Regression Lab

Liam Niehus-Staab

11/15/2018

## 1a) Create a regression model to predict mpg from displacement, mpg=a + b\*displacement. What is the regression equation and R^2? y = 7.622 + 4.034x R^2 = .4055

## 1b) Look at residual vs. speed and residual vs. displacement plots. Describe any patterns in the residual plots.

Speed has a clear quadratic shape to its residual/speed points, so speed is not a good indicator for the model Displacement has no clear shape, meaning it could be a good predictor for the model.

## 1c) Describe any patterns in the residual normal probability plot.

qqnorm(linMPGresiduals) the normal probability plot shows a fairly normal distribution of the data. That means that it is not breaking assumptions to try to use this data as a predictor in a model.

## 2a) Conduct a regression analysis to predict mpg from speed and displacement, mpg=a + b*speed + c*displacement What is the regression equation and R^2?

linDouble = lm(mpg ~ speed + displacement, data = MPG)

y = 11.69 - .044(speed) + 4.176(displacement) R^2 = .5499

## 2b) Look at residual vs. speed and residual vs. displacement plots. Describe any patterns you see.

The same patterns we saw before are still present; Speed has a clear quadratic shape to its residual/speed points, so speed is not a good indicator for the model Displacement has no clear shape, meaning it could be a good predictor for the model.

## 2c) What does the residual normal probability plot show?

The data appears to be not quite normally distributed; the points form an 'S' shape, with the points coming off the normal line at the top and bottom of the plot.

## 2d) Is a better model than mpg = a + b\*speed? Use the residual plots and R^2 to explain why. The new 2 variable model has a higher R^2 value, but that is to be expected as there are more variables in the 2 variable model. Looking at the residulas though, we can see that there isn't much difference between the models. Since the residuals are more important for analyzing the validity of a model, I would say that the new 2 variable model isn't really much better than the model that included just speed.

## 3) Without creating any new models or graphs, predict which of the following regression equations would create a better model for your data? Use the residual plots in question 1) to explain why. mpg = a +b*speed^2+ c*displacement or mpg = a + b\* LN(speed) + c\*displacement Since the residuals vs. speed graph showed a negative quadratic shape, it might be smarter to try using a model that squares the speed variable; would most likely be the best choice of action.

## 4) Create a model with four new variables: speed^2, LN(speed), displacement^2, and a speed\*displacement. What is the regression equation and if all 6 terms are included in the model? View (but do not print) each of the residual plots. If a clear pattern exists in any of these plots, describe the pattern.

newVars = lm(mpg ~ speed + displacement + log(speed) + I(speed\*displacement) + I(speed^2) + I(displacement^(2)), data = MPG)

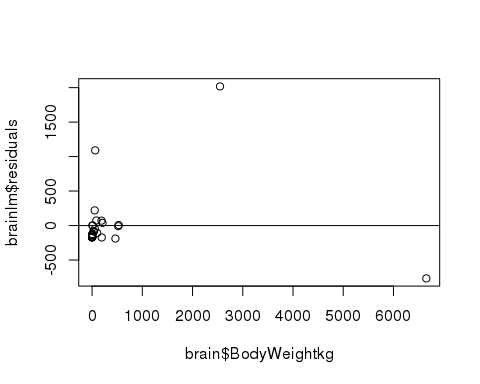
R^2 = .933 Where the speed variable was present, the same parabolic shape was visible. The transformations on the data appeared to have little effect on the overall shape.

## 5) Give the regression equation and R^2 value of the best model using some combination of all 6 terms. Note that including all 6 variables will give the highest R^2 value. But some of the terms really have very little impact on R^2. The idea is to find the simplest model (fewest variables) without significantly reducing the R^2 value.

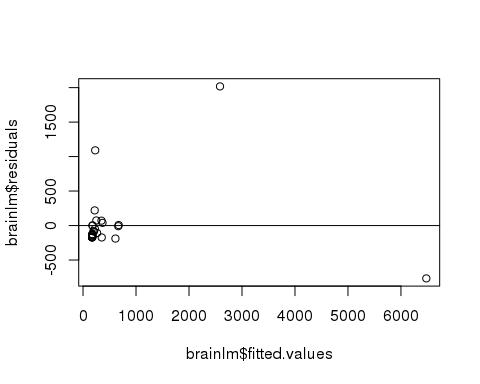
y = displacement + speed + speed2 R2 = .927

## 6a) Create a scatterplot to predict brain weight from body weight with a regression line, a plot of residuals versus the explanatory variable, a plot of residuals versus predicted values (i.e. fitted.values), and either a normal probability plot or a histogram of the residuals. Describe any patterns you see in the residual plots.

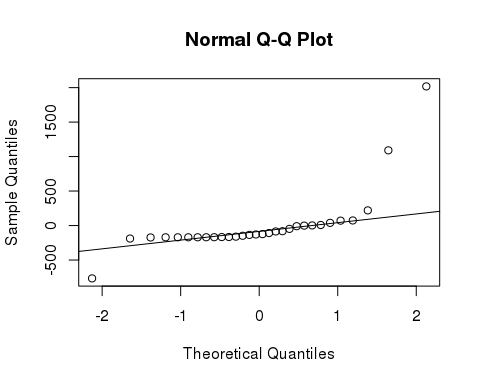
brainlm = lm(BrainWeightg ~ BodyWeightkg, data = brain)  
plot(brainlm$residuals ~ brain$BodyWeightkg)  
abline(h = 0)



plot(brainlm$residuals ~ brainlm$fitted.values)  
abline(h = 0)

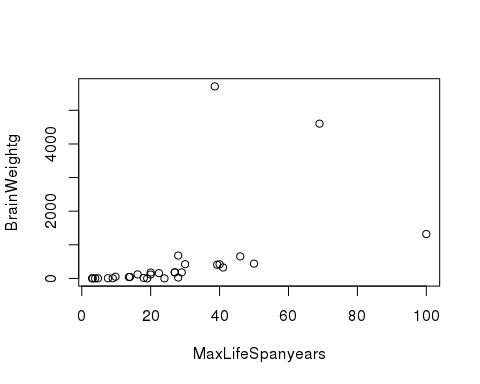
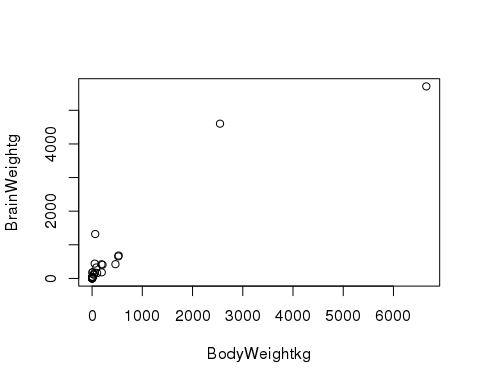


qqnorm(brainlm$residuals)  
qqline(brainlm$residuals)

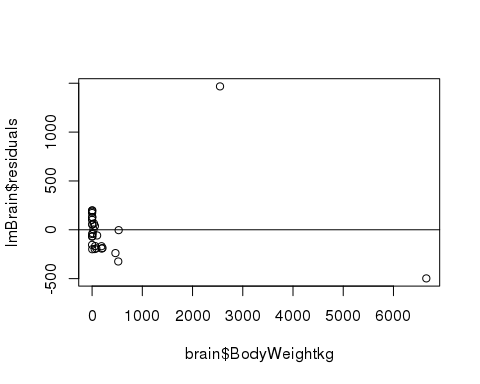
 Its hard to say if there is a pattern or not in the residual plots because the outliers spread the fram so much. There may be a slight logorithmic pattern to the data.

## 6b) Try various transformations of the explanatory and response variables to create a better linear regression model. Submit a fitted line plot and describe any patterns you see in the residual plots.

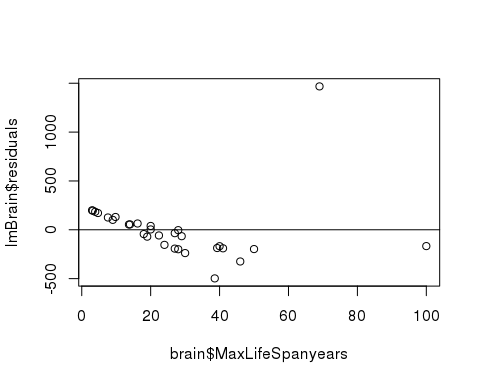
plot(BrainWeightg ~ BodyWeightkg + MaxLifeSpanyears, data = brain)



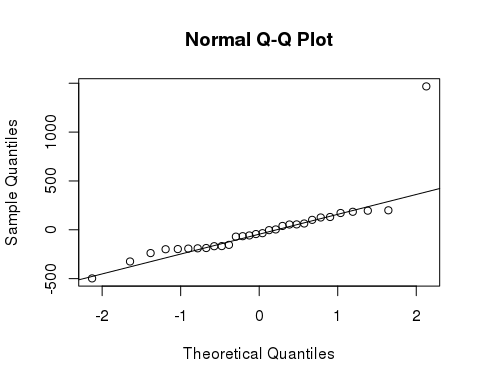
lmBrain = lm(BrainWeightg ~ BodyWeightkg + MaxLifeSpanyears, data = brain)  
#residuals  
plot(lmBrain$residuals~brain$BodyWeightkg)  
abline(h = 0)



plot(lmBrain$residuals~brain$MaxLifeSpanyears)  
abline(h = 0)



#normal plot  
qqnorm(lmBrain$residuals)  
qqline(lmBrain$residuals)



lmBrain

##   
## Call:  
## lm(formula = BrainWeightg ~ BodyWeightkg + MaxLifeSpanyears,   
## data = brain)  
##   
## Coefficients:  
## (Intercept) BodyWeightkg MaxLifeSpanyears   
## -248.9931 0.8731 16.8189

summary(lmBrain)

##   
## Call:  
## lm(formula = BrainWeightg ~ BodyWeightkg + MaxLifeSpanyears,   
## data = brain)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -497.68 -182.07 -39.33 92.23 1467.76   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -248.99311 100.85646 -2.469 0.0202 \*   
## BodyWeightkg 0.87308 0.05086 17.166 4.72e-16 \*\*\*  
## MaxLifeSpanyears 16.81888 3.11451 5.400 1.04e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 335.7 on 27 degrees of freedom  
## Multiple R-squared: 0.9374, Adjusted R-squared: 0.9328   
## F-statistic: 202.2 on 2 and 27 DF, p-value: < 2.2e-16

The normal plot shows one outlier, the rest of the points seem to closesly follow the normal distribution line. The residual plots show some strong patterns. Residuals vs body weight shows that most of the points are perpendicular to the abline, although they are closely packed together beacuse the outliers stretch the plot. Residuals vs max life span shows a negative linear trend in the data points, crossing the abline.