

Regression Effect/Fallacy

Regression Effect: In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test and the top group will on average fall back. This effect is known as the *regression effect*.

Regression Fallacy: The *regression fallacy* is thinking that the regression effect must be due to something important, not just due to spread about the regression line.

Regression Diagnostics

Recall, a **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\text{residual}_i = \text{observed } y_i - \text{predicted } y_i = y_i - \hat{y}_i$$

- Sum of the residuals is always 0.
- Mean of the residuals is always 0.

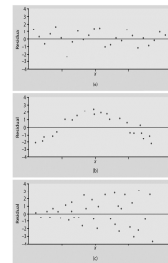
Residual Plot - scatterplot of the residuals against the explanatory variable => help us assess the fit of the regression line.

- Since the mean of the residuals is 0, we want the points on the residual plot to be evenly distributed on either side of the horizontal line with height equal to 0. Always add the horizontal line at 0 to the residual plot.
- If the regression line catches the overall pattern of the data, there should be *no pattern* in the residuals.

1) No Pattern - good

2) Curved Pattern - evidence that the relationship between x and y is rather than linear

3) Fan Shape - variation in y increases as x increases (heteroskedasticity)



Other Residual Plots

Normality Assumption (don't need for least squares regression)

- Normal quantile plot of residuals (straight line?)
- Histogram of residuals (bell-shaped?)

Measurement Problems

- Residuals versus observation number

Outliers and Influential Points

Outliers: An outlier is an observation which lies outside the pattern of the rest of the data

- Not all outliers have large residuals.

Influential Point: An influential point is an observation which affects the regression results if that value is removed from the dataset.

- Not all influential observations are outliers.
- How can you test whether an observation is influential?

Fit the model with and without the observation. Look at the regression lines and the r^2 values.