STA 674

Regression Analysis And Design Of Experiments

Assessing Model Assumptions – Lecture 3

STA 674, RADOE:
Assessing Model Assumptions

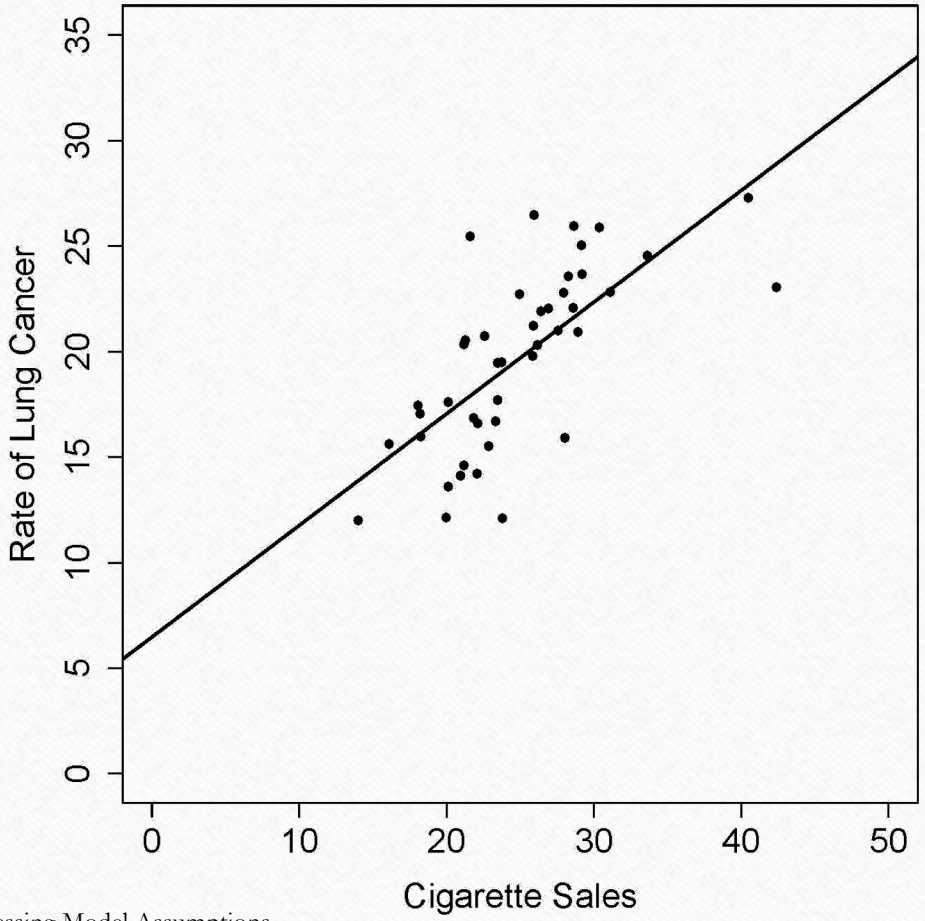
- Last time, we covered the last assumption—normality of the errors and did an example.
- This time, we talk about two kinds of "extreme" observations—outliers and influential observations.

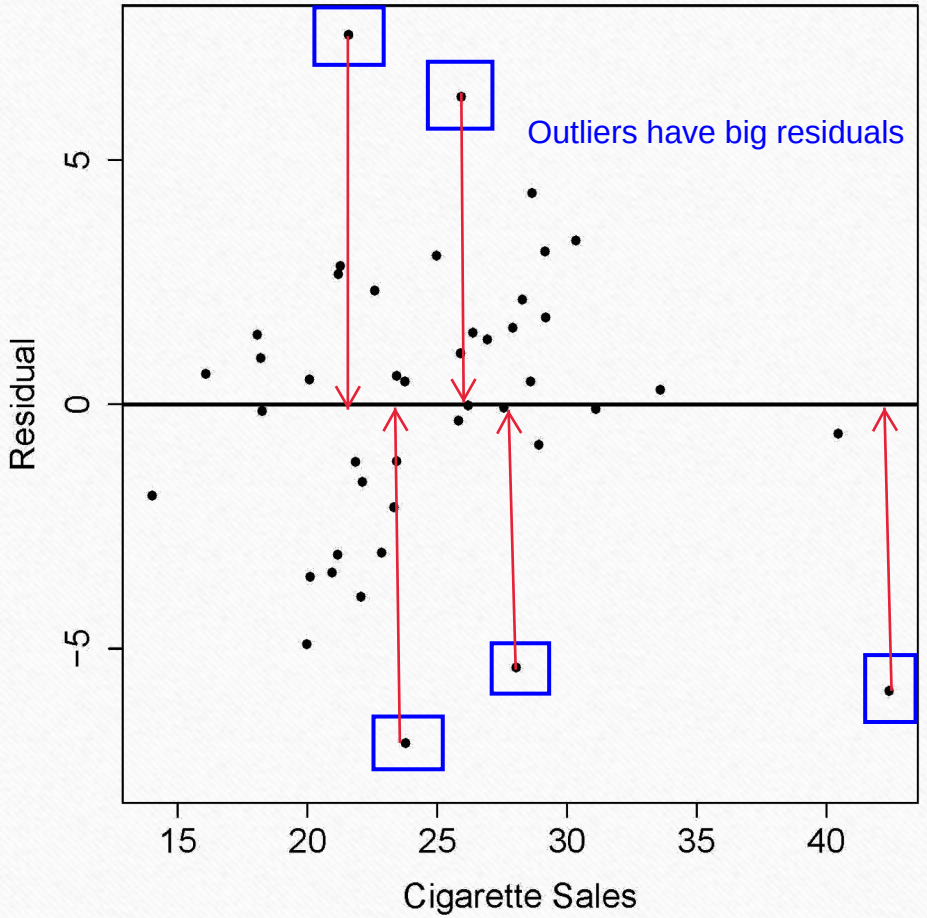
STA 674, RADOE:
Assessing Model Assumptions

Definition

- A data point is an **outlier** if the *response* at this point is far from the response predicted by the fitted model.
- *Outlier* refers to a point with an unusual y value.

Example – Cigarette Sales and Lung Cancer: Rate and Residuals versus Cigarette Sales





STA 674, RADOE:
Assessing Model Assumptions

Definition

- The standardized residual for the i^{th} data point is:

$$e_{is} = \frac{e_i}{s_e}.$$

error of residuals divided by standard error of residuals

- The distribution of the residuals is approximately normal with mean 0 and variance σ_e^2:

$$e_i \sim N(0, \sigma_e^2), \text{ approximately}$$

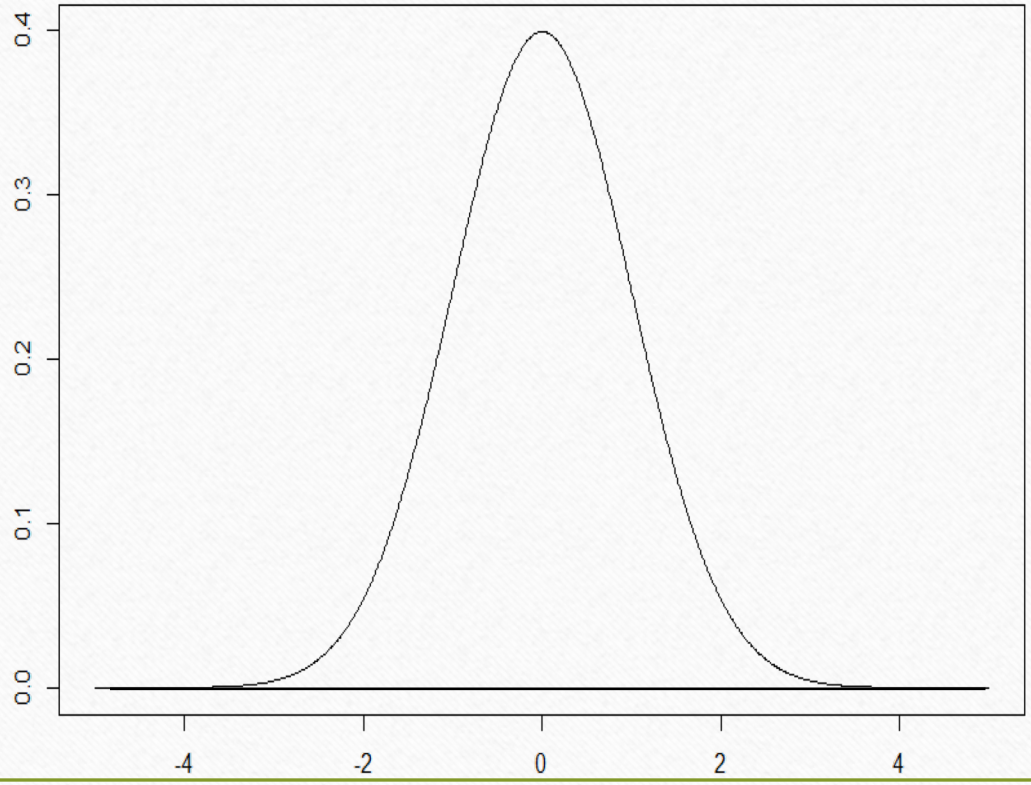
- The distribution of the standardized residuals is approximately normal with mean 0 and variance 1.

$$e_{is} \sim N(0,1), \text{ approximately}$$

STA 674, RADOE:
Assessing Model Assumptions

Standardized Residuals

If the assumptions about the errors are satisfied then the standardized residuals follow a standard normal distribution.

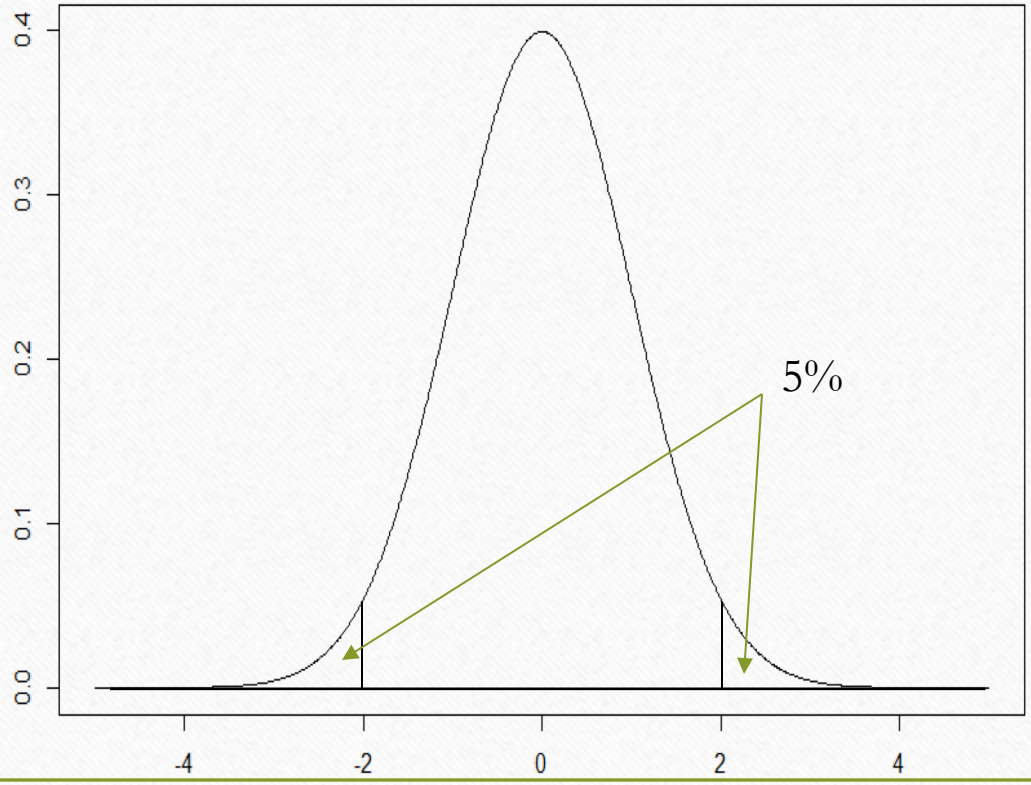


STA 674, RADOE:
Assessing Model Assumptions

Standardized Residuals

If the assumptions about the errors are satisfied then the standardized residuals follow a standard normal distribution.

We expect about 5% of the residuals to be < – 2 or > 2.



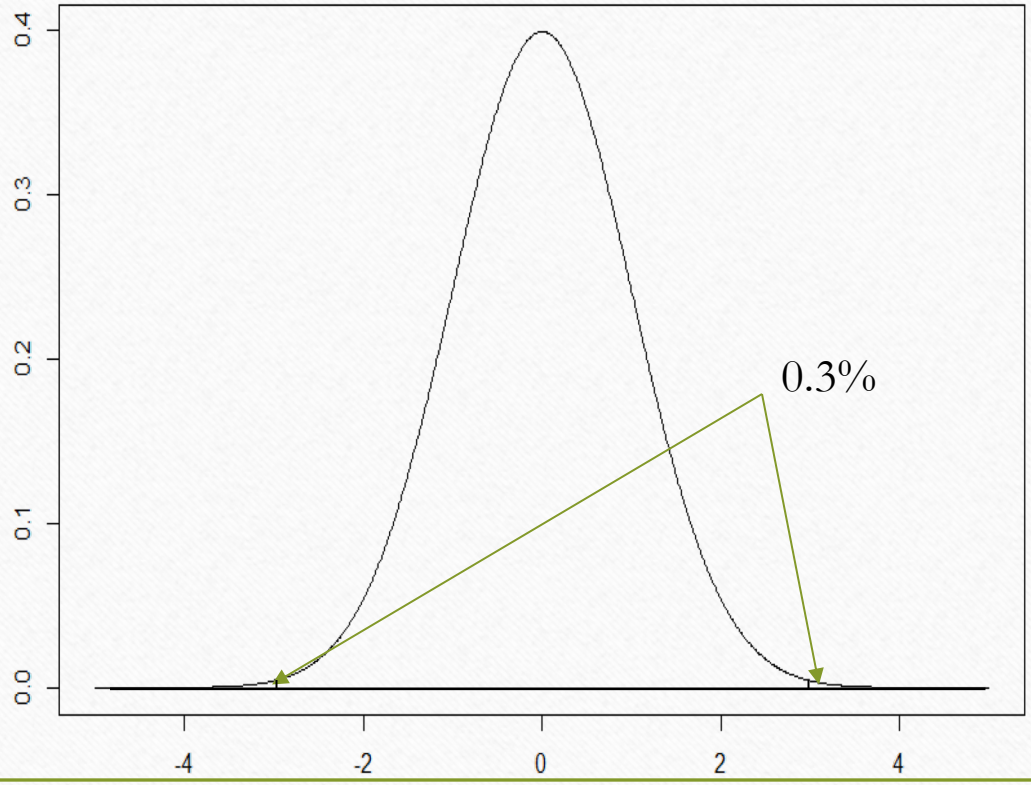
STA 674, RADOE:
Assessing Model Assumptions

Standardized Residuals

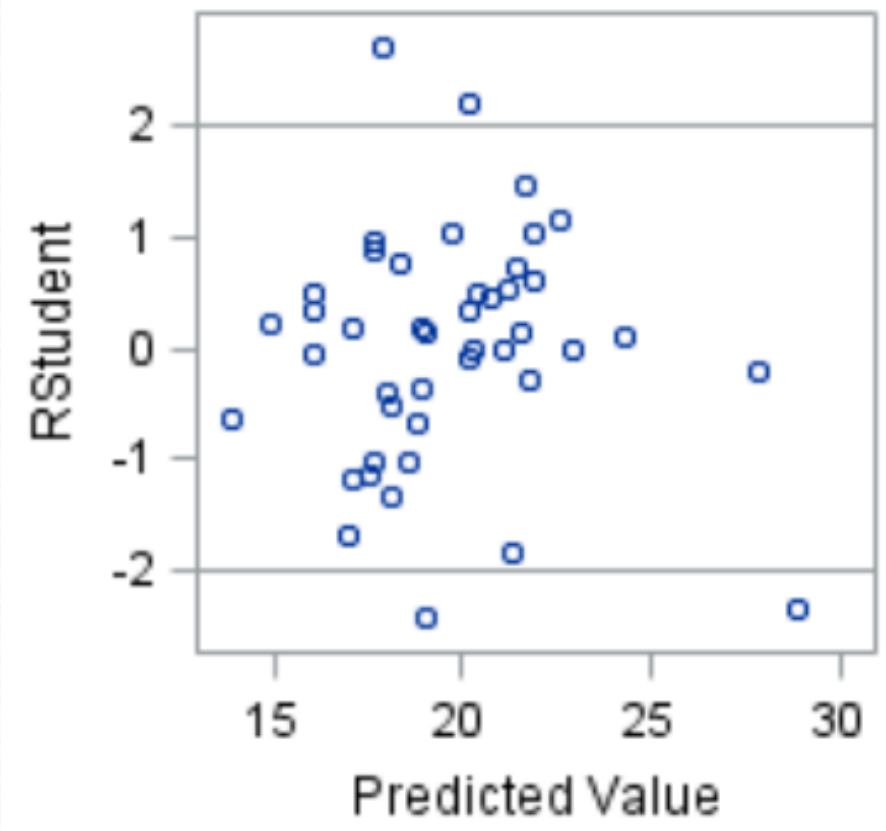
If the assumptions about the errors are satisfied then the standardized residuals follow a standard normal distribution.

We expect about 5% of the residuals to be < – 2 or > 2.

We expect about 0.3% of the residuals to be < – 3 or > 3.



Example – Cigarette Sales and Lung Cancer



Standardized residuals...or...studentized

here n=42...we would expect ~2 points outside -2 to 2...here there are 4 which is concerning...we don't necessarily delete these, but we examine further

Outlier: unusual y
ei = yi - yi hat
data points are either outliers
or they are not

STA 674, RADOE:
Assessing Model Assumptions

Definition

- A point is an **influential observation** if the value(s) of the *predictor*(*s*) for this point are different from the value(s) for the other points.
- ***Influential observation*** refers to a point with unusual x values.

Influential points refers to unusual x...with MLR a data point can either be influential in one predictor variable, but not others