Inference for linear regression

Overview

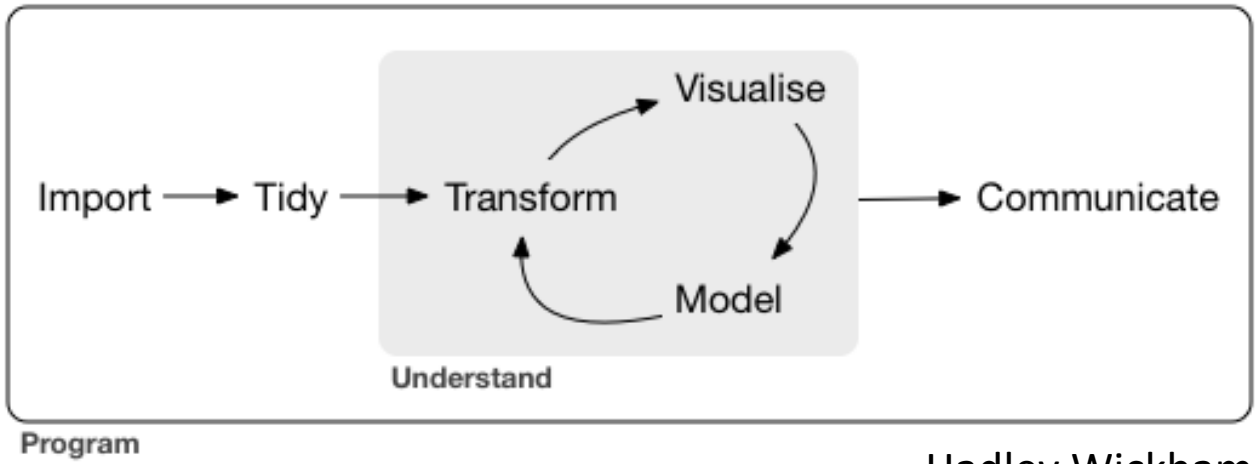
Review of regression models and regression inference

Regression diagnostics

Statistics for identifying unusual observations

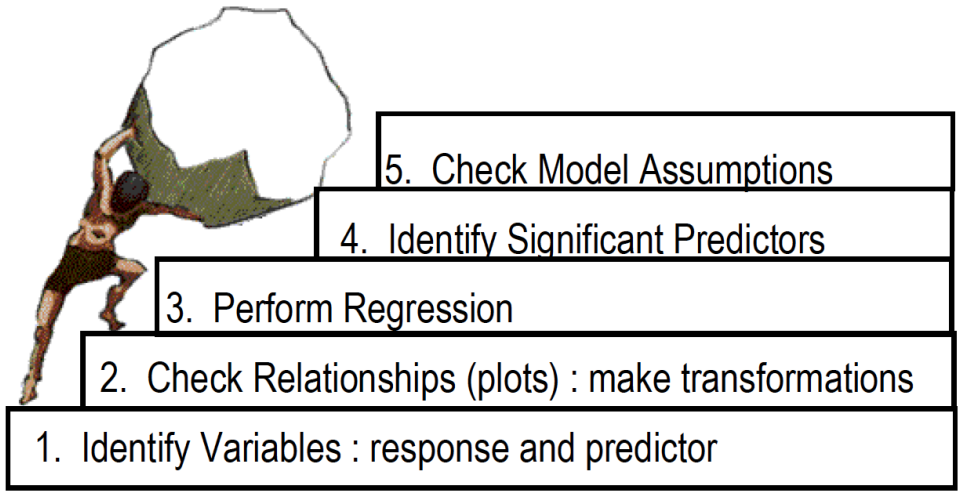
Linear regression continued…

The process of building regression models



Hadley Wickham

Sisyphus' Five Steps for Simple Linear Regression



Jonathan Reuning-Scherer

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 RSS)

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

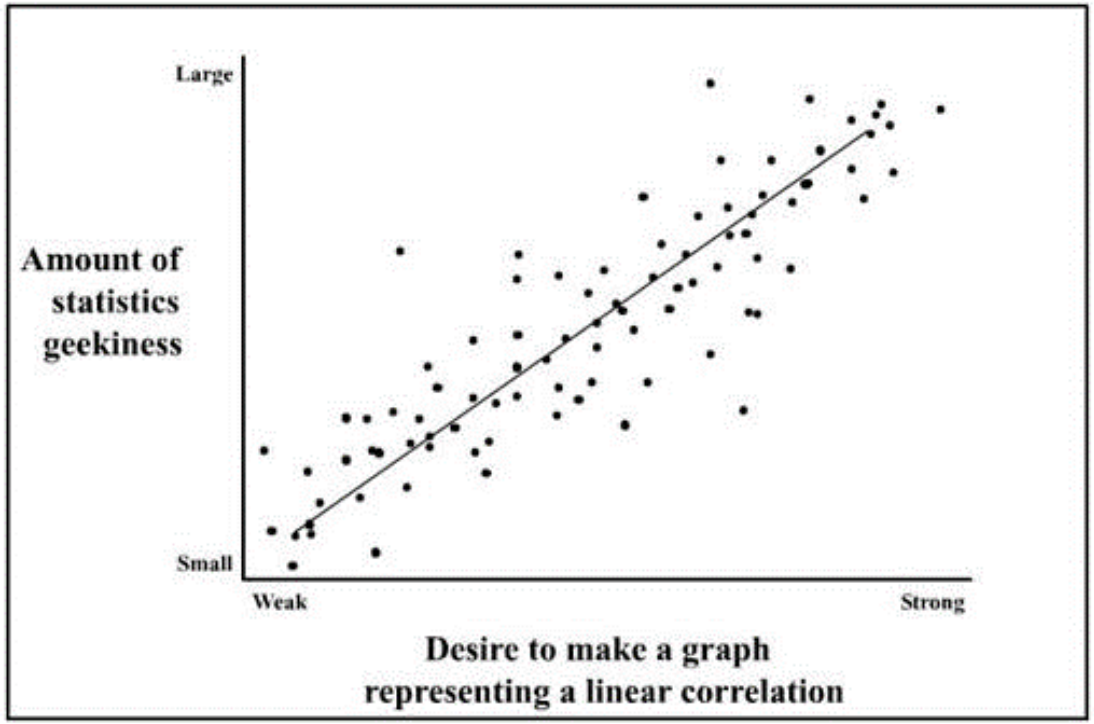
- Make predictions
- Explore relationships between response variable (y) and explanatory variables (x)
- Keep in mind limitations of the model
  - e.g., can be difficult to make the claim that changes in y *cause* changes in x from *observational data*

Review of underlying models and inference

Linear regression

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

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



Linear regression underlying model

Intercept Slope } *Parameters*

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

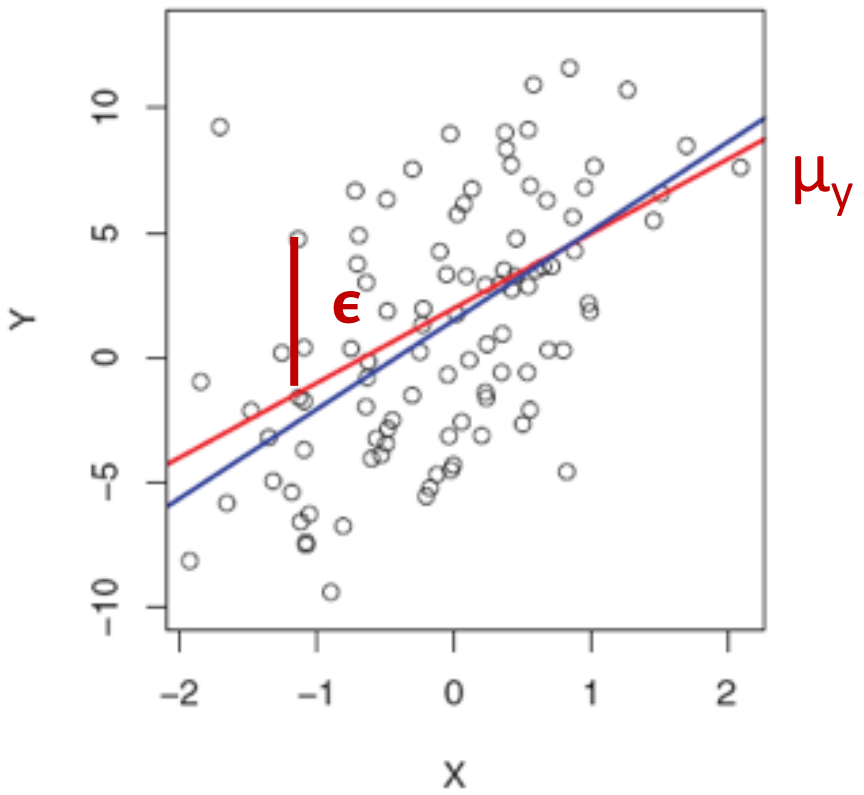
Error

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

- **ϵ** = Y - μ_y



Linear regression underlying model

Intercept Slope } *Parameters*

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

Error

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

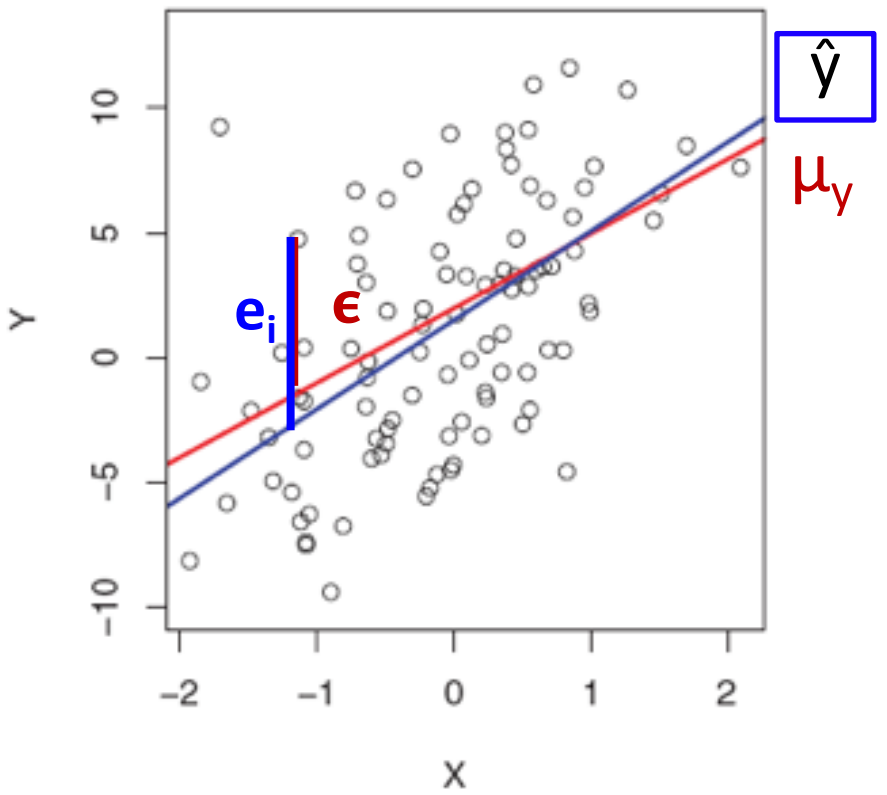
Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

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

- $\epsilon = Y - \mu_y$

Residuals e_i are the difference between the **estimated regression line** $\hat{y}$ and observed data points Y

- $e_i = Y - \hat{y}$



Standard deviation of the errors: σ_ε

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

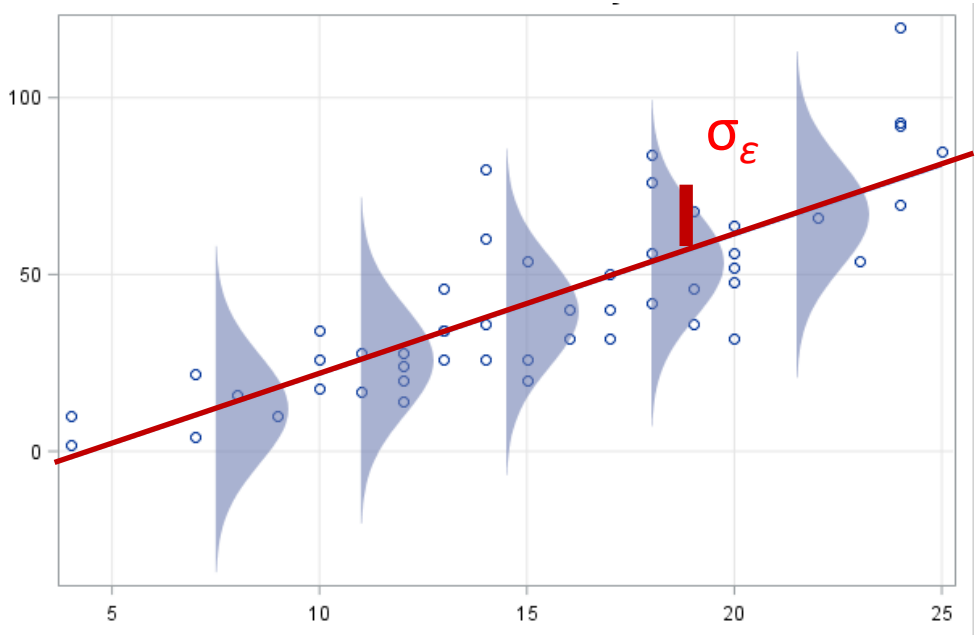
We can use the **standard deviation of residuals** as an estimate standard deviation of the errors σ_ε. This is known as the…

- **regression standard error (RSE)**

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

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

We will *assume* that the errors are **normally distributed**



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: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

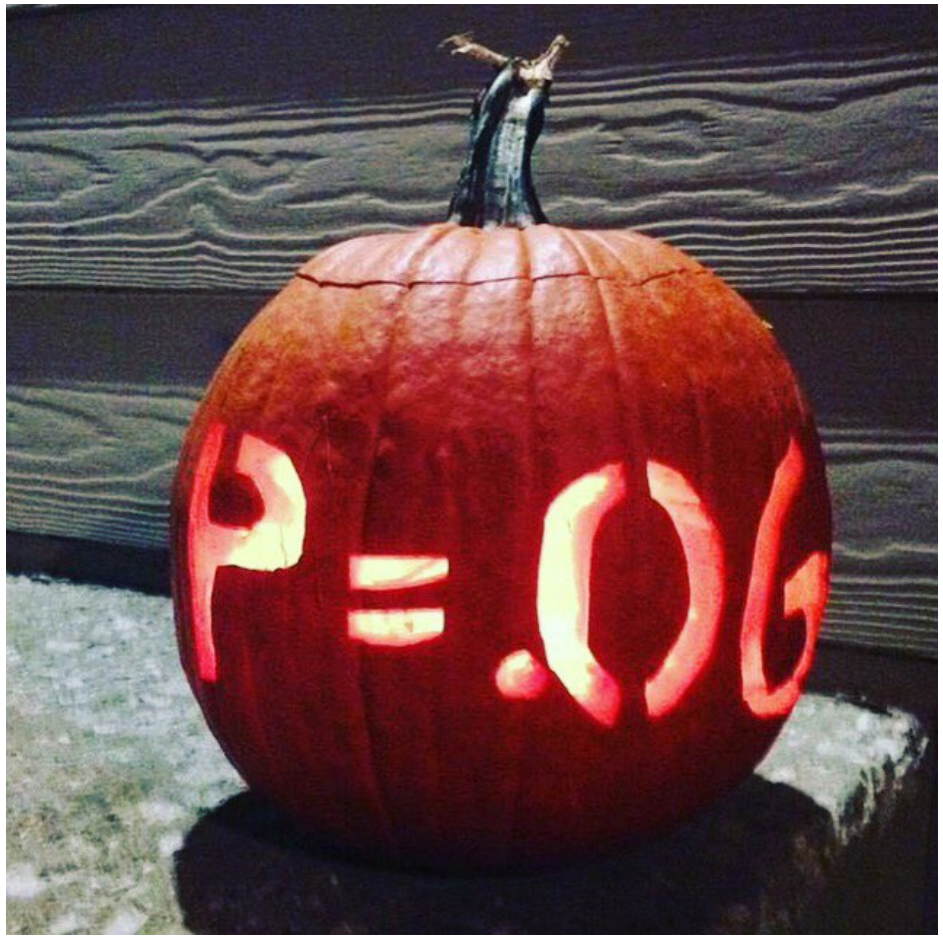
One type of hypothesis test we can run is based on a t-statistic: $t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$

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

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

$$SE_{\hat{\beta}_0} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2)}}$$

Hypothesis test for regression coefficients



Confidence and prediction intervals

1. CI for Slope β

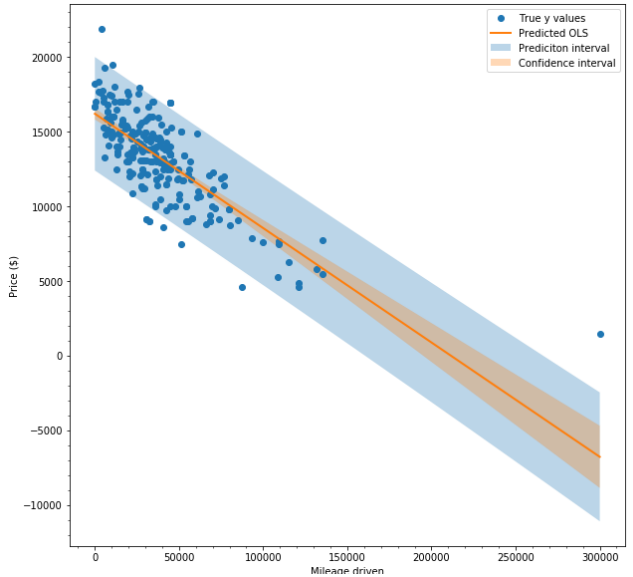
$$\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1} \qquad SE_{\hat{\beta}_1} = \sigma_\epsilon \sqrt{\frac{1}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$

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

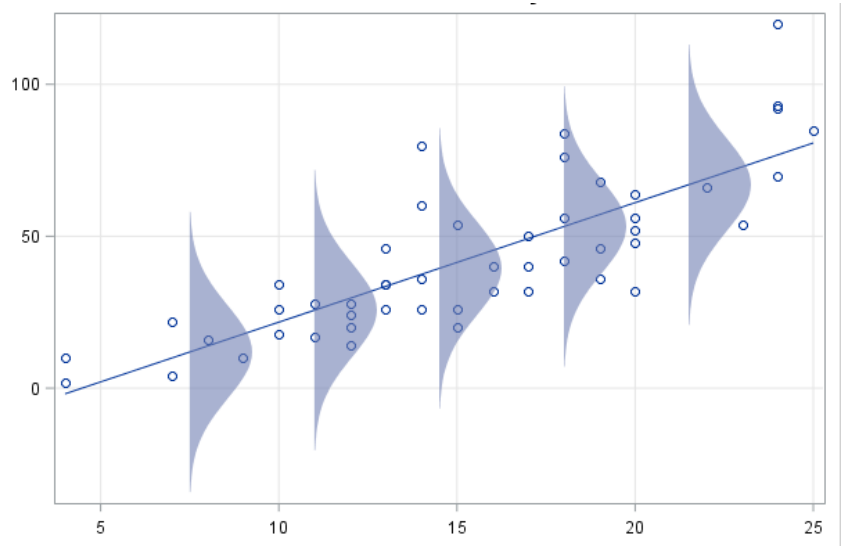
$$\hat{y} \pm t^* \cdot SE_{\hat{y}} \qquad SE_{\hat{y}} = \sigma_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$



Inference using parametric methods

When using parametric methods, we usually make the following assumptions:

- **Normality**: residuals are normally distributed around the predicted value ŷ
- **Homoscedasticity**: constant variance over the whole range of x values
- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points



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

Questions?

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:

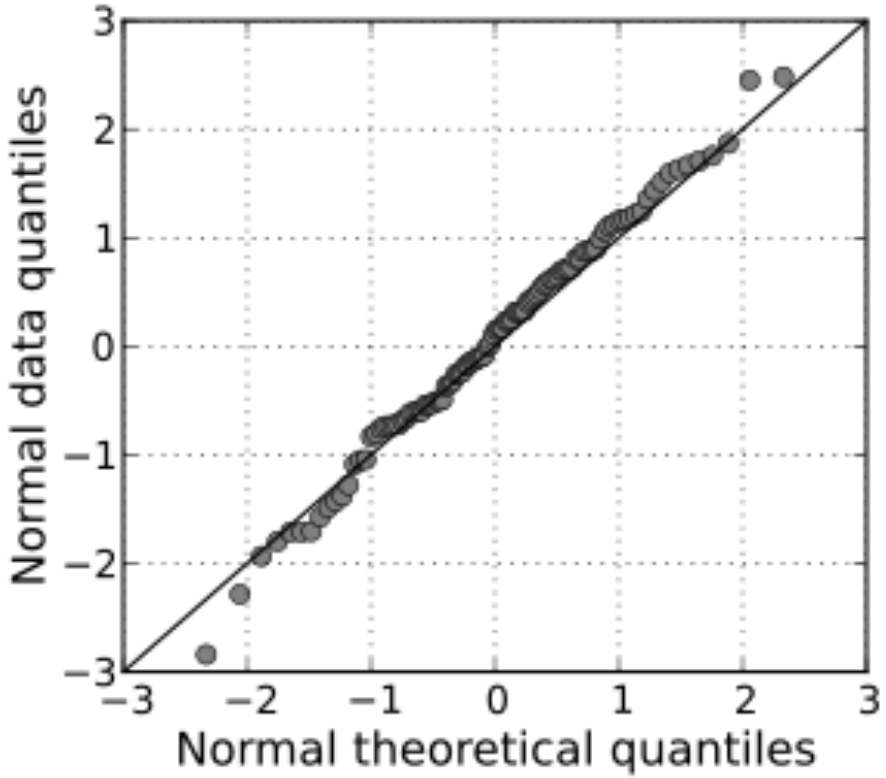
- **Normality**: residuals are normally distributed around the predicted value $\hat{y}$
- **Homoscedasticity**: constant variance over the whole range of x values
- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points

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 homoscedasticity and linearity

Homoscedasticity: constant variance over the whole range of x values

Linearity: A line can describe the relationship between x and y

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

Checking homoscedasticity and linearity

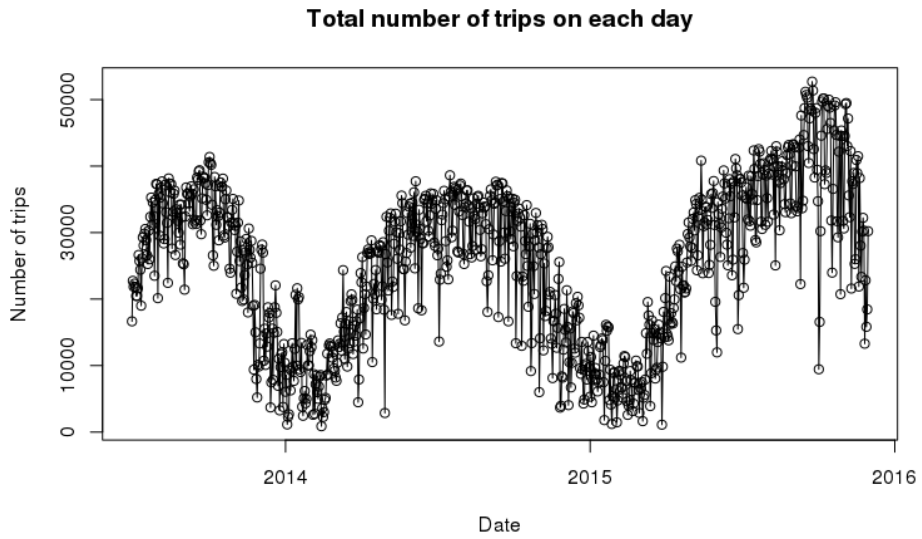
Checking independence

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

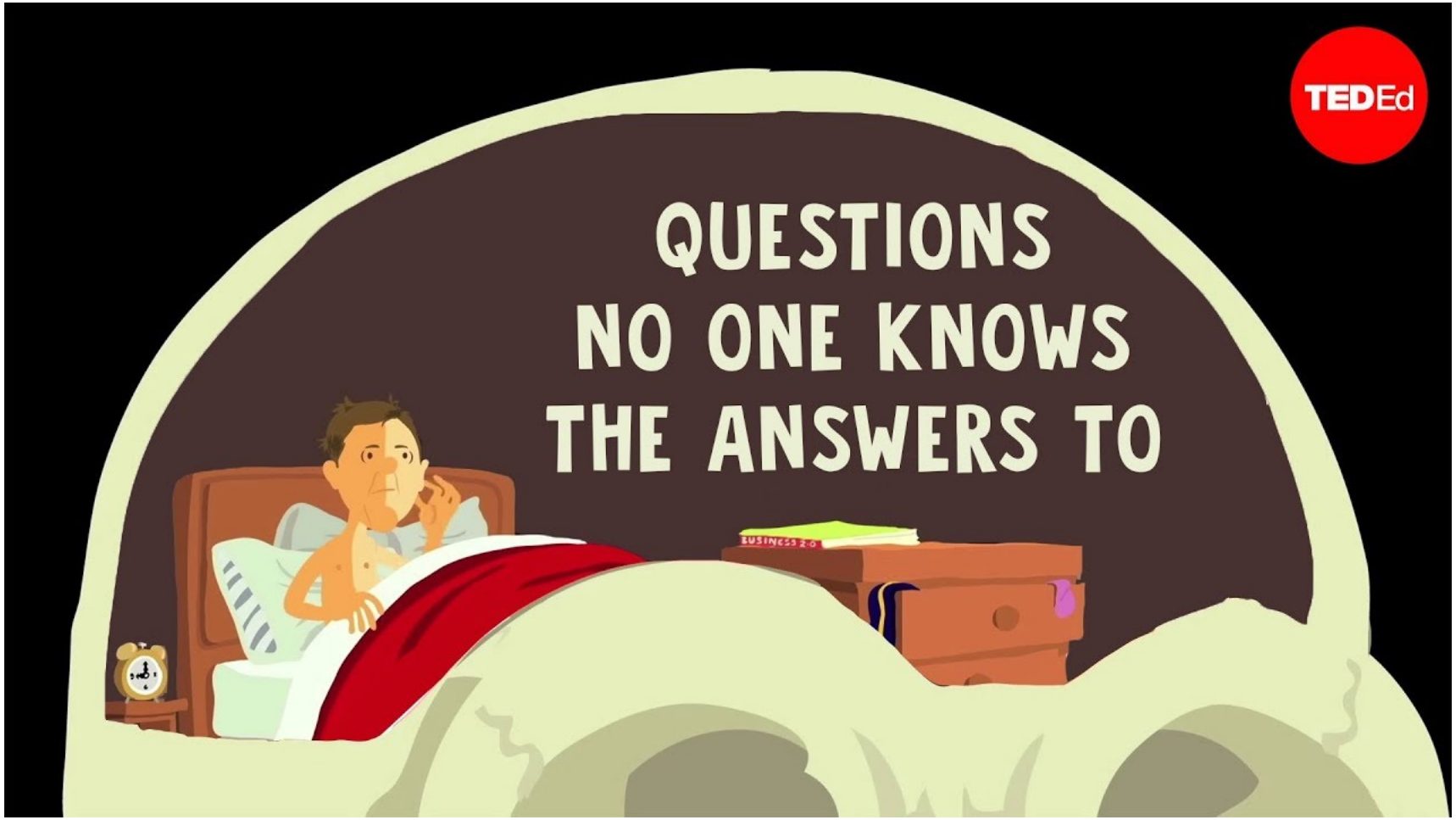
- Simple random sample from the population is likely independent
- Time often are not independent

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

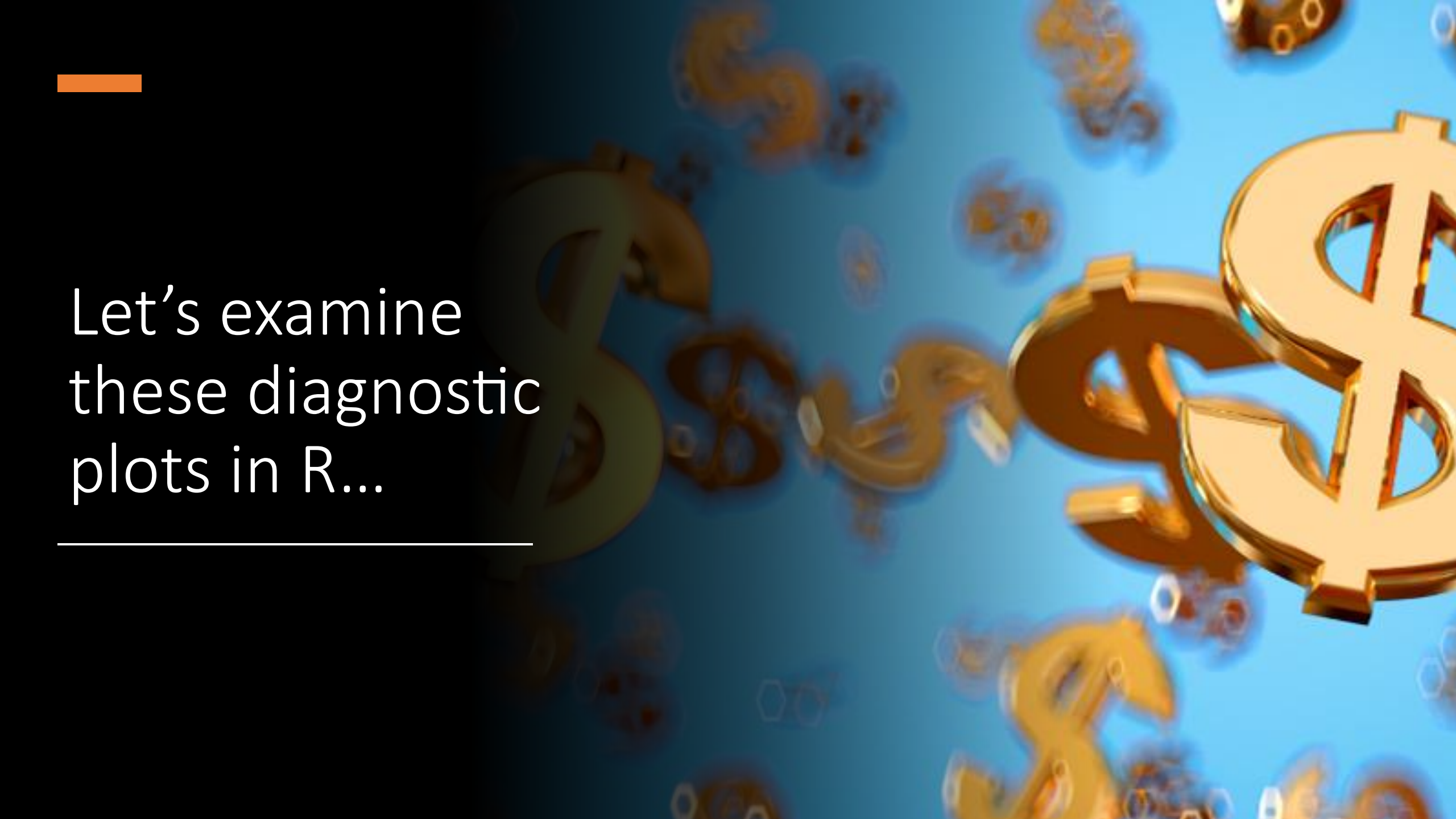
- i.i.d. independent and identically distributed



Questions?

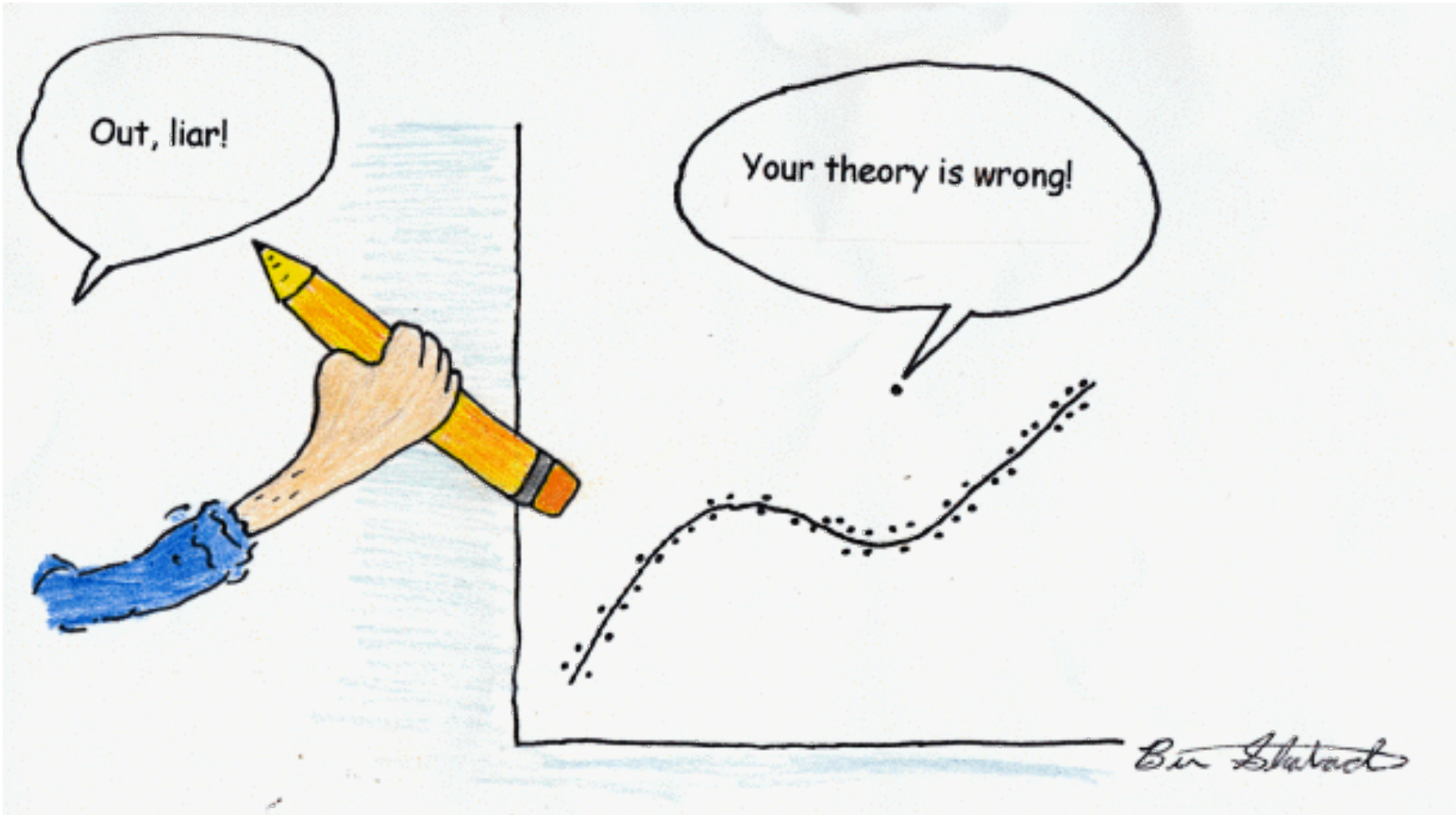


Let’s examine these diagnostic plots in R…



Start at part 4 of the R markdown document

Statistics for unusual observations



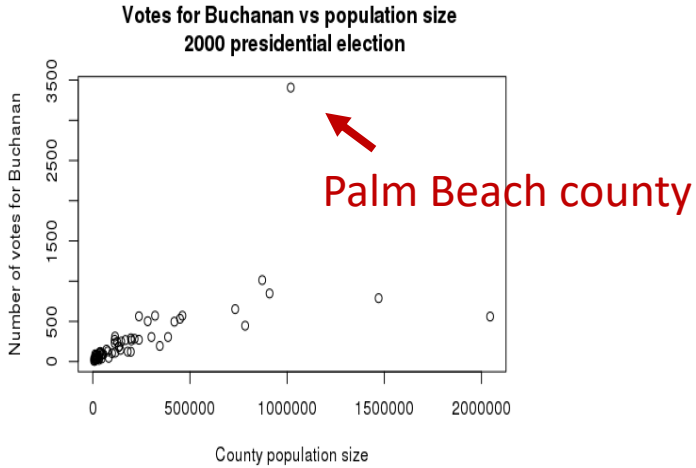
Statistics for unusual observations

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

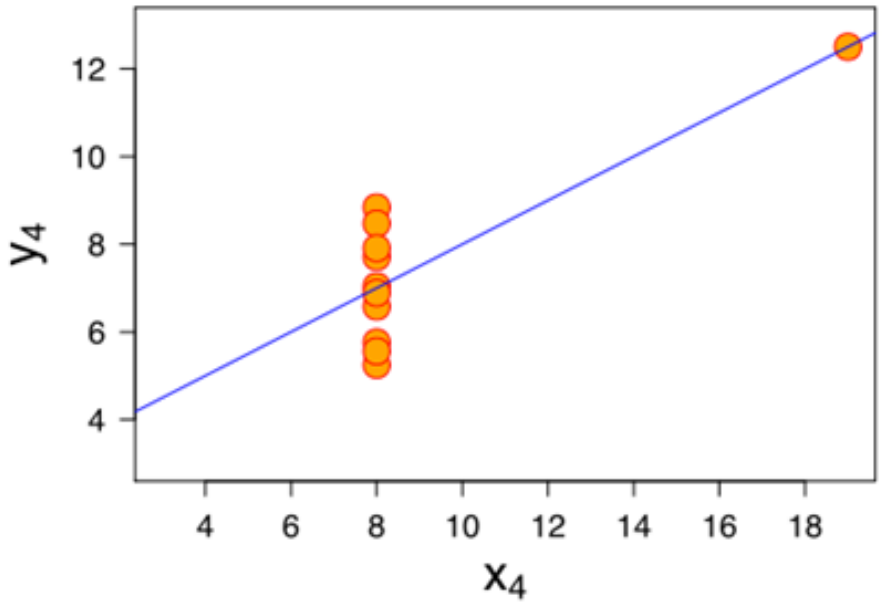
High leverage points are predictors **x** that are far from the mean

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

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

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

R: hatvalues()



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

Typical:	$h_i = 2/n$
High:	$h_i = 4/n$
Very high:	$h_i = 6/n$

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

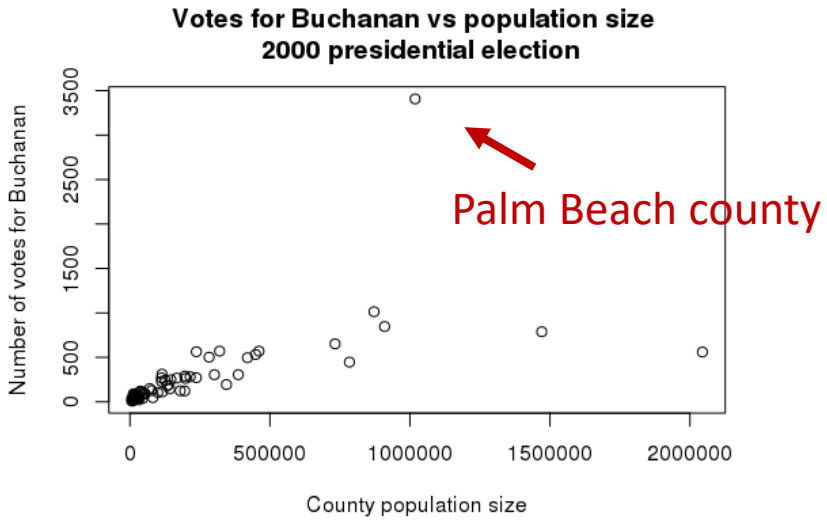
Makes residuals at the ends a bit larger to deal with the fact that they are 'overfit'

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)

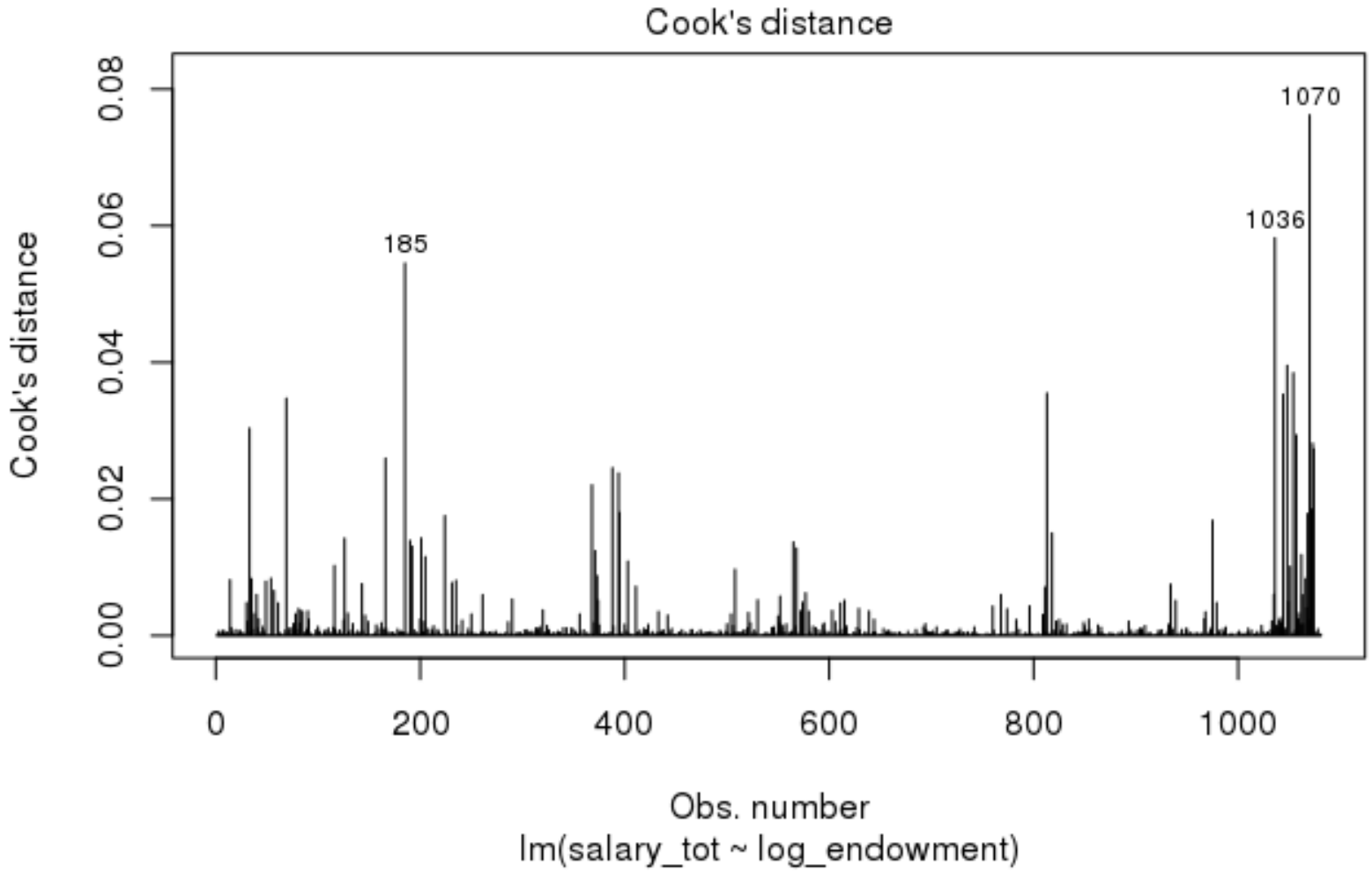
Larger for high leverage points

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

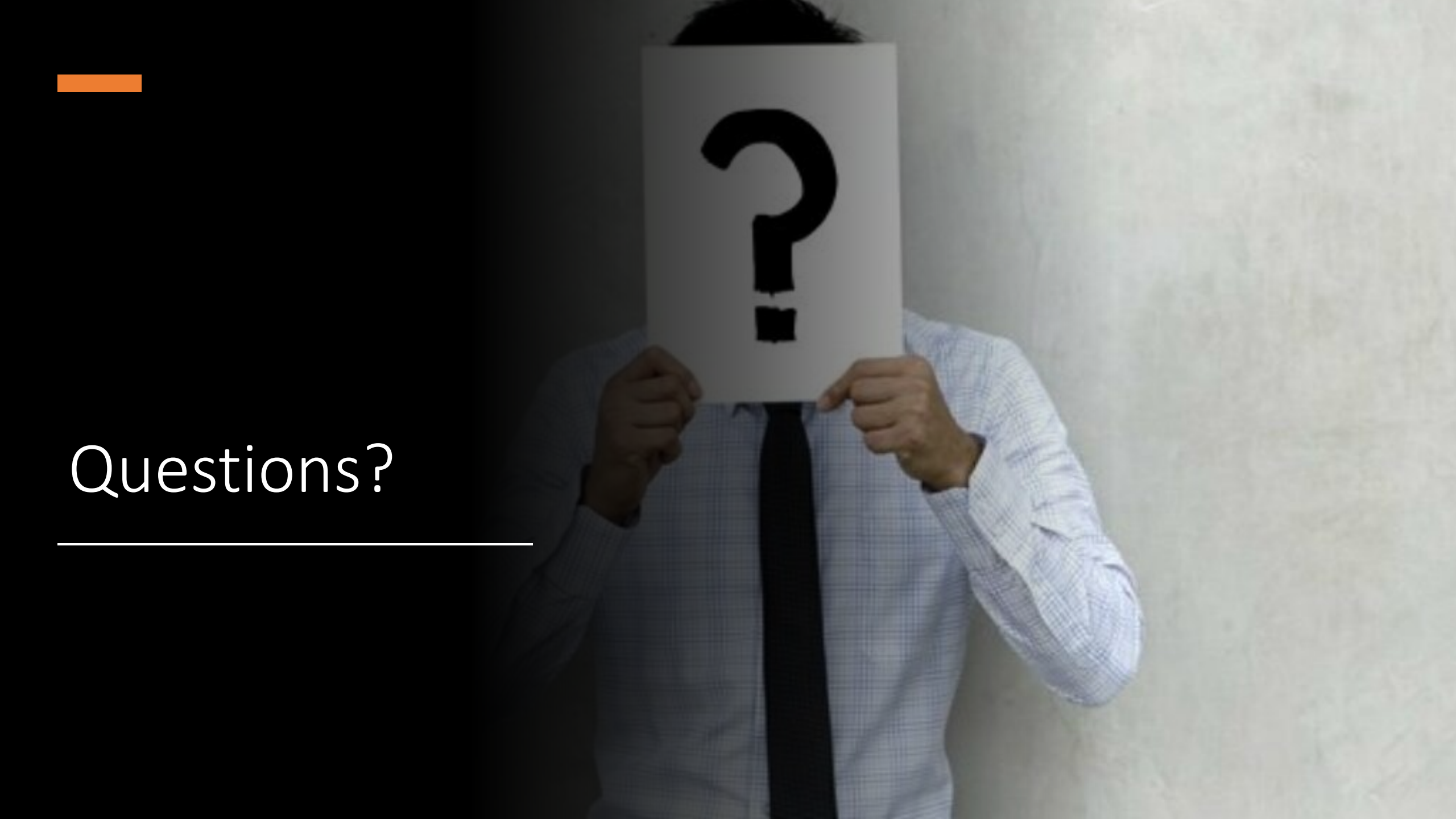
Statistic	Moderately unusual	Very unusual
Leverage, h_i	Above $2(k + 1)/n$	Above $3(k + 1)/n$
Standardized residual	Beyond ± 2	Beyond ± 3
Studentized residual	Beyond ± 2	Beyond ± 3
Cook's D	Above 0.5	Above 1.0

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?



Let’s examine these statistics in R…

Homework 7

edmunds.com SM

price_bought ≈ $\beta_0 + \beta_1$ x mileage_bought + ε

Questions?