STAT 614 Lab 7

Amy Mailo

2022-10-28

library(Sleuth3) # all textbook data sets  
library(mosaic) # More summary stat and plots.

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.

##   
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

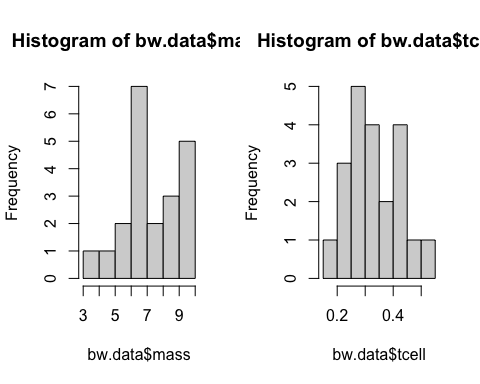
library(ggplot2) # More plot options.

bw.data <- read.csv("Desktop/R files/blackwheatears.csv", header=TRUE)  
head(bw.data, 3)

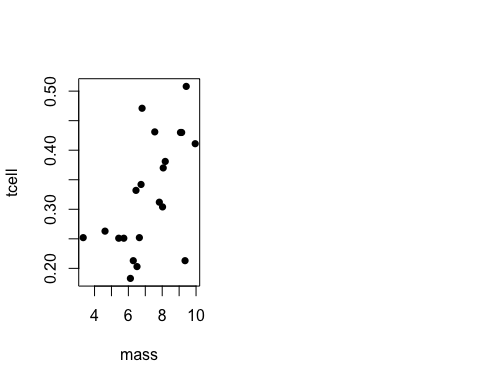
## mass tcell  
## 1 3.33 0.252  
## 2 4.62 0.263  
## 3 5.43 0.251

#1 What were the overall goals of the study?   
#GOAL: to identify if stone carrying reflected male health status for black wheaters.   
  
#2 What was the study design?   
#A randomized experiment AND observational study. For the experiment, they randomly allocated males to one of the two treatment groups they created (1) wing cutting or (2) NOT wing cutting to see if this impacted their ability to carry stones   
  
#3 What variables were measured on the birds?   
#Original article: researchers measured health status by looking at t-cell response, hematocrit, and leukocyte concentration, stone mass, and parasite abundance.   
#Data set, Figure 1 of article: measured T-cell and stone mass  
  
#4 Did the authors state any specific hypotheses they wanted to test?   
#Hypothesis that the males signal their current health status by carrying many heavy stones. More specifically "If male black wheatears carrying many heavy stones are healthy, we would expect them to have stronger T-cell responses than males carrying few light stones.”  
  
#5 What statistical methods/models were used to address the research questions in the paper?  
#ANOVA, linear regression, t-tests, and Bonferroni adjustments were used for multiple comparisons.  
  
#5 Conduct a simple linear regression analysis to predict health, as measured by T-cell response, from the average stone mass carried.  
  
#6 What is the response variable? What is the predictor (aka explanatory) variable? Are these the same as in the paper? If not, what did the authors do? Why?  
#Response variable should be T-cell responses since this is what is changing in response to the explanatory variable/predictor (which is the mean stone mass). We need to build a model in which the 'stone mass' helps us predict 'T-cell'.  
#in the paper, however, the authors predict the amount of stone mass carried from t-cell response.  
#thus, with different research goals, the choice of predictor (explanatory variable) and response variable may end up being different.

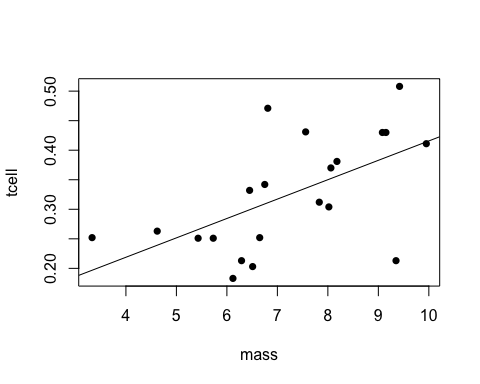
#7 Prepare graphical summary of the variables. Describe the nature of the relationship displayed in this graph  
par(mfrow=c(1, 2))  
hist(bw.data$mass)  
hist(bw.data$tcell)



plot(tcell ~ mass, data=bw.data, pch=16)  
  
#The histograms show that stone mass is slighted skewed to the left, while the t-cell is approximately symmetrical. Neither variable has extreme outliers.   
#The scatter plot looks similar to paper's Figure 1 aside from x and y variables being reversed. We need to ensure that T-cell is on the vertical axis because we designated it as a response variable.   
#It appears to show a moderate linear association between the variables. As mass increases, the mean of t-cell increases.



#8 Does a linear regression model seem appropriate?  
bw.reg <- lm(tcell ~ mass, data=bw.data)  
plot(tcell ~ mass, data=bw.data, pch=16)  
abline(bw.reg)



#Yes, the scatter plot suggests that a linear association seems appropriate.

#9 State the estimated mean response function (i.e., the regression function). Did you get the same fitted line as in the paper? Why not?!  
bw.reg$coefficients

## (Intercept) mass   
## 0.08749698 0.03282149

#The estimated regression function is: ˆtcell = 0.087 + 0.033(mass).  
#This regression model is different from that in the article (ˆy = 4.00 + 9.98x) because the article used “stone mass” as the response variable (y) and “t-cell” as the predictor x.  
#x and y's were switched- that's the difference

#10 Set-up and test the hypothesis that the slope parameter is zero. Interpret the results of this test in the context of this problem.  
#H0 : β1 = 0, vs. HA : β1 6= 0.  
#The above test can be tested using either t-test for the slope, or the F-test from the ANOVA table (for simple linear regression.)-- these are our options to help us test  
summary(bw.reg)

##   
## Call:  
## lm(formula = tcell ~ mass, data = bw.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.18138 -0.04673 0.01796 0.04219 0.15999   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.08750 0.07868 1.112 0.27996   
## mass 0.03282 0.01064 3.084 0.00611 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.08102 on 19 degrees of freedom  
## Multiple R-squared: 0.3336, Adjusted R-squared: 0.2986   
## F-statistic: 9.513 on 1 and 19 DF, p-value: 0.006105

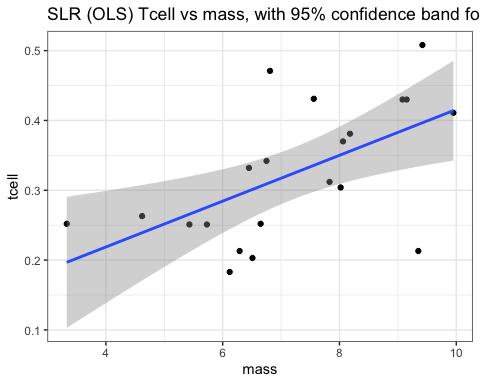
#The t-test statistic for this hypothesis t = 3.084 and the resulting p-value is p = 0.006. The data provides significant evidence to support that the birds‘ t-cell response is associated with stone mass (when assuming a straight-line association). So we are led to believe that birds' t-cell response may have a relationship with stone mass.

#11 Find the estimated mean T-cell response and corresponding 95% point-wise confidence intervals for stone masses of x = 0, x = 3g, x = 7.2g, x = 10g, and x = 20g. Interpret 2 sets of the CIs in the context of the problem.   
predict(bw.reg, newdata=data.frame(mass=c(0, 3, 7.2, 10, 20 )), interval = "confidence", level=0.95)

## fit lwr upr  
## 1 0.08749698 -0.07717487 0.2521688  
## 2 0.18596146 0.08527344 0.2866495  
## 3 0.32381172 0.28680714 0.3608163  
## 4 0.41571189 0.34327770 0.4881461  
## 5 0.74392680 0.45653698 1.0313166

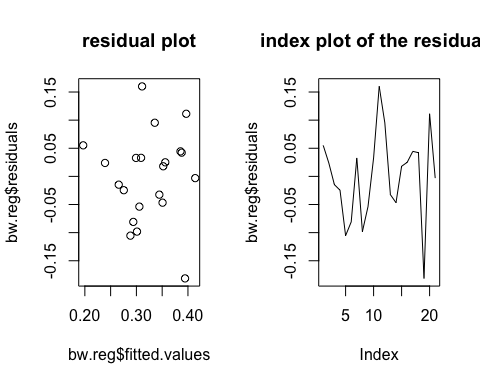
#For ALL male black wheatears who carry 3g of stones, the mean t-cell response is estimated to be 0.18596. With 95% confidence, the mean t-cell response is estimated to be between 0.08527 and 0.28665.  
#For ALL male black wheatears who carry 7.2g of stones, the mean t-cell response is estimated to be 0.32381. With 95% confidence, the mean t-cell response is estimated to be between 0.28681 and 0.36082.

#12 Should we be concerned about any of the above prediction (i.e., estimations)?  
#Yes, we should be concerned about extrapolation because of the predictions at x = 0 and x = 20. These x-values are outside the range of our data which shows that there is extrapolation.  
  
#13 Plot the 95% point-wise confidence band for the estimated mean response (t-cell). Does this plot take into consideration the fact that you are conducting multiple intervals (one for each x-value)? Why or why not? How would the plot be different if you were (or weren’t, as the case may be)?  
  
#use ggplot   
ggplot(bw.data, aes(x=mass, y=tcell)) +  
geom\_point() +  
geom\_smooth(method=lm, se=TRUE, formula = y ~ x, level=0.95) +  
ggtitle("SLR (OLS) Tcell vs mass, with 95% confidence band for mean of Tcell")+  
theme\_bw()

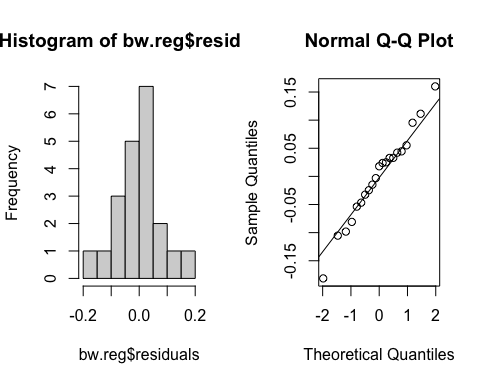


#The confidence band in the above plot does NOT adjust for multiple intervals. At each x-value,the point-wise confidence intervals for mean of y is at 95% confidence level.  
#BUT, If we adjust for multiple intervals, we will need to use a HIGHER confidence level for each point-wise interval, OR to use a DIFFERENT critical value. Both will result in wider interval if we did make an adjustment (they would then be called “joint”, or “family”, or “simultaneous”, confidence intervals).

#14 Would a 95% prediction band be wider or narrower? Why?  
#The 95% prediction band would be WIDER The prediction band predicts an individual y-value at a given x-value. At the same x-value, the standard error of an individual y is ALWAYS GREATER than the standard error of the mean of y.  
  
#15 Plot the residuals to evaluate the assumptions of the linear regression model? Do you think they are met for this data?  
par(mfrow=c(1, 2))  
plot(bw.reg$fitted.values, bw.reg$residuals, main="residual plot")  
plot(bw.reg$residuals, type="l", main="index plot of the residuals")



hist(bw.reg$residuals)  
qqnorm(bw.reg$residuals)  
qqline(bw.reg$residuals)



shapiro.test(bw.reg$residuals)

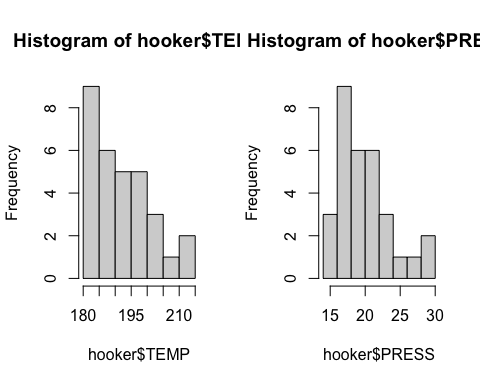
##   
## Shapiro-Wilk normality test  
##   
## data: bw.reg$residuals  
## W = 0.98439, p-value = 0.974

#The model assumptions appear to be satisfied.  
  
#The residual plot (residuals vs ˆy) shows a “rectangular box”. Because there is no obvious signs of curved pattern or signs of non-constant variance, the linear assumption and constant variance assumptions appear to be valid.  
#The index plot of the residuals doesn’t have special pattern or tread. Such finding and the descriptions from the article suggest that the samples are randomly and independently selected from the population.  
#The histogram and the Normal QQ plot of the residuals suggest that the residuals follow Normal distribution. This is also verified by the Shapiro Test of Normality.  
#The histogram and the residual plot do not show extreme outliers.  
#The “Test of Linearity” (aka. Lack-of-fit test) does not apply this data, because we would have to take extra steps in order to use this test.

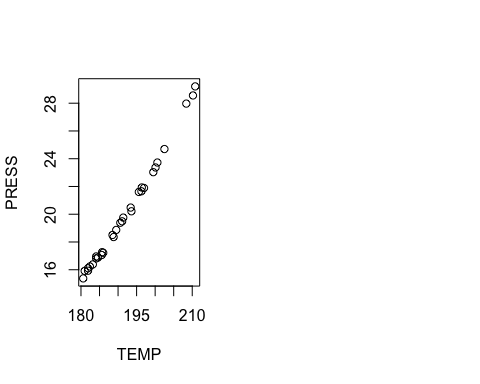
#Exercise 2  
hooker <- read.table("Desktop/R files/hooker.txt", header=T)  
head(hooker, 5)

## TEMP PRESS  
## 1 210.8 29.211  
## 2 210.2 28.559  
## 3 208.4 27.972  
## 4 202.5 24.697  
## 5 200.6 23.726

#1 In physics, boiling point varies as a result of altitudes and pressure changes. Why do we use the baromeric pressure as the response variable (y) in this study?  
#19th century barometers were extremely weak and hard to transport. But, boiling water was WAY easier.   
#Hooker thoughg that it would be easier to determine their relationship by exploiting this relationship between barometric pressure and boiling point of water. So the goal became: build a model so that we could use the boiling point (temperature) to predict or estimate the barometric pressure and altitude.  
  
#the higher the pressure, the boiling temperature is higher   
#temperature used as x variable, pressure is used as y variable- he wants to understand the associations so that he can use the boiling temperature to predict the air pressure and thus the altitude   
#use temp (explanatory) to predict air pressure (response)  
  
  
  
#2 Plot the data. Does a straight line association appear to be appropriate.   
par(mfrow=c(1, 2))  
hist(hooker$TEMP)  
hist(hooker$PRESS)



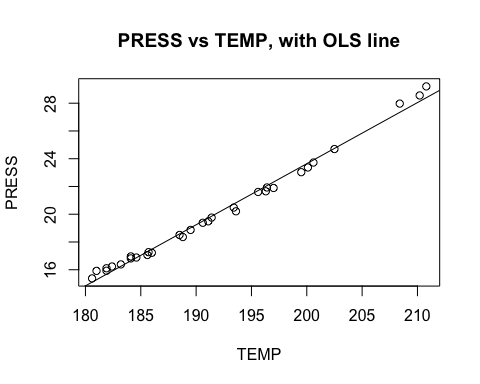
plot(PRESS ~ TEMP, data=hooker)  
  
#With the scatter plot shows us that there appears to be a linear association between the two variables- boiling temperature and pressure curvature, but it is not clear.



#3 Fit a preliminary linear regression model. Plot the estimated regression line with the data.  
hooker.lm <- lm(PRESS ~ TEMP, data=hooker)  
hooker.lm$coefficients

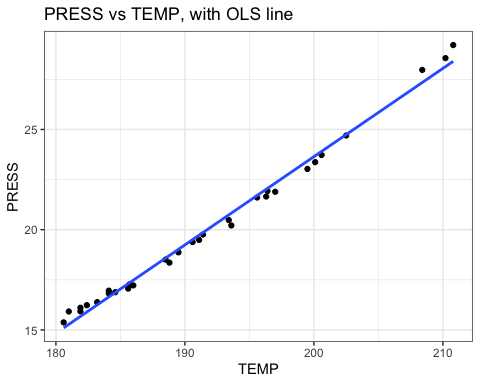
## (Intercept) TEMP   
## -64.4127512 0.4402819

plot(PRESS ~ TEMP, data=hooker, main="PRESS vs TEMP, with OLS line")  
abline(hooker.lm)



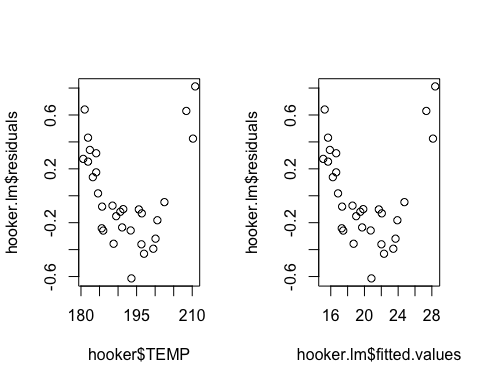
# Another plot example for Tidyverse users:  
ggplot(hooker, aes(x=TEMP, y=PRESS)) +  
geom\_point() +  
geom\_smooth(method="lm", se=FALSE) +  
theme\_bw() +  
ggtitle("PRESS vs TEMP, with OLS line")

## `geom\_smooth()` using formula 'y ~ x'

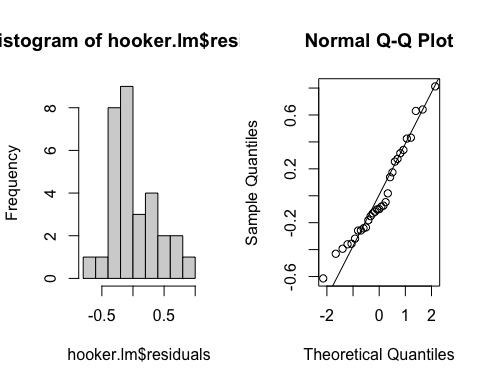


#The estimated linear regression line is ̂Pressure = −64.41 + 0.44(Temperature)  
#The straight line does seem to fit the data. But if we observe more carefully, we’ll notice the curvature!

#4. Use residual plot, histogram, and Q-Q plot to check the model assumptions.  
par(mfrow=c(1, 2))  
plot(hooker$TEMP, hooker.lm$residuals)  
plot(hooker.lm$fitted.values, hooker.lm$residuals)



#or   
hist(hooker.lm$residuals)  
qqnorm(hooker.lm$residuals)  
qqline(hooker.lm$residuals)



#We can use x or ˆy (fitted value, aka. the estimated value, the predicted value) in the residual plot-- which will give the same information in simple linear regression. We do NOT use the observed y.  
#The residual plot shows a curved pattern. Thus, the linearity assumption is NOT valid because curved pattern = linearity assumption NOT met.   
#The residual plot does NOT show signs of non-constant variance or extreme outliers It has a constant “band width” vertically.  
#The histogram and the Q-Q plot do NOT show strong signs of Non-Normality

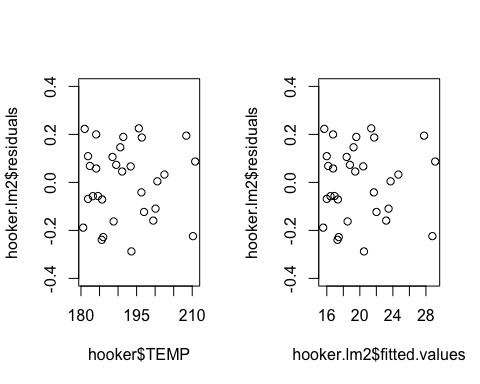
#5 Create a new variable that equals to (TEMP)2. Refit the model with both TEMP and (TEMP)  
## The "I" in the following code tells R to compute a new variable as predictor.  
hooker.lm2 <- lm(PRESS ~ TEMP + I(TEMP^2), data=hooker)  
hooker.lm2$coefficient

## (Intercept) TEMP I(TEMP^2)   
## 88.016622655 -1.129540986 0.004033036

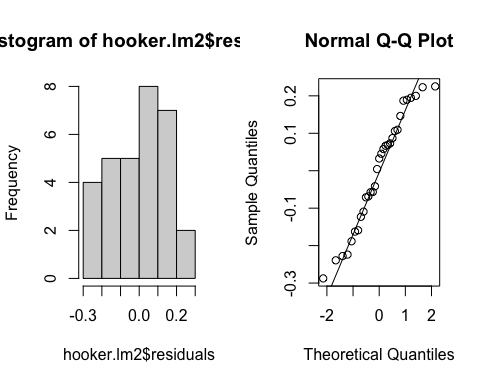
# The above code is equivalent to the following  
hooker$TEMP2 <- hooker$TEMP^2  
hooker.lm2b <- lm(PRESS ~ TEMP + TEMP2, data=hooker)  
hooker.lm2b$coefficient

## (Intercept) TEMP TEMP2   
## 88.016622655 -1.129540986 0.004033036

#6 Use plots to check the model assumptions for the model based on the transformed data  
  
# Either ONE of them will do.  
par(mfrow=c(1, 2))  
plot(hooker$TEMP, hooker.lm2$residuals, ylim=c(-0.4, 0.4))  
plot(hooker.lm2$fitted.values, hooker.lm2$residuals, ylim=c(-0.4, 0.4))

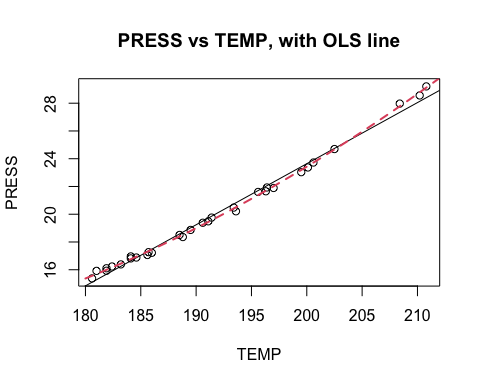


hist(hooker.lm2$residuals)  
qqnorm(hooker.lm2$residuals)  
qqline(hooker.lm2$residuals)



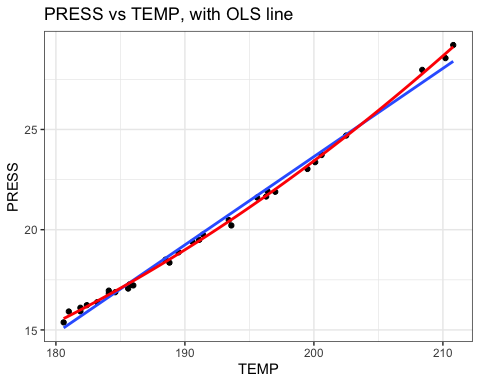
#The residual plot shows a horizontal band WITHOUT other special pattern. Special patterns can include but are not limited to: curvature, or varying vertical “band width”.  
#The histogram and Q-Q plot suggest we can assume Normality as the graphs depict the shape of the data points being normally distributed

#7 Plot the data with both estimated lines.   
plot(PRESS ~ TEMP, data=hooker, main="PRESS vs TEMP, with OLS line")  
abline(hooker.lm)  
newX <- seq(180, 220, by=0.5)  
yhat2 <- predict(hooker.lm2, newdata=data.frame(TEMP = newX))  
lines(newX, yhat2, col=2, lwd=2, lty=2)

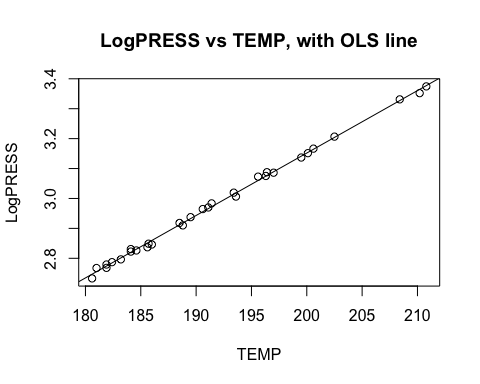


# Another plot example for Tidyverse users:  
ggplot(hooker, aes(x=TEMP, y=PRESS)) +  
geom\_point() +  
geom\_smooth(method="lm", se=FALSE) +  
geom\_smooth(method="lm", formula = y ~ x + I(x^2), se=FALSE, color="red") + theme\_bw() +  
ggtitle("PRESS vs TEMP, with OLS line")

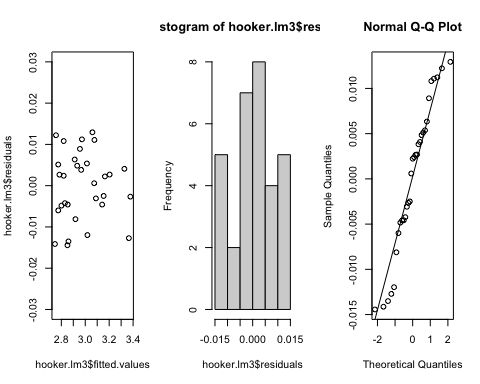
## `geom\_smooth()` using formula 'y ~ x'



#8 Note that there often can be more than 1 possible ways to transform the data, so that the model assumptions are valid on the transformed data. In this case, a log-transformation on the response variable (PRESS) may also work. In practice, we should consider both the numeric results (e.g., model assumptions, model fitness, etc.) and the ability to interpret the result in the context of the problem to decide the “final” model(s). Run the following code, write down the estimated regression function, and comment on the model assumptions.  
hooker$LogPRESS <- log(hooker$PRESS)  
hooker.lm3 <- lm(LogPRESS ~ TEMP, data=hooker)  
plot(LogPRESS ~ TEMP, data=hooker, main="LogPRESS vs TEMP, with OLS line")  
abline(hooker.lm3)



# Either ONE of them will do.  
par(mfrow=c(1, 3))  
plot(hooker.lm3$fitted.values, hooker.lm3$residuals, ylim=c(-0.03, 0.03))  
hist(hooker.lm3$residuals)  
qqnorm(hooker.lm3$residuals)  
qqline(hooker.lm3$residuals)



#The estimated regression model is: ̂ln(pressure) = −1.02 + 0.02(temperature). It can also be written as ̂pressure = e(−1.02+0.02(temperature)).  
#The model assumptions (Normality, constant variance, linearity, no outliers) appear to be met when using log(PRESS) as the response in the residual plot and qqplot for the residual. The indepdnence assumption, on the other hand, can be looked at when you look at the data collection/sequence plot part of the study's design.