CS 147:
Computer Systems Performance Analysis
Multiple and Categorical Regression

CS 147:
Computer Systems Performance Analysis
Multiple and Categorical Regression

Overview

Overview
Multiple Linear Regression
Basic Formulas
Example
Quality of the Example
Categorical Models

Multiple Linear Regression

- Develops models with more than one predictor variable
- But each predictor variable has linear relationship to response variable
- Conceptually, plotting a regression line in *n*-dimensional space, instead of 2-dimensional

Multiple Linear Regression

- Develops models with more than one predictor variable
- But each predictor variable has linear relationship to response variable
- Conceptually, plotting a regression line in *n*-dimensional space, instead of 2-dimensional

Basic Multiple Linear Regression Formula

Response y is a function of k predictor variables $x_1, x_2, \ldots, x_k$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

Basic Multiple Linear Regression Formula

Response y is a function of k predictor variables $x_1, x_2, \ldots, x_k$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

A Multiple Linear Regression Model

Given sample of n observations

$$\{(x_{11}, x_{21}, \ldots, x_{k1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)\}$$

model consists of n equations (note possible $+$ vs. $-$ typo in book):

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{11} + b_2 x_{21} + \cdots + b_k x_{k1} + e_1 \\
y_2 &= b_0 + b_1 x_{12} + b_2 x_{22} + \cdots + b_k x_{k2} + e_2 \\
&\vdots \\
y_n &= b_0 + b_1 x_{1n} + b_2 x_{2n} + \cdots + b_k x_{kn} + e_n
\end{aligned}$$

A Multiple Linear Regression Model

Given sample of n observations

$$\{(x_{11}, x_{21}, \ldots, x_{k1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)\}$$

model consists of n equations (note possible $+$ vs. $-$ typo in book):

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{11} + b_2 x_{21} + \cdots + b_k x_{k1} + e_1 \\
y_2 &= b_0 + b_1 x_{12} + b_2 x_{22} + \cdots + b_k x_{k2} + e_2 \\
&\vdots \\
y_n &= b_0 + b_1 x_{1n} + b_2 x_{2n} + \cdots + b_k x_{kn} + e_n
\end{aligned}$$

Looks Like It's Matrix Arithmetic Time

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Note that:

- $\mathbf{y}$ and $\mathbf{e}$ have n elements
- $\mathbf{b}$ has $k + 1$
- $\mathbf{x}$ is k by n

Analysis of Multiple Linear Regression

- Listed in box 15.1 of Jain
- Not terribly important (for our purposes) how they were derived
  - This isn't a class on statistics
- But you need to know how to use them
- Mostly matrix analogs to simple linear regression results

Analysis of Multiple Linear Regression

- Listed in box 15.1 of Jain
- Not terribly important (for our purposes) how they were derived
  - This isn't a class on statistics
- But you need to know how to use them
- Mostly matrix analogs to simple linear regression results

Example of Multiple Linear Regression

- IMDB keeps numerical popularity ratings of movies
- Postulate popularity of Academy Award-winning films is based on two factors:
  - Year made
  - Running time
- Produce a regression

$$\text{rating} = b_0 + b_1(\text{year}) + b_2(\text{length})$$

Example of Multiple Linear Regression

- IMDB keeps numerical popularity ratings of movies
- Postulate popularity of Academy Award-winning films is based on two factors:
  - Year made
  - Running time
- Produce a regression

$$\text{rating} = b_0 + b_1(\text{year}) + b_2(\text{length})$$

Some Sample Data

Title	Year	Length	Rating
Silence of the Lambs	1991	118	8.1
Terms of Endearment	1983	132	6.8
Rocky	1976	119	7.0
Oliver!	1968	153	7.4
Marty	1955	91	7.7
Gentleman's Agreement	1947	118	7.5
Mutiny on the Bounty	1935	132	7.6
It Happened One Night	1934	105	8.0

Some Sample Data

Title	Year	Length	Rating
Silence of the Lambs	1991	118	8.1
Terms of Endearment	1983	132	6.8
Rocky	1976	119	7.0
Oliver!	1968	153	7.4
Marty	1955	91	7.7
Gentleman's Agreement	1947	118	7.5
Mutiny on the Bounty	1935	132	7.6
It Happened One Night	1934	105	8.0

Now for Some Tedious Matrix Arithmetic

- We need to calculate $\mathbf{X}$, $\mathbf{X}^\mathsf{T}$, $\mathbf{X}^\mathsf{T}\mathbf{X}$, $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, and $\mathbf{X}^\mathsf{T}\mathbf{y}$
- Because $\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$
- We will see that $\mathbf{b} = (18.5430, -0.0051, -0.0086)$
- Meaning the regression predicts:

$$\text{rating} = 18.5430 - 0.0051(\text{year}) - 0.0086(\text{length})$$

Now for Some Tedious Matrix Arithmetic

- We need to calculate $\mathbf{X}$, $\mathbf{X}^\mathsf{T}$, $\mathbf{X}^\mathsf{T}\mathbf{X}$, $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$, and $\mathbf{X}^\mathsf{T}\mathbf{y}$
- Because $\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$
- We will see that $\mathbf{b} = (18.5430, -0.0051, -0.0086)$
- Meaning the regression predicts:

$$\text{rating} = 18.5430 - 0.0051(\text{year}) - 0.0086(\text{length})$$

X Matrix for Example

$$\mathbf{X} = \begin{bmatrix} 1 & 1991 & 118 \\ 1 & 1983 & 132 \\ 1 & 1976 & 119 \\ 1 & 1968 & 153 \\ 1 & 1955 & 91 \\ 1 & 1947 & 118 \\ 1 & 1935 & 132 \\ 1 & 1934 & 105 \end{bmatrix}$$

X Matrix for Example

$$\mathbf{X} = \begin{bmatrix} 1 & 1991 & 118 \\ 1 & 1983 & 132 \\ 1 & 1976 & 119 \\ 1 & 1968 & 153 \\ 1 & 1955 & 91 \\ 1 & 1947 & 118 \\ 1 & 1935 & 132 \\ 1 & 1934 & 105 \end{bmatrix}$$

Transpose to Get $\mathbf{X}^\mathsf{T}$

$$\mathbf{X}^\mathsf{T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1991 & 1983 & 1976 & 1968 & 1955 & 1947 & 1935 & 1934 \\ 118 & 132 & 119 & 153 & 91 & 118 & 132 & 105 \end{bmatrix}$$

Transpose to Get $\mathbf{X}^\mathsf{T}$

$$\mathbf{X}^\mathsf{T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1991 & 1983 & 1976 & 1968 & 1955 & 1947 & 1935 & 1934 \\ 118 & 132 & 119 & 153 & 91 & 118 & 132 & 105 \end{bmatrix}$$

Multiply To Get $\mathbf{X}^\mathsf{T}\mathbf{X}$

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 8 & 15689 & 968 \\ 15689 & 30771385 & 1899083 \\ 968 & 1899083 & 119572 \end{bmatrix}$$

Multiply To Get $\mathbf{X}^\mathsf{T}\mathbf{X}$

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 8 & 15689 & 968 \\ 15689 & 30771385 & 1899083 \\ 968 & 1899083 & 119572 \end{bmatrix}$$

Invert to Get $\mathbf{C} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$

$$\mathbf{C} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{bmatrix} 1207.7585 & \text{-}0.6240 & 0.1328 \\ \text{-}0.6240 & 0.0003 & \text{-}0.0001 \\ 0.1328 & \text{-}0.0001 & 0.0004 \end{bmatrix}$$

Invert to Get $\mathbf{C} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$

$$\mathbf{C} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{bmatrix} 1207.7585 & \text{-}0.6240 & 0.1328 \\ \text{-}0.6240 & 0.0003 & \text{-}0.0001 \\ 0.1328 & \text{-}0.0001 & 0.0004 \end{bmatrix}$$

Multiply to Get $\mathbf{X}^\mathsf{T}\mathbf{y}$

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \begin{bmatrix} 60.1 \\ 117840.7 \\ 7247.5 \end{bmatrix}$$

Multiply to Get $\mathbf{X}^\mathsf{T}\mathbf{y}$

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \begin{bmatrix} 60.1 \\ 117840.7 \\ 7247.5 \end{bmatrix}$$

Multiply $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$ to Get $\mathbf{b}$

$$\mathbf{b} = \begin{bmatrix} 18.5430 \\ -0.0051 \\ -0.0086 \end{bmatrix}$$

Multiply $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}(\mathbf{X}^\mathsf{T}\mathbf{y})$ to Get $\mathbf{b}$

$$\mathbf{b} = \begin{bmatrix} 18.5430 \\ -0.0051 \\ -0.0086 \end{bmatrix}$$

How Good Is This Regression Model?

- How accurately does model predict film rating based on age and running time?
- Best way to determine this analytically is to calculate errors:

$$\mathrm{SSE} = \mathbf{yTy} - \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

or

$$\mathrm{SSE} = \sum e_i^2$$

How Good Is This Regression Model?
- How accurately does model predict film rating based on age and running time?
- Best way to determine this analytically is to calculate errors:
$$\mathrm{SSE} = \mathbf{yTy} - \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$
or
$$\mathrm{SSE} = \sum e_i^2$$

Calculating the Errors

Year	Length	Rating	Estimated Rating	e_i	e_i^2
1991	118	8.1	7.4	−0.71	0.51
1983	132	6.8	7.3	0.51	0.26
1976	119	7.0	7.5	0.45	0.21
1968	153	7.4	7.2	−0.20	0.04
1955	91	7.7	7.8	0.10	0.01
1947	118	7.5	7.6	0.11	0.01
1935	132	7.6	7.6	−0.05	0.00
1934	105	8.0	7.8	−0.21	0.04

Calculating the Errors

Year	Length	Rating	Estimated Rating	e_i	e_i^2
1991	118	8.1	7.4	−0.71	0.51
1983	132	6.8	7.3	0.51	0.26
1976	119	7.0	7.5	0.45	0.21
1968	153	7.4	7.2	−0.20	0.04
1955	91	7.7	7.8	0.10	0.01
1947	118	7.5	7.6	0.11	0.01
1935	132	7.6	7.6	−0.05	0.00
1934	105	8.0	7.8	−0.21	0.04

Calculating the Errors, Continued

- SSE $= 1.08$
- SSY $= \sum y_i^2 = 452.91$
- SS0 $= n\overline{y}^2 = 451.5$
- SST $=$ SSY $-$ SS0 $= 452.9 - 451.5 = 1.4$
- SSR $=$ SST $-$ SSE $= 0.33$
- $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{0.33}{1.41} = 0.23$
- In other words, this regression stinks

Calculating the Errors, Continued

- SSE $= 1.08$
- SSY $= \sum y_i^2 = 452.91$
- SS0 $= n\overline{y}^2 = 451.5$
- SST $=$ SSY $-$ SS0 $= 452.9 - 451.5 = 1.4$
- SSR $=$ SST $-$ SSE $= 0.33$
- $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{0.33}{1.41} = 0.23$
- In other words, this regression stinks

Why Does It Stink?

- Let's look at properties of the regression parameters

$$s_e = \sqrt{\frac{\text{SSE}}{n-3}} = \sqrt{\frac{1.08}{5}} = 0.46$$

- Now calculate standard deviations of the regression parameters (These are estimations only, since we're working with a sample)
- Estimated stdev of

$$b_0 \text{ is } s_e\sqrt{c_{00}} = 0.46\sqrt{1207.76} = 16.16$$
$$b_1 \text{ is } s_e\sqrt{c_{11}} = 0.46\sqrt{0.0003} = 0.0084$$
$$b_2 \text{ is } s_e\sqrt{c_{22}} = 0.46\sqrt{0.0004} = 0.0097$$

Why Does It Stink?
- Let's look at properties of the regression parameters
$$s_e = \sqrt{\frac{\text{SSE}}{n-3}} = \sqrt{\frac{1.08}{5}} = 0.46$$
- Now calculate standard deviations of the regression parameters (These are estimations only, since we're working with a sample)
- Estimated stdev of
$$b_0 \text{ is } s_e\sqrt{c_{00}} = 0.46\sqrt{1207.76} = 16.16$$
$$b_1 \text{ is } s_e\sqrt{c_{11}} = 0.46\sqrt{0.0003} = 0.0084$$
$$b_2 \text{ is } s_e\sqrt{c_{22}} = 0.46\sqrt{0.0004} = 0.0097$$

Calculating Confidence Intervals of STDEVs

- We will use 90% level
- Confidence intervals for

$$
\begin{aligned}
b_0 &\text{ is } 18.54 \mp 2.015(16.16) = (-14.02, 51.10) \\
b_1 &\text{ is } 0.005 \mp 2.015(0.0084) = (-0.022, 0.012) \\
b_2 &\text{ is } 0.009 \mp 2.015(0.0097) = (-0.028, 0.011)
\end{aligned}
$$

- None is significant at this level

Calculating Confidence Intervals of STDEVs
- We will use 90% level
- Confidence intervals for
b_0 is $18.54 \mp 2.015(16.16) = (-14.02, 51.10)$
b_1 is $0.005 \mp 2.015(0.0084) = (-0.022, 0.012)$
b_2 is $0.009 \mp 2.015(0.0097) = (-0.028, 0.011)$
- None is significant at this level

Analysis of Variance

- So, can we really say that none of the predictor variables are significant?
  - Not yet; predictors may be correlated
- F-tests can be used for this purpose
  - E.g., to determine if the SSR is significantly higher than the SSE
  - Equivalent to testing that y does not depend on any of the predictor variables
- Alternatively, that no b_i is significantly nonzero

Analysis of Variance

- So, can we really say that none of the predictor variables are significant?
  - Not yet; predictors may be correlated
- F-tests can be used for this purpose
  - E.g., to determine if the SSR is significantly higher than the SSE
  - Equivalent to testing that y does not depend on any of the predictor variables
- Alternatively, that no b_i is significantly nonzero

Running an F-Test

- Need to calculate SSR and SSE
- From those, calculate mean squares of regression (MSR) and errors (MSE)
- MSR/MSE has an F distribution
- If $\mathrm{MSR}/\mathrm{MSE} > \mathrm{F}_{\mathrm{table}}$, predictors explain significant fraction of response variation
- Note typos in book's table 15.3
  - SSR has k degrees of freedom
  - SST matches $y - \bar{y}$, not $y - \hat{y}$

Running an F-Test

- Need to calculate SSR and SSE
- From those, calculate mean squares of regression (MSR) and errors (MSE)
- MSR/MSE has an F distribution
- If $\mathrm{MSR}/\mathrm{MSE} > \mathrm{F}_{\mathrm{table}}$, predictors explain significant fraction of response variation
- Note typos in book's table 15.3
  - SSR has k degrees of freedom
  - SST matches $y - \bar{y}$, not $y - \hat{y}$

F-Test for Our Example

- $\text{SSR} = .33$
- $\text{SSE} = 1.08$
- $\text{MSR} = \text{SSR}/k = .33/2 = .16$
- $\text{MSE} = \text{SSE}/(n - k - 1) = 1.08/(8 - 2 - 1) = .22$
- $\text{F-computed} = MSR/MSE = .76$
- $\text{F}[90; 2, 5] = 3.78$
- So it fails the F-test at 90% (miserably)

F-Test for Our Example
- SSR = .33
- SSE = 1.08
- MSR = SSR/k = .33/2 = .16
- MSE = SSE/(n − k − 1) = 1.08/(8 − 2 − 1) = .22
- F-computed = MSR/MSE = .76
- F[90; 2, 5] = 3.78
- So it fails the F-test at 90% (miserably)

Multicollinearity

- If two predictor variables are linearly dependent, they are collinear
  - Meaning they are related
  - And thus second variable does not improve regression
  - In fact, it can make it worse
- Typical symptom is inconsistent results from various significance tests

Multicollinearity

- If two predictor variables are linearly dependent, they are collinear
  - Meaning they are related
  - And thus second variable does not improve regression
  - In fact, it can make it worse
- Typical symptom is inconsistent results from various significance tests

Finding Multicollinearity

- Must test correlation between predictor variables
- If it’s high, eliminate one and repeat regression without it
- If significance of regression improves, it’s probably due to collinearity between the variables

Finding Multicollinearity

- Must test correlation between predictor variables
- If it’s high, eliminate one and repeat regression without it
- If significance of regression improves, it’s probably due to collinearity between the variables

Is Multicollinearity a Problem in Our Example?

- Probably not, since significance tests are consistent
- But let's check, anyway
- Calculate correlation of age and length
- After tedious calculation, 0.25
  - Not especially correlated
- Important point—**adding a predictor variable does not always improve a regression**
  - See example on p. 253 of book

2015-06-15

CS147
└Multiple Linear Regression
 └Quality of the Example
  └Is Multicollinearity a Problem in Our Example?

Is Multicollinearity a Problem in Our Example?

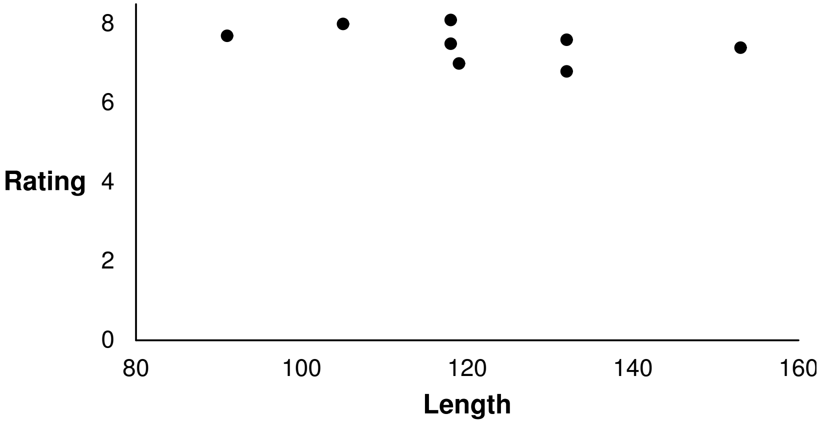
- Probably not, since significance tests are consistent
- But let's check, anyway
- Calculate correlation of age and length
- After tedious calculation, 0.25
  - Not especially correlated
- Important point—**adding a predictor variable does not always improve a regression**
  - See example on p. 253 of book

Why Didn't Regression Work Well Here?

- Check scatter plots
  - Rating vs. year
  - Rating vs. length
- Regardless of how good or bad regressions look, **always check the scatter plots**

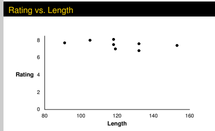
Why Didn't Regression Work Well Here?
- Check scatter plots
  - Rating vs. year
  - Rating vs. length
- Regardless of how good or bad regressions look, **always check the scatter plots**

Rating vs. Length

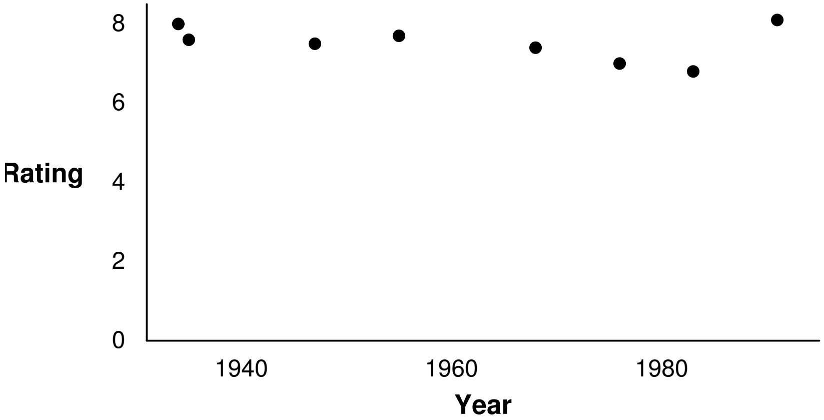


2015-06-15

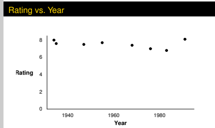
CS147
Multiple Linear Regression
Quality of the Example
Rating vs. Length



Rating vs. Year



2015-06-15
CS147
Multiple Linear Regression
Quality of the Example
Rating vs. Year



Regression With Categorical Predictors

- Regression methods discussed so far assume numerical variables
- What if some of your variables are categorical in nature?
- If *all* are categorical, use techniques discussed later in the course
- **Levels**: number of values a category can take

Regression With Categorical Predictors

- Regression methods discussed so far assume numerical variables
- What if some of your variables are categorical in nature?
- If *all* are categorical, use techniques discussed later in the course
- **Levels**: number of values a category can take

Handling Categorical Predictors

- If only two levels, define b_i as follows
  - $x_i = 0$ for first value
  - $x_i = 1$ for second value
- (This definition is missing from book in section 15.2)
- Can use +1 and -1 as values, instead
- Need $k - 1$ predictor variables for k levels
  - To avoid implying order in categories

Handling Categorical Predictors

- If only two levels, define b_i as follows
  - $x_i = 0$ for first value
  - $x_i = 1$ for second value
- (This definition is missing from book in section 15.2)
- Can use +1 and -1 as values, instead
- Need $k - 1$ predictor variables for k levels
  - To avoid implying order in categories

Categorical Variables Example

Which is a better predictor of a high rating in the movie database?

- Winning an Oscar?
- Winning the Golden Palm at Cannes?
- Winning the New York Critics Circle?

Categorical Variables Example

Which is a better predictor of a high rating in the movie database?
- Winning an Oscar?
- Winning the Golden Palm at Cannes?
- Winning the New York Critics Circle?

Choosing Variables

- Categories are not mutually exclusive
- $x_1 = 1$ if Oscar, 0 otherwise
- $x_2 = 1$ if Golden Palm, 0 otherwise
- $x_3 = 1$ if Critics Circle Award, 0 otherwise
- $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

Choosing Variables

- Categories are not mutually exclusive
- $x_1 = 1$ if Oscar, 0 otherwise
- $x_2 = 1$ if Golden Palm, 0 otherwise
- $x_3 = 1$ if Critics Circle Award, 0 otherwise
- $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

A Few Data Points

Title	Rating	Oscar	Palm	NYC
Gentleman's Agreement	7.5	X		X
Mutiny on the Bounty	7.6	X		
Marty	7.4	X	X	X
If	7.8		X	
La Dolce Vita	8.1		X	
Kagemusha	8.2		X	
The Defiant Ones	7.5			X
Reds	6.6			X
High Noon	8.1			X

A Few Data Points

Title	Rating	Oscar	Palm	NYC
Gentleman's Agreement	7.5	X		X
Mutiny on the Bounty	7.6	X		
Marty	7.4	X	X	X
If	7.8		X	
La Dolce Vita	8.1		X	
Kagemusha	8.2		X	
The Defiant Ones	7.5			X
Reds	6.6			X
High Noon	8.1			X

And Regression Says. . .

- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?

And Regression Says. . .

- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?

And Regression Says...

- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?
- R^2 is 34% of variation
  - Better than age and length
  - But still no great shakes

And Regression Says...
- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?
- R^2 is 34% of variation
  - Better than age and length
  - But still no great shakes

And Regression Says...

- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?
- R^2 is 34% of variation
  - Better than age and length
  - But still no great shakes
- Are regression parameters significant at 90% level?

And Regression Says...

- $\hat{y} = 7.8 - 0.1x_1 + 0.2x_2 - 0.4x_3$
- How good is that?
- R^2 is 34% of variation
  - Better than age and length
  - But still no great shakes
- Are regression parameters significant at 90% level?