

Big Data Analytics

Session 3

Simple Linear Regression

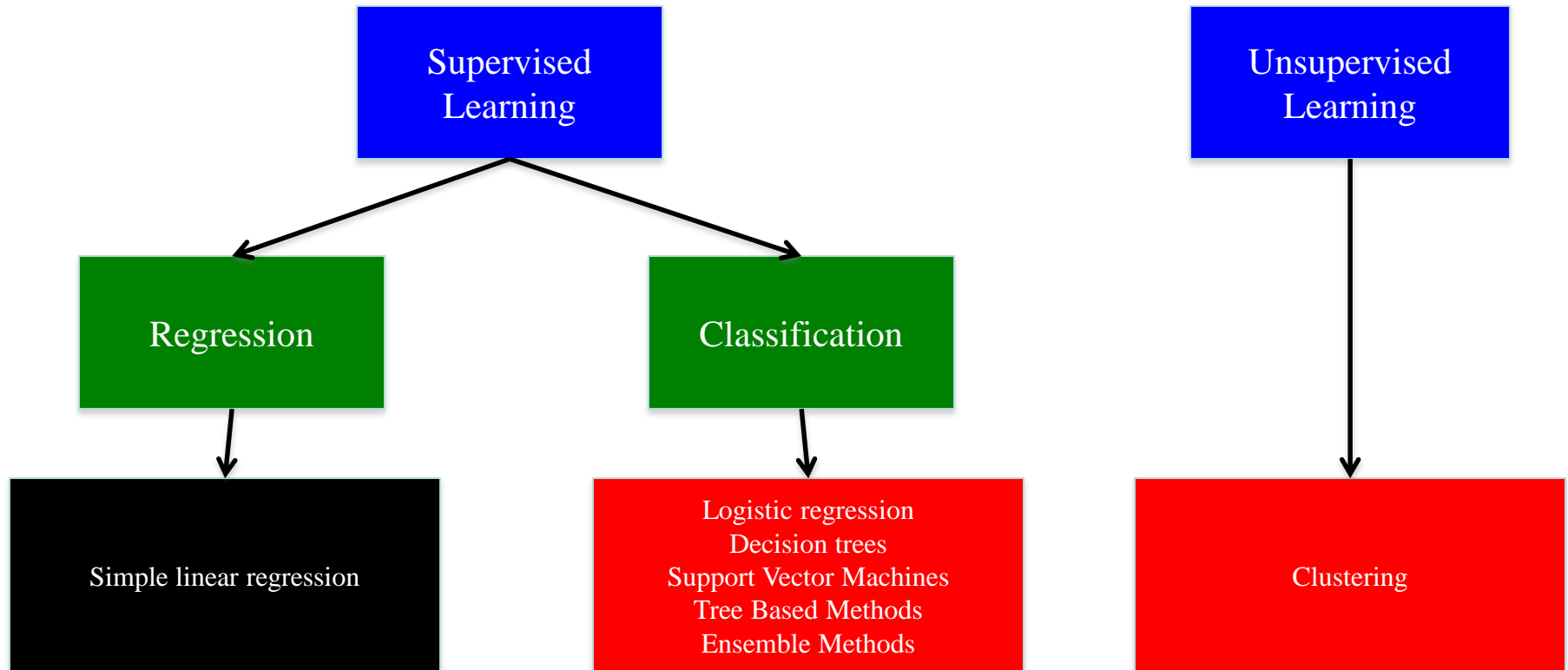
Where were we last week?



- Data: Scale of measurement
 - Nominal, Ordinal, Interval, Ratio
- Univariate analysis: describing the distribution of a single variable
 - Measures of central tendency: Mean, Median, Mode
 - Measures of spread: Variance, Standard Deviation
 - Measures of dispersion: Range, Quartiles, Interquartile Range
- Bivariate analysis: describing the relationship between pairs of variables
 - Quantitative measures of dependence: Correlation, Covariance
- Tabular and graphical presentation
 - Frequency distribution, Histogram, Box plot, Scatter plot

Today: Linear Regression

- Predicting a **quantitative** response

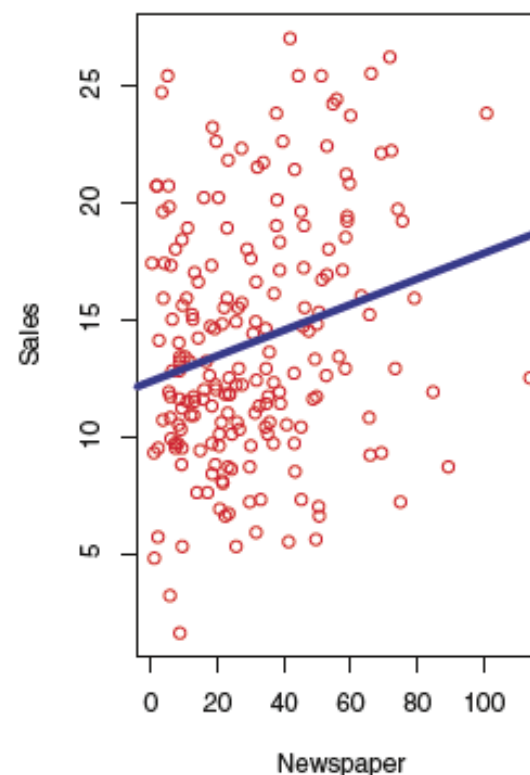
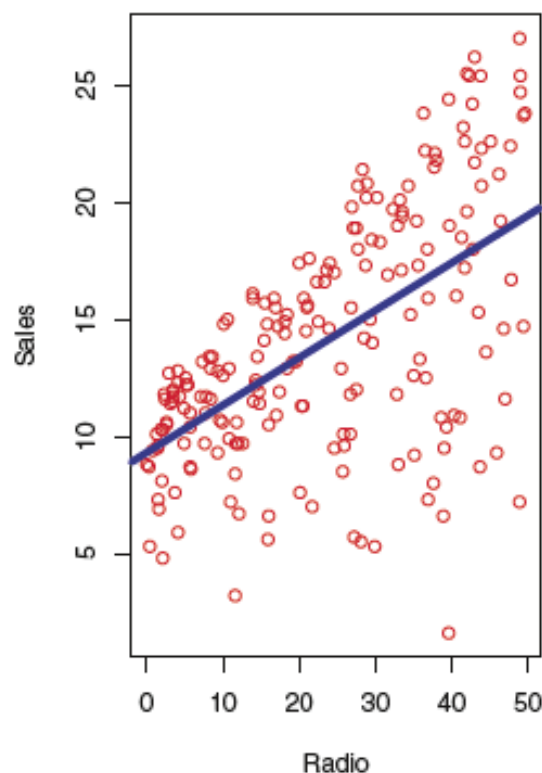
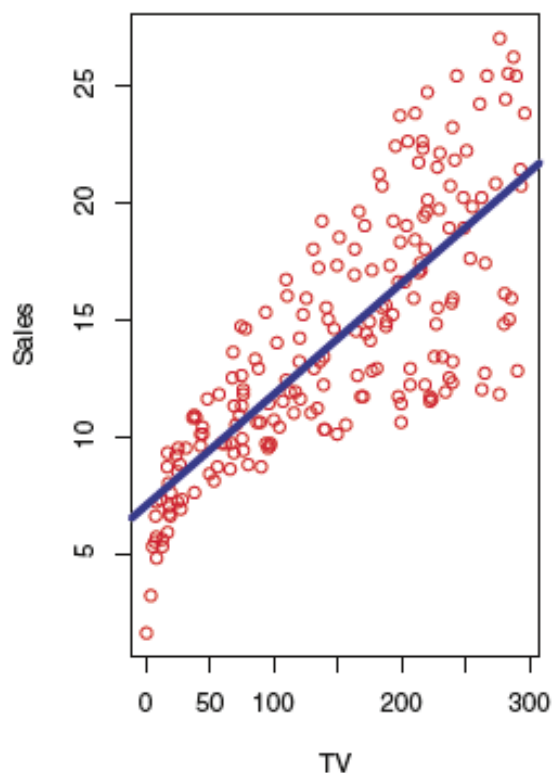


Choosing the best methods for a given application: Cross-validation

Applications: e.g., Social Networks.

Example: Advertising

- Sales for a particular product as a function of advertising budgets for TV, radio and newspaper media



Linear Functions

- **Linear** functions refer to equations such as:
 - Linear functions are linear with respect to the **variables**
 - $f(x) = -0.4x - 2$
 - $f(x_1, x_2) = 4x_1 + 5x_2 + 3$
 - $f(x_1, x_2, x_3) = -7x_1 + 5x_2 - \sqrt{2} x_3 - 1$
- **Non-linear** functions refers to equations such as:
 - $f(x_1, x_2) = 2x_1^2 + 3x_2$
 - $f(x_1, x_2, x_3) = -2x_1^{1/2} + 3x_2^5 - 0.7x_3^3$
 - $f(x_1, x_2) = 2x_1 + 3x_2 + 3x_1x_2$
- If we assume x_1^2 and x_2 are **known and fixed**:
 - Is $f(a, b) = ax_1^2 + bx_2$ linear or non-linear?

First-Order Linear Functions

A first-order linear function is a straight line of the form:

$$y = \beta_0 + \beta_1 x$$

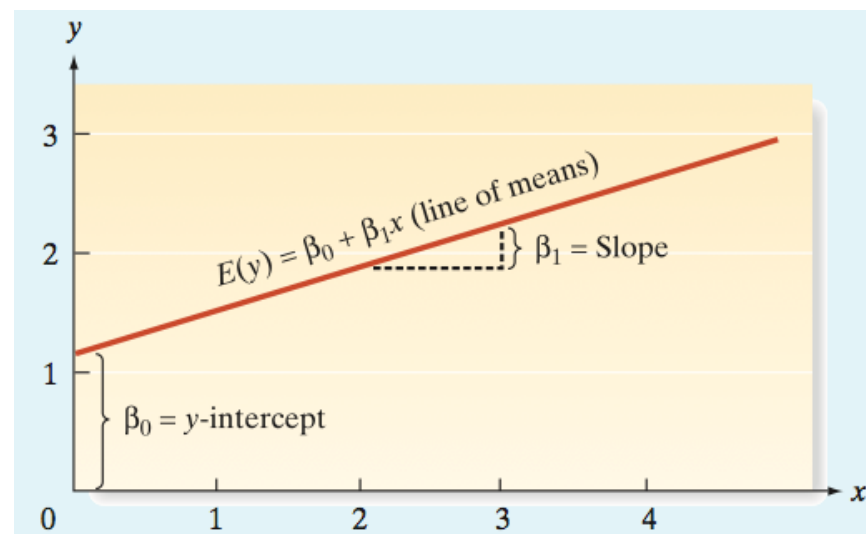
where

β_0 = **y-intercept of the line**

the point at which the line *intercepts or cuts through the y-axis*

β_1 = **slope of the line**

the change (amount of increase or decrease) in the deterministic component of y for every 1-unit increase in x



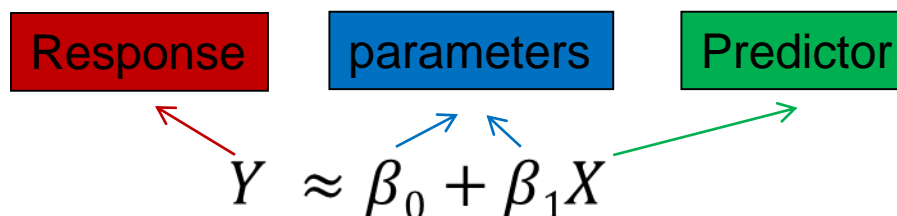
Outline



- **Simple** linear regression
 - a *single predictor* variable: $Y \sim X$
 - E.g., The relationship between **sales** and **TV** advertising budget
- **Multiple** linear regression (self-study, selective)
 - *More than one* predictor variable: $Y \sim X_1, X_2, \dots$
 - E.g., The relationship between **sales** and **TV, radio and newspaper** advertising budgets

Simple Linear Regression

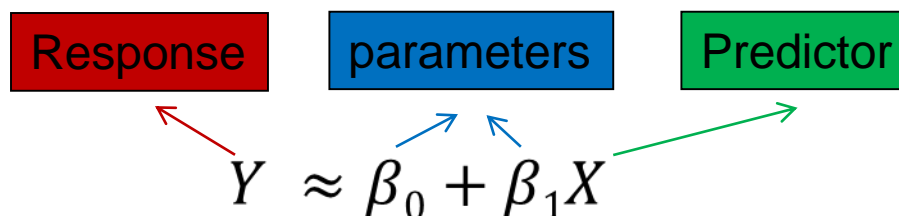
To predict a quantitative response Y on the basis of a single predictor variable X .



We are regressing Y on X .

Simple Linear Regression

To predict a quantitative response Y on the basis of a single predictor variable X .



We are regressing Y on X .

Step1:

Use the training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Step2:

Use $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ to predict Y (as \hat{y}) on the basis of $X = x$

Overview of Step 1



- Step 1: use training data to estimate coefficients (parameters)
 - How to estimate?
 - Assessing the accuracy of the coefficient estimates
 - Assessing the accuracy of the model

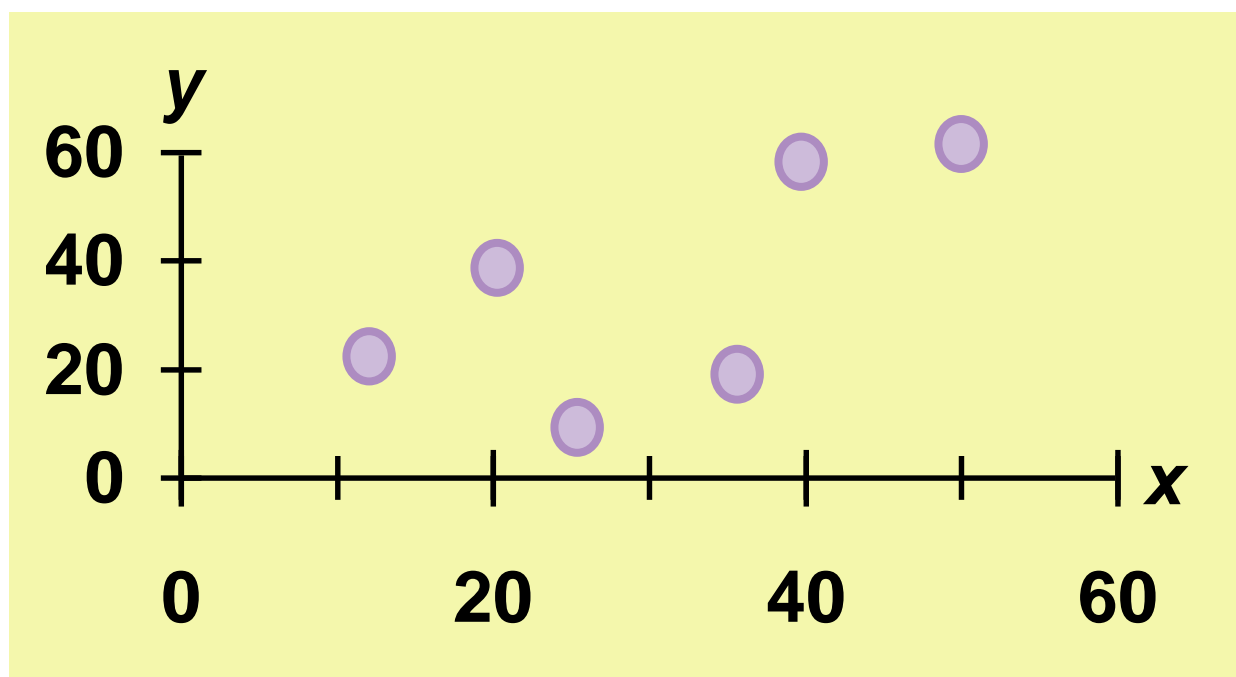
Overview of Step 1



- Step 1: use training data to estimate coefficients (parameters)
 - How to estimate?
 - Assessing the accuracy of the coefficient estimates
 - Assessing the accuracy of the model

Plotting Training Data

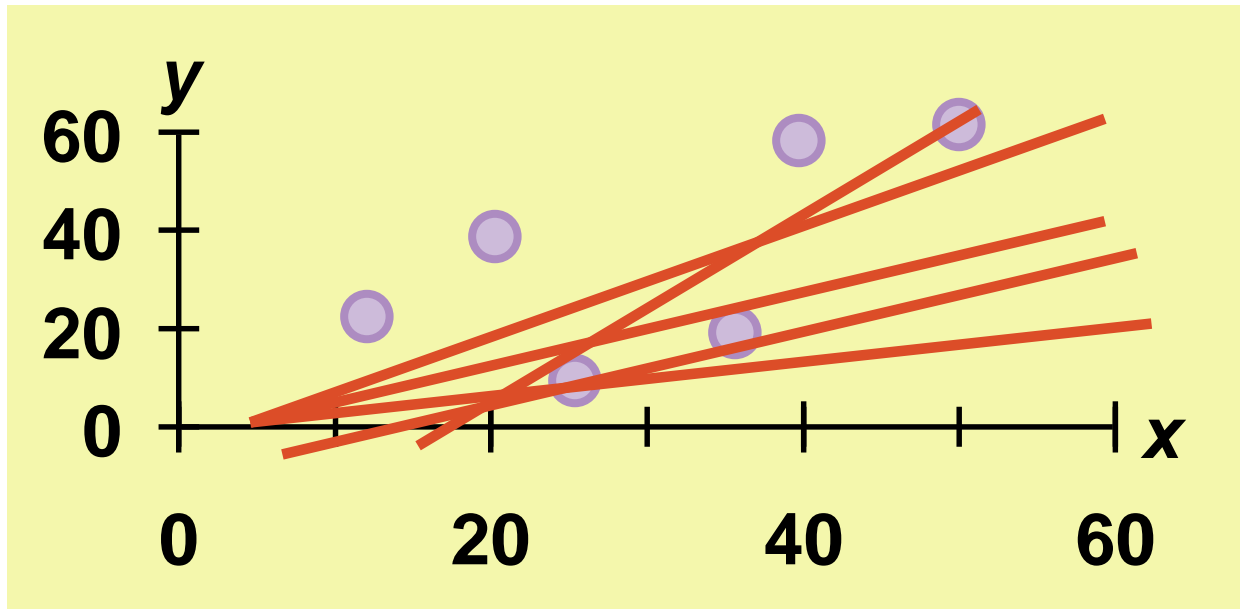
- Given n observations $(x_1, y_1), \dots, (x_n, y_n)$, plot all (x_i, y_i) pairs by **scatter plots**



- Remember the cigarette and lung capacity example last week?

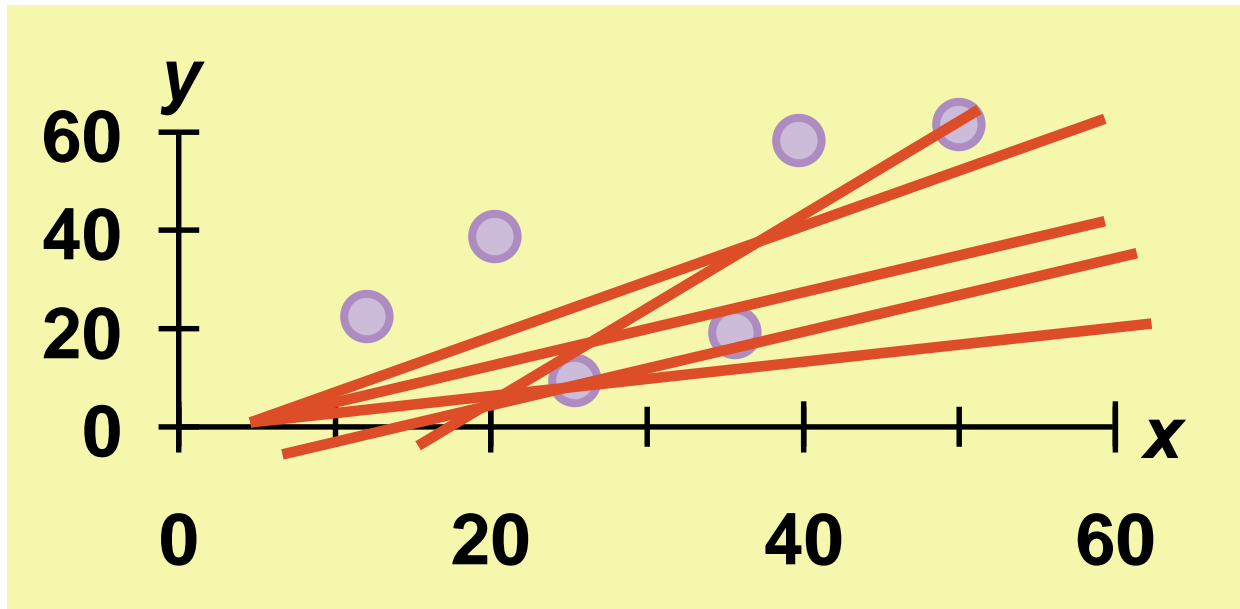
How to fit?

- How would you draw a line through the points?



How to fit?

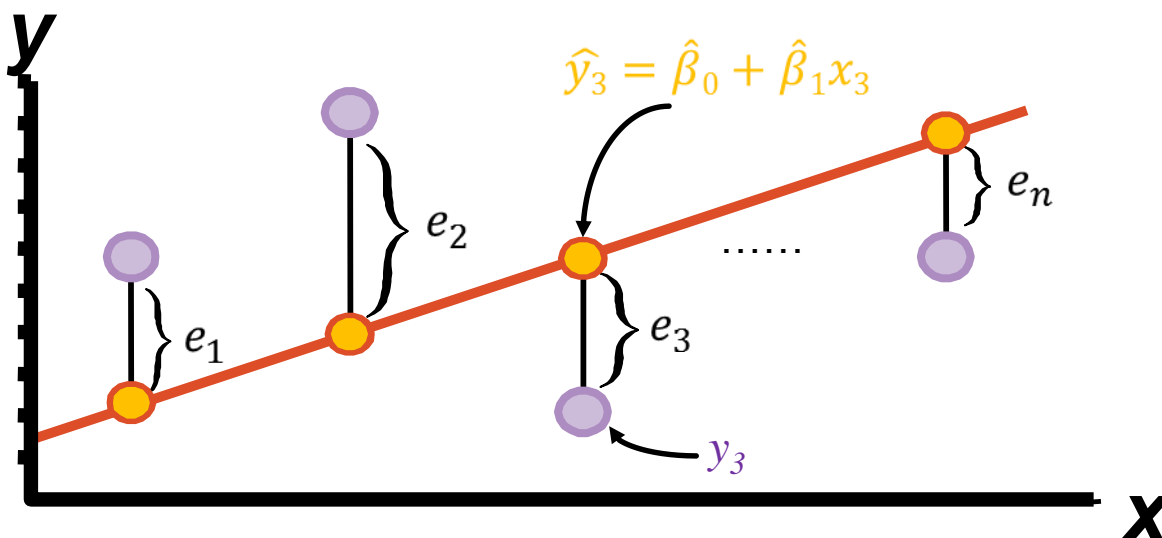
- How would you draw a line through the points?
- How do you determine which line 'fits best'?



Residual Sum of Squares

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the prediction of Y based on the i th value of X
- y_i is the observed value ← Real value!
- $e_i = y_i - \hat{y}_i$ is the i th residual (residual = observed – predicted)
- Residual sum of squares (RSS)
- $RSS = e_1^2 + e_2^2 + \dots + e_n^2$

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$



Least Squares Line



- The **least squares line** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is one that has the following two properties:
 - The sum of the residuals equals 0, that is, mean residual = 0
 - The residual sum of squares is minimised

Least Squares Line

- The **least squares line** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is one that has the following two properties:
 - The sum of the residuals equals 0, that is, mean residual = 0
 - The residual sum of squares is minimised
- Using some calculus, one can show that the **minimisers** are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$$

- In other words, the above equation defines the least squares coefficient estimates for simple linear regression.

Least Squares Line

- The **least squares line** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is one that has the following two properties:
 - The sum of the residuals equals 0, that is, mean residual = 0
 - The residual sum of squares is minimised
- Using some calculus, one can show that the **minimisers** are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$$

- In other words, the above equation defines the **least squares coefficient estimates** for simple linear regression.

Least Squares Example

You're a marketing analyst for Hasbro Toys.
You gather the following data:

<u>Ad Expenditure (100 £)</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

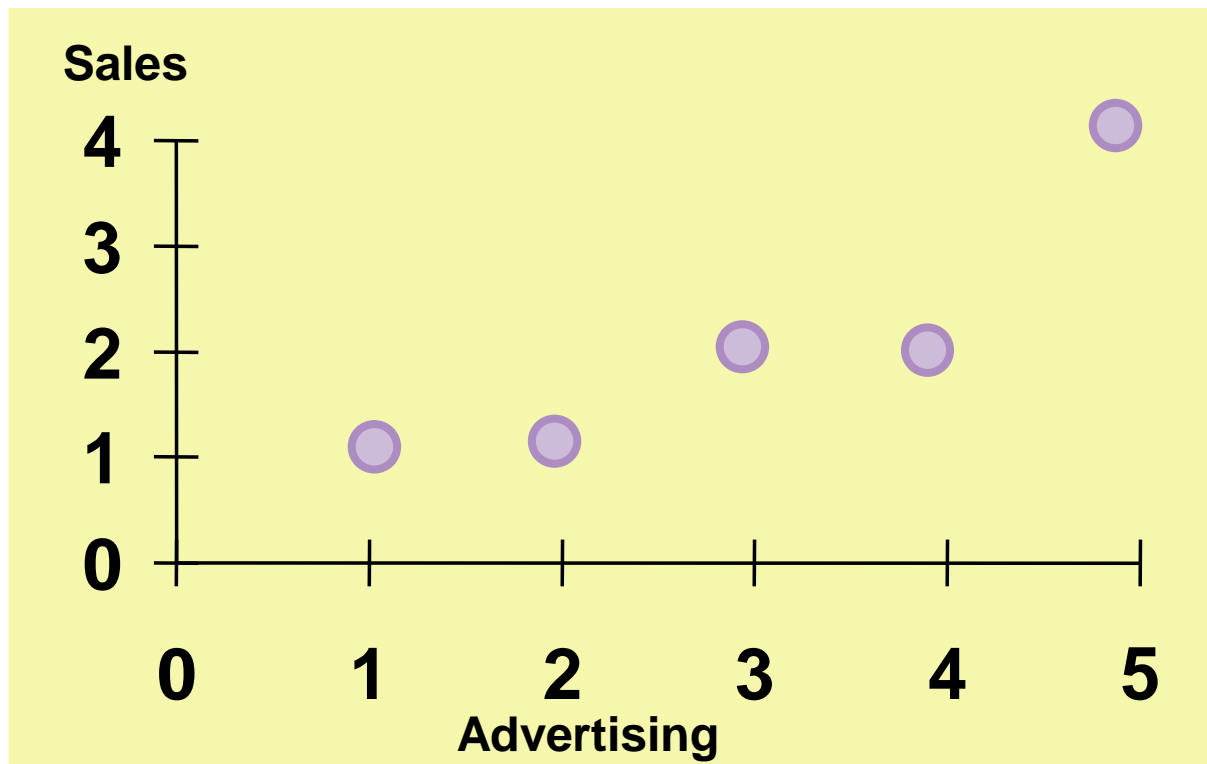
Find the **least squares line** relating sales and advertising.



Scatter Plot -- Sales vs. Advertising

- Plot it

<u>Ad Expenditure (100 £)</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4



Minimising RSS

- Recall:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$$

Minimising RSS

<u>Ad Expenditure (100 £)</u>	<u>Sales (Units)</u>
1	1
2	1
3	2
4	2
5	4

- $\bar{x} = \frac{1+2+3+4+5}{5} = 3$
- $\bar{y} = \frac{1+1+2+2+4}{5} = 2$
- $\hat{\beta}_1 = \frac{(1-3)(1-2)+(2-3)(1-2)+(3-3)(2-2)+(4-3)(2-2)+(5-3)(4-2)}{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2} = 0.7$
- $\hat{\beta}_0 = 2 - 0.7 * 3 = -0.1$

- Least Squares Line:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = -0.1 + 0.7x_i$$

- Recall:

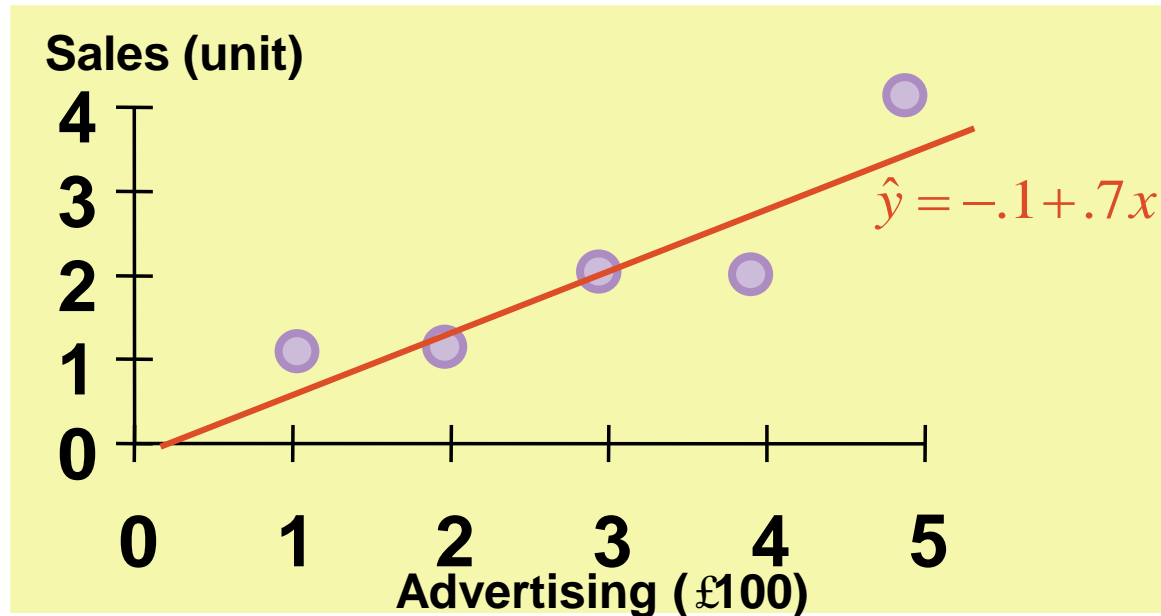
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$$

Regression Line Fitted to the Data



1. Slope (β_1)

- Sales Volume (y) is expected to increase by 0.7 unit for each £100 increase in advertising (x), *over the sampled range of advertising expenditures from £100 to £500*

2. y -Intercept (β_0)

- Since 0 is outside of the range of the sampled values of x , the y -intercept has no meaningful interpretation

Overview of Step 1



- Step 1: use training data to estimate coefficients (parameters)
 - How to estimate?
 - Assessing the accuracy of the coefficient estimates
 - Assessing the accuracy of the Model

Assessing the accuracy of coefficient estimates



- Three different lines:

- True relationship:

$$Y = f(X) + \epsilon$$

- ϵ is a mean-zero random error term

Assessing the accuracy of coefficient estimates



- Three different lines:

- True relationship: $Y = f(X) + \epsilon$

- ϵ is a mean-zero random error term

- Population regression line: $Y = \beta_0 + \beta_1 X + \epsilon$

- f is to be approximated by a linear function
 - ϵ is a catch-all for what we miss with this simple model:
 - The true relationship is probably not linear; (reducible error)
 - There may be other variables that cause variation in Y
 - There may be measurement error
 - Assume that ϵ is independent of X
 - The best *linear* approximation to the true relationship between X and Y

Assessing the accuracy of coefficient estimates



- Three different lines:

- True relationship: $Y = f(X) + \epsilon$

- ϵ is a mean-zero random error term

- Population regression line: $Y = \beta_0 + \beta_1 X + \epsilon$

- f is to be approximated by a linear function
 - ϵ is a catch-all for what we miss with this simple model:
 - The true relationship is probably not linear; (reducible error)
 - There may be other variables that cause variation in Y
 - There may be measurement error
 - Assume that ϵ is independent of X
 - The best *linear* approximation to the true relationship between X and Y

- Least squares line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

- With the least squares regression coefficient estimates

Sample Mean and Population Mean



- Recall in Session 2:

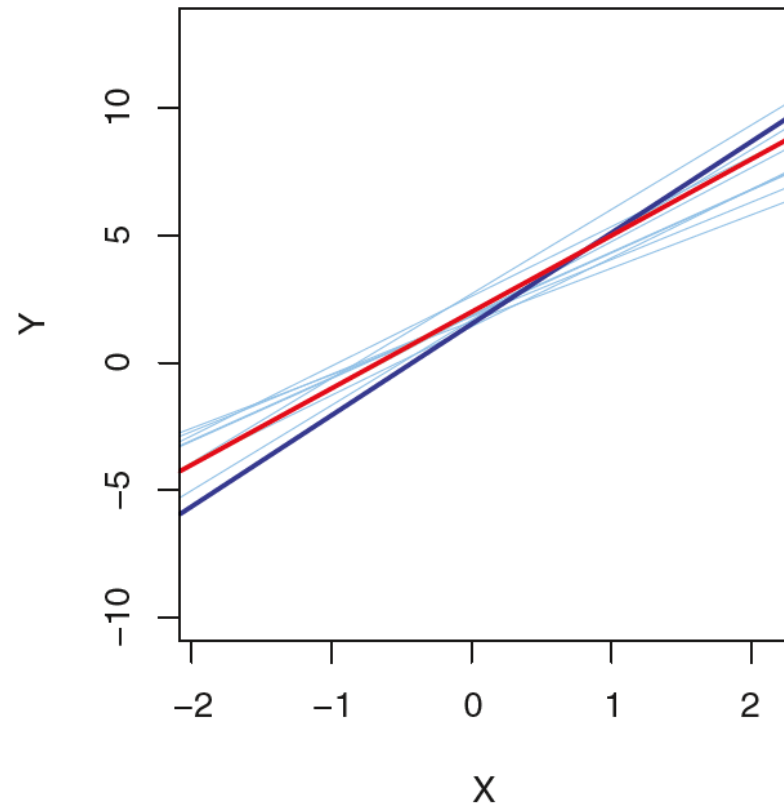
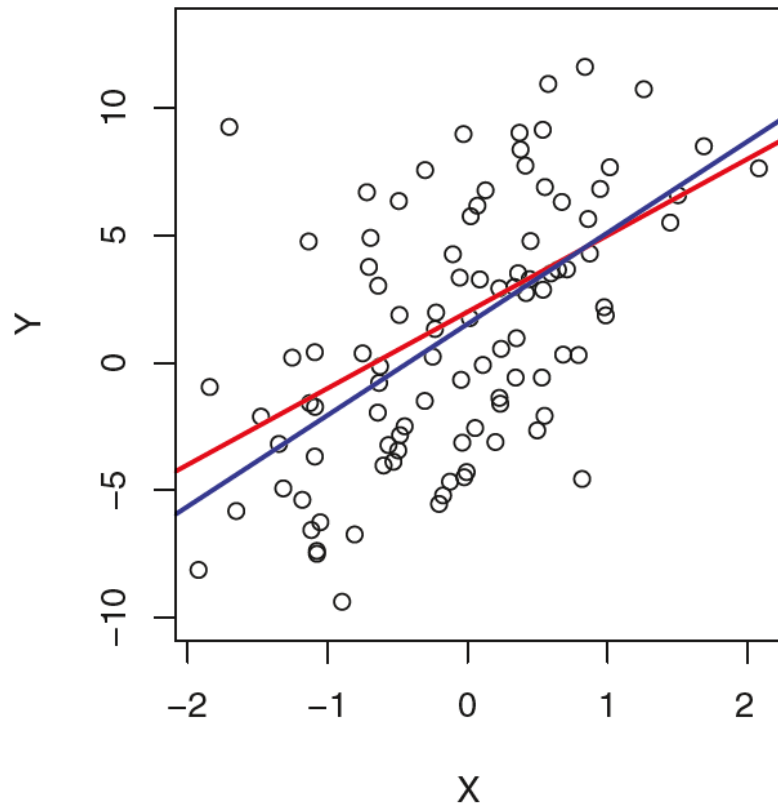
- Sample mean $\bar{x} = \frac{\sum x_i}{n}$ - population mean $\mu = \frac{\sum x_i}{N}$

- Use \bar{X} to estimate $\mu \rightarrow$ write $\hat{\mu} = \bar{X}$

- If $\hat{\mu}$ is based on one particular set of observations, $\hat{\mu}$ may be over or under estimate μ

- If we could average a huge number of sample means, then $\hat{\mu}$ will be the accurate population mean

An Analogue



Red line: population regression line $f(X) = 2+3X$, usually unknown

Dark blue line: least square line – based on one set of observations

Light blue lines: least square lines – each based on a separate random set of obs.

An Analogue

- Population regression line:
- Least squares line:

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

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

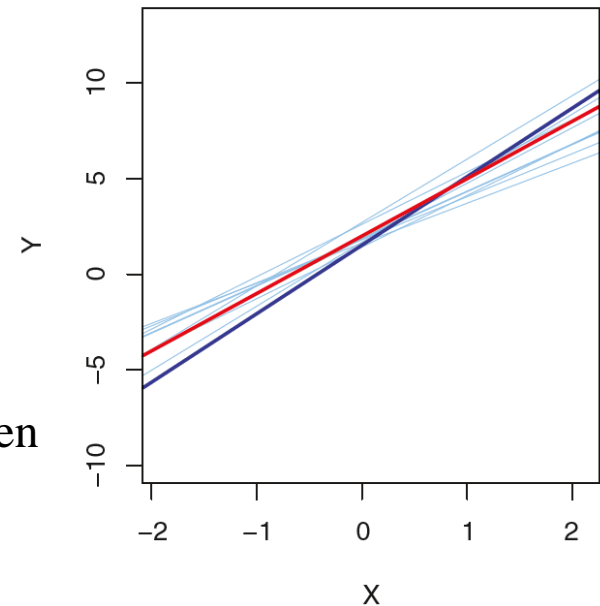
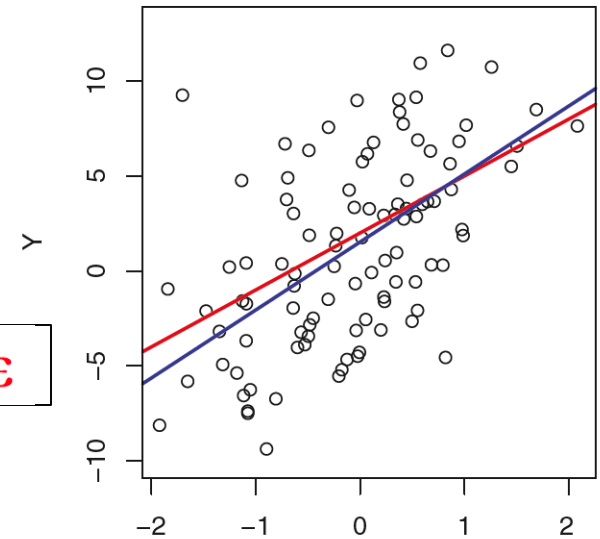
- Use $\hat{\beta}_0$ and $\hat{\beta}_1$ to estimate β_0 and β_1
- If $\hat{\beta}_0$ and $\hat{\beta}_1$ are based on one particular set of observations, $\hat{\beta}_0$ and $\hat{\beta}_1$ may under or over estimate β_0 and β_1
- If we could average large number of the parameters, then the resulting $\hat{\beta}_0$ and $\hat{\beta}_1$ will be the accurate population

regression line parameters

Red line: population regression line $f(X) = 2+3X$, usually unknown

Dark blue line: least square line – based on one set of observations

Light blue lines: least square lines – each based on a separate random set of obs.



Standard Error

- How close is a single sample mean $\hat{\mu}$ to the population mean μ ?
 - Use standard error: the average amount that this estimate $\hat{\mu}$ differs from μ
 - $$\text{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$
 - ← σ : the standard deviation, σ^2 : variance
 - ← the more observations we have, the smaller the SE is
- When sample size increases
 - the standard error of the sample will tend to 0
 - because the estimate of the population mean will improve

An Analogy

- Population regression line: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Least squares line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- How close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true value β_0 and β_1 ?
 - This can be calculated by the **standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$**

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

- Standard error can be used to
 - Compute confidence interval
 - Perform hypothesis tests

Confidence Interval



- **Standard errors** can be used to compute **confidence intervals (CI)**
 - E.g., A 95% CI for β_1 means with 95% probability, the range will contain the true unknown value of β_1
 - E.g., A 2.5% CI for β_0 means with 2.5% probability, the range will contain the true unknown value of β_0
- For linear regression, the 95% CI for β_1 approximately takes the form $\left[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1) \right]$
- The same applies to β_0
- Example: in the TV advertising data
 - 95% CI for β_0 is [6.130,7.935], 95% CI for β_1 is [0.042,0.053]

Hypothesis Tests



- Is $\beta_1=0$ or not? If we can't be sure that $\beta_1 \neq 0$ then there is no point in using X as our predictor
 - Use a hypothesis test to answer this question

Hypothesis Tests



- Is $\beta_1=0$ or not? If we can't be sure that $\beta_1 \neq 0$ then there is no point in using X as our predictor
 - Use a hypothesis test to answer this question
- Hypothesis tests
 - Null hypothesis
 - H_0 : There is no relationship between X and Y ($H_0: \beta_1 = 0$)
 - Alternative hypothesis
 - H_a : There is some relationship between X and Y ($H_a: \beta_1 \neq 0$)

Hypothesis Tests



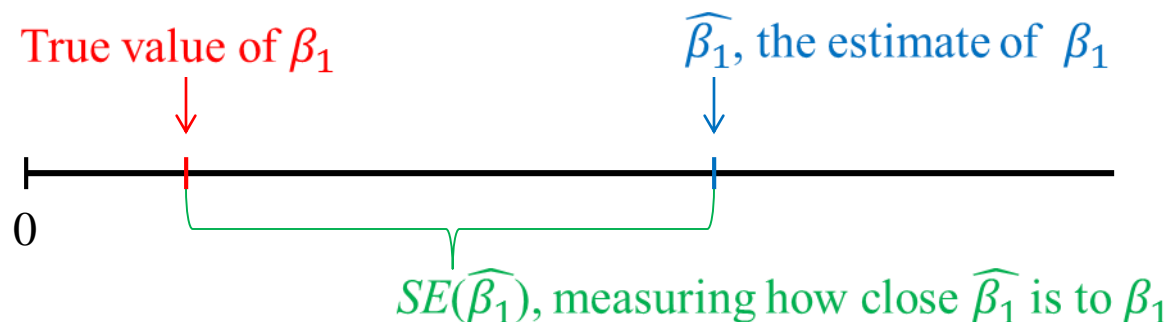
- Is $\beta_1=0$ or not? If we can't be sure that $\beta_1 \neq 0$ then there is no point in using X as our predictor
 - Use a hypothesis test to answer this question
- Hypothesis tests
 - Null hypothesis
 - H_0 : There is no relationship between X and Y ($H_0: \beta_1 = 0$)
 - Alternative hypothesis
 - H_a : There is some relationship between X and Y ($H_a: \beta_1 \neq 0$)
 - To test whether $\hat{\beta}_1$, the estimate of β_1 , is sufficiently far from 0
 - How far is far enough? Compute t-value

t-value

- How far is $\widehat{\beta}_1$, the estimate of β_1 , sufficiently far from 0?
 - This depends on the accuracy of $\widehat{\beta}_1$, that is, the standard error of β_1 .
 - Recall: $SE(\widehat{\beta}_1)$ measures how close $\widehat{\beta}_1$ is to the true value β_1 .

t-value

- How far is $\widehat{\beta}_1$, the estimate of β_1 , sufficiently far from 0?
 - This depends on the accuracy of $\widehat{\beta}_1$, that is, the standard error of β_1 .
 - Recall: $SE(\widehat{\beta}_1)$ measures how close $\widehat{\beta}_1$ is to the true value β_1 .
 - If $SE(\widehat{\beta}_1)$ is small, then even relatively small values of $\widehat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$, and hence there is a relationship between X and Y.
 - If $SE(\widehat{\beta}_1)$ is large, then $\widehat{\beta}_1$ must be large in absolute value in order to claim that there is a relationship between X and Y.



$$t = \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$$

- The higher t-value is, the more possible X and Y are related

Understanding p -value



- Example: to test the hypothesis
 H : the average height of the male students at BBK is 178cm
 - Sample: 100 students with sample mean 185cm
 - p -value is, for instance, 0.06
- How to interpret $p=0.06$?
 - Assume the hypothesis H is true
 - Compute the probability that the sample mean is greater than 185cm given the hypothesis is true, i.e.,
$$P(\text{sample mean} \geq 185\text{cm} \mid \text{population mean} = 178\text{cm}) = 0.06$$

Understanding p -value



- $P(\text{sample mean} \geq 185\text{cm} \mid \text{population mean} = 178\text{cm}) = 0.06$
 - If we were repeat our experiments many, many times
 - each time 100 students selected randomly and calculate sample mean
 - 6 times out of 100 we can expect to see a sample mean $\geq 185\text{cm}$
- $p=0.06$ indicates one of the two things have happened:
 - (A) Either our hypothesis is correct and an extremely unlikely event has occurred (e.g., all 100 students are student athletes), or
 - (B) Our assumption is incorrect and the sample we have obtained is not that unusual.
- Traditionally, we choose a cut-off for p to be 0.05
 - (A) if $p > 0.05$ and (B) if $p \leq 0.05$

***t*-value and *p*-value**

- $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
- t-value (or t-statistics) measures the number of standard deviations away from 0
- p-value measures the probability of observing any value $\geq |t|$, assuming $\beta_1 = 0$
- If t is large (equivalently p-value is small) we can be sure that $\beta_1 \neq 0$ and that there is a relationship

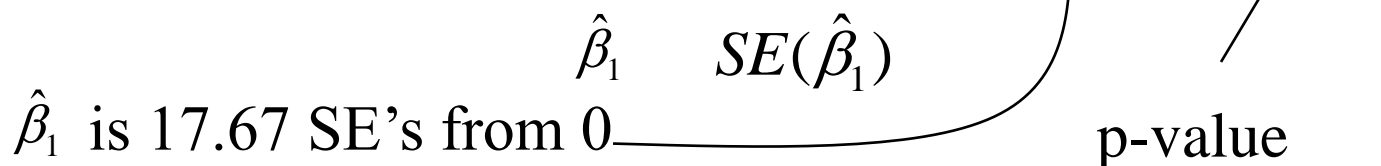
***t*-value and *p*-value**

- $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
- *t*-value (or *t*-statistics) measures the number of standard deviations away from 0
- *p*-value measures the probability of observing any value $\geq |t|$, assuming $\beta_1 = 0$
- If *t* is large (equivalently *p*-value is small) we can be sure that $\beta_1 \neq 0$ and that there is a relationship

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$ is 17.67 SE's from 0 ————— *p*-value



***t*-value and *p*-value**

- $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
- *t*-value (or *t*-statistics) measures the number of standard deviations away from 0
- *p*-value measures the probability of observing any value $\geq |t|$, assuming $\beta_1 = 0$
- If *t* is large (equivalently *p*-value is small) we can be sure that $\beta_1 \neq 0$ and that there is a relationship

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$

$SE(\hat{\beta}_1)$

$\hat{\beta}_1$ is 17.67 SE's from 0

p-value

How far is far enough?

Typical *p*-value cutoffs for rejecting the null hypothesis are 5 or 1%.

t-value and p-value



- The t-test produces a single value, t , which grows larger as the difference between the means of two samples grows larger;
- t does not cover a fixed range such as 0 to 1 like probabilities do;
- You can convert a t-value into a probability, called a p-value;
- The p-value is always between 0 and 1 and it tells you the probability of the difference in your data being due to sampling error;
- The p-value should be lower than a chosen significance level (0.05 for example) before you can reject your null hypothesis.

Overview of Step 1



- Step 1: use training data to estimate coefficients
 - How to estimate?
 - Assessing the accuracy of the coefficient estimates
 - Comparing coefficients only
 - Assessing the accuracy of the model
 - Quantifying the extent to which the model fits the data

Measures of Fit: RSE

- Recall:

Population regression line:

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

Least squares line:

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

Measures of Fit: RSE

- Recall:

Population regression line:

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

Least squares line:

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

- Measuring the extent to which the model fits the data
 - **Residual Standard Error (RSE)**
 - Even if it is a true regression line ($\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$), we would not be able to perfectly predict Y from X due to the *error term ε*

Measures of Fit: RSE

- Recall:

Population regression line:

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

Least squares line:

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

- Measuring the extent to which the model fits the data
 - **Residual Standard Error (RSE)**
 - Even if it is a true regression line ($\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$), we would not be able to perfectly predict Y from X due to the *error term ε*
 - RSE is the standard deviation of ε
 - Quantifies average amount that the response will deviate from the true regression line

Measures of Fit: RSE



- Measuring the extent to which the model fits the data
 - Residual Standard Error (RSE)
 - Example: regressing number of units sold on TV advertising budget
 - $RSE = 3.26$
 - Even if the model were correct, any prediction on sales on the basis of TV advertising budget would still be off by about 3260 units on average
 - An absolute measure of lack of fit of the model to the data
 - Measured in the units of Y
 - Not always clear whether it is a good fit

Measures of Fit: R^2



- Measuring the extent to which the model fits the data
 - R^2 statistic
 - Some of the variation in Y can be explained by variation in the X 's and some cannot.
 - R^2 tells you the proportion of variance that can be explained by X .

$$R^2 = 1 - \frac{RSS}{\sum (Y_i - \bar{Y})^2} \approx 1 - \frac{\text{Ending Variance}}{\text{Starting Variance}}$$

- Starting variance: the amount of variability inherent in the response before the regression is performed
- Ending variance: the amount of variability that is left unexplained after performing regression

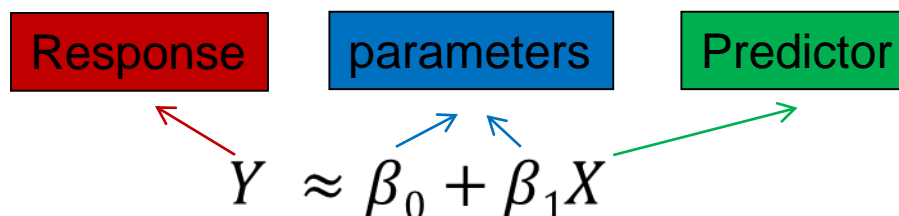
Measures of Fit: R^2



- Measuring the extent to which the model fits the data
 - R^2 statistic
 - R^2 is always between 0 and 1.
 - Zero means no variance has been explained.
 - One means it has all been explained (perfect fit to the data).
 - In simple linear regression, $R^2 = \text{Cor}(X, Y)^2$
 - Both measure the linear relationship between X and Y

Simple Linear Regression

To predict a quantitative response Y on the basis of a single predictor variable X .



We are regressing Y on X .

Step1: ← Done!

Use the training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Step2: ← Now!

Use $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ to predict Y (as \hat{y}) on the basis of $X = x$

An Example: Body Fat and Waist Size

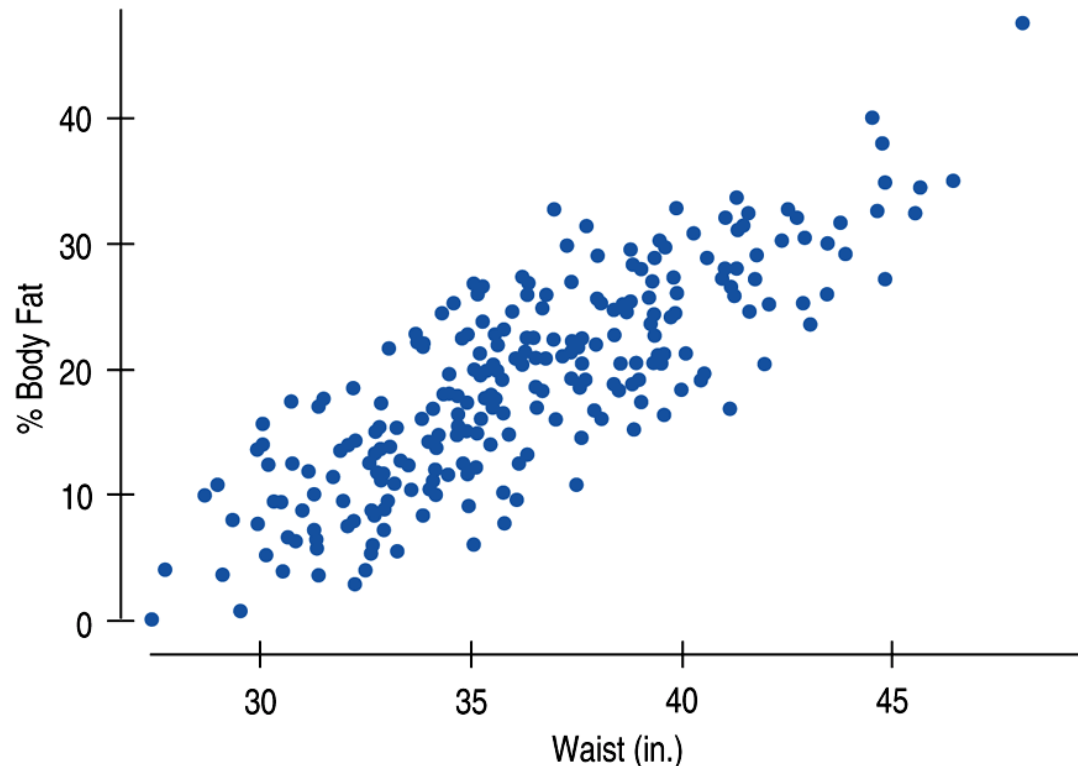
- Investigating the relationship in adult males between
 - Y : % *Body Fat* and X : *Waist size* (in inches).



An Example: Body Fat and Waist Size



- Investigating the relationship in adult males between
 - *Y: % Body Fat* and *X: Waist size* (in inches).
- Here is a scatterplot of the data for 250 adult males of various ages:



Confidence Intervals and Prediction Intervals for Predicted Values



- For our *%body fat* and *waist size* example, there are two questions we could ask:
 1. Do we want to know the mean *%body fat* for *all men* with a *waist size* of, say, 38 inches? → predicting for a mean
 2. Do we want to estimate the *%body fat* for *a particular man* with a 38-inch *waist*? → predicting for an individual
- **The predicted *%body fat* is the same in both questions**, but we can predict the *mean %body fat* for *all men* whose *waist size* is 38 inches with **a lot more precision** than we can predict the *%body fat* of *a particular individual* whose *waist size* happens to be 38 inches.

Confidence Intervals and Prediction Intervals for Predicted Values

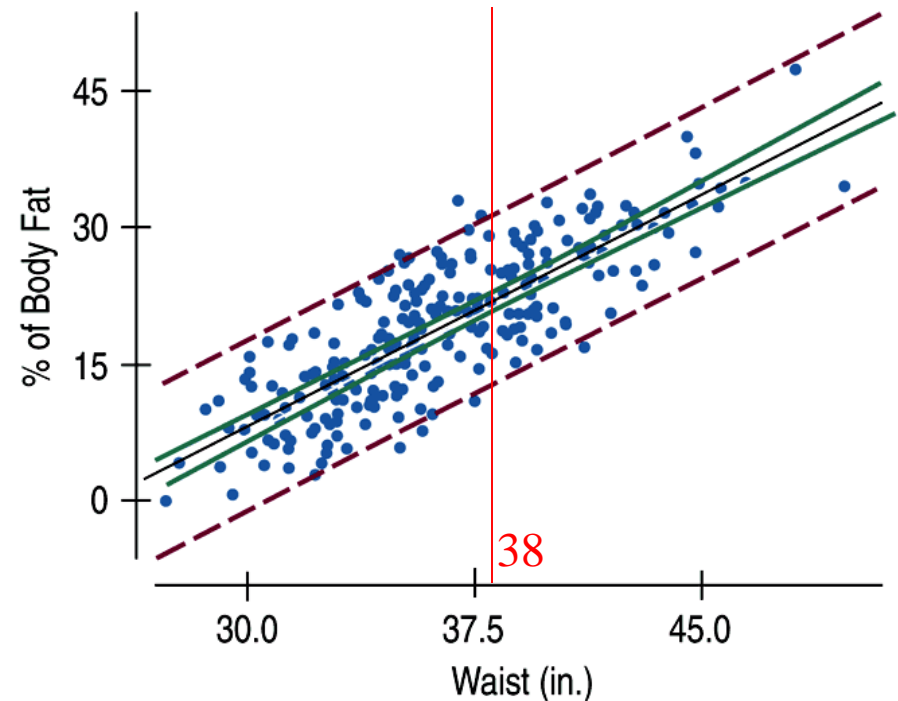


- We start with the same prediction in both cases.
 - We are predicting for a new individual, one that was not in the original data set.
 - Call his x -value x_v .
 - The regression predicts *%body fat* as

$$\hat{y}_v = \hat{\beta}_0 + \hat{\beta}_1 x_v$$

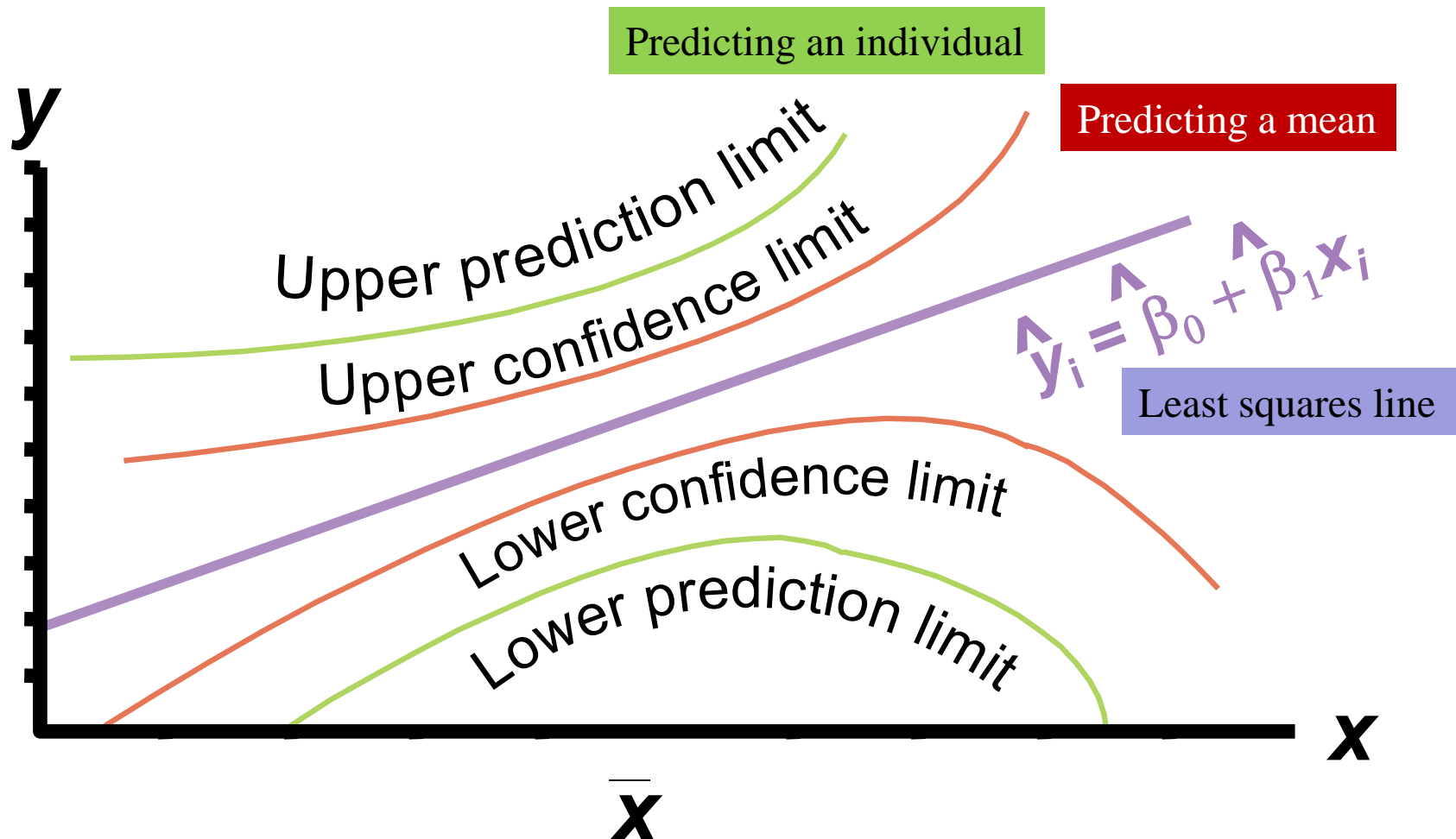
Confidence/Prediction Intervals for Predicted Values

- Here's a look at the difference between **predicting for a mean** and **predicting for an individual**.
- The solid green lines near the regression line show the 95% **confidence intervals for the mean** predicted value, and the dashed red lines show the **prediction intervals for individuals**.
- The solid green lines and the dashed red lines curve away from the least squares line as x moves farther away from \bar{x} .



Prediction interval (PI) is an estimate of an interval in which **future observations (particular individuals)** will fall, with a certain probability, given what has already been observed.

Confidence Intervals vs. Prediction Intervals



Conclusion



- Simple Linear Regression
 - Supervised Learning
 - Prediction
 - Parameterised method
- Variables
 - y = **Dependent** variable (quantitative)
 - x = **Independent** variable (quantitative)
- Least Squares Line
 - average error of prediction = 0
 - sum of squared errors is minimum

Conclusion



- Practical Interpretation of y -intercept
 - predicted y value when $x = 0$
 - no practical interpretation if $x = 0$ is either nonsensical or outside range of sample data
- Practical Interpretation of Slope
 - Increase or decrease in y for every 1-unit increase in x
- Analysis of Regression
 - RSE, R^2 -statistic, p -value, Confidence Interval, Prediction Interval

LAB

Simple Linear Regression

Load libraries



- `library()` function loads libraries, or groups of functions and data sets that are not included in the base `R` distribution.
 - Basic functions for least squares linear regression and other simple analysis → included in the base distribution
 - `MASS` package, which is a very large collection of data sets and functions
 - `ISLR` package, includes the data sets associated with the textbook

```
> library(MASS)
```

```
> library(ISLR)
```

```
Error in library(ISLR) : there is no package called 'ISLR'
```

```
> install.packages("ISLR")
```

```
# or select the Install package option under the Package tab
```

```
> library(ISLR)
```

The Boston House Data



- The data set records median house value (**medv**) for 506 neighbourhoods around Boston.
- We will seek to predict **medv** using 13 predictors such as
 - **rm**: average number of rooms per house
 - **age**: average age of houses
 - **lstat**: percentage of households with low socio-economic status

```
> fix(Boston)
> names(Boston)
[1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"
[10] "tax"     "ptratio" "black"   "lstat"   "medv"
> ?Boston
> # open the web page to find out about the data set
```

lm() to Fit Simple LR Models



- Using `lm()` to fit a simple linear regression model
 - The response (y): `medv`
 - The predictor (x): `lstat`
 - Basic syntax: `lm(y~x, data)`

```
> lm.fit=lm(medv~lstat)
```

```
Error in eval(expr, envir, enclos) : object 'medv' not found
```

```
# we need to let R know where to find the variables medv and lstat
```

```
# we have two ways to solve this:
```

```
# first way: indicate where the variables are in the lm func
```

```
> lm.fit=lm(medv~lstat,data=Boston)
```

```
# second way: attach the dataset
```

```
> attach(Boston)
```

```
> lm.fit=lm(medv~lstat)
```


Check model details



```
> lm.fit                # basic information
Call:
lm(formula = medv ~ lstat)
Coefficients:
(Intercept)          lstat
      34.55         -0.95          # medv = -0.95 * lstat + 34.55
```

```
> summary(lm.fit)      # more details
```

Call:

```
lm(formula = medv ~ lstat)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.168	-3.990	-1.318	2.034	24.500

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	34.55384	0.56263	61.41	<2e-16 ***
lstat	-0.95005	0.03873	-24.53	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: **6.216** on 504 degrees of freedom

Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432

F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

How to read the results?

Extract Quantities



- Use `names(lm.fit)` to find out what other pieces of information are stored in `lm.fit`

```
> names(lm.fit)
[1] "coefficients" "residuals" "effects" "rank" "fitted.values" "assign"
[7] "qr" "df.residual" "xlevels" "call" "terms" "model"
```

- How to extract the quantities?
 - By name: e.g., `lm.fit$coefficients`
 - By the extractor functions: e.g., `coef(lm.fit)`

```
> lm.fit$coefficients
(Intercept)          lstat
 34.5538409   -0.9500494
```

```
> coef(lm.fit)
(Intercept)          lstat
 34.5538409   -0.9500494
```

Obtaining CI and PI



- To obtain a confidence interval for the coefficient estimates:

```
> confint(lm.fit)
                2.5 %      97.5 %
(Intercept) 33.448457 35.6592247
lstat      -1.026148 -0.8739505
```

- To obtain a confidence and prediction interval for the prediction of medv for a given value of lstat.

```
> predict(lm.fit, data.frame(lstat=c(5,10,15))), interval="confidence")
```

```
      fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
```

How to read the results?

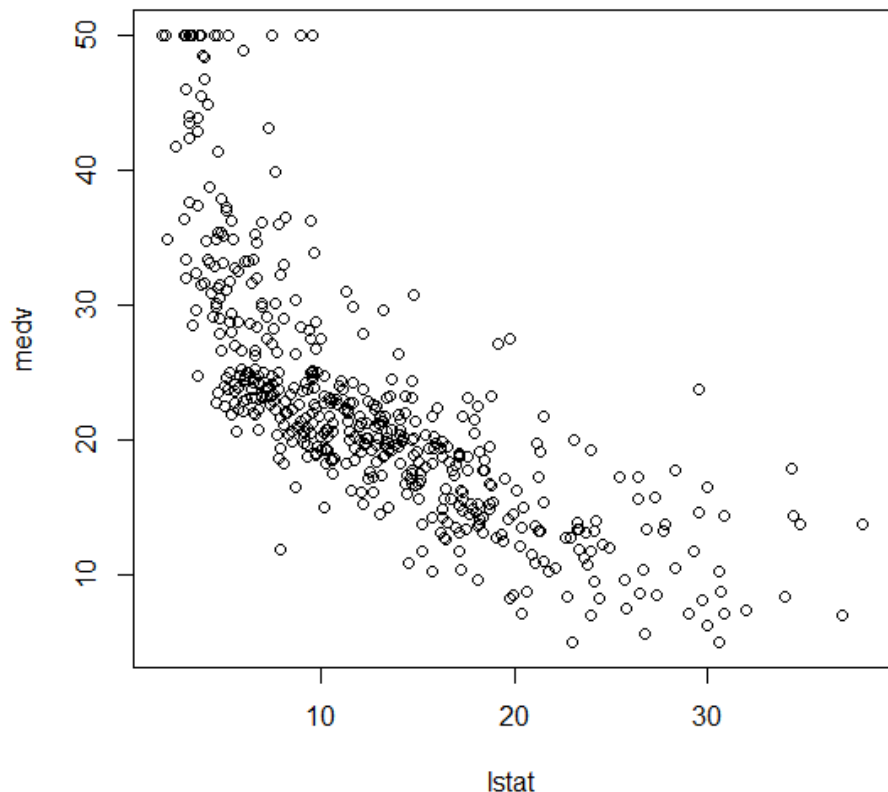
```
> predict(lm.fit, data.frame(lstat=c(5,10,15))), interval="prediction")
```

```
      fit      lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

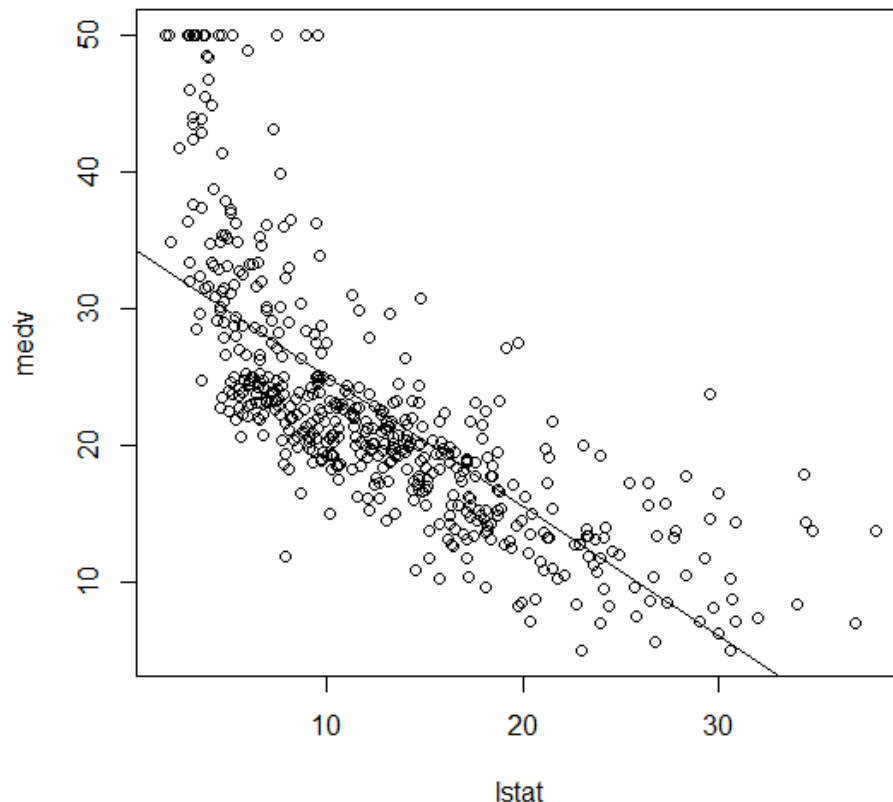
Which interval is wider?

Plot the results

```
> plot(lstat,medv)
```



```
> abline(lm.fit)
```

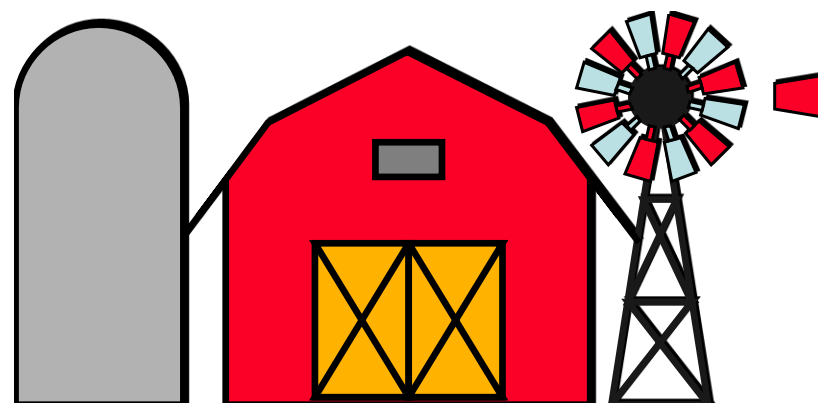


Try out other options on the width of the regression line, colour, symbols, etc
`abline(lm.fit, lwd=3,col="red", pch="+")`, ...

Least Squares - Exercise

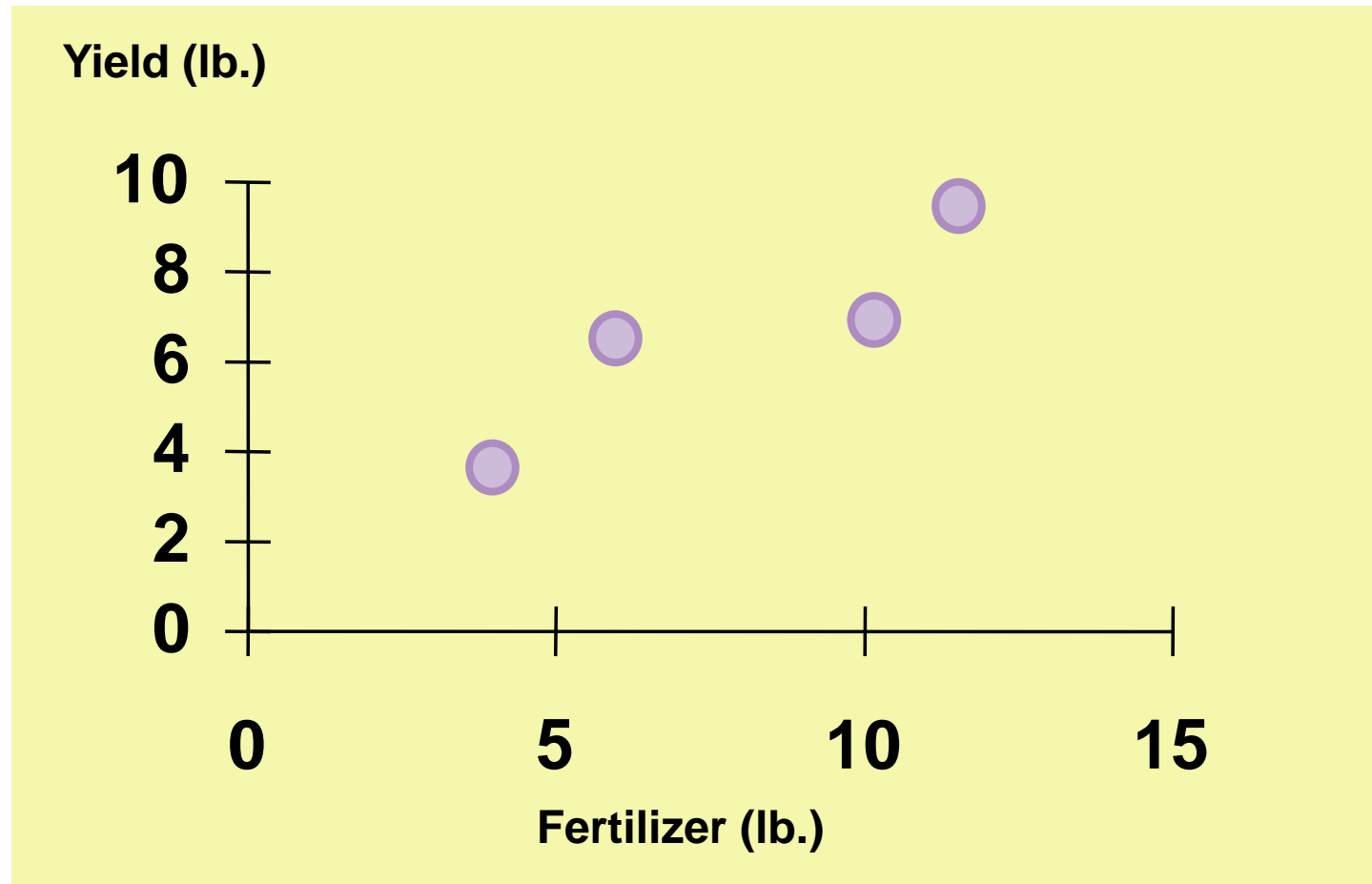
You're an economist for the county cooperative. You gather the following data:

<u>Fertilizer (lb.)</u>	<u>Yield (lb.)</u>
4	3.0
6	5.5
10	6.5
12	9.0



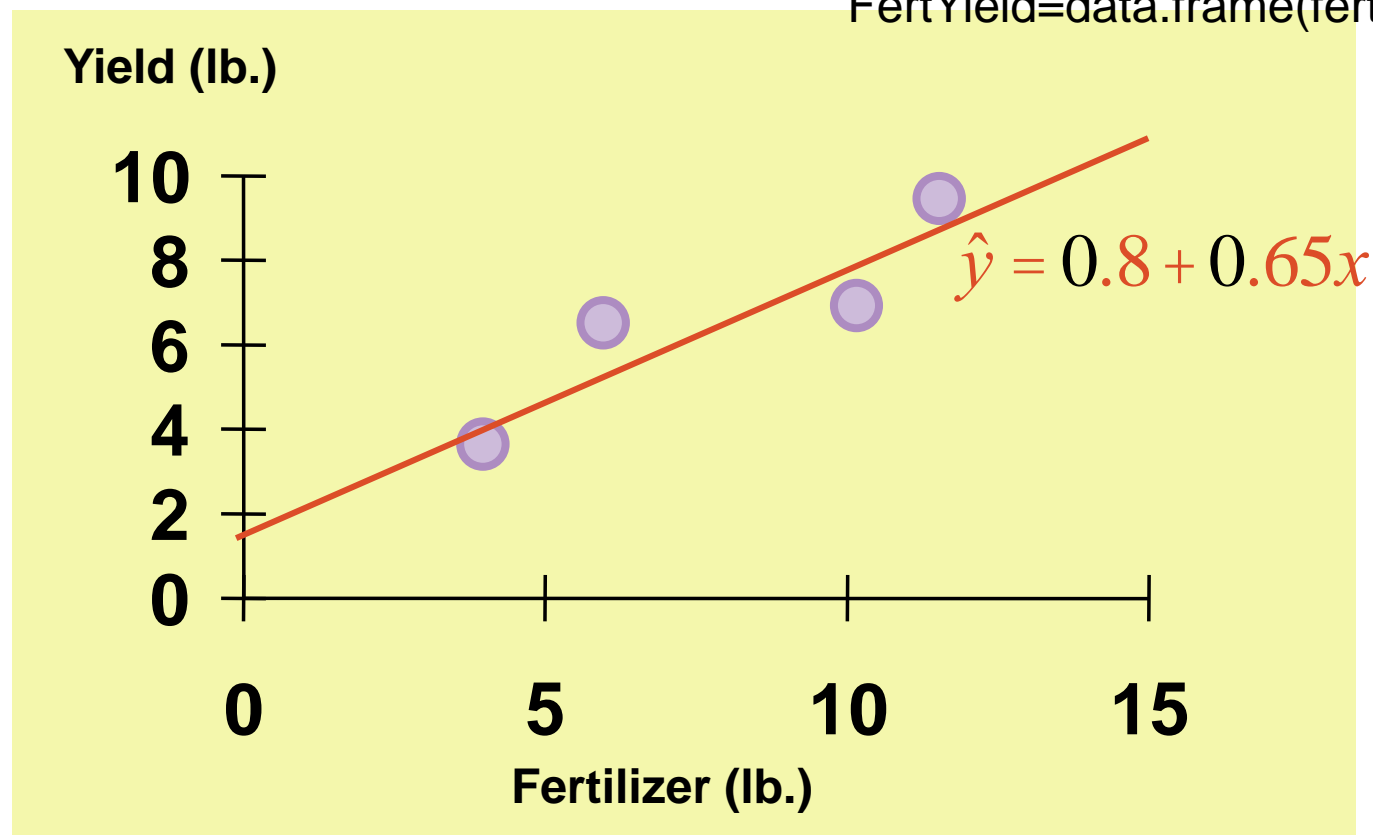
Find the **least squares line** relating crop yield and fertilizer.

Scatter Plot Crop Yield vs. Fertilizer



Regression Line Fitted to the Data

```
fert=c(4,6,10,12)  
yield=c(3.0,5.5,6.5,9.0)  
FertYield=data.frame(fert,yield)
```



Predict



- Predict the yield when 2.5, 5.5 and 8.5 lb of fertilizer are used
- What is the 95% CI and PI?
 - for the coefficients
 - for the prediction of yield given 2.5, 5.5 and 8.5 lb of fertilizer
- Find the following measures:
 - p value,
 - t value,
 - the RSE,
 - the R^2
- Do you think fert is related with yield? Why?

Exercise:

A Complete Example

-- Try to obtain the results and plots yourself

Example

Suppose a fire insurance company wants to relate **the amount of fire damage** in major residential fires to **the distance** between the burning house and the nearest fire station.

The study is to be conducted in a large suburb of a major city; a sample of **15 recent fires** in this suburb is selected.

The **amount of damage, y** , and the **distance between the fire and the nearest fire station, x** , are recorded for each fire.

Amount of
damage y



distance x



Example

Table 10.5 Fire Damage Data

Distance from Fire Station, x (miles)	Fire Damage, y (thousands of dollars)
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1

Example



Step 1: First, we hypothesise a model to relate fire damage, y , to the distance from the nearest fire station, x .

We hypothesise a simple linear model (*population regression line*):

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Example

Step 2: Use R to estimate the unknown parameters of the hypothesised model. The least squares estimates of the slope β_1 and intercept β_0 , highlighted on the printout, are

$$\hat{\beta}_1 = 4.9193, \hat{\beta}_0 = 10.2779$$

```
Call:
lm(formula = damage ~ distance)

Residuals:
    Min       1Q   Median       3Q      Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915

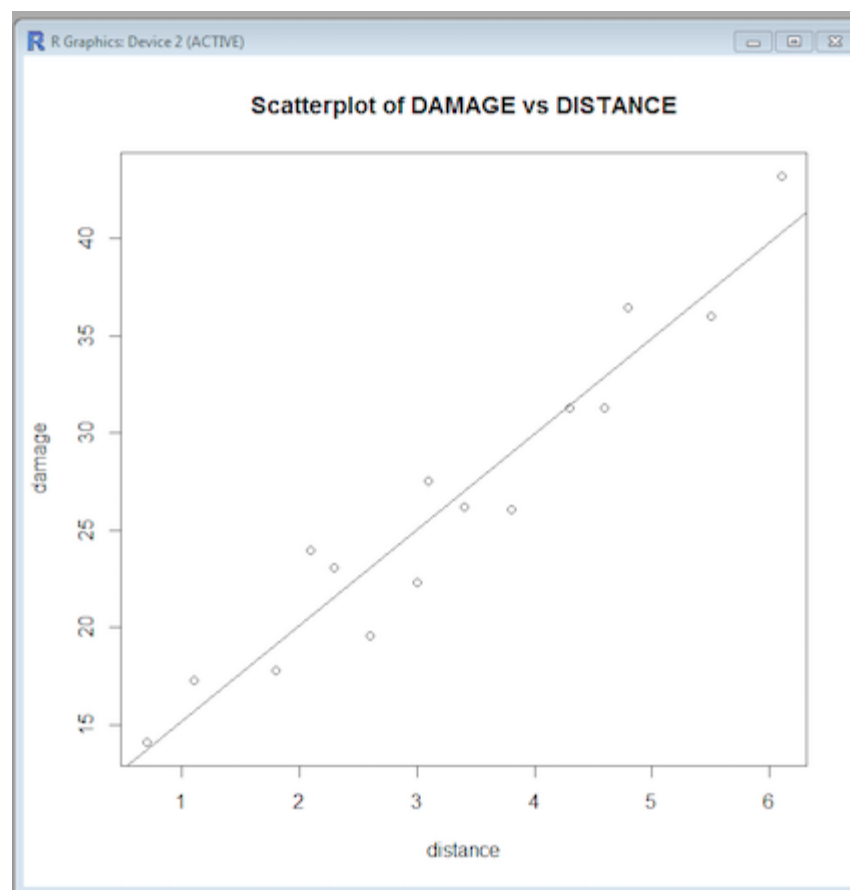
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
distance      4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared:  0.9235,    Adjusted R-squared:  0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Example

Least squares equation: $\hat{y} = 10.278 + 4.919x$

This prediction equation is plotted as:



Example



The least squares estimate of the slope, $\hat{\beta}_1 = 4.919$, implies that the estimated mean damage increases by \$4,919 for each additional mile from the fire station.

This interpretation is valid over the range of x , or from .7 to 6.1 miles from the station.

The estimated y -intercept, $\hat{\beta}_0 = 10.278$, has the interpretation that a fire 0 miles from the fire station has an estimated mean damage of \$10,278.

Example



Step 3: Measuring the extent to which the model fits the data

1) RSE: The estimate of the standard deviation σ of ε , is

$$RSE = 2.31635$$

This implies that any prediction on the observed fire damage (y) values based on the distance would be off by about 2.32 thousand dollars on average when using the least squares line.

2) $R^2 = .9235$, which implies that about 92% of the sample variation in fire damage (y) is explained by the distance (x) between the fire and the fire station.

Example



Step 4: First, test the **null hypothesis** that the slope β_1 is 0 –that is, that there is no linear relationship between fire damage and the distance from the nearest fire station, against the **alternative hypothesis** that fire damage increases as the distance increases.

We test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0$$

t-value = 12.525, p-value = 1.248e-08

t is large and p is small. We can conclude that $\beta_1 \neq 0$ and there is a relationship between fire damage and the distance from the nearest fire station.

Example

The 95% confidence interval yields (4.070, 5.768).

```
> confint(lm.fit, level=0.95)
      (1-level)/2 = 2.5 %      97.5 % = 1-(1-level)/2
(Intercept)          7.209605    13.346252
distance             4.070851    5.767811
```

We estimate (with 95% confidence) that the interval from \$4,070 to \$5,768 encloses the mean increase (β_1) in fire damage per additional mile distance from the fire station.

Example

- To obtain a confidence interval and prediction interval for the prediction of damage for a given value of distance.

```
> predict(lm.fit, data.frame(distance=c(5,10,15)), interval="confidence")
```

	fit	lwr	upr
1	29.80359	29.00741	30.59978
2	25.05335	24.47413	25.63256
3	20.30310	19.73159	20.87461

```
> predict(lm.fit, data.frame(distance=c(5,10,15)), interval="prediction")
```

	fit	lwr	upr
1	29.80359	17.565675	42.04151
2	25.05335	12.827626	37.27907
3	20.30310	8.077742	32.52846