Normality, Hypothesis Tests, and Inference

D. Alex Hughes

October 21, 2014

1. Power
2. Parameters, Estimators, and Propertites
3. Building Blocks
   - Scatterplots, carefully
   - Covariance & Correlation
4. Regression
   - Impression of Models
   - Interpretation of OLS
5. Bringing it Back
6. Linear Algebra Representation

Power

We have constructed *all* of our tests in terms of α. Remind me what we're accomplishing here?

But, this is only *one* of the ways that our test could (with probability) be wrong.

Definition

The **Power**, frequently denoted $(1 - \beta)$ of a test is the probability that we correctly reject the null, *given the (unknown) true state of the world should reject the null.*

Canonically, this is shown with the following table:

Power

	Truth:	
Action	H_0	H_A
DN Reject	Correct	Type II Error "False Negative" β
Reject	Type I Error "False Positive" α	Correct

Power

Draw on Board...

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Properties of Estimators

Many ways to skin a cat (...or to measure a parameter).

- **Central Moment**:
  - mean
  - median
  - mean(max,min)
- **Second Central Moment**
  - sd
  - median of $x_i - x_j$
  - sd(any two values)

Which is the "best" estimator?

Properties of Estimators

What do we want in an estimate? **1. Unbiasedness:**

- On average, it is right!
- $E(\hat{\theta} - \theta) = 0$
- max(x) versus mean(x)

This is for fixed n. There is also asymptotic unbiasedness, as n becomes large.

Properties of Estimators

2. Efficiency

- How close do we think we are?
- Mean-squared error: $MSE = E[(\hat{\theta} - \theta)^2]$
- More efficient = smaller MSE

Properties of Estimators

3. Consistency: As sample size gets bigger and bigger, the estimate approaches the true value.

- With an infinite sample size, we know everything.
- Limit

Properties of Estimators

4. Sufficiency: The estimate contain all the information about the parameter that is in the data.

- n or sample $\frac{n}{2}$?
- Mean versus median
- Learn more: Bickel and Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*

Compare Estimators

Consider Four Estimators of μ

- Mean
- Median
- Average of max and min
- Maximum

Compare Estimators When we actually use these estimators, can we talk intelligently about how *"good"* they are?

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Conditional Distribution

Regression Analysis traces the distribution of some response variable we're interested in as a function of one or more explanatory variables.

- **Response Variable**: The quantity we want to know more about, the policy outcome, election result, voting pattern, etc. Typically denoted as Y. Sometimes called the LHS variables.
- **Explanatory Variables**: The quantity for which changes cause – or more weakly, are associated with – changes in the response variable. Typically denoted $(X_1, \ldots, X_k)$.

Definition

The **conditional distribution** of Y on x, written $P[y|x_1, \ldots, x_n]$, represents the probability density of observing the specific value y *conditional* on a set of specific values for the Xs.

Conditional Distribution

Example

Wages and Earnings. Is there some relationship between years of education and wages demanded in the workforce?

This is plausible, and is *the* classic example in econ. See code here.

What might we have learned?

- No information about Xs, or no information about relationships between Xs and Ys, then what is our best guess for some y_i?
- Information on Xs let us make *different*, and better, predictions about any y_i.
- What if there were no stochastic component?
- ... and we had measurement on all Xs?

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Scatterplots

Definition

A **scatterplot** is a visual display of data in the plane defined by the variable.

- Show the relationship between two (or more variables)
- Frequently the RHS (x) variable displayed on the x-axis; LHS (y) variable on the y-axis
- Each point represents a single observation
- The x and y axis show the corresponding values of x and y for each variable. What can we learn?
  - Relationship? Type? Direction? Strength?
  - Anything weird?

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Covariance

If we have two variables, how does one change *with respect to the other*?

- If X increases, what happens to Y?
- What could be a null hypothesis here?

Definition

Let X and Y be random variables with means μ_X and μ_y. Then, the **covariance** of X and Y, written $COV(X, Y)$, is given by:

$$COV(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

What is the relationship between covariance and variance?

Theorem

For an random variables X and Y,

$$COV(X, Y) = E(XY) - \mu_X \mu_Y$$

Covariance Example

Example

Suppose two variables jointly distributed according to the pdf

$$f_{X,Y}(x, y) = 8xy, 0 \leq x \leq 1.$$

What is their covariance?

Covariance Example

Answer

The marginal pdf's are

$$f_X(x) = \int_{y=0}^{x} 8xy\,dy = 4x^3, 0 \leq x \leq 1$$

and

$$f_Y(y) = \int_{x=y}^{1} 8xy\,dx = 4y - 4y^3, 0 \leq y \leq 1.$$

Therefore,

$$\mu_X = \int_0^1 x * 4x^3\,dx = \frac{4}{5}$$

, and

$$\mu_Y = \int_0^1 y * (4y - 4y^3)\,dy = \frac{8}{15}...$$

Covariance Example

Answer

... Then, since

$$
\begin{aligned}
COV(X,Y) &= E(XY) - \mu_X \mu_Y \\
&= \iint_{x=0,y=0}^{1,x} xy * 8xy \, dy \, dx - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) \\
&= \int_0^1 \frac{8x^5}{3} \, dx - \frac{32}{75} \\
&= \frac{8}{450}.
\end{aligned}
$$

Is the covariance useful?

Remember previously when I forgot to tell you that X and Y were independent? That was problematic...

Theorem

If any X and Y are independent,

$$COV(X, Y) = 0$$

Proof.

If X and Y are independent, $E[XY] = E[X] * E[Y] = \mu_X \mu_Y$. □

Is the covariance useful?

Theorem

Let $W_1, W_2, \ldots, W_n$ be any set of random variables. Let $S = W_1 + W_2 + \ldots + W_n$. Then,

$$VAR(S) = \sum_{i=1}^{n} VAR(W_i) + 2\sum_{j \neq k} COV(W_j W_k)$$

Proof.

$$\begin{aligned} Var(S) &= E[S^2] + [E[S]]^2 \\ &= E\left[\left(\sum_{i=1}^{n} W_i\right)^2\right] - \left[\sum_{i=1}^{n} E[W_i]\right]^2 \end{aligned}$$

Is the covariance useful?

Proof.

$$
\begin{aligned}
&= E\left[\sum W_i^2 + 2\sum W_j W_k\right] - \left[\sum [E[W_i]]^2 + 2\sum E[W_j]E[W_k]\right] \\
&= \sum \left[E[W_i^2 - [E[W_i]]^2\right] + 2\sum [E[W_j W_k] - E[W_j]E[W_k]] \\
&= \sum Var(W_i) + 2\sum Cov(W_j, W_k)
\end{aligned}
$$

□

Correlation Coefficient

What are the units of covariance?

- What are the units of variance?
- So, the covariance are in units of *both* variables

So, just like we sometimes scale the variance in SD, we can scale the covariance into **correlation** by dividing the sample standard deviation of the two vectors – $\sigma_X \sigma_Y$.

Definition

Let X and Y be two random variables. The **correlation coefficient** of X and Y, written $\rho(X, Y)$, is given by

$$\rho(X, Y) = \frac{COV(X, Y)}{\sigma_X \sigma_Y}.$$

Correlation Coefficient

Definition

This can also be represented as $COV(X^*, Y^*)$, where

$$X^* = \frac{X - \mu_X}{\sigma_X}$$
$$Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

Note that the correlation relies on the asymptotic properties of covariance. Since we don't typically have *infinite* data, we might want a sample correlation coefficient.

Sample Correlation Coefficient

Think of the correlation coefficient in terms of theoretical moments:

$$\rho(X, Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

Then, it wouldn't be challenging to think of these parts in terms of their conjugate *sample* counterparts.

- $E[X] = \bar{X}$ and $E[Y] = \bar{Y}$
- $E[XY] = \frac{1}{n}\sum X_i Y_i$
- $Var(X) = \frac{1}{n}\sum(X_i - \bar{X})^2$ and the same for Y

Definition

The **sample correlation coefficient** is

$$\hat{\rho} = \frac{\frac{1}{n}\sum X_i Y_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}}$$

Sample Correlation Coefficient

This value $\hat{\rho}$ is called the **Pearson product-moment correlation coefficient**, or the Pearson correlation coefficient.

But, what does it do for us, Alex?

- A single number that measures the strength and direction of a *straight line* relationship
- Bounded from $-1 \leq \rho \leq +1$
- Susceptible to outliers
- Correlation $\neq$ causation

See code

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Linear Regression: A Special Case

At the top, we set-up the notion of the *conditional expectation* and *regression analysis* very generally. We now turn to a special case of this, the **simple linear model.**

Four Assumptions:

1. $f_{Y|x}(y)$ is a normal pdf for all x
2. The standard deviation, σ associated with $f_{Y|x}(y)$ is known, and the same for all the $x's$
3. The means of all the conditional Y distributions are linear-additive:

$$y = E(Y|x) = \beta_0 + \beta_1 x_1$$

4. All of the conditional distributions represent independent random variables

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

The line

We're giving a lot of flexibility up when we make *parametric* assumptions about the world. But:

- Lots and lots of times, we want to talk about the world in terms of *comparative statics*
- The line formulate works pretty well for this, and captures much of the thrust of our theory

$$y = \beta_0 + \beta_1 x_1$$

- But, since we don't believe in a deterministic world, we "fudge" it a bit and add an error (ϵ) or deviation from the theory line.
- β_1 gives us the change in our RHS variable as our LHS variable changes – the *slope* or *rise-over-run*
- β_0 give the value of the RHS variable when our other LHS variables are zero

Picking a line from a scatterplot

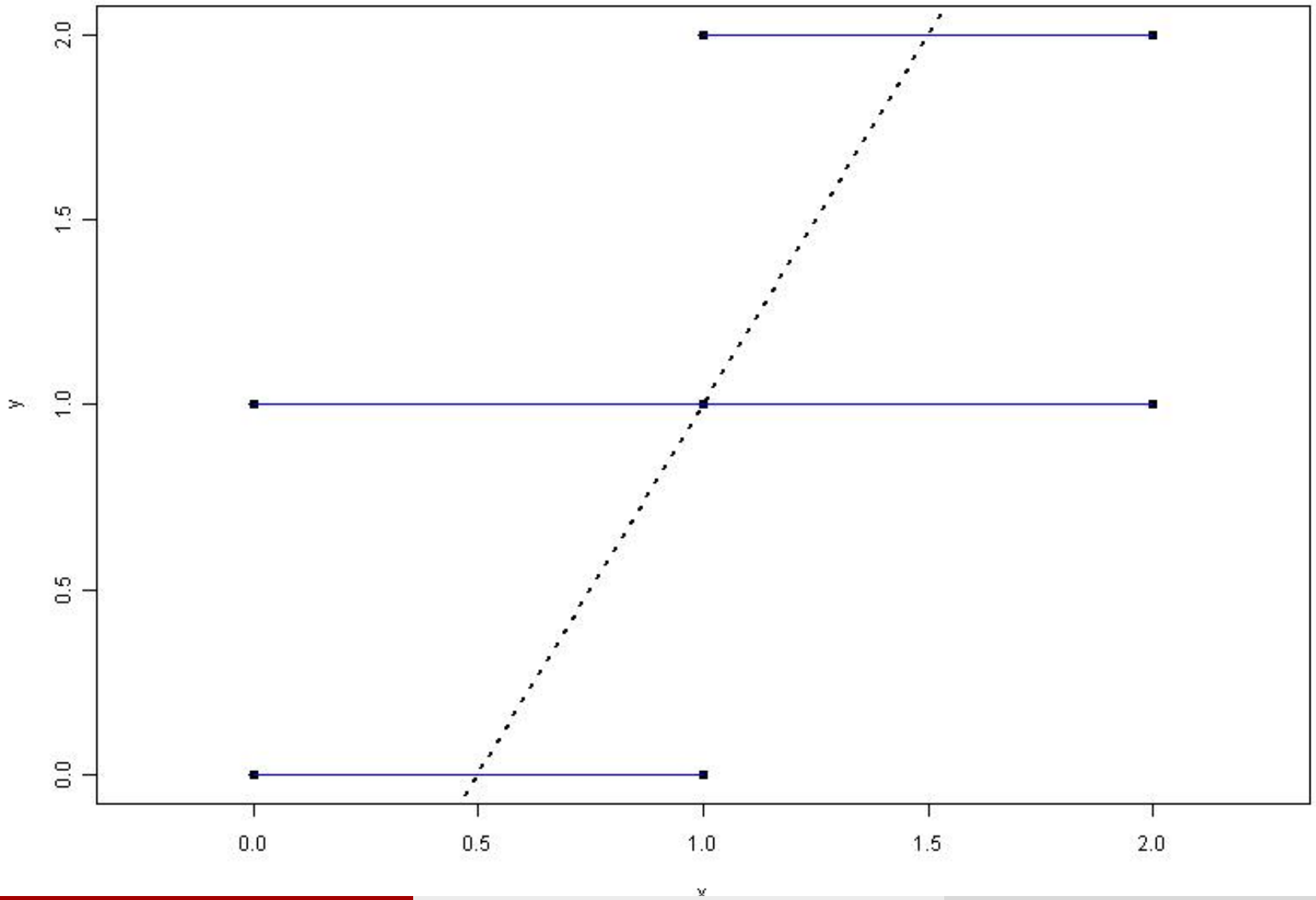
It turns out, when there isn't much data, humans are pretty good at this

- Draw a line
- Slope is "rise-over-run"
- Intercept is height when $x = 0$

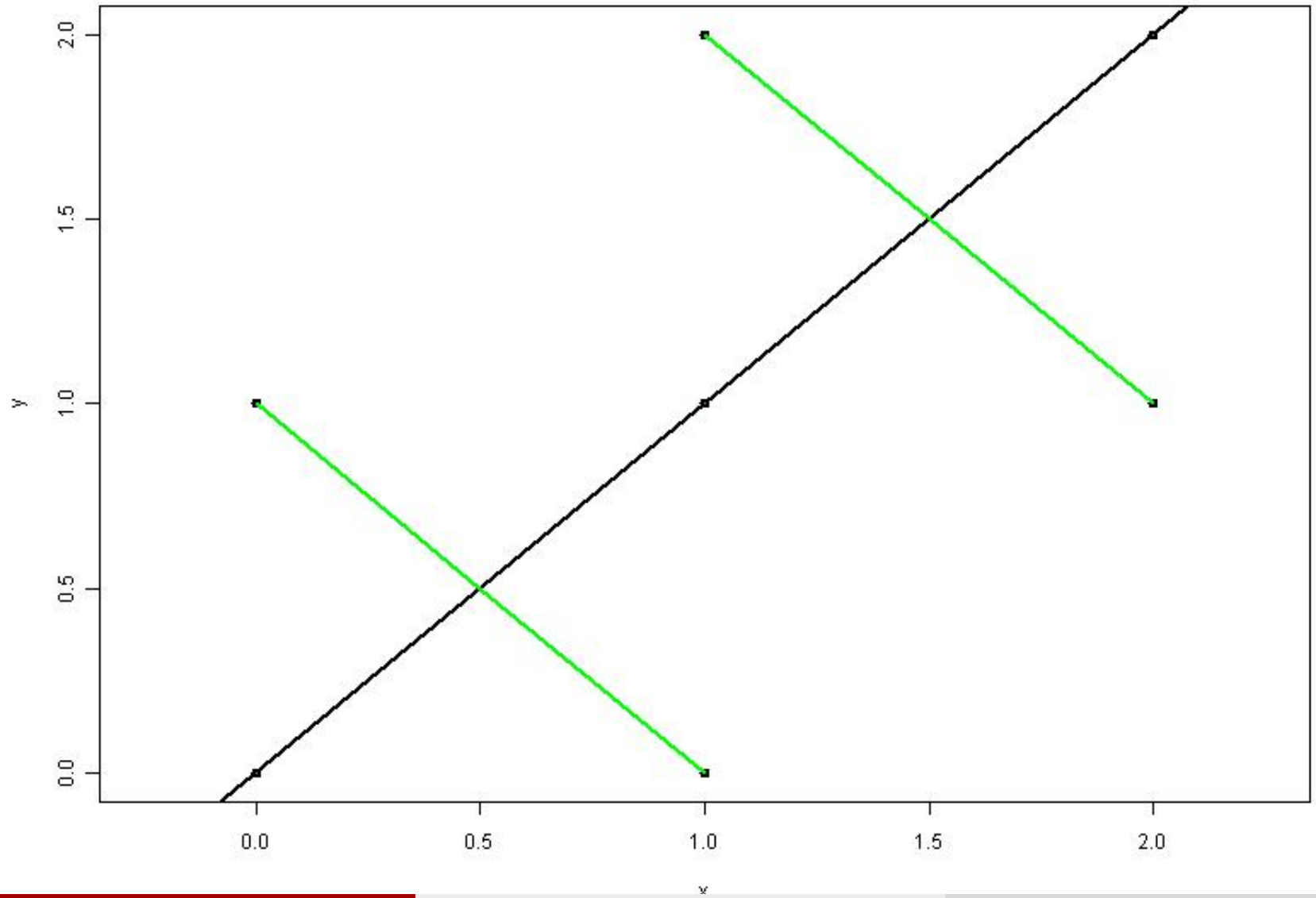
What criteria did you use to pick your line?

- Delegation to Committee?
- Casting Lots?
- Optimization Criteria?
  - Mean distance from the line
  - Absolute value of the distance from the line
  - Squared value of the distance from the line
- Distance
  - Perpindicular
  - Horizontal
  - Vertical

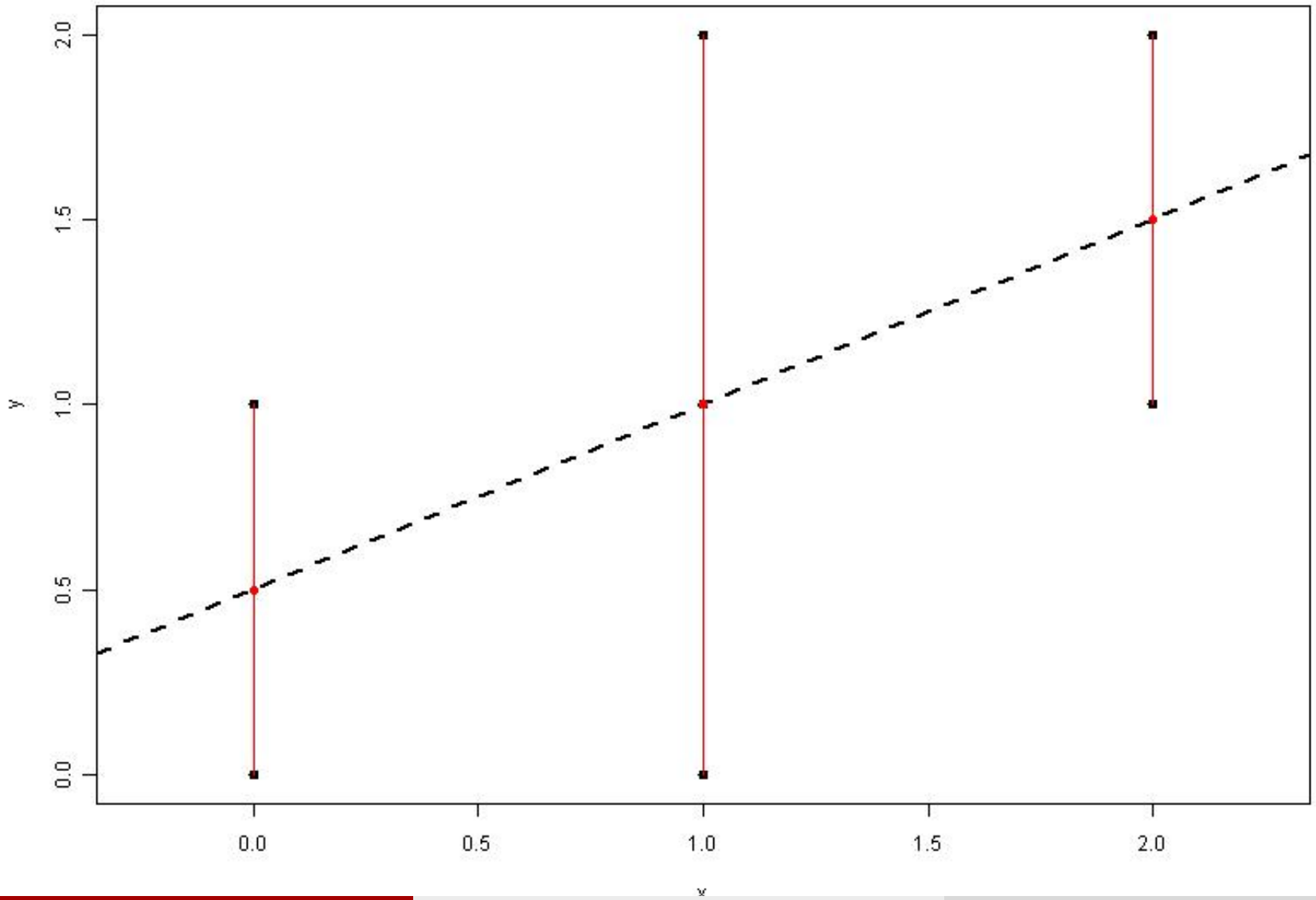
Horizontal Distance



Absolute (Perpindicular) Distance



Vertical Distance



Best Method: Sum of Squared Residuals

Minimize *sum* of the squared vertical deviation from the fit regression line

$$\begin{aligned} Y &= A + BX_i + E_i \\ &= \hat{Y}_i + E_i \end{aligned}$$

Lightly re-writing, we can consider any residual E_i to be

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$$

- Why not just one of the residuals?
- Why square the residuals?

Example

Consider the following function

$$S(A, B) = \sum E_i^2 = \sum (Y_i - A - BX_i)^2$$

Every $(A, B) \rightarrow$ a unique sum of squared residuals. See code.

The Normal Equations

From S(A,B), we can set up a tedious, but straightforward optimization problem.

$$\frac{\partial S(A,B)}{\partial A} = \sum (2)(Y_i - A - BX_i)(-1) \tag{1}$$

$$\frac{\partial S(A,B)}{\partial B} = \sum (2)(Y_i - A - BX_i)(-X_i) \tag{2}$$

Set to be zero and solve (which will either maximize or minimize the function) the system of two equations with two unknowns.

Demonstrate on board.

Least Squares Regression

This yields the least squares regression line, and estimated coefficients

$$\hat{\alpha} = \bar{Y} - \beta\bar{X}$$

$$\hat{\beta} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

But, "Did our regression get an A-...

"because it needs at least a A- to get into Stanford Law."

- How closely did our model fit the scatterplot
- Lots of ways
  - *Standard Error of the Regression*, also called the *Residual Standard Error*
  - *Explained Variance*

Definition

The **variance of the residuals**, S_E^2, is just

$$\frac{\sum E_i^2}{n-2}$$

And, under a familiar transformation, the **Residual Standard Error** is

$$S_E = \hat{\sigma} = \sqrt{\frac{\sum E_i^2}{n-2}}$$

Interpreting Residual Standard Error

- Measured in units of response variable
- Tells us that, *on average* the least-squares regression line we will be about S_E away from the observed data.
- By assumption (and required for OLS to be an appropriate estimator), E_I are distributed according to the Gaussian distribution.

Another Way to Measure Fit

1. Estimate, using the same technique, a "null" model where we restrict the slope ($\beta = 0$). That is, estimate the equation

$$Y_i = A' + E_i'$$

(It should be clear that under this restriction $A = \bar{Y}$)

2. Square and sum these residuals and call them the **Total Sum of Squares** (TSS) because these are the *total* distance a very naïve estimator is from the data.

$$TSS = E_i'^2 = \sum (Y_i - \bar{Y})^2$$

3. Compare to the **Residual Sum of Squares** (RSS), which are the residuals left after fitting a regression.

$$RSS = E_i^2 = \sum (Y_i - \hat{Y})^2$$

Another way to Measure Fit

In particular, compare the percent improvement in fit for estimating $\hat{\beta}$:

$$r^2 = \frac{TSS - RSS}{TSS}$$

Definition

1. The **Regression Sum of Squares** is the difference between the *Total Sum of Squares* under the null model and the *Residual Sum of Squares* in the fit model.

$$RegSS \equiv TSS - RSS$$

2. And so, $r^2 = \frac{TSS - RSS}{TSS} \equiv \frac{RegSS}{TSS}$

Then, R^2 is the percentage of varaince in Y that is explained by $X\beta$

Let's do one by hand

Name	Education	Income
April	12	30
Bob	12	25
Christie	16	50
Dan	16	45

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Interpretation of OLS

1. A one-unit change in X produces a $\hat{\beta}$ unit increase in Y
2. When X is zero, $E[Y]$ is $\hat{\alpha}$
3. The typical data-point is about $\hat{\sigma}$ away from the regression line.

But, if we've got this formula ($\hat{Y} = \hat{\alpha} + \hat{\beta}X$), why don't we use it?

- In particular, we can predict an expected $\hat{Y}$ that is an outcome for a realization of X that we might not have observed.

Empirical Tests of our Theories

- Inflation = 2% + 1% * Deficit
- P(War) = 3.25% - 3% * Democracy
- Income = \$11k + \$5k * Education
- Democracy Therm = 8 + .25 * Income
- Ideological Variance = 0.25 + 0.25 * Poor

For each of these, we can clearly state

1. Comparative static for Δ X-var
2. Predicted Value for some fixed X-var

Let's do one by hand (suite)

From our previous example that we calculated by hand

1. Image that we identify a new person via the same simple random sample that we set up for the others. She has 14 years of education – started a four-year degree before the *dot-com* era, and dropped out to start Gecko insurance.
2. When we spoke with April, she was part way though college. She will finish in 6 mores years. How much more would you expect she will earn?

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Conditional Expectation → Regression → Conditional Expectation

Tradeoff between flexibility and efficiency

- If assumptions met (1-4) OLS is actually the *best* that can be done. In this estimation problem.
- By the *Gauss-Markov* theorem, which we will examine at greater length later in the course, if:
  1. $E(\epsilon) = 0$
  2. $V(\epsilon) = \sigma^2 < \infty$
  3. $COV(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$
  4. And if Y is a linear combination of Xs

  Then, the OLS estimate is actually the **B**est (lowest-variance), **L**inear, **U**nbiased, **E**stimator of the vector of β.
- Under these assumptions, we don't need anything more flexible

Recall Propertites of Estimators

We had a set of properties that are desirable for estimators of test statistics. Remember them?

1. Unbiasedness: $E(\hat{\beta} - \beta) = 0$
2. Efficiency: $minMSE = minE[(\hat{\beta} - \beta)^2]$
3. Consistency: In the limit, $\hat{\beta} \rightarrow \beta$
4. Sufficiency: Contains all the data we want

1 Power

2 Parameters, Estimators, and Propertites

3 Building Blocks
- Scatterplots, carefully
- Covariance & Correlation

4 Regression
- Impression of Models
- Interpretation of OLS

5 Bringing it Back

6 Linear Algebra Representation

Linear Algebra and Regresion

Write the fitted model in terms of linear algebra:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Then, under the same setup as for the univariate case, minimize $\sum E_i^2$.

$$\begin{aligned} S(\mathbf{b}) = \sum E_i^2 = \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{x}\beta) \\ &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta \\ &= y'y - (2y'X)\beta + \beta'(X'X)\beta \end{aligned}$$

Again, this is an optimization problem. Objective function is $S(\beta)$, wrt β.

Linear Algebra Derivation

$$\begin{aligned}\frac{\partial S(\beta)}{\partial \beta} &= 0 - 2X'y + 2X'X\beta = 0 \\ &= X'X\beta = X'y \\ \beta &= (X'X)^{-1}X'y\end{aligned}$$