Big Data Analytics

Session 3

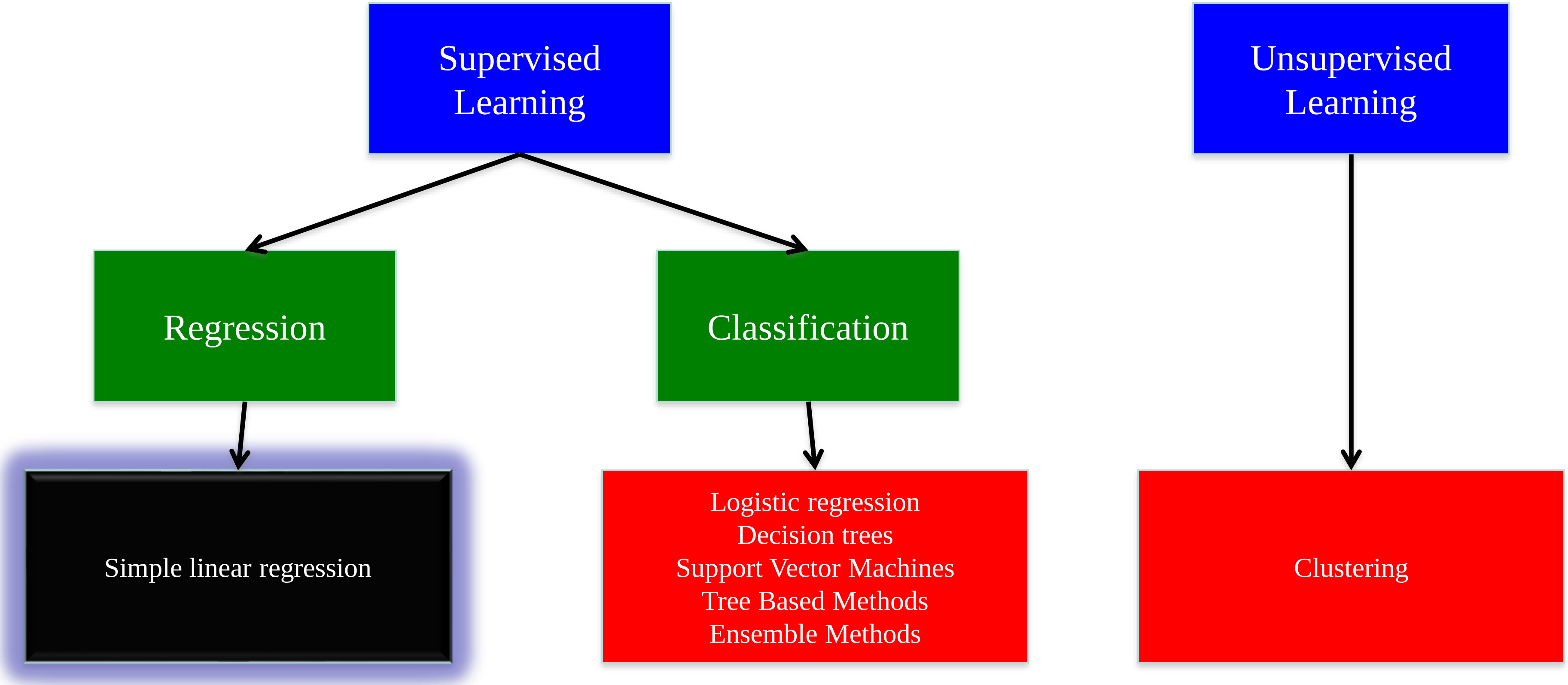
Simple Linear Regression

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



Supervised Learning

Unsupervised Learning

Regression

Classification

Simple linear regression

Logistic regression
Decision trees
Support Vector Machines
Tree Based Methods
Ensemble Methods

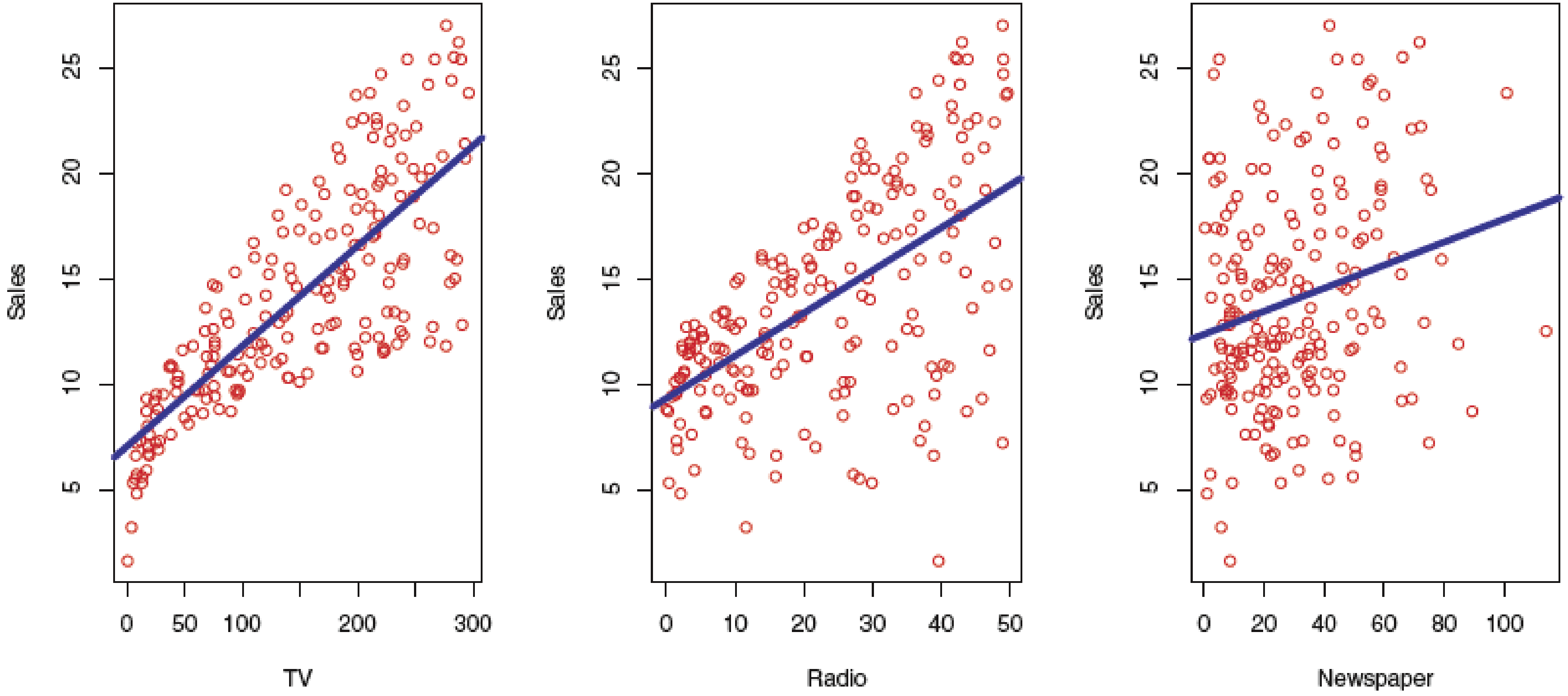
Clustering

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
  - $f(x) = -0.4\, x - 2$
  - $f(x_1, x_2) = 4\, x_1 + 5^3 x_2 - 7$
  - $f(x_1, x_2, x_3) = -7\, x_1 + 5\, x_2 - \sqrt{2}\, x_3 - 1$
- 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?
  - Yes, let's assume $x_1^2 = 4$ and $x_2 = 3$. Then $f(a,b)=4a+3b$

First-Order Linear Functions

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

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

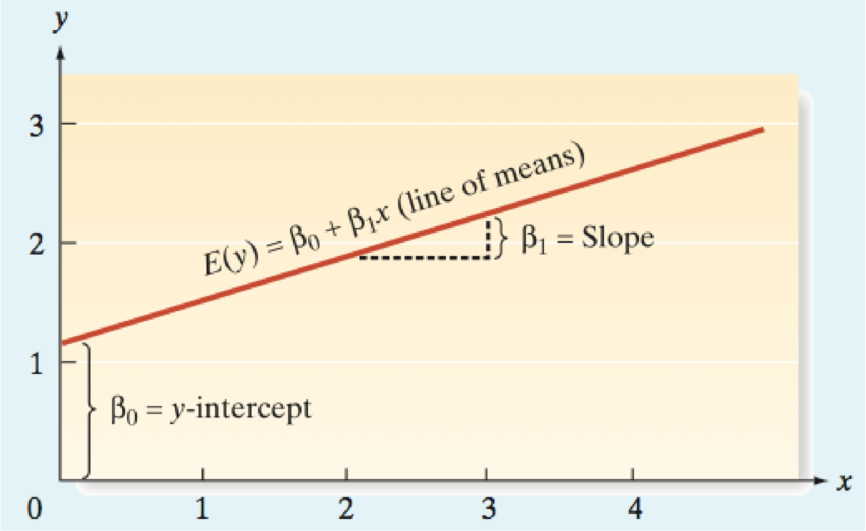
where

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*



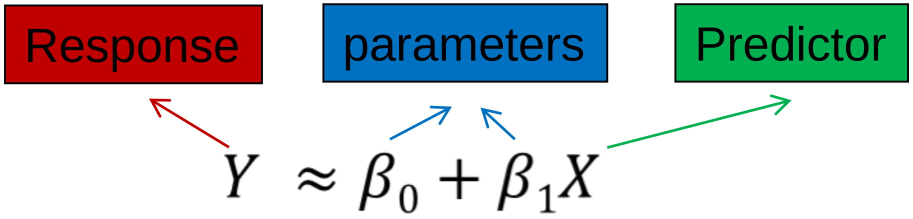
Outline

- 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, optional)
  - *More than one predictor variable*: $Y \sim X_1, X_2, \ldots$
  - 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.



Response

parameters

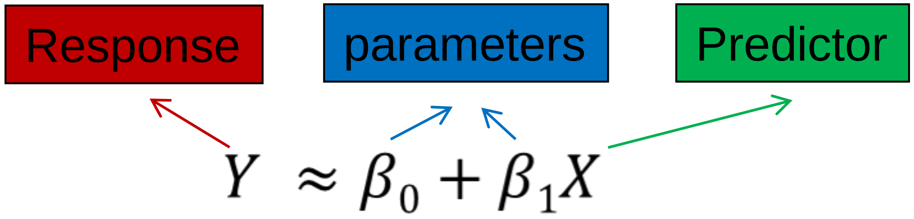
Predictor

$$Y \approx \beta_0 + \beta_1 X$$

We are regressing Y on X.

Simple Linear Regression

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



Response

parameters

Predictor

$$Y \approx \beta_0 + \beta_1 X$$

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

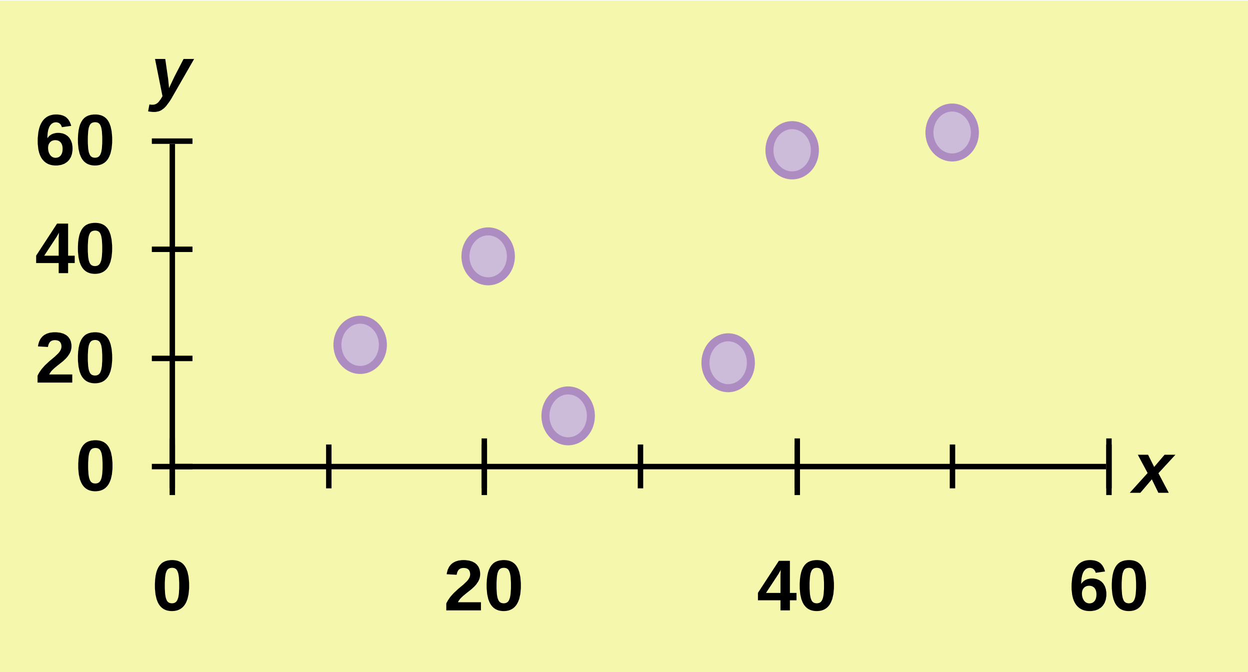
- Step 1: use training data to estimate coefficients (parameters)
  - How to estimate?
  - Assessing the accuracy of the coefficient estimates
  - Assessing the accuracy of the model

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

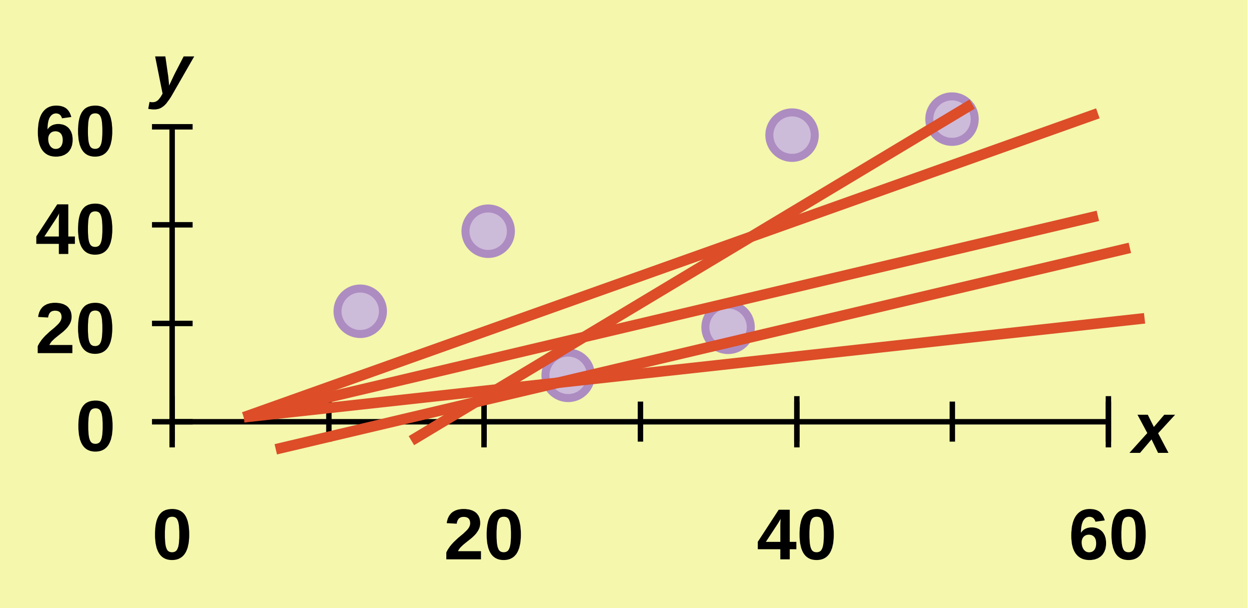
Plotting Training Data

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



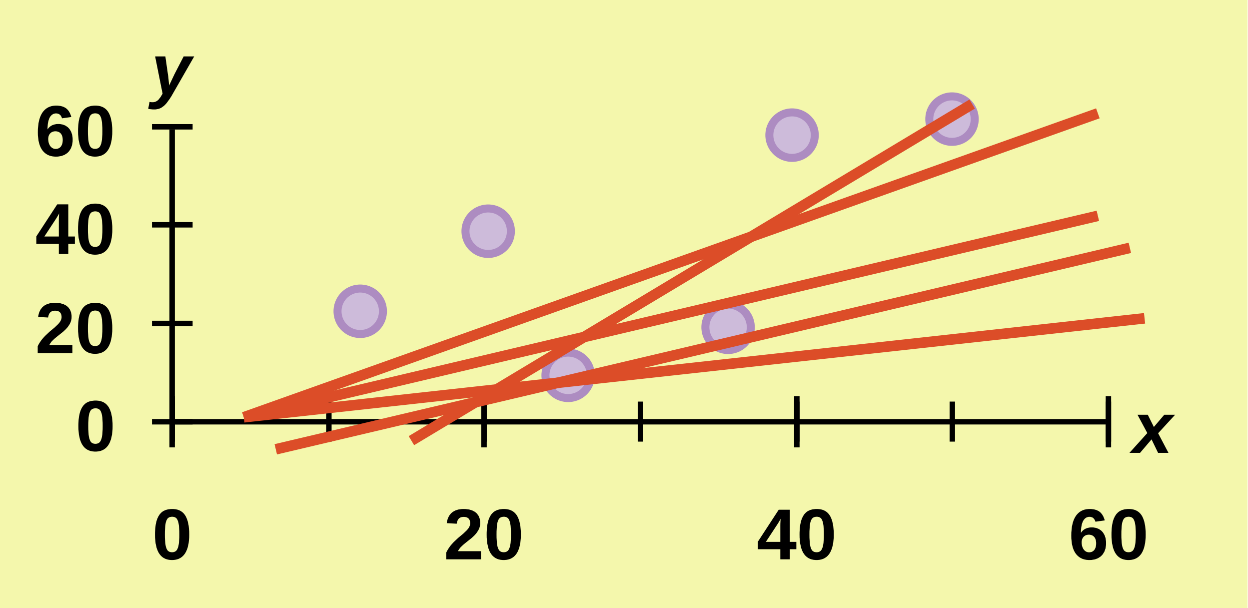
How to fit?

- How would you draw a line through the points?



How to fit?

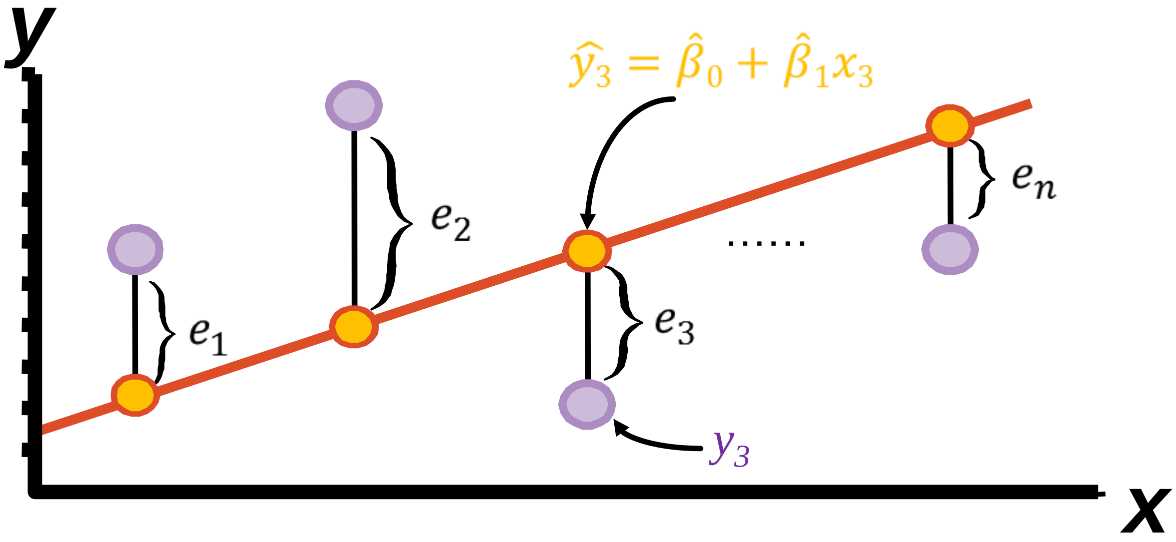
- How would you draw a line through the points?
- How do you determine which line ‘fits best’?



Residual Sum of Squares

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

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



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

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

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

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

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

- 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:

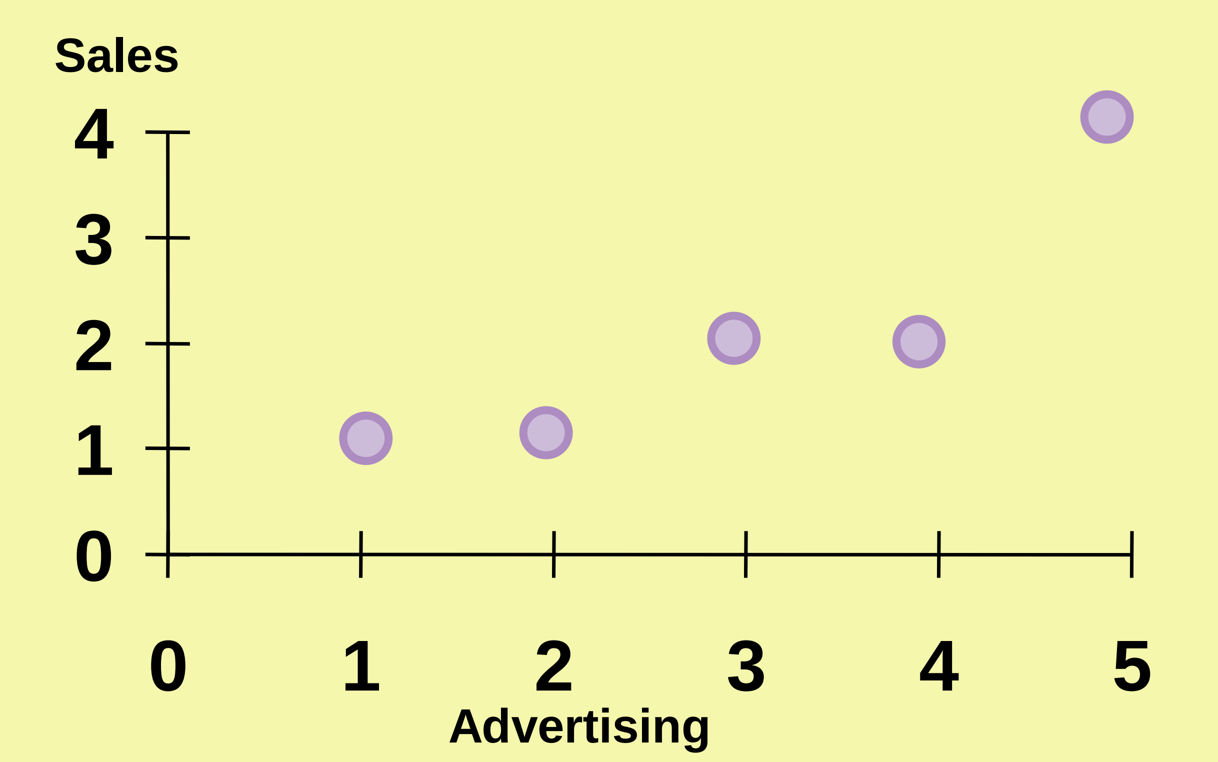
Ad Expenditure (100£)	Sales (Units)
1	1
2	1
3	2
4	2
5	4

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

Scatter Plot -- Sales vs. Advertising

- Plot it

Ad Expenditure (100£)	Sales (Units)
1	1
2	1
3	2
4	2
5	4



Minimising RSS

- Recall:

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

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

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

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

Minimising RSS

Ad Expenditure (100£)	Sales (Units)
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:

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

- Recall:

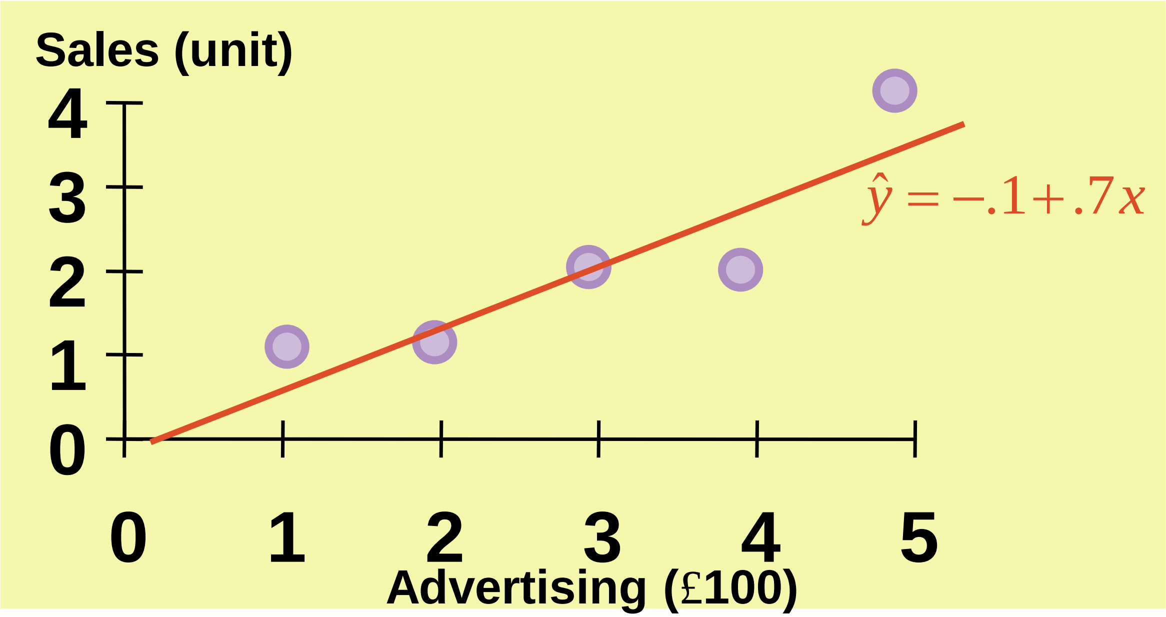
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

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

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

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: $Y = \beta_0 + \beta_1 X + \varepsilon$
    - 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; (reducible error)
      - There may be measurement error (irreducible error)
    - Assume that ε is independent of X
    - *The best linear approximation to the true relationship between X and Y*

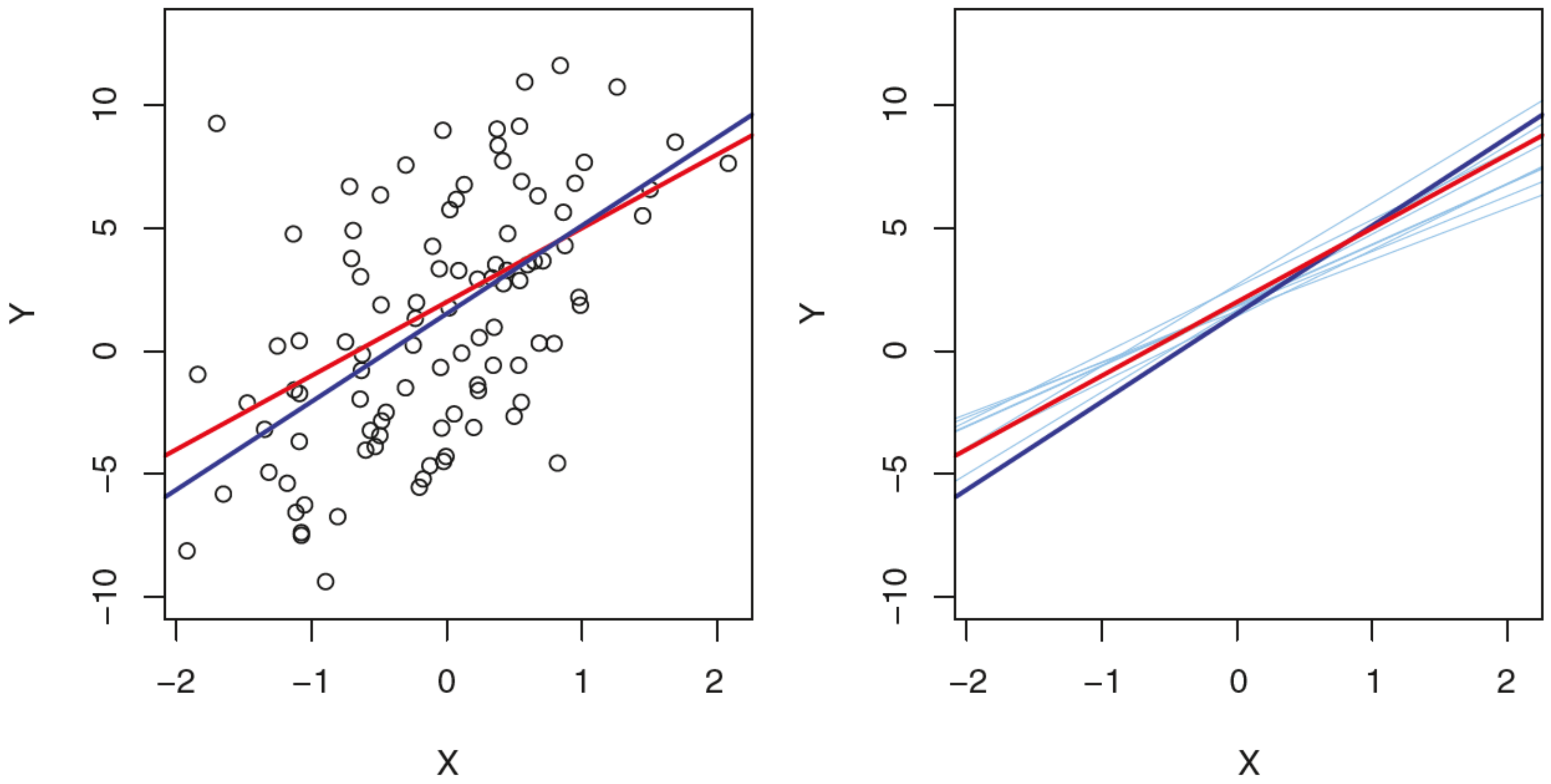
Assessing the accuracy of coefficient estimates

- Three different lines:
  - True relationship: $Y = f(X) + \epsilon$
    - ϵ is a mean-zero random error term
  - Population regression line: $Y = \beta_0 + \beta_1 X + \varepsilon$
    - 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; (reducible error)
      - 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 μ ➔ write $\hat{\mu} = \bar{x}$
  - $\hat{\mu}$ is the estimate of μ

An Analogue



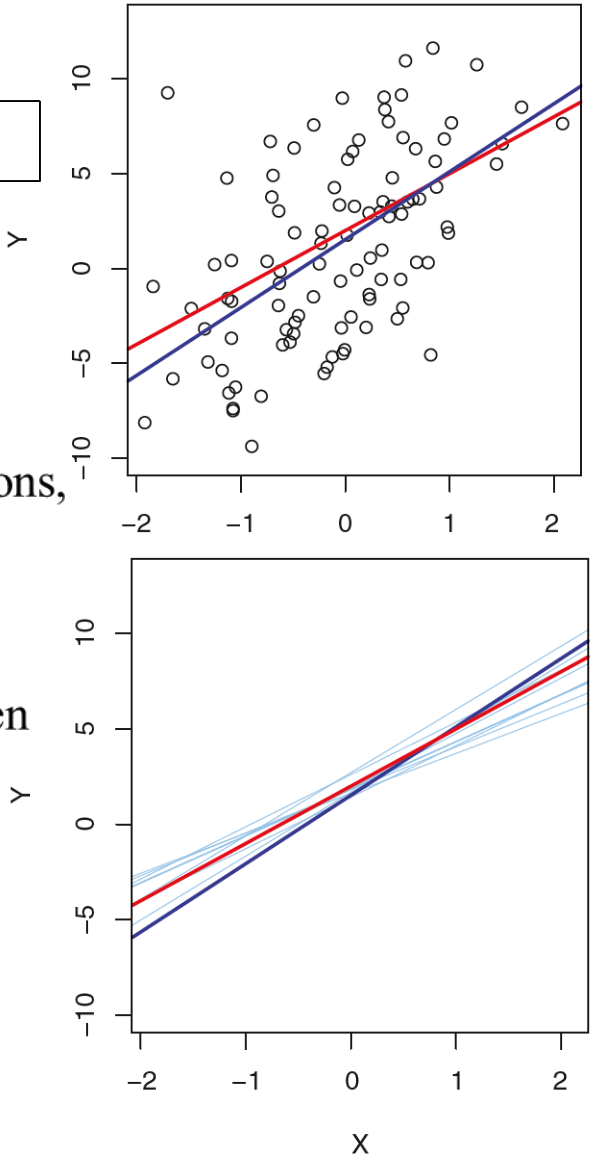
Red line: population regression line *f*(X) = 2+3X, usually unknown

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: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Least squares line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



- 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 a huge number of the parameters, then the resulting $\hat{\beta}_0$ and $\hat{\beta}_1$ will be the accurate population regression line parameters

Standard Error

- How close is a single sample mean $\hat{\mu}$ to the population mean μ?
  - Use standard error (SE): the average amount that this estimate $\hat{\mu}$ differs from μ
  - $$\mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$
    ← 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: $\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$
- Least squares line: $\boxed{\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X}$

- How close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true value β_0 and β_1?

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

Overview of Step 1

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

Hypothesis Tests

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

- 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

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

- 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

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

- 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: $\beta_1 \neq 0$)

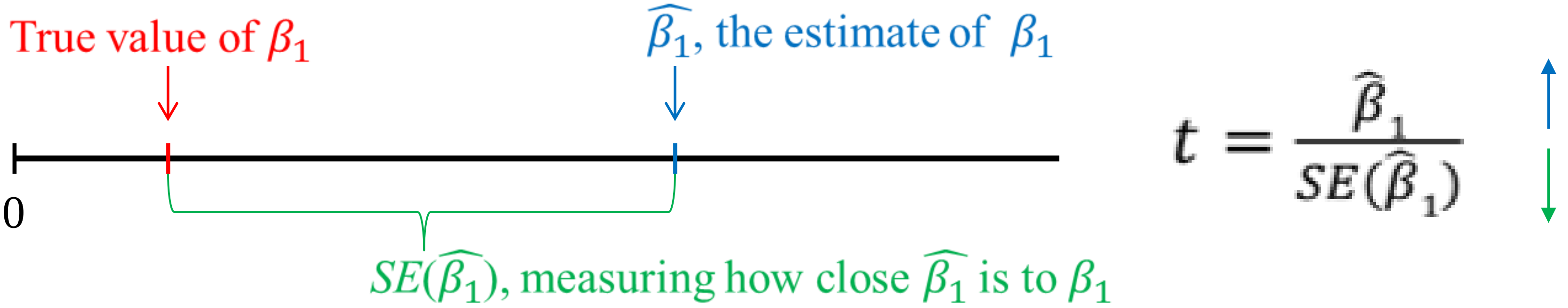
  - To test whether $\hat{\beta}_1$, the estimate of β_1, is sufficiently far from 0
    - How far is far enough? Compute t-value

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

t-value does not have a fixed range! Convert it to a p-value

***P*-value**

- Given a t-value, we can calculate a p-value (a probability, between 0 and 1).

- P values address only one question: how likely are your data, assuming a true null hypothesis?

- P values evaluate how well the sample data support that the null hypothesis is true. It measures how compatible your data are with the null hypothesis
  - A small *p*-value (typically ≤ 0.05) indicates your sample provides strong evidence against the null hypothesis, so you reject the null hypothesis.
  - A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
  - *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

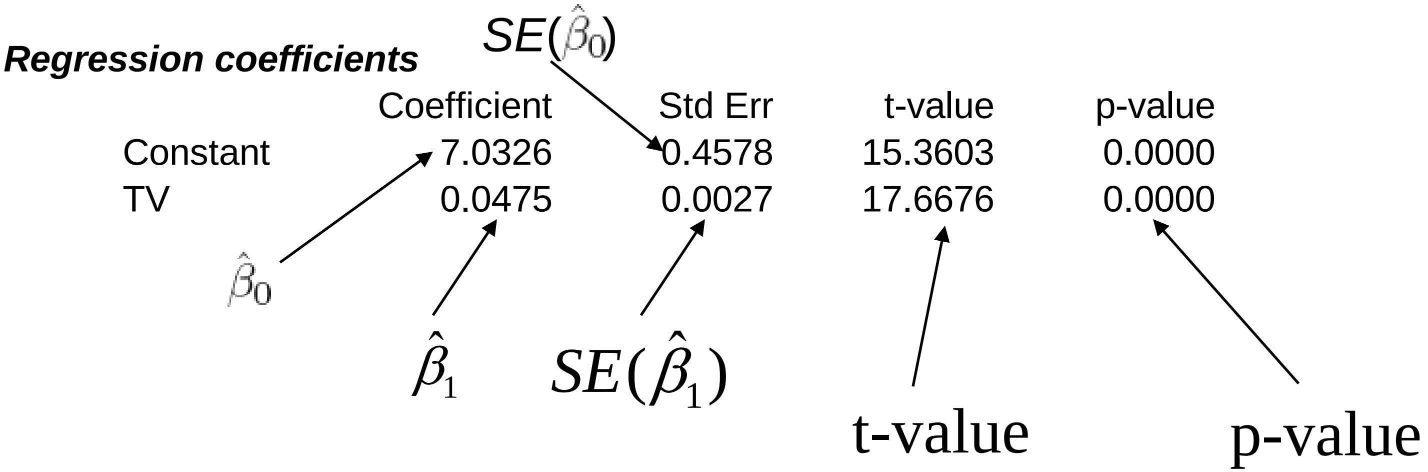
- P values <u>do not</u> measure support for the alternative hypothesis.

t-value and *p*-value

If t is large (equivalently p-value is small), we can be sure that $\hat{\beta}_1$ is not 0.

- ➔ We reject the Null Hypothesis.
- ➔ We declare a relationship to exist between X and Y.

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



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

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

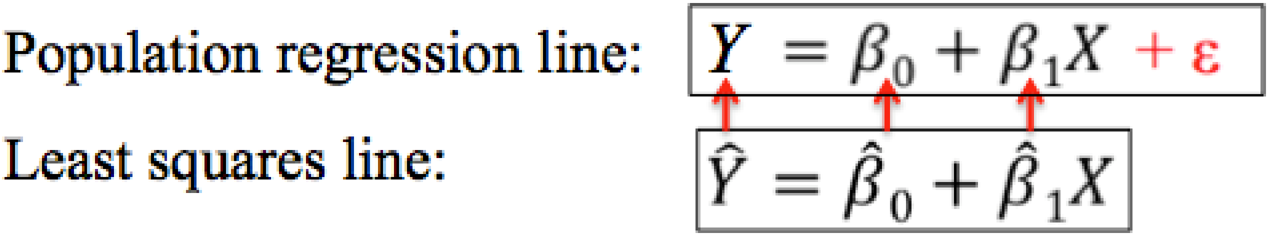
Measures of Fit: RSE

- Recall:

$$\begin{array}{ll} \text{Population regression line:} & \boxed{Y = \beta_0 + \beta_1 X + \varepsilon} \\ & \uparrow \qquad \uparrow \qquad \uparrow \\ \text{Least squares line:} & \boxed{\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X} \end{array}$$

Measures of Fit: RSE

- Recall:



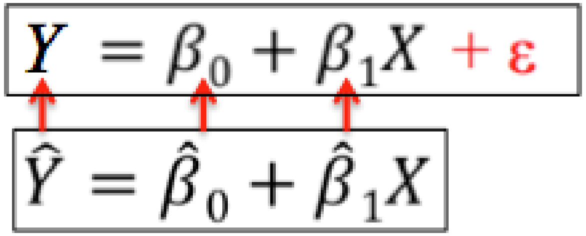
Population regression line: $Y = \beta_0 + \beta_1 X + \varepsilon$

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 estimate of the standard deviation of **ε**
      - Quantifies average amount that the response will deviate from the population 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, $R^2 = \mathrm{Cor}(X,Y)^2$
      - Both measure the linear relationship between X and Y

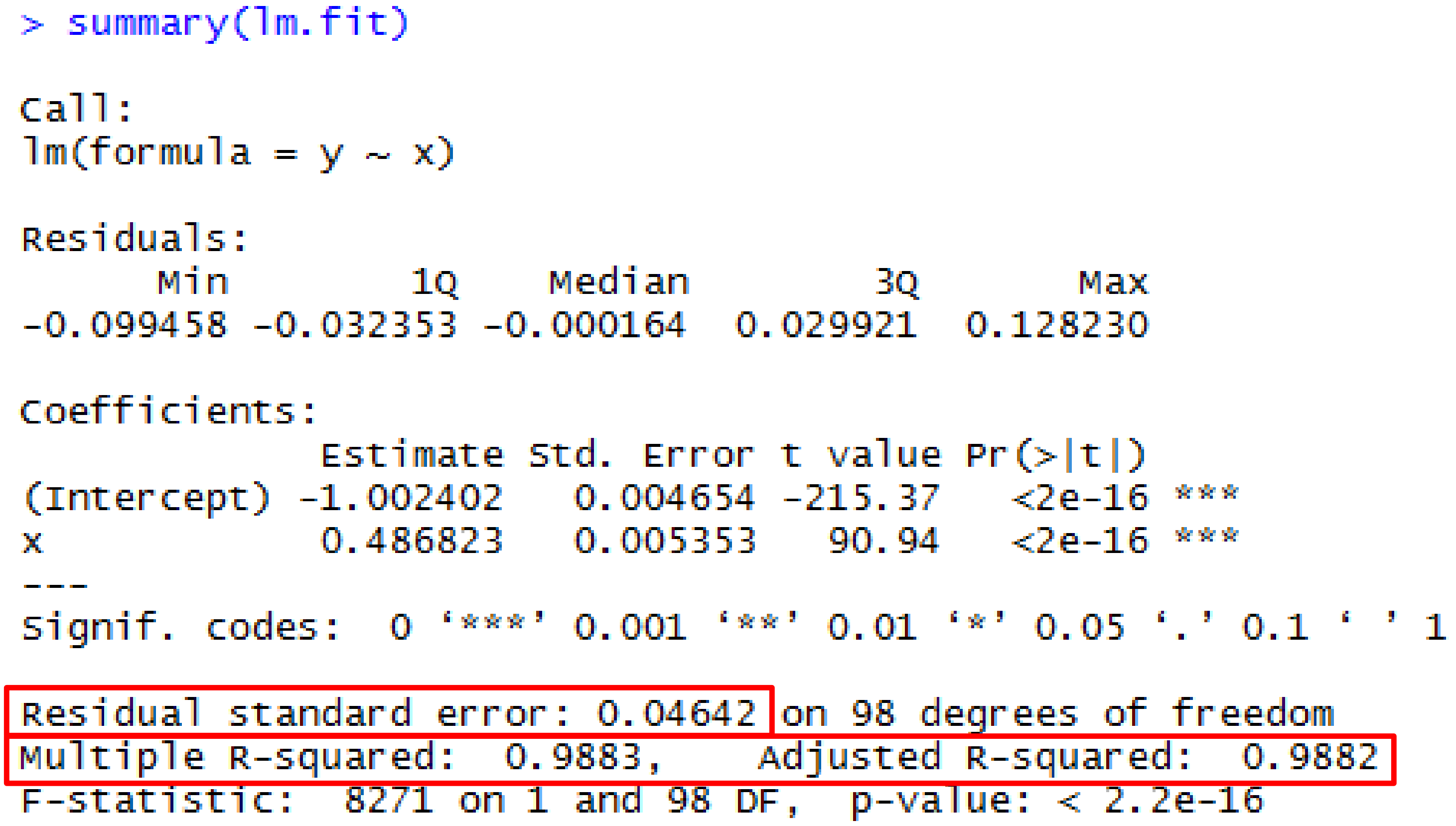
Remark: Cor(X,Y) = 0 means there is no linear relationship between X and Y, but there could be other relationship.

Example: X <- c(-3, -2, -1, 0, 1, 2, 3)

```
Y <- c(9, 4, 1, 0, 1, 4, 9)
# cor(X,Y) = 0
```

#But $Y = X^2$ ➔ Y and X has quadratic relationship

Measure of Fit

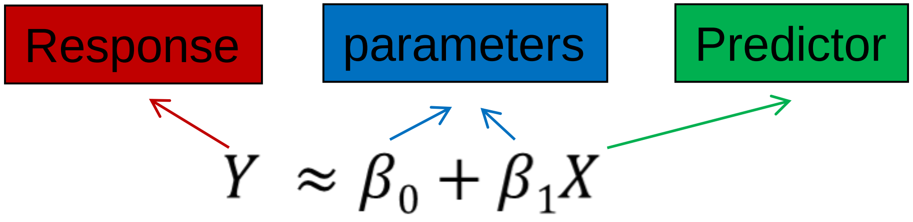


```
> summary(lm.fit)

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.099458 -0.032353 -0.000164  0.029921  0.128230 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.002402   0.004654 -215.37   <2e-16 ***
x            0.486823   0.005353   90.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04642 on 98 degrees of freedom
Multiple R-squared:  0.9883,    Adjusted R-squared:  0.9882 
F-statistic:  8271 on 1 and 98 DF,  p-value: < 2.2e-16
```

Adjusted R-squared: penalize for adding irrelevant/non-significant variables
Model with multiple variables: use adjusted R-squared
Model with single variable: use R squared and adjusted R squared interchangably

Simple Linear Regression

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



Response

parameters

Predictor

$$Y \approx \beta_0 + \beta_1 X$$

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 $X = x$

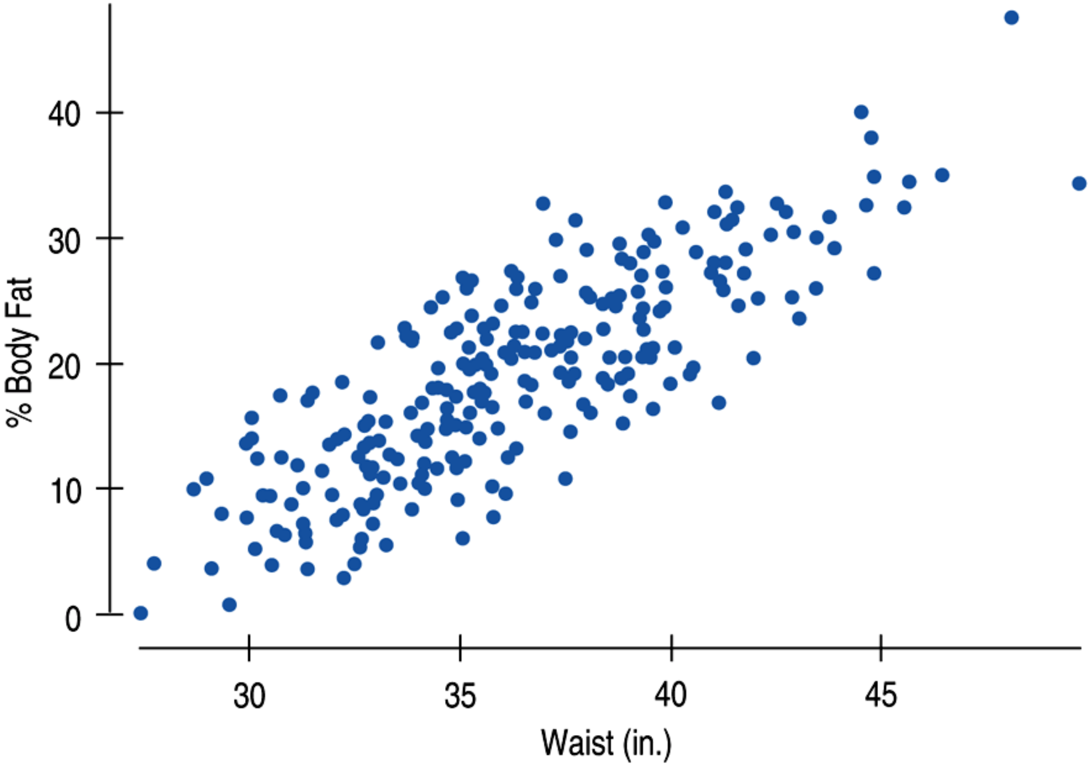
But how confident we are with the predicted $\hat{y}$?

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

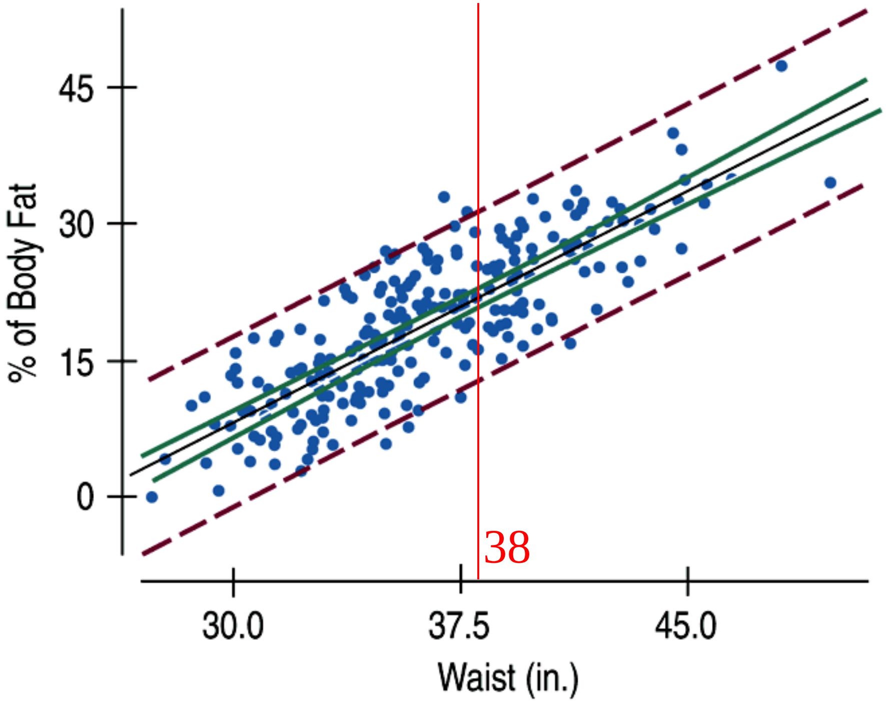
- **The predicted *%body fat* is the same in both questions**, but we can predict the *mean %body fat* for *all* men whose *waist size* is 38 inches with a lot more precision than we can predict the *%body fat* of a *particular individual* whose *waist size* happens to be 38 inches.

Confidence/Prediction Intervals for Predicted Values

- Here’s a look at the difference between predicting for a mean and predicting for an individual.

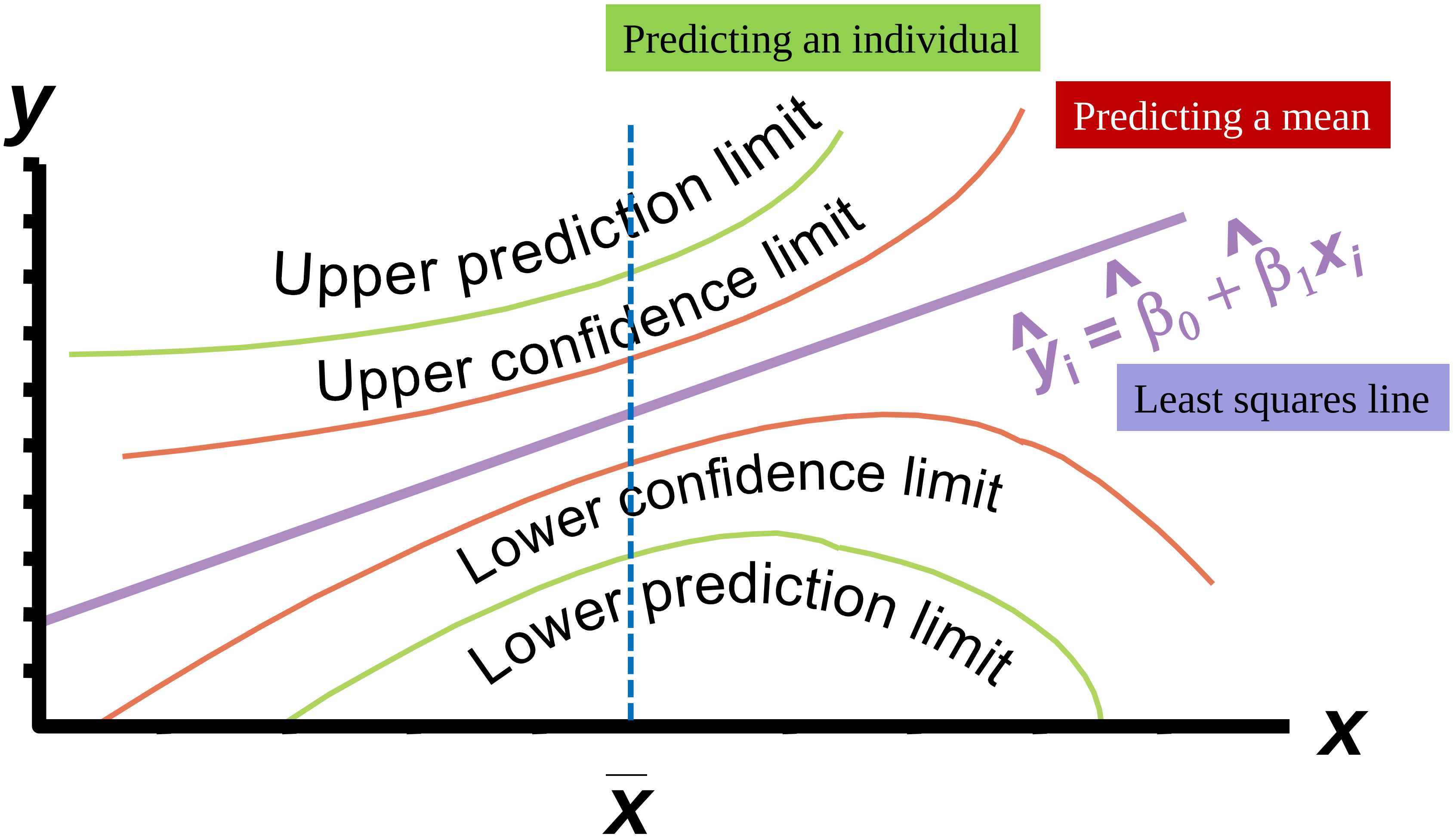
- The solid green lines near the regression line show the 95% confidence intervals for the mean predicted value, and the dashed red lines show the prediction intervals for individuals.

- 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



Predicting an individual

Predicting a mean

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

Least squares line

Conclusion

- Simple Linear Regression
  - Supervised Learning
  - Prediction
  - Parameterised method

- Variables
  - y = **Dependent** variable (quantitative)
  - x = **Independent** variable (quantitative)

- Least Squares Line
  - mean error = 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

Install packages/Load libs

- `install.package()` function downloads and installs packages from CRAN-like repositories or from local files.
- `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（a.k.a. towns）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 (not recommended)
> 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
```

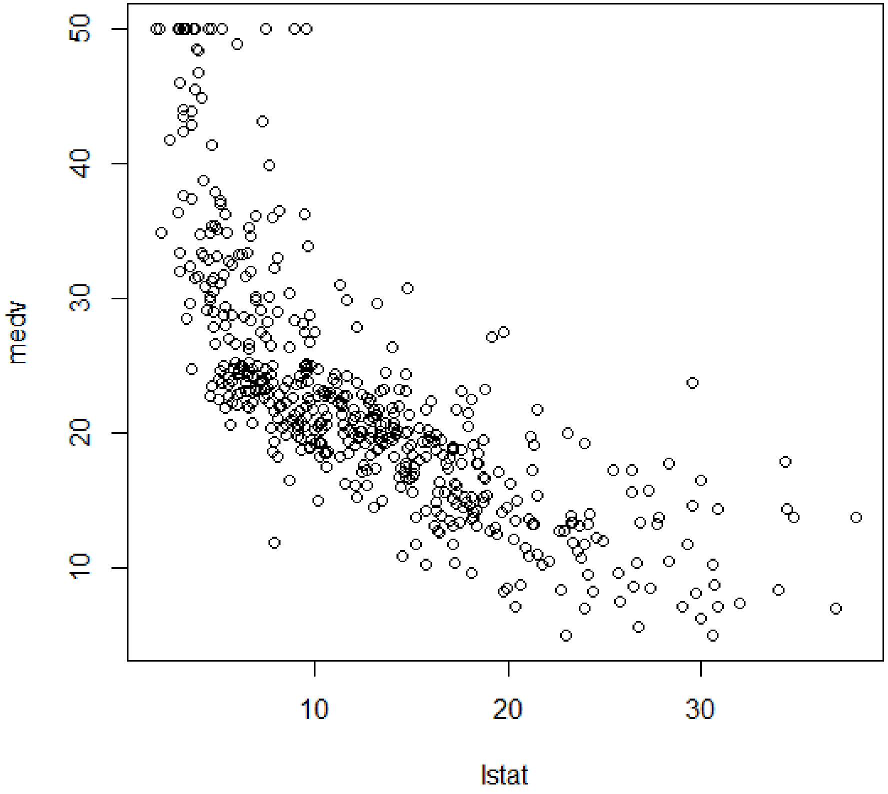
How to read the results?

```
> 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
```

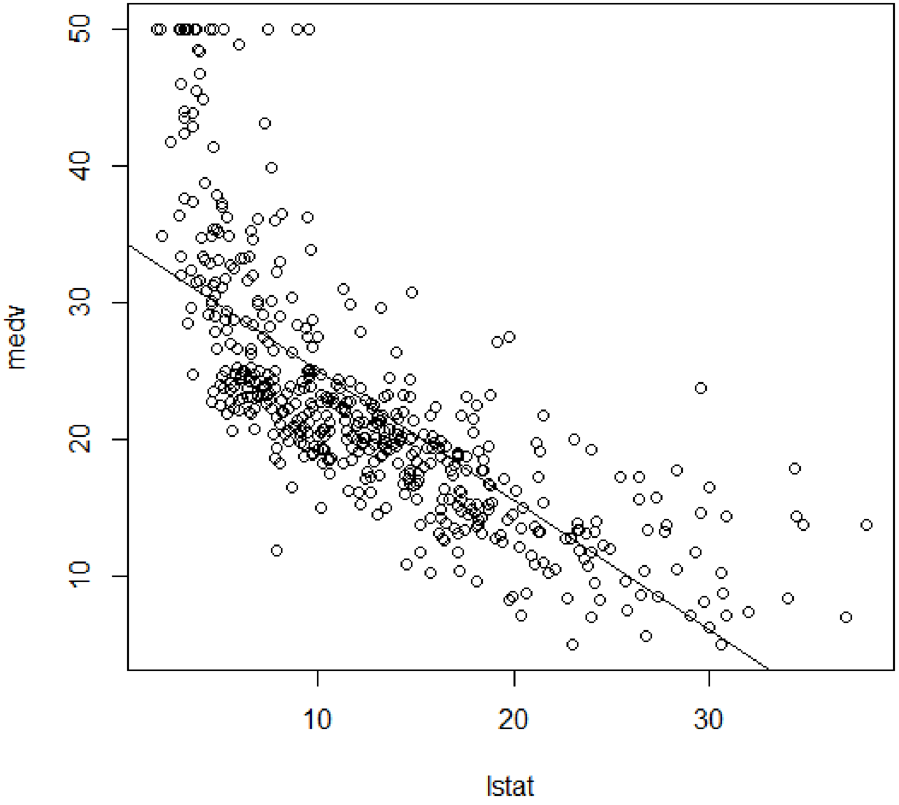
Which interval is wider?

Plot the results

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



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

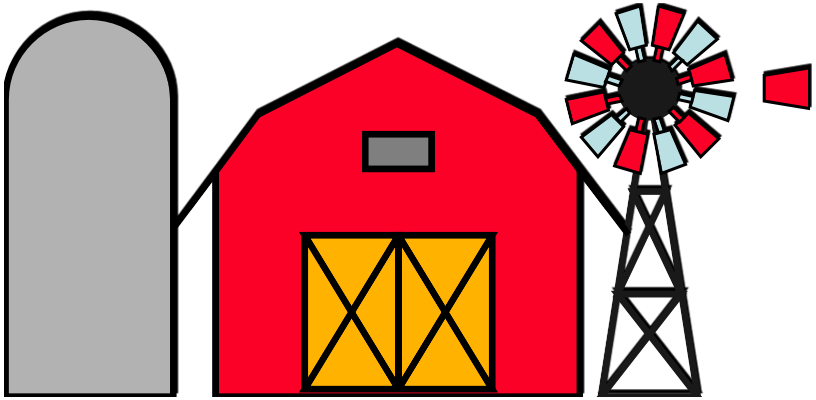


Try out other options on the width of the regression line, colour, symbols, etc
abline(lm.fit, lwd=3,col="red", pch="+"), …

Least Squares - Exercise

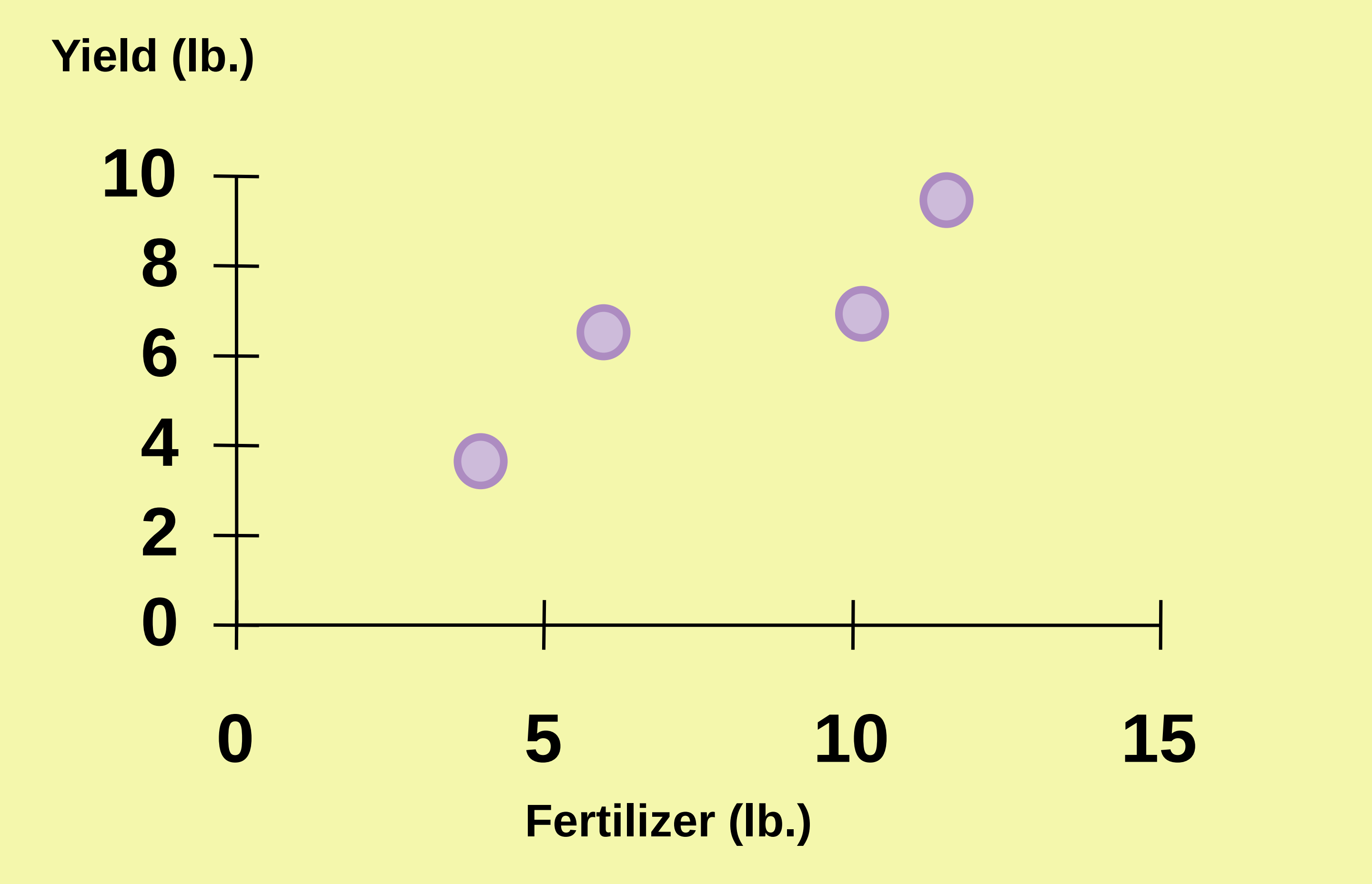
You're an economist for the county cooperative. You gather the following data:

Fertilizer (lb.)	Yield (lb.)
4	3.0
6	5.5
10	6.5
12	9.0



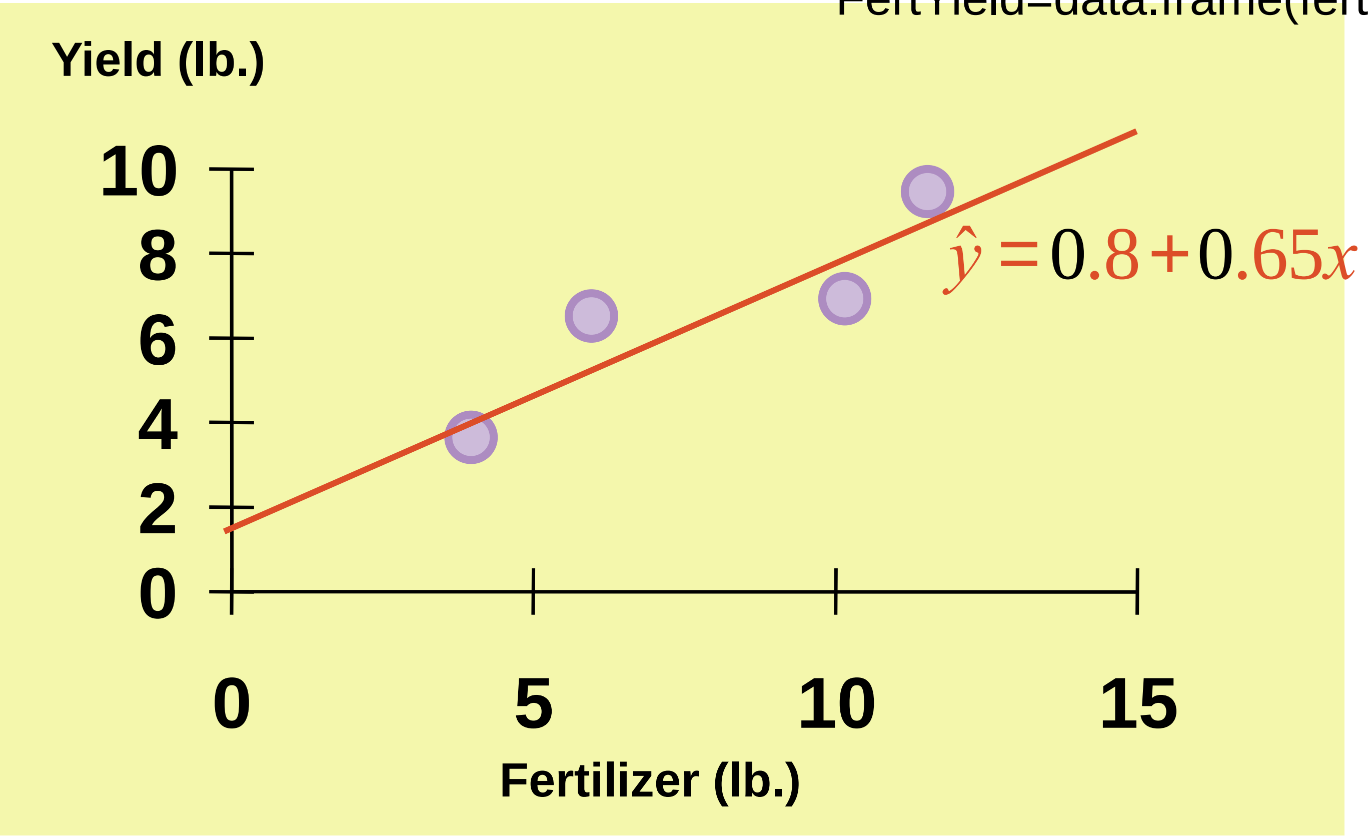
Find the **least squares line** relating crop yield and fertilizer.

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)
```



$\hat{y} = 0.8 + 0.65x$

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?

How to draw the CI/PI Curves?

```
lm.fit.Fert=lm(yield~fert,data=FertYield)
nd <- data.frame(fert=seq(2,8,length=51))
p_conf <- predict(lm.fit.Fert,interval="confidence",newdata=nd)
p_pred <- predict(lm.fit.Fert,interval="prediction",newdata=nd)

plot(fert,yield,data=FertYield,ylim=c(-5,12),xlim=c(0,15)) ## data
abline(lm.fit.Fert) ## fit
lines(nd$fert, p_conf[,"lwr"], col="red", type="b", pch="+")
lines(nd$fert, p_conf[,"upr"], col="red", type="b", pch="+")
lines(nd$fert, p_pred[,"upr"], col="blue", type="b", pch="*")
lines(nd$fert, p_pred[,"lwr"], col="blue", type="b", pch="*")
```

The CI/PI Plot