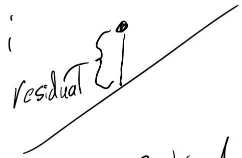
Multivariate Regression: Transformations and Diagnostics

Diagnostics

When doing regression we assume follows a linear trend and that residuals have no pattern and are normally distributed.

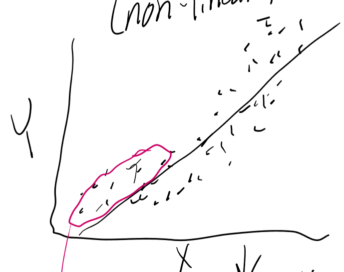
$(y_i - \hat{y}_i) = \text{resid}_i$

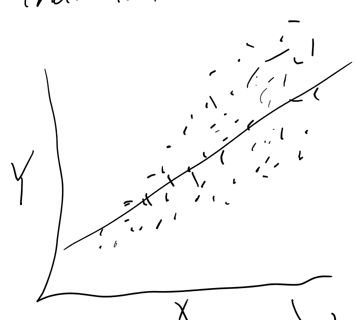


One of the best ways to confirm assumptions is to plot residuals against fitted value.

What if we have data that looks like....

(non-linear)

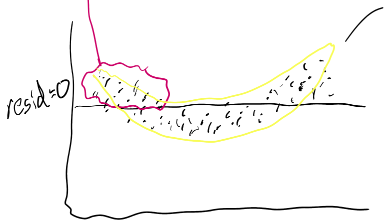




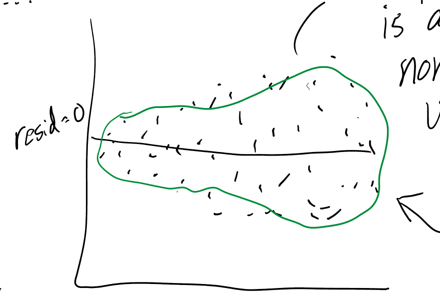
non-homoscedastic or heteroscedastic

Scedasticity refers to randomness

Our residuals will look like this....



This banana shape is an indicator of non-linearity



This pear shape is an indicator non-constant variance.

We want no pattern

gaps are okay but maybe indicate different groups

Diagnosing normality of residuals can be done with a simple histogram or more methodically with a quartile-quartile plot (qq plot)

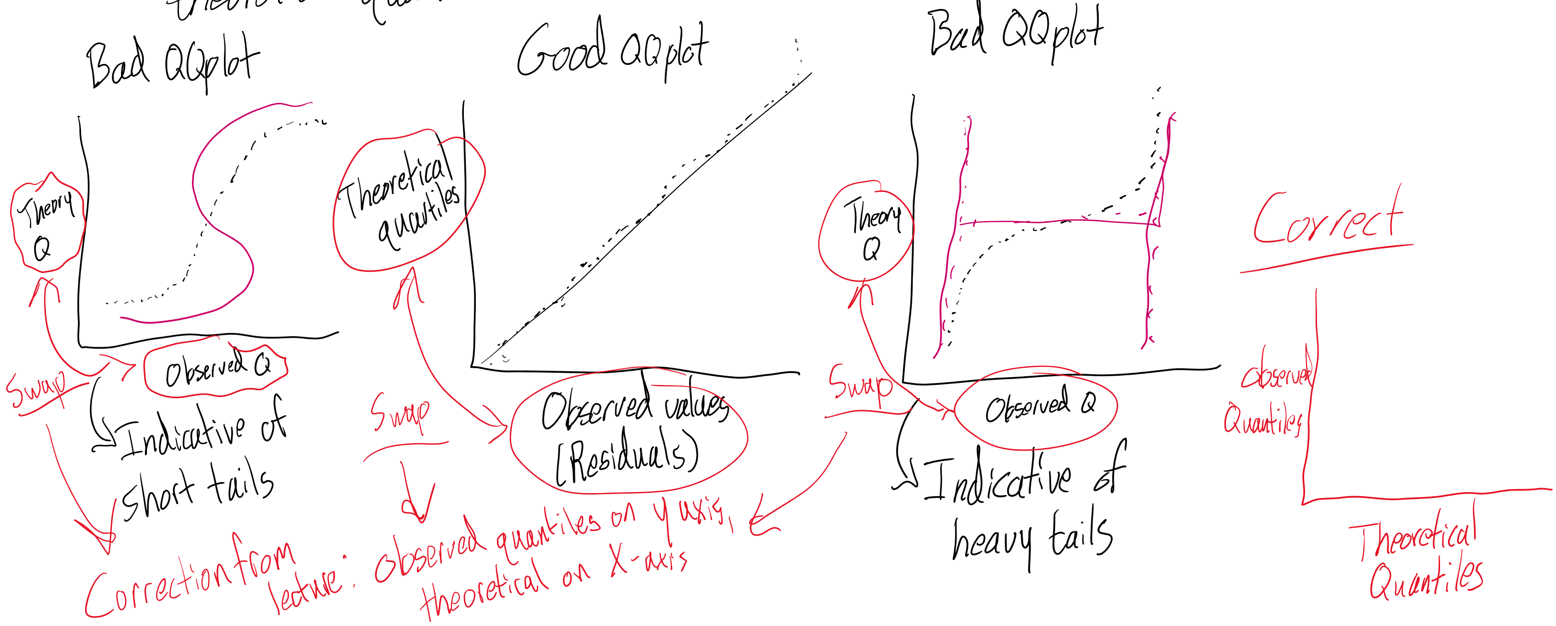
A qqplot will find the quartiles of the data and plot them against theoretical quartiles.

Bad QQplot

Good QQplot

Bad QQplot

theoretical quantiles.



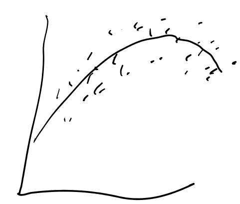
We can sometimes resolve these issues by transforming the data and/or adding more terms to the model

rather than $y_i = b_0 + b_1 X + \varepsilon_i$ ← resid...

We can add a squared term for X.

we might consider

$$y_i = b_0 + b_1 X + b_2 X^2$$



But we can all sorts of transformations on X or Y.

$$\log(y_i) = b_0 + b_1 X + b_2 X^2$$

$$\sqrt{y_i} = b_0 + b_1 X + b_2 \log(X) + b_3 \sqrt{X}$$

We want to find the $\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots$ that best describe the data with linear relations between outcome and predictors.

IE: The relationship between $\sqrt{y_i}$ and $\log(X)$ is still linear!

→ IE: The relationship between $\sqrt{y_i}$ and $\log(X)$ is still linear!

One transformation that is common but has "little impact" on the model

We can take X and Y and "normalize" or "standardize" them.

$$Z_{Y_i} = \frac{Y_i - \bar{Y}}{S_Y} \qquad Z_{X_i} = \frac{X_i - \bar{X}}{S_X}$$

mean 0, sd 1 mean 0, sd 1.

Benefit: Slope interpretation becomes: "For a 1 standard deviation increase in X, we expect a b_1 standard deviation increase in Y"

Benefit: $b_0 = 0$

Original Data