

# Economics 103 – Statistics for Economists

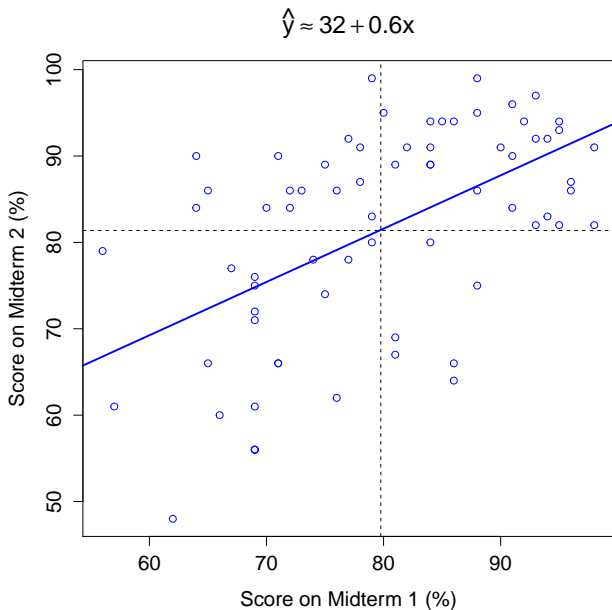
Francis J. DiTraglia

University of Pennsylvania

Lecture 24

# Regression – Part II

## Recall: “Best Fitting” Line Through Cloud of Points



# Recall: Regression as a Data Summary

## Linear Model

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

Choose  $a, b$  to Minimize Sum of Squared Vertical Deviations

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

## The Prediction

Predict score  $\hat{y} = a + bx$  on second midterm for someone with score  $x$  on first.

# Recall: Regression as a Data Summary

## Problem

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

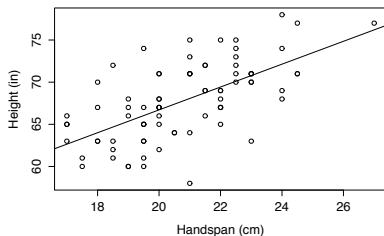
## Solution

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

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

## Beyond Regression as a Data Summary

Based on a sample of Econ 103 students, we made the following graph of handspan against height, and fitted a linear regression:



The estimated slope was about 1.4 inches/cm and the estimated intercept was about 40 inches.

What if anything does this tell us about the relationship between height and handspan *in the population*?

# 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

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

$\beta_0, \beta_1$  are two unknown population parameters (constants).

## Assumption II: Error Term $\epsilon$

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

# Predictive Interpretation of Regression

Under Assumptions I and II

$$E[Y|X] = \beta_0 + \beta_1 X$$

- ▶ “Best guess” of  $Y$  having observed  $X = x$  is  $\beta_0 + \beta_1 x$
- ▶ If  $X = 0$ , we predict  $Y = \beta_0$
- ▶ If two people differ by one unit in  $X$ , we predict that they will differ by  $\beta_1$  units in  $Y$ .

The only problem is, we don't know  $\beta_0, \beta_1 \dots$



## Estimating $\beta_0, \beta_1$

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

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

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

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

## Sampling Uncertainty: Pretend the Class is our Population

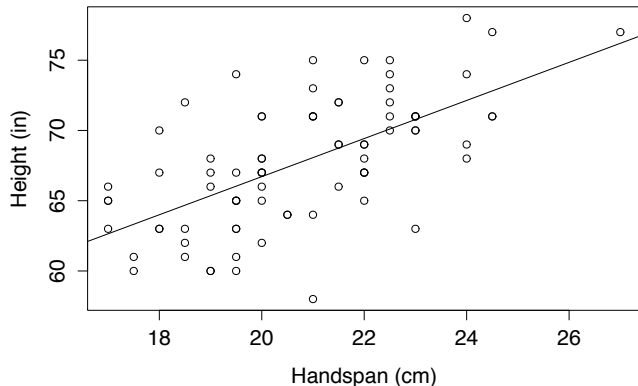
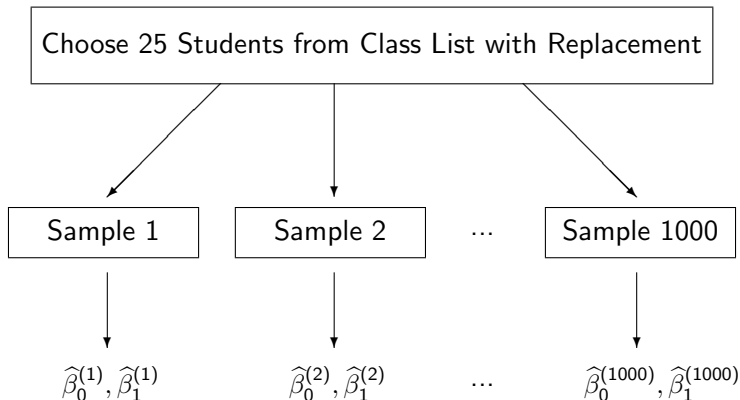


Figure : Estimated Slope = 1.4, Estimated Intercept = 40

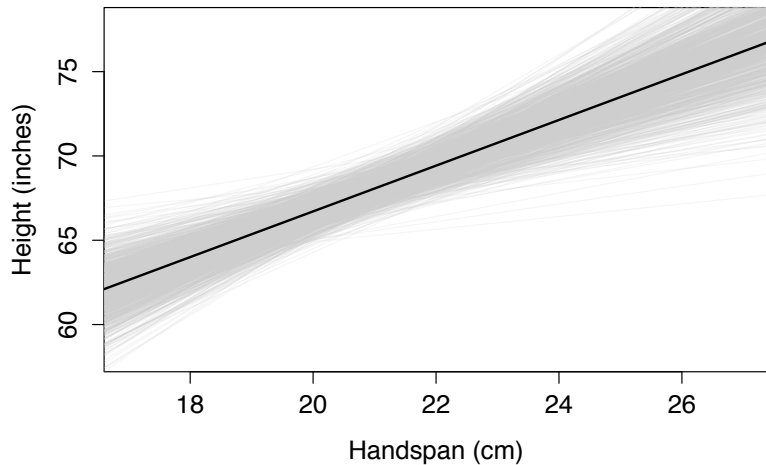
# Sampling Distribution of Regression Coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$



Repeat 1000 times → get 1000 different pairs of estimates

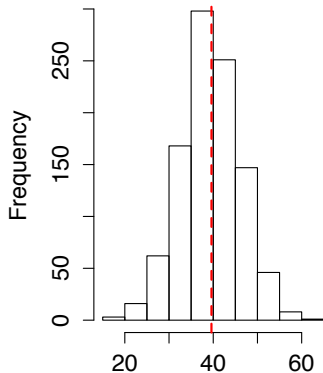
Sampling Distribution: long-run relative frequencies

1000 Replications,  $n = 25$

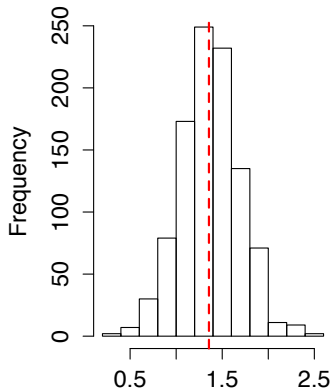


Population: Intercept = 40, Slope = 1.4

**Intercept Estimates**



**Slope Estimates**



Based on 1000 Replications,  $n = 25$

# Inference for Linear Regression

## Central Limit Theorem

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

## How to calculate $\widehat{SE}$ ?

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

## Intuition for What Effects $SE(\hat{\beta}_1)$ for Simple Regression

$$SE(\hat{\beta}_1) \approx \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{s_X}$$

- ▶  $\sigma = SD(\epsilon)$  – inherent variability of the  $Y$ , even after controlling for  $X$
- ▶  $n$  is the sample size
- ▶  $s_X$  is the sampling variability of the  $X$  observations.

I treated the class as our population for the purposes of the simulation experiment but it makes more sense to think of the class as a sample from some population. We'll take this perspective now and think about various inferences we can draw from the height and handspan data using regression.



$$\text{Height} = \beta_0 + \epsilon$$

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

$$\text{Height} = \beta_0 + \epsilon$$

```
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
```

$$\text{Height} = \beta_0 + \epsilon$$

```
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
```

```
> sd(student.data$height)/sqrt(length(student.data$height))
```

```
[1] 0.5080814
```

## 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.

$$\text{Height} = \beta_0 + \beta_1 \text{ Male} + \epsilon$$

```
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
```

$$\text{Height} = \beta_0 + \beta_1 \text{ Male} + \epsilon$$

```
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
```

$$\text{Height} = \beta_0 + \beta_1 \text{ Male} + \epsilon$$

```
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
```

```
> sqrt(var(male$height)/length(male$height) +  
      var(female$height)/length(female$height))
```

```
[1] 0.7463796
```

$$\text{Height} = \beta_0 + \beta_1 \text{ Male} + \epsilon$$



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
```



$$\text{Height} = \beta_0 + \beta_1 \text{ Handspan} + \epsilon$$

```
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
```

$$\text{Height} = \beta_0 + \beta_1 \text{ Handspan} + \epsilon$$



What is the ME for an approximate 95% CI for  $\beta_1$ ?

```
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:

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

- ▶ In both cases  $\epsilon_1, \epsilon_2, \dots, \epsilon_n \sim \text{iid}(0, \sigma^2)$
- ▶ Multiple regression coefficient estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  calculated by minimizing sum of squared vertical deviations, but formula requires linear algebra so we won't cover it.

# Interpreting Multiple Regression

## Predictive Interpretation

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

$\beta_j$  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  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  for us, R will calculate the corresponding standard errors. It turns out that

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

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

$$\text{Height} = \beta_0 + \beta_1 \text{Handspan} + \epsilon$$

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 $\hat{y}_i$

The value of the  $Y$ -variable that we would *predict* for person  $i$  using our estimated regression coefficients, given her values for all of the  $X$ -variables:  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}$

## Residual $\hat{\epsilon}_i$

The difference between the actual value of the  $Y$ -variable that we observed for person  $i$  ( $y_i$ ) and the value predicted from our regression model:  $\hat{\epsilon}_i = y_i - \hat{y}_i$ . The residuals are *stand-ins* for the errors  $\epsilon_i$ , which we don't observe.

## Residual Standard Deviation: $\hat{\sigma}$

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

- ▶ Average distance observations fall from regression line.
- ▶ Summarizes scale of the residuals  $\hat{\epsilon}_i$ 
  - ▶ Suppose model predicting children's test scores has  $\hat{\sigma} = 16$ .  
This model predicts to an accuracy of about 16 points.
- ▶ A measure of “unexplained variation” in  $Y$ 
  - ▶ Higher values means there is more unexplained variation in  $Y$
- ▶ Same units as  $Y$  (Exam practice: verify this)
- ▶ Denominator  $(n - k) =$  “degrees of freedom” of regression
  - ▶ ( $\#$  Datapoints -  $\#$  of  $X$  variables)



# Proportion of Variance Explained: $R^2$

aka Coefficient of Determination

$$R^2 = 1 - \left[ \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right] \approx 1 - \frac{\widehat{\sigma^2}}{s_y^2}$$

- ▶ Measures proportion of variance in  $Y$  explained by the model.
  - ▶ Higher value means greater proportion explained
- ▶ Unitless, between 0 and 1
- ▶ People often claim “you want  $R^2$  to be as high as possible” but this is not quite accurate...
- ▶ For simple linear regression  $R^2 = (r_{xy})^2$  and this where its name comes from!

$$\text{Height} = \beta_0 + \beta_1 \text{ Handspan} + \epsilon$$

```
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
```

Bring Your Laptop Next Time:  
We'll be Using R