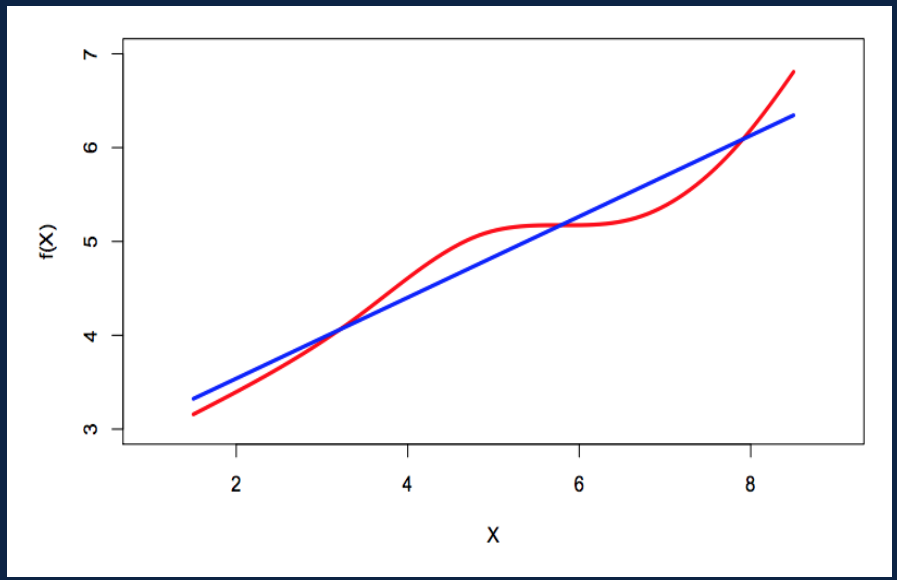
Simple Linear Regression

UBCO MDS — DATA 570

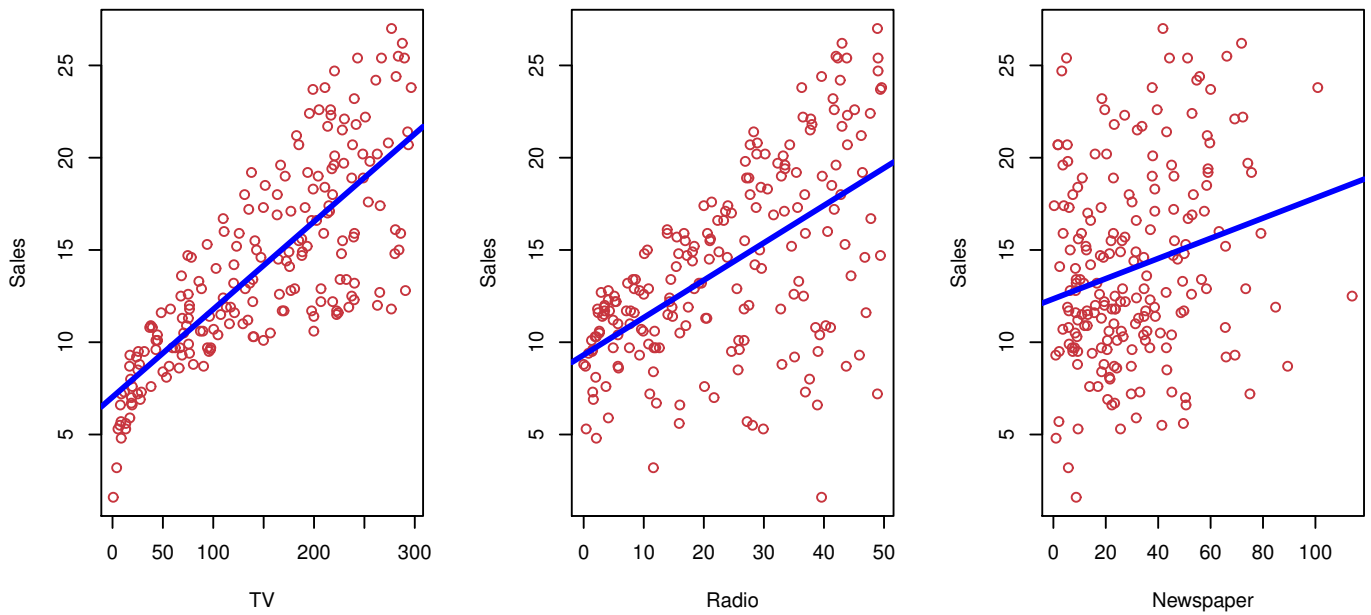
The Simple Linear Regression Model

- Linear regression is a simple approach to *supervised* learning.
- It assumes that the dependence of Y on $X_1, X_2, \ldots, X_n$ is linear.
- While the underlying distribution is never truly linear, linear regression is extremely useful both conceptually and practically.



Advertising data set

- The Advertising data set consists of the sales of a certain product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper
- Below plots the Sales vs TV, Radio and Newspaper, with a blue linear-regression line fit separately to each.



Advertising data set

- Questions we might ask:
  - Is there an association between advertising and sales ?
  - How strong is the relationship between advertising budget and sales?
  - Which media contribute to sales?
  - Can we predict Sales using these three?
  - How accurately can we predict future sales?
  - Is the relationship linear?
  - Is there synergy among the advertising media?

The Simple Linear Regression Model

- We assume that we can write the following model

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

- Y is the random, quantitative response variable
- X is the random predictor variable
- β_0 is the true intercept (unknown)
- β_1 is the true slope (unknown)
- ϵ is the true error

Assumptions

- For inferential purposes (which we will discuss momentarily), the model assumes the following.

1. There exists some (approximately) linear relationship between Y and X.
2. The distribution of ϵ_i has constant variance.
3. ϵ_i is normally distributed.
4. ϵ_i are independent of one another. For example, ϵ_2 is not affected by ϵ_1.

The Observed Line

- Given the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we can predict y using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{y}$ is the estimate of the response variable
- x is the observed predictor variable
- $\hat{\beta}_0$ is the estimated intercept
- $\hat{\beta}_1$ is the estimated slope

Simple Linear Regression

- Fitting a line to data is a special case of a very versatile technique called **regression**.
- Regression, in general, involves modeling some 'dependent' (or response) variable using a number of 'explanatory' (or predictor) variables.
- The equation of a line involves only one explanatory variable and hence the term **simple** linear regression (SLR).
- But why the term 'regression'?

Historical Comment

- Sir Francis Galton studied the height of grown sons in comparison to their fathers[1].
- The author thought he found that children tended to have a more moderate height (i.e. closer to average height) than their parents.
- In other words: the sons of tall men tended to be shorter than their father but taller than average, and the sons of short men tended to be taller than their fathers, but shorter than average.
- Lets look at the visual representation of this linear relationship. . .

[1] Galton, F. (1886), *Regression Towards Mediocrity in Hereditary Stature*. Journal of the Anthropological Institute, 246-263.

Galton's original data

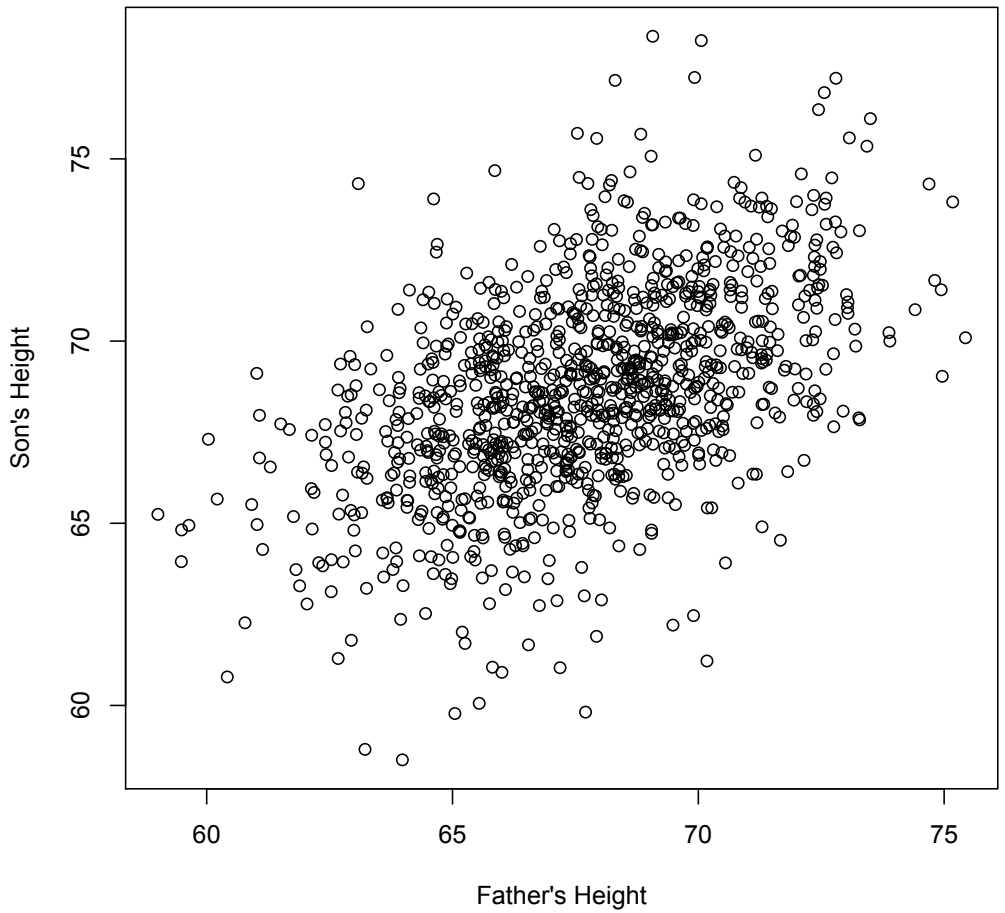


Figure: This plots the grown up sons height against the height of their fathers (i.e. each point is of the form $(x, y) =$ (height of son, height of father).

Galton's original data

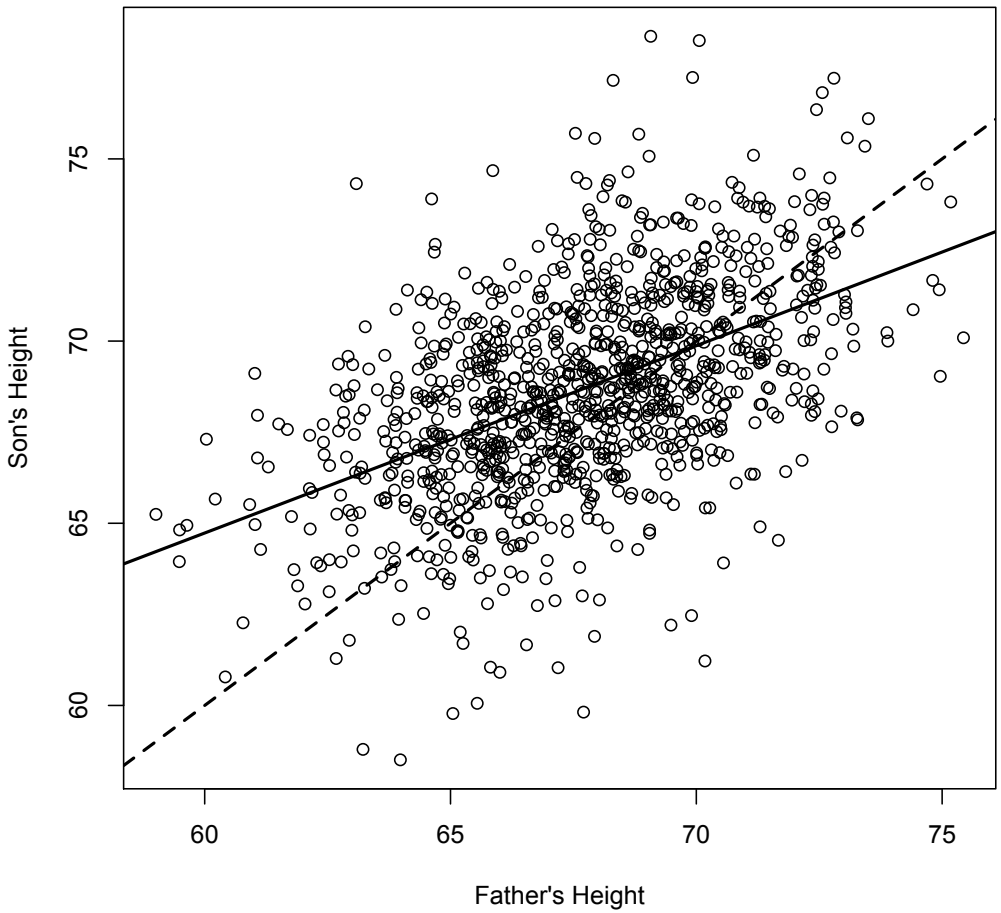


Figure: The dotted line corresponds to the slope of 1 (i.e. if the predicted height of the son was the same height as the father). The solid line represents the least squares regression line.

Regression

- Naturally, if the heights of sons were more moderate than their fathers, then the height of fathers should be more extreme than their sons.
- In other words: Tall men should have even taller fathers, and short men, even shorter fathers.
- To check this we flip the role of x and y: i.e. let the height of a son be the INdependent variable and the height of the fathers be the dependent variable.

Galton's original data

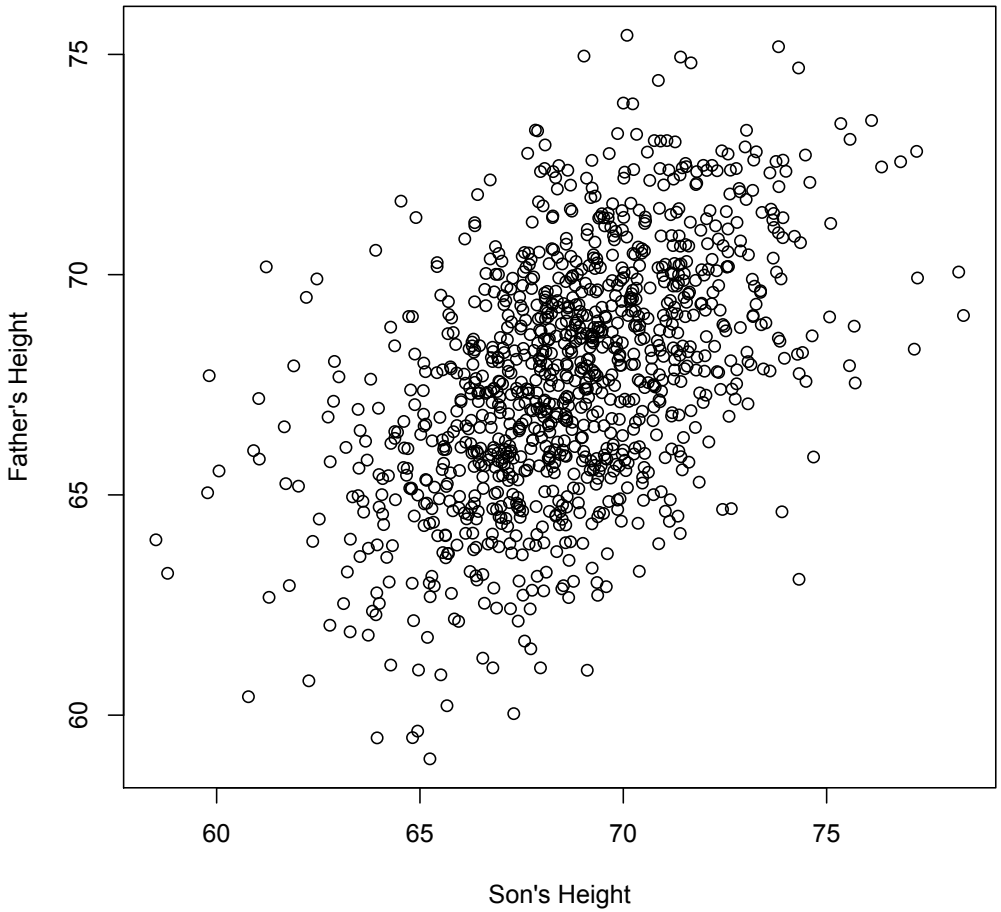


Figure: This plots the height of their fathers against the height of their sons (i.e. each point is of the form $(x, y) =$ (height of father, height of son).

Galton's original data

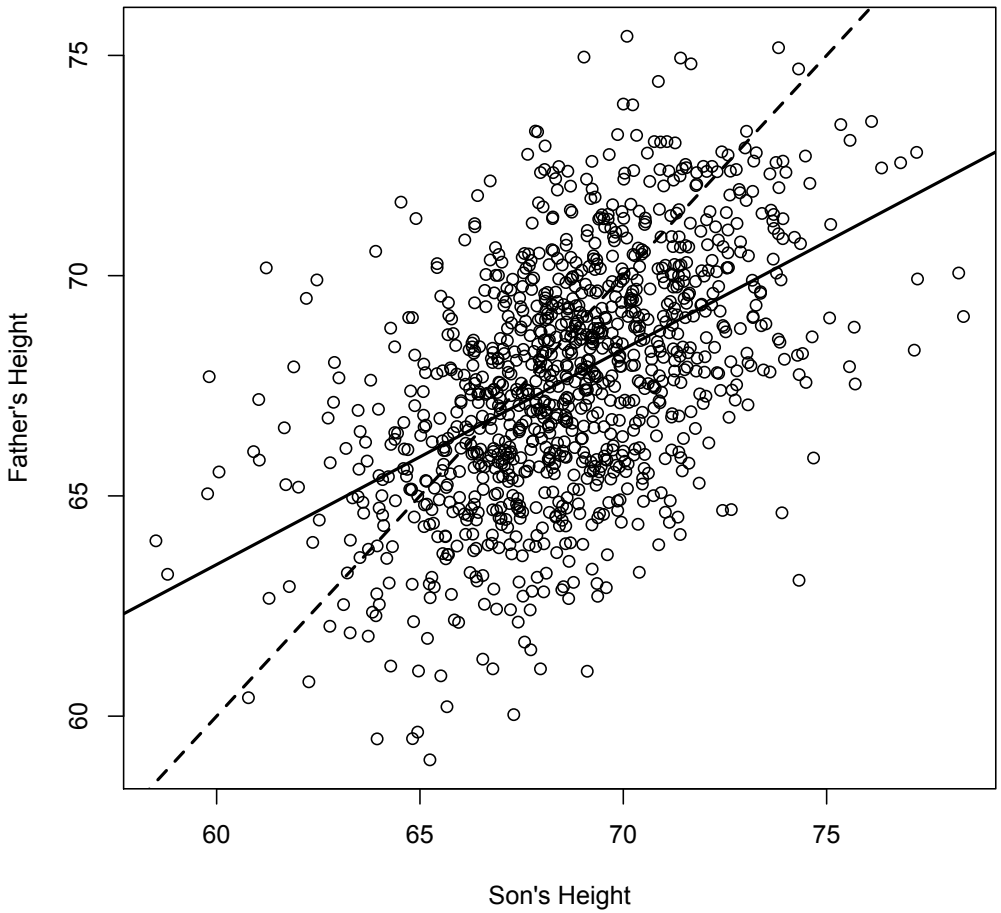


Figure: The dotted line corresponds to the slope of 1 (i.e. if the predicted height of the father was the same height as the son). The solid line represents the least squares regression line.

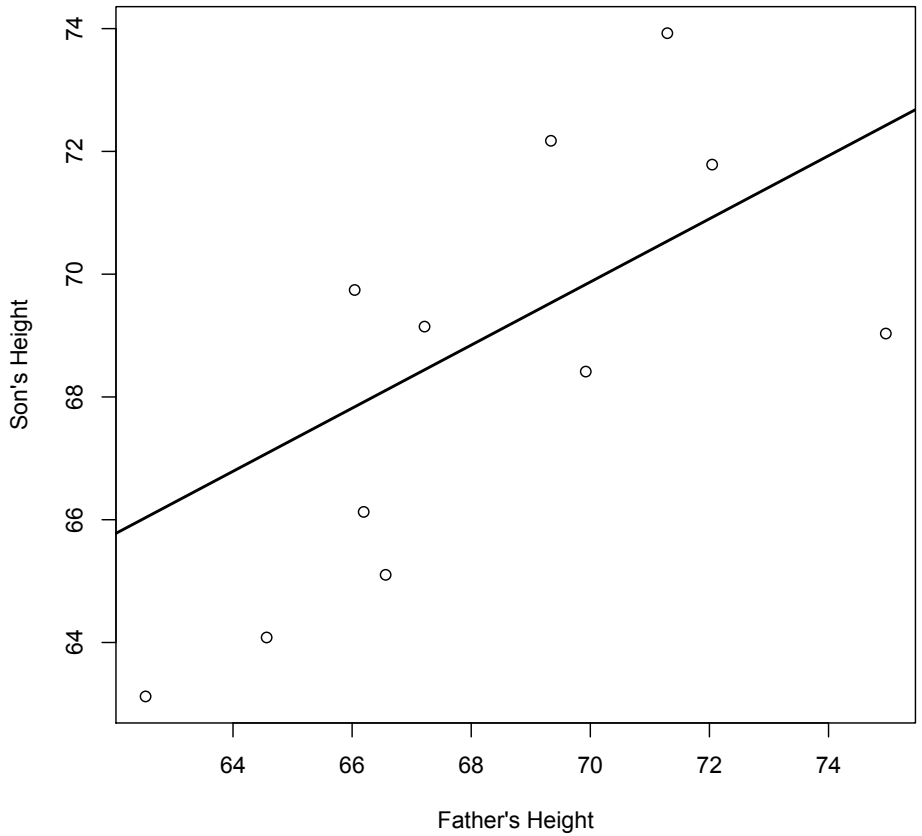
Regression

- Flipping the role of x and y shows that *fathers* tended to have a more moderate height than that of their *sons* (the contradictory argument).
- This is an artifact of the way in which we construct on least squares regression line (we are minimizing vertical distances, more on this to come . . .)
- This displayed that the height of sons *regressed* towards the average height
- Although this is a spurious conclusion this is how the term regression—in the context of how we use it in Statistics—was born.

Regression

- Today, *regression* is taken to mean that one of a wide class of models is fitted to data; usually for the purpose of analysis.
- Today we will discuss a simple linear regression (SLR) model used to find the **linear** relationship between two correlated variables.
- Here is the gist . . .

Regression



- This plots the grown up sons height against the height of their fathers for eleven randomly selected points from the Galton data set.
- There are infinitely many choices for the lines that we can fit to this data.
- We aim to find the line (i.e. the equation of the line) that minimizes the sum of square vertical distance between the point and the line.

Fitting the Model I

- In real terms, we do not know the theoretical values of the model parameters (β_0, β_1).
- To fit our line, we mean that we need to find *estimates* (i.e. good guesses) for (β_0, β_1).
  - We denote the estimate of the y-intercept β_0 by $\hat{\beta}_0$.
  - We denote the estimate of slope β_1 by $\hat{\beta}_1$.
- While there are other methods for finding these *estimators*, the least squares technique is typically implemented.

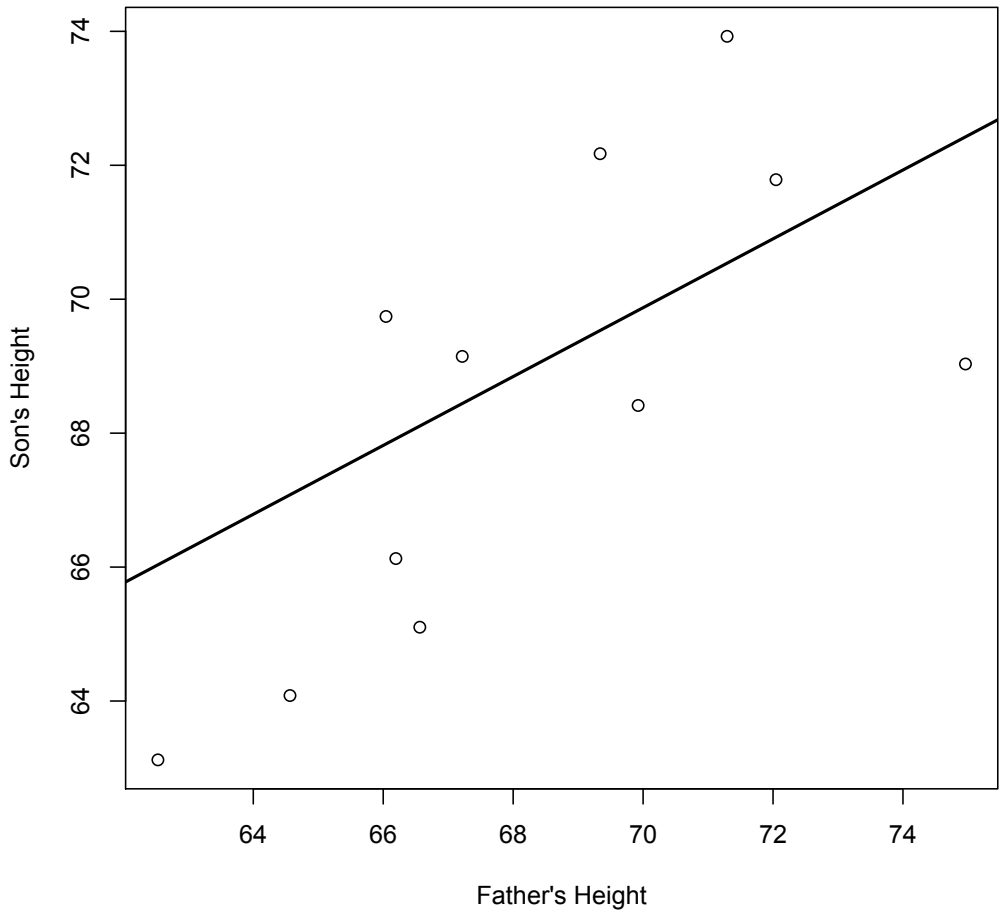
The Least Squares Technique

- This *least squares* technique involves minimizing the sum of the squared vertical distance from each point to the line, or curve.
- You can imagine this as 'wiggling' the line about until the sum of the squared vertical distances is minimized.
- This minimization can be achieved using some basic calculus.

Fitting the Model I

- For any line, we can define the observed error as $e_i = y_i - \hat{y}_i$ (these are often called residuals).
- It's a mathematical truth that $\sum e_i = 0$ for any line that passes through the point $(\bar{x}, \bar{y})$.
- To find the "best" line, we instead minimize the residual sum of squares $RSS = \sum e_i^2$.

Lets identify what the predicted values and residuals on the subset of the Galton Data...



Fitting the Model II

- In the least squares paradigm, we are minimizing the square residuals i.e. minimize $\sum_i e_i^2$.
- First some notation. Let

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

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

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

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

Fitting the Model III

$\hat{\beta}_0$ is found by setting $\partial Q/\partial \hat{\beta}_0$ and solving for $\hat{\beta}_0$.

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad (1)$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \quad (2)$$

Add $-(1) \times \bar{X}$ to (2).

$$\sum_{i=1}^{n}[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)](X_i - \bar{X}) = 0 \quad (3)$$

Substitute $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ in (3) from (2).

$$\sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})](X_i - \bar{X}) = 0 \quad (4)$$

Then, we have $\hat{\beta}_1 = S_{xy}/S_{xx}$.

Fitting the Model IV

$\hat{\beta}_1$ is found by setting $\partial Q/\partial \hat{\beta}_1$ and solving for $\hat{\beta}_1$.

Fitting the Model V

- Hence, the resulting least squares estimators are

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{xy}}{S_{xx}}$$

- Note that,

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

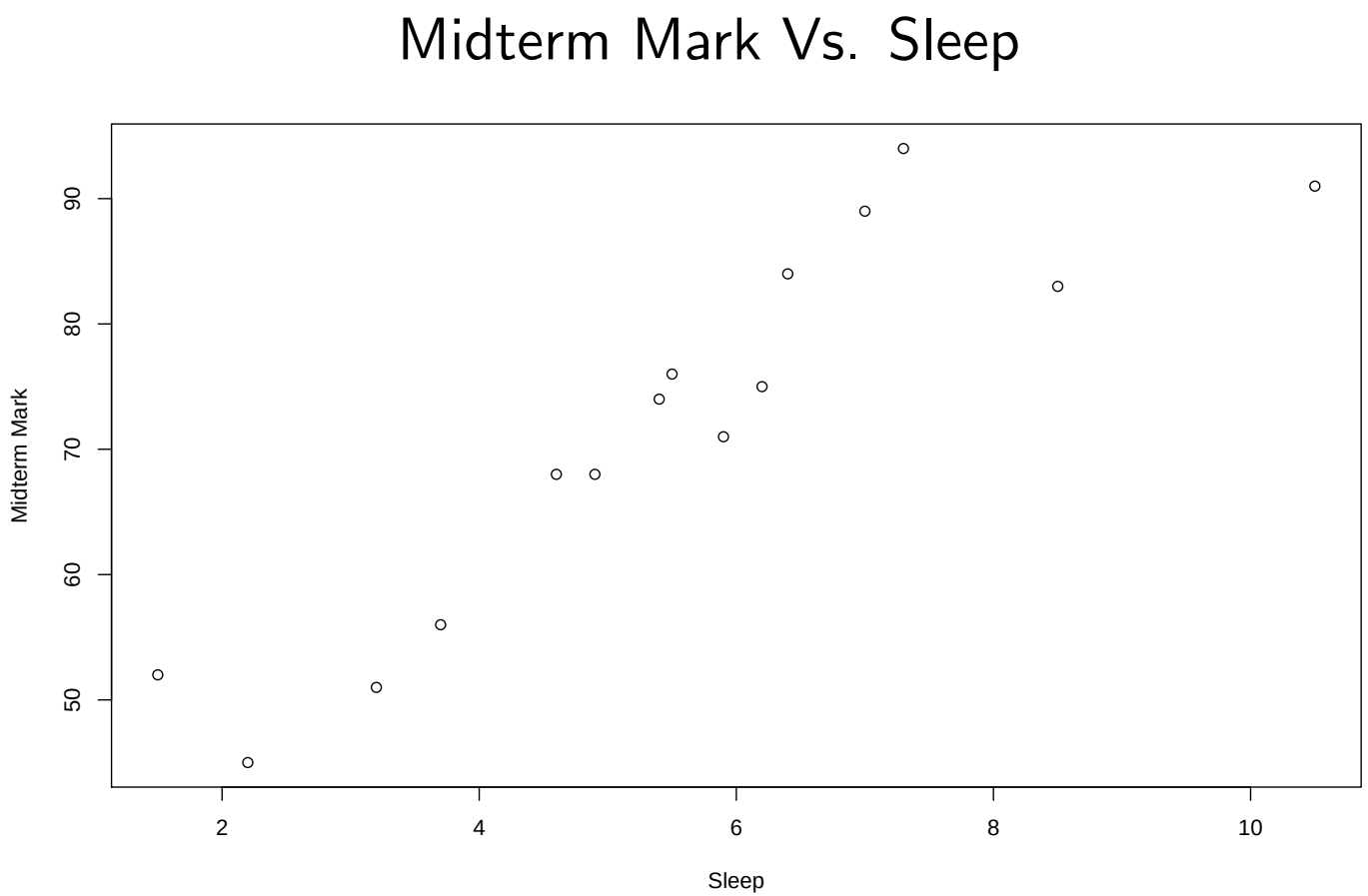
  i.e. our least squares regression line will always pass through the point $(\overline{X}, \overline{Y})$

Example I: Background

- X, the number of hours of sleep the night before a midterm, and Y, the subsequent midterm mark, were recorded for 15 randomly selected first year students.

Sleep (hours)	Midterm Mark
4.6	68
8.5	83
3.2	51
5.5	76
7.0	89
6.2	75
6.4	84
5.9	71
1.5	52
10.5	91
7.3	94
5.4	74
4.9	68
3.7	56
2.2	45

Example I: A visual



- There appears to be a positive linear relationship between the midterm mark and the number of hours of sleep.

Example I: Estimating β_0 and β_1

- Summary statistics: $\bar{x} = 5.52$ and $\bar{y} = 71.8$.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	y_i	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
	4.6	(4.6 - 5.52)	$(-0.92)^2$	68	(68-71.8)	3.496
	8.5	(8.5 - 5.52)	$(2.98)^2$	83	(83-71.8)	33.376
	3.2	(3.2 - 5.52)	$(-2.32)^2$	51	(51-71.8)	48.256
	5.5			76		
		⋮	⋮		⋮	⋮
	7.0	⋮	⋮	89	⋮	⋮
	6.2			75		
	6.4			84		
	5.9			71		
		⋮	⋮		⋮	⋮
	1.5	⋮	⋮	52	⋮	⋮
	10.5			91		
	7.3			94		
	5.4			74		
		⋮	⋮		⋮	⋮
	4.9	⋮	⋮	68	⋮	⋮
	3.7			56		
	2.2	(2.2 - 5.52)	$(-3.32)^2$	45	(45-71.8)	88.976
Sum			77.544			457.66

Example I continued

What is the least square regression line?

Interpreting the parameters

The slope: $\hat{\beta}_1$

- In general, for every 1 unit increase in X, the response variable Y increases by $\hat{\beta}_1$ units.
- For this example: for each additional hour of sleep, a student's midterm mark increases by 5.902% on average.

The y-intercept: $\hat{\beta}_0$

- In general, when X is 0 the predicted value of Y is $\hat{\beta}_0$.
- For this example: the predicted midterm mark of students who get no sleep the night before is 39.221%.

Note: Use caution when interpreting the y-intercept because it may have no practical meaning.

Example I: Prediction I

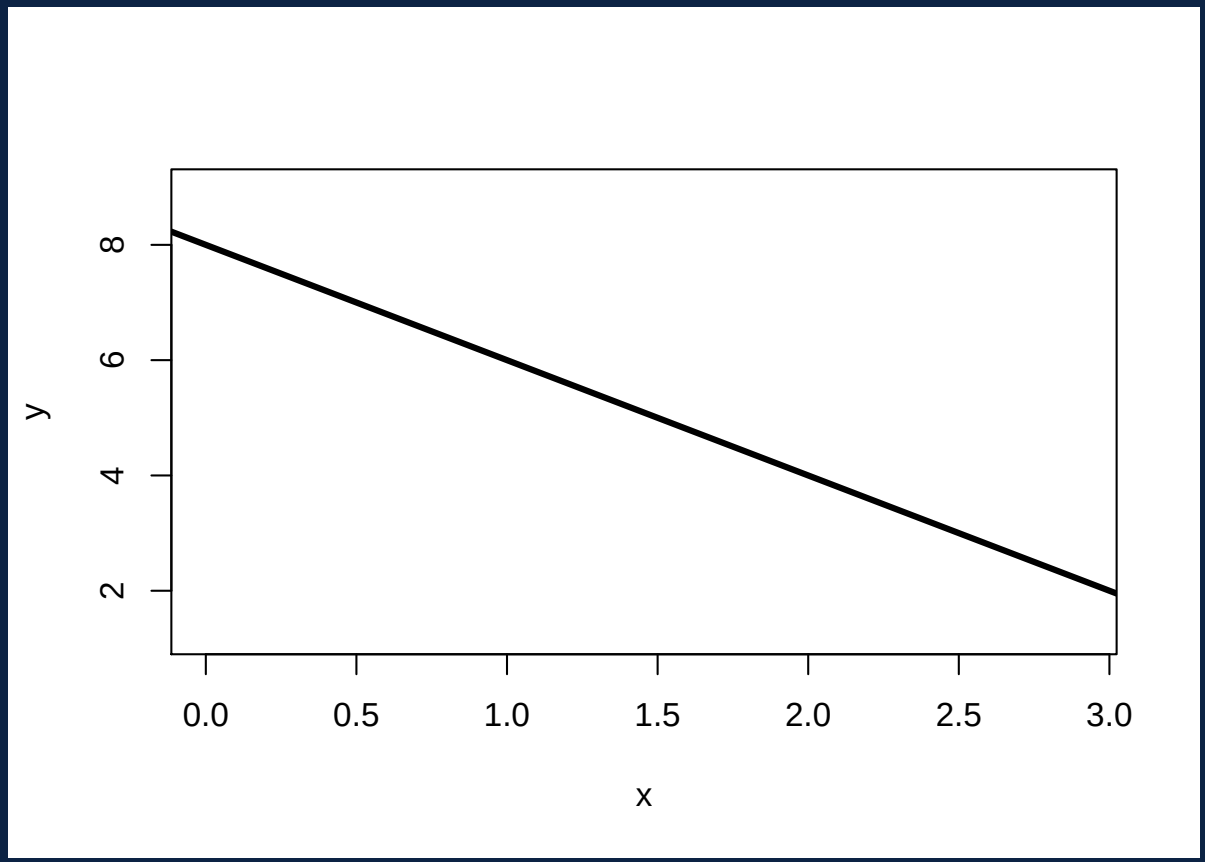
- What is the predicted midterm mark for a student who got 4 hours of sleep the night before the midterm?

Example I continued

We could calculate the predicted value for each of the observed x_i values and obtain their residuals.

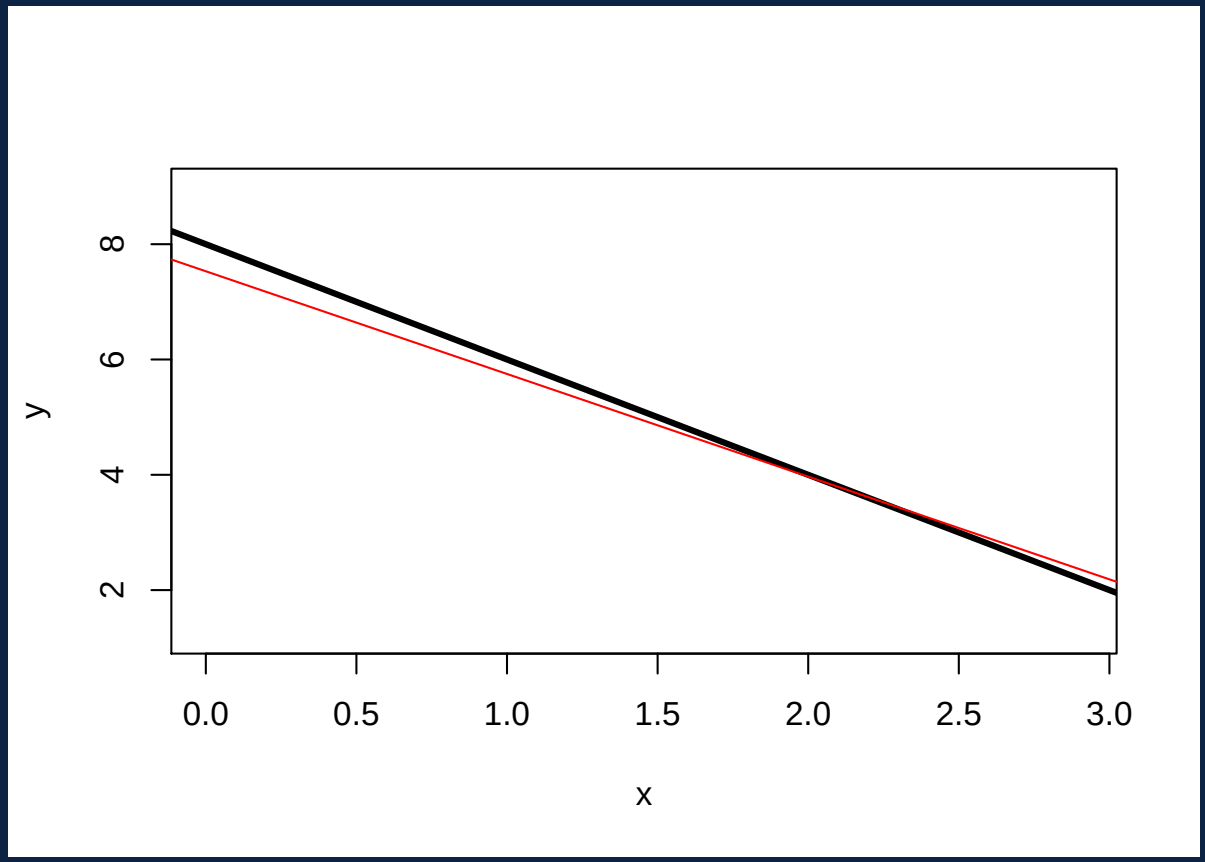
	x_i	y_i	$\hat{y}_i$	e_i	e_i^2
	4.6	68	66.37	1.63	2.66
	8.5	83	89.39	-6.39	40.80
	3.2	51	58.11	-7.11	50.52
	5.5	76	71.68	4.32	18.65
	7.0	89	80.53	8.47	71.66
	6.2	75	75.81	-0.81	0.66
	6.4	84	76.99	7.01	49.09
	5.9	71	74.04	-3.04	9.26
	1.5	52	48.07	3.93	15.41
	10.5	91	101.19	-10.19	103.87
	7.3	94	82.31	11.69	136.76
	5.4	74	71.09	2.91	8.46
	4.9	68	68.14	-0.14	0.02
	3.7	56	61.06	-5.06	25.59
	2.2	45	52.21	-7.20	51.92
Sum					585.32

Example: Simulation



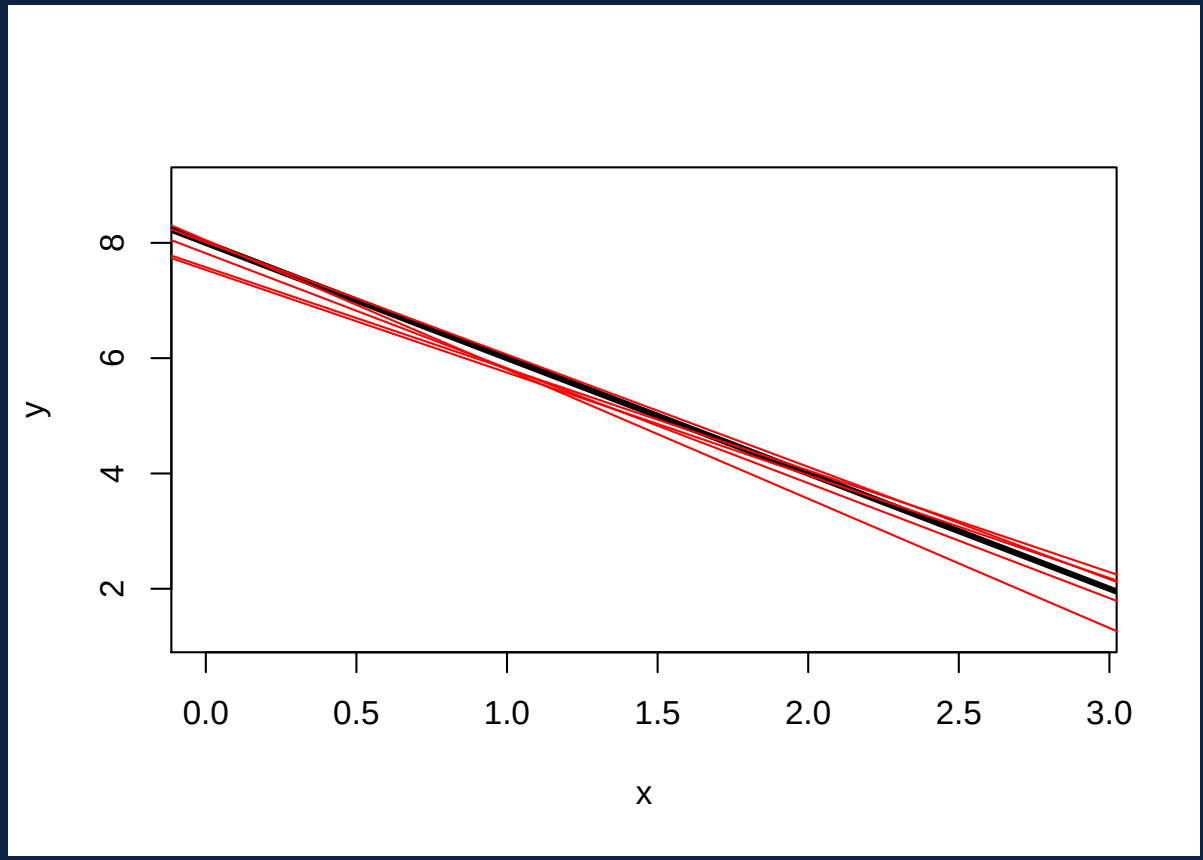
- The true model is in black: $Y = -2x + 8$
- The error standard normal (mean 0, variance 1).

Example: Simulation



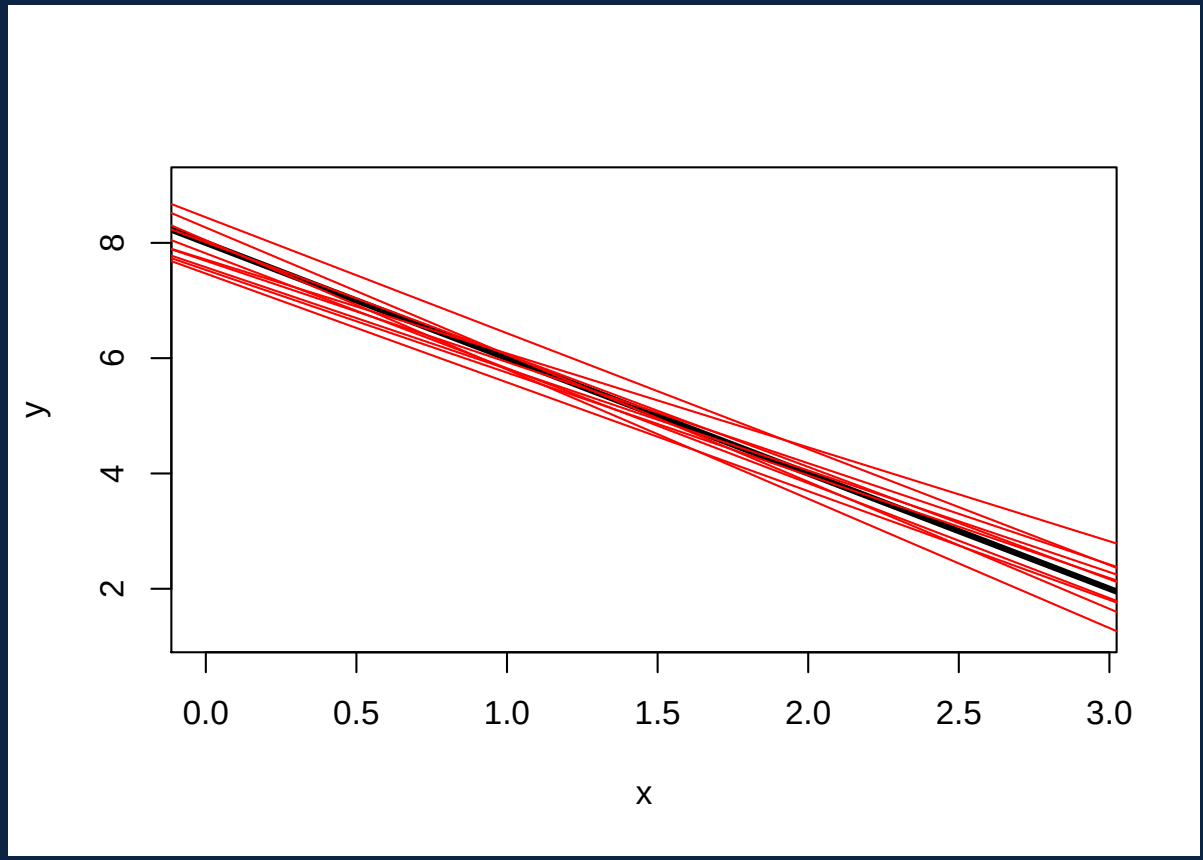
- In red is a line fit to a sample of size 30 from the model.
- Next, we repeat...

Example: Simulation



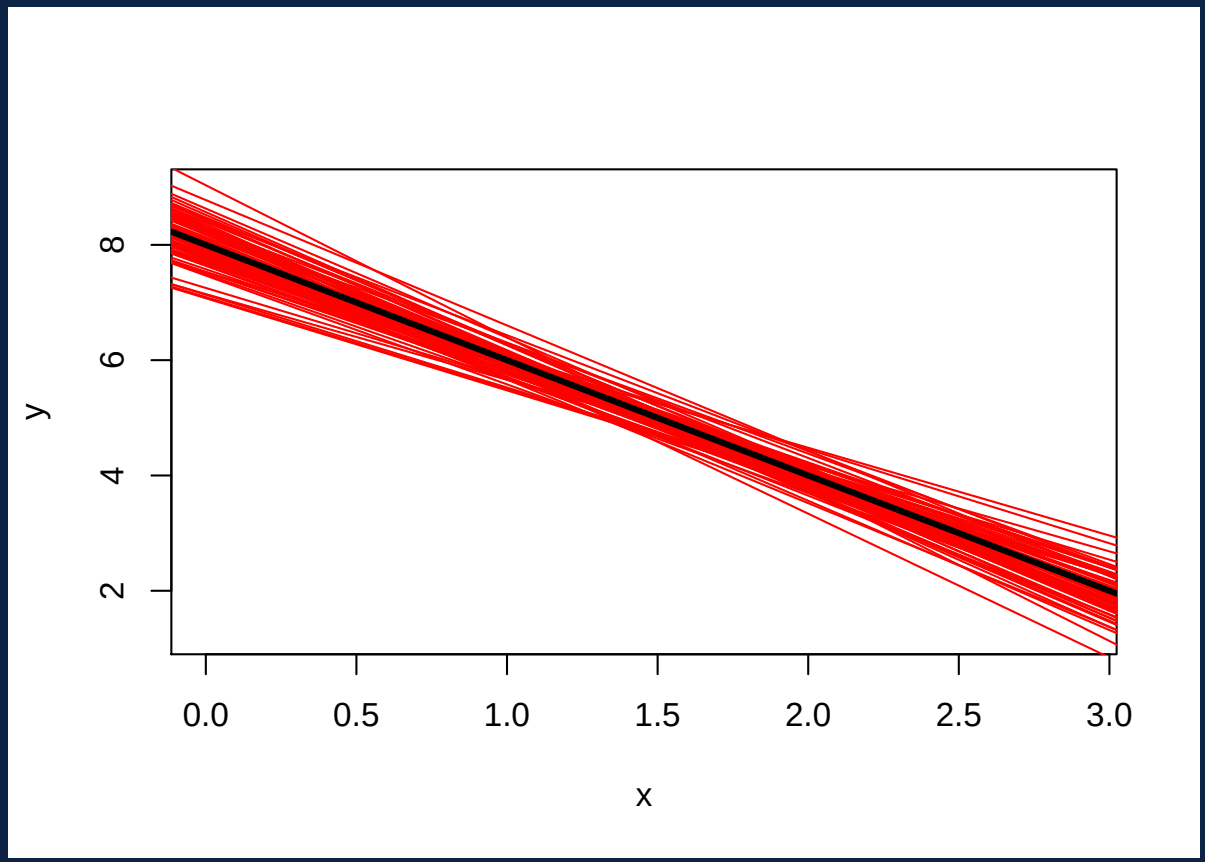
- Repeated 5 times.
- Next, more...

Example: Simulation



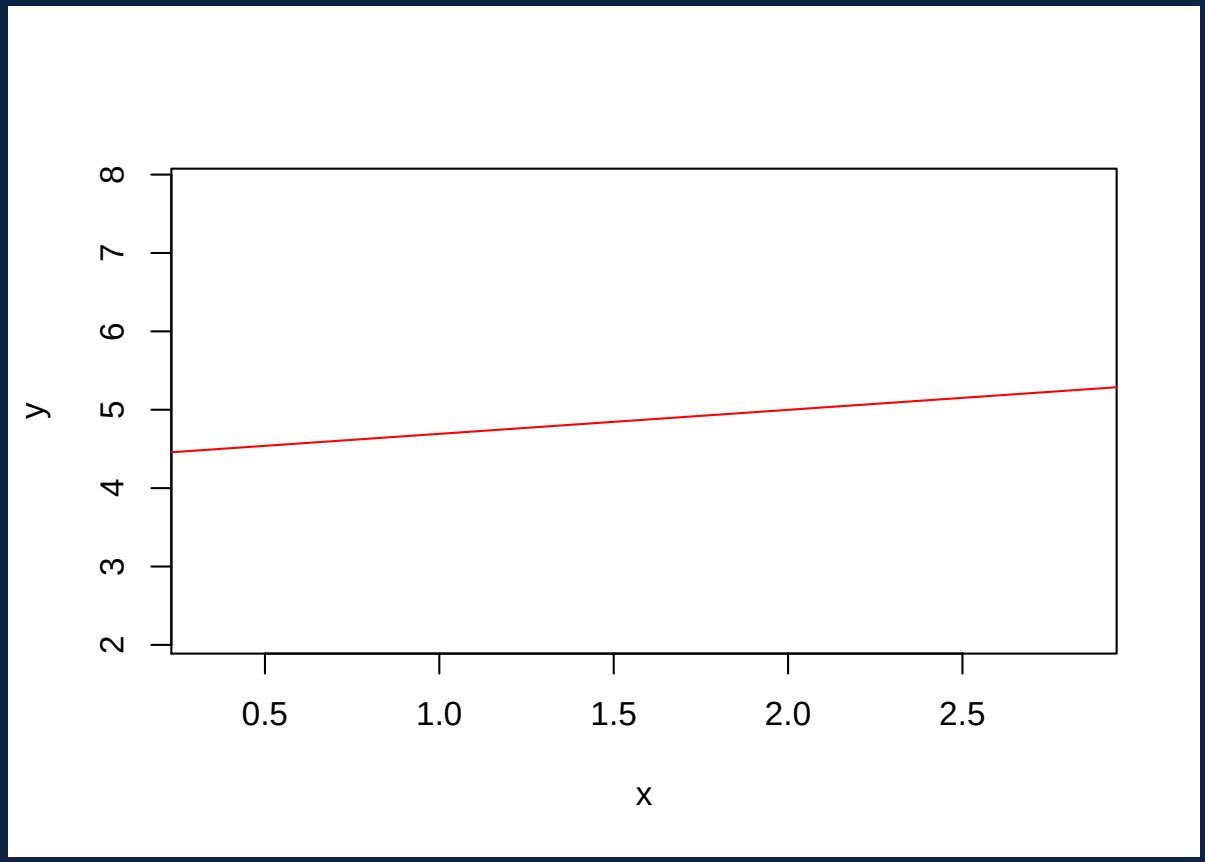
- Repeated 10 times.
- Next, more

Example: Simulation



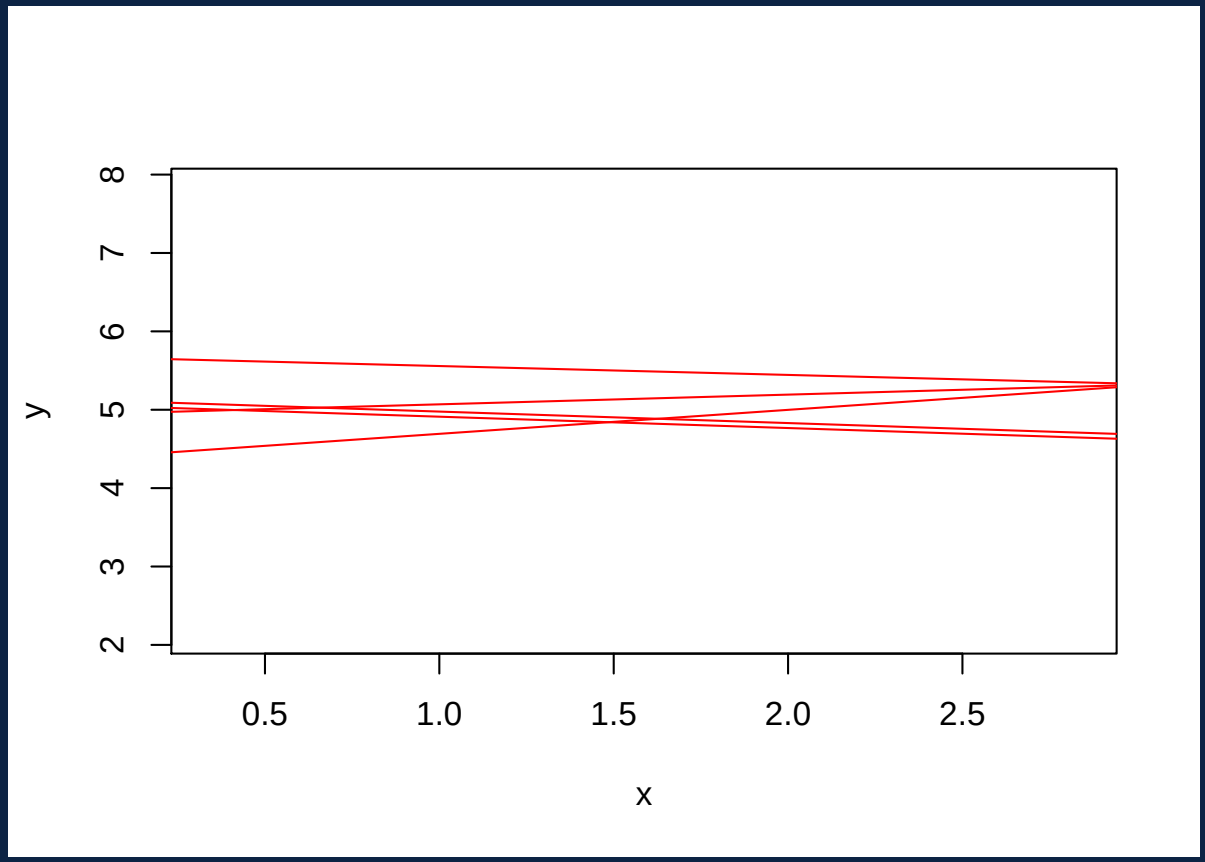
- I stopped counting...
- What does this illustrate?

Example: Simulation2



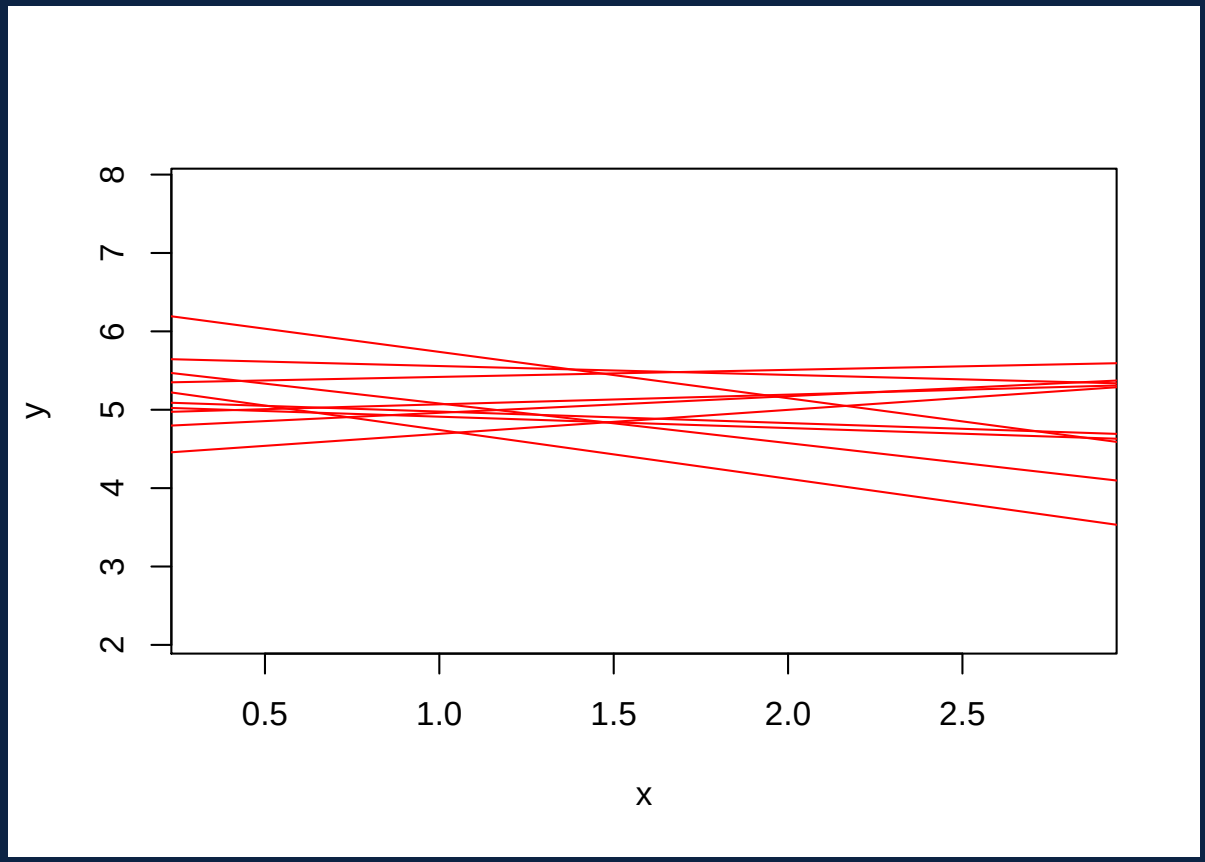
- The true model is 'unknown'
- Sample size 30, 1 run

Example: Simulation2



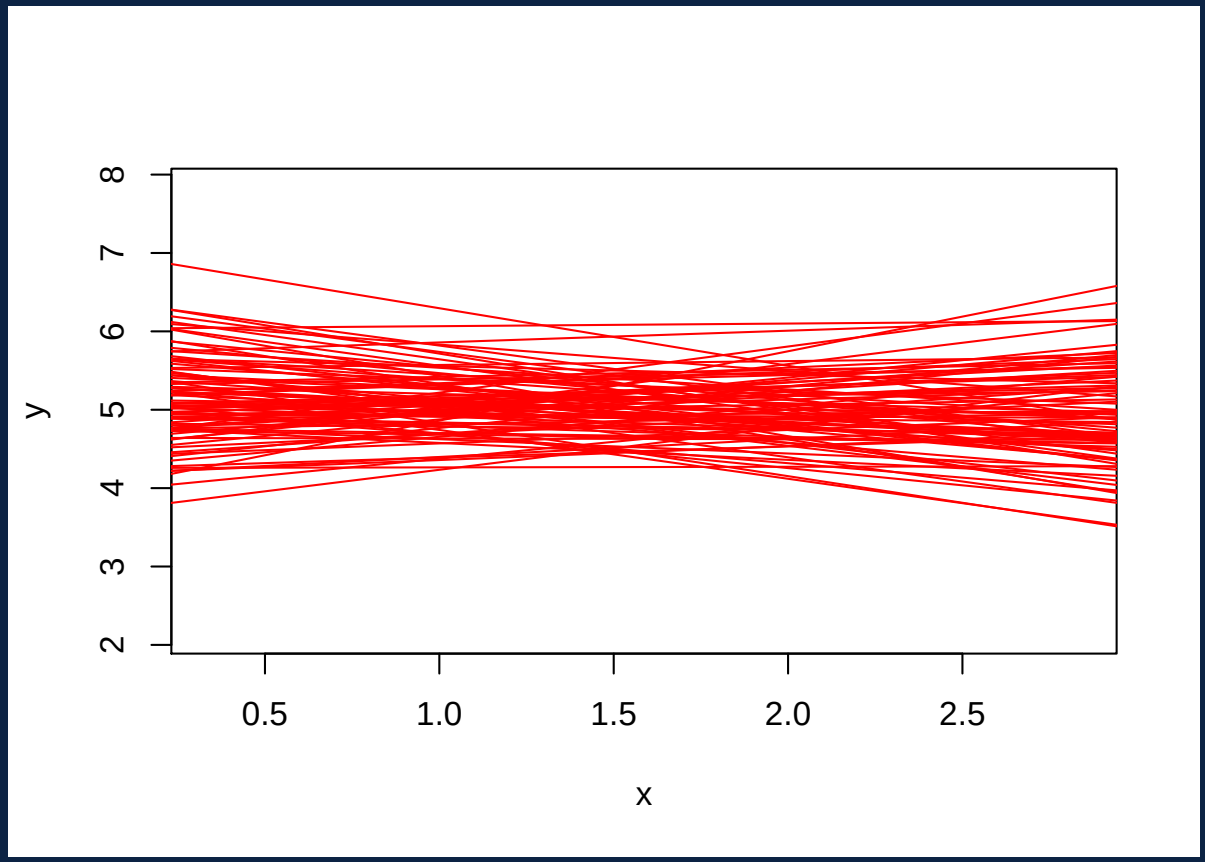
- The true model is 'unknown'
- Sample size 30, 5 runs

Example: Simulation2



- The true model is 'unknown'
- Sample size 30, 10 runs

Example: Simulation2



- The true model is 'unknown'
- Sample size 30, many runs

Inference in Regression

- Now that we have a way of finding estimates for our parameters, we might begin to ask other types of questions:
  - How precise are our estimates?
  - Is the slope significantly different from 0?
- To help answer these types of questions we need knowledge of sample distributions, . . .

Sampling Distribution for $\hat{\beta}_1$

- β_1 is our true, unknown slope that we are estimating with $\hat{\beta}_1$.
- As with any estimate, $\hat{\beta}_1$ is following some distribution. In fact, under the assumptions previously noted, $\hat{\beta}_1$ comes from a normal distribution with
  - $\mu_{\hat{\beta}_1} = \beta_1$ (that is, $\hat{\beta}_1$ is an unbiased estimate of β_1)
  - $SE(\hat{\beta}_1) = \dfrac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}$
- So the more noise around the line the less precise the slope
- and the more spread out our x's, the more precise the slope

Sampling Distribution for $\hat{\beta}_1$

- The σ used in the standard error formula on the previous slide represents the standard deviation of the error terms, ie. $\sigma^2 = \text{Var}(e_i)$
- σ is generally unknown and is estimated with

$$RSE = \sqrt{\frac{RSS}{n-2}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}$$

- This estimate stands for the residual standard error.

Confidence Interval for β_1

- The standard errors can be used to build confidence intervals
- An approximate 95% CI for β_1 can be calculated as

$$\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

- Interpretation: we are 95% confidence that the true value of β_1 lies within the numbers $[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)]$.

Hypothesis Test for β_1

- Confidence intervals are closely related to hypothesis testing.
- A hypothesis test seeks to determine whether or not evidence supplied by the data indicate that a particular *hypothesis* is supported at a predetermined level of confidence.
- In the context of this lecture, our null hypothesis sets the population parameter or interest to a certain value. This is tested against a competing statement called the alternative hypothesis.

Hypothesis Test for β_1

- Like confidence intervals, we assert a level of confidence in our test.
- For hypothesis testing, this confidence is given by the significance level, denoted by α,
- The significance level is equal to 1 - (the confidence level).
- Hence, a confidence level is 0.95, then our significance level is 0.05.

Hypothesis Test for β_1

- Hypotheses:
  - $H_0 : \beta_1 = 0$, in words: There is no relationship between X and Y
  - $H_a : \beta_1 \neq 0$ in words: There is some relationship between X and Y

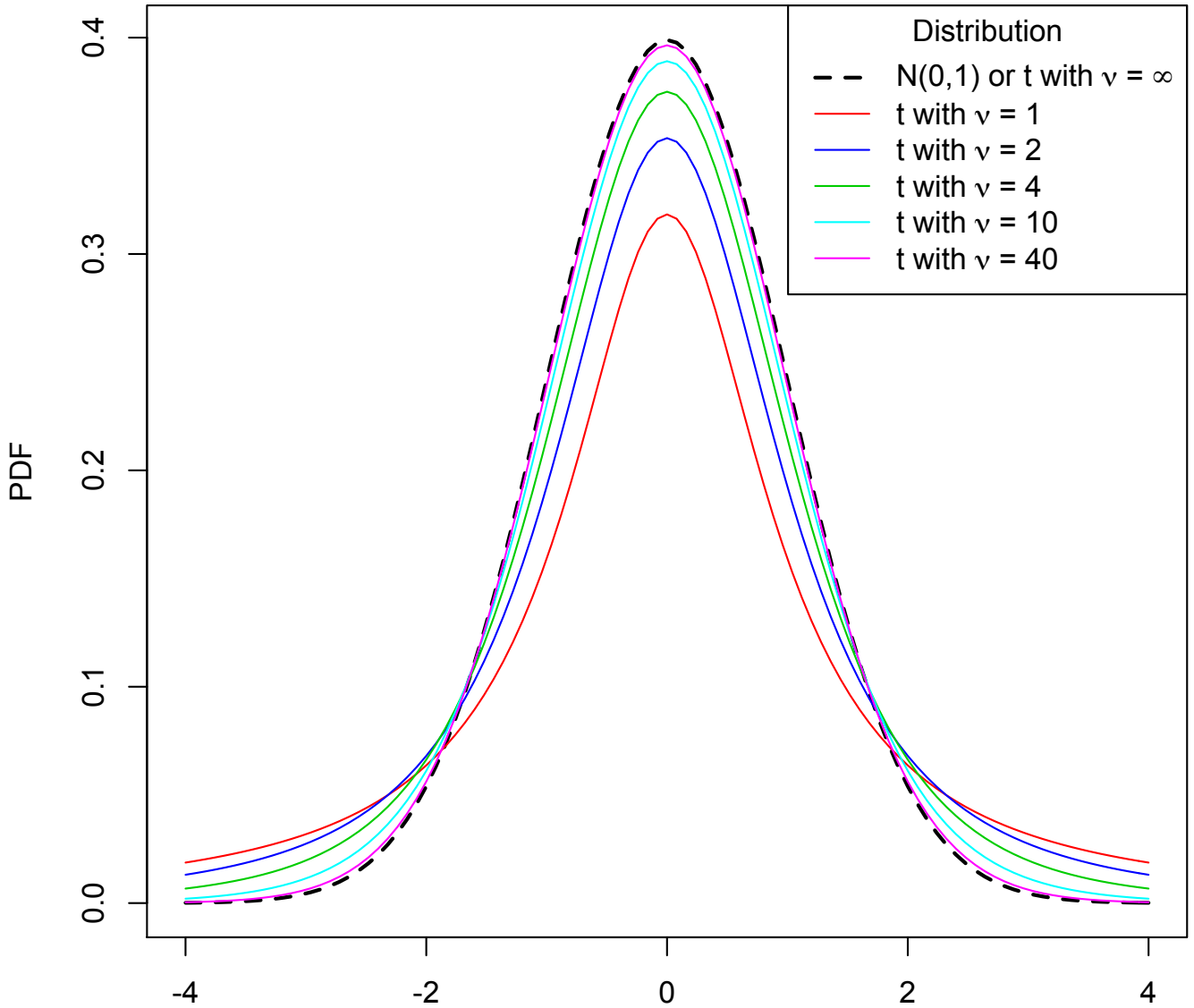
  How far from zero does $\hat{\beta}_1$ need to be?

- To test the null hypothesis, we compute a test statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- Assuming $H_0 : \beta_1 = 0$ is true, this follows a t-distribution with degrees of freedom $\nu = n - 2$.
- These tests are often referred to as t-tests.

Student's t Distribution



Hypothesis Test for β_1

- To make decisions based on this test statistic, we either compute a p-value or compare to a critical value
- A p-value calculates the probability of observing any value equal to or more extreme than t.
- If the p-value is smaller than our significance level α, the null hypothesis is rejected.
- If t is more extreme than the critical value, denoted $t^{\nu}_{\alpha/2}$, the null hypothesis is rejected

p-value

- Roughly speaking, *p*-values give the probability of seeing such a substantial associate between the predictor and the response due to chance.
- If there really is no relationship between X and Y, there should be a small chance that we observe a $\hat{\beta}_1$ that deviates too far from zero.
- The smaller the *p*-value, the less likely it is that we should sample data producing a slope steeper than the one we observed.
- If this probability gets small enough ($< \alpha$) we can infer that there is an association between the predictor and the response

Sampling Distribution for $\hat{\beta}_0$

- What about $\hat{\beta}_0$?
- $$\frac{\hat{\beta}_0 - 0}{RSE\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t(n-2)$$

Sampling Distribution for the mean response

- Let $x = x_0$, any value of the x within the range of the original data on x used to fit the model.
- We wish to estimate the mean response, say $E(y) = \beta_0 + \beta_1 x_0$.
- What about $E(y)$?
- $$\frac{\bar{y} + \hat{\beta}_1(x_0 - \bar{x}) - E(y)}{RSE\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$
- R command:
  predict(fit, newdata=data.frame(x=x0),
  interval="confidence",level=0.9)

Sampling Distribution for the prediction of new observations

- For a new observation $x = x_0$, we have $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$.
- We wish to estimate y_0.
- What about y_0?
- $$\frac{\bar{y} + \hat{\beta}_1(x_0 - \bar{x}) - y_0}{RSE\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$
- R command:
  predict(fit, newdata=data.frame(x=x0),
  interval="prediction",level=0.9)

Measuring Fit

- Once we have determined that it is likely that there is a significant relationship between X and Y, we can ask: how well does the linear model fit the data?
- The quality of a linear regression fit is typically assessed using two related quantities:
  1. the residual standard error (RSE) and
  2. the R^2 statistic

Measuring Fit

- The **coefficient of determination** R^2 is the proportion of the total variation that is explained by the regression line and is given by

$$R^2 = \frac{\text{TSS-RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

- R^2 is a measure of how close to the fitted line is to the observed data.

- It is easily interpretable and always lies between 0 and 1.

Measuring Fit

- Recall that RSE is an estimate of the standard deviation of e_i.
- RSE serves as a measure of the amount of training error in the model.
- Roughly speaking, it is the average amount that the response will deviate from the true regression line.
- Unlike R^2, this is not bounded by 1 so it is not alwaysclear what constitutes a good RSE.

Linear Correlation

- We can also find the **Pearson's product-moment correlation coefficient**(or simply the Pearson's r, or correlation coefficient) through the following formula

$$r = Cor(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

- Correlation coefficient tells us if there is a negative or positive linear relationship between x and y (and measures the strength of that relationship).

- It is a value between -1 and 1.

- Note: for simple linear regression, $r = \sqrt{R^2}$

Pearson's r

- Values of r close to 1 or -1 reflect a strong linear relationship between y and x.
- The sign of r indicates direction of association (i.e positive or negative correlation).
- If $r = 1$ or -1 then all points fall directly on a line.
- The closer r is to 0, the weaker the linear relationship between x and y.
- r is unitless, (i.e. r not affected by units of x and y)

A Word of Caution

- Neither the coefficient of determination or the correlation coefficient should be used alone to suggest a relationship.

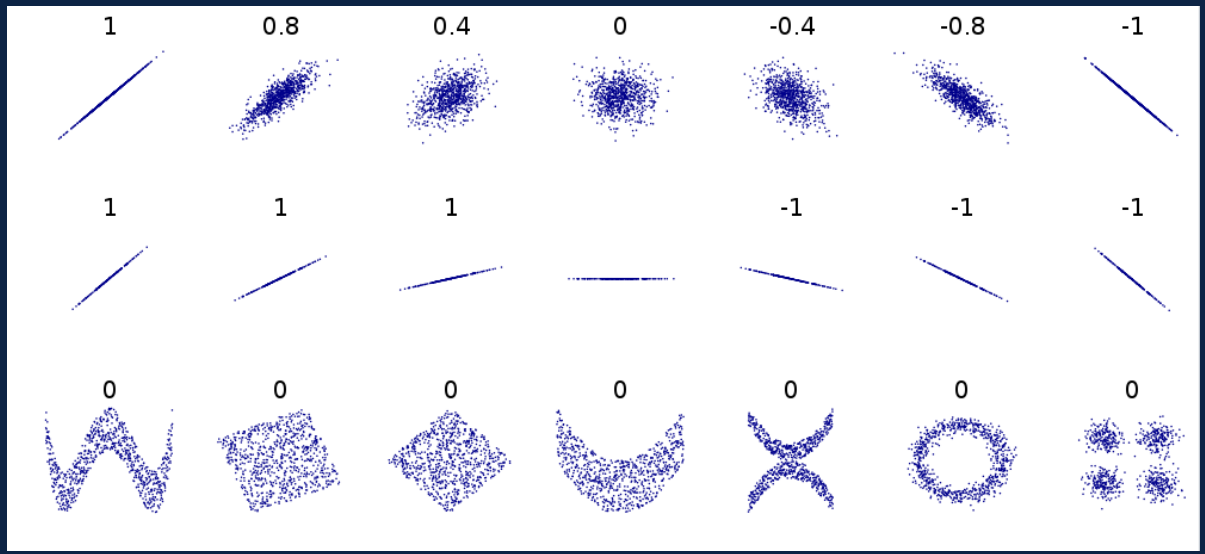
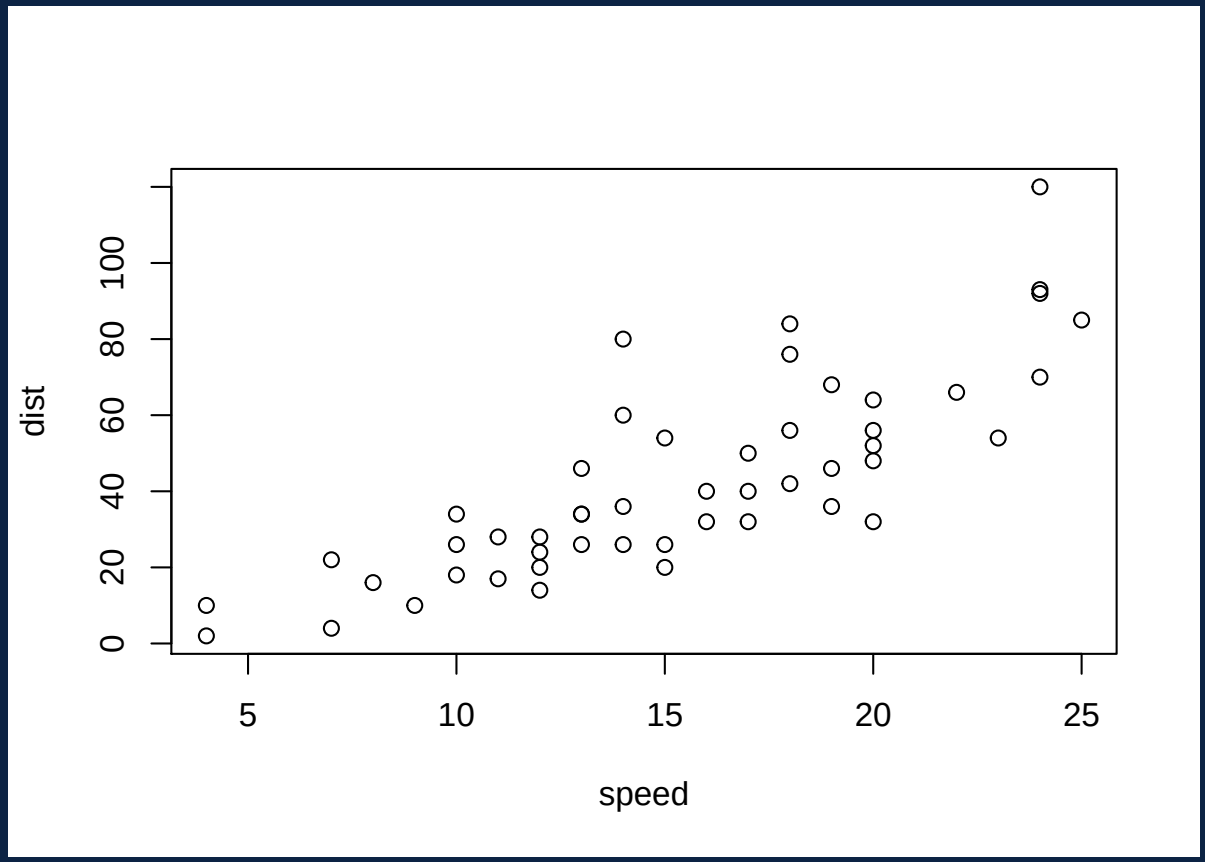


Image from `http://en.wikipedia.org/wiki/Correlation_and_dependence`

- A scatter plot of X versus Y is the most effective way to do this.

Example: Cars



Example: Cars

```
> require(stats); require(graphics)
> y=cars$dist;x=cars$speed
> carlm <- lm(y~x)
> summary(carlm)

Call:
lm(formula = dist ~ speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

SLR R Output

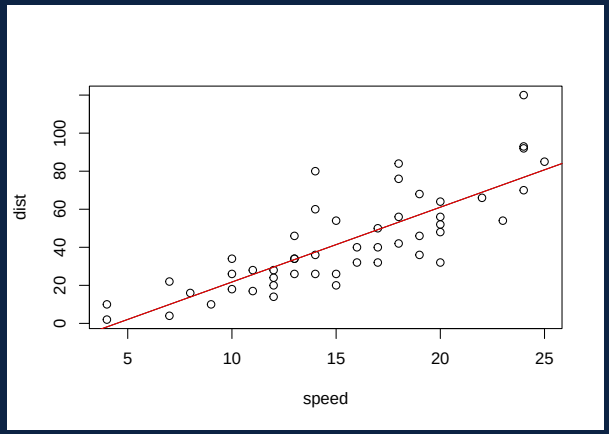
In general, the R output for a SLR model looks like this:

```
Call:
lm(formula = y ~ x)

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	(Aside) Tests for:
(Intercept)	b_0	$SE(b_0)$	$t_{obs} = \frac{b_0}{SE(b_0)}$	$2 * P(T > \|t_{obs}\|)$	$H_0 : \beta_0 = 0$
X variable	b_1	$SE(b_1)$	$t_{obs} = \frac{b_1}{SE(b_1)}$	$2 * P(T > \|t_{obs}\|)$	$H_0 : \beta_1 = 0$

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

Residual standard error: RSE (estimate of σ) on (n − 2) degrees of freedom
Multiple R-squared:  r^2 , Adjusted R-squared:
```

Example: Cars




```
> plot(x,y,xlab="speed",ylab="dis");abline(carlm,col="red")
> predict(carlm, newdata=data.frame(x=15),
 interval="confidence",level=0.9)
       fit      lwr      upr
1 41.40704 37.74843 45.06564
> predict(carlm, newdata=data.frame(x=30),
 interval="prediction",level=0.9)
       fit    lwr      upr
1 100.3932 72.425 128.3613
```


THE UNIVERSITY OF BRITISH COLUMBIA