STA255: Statistical Theory

Chapter 11: Linear Models

Summer 2017

Regression Analysis

- One is interested in the relationship between a single variable (called response, dependent variable, output, etc.) and a single variable or a set of variables (called predictors, explanatory variables, independent variables, inputs, regressors, etc.), described by an unknown mathematical function (called a regression function).

- One predictor; Simple linear regression; Two or more predictors; Multiple regression

Regression analysis

- Variables in regression analysis:
  - The response is a continuous variable (e.g., body weight).
  - The predictors can be continuous (e.g., temperature), ordinal (e.g., letter grade), and nominal (e.g., citizenship).

- The objectives of regression analysis:
  - Describe the relationship between the response and predictors.
  - Predict the observations of the response given predictors.
  - Assess the effects of the predictors on the response.

Some Regression Applications

- Predicting the monthly sales of a company.
- Determining the factors that influence energy consumption in a detergent plant.
- Measuring the volatility of financial securities.
- Determining the factors affect the salary of employees in a company.
- Measuring the fairness of CEO compensation.

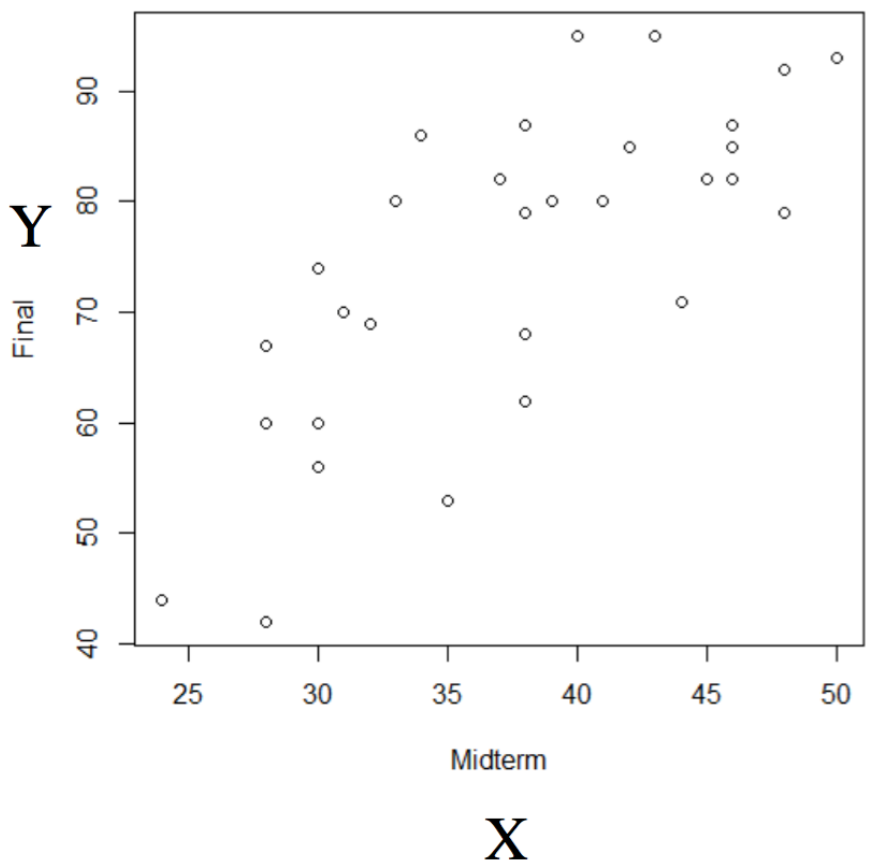
Example: One predictor (SLR)

Can midterm exam score (out of 50) be used to predict the final exam score (out of 100)?

Exam Data: 31 observations

Student	Midterm	Final
Katelin	35	53
Phoenix	38	68
Maria	50	93
Adam	39	80
Michelle	37	82
Allyssa	46	87
Chrissy	32	69
Jessica	24	44
Ryan	48	79
Gabe	41	80

Scatterplot



Examining a scatterplot:

- In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.
- You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.
- An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.
- **Clusters** in a graph suggest that the data describe several distinct kinds of individuals.

Examining a scatterplot:

- Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.
- Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa.
- The **strength** of a relationship in a scatterplot is determined by how closely the points follow a clear form.

Correlation:

- We say a linear relationship is strong if the points lie close to a straight line, and weak if they are widely scattered about a line.
- We use correlation to measure the relationship.

Correlation:

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.

Suppose that we have data on variables x and y for n individuals. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

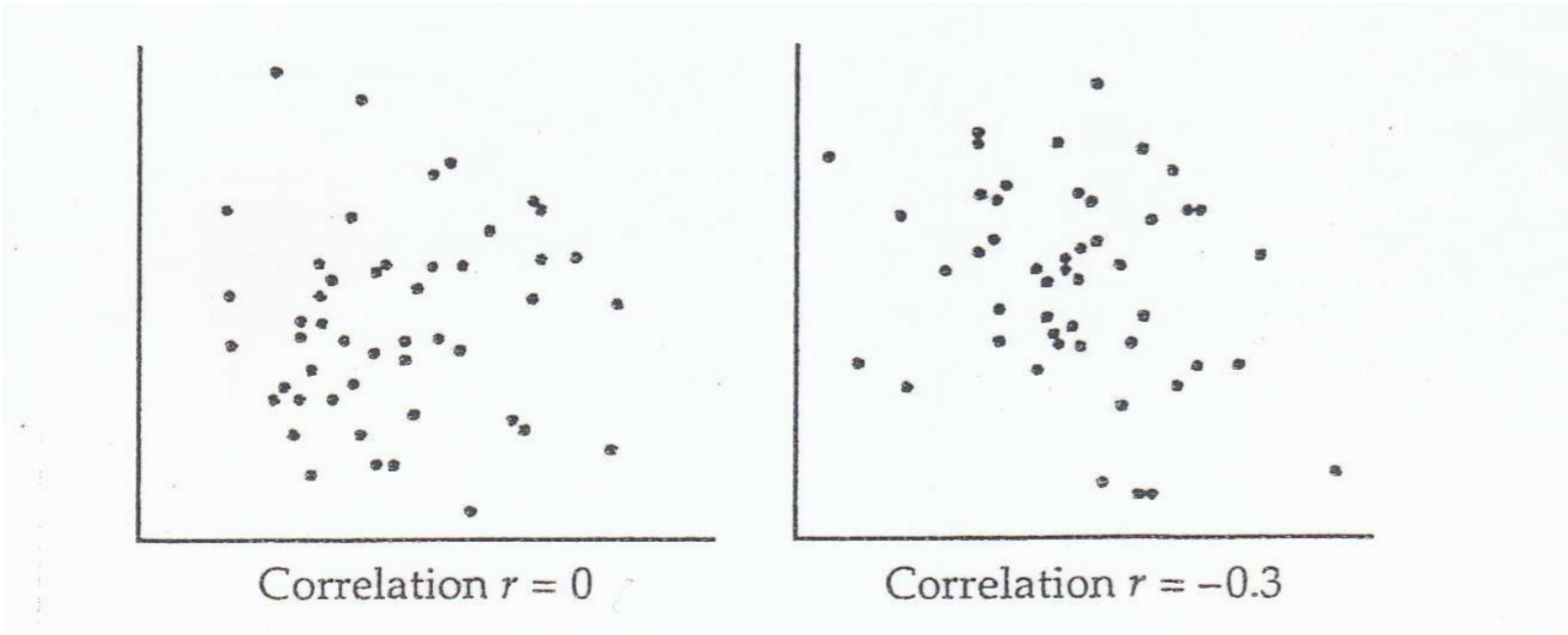
$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) .$$

Properties of Correlation:

- Correlation makes no use of the distinction between explanatory and response variables.
- Correlation requires that both variables be quantitative.
- The correlation r has no unit of measurement.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation is always a number between -1 and 1. Values near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 toward either -1 or 1.
- Values $r = \pm 1$ occur only when the points in a scatterplot lie exactly along a straight line.
- Correlation measures the strength of only the linear relationship between two variables.
- The correlation is not resistant: r is strongly affected by outliers.

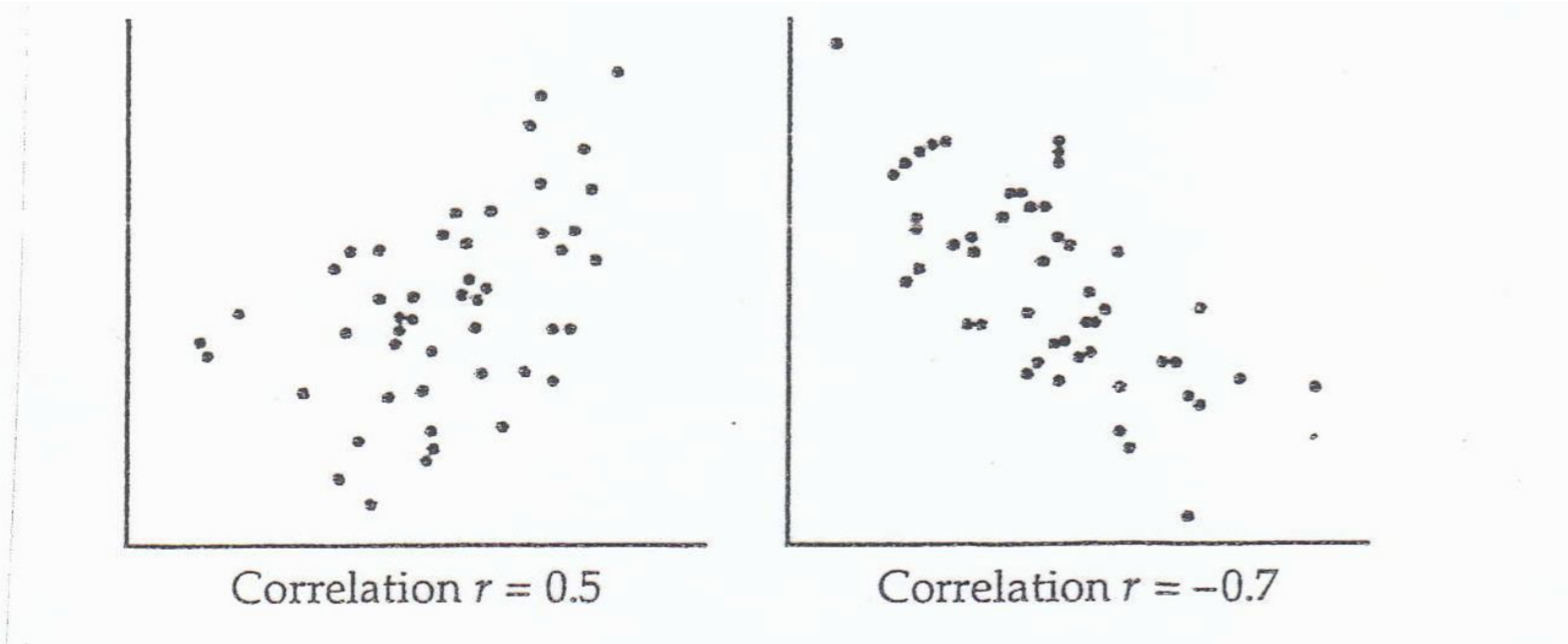
Properties of Correlation:

Here is how correlation r measures the direction and strength of a linear association:



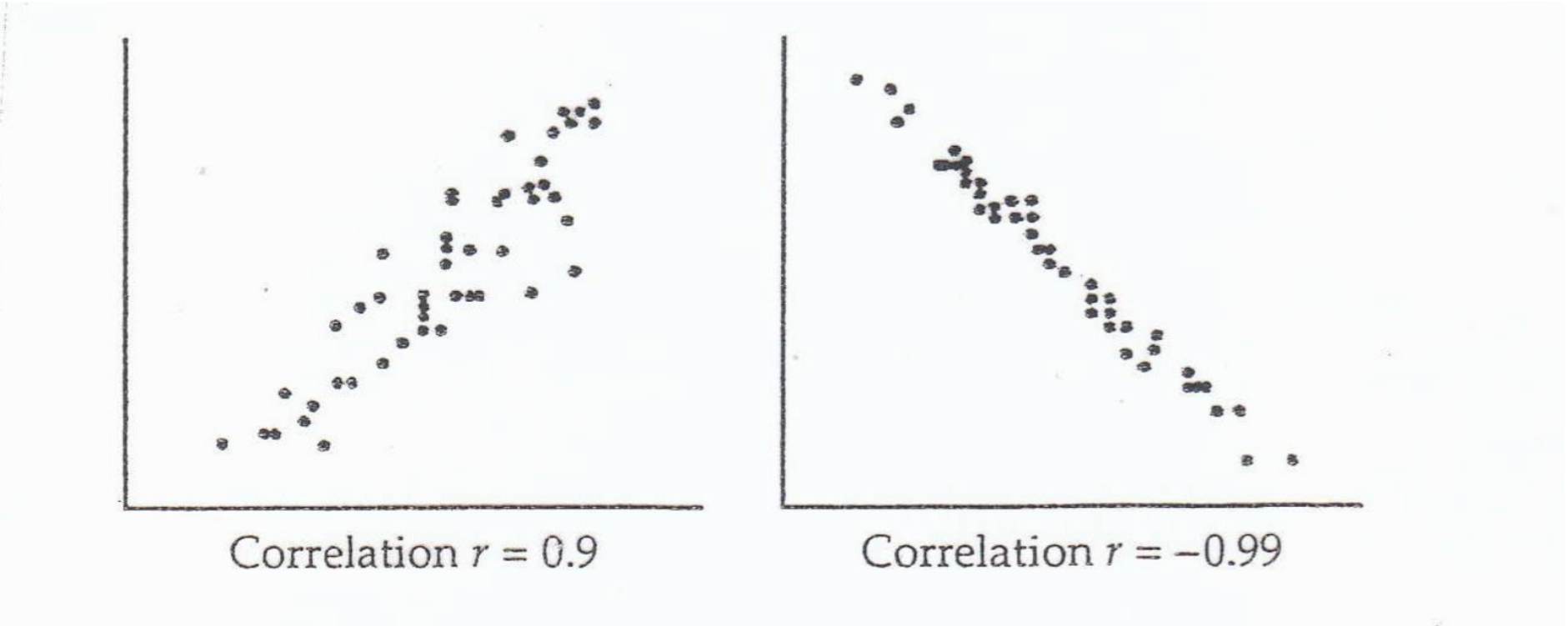
Properties of Correlation:

Here is how correlation r measures the direction and strength of a linear association:



Properties of Correlation:

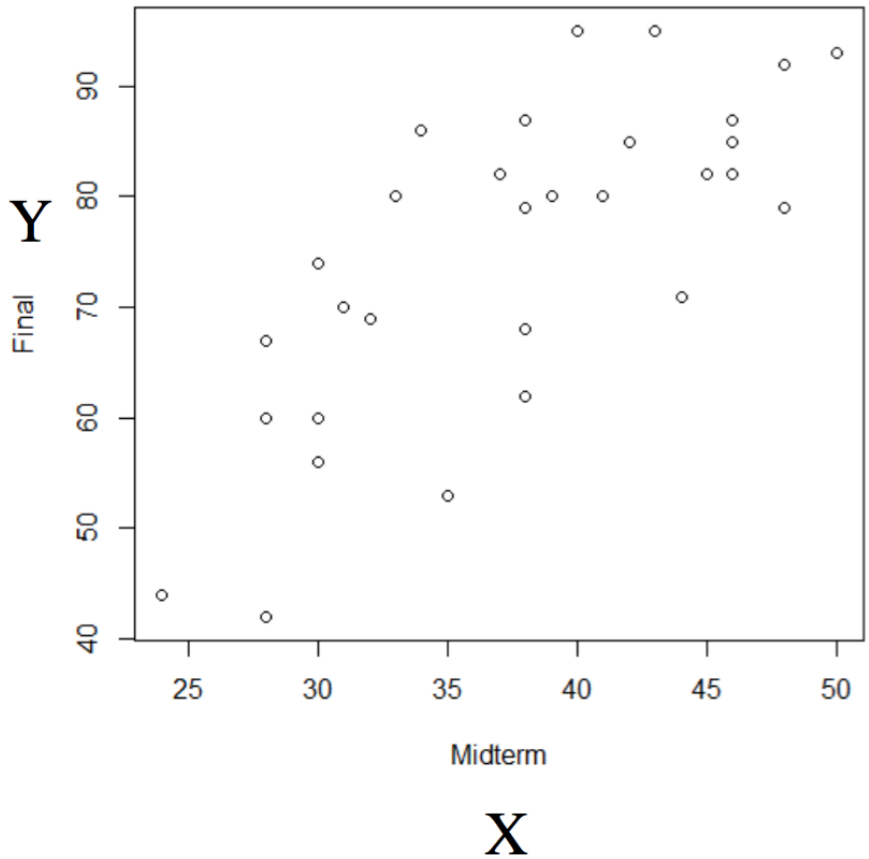
Here is how correlation r measures the direction and strength of a linear association:



First: Plot the data (Scatterplot)

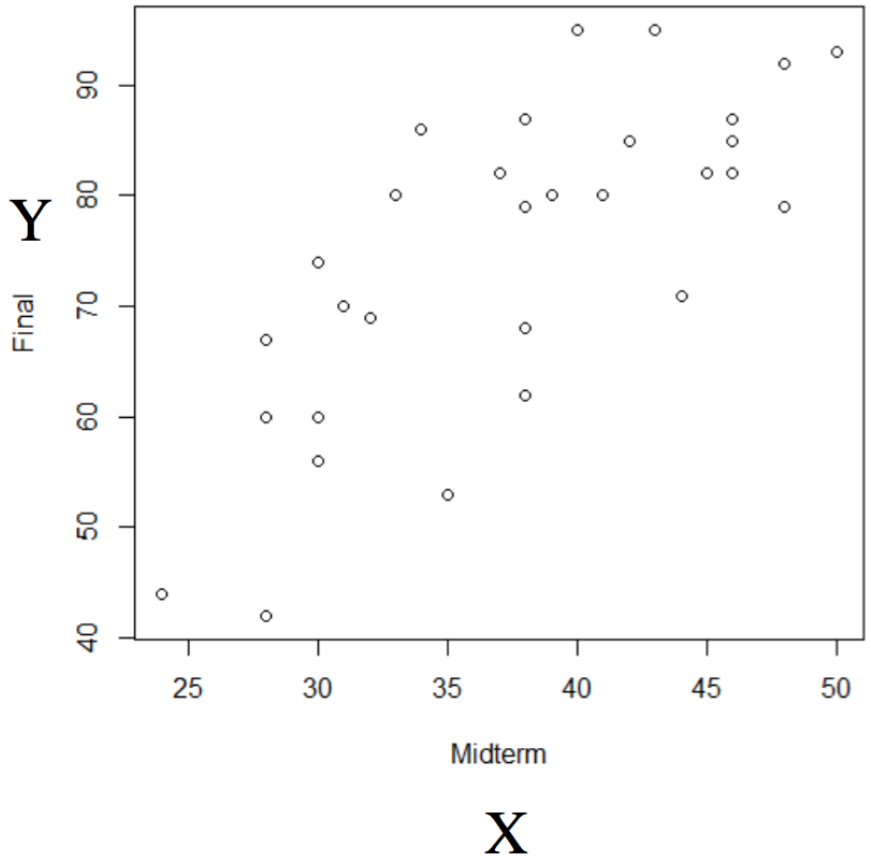
Can we use the midterm score (X) to predict the final score (Y)?

Many times we attempt to do so by fitting a well-fitting line through those points and using that line for prediction.

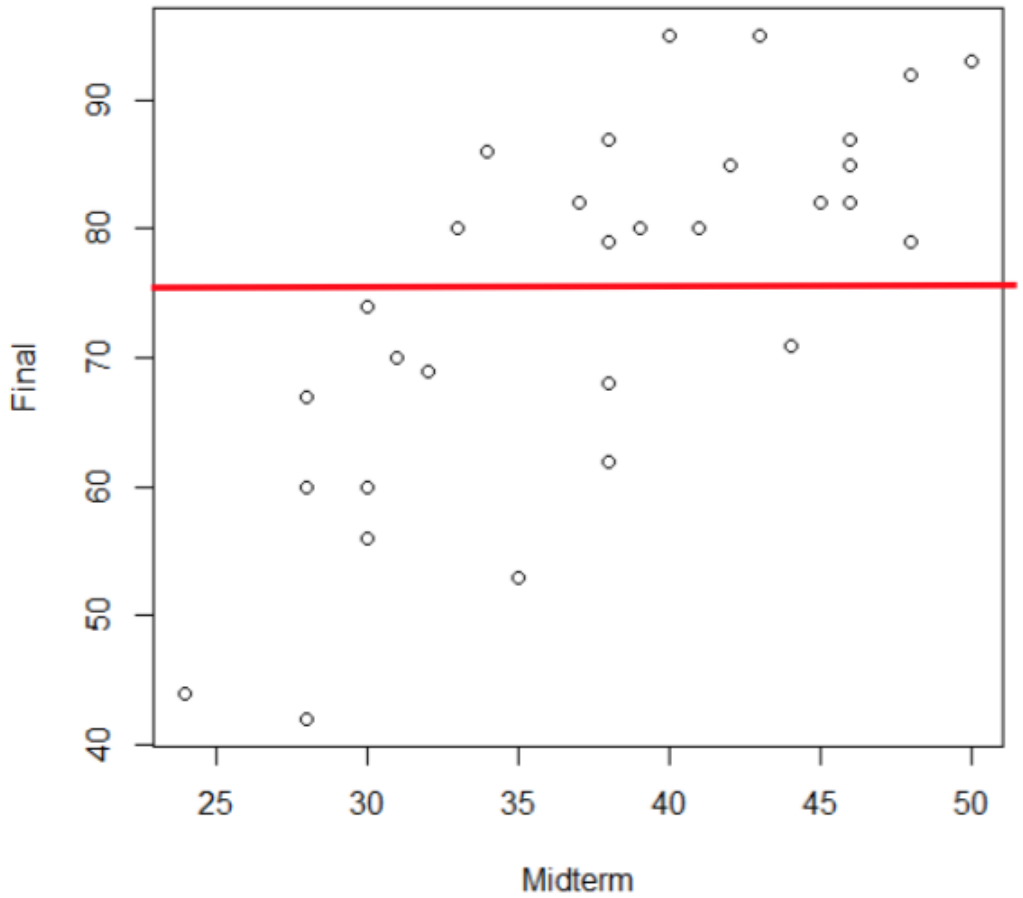


First: Plot the data (Scatterplot)

Question: Where's the "best fitting" line through the data?

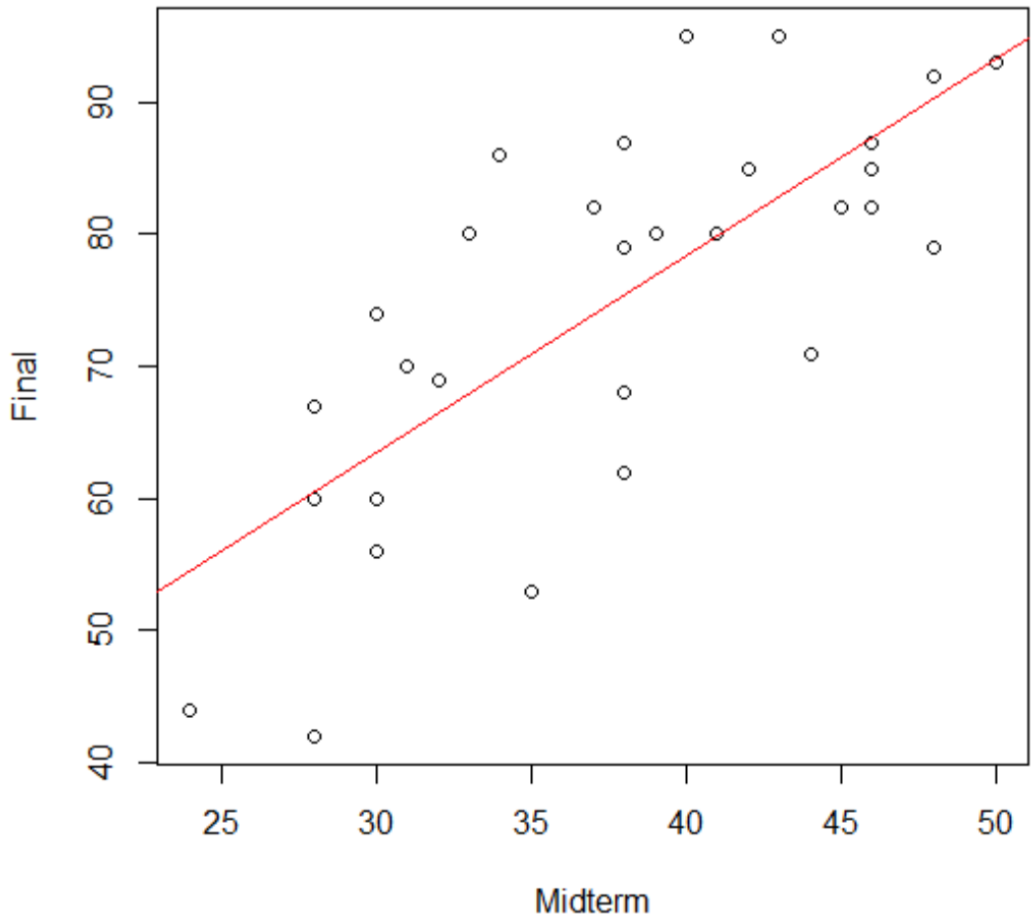


One Possible Model



Final mean $\bar{y} = 75.4$

Another Possible Model



Is it better than the previous model?

Simple Linear Regression

- We will assume the true linear relationship between X and Y is given by:

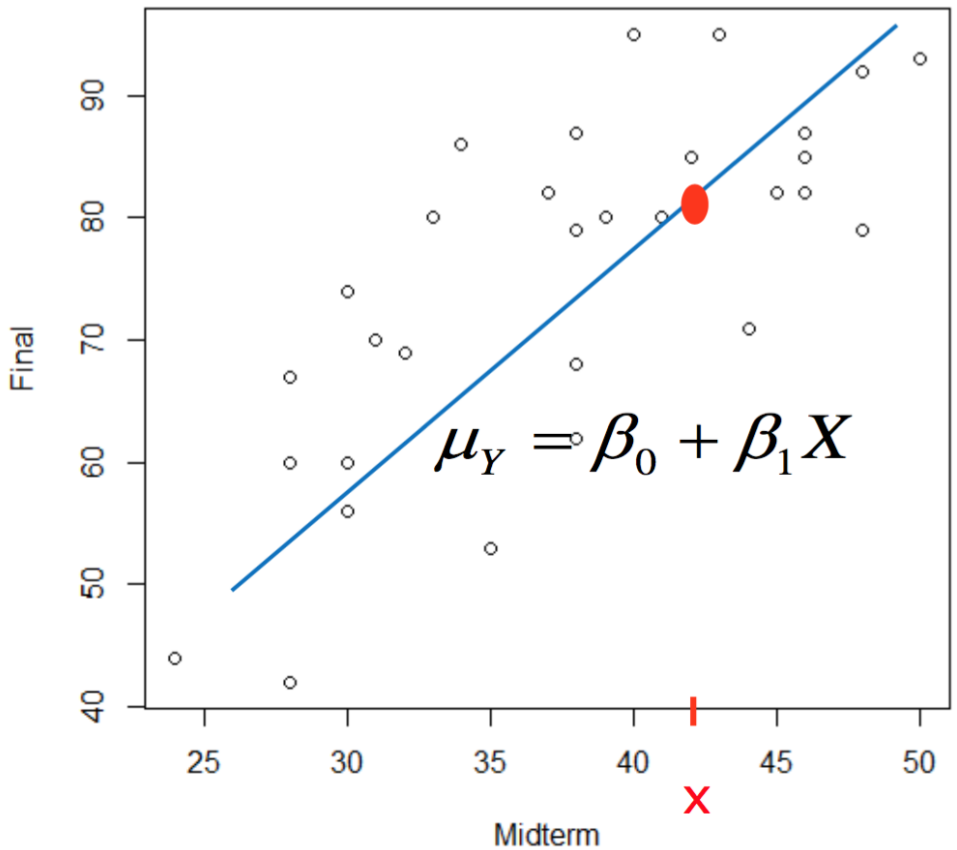
$$\mu_y = \beta_0 + \beta_1 X$$

$\beta_0 = Y -$ intercept

$\beta_1 =$ slope

- β_0 and β_1 parameters. Typically unknown values and need to be estimated.

Line of the true relationship



For a given value of x, true mean of y for this value this value of x falls precisely on the line.

The actual observed values of Y will not fall precisely on that line because there are some variability involved.

Line of the true relationship

- The observed value of Y vary about the line.
- Thus, we can write:

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

where ϵ is random error component.

Simple Linear Regression

- The models is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon, \quad i = 1, \ldots, n$$

- Y_i is the value of the response variable in the i-th trial.
- X_i is a known constant, the value of the predictor variable in the ith trial.
- ϵ_i is a random error term with $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$
- ϵ_i and ϵ_j are uncorrelated.

Later: we will replace the last assumption by:

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Important Features of Model

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \beta_0 + \beta_1 + E(\epsilon_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

$$\begin{aligned} Var(Y_i) &= Var(\beta_0 + \beta_1 X_i + \epsilon_i) \\ &= Var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

i.e., responses Y_i have the same constant variance

Notation

$\hat{Y}_i$ is the predicted response (or fitted value) for the i-th experimental unit.

Equation of best fitting line: $\hat{Y}_i = b_0 + b_1 X_i$.

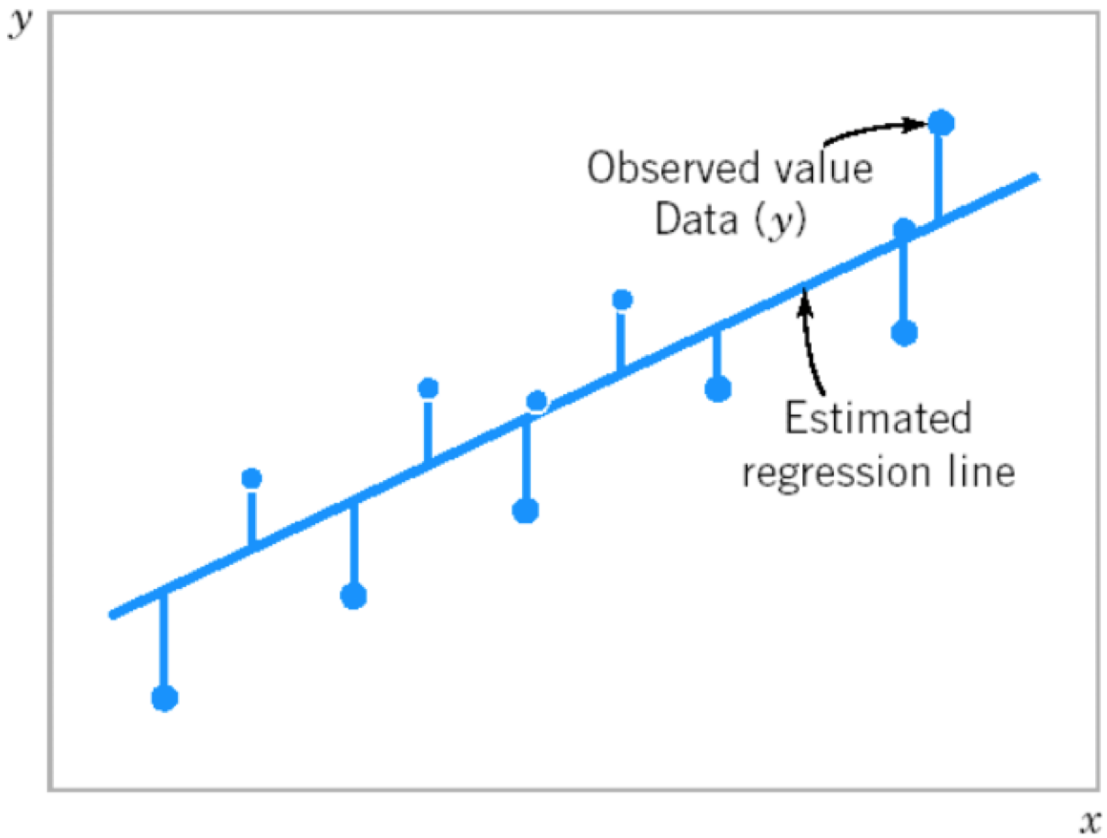
b_0 and b_1 : point estimators of β_0 and β_1.

$$\hat{\beta}_0 = b_0 \quad \text{and} \quad \hat{\beta}_1 = b_1$$

How to choose the values b_0 and b_1?

Methods of Least squares

In using $\hat{Y}_i$ to predict the actual response Y_i, we make a prediction error (or a residual error) of size $e_i = Y_i - \hat{Y}_i$



Method of Least squares

Proposal 1: Minimize

$$\sum_{i=1}^{n}[Y_i - (b_0 + b_1 X_i)]$$

Proposal 2: Minimize

$$\sum_{i=1}^{n}|Y_i - (b_0 + b_1 X_i)|$$

Final proposal: choose b_0 and b_1 that minimize

$$\begin{aligned} Q = Q(b_0, b_1) &= \sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \end{aligned}$$

The least squares regression line

Using calculus, minimize (take derivative with respect to b_0 and b_1, set to 0, and solve for b_0 and b_1):

$$Q = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

and get the least squares estimates b_0 and b_1:

$$b_1 = r\frac{S_y}{S_x} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - b_1\bar{X}$$

The least squares regression line

$$r = \frac{1}{n-1}\sum\left(\frac{X_i - \bar{X}}{S_x}\right)\left(\frac{Y_i - \bar{Y}}{S_y}\right)$$

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

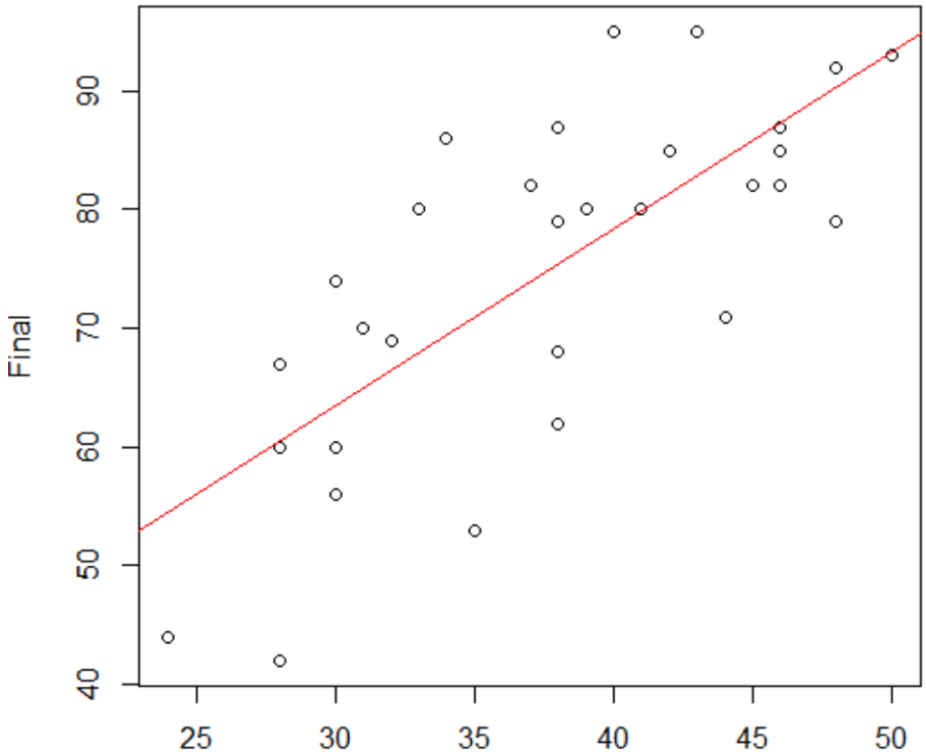
$$S_y^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

$$S_{xy} = \frac{1}{n-1}\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

Fitted line plot in R

The estimated regression line:

$$\hat{Y}_i = 18.672 + 1.492X_i$$



R code:

```
> Fit=lm(Final~Midterm)
> Fit$coefficients
(Intercept)     Midterm
    18.6721      1.4925
```

Prediction of future responses

Predict mean final exam score if midterm exam scores are 20, 40 and 45.

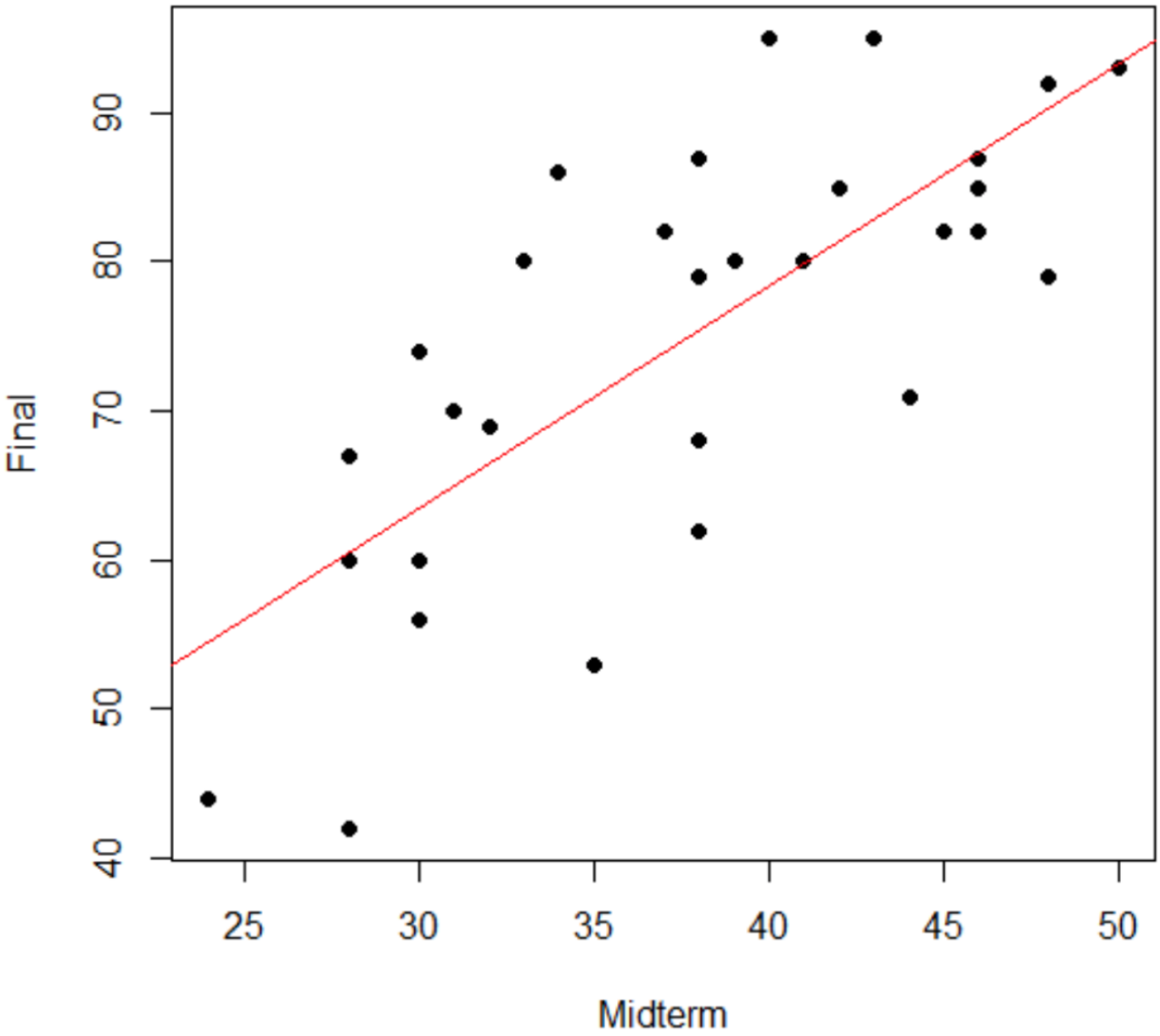
$$\hat{Y}_i = 18.6721 + 1.4925(20) = 48.5221$$

$$\hat{Y}_i = 18.6721 + 1.4925(40) = 78.3721 \quad \text{Ramon: (40, 95)}$$

$$\hat{Y}_i = 18.6721 + 1.4925(50) = 93.2971 \quad \text{Maria: (50, 93)}$$

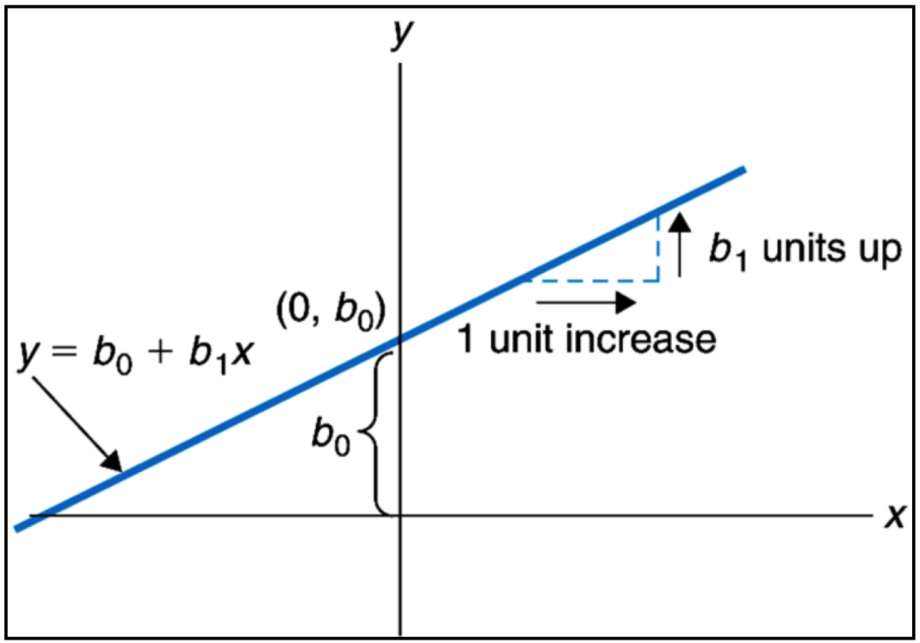
R code:

```
predict(Fit, newdata=data.frame(Midterm=c(20,40,50)))
48.5221 78.3721 93.2971
```


Final
40
50
60
70
80
90
25
30
35
40
45
50
Midterm

Interpretation of regression coefficients

- We can expect the mean response to increase or decrease by b_1 units for every unit increase in x.
- If the "scope of the model" includes $x = 0$, then b_0 is the predicted mean response when $x = 0$. Otherwise, b_0 is not meaningful.



Example: Interpretation of regression coefficients

$$\hat{Y}_i = 18.6721 + 1.4925\, X_i$$

- We predict the mean final score to increase by 1.492 for every additional score increase in the midterm exam.
- If a student get 0 in the midterm, then the predicted men final score will be 18.672.

Properties of the Fitted Regression Line

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)$$
$$= \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0$$

- The sum of the observed values equal the sum of the fitted values. That is,

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

Properties of the Fitted Regression Line

- The sum of the weighted residuals is zero when the residual in the i-th trial is weighted by the level of the predictor variable in the i-th trial:

$$\begin{aligned}\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i \left(Y_i - b_0 - b_1 X_i\right) \\ &= \sum_{i=1}^{n} X_i Y_i + b_0 \sum_{i=1}^{n} X_i - b_0 \sum_{i=1}^{n} X_i^2 \\ &= 0\end{aligned}$$

Properties of the Fitted Regression Line

The sum of the weighted residuals is zero when the residual in the i-th trial is weighted by the fitted value of the response variable in the i-th trial:

$$\begin{aligned}\sum_{i=1}^{n} \hat{Y}_i e_i &= \sum_{i=1}^{n} (b_0 + b_1 X_i) e_i \\ &= b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i \\ &= 0\end{aligned}$$

Properties of the Fitted Regression Line

The regression line always goes through the point $(\bar{x}, \bar{y})$:

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X \\ &= \bar{Y} - b_1 \bar{X} + b_1 X \\ &= \bar{Y} + b_1 (X - \bar{X})\end{aligned}$$

If $X = \bar{X}$, then $\hat{Y} = \bar{Y}$.

Regression Assumptions

- The mean of the responses, $E(Y_i)$, is a linear function of x_i.
- The errors, ϵ_i, and hence the responses Y_i, are independent.
- The errors, ϵ_i, and hence the responses Y_i, are normally distributed.
- The errors, ϵ_i, and hence the responses Y_i, have equal variances (σ^2) for all x values.

Properties of the Least Squares Estimates

b_0 and b_1 are unbiased estimators of β_0 and β_1:

- $E(b_0) = \beta_0, \qquad E(b_1) = \beta_1$
- $Var(b_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$
- $Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$
- $b_1 \sim N(E(b_1), Var(b_1))$
- $b_0 \sim N(E(b_0), Var(b_0))$

Estimating σ^2 in regression setting

- It quantifies how much the responses (y) vary around the (unknown) mean regression line $E(Y) = \beta_0 + \beta_1 X$.
- The estimate of σ^2 is called the mean square of errors (MSE) and given by:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- MSE is unbiased estimator of σ^2 : $E(MSE) = \sigma^2$
- R code: summary(Fit)$sigma

  9.651273

Confidence Intervals for β_1 and β_0

- $(1-\alpha)100\%$ CI for β_1

$$b_1 \pm t_{(1-\alpha/2,n-2)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}} \right)$$

- $(1-\alpha)100\%$ CI for β_0

$$b_0 \pm t_{(1-\alpha/2,n-2)} \times \sqrt{MSE} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sqrt{\sum(x_i - \bar{x})^2}}}$$

Example

```
confint(Fit)
                2.5 %      97.5 %
(Intercept) -0.4122049 37.756398
Midterm 0.9990233 1.985977
```

95% CI for β_1 : [0.9990233, 1.985977]

95% CI for β_0 : [-0.4122049,37.756398]

Testing Coefficients

- Testing whether the slope is equal to β_1^*
  - $H_0 : \beta_1 = \beta_1^*$ vs. $H_a : \beta_1 \neq \beta_1^*$
  - Test statistic $t = (b_1 - \beta_1^*)/s(b_1)$
  - which follows t_{n-2} under H_0 where $s(b_1) =$ standard error of b_1
- Testing whether the intercept is equal to β_0^*
  - $H_0 : \beta_0 = \beta_0^*$ vs. $H_a : \beta_0 \neq \beta_0^*$
  - Test statistic $t = (b_0 - \beta_0^*)/s(b_0)$
  - which follows t_{n-2} under H_0 where $s(b_0) =$ standard error of b_0

Testing Coefficients

- Testing whether the slope is significant
  - $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
  - Test statistic $t = b_1/s(b_1)$
  - which follows t_{n-2} under H_0 where $s(b_1) =$ standard error of b_1
- Testing whether the intercept is significant
  - $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$
  - Test statistic $t = b_0/s(b_0)$
  - which follows t_{n-2} under H_0 where $s(b_0) =$ standard error of b_0

Example

In previous example, does the midterm exam score seem a useful predictor for the final?

- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- $t = 6.186$
- P-value $= 0.000 < 0.05$
- We reject H_0 and conclude that midterm exam score is a useful predictor for the final exam score.

```
Midterm = c(35,38,50,39,37,46,32,24,48,41,40,44,48,28,46,48,43,33,
30,28,45,34,28,42,31,46,30,38,38,30,38)
Final = c(53,68,93,80,82,87,69,44,79,80,95,71,92,67,85,92,95,80,56,
60,82,86,42,85,70,82,60,87,62,74,79)
plot(Midterm,Final)
Fit=lm(Final ~ Midterm)
abline(Fit, col="red") # regression line (y ~ x)
summary(Fit)
```

Exercise 1, # 11.4 (Data)

For the cars example:

(1) Calculate the correlation coefficient and comment on the relationship between the variables.

(2) Develop a scatterplot and comment on the relationship relationship between the variables.

(3) Determine the regression equation for the data.

(4) Interpret carefully the regression coefficients

5) Compute and interpret coefficient of determination R^2.

(6) Does the predictor seem a good predictor for the response? Use $\alpha = 0.05$.

(7) Obtain the residuals and check the regression assumptions.

(8) Predict the mean of response for book value 12.

(9) Predict with 90% confidence the mean of response for book value 12.

(10) Predict with 90% confidence the response for book value 12.

(11) Check the assumptions of the fitted model.

Exercise 2, # 11.9 (Data)

Do all parts of Exercise 1 for Data set of exercise # 11.9 ?