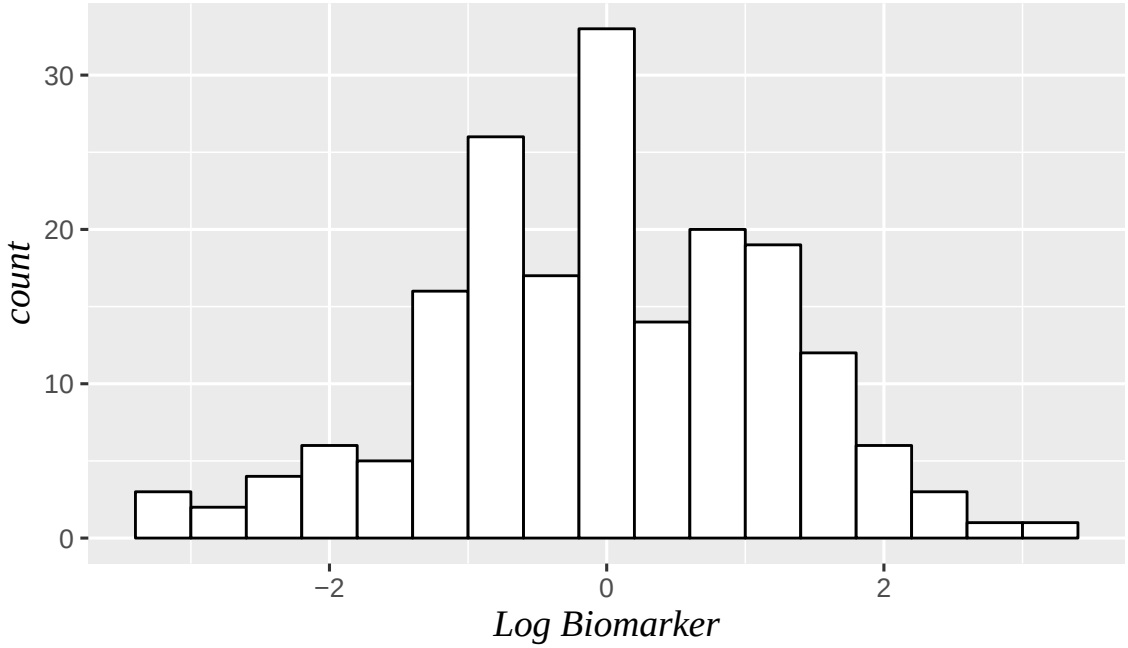

Figure 33: Histogram of Residuals

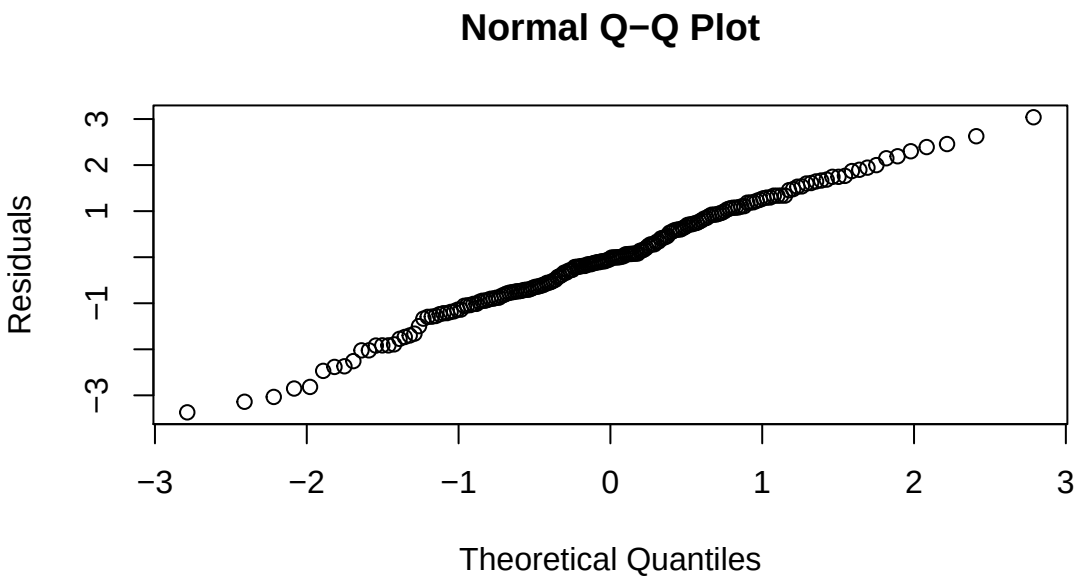

Figure 34: Normal Probability Plot

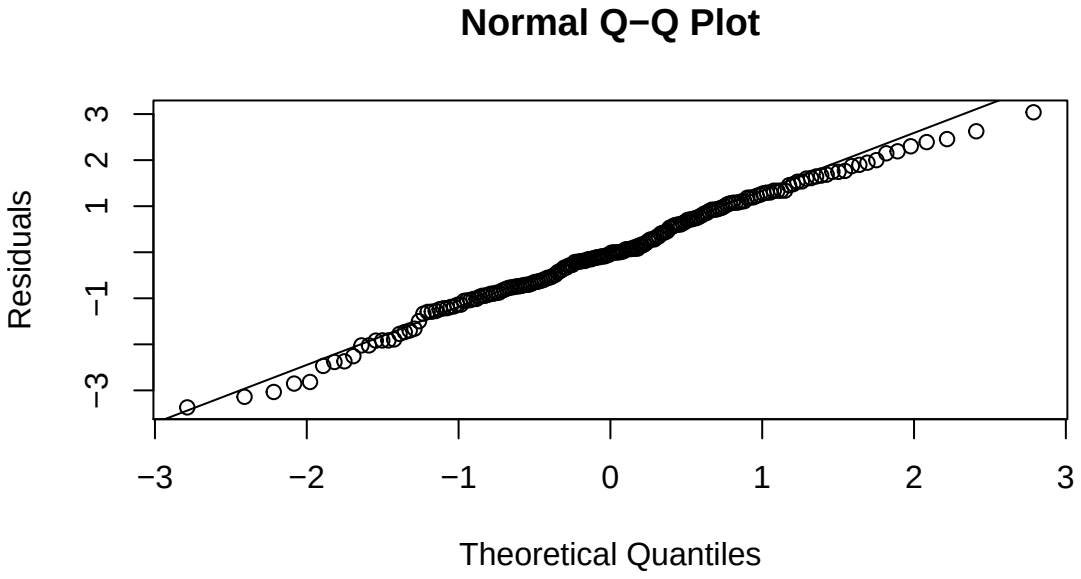

Figure 35: Normal Probability Plot

Now, we have a new weighted least squares model, where the normality assumption holds (as observed from the Normal QQ plot, histogram of the residuals, and Shapiro-Wilk p-value of 0.524). we will conduct our constant variance tests:

Table 9: Constancy of Variance Tests

BP test	BF test
0.1300218	0.9258172

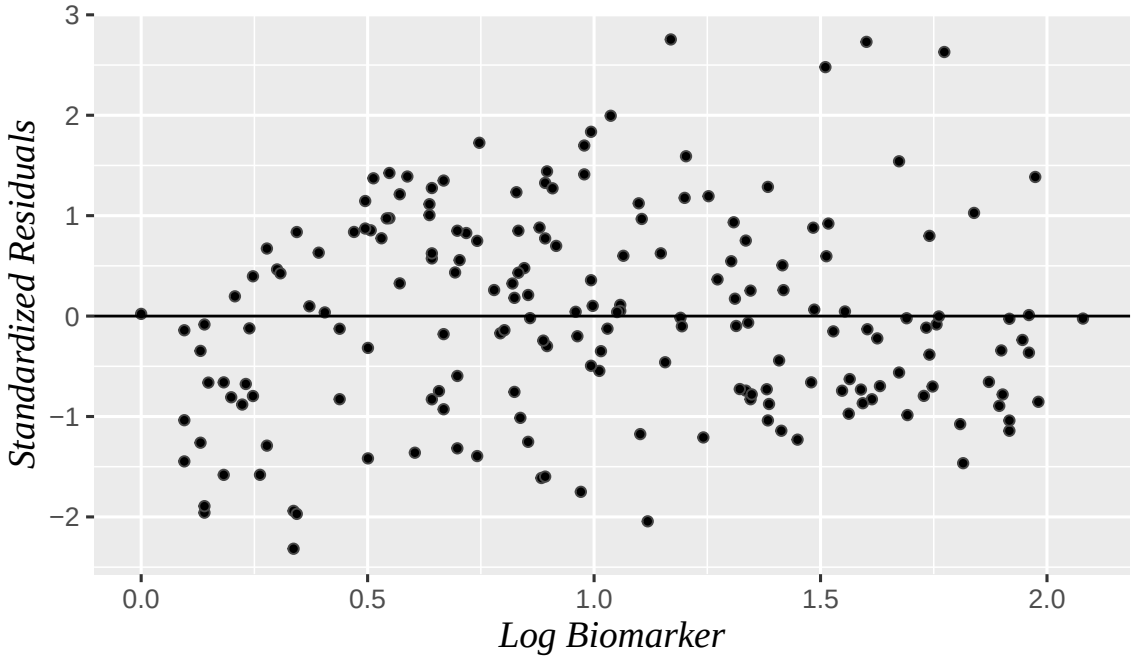

Figure 36: Residual Plot

Our results now point toward the constant variance assumption being satisfied (we note however the strong difference between the p-values of both tests which may indicate an error in the testing procedure for the Brown-Forsyth test, or prehaps this is due to the robustness of the Brown-Forsyth test against outliers). However, we note that the scatterplot does not seem to show constant variance, therefore we must be cautious when making statements that take the variance of the residuals into account.

We will now conduct a Runs Test to check the independence of the error terms. The results displayed below, indicate that our errors are independently distributed.

```
##
##  Runs Test - Two sided
##
## data:  residuals(lmfit1t_wls2)
## Standardized Runs Statistic = 0, p-value = 1
```

Now that we have a model that meets our critical assumptions, we begin the prediction process. First we note that we have applied two transformations to our model, the first being the Log-transformation to both the dependent and independent variables; the second being to weight our variables to conduct weighted least squares regression.Therefore, after we obtain our predicted value, we will need to carry out the anti-log transformation to find our desired prediction. We also present the confidence interval for our coefficient below.

Table 10: Confidence Interval for Coefficients

	Lower Bound CI	Upper Bound CI
Log Biomarker	-2.637541	-2.010436

Table 11: Prediction and Prediction Interval

	Predicted	Lower Bound PI	Upper Bound PI
Log Gene Expression	-13.2165889	-16.86172	-9.5714600
Gene Expression	0.0000018	0.00000	0.0000697

Inverse Polynomial Model

Now, we will attempt to build a second model, this time for the prediction of Biomarker value from Gene Expression.

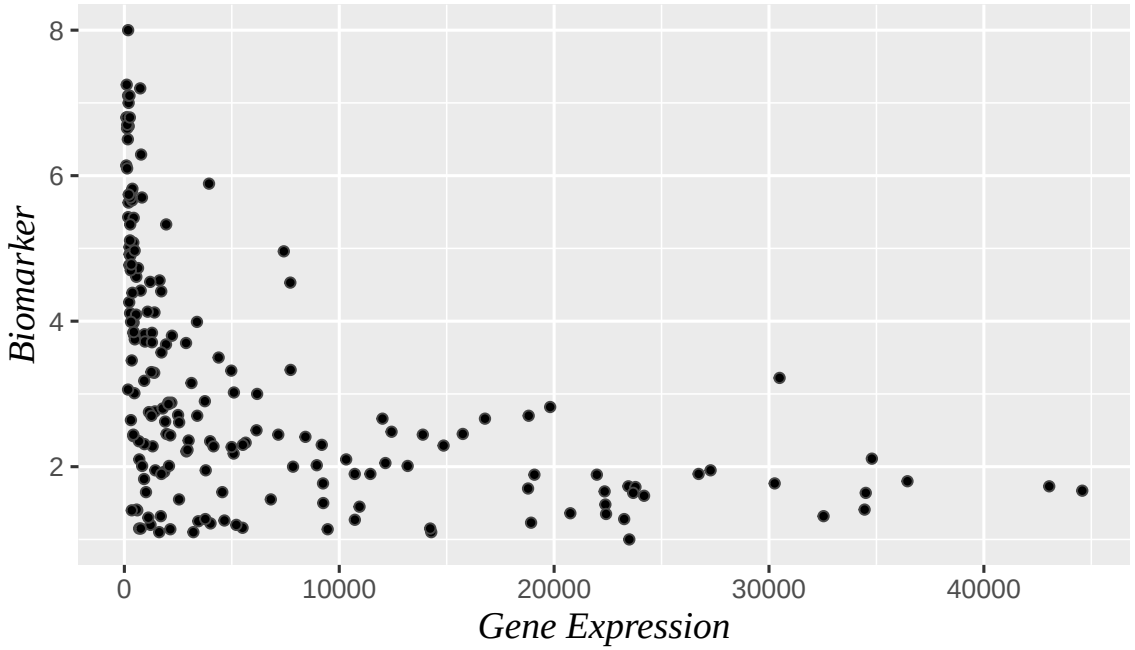

Figure 37: Scatterplot of Biomarker vs. Gene Expression

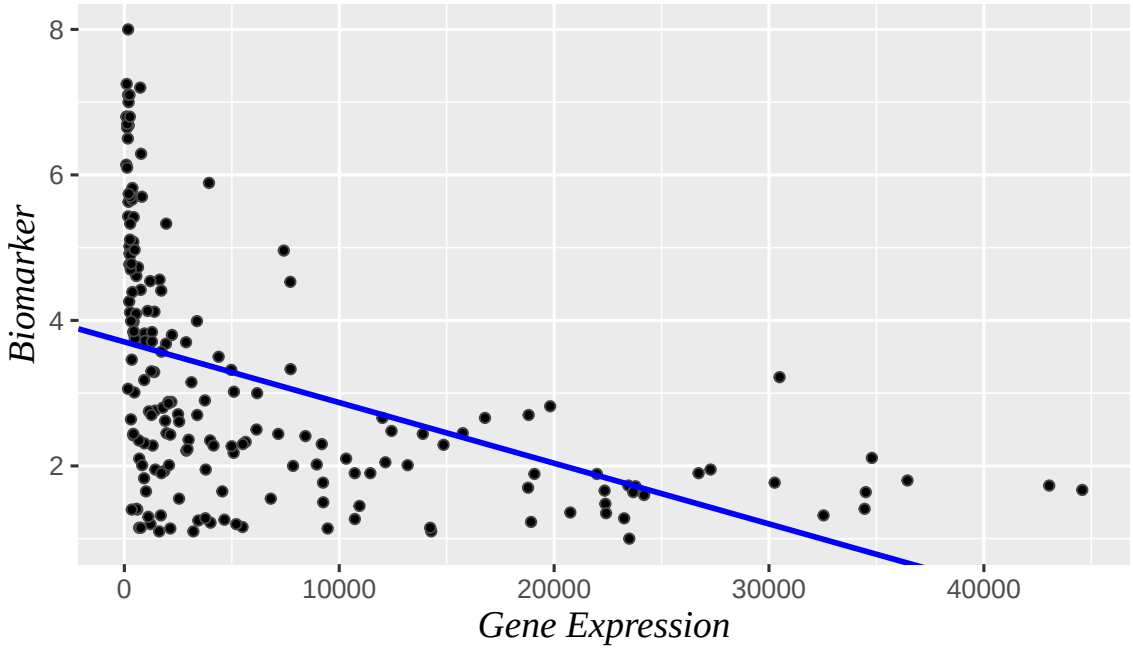

Figure 38: Scatterplot with fitted line

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                Biomarker
## ------------------------------------------------
## Gene Expression              -0.0001***
##                               (0.00001)
##
## Constant                      3.705***
##                                (0.135)
##
## ------------------------------------------------
## Observations                     188
## R2                              0.214
## Adjusted R2                     0.210
## Residual Std. Error       1.516 (df = 186)
## F Statistic            50.676*** (df = 1; 186)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Table 12: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	116.4794	116.479441	50.676	0
Residuals	186	427.5261	2.298528		

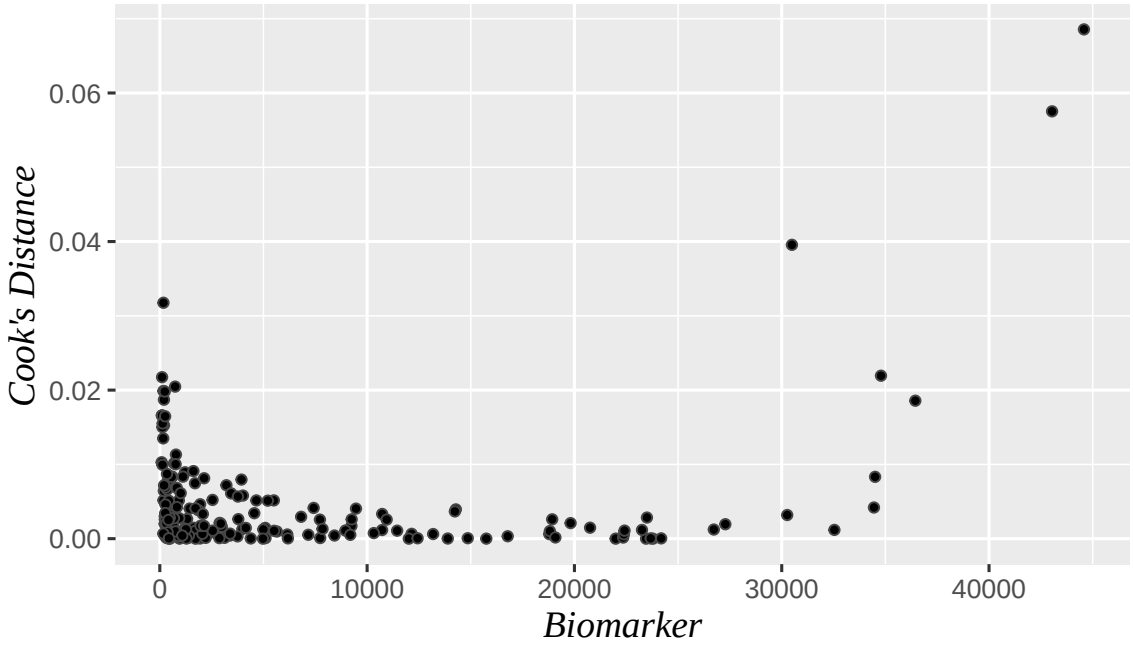

Figure 39: Cook's Distance

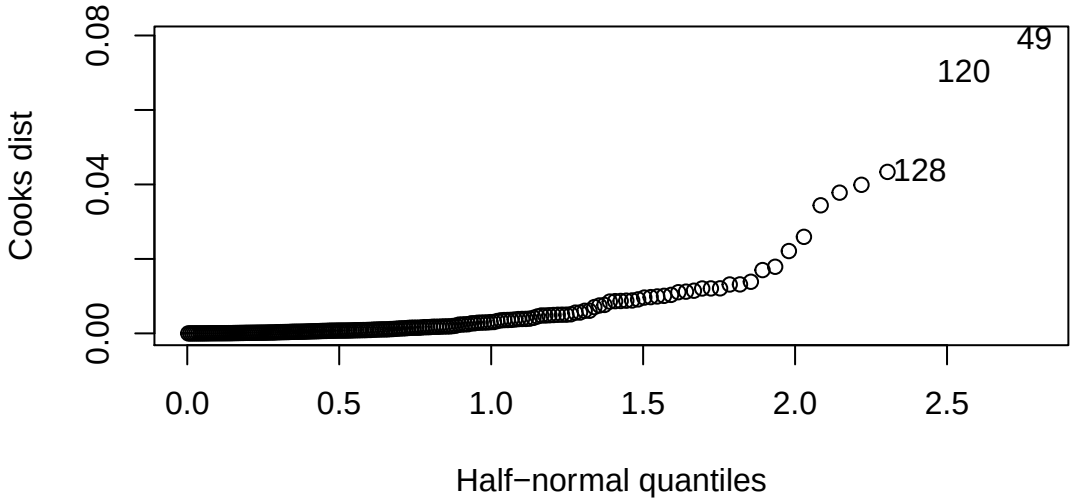

Figure 40: Cook's Distance

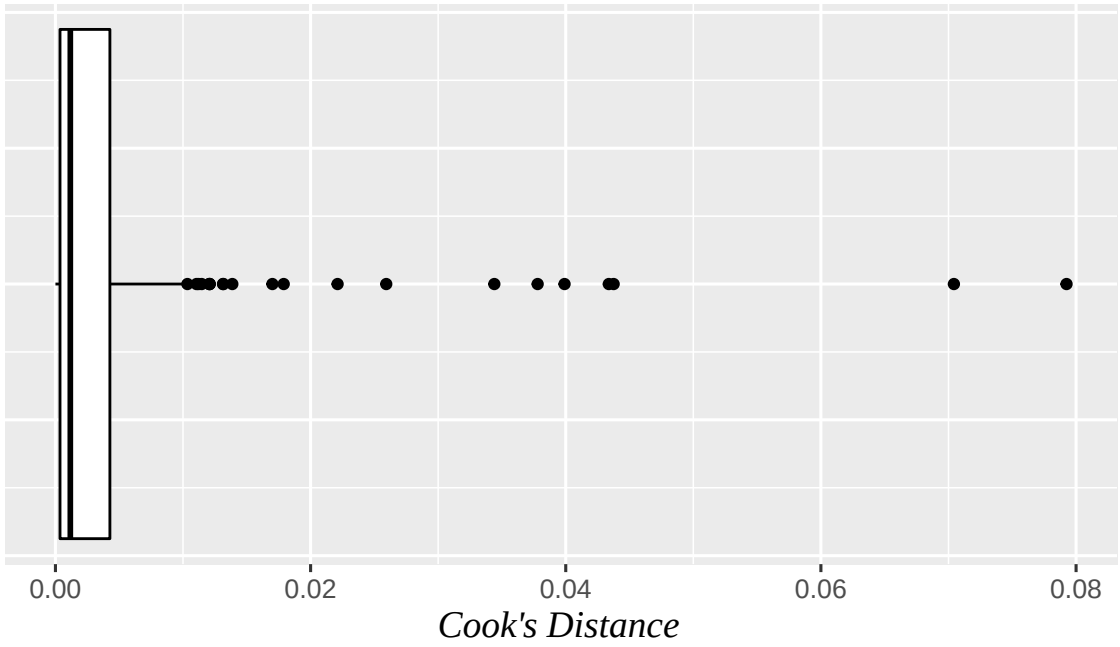

Figure 41: Cook's Distance

Here, we went through the same process as in the weighted least squares section; we began by fitting the simple linear model once more, this time with Biomarker being the dependent variable. We used the resulting residuals to compute and plot Cook's distance. From our cook's distance plots, we have multiple influential points. We will remove the points that are farthest from all other observations; namely observations 49, 120, & 140. We now, once again, fit a linear model to the data, after removing the aforementioned observations.

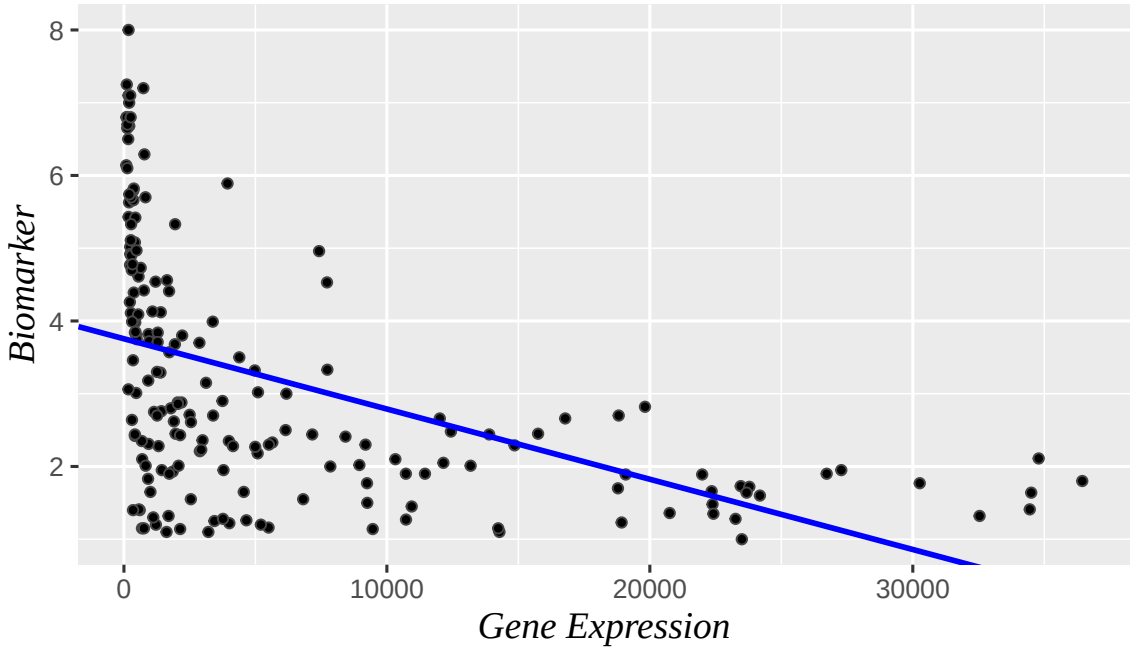

Figure 42: Scatterplot with fitted line

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                Biomarker
## ------------------------------------------------
## Gene Expression              -0.0001***
##                               (0.00001)
##
## Constant                      3.755***
##                                (0.136)
##
## ------------------------------------------------
## Observations                    185
## R2                             0.230
## Adjusted R2                    0.226
## Residual Std. Error      1.507 (df = 183)
## F Statistic           54.761*** (df = 1; 183)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Table 13: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	124.3031	124.303129	54.761	0
Residuals	183	415.3921	2.269902		

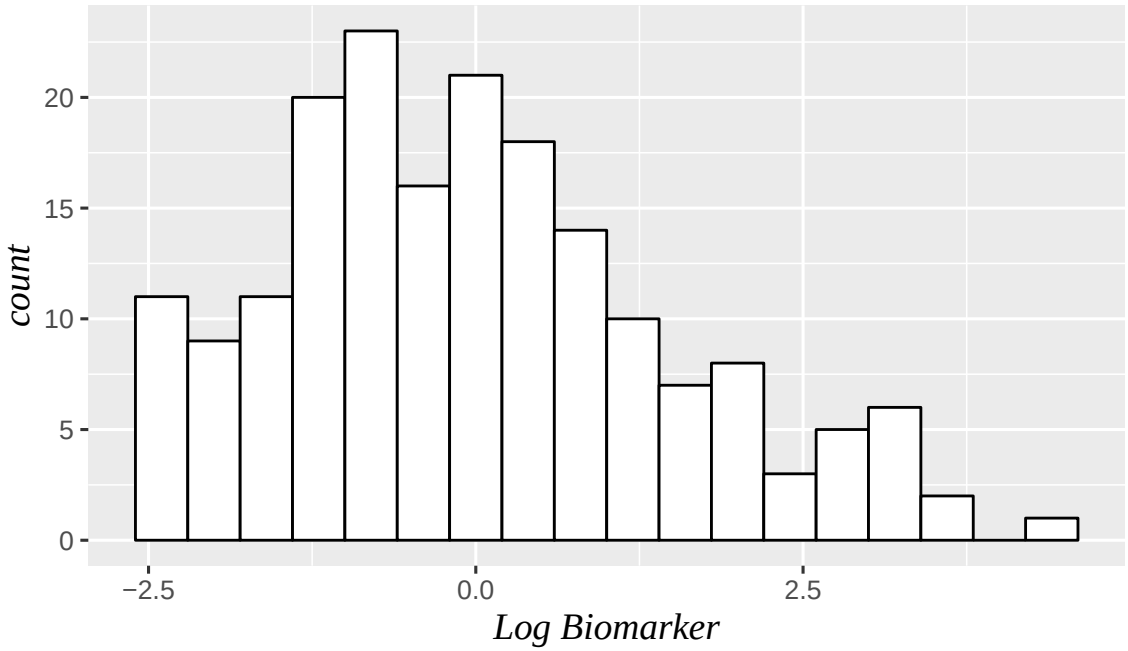

Figure 43: Histogram of Residuals

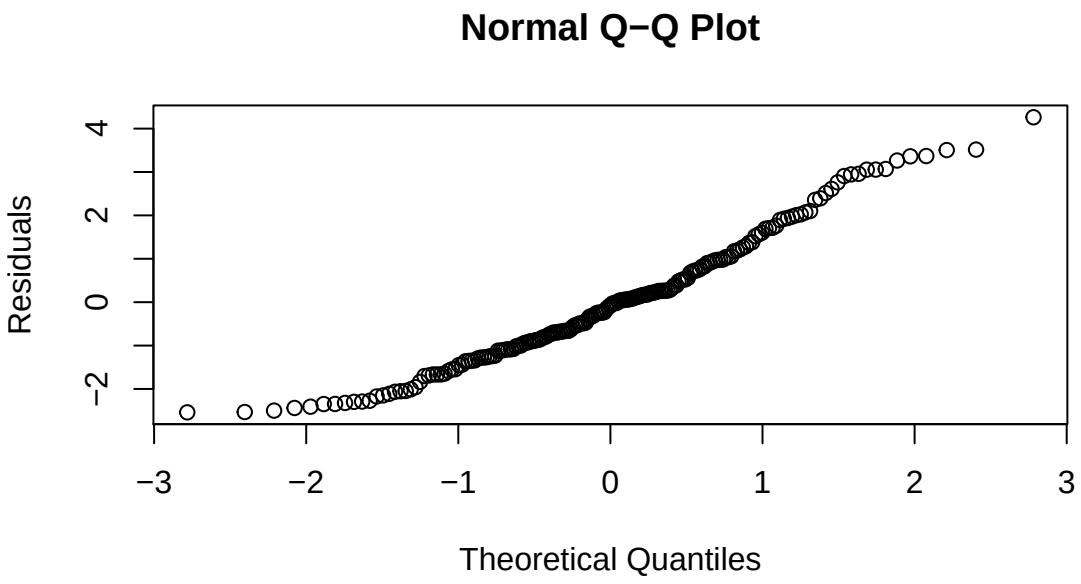

Figure 44: Normal Probability Plot

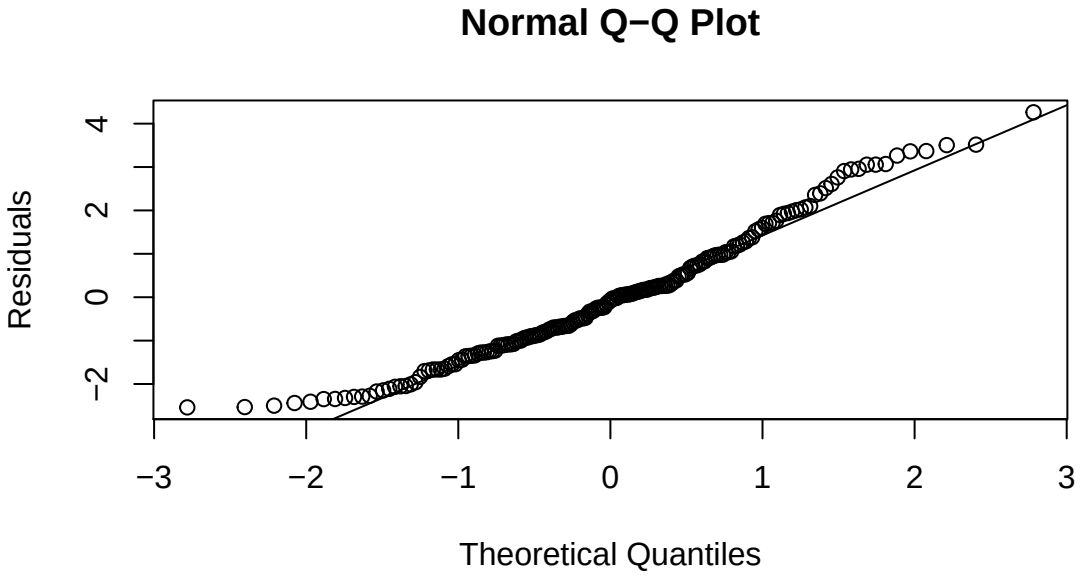

Figure 45: Normal Probability Plot

Once more, the linear relationship is not very apparent from the scatterplot of the fitted linear model. And seeing as how our residuals in the non-transformed model do not appear to follow a normal distribution as observed i the Normal QQ plot, the histogram of residuals and a Shapiro-Wilk p-value of 0.001, it would be inappropriate to carry out a t-test to determine whether our slope coefficient is different from zero.

Constructing a residual plot of the model:

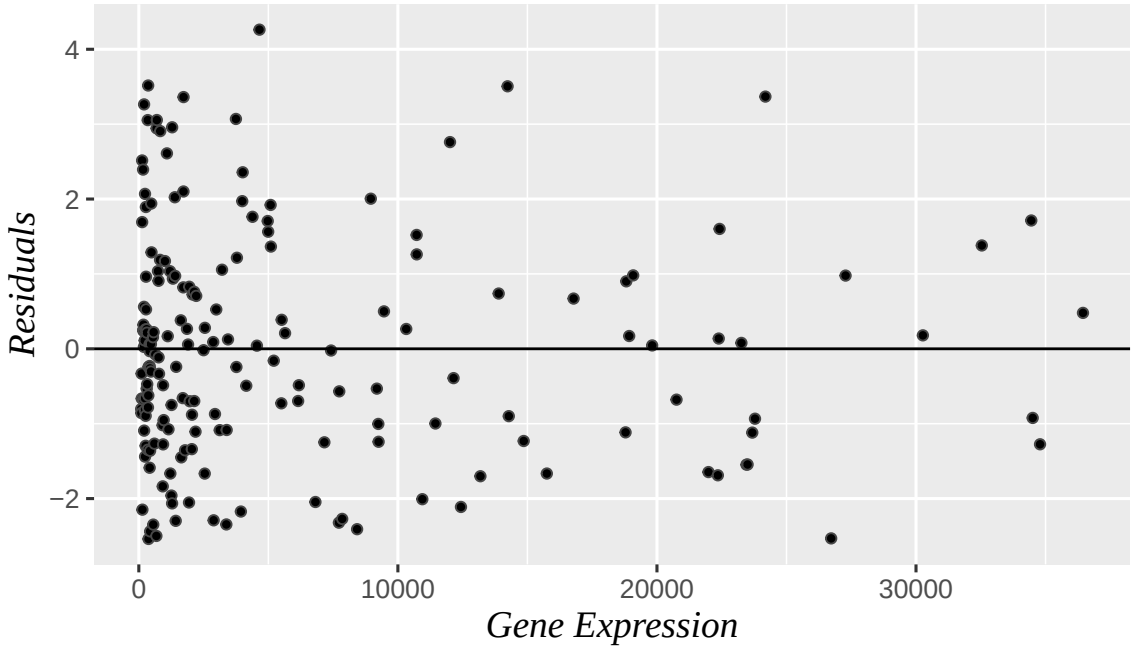

Figure 46: Residual Plot

From the residual plot, it doesn't seem like we have a problem with non-constant variance. To confirm we will perform the Breusch-Pagan and the Brown-Forsythe tests.

Table 14: Constancy of Variance Tests

BP test	BF test
2.75e-05	4e-07

From the results displayed in the table above and the residual plot, we cannot rule out non-constant variance for the model we are working on. Therefore, rather than applying a transformation as we had done earlier, we will attempt to fit a inverse polynomial model, since the relationship here seems to have a curvilinear shape that could be well approximated by an inverse polynomial model.

$$Y_i = \beta_0 + \frac{1}{X_i}.\beta_1 + \frac{1}{X_i^2}.\beta_2 + \epsilon_i$$

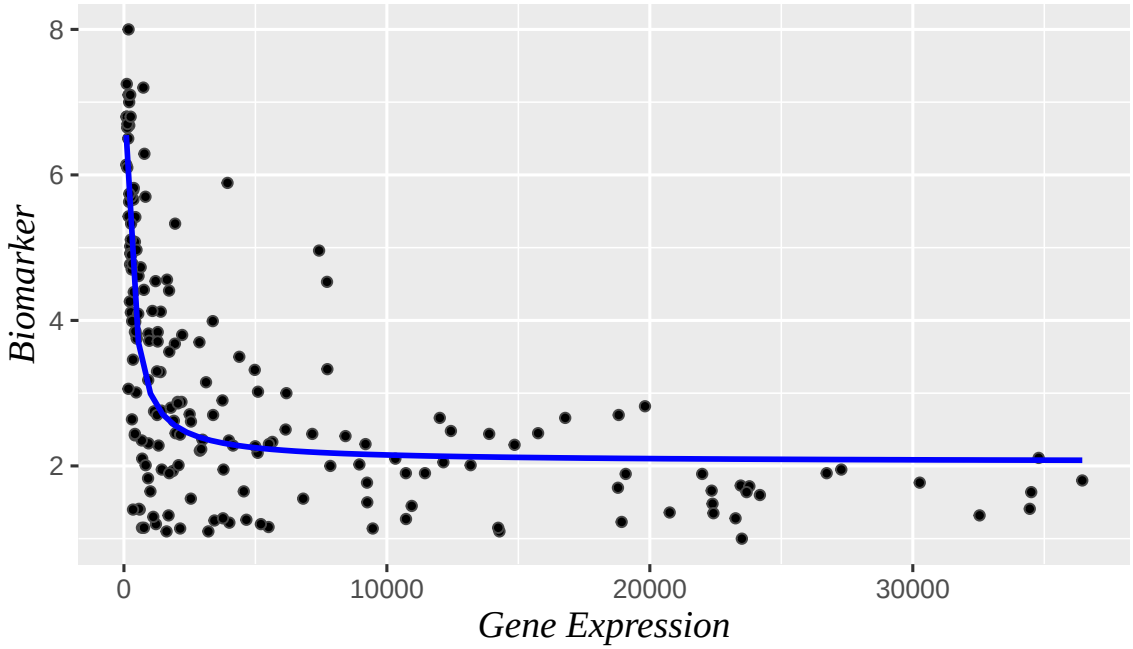

Figure 47: Scatterplot with fitted line

```
##
## ======================================================
##                              Dependent variable:
##                          ---------------------------
##                                    Biomarker
## ------------------------------------------------------
## 1/Gene Expression                  17.081***
##                                     (1.110)
##
## poly(1/GeneExpression, 2)2         -4.889***
##                                     (1.110)
##
## Constant                           3.175***
##                                     (0.082)
##
## ------------------------------------------------------
## Observations                          185
## R2                                   0.585
## Adjusted R2                          0.580
## Residual Std. Error            1.110 (df = 182)
## F Statistic                128.209*** (df = 2; 182)
## ======================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

Table 15: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	124.3031	124.303129	54.761	0
Residuals	183	415.3921	2.269902		

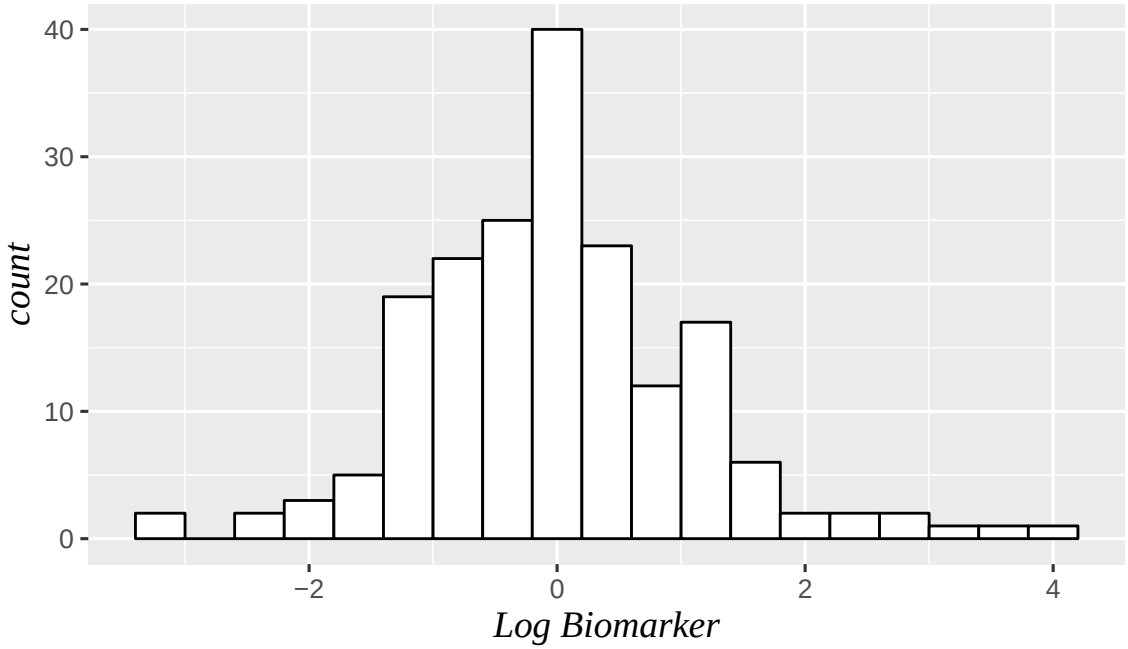

Figure 48: Histogram of Residuals

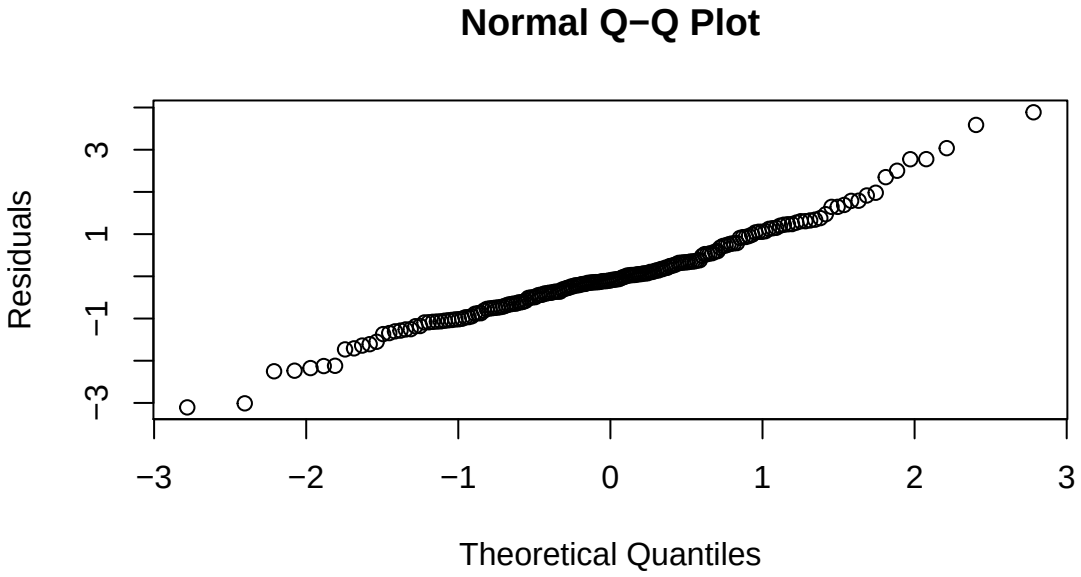

Figure 49: Normal Probability Plot

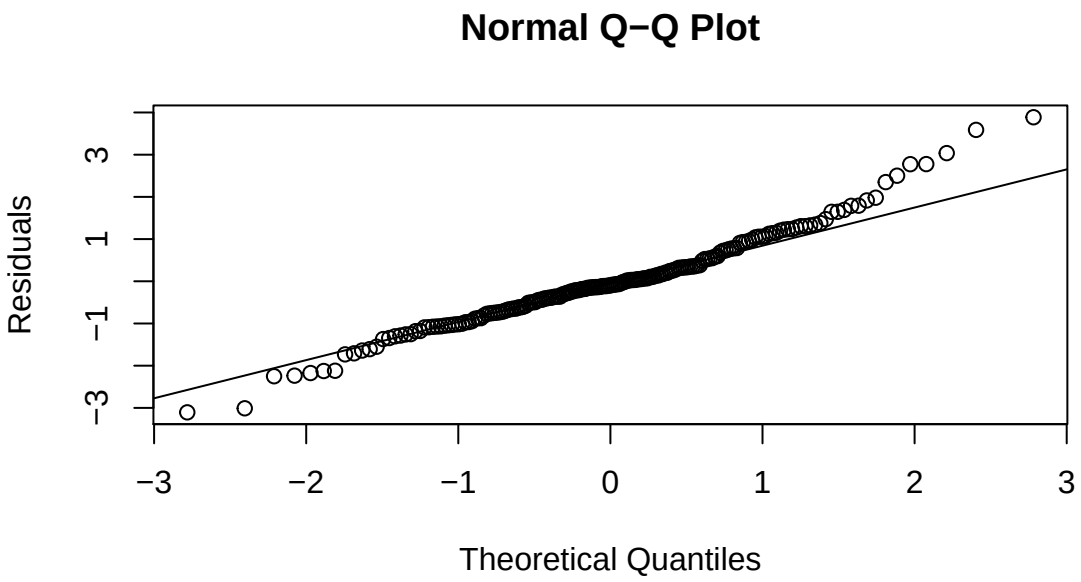

Figure 50: Normal Probability Plot

Our Polynomial model seems to offer a nice looking fit of the data. However, it seems the normality assumption of our residuals is violated, as evidenced by the heavy tail of the Normal QQ plot and the low p-value of the Shapiro-Wilk test 0.002. Next we will examine, the residual plot resulting from the above fit.

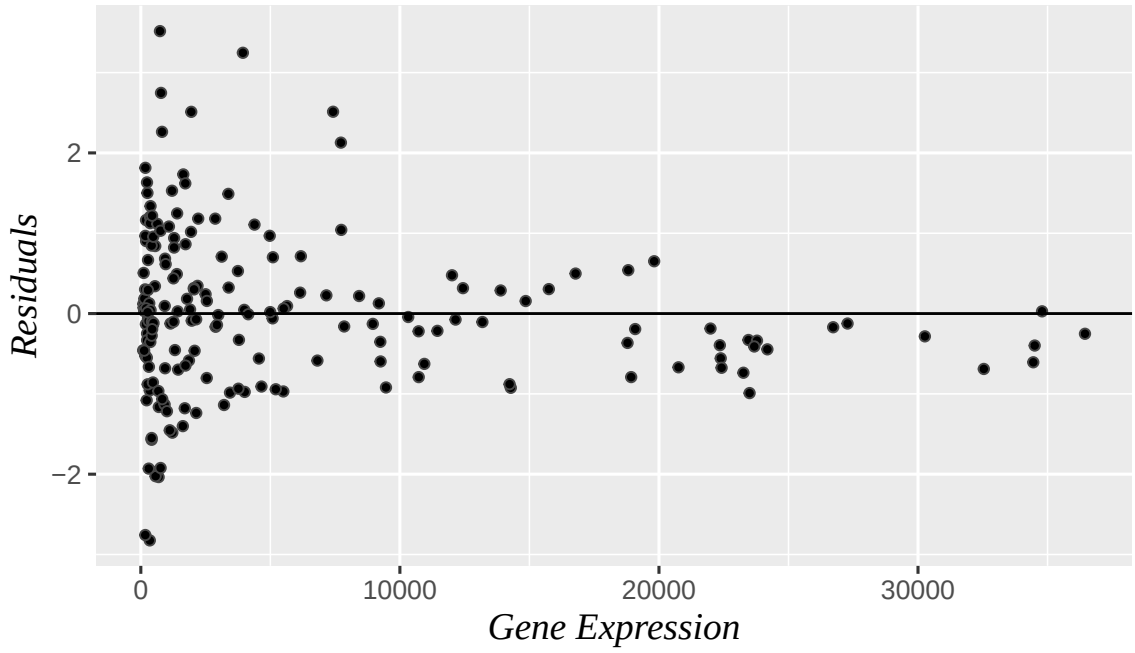

Figure 51: Residual Plot

From the residual plot, we can expect that our constant variance assumption will be violated. We verify this

by using the BP and BF tests.

Table 16: Constancy of Variance Tests

BP test	BF test
0.0347859	0.0509065

While the p-value here is higher than the problematic p-value we observed for the first model (which was remedied by weighted least squares), our tests point towards non-constant variance at the 5% level of type I error. This implies that the standard errors of our estimates may actually be higher than advertised. Therefore our constant variance and normality assumptions are violated despite the nice looking fit of the model. We will attempt to remedy this by weighting our observations using a similar weighting methodology to the one used in the earlier section.

```
##
## ================================================
##                              Dependent variable:
##                          ---------------------
##                                  Biomarker
## ------------------------------------------------
## 1/Gene Expression                18.100***
##                                   (1.123)
##
## poly(1/GeneExpression, 2)2       -5.900***
##                                   (1.134)
##
## Constant                         3.069***
##                                   (0.072)
##
## ------------------------------------------------
## Observations                        185
## R2                                 0.642
## Adjusted R2                        0.638
## Residual Std. Error          1.306 (df = 182)
## F Statistic                163.045*** (df = 2; 182)
## ================================================
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Table 17: ANOVA Table

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	1	124.3031	124.303129	54.761	0
Residuals	183	415.3921	2.269902		

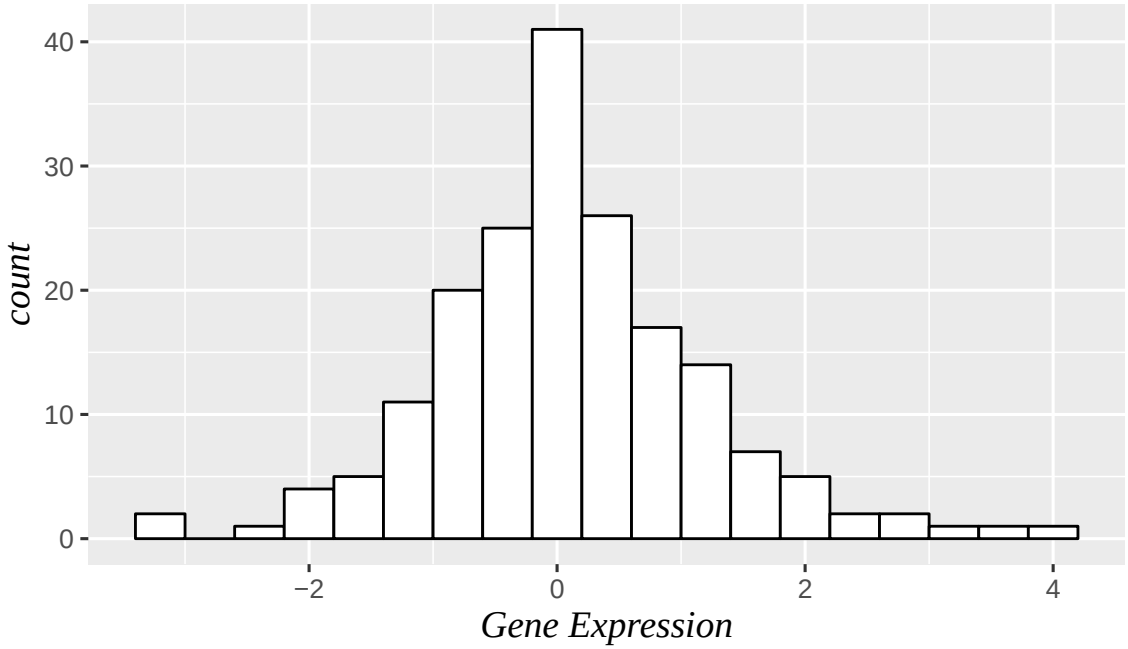

Figure 52: Histogram of Residuals

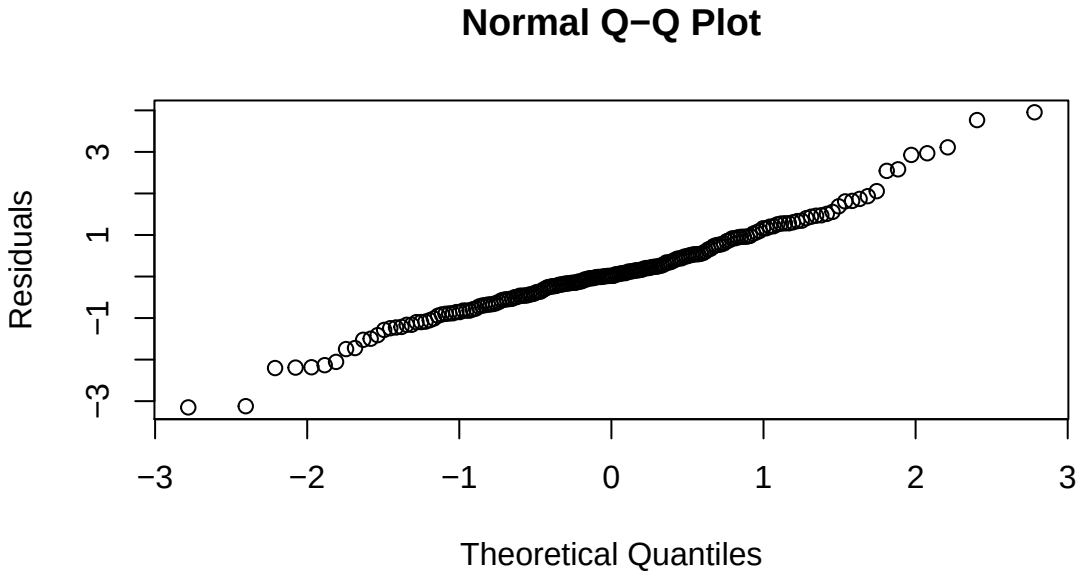

Figure 53: Normal Probability Plot

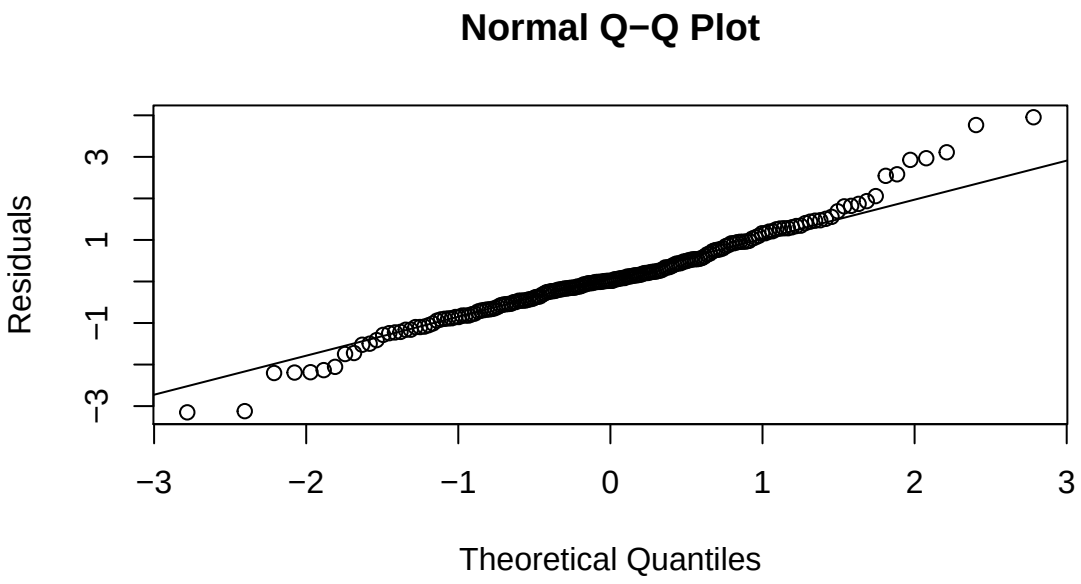

Figure 54: Normal Probability Plot

Now, we have a new weighted least squares model from the inverse polynomial model, and while the normality assumption does not hold (Shapiro-Wilk p-value of 0.001), the departure from normality does not seem to be great by observing the residual histogram and the Normal QQ plot. Before we begin to conduct our constant variance tests, we will first check that this is in fact the best polynomial model among all possible models of the inverse polynomial form we have proposed by presenting some model selection criteria for the 1st up to the 5th degree inverse polynomial models (without applying our weights):

Table 18: Best Subsets with Multiple Criteria

Intercept	X1	X2	X3	X4	X5	R.sq.adj	Cp	AIC	BIC
1	1	1	0	0	0	0.5803092	1.830442	-153.4254	-146.9847
1	1	1	1	0	0	0.5780236	3.816249	-151.4399	-141.7788
1	1	1	1	1	1	0.5775948	6.000000	-149.3076	-133.2058
1	1	1	1	1	0	0.5773554	5.102020	-150.1721	-137.2907

Now that we have determined our model is the best among this family of models (as displayed by **ALL** criteria) we proceed with carrying out our constancy of variance tests:

```
##
##  studentized Breusch-Pagan test
##
## data:  lmfit2poly_wls
## BP = 6.7171, df = 2, p-value = 0.03479
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  2.4007  0.123
##       183
```

Table 19: Constancy of Variance Tests

BP test	BF test
0.0347859	0.1230114

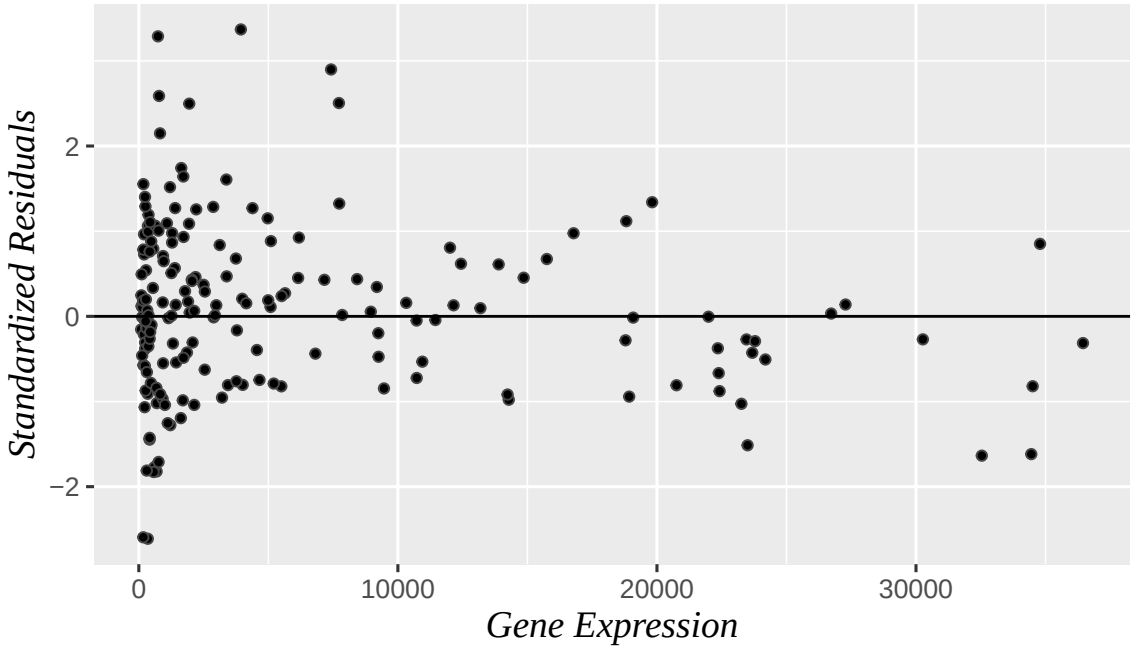

Figure 55: Residual Plot

```
##
##  Runs Test - Two sided
##
## data:  residuals(lmfit2poly_wls)
## Standardized Runs Statistic = 0.81137, p-value = 0.4172
```

Our results regarding the constant variance assumption bare inconclusive, with the BF test failing to reject the null hypothesis, and the BP test barely rejecting the null hypothesis. Therefore we cannot claim to have constant variance, however, we will proceed with predicting the value of Biomarker from Gene Expression while cautioning that due to the violation of our assumptions, our results may not be accurate as the prediction interval functions best with an assumption of Normality, and the standard errors of the coeffecients we used to predict Biomarker may be inaccurate. Before we present our predicted values, we will present confidence intervals for our estimated coefficients:

Table 20: Confidence Interval for Coefficients

	Lower Bound CI	Upper Bound CI
1/Gene Expression	15.885283	20.31487
1/(Gene Expression^2)	-8.137351	-3.66299

However, seeing as how our assumption of normality is not supported by evidence from our statistical tests, we will also present bootstrapped confidence intervals. We observe that for our first predictor variable, the bootstrap confidence intervals are very close to those obtained under the assumption of normality. For our second covariate, the upper confidence bound is noticeably larger.

Table 21: Bootstrap Confidence Interval for Coefficients

	Lower Bound CI	Upper Bound CI
1/Gene Expression	13.603928	20.107324
1/(Gene Expression^2)	-7.962349	-1.062106

Table 22: Prediction and Prediction Interval

	Predicted	Lower Bound PI	Upper Bound PI
Biomarker	6.572598	3.78451	9.360686

Conclusion:

Throughout the course of this paper we have fitted four different models: a log-transformed model, a weighted least squares log-transformed model, an inverse polynomial model, and finally, a weighted least squares inverse polynomial model. We mainly faced problems regarding the violation of the constant variance assumption; we have attempted to remedy these violations, though not all models were equally successful. Our weighted least squares inverse polynomial model provided an 11% increase in explained variability over the weighted least squares log-transformed model, however, tests of the constancy of the variance of its residuals were inconclusive. We have proposed these two models to account for the shape of the observed data, and their suitability is evidenced by the increase in explained variability over the simple linear model, as well as their maintenance of some critical assumptions in some cases. With regards to the transformed model, we believe it utilizes the most appropriate transformation in this case due to the evidence provided by the Box-Cox methodology. For the polynomial model, we believe the model captures the inverted curviliear shape of the data by its inclusion of the inverse square term, which, if the normality assumption holds, has a coefficient that is, with high probability, not equal to zero. It also happens to provide the best selection criteria for this family of models. Therefore we argue that in this case, the models presented here, may provide one of the best methods for understtanding the relationship between Biomarker value of colorectal cancer and Gene Expression value.