Inference for linear regression

Halloween edition...

Overview

Review of regression models

Inference on regression models

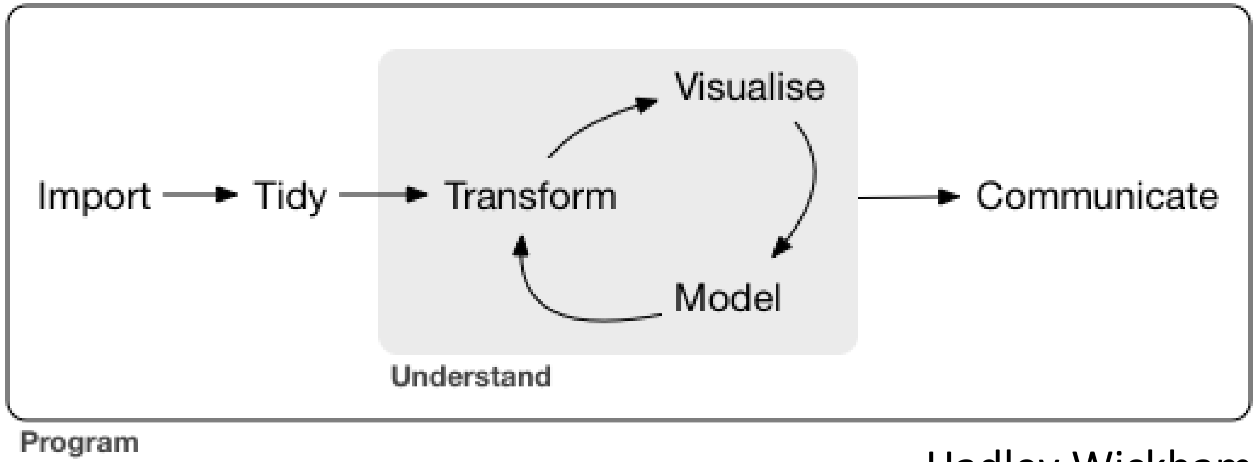
- Hypothesis tests
- Confidence intervals and predictions intervals

Regression diagnostics

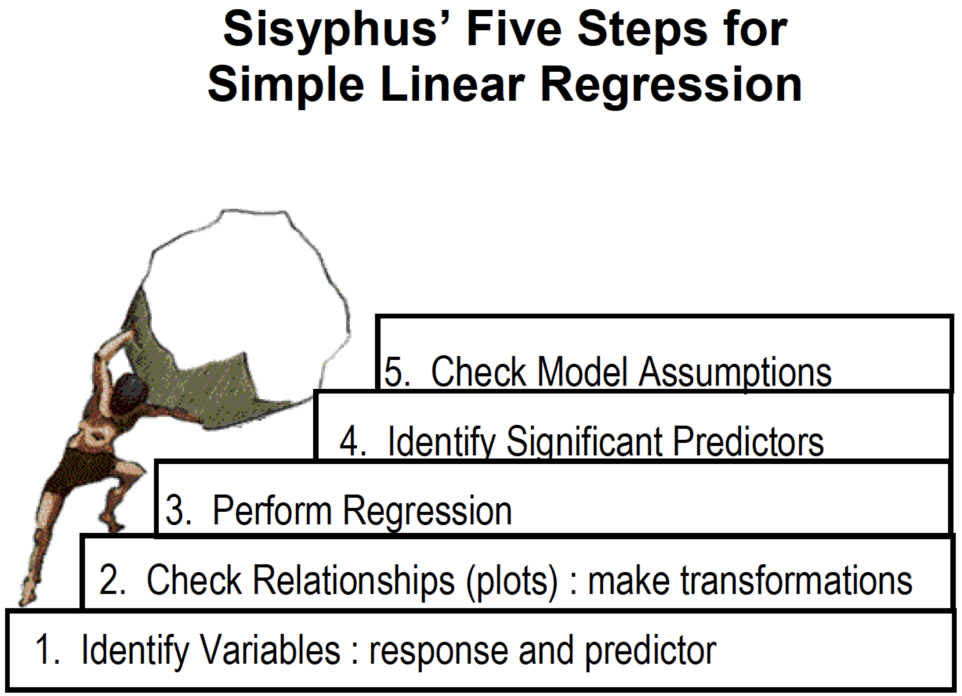
Next class: statistics for identifying unusual observations

Linear regression continued…

The process of building regression models



Hadley Wickham



Jonathan Reuning-Scherer

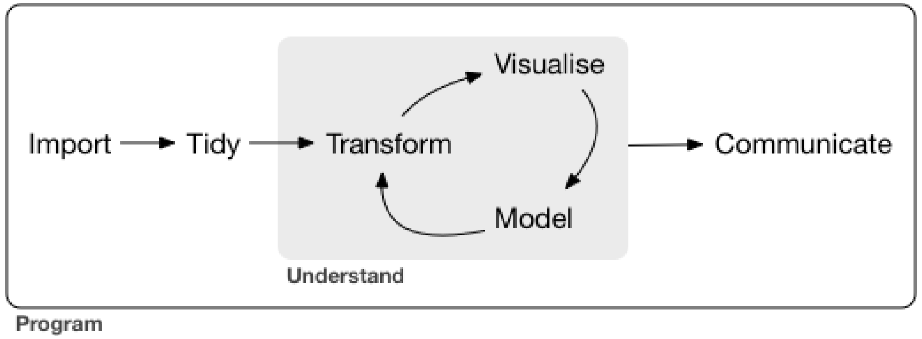
"All models are wrong, but some are useful"
- George Box

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



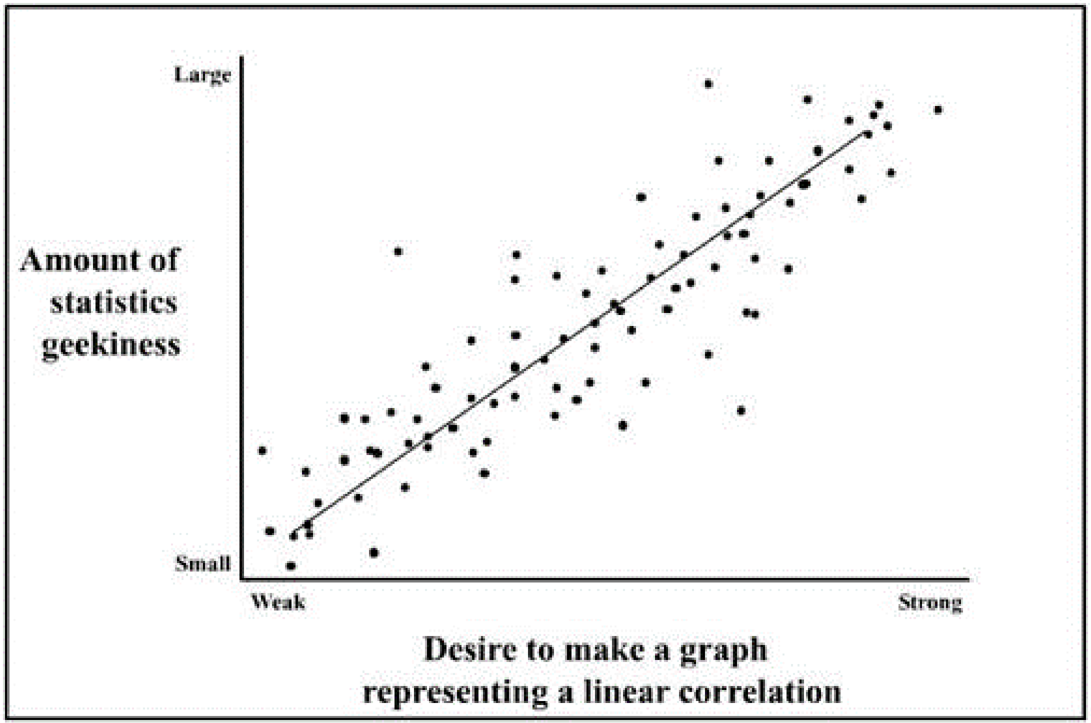
Use the model to address questions of interest
- Make predictions
- Explore relationships between response variable (y) and explanatory variables (x)
- Keep in mind limitations of the model
  - e.g., can be difficult to make the claim that changes in y *cause* changes in x from *observational data*

Review of underlying models and inference

Linear regression

In **linear regression** we fit a <u>regression line</u> 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: $$\mu_Y = \beta_0 + \beta_1 x$$

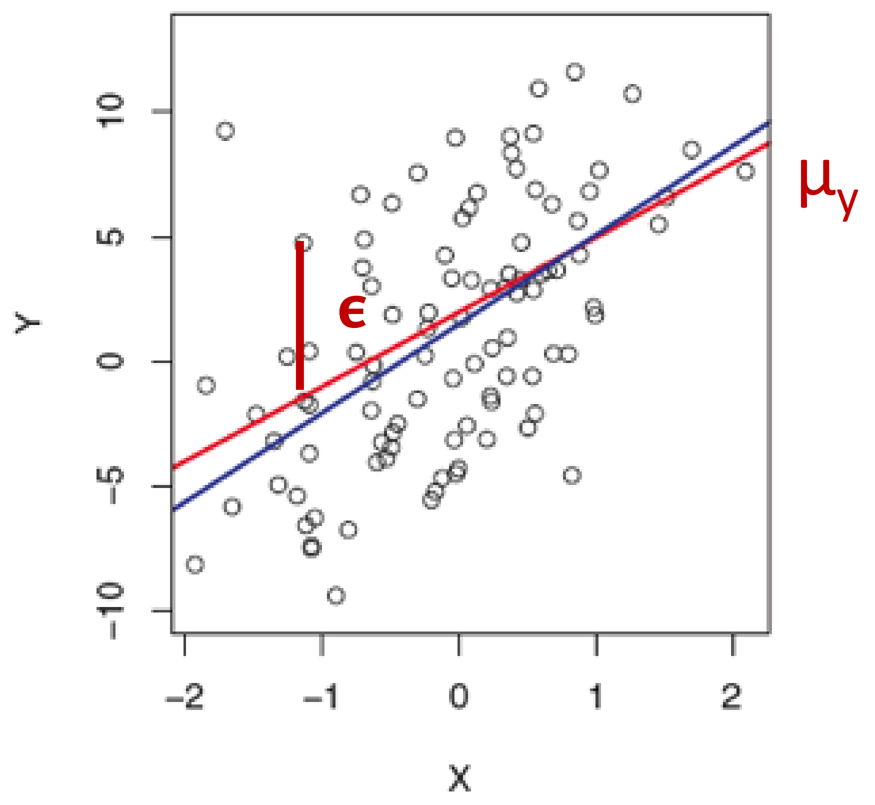
Error

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: $$Y = \beta_0 + \beta_1 x + \epsilon$$

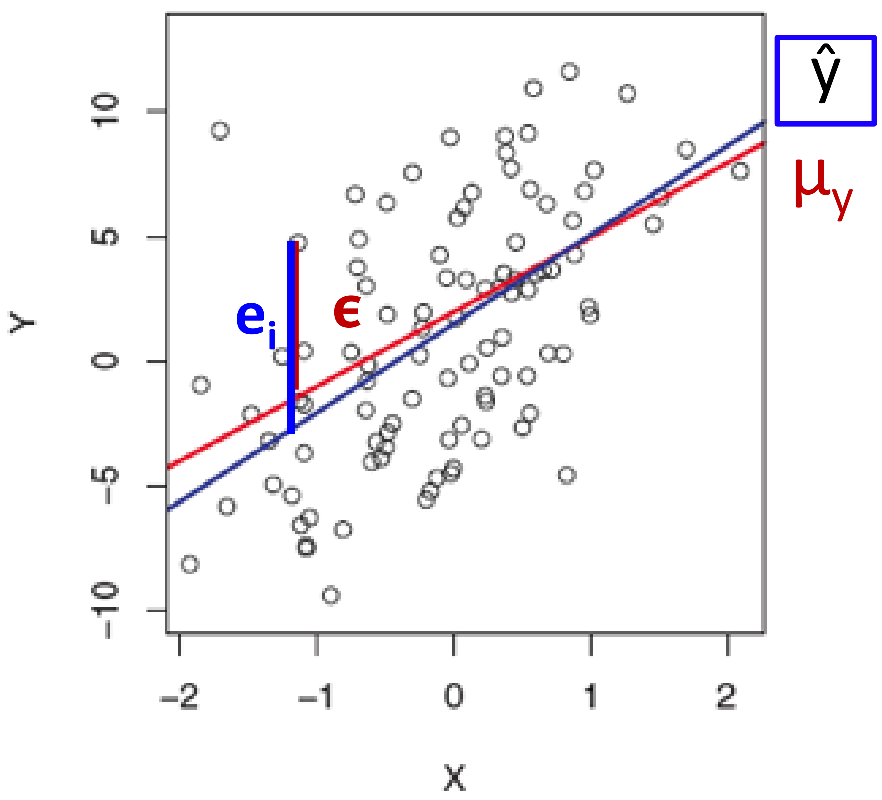
Estimated regression line: $$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

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

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



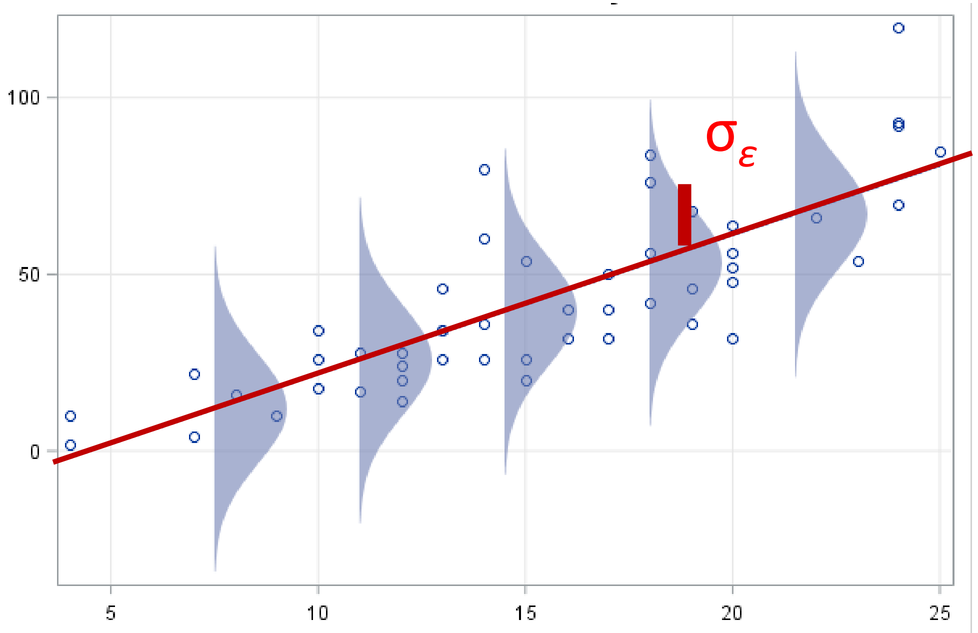
Standard deviation of the errors: σ_ε

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…

- residual **standard error (RSE)**

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



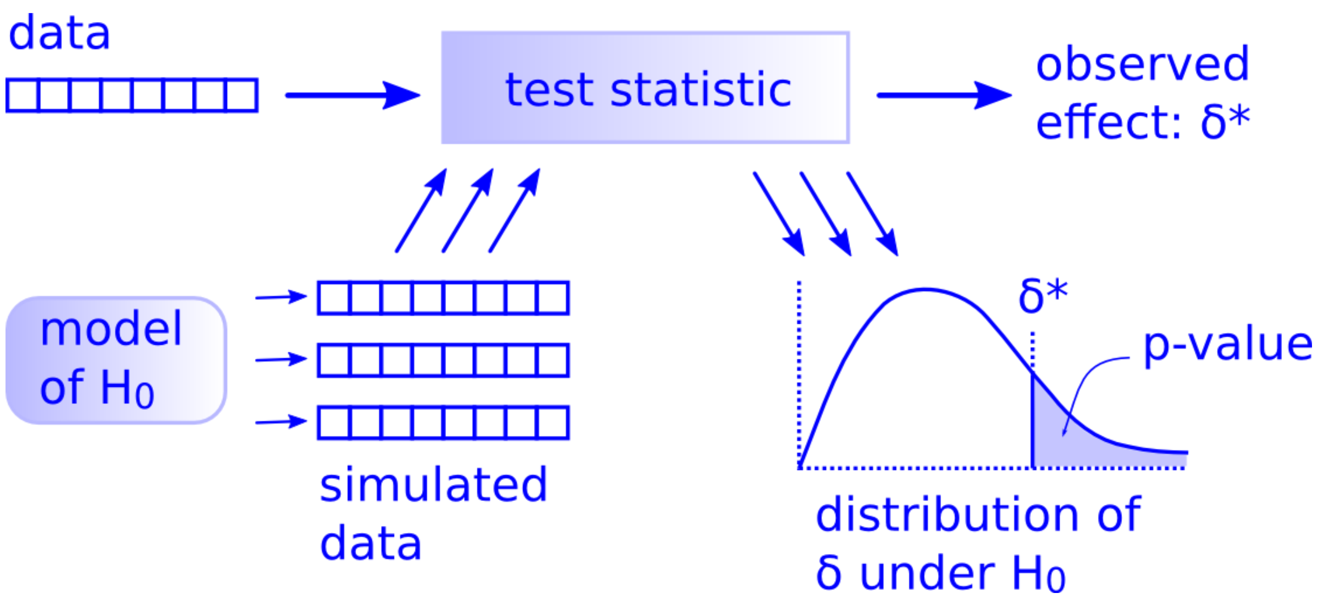
$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2} SSRes} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} e_i^2}$$

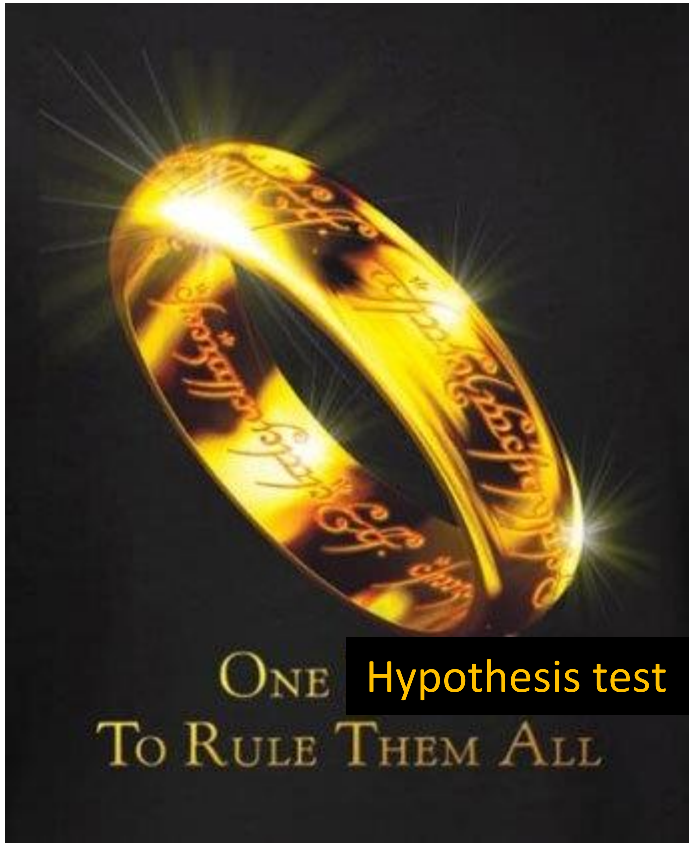
$$= \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Inference for linear regression: hypothesis tests

Hypothesis test for regression coefficients

There is only one hypothesis test!





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

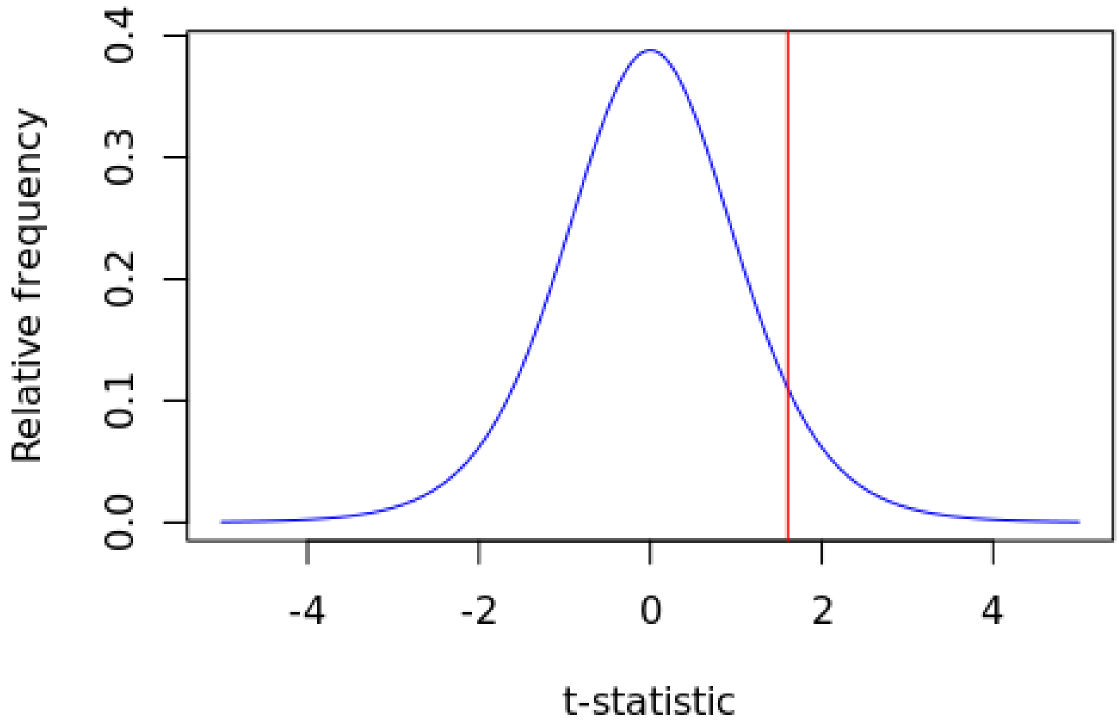
- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

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

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

Hypothesis test for regression coefficients

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



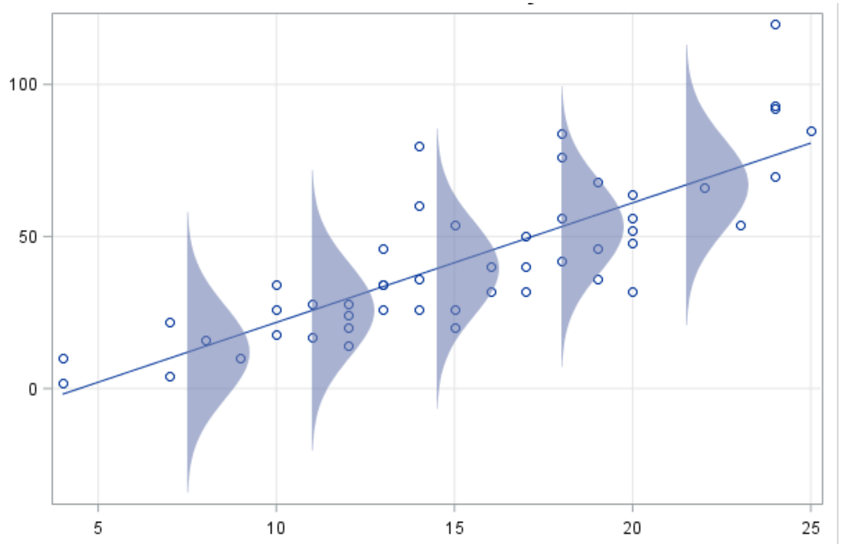
Inference using parametric methods

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

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

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



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

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

Back to faculty salaries

Start at part 2 of the class code…

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

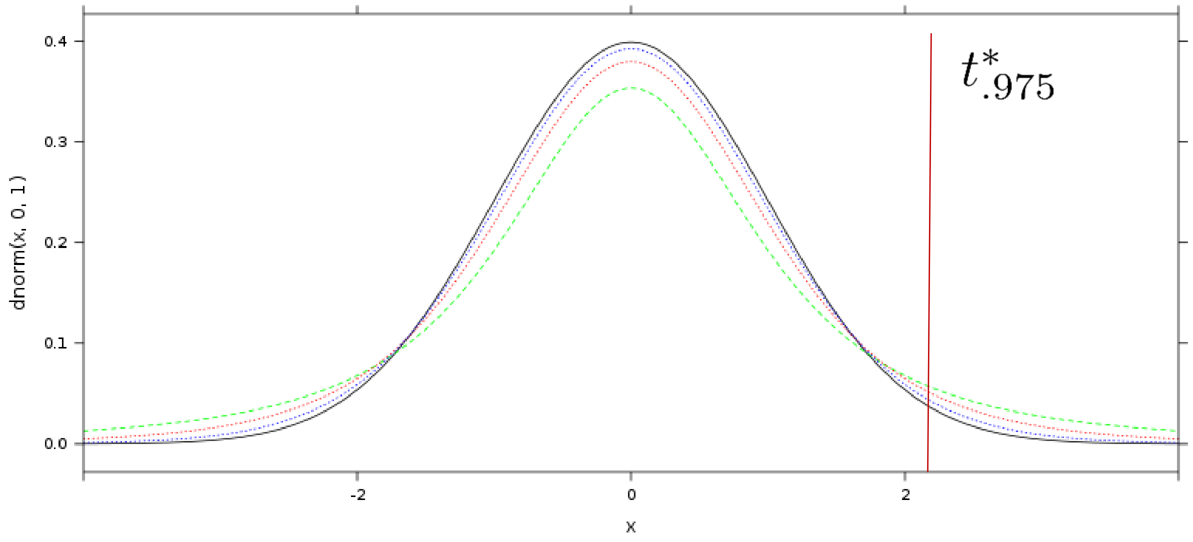
For the slope coefficient , the confidence interval is: $\hat{\beta}_1 \pm t^* \cdot SE_{\hat{\beta}_1}$

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

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

qt(.975, df)

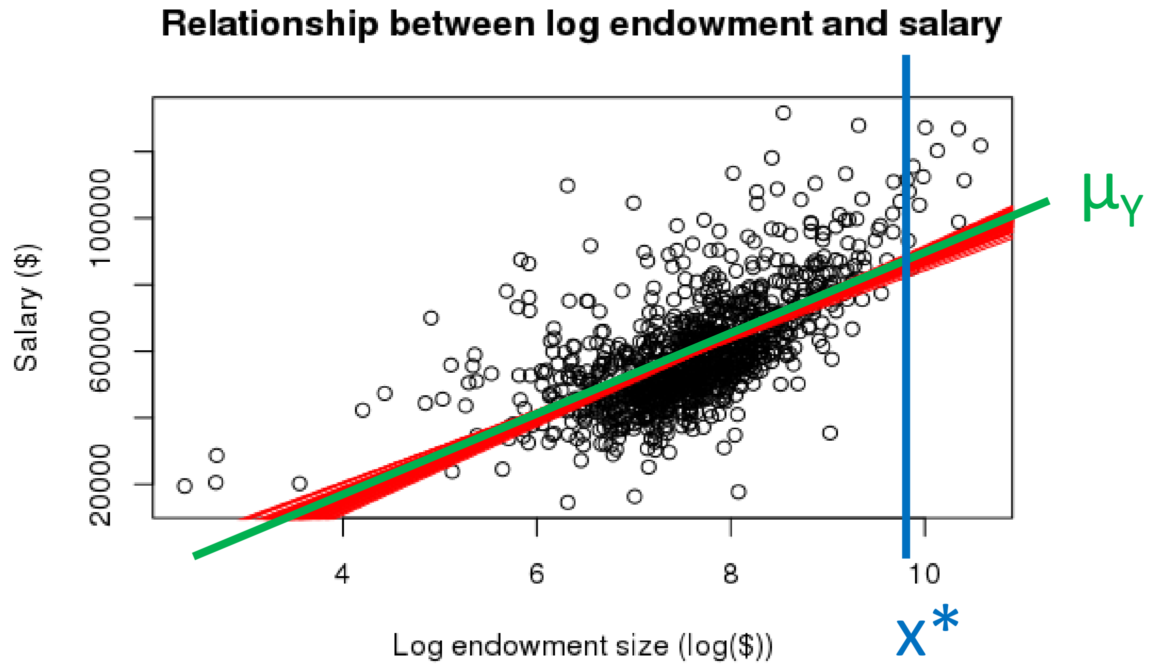
N(0, 1)
df = 2
df = 5
df = 15



Confidence intervals for the regression line μ_Y

A confidence interval for the mean response for the ***true regression line*** μ_Y when X = 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

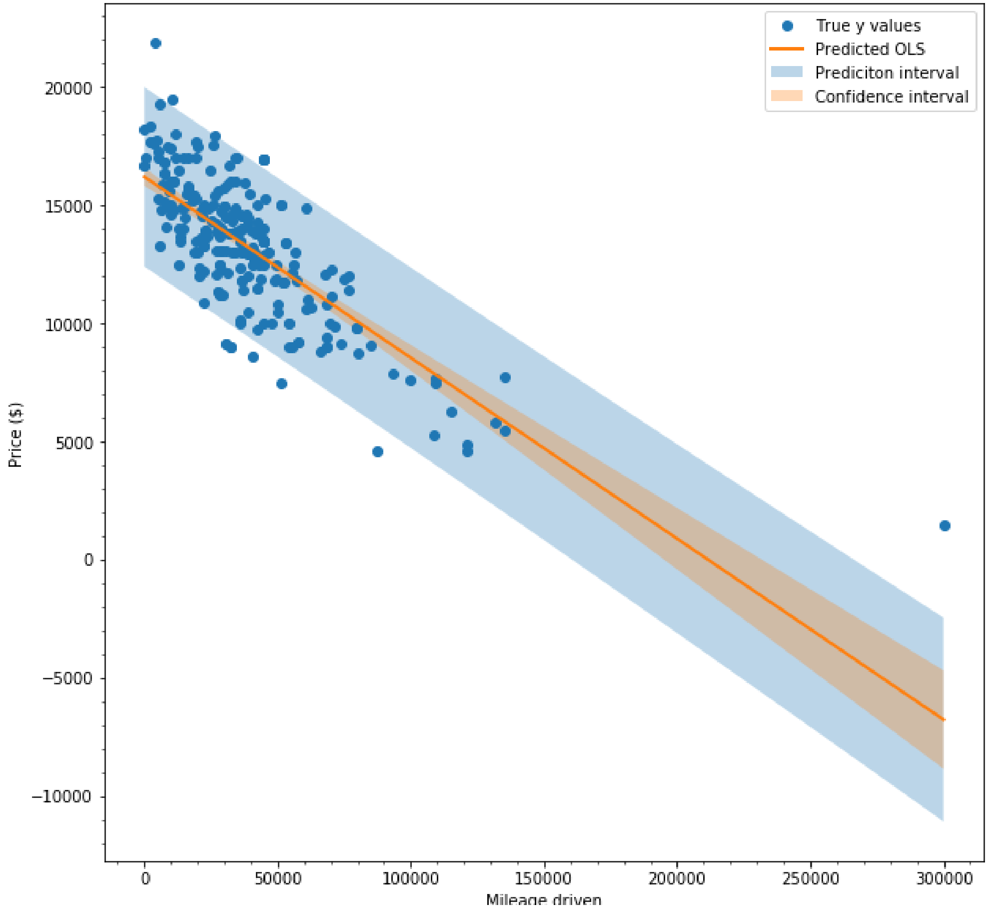
Prediction intervals

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:

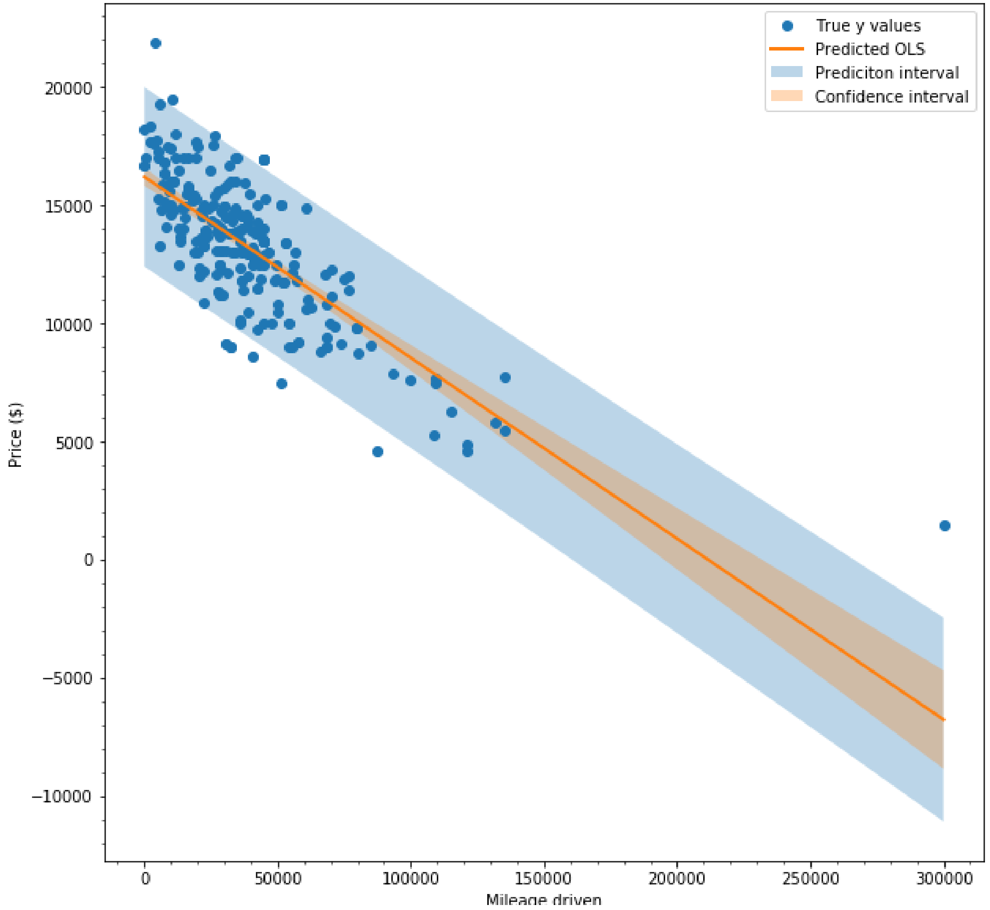
$$\hat{y} \ \pm \ t^* \cdot SE_{\hat{y}}$$

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

2. CI for regression line μ_Y at point x*

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

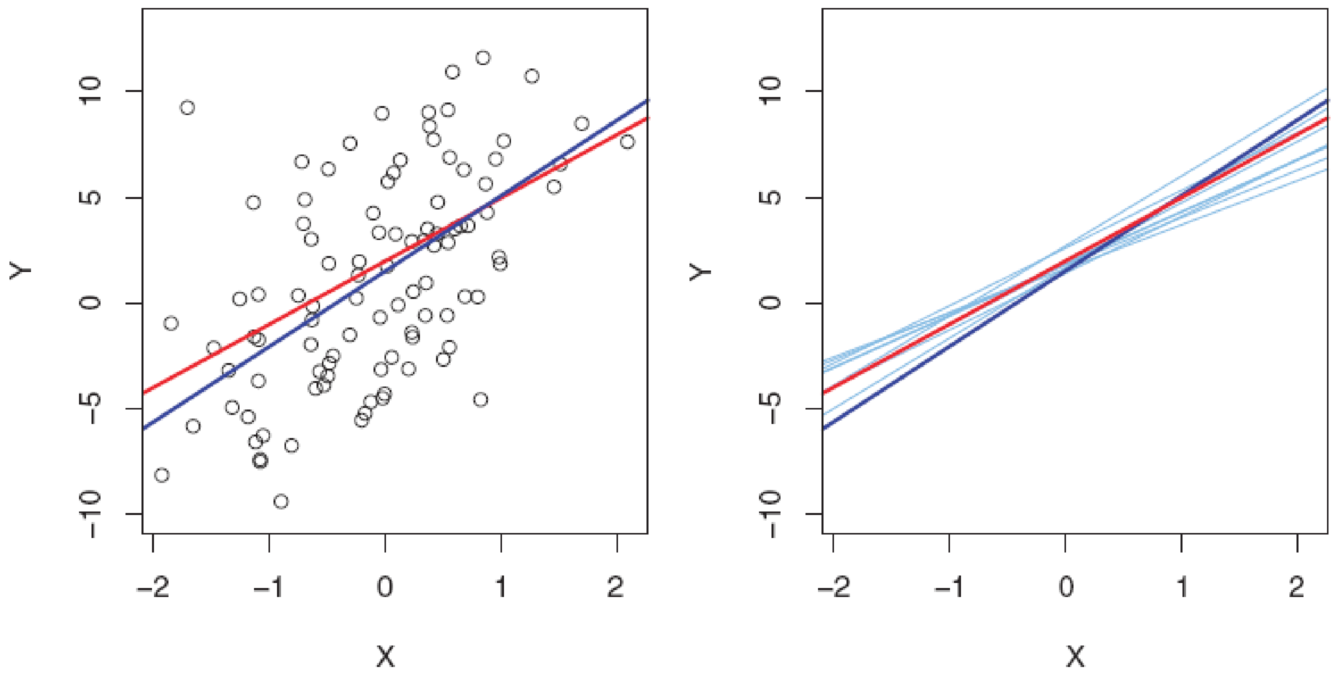
3. Prediction interval y

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

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

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

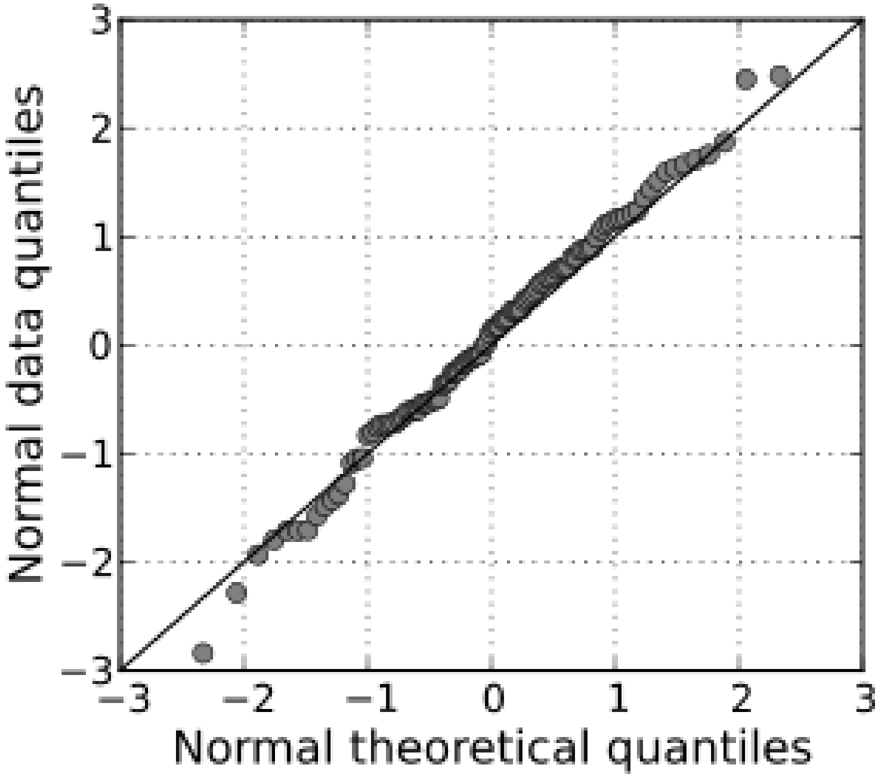
Checking linearity and homoscedasticity

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()



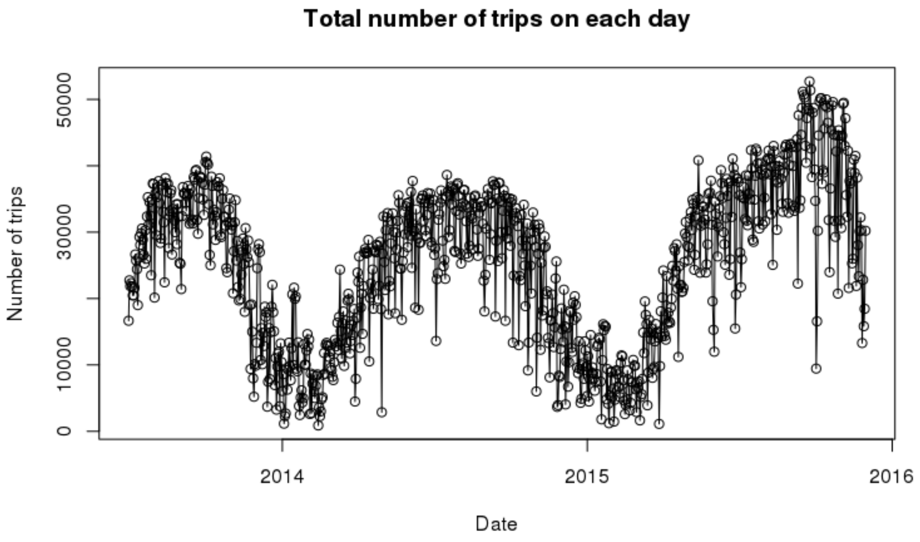
Checking Independence

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 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