NMAI061-22-EX5

Matej Nemec

Visualization of multi-dimensional data and Principal Component Analysis

1)Select appropriate data and try to select predictor variables (potentially with applied transformations) in order for predicted values to fit the observed data as good as possible.

We decided to go with the Animals2 dataset which only has 2 variables and we will be trying to predict brain wight based on animals body weight. Right away it makes intuitive sense not expect a completely linear relationship. But let us test this assumption a bit.

```
data=Animals2
cor.test(data$body,data$brain)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$body and data$brain
## t = 0.41062, df = 63, p-value = 0.6827
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1946886  0.2918848
## sample estimates:
##        cor
## 0.05166368
```

As we can see when not trasforming the variables at all there seems to be no correlation. Because the data is on mamals what we would intuitively expect is that there is both a point under which the brain size doesn't really decrease (or not nearly as much) with decreasing body size and at the same time we would expect brain size increase to plateau with very large body sizes. When we think about which functions could model such behavior, log, tanh and sigmoid come to mind.

```
cor.test(log(data$body),log(data$brain))
```

```
##
##  Pearson's product-moment correlation
##
## data:  log(data$body) and log(data$brain)
## t = 14.367, df = 63, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.8027970 0.9223075
## sample estimates:
##       cor
## 0.8753092
```

```
cor.test(tanh(data$body),tanh(data$brain))
```

```
##
##  Pearson's product-moment correlation
##
## data:  tanh(data$body) and tanh(data$brain)
## t = 7.3446, df = 63, p-value = 5.002e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5216904 0.7918941
## sample estimates:
##       cor
## 0.6791718
```

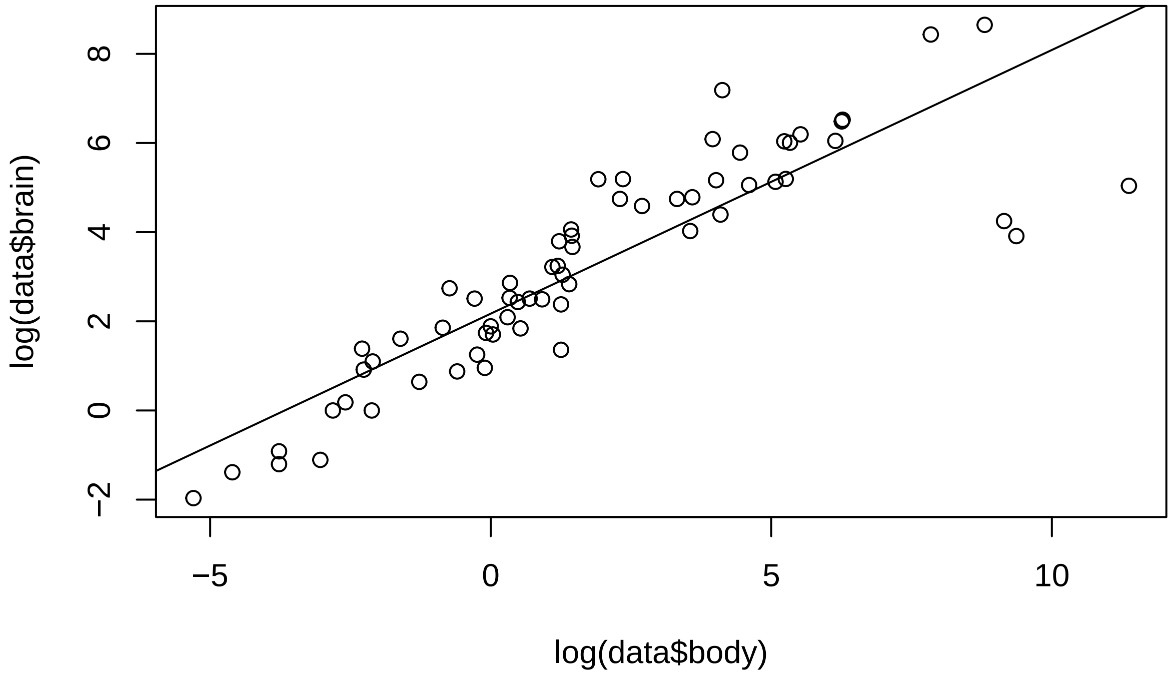
```
cor.test(sigmoid(data$body),sigmoid(data$brain))
```

```
##
##  Pearson's product-moment correlation
##
## data:  sigmoid(data$body) and sigmoid(data$brain)
## t = 7.6462, df = 63, p-value = 1.484e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5415126 0.8019757
## sample estimates:
##      cor
## 0.693781
```

We can see that best correlation is achieved with log transform.

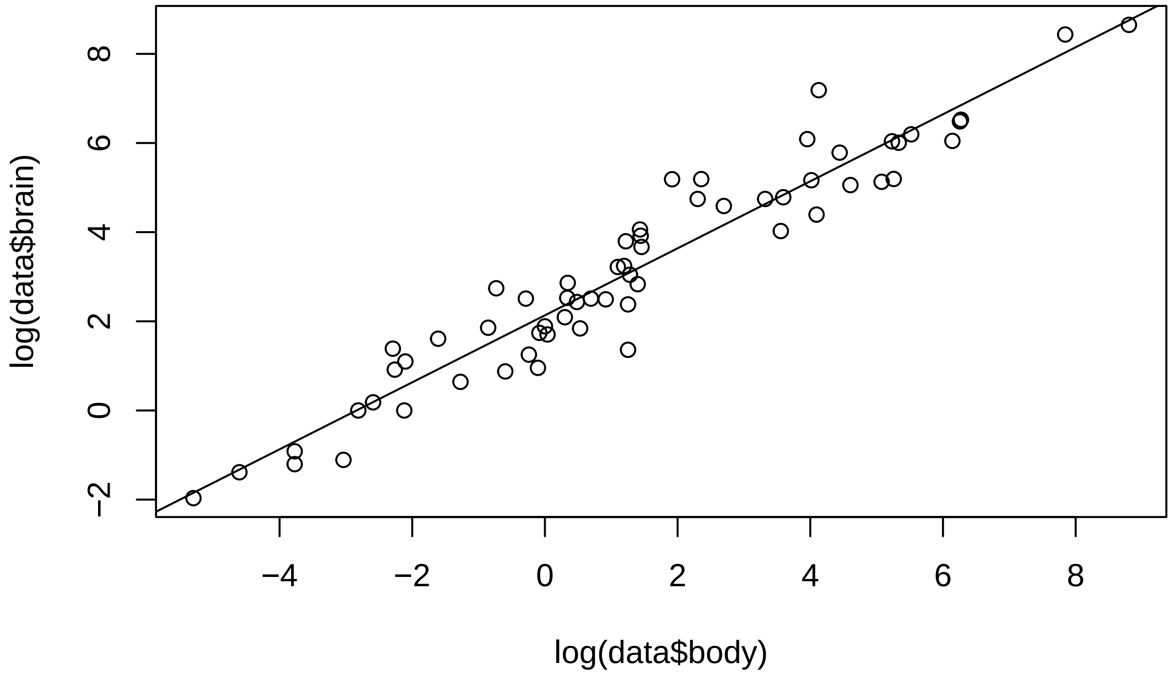
2)Visualize the data including fitted values from a selected linear model.

```
m1=lm(log(brain)~log(body),data=data)
plot(log(data$body),log(data$brain))
abline(m1)
```



We can see that there are some big outliers. The most problematic would be the three dinosaurs. We can remove them to improve our model.

```
data=data[data$body<data['Triceratops','body'],]
m2=lm(log(brain)~log(body),data=data)
plot(log(data$body),log(data$brain))
abline(m2)
```



As exoected this improves our model.

3)Comment on coefficient statistical significance. Test the null hypothesis of all coeffiecients being equal to 0.

```
summary(m1)
```

```
##
## Call:
## lm(formula = log(brain) ~ log(body), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8592 -0.5075  0.1550  0.6410  2.5724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.17169    0.16203   13.40   <2e-16 ***
## log(body)    0.59152    0.04117   14.37   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.172 on 63 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7625
## F-statistic: 206.4 on 1 and 63 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = log(brain) ~ log(body), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23   <2e-16 ***
## log(body)    0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Coefficients are statistically significant for both models, with dinosaurs and without. Rˆ2 is markedly improved when leaving out the dinosaurs.

We perform anova.

```
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: log(brain)
##           Df Sum Sq Mean Sq F value    Pr(>F)
## log(body)  1 336.19  336.19  697.42 < 2.2e-16 ***
## Residuals 60  28.92    0.48
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
coef=summary(m2)$coefficients[2,1]
err=summary(m2)$coefficients[2,2]
coef + c(-1,1)*err*qt(0.975, 42)
```

```
## [1] 0.6942441 0.8091277
```

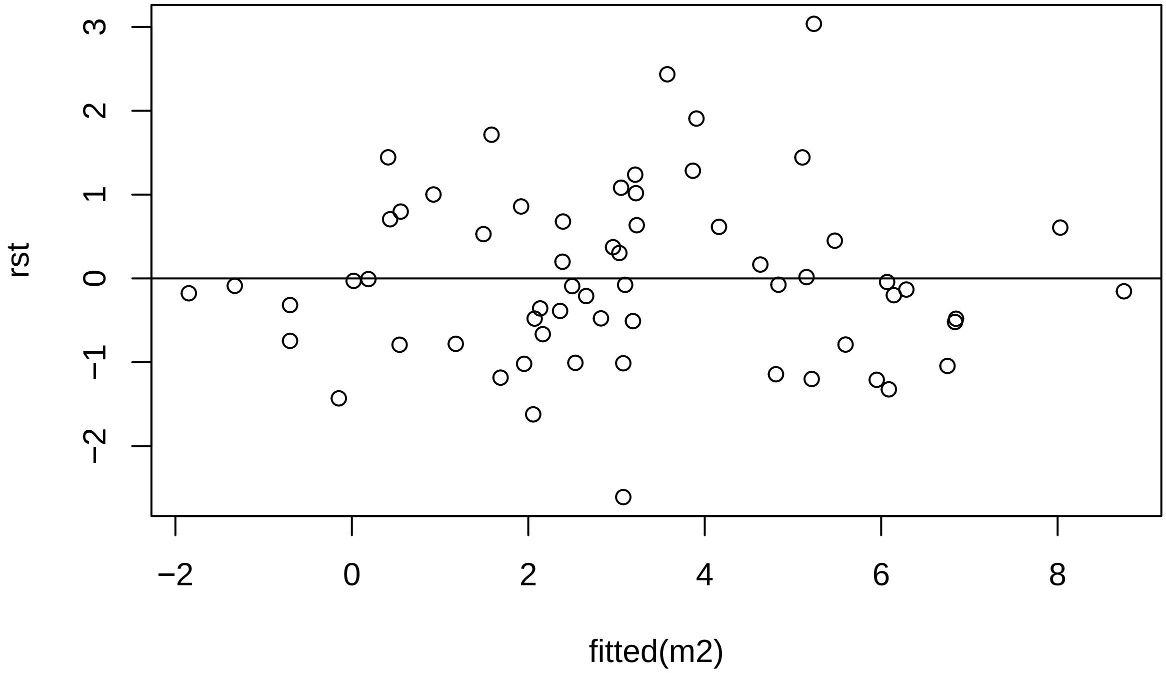
We can reject the hypothesis that coeffiecient is 0. The entire 95% CI lies very far from 0. ## 3)Calculate residuals a based on appropriate visualizations comment on fulfilling the prerequisites of the selected model (heteroskedascity, normality, regression model shape, etc.) We use the appropriate function to get residuals and test heteroskedacity using studentized Breusch-Pagan test.

```
rst=rstudent(m2)
bptest(m2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m2
## BP = 0.27849, df = 1, p-value = 0.5977
```

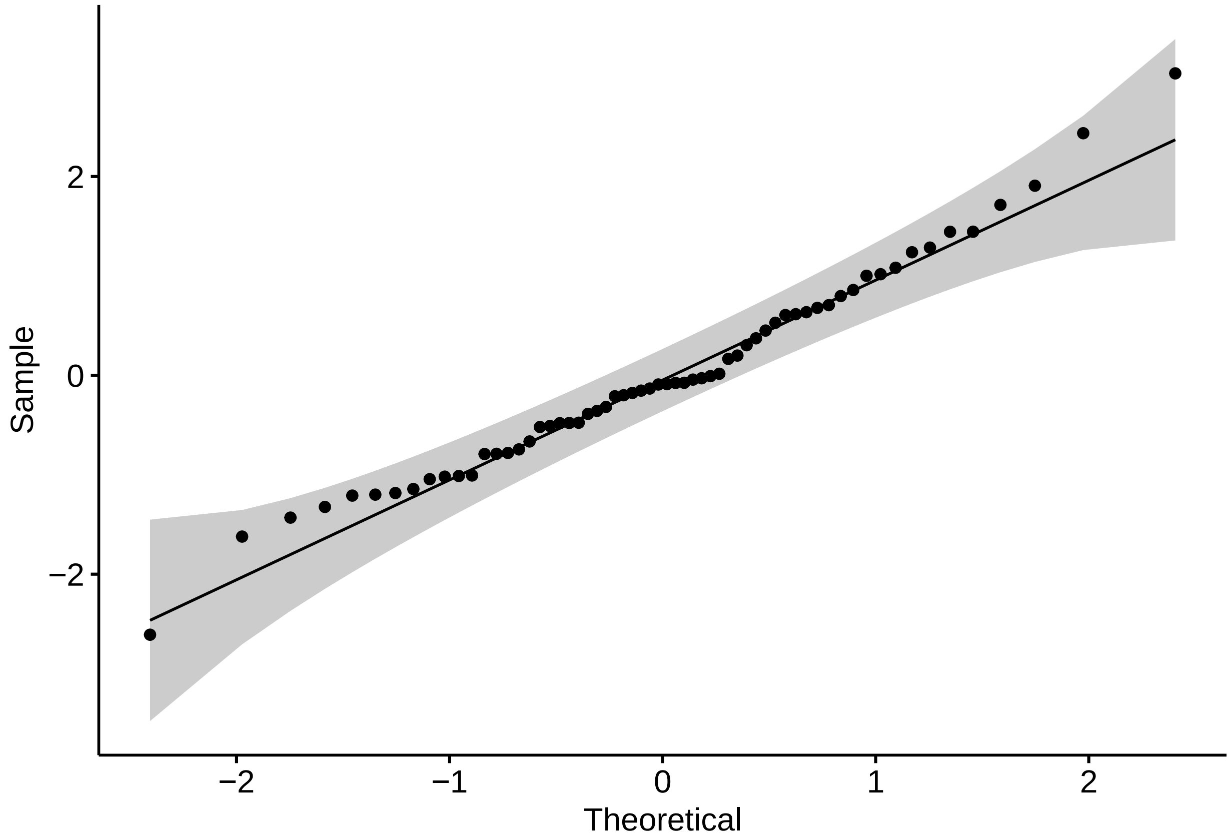
We cannot reject the null (homoskedacity) therefore hetoroskedacity is not present.

```
plot(rst~fitted(m2))
abline(0,0)
```



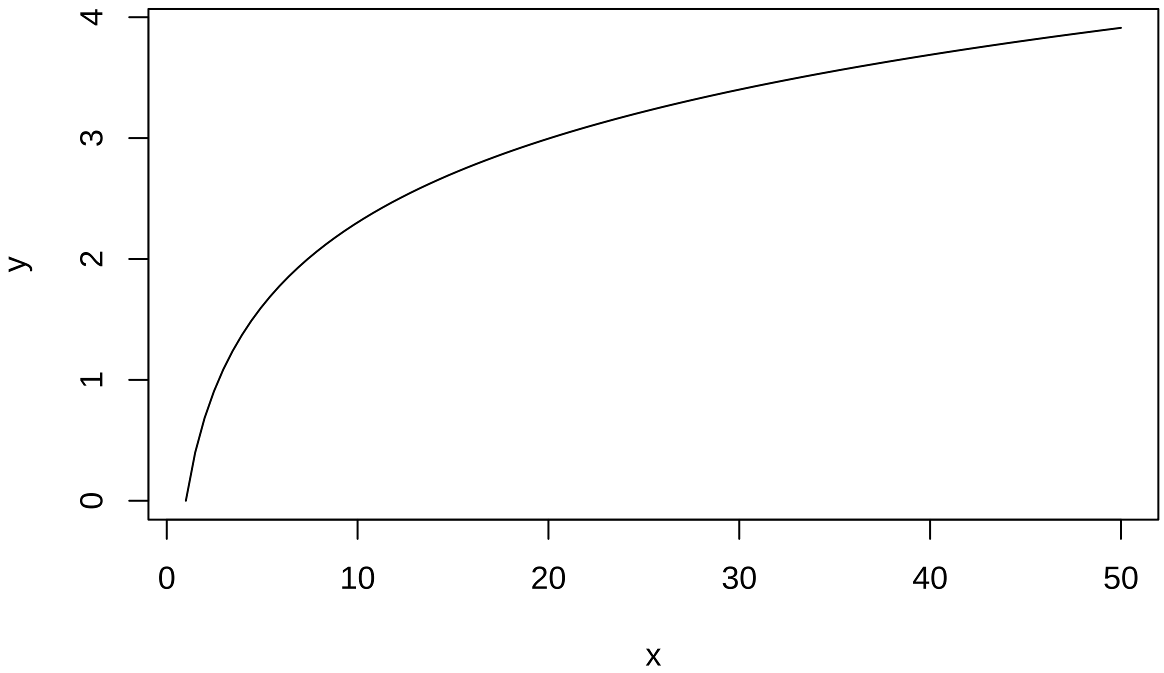
If we look at the visual representation this conclusion seems reasonable.

```
ggqqplot(rst)
```



There seem to be no major outliers in our residuals. Formula for our model is $log(Y) = log(X)$ shape of $log(X)$ is:

```
curve(log(x), from=1, to=50, , xlab="x", ylab="y")
```


y
0
1
2
3
4
x
0
10
20
30
40
50