Multiple regression

Overview

Analysis of variance for regression

Multiple regression

- Basic ideas
- Categorical predictors
- Interactions
- Adding non-linear variable transformations

Review of simple linear regression and analysis of variance for 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: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Hypothesis tests and confidence intervals

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

Regression diagnostics

- Normality, Homoscedasticity, Linearity and Independence

Identifying and examining usual observations

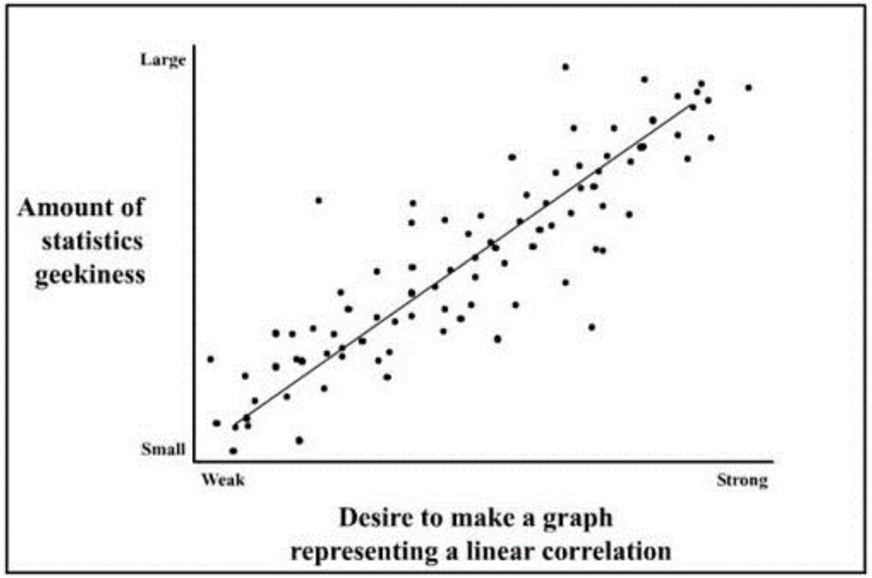
- Leverage, studentized residuals, influential points (Cook's distance)

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:

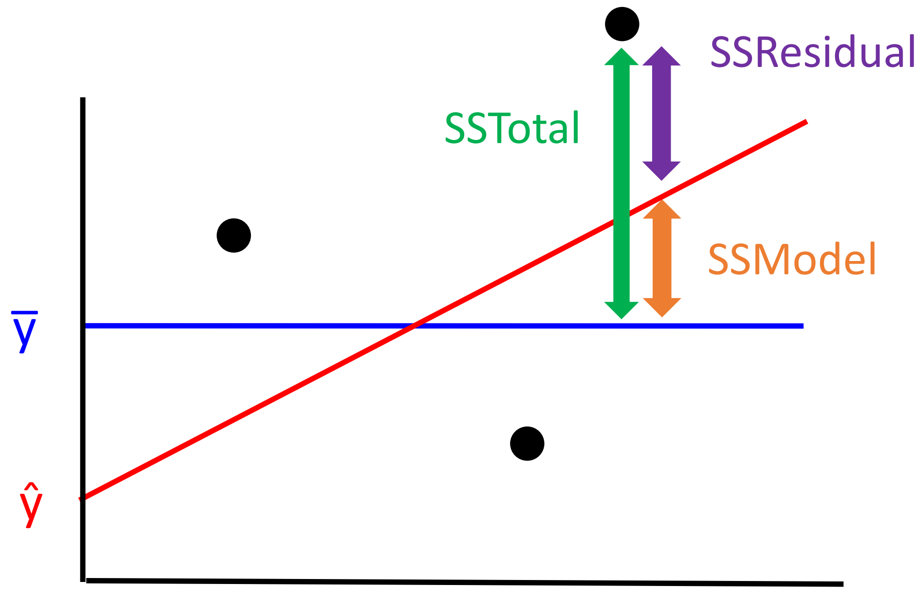
- 1. the variability explained by the model
- 2. the variability not explained by the model
  - i.e., the residuals



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

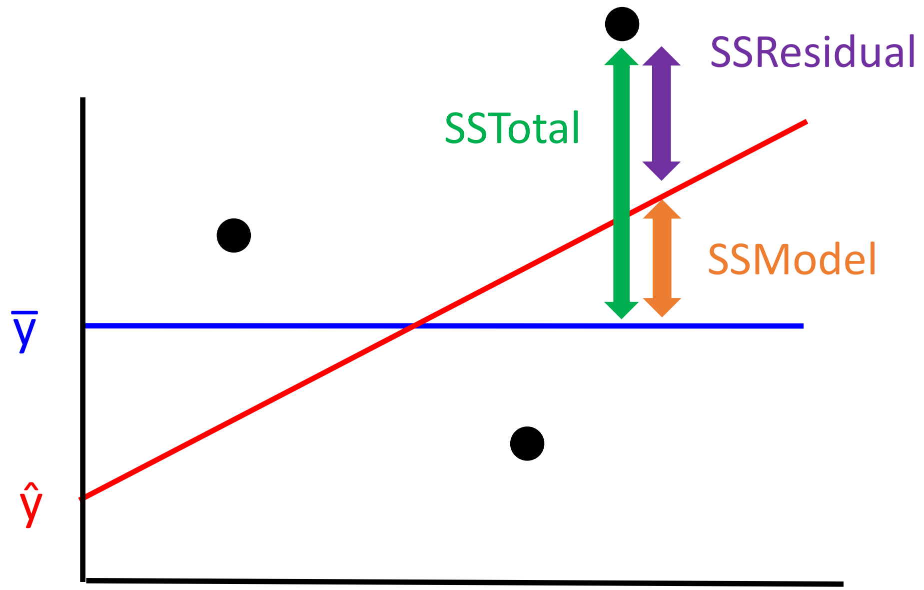
This equal 0 (proof via algebra)

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

The coefficient of determination r^2

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

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



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

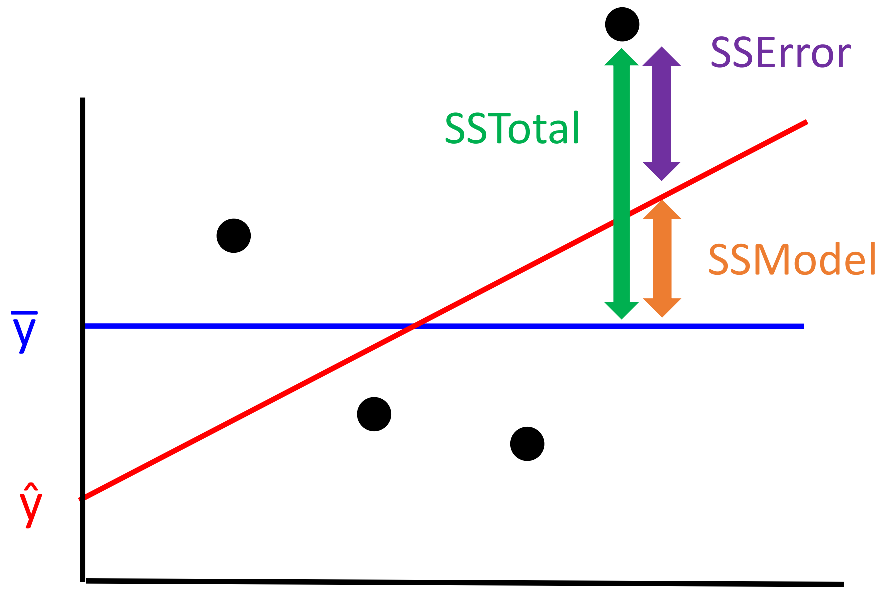
$$F = \frac{SSModel/df_{model}}{SSResidual/df_{error}}$$

$df_{model} = 1$

$df_{error} = n - 2$

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 8!

edmunds.com

Multiple regression

Multiple regression

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

For multiple linear regression, the underlying model is:

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

We estimate coefficients using a data set to make predictions ŷ

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

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:

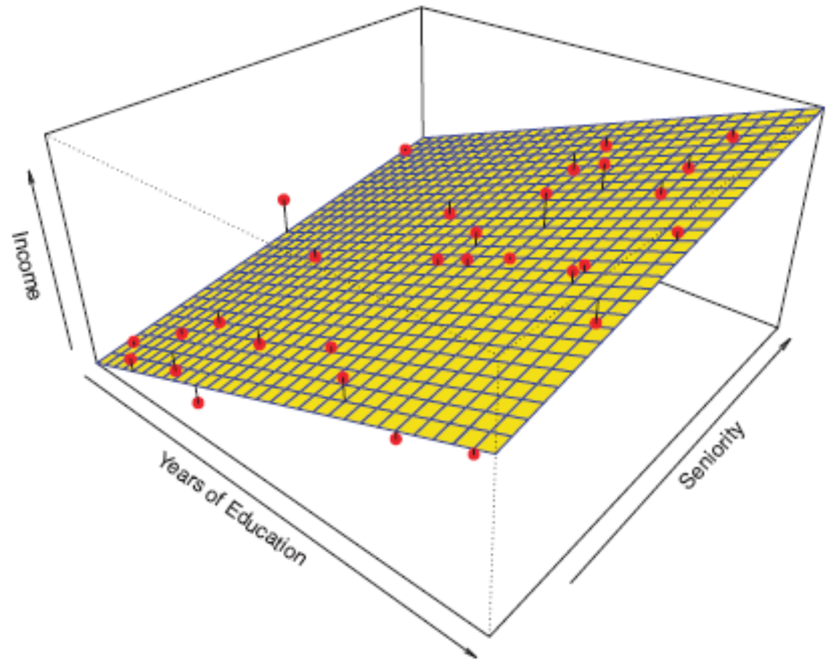
- To make predictions as accurately as possible

- 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…



Categorical predictors and interactions

Categorical predictors

Predictors can be categorical as well as quantitative

If a predictor only has two levels, we can use a single 'dummy variable' to encode these two levels:

- E.g., Assistant or Full Professor

$$x_i = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

Assistant Professors have an additional value added β_1 to their y-intercepts

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Categorical predictors

When a qualitative predictor has k levels, we need to use k -1 dummy variables to code it

- e.g., we would need two dummy variables to have different intercepts for Assistant, Associate and Full Professors

$$x_{i1} = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if Associate Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

$$y_i \; = \; \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \; = \; \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Interaction terms

The models we have looked at the relationship between the response and the predictors has been ***additive*** and ***linear***

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

These models assume that each predictor acts independently on the response y and that the relationship is linear

We can relax both of these assumptions

Interaction terms

An ***interaction effect*** occurs when the response variable y is influenced by the levels of two or more predictors in a non-additive way

For example, a professor's salary might be more effected by the size of a school's endowment depending on the number of students who attend the school

We can model this using an equation with an interaction term

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_3 (x_1 \cdot x_2) + \epsilon$$

Interaction terms

When using categorical variables, the interaction corresponds to different slopes depending for the quantitative variable depending on the value of the categorical variable

- e.g., professor's salary might be more effected by the size of a school's endowment depending whether she is an Assistant or a Full Professor

If Full Professor: salary $\approx \beta_0 + \beta_1 \cdot$ endowment

If Assistant Professor: salary $\approx (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot$ endowment

Additive term if Assistant Professor

Change in slope if Assistant Professor

Interaction terms

$$
\begin{aligned}
\text{salary} \approx \beta_0 &+ \beta_1 \cdot \text{endowment} \\
&+ \beta_2 \cdot \text{assistant\_rank\_dummy} \\
&+ \beta_3 \cdot (\text{assistant\_rank\_dummy} \cdot \text{endowment})
\end{aligned}
$$

Let’s try it in R…

Non-linear relationships

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$
\begin{aligned}
\text{salary} = \ & \beta_0 + \beta_1 \cdot \text{endowment} \\
& + \beta_2 \cdot (\text{endowment})^2 + \\
& + \beta_3 \cdot (\text{endowment})^3 + \varepsilon
\end{aligned}
$$

Still a linear equation but non-linear in original predictors

Non-linear relationships

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of covariates

We can compare model fits by:

- Assessing if higher order terms are statistically significant
- Looking at the r^2 values
- Running hypothesis tests comparing nested models
- Etc.

Let’s try it in R…