

# 1 Non-normal and Heteroscedastic Errors (Soc 504)

The Gauss-Markov Theorem says that OLS estimates for coefficients are BLUE when the errors are normal and homoscedastic. When errors are nonnormal, the ‘E’ property (Efficient) no longer holds for the estimators, and in small samples, the standard errors will be biased. When errors are heteroscedastic, the standard errors become biased. Thus, we typically examine the distribution of the errors to determine whether they are normal.

Some approaches to examining nonnormality of the errors include constructing a histogram of the error terms and constructing a Q-Q plot (Quantile-Quantile plot). Certainly, there are other techniques, but these are two simple and effective methods. Constructing a histogram of errors is trivial, so I don’t provide an example of it. Creating a Q-Q plot, on the other hand, is a little more difficult. The objective of a Q-Q plot is to compare the empirical error distribution to a theoretical distribution (i.e., normal). If the errors are normally distributed, then the empirical and theoretical distributions will look identical, and a scatterplot of the two will fall along a straight line. The observed errors are obtained from the regression analysis. The steps to computing the theoretical distribution are as follows:

1. Order the errors from smallest to largest so that  $X_1 < X_2 < \dots < X_n$ .
2. Compute the statistic:  $CDF_{empirical}(i) = \frac{i - \frac{1}{2}}{n}$ , where  $i$  is the rank of the empirical error after step 1. This gives us the empirical  $CDF$  for the data.
3. Compute the inverse of this  $CDF$  value under the theoretical distribution. So take  $z_i = \Phi^{-1}(CDF_{empirical}(i))$ .
4. Plot  $X$  against  $z$ .

The figure below provides an example of this plot. I simulated 100 observations (errors) from a  $N(5, 2)$  distribution:

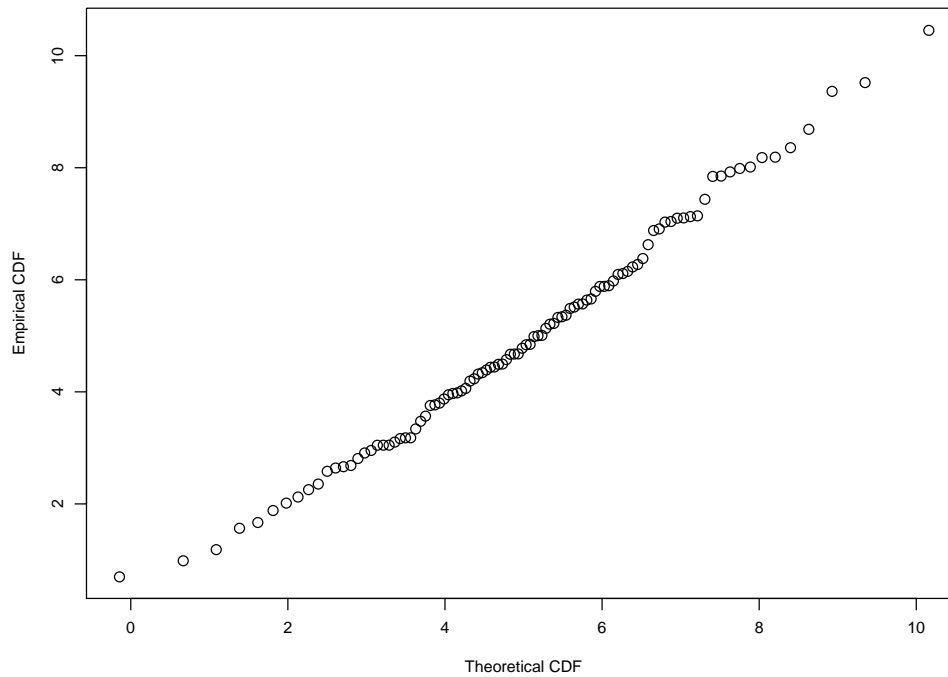


Figure 1. Q-Q Plot for 100 random draws( $X$ ) from a  $N(5,2)$  distribution.

Notice that in this case, the distribution clearly falls on a straight line, indicating the errors are probably normally distributed. The following figure is a histogram of  $\exp(X)$ . This distribution is clearly right skewed.

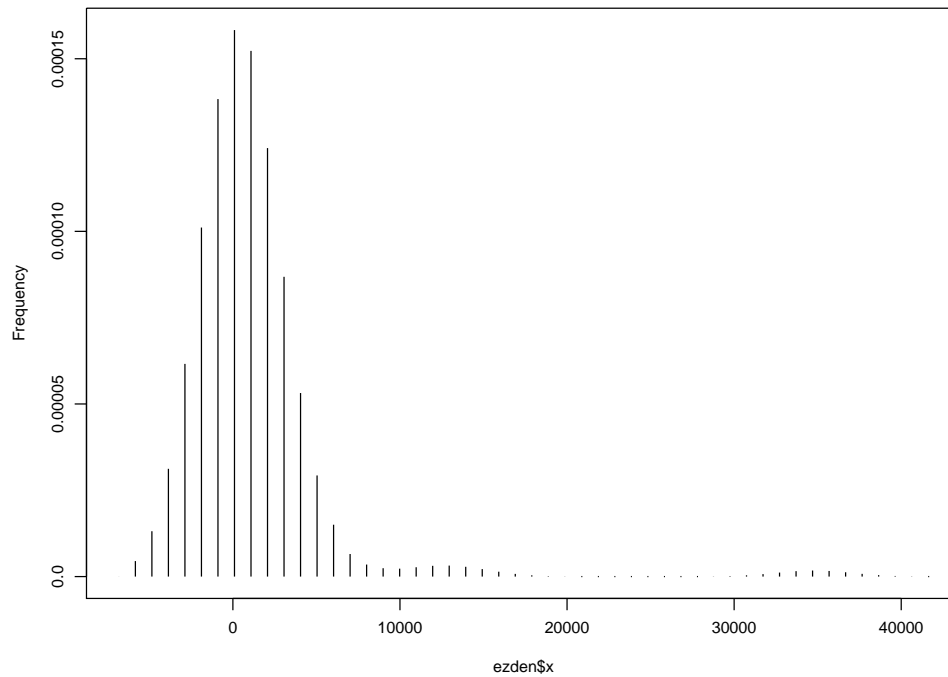


Figure 2. Histogram of  $\exp(X)$  from above.

The resulting Q-Q plot below reveals this skew clearly. The scatterplot points are not on a straight line. Rather, the plot follows a curve. The plot reveals that there is too much mass at the far left end of the distribution relative to what one expect if the distribution were normal.

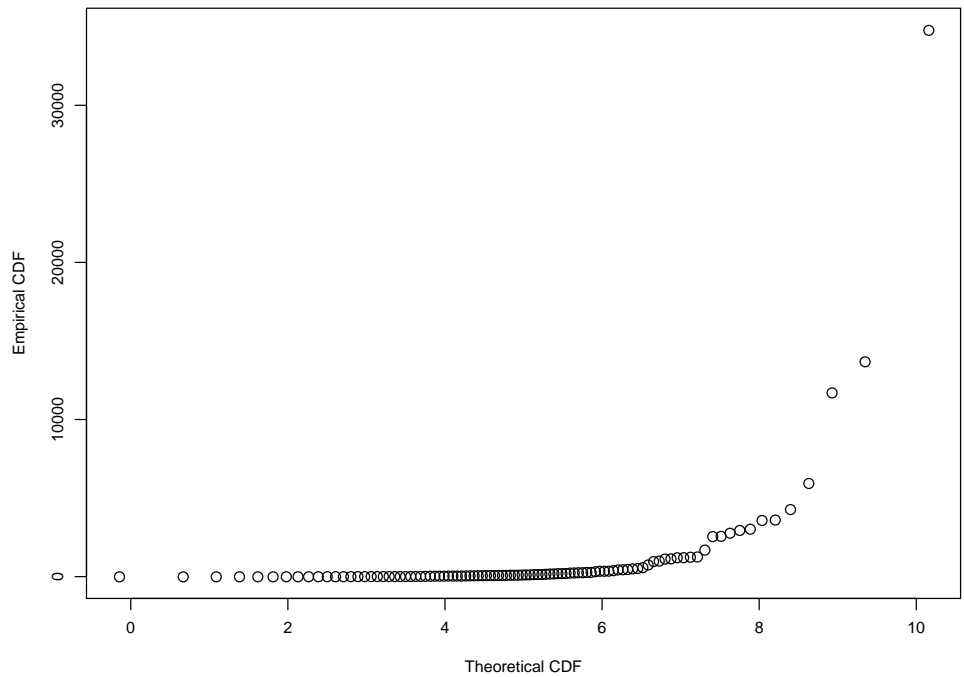


Figure 3. Q-Q Plot for  $\exp(X)$ .

The next histogram shows the distribution for the distribution of  $\ln(X)$ , which has a clear left skew.

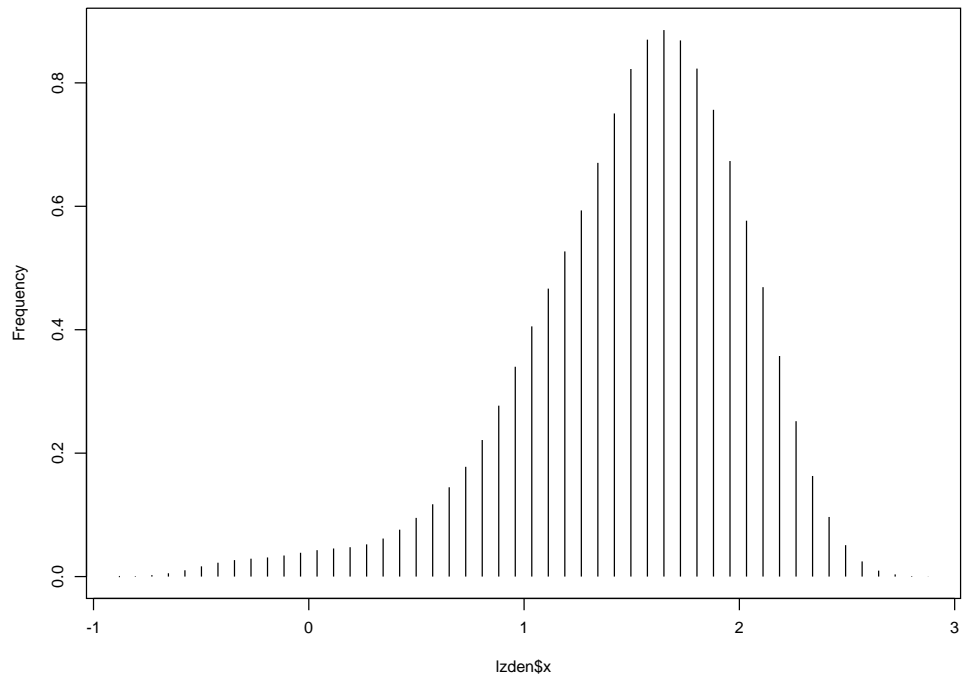


Figure 4. Histogram of  $\text{Log}(X)$ .

The resulting Q-Q plot also picks up this skew, as evidenced by the curvature at the top edge of the plot. This curvature indicates that there is not enough cumulative mass in the middle of the distribution relative to what would be expected under a normal distribution.

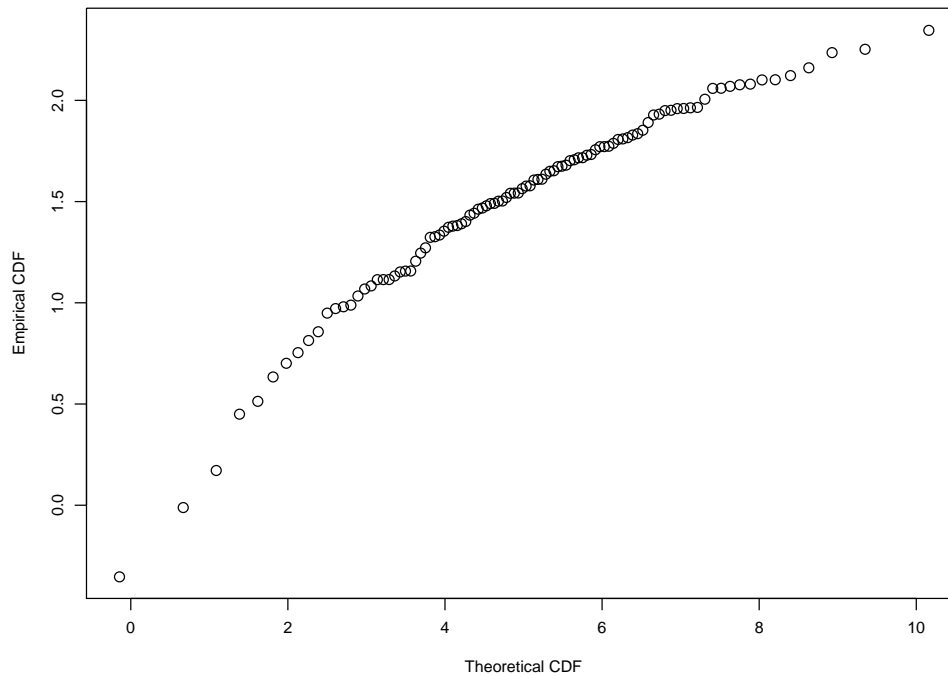


Figure 5. Q-Q Plot of  $\text{Log}(X)$ .

## 1.1 Heteroscedasticity

The above plots are helpful diagnostic tools for nonnormal errors. What about heteroscedasticity? Heteroscedasticity means non-constant variance of the errors across levels of  $X$ . The main consequence of heteroscedasticity is that the standard errors of the regression coefficients can no longer be trusted. Recall that the standard errors are a function  $\sigma_e$ . When  $\sigma_e$  is not constant, then, the typical standard error formula is incorrect.

The standard technique for detecting heteroscedasticity is to plot the error terms against either the predicted values of  $Y$  or each of the  $X$  variables. Below is a plot of errors that evidences heteroscedasticity across levels of  $X$

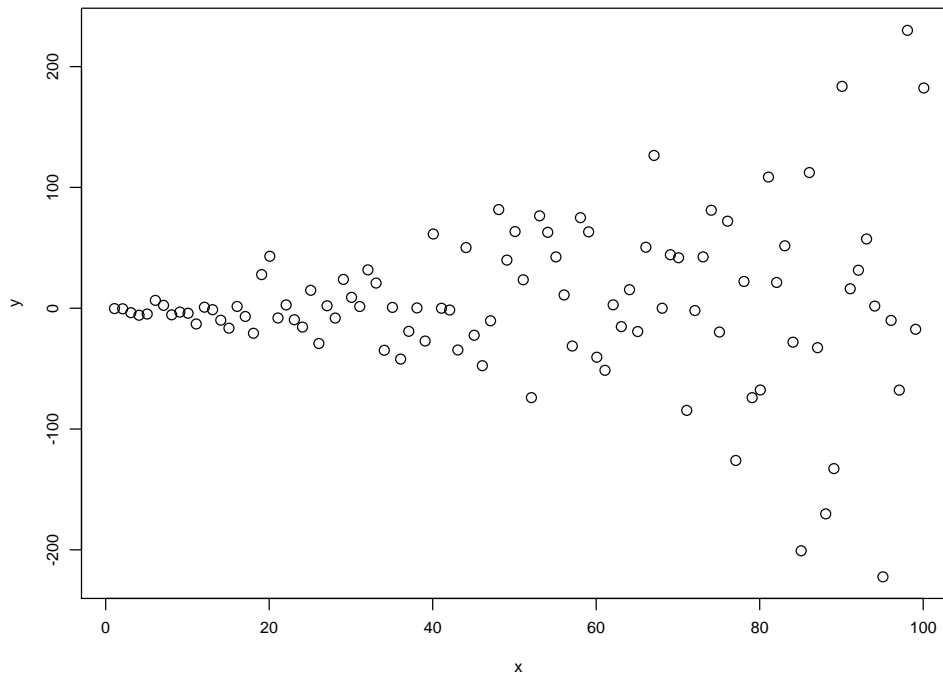


Figure 6. Example of Heteroscedasticity.

The plot has a classic ‘horn’ shape in which the variance of the errors appears to increase as  $X$  increases. This is typical of count variables, which generally follow a poisson distribution. The poisson distribution is governed by one parameter,  $\lambda$  (density is  $p(X) = \frac{1}{X!} \lambda^X \exp(-\lambda)$ ), which is both the mean and variance of  $X$ . When  $X$  is large, it implies that  $\lambda$  is probably large, and that the variance of  $X$  is also large. Heteroscedasticity is not limited to this horn-shaped pattern, but it often follows this pattern.

## 1.2 Causes of, and Solutions for, Nonnormality and Heteroscedasticity

Nonnormal errors and heteroscedasticity may be symptoms of an incorrect functional form for the relationship between  $X$  and  $Y$ , an inappropriate level of measurement for  $Y$ , or an omitted variable (or variables). If the functional form of the relationship between  $X$  and  $Y$  is misspecified, errors may be nonnormal because there may be clusters of incredibly large errors where the model doesn’t fit the data. For example, the following plot shows the observed and predicted values of  $Y$  from a model in which an incorrect functional form was

estimated.

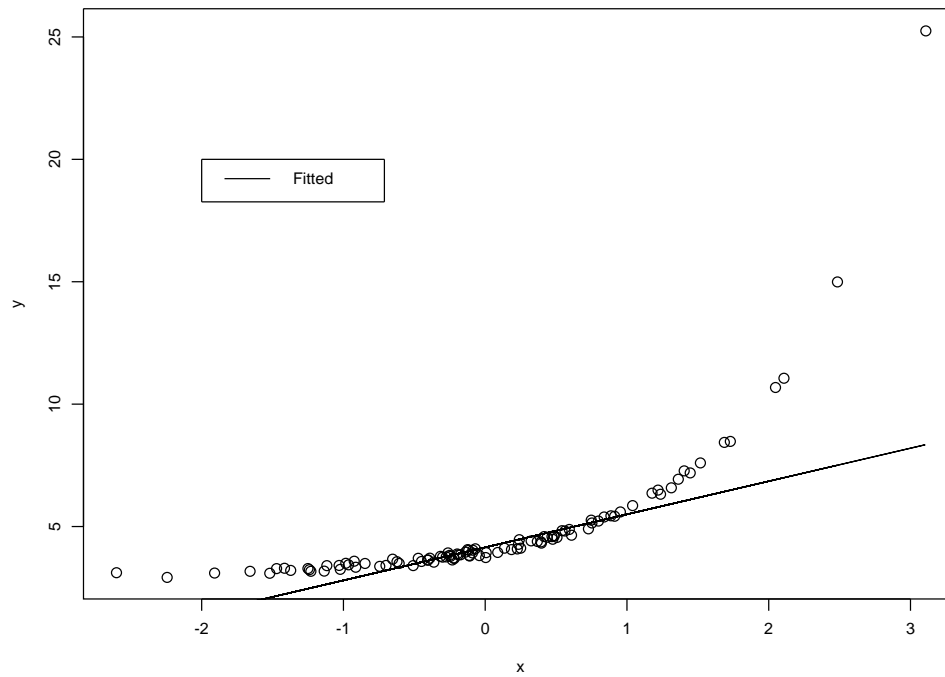


Figure 7. Observed and (Improperly) Fitted Values.

In this case, the correct model is  $y = b_0 + b_1 \exp(X)$ , but the model  $y = b_0 + b_1 X$  was fitted. The histogram of the errors is clearly nonnormal:

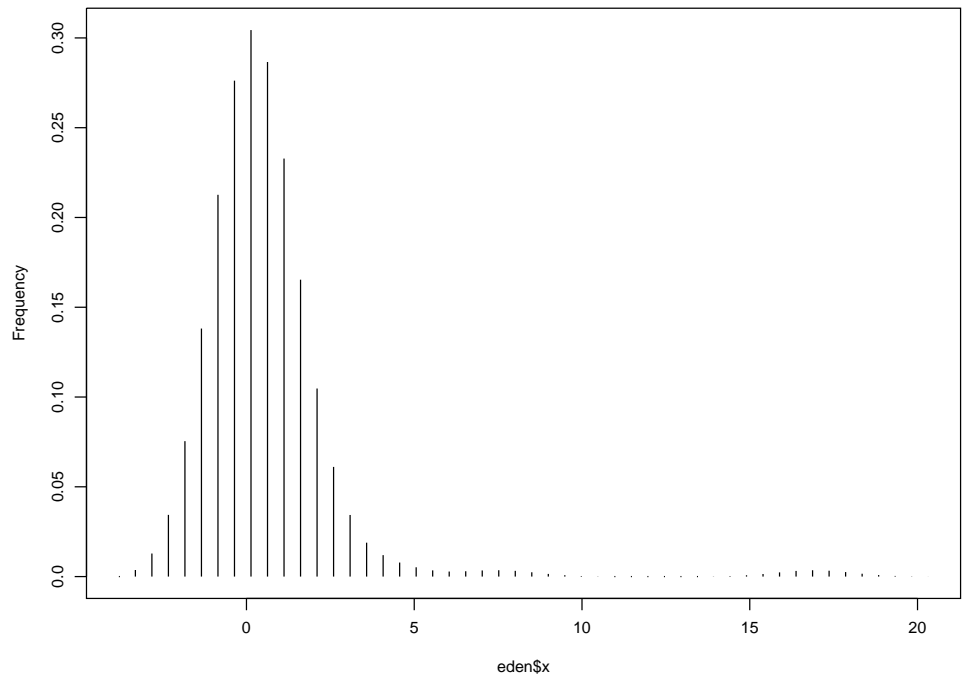


Figure 8. Histogram of Errors from Example.

and the Q-Q plot reveals the same problem:

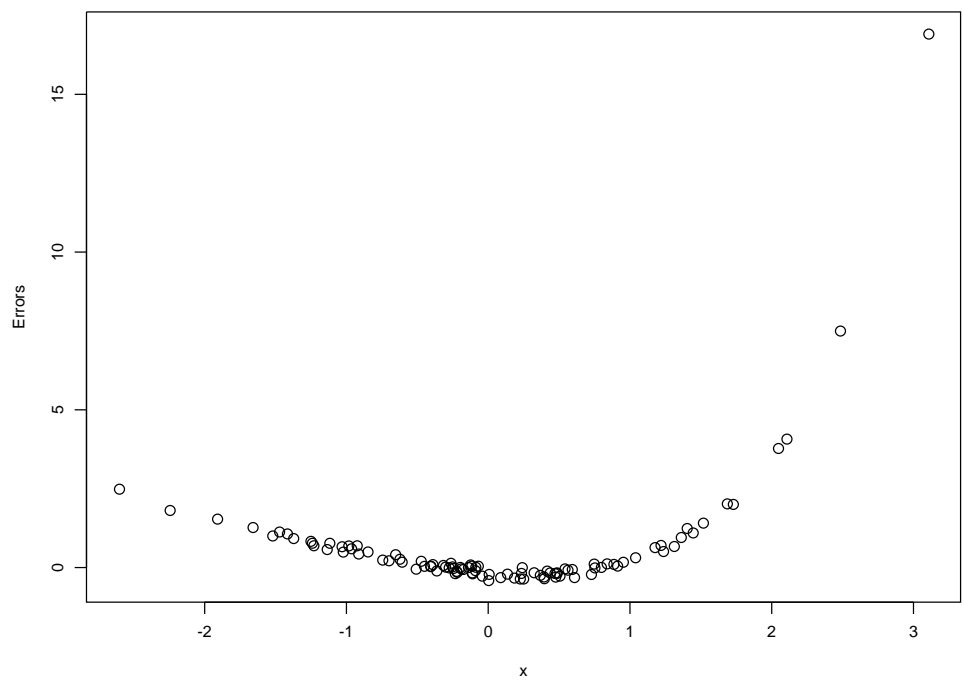


Figure 9. Q-Q Plot of Errors in Example.

A plot of the errors against the X values is:

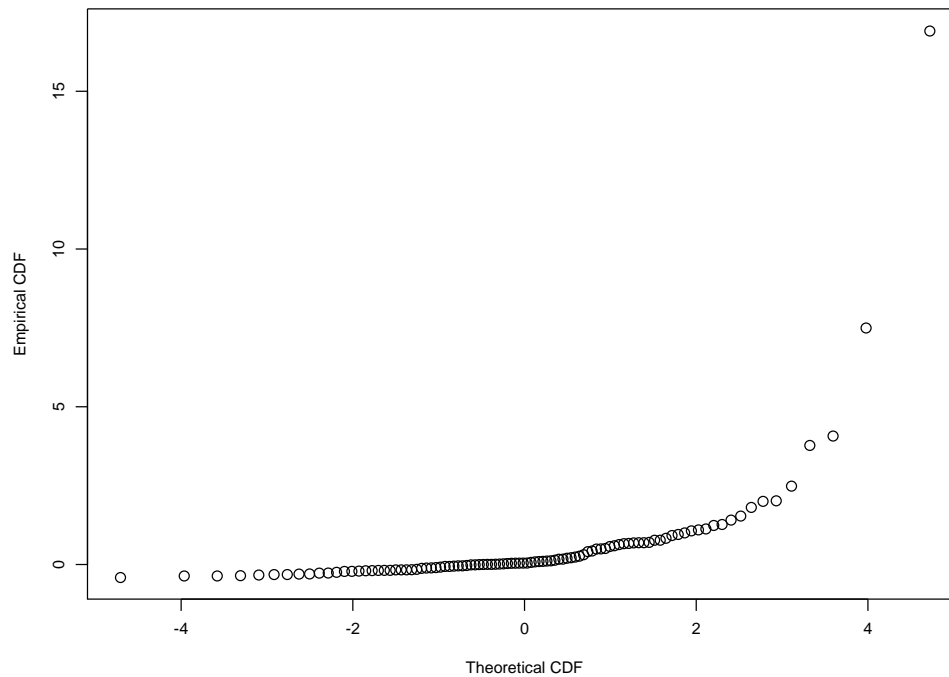


Figure 10. Plot of Errors against X in Example.

This plot suggests that heteroscedasticity is not a problem, because the range of the errors doesn't really appear to vary across levels of  $X$ . However, the figure shows there is clearly a problem with the functional form of the model—the errors reveal a clear pattern. This implies that we may not only have the wrong functional form, but we also have a problem with serial autocorrelation (which we will discuss in the context of time series analyses later).

If we do the appropriate transformation, we get the following.

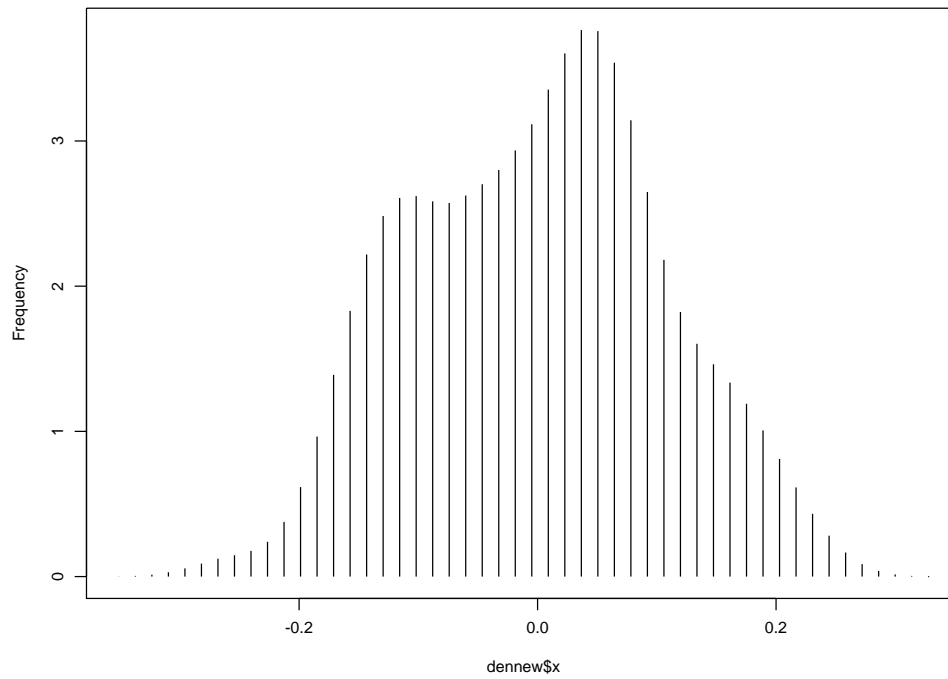


Figure 11. Histogram of Errors in Revised Example.

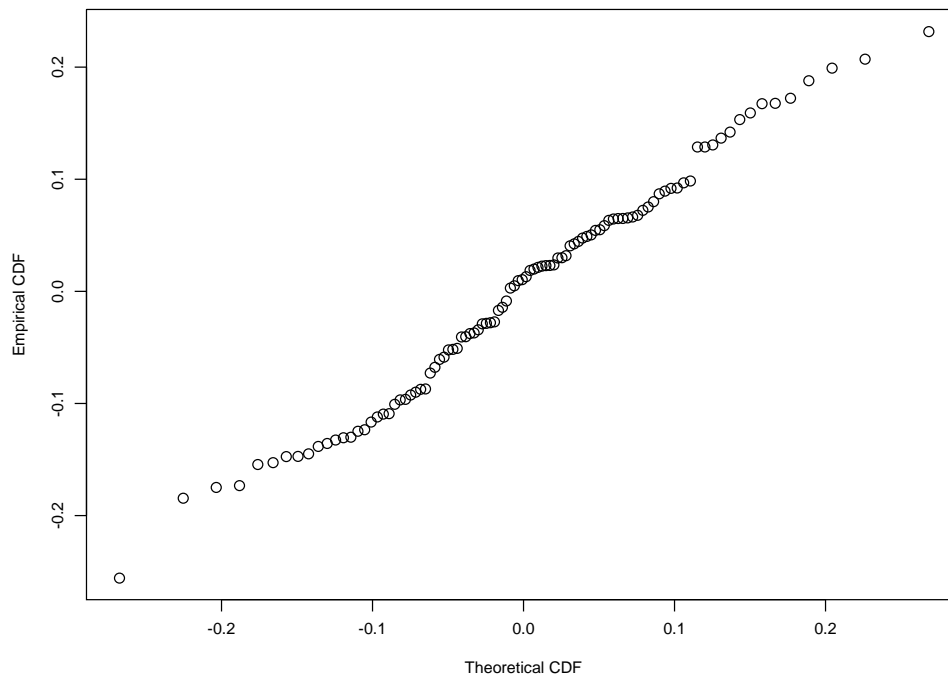


Figure 12. Q-Q Plot of Errors in Revised Example.

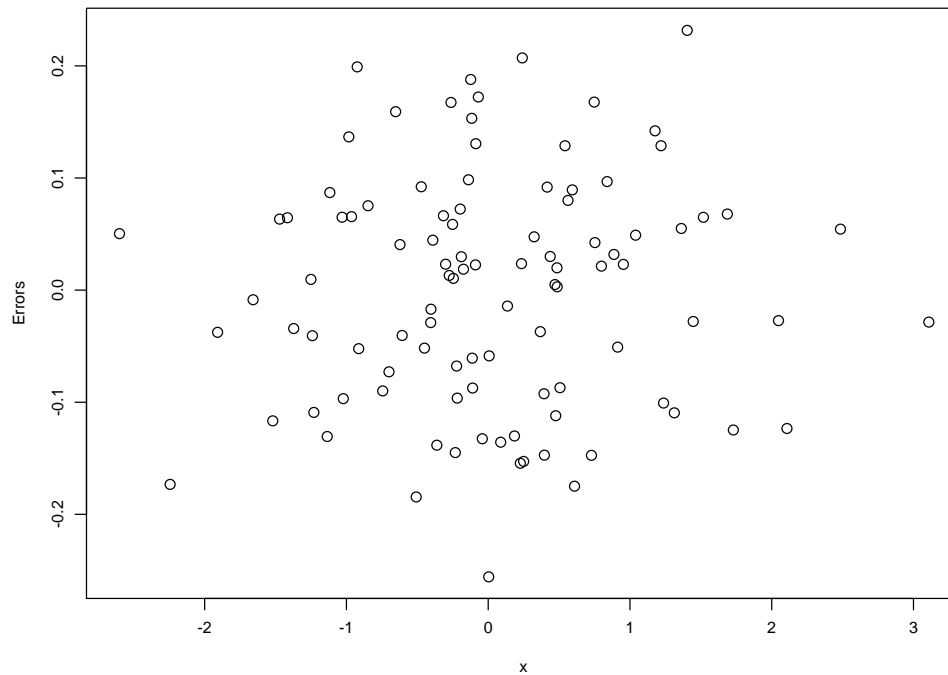


Figure 13. Plot of Errors Against X in Revised Example.

These plots show clear normality and homoscedasticity (and no autocorrelation) of the errors.

In other cases, a simple transformation of one or more variables may not be sufficient. For example, if we had a discrete, dichotomous outcome, the functional form of the model would need to be significantly respecified. Nonnormal and heteroscedastic errors are very often the consequence of measurement of the dependent variable at a level that is inappropriate for the linear model.

As stated above, another cause of heteroscedasticity is omitting a variable that should be included in the model. The obvious solution is to include the appropriate variable. An additional approach to correcting for heteroscedasticity (weighted least squares) will be discussed soon.