Lab 3 - Homework

setwd('C:/Users/Owner/Desktop/UT\_S15/GEOG415/snmarkley/homework3/')

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.2

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.1.2

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.1.2

library(magrittr)

## Warning: package 'magrittr' was built under R version 3.1.2

precip <- read.csv('C:/Users/Owner/Desktop/UT\_S15/GEOG415/snmarkley/homework2/data/precip.csv')  
rwi <- read.csv('C:/Users/Owner/Desktop/UT\_S15/GEOG415/snmarkley/homework2/data/rwi.csv')  
precip.df <- precip %>% gather(key=month, value=precip, -year) %>% arrange(year, month)  
precip.df <- precip.df %>% mutate(water\_year = lead(year, 6))  
annual\_precip <- precip.df %>% group\_by(water\_year) %>%   
 summarize(precip=sum(precip)) %>%  
 filter(water\_year>1895)  
rwi.precip.df <- left\_join(rwi, annual\_precip, by=c('year'='water\_year'))  
model <- lm(rwi~precip, data=rwi.precip.df)  
model2 <- lm((rwi)~(precip), data=rwi.precip.df, subset=(precip<20))

1. For the first three diagnostic plots, describe which assumptions they help to visually evaluate.

By examining the "Residuals vs Fitted" diagnostic plot, I can check whether or not the errors have a constant variance (i.e. homoskedasticity). If the points along the line fan outward, the data may have nonconstant variance, violating one of assumptions of linear regression. In addition, from this same plot I can visually evaluate whether or not the errors are correlated. To satisfy this assumption, the errors should be randomly assembled around the horizontal line at y=0.

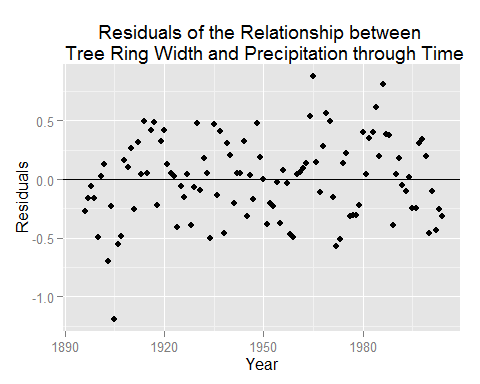
The "Normal Q-Q Plot" displays the degree to which the errors are normally distributed. According to this assumption, the errors should roughly follow along the y=x line, or the number of observations should be sufficiently large.

Similar to the first plot, the "Scale-Location" plot can help identify heteroskedasticity.

1. These data are time series data. In general, data are correlated in time, and this would be a problem for simple linear regression. Plot the regression residuals vs time, and visually assess whether there is evidence of temporal (serial) correlation.

model <- lm(rwi ~ precip, data=rwi.precip.df, na.action = na.exclude)  
rwi.precip.df$resid <- residuals(model)  
filter(rwi.precip.df,year >= 1895) %>% ggplot(data=.) + geom\_point(aes(x=year, y=resid)) +   
 labs(x="Year",y="Residuals") + geom\_abline(aes(slope=0, intercept=0)) +   
 ggtitle("Residuals of the Relationship between \n Tree Ring Width and Precipitation through Time")

## Warning: Removed 1 rows containing missing values (geom\_point).



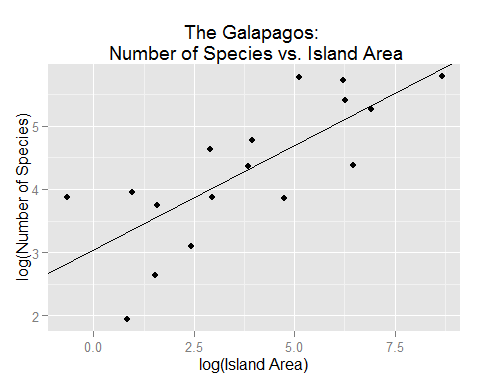
From this plot there does appear to be some relationship between time and the residuals. For instance, residuals near the beginning of the study period are disproportionately negative, and some residuals near the later years of the study period are large positives. It also appears that the points could follow a wave pattern, suggesting that temporal autocorrelation may be influencing these results.

1. In the data folder is the dataset GalapagosData.txt. The species data represents the number of species recorded from each of the Galapagos islands. A fundamental 'law' of island biogeography is that species diversity tends to follow a power law relationship with island area, i.e.
2. This is not linear, but it suggests that the following regression might make sense:
3. is not quite , rather .  
   Fit this regression, and present a brief write-up that a) describes the results in words, and b) summarizes your conclusions from diagnostic model checking.

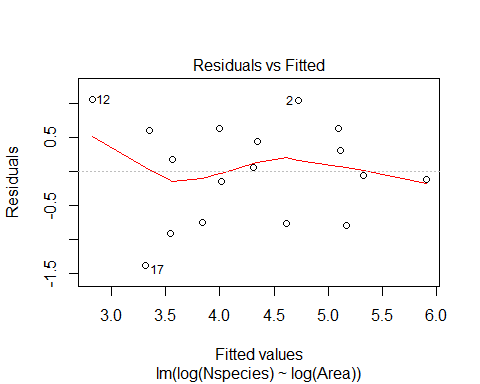
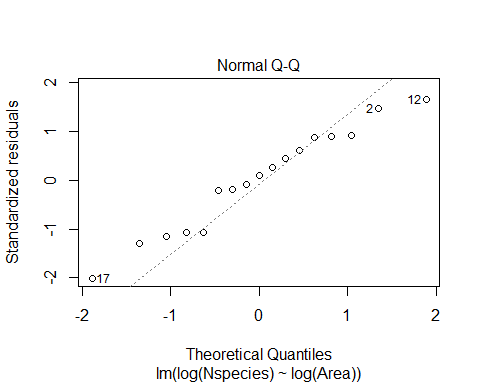
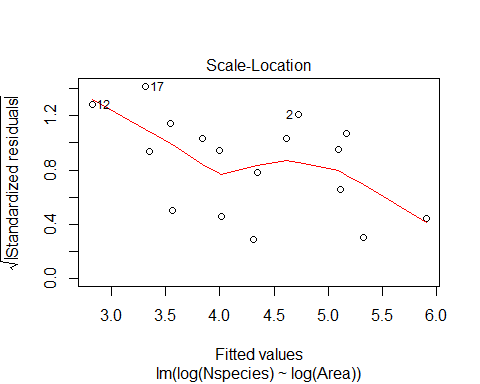
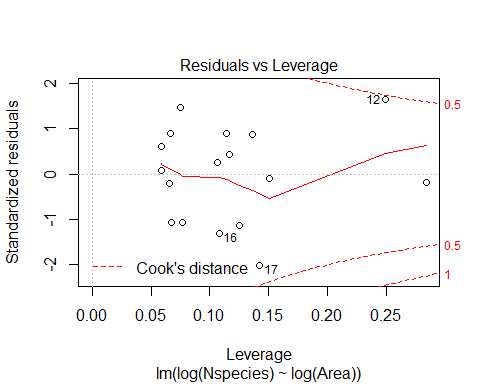
gal.data <- read.table("C:/Users/Owner/Desktop/UT\_S15/GEOG415/snmarkley/homework3/data/GalapagosData.txt", header=TRUE, quote="\"")  
#fitting regression  
fit.gal <- lm(log(Nspecies)~log(Area),data=gal.data)  
summary(fit.gal)

##   
## Call:  
## lm(formula = log(Nspecies) ~ log(Area), data = gal.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.37282 -0.75233 0.06034 0.59768 1.04971   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.03895 0.32728 9.286 1.31e-07 \*\*\*  
## log(Area) 0.33059 0.07194 4.595 0.00035 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7378 on 15 degrees of freedom  
## Multiple R-squared: 0.5847, Adjusted R-squared: 0.557   
## F-statistic: 21.12 on 1 and 15 DF, p-value: 0.0003501

# alpha = 3.03895, beta = 0.33059  
#Plot relationship  
ggplot(data=gal.data) + geom\_point(aes(x=log(Area), y=log(Nspecies))) +   
 geom\_abline(aes(slope=0.33059, intercept=3.03895)) + ggtitle("The Galapagos: \n Number of Species vs. Island Area") + labs(x="log(Island Area)", y="log(Number of Species)")



#Diagnostic plots  
plot(fit.gal)

Plotting the regression of the log of the number of species against the log of the island area reveals a positive linear relationship, suggesting that larger islands in the Galapagos are associated with a larger number of species. The R-squared value of 0.5847 suggests that the area of the islands explains about 58.47 percent of the variation in the number of species that inhabit them. However, the diagnostic plots reveal that there may be some issues with the data. The first and third plots ("Residuals vs Fitted" and "Scale-Location") reveal a possible trend in the residuals, which would violate one assumption of linear regression. Additionally, the Quantile-Quantile plot shows that the standardized residuals do not quite follow a normal distrution, though it is close. One problem contributing to these issues appears to be the low sample size (n=17), limiting the amount of information that can be interpreted from this regression model.