

# **Linear Regression**

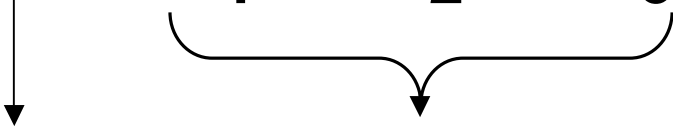
# Goals

- 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

# Regression Lingo

$$\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$$


Dependent Variable

Independent Variable

Outcome Variable

Predictor Variable

Response Variable

Explanatory Variable

# Why Linear Regression?

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

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

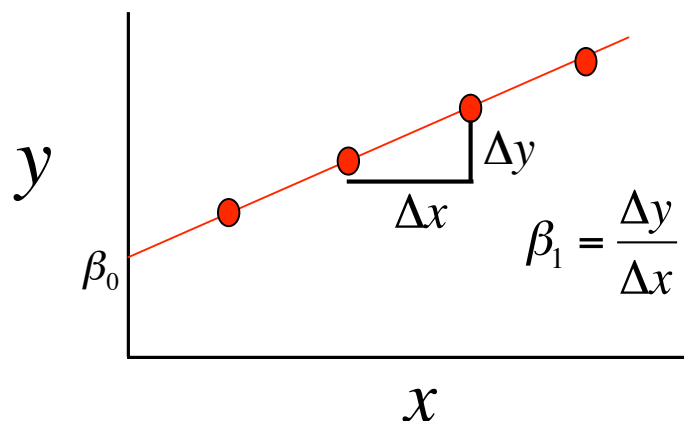
- Typically will not have enough data to try and directly estimate  $f$
- Therefore, we usually have to assume that it has some restricted form, such as linear

$$Y = X_1 + X_2 + X_3$$

# Linear Regression is a Probabilistic Model

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

$$y = \beta_0 + \beta_1 x$$



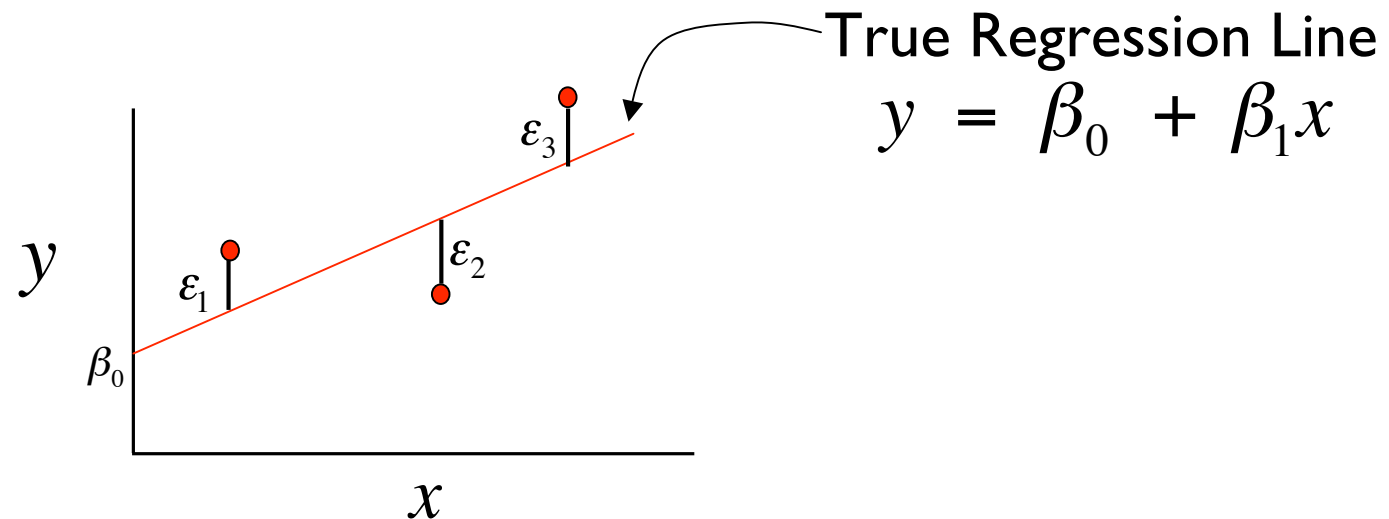
- But we're interested in understanding the relationship between variables related in a nondeterministic fashion

# A Linear Probabilistic Model

- Definition: There exists parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ , such that for any fixed value of the independent variable  $x$ , the dependent variable is related to  $x$  through the model equation

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- $\varepsilon$  is a rv assumed to be  $N(0, \sigma^2)$



# Implications

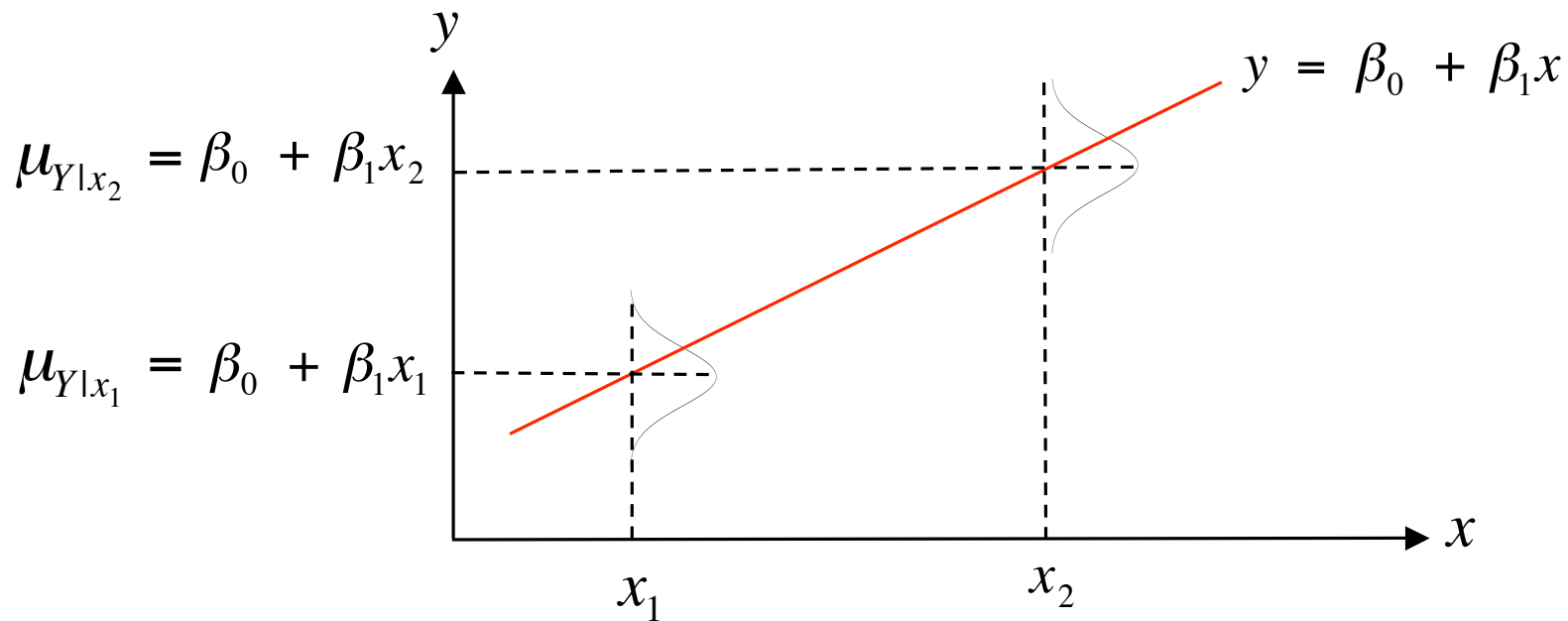
- The **expected** value of  $Y$  is a linear function of  $X$ , but for fixed  $x$ , the variable  $Y$  differs from its expected value by a random amount
- Formally, let  $x^*$  denote a particular value of the independent variable  $x$ , then our linear probabilistic model says:

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

$$V(Y \mid x^*) = \sigma_{Y \mid x^*}^2 = \text{variance of } Y \text{ when } x \text{ is } x^*$$



# Graphical Interpretation



- For example, if  $x$  = 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 independent variable height (x) and dependent variable weight (y) is described by a simple linear regression model with true regression line  
$$y = 7.5 + 0.5x \text{ and } \sigma = 3$$

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

The expected change in height associated with a 1-unit increase in weight

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

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

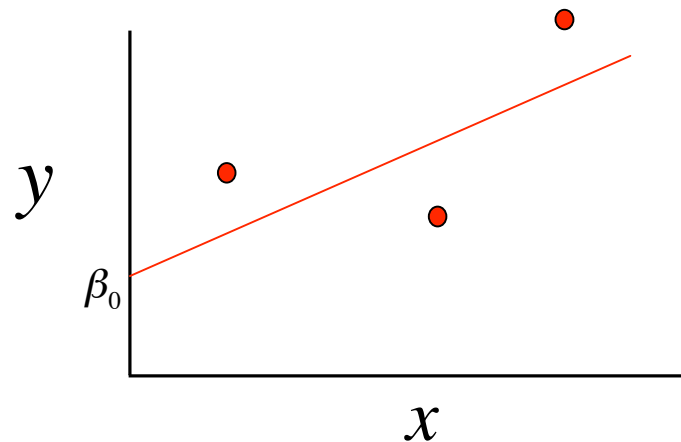
- Q3: If  $x = 20$  what is  $P(Y > 22)$ ?

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

# Estimating Model Parameters

- Point estimates of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are obtained by the principle of least squares

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



- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

# Predicted and Residual Values

- **Predicted**, or fitted, values are values of  $y$  predicted by the least-squares regression line obtained by plugging in  $x_1, x_2, \dots, x_n$  into the estimated regression line

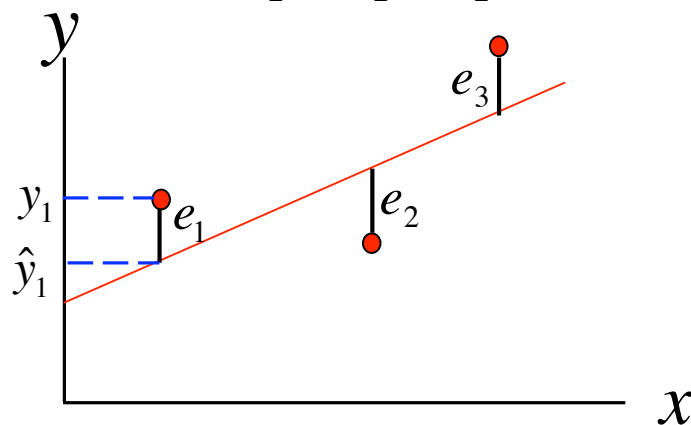
$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$$

- **Residuals** are the deviations of observed and predicted values

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$



# Residuals Are Useful!

- They allow us to calculate the error sum of squares (SSE):

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

- Which in turn allows us to estimate  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

- As well as an important statistic referred to as the coefficient of determination:

$$r^2 = 1 - \frac{SSE}{SST} \qquad SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

# Multiple Linear Regression

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

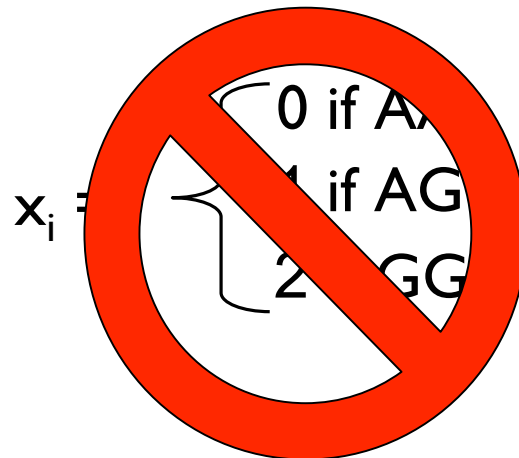
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

Expression = Baseline + Age + Tissue + Sex + Error

- Partial Regression Coefficients:  $\beta_i \equiv$  effect on the dependent variable when increasing the  $i^{\text{th}}$  independent 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:



# Categorical Independent Variables

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

$$x_1 = \begin{cases} 1 & \text{if AA} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if AG} \\ 0 & \text{otherwise} \end{cases}$$

	$x_1$	$x_2$
AA	1	0
AG	0	1
GG	0	0



# Hypothesis Testing: Model Utility Test (or Omnibus Test)

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

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_A : \text{At least one } \beta_1 \neq 0$$

$$f = \frac{R^2}{(1 - R^2)} \cdot \frac{k}{n - (k + 1)}$$

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

# Equivalent ANOVA Formulation of Omnibus Test

- We can also frame this in our now familiar ANOVA framework
  - partition total variation into two components: **SSE** (unexplained variation) and **SSR** (variation explained by linear model)

# Equivalent ANOVA Formulation of Omnibus Test

- We can also frame this in our now familiar ANOVA framework
  - partition total variation into two components: **SSE** (unexplained variation) and **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-2	$SSE = \sum (y_i - \hat{y}_i)^2$	$\frac{SSE}{n-2}$	
Total	n-1	$SST = \sum (y_i - \bar{y})^2$		

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

# F Test For Subsets of Independent Variables

- A powerful tool in multiple regression analyses 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

$$f = \frac{(SSE_R - SSE_F)/(k - l)}{SSE_F / ([n - (k + 1)])}$$

# Example of Model Comparison

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

Full Model: Trait = Mean + M1 + M2 + (M1\*M2) + error

Reduced Model: Trait = Mean + M1 + M2 + error

$$f = \frac{(SSE_R - SSE_F)/(3 - 2)}{SSE_F / ([100 - (3 + 1)])} = \frac{(SSE_R - SSE_F)}{SSE_F / 96}$$

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

# Hypothesis Tests of Individual Regression Coefficients

- Hypothesis tests for each  $\hat{\beta}_i$  can be done by simple t-tests:

$$H_0 : \hat{\beta}_i = 0$$

$$H_A : \hat{\beta}_i \neq 0$$

$$T = \frac{\hat{\beta}_i - \beta_i}{se(\beta_i)}$$

Critical value :  $t_{\alpha/2, n-(k-1)}$

- Confidence Intervals are equally easy to obtain:

$$\hat{\beta}_i \pm t_{\alpha/2, n-(k-1)} \bullet se(\hat{\beta}_i)$$

# Checking Assumptions

- Critically important to examine data and check assumptions underlying the regression model
  - Outliers
  - Normality
  - Constant variance
  - Independence among residuals
- Standard diagnostic plots include:
  - scatter plots of  $y$  versus  $x_i$  (outliers)
  - qq plot of residuals (normality)
  - residuals versus fitted values (independence, constant variance)
  - residuals versus  $x_i$  (outliers, constant variance)
- We'll explore diagnostic plots in more detail in R

# Fixed -vs- Random Effects Models

- In ANOVA and Regression analyses our independent variables can be treated as **Fixed** or **Random**
- **Fixed Effects**: variables whose levels are either sampled exhaustively or are the only ones considered relevant to the experimenter
- **Random Effects**: variables whose levels are randomly sampled from a large population of levels
- Example from our recent AJHG paper:

Expression = Baseline + Population + Individual + Error