Influential points, ANOVA for regression and multiple regression

Halloween edition…

Overview

Review of inference for simple linear regression

Examining unusual data points

Analysis of variance for regression

If there is time: multiple regression

- Basic ideas
- Nested model comparison
- Related sampling and multiple regression coefficients

Quick review of simple linear regression

The process of building regression models

Choose the form of the model

- Identify and transform explanatory and response variables

Fit the model to the data

- Estimate model parameters

Assess how well the model describes the data

- Analyze the residuals, evaluate unusual points, etc.

Use the model to address questions of interest

- Make predictions, explore relationships, etc.

All models are wrong, but some models are useful

Simple linear regression concepts

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

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

Inference for simple linear regression models

- Hypothesis tests for intercept and slope
- Confidence intervals for slope and line; prediction intervals

Inference is valid if these conditions are met:

- Linearity, Independence, Normality, Equal variance of errors

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

$$\hat{SE}_{\hat{\beta}_1} = \frac{\hat{\sigma}_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \hat{SE}_{\hat{\beta}_0} = \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Summary of confidence and prediction intervals

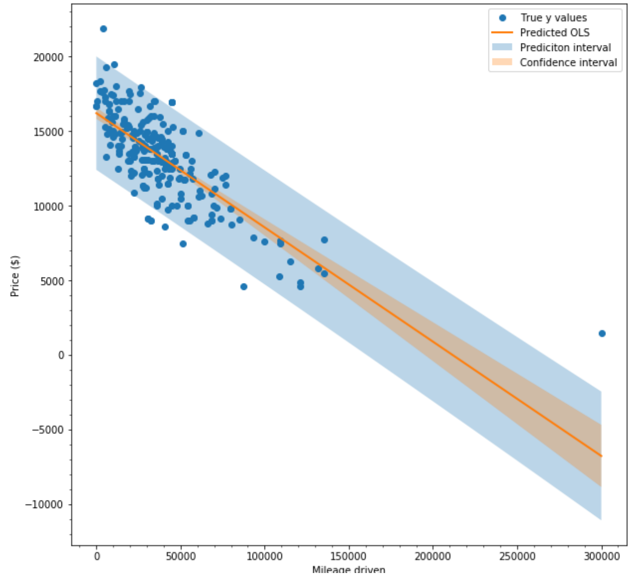
1. CI for slope β

$$\hat{\beta}_1 \pm t^* \cdot \hat{SE}_{\hat{\beta}_1} \qquad \hat{SE}_{\hat{\beta}_1} = \hat{\sigma}_e \sqrt{\frac{1}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

β_1

2. CI for regression line μ_Y at point x*

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



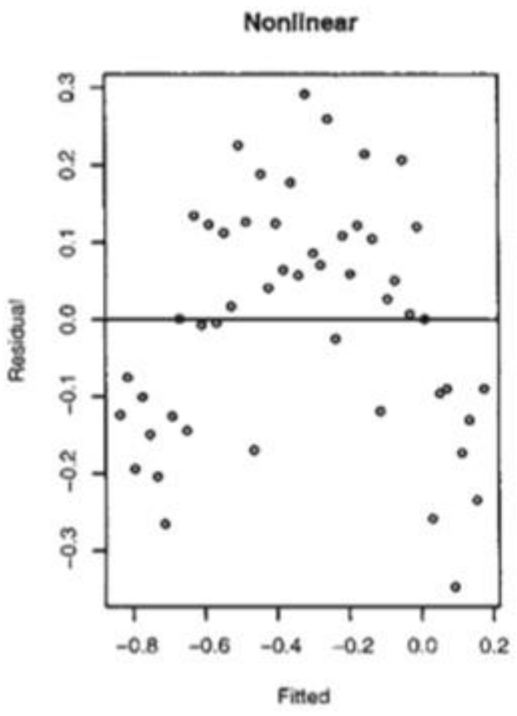
3. Prediction interval y

$$\hat{y} \pm t^* \cdot \hat{SE}_{\hat{y}} \qquad \hat{SE}_{\hat{y}} = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_{i=1}^n (x_i - \bar{x})^2}}$$

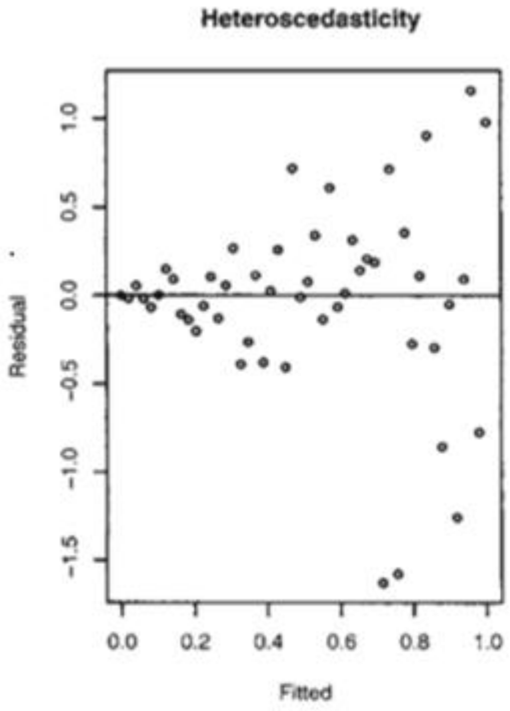
Regression diagnostics

Linearity, Independence, Normality, Equal variance of errors

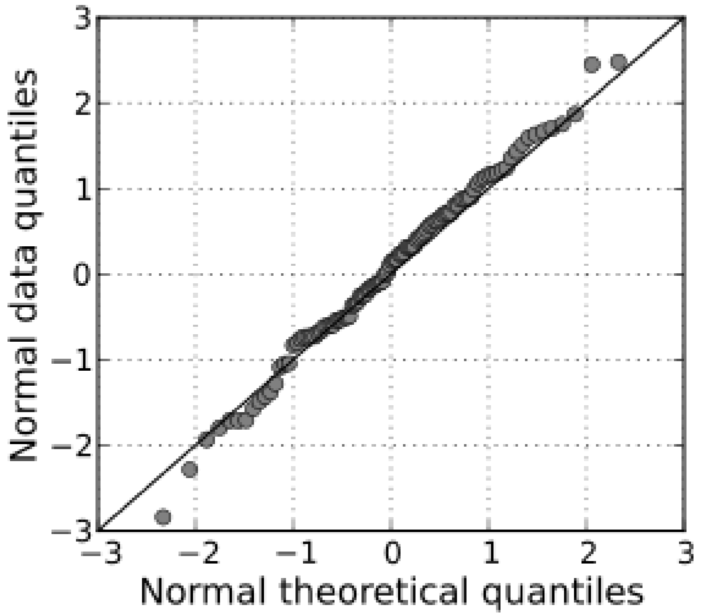
$e = y - \hat{y}$



$\hat{y}$



$\hat{y}$



Questions?

Calculating regression confidence intervals in R

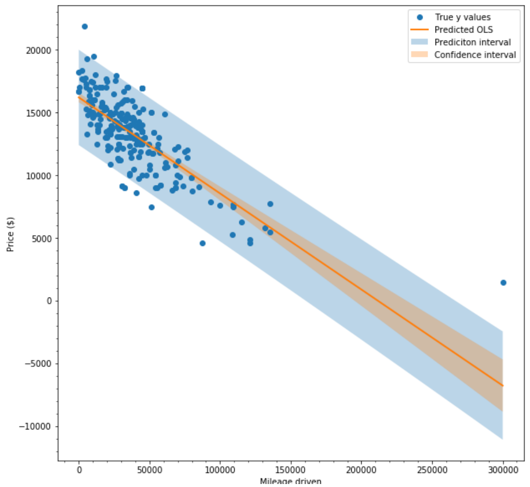
Since we ran out of time last class, let's quickly go through calculating:

1. Confidence intervals for regression coefficients (β_0 and β_1)

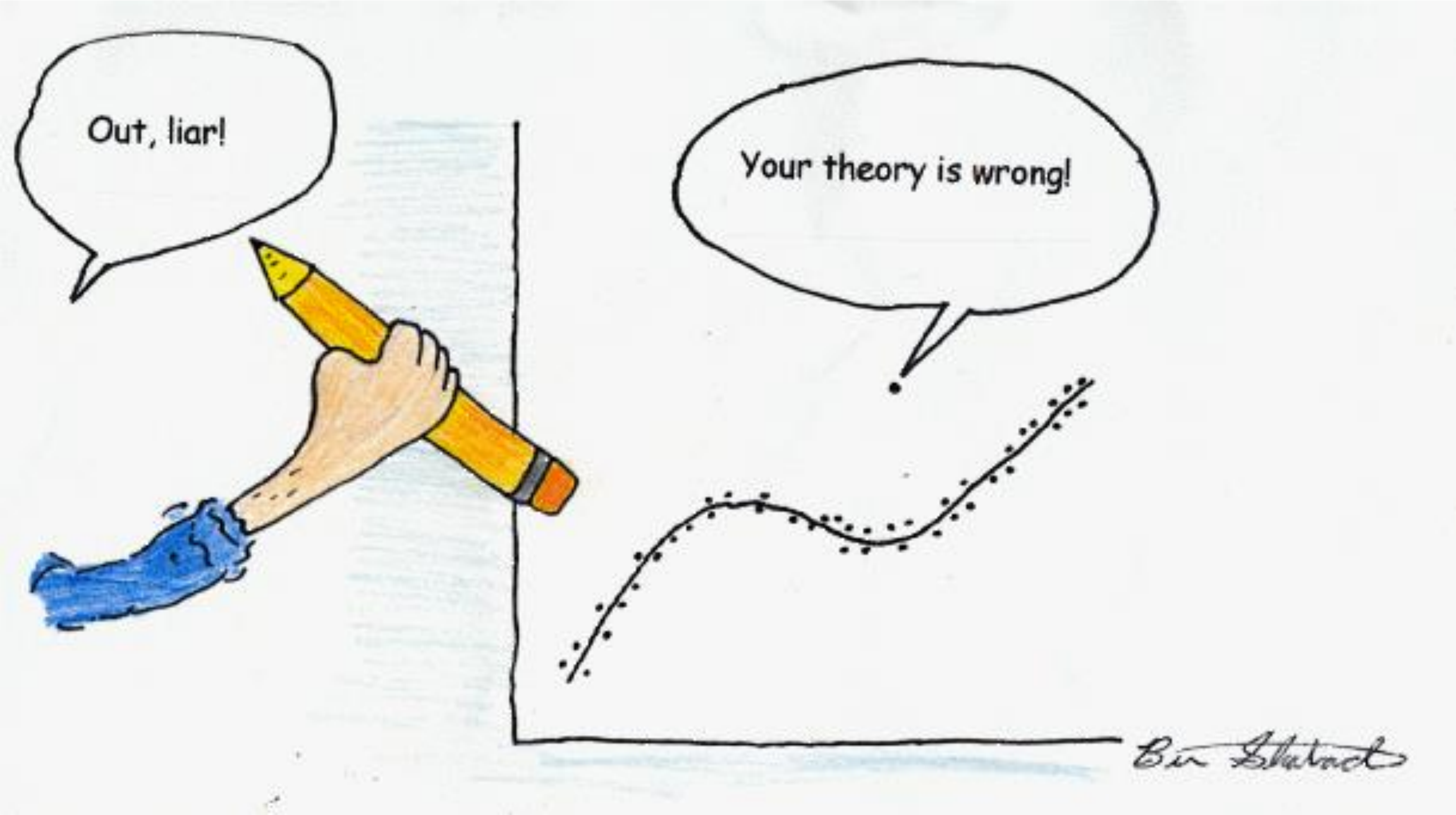
2. Confidence intervals for the regression line

3. Prediction intervals

β_1



Statistics for unusual observations



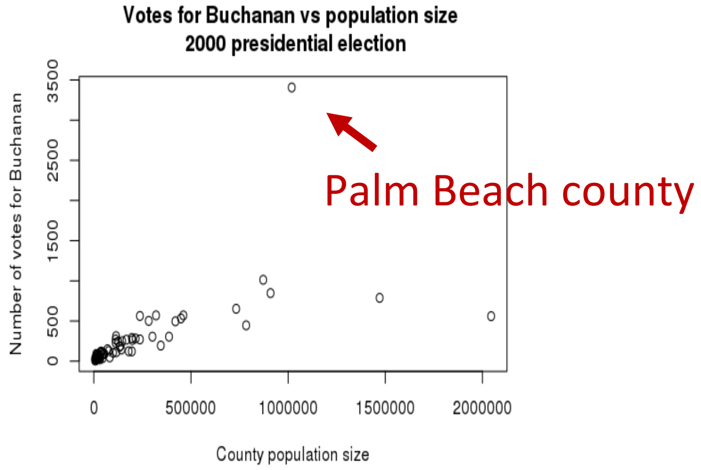
Statistics for unusual observations

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

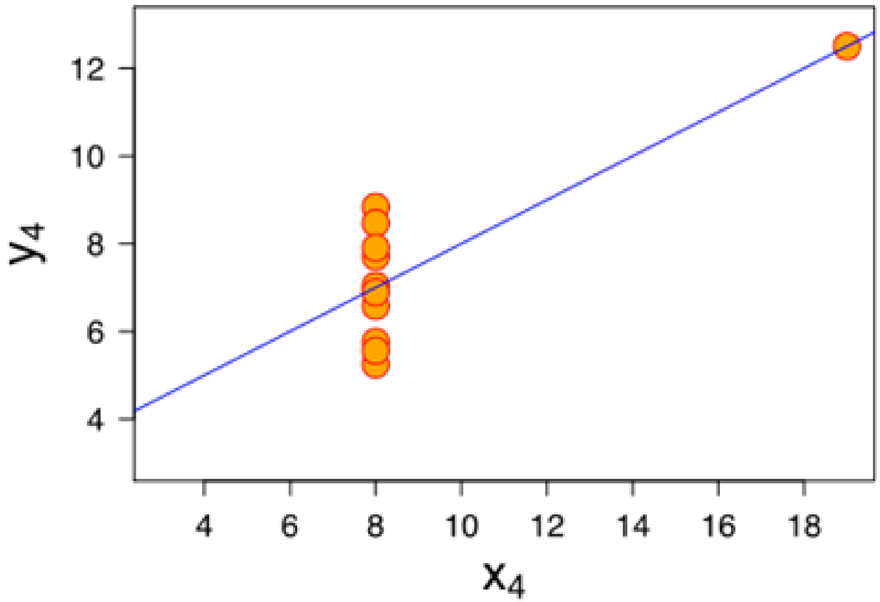


Unusual observations **can also have a big effect on the model fit**

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

Leverage

Unusual x



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

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

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

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

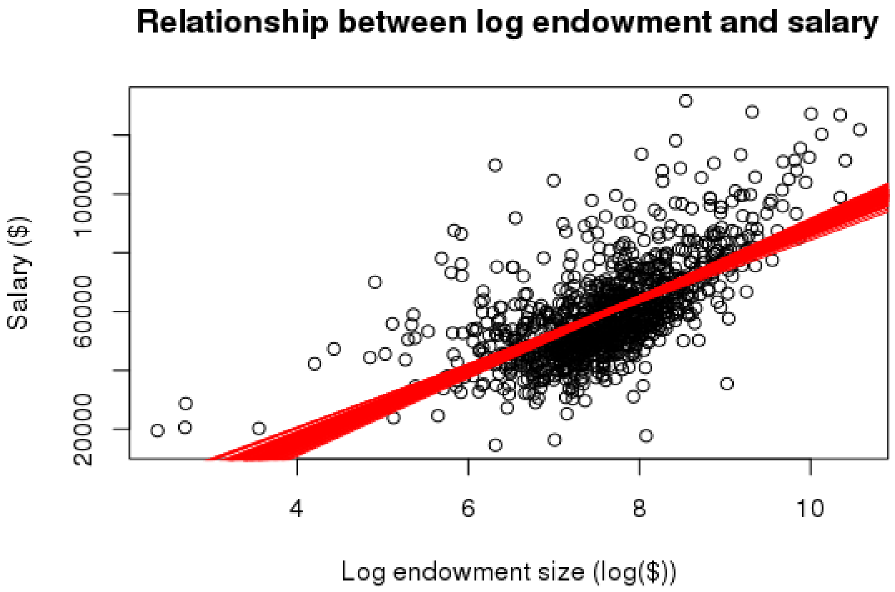
Typical:	$h_i = 2/n$
High:	$h_i = 4/n$
Very high:	$h_i = 6/n$

R: hatvalues()

Outliers: standardized residuals

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

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



Puts residuals on a 'normalized' scale

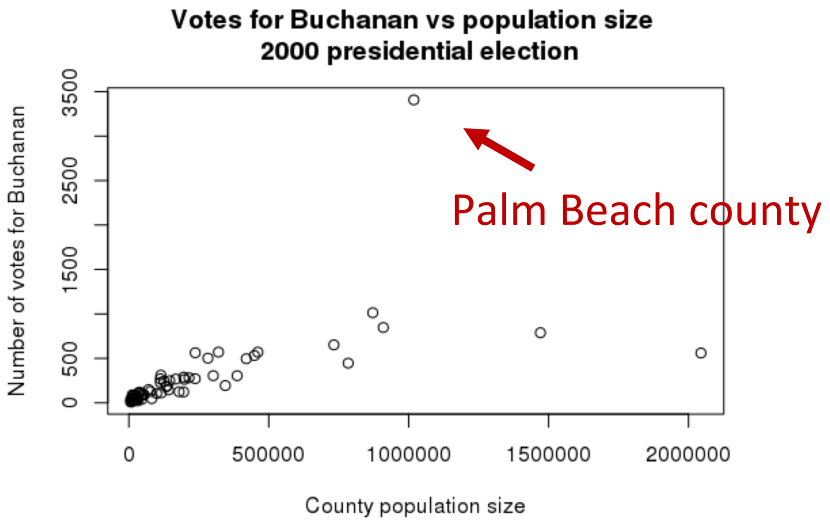
Makes residuals at the ends a bit larger to deal with the fact that they are 'overfit'

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 $\hat{\sigma}_e$ with the i^{th} point removed

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

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

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

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

R: cooks.distance ()

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 distance

Cook's distance can also be expressed as the how much the predicted values ŷ's would change if the i^{th} was not used when fitting the model

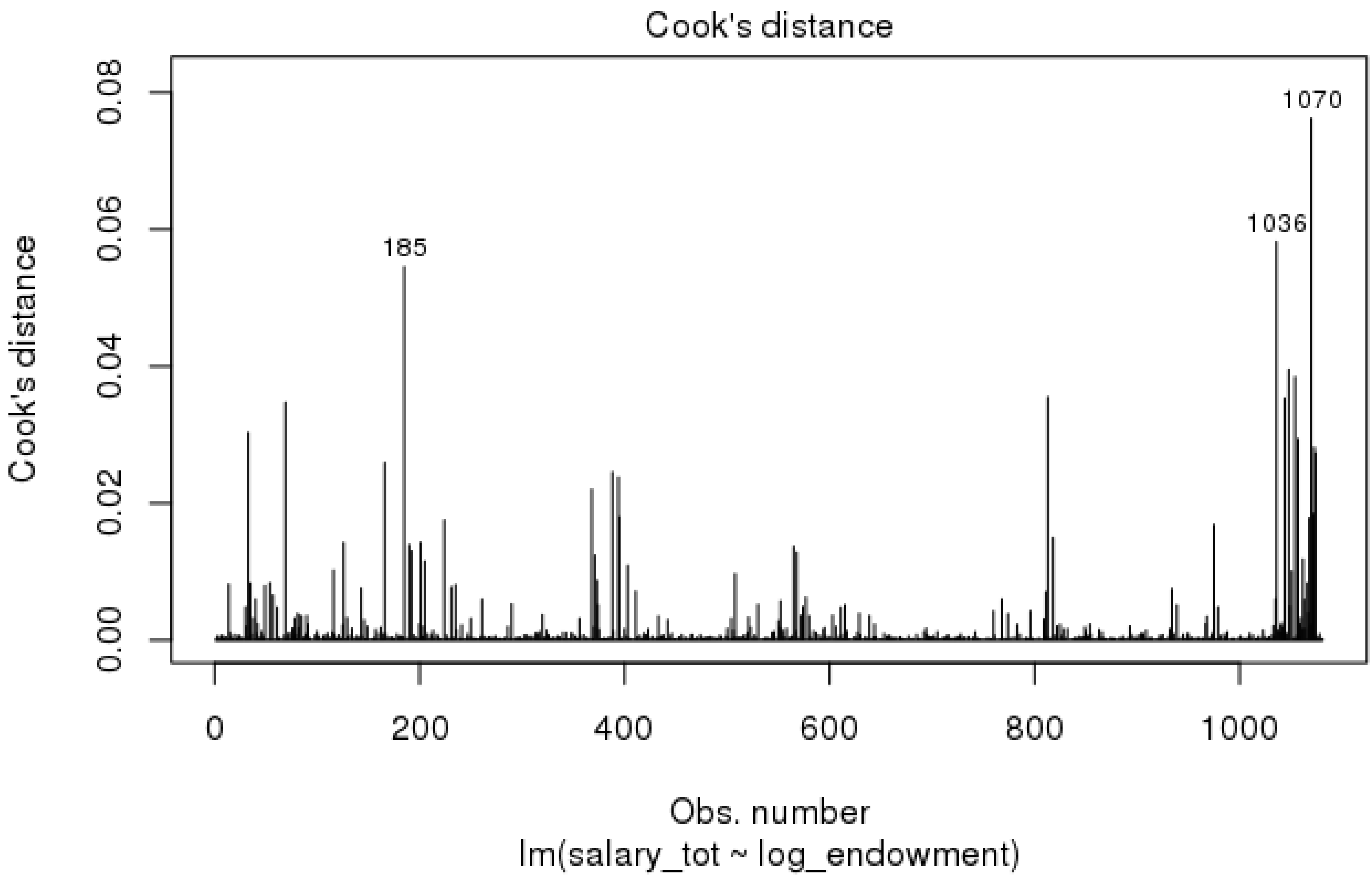
$$D_i = \frac{\Sigma_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1) \cdot \hat{\sigma}_e^2}$$

Number of predictors in the model

(i.e., k = 1 for simple linear regression)

The model fit with the i^{th} point removed

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

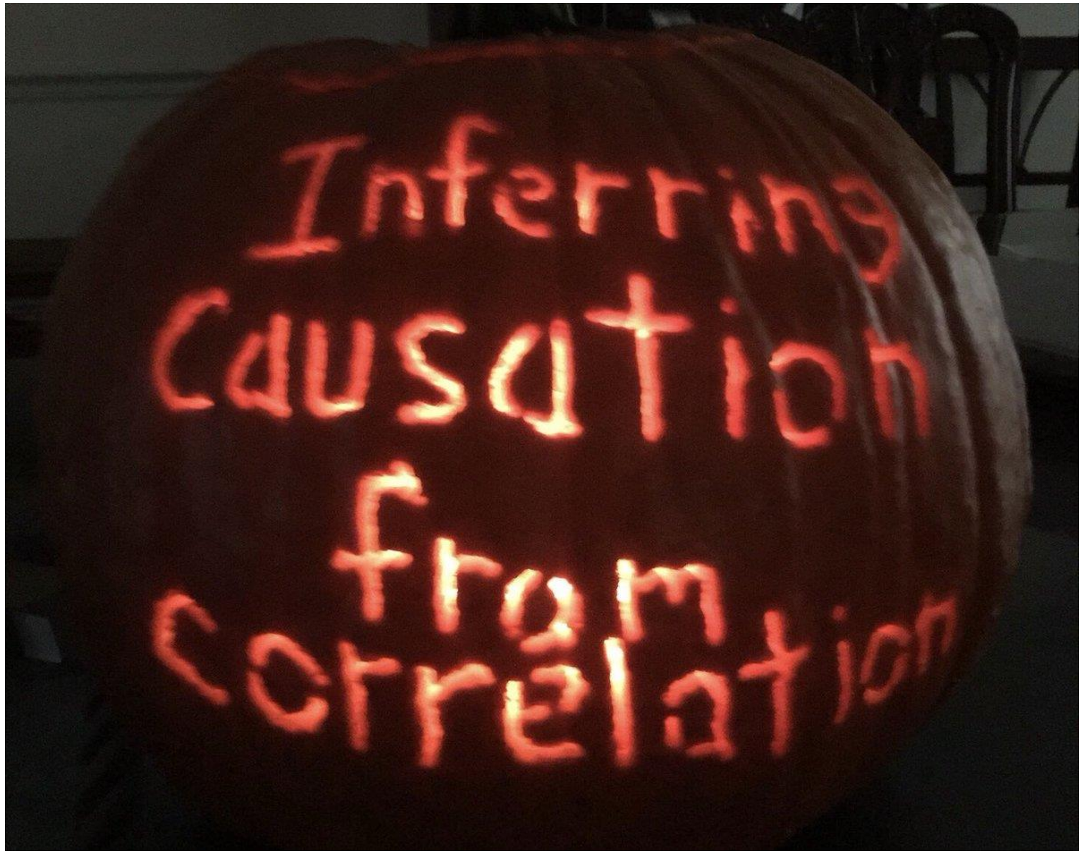
Where:

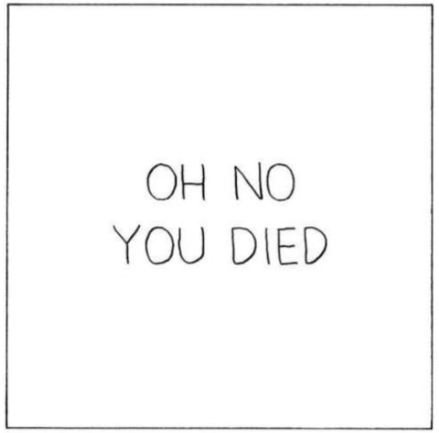
- k is the number of explanatory variables
- n is the number of data points

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

Questions?

Let's try it in R!



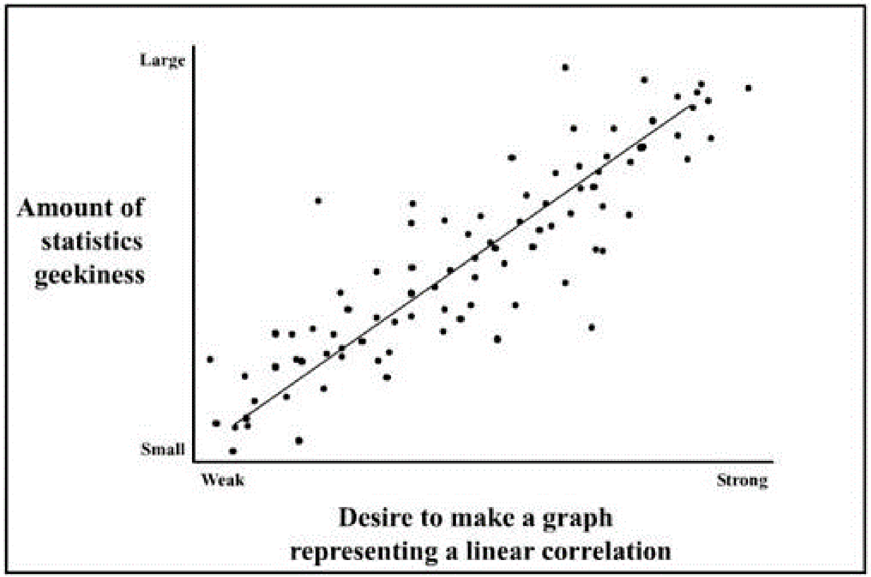


Analysis of Variance (ANOVA) for regression

Analysis of Variance (ANOVA) for regression

In an analysis of variance, we break down the **total variability** in the **response variable y** into:

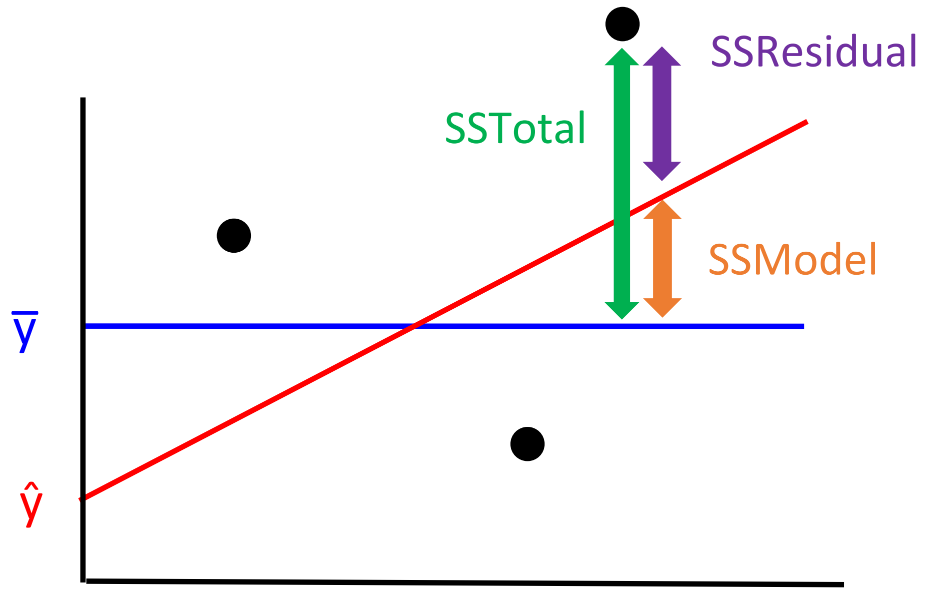
- 1. the variability explained by the model
- 2. the variability not explained by the model
    - i.e., the residuals



Analysis of Variance (ANOVA) for regression

In an analysis of variance, we break down the **total variability** in the **response variable y** into:

- 1. the variability explained by the model
- 2. the variability not explained by the model
  - i.e., the residuals



We can express this as:

- SSTotal = SSModel + SSResidual

$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Added and subtracted ŷ

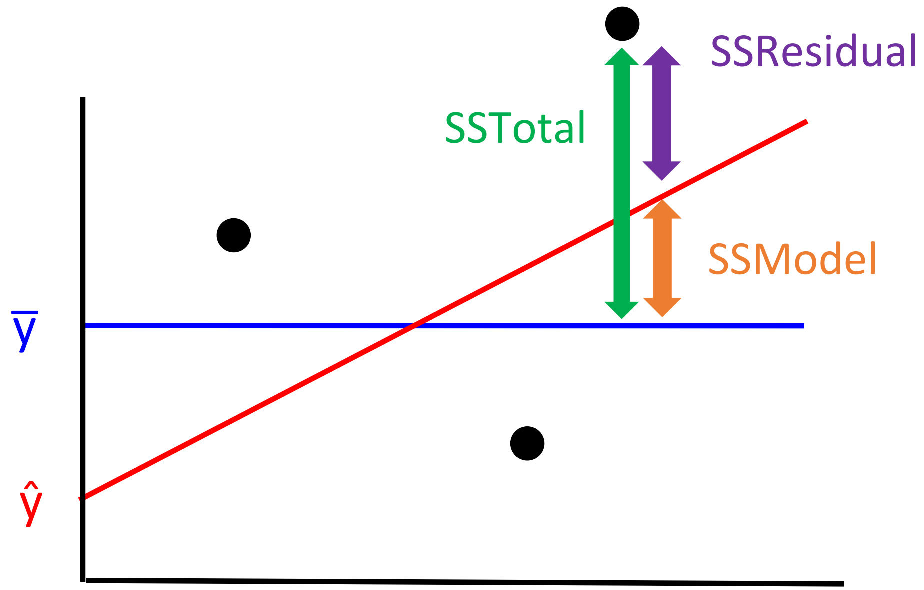
This equal 0 (proof via algebra)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + \cancel{2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}$$

The coefficient of determination r^2

The **percentage of the total variability explained by the model** is given by

$$r^2 = \frac{\text{SSModel}}{\text{SSTotal}} = 1 - \frac{\text{SSResidual}}{\text{SSTotal}}$$



We can express this as:

- SSTotal = SSModel + SSResidual

$$y - \bar{y} = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Added and subtracted ŷ

This equal 0 (proof via algebra)

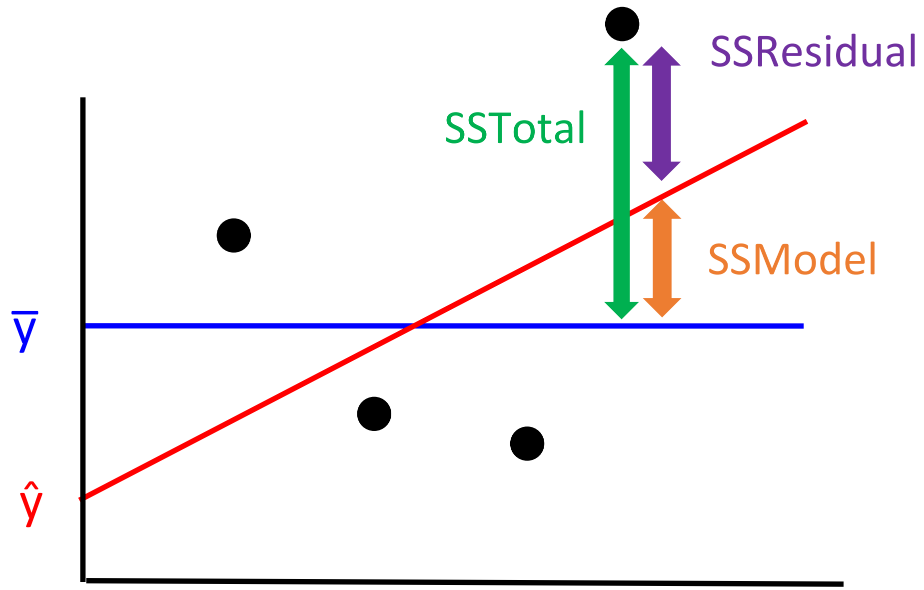
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 \cancel{+ 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}$$

Hypothesis test based on ANOVA for regression

$$F = \frac{SSModel/df_{model}}{SSResidual/df_{error}}$$

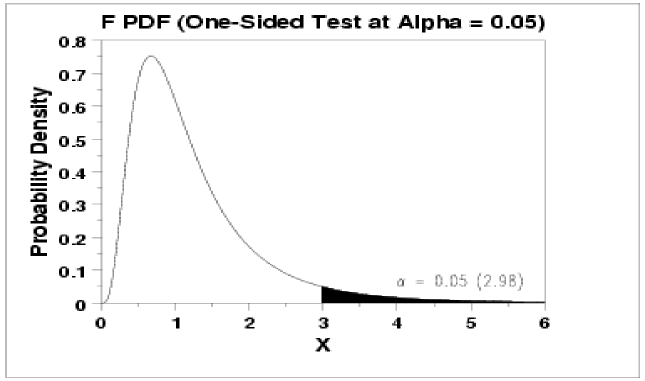
$df_{model} = 1$

$df_{error} = n - 2$



If the null hypothesis is true that $\beta_1 = 0$:

- Both the numerator and denominator are estimates of σ^2
- F comes from an F-distribution with df_{model}, df_{error} degrees of freedom
- For simple linear regression, this gives the same results as running a t-test.
  - $F = t^2$



Analysis of Variance (ANOVA) for regression in R

You can create an ANOVA table for regression relationships in R using:

- anova(lm_fit)

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

anova(lm_fit)
|
```
```

SSModel

SSResidual

F

```
Analysis of Variance Table

Response: salary_tot
                Df   Sum Sq   Mean Sq F value   Pr(>F)
log_endowment   1 132879258586 132879258586 764.29 < 0.00000000000000022 ***
Residuals   1173 203936190958   173858645

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
```

Analysis of Variance (ANOVA) for regression in R

You can create an ANOVA table for regression relationships in R using:

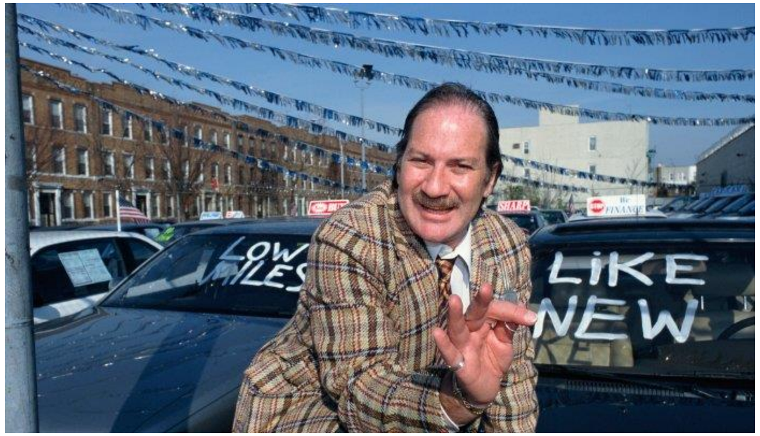
- anova(lm_fit)

We can check that the ANOVA relationships holds: SSTotal = SSModel + SSResidual using:

- The original data y values
- lm_fit$residuals
- lm_fit$fitted.values

You can also check that $F = t^2$ by comparing anova(lm_fit) and summary(lm_fit) values

Homework 7!



edmunds.com

Questions?

Multiple regression

Multiple regression

In multiple regression we try to predict a quantitative response variable *y* using several predictor variables $x_1, x_2, \dots, x_k$

For multiple linear regression, the underlying model is:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots \beta_k \cdot x_k + \epsilon$$

We estimate coefficients using a data set to make predictions ŷ

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \dots + \hat{\beta}_k \cdot x_k$$

Multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

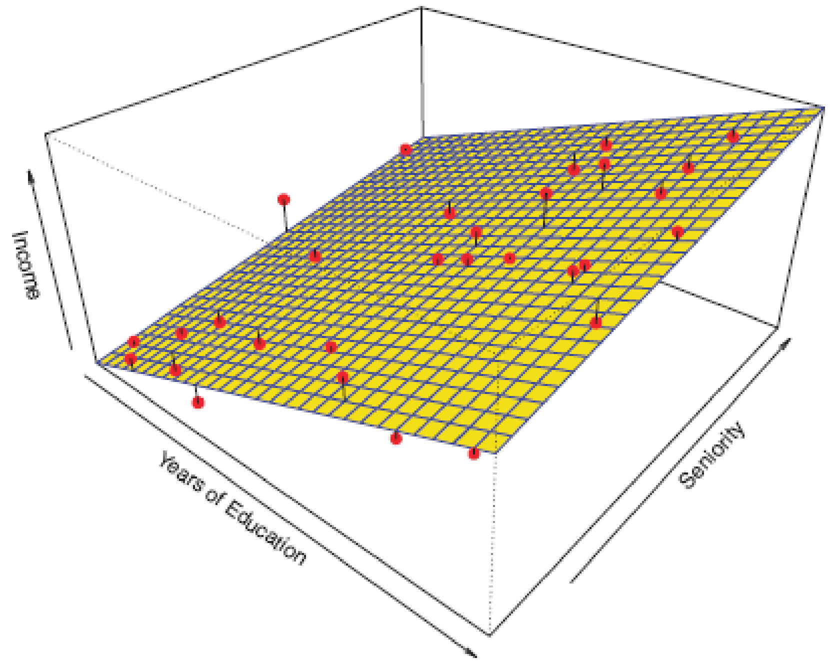
There are many uses for multiple regression models including:

- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)

Multiple regression

$$\text{salary} \;=\; \hat{\beta}_0 \;+\; \hat{\beta}_1 \cdot \text{f(endowment)} \;+\; \hat{\beta}_2 \cdot \text{g(enrollment)}$$

Let's explore this in R…



Nested model comparison

We can also assess whether a particular subset of q parameters is 0

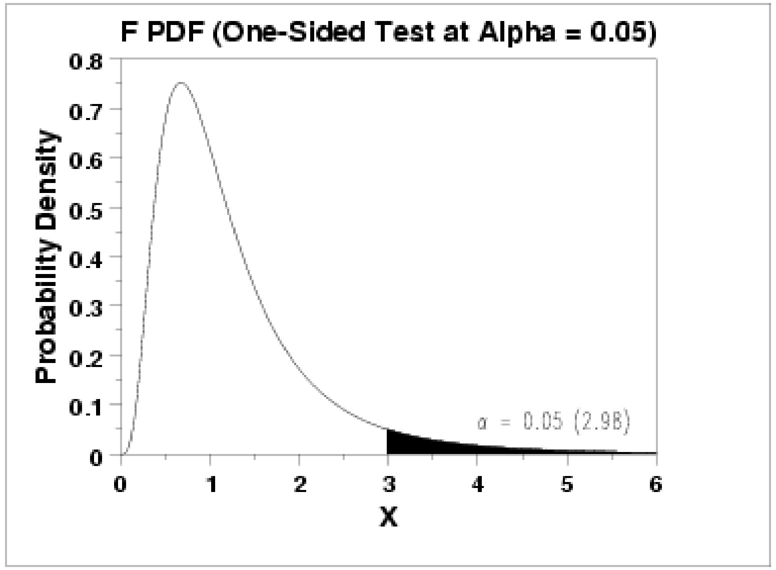
$$H_0: \beta_h = \beta_i = \ldots = \beta_g = 0$$

To do this we:

1. Fit the model without these features
2. Calculate the $SSRes_{Reduced}$ for the model without these predictors
3. Compare it to the full model $SSRes_{Full}$ with an F-statistic:

$$F = \frac{(SSRes_{Reduced} - SSRes_{Full})/q}{SSRes_{Full}/(n-k-1)}$$

where q is the number of additional terms in the full model



$df_1 = df_{Reduced} - df_{Full}$

$df_2 = df_{Full}$