

Agenda

1. Some thoughts related to the election and statistics
2. Wrap up of randomization
3. Unusual points

Randomization With your neighbor, discuss how you could use randomization to determine if the R^2 value of a simple linear model was greater than 0. Write out at least three steps that would be required (perhaps these steps would need to be repeated, as well).

Leverage Recall that points that have extreme x values can have a disproportionate influence on the slope of the regression line. The *leverage* of the i^{th} observation is defined by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

in simple linear regression. Note that $0 < h_i < 1$, $\sum_i h_i = 2$, and thus the *average* leverage is $2/n$. By convention we consider observations to have high leverage if $h_i > 4/n$.

The standard errors for confidence and prediction intervals can be rewritten as

$$SE_{CI}(x_i) = \hat{\sigma}_\epsilon \sqrt{h_i} \quad SE_{PI}(x_i) = \hat{\sigma}_\epsilon \sqrt{1 + h_i}$$

In multiple regression, the situation is more complicated.

Standardized & Studentized Residuals Recall that the residual associated with the i^{th} observation is defined by $e_i = y_i - \hat{y}_i$, and that the residual standard error is denoted $\hat{\sigma}_\epsilon$. Here we define two alternative measures:

- Standardized residual

$$stres_i = \frac{y_i - \hat{y}}{\hat{\sigma}_\epsilon} \cdot \frac{1}{\sqrt{1 - h_i}}$$

- Studentized residual

$$stures_i = \frac{y_i - \hat{y}}{\hat{\sigma}_{(i)}} \cdot \frac{1}{\sqrt{1 - h_i}}$$

where $\hat{\sigma}_{(i)}$ is the residual standard error of the same regression model with the i^{th} point omitted

Both follow a t -distribution under the standard assumption that the residuals follow a normal distribution. The latter helps avoid the case where a single point has extremely high leverage, which artificially decreases the residual.

```
require(mosaic)
require(Stat2Data)
data(PalmBeach)
m1 <- lm(Buchanan ~ Bush, data=PalmBeach)
rstandard(m1)
rstudent(m1)
```

Cook's Distance Recall that a point of high leverage is not necessarily influential. However, a point of high leverage *with a large standardized residual* is likely to be influential! Cook's distance captures both ideas:

$$D_i = \frac{(\text{stres}_i)^2}{k+1} \cdot \frac{h_i}{1-h_i}$$

We say that observations for which $D_i > 0.5$ are moderately influential, and observations for which $D_i > 1$ are very influential.

Caveat: Sometimes the most interesting analysis is in terms of the unusual observations!

```
cooks.distance(m1)
```

Activity

1. Use R to identify the six unusual counties given in the book on page 181.
2. Examine the fourth diagnostic plot for the model produced by `plot()`. Identify the two counties with Cook's distance greater than 1.