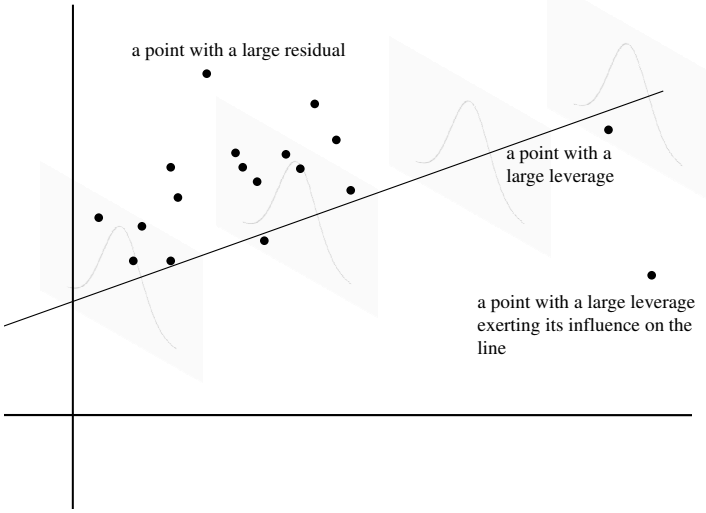
Lecture 17

Outliers and Leverage Points

In this lecture, we discuss the model assumptions for linear regression, how to assess the validity of those assumptions, how to detect outliers, how to determine if the outliers are important, and what to do with the outliers.

Recall that the regression model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$. Normality is not so critical for determining the regression line or for determining a confidence band around the regression line. It is critical for determining the prediction bands. So whether normality in the residuals is critical or not depends on the purpose of the regression line itself.



Outliers and Leverage Points

Minitab automatically flags residuals that are larger than two standard deviations. If you have 1000 data points, as we have with the normal fetuses when trying to predict femur length from biparietal diameter, you would expect $5\% \times 1000 = 50$ points to be flagged. There is no reason to become alarmed. Furthermore, a point with a large residual does not necessarily exert a large influence on the line. Points far away from the center of the predictor values have the ability to exert much more influence on the line than points in the center of the predictor values. The basic diagnostics Minitab provides are these:

1. The raw residual, ϵ_i. This is nearly worthless as a diagnostic tool.

2. The leverage values (called H_i in Minitab, but written $\hat{h}_i$ in statistics texts) that measure how far the predictor variable is from the center of the predictor values.

$$\hat{h}_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)s_X^2}$$

3. The standardized (also called studentized) residuals, $\epsilon_i/\hat{\sigma}_i$, where $\hat{\sigma}_i$ is the standard deviation in the i^{th} residual estimated from data which involves both the overall standard deviation $\hat{\sigma}$ and the leverage: $\hat{\sigma}_i = \hat{\sigma}\sqrt{1-\hat{h}_i}$.

4. The deleted t residuals (studentized deleted residuals). This essentially does the following: Remove the i^{th} data point from the data set. Estimate the regression line. Compute the studentized residual for the i^{th} point from this new regression line and report the results. If the point has exerted a great deal of influence on the line, its residual under this method is likely to be large in size.

5. DFITS - roughly the number of standard deviations between the predicted values with and without the i^{th} observation included in the regression.

6. Cook's Distance - USE THIS - this is a single measure that combines leverages and residuals to determine whether a point exerted its leverage and whether the residual was large enough to significantly influence the regression line. The formula is

$$D_i = \frac{1}{k+1}(\text{studentized residual}_i)^2 \left(\frac{h_i}{1-h_i}\right) = \sum_{j=1}^{n} \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{(k+1)\hat{\sigma}^2}$$

where $\hat{Y}_{j(i)}$ is the fit for the j^{th} value if observation i is removed from the data set, $\hat{Y}_j$ is the corresponding fit when all the data is used, $k+1$ is the number of parameters in the model (2 for simple linear regression), and $\hat{\sigma}$ is the estimate for the residual standard deviation. Flag any point with $D_i > 1$ and also examine any point with an outlying Cook's distance value.

We will examine these diagnostics for the regression line for predicting femur length from biparietal diameter. Eventually this will lead us to discover a clear data entry error. Removing the data entry error, however, makes little difference in our ultimate conclusion that using the regression is useless for helping to predict Down syndrome.

What to do with problem points

When you have a point that has a large residual, a large leverage, or is exerting a lot of influence on the line, can you simply remove it? Not really - not if the point is REAL and not a clear data entry error. If the point is real, you can conduct your analysis with it and without it. If your ultimate conclusions don't change, then you are in good shape. If removing the point changes your conclusions, then you are in bad shape and it is not possible to report a reliable conclusion. You ought to report that your results rely on a single observation or that your results are not reliable because they are unduly influenced by a single observation.

Is a line the correct model?

Minitab can perform two different tests to determine if a line is a good model for your data. One looks to see if a separate means model fits the data better. To conduct this test, you need repeated values for your predictor variable - something you can guarantee if you choose the predictor values but not if you don't. We don't get to choose the biparietal diameter of fetuses, but, with 1000 data points, we do have many biparietal diameters that are repeated in the data set. The other looks to see if their is curvature in the data. It divides the predictor values in half (or so) and looks to see if fitting a line to the first half and a line to the second half which meets up with the first half appropriately leads to lines with the same slop or different slopes. If the slopes are different, then the data has curvature which is reported in the output.

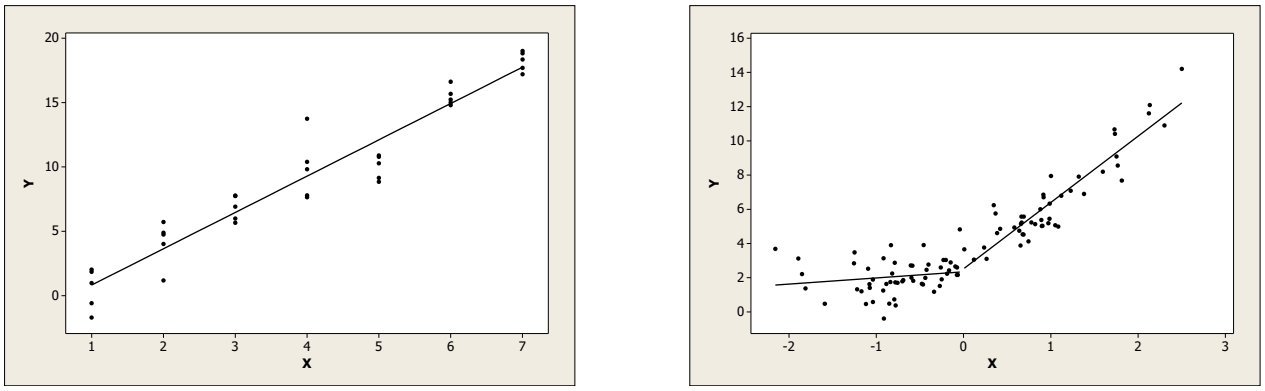


Figure 17.1: Examples when a separate means model may be better than a linear one and when there is curvature in the data that can be determined by the data subsetting procedure.

Exercises for Lecture 17

1. –

2. –