17.8: Assumptions and model diagnostics for simple linear regression

Introduction

The assumptions for all linear regression:

1. **Linear** model is appropriate.
   The data are well described (fit) by a linear model.
2. **Independent** values of Y and equal variances.
   Although there can be more than one Y for any value of X, the Y's cannot be related to each other (that's what we mean by independent). Since we allow for multiple Y's for each X, then we assume that the variances of the range of Y's are equal for each X value (this is similar to our ANOVA assumptions for equal variance by groups). Another term for equal variances is **homoscedasticity**.
3. **Normality**.
   For each X value there is a normal distribution of Y's (think of doing the experiment over and over).
4. **Error**
   The residuals (error) are normally distributed with a mean of zero.

Note the mnemonic device: **Linear, Independent, Normal, Error** or **LINE**.

Each of the four elements will be discussed below in the context of **Model Diagnostics.** These assumptions apply to how the model fits the data. There are other assumptions that, if violated, imply you should use a different method for estimating the parameters of the model.

Ordinary least squares makes the additional assumption about the quality of the independent variable that e that measurement of X is done without error. Measurement error is a fact of life in science, but the influence of error on regression differs if the error is associated with the dependent or independent variable. Measurement error in the dependent variable increases the **dispersion of the residuals** but will not affect the estimates of the coefficients; error associated with the independent variables, however, will affect estimates of the slope. In short, error in X leads to **biased estimates** of the slope.

The equivalent, but less restrictive practical application of this assumption is that the error in X is at least negligible compared to the measurements in the dependent variable.

Multiple regression makes one more assumption, about the relationship between the predictor variables (the X variables). The assumption is that there is no multicollinearity, a subject we will bring up next time (see Chapter 18).

Model diagnostics

We just reviewed how to evaluate the estimates of the coefficients of the model. Now we need to address a deeper meaning — how well the model explains the data. Consider a simple linear regression first. If $H_O : b = 0$ is not rejected, then the slope of the regression equation is taken to not differ from zero. We would conclude that if repeated samples were drawn from the population, on average, the regression equation would not fit the data well (lots of scatter) and it would not yield useful prediction.

However, recall that we assume that the fit is linear. One assumption we make in regression is that a line can, in fact, be used to describe the relationship between X and Y.

Here are two very different situations where the slope = 0.

> Example 1. Linear Slope = 0, no relationship between X and Y
>
> Example 2. Linear Slope = 0, a significant relationship between X and Y

But even if $H_O : b = 0$ is rejected (and we conclude that a linear relationship between X and Y is present), we still need to be concerned about the fit of the line to the data — the relationship may be more nonlinear than linear, for example. Here are two very different situations where the slope is not equal to 0.

> Example 3. Linear Slope > 0, a linear relationship between X and Y
>
> Example 4. Linear Slope > 0, curve-linear relationship between X and Y

How can you tell the difference? There are many **regression diagnostic tests**, many more than we can cover, but you can start with looking at the **coefficient of determination** (low R^2 means low fit to the line), and we can look at the pattern of residuals plotted

against the either the predicted values or the X variables (my favorite). The important points are:

1. In linear regression, you fit a model (the slope + intercept) to the data;
2. We want the usual hypothesis tests (are the coefficients different from zero?) and
3. We need to check to see if the model fits the data well. Just like in our discussions of chi-square, a "perfect fit would mean that the difference between our model and the data would be zero.

Graph options

Using residual plots to diagnose regression equations

Yes, we need to test the coefficients (intercept $H_O = 0$; slope $H_O = 0$) of a regression equation, but we also must decide if a regression is an appropriate description of the data. This topic includes the use of **diagnostic tests** in regression. We address this question chiefly by looking at

1. **scatterplots** of the independent (predictor) variable(s) vs. dependent (response) variable(s).
   what patterns appear between X and Y? Do your eyes tell you "Line"? "Curve"? "No relation"?
2. **coefficient of determination**
   closer to zero than to one?
3. **patterns of residuals** plotted against the X variables (other types of residual plots are used to, this is one of my favorites)

Our approach is to utilize graphics along with statistical tests designed to address the assumptions.

One typical choice is to see if there are patterns in the residual values plotted against the predictor variable. If the LINE assumptions hold for your data set, then the residuals should have a mean of zero with scatter about the mean. Deviations from LINE assumptions will show up in residual plots.

Here are examples of POSSIBLE outcomes:

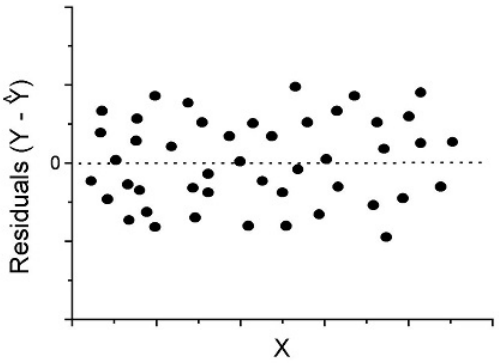

Figure 17.8.1: An ideal plot of residuals.

Solution: Proceed! Assumptions of linear regression met.

Compare to plots of residuals that differ from the ideal.

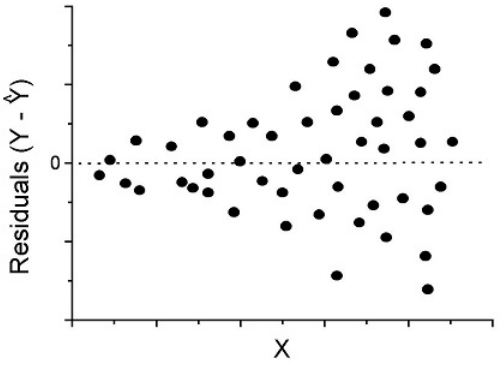

Figure 17.8.2: We have a problem. Residual plot shows **unequal variance** (aka **heteroscedasticity**).

Solution. Try a transform like the $\log_{10}$-transform.

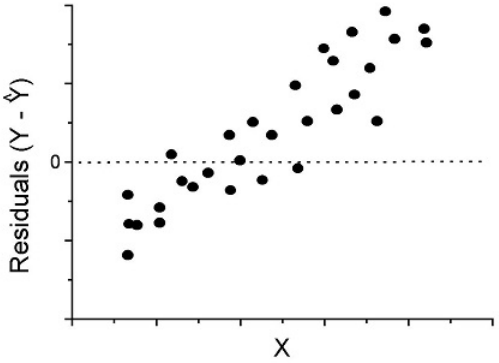

Figure 17.8.3: Problem. Residual plot shows **systematic trend**.

Solution. Linear model a poor fit; may be related to measurement errors for one or more predictor variables. Try adding an additional predictor variable or model the error in your general linear model.

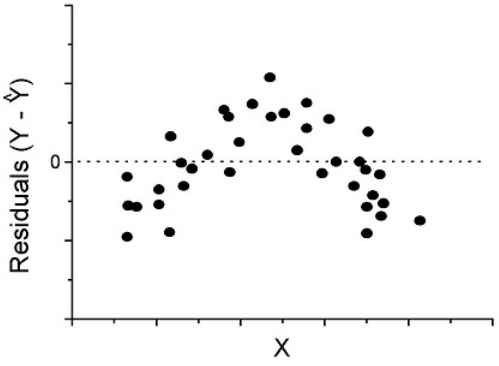

Figure 17.8.4: Problem. Residual plot shows **nonlinear trend**.

Solution. Transform data or use more complex model.

This is a good time to mention that in statistical analyses, one often needs to do multiple rounds of analyses, involving description and plots, tests of assumptions, tests of inference. With regression, in particular, we also need to decide if our model (e.g., linear equation) is a good description of the data.

Diagnostic plot examples

Return to our `example.Tadpole` dataset. To obtain residual plots, **Rcmdr: Models → Graphs → Basic diagnostic plots** yields four graphs.

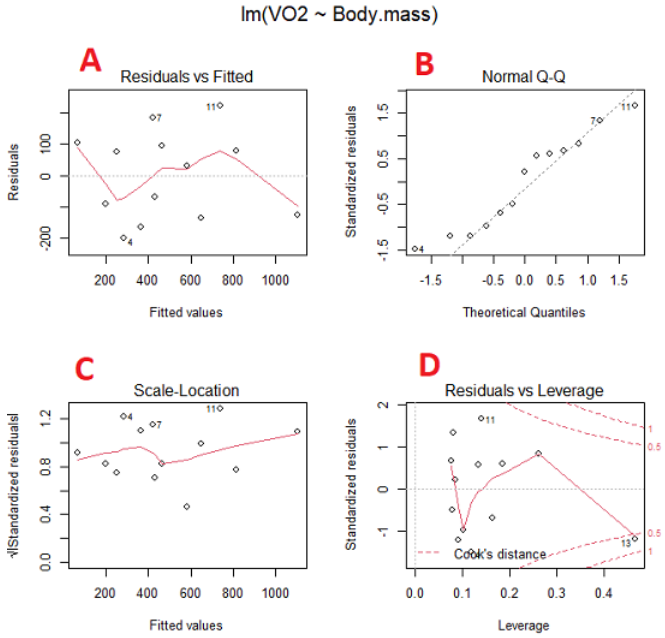

Figure 17.8.5: Basic diagnostic plots. A: residual plot; B: Q-Q plot of residuals; C: Scale-location (aka spread-location) plot; D: leverage residual plot.

In brief, we look at the plots:

A, the **residual plot**, to see if there are trends in the residuals. We are looking for a spread of points equally above and below the mean of zero. In Figure 17.8.5 we count seven points above and six points below zero so there's no indication of a trend in the residuals vs the fitted `VO2` (Y) values.

B, the **Q-Q plot** is used to see if normality holds. As discussed before, if our data are more or less normally distributed, then points will fall along a straight line in a Q-Q plot.

C, the **Scale-** or **spread-location plot** is used to verify equal variances of errors.

D, **Leverage plot** — looks to see if an outlier has leverage on the fit of the line to the data, i.e., changes the slope. Additionally, provides location of **Cook's distance** measure (dashed red lines). Cook's distance measures the effect on the regression by removing one point at a time and then fitting a line to the data. Points outside the dashed lines have influence.

Note:

A note of caution about over-thinking with these plots. R provides a red line to track the points. However, these lines are guides, not judges. We humans are generally good at detecting patterns, but with data visualization, there is the risk of seeing patterns where none exits. In particular, recognizing randomness is not easy. If anything, we may tend to see patterns where none exist, termed apophenia. So yes, by all means look at the graphs, but do so with a plan: red line more or less horizontal? Then there is no pattern and the regression model is a good fit to the data.

Statistical test options

After building linear models, run statistical diagnostic tests that compliment graphics approaches. These are available via

Rcmdr: Models → Numerical diagnostics

Variance inflation factors (VIF): used to detect multicollinearity among the predictor variables. If correlations are present among the predictor variables, then you can't rely on the the coefficient estimates — whether predictor A causes change in the response variable depends on whether the correlated B predictor is also included in the model. If correlation between predictor A and B, the statistical effect is increased variance associated with the error of the coefficient estimates. There are

VIF for each predictor variable. A VIF of one means there is no correlation between that predictor and the other predictor variables. A VIF of 10 is taken as evidence of serious multicollinearity in the model.

Breusch-Pagan test for heteroscedasticity… Recall that heteroscedasticity is another name for unequal variances. The test statistic can be calculated as $\chi^2 \sim nR^2$

Durbin-Watson for autocorrelation

RESET test for nonlinearity

Questions

1. Referring to Figures $17.8.1$– $17.8.4$on this page, which plot best suggests a regression line fits the data?
2. Return to the electoral college data set and your linear models of Electoral vs. POP_2010 and POP_2019. Obtain the four basic diagnostic plots and comment on the fit of the regression line to the electoral college data.
   - Residual plot
   - Q-Q plot
   - Scale-location plot
   - Leverage plot
3. With respect to your answers in question 2, how well does the electoral college system reflect the principle of one person, one vote?
