Course Materials May Not Be Distributed or Posted Electronically

These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service.

PSTAT 126
Regression Analysis

Dr. Todd Gross

Department of Statistics and Applied Probability

UCSB

Lecture 2

Lecture Outline

- Administrative Issues
- Introduction to Regression
- The Linear Regression Model
- Calculating and Interpreting the Regression Equation
- Goodness of Fit

Lecture Outline

- Administrative Issues
- Introduction to Regression
- The Linear Regression Model
- Calculating and Interpreting the Regression Equation
- Goodness of Fit

Assigned Readings for this Lecture

- Introduction to Statistical Learning (ISL)
  - Ch 3 Section 3.1 – Simple Linear Regression
- Introduction to Linear Regression
  - http://onlinestatbook.com/2/regression/intro.html
- Penn State Stat 501 – Lesson 1: Simple Linear Regression
  - https://onlinecourses.science.psu.edu/stat501/node/250

Introduction To Regression

Regression analysis is among the most useful and widely used statistical tools in practice.

Suppose we have

- **Y**: dependent (response or outcome) variable
- **X**: independent (predictor or explanatory) variable

We want to

- **Describe** the relationship between X and Y
- **Predict** future observations of Y using values of X

X and Y Relationships

- Does weight (Y) depend on height (X)?
- Does life expectancy (Y) depend on blood pressure level (X)?
- Does lung capacity (Y) decrease with the number of the cigarettes smoked per day (X)?
- Do sales (Y) increase with the advertising expenditure (X)?

Types of Relationships Between Variables

- Functional
  - Y can be directly calculated from X <u>without</u> error
    - Revenue (Y) from Units Sold (X)
    - Time for Light to Travel (Y) from Distance Between Points (X)
- Statistical
  - Y <u>may</u> be a function of X, but <u>with</u> some error
    - Weight (Y) from Height (X)
    - Annual Raise (Y) from Performance Rating (X)

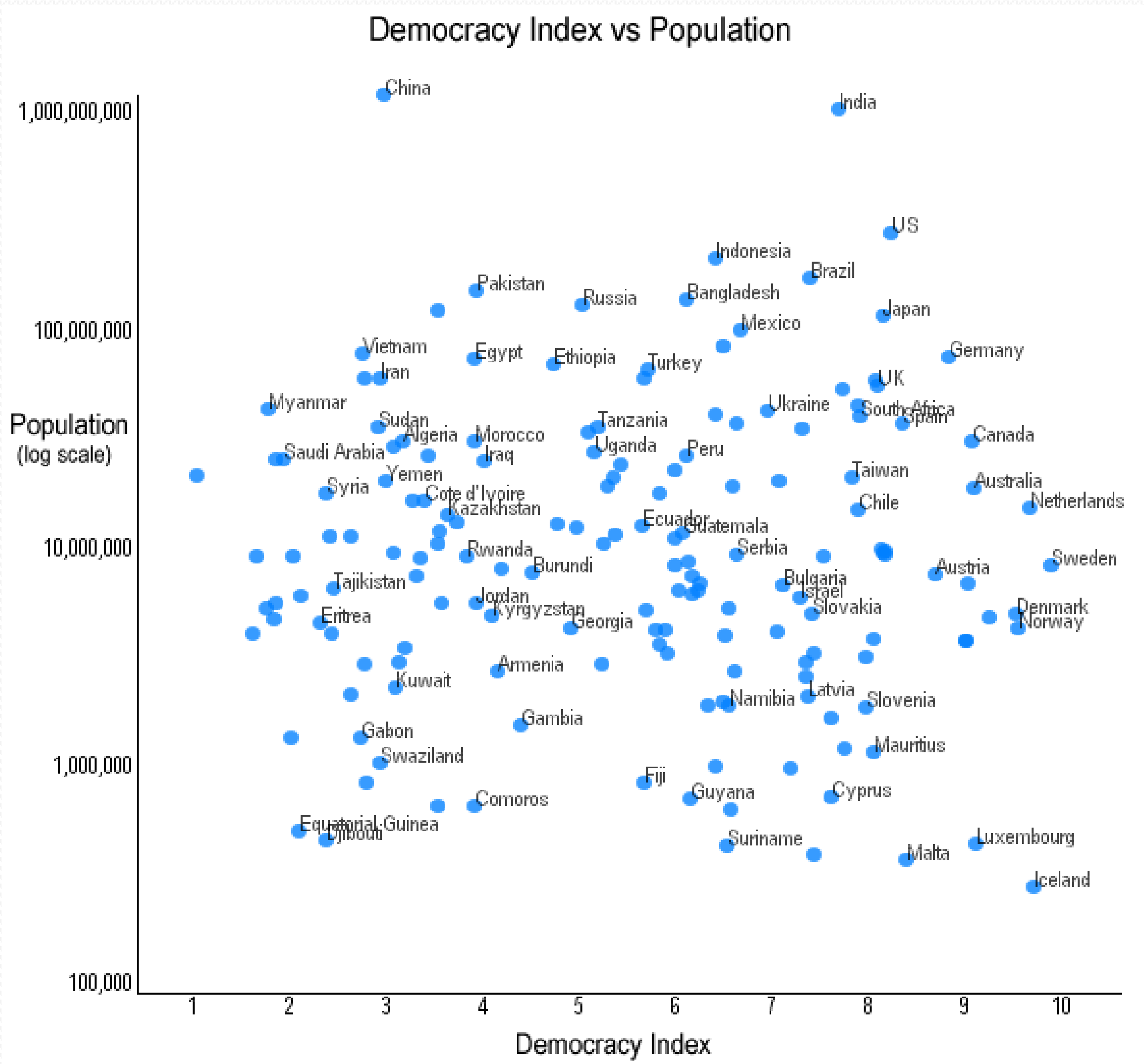
Types of Relationships (con't)

- No Relationship
  - Y cannot be predicted from X

- Linear
  - Y is a linear function of X
  - "straight line" relationship

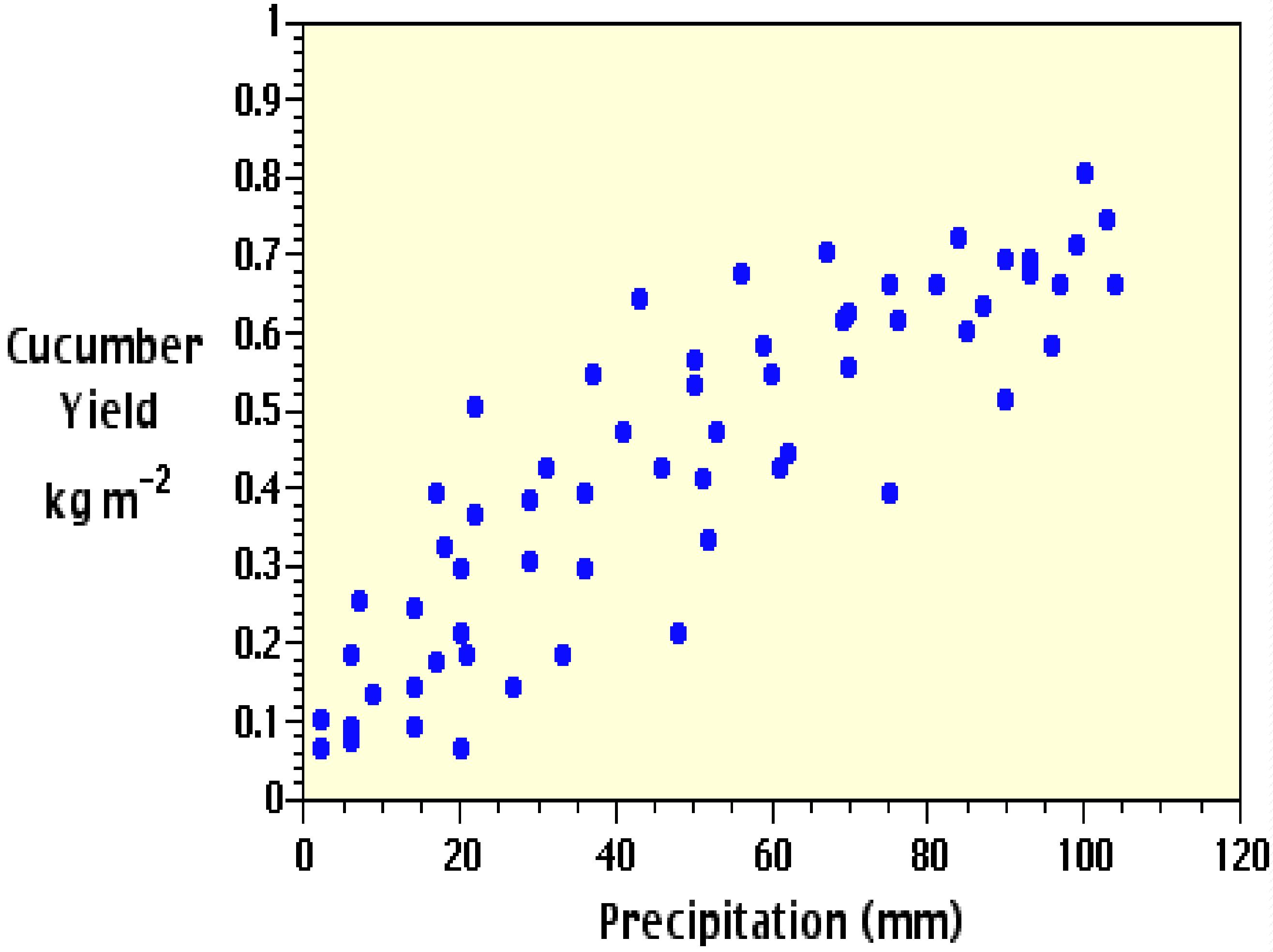
- Non-linear
  - Y is a non-linear function of X
  - Curvilinear, sinusoidal

No Relationship



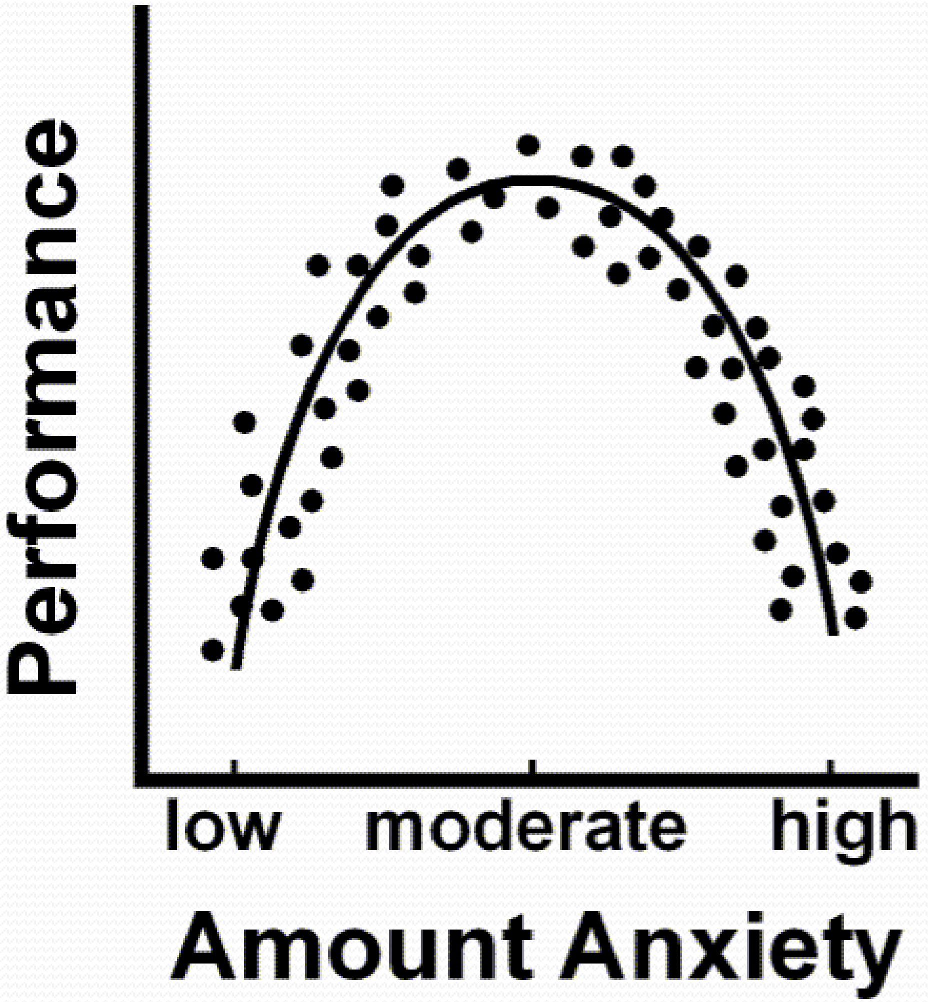
http://www.neoformix.com/2007/DemocracyIndex.html

Linear Relationship



http://www.physicalgeography.net/fundamentals/3h.html

Non-Linear Relationship



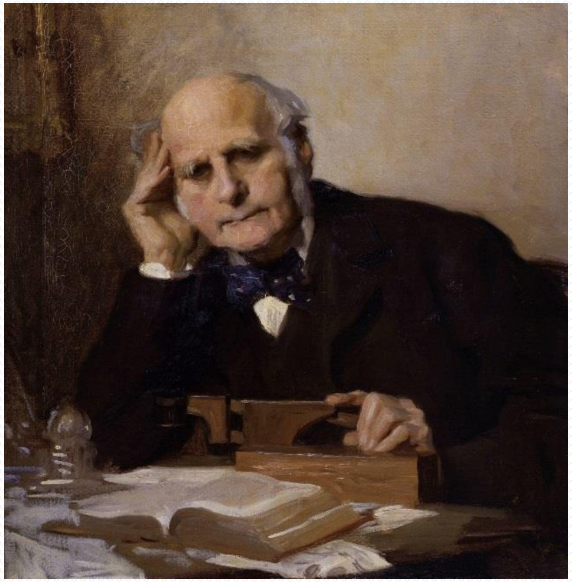
http://changingminds.org/explanations/motivation/yerkes-dodson.htm

How Do We Use Regression?

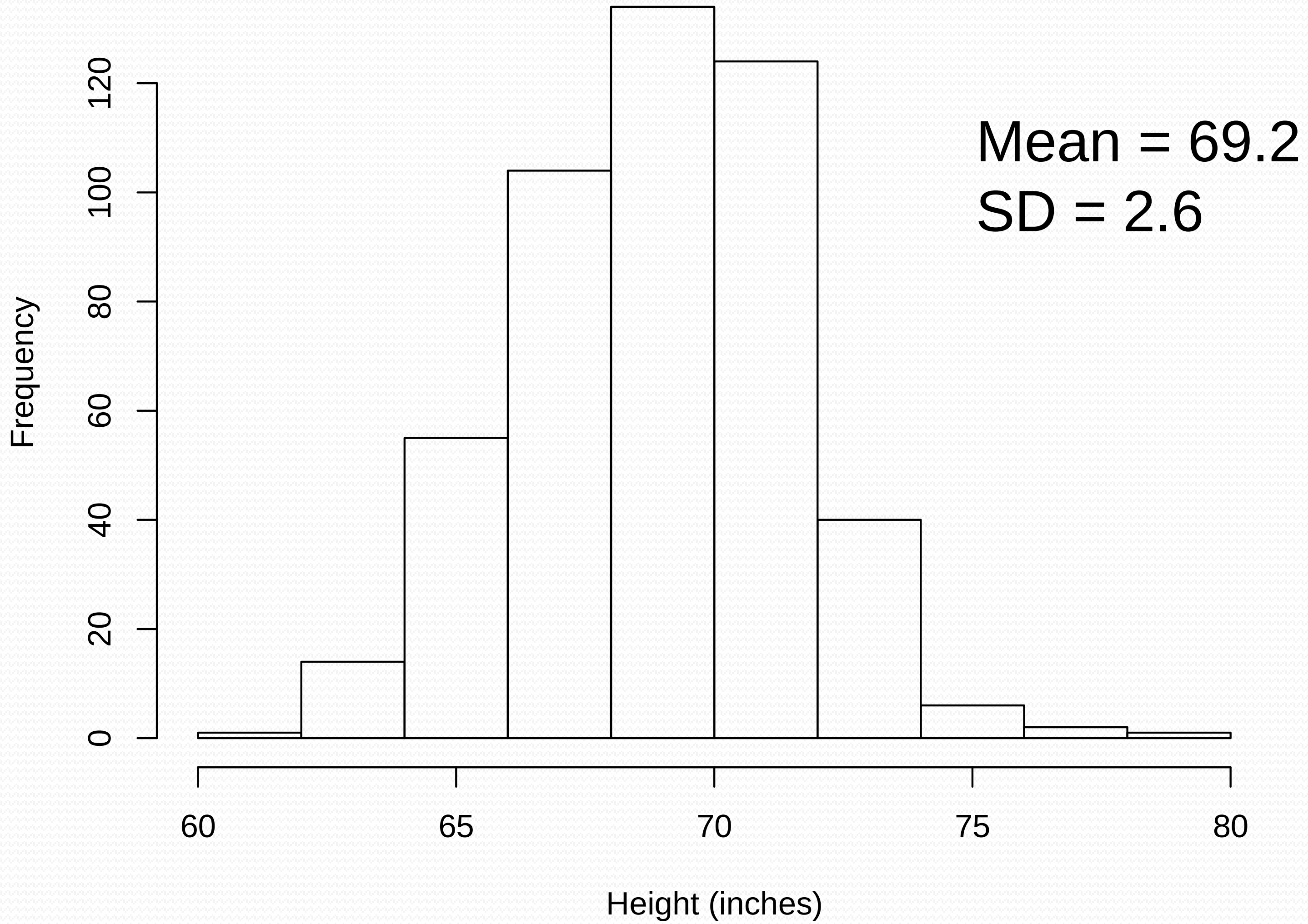
- Predict an unknown outcome based on known information
- Determine if prediction is better than chance (i.e., better than predicting the mean)
- Identify which variables are useful predictors
- Build a model to be tested in a future study

Brief History of Regression

- Developed by Sir Francis Galton (1822-1911)
  - Cousin of Charles Darwin
  - Credited with standard deviation (1888)
  - Karl Pearson’s doctoral advisor
- Interested in heredity
  - Described regression techniques
  - Identified “regression to the mean”
  - Criticized for proposing Eugenics

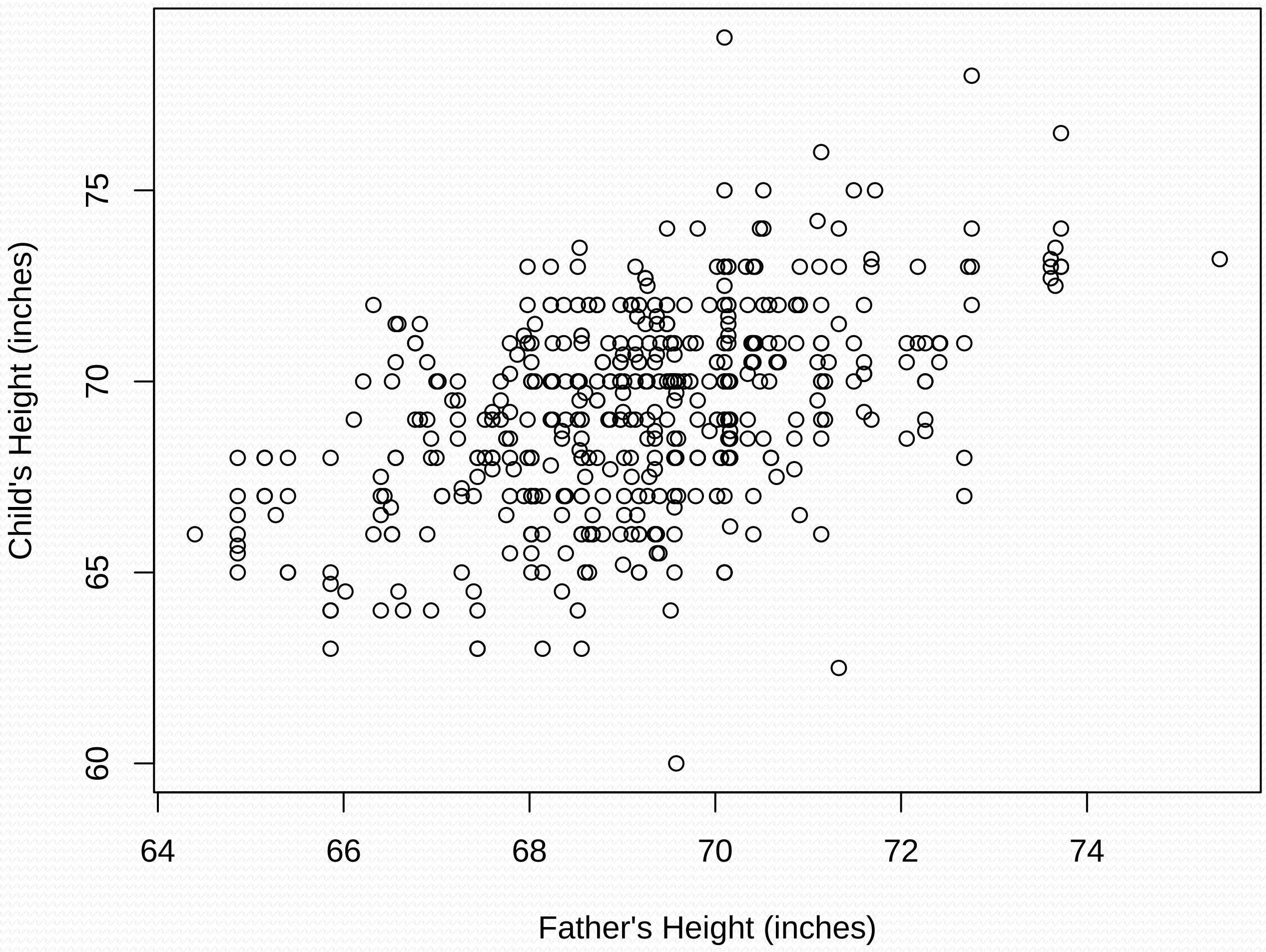


Distribution of Male Height



What height should we predict for a random individual?

Is There a Relationship between the Height of Parents and Children?

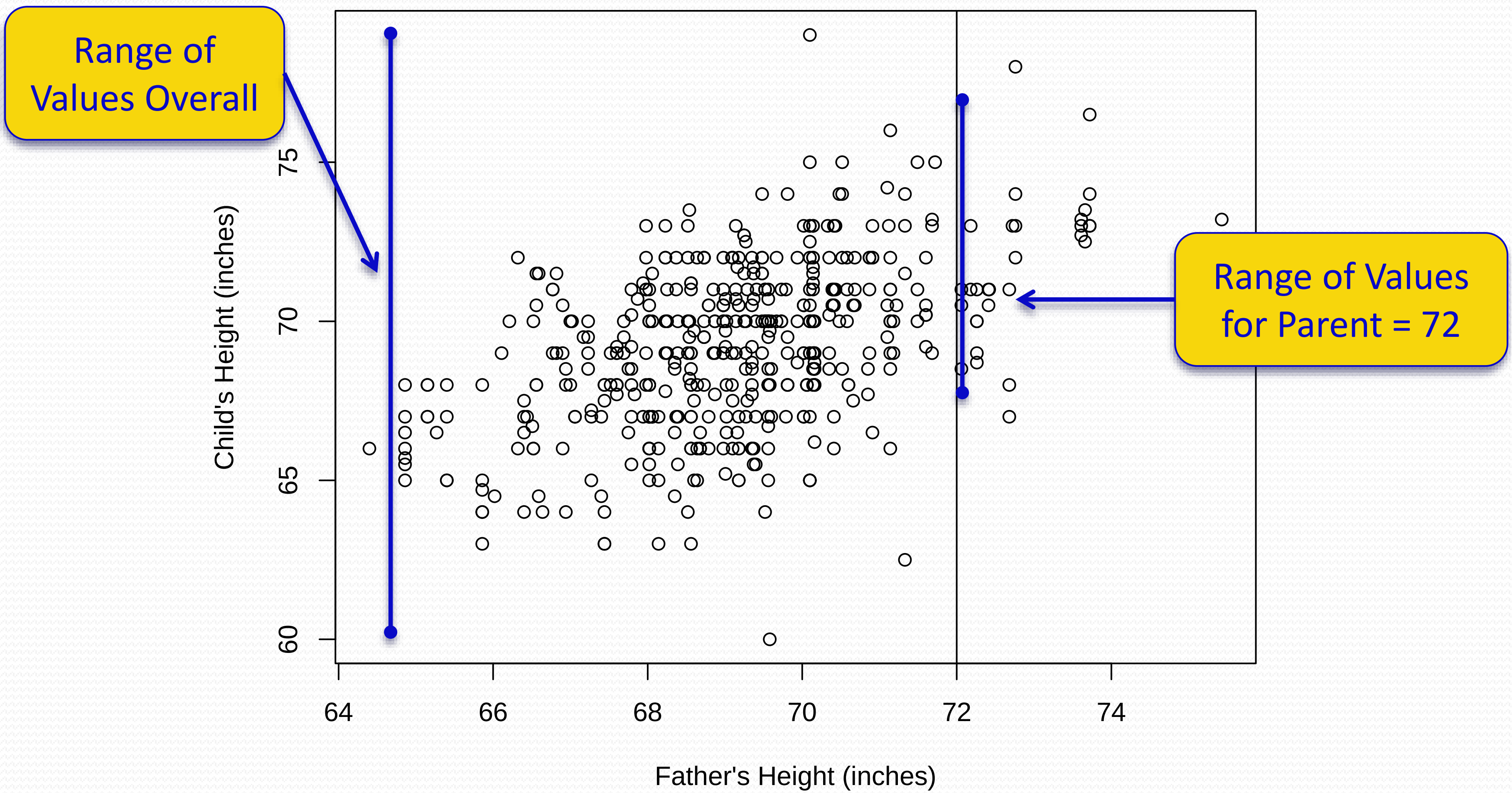


Describe the Relationship

We want to state 3 things:

- Strength of the relationship
- Direction of the relationship
- Is it Linear or Non-Linear

Is There a Relationship between the Height of Fathers and Sons?



Does knowing the Father's height improve our prediction of the son's height?

Regression Examples

- Predict first year post-college annual income based on college GPA
- Predict length of marriage based on age, hours of interaction, income
- Predict Response to Medical Treatment based on demographic and baseline characteristics (BMI, age, sex, race, baseline severity)

Lecture Outline

- Administrative Issues
- Introduction to Regression
- The Linear Regression Model
- Calculating and Interpreting the Regression Equation
- Goodness of Fit

Steps in a Regression Analysis

1. Identify the research question
2. Identify the target population
3. Collect a sample of subjects (or participants)
4. For each subject, measure values for X and Y
5. Calculate the regression equation

----

6. For individuals that were not in your sample, use the regression equation to predict a future value of the response using their value of the predictor

A Simple Example

- What variable (that is under your control) will predict your score on the midterm?

Hours of Study and Exam Score

Students Study for an Exam

For each student we measure:

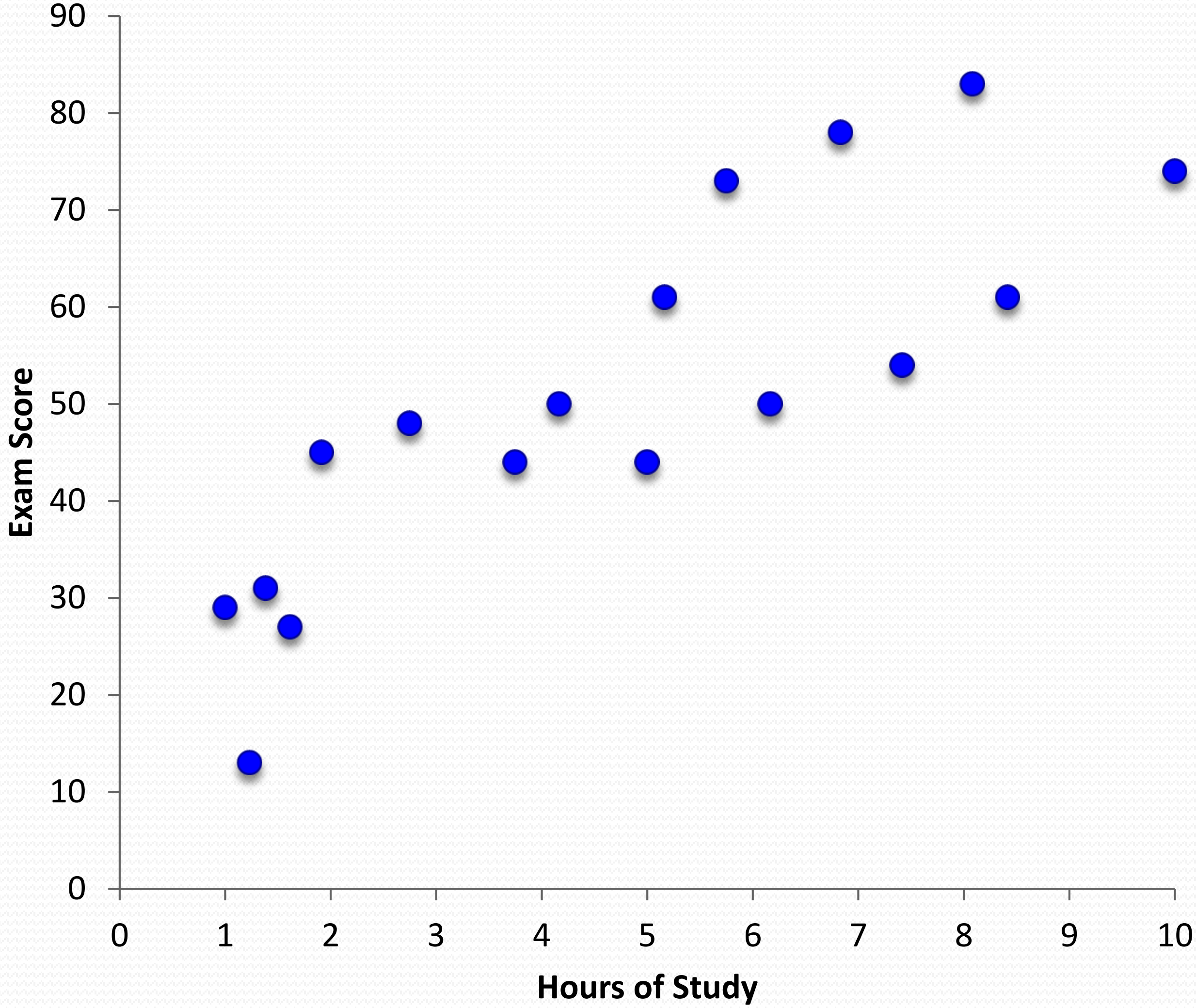
- the number of hours of study (X)
- the score on the exam (Y)

We want to know:

1. Is there a relationship between hours of study and score?
2. Is the relationship linear?
3. Describe the relationship

Student	Hours of Study	Exam Score
1	1.0	29
2	1.2	13
3	1.4	31
4	1.6	27
5	1.9	45
6	2.8	48
7	3.8	44
8	4.2	50
9	5.0	44
10	5.2	61
11	5.8	73
12	6.2	50
13	6.8	78
14	7.4	54
15	8.1	83
16	8.4	61
17	10.0	74
18	10.7	67
19	11.1	80

Is Exam Score Related to Study Hours?



Do students who study more hours than average score more than average on the exam?

The Foundation of Linear Regression

We collect *n* pairs of observations:

$$(Y_i ; X_i); \; i = 1; \ldots; n$$

We want to establish a statistical model that summarizes the relationship between X and Y.

The simplest relationship between X and Y is linear:

$$Y = B_o + B_1 X$$

But the relationship is not perfect:

$$Y - (B_o + B_1 X) = \varepsilon$$

where ε is a random error which contains all sources of variation <u>unexplained</u> by the linear relationship.

Linear Regression Model

- The formal regression model is:

$$Y_i = b_0 + b_1 X_i + e_i$$

- Y_i is the i^{th} observation of a response variable
- X_i is the i^{th} observation of the predictor variable
- B_0 and B_1 are regression parameters (constants)
- E_ι is the error for the i^{th} observation of Y

- This equation defines:
  - A linear relationship between X and Y
  - A line of best fit through a plot of X and Y

Linear Regression Model

The linear regression model assumes that

$$Y_i = B_0 + B_1 X_i + \varepsilon_i; \; i = 1; \ldots; n$$

where ε_i are random errors with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i; \varepsilon_j) = 0$ for all $i \neq j$.

- $(X_i; Y_i)$ are observed and provided in data.
- B_0, B_1 and σ^2 are unknown parameters of the model.
- Note that ε_i are unobservable.

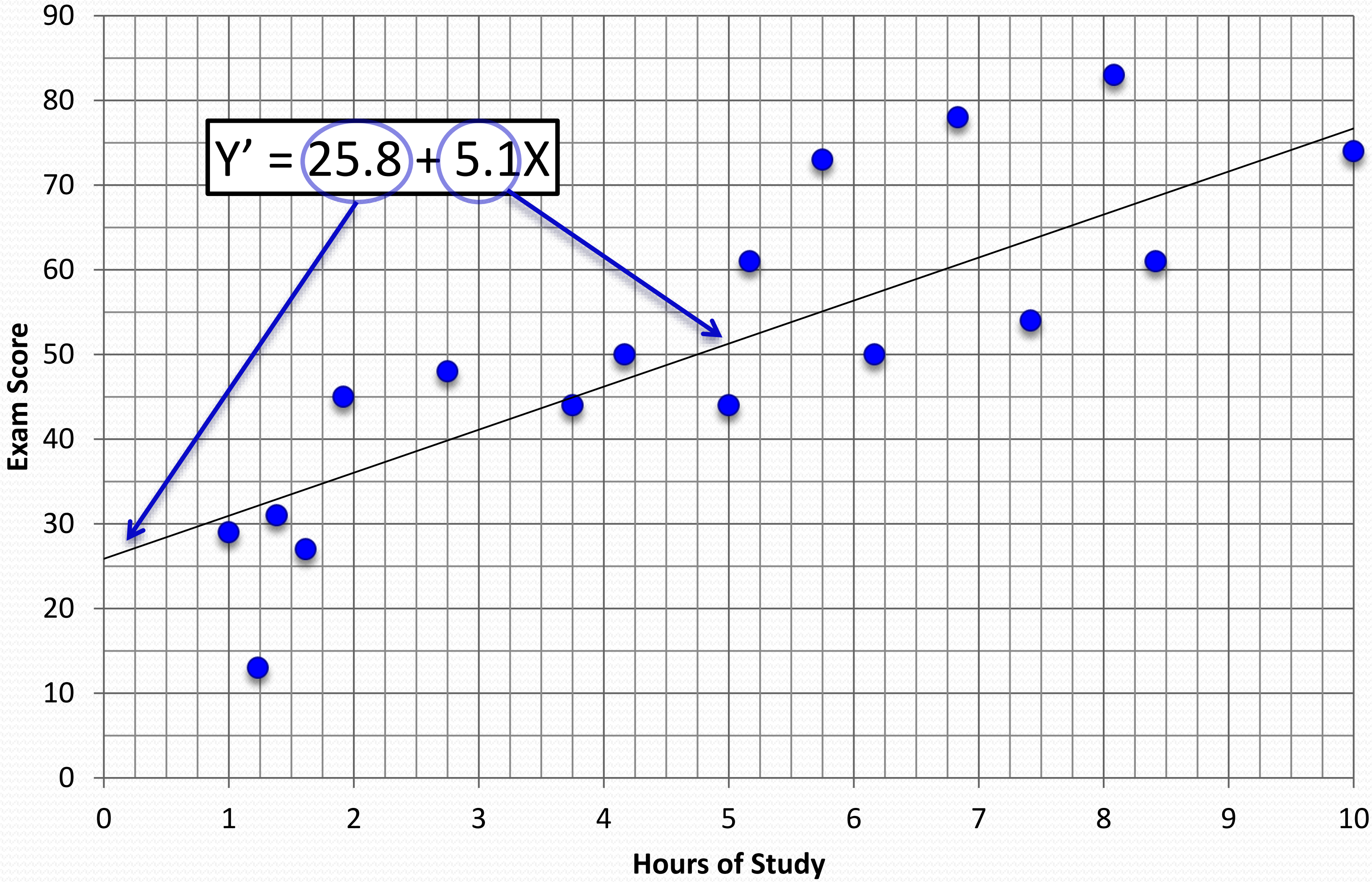
The Fitted Regression Model

$$Y_i' = b_o + b_1 X_i$$

- Y'_i is the value of Y we predict based on the value of X_i

- The predicted value (Y') may not be the same as the observed value (Y) due to random error (residuals)

Scatterplot and Regression Equation



Interpreting the Regression Parameters

Slope

- Amount of change in Y we predict per 1 unit change in X
- Exam Example:
  - For each additional hour of study we predict a 5.1 point increase in exam score

Intercept

- Value of Y we predict when X = 0
- Exam Example:
  - We predict an exam score of 25.8 if student doesn’t study at all

Interpreting (continued)

- Slope and Intercept are in the <u>same units</u> as the Y variable
  - Slope: 5.1 <u>exam points</u> per hour of study
  - Intercept: 25.8 <u>exam points</u> with no study

- Exam Example:

  "We predict that a student who does not study at all will score 25.8 points on the exam, and that the score will increase 5.1 points for every additional hour of study"

Using Regression to Predict Future Outcomes

- We can predict future observations using the regression equation.

$$Y' = b_0 + b_1(X)$$

- Example: predict the score for a student who studies 7 hours

$$Y' = 25.8 + 5.1(7) = 61.5$$

Assumptions of the Linear Regression Model

- Y is a random variable with mean

  $E(Y) = (B_o + B_1 X_{mean})$

- X is known and measured without error

- Values of Y are independent of each other:

  $Cov(Y_i, Y_j) = 0$

- Note: there is no assumption about the distribution of Y (i.e., normally distributed) – Not Yet!

Lecture Outline

- Types of Relationships between Variables
- The Linear Regression Model
- Calculating The Regression Equation
- Regression Analysis Using R
- Goodness of Fit

Solving for the Regression Coefficients

- If there is a relationship between X and Y, then we can use X to <u>predict</u> Y. The predicted value of Y is

  $$Y' = B_o + B_1 X$$

  - The error of this prediction is known as a residual: (Y – Y')
- The "best" estimates of B_o *and* B_1 will give the smallest residuals.
- Because we deal with variance, we want to minimize the squared residuals: $\Sigma(Y-Y')^2$

Principal of Least Squares

Our goal is to estimate parameters β_0, β_1 and σ^2. We discuss the estimation of β_0 and β_1 first. Denote b_0 and b_1 as estimates of β_0 and β_1. The fitted (predicted) response is

$$\hat{Y} = b_0 + b_1 X$$

For observed response Y_i, the fitted (predicted) response is

$$\hat{Y}_i = b_0 + b_1 X_i$$

The discrepancy between observed response Y_i and the the fitted response $\hat{Y}_i$ is $Y_i - \hat{Y}_i$. The overall squared discrepancy is

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

LS principal: find b_0 and b_1 such that Q is minimized

The Normal Equations

Taking derivative of Q with respect to b_0 and b_1 and setting them equal to zero, we have the `normal equations`:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) &= 0 \\ \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) &= 0\end{aligned}$$

Solving above two equations, we have the LS estimates:

$$\boxed{\begin{aligned}b_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ b_0 &= \bar{Y} - b_1 \bar{X}\end{aligned}}$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$

Properties of the LS estimates

- b_o and b_1 are linear functions of $Y_1, \ldots, Y_n$
- b_o and b_1 are <u>unbiased</u> estimates of B_0 and B_1
  - The $E(b_o) = B_{0\ \text{and}}$ and $E(b_1) = B_1$
- Gauss-Markov Theorem
  - The LS estimators b_o and b_1 are unbiased and have minimum variance among all unbiased linear estimators
- The slope and intercept <u>minimize the squared deviations</u> between the observed and predicted values!

Features of the Model

- E(Y) = E(Y') = E($B_o + B_1$X)
  - The mean of the predicted values is the same as the mean of observed values

- E(ε_i) = E(Y – Y') = 0
  - The expected valueof the residuals is zero

A Worked Example of Regression

We ask 5 students to tell us:

- Their year in college (X)
- the number of units they are taking (Y)

Year in School (X)	Number of Units (Y)
1	9
2	12
3	10
4	14
5	15

We want to know:

1. Do units increase with class standing?
2. Is the relationship linear?
3. Describe the relationship

A Worked Example (cont'd)

- Find the regression coefficients
- Plot the regression line
- Predict the number of units for a Junior (Year 3)
- Estimate s^2

A Worked Example Calculations

	Year in College (X)	Number of Units (Y)	X-Xbar	Y-Ybar	(X-Xbar)(Y-Ybar)	(X-mean)^2	(Y-mean)^2
	1	9	-2	-3	6	4	9
	2	12	-1	0	0	1	0
	3	10	0	-2	0	0	4
	4	14	1	2	2	1	4
	5	15	2	3	6	4	9
n	5	5	5	5	5	5	5
Sum	15	60	0	0	14	10	26
Sum of Squares	55	746	10	26	76	34	194
Mean or Mean Square	3	12	0	0	2.8	2	5.2
b1	1.4						
bo	7.8						

Lecture Outline

- Types of Relationships between Variables
- The Linear Regression Model
- Calculating the Regression Equation
- Regression Analysis Using R
- Goodness of Fit

R Statistical Software

- R is terminal software program (you enter commands at a command prompt)
- It is free (download at http://cran.r-project.org)
- It is simple and straightforward (manual at http://cran.r-project.org/doc/manuals/R-intro.html)

Basic R Tasks

- Entering Data

```
> x<-c(10.4, 5.6, 3.1, 6.4,…,21.7)
> y<-c(92, 103, 77, 85,…,112)
```

- Summarizing Data

```
> Summary(x)
> Summary(y)
```

- Regression parameters

```
> lm(y~x)
```

- Full Regression Analysis

```
> fit1<-lm(y~x)
> summary(fit1)
```

Using R to Perform Regression

```
x<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7,
11.1)
> y<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> fit1<-lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.909  -7.280  -1.925   8.355  17.703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
x             5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,     Adjusted R-squared:  0.7028
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06

> plot(x,y,xlab="Hours of Study",ylab="Exam Score")
> abline(fit1)
```

Input X,Y data

Calculate Regression

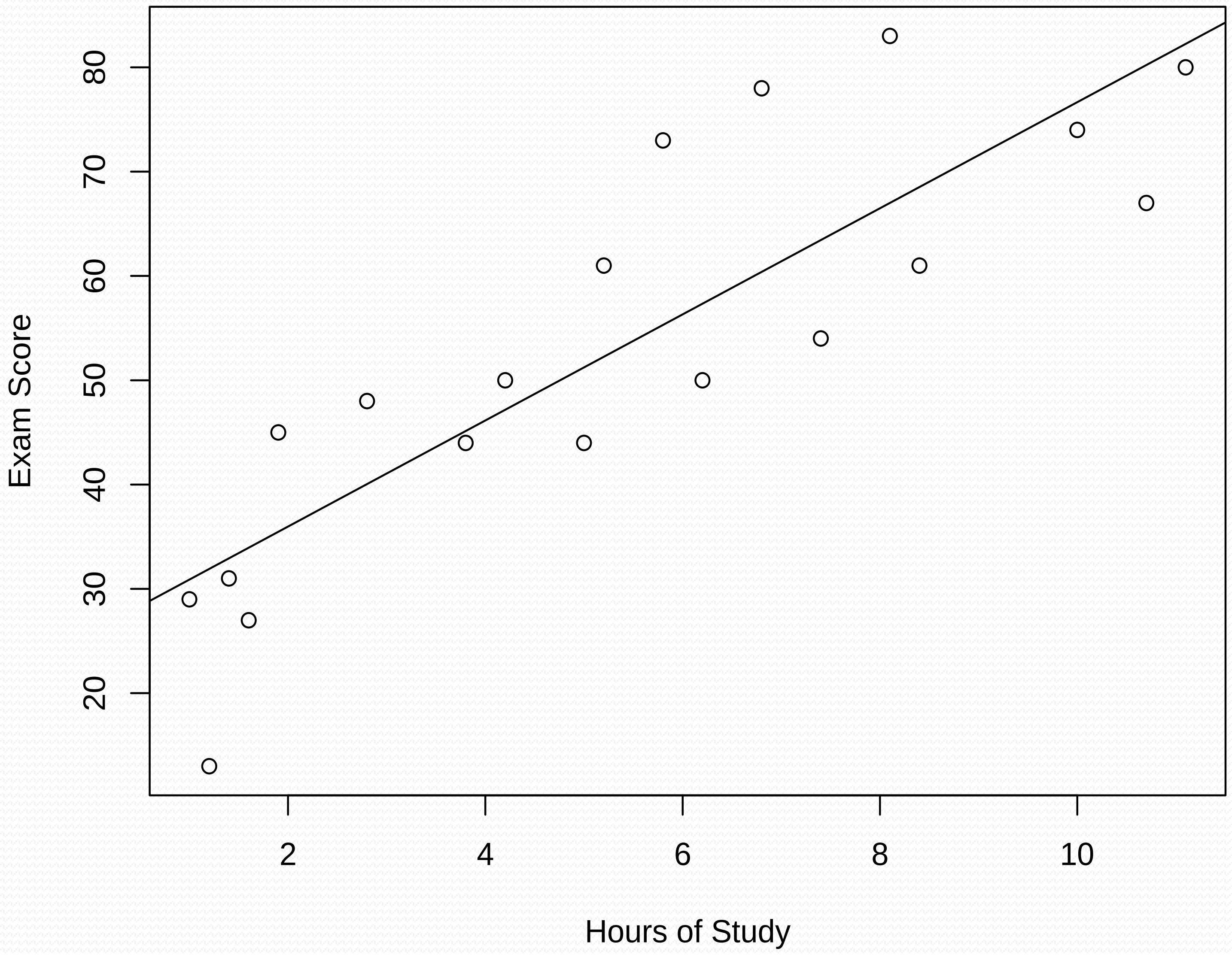
Create scatterplot

Using R to Perform Regression

```
x<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7,
11.1)
> y<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> fit1<-lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.909  -7.280  -1.925   8.355  17.703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
x             5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,     Adjusted R-squared:  0.7028
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06

> plot(x,y,xlab="Hours of Study",ylab="Exam Score")
> abline(fit1)
```

Regression parameters

R Scatterplot with Regression Line



Lecture Outline

- Types of Relationships between Variables
- The Linear Regression Model
- Calculating the Regression Equation
- Regression Analysis Using R
- Goodness of Fit

Error in Regression

- Regression error refers to the difference between the predicted value of Y and the observed values

- If we predict without regression, using the mean, the error equals the total variance of Y

- If we predict using the regression line the error equals the variance of Y around the regression line

Defining Regression Error

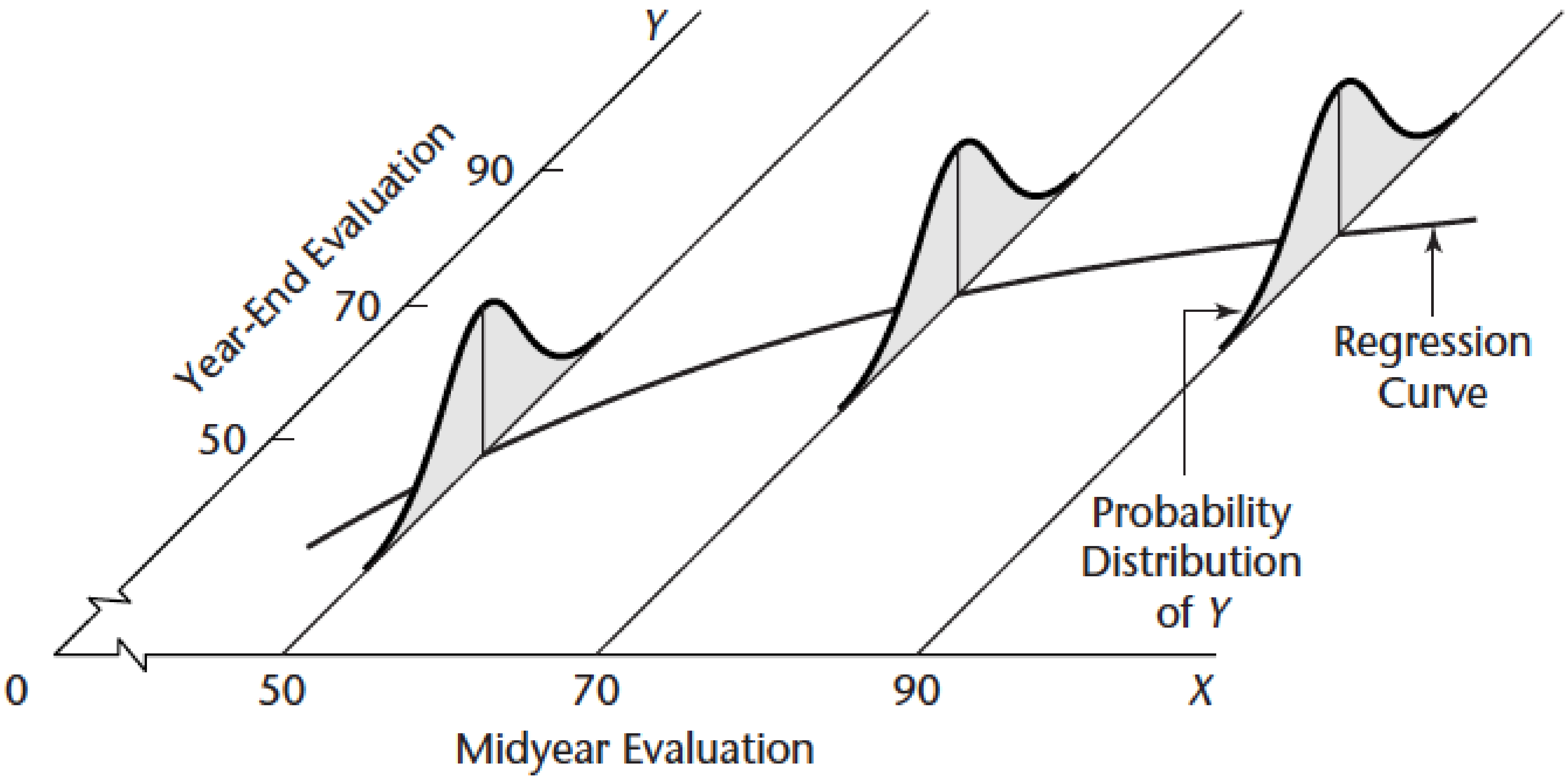
$Y = \beta_0 + \beta_1 X + \varepsilon$ (Normal Regression Equation)

$Y' = \beta_0 + \beta_1 X$ (Fitted Regression Equation)

$Y = Y' + \varepsilon$ (Observed = Predicted + Error)

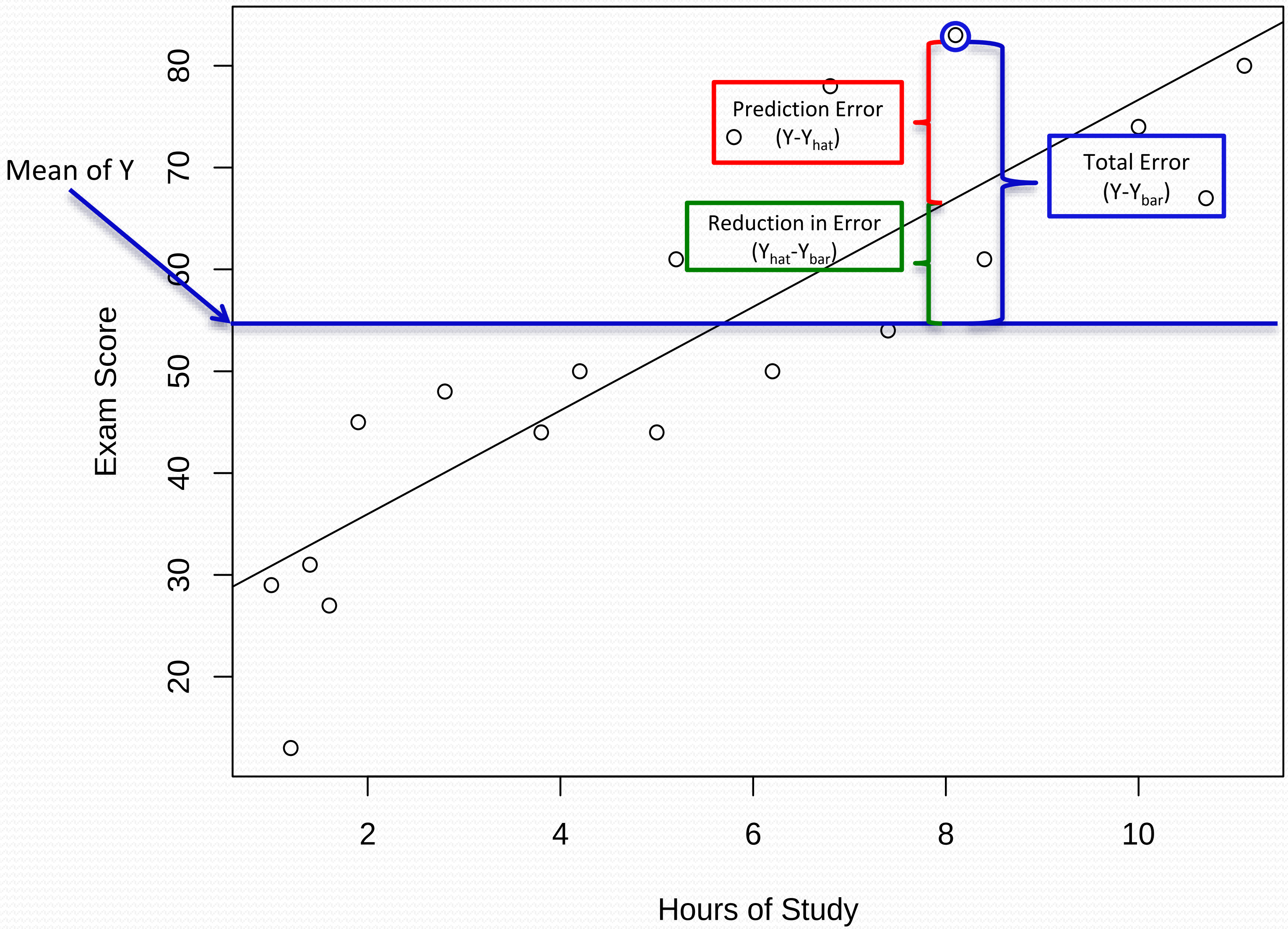
Estimated error (Residual) = $(Y - Y')$

Error around the Regression Line



- There is a distribution of observed values of Y around each predicted value of Y’

Regression Error (Continued)



Residuals and Partitioning

- Each value of Y has a residual

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

- The sum of each of these deviations equals zero

Variance of Y, Y' and (Y-Y')

- The total variance of Y

$$s_Y^2 = \frac{\sum (Y_i - \overline{Y})^2}{n - 1}$$

- The variance of the residuals (Y-Y')

$$s_{Y-\hat{Y}}^2 = \frac{\sum (Y_i - \hat{Y})^2}{n - 2}$$

- Variance = Mean Square (MS) = SS/df
- $E[MS_E]$ = population residual variance

Goodness of Fit

How well does the regression fit the data?

- Use the variance of residuals to describe the fit
  - but how do we know if this variance is large or small?
- R^2 (Coefficient of Determination)

$$R^2 = \frac{\sum (Y' - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{SSReg}{SSTotal} = 1 - \frac{SSError}{SSTotal}$$

Coefficient of Determination (R^2)

- R^2 indicates how much of the variance in Y can be explained by adding a predictor to the model
  - $0 \leq R^2 \leq 1$
  - R^2 can be derived by squaring the correlation coefficient of X,Y (i.e., $R^2 = r^2$)
  - However, we are not covering r for the bivariate case

- To calculate R^2, you need to know SSTO and SSE

- R^2 appears in the R output

- For Example #2, we can state: “72% of the variance in Exam Score can be explained by knowing Hours of Study”

Coefficient of Determination in R

```
> x<-c(1, 1.2, 1.4, 1.6, 1.9, 2.8, 3.8, 4.2, 5, 5.2, 5.8, 6.2, 6.8, 7.4, 8.1, 8.4, 10, 10.7,
11.1)
> y<-c(29, 13, 31, 27, 45, 48, 44, 50, 44, 61, 73, 50, 78, 54, 83, 61, 74, 67, 80)
> fit1<-lm(y~x)
> summary(fit1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.909  -7.280  -1.925   8.355  17.703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8073     4.8378   5.335 5.48e-05 ***
x             5.0844     0.7703   6.601 4.50e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.77 on 17 degrees of freedom
Multiple R-squared:  0.7193,    Adjusted R-squared:  0.7028
F-statistic: 43.57 on 1 and 17 DF,  p-value: 4.497e-06

> plot(x,y,xlab="Hours of Study",ylab="Exam Score")
> abline(fit1)
```

Coefficient of Determination