Lecture 8: Linear Regression

May 24, 2012
GENOME 560, Spring 2012

Su-In Lee, CSE & GS
suinlee@uw.edu

Goals

- Develop basic concepts of linear regression from a probabilistic framework
- Estimating parameters and hypothesis testing with linear models
- Linear regression in R

Regression

- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships

Why Linear Regression?

- Suppose we want to model the outcome variable Y in terms of three predictors, X_1, X_2, X_3

$$Y = f(X_1, X_2, X_3)$$

- Typically will not have enough data to try and directly estimate f
- Therefore, we usually have to assume that it has some restricted form, such as **linear**

$$Y = X_1 + X_2 + X_3$$

Regression Terminology

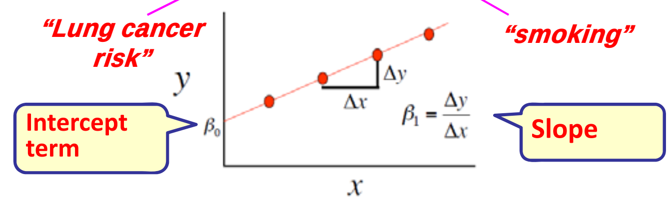
$$\mathbf{Y = X_1 + X_2 + X_3}$$

Dependent Variable	Independent Variable
Outcome Variable	Predictor Variable
Response Variable	Explanatory Variable
Lung cancer risk	Genetic factor, smoking, diet, etc
Expression level of gene X	Expression levels of X's TFs A, B and C

Linear Regression is a Probabilistic Model

- Much of mathematics is devoted to studying variables that are deterministically related to one another

$$Y = \beta_0 + \beta_1 X$$



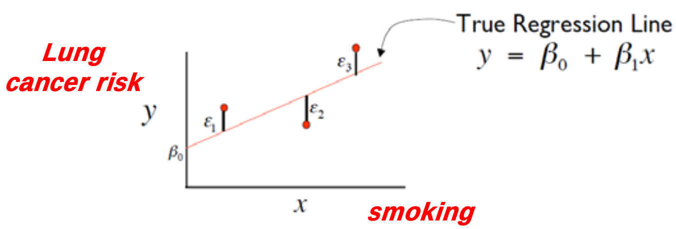
- But we're interested in understanding the relationship between variables related **in a nondeterministic fashion**

A Linear Probabilistic Model

- **Definition:** There exists parameters β_0, β_1 and σ^2, such that for any fixed value of the predictor variable X, the outcome variable Y is related to X through the model equation

$$Y = \beta_0 + \beta_1 X + \varepsilon\,,$$

where ε is a RV assumed to be $N(0, \sigma^2)$



Implications

- The **expected value of Y** is a linear function of X, but for fixed value x, the variable Y differs from its expected value by a *random amount*

Variables & Symbols: How is x different from X ?

Capital letter X: a random variable
Lower case letter x: corresponding values
(i.e. the real numbers the RV X map into)

For example,
X: Genotype of a certain locus
x: 0, 1 or 2 (meaning AA, AG and GG, respectively)

Implications

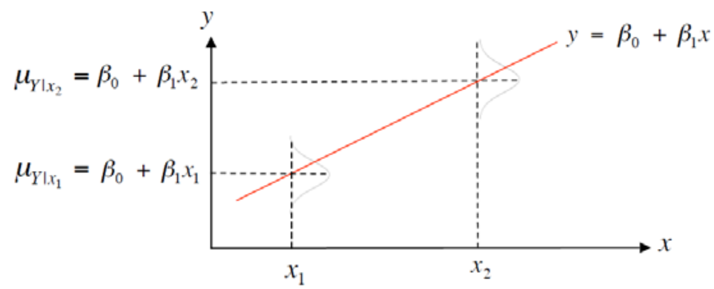
- The **expected value of Y** is a linear function of X, but for fixed value x, the variable Y differs from its expected value by a *random amount*

- Formally, let x^* denote a particular value of the predictor variable X, then our linear probabilistic model says:

$$E(Y \mid x^*) = \mu_{Y|x^*} = \text{mean value of } Y \text{ when } X \text{ is } x^*$$

$$V(Y \mid x^*) = \sigma^2_{Y|x^*} = \text{variance of } Y \text{ when } X \text{ is } x^*$$

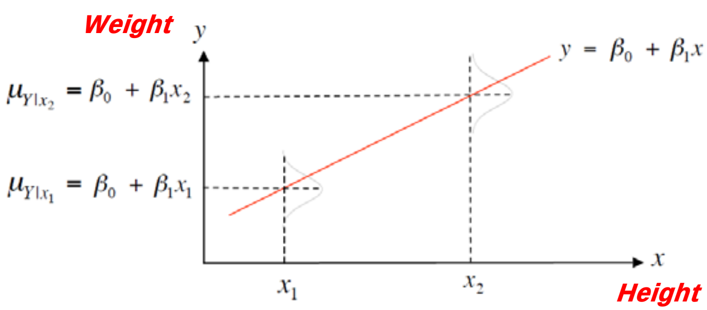
Graphical Interpretation



$$E(Y \mid x^*) = \mu_{Y|x^*} = \text{mean value of } Y \text{ when } X \text{ is } x^*$$

$$V(Y \mid x^*) = \sigma^2_{Y|x^*} = \text{variance of } Y \text{ when } X \text{ is } x^*$$

Graphical Interpretation



- Say that X = height and Y = weight
- Then $\mu_{Y|x=60}$ is the average weight for all individuals 60 inches tall in the population

One More Example

- Suppose the relationship between the predictor variable height (X) and outcome variable weight (Y) is described by a simple linear regression model with true regression line

$$Y = 7.5 + 0.5\,X, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2) \text{ and } \sigma = 3$$

- Q1: What is the interpretation of $\beta_1 = 0.5$?

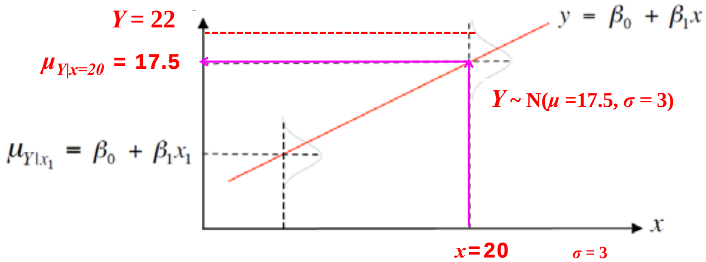
  The expected change in weight (Y) associated with a 1-unit increase in height (X)

- Q2: If $x = 20$, what is the expected value of Y?

  $\mu_{Y|x=20} = 7.5 + 0.5\,(20) = 17.5$

One More Example

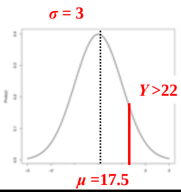
- Q3: If x = 20, what is P(Y>22) ?



- Given Y ~ N(μ =17.5, σ = 3),

$$P(Y > 22 \mid x = 20) = 1 - \phi\left(\frac{22 - 17.5}{3}\right) = 1 - \phi(1.5) = 0.067$$

where ϕ means the CDF of Normal dist. N(0,1)



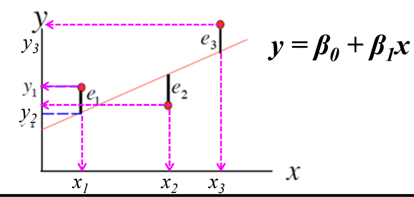
Estimating Model Parameters

- Where are the parameters β_0 and β_1 from?
- ***Predicted***, or fitted, values are values of y predicted by plugging $x_1, x_2, \ldots, x_n$ into the estimated regression line: $y = \beta_0 + \beta_1 x$

$$\hat{y}_1 = \beta_0 + \beta_1 x_1$$
$$\hat{y}_2 = \beta_0 + \beta_1 x_2$$
$$\hat{y}_3 = \beta_0 + \beta_1 x_3$$

- ***Residuals*** are the deviations of observed (red dots) and predicted values (red line)

$$e_1 = y_1 - \hat{y}_1$$
$$e_2 = y_2 - \hat{y}_2$$
$$e_3 = y_3 - \hat{y}_3$$

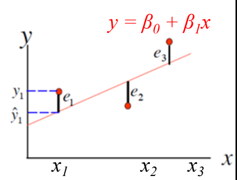


Residuals Are Useful!

- The error sum of squares (SSE) can tell us how well the line fits to the data

$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_1 = \beta_0 + \beta_1 x_1$$
$$\hat{y}_2 = \beta_0 + \beta_1 x_2$$
$$\hat{y}_3 = \beta_0 + \beta_1 x_3$$

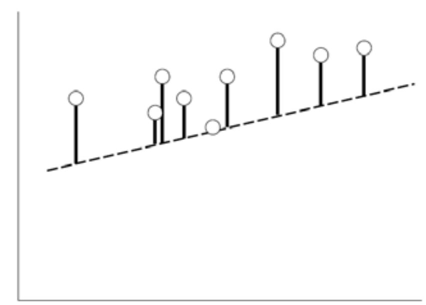


- ***Least squares***
  - Find β_0 and β_1 that minimizes SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

  - Denote the solutions by $\hat{\beta}_0$ and $\hat{\beta}_1$

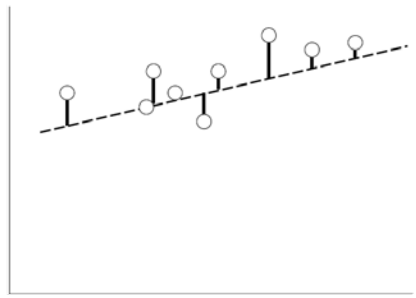
Least Squares



- ***Least squares***
  - Find β_0 and β_1 that minimizes SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

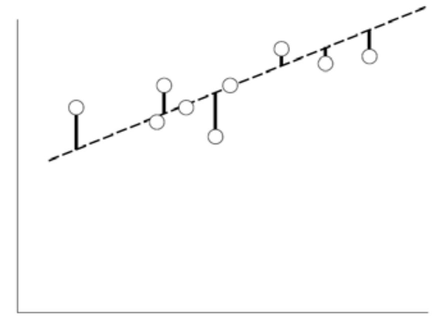
Least Squares



- ***Least squares***
  - Find β_0 and β_1 that minimizes SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

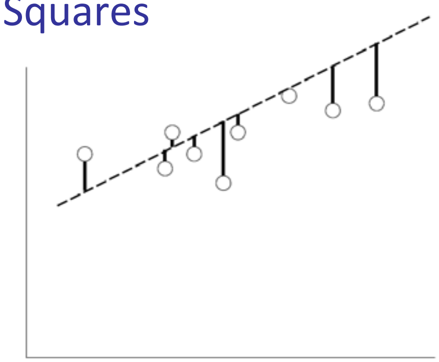
Least Squares



- ***Least squares***
  - Find β_0 and β_1 that minimizes SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Least Squares



- ***Least squares***
  - Find β_0 and β_1 that minimizes SSE

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Coefficient of Determination

- Important statistic referred to as the coefficient of determination (R^2):

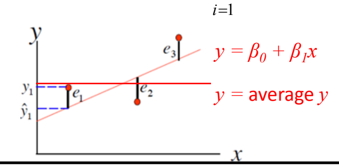
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Error Sum Squares

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Error Sum Squares, when β_0= avg(y) and β_1=0



Multiple Linear Regression

- Extension of the simple linear regression model to two or more independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Expression = Baseline + Age + Tissue + Sex + Error

- **Partial Regression Coefficients:**

  $\beta_i \equiv$ effect on the outcome variable when increasing the i^{th} predictor variable by 1 unit, **holding all other predictors constant**

Categorical Independent Variables

- Qualitative variables are easily incorporated in regression framework through ***dummy variables***
- Simple example: sex can be coded as 0/1
- What if my categorical variable contains three levels:

$$X_i = \begin{cases} 0 & \text{if AA} \\ 1 & \text{if AG} \\ 2 & \text{if GG} \end{cases}$$

  - NO!

Categorical Independent Variables

- Previous coding would result in ***collinearity***
- Solution is to set up a series of dummy variable. In general for k levels you need (k-1) dummy variables

$$X_1 = \begin{cases} 1 & \text{if AA} \\ 0 & \text{otherwise} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if AG} \\ 0 & \text{otherwise} \end{cases}$$

$$X_i = \begin{cases} 0 & \text{if AA} \\ 1 & \text{if AG} \\ 2 & \text{if GG} \end{cases}$$

	X_1	X_2
AA	1	0
AG	0	1
GG	0	0

Hypothesis Testing: Model Utility Test

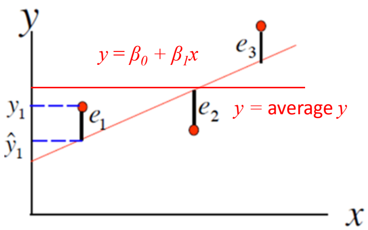
- The first thing we want to know after fitting a model is whether any of the predictor variables (X's) are significantly related to the outcome variable (Y):

$$\begin{aligned} &H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_A: \text{At least one } \beta_i \neq 0 \end{aligned}$$

- Let's frame this in our ANOVE framework
- In ANOVA, we partitioned total variance (SST) into two components:
  - SSE (unexplained variation)
  - SSR (variation explained by linear model)

Model Utility Test

- Partition total variance (SST) into two components:
  - SSE (unexplained variation)
  - SSR (variation explained by linear model)
- Let's consider n (=3) data points and k (=1) predictor model



$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

ANOVA Formulation of Model Utility Test

- Partition total variance (SST) into two components:
  - SSE (unexplained variation)
  - SSR (variation explained by linear model)

Source of Variation	df	Sum of Squares	MS	F
Regression	k	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$\frac{SSR}{k}$	$\frac{MS_R}{MS_E}$
Error	n-(k+1)	$SSE = \sum (y_i - \hat{y}_i)^2$	$\frac{SSE}{n\text{-}(k+1)}$	
Total	n-1	$SST = \sum (y_i - \bar{y})^2$		

data points – (# parameters in the model)

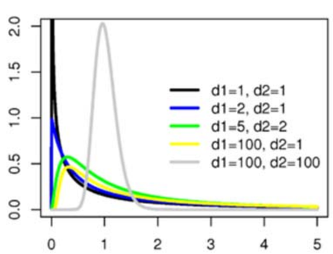
Rejection Region : $F_{\alpha,k,n-(k+1)}$

ANOVA Formulation of Model Utility Test

- F-test statistic

$$F = \frac{MS_R}{MS_E} = \frac{SSR/k}{SSE/[n-(k+1)]} = \frac{R^2}{1-R^2} \cdot \frac{n-(k+1)}{k}$$

Rejection Region : $F_{\alpha,k,n-(k+1)}$



- Pick the distribution function, based on k and n-(k+1).
- Choose the critical value based on α ($F_{\alpha,k,n-(k+1)}$)
  - Say that α =0.05
  - Prob(F>$F_{\alpha,k,n-(k+1)}$) = 0.05

F Test For Subsets of Independent Variables

- A powerful tool in multiple regression analysis is the ability to compare two models
- For instance say we want to compare

$$\text{Full Model}: \; y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

$$\text{Reduced Model}: \; y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Again, another example of ANOVA

SSE_R = error sum of squares for reduced model with l predictors

SSE_F = error sum of squares for full model with k predictors

$$F = \frac{(\text{SSE}_R - \text{SSE}_F)/(k-l)}{\text{SSE}_F/[n-(k+1)]}$$

Example of Model Comparison

- We have a quantitative trait and want to test the effects at two markers, M1 and M2.

$$\text{Full Model: Trait} = \text{Mean} + \text{M1} + \text{M2} + (\text{M1} \times \text{X2}) + \varepsilon$$

$$\text{Reduced Model: Trait} = \text{Mean} + \text{M1} + \text{M2} + \varepsilon$$

$$F = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})/(k-l)}{\text{SSE}_\text{F}/[n-(k+1)]} = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})/(3-2)}{\text{SSE}_\text{F}/[100-(3+1)]}$$

$$= \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})}{\text{SSE}_\text{F}/96}$$

$$\text{Rejection Region: } F_{\alpha,1,96}$$

How To Do In R

- You can fit a least-squares regression using the function
  - mm <- lsfit(x,y)
- The coefficients of the fit are then given by
  - mm$coef
- The residuals are
  - mm$residuals
- And to print out the tests for zero slope just do
  - ls.print (mm)

Input Data

- http://www.cs.washington.edu/homes/suinlee/genome560/data/cats.txt
- Data on fluctuating proportions of marked cells in marrow from heterozygous Safari cats
- Proportions of cells of one cell type in samples from cats (taken in our department many years ago). Column 1 is the ID number of the particular cat. You will want to plot the data from one cat.
  - For example cat 40004 is rows 1:17, 40005a is 18:31, 40005b is 32:47, 40006 is 48:65, 40665 is 66:83 and so on.

Input Data

- **2nd column:** Time, in weeks from the start of monitoring, that the measurement from marrow is recorded.
- **3rd column:** Percent of domestic-type progenitor cells observed in a sample of cells at that time.
- **4th column:** Sample size at that time, i.e. the number of progenitor cells analyzed.

Cat #1	40004	11	33	72
	40004	13	49	67
	40004	19	46	56
	40004	25	42	19
	40004	28	68	59
	40004	31	55	64
	40004	33	38	61
	40004	36	23	73
	40004	41	32	170
	40004	45	41	120
	40004	48	50	70
	40004	50	54	39
	40004	52	30	143
	40004	54	30	56
	40004	56	32	78
	40004	58	18	74
	40004	62	36	81
Cat #2	40005a	14	34	65
	40005a	17	26	74
	40005a	23	21	73
	40005a	26	11	72
	40005a	29	19	77
	40005a	31	20	70
	40005a	34	13	56
	40005a	37	17	65