Linear Modeling:

The Normality Assumption

Carly Lundgreen

December 10, 2020

Motivation

When fitting a linear model, multiple assumptions about the nature of the data, normally understood in the form of model diagnostics, are made. One of these assumptions is that the calculated residuals from the linear model are normally distributed. We are interested in discovering how a violation of this normality condition may affect the model's average prediction error, the width of 95% confidence intervals, and the proportion of these intervals that would actually contain the parameter of interest–which we refer to as the confidence interval coverage.

Data Introduction

The National Cancer Institute reports that the average cost of a pack of cigarettes in the United States is $6.28. This price tag implies that someone with a pack-a-day habit is set back $188 every month, or $2,292 per year, on average. This is not a small cost–some used cars can be purchased with that sum of money. For others, it could equate to multiple months of rent or a downpayment on a house or apartment.The dataset that we will discuss reports the monthly income (`Income`) of 67 daily smokers and the dollar amount they spent the previous month on cigarettes (`Sales`). As part of the analysis, we fit the following linear model to these data:

$$Sales = \beta_0 + \beta_1 Income + \epsilon$$
$$\epsilon \sim \ N(0, \sigma^2)$$

Assessing Normality

The model assumptions must be investigated before one proceeds with an analysis of the effect of `Income` on `Sales` (β_1). Our interest is in the normality assumption in particular, and we will approach an analysis of this assumption in two ways. One way to get a quick idea of the shape of the residuals' distribution is to simply plot a histogram. Oftentimes, when investigating model assumptions, it can be helpful to standardize the residuals before plotting. Standardization estimates the error associated with a particular observation while considering the leverage of that point in the calculation.

Instead of standardizdation, we may want to calculate R-Studentized residuals. R-Studentized residuals scale the difference in observation and prediction by dividing by the standard deviation computed without that particular observation. Studentized residuals are useful in the identification of outliers and have the property that the studentized residual for each point will follow a t-distribution if the normality assumption of the SLR model is met. This analysis is not particularly interested in identifying outliers, and we are testing whether the normality assumption is met with this data, so we will calculate standardized residuals. A histogram of these residuals is shown below.

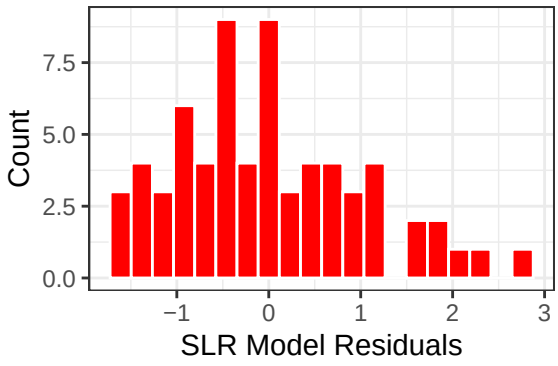

Figure 1: Histogram of Standardized Residuals

We can see that there does appear to be a right-skew to the residuals, so we can infer that the normality assumption is likely violated, but generally a histogram itself is not a perfect diagnostic for a departure from normality in the residuals. Hypothesis testing is another method to to evaluate this condition.

There are multiple hypothesis tests that may be used to evaluate the normality assumption. One test, the Kolmogorov-Smirnov (KS) test, can be used to compare the empirical (data) CDF (or residuals in our case) to the CDF of a normal distribution. The null hypothesis test of the KS test is that the data follow a normal distribution, and the alternative is that they do not follow a normal distribution. Another hypothesis test that can be performed is the Shapiro-Wilk (SW) test. This test evaluates whether a sample follows a normal distribution specifically. Monte Carlo simulations have shown that the SW test is generally more powerful than the KS test for normality, so we will opt for an SW test here.

The null hypothesis of the Shapiro-Wilk test is that the sample follows a normal distribution (the sample will be the residuals in this case). The p-value of the SW test was found to be 0.039, and at our significance level of 0.05, we reject the null hypothesis that the residuals are normally distributed. However, like most hypothesis testing procedures, if the sample size is very large this test may detect even trivial departures from the null hypothesis, so one more diagnostic by which to identify the shape of the residuals is a Q-Q plot. The Q-Q plot displays two sets of quantiles against one another. In our case, because we are evaluating a normal assumption, if both sets of quantiles came from a normal distribution, then the points should follow (roughly) a straight line.

The Q-Q plot for this data is displayed below. 95% confidence interval bands are included for reference.

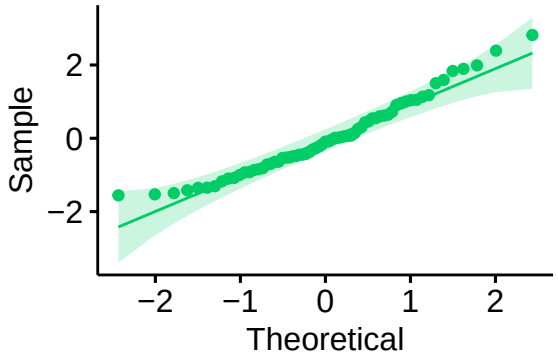

Figure 2: Q-Q plot for Standardized Residuals

We can see some curvature in the Q-Q, which is a departure from roughly straight line that would be seen if the normality condition was met. Though the standardized residuals do not fall out of the confidence interval

bounds, we can still see that the scatterplot does not form a sufficiently straight line. Due to the histogram of the residuals, the result of the Shapiro-Wilk test for normality, and the Q-Q plot, we can assume that the normality condition is not sufficiently met by these data.

Simulation Study

We conducted a simulation study in order to investigate the effects of a normality assumption violation on confidence interval width, interval coverage, and root predictive mean squared error (RPMSE). As an overview, we generated n random values from 3 different skewed distributions, in addition to a normal distribution, for three different sample sizes (n): 10, 100, and 1000. These datapoints were utilized as the errors in a fitted SLR model. We calculated interval width and coverage, as well as RPMSE, and compared these metrics for the different distributions and sample sizes. The distributions that we compared were:

- Normal with $\mu = 1$ and $\sigma^2 = 1$
- Weibull with $\lambda = 1$ and $k = 1$ where λ is the scale parameter and k is the shape parameter
- Gamma with $k = 1$ and $\theta = 1$, where k is the scale parameter and θ is the shape parameter
- Exponential with $\lambda = 1$, where λ is the rate parameter

For a visual reference, we generated 500 values from each of these distributions and created histgorams for comparison. We can see that the Weibull, Gamma, and Exponential distributions all have a significant right-skewness. By using each of these distributions as error distributions within an SLR model, we can effectively compare the effects of a normality assumption violation on the metrics we have chosen to analyze.

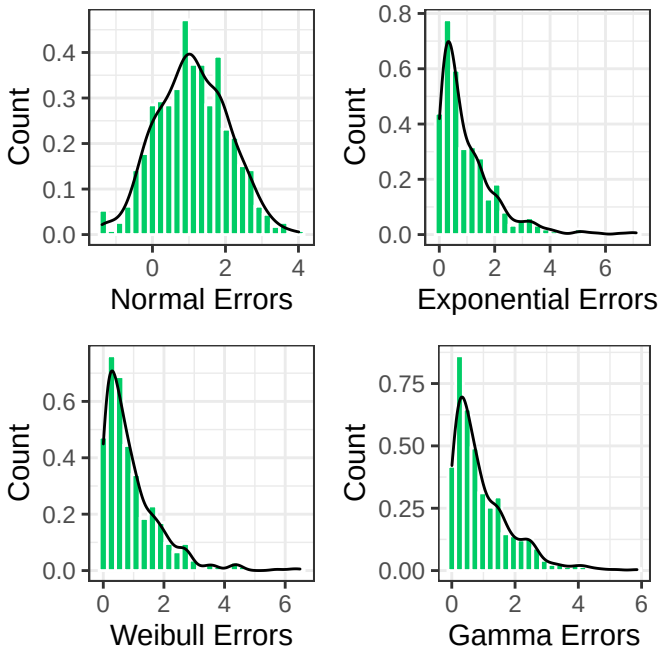

Figure 3: Histograms for 500 draws from the Comparison Error Distributions

The shape/scale/rate parameters for the skewed distribution examples were carefully selected so that the mean and variance of each of the distributions was equal to the mean and variance of the comparison Normal distribution. In order to calculate confidence intervals and the resulting width and coverage, as well

as RPMSE, we fit the following simple linear model for each of the above error distributions, using the n datapoints as the ϵ vector in the SLR model outlined below.

$$y = X\beta + \epsilon$$

We have set the X matrix and β vector to be:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

For the study, we first created a function,`setup`, that takes only n, the sample size, as an argument. The function first generates n random datapoints from the distributions outlined above and combine them in an Nx4 matrix. Then, it creates the β vector and X matrix as outlined above. It then calculates y-values for each of the 4 distributions according to the SLR model $y = X\beta + \epsilon$, as well as $\hat{\beta}$ according to $\hat{\beta} = (X'X)^{-1}X'y$ for each of the error distributions. Within the function, another function is created that calculates 95% confidence interval bounds for β_1, the effect of `Income` on `Sales`. The interval is calculated according to the formula

$$C\beta = C\hat{\beta} \pm t\sqrt{s^2C(X'X)^{-1}C'}$$

Where C is a 1x2 vector with respective elements 0 and 1, so that $C\beta = \beta_1$. Finally, the function calculates each of the $\hat{y}$ (prediction) vectors for the four error distributions according to $\hat{y} = X\hat{\beta}$ and returns the predictions, y-values, and 95% confidence intervals for β_1.

We then created another function, `replicate_sim`, which takes as input the output from the previous function as well as an argument `Nsim`, the number of simulations/draws from each of the above error distributions. This function accomplishes the following:

- Replicates the `setup` function `Nsim` times to generate `Nsim` $\hat{y}$ vectors of length N, `Nsim` y vectors of length n, and `Nsim` intervals for β_1
- Extracts the `Nsim` interval bounds for each of the distributions used in the error terms and calculates the coverage of the intervals according to whether 2 (our chosen value of β_1) was included within the interval.
- If 2 was contained in the interval, an alternate vector `is_in` was assigned a 1, and if not, `is_in` was assigned a zero. The mean of `is_in` was then calculated to return the proportion of intervals that contain the true value of β_1
- Calculates the average width of intervals by returning, for each of the `Nsim` intervals (for each of the four distributions), the difference between the upper bound and lower bound of the interval.
- The mean of these differences was then calculated to return the average width of the 95% confidence intervals.
- For each of the four distributions, the `Nsim` $\hat{y}$ and y vectors are used to calculate `Nsim` RPMSE values according to the following formula:

$$RPMSE \ = \ \sqrt{\frac{1}{n}\sum_{i=1}^{Nsim}(y_i - \hat{y}_i)^2}$$

These two functions were then utilized to find the interval width, coverage, and RPMSE values for each of the sample sizes we are investigating (n=10, 100, and 1000). We ran the `setup` function three times, once for each sample size. Then, saving the output from each of the function calls, we ran the `replicate_sim` function using `Nsim` = 500.

Simulation Study Results

The results of the simulation study are illustrated in the below tables.

Table 1: RPMSE Values from 500 Simulations

	Normal	Gamma	Exponential	Weibull
n = 10	0.944	0.921	0.783	0.536
n = 100	1.064	1.322	1.057	0.899
n = 1000	0.974	2.083	1.019	0.999

Table 2: Mean 95% Confidence Interval Coverage

	Normal	Gamma	Exponential	Weibull
n = 10	0.950	0.948	0.952	0.954
n = 100	0.948	0.948	0.962	0.948
n = 1000	0.964	0.950	0.940	0.932

Table 3: Mean 95% Confidence Interval Width

	Normal	Gamma	Exponential	Weibull
n = 10	0.4930	0.4681	0.4522	0.4715
n = 100	0.0138	0.0136	0.0136	0.0135
n = 1000	0.0004	0.0009	0.0004	0.0004

We can see from Table 1 that the RPMSE values are similar for when each of the four different distributions are used for ϵ in the SLR model. The highest RPMSE comes from the Gamma distribution at n = 1000. However, if we consider the range of the covariate column in the X matrix for a sample size of 1000, which is n-1 (999), being off in predictions by about 2 doesn't feel very significant. We see a similar pattern in Table 2, where all of the coverages appear to be approximately 95%. This is what we would expect to see if calculating 95% confidence intervals. It does not appear that changing either the sample size or the error distribution affects the coverage of the intervals. Finally, we can see in Table 3 that the width of 95% confidence intervals changes only with the sample size. The width of intervals are very similar for all four error distributions at constant sample size.

Advice for Statistical Practice

The simulation study showed that a violation of the normality assumption doesn't appear to have a large effect on prediction error, and hardly any effect on the width and coverage of 95% confidence intervals. For this reason, if the purpose of the analysis is to predict Sales from Income, assuming all other conditions of the SLR model are met, the normality assumption in this case would not have a very large effect on the accuracy of the predictions with respect to the RPMSE metric. Confidence intervals for β_1, the effect of Income on Sales, are also going to be about the same width regardless of the error distribution.

At very large sample sizes, the width of these intervals becomes very close to zero. The coverage of 95% intervals appears to be independent of sample size and error distribution, so we can be confident that if we calculate 95% confidence intervals for β_1, approximately 95% of those intervals would contain the true effect of income on sales in repeated trials. It appears that the comparative success of t-procedures (like the one-sample t confidence interval for β_1 and predictions from linear regression models in sufficiently large samples is not very dependent upon the normality of the residuals. Confidence intervals behave how we expect them to behave in coverage, and width and RPMSE seem to be relatively unaffected as well. However, there are often simple solutions to taking care of the assumption of normality if the departure is extreme, and these

solutions would be necessary to incorporate for a normality departure in a dataset with a very small sample size.

A common method for dealing with a violation of the normality assumption is to perform a transformation on the response variable. Oftentimes, a log transformation sufficiently takes care of the skewness in the residuals. Note however, that the interpretation of β_0 and β_1 change with this transformation. We will calculate RPMSE, interval width, and coverage for our data, and compare with metrics from a log-transformed model. We can see based on the histogram below that the log-transformation does seem to have taken care of the skewness. A Shapiro-Wilk test on the standardized residuals reported a p-value of 0.839, indicating that the standardized do now follow a normal distribution.

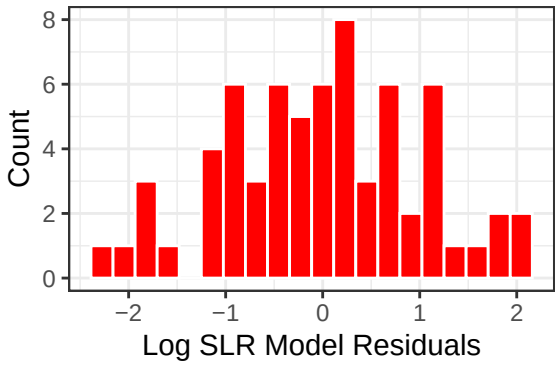

Figure 4: Histogram of (Log) Standardized Residuals

If the goal of the analysis is to retrieve a confidence interval for β_1, the skewness of the dataset will not significantly affect the width of that confidence interval or the expected 95% confidence level. The dataset has a sufficiently large number of observations (generally, with the normality assumption, 'small' sample sizes would be less than approximately 10 observations, although this that is a different simulation study). If the goal is predicting a daily smoker's monthly amount spent on cigarettes based only on their income, the skewness of the data will not change the RPMSE very significantly. A 95% confidence interval for β_1 for the SLR model applied to the untransformed dataset is found to be (0.025,0.042).

Note that, as part of this analysis, we did not check the assumptions of linearity, homeogeneity of variance, or independence of observations. These assumptions are also necessary to check before performing a full regression analysis and reporting results. We also note that there are other variables that very likely affect the amount of money spent on cigarettes (i.e. the price of cigarettes, length of the addiction, age, etc.), so if we want to increase the predictive power of this model, the addition of new and informative data would be most useful.