A Simple Example Sweave Report

Elizabeth G. Hill
Hollings Cancer Center Biostatistics Shared Resource

April 10, 2012

1 Airquality data desciption

The data set `airquality` provides air quality measurements in New York from May to September in 1973. We wish to investigate the association between month of the year and ozone levels. The median ozone level for the five-month period was 31.5 (interquartile range = 18 to 63.25). Average ozone for the five month period was 42.13 (SD = 32.99). Figure 1 shows the distribution of ozone is positively skewed. (Of course, we could tell that anyway by inspection of the mean and standard deviation.) Figure 2 shows a quantile-quantile plot of ozone indicating the data are not normally distributed. Therefore, we will use a non-parametric approach to assess the association between ozone level and month.

2 Association of ozone with month

Figure 3 shows boxplots of ozone by month.

2.1 Analysis using Kruskal-Wallis test

```
> kw.out <- kruskal.test(Ozone ~ Month, data = airquality)
> kw.out

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 6.901e-06

> pval <- round(kw.out$p.value, 6)
```

Based on the results of the Kruskal-Wallis test, ozone differs significantly by month (p = 7e-06).

2.2 Analysis using ANOVA

Let's revisit this analysis using ANOVA. We'll perform a logarithmic transformation of the ozone values to facilitate transformation to approximate normality.

Figure 4 shows histogram and quantile-quantile plots of log-transformed ozone values. Based on these empirical results, we are comfortable with approximate normality and proceed with inference using ANOVA.

```
> lm.out <- lm(log(Ozone) ~ as.factor(Month), data = airquality)
> summary(lm.out)
```

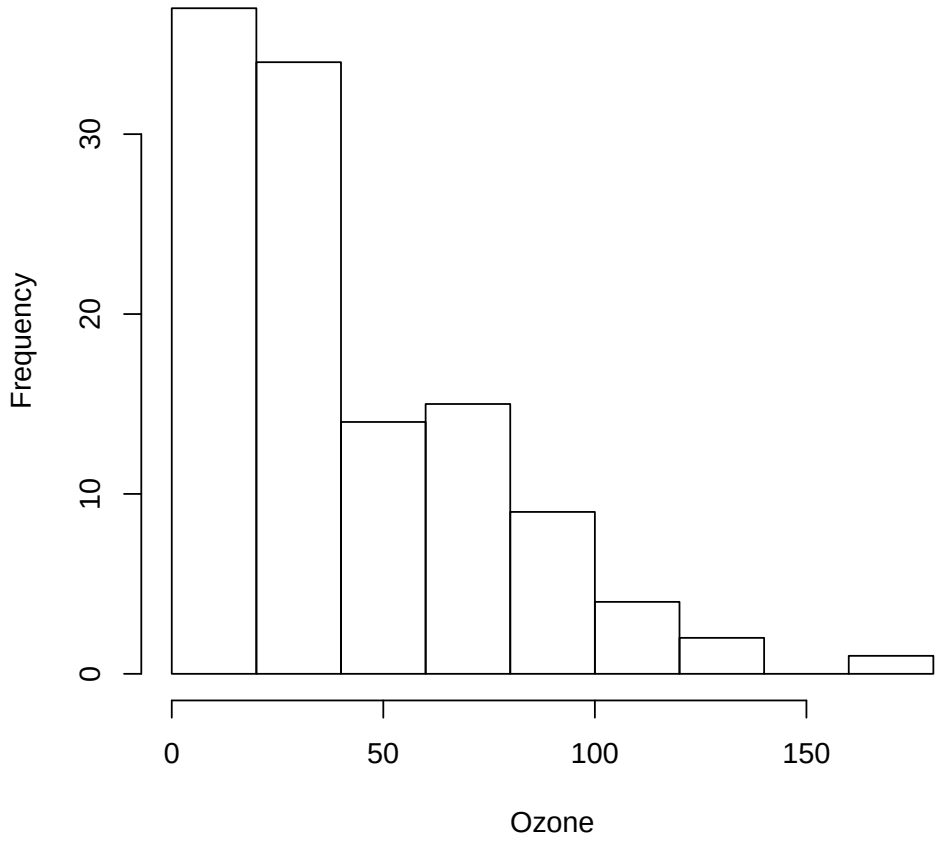

Figure 1: Histogram of ozone.

```
Call:
lm(formula = log(Ozone) ~ as.factor(Month), data = airquality)

Residuals:
     Min       1Q   Median       3Q      Max
-2.81208 -0.41418  0.01809  0.52082  1.93286

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.8121     0.1498  18.773  < 2e-16 ***
as.factor(Month)6   0.4246     0.2954   1.437   0.1534
as.factor(Month)7   1.0718     0.2118   5.059 1.68e-06 ***
as.factor(Month)8   1.0333     0.2118   4.878 3.61e-06 ***
as.factor(Month)9   0.4067     0.2063   1.972   0.0511 .
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
```

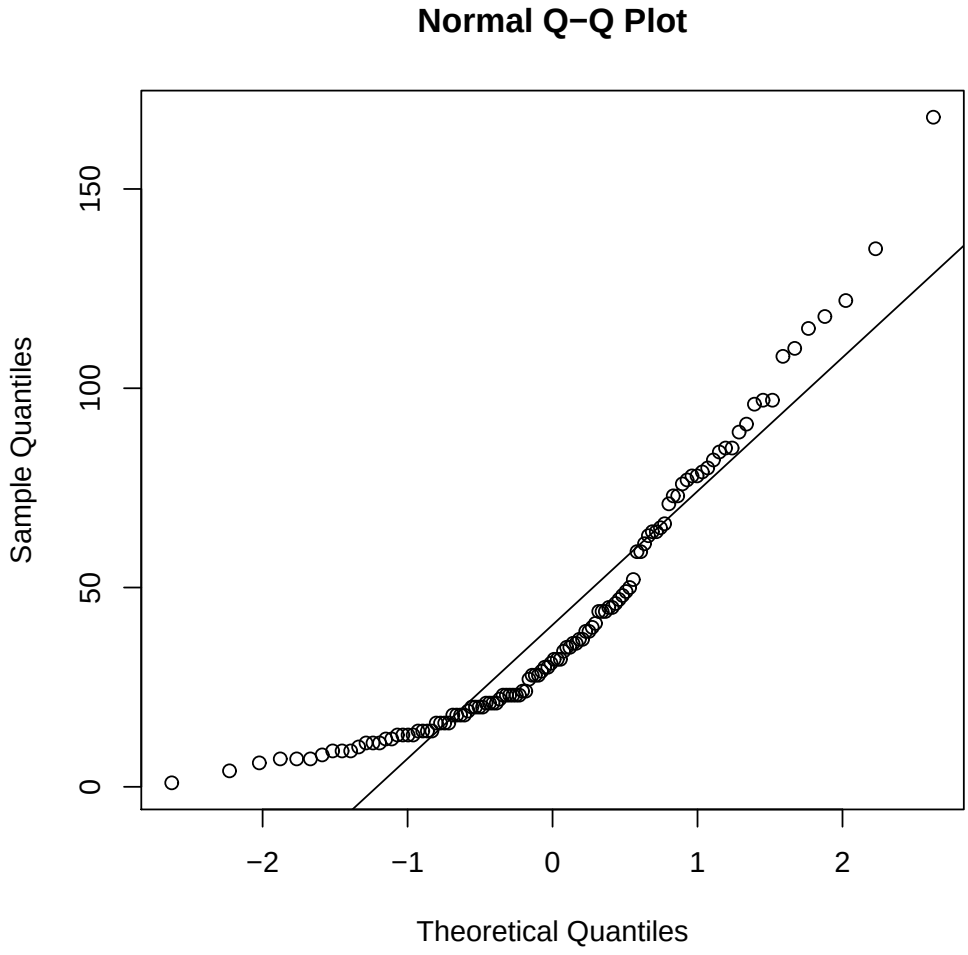

Figure 2: Quantile-quantile plot of ozone.

```
Residual standard error: 0.7638 on 111 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared: 0.2482,        Adjusted R-squared: 0.2211
F-statistic: 9.163 on 4 and 111 DF,  p-value: 1.96e-06
```

By default, comparisons are performed relative to May ozone levels, the first level of the factor. Suppose we'd like to perform all pairwise comparisons. Table 1 shows p-values associated with all pairwise comparisons of ozone levels across months.

3 Source code for this report

```
\documentclass[11pt]{article}
%TO RUN THIS FILE, WITHIN R USE THE COMMAND
%>Sweave("Sweave_example.Rnw")
\usepackage{url}
```

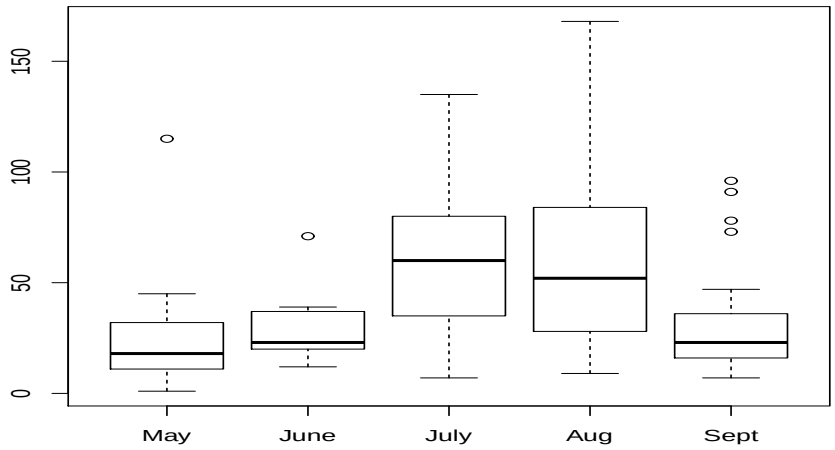

Figure 3: Boxplot of ozone by month of year.

Table 1: All pairwise comparisons.

Month	May	June	July	August	September
May		0.153	0	0	0.051
June	0.153		0.031	0.042	0.042
July	0	0.031		0.856	0.002
August	0	0.042	0.856		0.003
September	0.051	0.042	0.002	0.003	

```
\usepackage{moreverb}
%One common problem that people encounter is that the Sweave.sty  file
%is missing. It is located at ${RHOME}/share/texmf/Sweave.sty; just save
%it in the same directory as your LaTeX document.
\textwidth 6.75in
\textheight 9.25in
\topmargin -0.875in
\oddsidemargin -.125in
\evensidemargin -.125in

\SweaveOpts{prefix.string=CFRIeg1}
\SweaveOpts{eps=TRUE,pdf=FALSE}

\title{A Simple Example Sweave Report}
\author{Elizabeth G. Hill \\ Hollings Cancer Center Biostatistics Shared Resource}
\date{\today}
\begin{document}
```

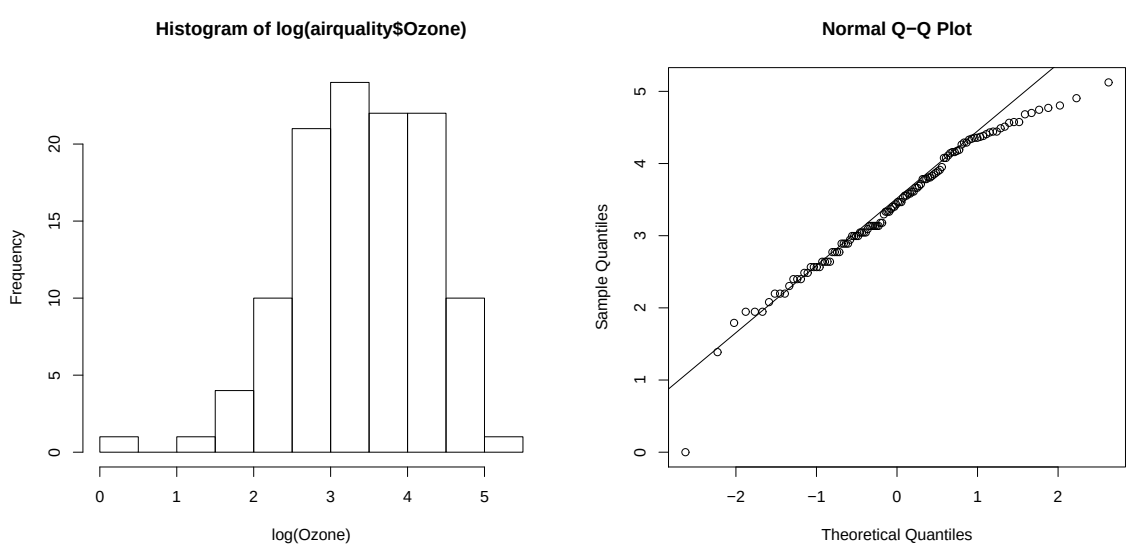

Figure 4: Histogram of log transformed ozone.

```
\maketitle

\section{Airquality data desciption}
The data set \texttt{airquality} provides air quality measurements
in New York from May to September in 1973.  We wish to investigate
the association between month of the year and ozone levels.
The median ozone
level for the five-month period was \Sexpr{median(airquality$Ozone,na.rm=TRUE)}
(interquartile range = \Sexpr{quantile(airquality$Ozone,probs=0.25,na.rm=TRUE)}
to \Sexpr{quantile(airquality$Ozone,probs=0.75,na.rm=TRUE)}).
Average ozone for the five month
period was \Sexpr{round(mean(airquality$Ozone,na.rm=TRUE),2)}
(SD = \Sexpr{round(sd(airquality$Ozone,na.rm=TRUE),2)}).
Figure \ref{ozhist} shows the distribution of ozone is positively skewed.
(Of course, we could tell that anyway by inspection of the
mean and standard deviation.)  Figure \ref{ozqq} shows a quantile-quantile
plot of ozone indicating the data are not normally distributed.
Therefore, we will use a non-parametric approach to assess the association
between ozone level and month.

\begin{figure}[ht]
\begin{center}
<<ozonehist, echo=FALSE, fig=TRUE>>=
hist(airquality$Ozone,main="",xlab="Ozone")
@
\caption{Histogram of ozone.}
\label{ozhist}
\end{center}
\end{figure}
```

```
\begin{figure}[ht]
\begin{center}
<<ozoneqq, echo=FALSE, fig=TRUE>>=
qqnorm(airquality$Ozone)
qqline(airquality$Ozone)
@
\caption{Quantile-quantile plot of ozone.}
\label{ozqq}
\end{center}
\end{figure}

\section{Association of ozone with month}

<<ozboxplots,echo=FALSE,fig=TRUE,include=FALSE>>=
boxplot(Ozone ~ Month, data = airquality,xaxt="n")
axis(side=1,at=1:5,labels=c("May","June","July","Aug","Sept"))
@
Figure \ref{newozbox} shows boxplots of ozone by month.

\begin{figure}[ht]
\begin{center}
\includegraphics[width=5in,height=3in]{CFRIeg1-ozboxplots}
\caption{Boxplot of ozone by month of year.}
\label{newozbox}
\end{center}
\end{figure}

\subsection{Analysis using Kruskal-Wallis test}

<<kw>>=
kw.out <- kruskal.test(Ozone ~ Month, data=airquality)
kw.out
pval <- round(kw.out$p.value,6)
@
Based on the results of the Kruskal-Wallis test, ozone differs
significantly by month (p = \Sexpr{pval}).

\subsection{Analysis using ANOVA}

Let's revisit this analysis using ANOVA.  We'll perform a logarithmic
transformation of the ozone
values to facilitate transformation to approximate normality.

<<histtransf,echo=FALSE,fig=TRUE,include=FALSE>>=
hist(log(airquality$Ozone),xlab="log(Ozone)")
@

<<qqtransf,echo=FALSE,fig=TRUE,include=FALSE>>=
```

```
qqnorm(log(airquality$Ozone))
qqline(log(airquality$Ozone))
@

Figure \ref{transf} shows histogram and quantile-quantile plots of log-transformed
ozone values.  Based on these empirical results, we are comfortable with approximate
normality and proceed with inference using ANOVA.

\begin{figure}[ht]
\begin{center}
\includegraphics[width=3in,height=3in]{CFRIeg1-histtransf}
\includegraphics[width=3in,height=3in]{CFRIeg1-qqtransf}
\caption{Histogram of log transformed ozone.}
\label{transf}
\end{center}
\end{figure}

<<>>=
lm.out <- lm(log(Ozone)~as.factor(Month),data=airquality)
summary(lm.out)
@

By default, comparisons are performed relative to May ozone levels,
the first level of the factor.  Suppose we'd like to perform all
pairwise comparisons.  Table \ref{pairwise} shows p-values associated
with all pairwise comparisons of ozone levels across months.

<<contrasts,echo=FALSE,results=hide>>=
#MAY COMPARISONS
p.may.june <- summary(lm.out)$coefficients[2,4]
p.may.july <- summary(lm.out)$coefficients[3,4]
p.may.aug <- summary(lm.out)$coefficients[4,4]
p.may.sep <- summary(lm.out)$coefficients[5,4]
#JUNE COMPARISIONS
library(gmodels)
contrast.june.july <- estimable(lm.out, c(0, 1, -1, 0, 0))
p.june.july <- contrast.june.july$Pr

contrast.june.aug <- estimable(lm.out, c(0, 1, 0, -1, 0))
p.june.aug <- contrast.june.aug$Pr

contrast.june.sep <- estimable(lm.out, c(0, 1, 0, 0, -1))
p.june.sep <- contrast.june.aug$Pr
#JULY COMPARISONS
contrast.july.aug <- estimable(lm.out, c(0, 0, 1, -1, 0))
p.july.aug <- contrast.july.aug$Pr

contrast.july.sep <- estimable(lm.out, c(0, 0, 1, 0, -1))
p.july.sep <- contrast.july.sep$Pr
```

```
#AUGUST COMPARISONS
contrast.aug.sep <- estimable(lm.out, c(0, 0, 0, 1, -1))
p.aug.sep <- contrast.aug.sep$Pr
@

\begin{table}
\caption{All pairwise comparisons.}
\begin{center}
\begin{tabular}{rccccc} \\ \hline
Month &May &June &July &August &September \\ \hline
May &&\Sexpr{round(p.may.june,3)} &\Sexpr{round(p.may.july,3)}
&\Sexpr{round(p.may.aug,3)} &\Sexpr{round(p.may.sep,3)} \\
June &\Sexpr{round(p.may.june,3)} &&\Sexpr{round(p.june.july,3)}
&\Sexpr{round(p.june.aug,3)} &\Sexpr{round(p.june.sep,3)}\\
July &\Sexpr{round(p.may.july,3)} &\Sexpr{round(p.june.july,3)}
&&\Sexpr{round(p.july.aug,3)} &\Sexpr{round(p.july.sep,3)}\\
August &\Sexpr{round(p.may.aug,3)} &\Sexpr{round(p.june.aug,3)}
&\Sexpr{round(p.july.aug,3)} &&\Sexpr{round(p.aug.sep,3)}\\
September &\Sexpr{round(p.may.sep,3)} &\Sexpr{round(p.june.sep,3)}
&\Sexpr{round(p.july.sep,3)} &\Sexpr{round(p.aug.sep,3)} & \\ \hline
\end{tabular}
\end{center}
\label{pairwise}
\end{table}

\section{Source code for this report}
\verbatimtabinput{Sweave_example.Rnw}

\end{document}
```