|  |
| --- |
| title: "Statistical Foundations for Data Science - Assignment8 " |
| author: "Travis Deason" |
| date: "November 18th 2017" |
| output: word\_document |

# MSDS 6371 UNIT 11 HW

## Problem 26 Chapter 8 The Metabolic data has the average mass, metabolic rate and average lifespan of 95 different species of mammals. Kleiber’s law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of ¾. Judge the adequacy of this theory with these data.  
### Be sure and provide: >\* A Scatterplot with confidence intervals of the regression line and prediction intervals of the regression line. In SAS and R!

metdata$massthreeqtr <- metdata$mass \*\* .75  
model <- lm(metab~massthreeqtr, data=metdata)  
coeff <- coefficients(model)  
coefficients(model)

## (Intercept) massthreeqtr   
## -481.3456 395.0158

plot <- ggplot(data=metdata, aes(x=massthreeqtr, y=metab)) + geom\_point()  
plot <- plot + labs(title="Weight vs Metabolism", x='mass^(3/4)', y='Metabolism')  
plot + geom\_abline(intercept= coeff[1],1, slope= coeff[2],1)

|  |  |
| --- | --- |
|  |  |

hist(model$residuals, breaks=10)

summary(model)

##   
## Call:  
## lm(formula = metab ~ massthreeqtr, data = metdata)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11712.7 -117.4 368.5 474.3 6598.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -481.346 213.467 -2.255 0.0265 \*   
## massthreeqtr 395.016 4.299 91.895 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1992 on 93 degrees of freedom  
## Multiple R-squared: 0.9891, Adjusted R-squared: 0.989   
## F-statistic: 8445 on 1 and 93 DF, p-value: < 2.2e-16

* The Plot eplot above shows the comparasin of average animal weight to the 3/4 power compared to metabolisim for 95 animals. A linear regression of the data shows the data is coorelated with a R^2 of .9891, meaning 98.9 percent of the variance in metabolisim (in Joules) can be explained by the relationship with weight in KG to the 3/4 power. The p-value on the slope is .0265 on a t-value of -2.255, and ~0 on a t-value of 91.895 for the intercept term; since the residuals are normal and of equal variance, we would consider this relationship to be valid at a .95 significence level.
* The equation for the line above has a slope of 395.016, meaning the metaboliic rate of an animal increases by 400 joules for each additional KG of animal mass. The intercept of the line is -481.346J meaning if an animal had a theoritical mass of 0, that animal would have a metabolic rate below 0. This of course is not possible, but since an animal cannot weigh zero, we will not assume this is an error in the model.

## Problem 29 Chapter 8

The Autism data shows the prevalence of autism per 10,000 ten-year old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year in this time period.  
A Scatterplot with confidence intervals of the regression line and prediction intervals of the regression line. In SAS OR R! A table showing the t-statistics and p-values for the significance of the regression parameters: β\_(0 ) and β\_1. In SAS OR R! The regression equation. Interpretation of the model given any transformation you may use. Hint Hint ;) A scatterplot of residuals. In SAS OR R! A histogram of residuals with normal distribution superimposed. (from SAS). Provide a measure of the amount of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly indicating the units of the response and the explanatory variables.

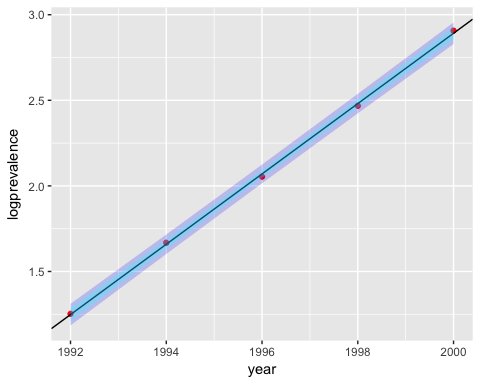
autisim <- read.csv('data/AutismDataProb29\_2\_2.csv')  
names(autisim) <- sapply(names(autisim), tolower)  
autisim$logprevalence <- log(autisim$prevalence)  
model <- lm(logprevalence~year, data=autisim)  
coeff <- coefficients(model)  
coefficients(model)

## (Intercept) year   
## -407.9752368 0.2054334

#plot <- ggplot(data=autisim, aes(x=year, y=logprevalence)) + geom\_point()  
#plot <- plot + labs(title="Autisim Prevelance per Year", x='Year', y='log(Autisim Prevelance)')  
#plot + geom\_abline(intercept= coeff[1],1, slope= coeff[2],1)  
slope <- model$coefficients['year']  
intercept <- model$coefficients[1]  
#Prediction intervals  
pred.int = predict(model, interval="prediction")

## Warning in predict.lm(model, interval = "prediction"): predictions on current data refer to \_future\_ responses

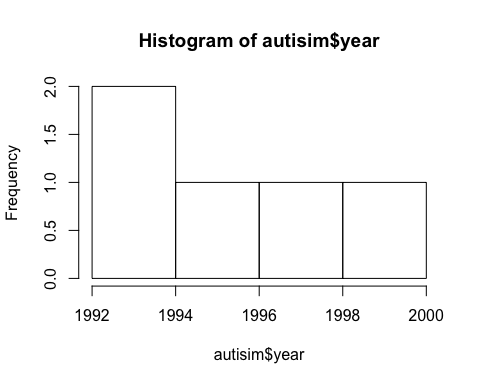
#Confidence intervals  
conf.int = predict(model, interval="confidence")  
# Add predictions and CI to original data  
autisim$pred.lower <- pred.int[,2]  
autisim$pred.upper <- pred.int[,3]  
autisim$ci.upper <- conf.int[,2]  
autisim$ci.lower <- conf.int[,3]  
## plot  
ggplot(data=autisim, aes(x=year, y=logprevalence, main='Scatterplot of data with regression line and Confidence intervals')) +   
 geom\_point(color= 'red') +  
 geom\_abline(intercept=intercept, slope=slope, color='black') +  
 geom\_ribbon(data=autisim, aes(ymin= pred.lower, ymax= pred.upper), fill = "blue", alpha = 0.2) +  
 geom\_ribbon(data=autisim, aes(ymin= ci.lower, ymax= ci.upper), fill = "cyan", alpha = 0.3)



summary(model)

##   
## Call:  
## lm(formula = logprevalence ~ year, data = autisim)  
##   
## Residuals:  
## 1 2 3 4 5   
## 0.004578 0.008655 -0.015795 -0.012686 0.015248   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.080e+02 4.953e+00 -82.38 3.94e-06 \*\*\*  
## year 2.054e-01 2.481e-03 82.79 3.88e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01569 on 3 degrees of freedom  
## Multiple R-squared: 0.9996, Adjusted R-squared: 0.9994   
## F-statistic: 6855 on 1 and 3 DF, p-value: 3.884e-06

hist(autisim$year)



* The Plot eplot above shows the comparasin of the year compared with the log of autisim prevalance for that year. A linear regression of the data shows the data is coorelated with a R^2 of .994, meaning 99.4 percent of the variance in autisim prevelanve (in autistic children per 10000) can be explained by the relationship with weight in e to the year. The p-value on the slope is 3.94e-6 on a t-value of -82.38, and 3.88e-6 on a t-value of 82.79 for the intercept term; since the residuals are of equal variance, we would consider this relationship to be valid at a .95 significence level.
* The equation for the line above has a slope of .2054, meaning the log of the rate of an autisim increases by .2054 children per 10,000 children born every year. The intercept of the line is -4.080e-2 meaning for year zero there were theoretically the log of -400 children diagnoised with autisim for every 10,000 children; Since it is impossible to have less then zero children diagnoised with autisim, and data has not been tracked back to the year 0, this data can be considered out of scope for the dataset.