Introduction to Linear Regression

Statistical inference

- Deriving predictions on the population from sample outcomes
- Parameters
  - Unknown mathematical characteristics of the population
- Statistics
  - Sample characteristics
  - Estimators and test statistics

Developing models

- Specification
  – Dependent and independent variables
  – Functional form
- Estimation
  – Finding unknown parameters of the model
- Application / prediction

Example

- Specification

$$p(i \mid k) \qquad \text{(why?)}$$

- Estimation

$$\hat{p}(i=1 \mid k=1) = \frac{60}{150} = 0.4$$

$$\hat{p}(i=1 \mid k=2) = \frac{200}{300} = 0.667$$

$$\hat{p}(i=1 \mid k=3) = \frac{140}{150} = 0.933$$

Prediction

- Impact of changes in the independent variables
  - Assume model parameters are stable

Income				
	Low k=1	Medium k=2	High k=3	Total
Population	10%	45%	45%	100%
Yes i=1	0.4x10 =4%	0.667x45 =30%	0.933x45 =42%	76%

The Estimation Problem

- Our a priori knowledge about the travel demand process is limited
- There are parameters in the models whose values we do not know
- The simple linear regression model

$$y = \beta_1 + \beta_2 x + \varepsilon$$

where

y = the dependent variable

β_1, β_2 = unknown parameters

x = the independent variable

ε = the disturbance term

We assume that the function form is known.

However we do not know the parameters β_1, β_2.

The goal of model estimation is to make inferences about their value.

The Estimation Problem (continued)

- The general model

$$y \approx f(x, \theta)$$

where

y - a random variable

x - a vector of known variables that influence the distribution of y

$Y \sim f(x, \theta)$

f - the distribution of y

θ- a vector of parameters, at least some of which are unknown apriori

Using a sample of observation from the process being modeled, drawn in some known way from the whole population, a function of the observations is constructed to estimate the unknown parameters. Such a function is called an *estimator*

Estimators

- Sample statistics to indicate on population parameters

Sample $Y_1, Y_2, \ldots, Y_N$

Average $\bar{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n = \hat{\mu}$

Variance $s^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(Y_n - \bar{Y}\right)^2 = \hat{\sigma}^2$

Model estimation

- Unknown population parameter values

$$Y = f(X, \theta) + \varepsilon$$

- Use sample of observations to infer about unknown parameters
- Estimator
  - Function of observations
- Estimate
  - Realized value of the estimator for a given sample

Estimation

- **Estimator:** Statistic whose calculated value is used to estimate a population parameter, θ

- **Estimate**: A particular realization of an estimator, $\hat{\theta}$

- **Types of Estimators**:

  - point estimate: single number that can be regarded as the most plausible value of θ

  - interval estimate: a range of numbers, called a confidence interval indicating, can be regarded as likely containing the true value of θ

Examples for Estimators

$$\hat{\mu}_N^1 = g_1(z_1, z_2, \ldots, z_n) = \frac{1}{N}\sum_{n=1}^{N} z_n$$

$$\hat{\mu}_N^2 = g_2(z_1, z_2, \ldots, z_n) = \frac{\max(z_n) + \min(z_n)}{2}$$

$$\hat{\mu}_N^3 = g_3(z_1, z_2, \ldots, z_n) = \mathrm{median}(z_1, z_2, \ldots, z_n)$$

$$\hat{\mu}_N^4 = g_4(z_1, z_2, \ldots, z_n) = \frac{1}{N-1}\sum_{n=1}^{N} z_n$$

Properties of Good Estimators

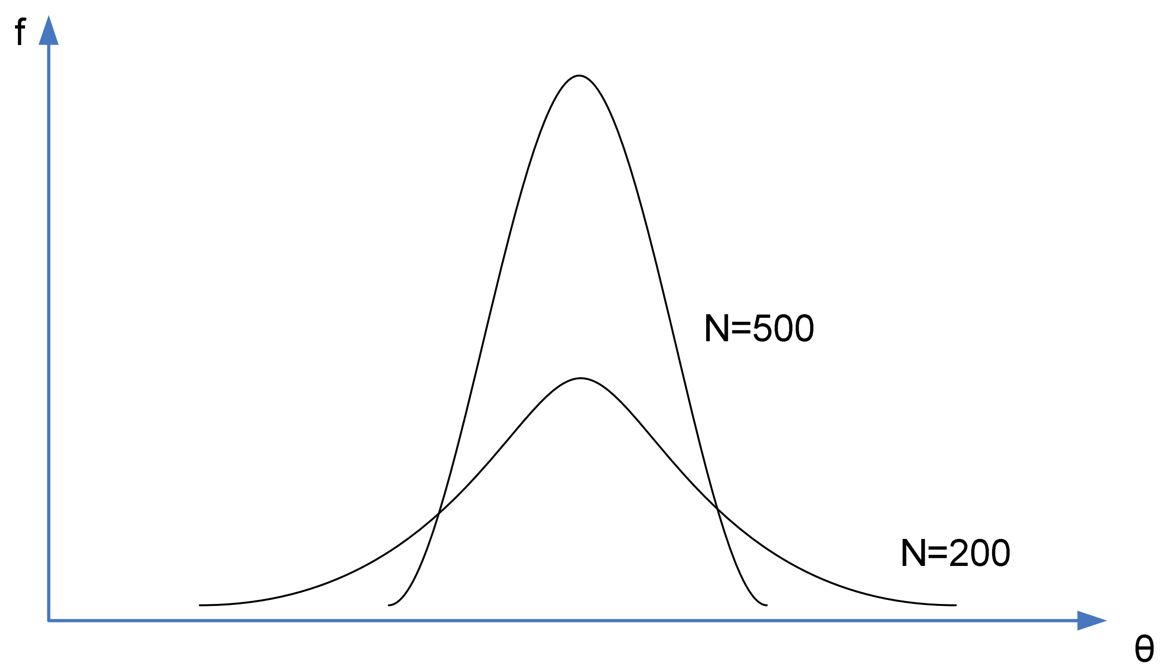
- In the ***Frequentist*** world view parameters are fixed, statistics are rv and vary from sample to sample (i.e., have an associated sampling distribution)
- In theory, there are many potential estimators for a population parameter
- What are characteristics of good estimators?

Sampling distribution

- Statistics are RV's. Why?

- Distribution depends on sample size

$$f_N\left(\hat{\theta}\right)$$



Properties of estimators

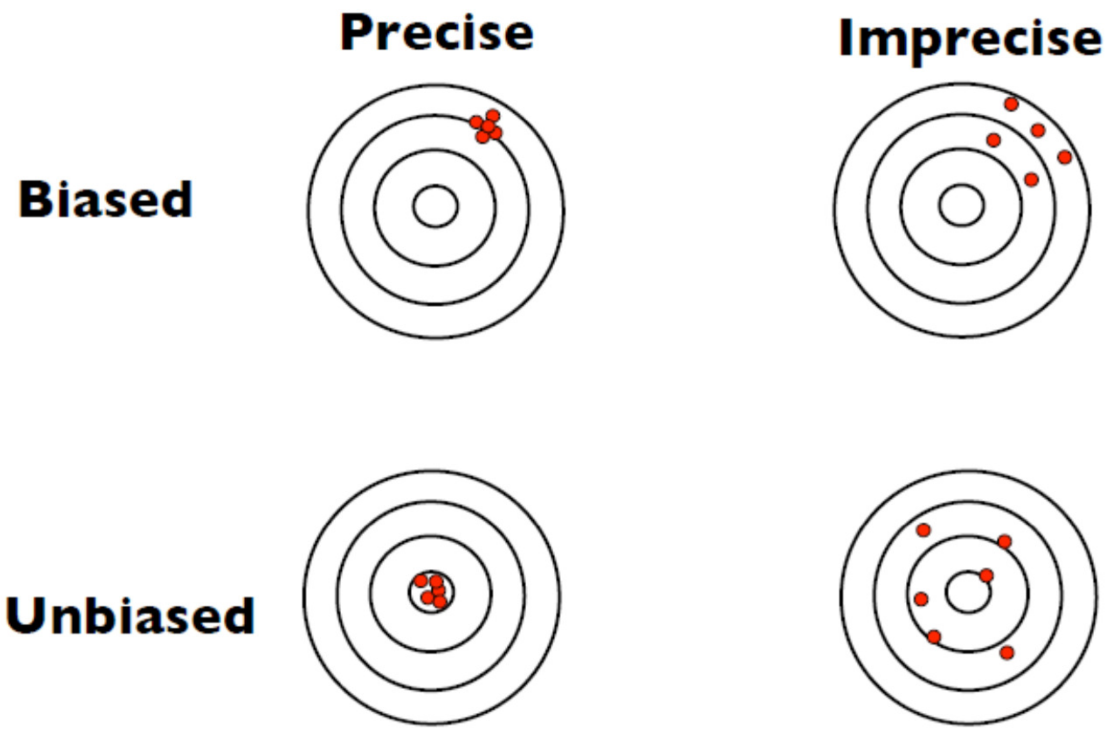
- Unbiasedness $$E\left(\hat{\theta}\right)=\theta$$
- Efficiency $$Var\left(\hat{\theta}_1\right)<Var\left(\hat{\theta}_2\right)$$
  - Unbiased
  - No other unbiased estimator has smaller variance

- *Precise*: Sampling distribution of $\hat{\theta}$ should have a small standard error

  - Cramer-Rao lower bound $$Var\left(\hat{\theta}\right)\geq\left[-E\left(\frac{\partial^2 L}{\partial\theta^2}\right)\right]$$

- Asymptotic properties $$\lim_{N_1\to\infty} P\left(\left|\hat{\theta}_1-\theta_1\right|<\delta\right)=1 \quad \delta>0$$
  - Asymptotic unbiasedness
  - Consistency

Bias Versus Precision



Asymptomatic Unbiaseness

$$\lim_{N\to\infty} E[\hat{\mu}_N] = \mu$$

$$E\left[\hat{\mu}_N^1\right] = E\left[\frac{1}{N}\sum_{n=1}^{N} z_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[z_n] = \frac{1}{N}\cdot N\mu = \mu$$

$$E\left[\hat{\mu}_N^4\right] = E\left[\frac{1}{N-1}\sum_{n=1}^{N} z_n\right] = \frac{1}{N-1}\sum_{n=1}^{N} E[z_n] =$$

$$= \frac{1}{N-1}\cdot N\cdot\mu = \left(\frac{N}{N-1}\right)\mu$$

$$\lim_{N\to\infty}\left(\frac{N}{N-1}\right)\mu = \mu$$

Consistency

- Consistency
  - *A large number of consistent estimators will often be available, some of which may be very biased or inefficient*

As the sample size increases $\hat{\theta}$ gets closer to θ

$\hat{\theta}_N$ is a consistent estimator for θ if

$$\lim_{N\to\infty}\left[Pr\left(\theta - \mathbf{q} \le \hat{\theta}_N \le \theta + \mathbf{q}\right)\right] = 1$$

where q is small constant

$$\operatorname*{Plim}_{N\to\infty} \hat{\theta}_N = \theta$$

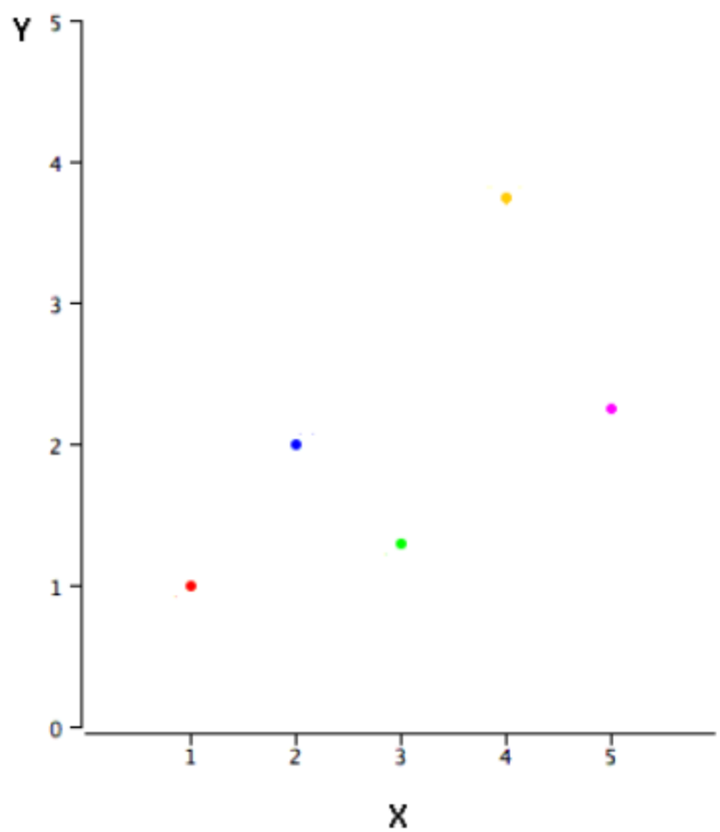
- Asymptotically normal
  - *Estimators are asymptotically normal if their distribution (which may be unknown) converge to normal multivariate one as n get larger and larger*

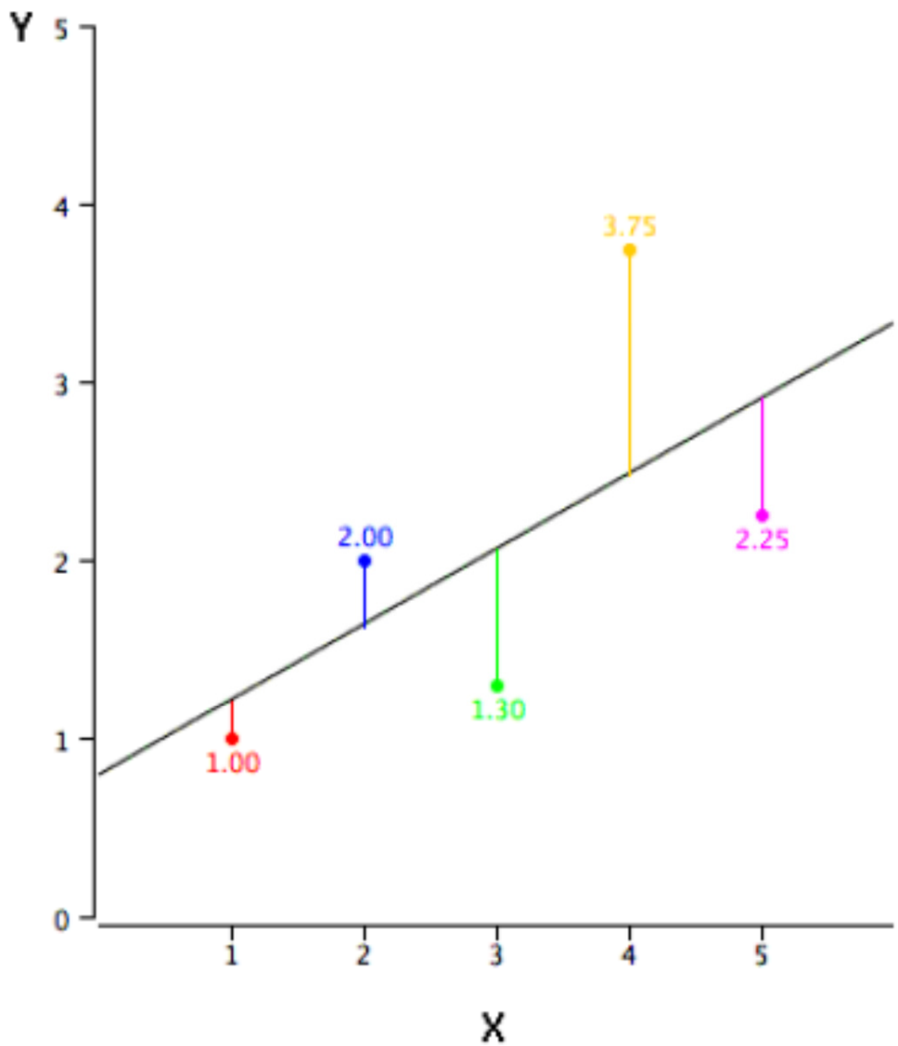
Estimation methods

- Least squares
- Maximum likelihood
- Method of moments

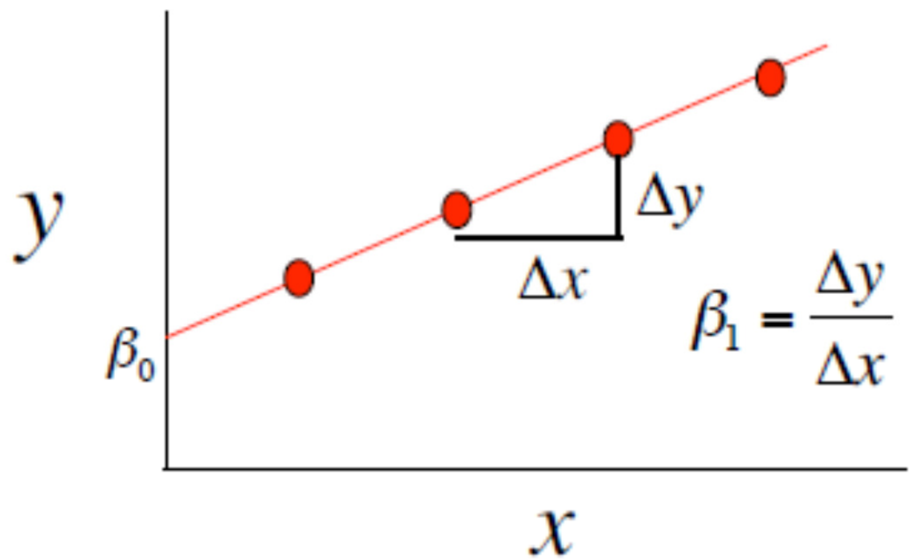
Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25




Y
5
4
3
2
1
0
3.75
2.00
2.25
1.30
1.00
1
2
3
4
5
X

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

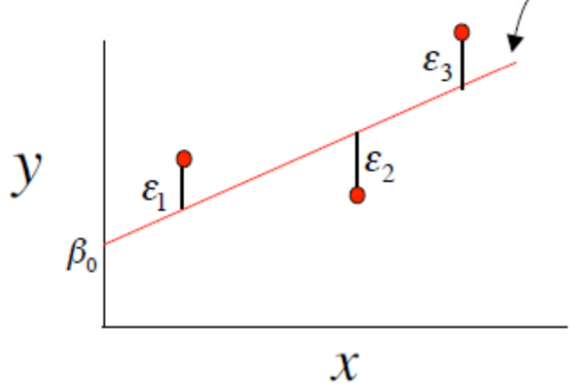


A Linear Probabilistic Model

- Definition: There exists parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is related to x through the model equation

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

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



True Regression Line

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

Implications

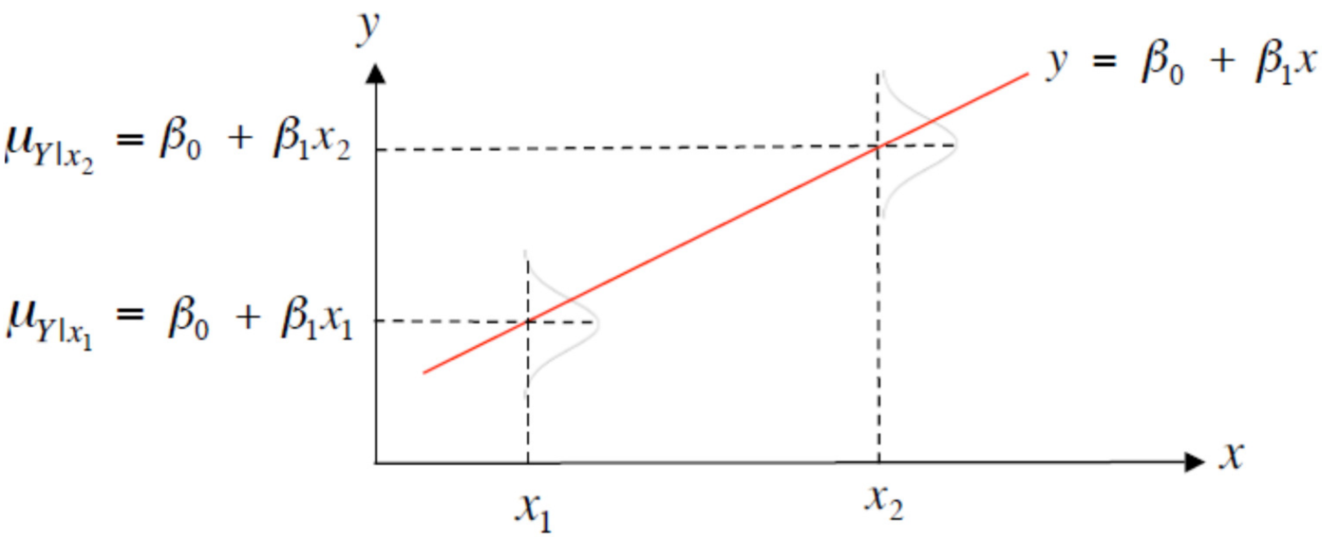
- The ***expected*** value of Y is a linear function of X, but for fixed x, the variable Y differs from its expected value by a random amount

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

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

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

Graphical Interpretation



- For example, if x = height and y = weight then $\mu_{Y|\mathbf{x}=60}$ 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
$y = 7.5 + 0.5x$ 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?

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

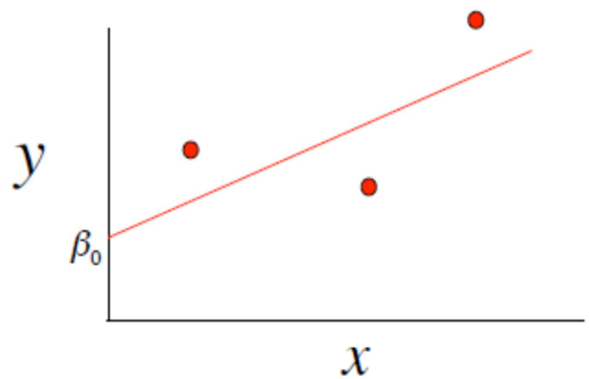
- Q3: If x = 20 what is P(Y > 22)?

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

Estimating Model Parameters

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

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



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

Least Squares Procedure

- **The Least-squares procedure obtains estimates of the linear equation coefficients β_0 and β_1, in the model**

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- **by minimizing the sum of the squared residuals or errors (e_i)**

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- **This results in a procedure stated as**

$$SSE = \sum e_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Choose β_0 and β_1 so that the quantity is minimized.

Least Square

$$L = \sum_{i=1}^{n} \Big[\underset{\text{obs}}{y_i} - \underbrace{(a + bx_i)}_{E(x)} \Big]^2$$

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{n} (-2) \big[y_i - (a + bx_i) \big]$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} (-2) x_i \big[y_i - (a + bx_i) \big]$$

$$na + \left(\sum_{i=1}^{n} x_i \right) b = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i \right) a + \left(\sum_{i=1}^{n} x_i^2 \right) b = \sum_{i=1}^{n} x_i y_i$$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n \left(\sum_{i=1}^{n} x_i^2 \right) - \left(\sum_{i=1}^{n} x_i \right)^2}$$

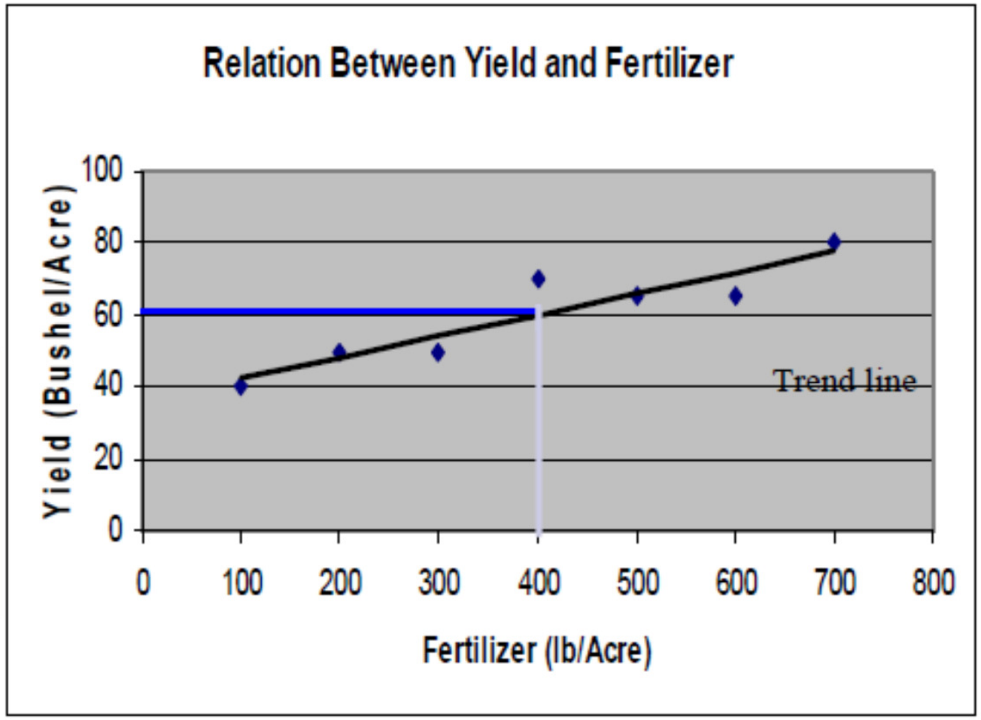
$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- Note that the regression line always goes through the mean X, Y.
- Think of this regression line as the expected value of Y for a given value of X.

That is, for any value of the independent variable there is a single most likely value for the dependent variable



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 σ^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

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

Expression = Baseline + Age + Tissue + Sex + Error

- Partial Regression Coefficients: $\beta_i \equiv$ effect on the dependent variable when increasing the i^{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:
- 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):

$$\mathrm{H}_0 \ : \ \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$\mathrm{H}_\mathrm{A} \ : \ \text{At least one } \beta_1 \neq 0$$

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

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

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)

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	$SST = \sum(y_i - \bar{y})^2$		

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

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

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

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

Hypothesis Tests of Individual Regression Coefficients

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

$$\mathrm{H}_0 \ : \ \hat{\beta}_i = 0$$

$$\mathrm{H}_\mathrm{A} \ : \ \hat{\beta}_i \neq 0$$

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

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

- Confidence Intervals are equally easy to obtain:

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

Hypothesis testing

H_0 null hypothesis to be tested

H_1 alternative hypothesis

significance level

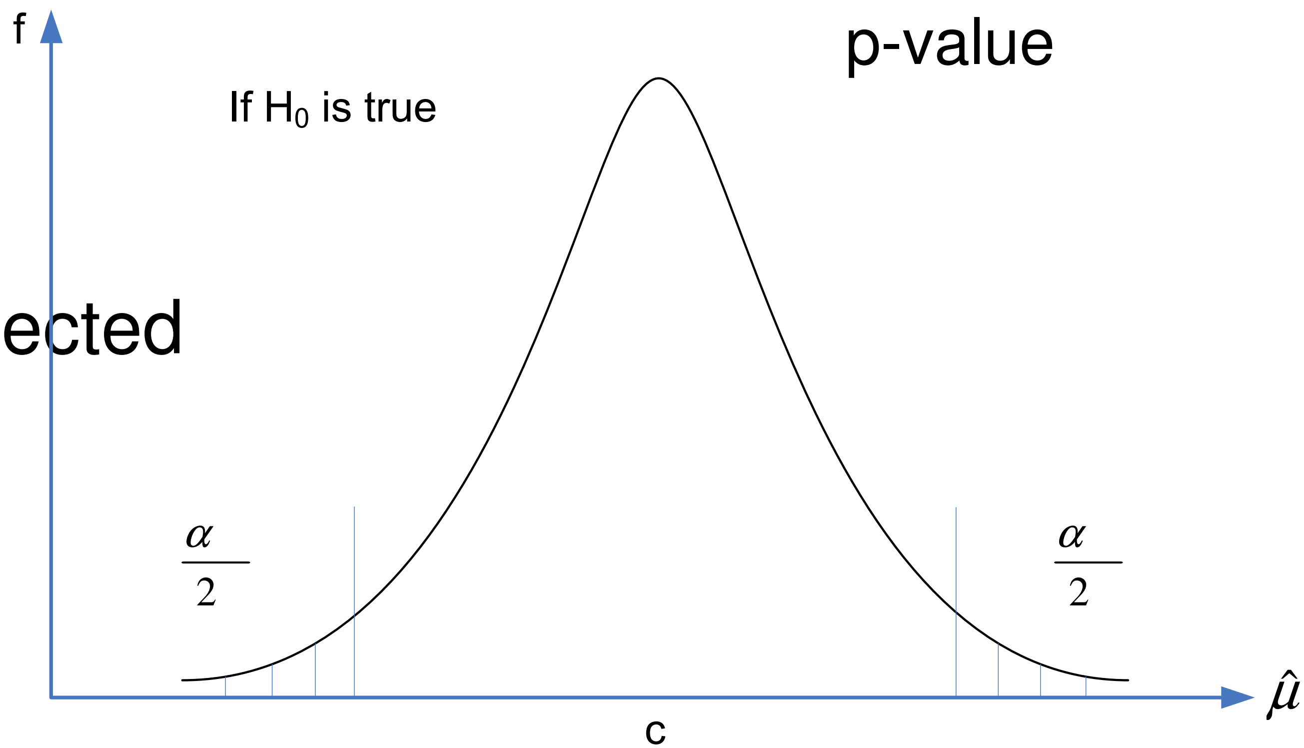
p-value

$H_0 : \mu = c$

$H_1 : \mu \neq c$

Null is rejected

or not



One and two sided tests

- Two sided

$$H_0 : \mu = c$$
$$H_1 : \mu \neq c$$

- One sided

$$H_0 : \mu = c$$
$$H_1 : \mu < c$$

Test procedure

- Define test statistics
- Define critical value to reject null
  - Distribution of test statistic
  - Significance level
  - Probability that “true” test statistics is zero

Checking Assumptions

- Critically important to examine data and check assumptions underlying the regression model
  - Outliers
  - Normality
  - Constant variance
  - Independence among residuals

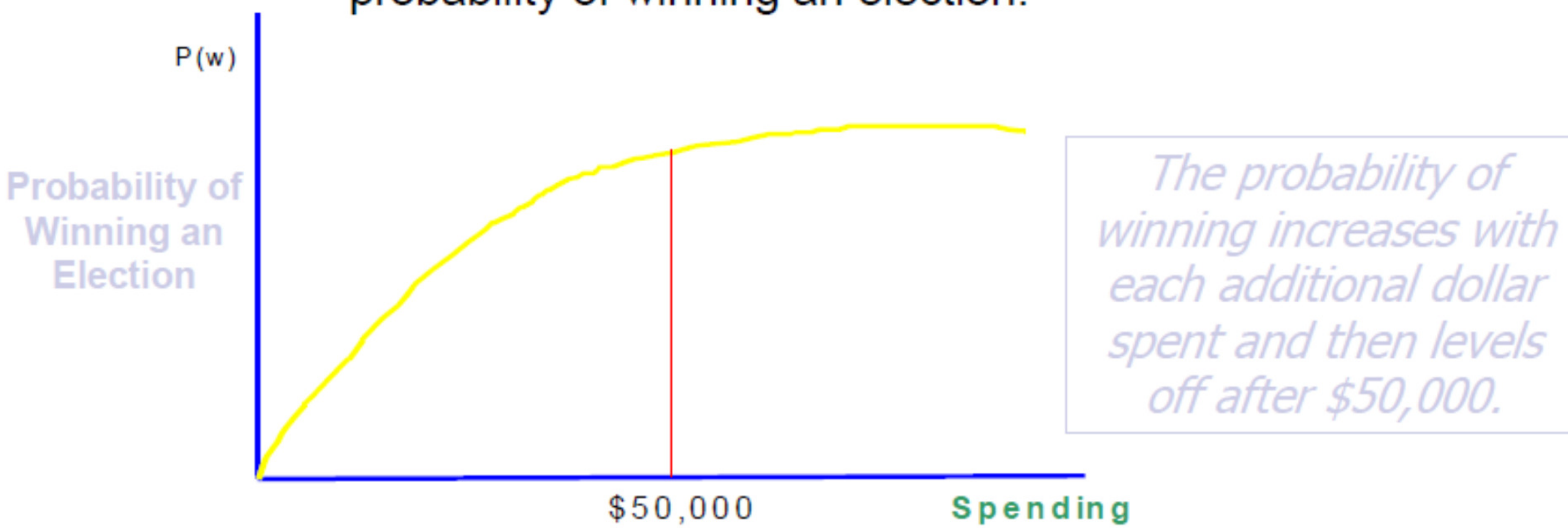
- Standard diagnostic plots include:
  - scatter plots of y versus x_i (outliers)
  - qq plot of residuals (normality)
  - residuals versus fitted values (independence, constant variance)
  - residuals versus x_i (outliers, constant variance)

Assumptions of Linear Regression

- A linear regression model assumes:
  - Linearity:
    - $\mu \{Y|X\} = \beta_0 + \beta_1 X$
  - Constant Variance:
    - $var\{Y|X\} = \sigma^2$
  - Normality
    - Dist. of Y's at any X is normal
  - Independence
    - Given X_i's, the Y_i's are independent

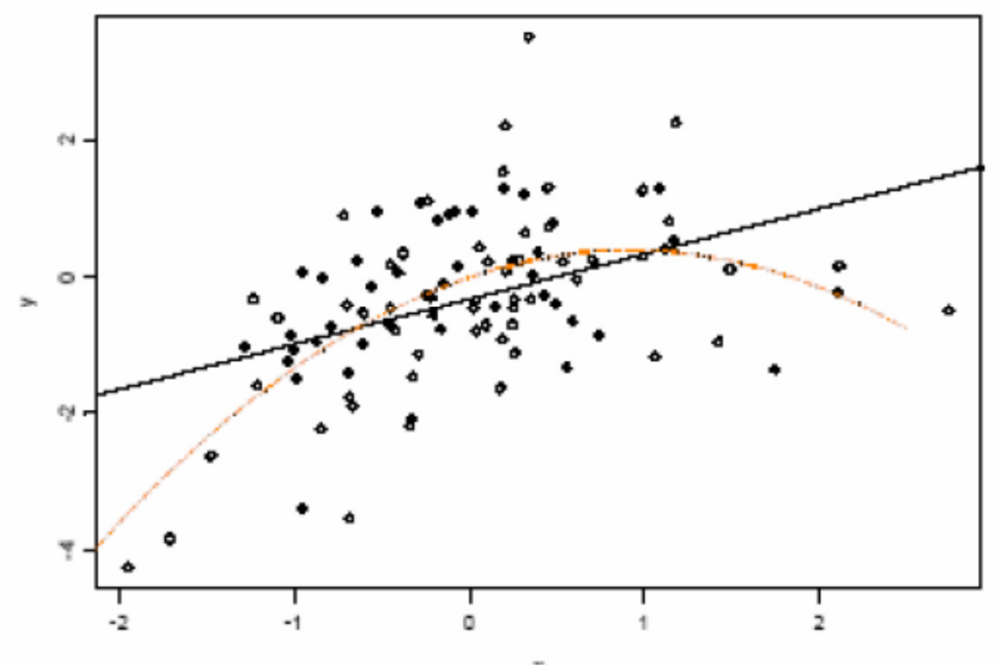
Examples of Violations

- Non-Linearity
  - The true relation between the independent and dependent variables may not be linear.
    - For example, consider campaign fundraising and the probability of winning an election.



Consequences of violation of linearity

- If "linearity" is violated, misleading conclusions may occur (however, the degree of the problem depends on the degree of non-linearity)

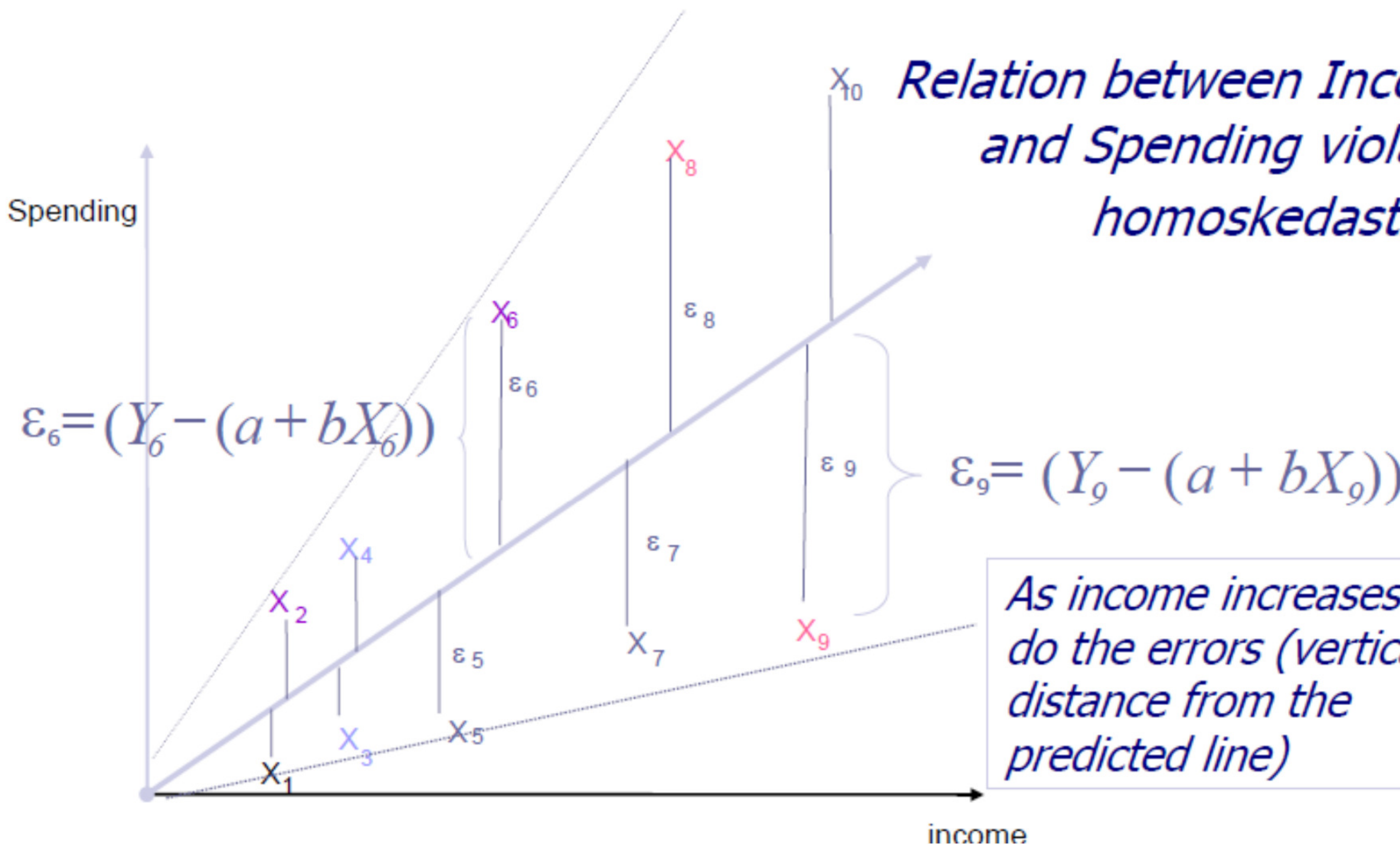


•Linear in parameters vs. linear variables

Examples of Violations: Constant Variance

- Constant Variance or Homoskedasticity
  - The Homoskedasticity assumption implies that, on average, we do *not expect* to get larger errors in some cases than in others.
    - Of course, due to the luck of the draw, some errors will turn out to be larger then others.
    - But homoskedasticity is violated only when this happens in a predictable manner.
  - Example: income and spending on certain goods.
    - People with higher incomes have more choices about what to buy.
    - We would expect that there consumption of certain goods is more variable than for families with lower incomes.

Violation of constant variance



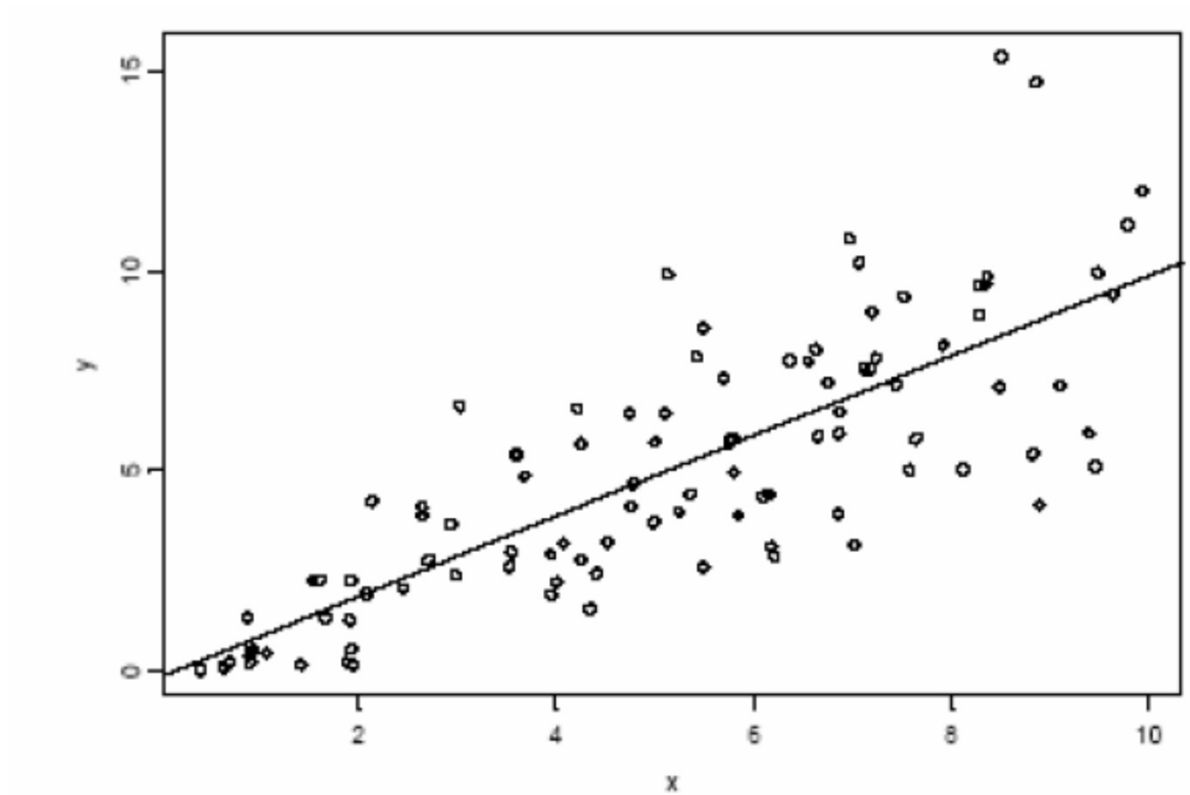
Relation between Income and Spending violates homoskedasticity

As income increases so do the errors (vertical distance from the predicted line)

Consequences of non-constant variance

- If "constant variance" is violated, LS estimates are still unbiased but SEs, tests, Confidence Intervals, and Prediction Intervals are incorrect

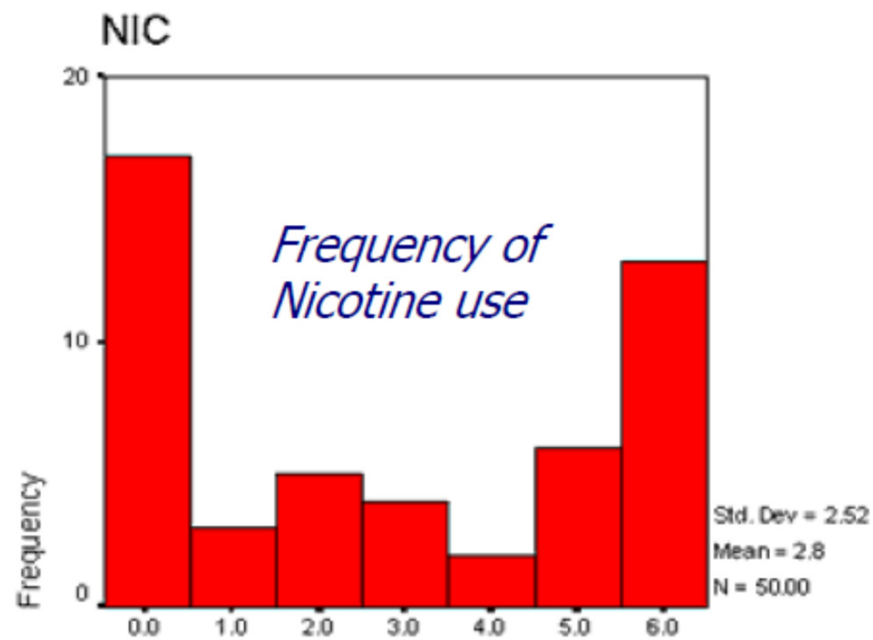
- However, the degree depends…



Violation of Normality

- Non-Normality

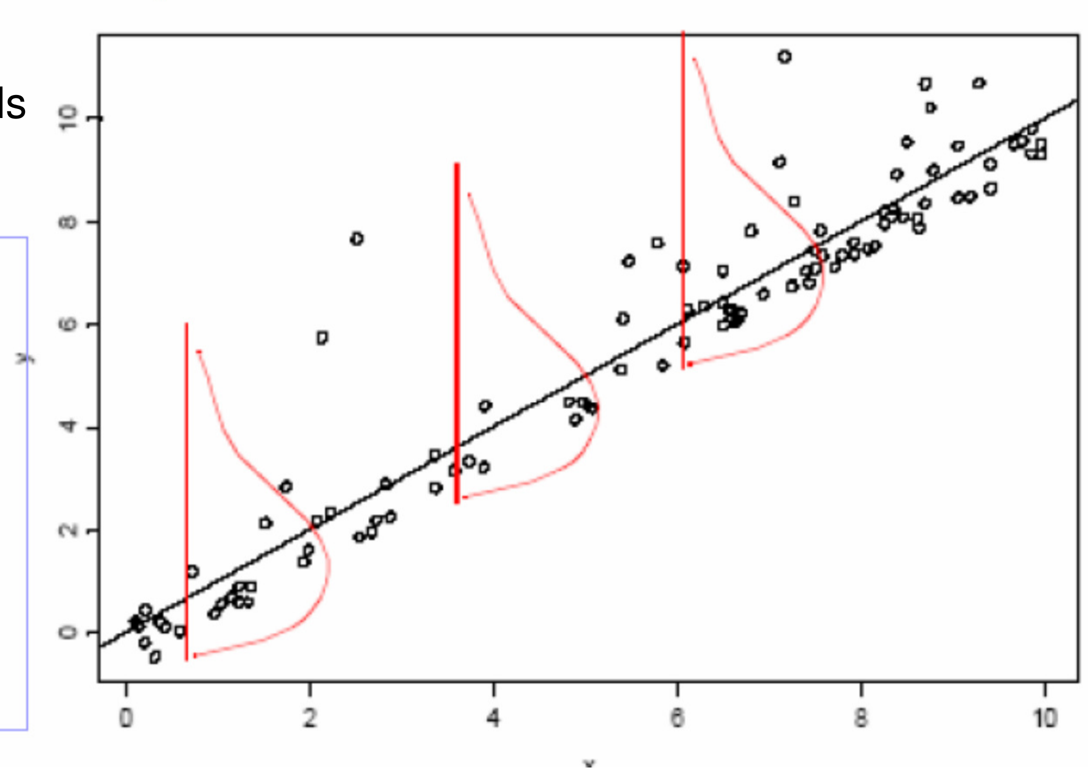
Nicotine use is characterized by a large number of people not smoking at all and another large number of people who smoke every day.



An example of a bimodal distribution

Consequence of non-Normality

- If "normality" is violated,
  - LS estimates are still unbiased
  - tests and CIs are quite robust
  - PIs are not
    - •Prediction intervals

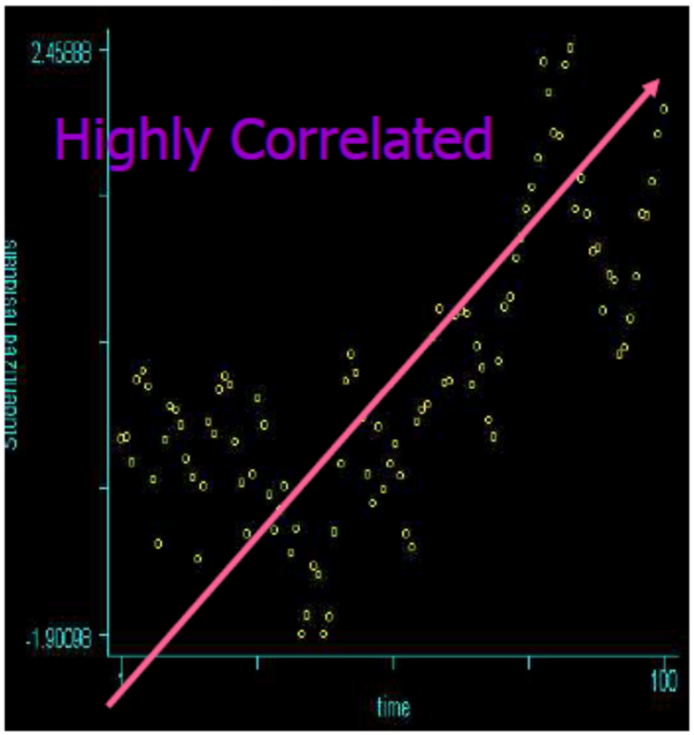


Of all the assumptions, this is the one that we need to be least worried about violating.

Why?

Violation of Non-independence

Residuals of GNP and Consumption over Time

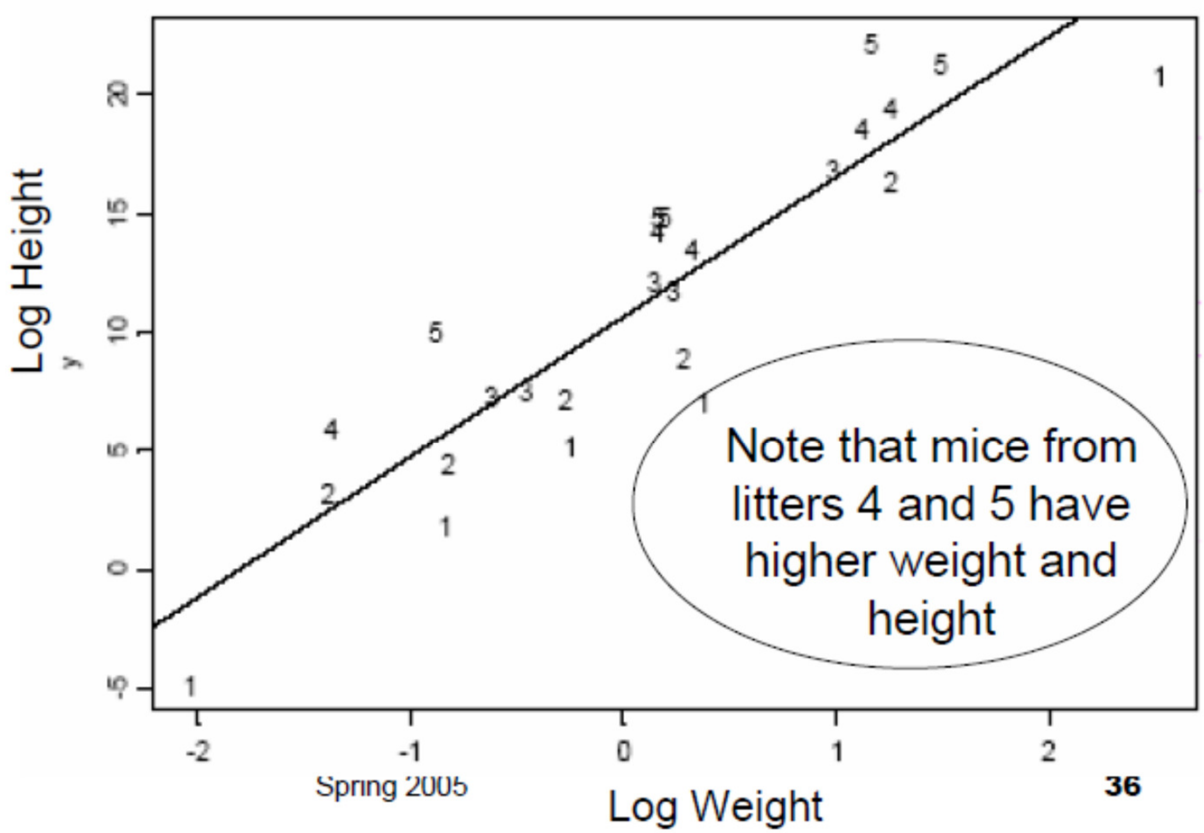


- Non-Independence
  - The independence assumption means that errors terms of two variables will not necessarily influence one another.
    - Technically, the RESIDUALS or error terms are uncorrelated.
  - The most common violation occurs with data that are collected over time or time series analysis.
    - Example: high tariff rates in one period are often associated with very high tariff rates in the next period.
    - Example: Nominal GNP and Consumption

Consequence of non-independence

- If "independence" is violated:
  - LS estimates are still unbiased
  - everything else can be misleading

Plotting code is litter (5 mice from each of 5 litters)



Robustness of least squares

- The "constant variance" assumption is important.
- Normality is not too important for confidence intervals and p-values, but is important for prediction intervals.
- Long-tailed distributions and/or outliers can heavily influence the results.

- Check:
- Scatterplot of Y vs. X
- Scatterplot of residuals vs. fitted values
- Look for curvature, non-constant variance and outlier