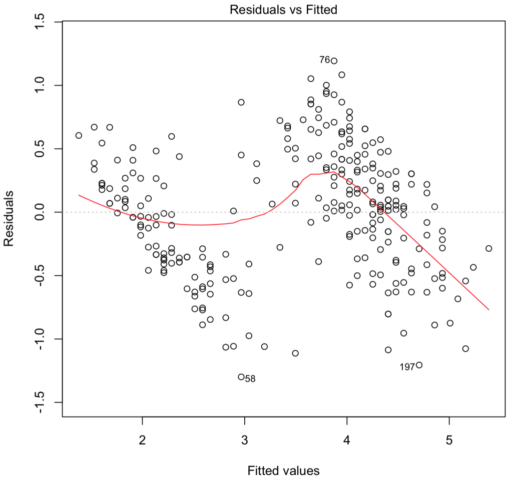
Categorical Predictors and Leverage

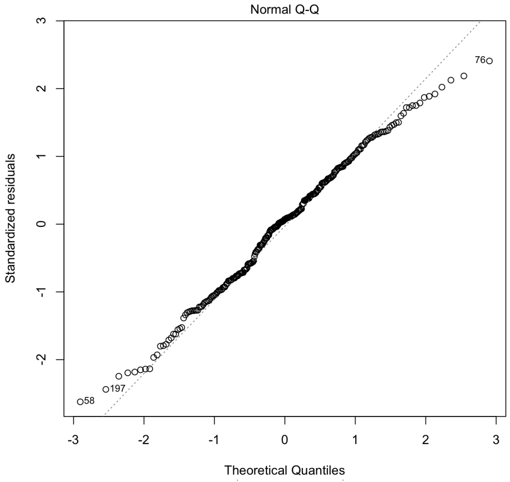
November 4, 2019

More Regression Diagnostics



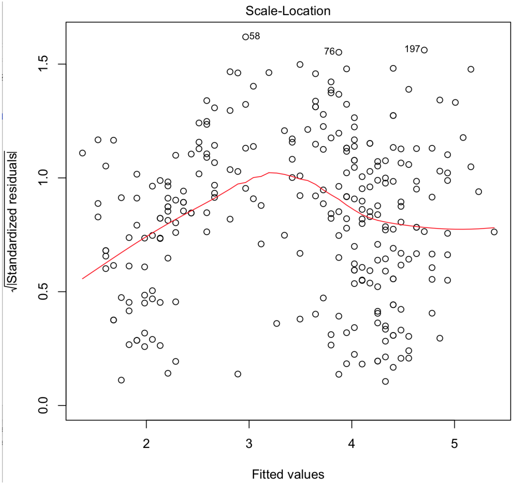
Residuals vs. fitted values in R for the `faithful` data.

The Normal Q-Q Plot



The normal quantile-quantile (QQ) plot for the `faithful` data.

The Scale-Location Plot

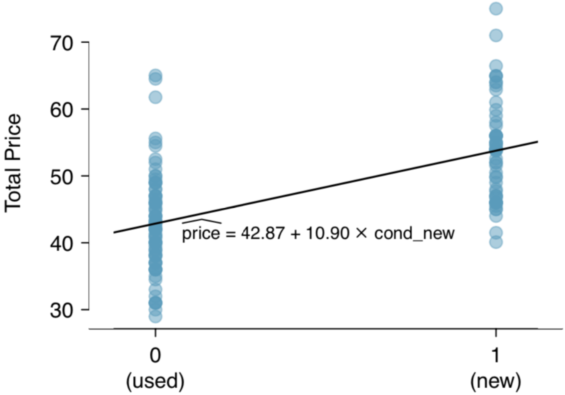


The scale-location plot for the `faithful` data.

Categorical Predictors with Two Levels

- We can also use categorical variables to predict outcomes!
- Under our current set up, we can use a categorical predictor with two levels.
- Later:
  - We will examine predictors with multiple levels.
  - We will examine response variables with two levels.

Example



- Consider Ebay auctions for Mario Kart Wii.
- We want to know how game `condition` affects selling `price`.

Example

To use `condition` in a regression, we use a **indicator variable**.

- An indicator variable always takes the values 0 or 1.
- Let $x = 0$ when `condition` is `used`.
- Let $x = 1$ when `condition` is `new`.
- We are *indicating* whether the game is new.

Example

Using our indicator variable for `condition`,

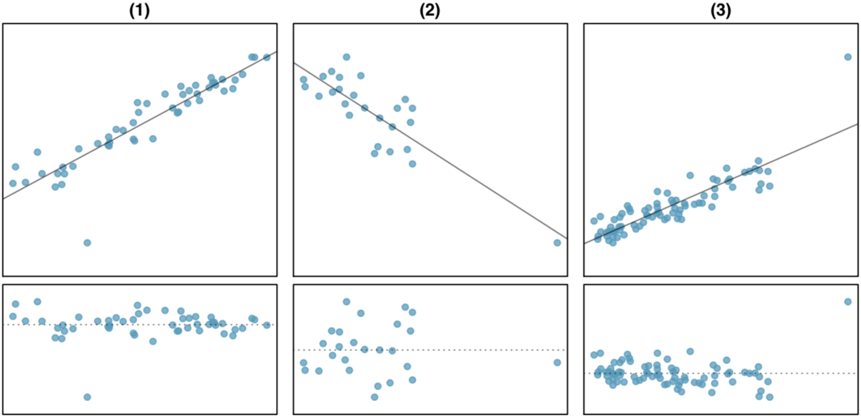
$$\begin{aligned}\hat{\text{price}} &= b_0 + b_1 x \\ &= 42.87 + 10.90x\end{aligned}$$

Interpret the model parameters.

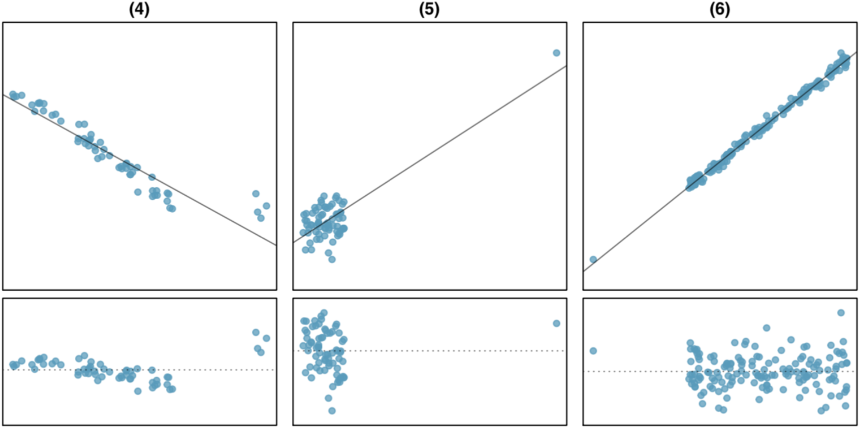
Outliers in Linear Regression

- We want to think about which points can be considered outliers.
- We also want to think about how influential these points are.

Example



Example



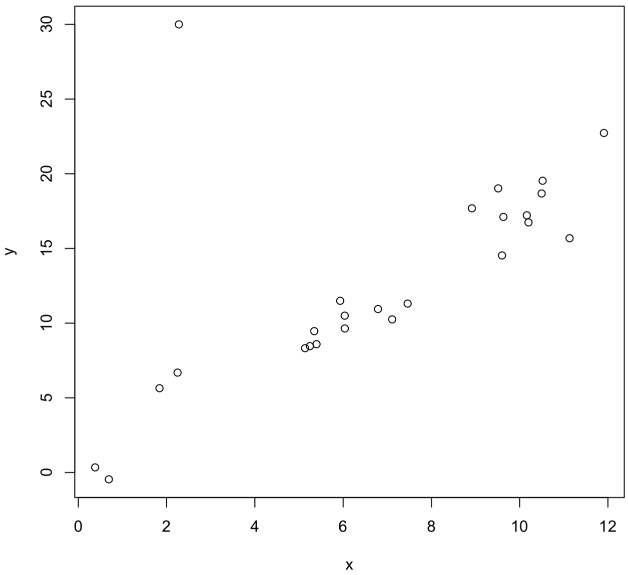
Leverage

Points that fall away horizontally from the center of the cloud tend to pull harder on the line. We refer to these points as **high leverage**.

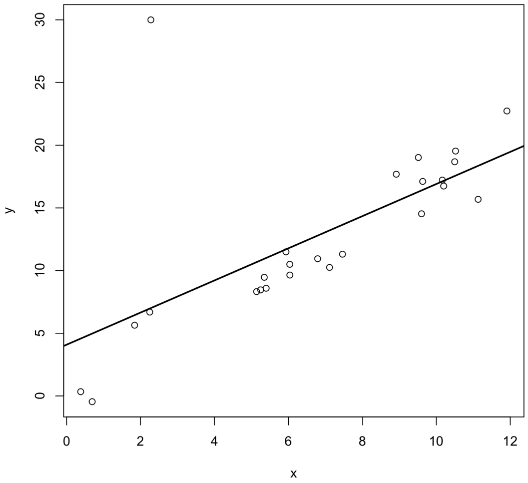
Influential Points

- We conclude that a point is **influential** if, had we fit the line without it
    - the line would have been very different.
    - the point would have been far from the line.

Example



Example



The least squares regression line is $\hat{y} = 4.0886 + 1.2817x$.

Example

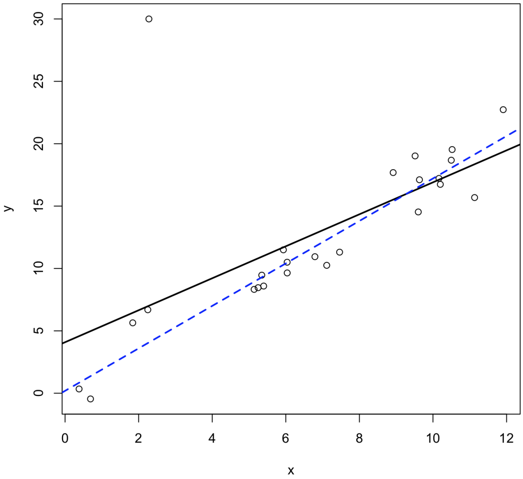
If we remove this point and rerun the regression, we get the line

$$\hat{y} = 0.1923 + 1.7021x$$

a significant deviation from the original line,

$$\hat{y} = 4.0886 + 1.2817x$$

Example



The blue dashed line is the regression line with the extreme point removed.

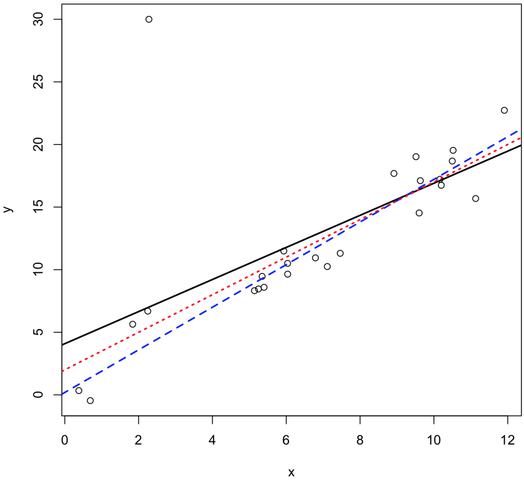
Example

I actually simulated 25 data points under

$$y = 2 + 1.5x + \epsilon$$

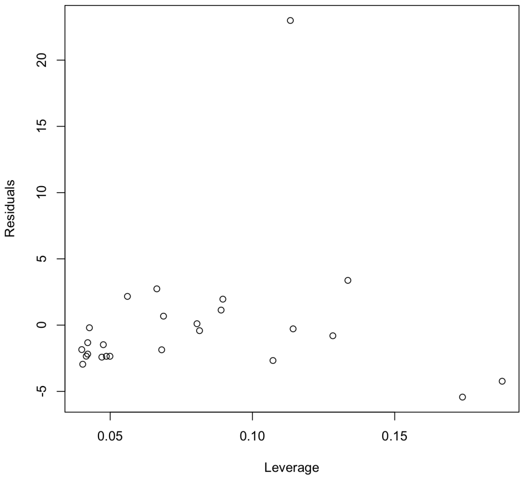
and then changed one of the points to create an outlier.

Example



The red dotted line is the truth.

Diagnosing Problematic Points

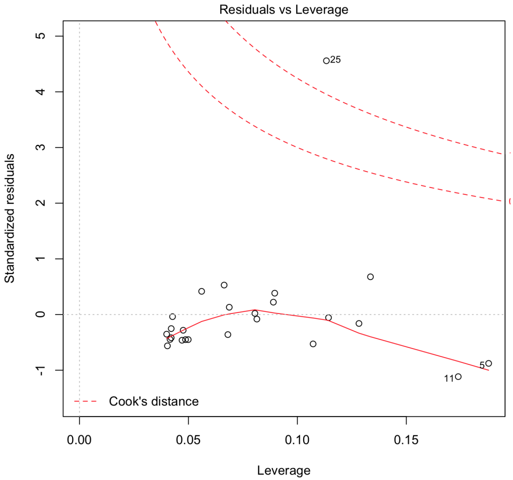


We are interested in points with high leverage *and* extreme residuals.

Cook's Distance

- We're not too concerned about outliers if they are low leverage.
- We're also not too concerned about high leverage points if they are not outliers.
- When is a point an outlier and high leverage? Enter Cook's distance.

Residuals vs Leverage



This is the final diagnostic plot automatically generated by R.

Removing Outliers

- It may be temping to remove outliers.
- However, we don't want to remove outliers for purely mathematical reasons!
- Outliers should be removed for good scientific reasons.
    - Faulty equipment, mis-entered data, etc.
- Sometimes outliers are the most interesting part of the data!