Linear Regression: Validation

Connor K. Brubaker

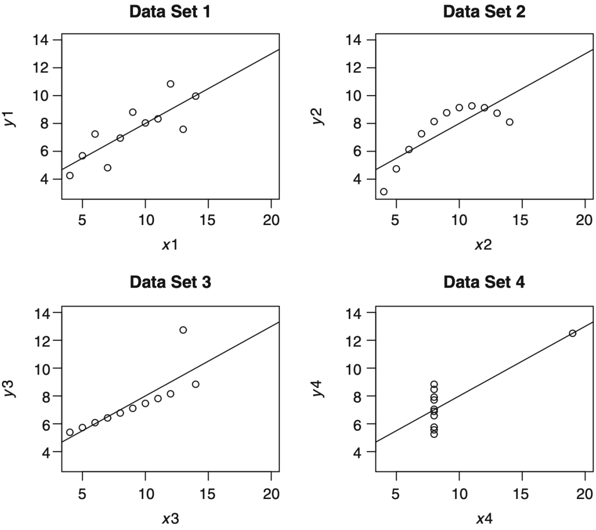
Department of Statistics
Texas A&M University

Model Validation

Inferences and predictions made from a fitted model only make sense if the assumptions of that model are fulfilled.

Model validation is the process of evaluating if the assumptions are met.

Anscombe's Data Sets



Data Set 1

Data Set 2

Data Set 3

Data Set 4

Anscombe's Data Sets

Each of these four data sets results in the same fitted regression line,

$$\hat{Y} = 3 + 0.5X$$

but only one (Data Set 1) satisfies the assumptions of the model–only one of these results in a valid model!

Don't just look at numerical output of a model. Always check it visually!

Assumptions of the Linear Model

The simple linear model makes four assumptions:

1. X and Y are linearly related,
2. the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent of each other,
3. the errors $\varepsilon_1, \ldots, \varepsilon_n$ have a common variance σ^2 (homoscedasticity), and
4. the errors $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed with a mean of 0 and variance σ^2.

We will look at ways of visually evaluating these.

Linearity and Constant Variance

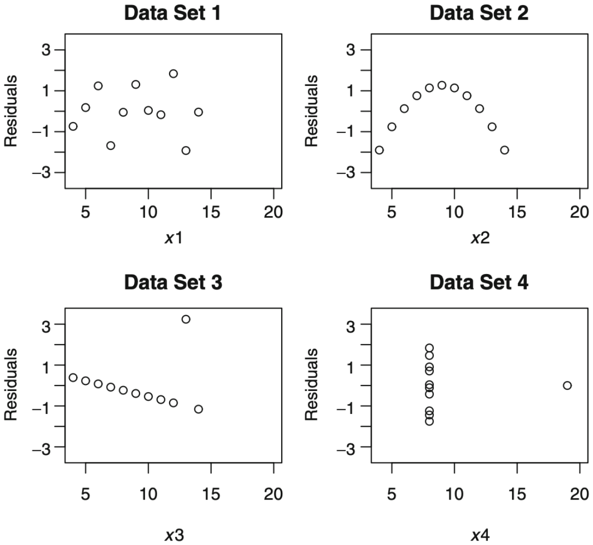
Linearity and homoscedasticity is assessed by looking at scatter plots of the fitted residual

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

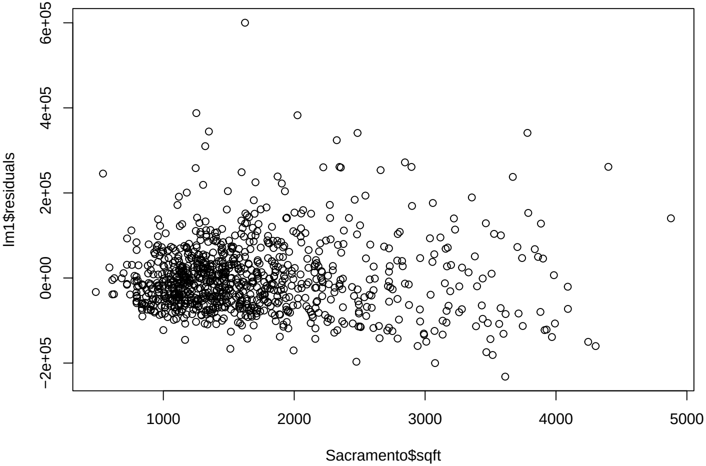
against X_i (called a residual plot).

A residual plot should have no discernable pattern.

Residual Plots for Anscombe's Data Sets

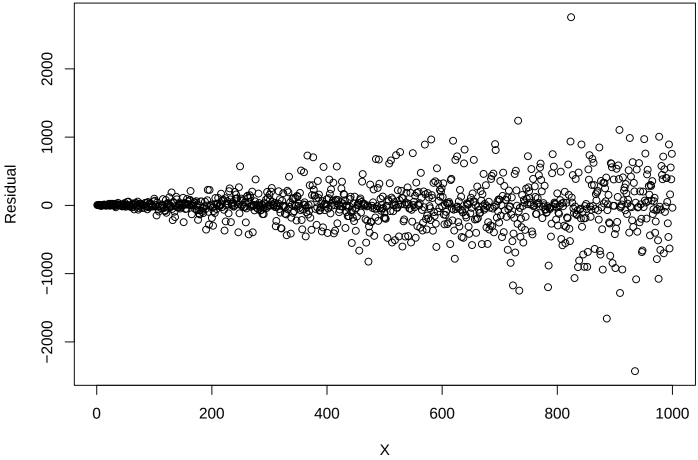


Residual Plot for Housing Data



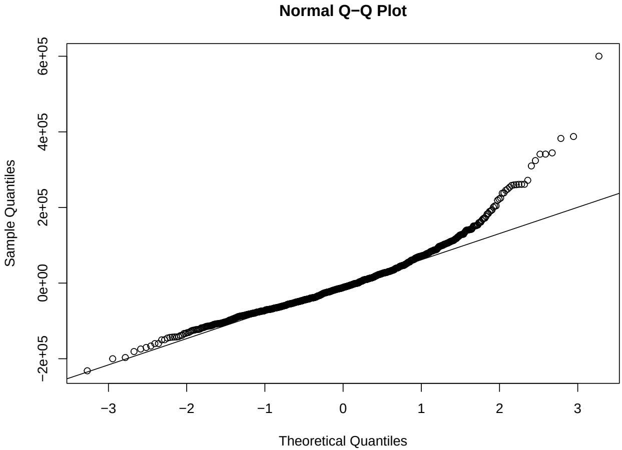
An Example of Heteroscedasticity

A residual plot like this indicates that the homoscedasticity assumption has been violated.



Normality Assumption

The normality of errors assumption can be checked visually with a normal Q-Q plot like we've seen before. Here is the Q-Q plot of the residuals from the housing data.

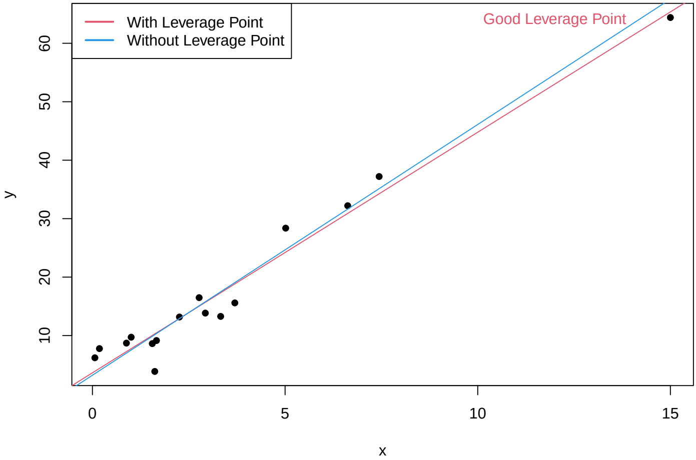


LEVERAGE POINTS

- Data points which have considerable influence on the fitted model are called **leverage points**.
- They have X values that lie away from the rest of the data points.
- Leverage points can be either “good” or “bad”.
- The leverage of a point can be examined by removing it from the data, refitting the model, and evaluating the effect.

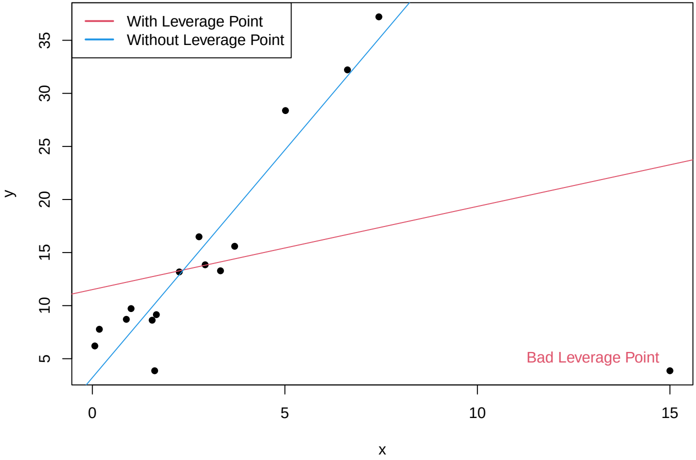
Good Leverage Points

A leverage point that follows the linear pattern of the data is a good leverage point.



Good Leverage Points

A leverage point that is also an outlier is a bad leverage point.



Dealing With Bad Leverage Points

- Bad leverage points need investigation. They may be candidates for removal from the data due to wrong data entry, bad experimental conditions, etc.
- If valid, the existence of bad leverage points may suggest a different model is needed (e.g., non-linear model)
- Even if a leverage point is “good”, they do affect standard errors and the value of R^2, so they always warrant investigation to check they are valid.