Economics 103 – Statistics for Economists

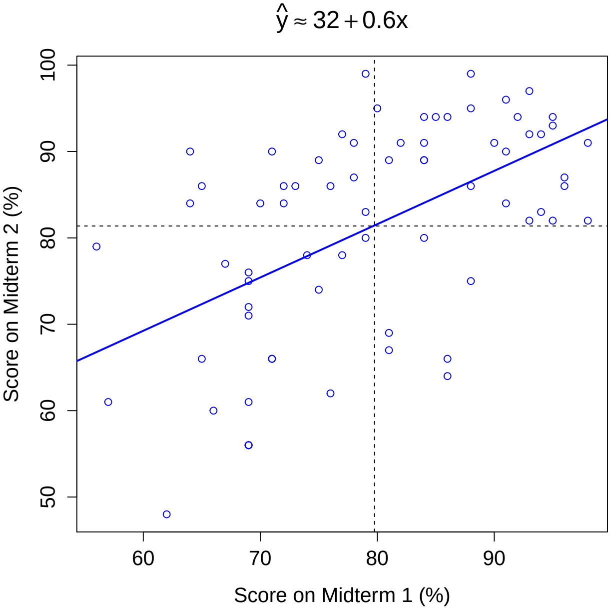
Minsu Chang

University of Pennsylvania

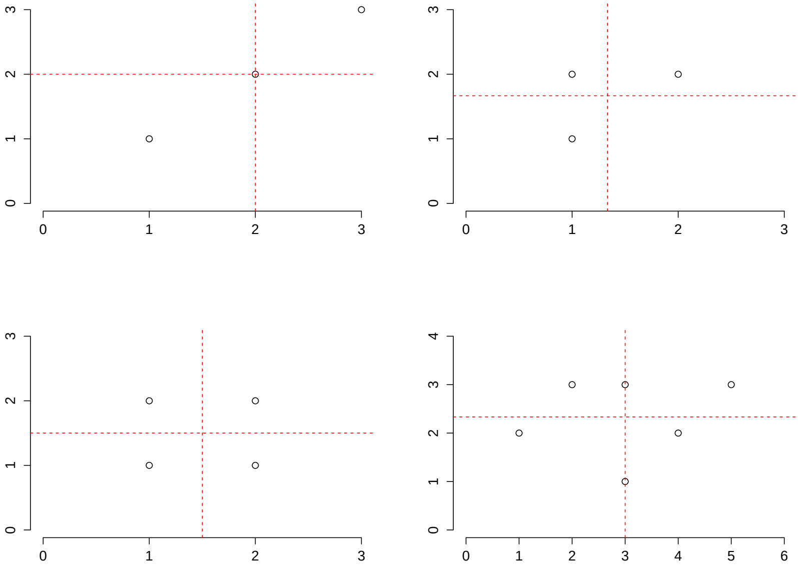
Lecture # 23

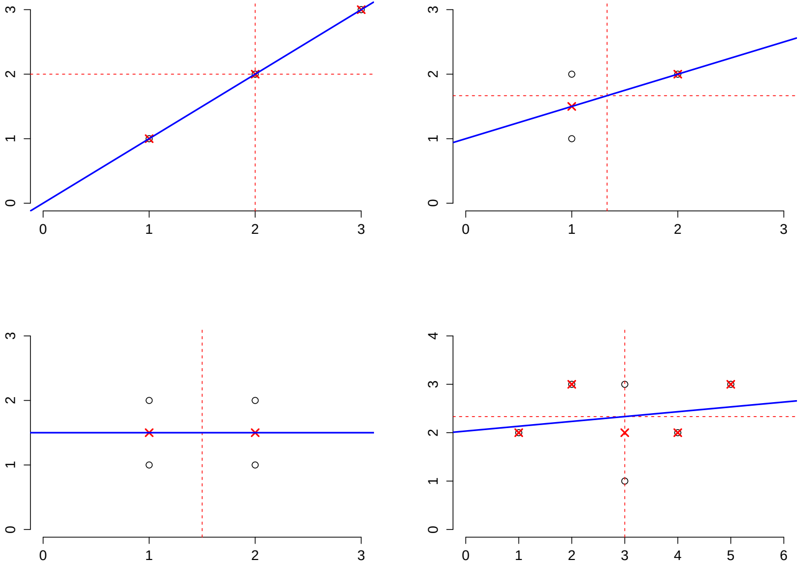
Introduction to Regression

Regression: "Best Fitting" Line Through Cloud of Points



Fitting a Line by Eye




0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3
0 1 2 3 4
0 1 2 3 4 5 6

But How to Do this Formally?

Least Squares Regression – Predict Using a Line

The Prediction

Predict score $\hat{y} = a + bx$ on 2nd midterm if you scored x on 1st

How to choose (a, b)?

Linear regression chooses the slope (b) and intercept (a) that

minimize the sum of squared vertical deviations

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Why Squared Deviations?

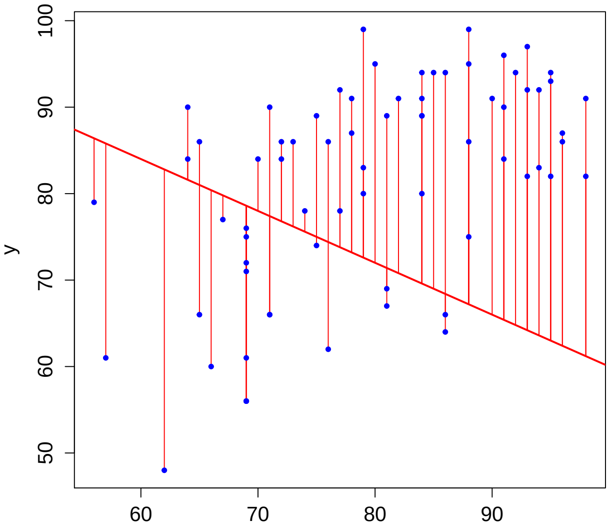
Important Point About Notation

$$\boxed{\underset{a,b}{\text{minimize}} \ \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2}$$

$$\boxed{\hat{y} = a + bx}$$

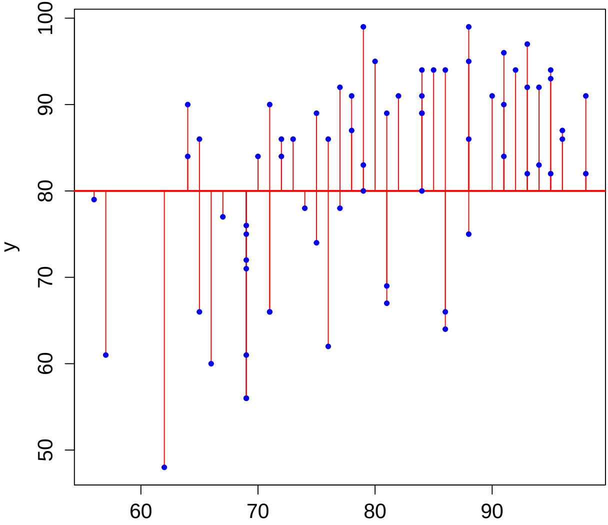
- $(x_i, y_i)_{i=1}^n$ are the observed data
- $\widehat{y}$ is our prediction for a given value of x
- Neither x nor $\widehat{y}$ needs to be in our dataset!

$$\sum d^2 = 25596.88$$



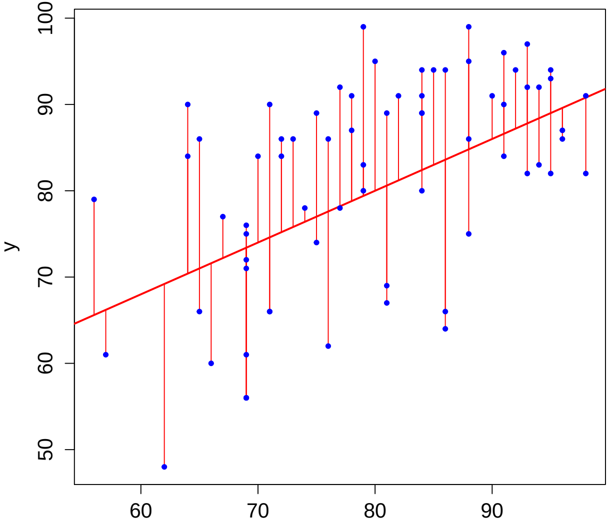
$\hat{y} = 120 + (-0.6)x$

$$\sum d^2 = 10728$$



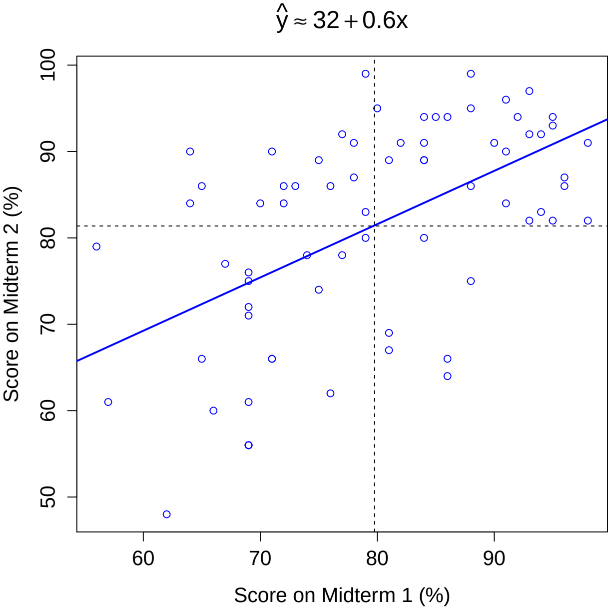
$\hat{y} = 80 + (0)x$

$$\sum d^2 = 7650.48$$

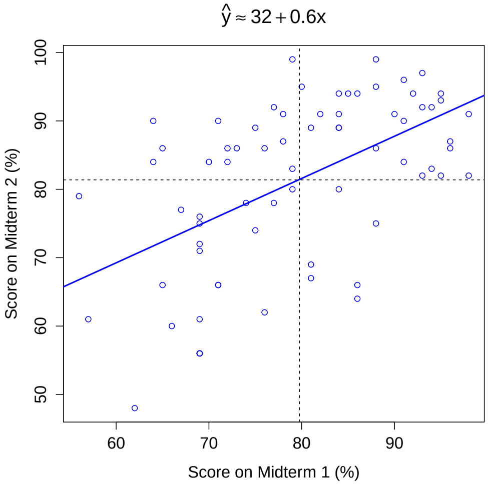


$$\hat{y} = 32 + (0.6)x$$

Prediction given 89 on Midterm 1?



Prediction given 89 on Midterm 1?



$$32 + 0.6 \times 89 = 32 + 53.4 = 85.4$$

How Can We Solve for a, b?

Minimize the sum of squared vertical deviations from the line:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

How should we proceed?

(a) Differentiate with respect to x

(b) Differentiate with respect to y

(c) Differentiate with respect to x, y

(d) Differentiate with respect to a, b

(e) Can't solve this with calculus.

Simple Linear Regression

Problem

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Solution

$$b = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Regression Line Goes Through the Means!

$$\boxed{\bar{y} = a + b\bar{x}}$$

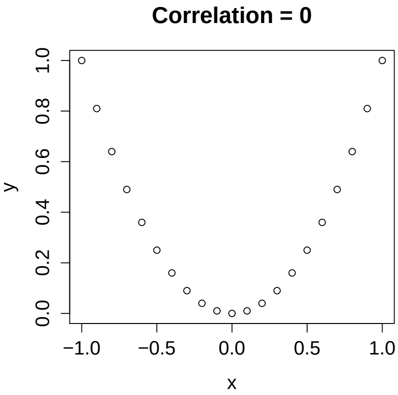
Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = b\frac{s_x}{s_y}$$

EXTREMELY IMPORTANT

- Regression, Covariance and Correlation: linear association.
- Linear association $\neq$ causation.
- Linear is not the only kind of association!



The Population Regression Model

How is Y (height) related to X (handspan) in the population?

Assumption I: Linearity

The random variable Y is linearly related to X according to

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

β_0, β_1 are two unknown population parameters (constants).

Assumption II: Error Term ϵ

$E[\epsilon] = 0$, $Var(\epsilon) = \sigma^2$ and ϵ is indpendent of X. The error term ϵ measures the unpredictability of Y *after controlling for* X

Estimating β_0, β_1

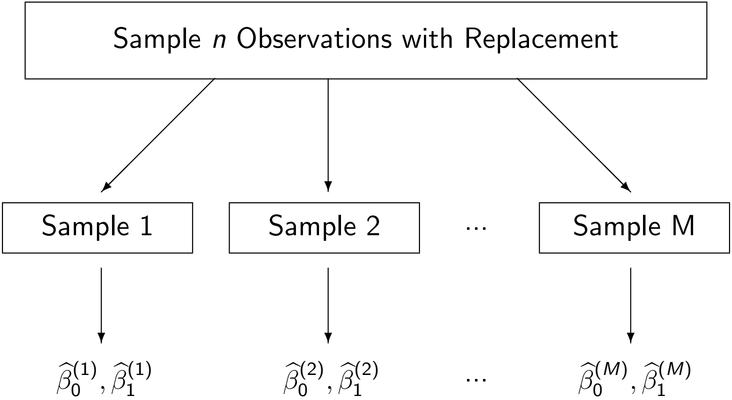
Suppose we observe an iid sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ from the population: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Then we can *estimate* β_0, β_1:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

$$\widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{X}_n$$

Once we have estimators, we can think about sampling uncertainty...

Sampling Distribution of Regression Coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$



Repeat M times $\rightarrow$ get M different pairs of estimates

Sampling Distribution: long-run relative frequencies

Inference for Linear Regression

Central Limit Theorem

$$\frac{\widehat{\beta} - \beta}{\widehat{SE}(\widehat{\beta})} \approx N(0,1)$$

How to calculate $\widehat{SE}$?

- Complicated
  - Depends on variance of errors ϵ and all predictors in regression.
  - We'll look at a few simple examples
  - R does this calculation for us
- Requires assumptions about population errors ϵ_i
  - Simplest (and R default) is to assume $\epsilon_i \sim iid(0, \sigma^2)$
  - Weaker assumptions in Econ 104

Let's consider various inferences we can draw from the height and handspan data using regression in R.

Height $= \beta_0 + \epsilon$

```
lm(formula = height ~ 1, data = student.data)
            coef.est coef.se
(Intercept) 67.74    0.51
---
n = 80, k = 1
```

Height $= \beta_0 + \epsilon$

```
lm(formula = height ~ 1, data = student.data)
            coef.est coef.se
(Intercept) 67.74     0.51
---
n = 80, k = 1

> mean(student.data$height)
[1] 67.7375
```

Dummy Variable (aka Binary Variable)

A predictor variable that takes on only two values: 0 or 1. Used to represent two categories, e.g. Male/Female.

Height $= \beta_0 + \beta_1$ Male $+\epsilon$

```
lm(formula = height ~ sex, data = student.data)
            coef.est coef.se
(Intercept) 64.46     0.56
sexMale      6.10     0.76
---
n = 80, k = 2
residual sd = 3.38, R-Squared = 0.45
```

Height $= \beta_0 + \beta_1$ Male $+\epsilon$

```
lm(formula = height ~ sex, data = student.data)
            coef.est coef.se
(Intercept) 64.46    0.56
sexMale      6.10    0.76
---
n = 80, k = 2
residual sd = 3.38, R-Squared = 0.45

> mean(male$height) - mean(female$height)
[1] 6.09868
```

Height $= \beta_0 + \beta_1$ Male $+\epsilon$

What is the ME for an approximate 95% confidence interval for the difference of population means of height: (men - women)?

```
lm(formula = height ~ sex, data = student.data)
            coef.est coef.se
(Intercept) 64.46     0.56
sexMale      6.10     0.76
---
n = 80, k = 2
residual sd = 3.38, R-Squared = 0.45
```

Height $= \beta_0 + \beta_1$ Handspan $+\epsilon$

```
lm(formula = height ~ handspan, data = student.data)
            coef.est coef.se
(Intercept) 39.60     3.96
handspan     1.36     0.19
---
n = 80, k = 2
residual sd = 3.56, R-Squared = 0.40
```

Height $= \beta_0 + \beta_1$ Handspan $+\epsilon$

What is the ME for an approximate 95% CI for β_1?

```
lm(formula = height ~ handspan, data = student.data)
            coef.est coef.se
(Intercept) 39.60     3.96
handspan     1.36     0.19
---
n = 80, k = 2
residual sd = 3.56, R-Squared = 0.40
```

Simple vs. Multiple Regression

Terminology

Y is the "outcome" and X is the "predictor."

Simple Regression

One predictor variable: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Multiple Regression

More than one predictor variable:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$

- In both cases $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \sim \text{iid}(0, \sigma^2)$
- Multiple regression coefficient estimates $\widehat{\beta}_1, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ calculated by minimizing sum of squared vertical deviations, but formula requires linear algebra so we won't cover it.

Interpreting Multiple Regression

Predictive Interpretation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

β_j is the difference in Y that we would predict between two individuals who differed by one unit in predictor X_j *but who had the same values for the other X variables.*

What About an Example?

In a few minutes, we'll work through an extended example of multiple regression using real data.

Inference for Multiple Regression

In addition to estimating the coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ for us, R will calculate the corresponding standard errors. It turns out that

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}(\widehat{\beta}_j)} \approx N(0,1)$$

for *each* of the $\widehat{\beta}_j$ by the CLT provided that the sample size is large.

Height $= \beta_0 + \beta_1$ Handspan $+\epsilon$

What are residual sd and R-squared?

```
lm(formula = height ~ handspan, data = student.data)
            coef.est coef.se
(Intercept) 39.60    3.96
handspan     1.36    0.19
---
n = 80, k = 2
residual sd = 3.56, R-Squared = 0.40
```

Fitted Values and Residuals

Fitted Value $\widehat{y}_i$

Predicted y-value for person i given her x-variables using estimated regression coefficients: $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}$

Residual $\widehat{\epsilon}_i$

Person i's *vertical deviation* from regression line: $\widehat{\epsilon}_i = y_i - \widehat{y}_i$.

The residuals are *stand-ins* for the unobserved errors ϵ_i.

Residual Standard Deviation: $\widehat{\sigma}$

- Idea: use residuals $\widehat{\epsilon}_i$ to estimate σ

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \widehat{\epsilon}_i^2}{n-k}}$$

- Measures avg. distance of y_i from regression line.
  - E.g. if Y is points scored on a test and $\widehat{\sigma} = 16$, the regression predicts to an accuracy of about 16 points.
- Same units as Y
- Denominator $(n-k) = (\#$ Datapoints - $\#$ of X variables)

Proportion of Variance Explained: R^2

aka Coefficient of Determination

$$R^2 \approx 1 - \frac{\widehat{\sigma^2}}{s_y^2}$$

- $R^2 =$ proportion of $Var(Y)$ "explained" by the regression.
  - Higher value $\implies$ greater proportion explained
- Unitless, between 0 and 1
- Generally harder to interpret than $\widehat{\sigma}$, but...
- For simple linear regression $R^2 = (r_{xy})^2$ and this is where its name comes from!

Height $= \beta_0 + \beta_1$ Handspan $+\epsilon$

```
lm(formula = height ~ handspan, data = student.data)
            coef.est coef.se
(Intercept) 39.60     3.96
handspan     1.36     0.19
---
n = 80, k = 2
residual sd = 3.56, R-Squared = 0.40
> cor(student.data$height, student.data$handspan)^2
[1] 0.3954669
```

Which Gives Better Predictions: Sex (a) or Handspan (b)?

```
lm(formula = height ~ sex, data = student.data)
            coef.est coef.se
(Intercept) 64.46    0.56
sexMale      6.10    0.76
---
n = 80, k = 2
residual sd = 3.38, R-Squared = 0.45

lm(formula = height ~ handspan, data = student.data)
            coef.est coef.se
(Intercept) 39.60    3.96
handspan     1.36    0.19
---
n = 80, k = 2
residual sd = 3.56, R-Squared = 0.40
```