Chapter 9.1: Linear regression explanatory variable fixed by experiment

Amy Hurford ([ahurford@mun.ca](mailto:ahurford@mun.ca))

2023-07-11

Here, we show the R code to analyze the Pcorn data. We test the assumptions of the linear regression. As we find an assumption violation, we then perform a randomization to determine if the assumption violation matters.

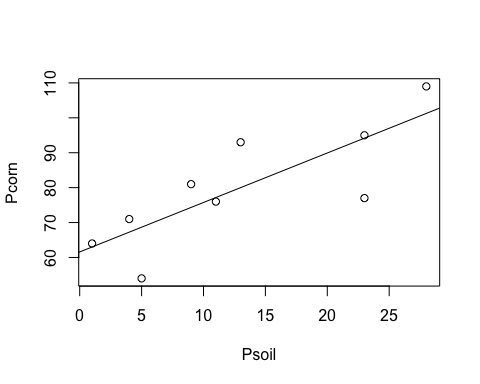
This .Rmd file can be downloaded here: <https://github.com/ahurford/biol-4605-data/blob/main/code/Ch9-1.Rmd>

## Loading the data and doing the linear regression

# This imports the data into R from a website without needing to download  
# Generally, click 'Raw' on the github website and copy url  
data <- read.csv('https://raw.githubusercontent.com/ahurford/biol-4605-data/main/data/corn.csv', fill=TRUE)  
  
# Give the variables shorter names:   
# Response variable  
Pcorn = data$Pcorn  
# Explanatory variable  
Psoil = data$Psoil  
  
# Do the linear regression:  
reg <- lm(Pcorn~Psoil)  
# see the results of your regression  
summary(reg)

##   
## Call:  
## lm(formula = Pcorn ~ Psoil)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.169 -1.166 1.003 6.668 13.000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 61.5804 6.2477 9.857 2.35e-05 \*\*\*  
## Psoil 1.4169 0.3947 3.590 0.00886 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.69 on 7 degrees of freedom  
## Multiple R-squared: 0.648, Adjusted R-squared: 0.5977   
## F-statistic: 12.89 on 1 and 7 DF, p-value: 0.008859

# Plot of the regression (line) and the data (open circles)  
plot(Psoil,Pcorn)  
abline(reg)



## Calculating the likelihood ratio

Calculate the likelihood ratio for the evidence supporting the linear regression relative to a null model that for all values of phosphorous in the soil (Psoil), the phosphorous in the corn (Pcorn) is predicted as the mean recorded value.

# NULL MODEL  
# A dataframe with the residuals and the residuals^2  
# Model values are equal to mean(Pcorn)  
data.eq\_null = data.frame(Psoil = Psoil, Data = Pcorn, Model = rep(mean(Pcorn),length(Pcorn)), res = Pcorn-mean(Pcorn), res2 = (Pcorn-mean(Pcorn))^2)  
# Print this dataframe with the residuals and the residuals^2 for the null model  
data.eq\_null

## Psoil Data Model res res2  
## 1 1 64 80 -16 256  
## 2 4 71 80 -9 81  
## 3 5 54 80 -26 676  
## 4 9 81 80 1 1  
## 5 13 93 80 13 169  
## 6 11 76 80 -4 16  
## 7 23 77 80 -3 9  
## 8 23 95 80 15 225  
## 9 28 109 80 29 841

#The sum of the residuals should be 0.  
# SS.total is the total sum squares  
sum(data.eq\_null$res)

## [1] 0

SS.total = sum(data.eq\_null$res2)  
SS.total #prints the value

## [1] 2274

# FULL MODEL  
# Similar to the data frame for the null model, but for the full model the model values are equal to reg$fitted.values (were reg is the linear regression model object)  
data.eq\_full = data.frame(Psoil = Psoil, Data = Pcorn, Model = reg$fitted.values, res = Pcorn-reg$fitted.values, res2 = (Pcorn-reg$fitted.values)^2)  
# Print this dataframe with the residuals and the residuals^2 for the regression model  
data.eq\_full

## Psoil Data Model res res2  
## 1 1 64 62.99728 1.0027248 1.0054570  
## 2 4 71 67.24796 3.7520436 14.0778312  
## 3 5 54 68.66485 -14.6648501 215.0578295  
## 4 9 81 74.33243 6.6675749 44.4565555  
## 5 13 93 80.00000 13.0000000 169.0000000  
## 6 11 76 77.16621 -1.1662125 1.3600517  
## 7 23 77 94.16894 -17.1689373 294.7724090  
## 8 23 95 94.16894 0.8310627 0.6906652  
## 9 28 109 101.25341 7.7465940 60.0097187

#The sum of the residuals should be 0.  
#SS.res is the sum squares of the residuals  
sum(data.eq\_full$res)

## [1] 1.421085e-14

SS.res = sum(data.eq\_full$res2)  
SS.res #prints the value

## [1] 800.4305

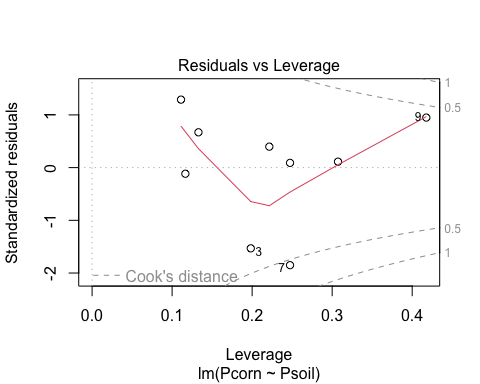
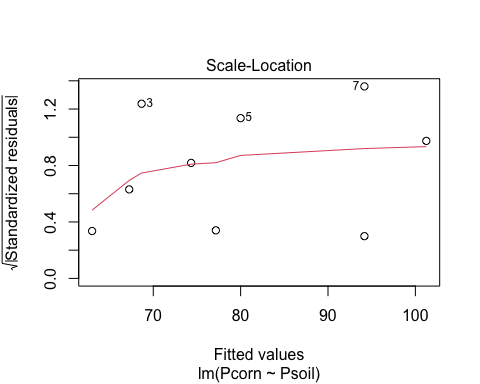
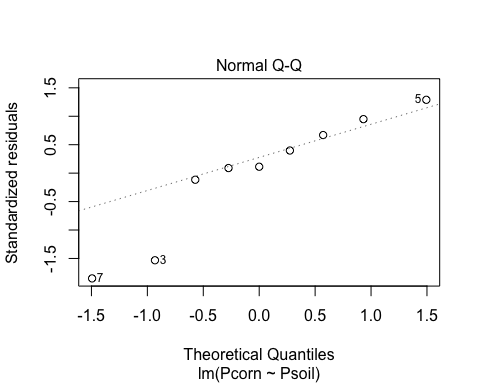
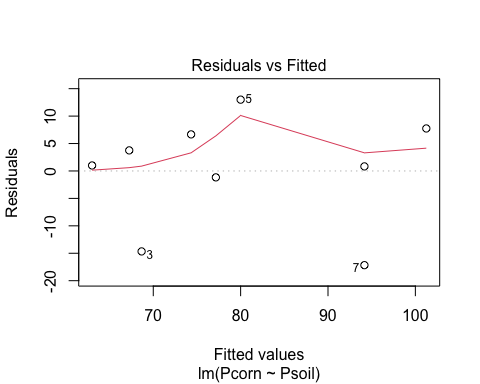
# Calculate the Likelihood ratio  
n = length(Pcorn)  
LR = (SS.res/SS.total)^(-n/2)  
LR # prints

## [1] 109.7996

## Model assumptions

The plot function plot() applied to a model object (here reg) prints a series of graphs that visualize the agreement of the fitted model with the assumptions of the linear regression.

# A series of plots to check model assumptions  
plot(reg)



The Q-Q plot shows a poor agreement of the linear model residuals relative to a normal distribution (an assumption of the general linear model). In particular, observations 3 and 7 are much further from the model predicted values than would be expected if the residuals were normally distributed.

## Randomization

Does this possible violation of the model assumptions change our conclusion that there is a relationship between the phosphorous in the soil and in the corn?

The regression slope coefficient was 1.42 with a p-value of 0 suggesting that the probability of observing these data, if there truly was no relationship between the phosphorous in the soil and in the corn (i.e., a Type II error) is 88.6 out of 10000 *given that the assumptions of the general linear model are met*.

From our Q-Q plot, we know that the assumptions of the general linear model are not met, but does this matter?

To test this, we randomly assign values of Pcorn to values of Psoil as new hypothetical data. Conceptually, if there really is no relationship between Psoil and Pcorn, then any of the Pcorn values could have been reported for any of the Psoil measurements. How does an F-value calculated for the observed data, compare with an F-value calculated for data where Pcorn values are randomly assigned to any Psoil measurement?

The function below completes 10,000 replicates of hypothetical data that might be observed if there was truly no effect of Psoil: this is sample(y,length(y),TRUE). Here y is the response variable supplied to the function, and we sample from y to make new response variable data of length length(y). The sample is performed with replacement (=TRUE).

The explanatory variable (in the function is x). In our application, this will be Psoil, and this remains the same for all the linear regressions with the randomly associated response variable data. From the regression, we extract the F-value with $F value``.

Below is a function, F.rand(), that is defined for any supplied x and y.

F.rand = function(x,y){  
result = replicate(10000,anova(lm(sample(y,length(y),TRUE)~x))$`F value`[1])  
}

Note that randomly choosing an observed Pcorn value to be associated with an observed Psoil value makes different assumptions than the assumptions of the general linear model, so this approach serves to test the robustness of our conclusions to the assumptions of the general linear model (which our error diagnostic plots showed to be violated).

# We use the function F.rand() and calculate 10,000 F-values for random associations between Pcorn and Psoil.  
rand = F.rand(Psoil,Pcorn)  
# Prints the F-values for the first 6 random data sets:  
head(rand)

## [1] 3.71860567 0.67349406 0.22804592 1.89654952 0.08030211 4.52348016

The F-value that we calculated from the data was quite a bit larger than the first 6 F-values for the data generated by random associations between the explanatory and response variables:

F.data <- anova(lm(Pcorn~Psoil))$`F value`[1]  
F.data # prints

## [1] 12.8868

How many F-values from the randomizations are bigger than the F-value for the real data? We calculate this as:

n.rand = length(which(rand>F.data))  
n.rand #prints

## [1] 96

Under the assumptions of the general linear model, we expected failure to reject the null hypothesis 88.6 times out of 10,000. For the different assumptions of the randomization, we failed to reject the null hypothesis 96 times out of 10,000.

However, in either case, it is extremely unlikely that there is no relationship between Psoil and Pcorn: for the randomization this occurs 0.96% of the time. Therefore, the violation of the general linear model assumptions, mean that we might not be able to interpret the reported p-vale as the probability of the Type II error, however, the conclusion that there is a relationship between Psoil and Pcorn is justified, even given the assumption violation shown in the QQ-plot. This result makes sense, since the likelihood ratio (the calculated value was 109.8) suggested very strong support for this relationship also.