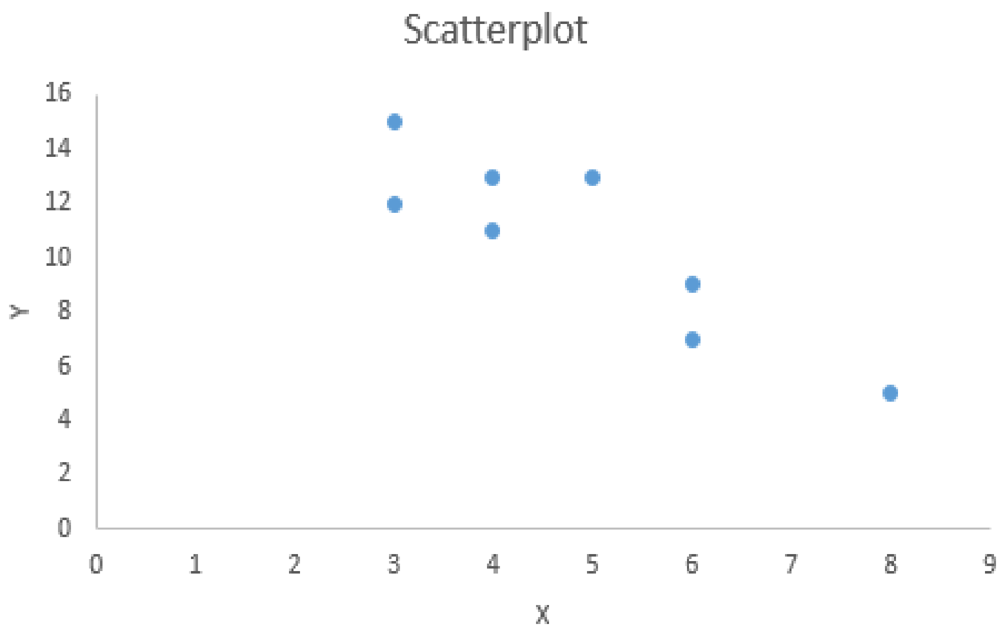
Week 8
Bivariate Data

Scatterplot

- A scatterplot is a graph that shows the relationship between two quantitative data sets. The two sets of data are graphed as ordered pairs in a coordinate plane.
- When 2 variables are measured for each member of a group, the data are called two dimensional or *bivariate* such as height and weight of students, study time and the exam results.
- The values of the *independent variable* (that stands alone and is not changed by the other variables you are trying to measure) are plotted on the horizontal axis, and the values of the *dependent variable* (the response that is being predicted or measured) on the vertical axis.



Covariance

- Covariance measures the co-variation between 2 quantitative variables that have different units of measurement. It tells us whether units are increasing or decreasing. In other words, it tells us whether two variables are positively or inversely related.
- A positive covariance means the variables are positively related since they move together in the same direction.
- A negative covariance means the variables are inversely related since they move together in the opposite direction.
- However, it is not possible to measure the degree to which the variables move together because covariance does not use one standard unit of measurement.
- The formula for calculating the sample covariance is shown below:

$$\mathbf{C(x, y) = \frac{S_{xy}}{n-1}}$$

where

C(x, y) = covariance of x and y

$$\mathbf{S_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - \frac{\sum x \sum y}{n} = \sum xy - n\bar{x}\bar{y}}$$

Correlation

- Correlation is another way to determine how two quantitative (metric) variables are related.

- Not only does it tell us the direction of a linear relationship between two variables, but it also tells us the degree to which the variables tend to move together. Since correlation standardizes the measure of interdependence between two variables, it tells us how closely they move.

- The correlation coefficient, **r**, is calculated as

$$\mathbf{r} = \frac{\mathbf{Cov(x,y)}}{\mathbf{Sd(x)*Sd(y)}} = \frac{\mathbf{S_{xy}}}{\sqrt{\mathbf{S_{xx}S_{yy}}}}$$

where
Cov(x, y) = covariance of x and y

Sd(x) = standard deviation of x = $\sqrt{\frac{S_{xx}}{n-1}}$

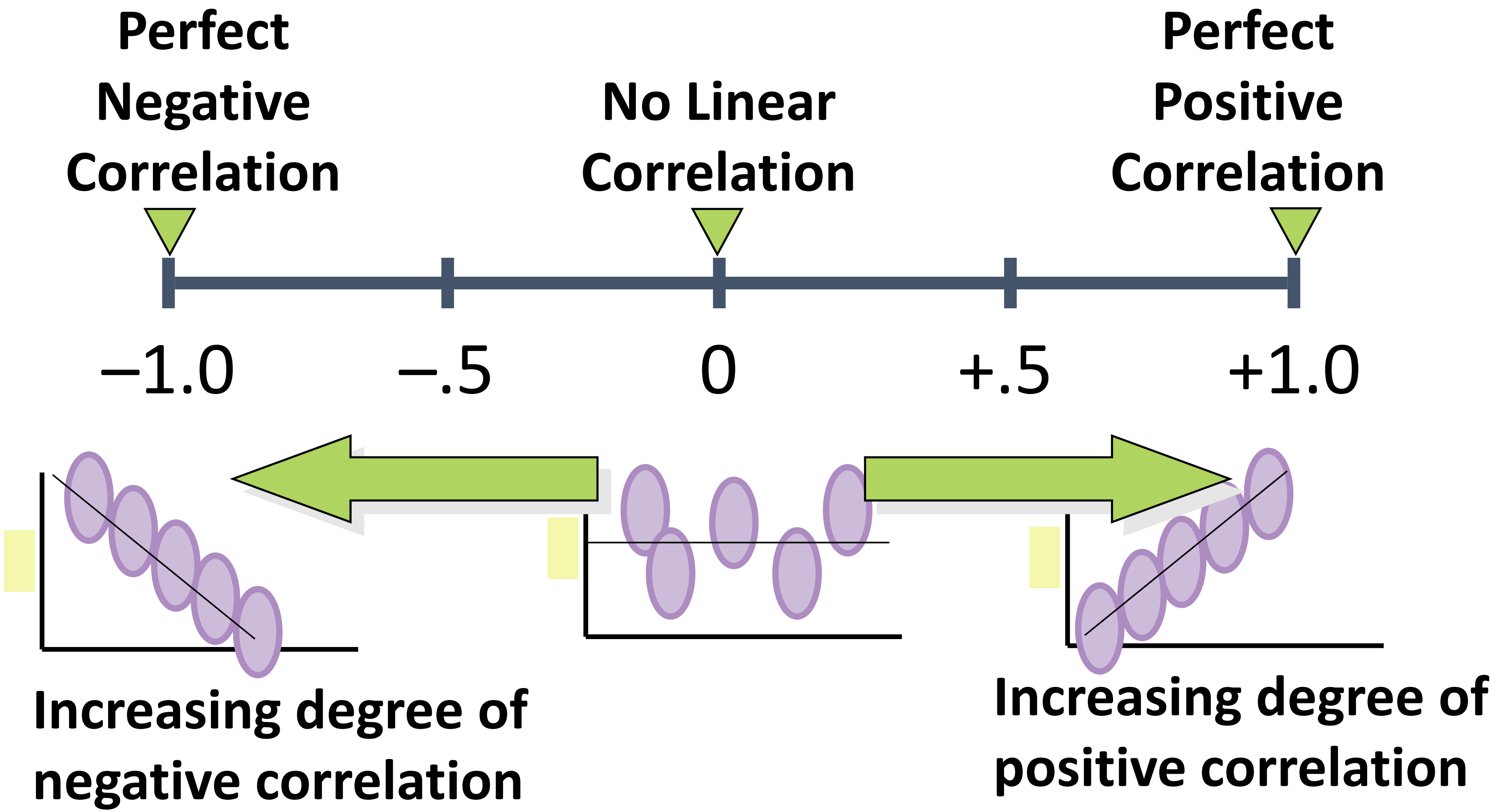
Sd(y) = standard deviation of y = $\sqrt{\frac{S_{yy}}{n-1}}$

$$S_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = \sum x^2 - n\bar{x}^2$$

$$S_{yy} = \sum(y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = \sum y^2 - n\bar{y}^2$$

- It is the sign of the covariance that determines the sign of **r**.

Coefficient of Correlation Values

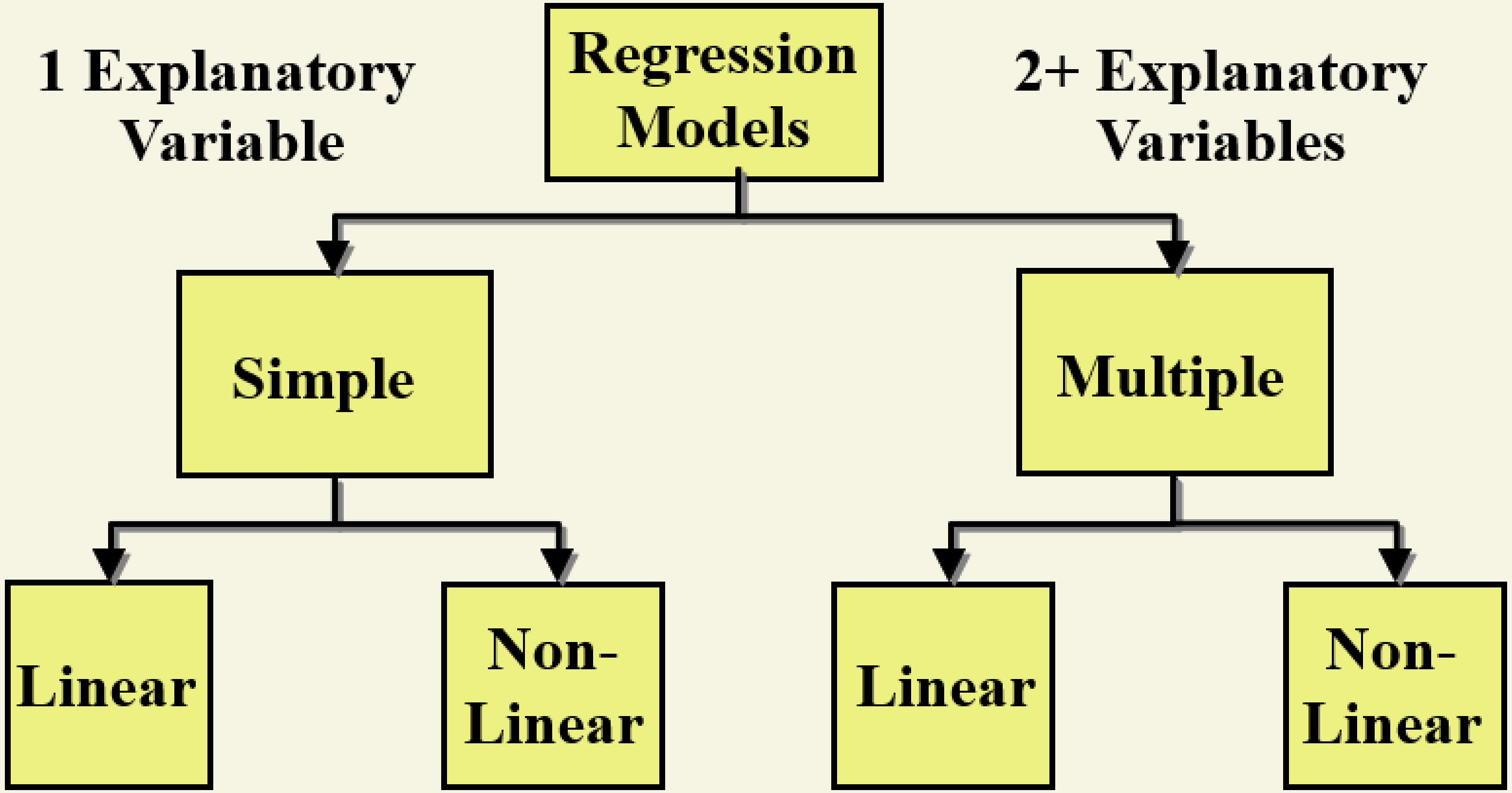


Correlation

- For any data set, $-1 \leq r \leq 1$
- *r* close to -1 => strong negative linear relationship
- *r* close to 0 => weak or non-existent linear relationship
- *r* close to 1 => strong positive linear relationship
- Rules of thumb about the strength of correlation coefficients:

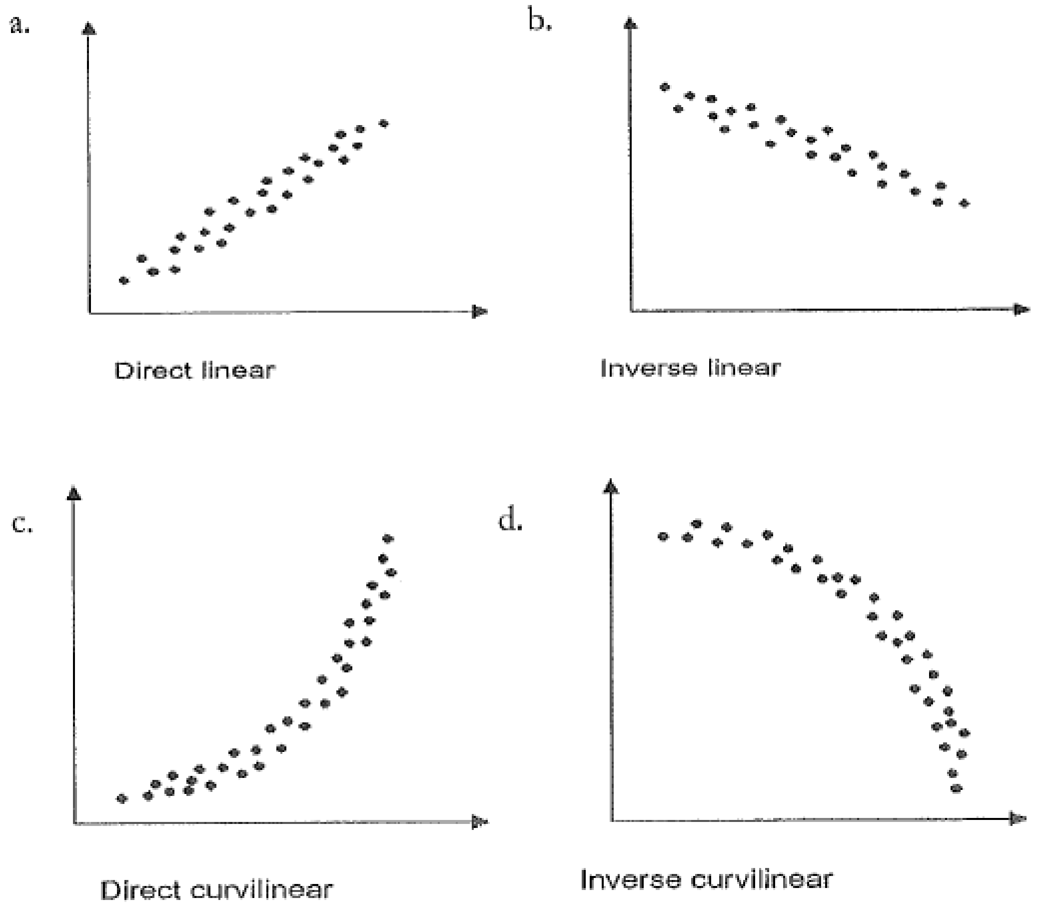
Range of coefficient	Description of strength
± 0.81 to ± 1.00	Very strong
± 0.61 to ± 0.80	Strong
± 0.41 to ± 0.60	Moderate
± 0.21 to ± 0.40	Weak
± 0.00 to ± 0.20	Very weak

Types of Regression Models



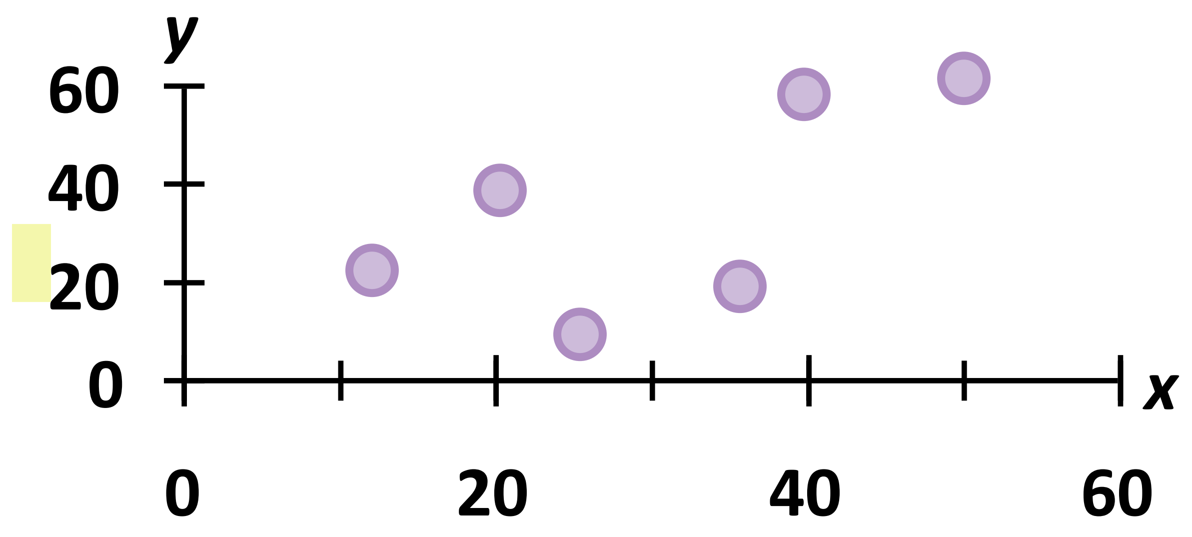
Simple linear regression

- If there is a strong relationship between two variables, knowing the value of one of them will help in predicting the value of the other.
- Regression is the technique of using values of one variable to predict the values of another related variable.
- Possible types of relationships between variables:
  - If the two variables are related, the actual pattern that emerges in the scatterplot will indicate the type of relationship.
  - The diagram to the right show the possible relationships between x and y in scatterplots:



Scattergram

1. Plot of all (x_i, y_i) pairs
2. Suggests how well model will fit

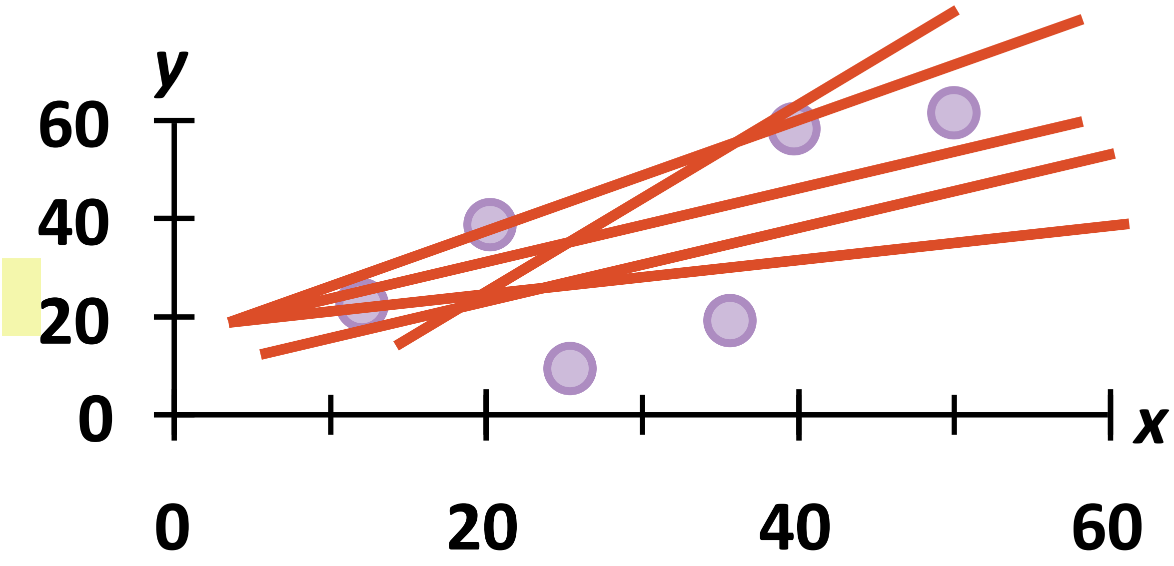


Line of best fit

- A line of best fit is a straight line drawn on a scatterplot and meant to mimic the trend of the data. Hence, it is also called a trend line.
- A line of best fit may pass through all of the data points, some of the data points, or none of the data points. The idea is to get a line that has equal numbers of points on either side. It can be used to estimate data on a graph.
- A line of best fit through the mean point is a straight line drawn through the mean point.
- When you use the line of best fit to estimate the values between the given points, you interpolate.
- When you use the line of best fit to estimate the values outside the given points, you extrapolate.
- You should extrapolate only when the linear relationship is thought to be appropriate.

Thinking Challenge

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



Simple Linear Regression

- Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

  Assume $E(\varepsilon|X) = 0$
- Population regression line: $E(Y|X) = \beta_0 + \beta_1 X$
- Using the least squares method, Least squares line / fitted regression line: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

- Random error versus residual
  - Random error (ε) = $Y - E(Y|X)$
  - Residual ($\hat{\varepsilon}$) = $Y - \hat{Y}$

The least squares regression line

- A method that results in a single, best line of best fit, is called the least squares principle. The least squares regression line is the best line of best fit that minimizes the sum of the vertical distances from the actual data points to the line of best; i.e., min $\sum(y - \hat{y})^2$
- The formula for the least squares regression line is $\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 \mathbf{x}$ where
- $\hat{y}$, read y hat, is the estimated value of the y variable for a selected x value.
- $\hat{\beta}_0$ is the y-intercept. It is the estimated value of y when x = 0.
- $\hat{\beta}_1$ is the slope of the line, or the average change in $\hat{y}$ for a one unit change in x.
- x is any value of the independent variable that is selected.

$$\hat{\boldsymbol{\beta}}_1 = \frac{\mathbf{Cov(x,y)}}{\mathbf{var(x)}} = \frac{\mathbf{S_{xy}}}{\mathbf{S_{xx}}} = \mathbf{r}\frac{\mathbf{S_y}}{\mathbf{S_x}}$$

$$S_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = \sum xy - n\bar{x}\bar{y}$$

$$S_{xx} = \Sigma(x - \bar{x})^2 = \sum x^2 - n\bar{x}^2$$

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

covariance $\mathrm{cov}_{XY} = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$

correlation $\mathrm{corr}_{XY} = \rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)]/(\sigma_X \sigma_Y),$

Least Squares

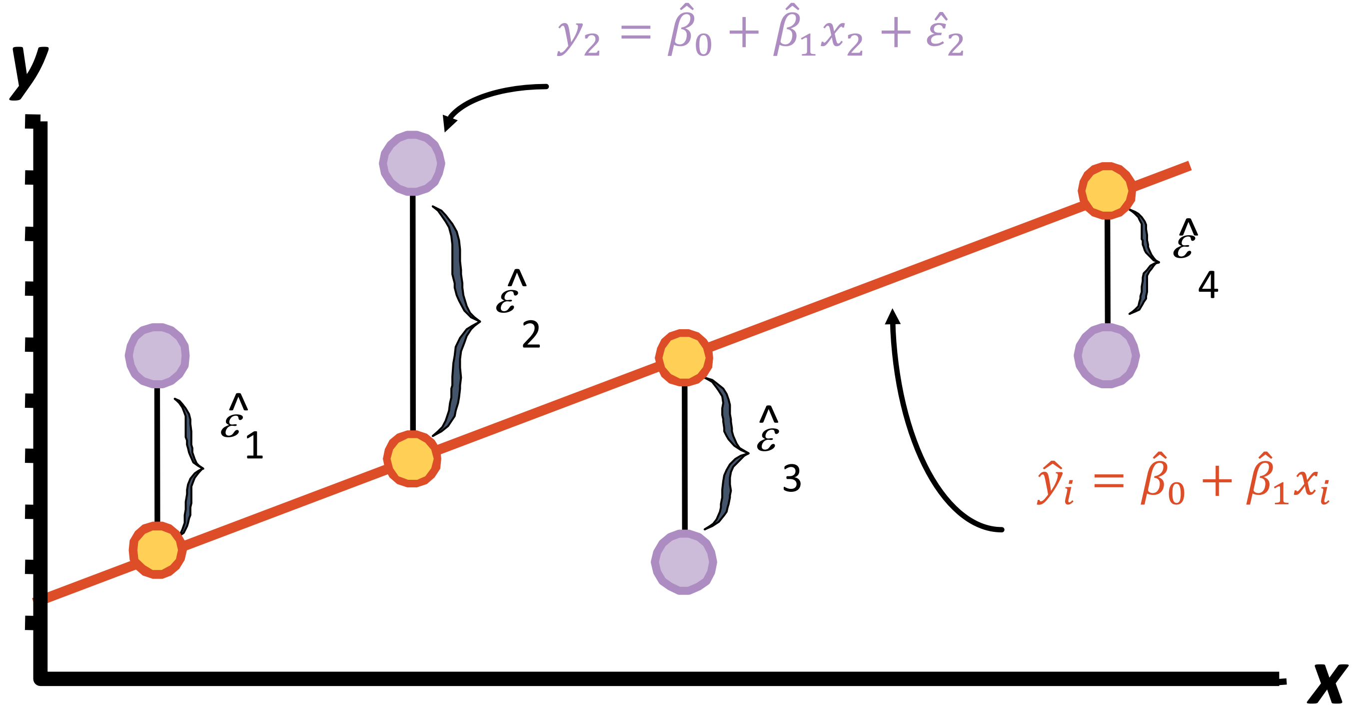
- ‘Best fit’ means difference between actual *y* values and predicted *y* values are a minimum
  - *But* positive differences off-set negative

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- Least Squares minimizes the Sum of the Squared Residuals (RSS)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Least Squares Numerically

To minimise $\mathrm{RSS} = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Differentiate RSS with respect to β_0 and β_1, we get

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Setting $\frac{\partial \mathrm{RSS}}{\partial \beta_0} = 0$ and $\frac{\partial \mathrm{RSS}}{\partial \beta_1} = 0$, we have

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \;\Rightarrow\; \sum_{i=1}^{n} \beta_0 + \sum_{i=1}^{n} \beta_1 x_i = \sum_{i=1}^{n} y_i \qquad (1)$$

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \;\Rightarrow\; \sum_{i=1}^{n} \beta_0 x_i + \sum_{i=1}^{n} \beta_1 x_i^{\,2} = \sum_{i=1}^{n} x_i y_i \qquad (2)$$

Coefficient Equations

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

Slope

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

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

Interpretation of Coefficients

1. Slope ($\hat{\beta}_1$)
   - Estimated y changes by $\hat{\beta}_1$ for each 1unit increase in x

2. Y-Intercept ($\hat{\beta}_0$)
   - Average value of y when $x = 0$

Simple Linear Regression Assumptions

1. The population model is given by $Y = \beta_0 + \beta_1 X + \varepsilon$
2. Values of X are treated as fixed / predetermined.
3. The mean value of ε given X is 0; i.e., $E(\varepsilon|X) = 0$. This implies that $E[E(\varepsilon|X)] = E(\varepsilon) = 0$ and $Cov(\varepsilon, X) = E(\varepsilon X) - E(\varepsilon)E(X) = 0$.
4. The sample size is greater than the number of parameters to be estimated.
5. There is a variation in the value of X.
6. Homoskedasticity means “same variance”. $Var(\varepsilon_i|X_i) = \sigma^2$; i.e., the conditional variance of the error term is the same given any value of X. Homoskedasticity implies that $Var(\varepsilon) = \sigma^2$. Note that heteroskedasticity means “different variance”. $Var(\varepsilon_i|X_i) = \sigma_i^2$; i.e., the variance of the error term depends on the value of X.
7. There is no autocorrelation in the random error term: $Corr(\varepsilon_i, \varepsilon_j|X_i, X_j) = 0$ for any i and j, $i \neq j$.

Random Error Variation

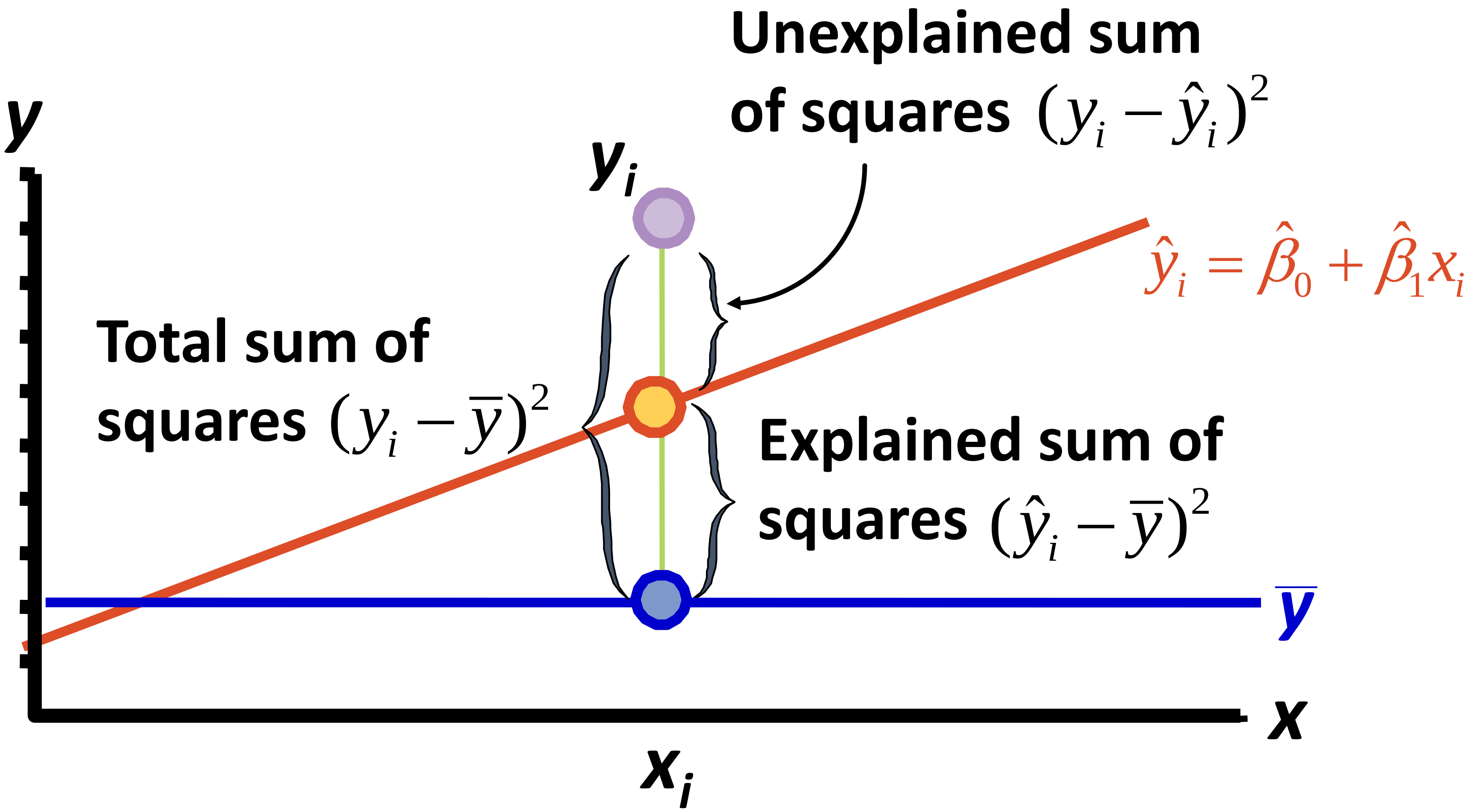
- Variation of actual y from predicted $\hat{y}$
- Measured by the standard error of regression model
  - Sample standard deviation of $\hat{\varepsilon}$ (SER)
- Affects several factors
  - Parameter significance
  - Prediction accuracy

Estimation of σ^2

$$\hat{\sigma}^2 = \mathrm{MSE} = \frac{\mathrm{RSS}}{n-p} \quad \text{where} \quad \mathrm{RSS} = \sum (y_i - \hat{y}_i)^2$$

$$\hat{\sigma} = \mathrm{SER} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\mathrm{RSS}}{n-p}} = \sqrt{\mathrm{MSE}}$$

Variation Measures



Decomposition of sample variation in Y

$S_y^2 = \frac{\sum(y - \bar{y})^2}{n-1} = \frac{TSS}{n-1}$. Therefore, TSS = (n – 1)S_y^2

TSS = RegSS + RSS

- Total sum of squares (TSS) captures the total variation in Y.

  TSS = $\sum(y - \bar{y})^2$
- The regression sum of squares (RegSS) captures the variation in Y that is explained by the model.

  RegSS = $\sum(\hat{y} - \bar{y})^2$
- The residual sum of squares (RSS) captures the variation in Y that is unexplained by the model.

  RSS = $\sum(y - \hat{y})^2$

ANOVA table

Source of Variation	SS	df	MS	F
Regression	RegSS	p – 1	MSR	F = $\frac{\text{MSR}}{\text{MSE}}$
Residual	RSS	n – p	MSE	
Total	TSS	n – 1		

RegSS + RSS = TSS

(p – 1) + (n – p) = (n – 1)

MSR + MSE ≠ S_y^2

Overall F test – test for the usefulness of the model

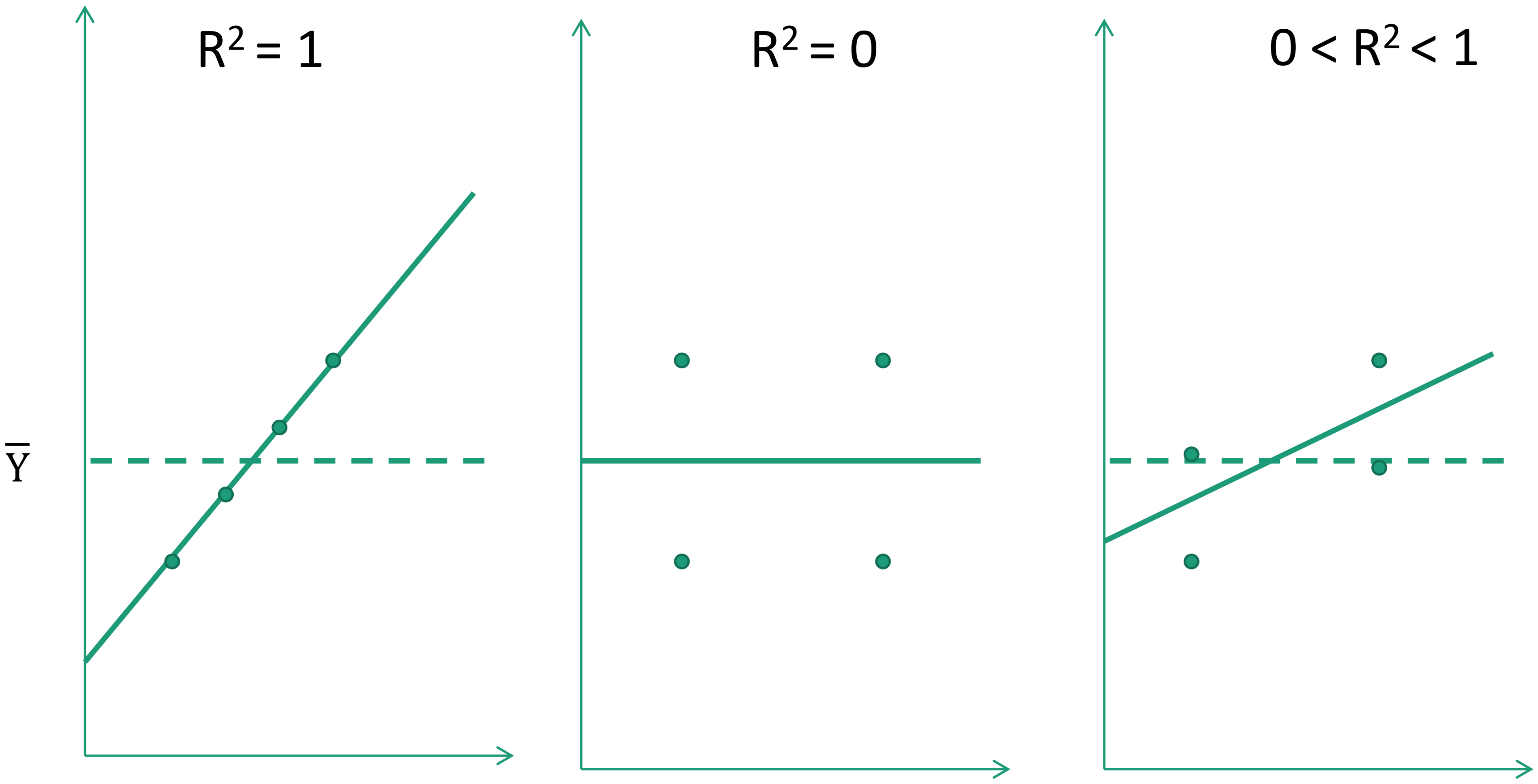
- $E(Y|X) = \beta_0 + \beta_1 X$
- H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$

- $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
- H_0: $\beta_1 = \beta_2 = 0 = \ldots = \beta_k = 0$ against H_1: At least one $\beta_j \neq 0$

Goodness-of-fit

- TSS = RegSS + RSS
- Dividing both sides of the equation by TSS, we have $1 = \frac{\text{RegSS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}}$
- Coefficient of determination (R^2) = $\frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$
- R^2 is a summary measure that indicates how well the sample regression line fits the data
- The higher the R^2, the more the explanatory power of the model
  - Perfect fit ($R^2 = 1$): All data points form a nonflat straight line
  - No fit ($R^2 = 0$): Data points do not form a straight line. The SRL is flat.
  - Some fit ($0 < R^2 < 1$): The higher the R^2, the closer are the data points clustered around the SRL. The SRL is not flat
- R^2 measures the proportion of the variation in Y that is explained by the model.

Goodness-of-fit: three cases

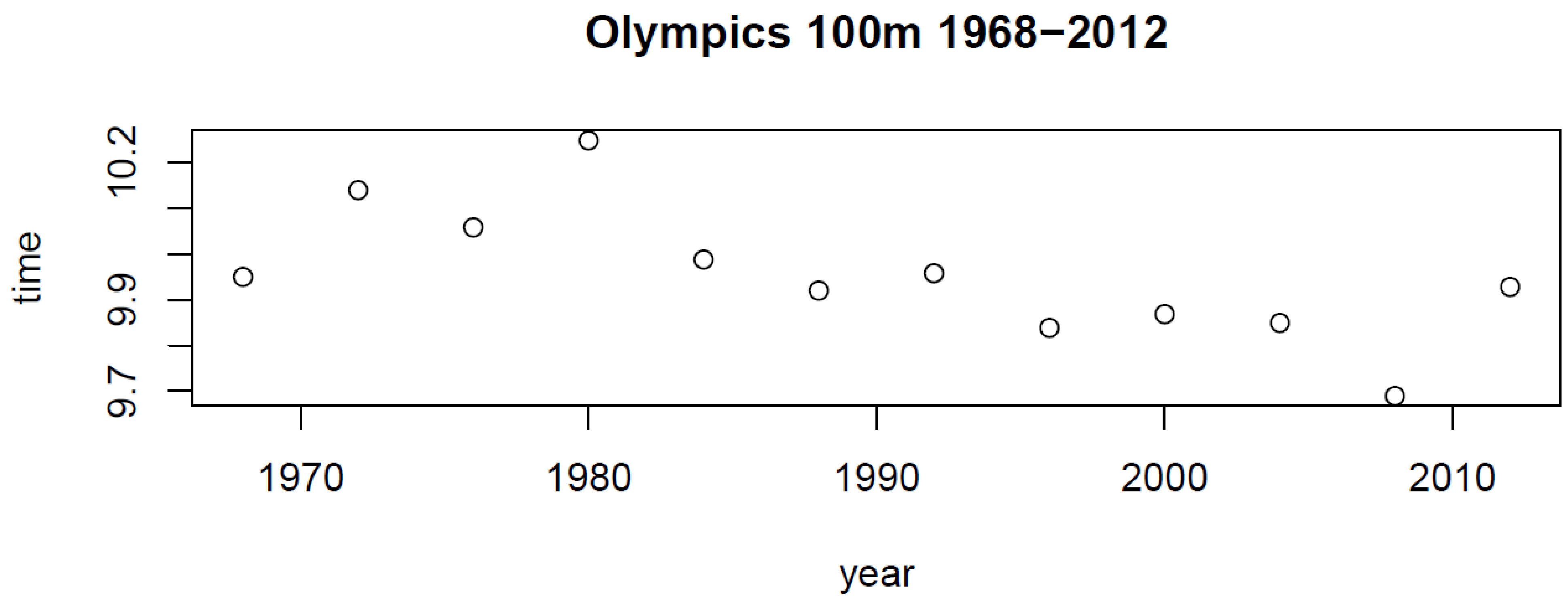


Theory – Residual Plot

- A residual plot is a scatter plot of the residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - (\hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x})$ vs. $\boldsymbol{x}$.
- "Does the plot look random, or is there any pattern?"
- If the plot is random, then the least squares line fit is good.
- If the plot shows a relationship between $\hat{\boldsymbol{\epsilon}}$ and $\boldsymbol{x}$, then the least squares line is not adequate and we may need to consider a more complex function or a transformation (e.g. $y = x^2$ or $y = \log(x)$).

Example – Scatter Plot

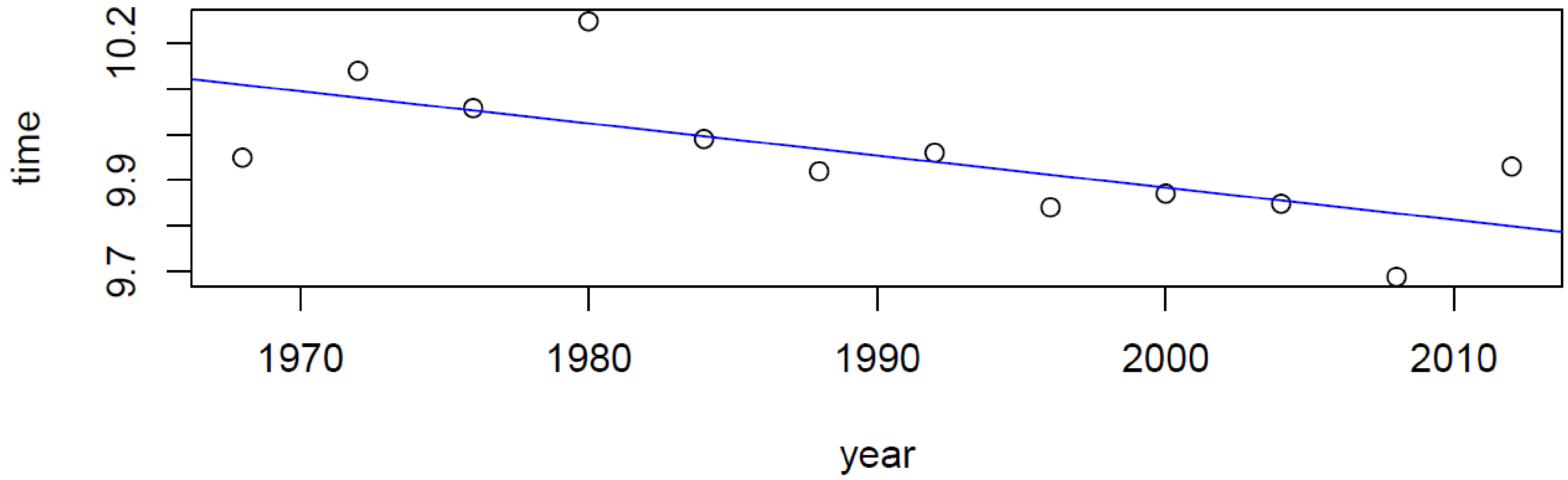
```
olympics = read.csv("../datasets/Olympics100m.csv", header=T)
plot(Time ~ Year, data=olympics, xlab="year", ylab="time",
        main="Olympics 100m 1968-2012")
```



Code – Least Squares Regression Line for Bivariate Data

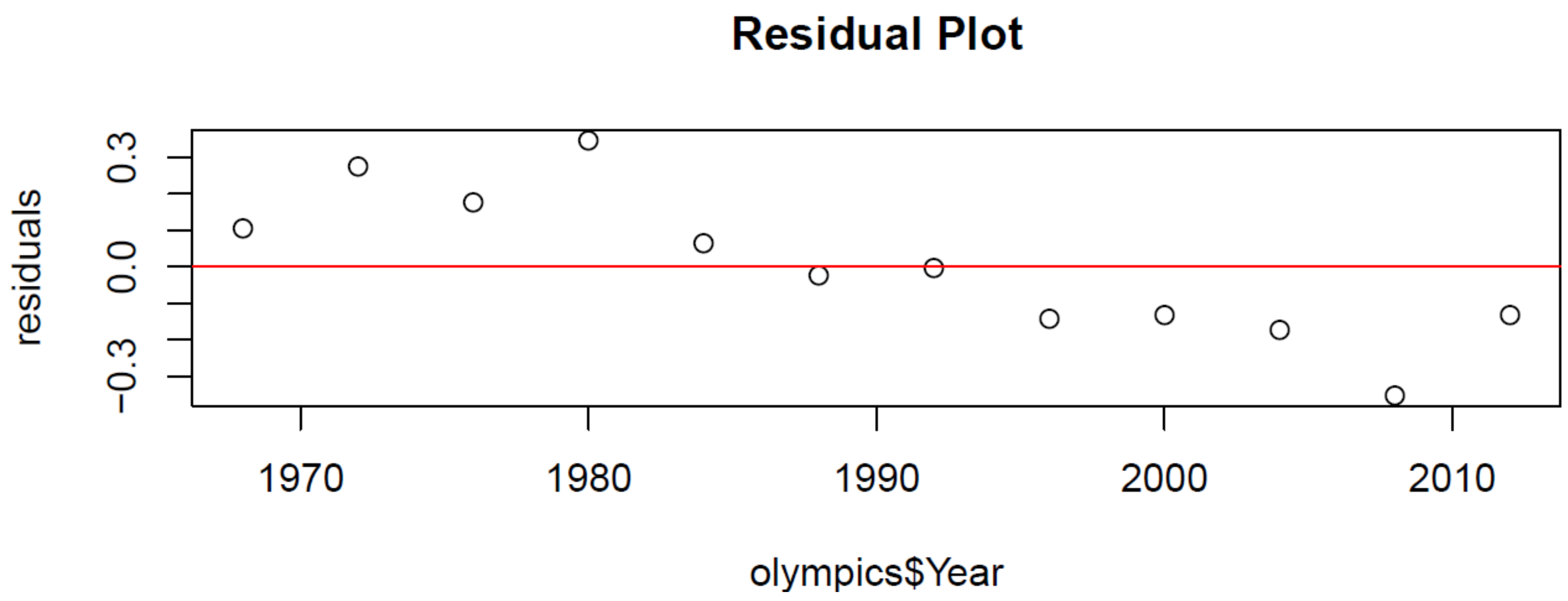
You can fit the least squares regression line in R using the `lm()` function.

```
model = lm(Time ~ Year, data=olympics)
model$coeff
plot(Time ~ Year, data=olympics, xlab="year", ylab="time")
abline(model$coeff, col="blue")
```



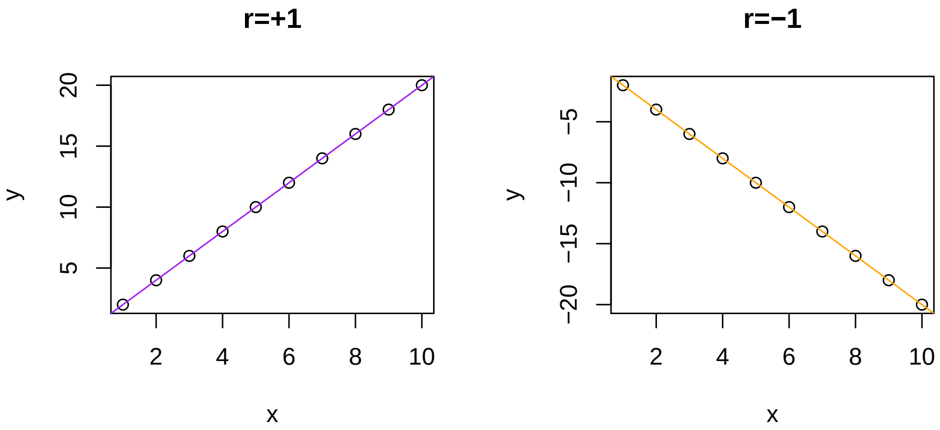
Example – Residual Plot

```
residuals = model$res
plot(olympics$Year, residuals, main="Residual Plot")
abline(h=0, col="red")
```



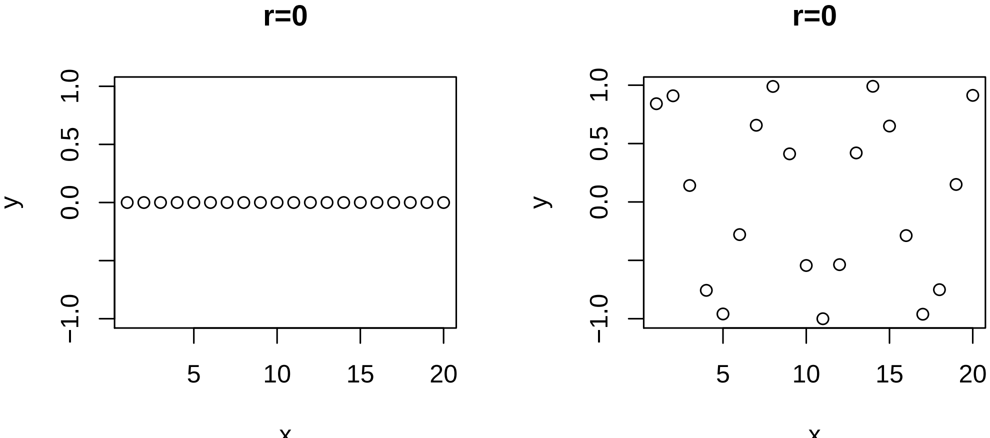
Theory – Pearson's Sample Correlation Coefficient

- $r = \pm 1$ corresponds to a perfect linear correlation, with all the data points lying on the least squares line.

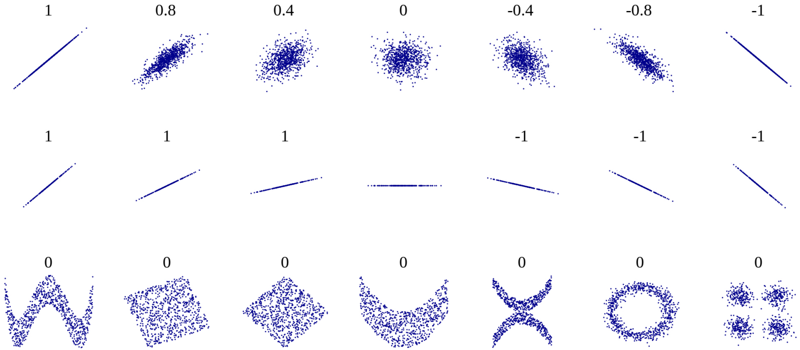


Theory – Pearson's Sample Correlation Coefficient

- $r = 0$ indicates no linear correlation, for example a line with zero slope, or a random scatter, or a non linear relationship.



Examples of Correlation Coefficients



▸ Guess the correlation game

Relationship between Correlation Coefficient and Slope

There is an interesting relationship between r and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r\frac{\sqrt{S_{yy}/(n-1)}}{\sqrt{S_{xx}/(n-1)}} = r\frac{s_y}{s_x}$$

Hence:

- The sign of r reflects the trend (slope) of the data.
- r is unaffected by a change of scale or origin.

Theory – Coefficient of Determination

The coefficient of determination is the proportion of variability of y explained by x for a model, or in our context, the proportion of variability explained by the linear regression.

$$\boxed{r^2 = \frac{s_y^2 - s_\epsilon^2}{s_y^2} = \frac{S_{yy} - \text{RSS}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}}$$

where s_ϵ^2 is the sample variance of the residuals and

$$\text{residual sum of squares} = \text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = (n-1)s_\epsilon^2.$$

Properties:

- $0 \leq r^2 \leq 1$
- $r^2 = 1$ arises when $\hat{\epsilon}_i = 0$ for all i and $s_\epsilon^2 = 0$, i.e. all of the variability of the model is associated with the linear regression – all points are on the regression line.

Theory – Coefficient of Determination

- $r^2 \approx 1$ arises when $\hat{\epsilon}_i = 0$ is small compared to s_y^2, i.e. most of the variability of the model is associated with the linear regression.
- $r^2 = 0$ arises when $s_\epsilon^2 = s_y^2$, i.e. none of the variability of the model is associated with the linear regression.
- $r^2 \approx 0$ arises when $s_\epsilon^2 \approx s_y^2$, i.e. almost none of the variability of the model is associated with the linear regression.
- Note that r^2 can be small and the model may still be "useful" as there may naturally be a low association between x and y.

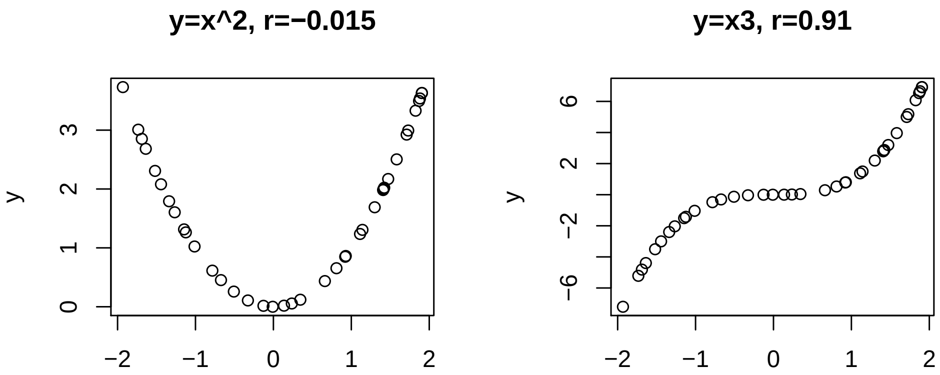
Example – correlation coefficient and coefficient of determination

```
cor(olympics$Year, olympics$Time)

## [1] -0.6912573

cor(olympics$Year, olympics$Time)^2

## [1] 0.4778366
```

Hence, the linear association between year and time for the Olympics 100m sprint is -0.7 (fairly high). 48% of the variation in times is explained by the variation in years.

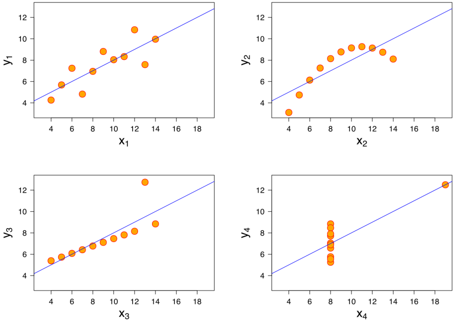
Correlation $\neq$ Causation

- Correlation does not imply causation.
- A high value of r does not necessarily imply a causal relationship between x and y. For example, December temperature and consumer spending. ▸ Spurious Correlation
- Likewise, causation does not imply high correlation.



Cautionary Tale – Anscombe's Quartet

- The same value of r can correspond to very different models.
- 'Anscombe's Quartet' – type `anscombe` in `R` to see the data



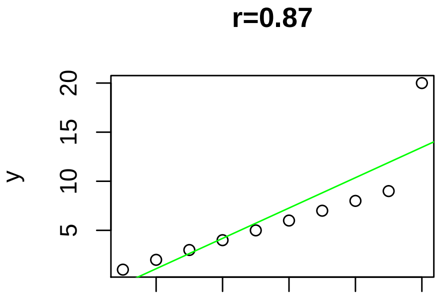
$\bar{x} = 9,\ s_x^2 = 11,\ \bar{y} = 7.5,\ s_y^2 = 4.127,\ r = 0.816$ and $y = 3 + 0.5x$.

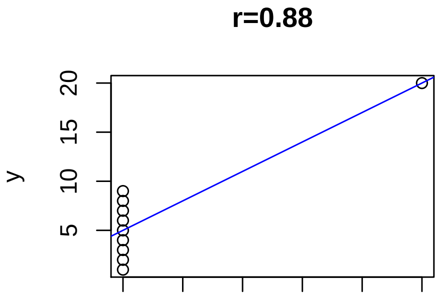
Image: Wikimedia Commons CC BY-SA 3.0
Anscombe, F. (1973) Graphs in Statistical Analysis, *American Statistician* **27**(1)

cook leverage polynomial datasaurus

Outlier

- An outlier is an observation that may arise from a different distribution to the bulk of the data.
- Even one outlier can distort the model.
- How do you find an outlier?
- There are a number of ways but we will use the Cook's distance and high-leverage points as guidelines for detecting outliers.





Example – Cook's distance

- We will *not* delve into the calculation of the Cook's distance in this course.
- You can find the Cook's distance in R for a linear model object M1 by using `cooks.distance(M1)`.
- If the Cook's distance is greater than one, then we say that the corresponding observation is an outlier.[1]

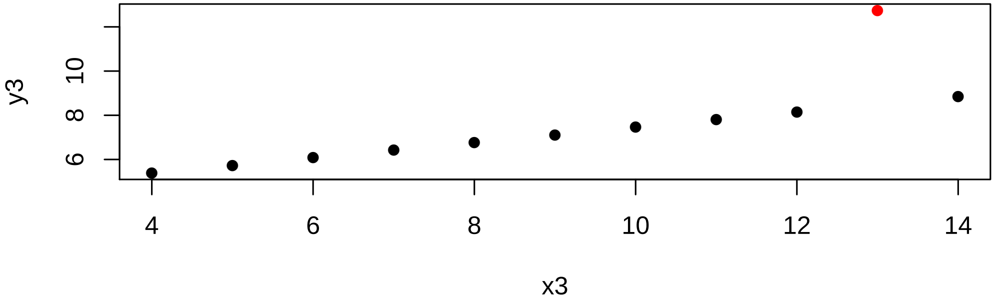
anscombe

```
M1 = lm(y3 ~ x3, data=anscombe)
round(cooks.distance(M1), 3)

##     1     2     3     4     5     6     7     8     9    10    11
## 0.012 0.002 1.393 0.005 0.026 0.301 0.001 0.034 0.060 0.000 0.007
```

[1] Others may use a different rule of thumb but we'll use this criteria for this course.

Example – Cook's distance

```
anscombe[3, c("x3", "y3")]
col <- rep("black", nrow(anscombe))
col[3] <- "red"
plot(y3 ~ x3, data=anscombe, col=col, pch=16)
```



Example – Leverage

- Again, we will *not* delve into the calculation of the leverage in this course.
- You can find the leverage of a point in `R` for a linear model object `M1` by using `lm.influence(M1)$h`.
- If the leverage is greater than $2p/n$ where p is the number of regression parameters in the model and n is the number of observations, then we say that the corresponding observation has a high leverage.[2]
- In a simple linear regression, we have two regression parameters: intercept and slope, so $p = 2$.

anscombe

```
M1 = lm(y4 ~ x4, data=anscombe)
lm.influence(M1)$h > 2 * 2 / 11

##     1     2     3     4     5     6     7     8     9    10    11
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
```

[2] Again, different rules of thumb maybe used elsewhere but we'll use this criteria for this course.

Example – Leverage

- Leverage is only dependant on explanatory variables, in other words, it is not affected by the response.
- It is a measure of how far away the independent variable values of the observations are from the other observations.

```
x = c(1, 2, 3, 4, 5, 10)
y1 = c(1, 2, 3, 4, 5, 10)
round(lm.influence(lm(y1 ~ x))$h, 3)

##     1     2     3     4     5     6
## 0.364 0.259 0.193 0.167 0.180 0.836

y2 = c(1, 2, 3, 4, 5, 1)
round(lm.influence(lm(y2 ~ x))$h, 3)

##     1     2     3     4     5     6
## 0.364 0.259 0.193 0.167 0.180 0.836
```

Summary – Fitting a Simple Linear Regression

Consider the following 3 steps:

1. Construct a Scatterplot: $\boldsymbol{y}$ vs $\boldsymbol{x}$.
2. If the plot looks linear, fit the Least Square's (Regression) Line: $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
3. Consider some model diagnostics such as
   - check the residual plot;
   - correlation coefficient or coefficient of determination;
   - check if there are outliers by using Cook's distance and leverage values.

Eye-balling the fit on Step 2 is always helpful than purely relying on Step 3. Why? It's a sanity check especially for the little bugs that like to creep into codes.

Theory – Inference for Linear Regression

Suppose we consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

$\boxed{\text{H}}$ We may be interested in $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

$\boxed{\text{A}}$ We assume $\epsilon_i \sim NID(0, \sigma^2)$, i.e. the errors are independent $N(0, \sigma^2)$ random variables.

$\boxed{\text{T}}$ Now

$$\hat{\beta}_1 \sim N\left(0, \sigma^2 \frac{1}{S_{xx}}\right) \quad \text{under } H_0 \text{ (proof not shown).}$$

So we use the t-test statistic

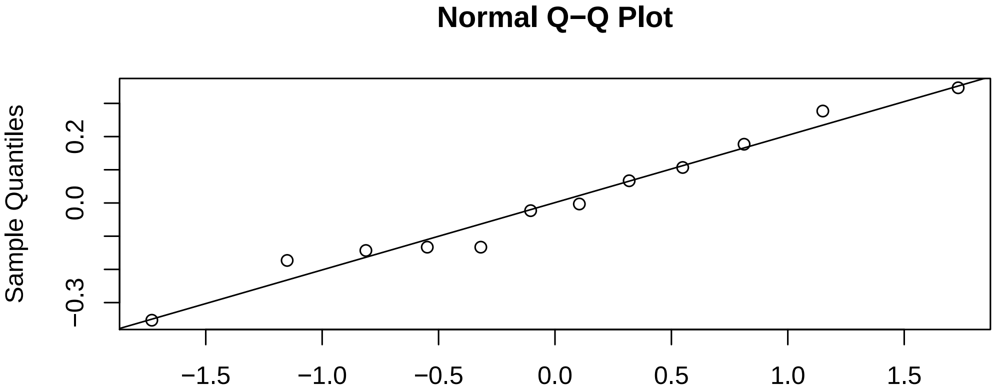
$$\tau = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2} \qquad \text{where} \qquad \hat{\sigma}^2 = \frac{\text{RSS}}{n-2}.$$

Theory – Q-Q Plot

- Quantile-Quantile plot or Q-Q plot is a graphical method to compare two probability distributions by plotting their quantiles against each other.
- If two distributions are similar, the points in the Q-Q plot will be approximately on the line $y = x$.
- For a normal Q-Q plot (where you compare with a standard normal distribution), if the points are roughly on a straight line $y = a + bx$ then the compared distribution is approximately $N(a, b^2)$.

Example – Q-Q Plot

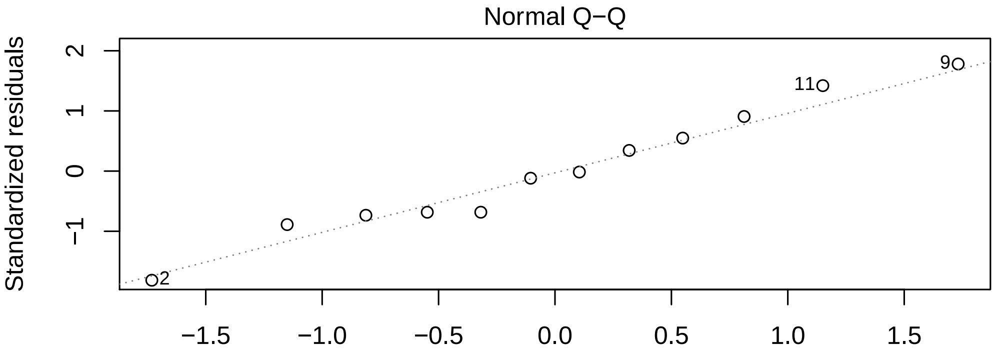
```r
q <- qqnorm(residuals)
abline(lm(y ~ x, data=q))
```



Example – Q-Q Plot

Or alternatively,

```
plot(model, which=2)
```



Theory – Polynomial Regression

- What if a linear trend doesn't seem sufficient (even after transformation)?
- A trend may be captured by a fit of the polynomial regression of order k:

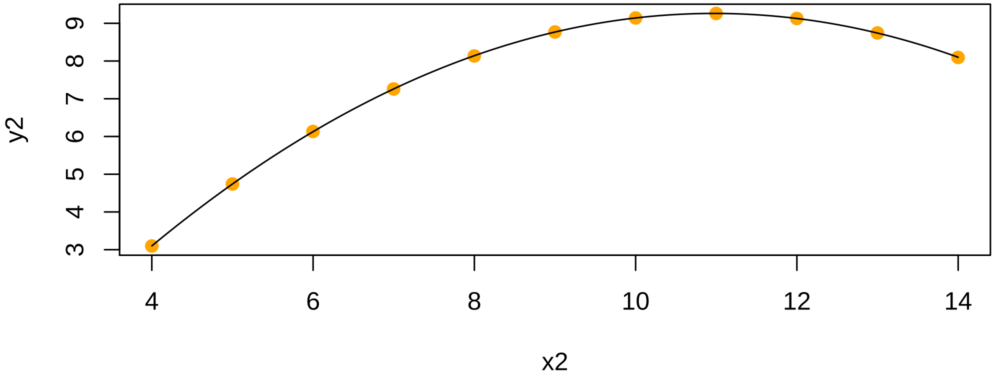
$$Y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_i^k + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim NID(0, \sigma^2)$.

anscombe

Example – Polynomial Fit

```
modelP2 = lm(y2 ~ 1 + x2 + I(x2^2), data=anscombe)
plot(y2 ~ x2, data=anscombe, pch=16, col="orange", cex=1.2)
beta = coef(modelP2)
yhat = function(x) beta[1] + beta[2] * x + beta[3] * x^2
curve(yhat, 4, 14, add=T)
```



Example – Polynomial Fit

You can fit the same model in the previous slide using `poly` as below.

```
modelP2 = lm(y2 ~ poly(x2, 2, raw=TRUE), data=anscombe)
plot(y2 ~ x2, data=anscombe, pch=16, col="orange", cex=1.2)
beta = coef(modelP2)
yhat = function(x) beta[1] + beta[2] * x + beta[3] * x^2
curve(yhat, 4, 14, add=T)
```