

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

# Statistics 202: Introduction to Data Mining

## Multiple Linear Regression

Jonathan Taylor  
Department of Statistics  
Stanford University

December 2, 2009

# Multiple linear regression

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Outline

- Specifying the model.
- Fitting the model: least squares.
- Interpretation of the coefficients.
- More on  $F$ -statistics.
- Matrix approach to linear regression.
- $T$ -statistics revisited.
- More  $F$  statistics.
- Tests involving more than one  $\beta$ .

# Job supervisor data

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

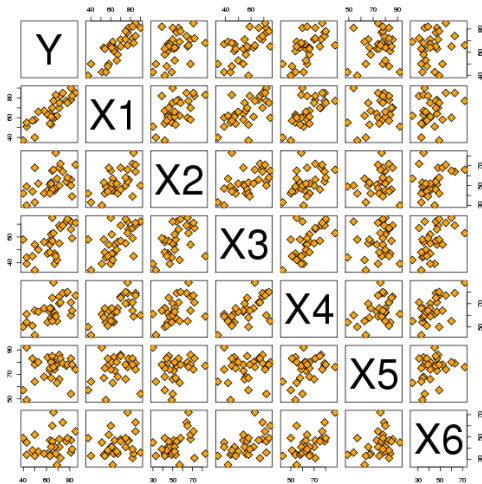
## Description

Variable	Description
$Y$	Overall supervisor job rating
$X_1$	How well do they handle complaints
$X_2$	Do they allow special priveleges
$X_3$	Give opportunity to learn new things
$X_4$	Raises based on performance
$X_5$	Too critical of poor performance
$X_6$	Good rate of advancement

# Job supervisor data

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Specifying the model

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Multiple linear regression model

- Rather than one predictor, we have  $p = 6$  predictors.



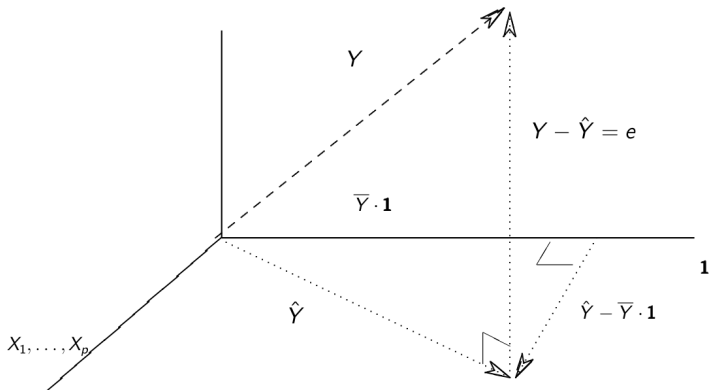
$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

- Errors  $\varepsilon$  are assumed independent  $N(0, \sigma^2)$ , as in simple linear regression.
- Coefficients are called (partial) regression coefficients because they “allow” for the effect of other variables.

# Geometry of Least Squares

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Fitting the model

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Least squares

- Just as in simple linear regression, model is fit by minimizing

$$\begin{aligned}SSE(\beta_0, \dots, \beta_p) &= \sum_{i=1}^n (Y_i - (\beta_0 + \sum_{j=1}^p \beta_j X_{ij}))^2 \\ &= \|Y - \hat{Y}(\beta)\|^2\end{aligned}$$

- Minimizers:  $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$  are the “least squares estimates”: are also normally distributed as in simple linear regression.

# Error component

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Estimating $\sigma^2$

- As in simple regression

$$\hat{\sigma}^2 = \frac{SSE}{n - p - 1} \sim \sigma^2 \cdot \frac{\chi_{n-p-1}^2}{n - p - 1}$$

independent of  $\hat{\beta}$ .

- Why  $\chi_{n-p-1}^2$ ? Typically, the degrees of freedom in the estimate of  $\sigma^2$  is  $n - \#$ number of parameters in regression function.

# Interpretation of $\beta_j$ 's

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Supervisor example

- Take  $\beta_1$  for example. This is the amount the average job rating increases for one “unit” of “Handles complaints”, keeping everything else constant.
- Units of “Handles complaints” are individual favorable responses, so on average for every extra person who rated the supervisor as good at handling complaints (other things being fixed), the average job rating increases by  $\beta_1$ .

# Goodness of fit for multiple regression

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Sums of squares

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$SSR = \sum_{i=1}^n (\bar{Y} - \hat{Y}_i)^2$$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$R^2 = \frac{SSR}{SST}$$

$R^2$  is called the *multiple correlation coefficient* of the model, or the *coefficient of multiple determination*.

# Adjusted $R^2$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Compensating for more variables

- As we add more and more variables to the model – even random ones,  $R^2$  will increase to 1.
- Adjusted  $R^2$  tries to take this into account by replacing sums of squares by *mean squares*

$$R_a^2 = 1 - \frac{SSE/(n - p - 1)}{SST/(n - 1)} = 1 - \frac{MSE}{MST}.$$

# Goodness of fit test

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Another $F$ -test

- As in simple linear regression, we measure the goodness of fit of the regression model by

$$F = \frac{MSR}{MSE} = \frac{\|\bar{Y} \cdot \mathbf{1} - \hat{Y}\|^2/p}{\|Y - \hat{Y}\|^2/(n-p-1)}.$$

- Under  $H_0 : \beta_1 = \dots = \beta_p = 0$ ,

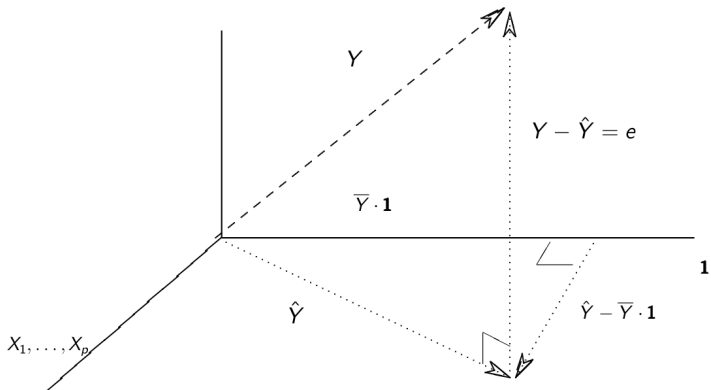
$$F \sim F_{p, n-p-1}$$

so reject  $H_0$  at level  $\alpha$  if  $F > F_{p, n-p-1, 1-\alpha}$ .

# Geometry of Least Squares

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Reasoning behind the $F$ test

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Measuring lengths

- The  $F$  statistic is a ratio of lengths of orthogonal vectors (divided by degrees of freedom).
- We can prove that our model implies

$$\mathbb{E}(MSR) = \sigma^2 + \underbrace{\|\boldsymbol{\mu} - \bar{\mu} \cdot \mathbf{1}\|^2 / p}_{(*)}$$

$$\mathbb{E}(MSE) = \sigma^2$$

$$\mu_i = \mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

so  $F$  should be not be too far from 1 if  $H_0$  is true, i.e.  
 $(*) = 0$ .

- If  $F$  is large, it is evidence that  $(*) \neq 0$ , i.e.  $H_0$  is false.

# F-test revisited

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Example in more detail

- *Full (bigger) model :*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \varepsilon_i$$

- *Reduced (smaller) model:*

$$Y_i = \beta_0 + \varepsilon_i$$

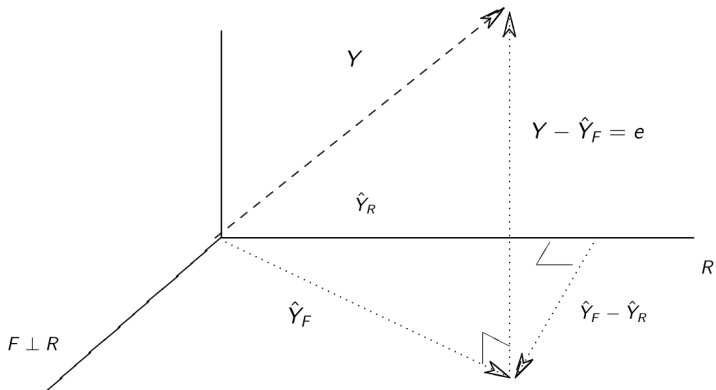
- The  $F$ -statistic has the form

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}.$$

# Geometry of Least Squares

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Matrix formulation

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Equivalent formulation

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (p+1)} \boldsymbol{\beta}_{(p+1) \times 1} + \boldsymbol{\varepsilon}_{n \times 1}$$

- $\mathbf{X}$  is called the *design matrix* of the model
- $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_{n \times n})$  is multivariate normal

## SSE in matrix form

$$SSE(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

# Matrix formulation

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Design matrix

- The design matrix is the  $n \times (p + 1)$  matrix with entries

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p} \end{pmatrix}$$

# Least squares solution

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Solving for $\hat{\beta}$

- Normal equations

$$\frac{\partial}{\partial \beta_j} SSE \Big|_{\hat{\beta}} = -2 \left( \mathbf{Y} - \mathbf{X}\hat{\beta} \right)^t \mathbf{X}_j = 0, \quad 0 \leq j \leq p.$$

- Equivalent to

$$\begin{aligned} (\mathbf{Y} - \mathbf{X}\hat{\beta})^t \mathbf{X} &= 0 \\ \hat{\beta} &= (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \end{aligned}$$

- Properties:

$$\hat{\beta} \sim N(\beta, \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}), \text{ indep. of } \hat{\sigma}^2$$

- R code

# Inference for multiple regression

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Regression function at one point

- One thing one might want to *learn* about the regression function in the supervisor example is something about the regression function at some fixed values of  $\mathbf{X}_1, \dots, \mathbf{X}_6$ , i.e. what can be said about

$$\beta_0 + 65 \cdot \beta_1 + 50 \cdot \beta_2 + 55 \cdot \beta_3 + 64 \cdot \beta_4 + 75 \cdot \beta_5 + 40 \cdot \beta_6 \quad (*)$$

roughly the regression function at “typical” values of the predictors.

- The expression (\*) is equivalent to

$$\sum_{j=0}^6 a_j \beta_j, \quad a = (1, 65, 50, 55, 64, 75, 40).$$

# Inference for $\sum_{j=0}^p a_j \beta_j$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Confidence interval for $\sum_{j=0}^p a_j \beta_j$

- Suppose we want a  $(1 - \alpha) \cdot 100\%$  CI for  $\sum_{j=0}^p a_j \beta_j$ .
- Just as in simple linear regression:

$$\sum_{j=0}^p a_j \hat{\beta}_j \pm t_{1-\alpha/2, n-p-1} \cdot SE \left( \sum_{j=0}^p a_j \hat{\beta}_j \right).$$

# Inference for $\sum_{j=0}^p a_j \beta_j$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## T-statistics revisited

- Suppose we want to test

$$H_0 : \sum_{j=0}^p a_j \beta_j = h.$$

As in simple linear regression, it is based on

$$T = \frac{\sum_{j=0}^p a_j \hat{\beta}_j - h}{SE(\sum_{j=0}^p a_j \hat{\beta}_j)}.$$

- If  $H_0$  is true, then  $T \sim t_{n-p-1}$ , so we reject  $H_0$  at level  $\alpha$  if

$$|T| \geq t_{1-\alpha/2, n-p-1}, \quad \text{OR}$$
$$p\text{-value} = 2 * (1 - \text{pt}(|T|, n - p - 1)) \leq \alpha.$$

# Inference for $\sum_{j=0}^p a_j \beta_j$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Standard error of $\sum_{j=0}^p a_j \hat{\beta}_j$

- Based on matrix approach to regression

$$SE \left( \sum_{j=0}^p a_j \hat{\beta}_j \right) = \sqrt{\hat{\sigma}^2 a (X^T X)^{-1} a^T}.$$

- Don't worry too much about implementation – R will do this for you in general, [R code](#)

# Inference for $\sum_{j=0}^p a_j \beta_j$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Prediction interval

- Identical to simple linear regression.
- Prediction interval at  $X_{1,new}, \dots, X_{p,new}$ :

$$\hat{\beta}_0 + \sum_{j=1}^p X_{j,new} \hat{\beta}_j \pm t_{1-\alpha/2, n-p-1} \times \sqrt{\hat{\sigma}^2 + SE \left( \hat{\beta}_0 + \sum_{j=1}^p X_{j,new} \hat{\beta}_j \right)^2}.$$

# Inference for multiple regression

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Questions about many (combinations) of $\beta_j$ 's

- In multiple regression we can ask more complicated questions than in simple regression.
- For instance, we could ask whether
  - $X_2$  : Do they allow special privileges
  - $X_3$  : Give opportunity to learn new thingsexplains little of the variability in the data, and might be dropped from the regression model.
- These questions can be answered by  $F$ -statistics.

# Inference for more than one $\beta$

## Dropping one or more variables

- Suppose we wanted to test whether how the supervisor handles special privileges, or allows employees to try new things explains a significant amount of the variability in the overall job rating. Formally, this is:

$$H_0 : \beta_2 = \beta_3 = 0, \quad \text{vs. } H_a : \text{one of } \beta_2, \beta_3 \neq 0$$

- This test is again an  $F$ -test based on two models

$$R : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + \varepsilon_i$$

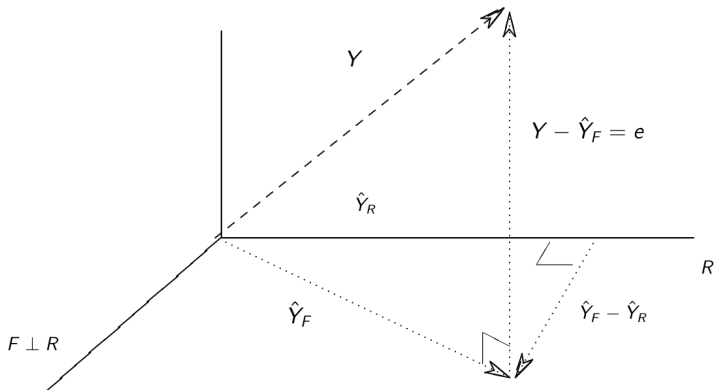
$$F : Y_i = \beta_0 + \sum_{j=1}^6 \beta_j X_{ij} + \varepsilon_i$$

- **Note:** The reduced model  $R$  must be a special case of (nested within) the full model  $F$  to use the  $F$ -test.

# Geometry of Least Squares

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## SSE of a model

- In the graphic, a “model”,  $\mathcal{M}$  is a subspace of  $\mathbb{R}^n =$  column space of  $\mathbf{X}$ .
- Least squares fit = projection onto the subspace of  $\mathcal{M}$ , yielding predicted values  $\hat{Y}_{\mathcal{M}}$
- Error sum of squares:

$$SSE(\mathcal{M}) = \|Y - \hat{Y}_{\mathcal{M}}\|^2.$$

# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## F-statistic for $H_0 : \beta_2 = \beta_3 = 0$

- 

$$F = \frac{\frac{SSE(R) - SSE(F)}{2}}{\frac{SSE(F)}{n-1-p}} \quad (1)$$

$\sim F_{2, n-p-1}$  (if  $H_0$  is true)

- Reject  $H_0$  at level  $\alpha$  if  $F > F_{1-\alpha, 2, n-1-p}$ .
- Here is an example [R code](#).

# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## Dropping an arbitrary subset

- For an arbitrary model, suppose we want to test

$$H_0 : \beta_{i_1} = \dots = \beta_{i_j} = 0$$

$$H_a : \text{one of } \beta_{i_1}, \dots, \beta_{i_j} \neq 0$$

for some subset  $\{i_1, \dots, i_j\} \subset \{0, \dots, p\}$ .

# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

*F*-statistic for  $H_0 : \beta_{i_1} = \dots = \beta_{i_j} = 0$

- You guessed it: it is based on the two models:

$$R : Y_i = \sum_{l=0, l \notin \{i_1, \dots, i_j\}}^p \beta_l X_{il} + \varepsilon_i$$

$$F : Y_i = \sum_{j=0}^p \beta_j X_{ij} + \varepsilon_i$$

where  $X_{i0} = 1$  for all  $i$ .

# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

$F$ -statistic for  $H_0 : \beta_{i_1} = \dots = \beta_{i_j} = 0$

- 

$$F = \frac{\frac{SSE(R) - SSE(F)}{j}}{\frac{SSE(F)}{n-p-1}}$$

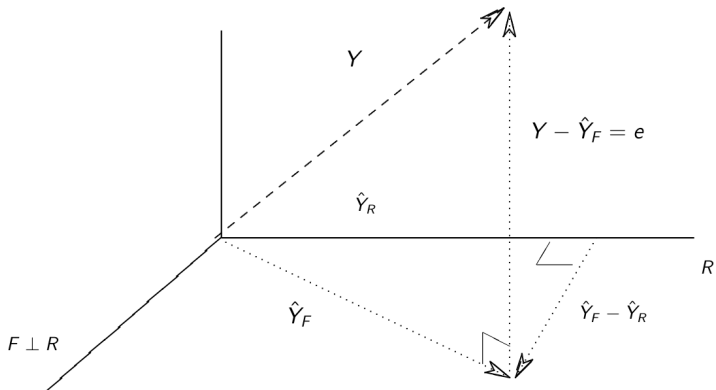
$\sim F_{j, n-p-1}$  (if  $H_0$  is true)

- Reject  $H_0$  at level  $\alpha$  if  $F > F_{1-\alpha, j, n-1-p}$ .

# Geometry of Least Squares

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University



# Inference for more than one $\beta$

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

## General $F$ -tests

- Given two models  $R \subset F$  (i.e.  $R$  is a subspace of  $F$ ), we can consider testing

$$H_0 : R \text{ is adequate (i.e. } \mathbb{E}(Y) \in R)$$

vs.

$$H_a : F \text{ is adequate (i.e. } \mathbb{E}(Y) \in F).$$

- The test statistic is

$$F = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$$

- If  $H_0$  is true,  $F \sim F_{df_R - df_F, df_R}$  so we reject  $H_0$  at level  $\alpha$  if  $F > F_{df_R - df_F, df_R, 1 - \alpha}$ .

# Constraints

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

Constraining  $\beta_1 = \beta_3$  (after deciding  $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ )

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$$

- Reduced model:

$$\begin{aligned} Y_i &= \beta_0 + \tilde{\beta}_1 X_{i1} + \tilde{\beta}_1 X_{i3} + \varepsilon_i \\ &= \beta_0 + \tilde{\beta}_1 (X_{i1} + X_{i3}) + \varepsilon_i \end{aligned}$$

R code.

# Constraints

Statistics 202:  
Introduction  
to Data  
Mining

Jonathan  
Taylor  
Department of  
Statistics  
Stanford  
University

Constraining  $\beta_1 + \beta_3 = 1$  (after deciding  
 $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ )

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \varepsilon_i$$

- Reduced model:

$$Y_i = \beta_0 + \tilde{\beta}_1 X_{i1} + (1 - \tilde{\beta}_1) X_{i3} + \varepsilon_i$$

$$Y_i - X_{i3} = \beta_0 + \tilde{\beta}_1 (X_{i1} - X_{i3}) + \varepsilon_i$$

R code.