Class 4. Leverage, residuals and influence

1 Today's material

An in depth look at

Residuals

Leverage

Influence

Jackknife

Masking

2 Residuals

Residuals are vital to regression because they establish the credibility of the analysis. Never accept a regression analysis without having checked the residual plots.

Residuals come in many flavors:

Plain vanilla residual: $e_i = (y_i - \hat{y}_i)$.

Standardized residual:

$$s_i = \frac{e_i}{\sqrt{\widehat{Var}(e_i)}}.$$

Studentized residual: t_i.

2.1 Plain vanilla

The plain residual e_i and its plot is useful for checking how well the regression line fits the data, and in particular if there is any systematic lack of fit, for example curvature.

But, what value should be considered as a big residual?

Problem: e_i retains the scale of the response variable (Y).

Answer: standardize by an estimate of the variance of the residual.

Know, $Var(y_i) = \sigma^2$ estimated by $(RMSE)^2$.

But, $e_i = (y_i - \hat{y}_i)$, which is more than just y_i.

Turns out, $Var(e_i) = \sigma^2(1 - h_{ii})$.

Use standardized residual, s_i.

The quantity, h_{ii} is fundamental to regression.

An heuristic explanation of h_{ii} (visually we are dragging a single point upward and measuring how the regression line follows):

Think about y_i the observed value, and $\hat{y}_i$ the estimated value (ie the point on the regression line).

For a fixed x_i perturb y_i a little bit, how much do you expect $\hat{y}_i$ to move?

If $\hat{y}_i$ moves as much as y_i then clearly y_i has the potential to drive the regression – so y_i is leveraged.

If $\hat{y}_i$ hardly moves at all then clearly y_i has no chance of driving the regression.

In other words h_{ii} is the measure of "leverage".

More precisely

$$h_{ii} = \frac{d\hat{y}_i}{dy_i},$$

and it depends only on the x-values.

Understanding leverage is essential in regression because leverage exposes the potential role of individual data points. Do you want your decision to be based on a single observation?

2.2 Standardized Residuals

Standardized residuals, allow the residuals to be compared on the "standard scale". Plus/Minus 2 indicates something unusual, Plus/Minus 3 indicates something really out of the ordinary and Plus/Minus 4 is something from outer space (it just shouldn't happen).

Subtle point

Problem. The standardized residuals, s_i, still start off with $y_i - \hat{y}_i$ and the problem is that if y_i is really leveraged then it will drag the regression line toward it, influencing the estimate of the residual itself.

Solution. Fit the regression line excluding y_i and base the residual on $y_i - \hat{y}_{i,(-i)}$, where $\hat{y}_{i,(-i)}$ denotes the fit based on a regression line estimated excluding y_i.

This idea leads to the studentized residuals.

2.3 Studentized

The studentized residuals are driven by the **leave one out** idea, which is the basis for much computationally intensive modern statistics. The leave one out idea is often called "jackknifing".

This "leave one out" residual can be used as a basis for judging the predictive ability of a model. Clearly the lower the residual the better, and the sum of the squares of the jackknifed residuals is called the PRESS statistics, or Predicted Sum of Squares.

The **studentized residual**, t_i, is just a standardized jackknifed residual. This is an extremely good way of judging how much of an outlier in the y-direction a point is.

From now on we will use the studentized residual plot to judge outliers in the y-direction.

A new plot. Leverage vs. studentized residual. Points that drive the regression have big leverage and extreme studentized residuals.

The delete one idea works pretty well, except when there is a second data point lying close by. In this case the second point can drive the regression line, **masking** the effect of the first point. This leads to the idea of "delete two" etc.