Lecture 8

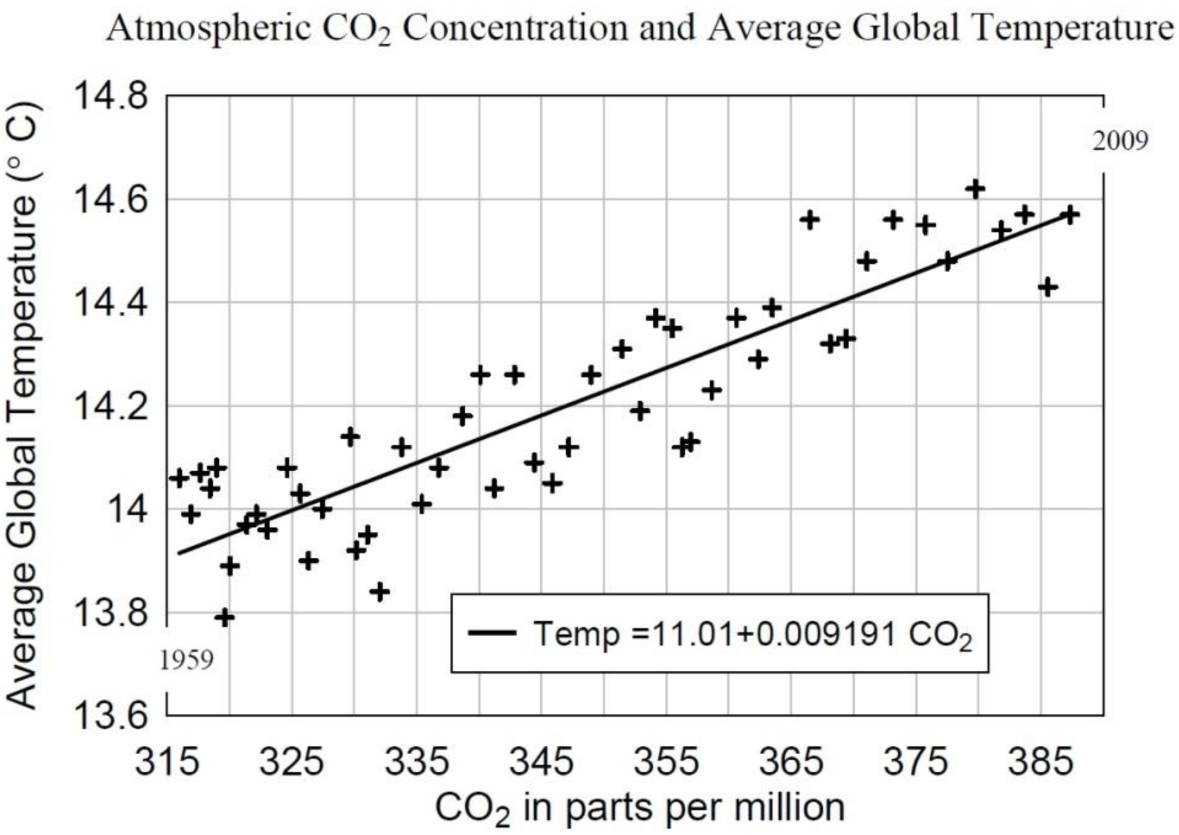
Correlation and linear regression

相关与线性回归

线性回归

Linear regressions

- It is the simplest case of statistical modelling
- Linear regression is a very important and popular model
  - global warming
  - trends in human longevity
  - etc.
- Despite all advances in statistical and computational methods, linear regression remains the option number 1 when trying to model or predict a variable as a function of another



试图将一个变量建模或预测为另一个变量的函数时，线性回归仍然是首选

变量之间的关联

Association between variables

变量可以在不同的级别相关联

- Variables may are associated at different levels
  - Covid-19 patients have been infected with coronavirus (always)
  - height is associated with weight (frequently)
  - fizzy drinks are associated with throat cancer (rarely)
  - ethnicity is not associated with IQ (ever)

相关分析 量化了变量之间的关联 1.确定变量是否关联 2.确定相关性是正还是负 3.量化关联水平

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

皮尔逊相关

正态分布

Pearson correlation

是两个变量之间线性相关性的度量=两个变量的变化相关联的程度

- 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 and two variables:
    - x = male height
    - y = male weight

  - calculate average weight $\underline{X}$ and height $\underline{Y}$

  - For each case i in the sample, calculate
    - difference between its height and average height 高度与平均高度之差
      - $= (x_i - \underline{X})$
    - difference between its height and average weight 身高与平均体重之差
      - $= (y_i - \underline{Y})$

协方差

Covariance

- The product of the two quantities 两个量的乘积

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

gives an idea of how height and weight covary in one individual

样本中所有这些产品的平均值是两个性状的协方差

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

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

Exercise:

Manually calculate the covariance of hight and weight in the sample of three cases

Individual	height	weight
1	173	61
2	182	76
3	165	58

Pearson correlation

协方差收到变量的尺度和计量单位的影响

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

将协方差除以两个变量的标准差，得到皮尔逊相关系数r

- 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表示正向绝对关联(实验中不可能出现的，如果出现就是数据错了)
    - *r*=-1 means absolute (but inverse) association r=-1表示负向绝对相关
    - *r*=0 means no association r=0表示无关联

Exercise:

Now manually calculate the Pearson correlation between height and weight

Individual	height	weight
1	173	61
2	182	76
3	165	58

相关性的显著性检验

Significance test of correlation

- But correlation may or may not be significant 相关性可能显著可能不显著
  - as in the case of differences between means,
    - sample may be too small etc.

- We want to test whether the two variables are significantly correlated
  - null hypothesis: *r*=0 (no correlation) 零假设：r=0 无相关性

假设xy是正态分布的，定义一个t检验，检验值是r=0

- Parametric correlation test: we assume that *x* and *y* are normally distributed and define a *t*-test where test value is a correlation of *r* = 0

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

- Correlation test calculates probability that t is significantly different from 0
  - since this is a t-test, 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: which are the three questions to ask?)
    - *t*= 319.7
    - *P* ~ 0 => correlation is not zero; variables ARE correlated p值接近零，拒绝零假设，两者是相关的
    - correlation is positive: head size increases with weight 相关性是正的
    - association (*r*=0.59) is relatively strong 相关性r=0.59，相对较强

斯皮尔曼相关

Spearman’s correlation ρ (rho)

代替皮尔逊相关的非参数检验办法

- = a nonparametric (rank) test alternative to Pearson’s correlation

- To be used when

当样本量过小或样本变量不符合正态分布

  - sample size is small
  - distribution of variables is not normal

- Procedure:
  - ranks the two variables 对两个变量进行排序
  - replaces values with ranks 将值替换成等级，如果x和y的等级是相关的
  - calculates Pearson correlation between the two rank distributions

那两个变量之间就是相关的(X是一级，Y也是一级)

Spearman's correlation ρ (rho)

- Running Spearman's correlation:
  > cor.test(variable 1, variable 2, method="spearman")

- File *Brains2*: brain structures
  - Small sample of 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
```

两个变量之间存在显著的、正的、强的关联

- Conclusion: significant, positive, strong association between the two variables

Exercise:

Calculate the corrections between

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

using the full *HDR2011* dataset

Which test do you use? Pearson or Spearman correlation?

```
hist(HDR2011$lifespan)
hist(as.numeric(HDR2011$schooling))
hist(as.numeric(HDR2011$income))
```

线性方程与线性回归

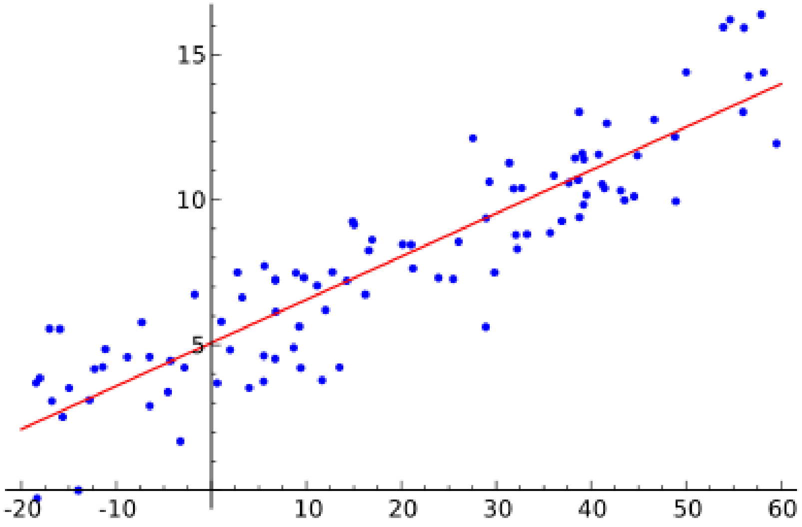
Linear equation and linear regression

- The linear equation

$$y = a + bx$$

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

- Simple linear regression uses the linear equation to *model (= predict) the dependent variable y from the independent variable x*



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

- *a* = intercept 截距
  - where line crosses *y* axis
- *b* = slope or regression coefficient 斜率或回归系数
  - change in *y* per unit change in *x*
- *ε* = residual error 剩余误差，观察Y和预测Y之间的差异
  - difference between observed *y* and predicted *y*

线性回归的估计

Estimation of linear regression

- Method of *least squares* estimates the 'best line' across sample of (x, y) points
- Best line is the one that minimises sum of squared differences (residuals) between observed y and predicted y:
  - SSres = Σ(observed y_i – predicted y_i)^2
    = $\Sigma(y_i - (a + bx_i))^2$
    (since predicted y = $a + bx$)
  - best line always includes point ($\underline{X}$, $\underline{Y}$),
    where $\underline{X}$ = mean x, $\underline{Y}$ = mean y

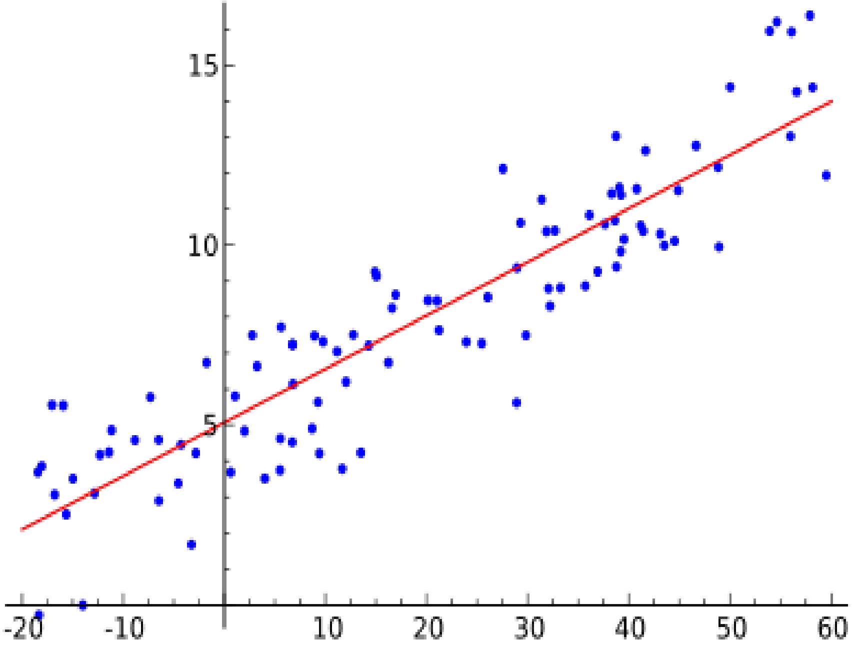
最佳线过(x,y)，xy是平均值

- Minimising SSres:

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

$$a = \underline{Y} - b\underline{X}$$

先算B
a = y的均值-b*x的均值



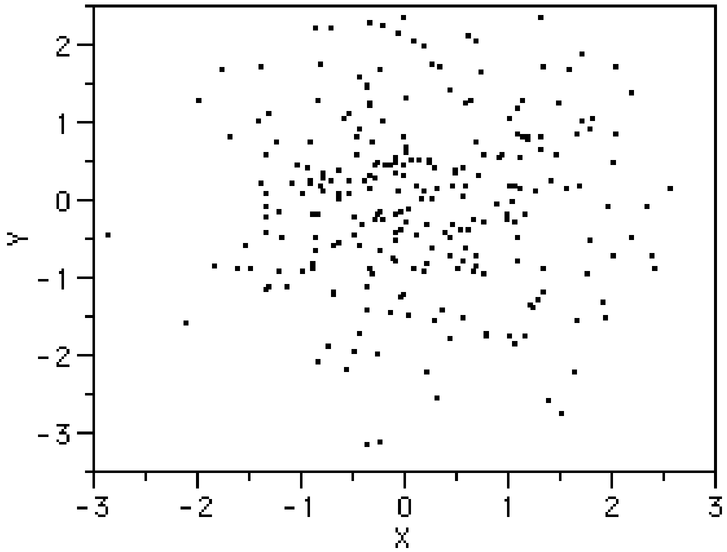
- Properties of the solution:
  - method rotates line around mean values (X, Y) to find combination of intercept and slope that reduces the sum of residuals
  - average residual = 0 平均残差=0

斜率检验

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; slope not different from 0; 'best line' is horizontal)
  - A horizontal line (*b*=0) means that x has no effect on y
    - you cannot predict y from x

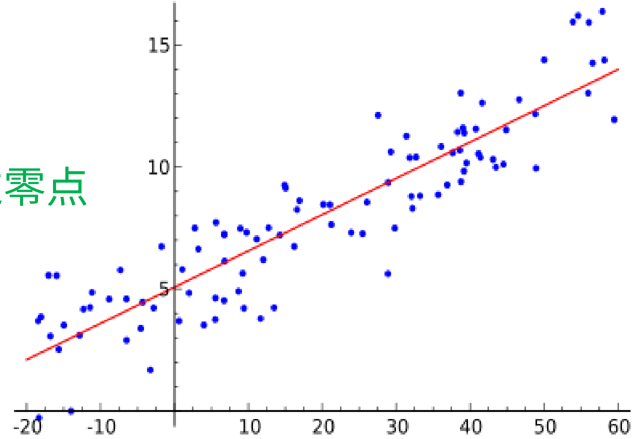
如果b=0，则变量之间不存在线性关系
表示x对y没有影响，不能从x预测y



- We use a *t*-test for slope b

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

- We test whether *b*/*sem*(*b*) is within a 95% CI around the tested slope *b*=0
  - null hypothesis: *b*=0 (=no regression) 零假设：b=0(无回归)
- Intercept is also tested, but result is less important α =0仍会得到回归，但不会过零点
  - If *a*=0, you still get a regression, but it does not cross origin (0, 0)
  - **Therefore, the regression test is the slope test!**



因此回归检验就是斜率检验

Running linear regression in *R*

Is the amount of white matter in brains affected by the amount of grey matter?

To run regression: function *lm()*

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

y~x

or

```
> lm(BrWhite ~ BrGray, data = Brains)
```

or create an object

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

Always run command *summary* either on *lm* command or object

```
> summary(lm(Brains$BrWhite ~ Brains$BrGray))
```

or

用smmary函数

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

残差

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)    -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
```

Residuals: 根据定义，平均值=0，中位数应该是0

- mean=0 (by definition)
  - median should be ~0

如果第一个和第三个四分位数，或者最小和最大残差在数量上相差太大，则x和y之间的关系可能不是线性的
可能存在异常值，有很多处理残差的技术(此处未提及)

- if 1st and 3rd quartile, or min and max residuals are too different in magnitude, relationship between *x* and *y* may not be linear
  - there may be outliers; many techniques to deal with residuals
  - (not addressed here)

截距

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: 零假设：α =0

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

Conclusion:

- a not different from 0
  - = curve goes through the origin
  - (as expected in this case)

α 与零无差异=曲线过原点

系数

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=0
- t-statistic=31.3
- P ~0

斜率b明显不同于0

Conclusion:
- slope b is significantly different from 0
- b > 0: there is a positive effect of grey matter 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 significant regression

我们有一个显著的回归

Confidence intervals

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

95%的置信区间不包括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

b明显不同于0
回顾是显著的

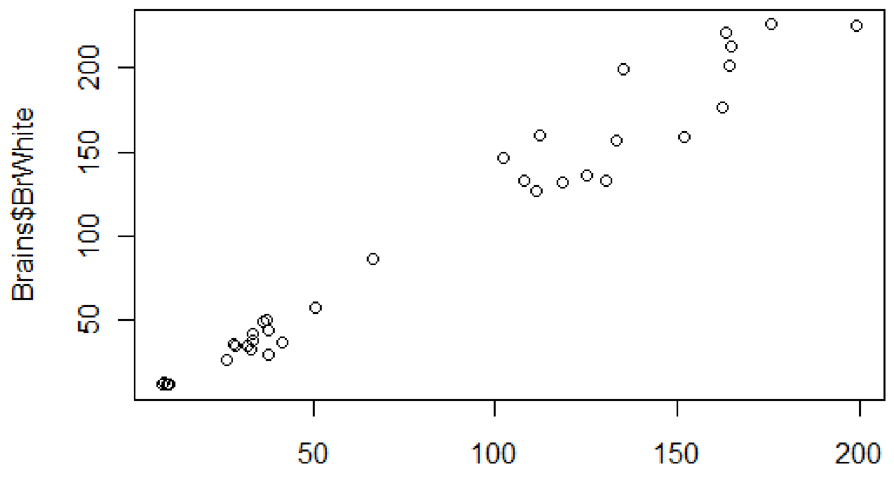
可视化回归

Visualising regression

首先绘制y和x的关系图

- First plot *y* against *x*

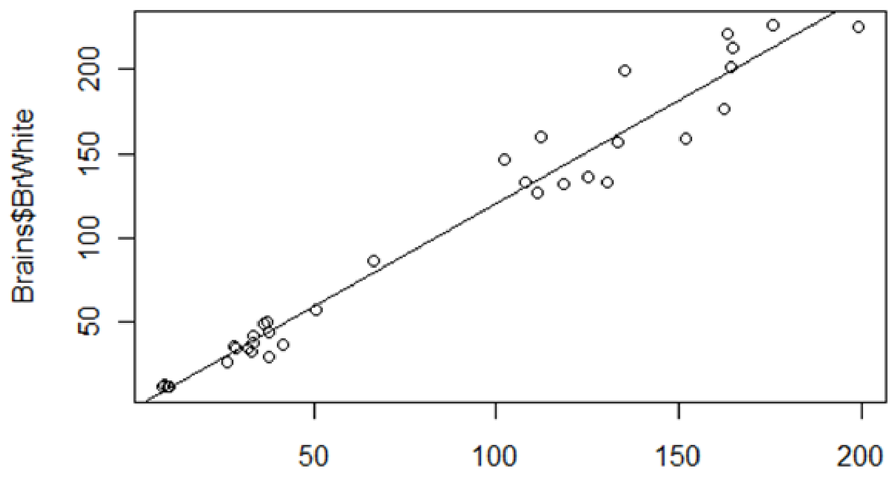
```
>plot(Brains$BrWhite ~ Brains$BrGray)
```



- Now plot line from the linear model
  - save your model as an object (in this case, *brainreg)*
  - plot regression line with command *abline* (=line defined by parameters *a*=intercept and *b*=slope)

用abline绘制回归线

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



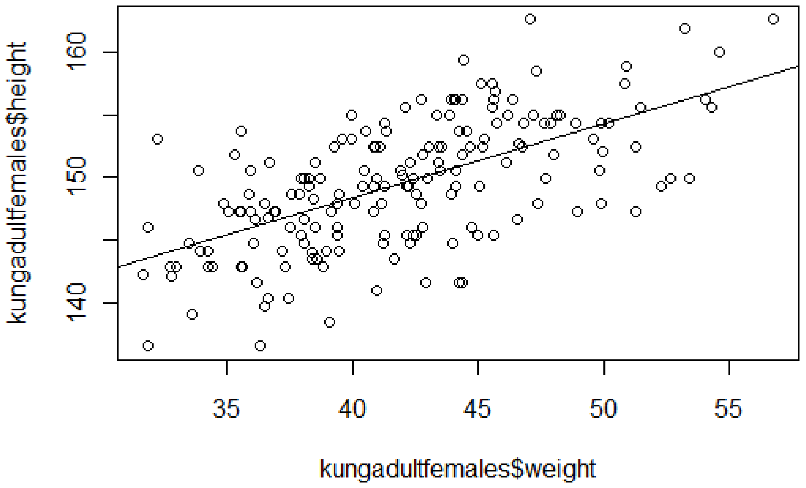
拟合度

Goodness of fit

两条回归线可能是显著的，但他们与观测数据的匹配程度可能不同=模型的拟合度

- Two regression lines may be significant, but they may differ in how closely they match observed data

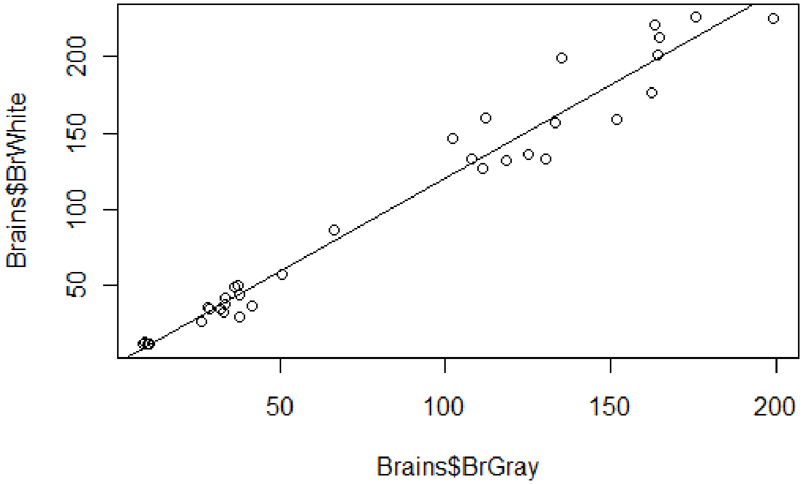
  = the ‘goodness of fit’ of the model



这反映了变量之间的线性关系=在回归线附近扩散

- This reflects how linear the relationship between the variables is
  - = dispersal around the regression line

拟合优度的主要测量是基于方差分析=R方



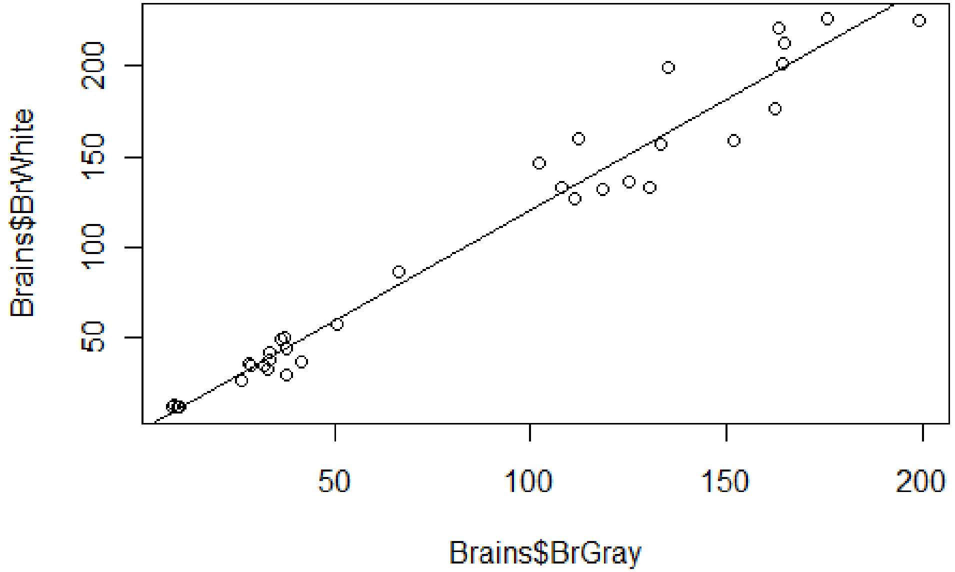
- Main measure of ‘goodness of fit’ is based on a generalisation of analysis of variance (ANOVA)

  = (Multiple) R^2

广义方差分析

Generalised ANOVA

ANOVA可以用于计算模型的决定系数



- ANOVA can be used to calculate a coefficient of determination (COD) of a model

- COD is the fraction of variance of Y explained by model (=by the independent variable X)

- COD is estimated after partition of total variance into:
  - *S = sum of squares explained by model* 模型解释的平方和

    = squared differences between predicted *y* and $\underline{Y}$ *(general Y mean)*
  - *R = residual sum of squares* 残差平方和=观测y与预测y之间的平方差

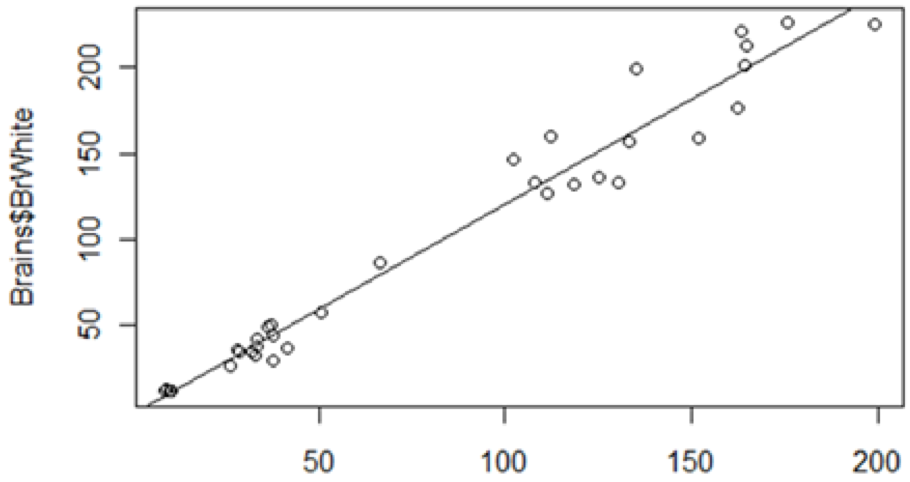
    = *squared* differences between observed *y* and *predicted y*

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

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:
- In linear regression analysis, COD is called Multiple R^2

COD被称为multiple R方

- Multiple R squared = R^2 = 0.9673
- Very high value! Almost all variance in y can be predicted by x through the regression

很高的值，y中几乎所有的方差都可以通过回归用x预测

R^2 and r^2

在线性回归中，决定系数是两个变量之间的皮尔逊相关系数的平方

- In 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(Brains$BrWhite, Brains$BrGray)
[1] 0.9835284
```

皮尔逊系数r可能是正的也可能是负的一平方就和R一样了

And its square:

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

- Squared Pearson coefficient = r^2 = COD = same value as R^2

r是标准化回归斜率

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*=0 (horizontal line)

如果相关性是完美的（r=1），则x和y的z分数对于所有情况都是相同的
如果不存在相关性，则结果是r=0(水平线)

Summary

To create a linear regression model:

绘制变量y和x并且目测检查数据：看起来是否有线性关系？

- Plot variables y and x and visually inspect data
  - Does it seem that there is a linear relationship?

回归斜率的显著性检验，斜率显著表示线性模型有效

- Test significance of regression slope;
  - significant slope means linear model is valid

- If slope is significant, write down model $y = a + bx$; interpret meaning of intercept and slope

- Report confidence intervals of slope *b* and goodness-of-fit R^2

报告斜率b和拟合优度R方的置信区间

- Exercises

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

y是weight,x是height

- Let's say you want to predict body weights of !Kung men from their heights. What is the dependent variable *y*?

```
plot(weight ~ height, data=Kungadultmales)
reg1 <- lm(weight ~ height, data=Kungadultmales)
summary(reg1)
0.429**0.5
cor.test(Kungadultmales$weight, Kungadultmales$height)
abline(reg1)
```

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

yes

- Run a linear regression of weight on height
  - is the regression significant? 看b=0否
  - how much of variance in data is explained by the model?
  - what is the correlation between weight and height? multiple r-square开根号
  - what is the final model? Write it as a line equation $y = a + bx$

数据中有多少方差是由模型解释的？

- Add regression line to points

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