Applied Bayesian Analysis : NCSU ST 540

Homework 1

Bruce Campbell

---

1) ESP Paper Commentary

Read the paper "Miller G (2011). ESP paper rekindles discussion about statistics. Science,21, 272-273" and write a one-paragraph summary.

In the Miler paper (Miller 2011) an argument is put forth that the use of null hypothesis significance testing [NHST] in the social sciences comes with risk of misinterpretation and provides a basis upon which spurious results may be published. The authors are discussing a recent publication (Bem 2011) purporting to have found statistically significant evidence in favor of ESP. It is (correctly) argued that the misinterpretation of what a p-value is and the meaning of statistically significant results led the author to make dubious conclusions about ESP. A call is made to the scientific community to bring Bayesian analysis back to the social sciences. It's argued that the ability to include prior information into model parameters, to calculate probabilities on the model parameter space, and to update Bayesian models with new data make Bayesian analysis easier to interpret and less prone to misinterpretation. The main objection to Bayesian analysis regarding the subjectivity of the prior is acknowledged in the paper. Of course Bayesian analysis is subject to being abused as well as NHST it's just harder to hide the abuse when using Bayesian analysis. The main argument that causality and reason be part of any statistical analysis is long overdue. Extraordinary claims require extraordinary analysis.

2) Ozone Data Analysis

Load the ozone data from the course webpage. This dataset has daily ozone values for $n = 1106$ monitoring stations and $m = 31$ days in July 2005.

```
rm(list = ls())
setwd("C:/E/brucebcampbell-git/bayesian-learning-with-R")
df <- read.csv("ozone.csv",header = TRUE)
```

(a) Create a table with the overall (across sites and days) mean, standard deviation, and

percent missing.

```
#Remove the id column and collapse the remaining columns into a single one.
df.data <- df
df.data$Station.ID <- NULL

raw.data <-unlist(df.data)
```

```r
mean.raw <- mean(raw.data,na.rm = TRUE)
sd.raw <- sd(raw.data,na.rm = TRUE)
nan.count.raw <- sum(is.na(raw.data))
nan.proportion.raw <- nan.count.raw / length(raw.data)

pander(data.frame(mean = mean.raw,
                  sd = sd.raw,
                  percent.missing =nan.proportion.raw*100), "ozone side-days merged summary sta
```

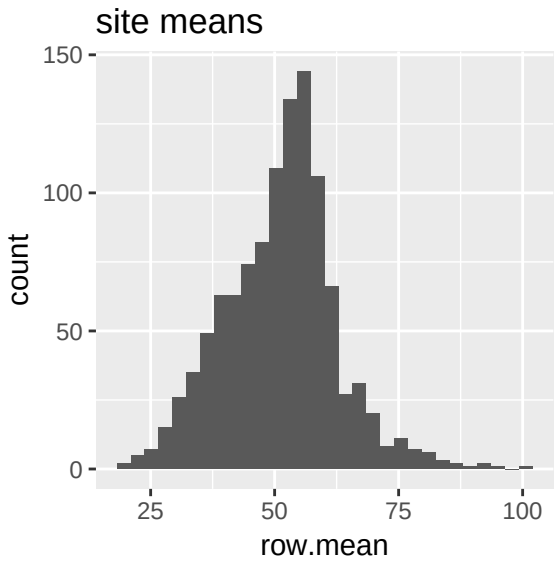
Table 1: ozone side-days merged summary stats

mean	sd	percent.missing
51.27	17.26	4.322

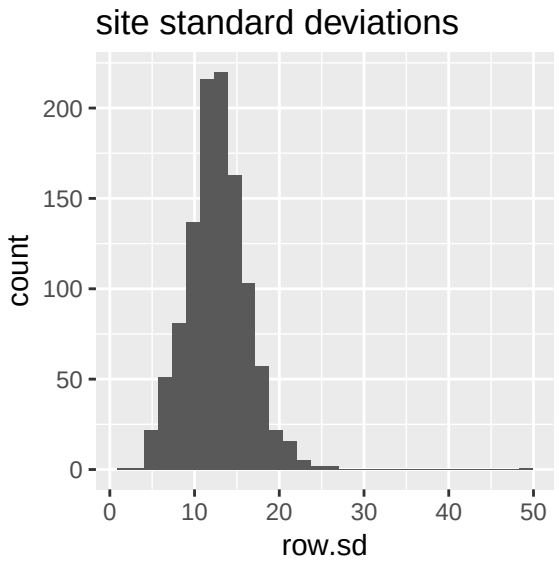
(b) Write a loop to compute the mean, variance, and percent missing for each of the n sites;

make a histogram of each variable (all three histograms should have n observations); create scatter plots of each pair of these variables (each of the three plots should have n points).

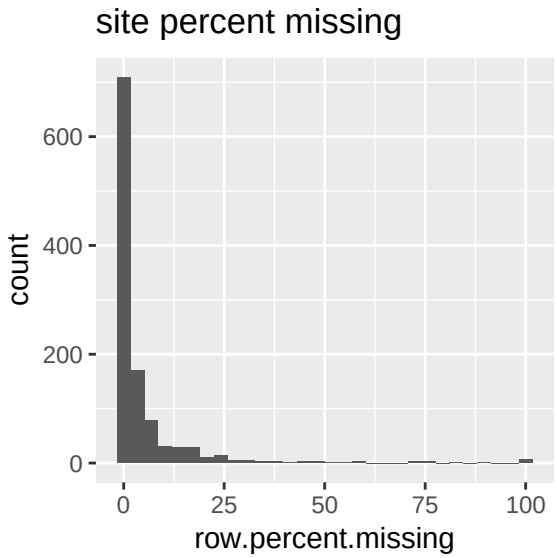
```r
#df.data is the data without the station id variable.
row.mean <- matrix(nrow = 1,ncol = nrow(df.data))
row.sd <- matrix(nrow = 1,ncol = nrow(df.data))
row.percent.missing <- matrix(nrow = 1,ncol = nrow(df.data))
for( i in 1:nrow(df.data))
{
  row.data <- df.data[i,]

  row.mean.no.nan <- df.data[i,is.na(df.data[i,])==FALSE]
  row.mean[1,i]<- mean(as.numeric(row.mean.no.nan))
  row.sd[1,i] <- sd(row.data,na.rm = TRUE)

  row.nan.count.raw <- sum(is.na(row.data))
  row.percent.missing[1,i] <- 100* row.nan.count.raw / length(df.data)
}

df.row.summary <- data.frame(row.mean=row.mean[1,],
                             row.sd=row.sd[1,],
                             row.percent.missing=row.percent.missing[1,])

ggplot(df.row.summary,aes(x=row.mean))+
  geom_histogram()+ggtitle("site means")
```




```
ggplot(df.row.summary,aes(x=row.sd))+
  geom_histogram() +ggtitle("site standard deviations")
```



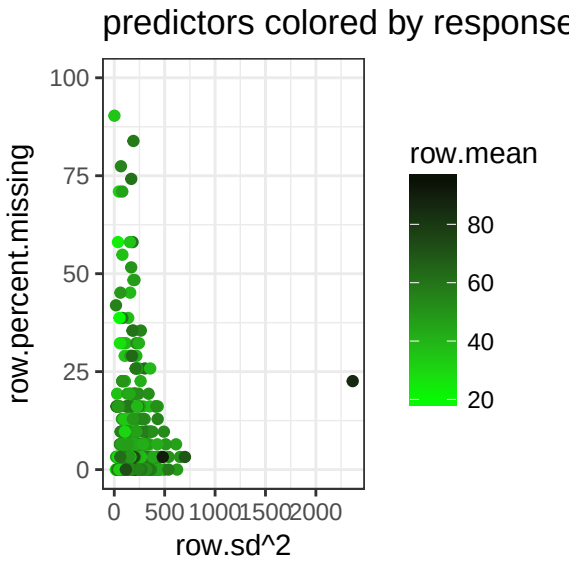

```
ggplot(df.row.summary,aes(x=row.percent.missing))+
  geom_histogram() +ggtitle("site percent missing")
```



(c) Conduct a linear regression with response equal to the site's mean and the site's variance

and percent missing as covariates

```
#First we plot the data
p <- ggplot(df.row.summary, aes(x = row.sd^2, y = row.percent.missing))+
  geom_point(aes(col = row.mean))
p + scale_colour_gradientn(colours=c("green","black")) +
  theme_bw()+ggtitle("predictors colored by response")
```




```
#Fit the model
lm.fit <- lm(row.mean ~ row.sd^2 + row.percent.missing, data = df.row.summary)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = row.mean ~ row.sd^2 + row.percent.missing, data = df.row.summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.839  -7.561   0.719   6.312  44.544
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         44.21631    1.23932  35.678  < 2e-16 ***
## row.sd               0.57628    0.09275   6.213 7.37e-10 ***
## row.percent.missing -0.09602    0.03607  -2.662  0.00787 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.15 on 1097 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.04121,    Adjusted R-squared:  0.03946
## F-statistic: 23.57 on 2 and 1097 DF,  p-value: 9.451e-11
```

```
#Calculate residual correlation
u<- residuals(lm.fit)[-length(residuals(lm.fit))]
v <- residuals(lm.fit)[-1]
cor.residuals <- cor(v,u)
pander(data.frame(cor.residuals=cor.residuals), caption = "residual correlation")
```

Table 2: residual correlation

cor.residuals
0.62

```
#Cookes distance and leverage plots.
#plot(lm.fit,which =4)
#plot(lm.fit,which = 5)
```

‘‘‘

Bibliography

Bem, D.J. 2011. “Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect.” *Journal of Personality and Social Psychology*. doi:doi: 10.1037/a0021524.

Miller, Greg. 2011. “ESP Paper Rekindles Discussion About Statistics.” *Science* 331 (6015). American Association for the Advancement of Science: 272–73. doi:10.1126/science.331.6015.272.