Lecture 8

Correlation and linear regression

Association between variables

- Variables may are associated at different levels
  - malaria patients host *Plasmodium* protozoan (always)
  - height is associated with weight (frequently)
  - drinking fizzy drinks is associated with throat cancer (rarely)
  - ethnicity is not associated with IQ (ever)

- Correlation analysis defines patterns of association between variables
  - 1. determine whether variables are associated
  - 2. establish whether correlation is positive or negative
  - 3. quantify levels of association

Pearson correlatio: r

Number between -1 to 1: [-1,1]

- Pearson (or linear) correlation is a measure of linear dependence between two variables
  - =the degree to which change in variable 1 is associated with change in variable 2

- How do we measure association?
  - take a sample (!Kung males) and two variables:
    - x = male weight
    - y = male height
  - calculate average weight $\underline{X}$ and height $\underline{Y}$
  - For each case i in the sample, calculate
    - difference between its weight and average weight
      - = $(x_i - \underline{X})$
    - difference between its height and average height
      - = $(y_i - \underline{Y})$

Covariance

- The product of the two quantities

$$(x_i - \underline{X})*(y_i - \underline{Y})$$

Variables could vary in three direction: pos, neg, and irrelevant at individual level

The result could be positive and negative or null

f.i. 5*(-5) = -25

gives an idea of how height and weight co-vary in one individual

- The average of all those products in a sample is the *covariance* of the two traits

$$cov_{x,y} = \sum \frac{(x_i - X)(y_i - Y)}{n}$$

and measures how two traits vary together in the sample

Exercise:

Manually calculate the covariance of lifespan and schooling years in this hypothetical sample of three countries

country	lifespan	schooling
country 1	81	12.6
country 2	82	12
country 3	78	12.3

Pearson correlation

- But covariance is affected by scale and measurement units of variables

- If we divide covariance by the standard deviations of the two variables, we obtain the Pearson correlation *r*

$$r = \frac{cov_{x,y}}{\sigma_x \sigma_y}$$

- i.e., correlation is the *standardised* covariance of *x* and *y*
  - for this reason, it varies between -1 and 1
    - *r*=1 means absolute association
    - *r*=-1 means absolute (but inverse) association
    - *r*=0 means no association

country	lifespan	schooling
country 1	81	12.6
country 2	82	12
country 3	78	12.3

Exercise:

Now manually calculate the Pearson correlation of lifespan and schooling years in sample of three countries

Significance test of correlation

- But correlation may or may not be significant
  - like a difference between means
    - sample may be too small etc.
    - (in the case of the three countries, P = 0.85)

- We want to test the null hypothesis that variables are not correlated
  - null hypothesis: r=0

- Parametric test: we assume that x and y are normally distributed to define a t-test

$$t = \frac{r - 0}{sem} = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Correlation test calculates probability that t is significantly different from 0
  - since this is a t-test, just look for t > 1.96 or t<-1.96 for a significant difference

Significance test of correlation

- Example: is newborn head circumference and newborn weight (Swedish Birth Record) significantly correlated?
  - null hypothesis = no correlation (*r*=0)
  - File *SBR*

```
> cor.test(SBR$size, SBR$head)
    Pearson's product-moment correlation
data:  SBR$size and SBR$head
t = 319.6791, df = 186873, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5916467 0.5975088
sample estimates:
      cor
0.5945857
```

- Interpreting output:
  - *t*= 319.7 (anything over 1.96 is good)
  - *P* ~ 0 => correlation is not zero; variables ARE correlated
  - Association (*r*=0.59) is relatively strong

  (how do we interpret *r*=0.59? Which are the three questions to ask?)

Spearman's correlation ρ (rho)

- This is a nonparametric (rank) test alternative to Pearson's correlation
- To be used when
  - distribution of variables is not normal
  - sample size is small
- Procedure:
  - ranks the two variables
  - replaces values with ranks
  - then calculates Pearson correlation between the two rank distributions

Spearman's correlation ρ (rho)

- Running Spearman's correlation:
  > cor.test(variable 1, variable 2, method="spearman")
- File *Brains2*: brain structures, ape species (n=18)
  - what is the correlation between prefrontal white matter and prefrontal grey matter?

```
> cor.test(Brains2$PreWhite, Brains2$PreGray, method="spearman")
        Spearman's rank correlation rho
data:  Brains2$PreWhite and Brains2$PreGray
S = 200, p-value = 0.0001219
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7936017
```

- Conclusion: strong association between the two variables

Exercise:

Check on histogram first >>> distribution
Then apply spearman

Calculate the corrections between

- Lifespan and schooling
- Lifespan and income
- Income and schooling

cor.test(HDR2011$lifespan,HDR2011$schooling, method="spearman")

using the full HDR2011 dataset

Linear equation and linear regression

- The linear equation

  $y = a + bx$

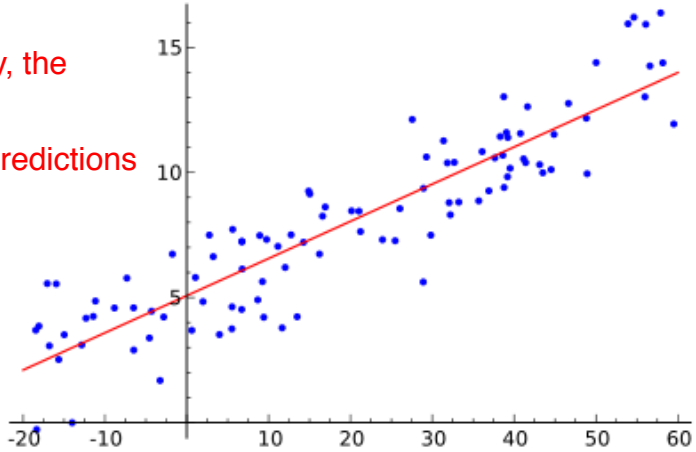
  relates variables *y* and *x* on the Cartesian plane

- Simple linear regression is the use of the linear equation to *'model' (=predict) dependent variable y from independent variable x*

$$y = a + bx + \varepsilon$$

blue points are reality, the observed values

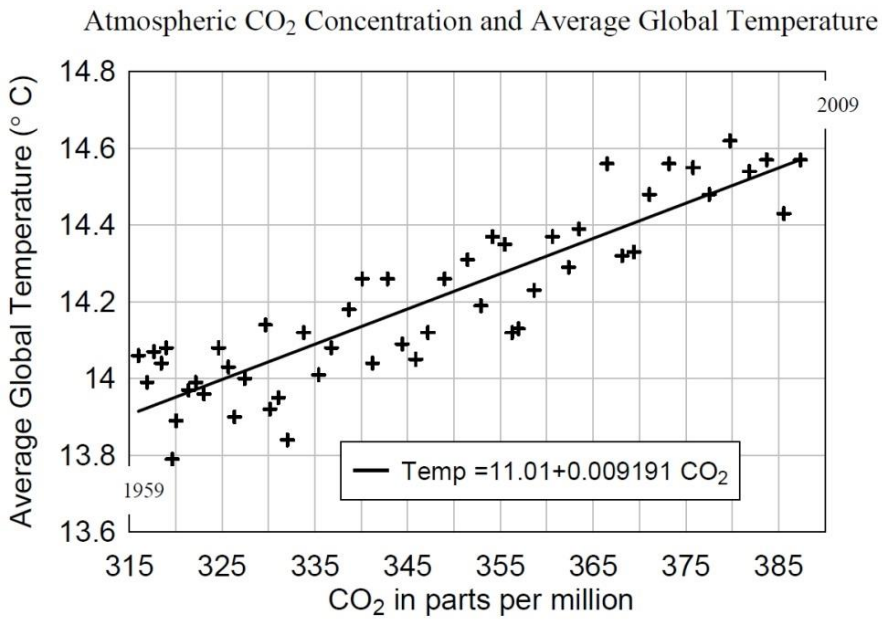
Red line shows the predictions



- *a* = intercept
  - where line crosses *y* axis
- *b* = slope or regression coefficient
  - change in *y* per unit change in *x*
- ε = residual error
  - difference between observed *y* and predicted *y*

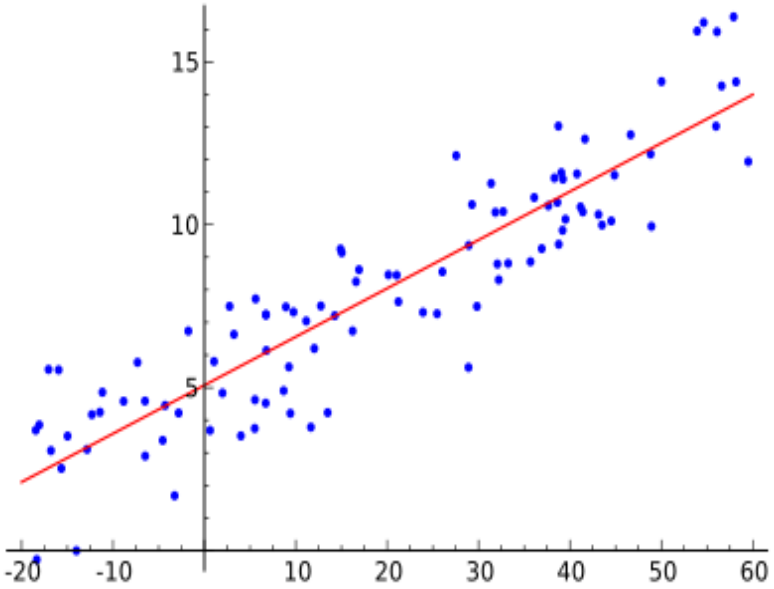
Examples

- Linear regression is a very important and popular model
  - global warming
  - trends in human longevity
  - etc.



Estimation of linear regression

- Method of least squares estimates the 'best line' across sample of (x, y) points as the line that minimises sum of squared differences (residuals) between observed y and predicted y:
  - SSres = Σ(observed y_i – predicted y_i)^2 minimize
  - = $\Sigma(y_i - (a + bx_i))^2$
    (since predicted y = $a + bx$)
  - best line always includes point (X, Y), i.e. averag



- Minimising SSres results in:

  $b = cov_{x,y} / var_x$

  $a = Y - bX$

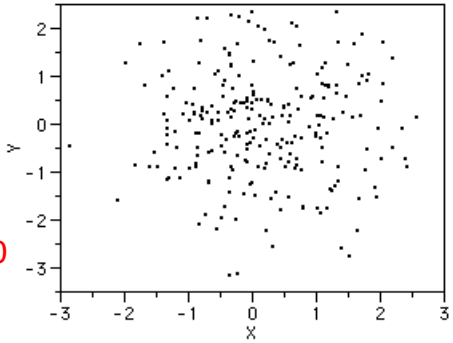
- Properties of the solution:
  - method rotates line around (X, Y) to find combination of intercept and slope that reduces the sum of residuals
  - average residual = 0

Regression test is a slope test, y is dependent on x, therefore sample size is less relevant

Significance of regression: slope test

- As in the case of means, proportions and correlations, significance of regression must be tested

- Key test is whether slope *b* is significantly different from 0
  - If *b* = 0, there is no linear relationship between variables! (i.e. there is no regression; 'best line' is horizontal)

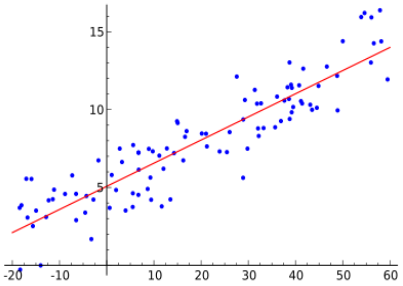


there is no correlation between x and y, thus null hypo is b=0 Meaning the slope is zero. This test is for how b varies from zero — significantly different from zero

- We use a *t*-test

$$t = \frac{b-0}{sem(b)} = \frac{b}{sem(b)}$$

- we test whether $b/sem(b)$, with df = n-2, is within a 95% CI around the tested slope b=0 (=null hypothesis)



- Intercept is also tested, but result is less important
  - If *a*=0, you still get a regression (that does not cross origin; i.e. curve has a 'height')
- **Therefore, the regression test is the slope test!**

Running linear regression in *R*

To run regression: lm: linear model

> lm(y ~ x, data=filename)

To obtain additional information, always run command *summary*

- directly on *lm* command

  > summary(lm(y ~ x))

- or on named object (i.e. the analysis you did)

  > lm(y ~ x) -> model

  > summary(model)

- In our example:

  >brainreg <- lm(Brains$BrWhite ~ Brains$BrGray)

  >summary(brainreg)

Regression statistics: residuals

```
> summary(brainreg)

Call:
lm(formula = Brains$BrWhite ~ Brains$BrGray)

Residuals:
    Min      1Q  Median      3Q     Max
-25.367  -6.760   0.504   4.675  35.780

Coefficients:
                 Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  a   -1.44510     3.91407   -0.369     0.714
Brains$BrGray    1.21928     0.03901   31.258   <2e-16 ***
---                  b
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 33 degrees of freedom
Multiple R-squared: 0.9673,  Adjusted R-squared: 0.9663
F-statistic:   977 on 1 and 33 DF,  p-value: < 2.2e-16
```

Residuals:

- mean=0 (by definition)
  - median should be ~0
- minimum and maximum residuals should be very similar
  - if they are not, large residual may be an outlier
- if 1st and 3rd quartile, or min and max residuals are too different in magnitude (not sign), relationship between *x* and *y* may not be linear

Regression statistics: intercept

```
> summary(brainreg)

Call:
lm(formula = Brains$BrWhite ~ Brains$BrGray)

Residuals:
    Min      1Q  Median      3Q     Max
-25.367  -6.760   0.504   4.675  35.780

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     -1.44510   3.91407    -0.369   0.714
Brains$BrGray    1.21928   0.03901    31.258   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 33 degrees of freedom
Multiple R-squared: 0.9673,     Adjusted R-squared: 0.9663
F-statistic:  977 on 1 and 33 DF,  p-value: < 2.2e-16
```

Intercept test:

- Null hypothesis: a=0
- t = -0.37
- P=0.714

Conclusion:

- a not different from 0
  - (as expected in this case)
- As discussed, high P value in intercept test does not mean regression is not significant
  - it simply means intercept is not distinct from 0

Regression statistics: coefficient

```
> summary(brainreg)

Call:
lm(formula = Brains$BrWhite ~ Brains$BrGray)

Residuals:
    Min      1Q  Median      3Q     Max
-25.367  -6.760   0.504   4.675  35.780

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     -1.44510     3.91407   -0.369     0.714
Brains$BrGray    1.21928     0.03901   31.258   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 33 degrees of freedom
Multiple R-squared: 0.9673,     Adjusted R-squared: 0.9663
F-statistic:   977 on 1 and 33 DF,  p-value: < 2.2e-16
```

Slope test:

- null hypothesis: b=o
- t-statistic=31.3
- P ~0

Conclusion:

- slope b is significantly different from 0 (=b>0)
- there is a positive effect of grey mater volume on white matter

Interpretation

- An extra gram of grey matter in primate brains predicts an extra 1.219 g of white matter

IMPORTANT

- Slope test is the regression test!
  - regression of white matter on grey matter IS significant
  - =we have a regression model

Confidence intervals

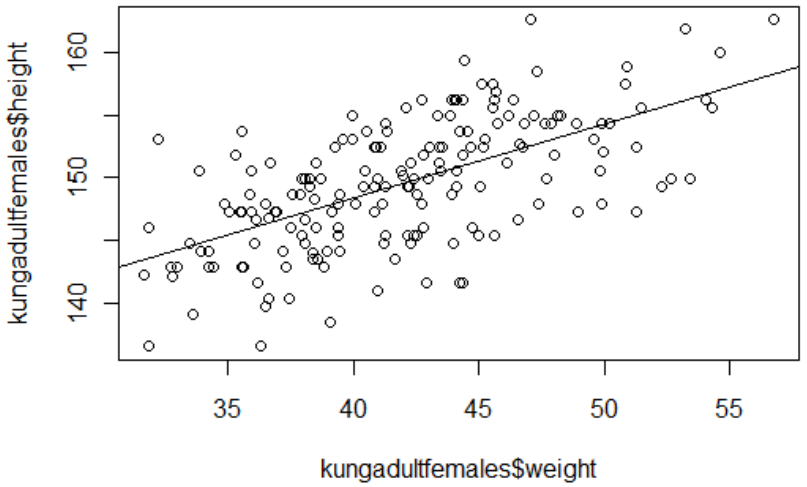
- Function *confint* calculates 95% confidence intervals of *a* and *b* estimates
  - Significant *b* => 95% CI excludes *b*=0

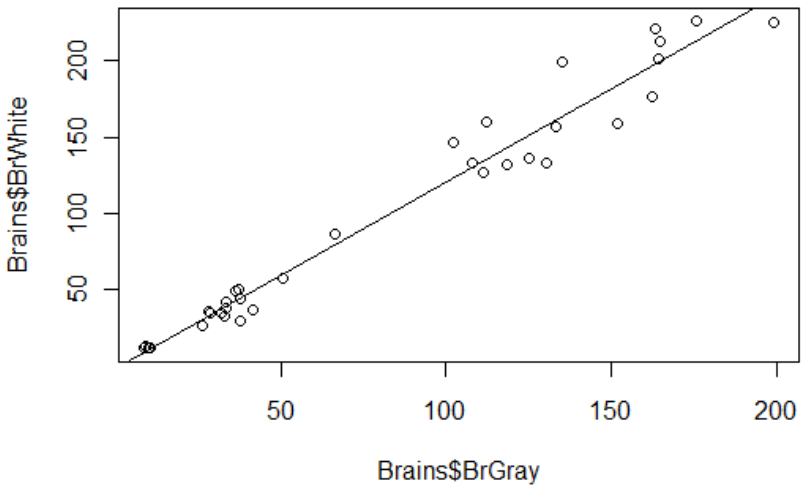
```
> confint(brainreg)
                    2.5 %       97.5 %
(Intercept)     -9.408335     6.518131
Brains$BrGray    1.139922     1.298644
```

- b is significantly different from 0
  - regression is significant

Goodness of fit

- Two regression lines may be significant, but they may differ in the extent to which they 'explain' observed data
- This reflects how linear the relationship between the variables is, or the level of dispersal around the regression line
- Main measures of 'goodness of fit' is based on a generalisation of analysis of variance (ANOVA)
  - (Multiple) R^2

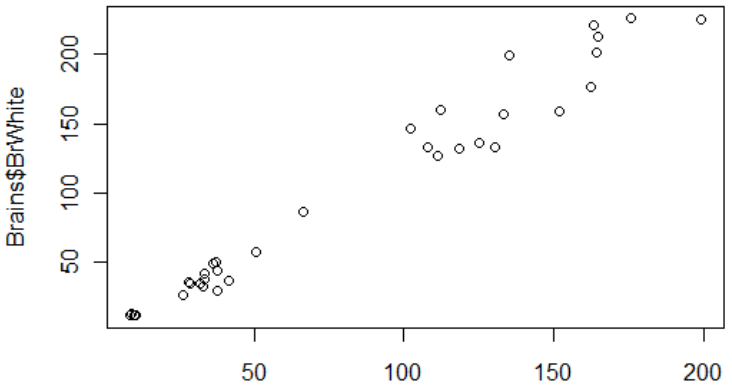




Visualising regression

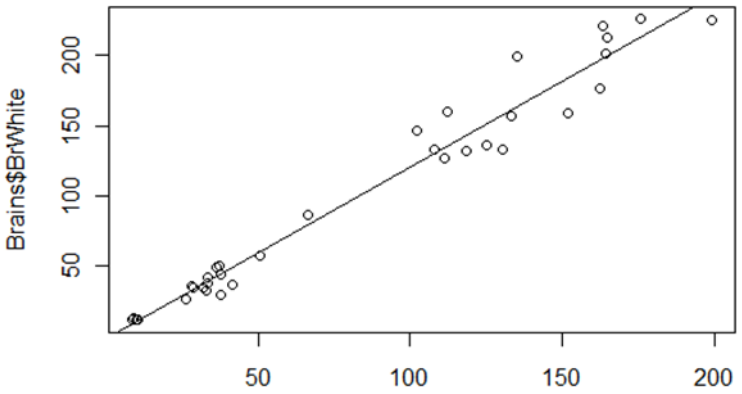
- To check whether linear regression model is appropriate, you may first want to look at the values of *x* and *y* on the plane
  - >plot(y~x)

```
>plot(Brains$BrWhite ~ Brains$BrGray)
```



- Now superimpose linear model:
  - save your model as an object; let's call it *brainreg*
  - plot regression line with command *abline* (=line defined by parameters *a*=intercept and *b*=slope)
  - Or just apply abline to the lm command

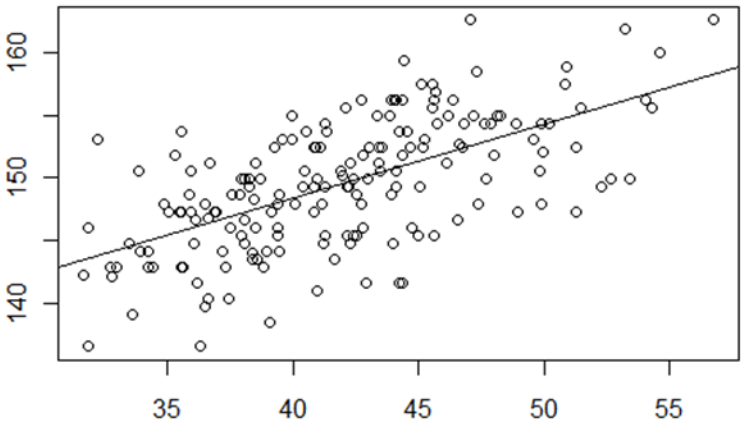
```
>brainreg <- lm(Brains$BrWhite ~ Brains$BrGray)
>abline(brainreg)
```



Generalised ANOVA

- ANOVA can also be used to calculate goodness of fit, i.e. the coefficient of determination:

$$\text{COD} = \frac{sum\ of\ squares\ explained\ by\ model}{total\ sum\ of\ squares}$$



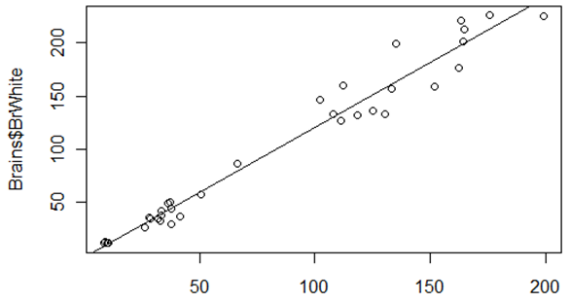
- COD is the % of variance in dependent variable y explained by model (=by the independent variable x);
- COD is estimated after partition of total variance into:
  - *sum of squares explained by model*: differences between predicted *y* and *Y (general Y mean)*
  - *residual sum of squares*: differences between observed *y* and *predicted y*

ANOVA table

- Let's calculate COD using the ANOVA table (which calculates how much of total data variance is explained by the model)

```
> anova(lm(Brains$BrWhite ~ Brains$BrGray))
Analysis of Variance Table
Response: Brains$BrWhite
                Df  Sum Sq Mean Sq F value    Pr(>F)
Brains$BrGray    1  184018  184018  977.05 < 2.2e-16 ***
Residuals       33    6215     188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Sum of squares explained by model (by *x*=*BrGray*):
  - 184018
- Total sum of squares (model + residuals):
  - 184018+6215=190233
- Goodness of fit = COD:
  - 184018/190233=0.9673



- Linear regression explains 96.7% of the data variance;
  - almost all variation in *y* is explained (predicted) by *x* (good linear model!)

Goodness-of-fit

```
> summary(brainreg)

Call:
lm(formula = Brains$BrWhite ~ Brains$BrGray)

Residuals:
    Min      1Q  Median      3Q     Max
-25.367  -6.760   0.504   4.675  35.780

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     -1.44510     3.91407   -0.369     0.714
Brains$BrGray    1.21928     0.03901   31.258    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.72 on 33 degrees of freedom
Multiple R-squared: 0.9673, Adjusted R-squared: 0.9663
F-statistic:   977 on 1 and 33 DF,  p-value: < 2.2e-16
```

- Back to our summary table:
- COD was already there!
- Multiple *R* squared= R^2=0.9673
- In linear regression analysis, COD is called R^2 or *R*-squared
- (Why?)

R^2 and r^2

- In simple linear regression, the coefficient of determination is the square of the Pearson correlation coefficient between the two variables

$$R^2 = r^2$$

- Calculating Pearson correlation *r* between x and y:

Cor.test( )

```
> cor(Brains$BrWhite, Brains$BrGray)
[1] 0.9835284
```

Cos of highly related linear relation, the predicted and the observed should be really close to each other.

[0,1]

And its square:

```
> (cor(Brains$BrWhite, Brains$BrGray))^2
[1] 0.9673282
```

- Squared Pearson coefficient = r2 =COD = R2

r is the standardised regression slope

- if *x* and *y* are expressed in standard deviation units (z-scores), regression slope is the Pearson coefficient *r*
  - if correlation is perfect (*r*=1), *z*-scores of *x* and *y* are the same for all cases
  - if there is no correlation, result is *r*=o (a horizontal line)

Summary

To create a linear model:

- Plot variables and visually inspect data
- Test significance of regression slope; this determines whether model is valid
- If slope is significant, write down model $y = a + bx$; interpret meaning of intercept and slope
- Report confidence intervals and goodness-of-fit

```
confint(brainreg)
```

- Exercises

Predicting !Kung adult male weight from height (file '*Kungadultmales*')

- application: you may have a sample of skeletons, and want to predict what their body weight was when individuals were alive

- What is the dependent variable?

plot

- Plot variables *y* against *x*
  - does the relationship look linear?

- Run analysis
  - is the regression significant?
  - how much of variance in data is explained by the model?
  - what is the correlation between weight and height?
  - what is the model?

Yes

Squared R = 0.49, meaning 49% data could be interpreted by this line

positive

- Add regression line to points

- Based on your model, what is the predicted weight of a !Kung man whose height is 165 cm?

Weight = -49.8 + 0.61 * height