STAT 515 - Section 11.3 Supplement

Brian Habing - University of South Carolina

Last Updated: November 14, 2016

S11.3 - Checking the Regression Assumptions

The equations for fitting the regression line in section 11.2 can be applied to any data set. However, that doesn't mean that the results of the regression are actually meaningful. In order to trust the results (the predictions, the p-values, etc...) of a regression analysis the four assumptions in section 11.3 (page 603) need to hold. Each of these four assumptions involves the errors (ε). But the errors depend on the true model (β_0 and β_1) that we don't have, and are thus parameters. Because of this we can't directly look at the errors to check the assumptions. Instead of looking at the errors themselves, we'll look at the estimated errors, called the residuals.

$$\text{errors}: \quad \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

$$\text{residuals}: \quad e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

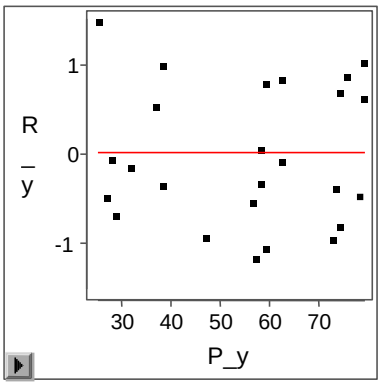
The two plots that we will use to check these assumptions appear are to use the residual versus predicted plot and the q-q plot of the residuals. In the plots in this supplement, `P_y` stands for "predicted *y* value" and `R_y` stands for "residual of *y*". In general the tests and confidence intervals for regression are fairly robust, and so the plots for checking the assumptions do not need to look perfect (but they should still look pretty good though).

Assumption 1: The mean of the errors is 0 at each *x*.

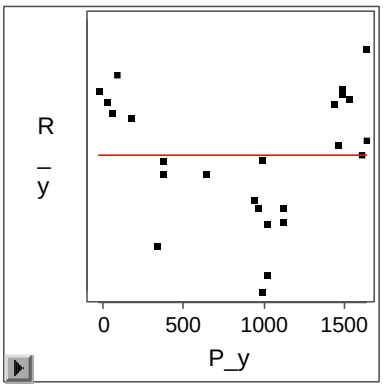
The obvious way to check this assumption would be to make a plot that had the residuals (the estimated errors) on one axis, and the *x* values (the independent variables) on the others. We would then look to see if the residuals seemed to be centered around zero at each value of x. This is basically what we do. The only difference is that we use the estimated values $\hat{\beta}_0 + \hat{\beta}_1 x_i$ instead of the *x*'s. The reason for this is because we'll still be able to use the same plot in STAT 516 where we will sometimes have several different *x*'s (e.g. predict weight from height and age instead of just height). We call this plot the ***residual versus predicted plot***, or sometimes just the ***residual plot***. We will say the assumption seems reasonable if the points are centered around 0 for each range of predicted value.

Figure 12.35a – Errors do not have mean 0. Figure 12.42 – Errors do not have mean zero or constant variance (bump in the middle and the huge outlier). Figure 12.39 – This does not have constant variance. Figure 12.40 – This one looks pretty good. Two of the points seem a bit lower than we might like, but not by much, and we might like a bit more data between -1.5 and -0.5, but that's about it.

a) Good Plot for Means



b) Bad Plot for Means

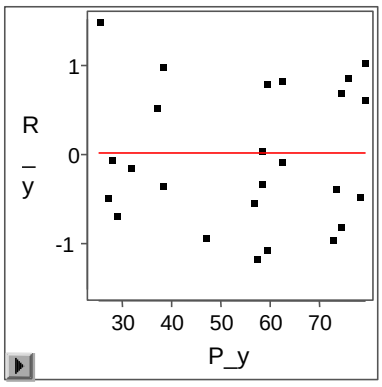


To see how to read the plot, take two index cards, placing one along the vertical axis, and another parallel to the vertical axis at round the P_y =40 line. In plot (a), as you slide the two index cards from the left to right you should see that the points are centered fairly well around R_y=0. They aren't exactly centered around zero because this is not the entire population of errors... it is a sample of the estimated errors. The single point at about P_y=47 and R_y=-1 might also be troubling. But that is because there are simply no other data points there to balance it out. (We are observing only one data point in that range of predicted values and it would have to either be above or below the line!) In plot (b), on the other hand, the residuals are positive for the low predicted values, high for the middle predicted values, and high for the high predicted values. This plot would say that a line is not appropriate and that a curve would have worked better.

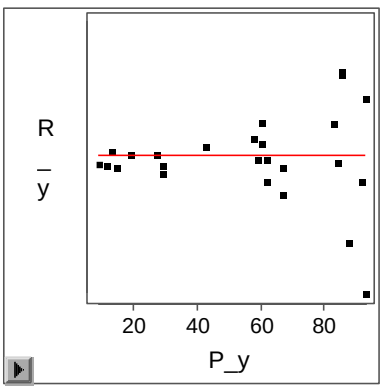
Assumption 2: The variance of the errors is constant across all *x*.

We will use the same plot to check the variances as we did the mean. On the residual versus predicted plot we will check to see that the points are spread equally up and down for each range of predicted value. We could again use two index cards, but instead of looking at the average of the points, we will look at the spread.

a) Good Plot for Variances



b) Bad Plot for Variances

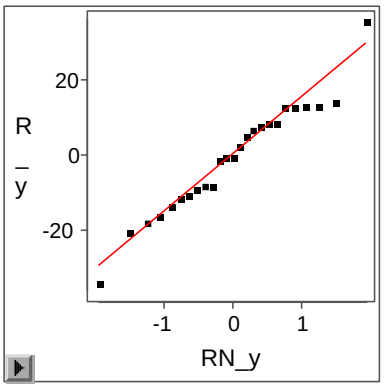


In plot (a) the points are spread out roughly between -1 and 1 all the way across (only one point is very far outside of that range.) In plot (b), on the other hand the residuals make a funnel shape. The residuals are not very spread out at the left of the plot, but are very spread out at the right side. This condition is called *heterogeneity* (changing spread). This could occur, for example, in a regression to predict weight from height. The taller you are the more variability you could have in weight (adults have a lot wider range of weights than infants!)

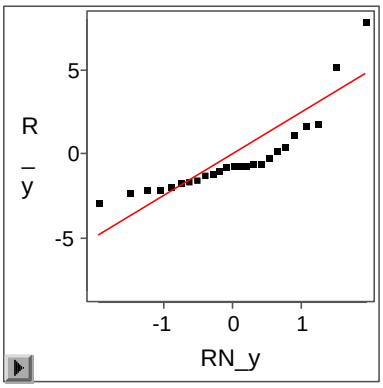
Assumption 3: The errors have a normal distribution at each *x*.

Ideally, to check that the errors are normally distributed at each *x*, we would make different q-q plots for the residuals in each range of *x*-values. Instead we usually just make one q-q plot for the entire set of residuals. We read this plot just like any other q-q plot.

a) Good Plot for Normality



b) Bad Plot for Normality



Assumption 4: The errors are independent.

In general there is no way to check this assumption from the plots alone, you need to know how the data was gathered. There are some plots that could be helpful. For example if your results were gathered in a certain order you could plot the residuals against the time when they were gathered. This would let you see if there was some extra relationship over time. Similarly if different people gathered the data you could make a separate box-plot of the residuals for each person to see if there was some relationship in the errors based on the person who gathered the data. In practice this assumption is often ignored or passed over, which is not a good thing!

Practice:

For practice, examine some of the residual plots in Section 12.11 and see which of the assumptions (if any) do not appear to be met. In particular look at Figure 12.35a (page 736), Figure 12.42 (page 741), Figure 12.39 (page 738), Figure 12.40 (page 739).

(The answers are upside-down on the bottom of the first page.)