Assumptions, Diagnostics, and Inferences for the Simple Linear Regression Model with Normal Residuals

4 December 2018

1 The Simple Linear Regression Model with Normal Residuals

In previous class sessions, we saw that we had to make assumptions about the population a sample resulted from in order to use the inferential procedures for the population mean. For example, to perform a hypothesis test on the population mean μ of some characteristic of a population using the test statistic $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$, we had to assume that:

1. The sample in-hand was a random sample from the population of interest.
2. The population standard deviation σ of the characteristic is known.
3. At least one of the two following requirements are met:
   (a) The characteristic in the population is normally distributed.
   (b) The sample size n is sufficiently large so that the Central Limit Theorem guarentees normality of Z.

In developing the simple linear regression model in the previous class session, nothing we did required statistical assumptions about the population. We determined the line of best fit

$$\hat{y} = b_0 + b_1 x \tag{1}$$

as a way to provide the 'best' predictions $\hat{y}_i$ of the responses y_i using the predictors x_i in our sample, where 'best' here means that we minimized the squared residual between our predictions $\hat{y}_i = b_0 + b_1 x_i$ and the actual responses y_i. If all we want to do is prediction, and we are satisfied with the performance of the linear prediction $\hat{y} = b_0 + b_1 x$, we can stop there. The regression function $\hat{y} = b_0 + b_1 x$ tells us how our prediction of y changes as the predictor x changes, and thus summarizes the data using a line.

Typically, however, we do not just care about how the values of x and y are associated in our sample, but rather how they are associated in the population from which they were sampled. In that case, we want to make an inference from the sample to the population, and a statistical model is required. The statistical model underlying all of the formulas in Chapter 9 of Triola & Triola is the linear regression model with normal residuals,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{2}$$

This model says that the response Y_i (now treated as a random variable) is constructed by multiplying the predictor x_i by β_1, adding that to β_0, and then adding a random residual term ϵ_i to $\beta_0 + \beta_1 x_i$. Here, β is the Greek letter 'beta' (the analog to the Roman letter b), and ϵ is the Greek letter 'epsilon' (the analog to the Roman letter e). Thus, the intercept β_0 and slope β_1 are the population analogs of b_0 and b_1, and the residual ϵ is the population analog of the in-sample residual / error e.

The simple linear regression model with normal residuals also makes assumptions on the distribution of ϵ. It assumes that the ϵ_i are all independent normal random variables with mean 0 and standard deviation σ_ϵ. Summing up, this means that the assumptions of the full simple linear regression model with normal residuals are the following:

Assumptions Needed for Inferences Derived from Simple Linear Regression Model with Normal Residuals:

1. The sample in-hand was a random sample from the population of interest.
2. The response Y can be modeled as a linear function of the predictor x by $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
3. The residuals ϵ_i are mutually independent. That is, each residual does not depend in any way on any of the other residuals.
4. The residuals ϵ_i have mean 0 and a standard deviation σ_ϵ, both of which do not depend on x_i.
5. At least one of the two following requirements are met:
   (a) In the population, the residuals ϵ are normally distributed.
   (b) The sample size n is sufficiently large that the distribution of the residuals washes out in a Central Limit Theorem-like result.

In the final section of this handout, we will discuss margins of error and hypothesis tests for the population intercept β_0 and slope β_1. For these to be valid, we need each of the above assumptions to hold. This means that before we perform an inference on β_0 and β_1, we should check that the assumptions hold. One way to check these assumptions is to use a set of diagnostic plots, which we turn to next.

2 Diagnostic Plots for the Simple Linear Regression Model with Normal Residuals

To test the assumptions of the Simple Linear Regression Model with Normal Residuals, we would like to have the population residuals $\epsilon = Y - (\beta_0 + \beta_1 x)$, which we could then test for each of the assumptions above. However, we do not have access to these residuals. Instead, we have access to the in-sample residuals $e = y - (b_0 + b_1 x)$ from each of our predictions. The in-sample residual e will not quite match ϵ since we are using our estimates of the intercept and slope, b_0 and b_1, in place of the population values of the intercept and slope, β_0 and β_1. If we've done a good job of estimating β_0 and β_1, however, then b_0 and b_1 will be close to the population values and $e \approx \epsilon$. This means that we can use the in-sample residuals to check the assumptions listed above. If the Simple Linear Regression Model with Normal Residuals is correct, then the sample residuals should approximately satisfy each of the assumptions in the box above.

When we perform a regression from the `Stat > Regression > Regression > Fit Regression Model ...` menu item in Minitab, we can tell Minitab to plot a set of diagnostic plots by clicking the `Graphs...` button in the dialog box, and selecting the `Four in one` radio button from the resulting window. Each of these diagnostic plots checks for one of the assumptions above. One such set of diagnostic plots is shown in Figure 1 below. I have labeled these plots 1-4, and will take them one-by-one, and explain the assumption they check.

Plot 1: Normality of the Residuals

The first plot is a plot we have seen before: it is a plot of the observed values of the residuals e_i (sorted from most negative to most positive) against the expected percentiles of those residuals if the residuals followed

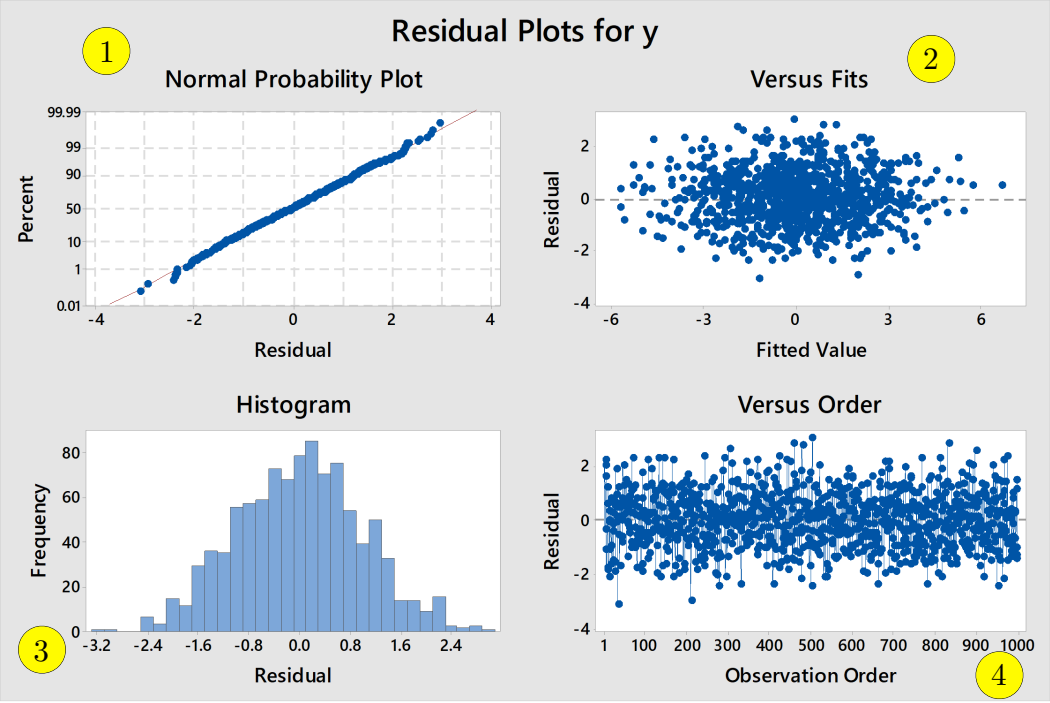

Figure 1: A sample diagnostic plot produced by Minitab when the Simple Linear Regression Model with Normal Residuals is appropriate. See the "Demo of Diagnostic Plots for Simple Linear Model" on the course website to generate more plots like these where the model is well-specified and misspecified.

a normal distribution. This is a **probability plot**, and if the residuals are approximately normal, then the points on the probability plot will approximately follow the expected values given by the red line.

Plot 2: Residuals Have Mean 0 and Constant Standard Deviation for Each Value of x

The second plot tests the assumption that the residual terms have mean 0 and standard deviation σ_ϵ, neither of which should depend on x. To check this, we plot each in-sample residual e_i against the fitted value of $\hat{y}_i$. So each point in the second plot corresponds to $(\hat{y}_i, e_i)$. Since $\hat{y} = b_0 + b_1 x$, this is just a rescaling the predictor x, so it is the same as plotting the residuals against the predictor x. If the assumptions hold, then this should look like a 'blob' without any discernible structure, and the center (mean) of that blob, along the vertical axis, should be zero for all values of x. Moreover, the vertical spread of the blob should be constant for each value of x, since the standard deviations of the residual terms are assumed to be fixed at σ_ϵ.

Plot 3: Normality of the Residuals

The third plot gives us the same information as the first plot: it is the frequency histogram of sample residuals. Thus, if the assumptions hold, the frequency histogram for the sample residuals should be bell-shaped, and if we overlay the frequency histogram with the density histogram of the appropriate normal distribution, they should match up.

Plot 4: Independence of the Residuals

The third assumption, that the residuals are mutually independent, is a very strong condition. One way to look for obvious deviations from mutual independence is to plot the residuals as a function of their order in the data set. This is especially useful if the data is ordered in some natural way, such as by time, by space, or by entrance into a study. This is precisely what the fourth plot shows: each sample residual is plotted against its order in the sample. If the residuals are mutually independent, then there should not be any pattern in this plot as a function of the observation index.

Conclusion from Diagnostic Plots in Figure 1

Based on the four plots in Figure 1, we see that all of the assumptions of the simple linear regression model with normal residuals seem to be satisfied. Plots 1 and 3 show that the sample residuals are approximately normally distributed. Plot 2 shows that there is no strong pattern in either the mean or the standard deviation of the sample residuals as a function of the predictor x: the plot is a 'blob.' Finally, Plot 4 shows no clear pattern in the sample residuals as a function of their index, so the sample residuals look approximately independent.

Demo on Course Website

You should experiment with the "Demo of Diagnostic Plots for Simple Linear Model" on the course website to generate more plots like these where the model is either well-specified and misspecified. Use the drop-down menu to change the population so that the population's distribution breaks one or more of the assumptions of the simple linear regression model with normal residuals. In this case, the simple linear regression model with normal residuals is 'misspecified': it does not match the actual properties of the population, and therefore inferences using the simple linear regression model with normal residuals will not be valid. For example, confidence intervals for the population characteristics will not have the desired coverage, and hypothesis tests will not have the desired Type I error rate. Whenever you find that your data do not match a pre-defined model, you should be very skeptical of the inferential statistics reported by Minitab.

3 Inferences from the Simple Linear Model with Normal Residuals

All of the inferential statistics (standard errors, T-statistics, P-values, etc.) reported by Minitab after running a regression rely on the assumptions outlined in the first section of this handout. This is why you should check these assumptions before proceeding. Otherwise, the numbers mean nothing: garbage in, garbage out.

Once we have found that the assumptions seem to hold in our data set, we can proceed to perform inferences from our sample to the population the sample comes from. Remember that, under the simple linear regression model with normal errors, we assume the response Y can be modeled as

$$Y = \beta_0 + \beta_1 x + \epsilon. \tag{3}$$

where β_0 is the population intercept and β_1 is the population slope. Notice that, if the model is true, then when $\beta_1 = 0$, our prediction of the response does not change per unit change in the prediction. Thus, the response and the predictor are not linearly associated. In fact, when the linear model with normal residuals is correct, zero linear association implies something much stronger: independence between the predictor and the response. For this reason, the most common hypothesis test to perform in the setting of a simple linear regression is:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

That is, we set up the null hypothesis as the boring case: there is no linear association between the predictor and the response. The alternative hypothesis is the opposite of this: there is some linear association between the predictor and the response.

To test this hypothesis, we will play the same hypothesis testing game as usual. We will compare b_1, our sample estimate of β_1, to the null value, in this case zero, and see how likely it was to get the observed value of b_1 when the null hypothesis that $\beta_0 = 0$ is actually true. So we set up a T-statistic under the null hypothesis that $\beta_1 = 0$, giving,

$$T = \frac{b_1 - \beta_1}{s_{b_1}} \tag{4}$$
$$= \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}} \tag{5}$$

where s_{b_1} is the standard error (what Triola & Triola call the margin of error) of b_1. Minitab reports this standard error as `SE Coef`: the standard error of the coefficient. Under the null hypothesis, this T-statistic is T-distributed with $n-2$ degrees of freedom. The P-value reported by Minitab thus uses the T-table with $n-2$ degrees of freedom to determine the probability of observing a value of t at least as extreme as $t = \frac{b_1}{s_{b_1}}$ when the null hypothesis is true. Notice that this is a two-tailed test. We interpret the P-value the usual way: we reject the null hypothesis at significance level α if the P-value is less than or equal to α.

We can also use the standard error of the coefficient to construct a $1 - \alpha$ confidence interval for β_1, namely

$$b_1 \pm t_{\alpha/2, n-2} s_{b_1}$$

Thus, a rough 95% confidence interval for β_1 for large enough n is $b_1 \pm 1.96 s_{b_1} \approx b_1 \pm 2 s_{b_1}$.

You can use the same construction to perform a hypothesis test or construct a confidence interval for the population slope β_0. However, this is typically of less interest, so I will not cover it in class or in these notes.