STATS 210P

Lecture 2

Sevan Koko Gulesserian

University of California, Irvine

Simple Linear Regression

The goal (for now) of a simple linear regression is to assess the relationship (if any) between two quantitative variables.

- Measure 2 quantitative variables on each sampled unit (e.g. subject, school, car, etc.).
- How strongly, if at all, are they related to each other?
- Given a new unit not in the sample, can we predict the value of one of the quantitative variables given the other?

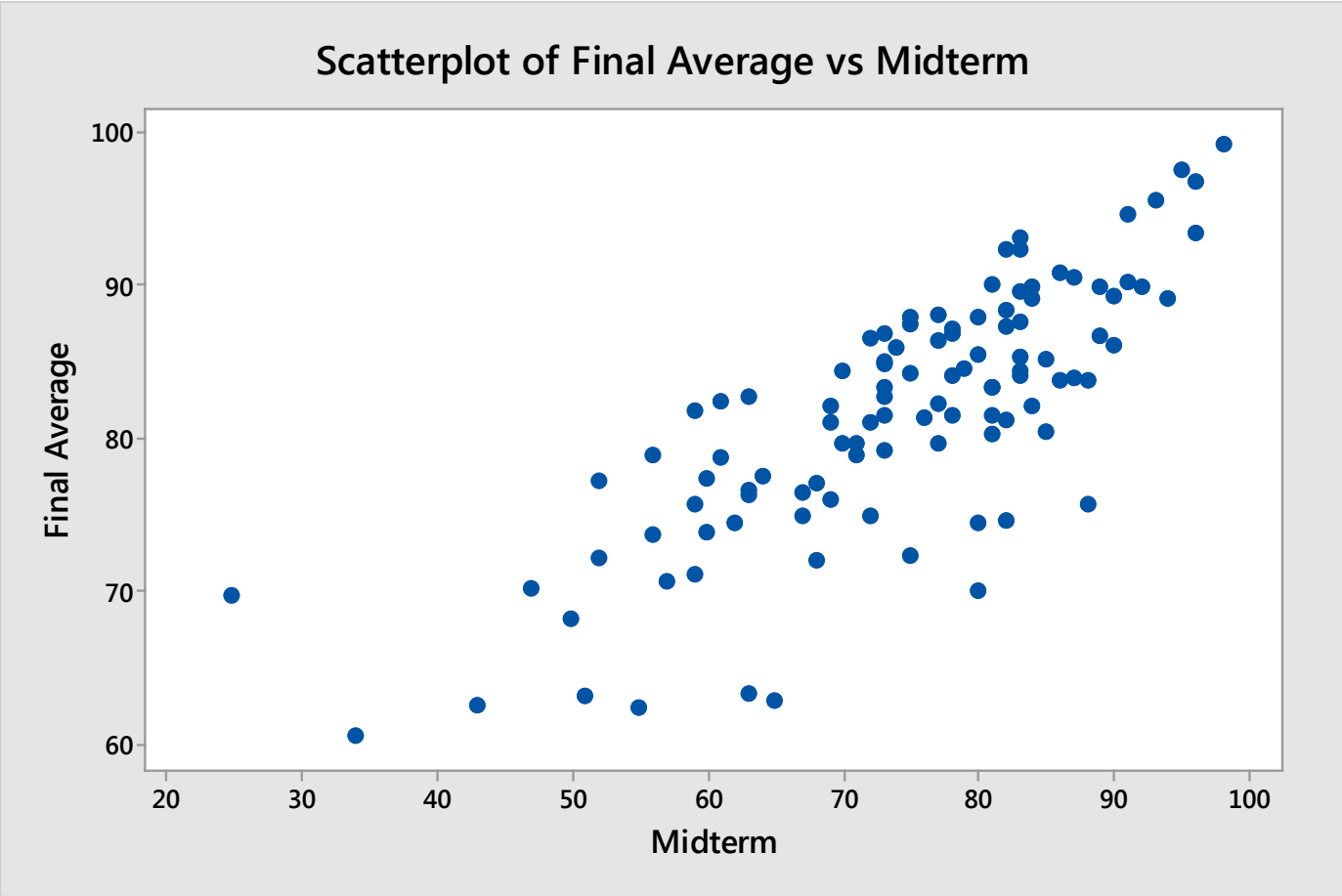
Simple Linear Regression

Example: After receiving a midterm score, how well can we predict the final average for the course?
Data: Previous statistics class where both midterm and final average are observed for each student.

- Let Y= Final average, be the response variable.
- Let X=Midterm score, be the explanatory variable.

Simple Linear Regression

A scatterplot plots each (X,Y) pair for all sampled units (students).



Quick Algebra Review

The equation for a linear line is: $Y = \beta_0 + \beta_1 X$.

- Y is the response variable.
- X is the explanatory variable.
- β_0 is the intercept. It is the value of Y when X=0.
- β_1 is the slope. It is the increases in Y when X increase by 1 unit.

Quick Algebra Review

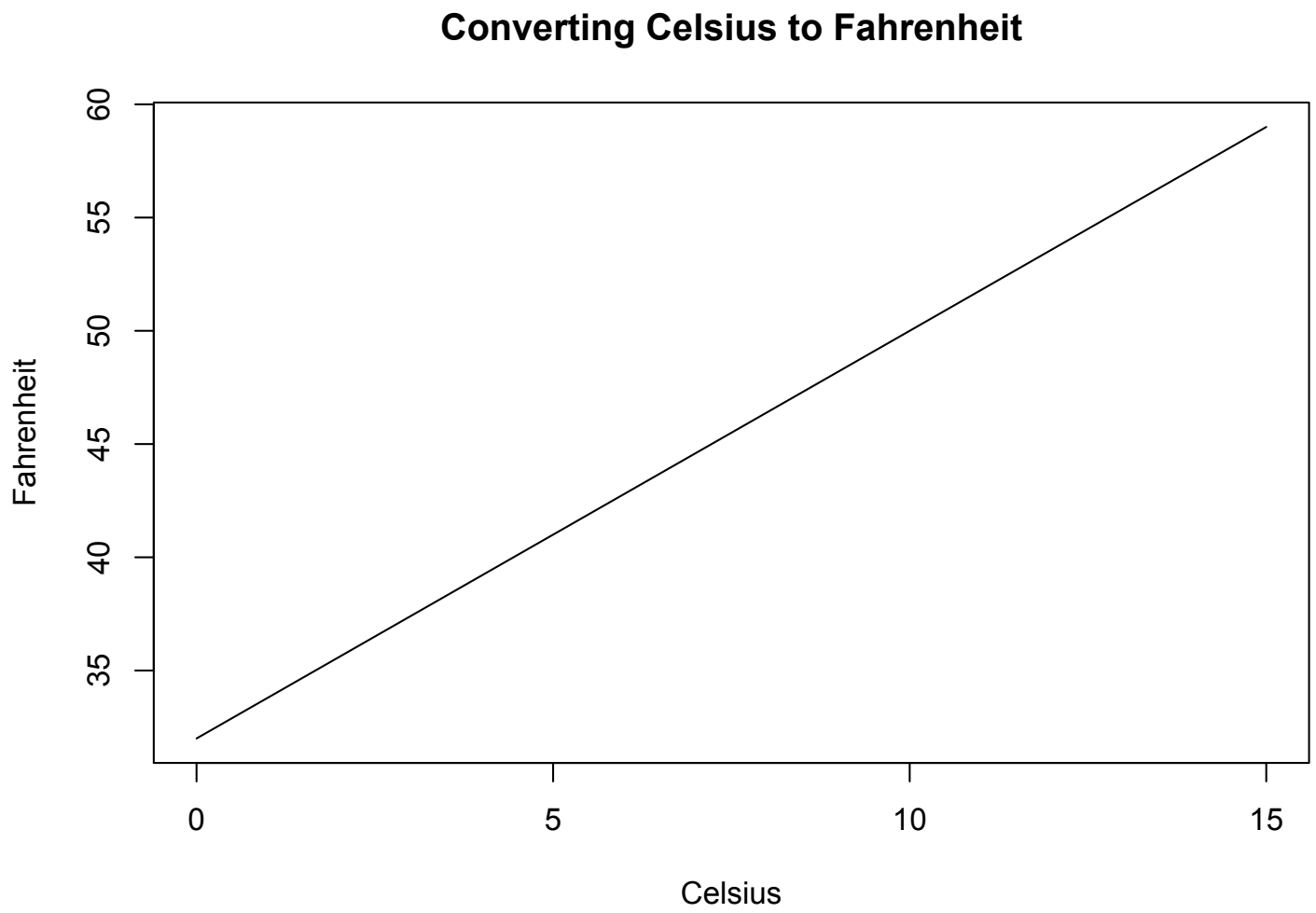
An example of a deterministic relationship is the equation that converts temperature from celsius (C) to fahrenheit (F) units.

The equation is: $F = 32 + 1.8C$.

- To convert a given celsius measurement to fahrenheit, just plug in the celsius value in for C.
- When C=0, then F=32+1.8*0=32.
- When C=10, then F=32+1.8*10=50.
- As C increase by 1 unit, then F increase by 1.8 units.

Quick Algebra Review

Plot of the equation $F = 32 + 1.8C$.



Quick Algebra Review

This has been an example of a *deterministic* relationship.

If you know X, then you can obtain the exact value of Y.

- In a statistical relationship, there is *variation* in the possible values of Y for each value of X.
- If you know the value of X, then you can obtain an *average* or *approximate* value for Y.
  - For example, two students can have the same shoe size (X), but have different values for their heights (Y).
  - Thus, what we can do is for a given value of X, we can only predict what Y is going to be, or obtain an average or approximate value for Y.

Some more examples of statistical relationships.

- The yearly income of a worker (Y) and the number of years of education obtained (X).
- The height, in inches, of a someone (Y) and the average height of their parents (X).
- How far away (in feet) can a driver read a sign (Y) and the age of the driver (X).

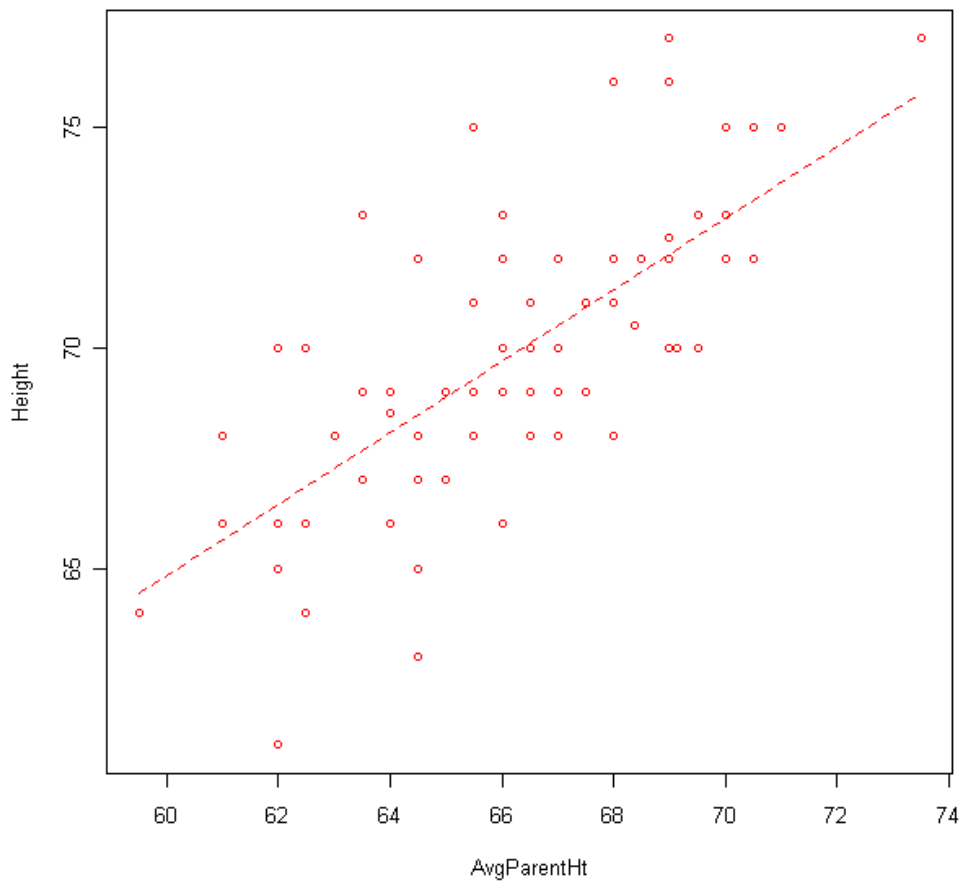
Simple Linear Regression

Relating two quantitative variables.

- Graph: Create a scatter plot to visually see the relationship.
- Regression equation: To describe the "best" straight line through the data, and to predict Y given a value of X.
- Correlation coefficient (ρ): Assess the strength and direction of the linear relationship.

Scatterplot

- 1. Create axis' with the appropriate ranges for X (the horizontal axis) and Y (the vertical axis).
- 2. Plot a dot for each (X,Y) pair in the data.



Scatterplot

What to look for in the scatterplot.

- 1. Is the average pattern linear, curved (quadratic), random, etc.?
- 2. Is the trend of the association positive or negative? In general, does Y increase as X increases (positive), or does Y decrease and X increases (negative).
- 3. How spread out are the Y-values at each value of X (the strength of the relationship).
- 4. Are there any outliers? E.g. any unusual combinations of (X,Y).

For the scatterplot in the previous slide (someone's height plotted against the average of their parents height), the average pattern looks linear, the association is positive, the Y values are fairly spread out at each value of X, and there does not seem to be any outliers.

Simple Linear Regression

The true simple linear regression equation that defines the statistical relationship between Y and X in the population is:

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

- The part $\beta_0 + \beta_1 X$ can be viewed as the model.
  - β_0 and β_1 are unknown and will be estimated using only the observed data.
- ε is the error. When these errors are not all equal to 0, then it is not a deterministic relationship between Y and X.
- For example when two people have different heights, but have the same average height for their parents, then ε will capture the statistical relationship aspect between X and Y.

Simple Linear Regression

The basic idea of the simple linear regression is to find the best fitting line that will:

- 1. *Estimate* the *average* value of Y at a given value of X.
- 2. *Predict* Y when X is known but the true Y is not.

The *least squares regression* line is the best straight line (linear) for the data.
Notation of the estimated least squares regression line is:

$$\widehat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated values of the unknown β_0 and β_1.

- We will denote estimates with the hat notation, $\widehat{\ }$.
- Say A is the population parameter we are interested in. Then $\widehat{A}$ is the estimate we obtain from the sample for the population parameter A.

Simple Linear Regression

A regression equation describes how the mean or expectation of the response variable relates to specific values of the predictor variable(s).

- The simple linear regression equation describes the mean of the response variable as a straight line function of a single predictor variable. This could be written as:

$$E(Y|X) = \beta_0 + \beta_1 X$$

- The simple linear regression population model can be written as:

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

where ε is the random error.

Simple Linear Regression

For individuals in the larger population from which the sample has been taken, the simple linear population regression model can be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

for $i = 1, 2, ..., n$.

Where n is the sample size, Y_i is the response value for the i-th observation, X_i is the covariate/predictor value for the i-th observation, and ε_i is the error for the i-th observation.

And we will call the β's (here we have β_0 and β_1) the Beta coefficients.

Linear Regression Equation

Continuing the example of someones height (Y) and the average height of their parents (X).

Say the regression equation is (we will learn how to get this equation soon enough):

$$\hat{Y} = 16.3 + 0.809X.$$

- If X=68 inches, then $\hat{Y} = 16.3 + 0.809 * 68 = 71.3$ inches.
  - This can be interpreted as the estimate of the average height of all people whose parents' average height is 68 inches.
  - Or can be interpreted as the predicted height of someone whose parent's average height is 68 inches.

Linear Regression Equation

Continuing the example of someones height (Y) and the average height of their parents (X). Say the regression equation is:

$$\hat{Y} = 16.3 + 0.809X.$$

- The interpretation of the intercept 16.3 is as follows. 16.3 inches is the predicted height of someone whose parents' average height is 0.
  - The intercept has a meaningful interpretation when X=0 is reasonable. In this case, X=0 is not reasonable as that would mean the average height of the parent's is 0 inches.
  - An example when X=0 is reasonable is when X is the number of years of secondary (beyond high school) education someone has. X=0 here means someone did not go to college, which is reasonable.

Linear Regression Equation

Slope interpretation.

- The interpretation of the slope 0.809 is as follows:
  When the average height of the parents (X) increases by 1 unit (that is to say increases 1 inch), we **expect** the height of the child to increase 0.809 inches.
- Note: We do NOT say that as X increases that Y **will** increase by a certain amount, because that would imply a deterministic relationship.

Linear Regression Equation

- Directly following this interpretation, we can compute the expected difference in heights of someone given a change in their parents heights.
  - What is the expected change in height of someone whose parents' average height changes/increases by 10 inches?
  - If a single unit change in X leads to a 0.809 inch change in the predicted Y, then a 10 unit change in X will lead to a 0.809*10=8.09 inch change in the predicted Y.
  - If it was a 8 inch change in X, then the predicted Y will change by 8*0.809=6.472.

Linear Regression Equation: Residuals

The linear equation at the population level is:

$$Y = \beta_0 + \beta_1 X + \varepsilon = \text{Model+Error}$$

- Therefore we can write the following:

$$\varepsilon = Y - (\beta_0 + \beta_1 X)$$

which is to say Error=Y-Model

Linear Regression Equation: Residuals

At the estimated sample level:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e = \text{Predicted value+Residual}$$

- Therefore: Residual=Y−Predicted value
- Predicted value is depicted as $\hat{Y}$.
- Thus: Residual=$Y - \hat{Y}$ (e.g. Observed Y−Predicted Y).
- Let r_i (or e_i) denote the residual for the i-th sampling unit, where we have a sample size of n (that is to say i=1,2,3,...,n) .
- $r_i = Y_i - \hat{Y}_i$, the observed value for the i-th sampling unit minus it's predicted value.
- Note that the residuals, r's are the empirical version of the true errors from the population model ε.

Linear Regression Equation: Residuals

Return to the example of someone's height and the average height of their parents.
Suppose someone's parents average height is 66 inches and the person is 69 inches tall.

- Observed data is Y=69 and X=66.
- $\hat{Y} = 16.3 + 0.809 * 66 = 69.7$. The predicted height for this person is 69.7 inches.
- Residual=69-69.7= -0.7 inches (overestimated the height by 0.7 inches).
- Y=Predicted value+Residual $\Rightarrow$ 69=69.7+(-0.7).

Can do this for each Y observation in the dataset, and thus have n many r_i's (for i=1,2,...,n).

Simple Linear Regression: Regression Equation

"Fitting" of a line to the data points:

- The line $\hat{\beta}_0 + \hat{\beta}_1 X$ obtained by using the least squares method is the best fitting line through the data in the scatterplot.
- What defines the "best" line is the line that minimizes the squared differences of $Y - \widehat{Y} = Y - (\hat{\beta}_0 + \hat{\beta}_1 X)$ (so what we focus on are the squared residuals, $(Y - (\hat{\beta}_0 + \hat{\beta}_1 X))^2$.
- That is why the term "least squares" is used when describing the method used to obtain the linear regression equation, because our goal is to minimize the sum of these squared residuals across all of our sampling units.

Simple Linear Regression: Regression Equation

Main idea: Minimize how far off the predicted value of Y is (using the regression equation) from the actual Y, across the observations, via what is called an *objective function*.

Compute the residual for each observation in the dataset (Residual=Y$-\hat{Y}$ for all Y in the data).

Thus we will have $r_i = Y_i - \widehat{Y}_i$ for i=1,2,3,...,n.

- The *least squares regression line* is the unique line that minimizes the sum of squared residuals.
- This sum is known as SSE (sum of squared errors) but is also called SSR (sum of squared residuals).
- It derives the values for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the following objective function:

  $$\text{SSE} = \sum_{all\ data} \text{residuals}^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

  - Notationally, $(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} r_i^2$

Simple Linear Regression: Regression Equation Estimation

We need to minimize our objective function (Q) with respect to the unknown beta coefficients. That is to say we take partial derivatives of the objective function with respect to β_0 and β_1, set those derivatives to 0, and solve for β_0 and β_1 respectively:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

$$-2\sum(Y_i - b_0 - b_1 X_i) = 0$$

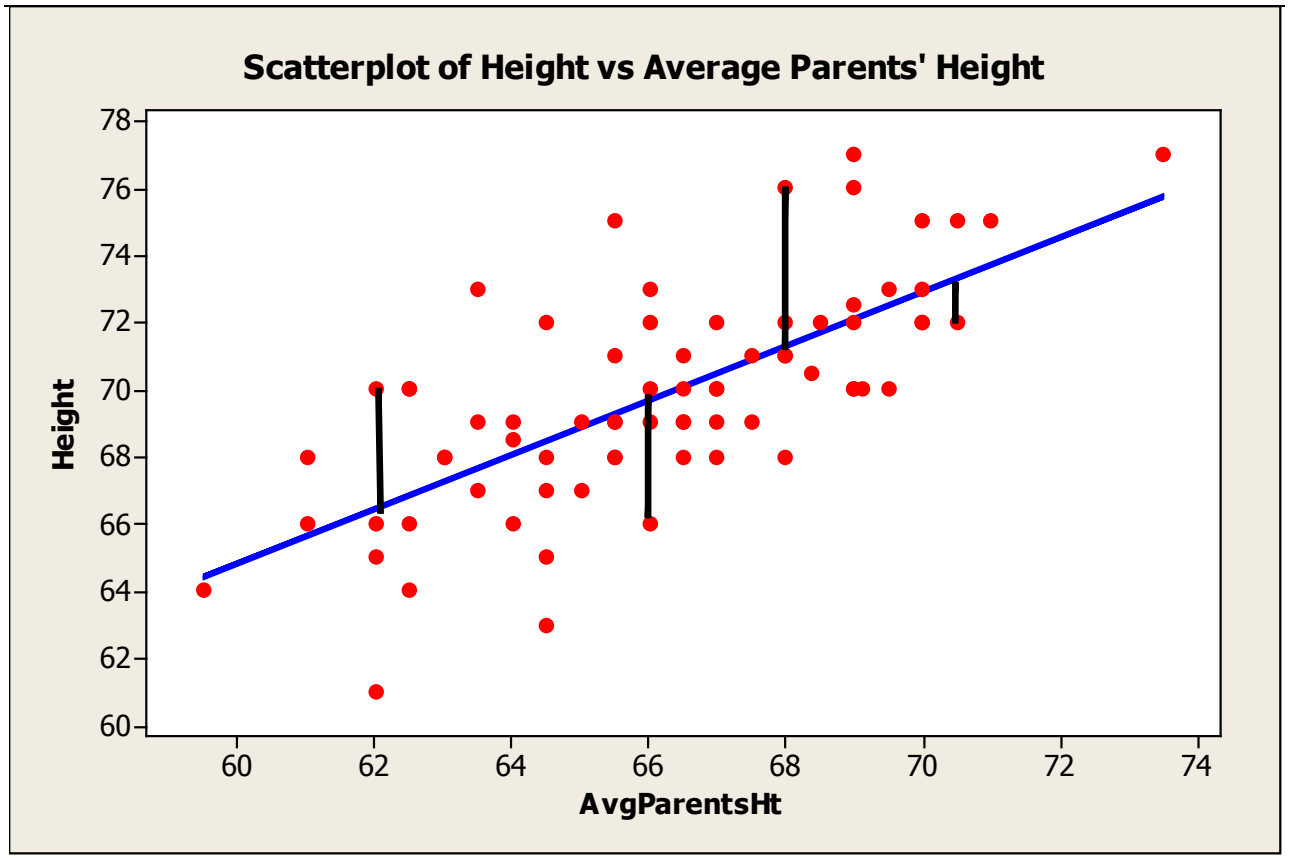
$$-2\sum X_i(Y_i - b_0 - b_1 X_i) = 0$$

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

We will use the notation $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$. And we can go as far as doing a second partial derivative test to show that the solutions to our estimates are minimizers.

Simple Linear Regression: Regression Equation



The black line depicts the residual (shown for 4 observations). The blue line (the linear regression line) will minimize the squared sum of all these black lines across all of the data points.

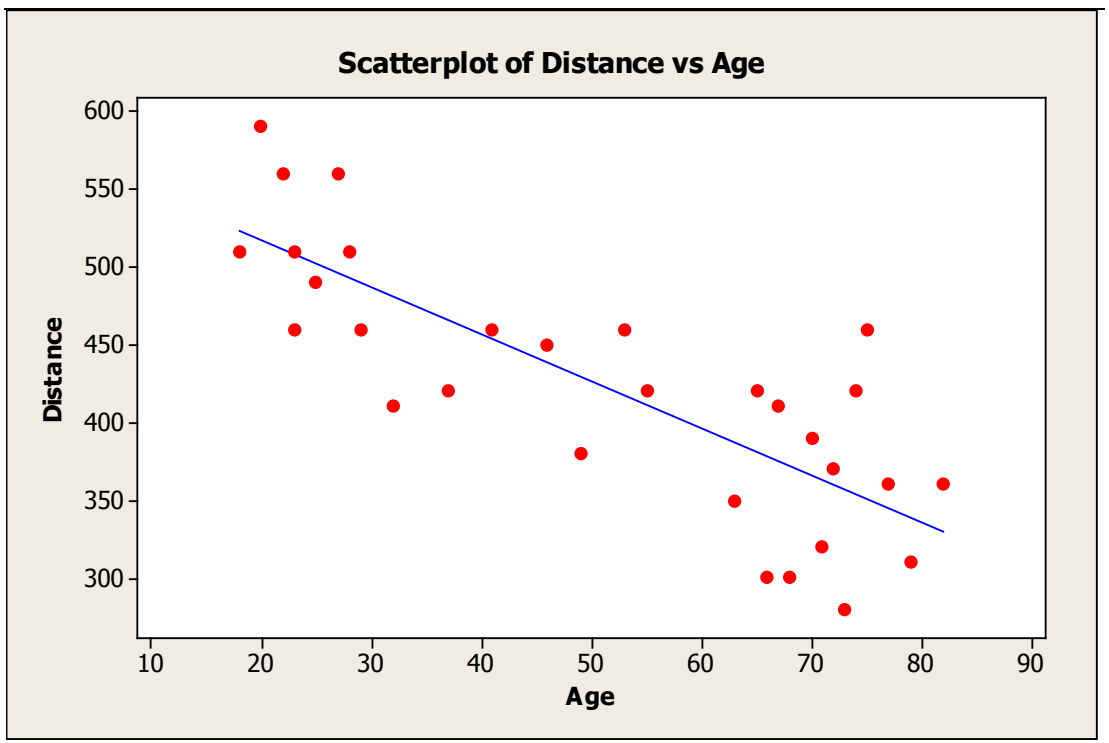
Simple Linear Regression: Regression Equation

- For this example, SSE = 376.9 (sum of squared errors/residuals).
- Given that the dataset contained 73 observations, this means that the average squared error per observation was 5.16.
- Taking square root, on average the predicted value of Y was off by 2.25 inches from the true Y.

Simple Linear Regression: Regression Equation

Another example: A study to see if the distance at which a driver could read a street sign at night changes with age.

- Data has $n = 30$ observations (X,Y pairs). X=age and Y=distance in feet the sign can be read.
- The scatterplot is below.



Simple Linear Regression: Regression Equation

The estimated linear regression equation is $\hat{Y} = 577 - 3X$.

- This represents a negative association (look at the sign of the slope).
  - As age (X) increases by 1 year, the predicted distance the person can read the sign decreases by 3 feet.
- For someone who is 50 years old, the predicted distance they can read the sign is $577 - 3 * 50 = 427$ feet.

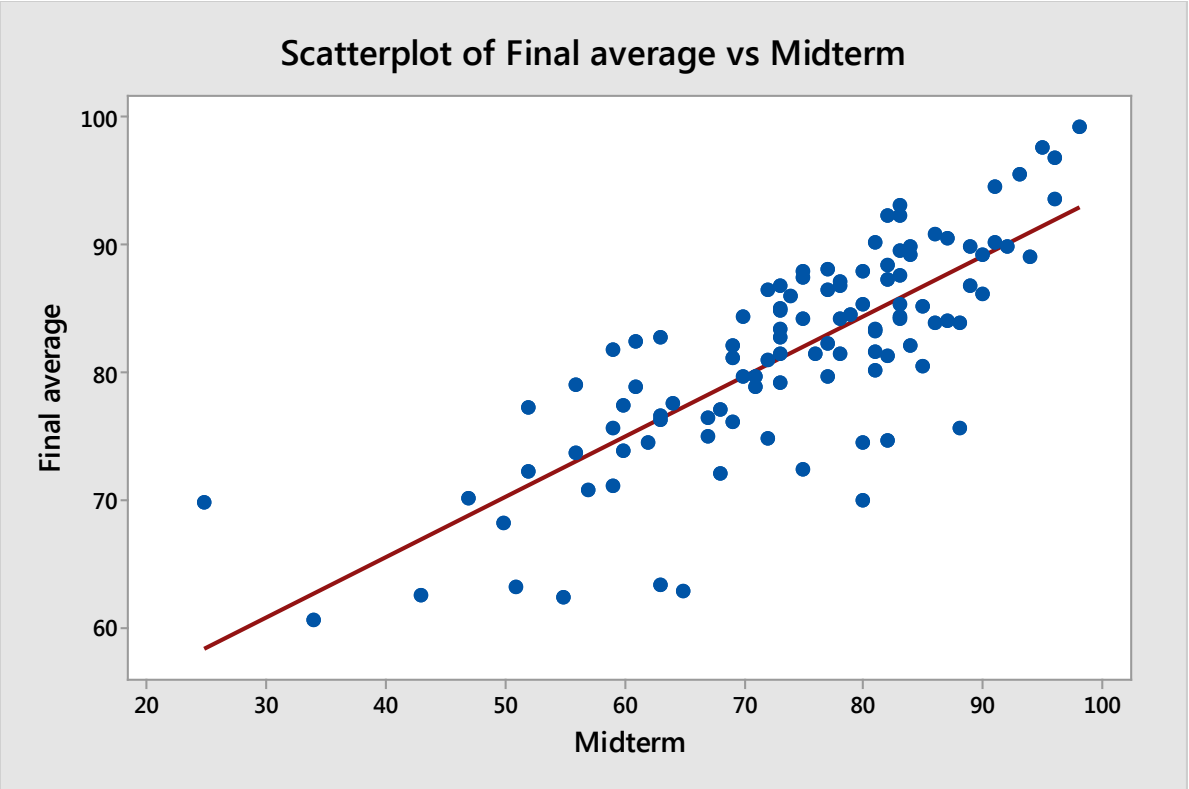
Simple Linear Regression: Regression Equation

The linear regression equation is $\hat{Y} = 577 - 3X$.

- Given two people who differ in age by 60 years, the predicted difference (in feet) of the distances they can read the sign is 3*60=180 feet (the older person will be predicted to read the sign a distance of 180 feet shorter than the younger person).
- Does the intercept have a reasonable interpretation? No, since X is age and a driver being 0 years old is not reasonable.

Simple Linear Regression: Regression Equation

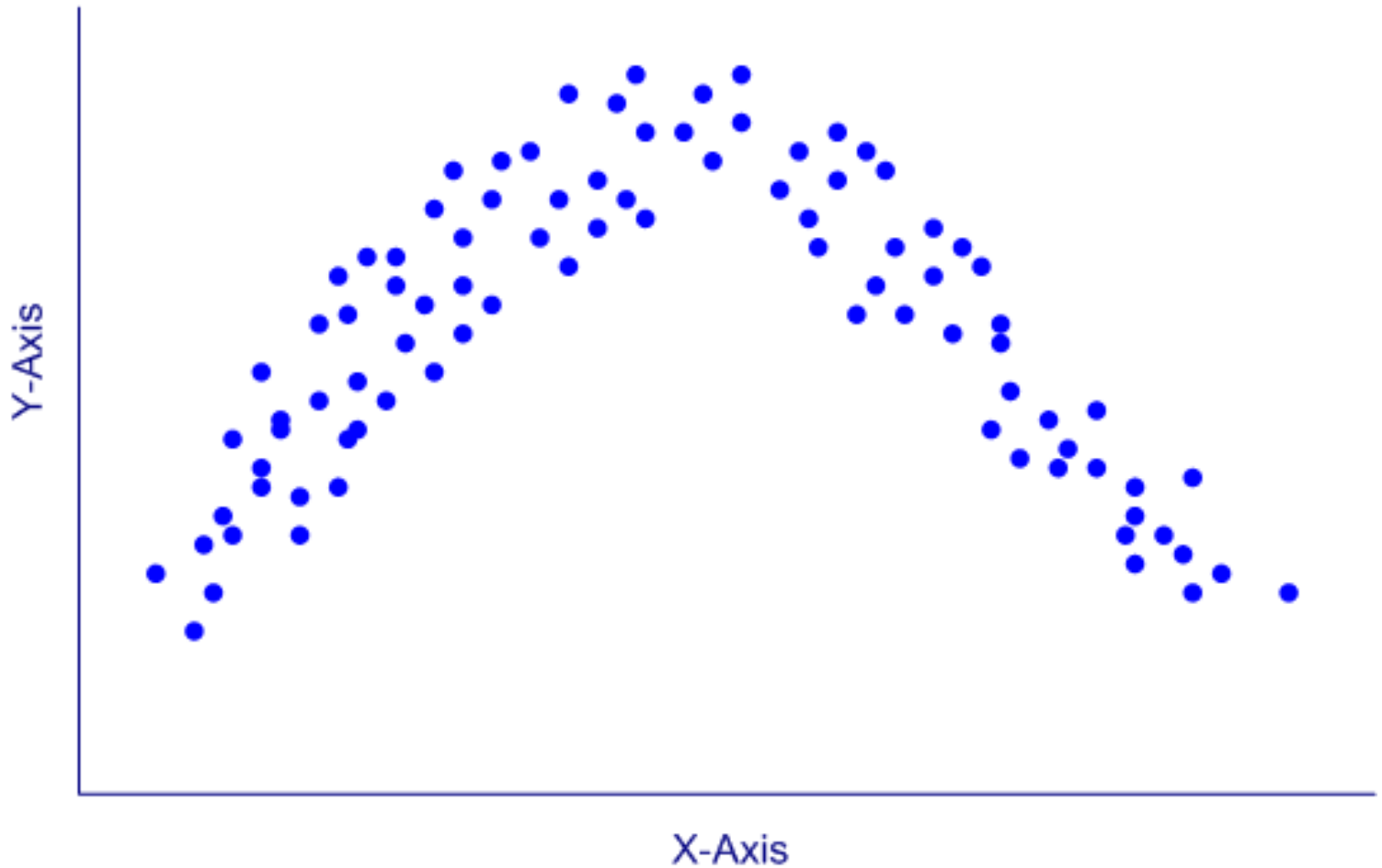
Another example: Predicting final average from midterm score.



- Regression equation is $\hat{Y} = 46.45 + 0.4744X$
  - This is a positive association. As midterm score increases, the predicted value of the final average increase by 0.4744 points.
  - Does the intercept have a reasonable interpretation?

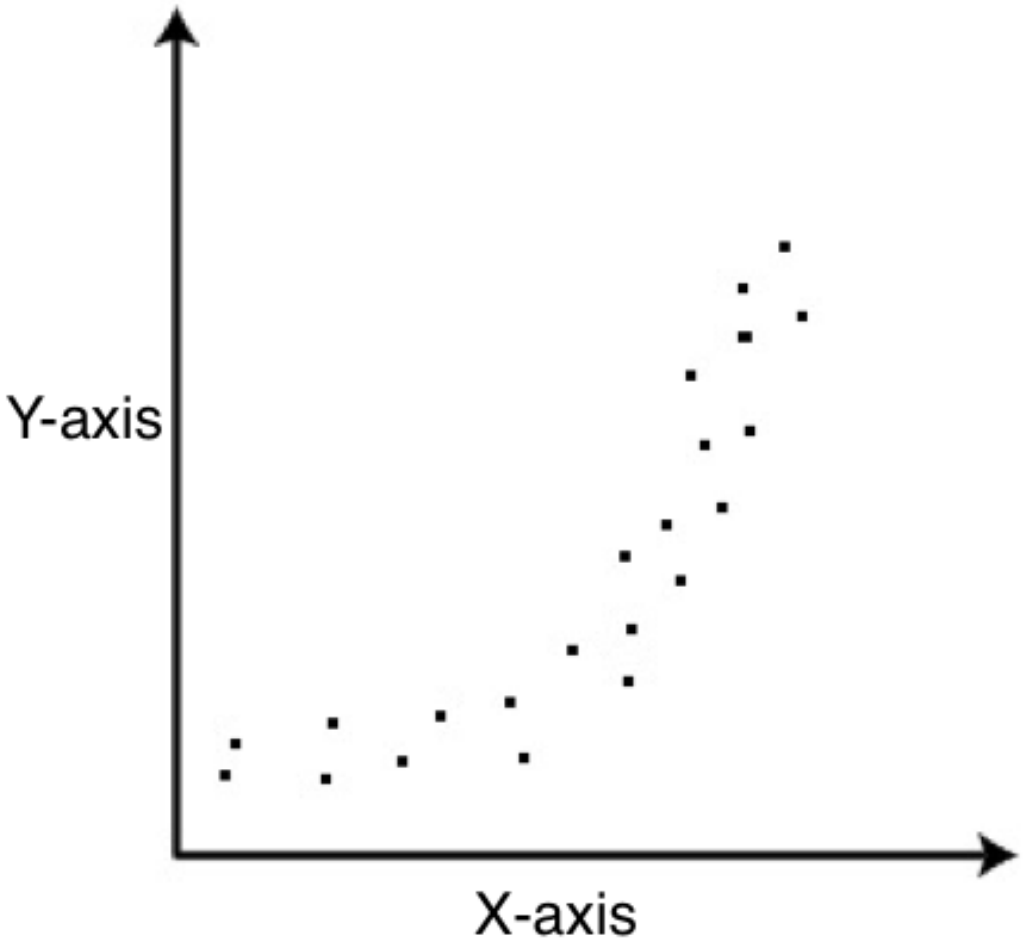
Simple Linear Regression: Non Linear Scatterplot

Example of a non-linear relationship.



Simple Linear Regression: Non Linear Scatterplot

Another example of a non-linear relationship.



Simple Linear Regression: Non Linear Scatterplot

- Can see that fitting a single linear line will not fit the data points well in the previous scatter plot.
- Also, the correlation coefficient does not have any useful meaning in these examples (since it measures the degree of linear association between X and Y).
- Will present a few extensions that can deal with such type of data later on in the course.

Simple Linear Regression: Notation

Notation review.

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}.\end{aligned}$$

Simple Linear Regression: Notation

Notation review.

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Simple Linear Regression: Notation

Notation review (let $e_i = r_i = Y_i - \hat{Y}_i$).

- $\sum_{i=1}^{n} e_i = 0$,
- $\sum_{i=1}^{n} e_i^2$ is a minimum,
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$,
- $\sum_{i=1}^{n} x_i e_i = 0$,
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

Simple Linear Regression: Notation

Notation review.

$$\begin{aligned} \mathrm{Cor}(X, Y) = \rho &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \end{aligned}$$