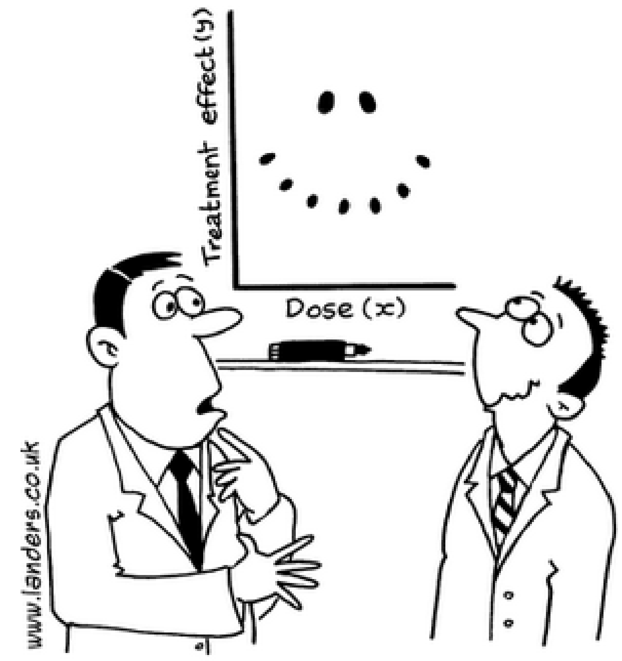
Inference for linear regression



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Overview

Quick review of regression models

Inference on regression models

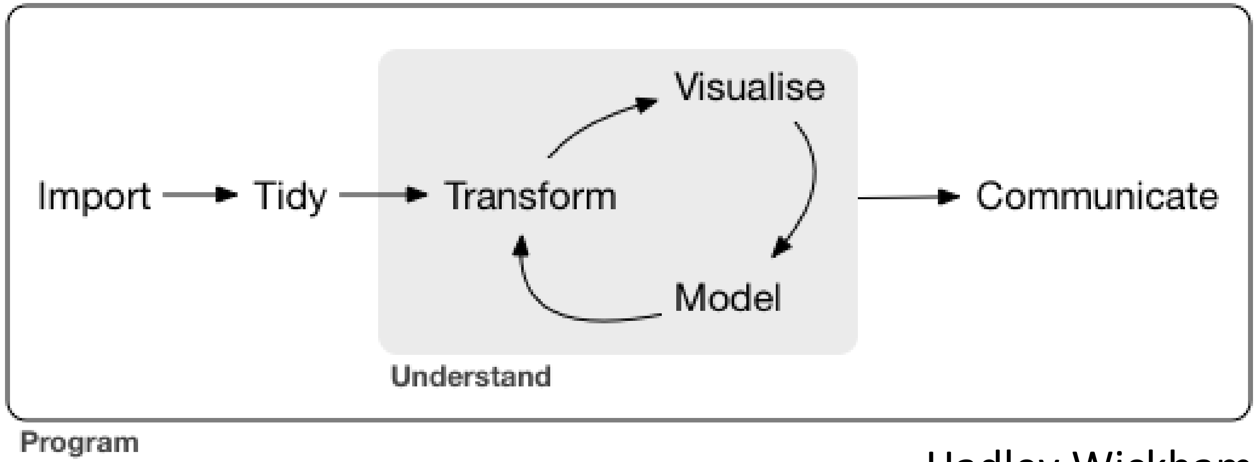
- Confidence intervals and predictions intervals

Regression diagnostics

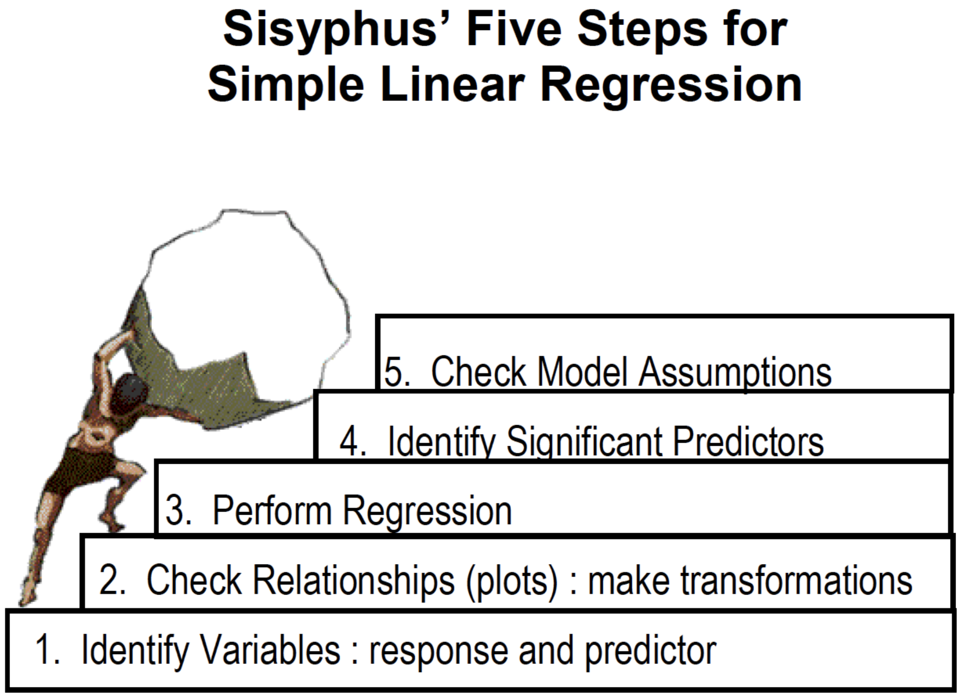
If there is time: statistics for identifying unusual observations

Linear regression continued…

The process of building regression models



Hadley Wickham



Jonathan Reuning-Scherer

"All models are wrong, but some are useful"
- George Box

The process of building regression models

Choose the form of the model

- Identify the response variable (y) and explanatory variables (x's)
- For exploratory analyses, graphical displays can help suggest the model form

Fit the model to the data

- Estimate model parameters, usually using least squares (minimize the SSRes)

Assess how well the model describes the data

- Analyze the residuals, compare to other models, etc.
- If model doesn't fit well, go to step 1.
  - This is as much an art as a science

Use the model to address questions of interest

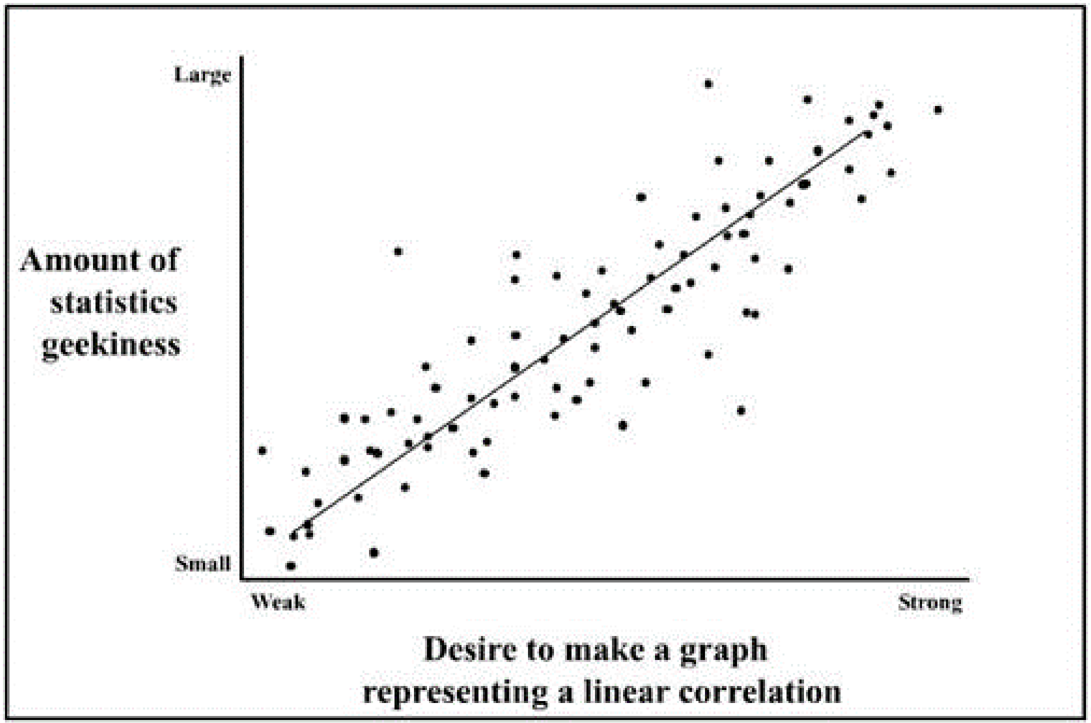
- Make predictions
- Explore relationships between response variable (y) and explanatory variables (x)
- Keep in mind limitations of the model
  - e.g., can be difficult to make the claim that changes in x *cause* changes in y from *observational data*

Review of underlying models and inference

Review: Linear regression

In **linear regression** we fit a regression line to the predict a variable y, from other variables x

- e.g., $\hat{y} = b_0 + b_1 \cdot x$



Review: Linear regression underlying model

Intercept Slope } *Parameters*

True regression line: $\mu_Y = \beta_0 + \beta_1 x$

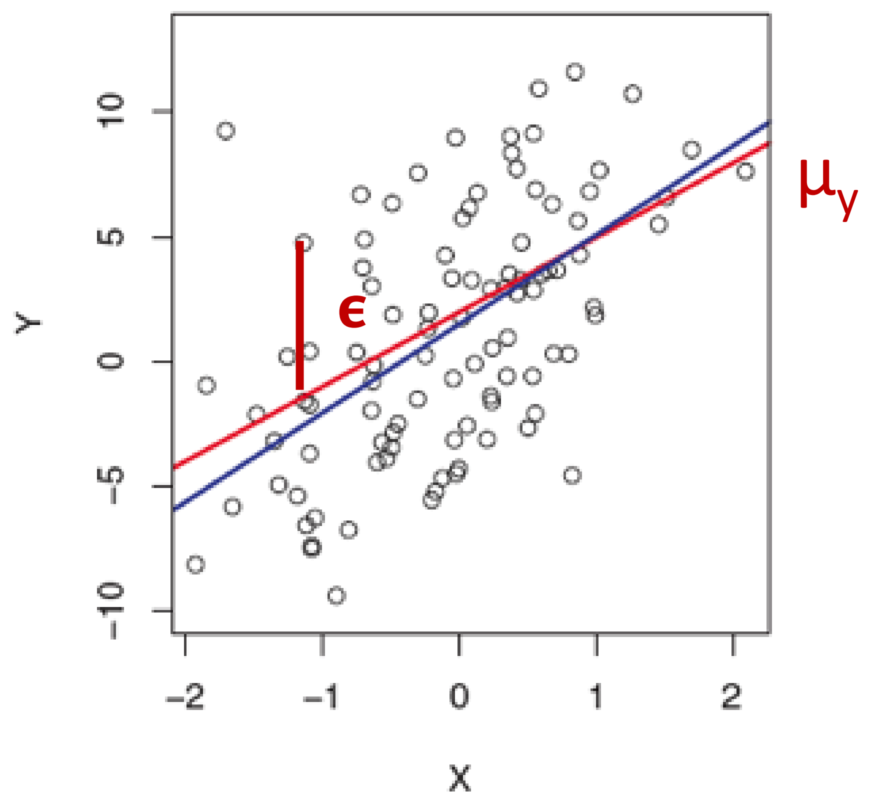
Error

Observed data point: $Y = \beta_0 + \beta_1 x + \epsilon$

$$= \mu_Y + \epsilon$$

Errors ϵ are the difference between the **true regression line** μ_y and observed data points Y

- $\epsilon = Y - \mu_y$



Review: Linear regression underlying model

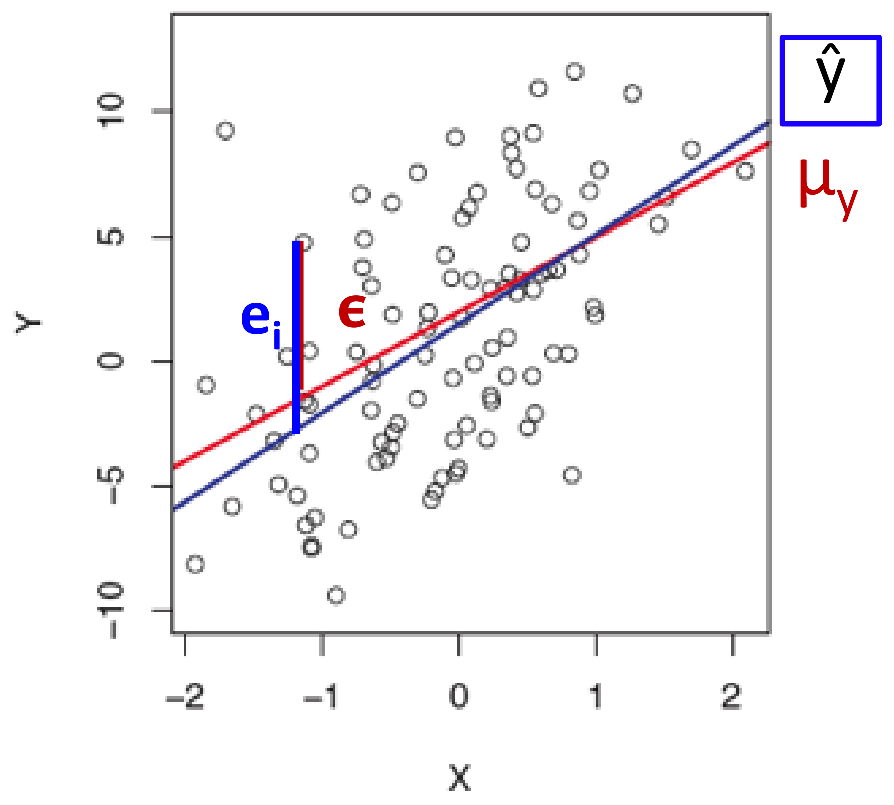
Intercept Slope } *Parameters*

True regression line: $$\mu_Y = \beta_0 + \beta_1 x$$

Error

Observed data point: $$Y = \beta_0 + \beta_1 x + \epsilon$$

Estimated regression line: $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Errors ϵ are the difference between the **true regression line** μ_y and observed data points Y

- **ϵ** = Y - μ_y

Residuals e_i are the difference between the **estimated regression line** ŷ and observed data points Y

- e_i = Y - ŷ

Review: Standard deviation of the errors: σ_ε

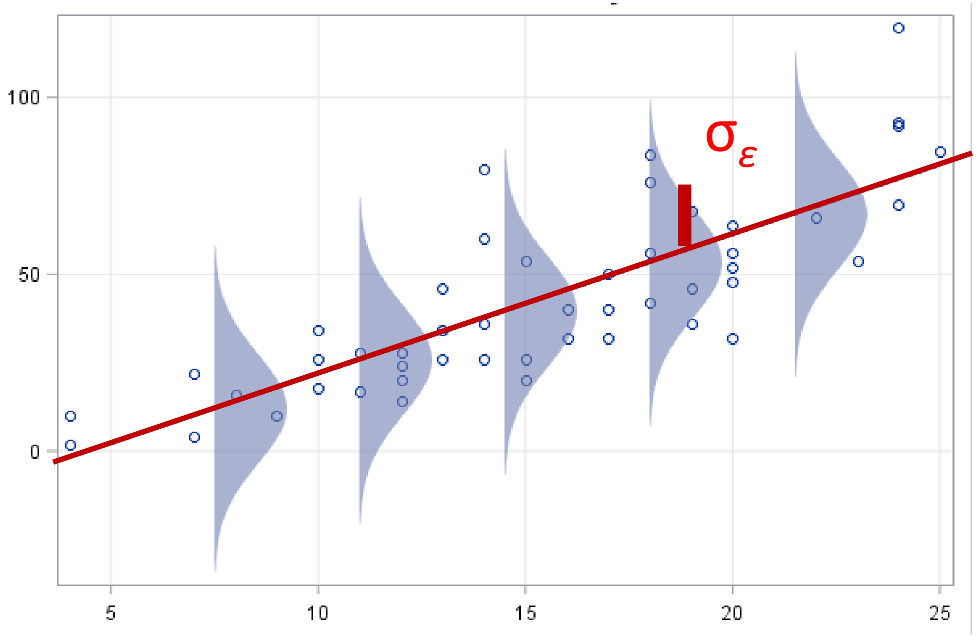
The standard deviation of the errors is denoted $\boldsymbol{\sigma_\varepsilon}$

We can use the **standard deviation of residuals** as an estimate standard deviation of the errors σ_ε .

- $\hat{\sigma}_\varepsilon$ often called the ~~"residual standard error"~~
- $\hat{\sigma}_\varepsilon$ we called it the "residual standard deviation"

$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2} SSRes}$$

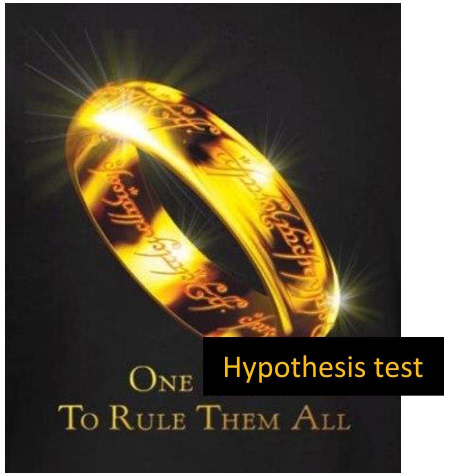
$$= \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$



Review: Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (no linear relationship between x and y)
- H_A: $\beta_1 \neq 0$



One type of hypothesis test we can run is based on a t-statistic:

- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

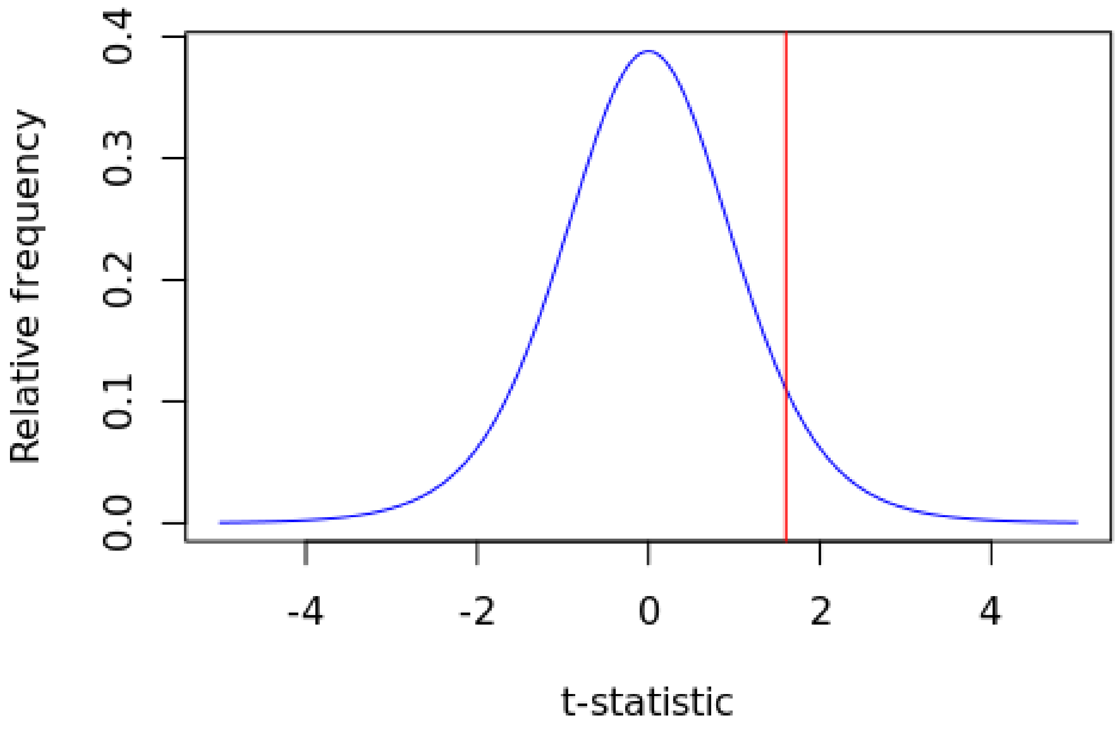
$$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}_{\hat{\beta}_1}}$$

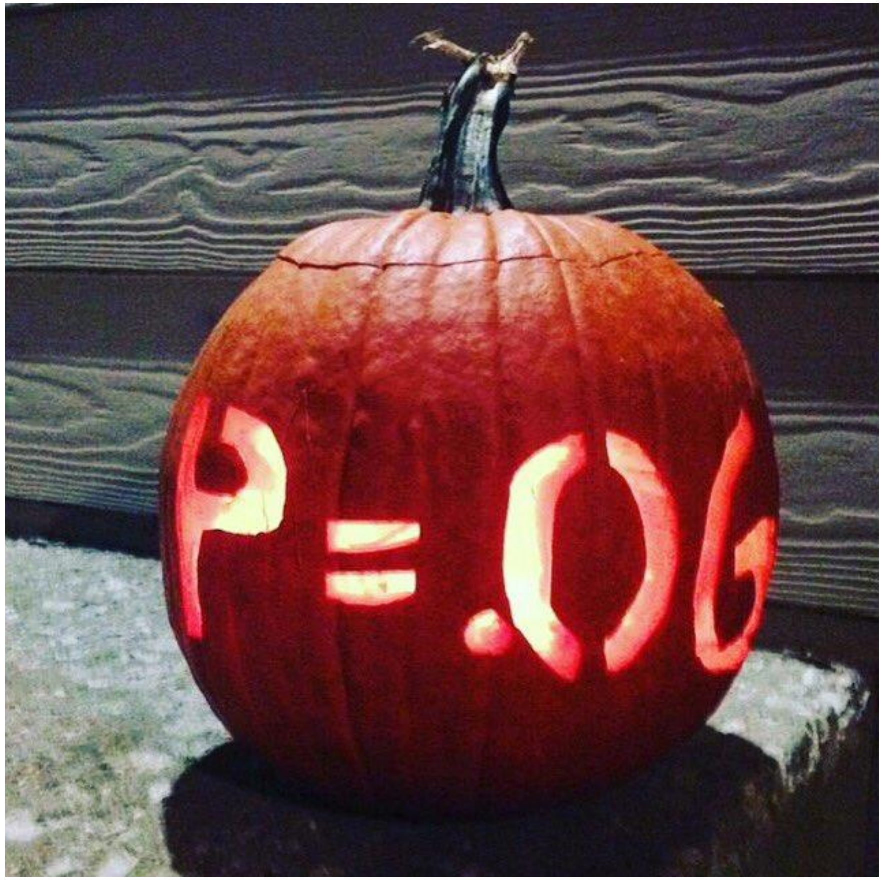
$$\hat{SE}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$\hat{SE}_{\hat{\beta}_0} = \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Review: Hypothesis test for regression coefficients

Step 4: Get a p-value by assessing whether our t-statistic comes from a null t-distribution





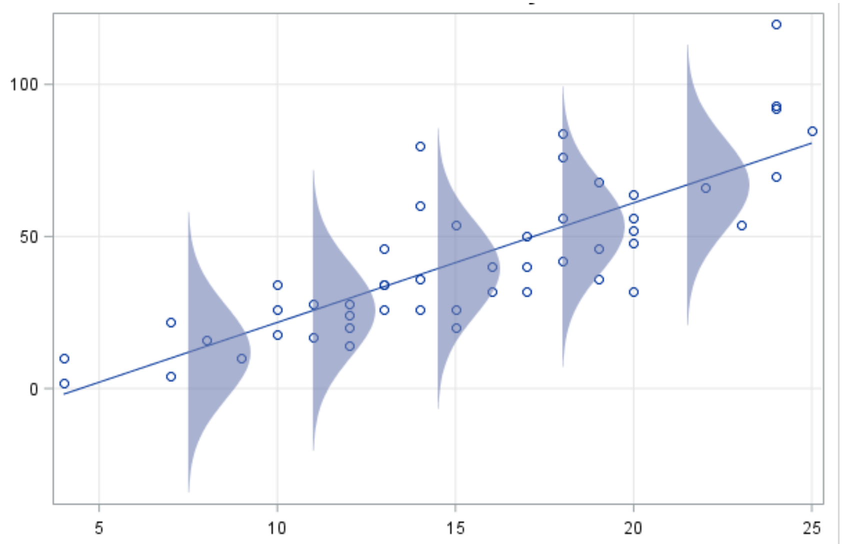
Review: Inference using parametric methods

When using parametric methods, we make the following (LINE) assumptions:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_\epsilon)$$



These assumptions are usually checked after the models are fit using 'regression diagnostic' plots.

Review: Simple linear regression in R

Faculty salaries…

```
lm_fit <- lm(salary_tot ~ log_endowment, data = assistant_data)
summary(lm_fit)
```

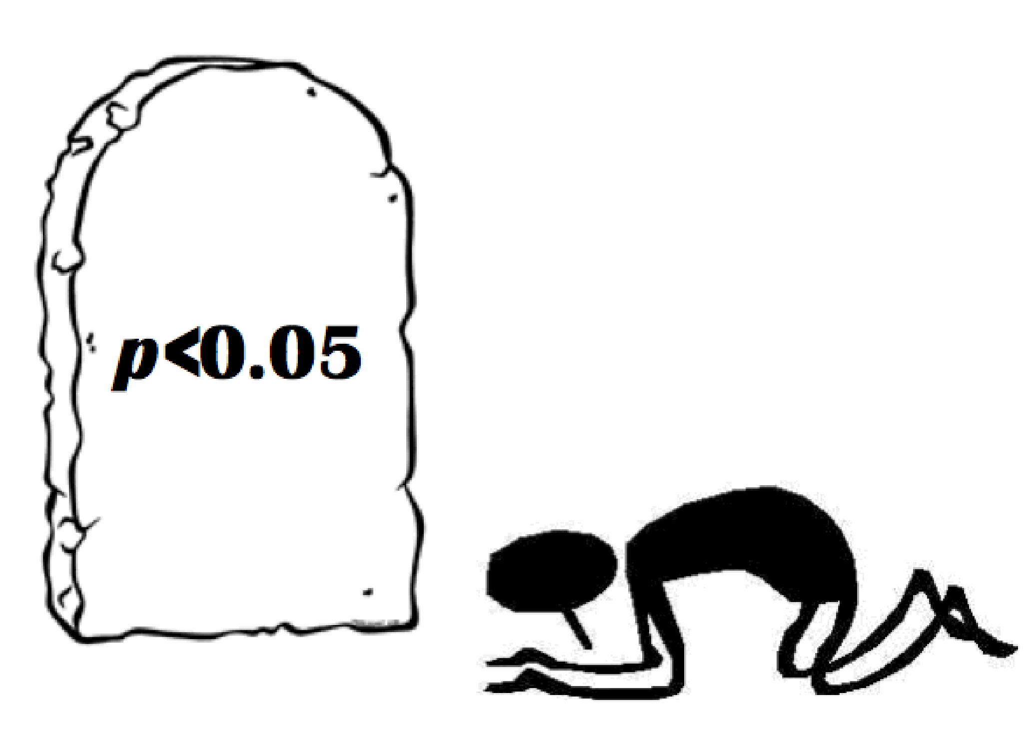
Coefficients:

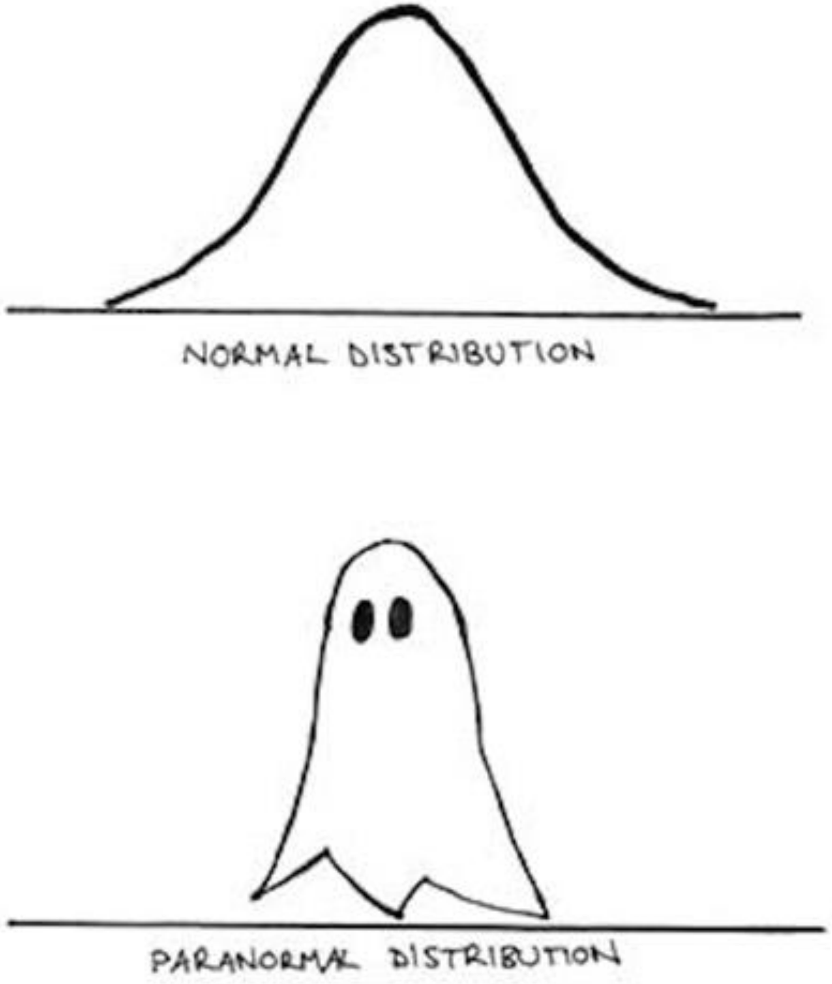
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-26761.7	3118.4	-8.582	<2e-16 ***
log_endowment	11350.1	410.6	27.646	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard ~~error~~: 13190 on 1173 degrees of freedom

Review: Step 5 - conclusions?





Inference for linear regression: confidence intervals

Inference for linear regression: confidence intervals

We can estimate three types of intervals for a regression:

1. Confidence intervals for the regression coefficients: β_0 and β_1

2. Confidence intervals for the full line $\mu_Y(x)$

3. Prediction intervals where most of the data is expected

Confidence intervals for regression coefficients

For the slope coefficient , the confidence interval is: $\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$

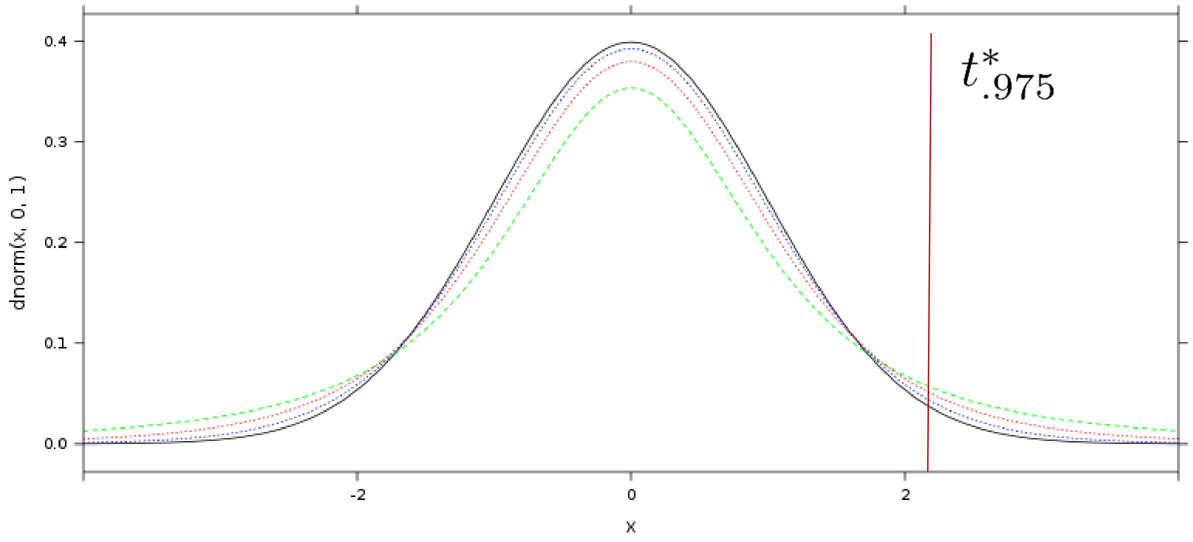
Where: $SE_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$

β_1

t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence level

qt(.975, df)

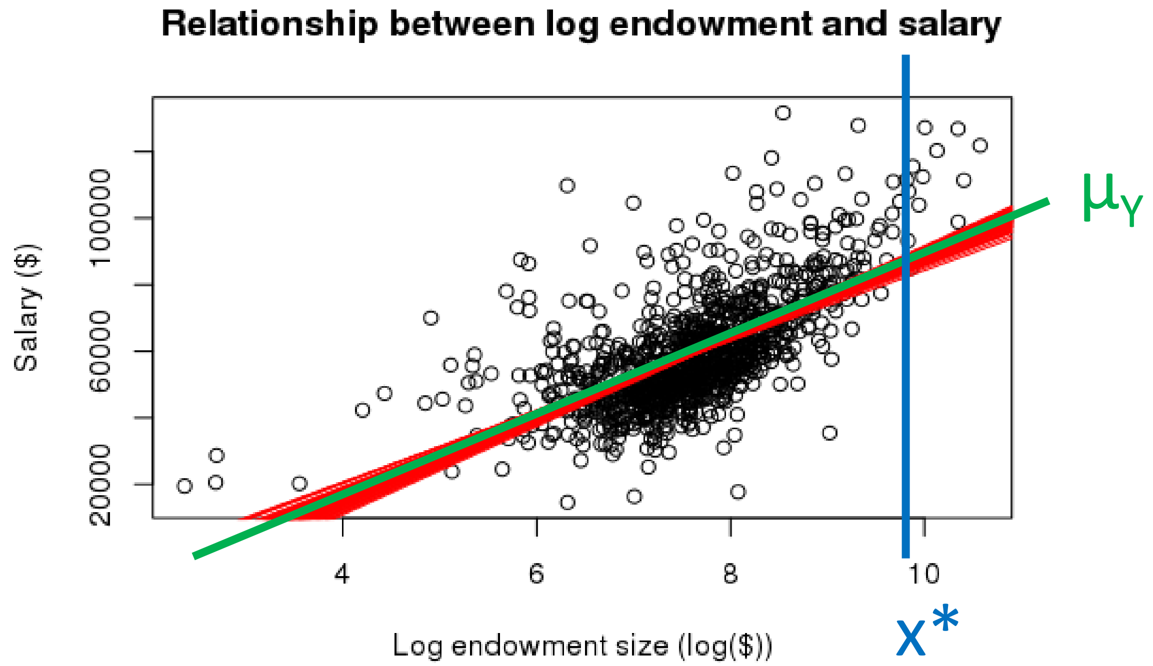
N(0, 1)
df = 2
df = 5
df = 15



Confidence intervals for the regression line μ_Y

A confidence interval for the mean response for the ***true regression line*** μ_Y at the value of x* is:

$$\hat{y} \pm t^* \cdot SE_{\hat{\mu}} \quad \text{where} \quad SE_{\hat{\mu}} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$



Note:

- There is more uncertainty at the ends of the regression line
- The confidence interval for the regression line μ_Y is different than the confidence interval for slope β_1

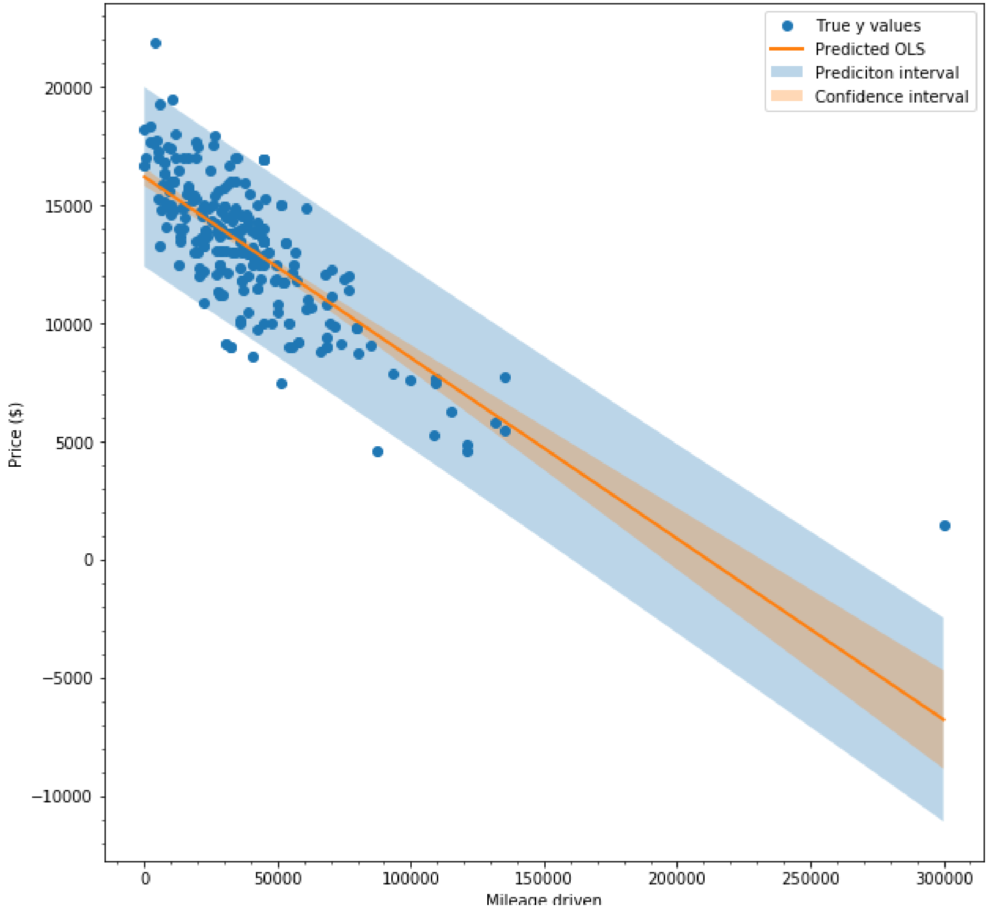
Prediction intervals

Confidence intervals give us a measure of uncertain about our the true relationship between x and y for:

- The true regression slope β_1
- The true regression line μ_Y

Prediction intervals give us a range of plausible values for y

- i.e., 95% of our y's with be within this range



Prediction intervals

A **prediction intervals** for the y can be calculated using:

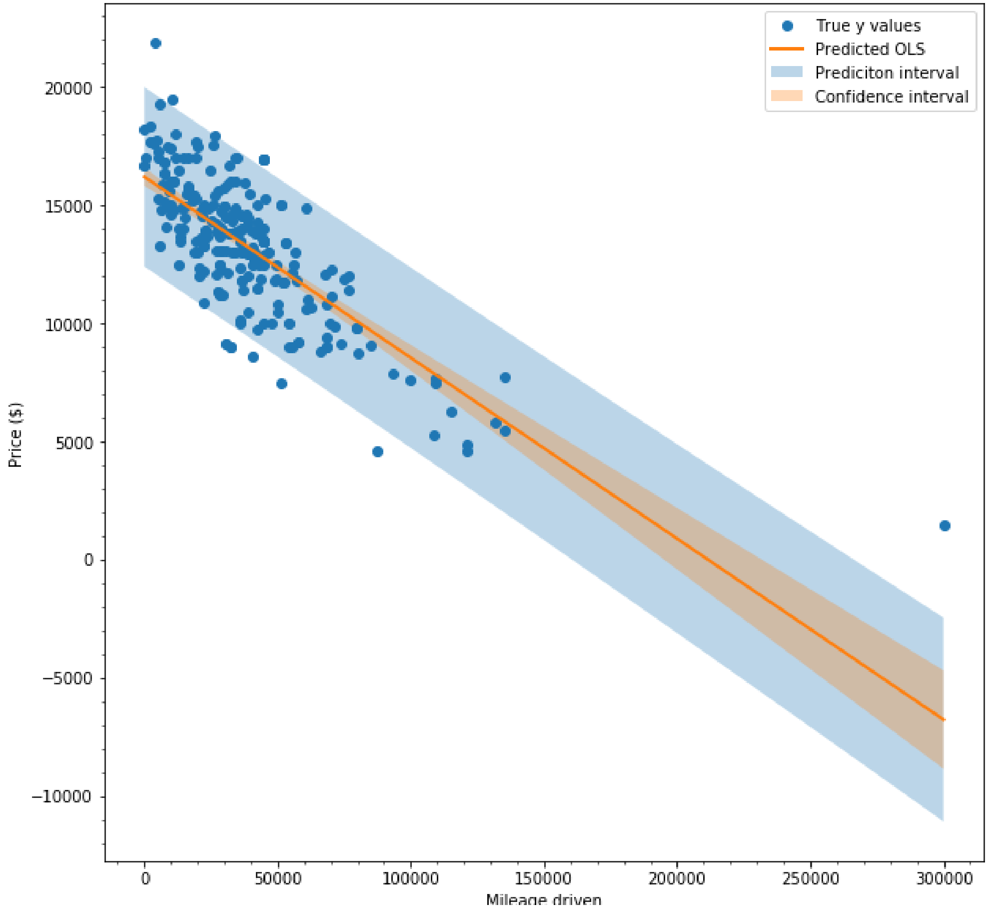
$$\hat{y} \ \pm \ t^* \cdot SE_{\hat{y}}$$

where

$$SE_{\hat{y}} = \sigma_{\epsilon}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$

Due to y's scattering around the true regression line

Due to uncertainty in where the true regression line is



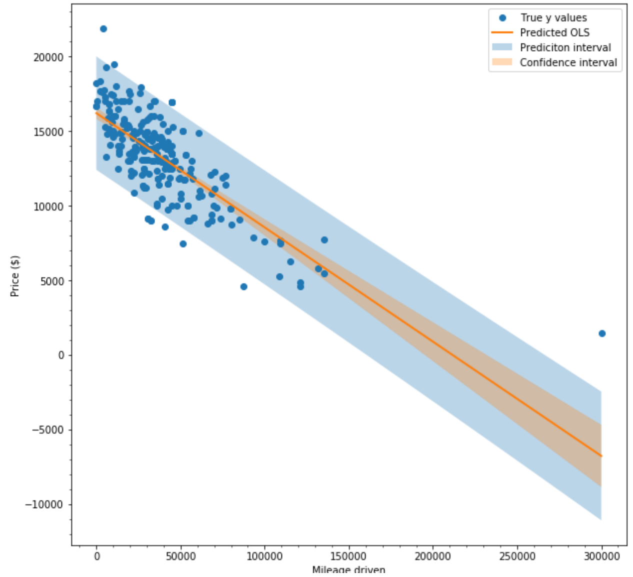
Summary of confidence and prediction intervals

1. CI for Slope β

$$\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1} \qquad SE_{\hat{\beta}_1} = \sigma_\epsilon \sqrt{\frac{1}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

2. CI for regression line μ_Y at point x*

$$\hat{y} \ \pm \ t^* \cdot SE_{\hat{\mu}} \qquad SE_{\hat{\mu}} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$



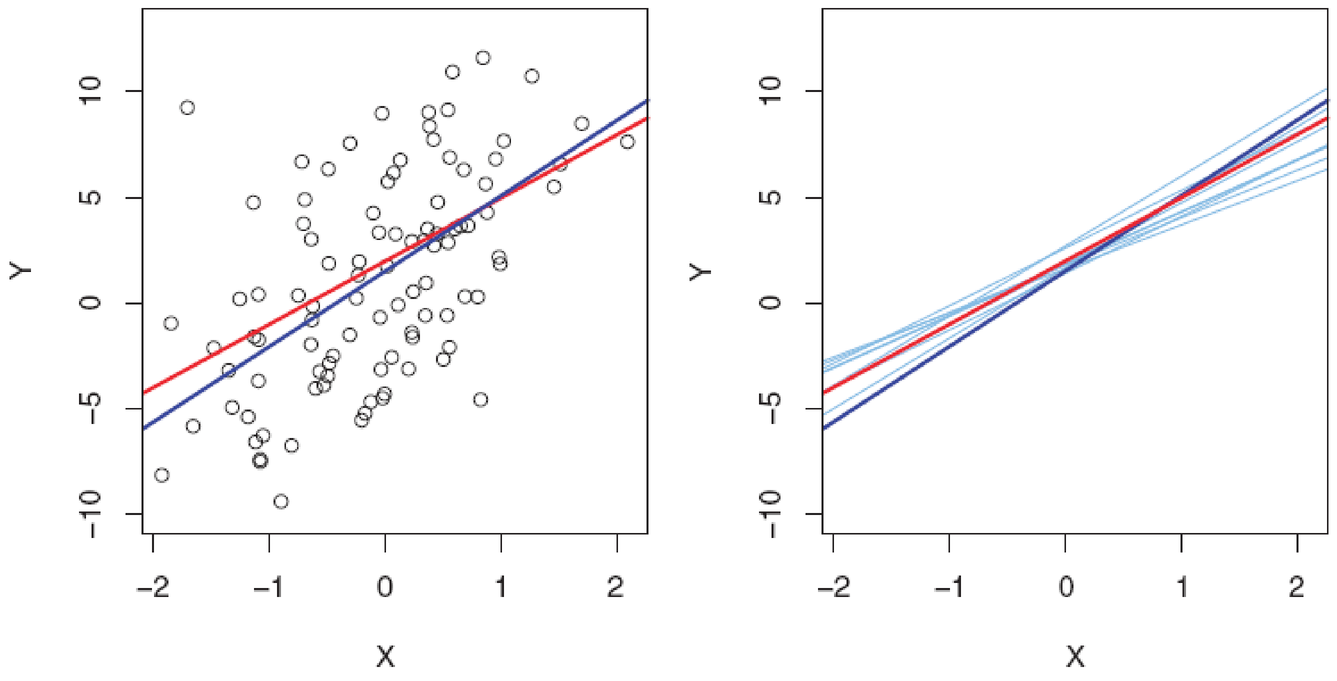
3. Prediction interval y

$$\hat{y} \ \pm \ t^* \cdot SE_{\hat{y}} \qquad SE_{\hat{y}} = \sigma_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

Resampling methods for inference in regression

We can also use resampling methods to estimate run hypothesis tests and create confidence intervals for the regression coefficients

- Bootstrap
- Permutation test



Let's look at inference for simple linear regression in R

More faculty salary data!

- We will start at part 3

Dall-e 2022

Dall-e 2023

Regression diagnostics

Regression diagnostics

We use diagnostics to see if the assumptions/conditions for inference are met

- If they aren't met, we can adjust the model and try again

Choose

Fit

Assess

Use

Regression diagnostics

Let's go through the 4 conditions that should be met when using parametric methods for inference:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

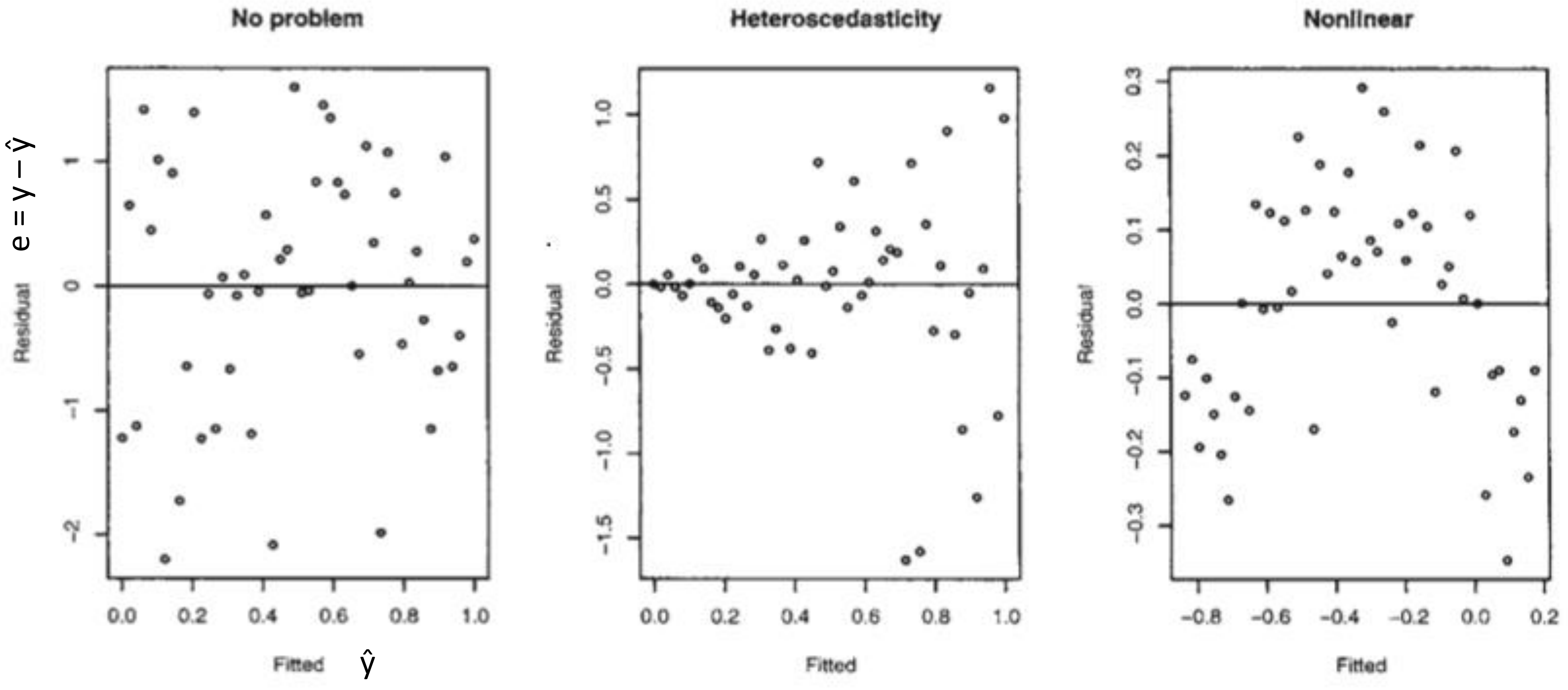
Regression diagnostics

Let's go through the 4 conditions that should be met when using parametric methods for inference:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

We can check linearity and homoscedasticity by plotting the residuals as a function of the fitted values

Checking linearity and homoscedasticity

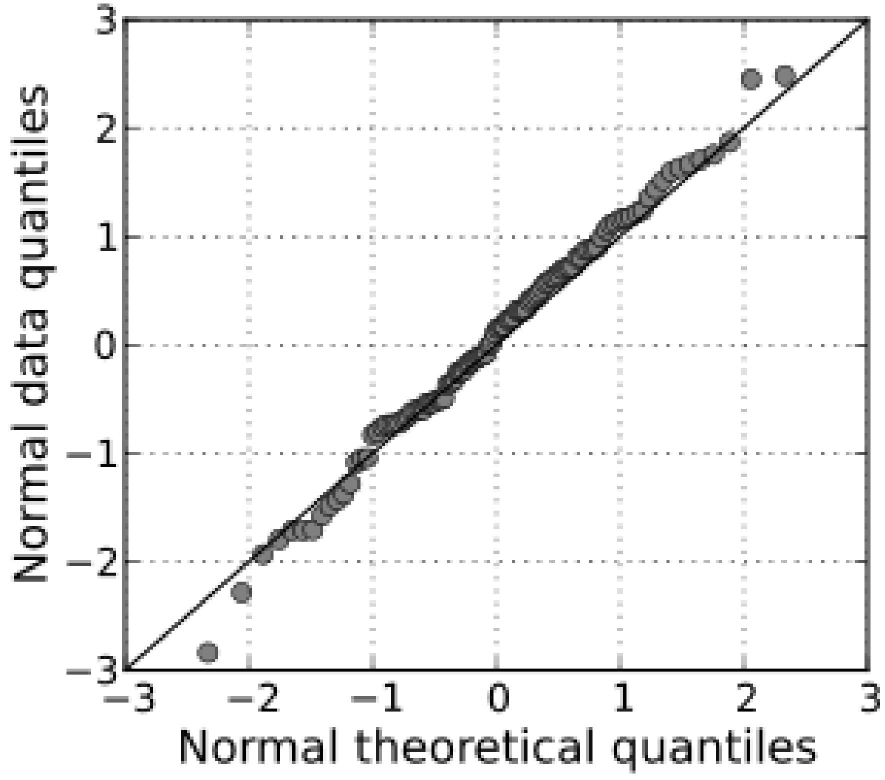


Checking normality

Normality: residuals are normally distributed around the predicted value ŷ

We can check this using a Q-Q plot

The ‘car’ package has a nice function for making qqplots called qqPlot()



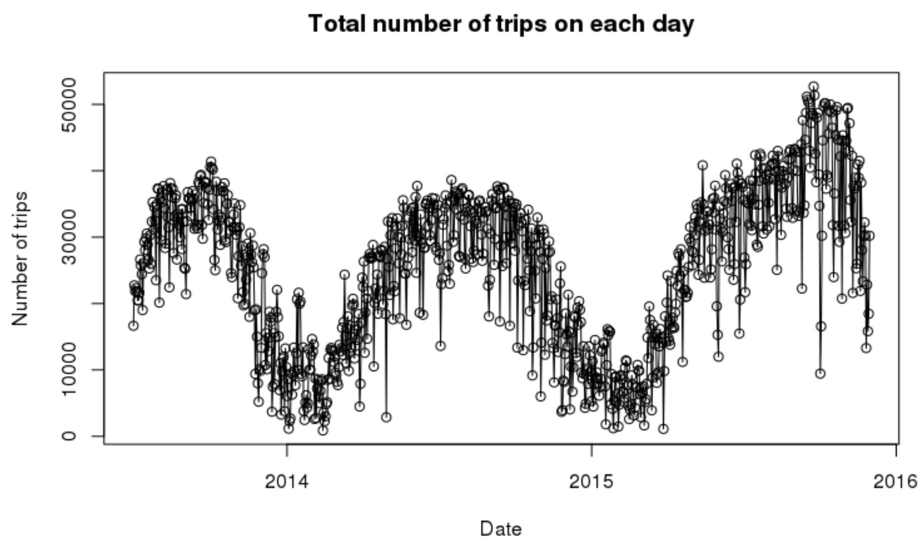
Checking Independence

To check whether each data point is independent requires knowledge of how the data was collected

- Simple random sample from the population is likely independent
- Time series often are not independent

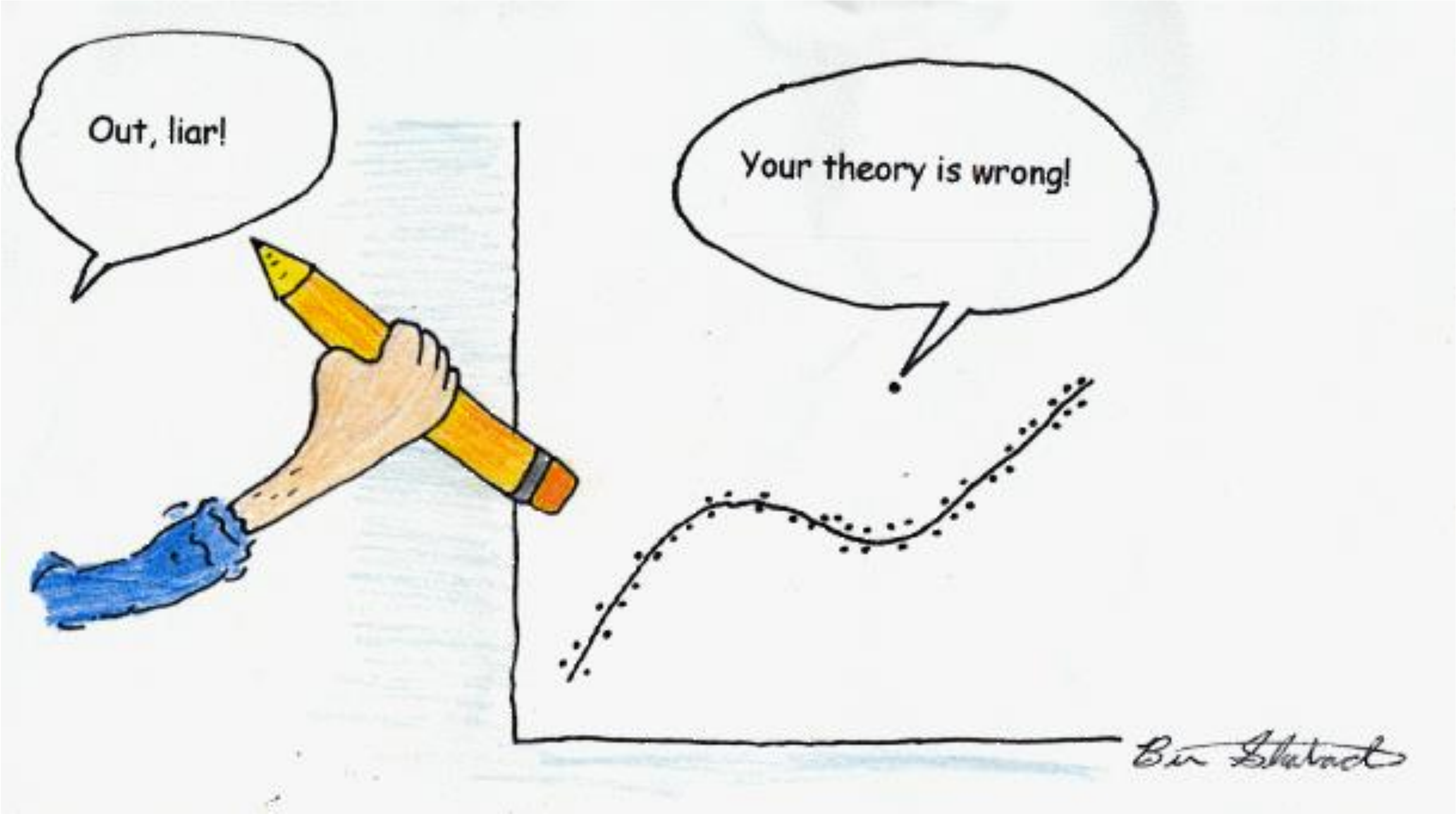
We have basically been assuming independence for everything we have done in this class

- i.i.d. independent and identically distributed



Let’s examine these diagnostic plots in R!

Statistics for unusual observations



Statistics for unusual observations

There are statistics that are useful for flagging usual observations

- **Outliers (large residuals):** unusual **y** values
- **High leverage points**: usual **x** values
- **Influential points**: both an outlier and a high leverage

Unusual observations can indicate:

- An error in data processing
- A need to modify the model
- An interesting phenomenon

Unusual observations **can also have a big effect on the model fit**

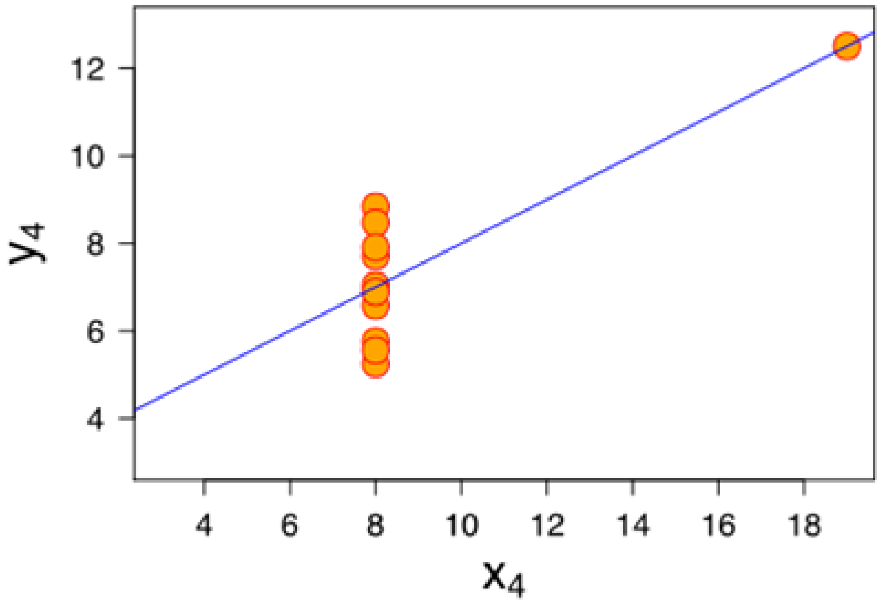
- E.g., a big effect on $\hat{\beta}_0$ $\hat{\beta}_1$

Leverage

High leverage points are predictors **x** that are far from the mean

We can calculate the leverage a data point has using the statistic:

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



$$\Sigma_{i=1}^{n} h_i = 2$$

High leverage points can have a big impact on the model that is fit!!!

Typical: $h_i = 2/n$

High: $h_i = 4/n$

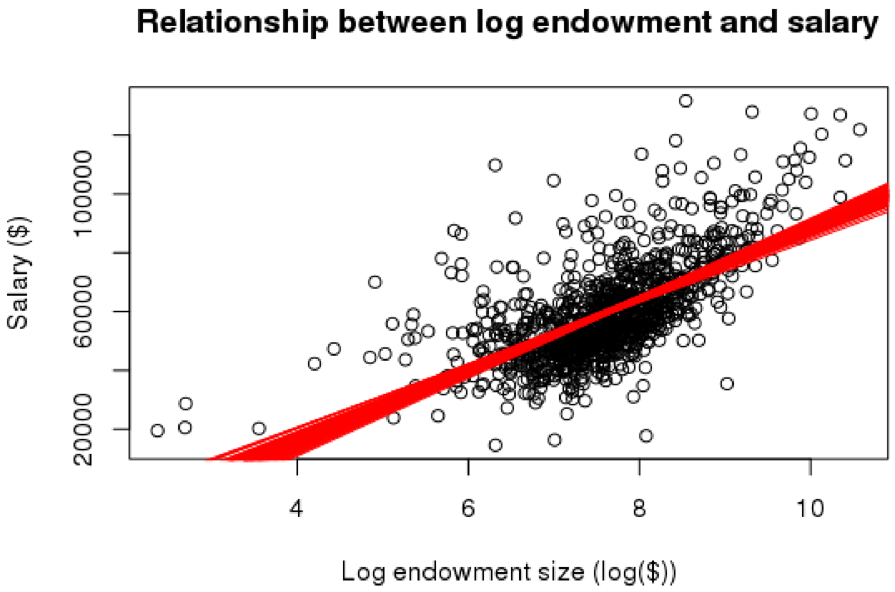
Very high: $h_i = 6/n$

R: hatvalues()

Outliers: standardized residuals

The **standardized residual** for the i^th data point in a regression model can be computed using:

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



Puts residuals on a 'normalized' scale

Makes residuals at the ends a bit larger to deal with the fact that they are 'overfit'

R: rstandard()

Outliers: studentized residuals

The **studentized residual** for the i^{th} data point in a regression model can be computed using:

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

Here $\hat{\sigma}_{(i)}$ is the an estimate of $\hat{\sigma}_{\epsilon}$ with the i^{th} point removed

Q: Why might we want to remove the i^{th} point when calculating $\hat{\sigma}_{\epsilon}$?

A: Outliers could have a big effect on our estimate of $\hat{\sigma}_{\epsilon}$

R: rstudent ()

Cook's distance

The amount of influence a point has on a regression line depends on:

- The size of the residual e_i
- The amount of leverage h_i

Cook's distance is a statistic that captures how much influence a point has on a regression line

$$D_i = \frac{(stdres_i)^2}{k+1} \frac{h_i}{1-h_i}$$

Larger for larger residuals (outliers)

Larger for high leverage points

Where *k* is the number of predictors in the model

R: cooks.distance ()

- For simple linear regression k = 1 (just a single predictor x)

Cook's distance

The amount of influence a point has on a regression line depends on:

- The size of the residual e_i
- The amount of leverage h_i

Cook's distance is a statistic that captures how much influence a point has on a regression line

$$D_i = \frac{(stdres_i)^2}{k+1} \frac{h_i}{1-h_i}$$

Larger for larger residuals (outliers)

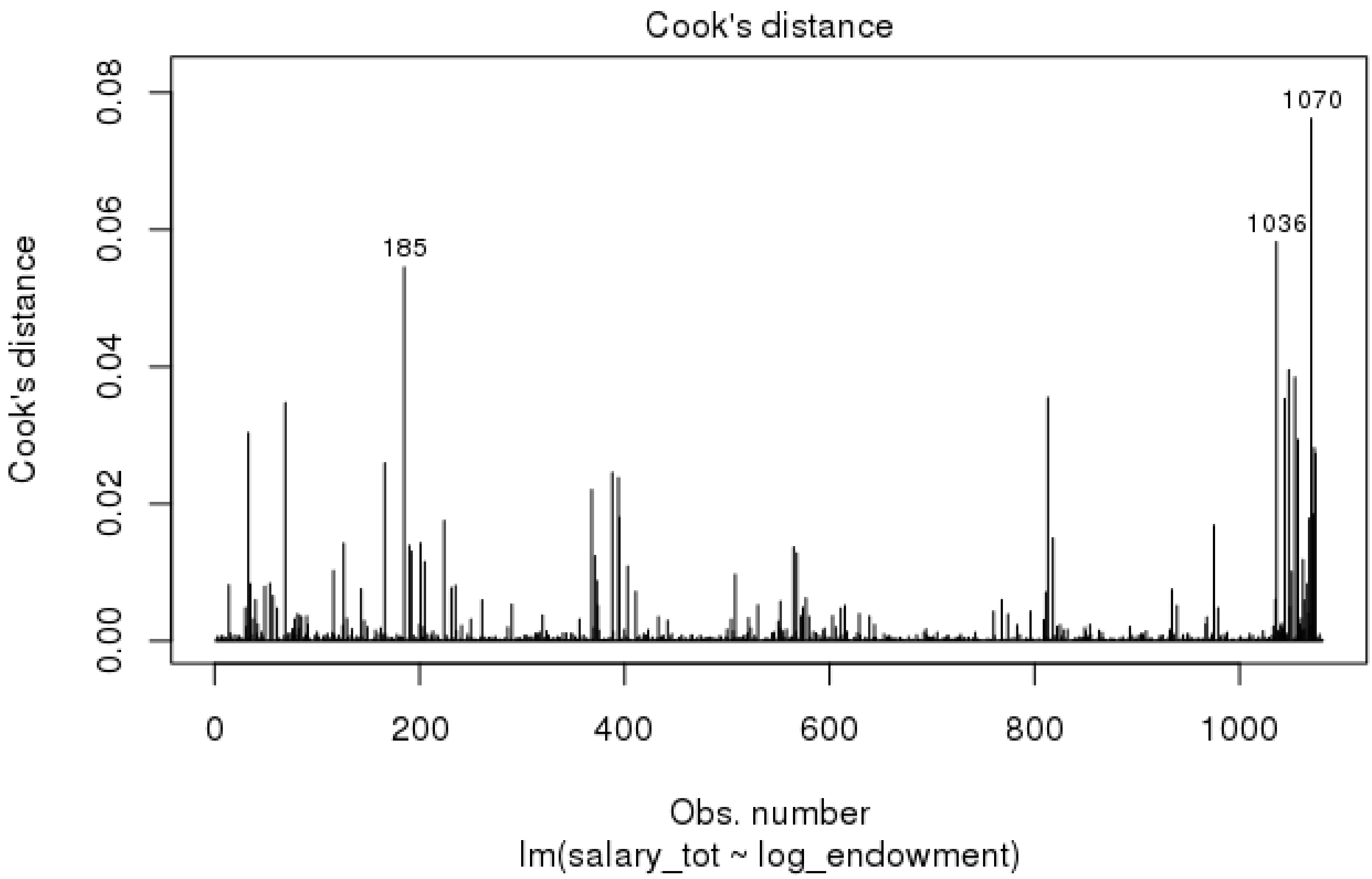
Larger for high leverage points

Rule of thumb:

- Moderately influential: $D_i > 0.5$
- Very influential: $D_i > 1$

R: cooks.distance ()

Cook's distances for salary ~ $\log_{10}$(endowment)



plot(lm_fit, 4)

Unusual points rules of thumb

Statistic	Moderately unusual	Very unusual
Leverage, h_i	Above $2(k + 1)/n$	Above $3(k + 1)/n$
Standardized residual	Beyond ± 2	Beyond ± 3
Studentized residual	Beyond ± 2	Beyond ± 3
Cook's D	Above 0.5	Above 1.0

Where:

- k is the number of explanatory variables
- n is the number of data points

From STAT2 by Cannon, Cobb, Hartlaub, Legler, Lock, Moore, Rossman, and Witmer

Questions?