Assignment 2

James La Fontaine

16/09/2021

## Question 3

### 3a)

The difference in scale between the raw brain and body measurements is too significant to fit a suitable model and to produce a suitable plot to compare their relationship

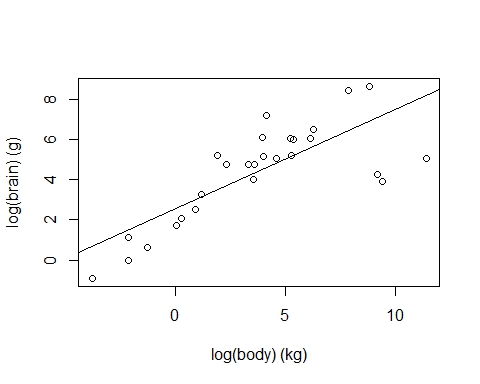
### 3b)

data(Animals, package = "MASS")  
  
brain = log(Animals$brain)  
body = log(Animals$body)  
  
model1 = lm(brain ~ body, data = log(Animals))  
  
summary(model1)

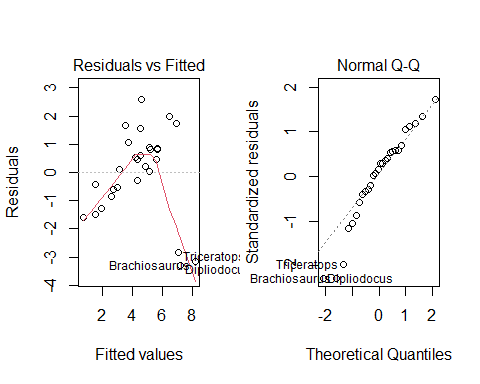
##   
## Call:  
## lm(formula = brain ~ body, data = log(Animals))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2890 -0.6763 0.3316 0.8646 2.5835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.55490 0.41314 6.184 1.53e-06 \*\*\*  
## body 0.49599 0.07817 6.345 1.02e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.532 on 26 degrees of freedom  
## Multiple R-squared: 0.6076, Adjusted R-squared: 0.5925   
## F-statistic: 40.26 on 1 and 26 DF, p-value: 1.017e-06

### 3c)

plot(body, brain, xlab = "log(body) (kg)", ylab = "log(brain) (g)")  
abline(model1)



par(mfrow = c(1, 2))  
plot(model1, 1:2)

 It seems quite plausible to assume that the residuals are normally distributed according to the QQ plot, and the model plot indicates that this linear regression model appears to represent the relationship between an animal’s body and brain moderately well. However, the residuals vs fitted values plot produces a systematic ‘U’ pattern which implies that the assumption of linearity isn’t holding, although this is clearly caused by the outliers in the data.

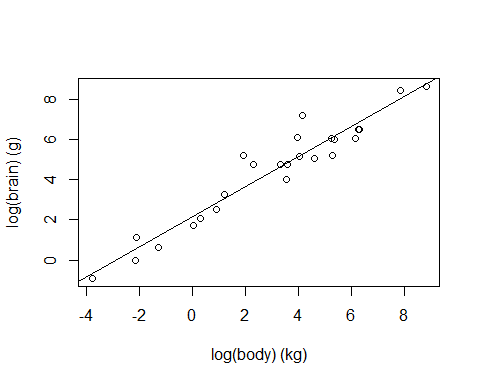
### 3d)

AnimalsNoDinosaurs = Animals[-c(6, 16, 26),]  
brain2 = log(AnimalsNoDinosaurs$brain)  
body2 = log(AnimalsNoDinosaurs$body)  
model2 = lm(brain2 ~ body2, data = log(AnimalsNoDinosaurs))  
  
summary(model2)

##   
## Call:  
## lm(formula = brain2 ~ body2, data = log(AnimalsNoDinosaurs))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9125 -0.4752 -0.1557 0.1940 1.9303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.15041 0.20060 10.72 2.03e-10 \*\*\*  
## body2 0.75226 0.04572 16.45 3.24e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7258 on 23 degrees of freedom  
## Multiple R-squared: 0.9217, Adjusted R-squared: 0.9183   
## F-statistic: 270.7 on 1 and 23 DF, p-value: 3.243e-14

### 3e)

plot(body2, brain2, xlab = "log(body) (kg)", ylab = "log(brain) (g)")  
abline(model2)



### 3f)

newdata = data.frame(body2 = 500)  
  
predict(model2, newdata, interval = "confidence")

## fit lwr upr  
## 1 378.2808 331.2781 425.2834

## Question 5

### 5a)

H0: p1 = p2

H1: p1 ≠ p2

### 5b)

x = c(120, 60)  
n = c(800, 600)  
  
p1 = prop.test(x, n)  
  
p1

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x out of n  
## X-squared = 7.2105, df = 1, p-value = 0.007248  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.01406774 0.08593226  
## sample estimates:  
## prop 1 prop 2   
## 0.15 0.10

z\_statistic = sqrt(p1$statistic)  
  
names(z\_statistic) = NULL  
  
cat('z\_statistic:', z\_statistic, '\n')

## z\_statistic: 2.68524

critical\_value = qnorm(1-0.05/2)  
  
cat('critical value:', critical\_value, '\n')

## critical value: 1.959964

The observed value of the Z-statistic is 2.6852. When α = 0.05, the rejection region for this test is |z| > 1.96. Therefore we reject H0 and can conclude that there is evidence suggesting that the rates of babies with low birth weight differ between Africa and the Americas. More specifically, there is evidence that the rate is higher in Africa than the Americas.

### 5c)

p2 = prop.test(x, n, conf.level = 0.99)  
  
p2

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x out of n  
## X-squared = 7.2105, df = 1, p-value = 0.007248  
## alternative hypothesis: two.sided  
## 99 percent confidence interval:  
## 0.003235253 0.096764747  
## sample estimates:  
## prop 1 prop 2   
## 0.15 0.10

z\_statistic = sqrt(p2$statistic)  
  
names(z\_statistic) = NULL  
  
cat('z\_statistic:', z\_statistic, '\n')

## z\_statistic: 2.68524

critical\_value = qnorm(1-0.01/2)  
  
cat('critical value:', critical\_value, '\n')

## critical value: 2.575829

If α = 0.01, then the rejection region for the test is |z| > 2.5758. Therefore we still reject H0 and conclude that there is evidence suggesting that the low birth weight rates differ between the two continents.

## 5d)

conf\_int = p1$conf.int  
  
cat('95% confidence interval for difference in rates: (',round(conf\_int[1], 4),',', round(conf\_int[2], 4),')')

## 95% confidence interval for difference in rates: ( 0.0141 , 0.0859 )

### Question 6

## 6a)

significance\_level = pgeom(4, 0.4, lower.tail = FALSE) + dgeom(4, 0.4)  
cat('The probability of committing a Type I Error is', significance\_level)

## The probability of committing a Type I Error is 0.1296

## 6b)

beta = pgeom(4, 0.2) - dgeom(4, 0.2)  
cat('The probability of committing a Type II Error is', beta)

## The probability of committing a Type II Error is 0.5904

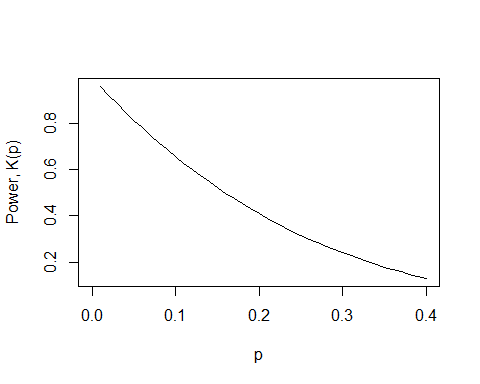
## 6c)

K1 = function(p)  
 1 - (pgeom(4, p) - dgeom(4, p))  
p = seq(0, 0.4, 0.01)  
K = K1(p)

## Warning in pgeom(4, p): NaNs produced

## Warning in dgeom(4, p): NaNs produced

plot(p, K, type = 'l', ylab = 'Power, K(p)')



## 6d)

cat('significance level for critical value = 4:', pgeom(4, 0.4, lower.tail = FALSE) + dgeom(4, 0.4), '\n')

## significance level for critical value = 4: 0.1296

cat('significance level for critical value = 5:', pgeom(5, 0.4, lower.tail = FALSE) + dgeom(5, 0.4), '\n')

## significance level for critical value = 5: 0.07776

cat('significance level for critical value = 6:', pgeom(6, 0.4, lower.tail = FALSE) + dgeom(6, 0.4), '\n')

## significance level for critical value = 6: 0.046656

cat('significance level for critical value = 7:', pgeom(7, 0.4, lower.tail = FALSE) + dgeom(7, 0.4), '\n')

## significance level for critical value = 7: 0.0279936

Therefore, a test where the null hypothesis is rejected if the observed value is greater than or equal to 6 gives an approximate significance level of 0.05. The actual significance level of this test is 0.046656.