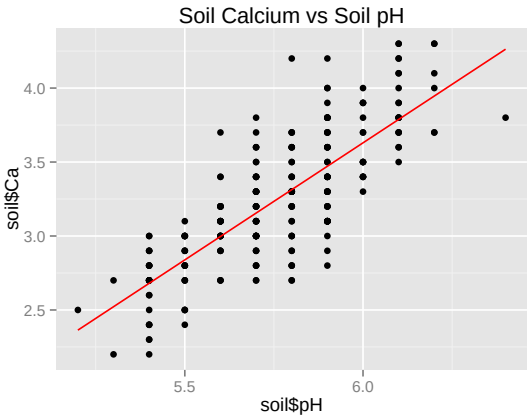
Stat 505 Assignment 6

October 10, 2014

1. In a study of soil properties, samples were taken on a 10 point by 25 point grid. We'll work with two variables: response Ca (calcium concentration) and predictor pH (low numbers are acidic, high numbers basic, 7 is neither).

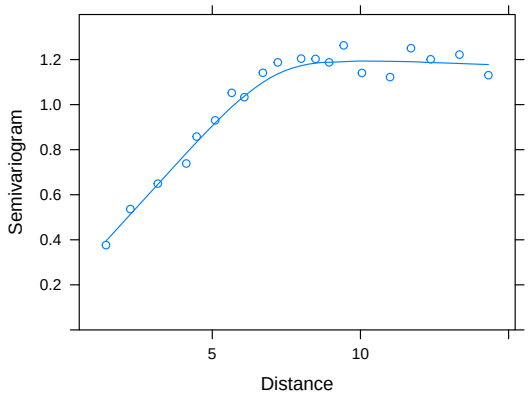
   (a) Make a scatterplot of the two variables and fit a model for Ca based on pH. (Choose the form of the model based on the scatterplot.) Print the estimated coefficients and discuss the relationship.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.8657	0.4546	-12.90	0.0000
pH	1.5826	0.0789	20.05	0.0000

I chose to model calcium concentration as a linear function of pH. I didn't see much curvature in the scatterplot and adding squared or cubed terms did not improve the fit of the model. For a one unit increase in pH, the calcium concentration is estimated to increase by 1.583 *ppm. This model is appropriate for the range of pH's seen in the sample, but I wouldn't use it for extrapolation.*

   (b) Plot the semivariogram of the residuals using euclidean distance on column and row. Does it appear that there is some spatial correlation? Make a guess about values for range and nugget.



There does appear to be some spatial correlation because the empirical semivariogram increases with distance. The nugget effect appears to be about 0.4*, the sill seems to be about* 1.2*, and the range seems to be about* 7*.*

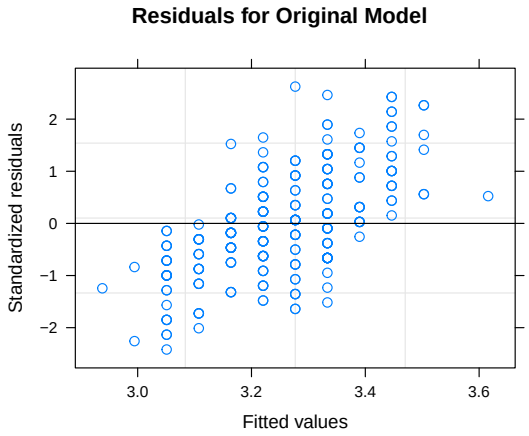
(c) Fit the five forms of spatial correlation available in the nlme library. Compare them with each other and with the original model. Do any of the spatial correlation fits improve AIC by more than 2 units? Are any of them "significant" improvements according to a LRT?

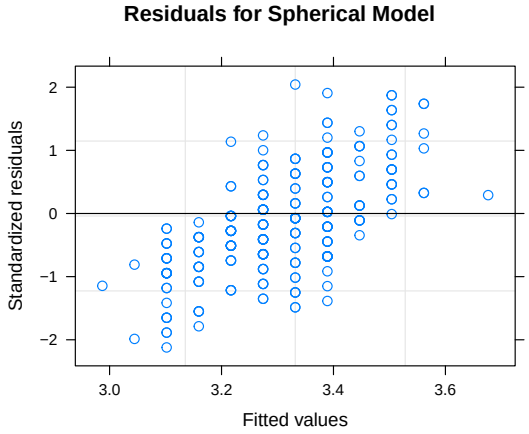
```
               Model df       AIC       BIC    logLik
gls.soil1          1  2 199.40574 206.44064 -97.70287
soil.glsSpher      2  4 -98.20318 -84.13337  53.10159
soil.glsRatio      3  4 -95.35160 -81.28179  51.67580
soil.glsLin        4  4 -46.11012 -32.04031  27.05506
soil.glsExp        5  4 -97.44184 -83.37203  52.72092
soil.glsGauss      6  4 -85.22073 -71.15092  46.61037

               Model df       AIC       BIC    logLik   Test  L.Ratio p-value
gls.soil1          1  2 199.40574 206.44064 -97.70287
soil.glsSpher      2  4 -98.20318 -84.13337  53.10159 1 vs 2 301.6089  <.0001
```

All of the models improve the AIC by more than 2 *units. All of them are significant improvements to the original, no spatial correlation model according to the likelihood ratio test. The spherical model seems to fit the best. It has the largest loglikelihood and the lowest AIC. The expoential model is a close second.*

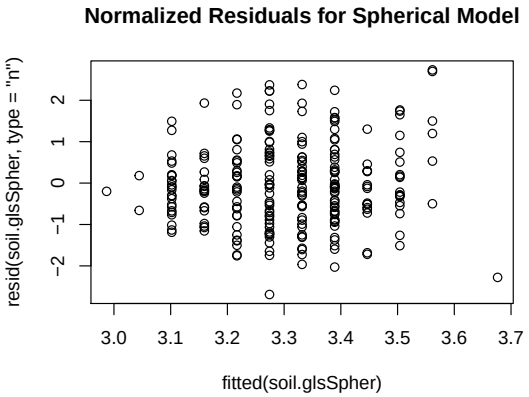
(d) Use the plot function on the first model with no spatial correlation and on the best of the spatial correlation models. Describe any problems you see.





The residuals appear to be the same or very similar. I suppose this makes sense because although the spherical model takes into account the spatial correlation structure, the fitted values have not changed. As a result, we see that the default residuals (Observed-Fitted Values) have not changed.

(e) Redo the second plot with normalized residuals instead of the default.



Read the help on `residuals.gls`. Write out a matrix equation to show how the normalized residuals are different from the default residuals.

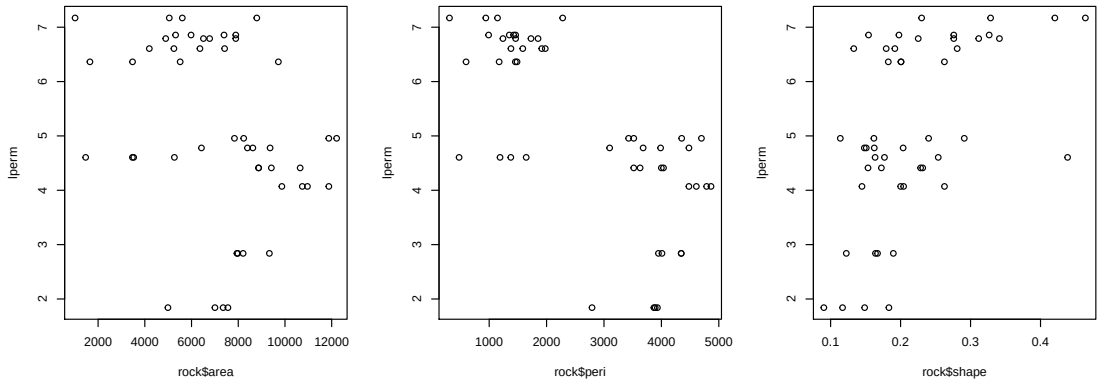
The default "response" residuals are $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

The normalized residuals are $\boldsymbol{\epsilon}_N = \sqrt{\Sigma^{-1}}\boldsymbol{\epsilon}^s$. Σ *is the estimated error correlation*

matrix and $\boldsymbol{\epsilon}^s$ is a vector of standardized residuals. These residuals take the correlation structure into account. That's why we no longer see the trend that we saw in the default residuals.

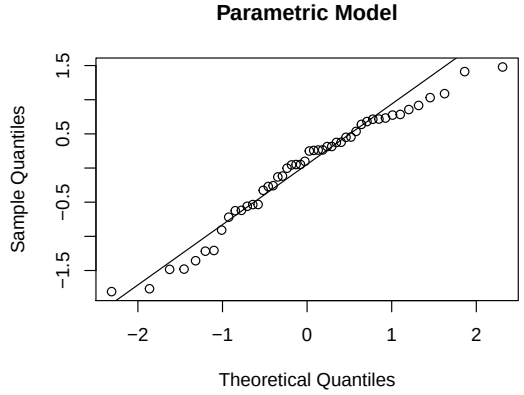
2. Examine the data on rock cores available in R as `data(rock)` (see help, too). The object is to model log(permeability) based on observations taken from image analysis of core cross sections. Explanatory variables are the total area, total perimeter, and shape of microscopic pores.

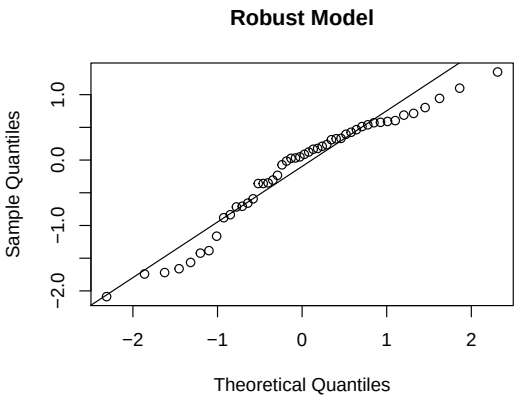
   (a) Plot each predictor against log(perm).



   (b) Fit a linear model and a robust linear model using all three predictors. Use qqnorm and qqline on the plain residuals.

   I chose to fit the first order model because I didn't find evidence of that any interaction terms were important. Below are the normal QQ plots for the parametric and bootstrap models.





(c) Due to some concern about normality, you are asked to give a non-normality-based estimate of the coefficient for shape.

i. Discuss whether model-based or case-based bootstrapping would be most appropriate in this setting. (Here each core is considered a random draw from some population of cores of interest. Ignore that fact that these observations are really repeated measurements since each core was measured optically 4 times, but the permeability only once.)
Case based bootstrapping is more appropriate because the area, perimeter, and shape of each core is random and can be considered a random draw from a multivariate distribution. We will resample 48 *cores from the original sample of cores to create the bootstrap distribution, keeping the area, perimeter, and shape of each core together.*

ii. Provide a 95% bootstrap CI.

```
Bootstrap Statistics :
         original        bias     std. error
t1*  5.3331449907 -0.2897861912 1.3439996069
t2*  0.0004849794 -0.0004860404 0.0001919930
t3* -0.0015266130  0.0015319319 0.0004241336
t4*  1.7565260120 -1.5108482975 4.5833694849

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 99 bootstrap replicates

CALL :
boot.ci(boot.out = rock.boot, index = 4)

Intervals :
Level      Normal              Basic
95\%   (-3.488, 10.993 )   (-3.612, 11.152 )

Level     Percentile            BCa
```

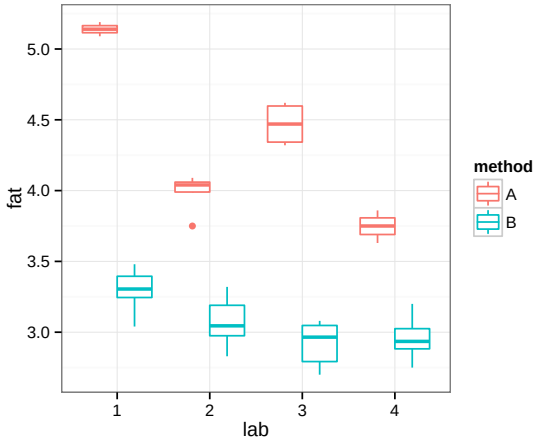
```
95\%   (-7.639,  7.125 )   (-3.826,  7.916 )
```

In the parametric model, a one unit increase in the shape parameter is estimated to be associated with a multiplicative change of 5.79 *in the permeability with a* 95% *confidence interval from* 0.168 *to* 199.41*.*

In the bootstrap model, a one unit increase in the shape parameter is estimated to be associated with a multiplicative change of 5.79 *in the permeability with a* 95% *confidence interval from* 0.022 *to* 2740.79*.*

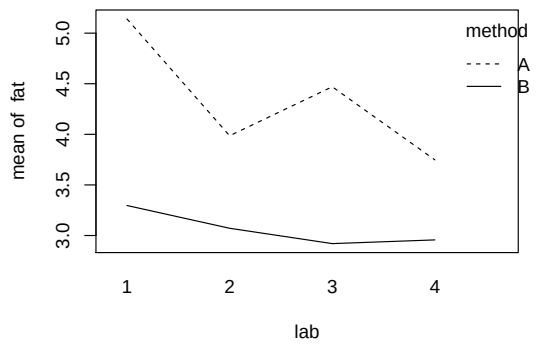
We can see that the bootstrap confidence interval is wider than the parametric confidence interval, although neither interval suggests that permeability depends on the shape of the core.

3. In a study of milkfat in yogurt, samples known to have 3% milkfat were sent to four labs, and the labs were told to use either method A or B (assigned at random) to measure the milkfat. The primary concern is that method A might give higher readings than method B. We need to look for an interaction between method and lab.

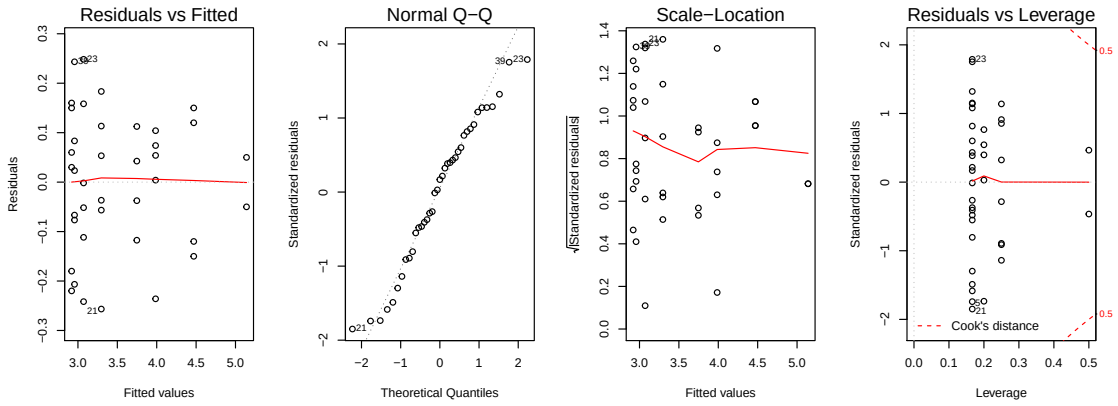


(a) Fit a 2-way ANOVA model with interaction and discuss the results (including diagnostics).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.1400	0.1075	47.81	0.0000
lab2	-1.1540	0.1272	-9.07	0.0000
lab3	-0.6700	0.1317	-5.09	0.0000
lab4	-1.3925	0.1317	-10.58	0.0000
lab1:methodB	-1.8433	0.1241	-14.85	0.0000
lab2:methodB	-0.9143	0.0921	-9.93	0.0000
lab3:methodB	-1.5500	0.0981	-15.79	0.0000
lab4:methodB	-0.7908	0.0981	-8.06	0.0000



There is strong evidence that the difference in measured milkfat between methods A and B depends on the laboratory (p=value< 0.0001 from F-stat=22.194 on 3 and 31 df). Within lab 1, method B is estimated to measure milkfat 1.84 percentage points lower than method A, with a 95% confidence interval from 2.10 to 1.59 percentage points lower. In lab 2, method B is estimated to measure milkfat 0.91 percentage points lower than method A, with a 95% confidence interval from 1.10 to 0.73 percentage points lower. The method effects in labs 3 and 4 are shown in the summary above. Also above is an interaction plot that illustrates the relationship between methods A and B for each lab. Below are the residual plots for this linear model. There is no obvious trend in the plot of standardized residuals and the points appear to follow the diagonal line on the normal QQ plot. I am not worried about the model assumptions.



(b) To avoid the normality assumption, we will also use a bootstrap approach. Decide whether case-based or model-based bootstrapping is more appropriate and explain why.

In this situation, I think model based bootstrapping is more appropriate because we are only interested in the two methods in the study, Methods A and B. The labs could be chosen from a larger population of labs, but model based bootstrapping is still appropriate because even if a different four labs were chosen in a subsequent experiment, the design matrix would still be the same.

(c) Use your chosen bootstrap method to test to

i. see if the interaction terms are needed? Use an approach similar to the extra sum of squares F test, but via bootstrapping.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.04    1.31    2.40    2.95    4.05   11.40
##      [,1]
## [1,] 66.6
```

The F-statistic for the original sample is much larger than the largest boot F-statistic, so there is evidence that the interaction terms are important (p-value< 0.002*).*

ii. see if the interaction is ignorable, refit without it.
The interaction is not ignorable.

iii. provide a confidence interval estimate of the Method effect after adjusting out lab effects.
The following are the bootstrap confidence intervals for the method effect for each lab. In all four labs, there is evidence that the mean measured percent of milkfat for method A is different than the mean measured percent of milkfat for method B. The probably need to refine their methods!

Method Effect(B-A)	BCa interval
Lab 1	(-2.107, -1.575)
Lab 2	(-1.1161, -0.7398)
Lab 3	(-1.733, -1.393)
Lab 4	(-0.9695, -0.6096)

Scope of Inference: Since the milk samples studied were not randomly selected, we cannot infer results beyond these milk samples. Since milk samples were randomly assigned to measurement method A or B, we can infer that the method caused the change in milkfat measurement.
Suppose we think of the four labs in the study as being randomly chosen from a larger population of labs. Since we found evidence of a method effect in all four labs in the study, we can infer that there is a method effect in the population of labs.

R Code

```
soil <-  read.table("~/Documents/Stat505/Homework/HW6/soils.dat", head = TRUE)
```

```
require(ggplot2)
curve.fit <- function(x)(-5.86569+1.58257*x)
qplot(x=soil$pH,y=soil$Ca, main="Soil Calcium vs Soil pH")+stat_function(fun = curve.fit
lm.fit <- lm(Ca~pH, data=soil)
require(xtable)
xtable(summary(lm.fit))
```

```
require(nlme)
gls.soil1 <- gls(Ca~pH -1, data=soil)
plot(Variogram(gls.soil1, form = ~ column+row, max=15), ylim=c(0, 1.4))
```

```
soil.glsSpher <- update(gls.soil1, corr=corSpher(c(7,.4), form = ~column+row, nugget=T))
soil.glsRatio <- update(gls.soil1, corr=corRatio(c(3,0.4), form=~column+row, nugget=T))
soil.glsLin = update(gls.soil1, corr=corLin(c(7,0.4), form=~column+row, nugget=T))
soil.glsExp = update(gls.soil1, corr=corExp(c(7,0.4), form=~column+row, nugget=T))
soil.glsGauss = update(gls.soil1, corr=corGaus(c(7,0.4), form=~column+row, nugget=T))
anova(gls.soil1, soil.glsSpher, soil.glsRatio, soil.glsLin, soil.glsExp, soil.glsGauss,
anova(gls.soil1, soil.glsSpher)
```

```
soil.glsSpher <- update(gls.soil1, corr=corSpher(c(7,.4), form = ~column+row, nugget=T))
plot(gls.soil1, main="Residuals for Original Model")
```

```
plot(soil.glsSpher, main="Residuals for Spherical Model")
```

```
data(rock)
lperm <- log(rock$perm)
par(mfrow=c(1,3))
plot(lperm~rock$area)
plot(lperm~rock$peri)
plot(lperm~rock$shape)
```

```
lm.rock <- lm(lperm~area+peri+shape, data=rock)
qqnorm(resid(lm.rock), main="Parametric Model")
qqline(resid(lm.rock))
```

```
require(MASS)
lm.rock.robust <- rlm(lperm~area+peri+shape, data=rock, method="MM", init="lts", maxit=5
qqnorm(resid(lm.rock.robust), main="Robust Model")
qqline(resid(lm.rock.robust))
```

```
##case based
library(boot)
rock.fun <- function(data,i){coef(lm(lperm~area+peri+shape,data=data[i,]))}
rock.boot <- boot(rock, rock.fun, R=99)
#boot.ci(rock.boot, index=4)
```

```
milkfat <- read.table("~/Documents/Stat505/Homework/HW6/milkfat.dat",head=TRUE)
require(ggplot2)
milkfat$lab <- factor(milkfat$lab)
milkfat$lab <- factor(milkfat$lab, levels=c(1,2,3,4))
qplot(x=lab, y=fat, data=milkfat, colour=method, geom = "boxplot") + theme_bw()
```

```
milk.lm <- lm(fat~lab+method:lab, data=milkfat)
require(xtable)
xtable(summary(milk.lm))
with(milkfat, interaction.plot(lab, method, fat))
```

```
par(mfrow=c(1,4))
plot(milk.lm)
```

```
#model based
milkfit <- lm(fat~lab*method, data=milkfat)
XtXinv <- summary(milkfit)$cov.unscaled
X <- model.matrix(milkfit)
Ct <- matrix(c(0,0,0,0,0,1,0,0, 0,0,0,0,0,0,1,0, 0,0,0,0,0,0,0,1),3,8, byrow=T)
#Ct ##contrast to test beta5=beta6=beta7=0
middle <- X%*%XtXinv %*% t(Ct) %*% solve(Ct%*%XtXinv%*%t(Ct)) %*% Ct %*% XtXinv %*% t(X)
fakeF <- function(e,i,middle = middle){
## e = residuals
## i = indices to use
em <- e[i] ## em is the current resample from e
n <- length(e)
sSqd <- var(em)/(n-1) * n
em %*% middle %*% em/sSqd }
bootFs <- rep(0,499) ## set up a vector for storage.
for(i in 1:499) bootFs[i] <- fakeF(rstudent(milkfit)*summary(milkfit)$sigma,sample(39,re
### sample picks a random sample uses integers 1 to 39 with replacement
summary(bootFs)
Cb <- Ct%*% coef(milkfit)
testF <- t(Cb)%*% solve(Ct%*%XtXinv%*%t(Ct)) %*% Cb / summary(milkfit)$sigma^2
testF
#66.58
# This is way bigger than the largest bootF, so reject H0: beta5=beta6=beta7=0
# at alpha = 1/500.
```

```
library(boot)
milk1 <- lm(fat ~ lab + method:lab, data=milkfat)
milk.fun <- function(data,i){
  y.star=fitted(milk1)+data[i]
```

```
  coef(lm(y.star~lab+method:lab, data=milkfat))}
milk.boot <- boot(rstudent(milk1)*summary(milk1)$sigma, milk.fun, R=99)
#boot.ci(milk.boot, index=5)
#boot.ci(milk.boot, index=6)
#boot.ci(milk.boot, index=7)
#boot.ci(milk.boot, index=8)
```