SAS and R for Statistical Analysis of Large Data Examples

J. D. Tubbs
Department of Statistical Science

Fall 2021
Spring 2022

Contents

Chapter 1

Introduction - Large Data Case Studies

This document contains a description of the larger "real-world" data sets that will be used in support of the Statistical Methods Lecture Notes for Stat 5380 and Stat 5381.

I have included five data sets. They are:

- Observational Data Set – Obesity in Texas preschool aged children.
- Observational Data Set – Exam results for certification.
- Observational Data Set – Consumer products.
- Clinical Trial – Placebo controlled Study for Urinary Incontinence.
- Observation Data Set – Body fat

1.1 Texas preschool aged Obesity

Piziak et. al. (2009)[1] investigated age/gender adjusted BMI for 18,462 children who participated in the Head Start program, which is funded and administered by the US Department of Health and Human Services Administration for Children and Families, from Fall 2003 through Spring 2008. Specifically, data were collected from Head Start centers in several South Texas border counties and one Central Texas county. The data from this study are used in two ways; results are compared to the cohort of the NHANES sample consisting of 2-5 year old children, presented by Ogden[2] and the prevalence of high BMI among pre-school children in South Texas exceeds that of the 2000 CDC growth curves is examined to determine if there are regional differences between the border counties of South Texas and a central Texas non-border county. The results suggest that prevalence estimates for high BMI children in the predominantly latino population exceeds those obtained by Ogden using the Mexican American subset of the 2-5 year cohort within the NHANES sample. Furthermore, the analysis suggest that there are some regional differences among the prevalence of high BMI between border and non-border counties in Texas.

This data set was expanded by adding data for another central Texas county – McLennan(Waco).

[1] Piziak, et. al.
[2] Ogden, et al.

References

1. Ogden CL, Carroll MG, Flegal KM. High body mass index for age among US children and adolescents, 2003-2006. *JAMA*. 2008; 299(20):2401-2405.

2. Pietrobelli A, Faith M, Allison D, Gallagher D, Chiumello G, Heymsfield, S. Body mass index as a measure of adiposity among children and adolescents : A validation study. *Journal of Pediatrics* 1998; 132(2);204-210.

3. Piziak, V., Morgan-Cox, M., and Tubbs, J. An Investigation of Obesity in Pre-School Children in South Texas

4. Kuczmarski, RJ et al. CDC growth charts for the United States: methods and development. *Vital Health Stat 11*. 2002;(246): 1-190.

5. Centers for Disease Control and Prevention National Health and Nutrition Examination Survey. http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/datafiles.htm. Accessed December 11, 2008.

1.1.1 Texas Preschool Aged Obesity Data Set

The SAS code for reading the data sets is:

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options nodate nonumber ps=200 ls=80 formdlim=' ';

   title 'Obesity - All Counties';
   title2 'Cleaned Combined Data ';
   title3 ' ';
/*
   There are two data sets
       1. CDC BMI age/gender adjusted by CDC 2000
       2. Combined Texas preschool data
*/

   * define work data for cdc data files;
   * the permanent dataset is sasuser.cdc_bmi;
data cdc; set sasuser.cdc_bmi;
if agemos=24 then agemos=23.9;
age=floor(agemos);
if age < 13 then N_AGE = 0;
if age > 12 and age < 25 then N_AGE = 1;
if age > 24 and age < 37 then N_AGE = 2;
if age > 36 and age < 49 then N_AGE = 3;
if age > 48 and age < 61 then N_AGE = 4;
if age > 60 then N_AGE = 5;
if sex=1 then Gender= 'M';
if sex=2 then Gender= 'F'; drop sex;
run;
proc contents data=cdc; run;

* define work data for texas_obesity data files;
* the permanent dataset is sasuser.new_combined_clean;
* year 2002 is removed due to the limited number of observations for that year;
DATA texas_obesity; SET ldata.new_combined_clean; IF YEAR > 2002;RUN;
```

```
proc contents data=ldata.texas_obesity;
run;
quit;
```

The SAS Datasets are;

```
cdc_bmi.sas7bdat
new_combined_clean.sas7bdat
texas_obesity.sas7bdat
```

1.2 Certification Exams

These data are from an international institute that provides certification training and examinations that are taken by specialist throughout the world. The exams are written in English but are taken by many candidates for certification from non-English speaking countries.

1.2.1 International Certification Data Set

These data are a subset of 4000 observations taken from a much larger data set (approx. 20,000). The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options center nodate pagesize=100 ls=80;

title 'Sample Certification Data';
data cert; set sasuser.certification;
run;

proc contents data=cert;
run;

quit;
```

The SAS data set is

```
certification.sas7bdat
```

1.3 Consumer Product Data

This data set involves a number of different consumer products. There will be several data sets and corresponding SAS programs available for the data analysis.

1.3.1 Kolache Sales

This example involves daily sales of frozen Kolaches at many locations for a large grocery outlet for a two week period. The number of units sold is the primary response variable. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options center nodate pagesize=100 ls=80;

title1 'Kolache Sales in Large Consumer Warehouses';
```

```
data kolache; set sasuser.kolache;
run;

proc contents data=kolache;
run;

quit;
```

The SAS data set is

```
kolache.sas7bdat
```

1.3.2 Egg Sales

This example involves monthly volume and price of eggs for a two year period at various distribution centers (DC) in the US. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options center nodate pagesize=100 ls=80;

title1 'Egg Sales';

data eggs; set sasuser.eggs;
run;

proc contents data=eggs;
run;

quit;
```

The SAS data set is

```
eggs.sas7bdat
```

1.4 Clinic Trial – Urinary Incontinence Study

This trial was a two-armed placebo controlled 12 week study where women with involuntary incontinence problems where given either an active treatment or a placebo. They were asked for the number of involuntary (not induced by laughter or coughing) incontinence events (for a specified period of time) at baseline and again after 12 weeks on their respective therapy (treatment). Additional measures were taken and the women were stratified based upon the number of incidence at baseline (severity of the problem). It is common for studies of this type to compute the percent change from baseline (in this case a negative number for effective therapy) as the response variable of interest. The literature indicates that the percent change from baseline is often highly skewed and as a result, non-parametric methods are used for the inference.

1.4.1 Urinary Incontinence Data

The data is a subset of a much larger data set. In these data one I have selected a random subset (with proportional allocation) of 1000 responses. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options center nodate pagesize=100 ls=80;
```

```
title1 'Urinary Incontinence Data';

data urinary; set sasuser.urinary;
run;

proc contents data=urinary;
run;

quit;
```

The SAS data set is

```
urinary.sas7bdat
```

1.5 Health Study - Body Fat

A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g. abdominal circumference) and/or skin-fold measurements. See, for instance, Behnke and Wilmore (1974), pp. 66-67; Wilmore (1976), p. 247; or Katch and McArdle (1977), pp. 120-132).

Percentage of body fat for an individual can be estimated once body density has been determined. Folks (e.g. Siri (1956)) assume that the body consists of two components - lean body tissue and fat tissue. Letting

```
D = Body Density (gm/cm^3)
A = proportion of lean body tissue
B = proportion of fat tissue (A+B=1)
a = density of lean body tissue (gm/cm^3)
b = density of fat tissue (gm/cm^3)
```

where

$$D = 1/[(A/a) + (B/b)]$$

solving for B we find

$$B = (1/D) * [ab/(a-b)] - [b/(a-b)].$$

Using the estimates $a = 1.10 gm/cm^3$ and $b = 0.90 gm/cm^3$ (see Katch and McArdle (1977), p. 111 or Wilmore (1976), p. 123) we come up with "Siri's equation":

$$\text{Percentage of Body Fat (i.e. 100*B)} = 495/D - 450.$$

Volume, and hence body density, can be accurately measured a variety of ways. The technique of underwater weighing "computes body volume as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight in water with the appropriate temperature correction for the water's density" (Katch and McArdle (1977), p. 113). Using this technique,

$$\text{Body Density} = \text{WA}/[(\text{WA-WW})/\text{c.f.} - \text{LV}]$$

where

```
WA = Weight in air (kg)
WW = Weight in water (kg)
c.f. = Water correction factor (=1 at 39.2 deg F as one-gram of water
       occupies exactly one cm^3 at this temperature, =.997 at 76-78 deg F)
LV = Residual Lung Volume (liters)
```

(Katch and McArdle (1977), p. 115). Other methods of determining body volume are given in Behnke and Wilmore (1974), p. 22 ff.

The variables listed below, from left to right, are:

```
Density determined from underwater weighing
Percent body fat from Siri's (1956) equation
Age (years)
Weight (lbs)
Height (inches)
Neck circumference (cm)
Chest circumference (cm)
Abdomen 2 circumference (cm)
Hip circumference (cm)
Thigh circumference (cm)
Knee circumference (cm)
Ankle circumference (cm)
Biceps (extended) circumference (cm)
Forearm circumference (cm)
Wrist circumference (cm)
```

(Measurement standards are apparently those listed in Benhke and Wilmore (1974), pp. 45-48 where, for instance, the abdomen 2 circumference is measured "laterally, at the level of the iliac crests, and anteriorly, at the umbilicus".)

These data are used to produce the predictive equations for lean body weight given in the abstract "Generalized body composition prediction equation for men using simple measurement techniques", K.W. Penrose, A.G. Nelson, A.G. Fisher, FACSM, Human Performance Research Center, Brigham Young University, Provo, Utah 84602 as listed in Medicine and Science in Sports and Exercise, vol. 17, no. 2, April 1985, p. 189. (The predictive equations were obtained from the first 143 of the 252 cases that are listed below). The data were generously supplied by Dr. A. Garth Fisher who gave permission to freely distribute the data and use for non-commercial purposes.

References

1. Bailey, Covert (1994). Smart Exercise: Burning Fat, Getting Fit. Houghton-Mifflin Co., Boston, pp. 179-186.

2. Behnke, A.R. and Wilmore, J.H. (1974). Evaluation and Regulation of Body Build and Composition, Prentice-Hall, Englewood Cliffs, N.J.

3. Siri, W.E. (1956), "Gross composition of the body", in Advances in Biological and Medical Physics, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.

4. Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

5. Wilmore, Jack (1976). Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process, Allyn and Bacon, Inc., Boston.

The SAS code is [3]

```
options center nodate pagesize=100 ls=80;
*ods pdf;
   ods graphics on;

title1 'Body Fat Data';
/*
Density determined from underwater weighing
 Percent body fat from Siri's (1956) equation
 Age (years)
 Weight (lbs)
 Height (inches)
 Neck circumference (cm)
 Chest circumference (cm)
 Abdomen 2 circumference (cm)
 Hip circumference (cm)
 Thigh circumference (cm)
 Knee circumference (cm)
 Ankle circumference (cm)
 Biceps (extended) circumference (cm)
 Forearm circumference (cm)
 Wrist circumference (cm)
*/
data bodyfat; set sasuser.bodyfat; run;

title2 'Simple Random Sampling of 120 values';
proc surveyselect data=bodyfat
   method=srs n=120 out=new_bfat seed = 54321;
run;
```

1.6 Crime Data in US

This data set was found on the Web at

http://www.statsci.org/data/general/uscrime.txt

My intent with this data set is to present a solution to a regression problem given with R-code and then to convert this data set into a SAS dataset and reproduce the results using SAS. The initial R-code is

```
dfC <- read.csv("http://www.statsci.org/data/general/uscrime.txt", sep="\t")
#print summary
summary(dfC)
```

[3]Note: This data is found on my machine, you will likely have it stored in another location on your machine. Make sure SAS can find the data. You can always place the data into the sasuser.v94 folder and chance ldata to your SAS

Chapter 2

Simple Descriptive Statistics

The descriptive statistics are in the form of either pictures or summary statistics (more numbers). In some cases these two types of descriptors are combined, e.g. box plots. This chapter presents some descriptive results for the data sets considered in the previous chapter.

2.1 Discrete Data

When the descriptive data is discrete and nominal one can only count the responses and the graphs either describe the number of such counts or their frequency as percentages or proportions. Examples are given below:

2.1.1 Kolache Sales

The SAS code is

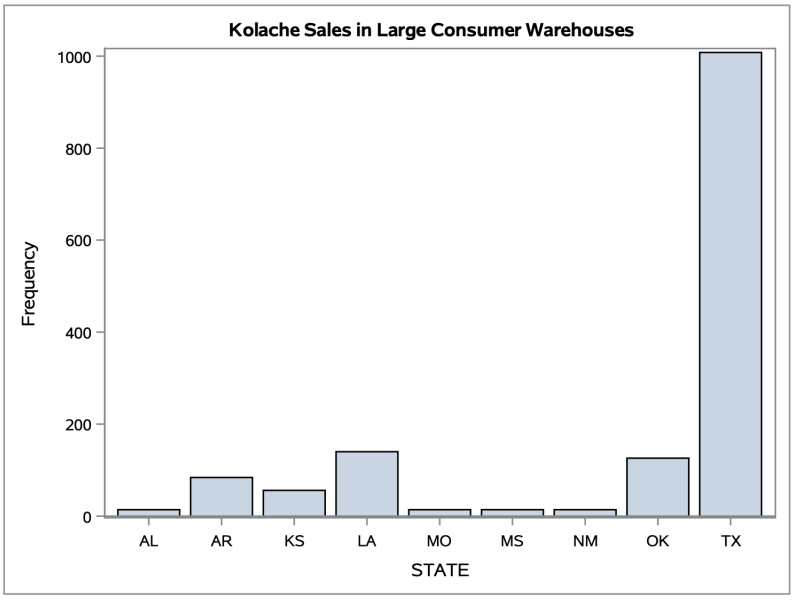
```
libname ldata "/folders/myfolders/Large Data Sets/SAS Data Sets";

options center nodate pagesize=100 ls=80;

title1 'Kolache Sales in Large Consumer Warehouses';

data kolache; set ldata.kolache; texas = (state eq 'TX');

proc freq data=kolache  ; table state; run;
proc sgplot;
vbar state;
run;
proc sgpie;
pie state;
run;


* Convert to readble R file;
proc export data=kolache
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/kolache.dbf"
replace dbms=dbf;
run;
quit;
```

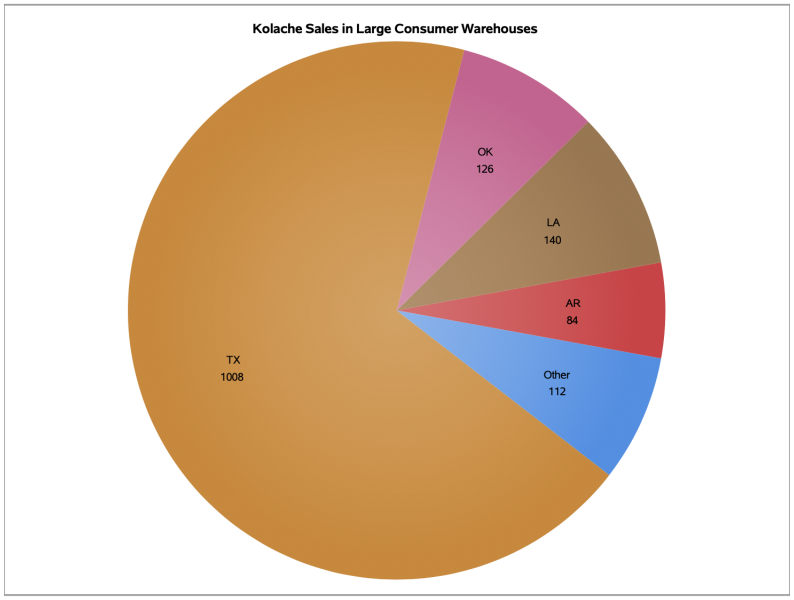
Two SAS example output tables and plots for the Kolache data are

Kolache Sales in Large Consumer Warehouses

The FREQ Procedure

STATE				
STATE	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
AL	14	0.95	14	0.95
AR	84	5.71	98	6.67
KS	56	3.81	154	10.48
LA	140	9.52	294	20.00
MO	14	0.95	308	20.95
MS	14	0.95	322	21.90
NM	14	0.95	336	22.86
OK	126	8.57	462	31.43
TX	1008	68.57	1470	100.00



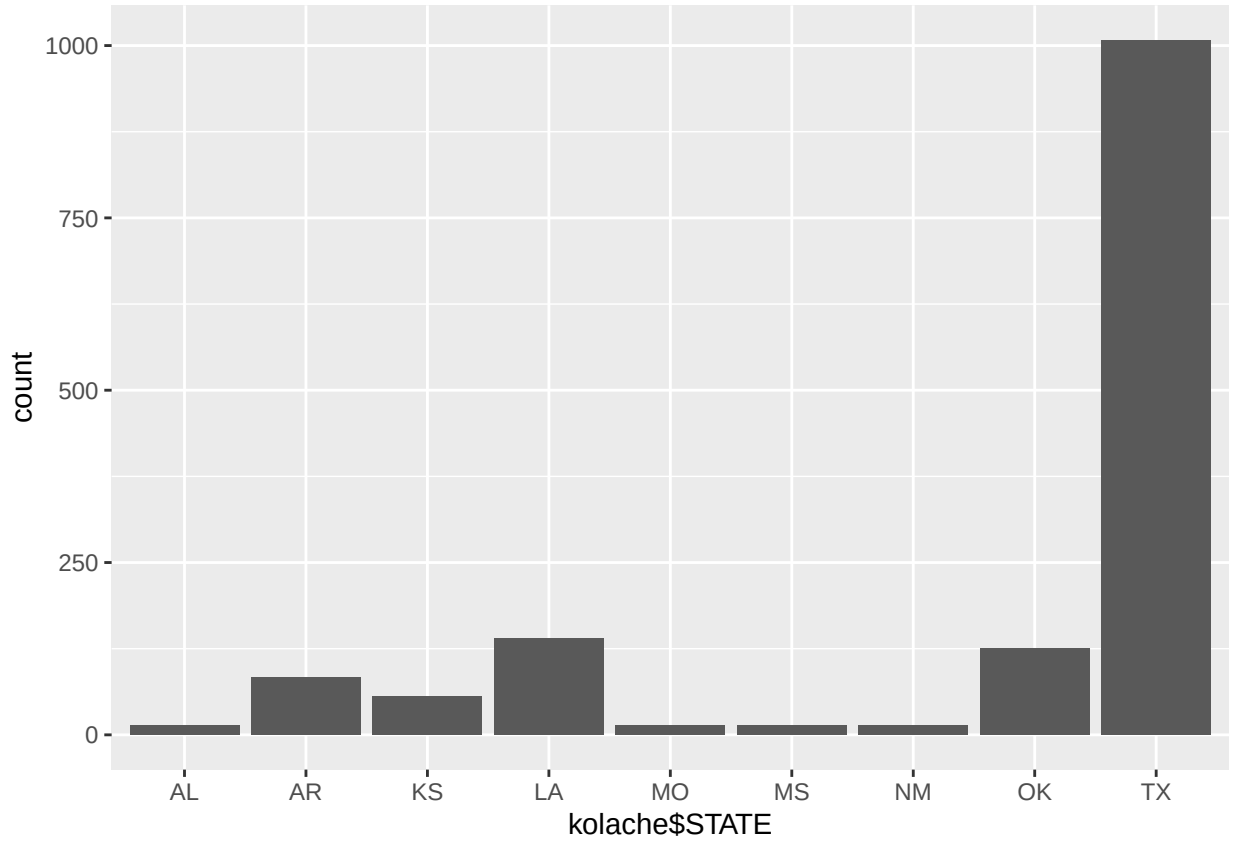


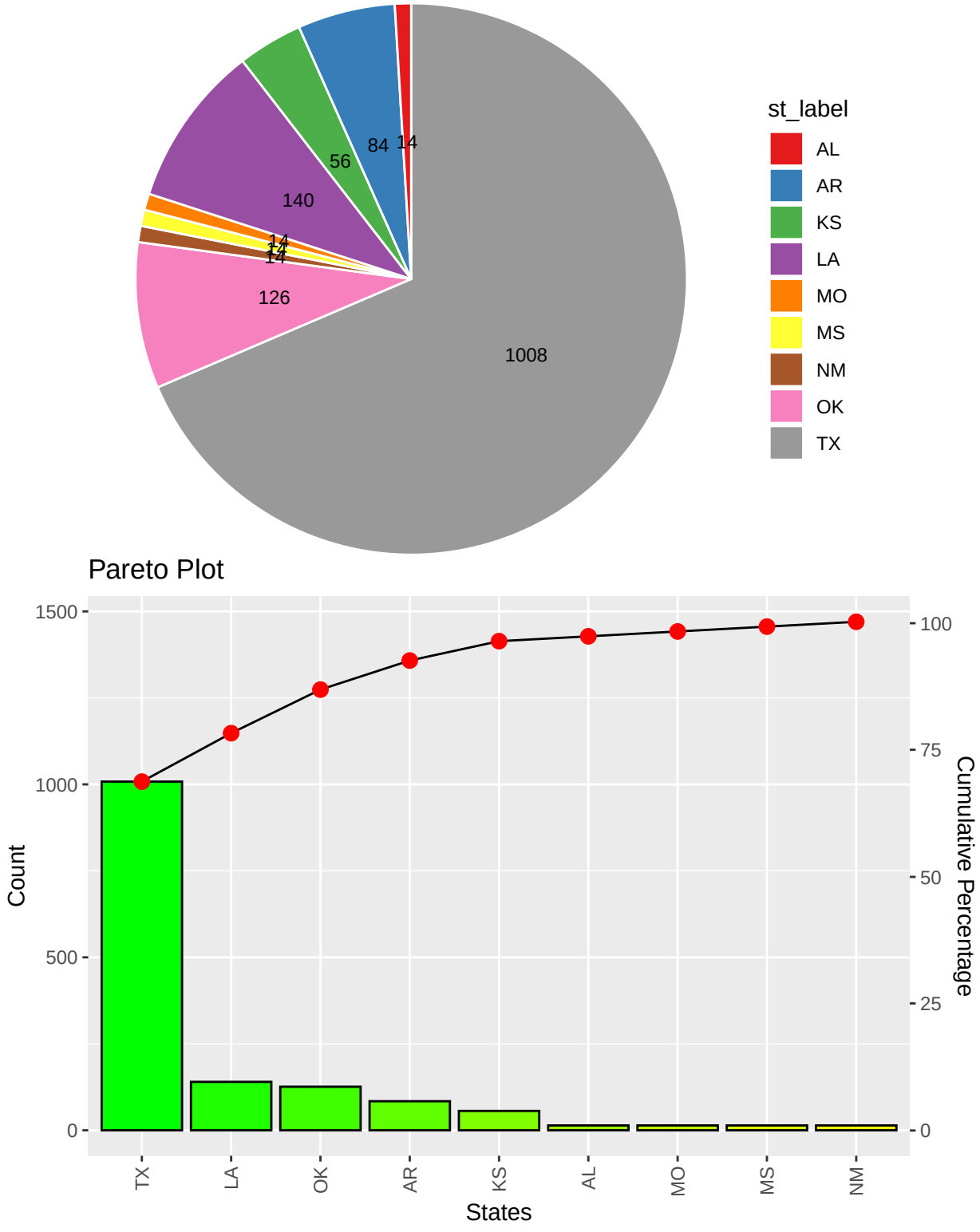
The R code is

```
library(foreign)
kolache = read.dbf("kolache.dbf")

library(ggplot2)
library(dplyr)
# Barplot
ggplot(kolache, aes(state)) + geom_bar()

st_label = c("AL", "AR", "KS", "LA", "MO", "MS", "NM", "OK", "TX")
st_count= c(14,   84,   56,  140,   14,   14,   14,  126, 1008 )
df = data.frame(st_label,st_count)

bp = ggplot(df, aes(x="", y=st_count, fill=st_label)) + geom_bar(width=1, stat = "identity")
pie = bp + coord_polar("y", start=0)


data = df %>%
 arrange(desc(st_label)) %>%
 mutate(prop = st_count/sum(df$st_count) * 100) %>%
 mutate(ypos = cumsum(prop) - 0.5*prop)



ggplot(data, aes(x="", y=prop, fill=st_label)) +
   geom_bar(stat="identity", width=1, color="white") +
   coord_polar("y", start=0) +
   theme_void() +
   theme(legend.position="right") +
   geom_text(aes(y = ypos, label = st_count), color = "black", size=3) +
   scale_fill_brewer(palette="Set1")
```

The R output is

```
> data[,1:3]
  st_label st_count      prop
1       TX     1008 68.571429
2       OK      126  8.571429
3       NM       14  0.952381
4       MS       14  0.952381
5       MO       14  0.952381
6       LA      140  9.523810
7       KS       56  3.809524
8       AR       84  5.714286
9       AL       14  0.952381
```





The R-markdown file is found on BOX for this course. The filename is

kolache_ch2.Rmd

2.1.2 Urinary Incontinence Trial

The SAS code is

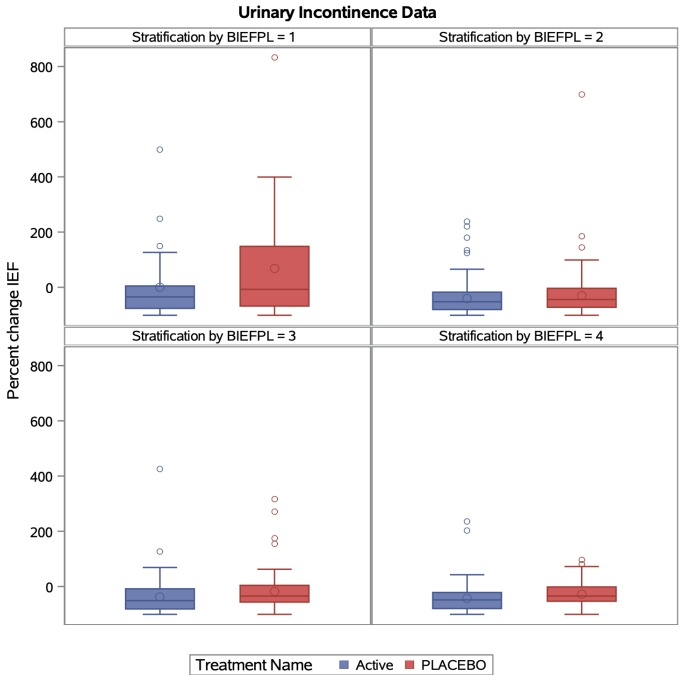
```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options center nodate pagesize=100 ls=80;

title1 'Urinary Incontinence Data';

data urinary; set sasuser.urinary;run;

* Perform descriptive methods;
proc freq data = urinary;
tables strata therapy strata*therapy ;
run;

ods graphics on;
proc sgpanel data=urinary;
panelby strata;
vbox pief/ group=therapy;
run;

*Create Output file to be read into R;
proc export data=urinary
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/urinary.dbf"
replace dbms=dbf;
run;
quit;
```

The SAS example output tables and plots for the urinary trial data are

The FREQ Procedure

Stratification by BIEFPL				
STRATA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	57	5.78	57	5.78
2	332	33.67	389	39.45
3	226	22.92	615	62.37
4	371	37.63	986	100.00
Frequency Missing = 14				

Treatment Name				
THERAPY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Active	525	52.50	525	52.50
PLACEBO	475	47.50	1000	100.00

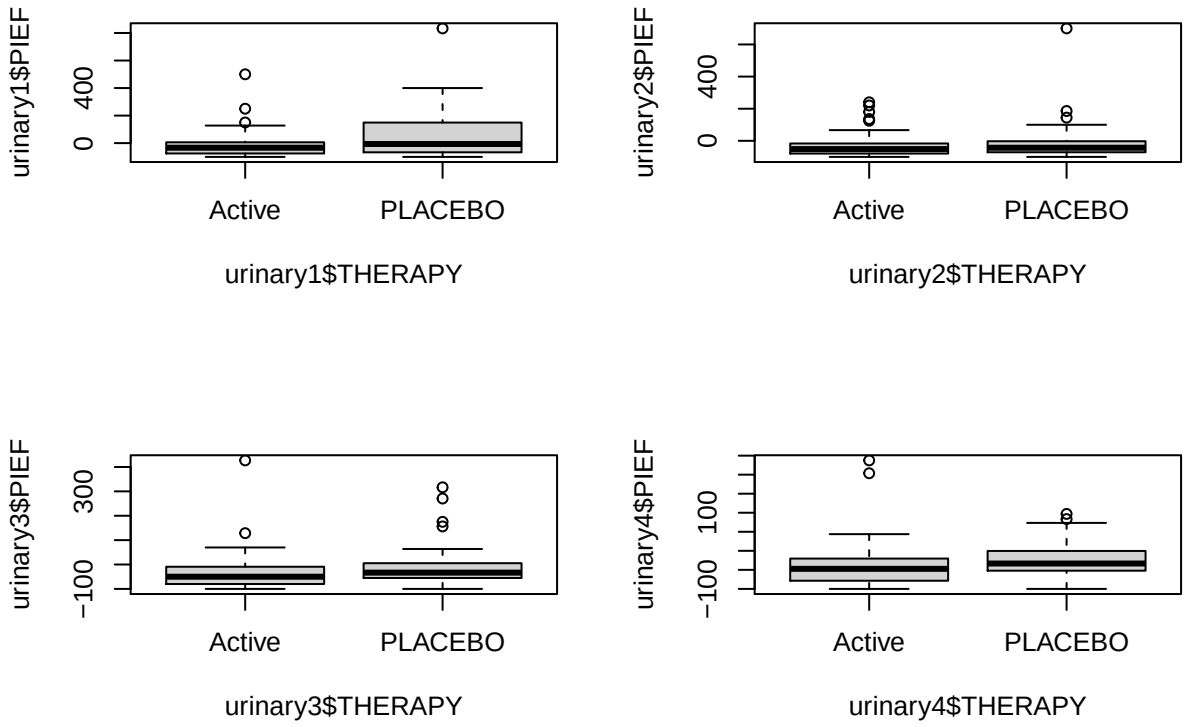


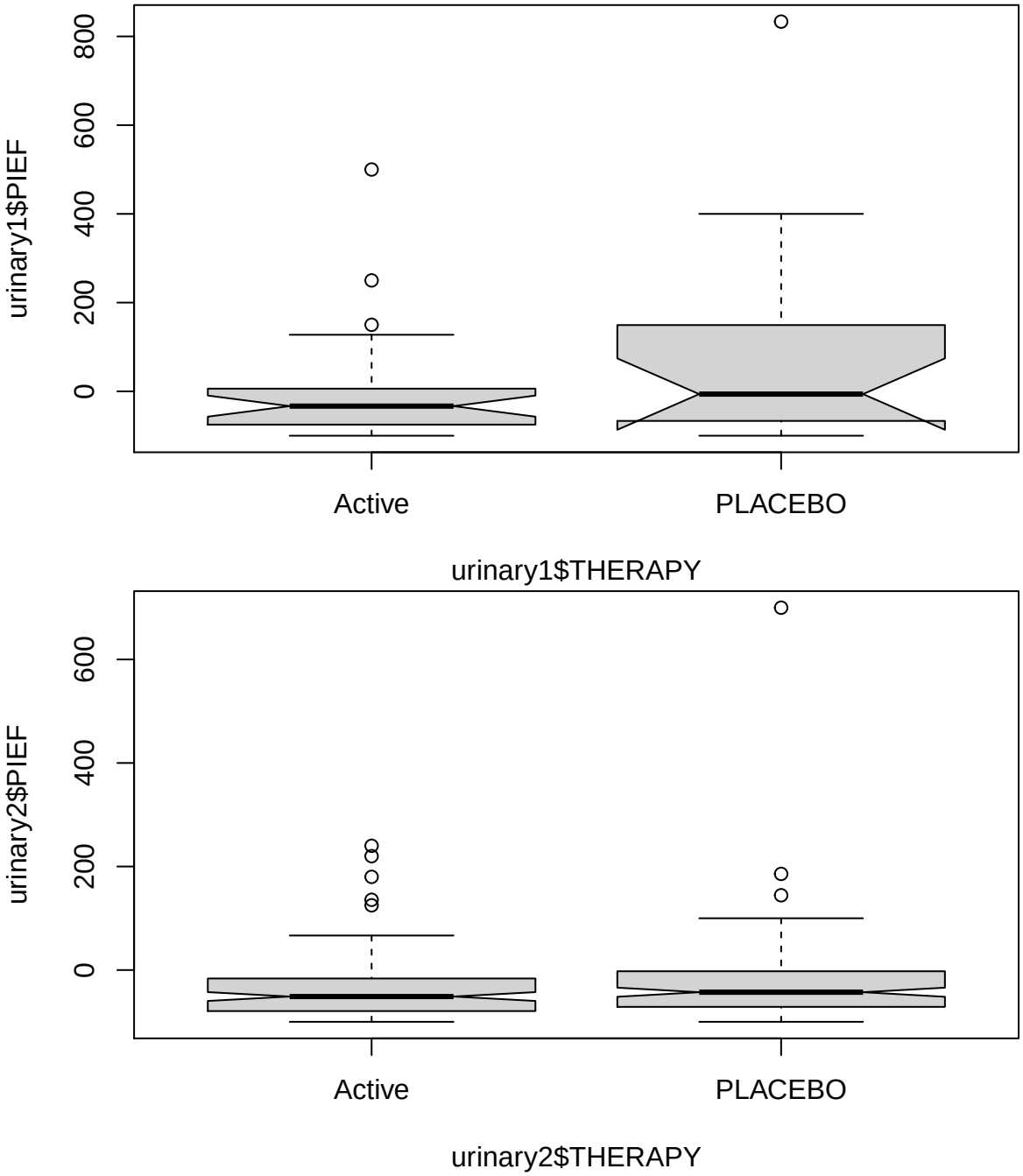
The R code is

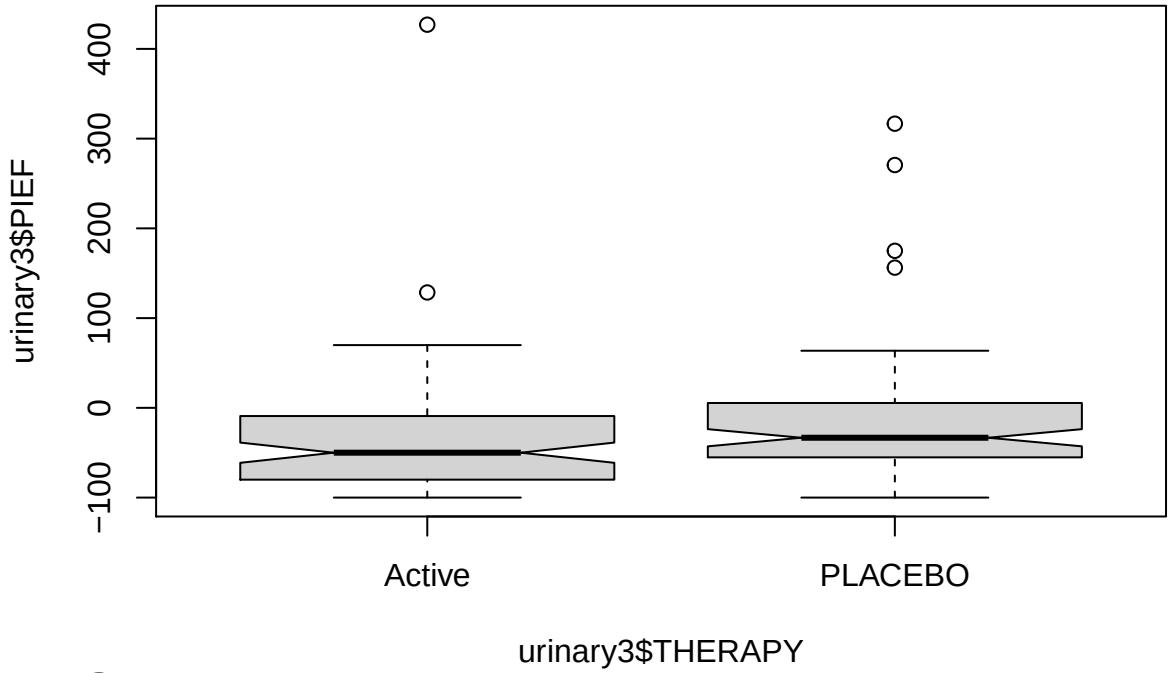
```
library(foreign)
urinary = read.dbf("urinary.dbf")
BMI = urinary$BMI
summary(BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NAs
  13.20   24.20   26.90   27.97   30.77   64.40       6

par(mfrow=c(2,2))

urinary1 = urinary[strata==1,]
urinary2 = urinary[strata==2,]
urinary3 = urinary[strata==3,]
urinary4 = urinary[strata==4,]

boxplot(urinary1$PIEF~urinary1$THERAPY, ds=urinary1)
boxplot(urinary2$PIEF~urinary2$THERAPY, ds=urinary2)
boxplot(urinary3$PIEF~urinary3$THERAPY, ds=urinary3)
boxplot(urinary4$PIEF~urinary4$THERAPY, ds=urinary4)

par(mfrow=c(1,1))
boxplot(urinary1$PIEF~urinary1$THERAPY,
        ds=urinary1, notch=TRUE)

boxplot(urinary2$PIEF~urinary2$THERAPY,
        ds=urinary2, notch=TRUE)

boxplot(urinary3$PIEF~urinary3$THERAPY,
        ds=urinary3, notch=TRUE)
```

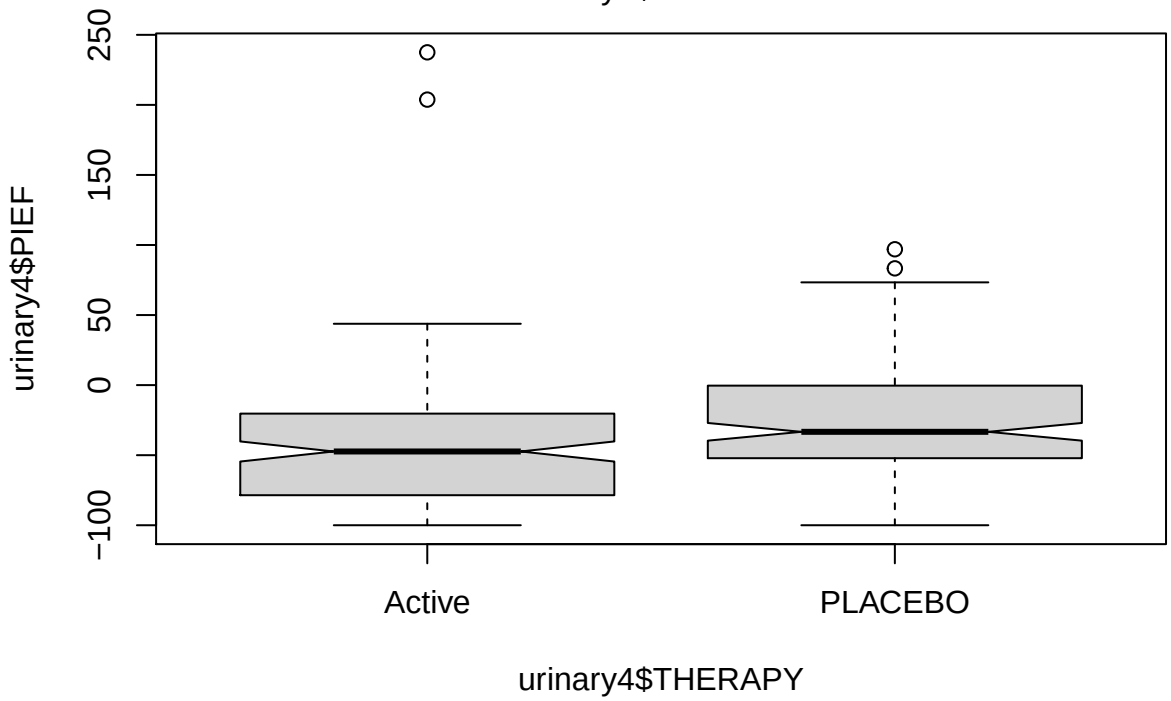
```
boxplot(urinary4$PIEF~urinary4$THERAPY,
        ds=urinary4, notch=TRUE)

library(ggplot2)
options(na.action = na.exclude)
ggplot(urinary1, aes(urinary1$THERAPY, urinary1$PIEF)) + geom_boxplot()
ggplot(urinary1, aes(urinary1$THERAPY, urinary1$PIEF)) + geom_boxplot(notch = TRUE)
ggplot(urinary1, aes(urinary1$THERAPY, urinary1$PIEF)) + geom_violin()
```

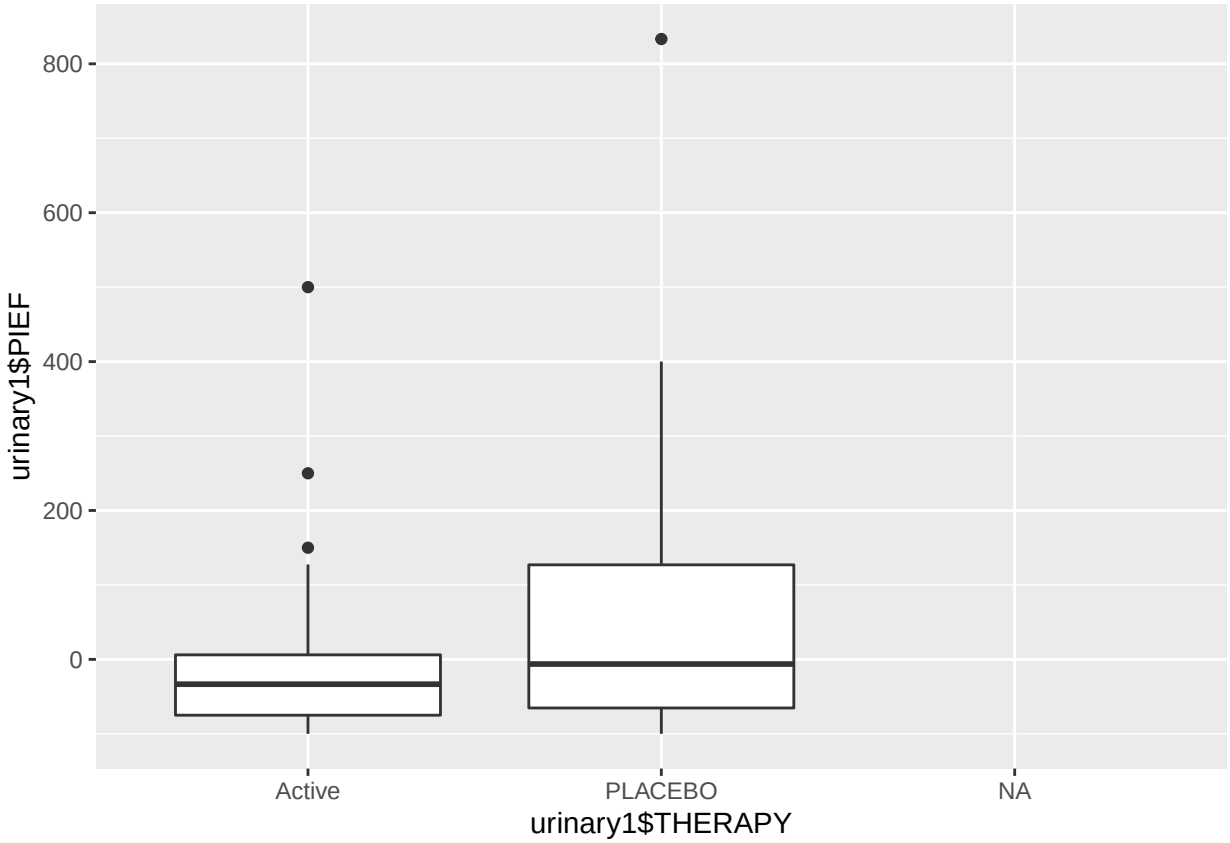
The R output is

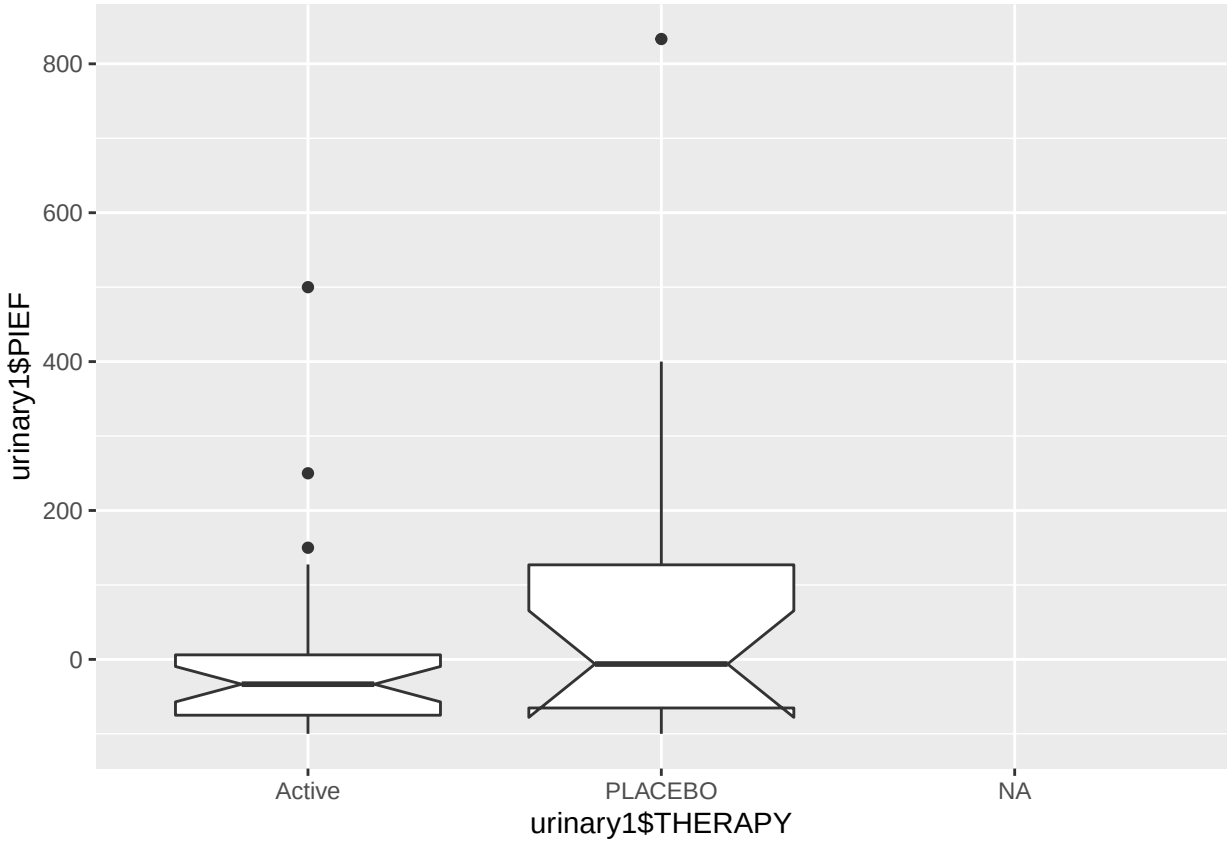


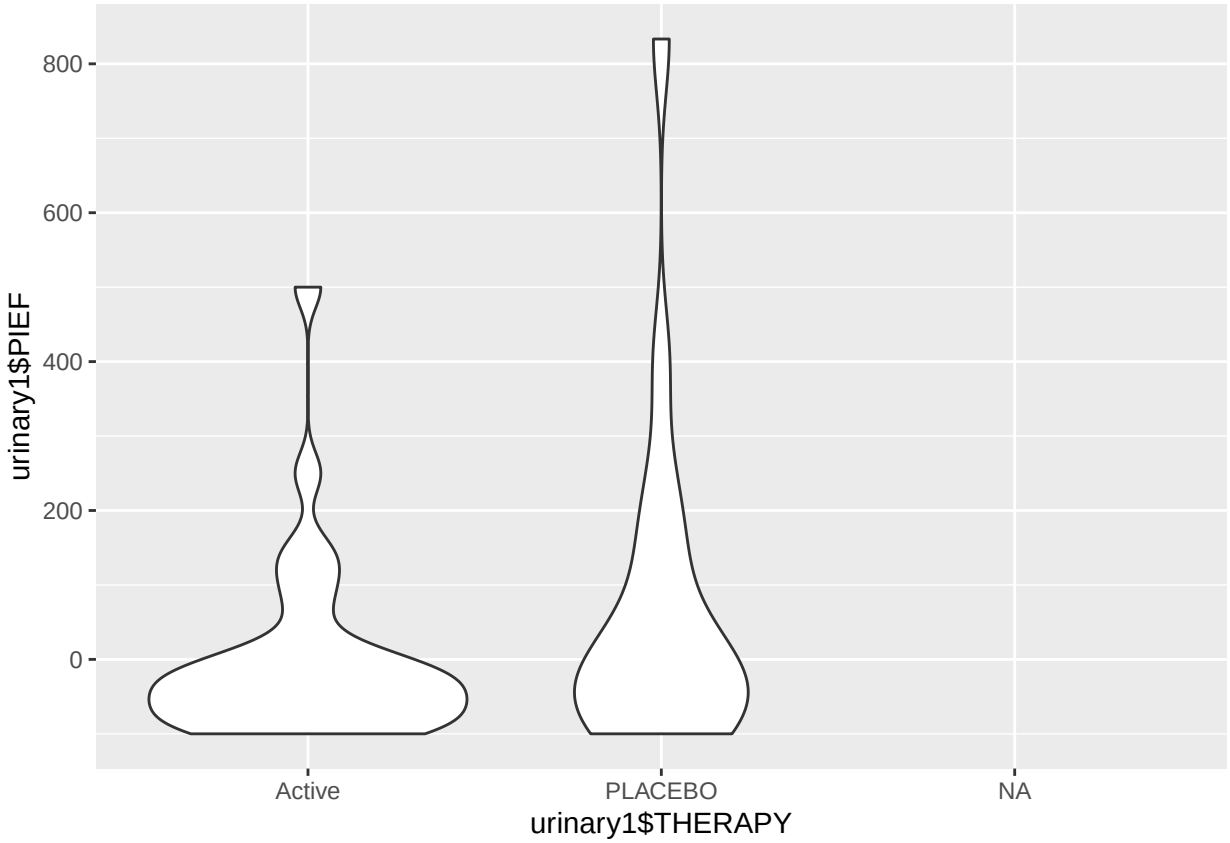

urinary1$PIEF
0
200
400
600
800
Active
PLACEBO
urinary1$THERAPY
urinary2$PIEF
0
200
400
600
Active
PLACEBO
urinary2$THERAPY


urinary3$PIEF
400
300
200
100
0
−100
Active
PLACEBO
urinary3$THERAPY


urinary4$PIEF
250
150
50
0
−100
Active
PLACEBO
urinary4$THERAPY


800
600
400
200
0
urinary1$PIEF
Active
PLACEBO
NA
urinary1$THERAPY


800
600
400
200
0
urinary1$PIEF
Active
PLACEBO
NA
urinary1$THERAPY



I have added some different styles of the simple boxplots that are available in R. The last three plots are created using ggplot2 which appears to be a favorite among the R users. I have included them for illustrative purposes.
The R-markdown file is found on BOX for this course. The filename is

```
Urinary_2.Rmd
```

2.2 Continuous Data

When the data are interval or continuous then the statistics and corresponding plots are more informative. Yet, the statistics are providing information about two main concepts for a distribution or probability density function; location and scale. The location is mainly concerned with the center or middle of the pdf. Whereas, the scale is addressing the spread or dispersion of the data – in most cases the spread about the center.

2.2.1 Certification

Some descriptive statistics SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options nodate nonumber ps=200 ls=80 formdlim=' ';

title 'Sample Certification Data';
data cert; set ldata.certification;
```

```
*proc means;
proc means data=cert; var written practica ; run;
proc sort data=cert; by year;
run;
title2 'Results for 2004 and 2008';
proc means data=cert;where year in (2004,2008);by year;
var written practica ;
run;

* US results;
title2 'Results for USA';
proc means data=cert;where country='USA' and
year in (2004, 2008); by year;
var written practica ;
run;

*proc univariate;
title3 'Passed Exams';
proc univariate data=cert normal; where country = 'USA'; var written;
   histogram / normal(color=(blue) mu=est sigma=est);
   qqplot;
run;

proc sgpanel data=cert; where country='USA' and
year in (2004, 2008);
panelby year;
histogram written;
density written/type=normal;
density written/type=kernel;
run;

* Create DBF file for R;
proc export data=cert
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/certification.dbf" dbms=dbf;
run;
quit;
```

The SAS output tables and plots are given below

Sample Certification Data
Results for 2004 and 2008

The MEANS Procedure

year=2004

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
WRITTEN	WRITTEN	135	86.2518519	6.4004647	69.0000000	100.0000000
PRACTICA	PRACTICAL	135	84.1555556	10.5899605	38.0000000	100.0000000

year=2008

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
WRITTEN	WRITTEN	392	85.2653061	8.4955846	45.0000000	98.0000000
PRACTICA	PRACTICAL	392	85.2984694	11.7170476	6.0000000	100.0000000

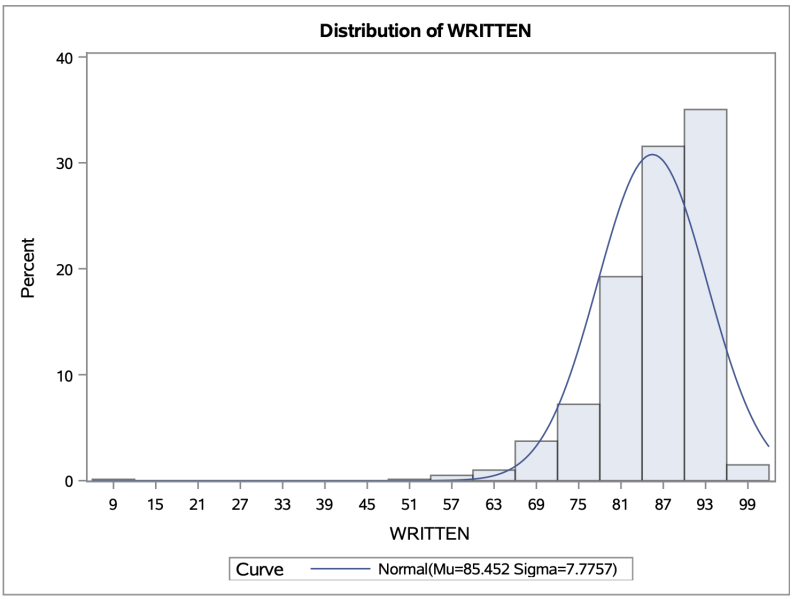
The UNIVARIATE Procedure
Variable: WRITTEN (WRITTEN)

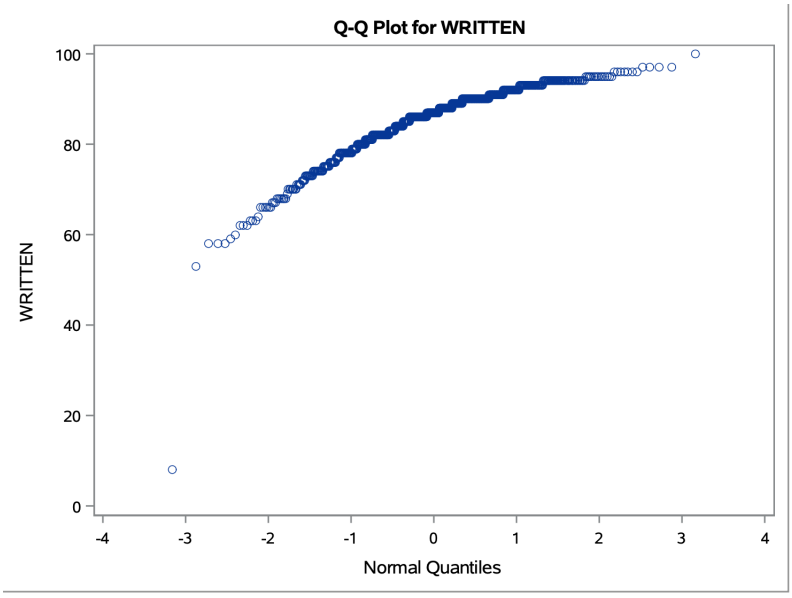
Moments			
N	805	Sum Weights	805
Mean	85.4521739	Sum Observations	68789
Std Deviation	7.77572833	Variance	60.4619511
Skewness	-2.1685634	Kurtosis	12.7948736
Uncorrected SS	5926781	Corrected SS	48611.4087
Coeff Variation	9.09950909	Std Error Mean	0.27405841

Basic Statistical Measures			
Location		Variability	
Mean	85.45217	Std Deviation	7.77573
Median	87.00000	Variance	60.46195
Mode	90.00000	Range	92.00000
		Interquartile Range	9.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	311.8028	Pr > |t|	<.0001
Sign	M	402.5	Pr >= |M|	<.0001
Signed Rank	S	162207.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.864404	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.14299	Pr > D	<0.0100
Cramer-von Mises	W-Sq	3.329751	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	19.74927	Pr > A-Sq	<0.0050



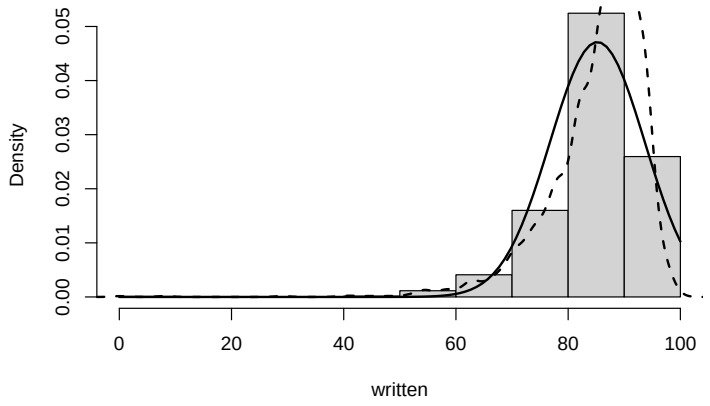


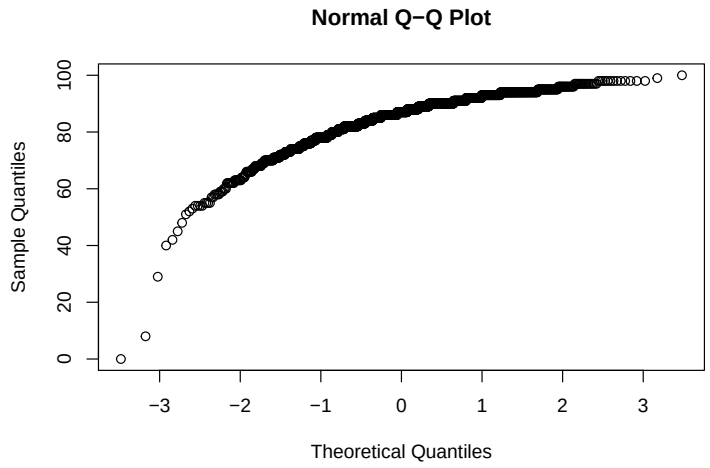
The R code is

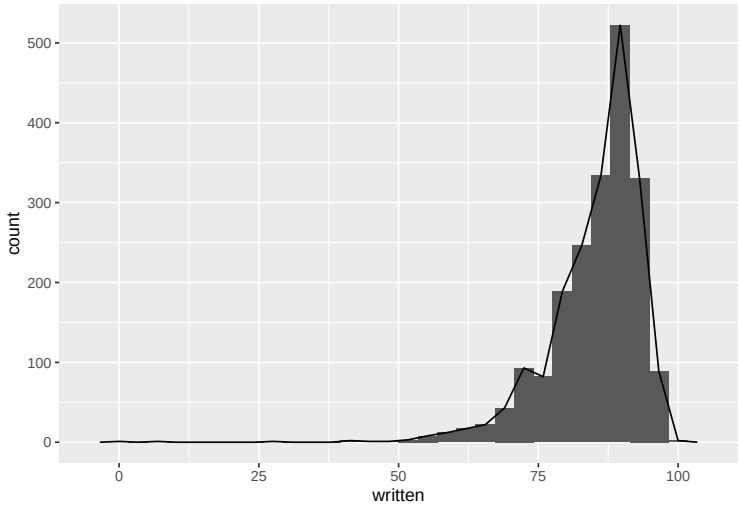
```
library(foreign)
cert = read.dbf("certification.dbf")

written = cert$WRITTEN
practica = cert$PRACTICA

library(mosaic)
favstats(written, data=cert)
 min Q1 median Q3 max   mean       sd    n missing
   0 82     87 91 100 85.211 8.466128 2000       0

mean(written, trim=.05)
[1] 85.93167
quantile(written, seq(from=.025, to= .975, by=.1))
 2.5% 12.5% 22.5% 32.5% 42.5% 52.5% 62.5% 72.5% 82.5% 92.5%
   64    76    81    83    86    87    89    90    92    94

t.test(written, mu=85.5)

  One Sample t-test
 data:  written
 t = -1.5266, df = 1999, p-value = 0.127
 alternative hypothesis: true mean is not equal to 85.5
 95 percent confidence interval:
  84.83974 85.58226
 sample estimates:
 mean of x
    85.211

library(nortest)
```

```
ad.test(written)

        Anderson-Darling normality test

data:  written
A = 52.682, p-value < 2.2e-16

cvm.test(written)

        Cramer-von Mises normality test

data:  written
W = 8.9833, p-value = 7.37e-10


lillie.test(written)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  written
D = 0.13713, p-value < 2.2e-16

pearson.test(written)

        Pearson chi-square normality test

data:  written
P = 2293, p-value < 2.2e-16

sf.test(written)

        Shapiro-Francia normality test

data:  written
W = 0.85814, p-value < 2.2e-16

with(cert, hist(written, main="", freq=FALSE))
with(cert, lines(density(written), main="WRITTEN", lty=2, lwd=2))
xvals = with(cert, seq(from=min(written), to=max(written), length=100))
with(cert, lines(xvals, dnorm(xvals, mean(written), sd(written)), lwd=2))

qqnorm(written)

library(ggplot2)
ggplot(cert, aes(written)) +
  geom_histogram() +
  geom_freqpoly()
```

The R output is







The R-markdown file is found on BOX for this course. The filename is

`certification_2.Rmd`

2.2.2 Kolache Sales

The SAS code is

```
libname ldata "/folders/myfolders/Large Data Sets/SAS Data Sets";

options center nodate pagesize=100 ls=80;

title1 'Kolache Sales in Large Consumer Warehouses';

data kolache; set ldata.kolache; texas = (state eq 'TX');

title2 'Desriptive Statistics for Texas Stores';

title2 'Total units sold in Texas';

proc univariate data=kolache trimmed=.05 winsor=.05;
where texas = 1; var unit_tot;
   histogram /normal;
   inset n = 'Number of Stores' / position=ne;
   probplot;
run;

proc sgplot data=kolache; where state='TX';
  vbox unit_tot/ category=week;
run;

proc sgplot data=kolache; where texas = 1;
  histogram unit_tot;
  density unit_tot/ type=kernel;
run;


* Convert to readble R file;
proc export data=kolache
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/kolache.dbf"
replace dbms=dbf;
run;
quit;
```

The SAS output is

Kolache Sales in Large Consumer Warehouses
Total units sold in Texas

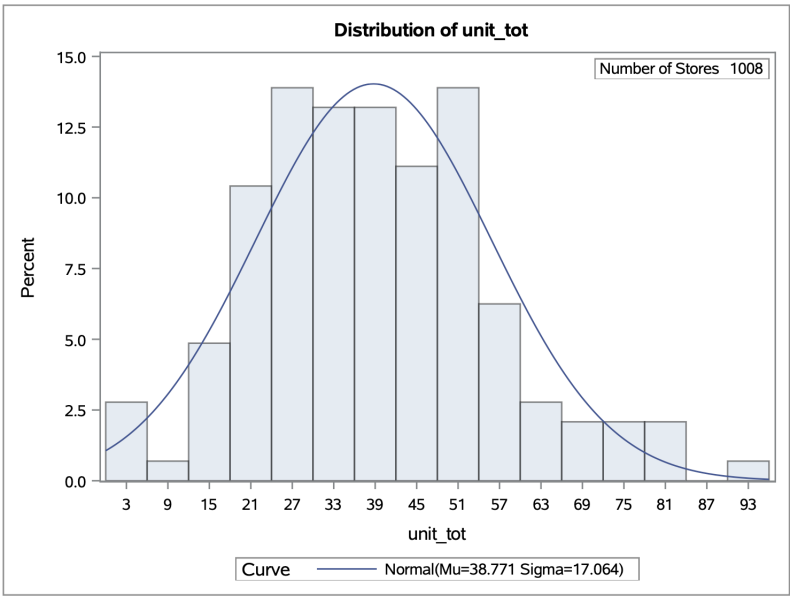
The UNIVARIATE Procedure
Variable: unit_tot

Moments			
N	1008	**Sum Weights**	1008
Mean	38.7708333	**Sum Observations**	39081
Std Deviation	17.0642242	**Variance**	291.187748
Skewness	0.41848734	**Kurtosis**	0.34429592
Uncorrected SS	1808429	**Corrected SS**	293226.062
Coeff Variation	44.0130447	**Std Error Mean**	0.53747254

Basic Statistical Measures			
Location		**Variability**	
Mean	38.77083	**Std Deviation**	17.06422
Median	38.50000	**Variance**	291.18775
Mode	41.00000	**Range**	94.00000
		Interquartile Range	23.00000

Trimmed Means								
Percent Trimmed in Tail	**Number Trimmed in Tail**	**Trimmed Mean**	**Std Error Trimmed Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.06	51	38.28146	0.530991	37.23934	39.32357	905	72.09439	<.0001

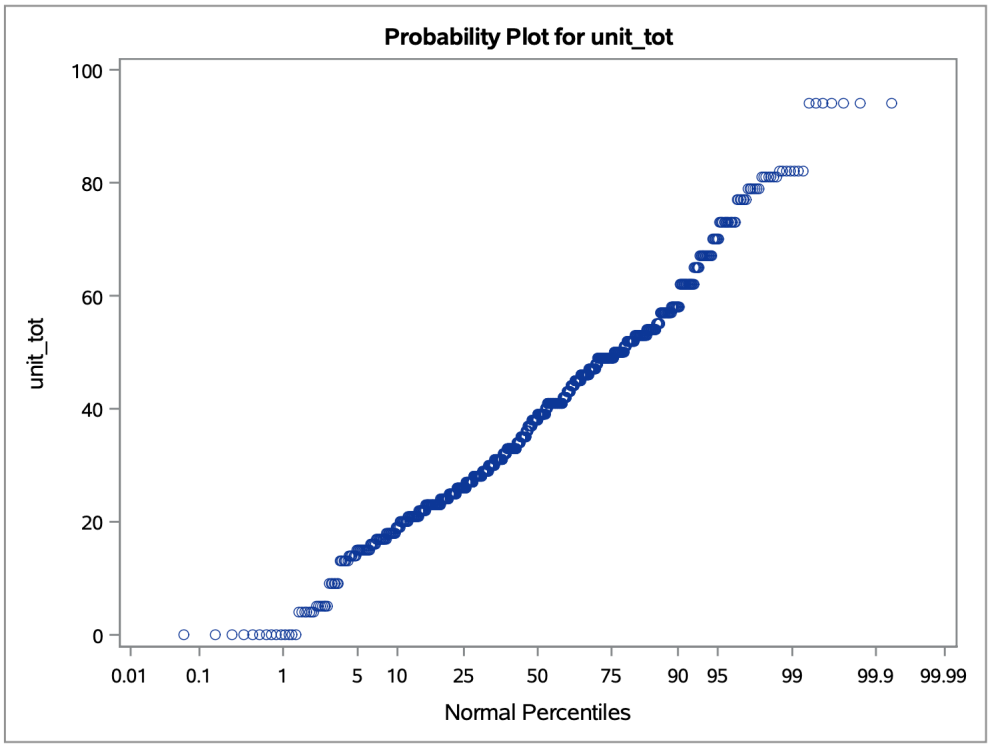
Winsorized Means								
Percent Winsorized in Tail	**Number Winsorized in Tail**	**Winsorized Mean**	**Std Error Winsorized Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.06	51	38.70833	0.531020	37.66616	39.75051	905	72.89424	<.0001



The UNIVARIATE Procedure
Fitted Normal Distribution for unit_tot

Parameters for Normal Distribution		
Parameter	**Symbol**	**Estimate**
Mean	Mu	38.77083
Std Dev	Sigma	17.06422

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.05599853	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.52724079	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	3.93492998	**Pr > A-Sq**	<0.005



The R code for these data would be very similar to the code used in the previous example with the certification data.

2.2.3 Assignment

Write your own Rmarkdown file to reproduce the above results with the Kolache Data

2.2.4 Body Fat - Density and Box-Cox

The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options center nodate pagesize=100 ls=80;

title1 'Body Fat Data';

data bodyfat; set ldata.bodyfat;
run;

title2 'Simple Random Sampling of 180 values';
proc surveyselect data=bodyfat
   method=srs n=180 out=new_bfat seed = 12345;
run;

title2 'Density estimation';
```

```
proc sgplot data=new_bfat;
histogram per_fat;
density per_fat;
run;
proc kde data=new_bfat;
univar per_fat;
run;
title3 'Abdomen Measurements';
proc sgplot data=new_bfat;
histogram abdomen;
density abdomen;
run;
proc kde data=new_bfat;
univar abdomen;
run;
proc univariate normal data=new_bfat;
var abdomen;
run;
/*
data new_bfat; set new_bfat; z=0; run;
proc transreg data=new_bfat;
model BoxCox(per_fat / convenient lambda=-2 to 2 by 0.05) =
         monotone(abdomen);
run;
*/
proc transreg data=new_bfat;
model BoxCox(abdomen / convenient lambda=-2 to 2 by 0.05) =
         monotone(density);
output out=new;
run;
title3 'Modified Abdomen measurement';

proc sgplot data=new;
histogram tabdomen;
density tabdomen;
run;
/*
proc kde data=new;
univar tabdomen;
run;
*/
proc univariate normal data=new;
var tabdomen;
run;

quit;
```

The SAS output is

The UNIVARIATE Procedure
Variable: per_fat

Basic Statistical Measures			
Location		**Variability**	
Mean	19.29278	**Std Deviation**	8.49765
Median	19.45000	**Variance**	72.21006
Mode	12.40000	**Range**	46.80000
		Interquartile Range	12.20000

Note: The mode displayed is the smallest of 4 modes with a count of 3.

Trimmed Means								
Percent Trimmed in Tail	**Number Trimmed in Tail**	**Trimmed Mean**	**Std Error Trimmed Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.00	9	19.19691	0.651435	17.91045	20.48337	161	29.46865	<.0001

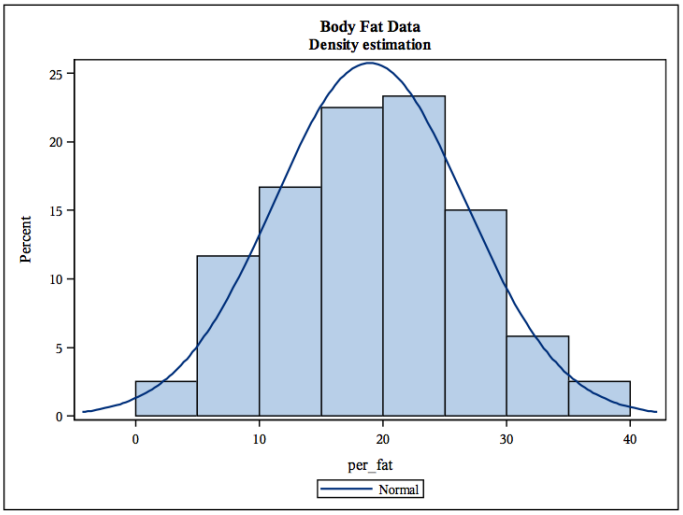
Winsorized Means								
Percent Winsorized in Tail	**Number Winsorized in Tail**	**Winsorized Mean**	**Std Error Winsorized Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.00	9	19.17222	0.651637	17.88536	20.45908	161	29.42161	<.0001

Robust Measures of Scale		
Measure	**Value**	**Estimate of Sigma**
Interquartile Range	12.20000	9.043870
Gini's Mean Difference	9.67682	8.575855
MAD	5.90000	8.747340
Sn	8.82524	8.825240
Qn	8.88760	8.703852

Body Fat Data
Test for Normality

The UNIVARIATE Procedure
Variable: per_fat

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.991185	**Pr < W**	0.3382
Kolmogorov-Smirnov	**D**	0.035802	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.029697	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.256514	**Pr > A-Sq**	>0.2500

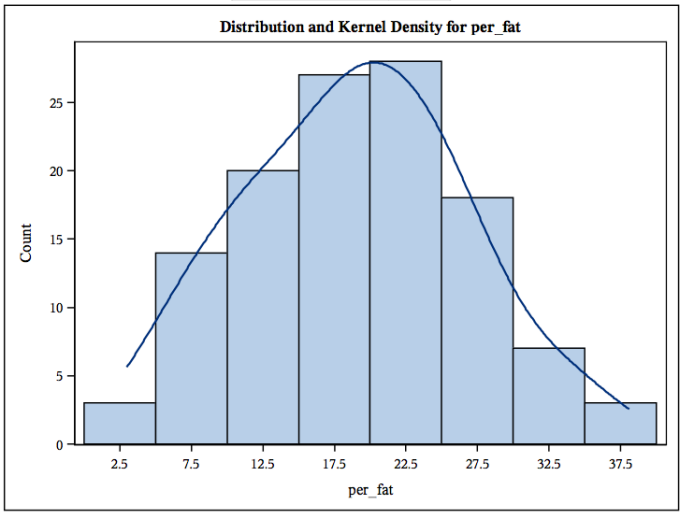


Body Fat Data
Density estimation

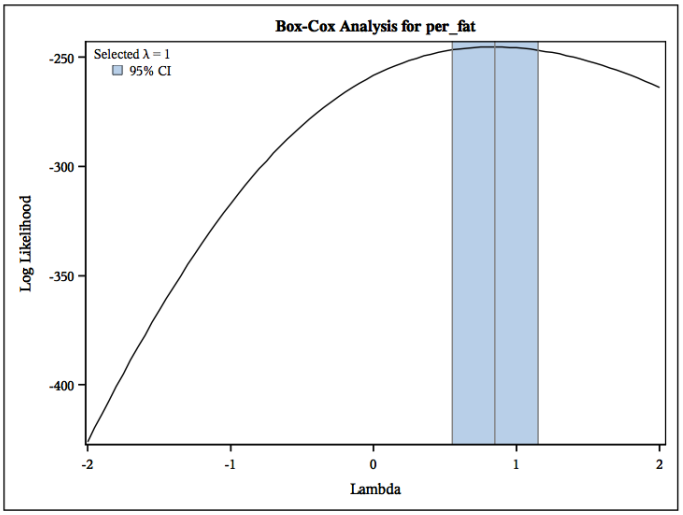
The KDE Procedure

Inputs	
Data Set	WORK.NEW_BFAT
Number of Observations Used	120
Variable	per_fat
Bandwidth Method	Sheather-Jones Plug In

Controls	
	per_fat
Grid Points	401
Lower Grid Limit	3
Upper Grid Limit	38.1
Bandwidth Multiplier	1



The TRANSREG Procedure



The SAS output for the abdomen measurement

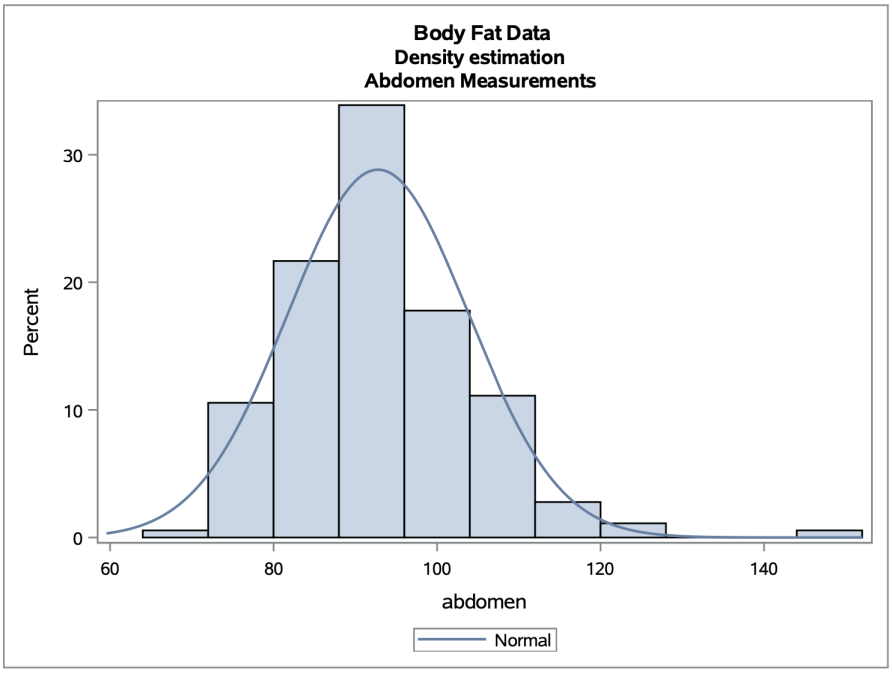
The UNIVARIATE Procedure
Variable: abdomen

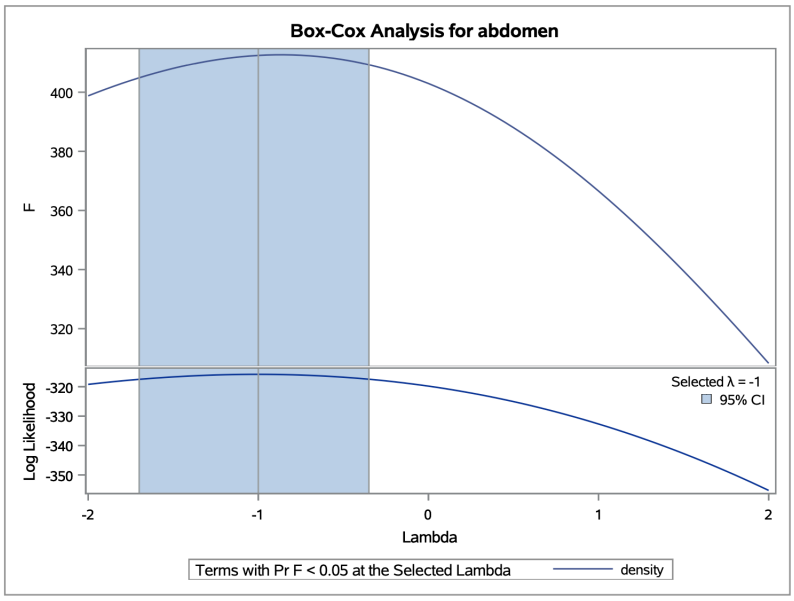
Moments			
N	180	**Sum Weights**	180
Mean	92.7805556	**Sum Observations**	16700.5
Std Deviation	11.072943	**Variance**	122.610067
Skewness	0.97950658	**Kurtosis**	2.82117509
Uncorrected SS	1571428.87	**Corrected SS**	21947.2019
Coeff Variation	11.9345513	**Std Error Mean**	0.82532844

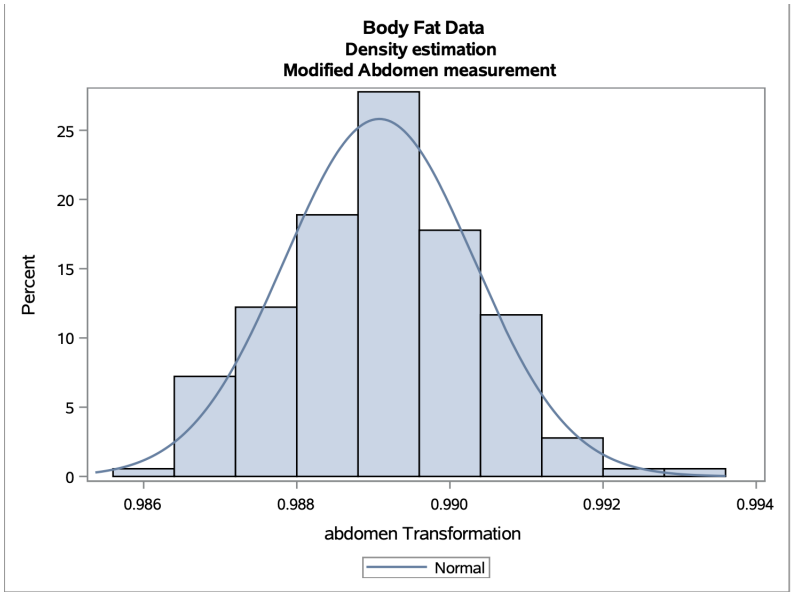
Basic Statistical Measures			
Location		**Variability**	
Mean	92.78056	**Std Deviation**	11.07294
Median	91.70000	**Variance**	122.61007
Mode	88.70000	**Range**	77.70000
		Interquartile Range	14.35000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	112.4165	**Pr > |t|**	<.0001
Sign	**M**	90	**Pr >= |M|**	<.0001
Signed Rank	**S**	8145	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.955412	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.05907	**Pr > D**	0.1265
Cramer-von Mises	**W-Sq**	0.132286	**Pr > W-Sq**	0.0424
Anderson-Darling	**A-Sq**	0.886939	**Pr > A-Sq**	0.0234







The UNIVARIATE Procedure
Variable: Tabdomen (abdomen Transformation)

Moments			
N	180	**Sum Weights**	180
Mean	0.98907771	**Sum Observations**	178.033988
Std Deviation	0.00123614	**Variance**	1.52805E-6
Skewness	0.04636421	**Kurtosis**	0.10561868
Uncorrected SS	176.089723	**Corrected SS**	0.00027352
Coeff Variation	0.12497933	**Std Error Mean**	0.00009214

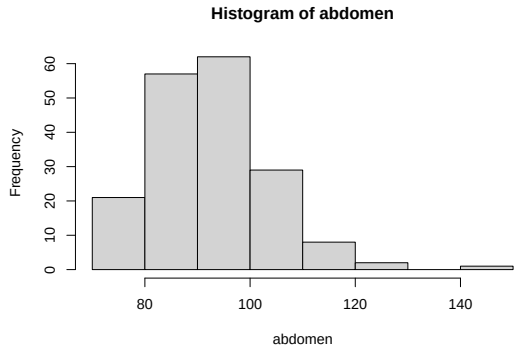
Basic Statistical Measures			
Location		**Variability**	
Mean	0.989078	**Std Deviation**	0.00124
Median	0.989095	**Variance**	1.52805E-6
Mode	0.988726	**Range**	0.00745
		Interquartile Range	0.00172

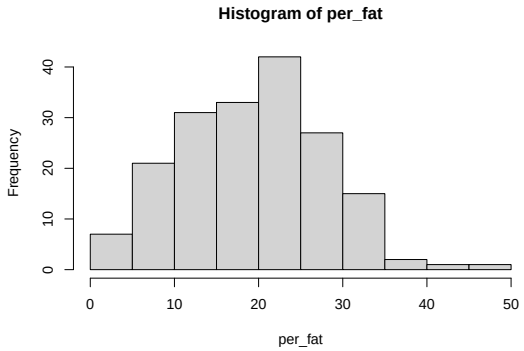
Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	10734.9	**Pr > |t|**	<.0001
Sign	**M**	90	**Pr >= |M|**	<.0001
Signed Rank	**S**	8145	**Pr >= |S|**	<.0001

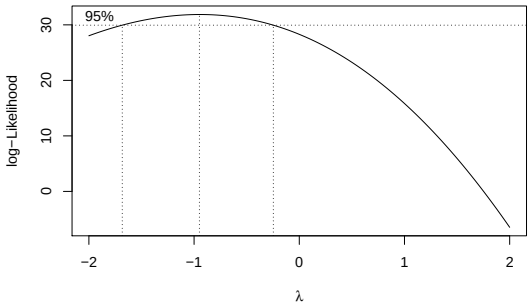
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.995666	**Pr < W**	0.8864
Kolmogorov-Smirnov	**D**	0.036668	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.024573	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.175484	**Pr > A-Sq**	>0.2500

The R code is

```
library(foreign)
bfat = read.dbf("new_bfat.dbf")
abdomen=bfat$abdomen
per_fat = bfat$per_fat
hist(abdomen)
hist(per_fat)

# Box Cox transformation
library(MASS)
boxcox(abdomen ~ per_fat, lambda=seq(-2, 2, length=100))
```







Chapter 3

Inference for Two Populations

3.1 Discrete Methods

3.1.1 Texas Childhood Obesity

In this example a new variable (**high_bmi** $= \mathbf{I}_{\text{bmi}>20}$) is analyzed. The SAS code is

```
options nodate nonumber ps=200 ls=80 formdlim=' ';

   title 'Obesity - All Counties';
   title2 'Cleaned Combined Data ';
   title3 ' ';
/*
   There are two data sets
       1. CDC BMI age/gender adjusted by CDC 2000
       2. Combined Texas preschool data
*/
   * define work data for cdc data files;
   * the permanent dataset is sasuser.cdc_bmi;
data cdc; set sasuser.cdc_bmi;
if agemos=24 then agemos=23.9;
age=floor(agemos);
if age < 13 then N_AGE = 0;
if age > 12 and age < 25 then N_AGE = 1;
if age > 24 and age < 37 then N_AGE = 2;
if age > 36 and age < 49 then N_AGE = 3;
if age > 48 and age < 61 then N_AGE = 4;
if age > 60 then N_AGE = 5;
if sex=1 then Gender= 'M';
if sex=2 then Gender= 'F'; drop sex;
run;
*proc contents data=cdc; run;

* define work data for texas_obesity data files;
* the permanent dataset is sasuser.new_combined_clean;
* year 2002 is removed due to the limited number of observations for that year;
DATA texas_obesity; SET sasuser.new_combined_clean;
IF YEAR > 2002;
```

```
if gender ne '.';RUN;
*proc contents data=texas_obesity; run;
/*
* Descriptive Statistics;
proc freq data = texas_obesity;
tables county*(year gender)/nocol nopercent ;
run;
*/

* Create a new data set where high_bmi = 1 if bmi > 20;
data high_bmi; set texas_obesity;
  high_bmi=(bmi > 20);
  central = 0;
  if county in ('Bastrop','McLenna') then central = 1;
  male=(gender='M');
run;
title3 'Chi Square for High BMI';
proc freq data=high_bmi;
  tables (county gender) * high_bmi/chisq nocol nopercent;
run;
title3 'Chi Square for High BMI';
proc freq data=high_bmi;
  tables central * high_bmi/chisq nocol nopercent;
run;

title3 'CMH Test for High BMI';
proc freq data=high_bmi;
  tables male* central * high_bmi/chisq cmh nocol nopercent;
run;

quit;
```

The SAS output is

Obesity - All Counties
Cleaned Combined Data
Chi Square for High BMI

The FREQ Procedure

Frequency Row Pct	Table of central by high_bmi		
	high_bmi		
central	0	1	Total
0	2676 90.56	279 9.44	2955
1	1719 93.02	129 6.98	1848
Total	4395	408	4803

Statistics for Table of central by high_bmi

Statistic	DF	Value	Prob
Chi-Square	1	8.8596	0.0029
Likelihood Ratio Chi-Square	1	9.0742	0.0026
Continuity Adj. Chi-Square	1	8.5458	0.0035
Mantel-Haenszel Chi-Square	1	8.8577	0.0029
Phi Coefficient		-0.0429	
Contingency Coefficient		0.0429	
Cramer's V		-0.0429	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2676
Left-sided Pr <= F	0.0016
Right-sided Pr >= F	0.9989
Table Probability (P)	0.0005
Two-sided Pr <= P	0.0029

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.7198	0.5792	0.8945
Relative Risk (Column 1)	0.9735	0.9571	0.9903
Relative Risk (Column 2)	1.3526	1.1069	1.6527

Sample Size = 4803

Frequency Row Pct	Table 1 of central by high_bmi		
	Controlling for Gender=F		
	high_bmi		
central	0	1	Total
0	1336 89.36	159 10.64	1495
1	870 93.65	59 6.35	929
Total	2206	218	2424

Statistics for Table 1 of central by high_bmi Controlling for Gender=F

Statistic	DF	Value	Prob
Chi-Square	1	12.8509	0.0003
Likelihood Ratio Chi-Square	1	13.4296	0.0002
Continuity Adj. Chi-Square	1	12.3328	0.0004
Mantel-Haenszel Chi-Square	1	12.8456	0.0003
Phi Coefficient		-0.0728	
Contingency Coefficient		0.0726	
Cramer's V		-0.0728	

The FREQ Procedure

Frequency Row Pct	Table 2 of central by high_bmi		
	Controlling for Gender=M		
	high_bmi		
central	0	1	Total
0	1340 91.78	120 8.22	1460
1	849 92.38	70 7.62	919
Total	2189	190	2379

Statistics for Table 2 of central by high_bmi Controlling for Gender=M

Statistic	DF	Value	Prob
Chi-Square	1	0.2783	0.5978
Likelihood Ratio Chi-Square	1	0.2798	0.5968
Continuity Adj. Chi-Square	1	0.2024	0.6528
Mantel-Haenszel Chi-Square	1	0.2782	0.5979
Phi Coefficient		-0.0108	
Contingency Coefficient		0.0108	
Cramer's V		-0.0108	

Summary Statistics for central by high_bmi
Controlling for Gender

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	8.8363	0.0030
2	Row Mean Scores Differ	1	8.8363	0.0030
3	General Association	1	8.8363	0.0030

Common Odds Ratio and Relative Risks				
Statistic	Method	Value	95% Confidence Limits	
Odds Ratio	Mantel-Haenszel	0.7202	0.5796	0.8950
	Logit	0.7264	0.5839	0.9037
Relative Risk (Column 1)	Mantel-Haenszel	0.9736	0.9571	0.9903
	Logit	0.9738	0.9573	0.9905
Relative Risk (Column 2)	Mantel-Haenszel	1.3522	1.1064	1.6525
	Logit	1.3398	1.0951	1.6390

Breslow-Day Test for Homogeneity of Odds Ratios	
Chi-Square	4.6577
DF	1
Pr > ChiSq	0.0309

The R-code is

```
library(foreign)
obesity = read.dbf("high_bmi.dbf")
summary(obesity)

central=obesity$central
high_bmi = obesity$high_bmi
male = obesity$male

table(central,high_bmi)
       high_bmi
central     0     1
      0 12038  1223
      1  4798   409

library(mosaic)
tally(~ central + high_bmi)
       high_bmi
central     0     1
      0 12038  1223
      1  4798   409
tally(~ central + high_bmi, format="percent")
       high_bmi
central         0         1
      0 65.183019  6.622266
      1 25.980074  2.214642
tally(~ central + high_bmi, data=obesity)
       high_bmi
central     0     1
```

```
      0 12038  1223
      1  4798   409
chisq.test(ob_tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  ob_tab
X-squared = 8.5131, df = 1, p-value = 0.003526

fisher.test(ob_tab)

        Fisher's Exact Test for Count Data'

data:  ob_tab
p-value = 0.003283
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7446432 0.9440423
sample estimates:
odds ratio
 0.8390762

mytab = table(central,high_bmi)
mytab
       high_bmi
central     0     1
      0 12038  1223
      1  4798   409
addmargins(mytab)
       high_bmi
central     0     1   Sum
    0   12038  1223 13261
    1    4798   409  5207
    Sum 16836  1632 18468
prop.table(mytab, 1)
       high_bmi
central          0          1
      0 0.90777468 0.09222532
      1 0.92145189 0.07854811

chisq.test(mytab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  mytab
X-squared = 8.5131, df = 1, p-value = 0.003526

table(male,central)
    central
male    0    1
   0 6785 2577
   1 6476 2630
mantelhaen.test(male, central, high_bmi)
```

```
        Mantel-Haenszel chi-squared test with continuity correction

data:  male and central and high_bmi
Mantel-Haenszel X-squared = 4.1509, df = 1, p-value = 0.04161
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 NA NA
sample estimates:
common odds ratio
         1.069502
```

3.1.2 2 x 2 Tables – Urinary Clinical

In this example, I have arbitrarily defined a new binary variable, benefit, where benefit =“affirmative” if the percent change from baseline < -25 and is “negative” if otherwise. The SAS code is given in

urinary_binary.sas

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options center nodate pagesize=100 ls=80;

title1 'Urinary Incontinence Data';

data urinary; set ldata.urinary;run;

*Define binary variable for pief < -25;
data urinary; set urinary; benefit = 'negative';
if pief < -25 then benefit = 'affirmative'; run;
* success means benefit = 1;
title2 'Benefit = 1 is a Success';
title3 'Combined results for ignoring Strata';
proc freq data=urinary ;
tables therapy *benefit/ or relrisk nocol nopercent;
run;
title3 'Results when accounting for Strata';
proc freq data=urinary;
tables strata*therapy*benefit / or   nocol nopercent cmh;
run;

proc export data=urinary
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/urinary.dbf" dbms=dbf;
run;
quit;
```

The SAS output is

Urinary Incontinence Data
Benefit = 1 is a Success
Combined results for ignoring Strata

The FREQ Procedure

Frequency Row Pct	Table of THERAPY by benefit		
THERAPY(Treatment Name)	benefit		
	affirmat	negative	Total
Active	387 73.71	138 26.29	525
PLACEBO	290 61.05	185 38.95	475
Total	677	323	1000

Statistics for Table of THERAPY by benefit

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.7890	1.3685	2.3386
Relative Risk (Column 1)	1.2074	1.1055	1.3186
Relative Risk (Column 2)	0.6749	0.5625	0.8098

Sample Size = 1000

Urinary Incontinence Data
Benefit = 1 is a Success
Results when accounting for Strata

The FREQ Procedure

Frequency Row Pct	Table 1 of THERAPY by benefit Controlling for STRATA=1		
THERAPY(Treatment Name)	benefit		
	affirmat	negative	Total
Active	20 58.82	14 41.18	34
PLACEBO	13 56.52	10 43.48	23
Total	33	24	57

Statistics for Table 1 of THERAPY by benefit
Controlling for STRATA=1

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.0989	0.3767	3.2055
Relative Risk (Column 1)	1.0407	0.6599	1.6413
Relative Risk (Column 2)	0.9471	0.5119	1.7522

Frequency Row Pct	Table 3 of THERAPY by benefit Controlling for STRATA=3		
THERAPY(Treatment Name)	benefit		
	affirmat	negative	Total
Active	87 73.11	32 26.89	119
PLACEBO	61 57.01	46 42.99	107
Total	148	78	226

Statistics for Table 3 of THERAPY by benefit
Controlling for STRATA=3

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	2.0502	1.1742	3.5799
Relative Risk (Column 1)	1.2824	1.0527	1.5622
Relative Risk (Column 2)	0.6255	0.4330	0.9037

Results when accounting for Strata

The FREQ Procedure

Summary Statistics for THERAPY by benefit
Controlling for STRATA

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	17.4414	<.0001
2	Row Mean Scores Differ	1	17.4414	<.0001
3	General Association	1	17.4414	<.0001

Common Odds Ratio and Relative Risks				
Statistic	Method	Value	95% Confidence Limits	
Odds Ratio	Mantel-Haenszel	1.7766	1.3554	2.3286
	Logit	1.7770	1.3552	2.3301
Relative Risk (Column 1)	Mantel-Haenszel	1.2041	1.1019	1.3157
	Logit	1.2008	1.0997	1.3113
Relative Risk (Column 2)	Mantel-Haenszel	0.6800	0.5665	0.8163
	Logit	0.6829	0.5689	0.8196

Breslow-Day Test for Homogeneity of Odds Ratios	
Chi-Square	1.0954
DF	3
Pr > ChiSq	0.7782

The R-code for this problem is very similar to that used in the previous example.

3.1.3 Assignment

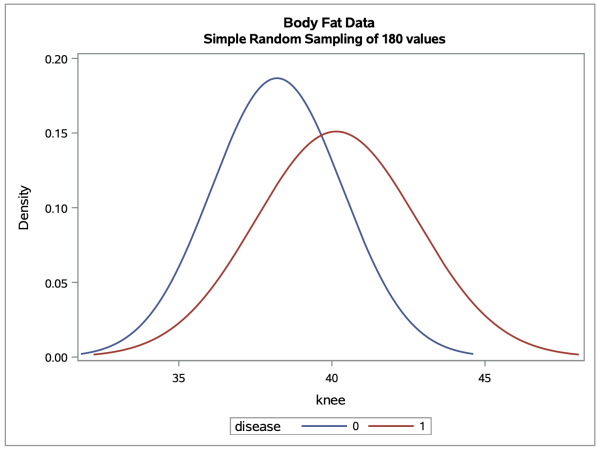
Create a R markdown file that reproduces the above results using the urinary incontinence data.

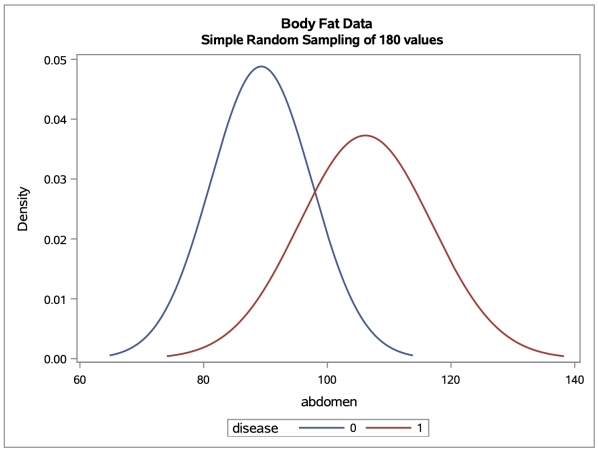
3.1.4 ROC Plots – Body Fat

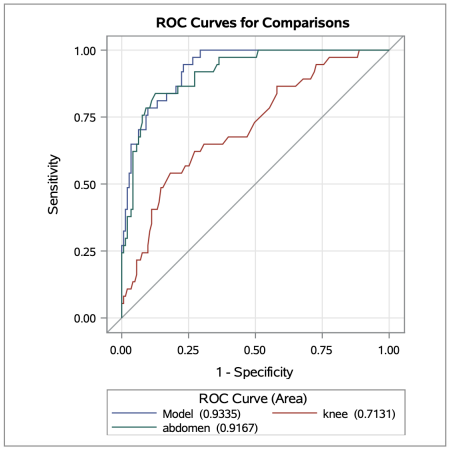
In this example, I have arbitrarily defined a new binary variable, disease, where disease = 1 if the percent fat > 28 and is 0 if otherwise. I am using two variables to model the event "disease = 1"; knee and abdomen measurements. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options center nodate pagesize=100 ls=80;

title1 'Body Fat Data';

data bodyfat; set ldata.bodyfat;
run;

title2 'Simple Random Sampling of 180 values';
proc surveyselect data=bodyfat
   method=srs n=180 out=new_bfat seed = 12345;
run;

* Create new data set with disease =1 if per_fat > 26;
data fat_roc; set new_bfat; disease = (per_fat > 26);

proc sgplot data=fat_roc;
density knee/ group=disease;
run;

proc sgplot data=fat_roc;
density abdomen/ group=disease;
run;

proc logistic data=fat_roc plots(only)=roc;;
  model disease (event='1') =knee abdomen /
                         scale=none
                         clparm=wald
                         clodds=pl
                         rsquare;
roc 'knee' knee;
roc 'abdomen' abdomen;
roccontrast reference('abdomen') / estimate e;
run;

proc export data=fat_roc
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/fat_roc.dbf"
replace dbms=dbf;
run;
proc export data=new_bfat
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/new_bfat.dbf"
replace dbms=dbf;
run;
quit;
```

The SAS output is







ROC Association Statistics

ROC Model	Mann-Whitney Area	Mann-Whitney Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
Model	0.9335	0.0185	0.8973	0.9697	0.8669	0.8669	0.2847
knee	0.7131	0.0476	0.6197	0.8065	0.4262	0.4316	0.1400
abdomen	0.9167	0.0234	0.8707	0.9626	0.8333	0.8346	0.2737

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = abdomen	2	32.5009	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model - abdomen	0.0168	0.0125	-0.00763	0.0413	1.8175	0.1776
knee - abdomen	-0.2036	0.0372	-0.2764	-0.1307	29.9633	<.0001

The R-code is

```
library(foreign)
bfat = read.dbf("new_bfat.dbf")

abdomen=bfat$abdomen
per_fat = bfat$per_fat
thigh = bfat$thigh


#Plots of Percent Fat and Abdomen Circumference

with(bfat, hist(per_fat, main="", freq=FALSE))
with(bfat, lines(density(per_fat), main="PERCENT FAT", lty=2, lwd=2))
xvals = with(bfat, seq(from=min(per_fat), to=max(per_fat), length=100))
with(bfat, lines(xvals, dnorm(xvals, mean(per_fat), sd(per_fat)), lwd=2))

with(bfat, hist(abdomen, main="", freq=FALSE))
with(bfat, lines(density(abdomen), main="ABDOMEN", lty=2, lwd=2))
xvals = with(bfat, seq(from=min(abdomen), to=max(abdomen), length=100))
with(bfat, lines(xvals, dnorm(xvals, mean(abdomen), sd(abdomen)), lwd=2))

#Box Cox transformation for Abdomen

library(MASS)
boxcox(abdomen ~ per_fat ,data=bfat, lambda=seq(-2.5, .5, length=200))
```

```
#ROC Curves for High Percent Fat
Define a binary variable for per_fat, disease_26, when the cutoff for abdomen is 95 and 101

disease = per_fat > 26
high_abdomen = abdomen > 95
disease_table = table(high_abdomen,disease)
addmargins(disease_table)

             disease
 high_abdomen FALSE TRUE Sum
        FALSE   107    5 112
        TRUE     36   32  68
        Sum     143   37 180

prop.out = prop.table(disease_table,1)
#chisq.test(disease_table)
specificity = prop.out[1,1]
sensitivity = prop.out[2,2]
```

```
prop.out
sensitivity
specificity
TPR = sensitivity
FPR = 1 - specificity
TPR
FPR
```

```
             disease
 high_abdomen      FALSE       TRUE
        FALSE 0.95535714 0.0446428
        TRUE  0.52941176 0.47058824
sensitivity
 [1] 0.4705882
specificity
 [1] 0.9553571
TPR = sensitivity
FPR = 1 - specificity
TPR
 [1] 0.4705882
FPR
 [1] 0.04464286
```

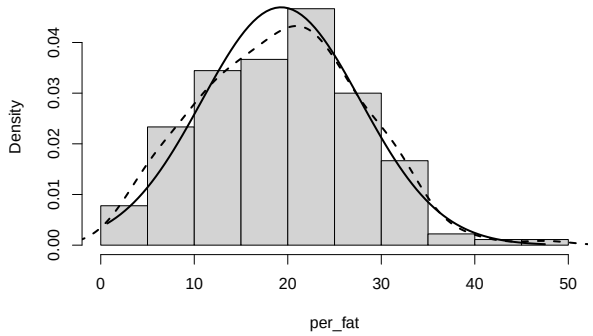
```
disease = per_fat > 26
high_abdomen = abdomen > 101
disease_table = table(high_abdomen,disease)
addmargins(disease_table)
```

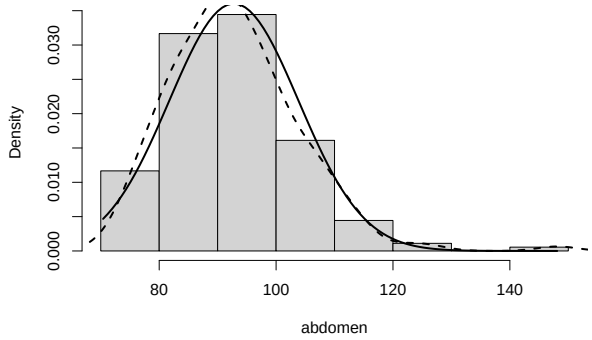
```
             disease
 high_abdomen FALSE TRUE Sum
        FALSE   133   11 144
        TRUE     10   26  36
        Sum     143   37 180
```

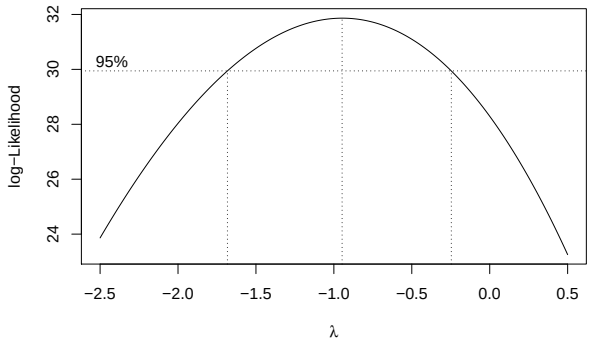
```
prop.out = prop.table(disease_table,1)
#chisq.test(disease_table)
specificity = prop.out[1,1]
sensitivity = prop.out[2,2]
prop.out
sensitivity
specificity
TPR = sensitivity
FPR = 1 - specificity
TPR
FPR
```

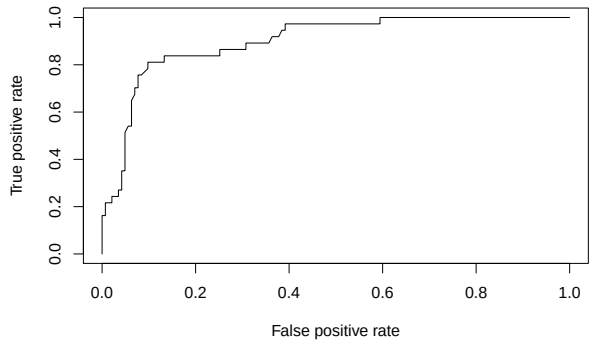
```
             disease
 high_abdomen      FALSE       TRUE
        FALSE 0.92361111 0.07638889
        TRUE  0.27777778 0.72222222
sensitivity
 [1] 0.7222222
specificity
 [1] 0.9236111
```

```
TPR = sensitivity
FPR = 1 - specificity
TPR
 [1] 0.7222222
FPR
 [1] 0.07638889

library(ROCR)
#The code for the ROC with abdomen
disease=(per_fat > 26)
pred = prediction(abdomen + thigh,disease)
perf=performance(pred, "tpr", "fpr")
plot(perf)


library(ROCR)
#The code for the ROC with abdomen
disease=(per_fat > 26)
pred = prediction(abdomen,disease)
perf=performance(pred, "tpr", "fpr")
plot(perf)

# ROC for abdomen + thigh
pred = prediction(abdomen + thigh,disease) perf=performance(pred, "tpr", "fpr")
plot(perf)
```









3.2 Continuous Methods

In this section, we will illustrate the inference methods for the two population case. I will use two data sets; the certification and the kolache data. In the certification data I will compute some correlations using the

methods described in the notes along with a two sample tests when the populations are independent. The kolache data will be used to illustrate the methods when the two populations are dependent. In this case the total unit sales for week 1 and week 2 for each store.

3.2.1 Independent Populations

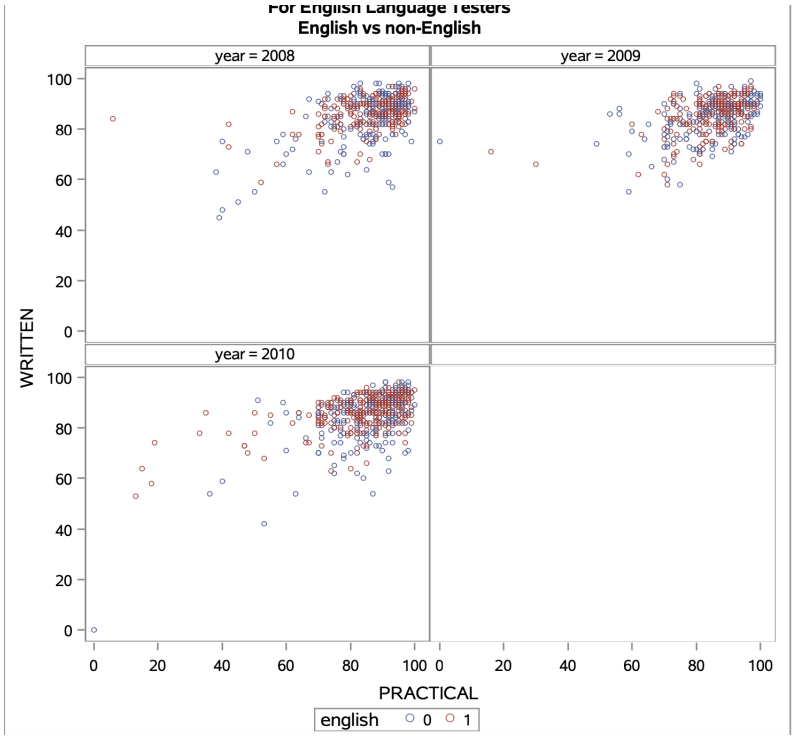
Certification Data

In this example, I have defined two new binary variables, *english* and *asian*, where english = 1 if the respondent is from an English speaking country and is 0 if otherwise. Likewise, asian =1 if the respondent is from an asian country and is 0 otherwise. The two parts to the exam; *written* and *practical* are the response variables of interest. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options nodate nonumber ps=200 ls=80 formdlim=' ';
title 'Sample Certification Data';
data cert; set ldata.certification;
english = (language = 'English');
asian = (language = 'Asian');
if year > 2007;
proc sort data=cert; by year; run;

ods graphics on;
title2 'For English Language Testers';
title3 'English vs non-English';
proc sgpanel data=cert;
panelby year;
scatter y=written x=practica/group=english;
run;

proc corr data=cert pearson spearman kendall; where english = 1;
var written practica;
run;

proc sgpanel data=cert;
panelby  english;
vbox written/category=year;run;

proc sgpanel data=cert;
panelby  english;
vbox written/category=status;run;

title4 'Year = 2008';
proc ttest data=cert; where year = 2008;
var written ;
class english;
run;
/*
proc ttest data=cert; where year = 2010;
var written practica;
class asian;
run;
```

```
*/
proc npar1way data=cert wilcoxon edf; where year=2008;
class english;
var written;
run;
/*
proc npar1way data=cert wilcoxon fp edf conover; where year = 2010;
class asian;
var written;
run;

*/
proc export data=cert
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/urinary.dbf"
replace dbms=dbf;

quit;
```

The SAS output is



Simple Statistics							
Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
WRITTEN	603	85.69320	7.07180	87.00000	53.00000	98.00000	WRITTEN
PRACTICA	603	84.10614	12.52660	87.00000	6.00000	100.00000	PRACTICAL

Pearson Correlation Coefficients, N = 603 Prob > |r| under H0: Rho=0		
	WRITTEN	PRACTICA
WRITTEN WRITTEN	1.00000	0.48990 <.0001
PRACTICA PRACTICAL	0.48990 <.0001	1.00000

Spearman Correlation Coefficients, N = 603 Prob > |r| under H0: Rho=0		
	WRITTEN	PRACTICA
WRITTEN WRITTEN	1.00000	0.39490 <.0001
PRACTICA PRACTICAL	0.39490 <.0001	1.00000

Kendall Tau b Correlation Coefficients, N = 603 Prob > |tau| under H0: Tau=0		
	WRITTEN	PRACTICA
WRITTEN WRITTEN	1.00000	0.28145 <.0001
PRACTICA PRACTICAL	0.28145 <.0001	1.00000

For English Language Testers
English vs non-English
Year = 2008

The TTEST Procedure

Variable: WRITTEN (WRITTEN)

english	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		213	84.7371	9.7693	0.6694	45.0000	98.0000
1		179	85.8939	6.6438	0.4966	59.0000	97.0000
Diff (1-2)	Pooled		-1.1568	8.4868	0.8605		
Diff (1-2)	Satterthwaite		-1.1568		0.8335		

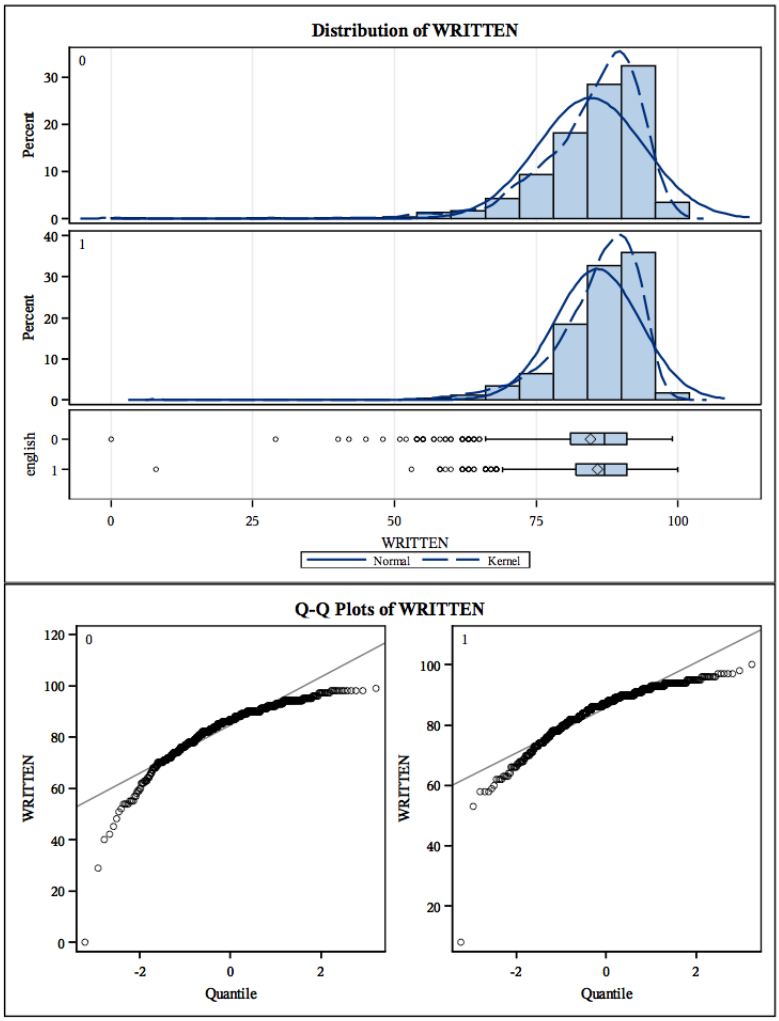
english	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		84.7371	83.4176	86.0566	9.7693	8.9213	10.7969
1		85.8939	84.9139	86.8738	6.6438	6.0195	7.4137
Diff (1-2)	Pooled	-1.1568	-2.8486	0.5351	8.4868	7.9308	9.1274
Diff (1-2)	Satterthwaite	-1.1568	-2.7956	0.4821			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	390	-1.34	0.1797
Satterthwaite	Unequal	374.47	-1.39	0.1660

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	212	178	2.16	<.0001

The TTEST Procedure

Variable: WRITTEN (WRITTEN)



Sample Certification Data
For English Language Testers
English vs non-English
Year = 2008

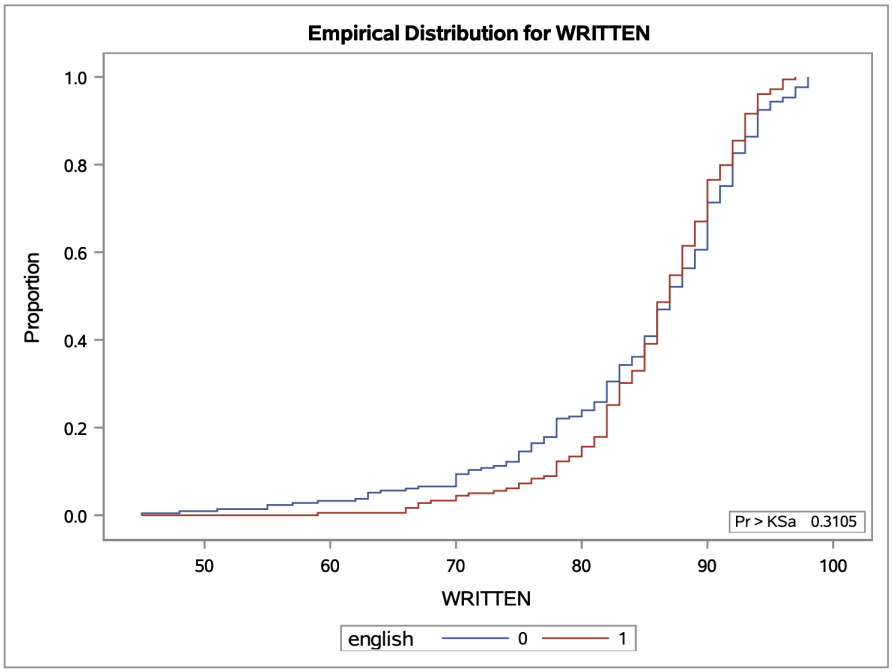
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable WRITTEN Classified by Variable english					
english	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	213	41843.50	41854.50	1115.69634	196.448357
1	179	35184.50	35173.50	1115.69634	196.561453
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
35184.50	0.0094	0.4962	0.9925	0.4962	0.9925
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
0.0001	1	0.9921

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.048692	D	0.097752
KSa	0.964051	Pr > KSa	0.3105



3.2.2 Assignment

Create a Rmarkdown file for the above analysis using the certification data

3.2.3 Dependent Populations

Kolache Sales Data

In this example, I defined a new variable, *diff* , where diff = week 2 total unit sales - week 1 total unit sales. The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options nodate nonumber ps=200 ls=80 formdlim=' ';
title1 'Kolache Sales in Large Consumer Warehouses';

data kolache; set ldata.kolache;run;

proc sort data=kolache; by week; run;

*Inference with two populations - dependent data;


title2 'Texas Data';
/*
proc univariate data = kolache normal; where state='TX'; by week; var unit_tot;
histogram / normal;
probplot;
run;
*/
title3 'Unit Total for Week 1 and Week 2';
proc sgplot data=kolache; where state='TX';
vbox unit_tot/ category=week;
run;

data a1; set kolache; week1=unit_tot; if week=1;run;
data a2; set kolache; week2=unit_tot; if week=2;run;
data paired; merge a1 a2; by store; run;

title3 'Test for UNIT_TOT FOR WEEK 1 VS WEEK 2';
proc ttest data=kolache;where state='TX';
class week;
var unit_tot;
run;

TITLE3 'Test for paired difference between Week 1 and Week 2';
proc ttest data=paired;where state='TX';
paired week1*week2;
run;

*independent non-parametric methods;
title3 'Nonparametric Tests for Week 1 vs Week 2';
proc npar1way data=kolache wilcoxon median edf;where state='TX';
class week;
var unit_tot;
```

```
run;

*dependent non-parametric methods;
title3 'Nonparametric Test for paired differences for Week 1 vs Week 2';
data kolache; set kolache; diff = w2_tot - w1_tot; run;
proc sgplot data=kolache; where state='TX';
vbox diff;
run;

title3 'Another parametric way to Test for Week 1 vs Week 2';
proc univariate data=kolache;where state='TX';
var diff;
run;
quit;
```

The SAS output is

The TTEST Procedure

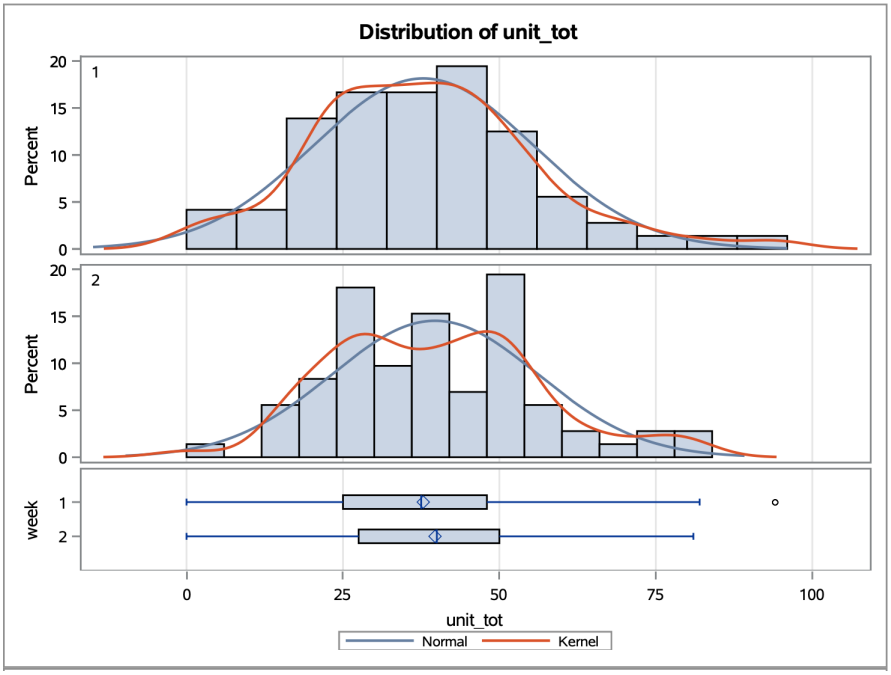
Variable: unit_tot

week	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		504	37.8333	17.5887	0.7835	0	94.0000
2		504	39.7083	16.4872	0.7344	0	81.0000
Diff (1-2)	Pooled		-1.8750	17.0469	1.0739		
Diff (1-2)	Satterthwaite		-1.8750		1.0739		

week	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		37.8333	36.2941	39.3726	17.5887	16.5658	18.7473
2		39.7083	38.2655	41.1512	16.4872	15.5284	17.5733
Diff (1-2)	Pooled	-1.8750	-3.9822	0.2322	17.0469	16.3335	17.8259
Diff (1-2)	Satterthwaite	-1.8750	-3.9823	0.2323			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1006	-1.75	0.0811
Satterthwaite	Unequal	1001.8	-1.75	0.0811

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	503	503	1.14	0.1473



Test for paired difference between Week 1 and Week 2

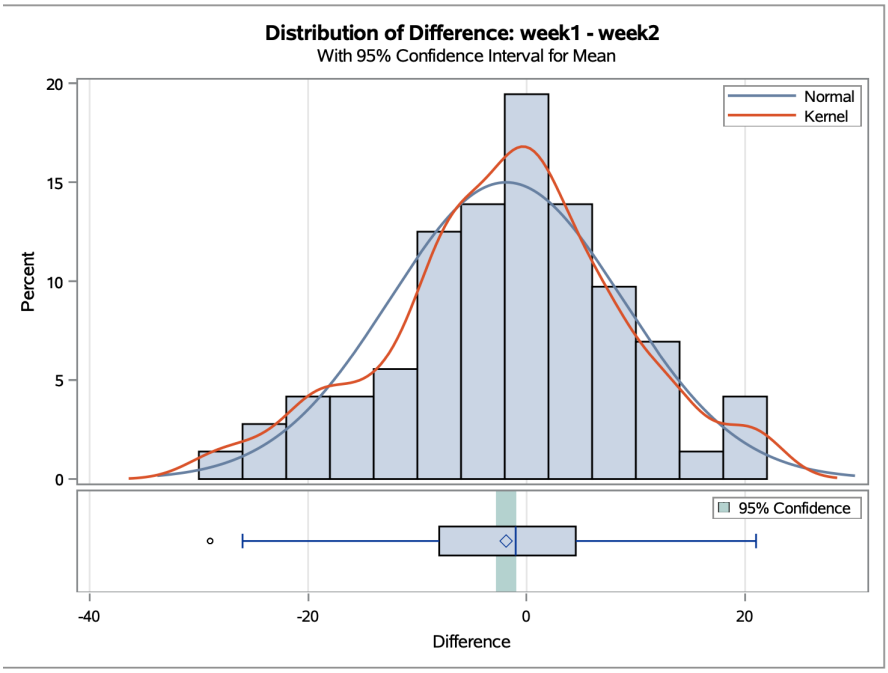
The TTEST Procedure

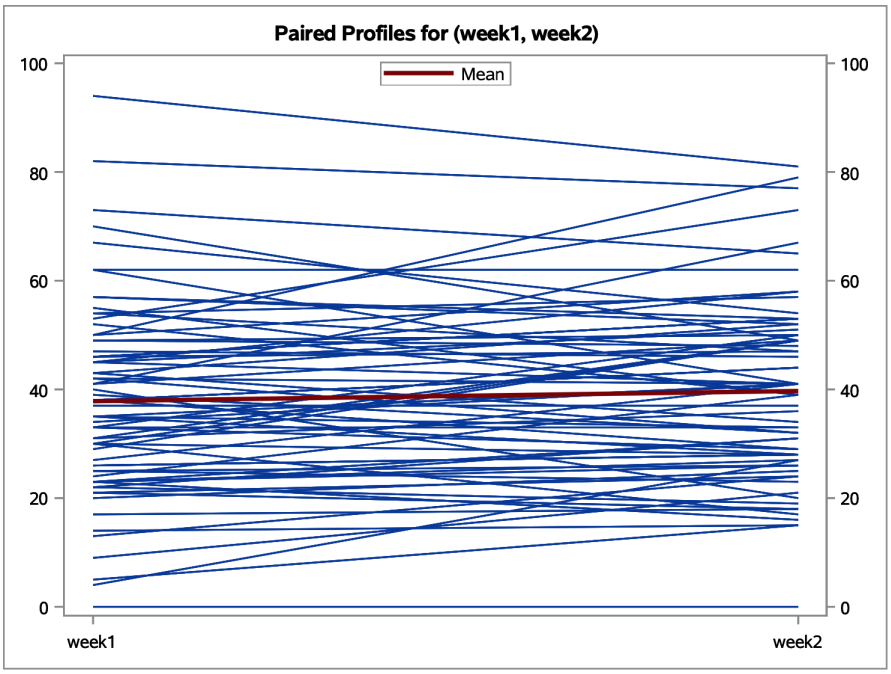
Difference: week1 - week2

N	Mean	Std Dev	Std Err	Minimum	Maximum
504	-1.8750	10.6432	0.4741	-29.0000	21.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-1.8750	-2.8064	-0.9436	10.6432	10.0243	11.3443

DF	t Value	Pr > |t|
503	-3.95	<.0001





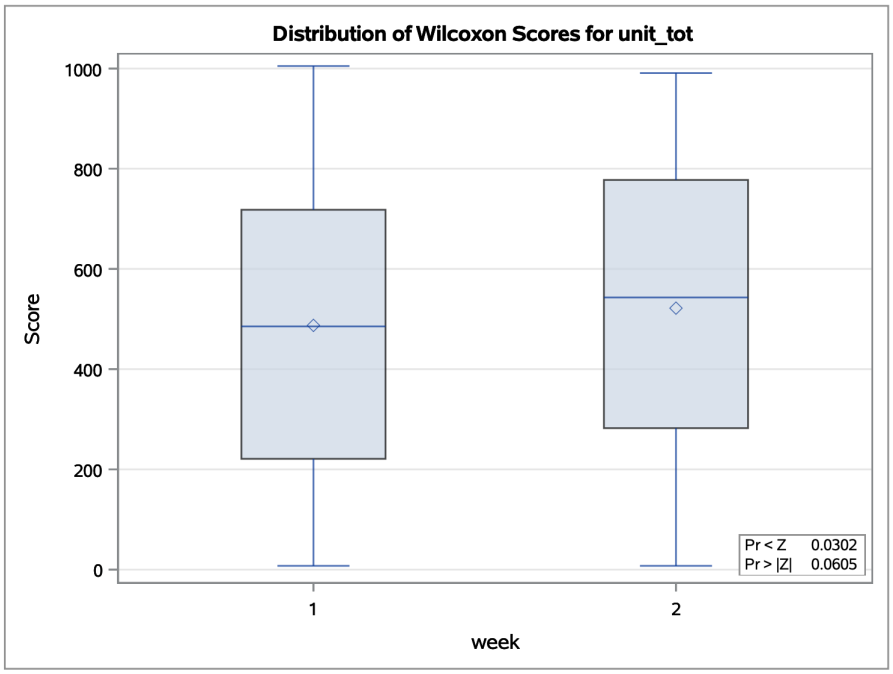
Nonparametric Tests for Week 1 vs Week 2

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable unit_tot Classified by Variable week					
week	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	504	245595.0	254268.0	4619.92711	487.291667
2	504	262941.0	254268.0	4619.92711	521.708333
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
245595.0	-1.8772	0.0302	0.0605	0.0304	0.0608
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
3.5243	1	0.0605



The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable unit_tot Classified by Variable week					
week	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	504	231.0	252.0	7.941194	0.458333
2	504	273.0	252.0	7.941194	0.541667
Average scores were used for ties.					

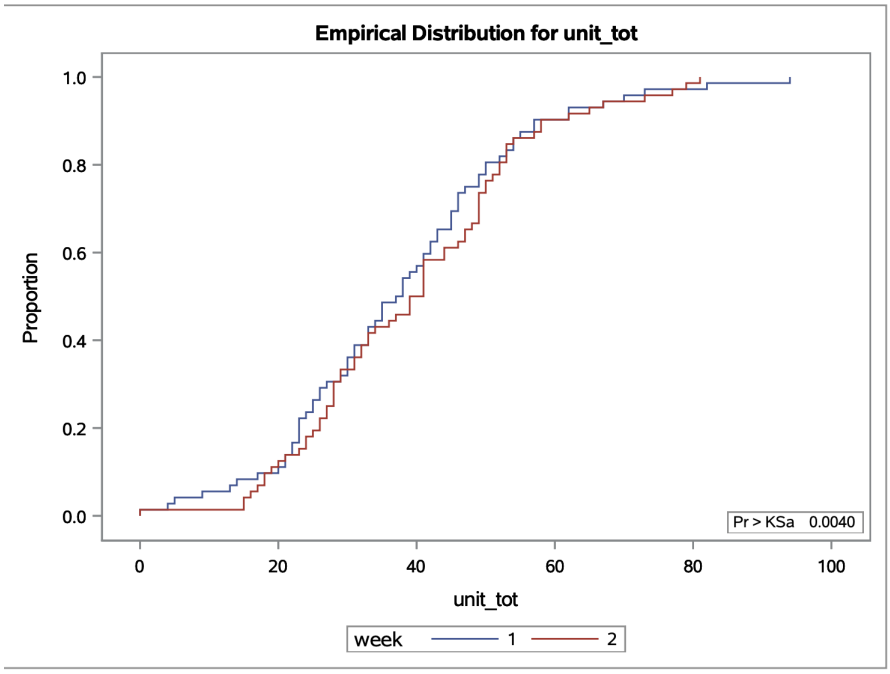
Median Two-Sample Test			
Statistic	Z	Pr < Z	Pr > |Z|
231.0000	-2.6444	0.0041	0.0082

Median One-Way Analysis		
Chi-Square	DF	Pr > ChiSq
6.9931	1	0.0082

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable unit_tot Classified by Variable week			
week	N	EDF at Maximum	Deviation from Mean at Maximum
1	504	0.736111	1.247219
2	504	0.625000	-1.247219
Total	1008	0.680556	
Maximum Deviation Occurred at Observation 568			
Value of unit_tot at Maximum = 46.0			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.055556	D	0.111111
KSa	1.763834	Pr > KSa	0.0040



3.2.4 Assignment

Create a Rmarkdown file for the above analysis using the Kolache Sales data

Chapter 4

Regression

In this chapter we consider the problem associated with finding linear models. I will use the body fat data set for most of the examples found in this chapter as the other data sets do not lend themselves to the simple linear models. I will use some of these when considering the analysis of covariance material.

4.1 Simple Linear, Polynomial, and Multiple Models

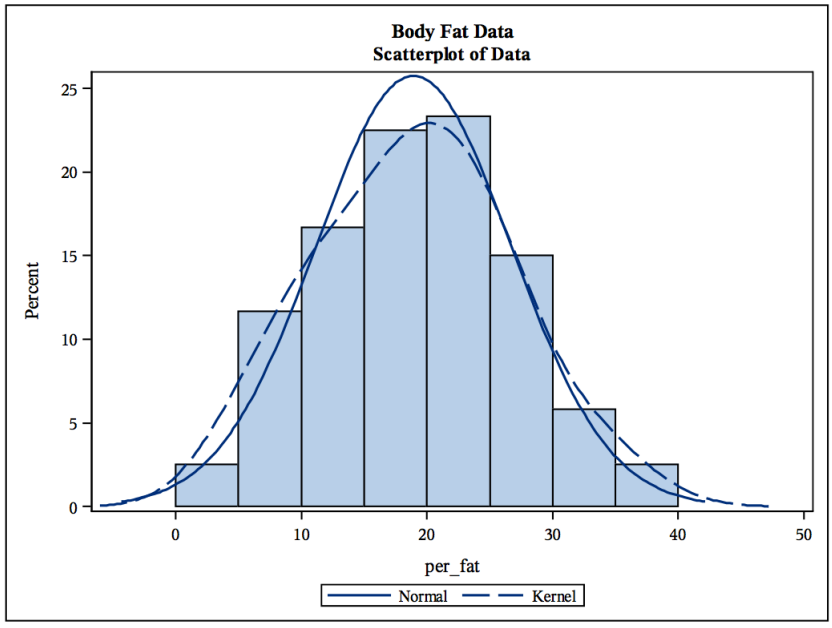
The SAS code is

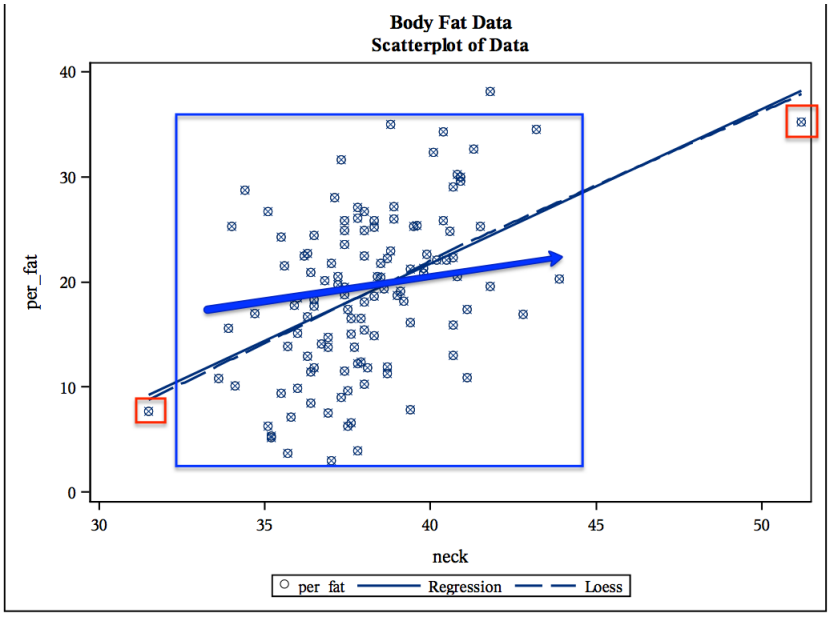
```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;
options center nodate pagesize=100 ls=80;

title1 'Body Fat Data';

data bodyfat; set ldata.bodyfat; run;

/*
Use per_fat or density as the dependent variable with a subset of the data
*/
title2 'Simple Random Sampling of size = 50';
proc surveyselect data=bodyfat
   method=srs n=50 out=new_bfat seed = 12345;
run;

data new_bfat; set new_bfat;
  neck2 = neck*neck;    abdomen2=abdomen*abdomen;
run;

title2 'Scatterplot of Data';
proc sgplot data=new_bfat;
  scatter y=per_fat x=neck ;
  reg y=per_fat x=neck;
  loess y=per_fat x=neck;
run;

proc sgplot data=new_bfat;
  scatter y=per_fat x=abdomen ;
```

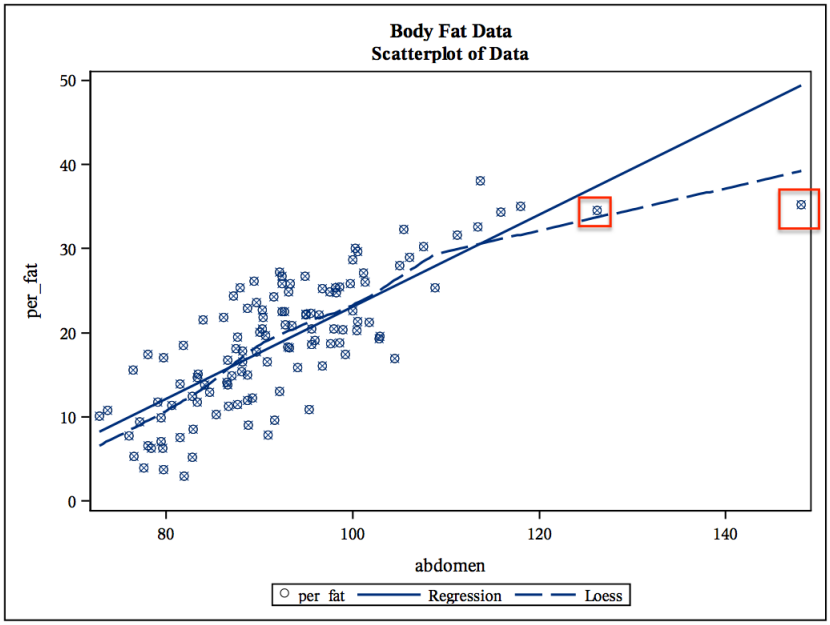
```
  reg y=per_fat x=abdomen;
  loess y=per_fat x=abdomen;
run;

proc sgplot data=new_bfat;
  histogram per_fat;
  density per_fat;
  density per_fat/ type=kernel;
run;

proc sgscatter data=new_bfat;
  matrix per_fat density thigh knee ankle biceps forearm
  wrist/diagonal=(histogram normal);run;

title2 'Simple Linear Model - Neck';
  proc reg data=new_bfat;*   plots = diagnostics;
  model per_fat=neck;
run;

title2 'Polynomial Regression';
  proc reg data=new_bfat;*  plots = (diagnostics partial);
  model per_fat=neck neck2/ press partial ss1;
run;

title2 'Multiple Regression';
  proc reg data=new_bfat;*  plots = (diagnostics partial);
  model per_fat=neck thigh abdomen/press partial ss1 ss2;
run;

title2 'Simple Linear Model - Abdomen';
  proc reg data=new_bfat   plots=(residuals DIAGNOSTICS);
  model per_fat=abdomen/ influence;
run;

title2 'Quadratic Linear Model - Abdomen';
  proc reg data=new_bfat   plots=(residuals DIAGNOSTICS);
  model per_fat=abdomen abdomen2/ ss1;

title2 'Multiple Regression - Abdomen, Thigh, Knee';
  proc reg data=new_bfat   plots = (diagnostics partial);
  model per_fat=abdomen thigh knee/ss1 ss2;
run;
quit;
```

4.1.1 Linear Least Squares Models

The SAS output is given below


Body Fat Data
Scatterplot of Data
Percent
25
20
15
10
5
0
0
10
20
30
40
50
per_fat
Normal
Kernel


Body Fat Data
Scatterplot of Data
per_fat
40
30
20
10
0
30
35
40
45
50
neck
per_fat
Regression
Loess



The REG Procedure
Model: MODEL1
Dependent Variable: per_fat

Number of Observations Read	120
Number of Observations Used	120

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	1543.41630	1543.41630	32.54	<.0001
Error	118	5597.48362	47.43630		
Corrected Total	119	7140.89992			

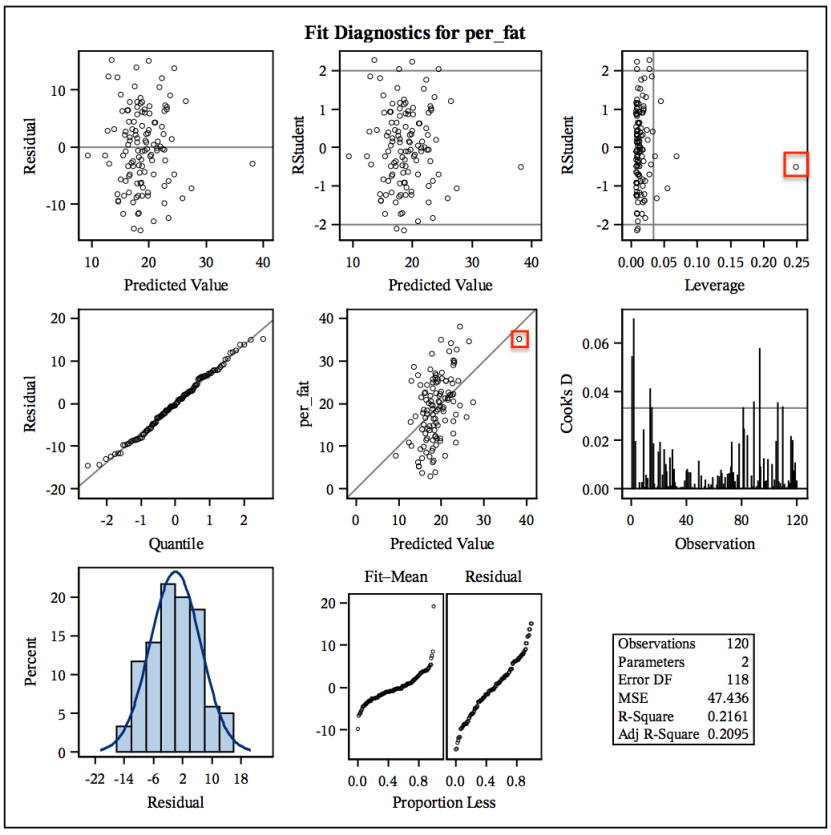
Root MSE	6.88740	**R-Square**	0.2161
Dependent Mean	18.92417	**Adj R-Sq**	0.2095
Coeff Var	36.39474		

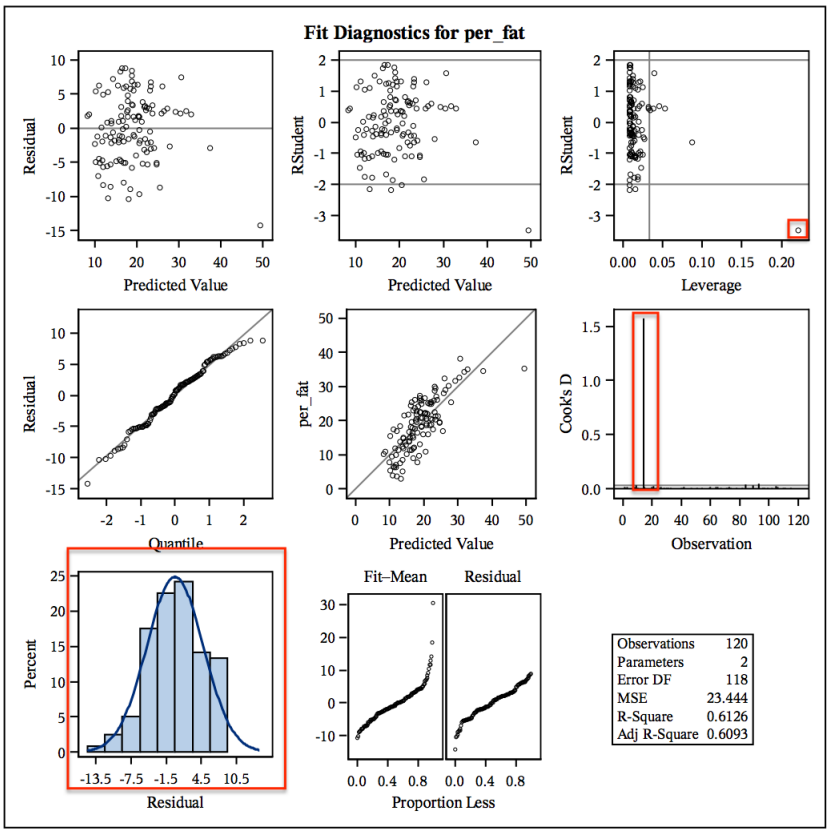
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-37.02592	9.82890	-3.77	0.0003
neck	1	1.46899	0.25753	5.70	<.0001

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	4374.53860	4374.53860	186.60	<.0001
Error	118	2766.36132	23.44374		
Corrected Total	119	7140.89992			

Root MSE	4.84187	**R-Square**	0.6126
Dependent Mean	18.92417	**Adj R-Sq**	0.6093
Coeff Var	25.58566		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-31.53197	3.72005	-8.48	<.0001
abdomen	1	0.54666	0.04002	13.66	<.0001


Fit Diagnostics for per_fat
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
per_fat
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 120
Parameters 2
Error DF 118
MSE 47.436
R-Square 0.2161
Adj R-Square 0.2095



4.1.2 Polynomial Regression

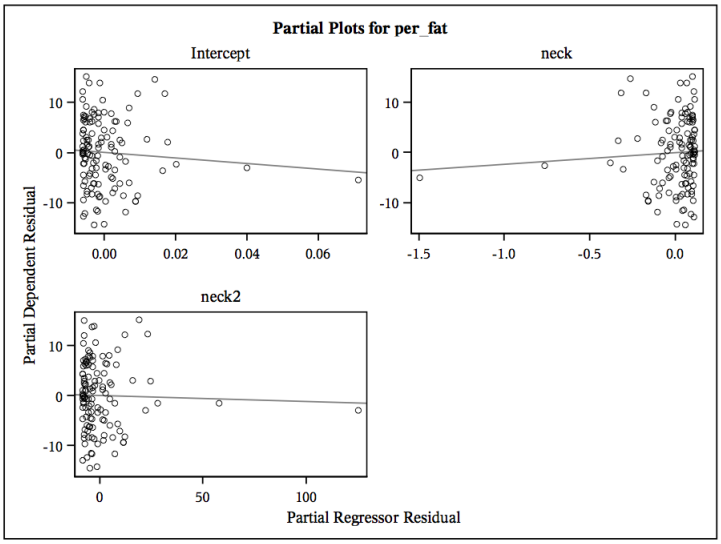
In this section I extended the above linear models by adding *neck2 = neck*neck* and *abdomen2 = abdomen*abdomen*. The results are found given below

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1547.03031	773.51515	16.18	<.0001
Error	117	5593.86961	47.81085		
Corrected Total	119	7140.89992			

Root MSE	6.91454	R-Square	0.2166
Dependent Mean	18.92417	Adj R-Sq	0.2033
Coeff Var	36.53814		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-55.31697	67.25622	-0.82	0.4125	42975
neck	1	2.39695	3.38508	0.71	0.4803	1543.41630
neck2	1	-0.01171	0.04258	-0.27	0.7839	3.61401

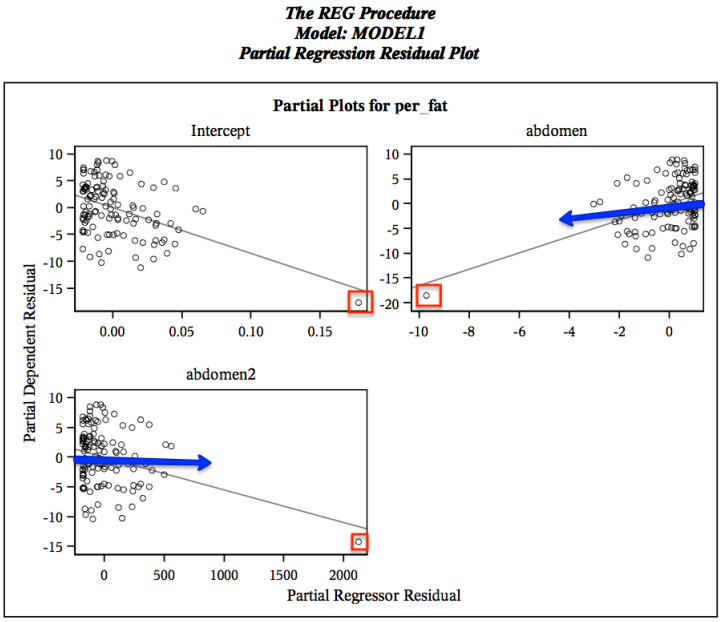
The REG Procedure
Model: MODEL1
Partial Regression Residual Plot



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	4626.13812	2313.06906	107.62	<.0001
Error	117	2514.76180	21.49369		
Corrected Total	119	7140.89992			

Root MSE	4.63613	**R-Square**	0.6478
Dependent Mean	18.92417	**Adj R-Sq**	0.6418
Coeff Var	24.49846		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**
Intercept	1	-85.61692	16.20434	-5.28	<.0001	42975
abdomen	1	1.64783	0.32413	5.08	<.0001	4374.53860
abdomen2	1	-0.00550	0.00161	-3.42	0.0009	251.59952

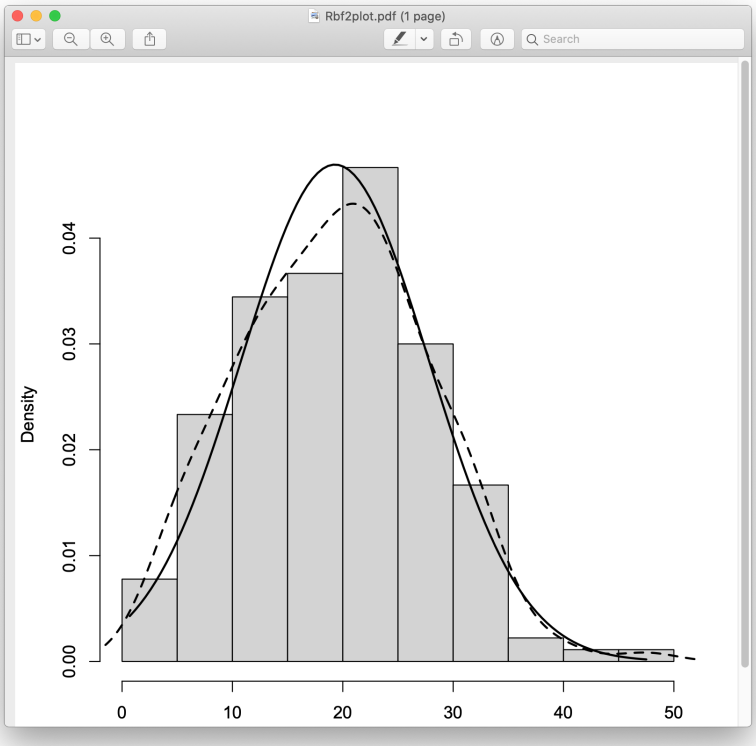


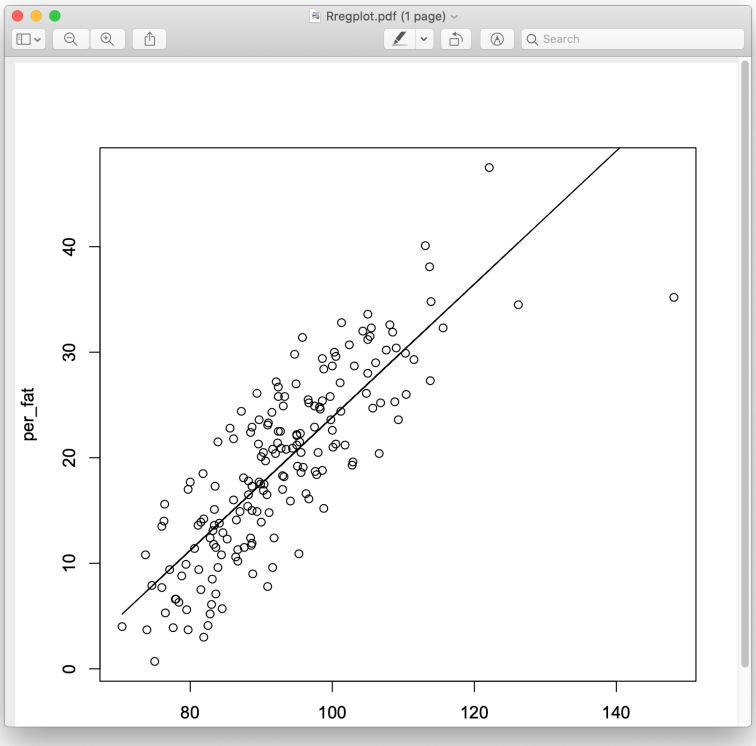
The R code is

```
#Using Body Fat Data with ds=bfat
with(bfat, hist(per_fat, main="", freq=FALSE))
with(bfat, lines(density(per_fat), main="PERCENT FAT", lty=2, lwd=2))
xvals = with(bfat, seq(from=min(per_fat), to=max(per_fat), length=100))
with(bfat, lines(xvals, dnorm(xvals, mean(per_fat), sd(per_fat)), lwd=2))

#Linear Regression
mod1 = lm(per_fat ~ abdomen, data=bfat)
summary(mod1)
Call:
lm(formula = per_fat ~ abdomen, data = bfat)

Residuals:
     Min       1Q   Median       3Q      Max
-18.9840  -3.6341  -0.0102   3.4709  10.2028

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) -39.22601    3.06172  -12.81   <2e-16 ***
abdomen       0.63072    0.03277   19.25   <2e-16 ***

Residual standard error: 4.854 on 178 degrees of freedom
Multiple R-squared:  0.6755,    Adjusted R-squared:  0.6736
F-statistic: 370.5 on 1 and 178 DF,  p-value: < 2.2e-16

covb = vcov(mod1)
coeff.mod1 = coef(mod1)
t = (coeff.mod1[2] + coeff.mod1[3] - 1)/
sqrt(covb[2,2] + covb[3,3] + 2*covb[2,3]) pvalue = 2*(1-pt(abs(t), df=mod1$df))
coef(mod1)
```

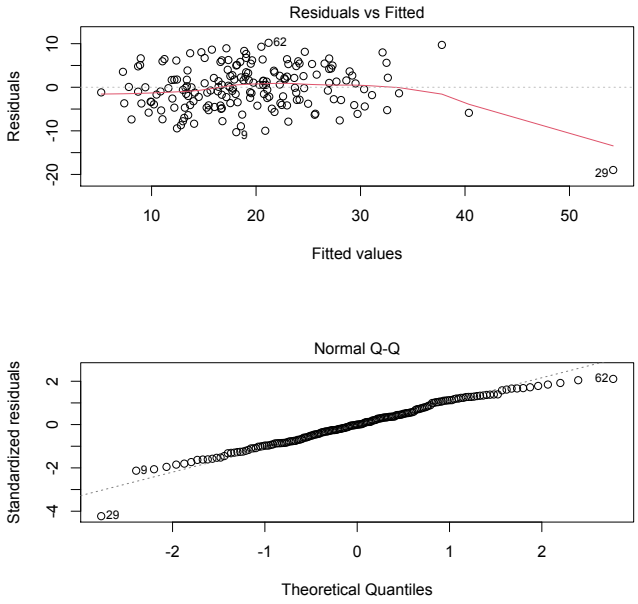
```
(Intercept)     abdomen
-39.2260089   0.6307225
covb = vcov(mod1)
covb
             (Intercept)      abdomen
(Intercept)  9.37412238 -0.099624319
abdomen     -0.09962432  0.001073763

pred.per_fat = predict(mod1)
res.per_fat = residuals(mod1)
summary(res.per_fat)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-18.98400  -3.63414  -0.01023   0.00000   3.47087  10.20279

# Residual Plots
par(mfrow=c(2,1))
> plot(mod1, which = c(1, 2))
> plot(mod1, which = c(3, 5))
```

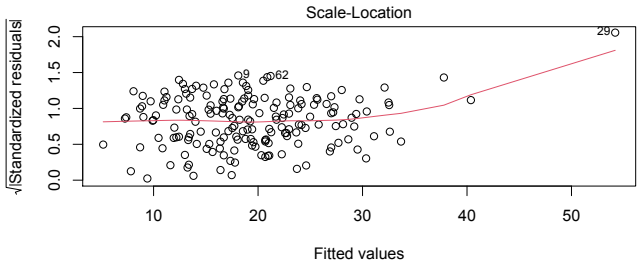
Some of the R output is

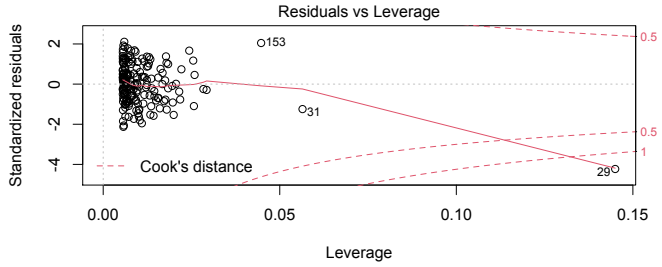




R Residual plots







4.1.3 Multiple Regression

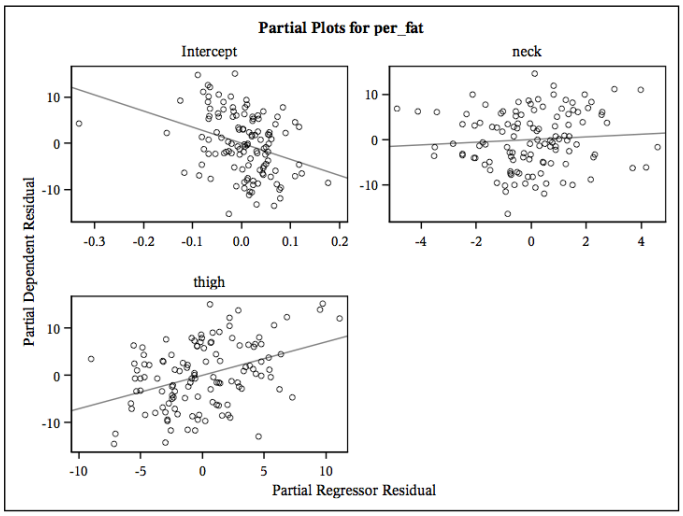
In this section I extended the above linear models by adding *thigh* to each model. The results are

Number of Observations Read	120
Number of Observations Used	120

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2402.86632	1201.43316	29.67	<.0001
Error	117	4738.03360	40.49601		
Corrected Total	119	7140.89992			

Root MSE	6.36365	R-Square	0.3365
Dependent Mean	18.92417	Adj R-Sq	0.3252
Coeff Var	33.62710		

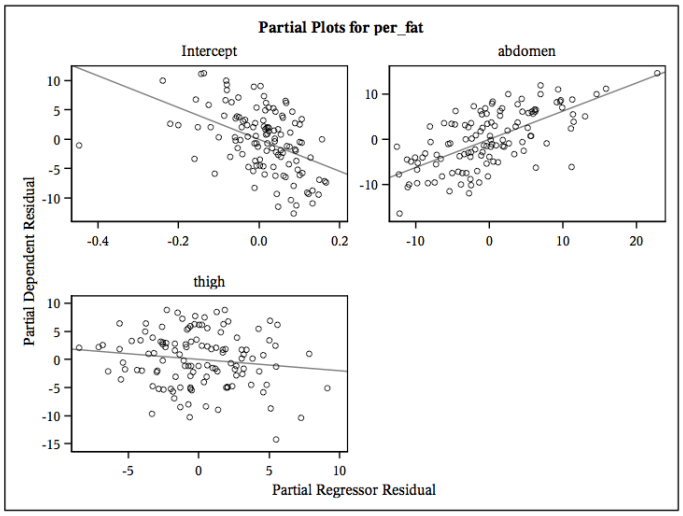
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-34.85650	9.09366	-3.83	0.0002	42975	594.97968
neck	1	0.30725	0.34672	0.89	0.3773	1543.41630	31.80188
thigh	1	0.70719	0.15351	4.61	<.0001	859.45002	859.45002



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	4425.37963	2212.68981	95.34	<.0001
Error	117	2715.52029	23.20958		
Corrected Total	119	7140.89992			

Root MSE	4.81763	**R-Square**	0.6197
Dependent Mean	18.92417	**Adj R-Sq**	0.6132
Coeff Var	25.45756		

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**	**Type II SS**
Intercept	1	-27.07782	4.77048	-5.68	<.0001	42975	747.77343
abdomen	1	0.62545	0.06648	9.41	<.0001	4374.53860	2054.31519
thigh	1	-0.19708	0.13316	-1.48	0.1416	50.84103	50.84103



The R-code is

```
# Multiple Regression
mod2 = lm(density ~ abdomen + thigh + neck)
> summary(mod2)

Call:
lm(formula = density ~ abdomen + thigh + neck)

Residuals:
      Min        1Q    Median        3Q       Max
-0.021148 -0.007087 -0.000568  0.007323  0.033771

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept)  1.1322911  0.0129757  87.262  < 2e-16 ***
abdomen     -0.0018588  0.0001274 -14.592  < 2e-16 ***
thigh        0.0001743  0.0002502   0.697    0.487
neck         0.0022306  0.0005301   4.208 4.11e-05 ***

Residual standard error: 0.01029 on 176 degrees of freedom
Multiple R-squared:  0.7102,    Adjusted R-squared:  0.7052
F-statistic: 143.8 on 3 and 176 DF,  p-value: < 2.2e-16

#Predicted values and Residuals
residual.density=residuals(mod2)
predicted.density=predict(mod2)
plot(residual.density, predicted.density)
plot(density, predicted.density)
```

The plots for the R output is not included as it is similar to the linear regression output.

4.2 Model Selection and Multicollinearity

In this section, an expanded multiple regression model is investigated by checking for collinearity and using several methods of subset selection. In the examples that follow I have used the *density measurement* as the

dependent variable. The SAS code is given in

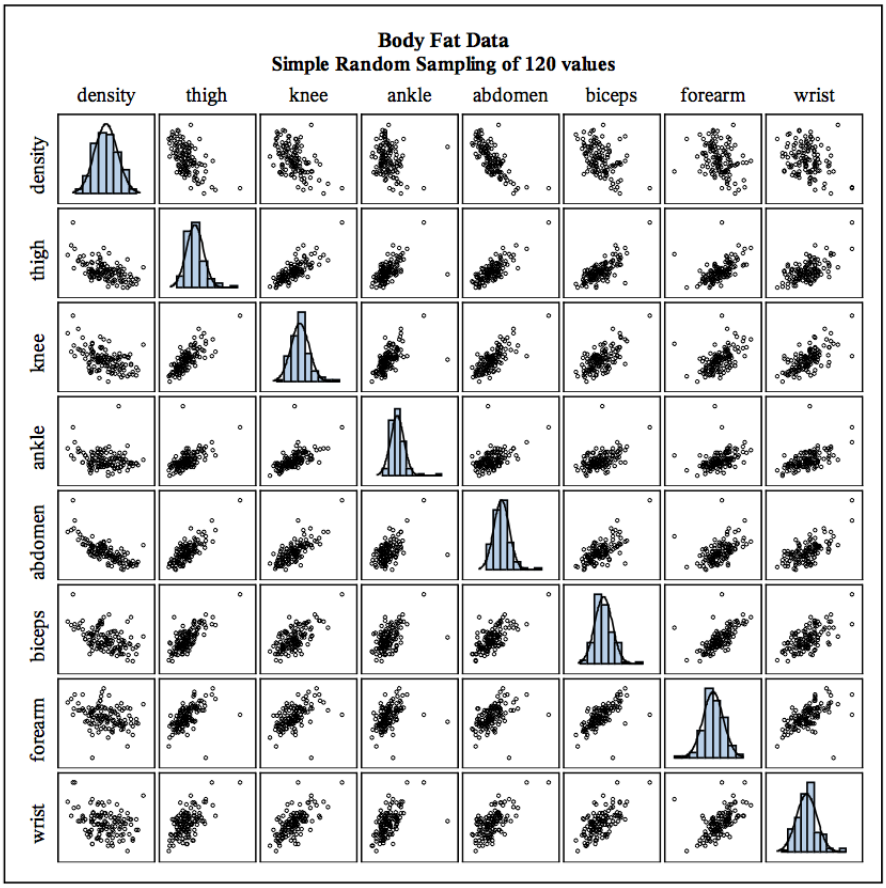
`body_fat_selection.sas`

```
options center nodate pagesize=100 ls=80;

title1 'Body Fat Data';

data bodyfat; set sasuser.bodyfat; run;

/*
Use per_fat or density as the dependent variable with a subset of the data
*/
title2 'Simple Random Sampling of 120 values';
proc surveyselect data=bodyfat
   method=srs n=120 out=new_bfat seed = 54321;
run;

proc sgplot data=new_bfat;
  histogram per_fat;
  density per_fat;
  density per_fat/ type=kernel;
run;

proc sgscatter data=new_bfat;
  matrix per_fat thigh knee ankle abdomen biceps forearm age wrist
  /diagonal=(histogram normal);
run;

title2 'Multipe Regression';
proc reg data=new_bfat  plots = (diagnostics partial);
  model per_fat=thigh knee ankle abdomen biceps forearm age wrist
        /partial   ss1 ss2;
run;

title2 'Multiple Regression';
proc reg data=new_bfat plots(only)=ridge(unpack VIFaxis=log)
         outest=b ridge=0 to 0.02 by .002;
  model per_fat=thigh knee ankle abdomen biceps forearm age wrist
         /  vif tol collin;
run;

title2 'Stepwise Regression';
proc reg data=new_bfat;*   plots = (diagnostics partial);
  model density=thigh knee ankle abdomen biceps forearm age wrist
        / selection=stepwise details=summary;
run;

title2 'Stepwise Regression';
proc reg data=new_bfat;*   plots = (diagnostics partial);
  model density=thigh knee ankle abdomen biceps forearm age  wrist
        / selection=cp best=5 details=summary;
run;
```

```
title2 'Stepwise Regression';
proc glmselect data=new_bfat   plots = all;
  model per_fat=thigh knee ankle abdomen biceps forearm  wrist
        / selection=stepwise choose=cp details=summary;
run;

title2 'Lasso Regression';
proc glmselect data=new_bfat   plots = all;
  model per_fat=thigh knee ankle abdomen biceps forearm age  wrist
        / selection=lasso choose=cp details=summary;
run;
quit;
```

4.2.1 Check for Multicollinearity

The SAS output is

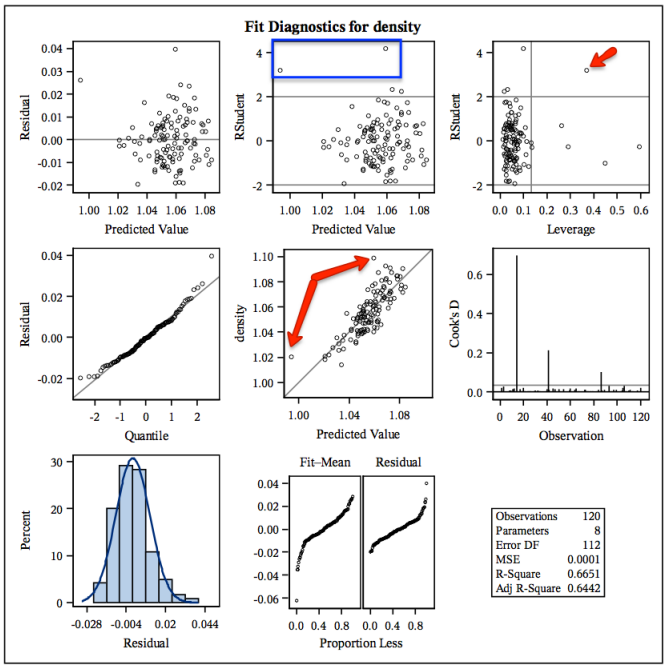


Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.02552	0.00365	31.78	<.0001
Error	112	0.01285	0.00011473		
Corrected Total	119	0.03837			

Root MSE	0.01071	R-Square	0.6651
Dependent Mean	1.05628	Adj R-Sq	0.6442
Coeff Var	1.01405		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Squared Semi-partial Corr Type I
Intercept	1	1.07033	0.02160	49.54	<.0001	133.88814	0.28161	.
thigh	1	0.00013146	0.00037077	0.35	0.7236	0.01196	0.00001442	0.31165
knee	1	0.00002183	0.00070828	0.03	0.9755	0.00030134	1.089614E-7	0.00785
ankle	1	0.00089312	0.00075687	1.18	0.2405	0.00122	0.00015976	0.03180
abdomen	1	-0.00168	0.00016847	-9.99	<.0001	0.00956	0.01146	0.24918
biceps	1	-0.00007921	0.00056478	-0.14	0.8887	0.00005384	0.00000226	0.00140
forearm	1	-0.00068757	0.00068487	-1.00	0.3176	0.00000222	0.00011564	0.00005775
wrist	1	0.00736	0.00160	4.60	<.0001	0.00242	0.00242	0.06317

Dependent Variable: density



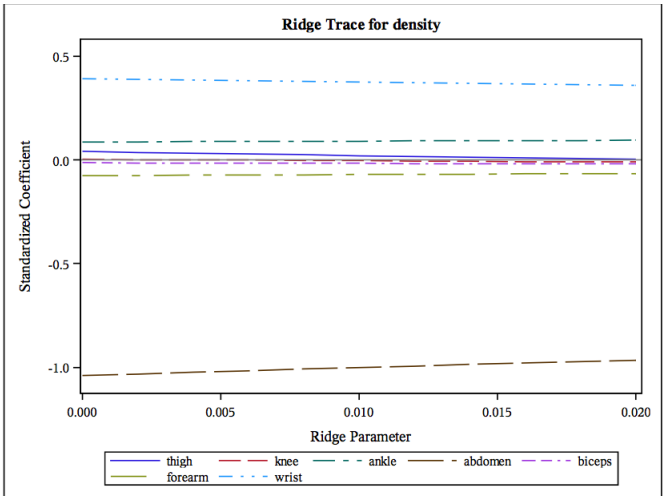
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	0.02552	0.00365	31.78	<.0001
Error	112	0.01285	0.00011473		
Corrected Total	119	0.03837			

Root MSE	0.01071	R-Square	0.6651
Dependent Mean	1.05628	Adj R-Sq	0.6442
Coeff Var	1.01405		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	1.07033	0.02160	49.54	<.0001	.	0
thigh	1	0.00013146	0.00037077	0.35	0.7236	0.22874	4.37174
knee	1	0.00002183	0.00070828	0.03	0.9755	0.28637	3.49197
ankle	1	0.00089312	0.00075687	1.18	0.2405	0.55708	1.79507
abdomen	1	-0.00168	0.00016847	-9.99	<.0001	0.27613	3.62142
biceps	1	-0.00007921	0.00056478	-0.14	0.8887	0.31279	3.19702
forearm	1	-0.00068757	0.00068487	-1.00	0.3176	0.51362	1.94695
wrist	1	0.00736	0.00160	4.60	<.0001	0.40928	2.44330

Collinearity Diagnostics		
Number	Eigenvalue	Condition Index
1	7.97832	1.00000
2	0.00937	29.17634
3	0.00397	44.80893
4	0.00299	51.66718
5	0.00181	66.36538
6	0.00165	69.45892
7	0.00106	86.85674
8	0.00082172	98.53570

Collinearity Diagnostics								
	Proportion of Variation							
Number	Intercept	thigh	knee	ankle	abdomen	biceps	forearm	wrist
1	0.00003207	0.00003053	0.00001996	0.00004853	0.00006092	0.00004505	0.00003888	0.00001754
2	0.05657	0.01860	0.00001371	0.01709	0.18896	0.01448	0.01340	0.00570
3	0.00758	0.00108	0.01401	0.15481	0.07132	0.28355	0.17053	0.00088675
4	0.09628	0.04095	0.00141	0.44888	0.21970	0.10572	0.02143	0.02104
5	0.10231	0.52357	0.03779	0.25907	0.24638	0.00539	0.07499	0.02757
6	0.10198	0.10289	0.00005287	0.01594	0.00039968	0.53192	0.69283	0.03644
7	0.29055	0.11652	0.87356	0.05479	0.12640	0.00023434	0.00049226	0.00739
8	0.34468	0.19635	0.07315	0.04937	0.14678	0.05866	0.02629	0.90095



The R-code for some of the output is

```
library(foreign)
bfat = read.dbf("new_bfat.dbf")
summary(bfat)
density = bfat$density
age=bfat$age
wt = bfat$wt
ht = bfat$ht
neck = bfat$neck
chest = bfat$chest
abdomen = bfat$abdomen
hip = bfat$hip
thigh = bfat$thigh
knee = bfat$knee
ankle = bfat$ankle
bicepts = bfat$biceps
forearm = bfat$forearm
wrist = bfat$wrist



pairs(data=bfat, ~density+hip+thigh+knee+ankle+biceps
      +forearm+wrist)
library(dplyr)
bfat.num = select(bfat,
                  density,
                  hip,
                  thigh,
                  knee,
                  ankle,
                  biceps,
                  forearm,
                  wrist)
model.null = lm(density ~ 1, data=bfat)
model.full = lm(density ~ hip+thigh+knee+abdomen+
                          biceps+ankle+forearm+wrist, data=bfat)
```

```
summary.aov(model.null)
             Df  Sum Sq   Mean Sq F value Pr(>F)
Residuals   179 0.06429 0.0003592
summary.aov(model.full)
             Df   Sum Sq  Mean Sq F value Pr(>F)
hip           1 0.025160 0.025160 250.953 <2e-16 ***
thigh         1 0.000115 0.000115   1.148 0.2855
knee          1 0.000049 0.000049   0.485 0.4871
abdomen       1 0.020780 0.020780 207.265 <2e-16 ***
biceps        1 0.000002 0.000002   0.018 0.8938
ankle         1 0.000489 0.000489   4.878 0.0285 *
forearm       1 0.000015 0.000015   0.145 0.7037
wrist         1 0.000541 0.000541   5.396 0.0214 *
Residuals   171 0.017144 0.000100
#Stepwise Model Selection
step(model.null,
     scope = list(upper=model.full),
     direction="both",
     data=bfat)
model.final = lm(formula = density ~ abdomen + hip + wrist
                 + thigh, data = bfat)
                 Start:  AIC=-1426.71
density ~ 1

          Df Sum of Sq      RSS     AIC
+ abdomen  1  0.043278 0.021016 -1626.0
+ hip      1  0.025160 0.039134 -1514.1
+ thigh    1  0.021818 0.042476 -1499.3
+ biceps   1  0.018826 0.045468 -1487.1
+ knee     1  0.016334 0.047959 -1477.5
+ wrist    1  0.009926 0.054368 -1454.9
+ forearm  1  0.009717 0.054576 -1454.2
+ ankle    1  0.004551 0.059743 -1437.9
<none>                 0.064294 -1426.7

Step:  AIC=-1625.98
density ~ abdomen

          Df Sum of Sq      RSS     AIC
+ hip      1  0.002302 0.018713 -1644.9
+ wrist    1  0.001691 0.019324 -1639.1
+ ankle    1  0.001659 0.019356 -1638.8
+ knee     1  0.001098 0.019918 -1633.6
+ thigh    1  0.000507 0.020508 -1628.4
<none>                 0.021016 -1626.0
+ biceps   1  0.000224 0.020792 -1625.9
+ forearm  1  0.000050 0.020965 -1624.4
- abdomen  1  0.043278 0.064294 -1426.7

Step:  AIC=-1644.87
density ~ abdomen + hip

          Df Sum of Sq      RSS     AIC
+ wrist    1 0.0008658 0.017848 -1651.4
```

```
+ ankle    1 0.0004826 0.018231 -1647.6
+ thigh    1 0.0003252 0.018388 -1646.0
<none>                 0.018713 -1644.9
+ knee     1 0.0000631 0.018650 -1643.5
+ biceps   1 0.0000205 0.018693 -1643.1
+ forearm  1 0.0000076 0.018706 -1642.9
- hip      1 0.0023024 0.021016 -1626.0
- abdomen  1 0.0204205 0.039134 -1514.1

Step:  AIC=-1651.39
density ~ abdomen + hip + wrist

          Df Sum of Sq      RSS     AIC
+ thigh    1 0.0003832 0.017464 -1653.3
+ biceps   1 0.0002255 0.017622 -1651.7
+ forearm  1 0.0001996 0.017648 -1651.4
<none>                 0.017848 -1651.4
+ ankle    1 0.0001110 0.017737 -1650.5
+ knee     1 0.0000071 0.017840 -1649.5
- wrist    1 0.0008658 0.018713 -1644.9
- hip      1 0.0014768 0.019324 -1639.1
- abdomen  1 0.0212597 0.039107 -1512.2

Step:  AIC=-1653.3
density ~ abdomen + hip + wrist + thigh

          Df Sum of Sq      RSS     AIC
+ ankle    1 0.0001987 0.017266 -1653.4
<none>                 0.017464 -1653.3
+ forearm  1 0.0000861 0.017378 -1652.2
+ biceps   1 0.0000626 0.017402 -1652.0
+ knee     1 0.0000120 0.017452 -1651.4
- thigh    1 0.0003832 0.017848 -1651.4
- wrist    1 0.0009238 0.018388 -1646.0
- hip      1 0.0016772 0.019142 -1638.8
- abdomen  1 0.0215229 0.038987 -1510.8

Step:  AIC=-1653.36
density ~ abdomen + hip + wrist + thigh + ankle

          Df Sum of Sq      RSS     AIC
<none>                 0.017266 -1653.4
- ankle    1 0.0001987 0.017464 -1653.3
+ forearm  1 0.0001014 0.017164 -1652.4
+ biceps   1 0.0000627 0.017203 -1652.0
+ knee     1 0.0000001 0.017265 -1651.4
- wrist    1 0.0004691 0.017735 -1650.5
- thigh    1 0.0004710 0.017737 -1650.5
- hip      1 0.0014644 0.018730 -1640.7
- abdomen  1 0.0199227 0.037188 -1517.2

Call:
lm(formula = density ~ abdomen + hip + wrist + thigh + ankle,
    data = bfat)
```

```
Coefficients:
(Intercept)      abdomen          hip        wrist
  1.1015316   -0.0020339    0.0011987    0.0025030
      thigh        ankle
 -0.0007388    0.0009059
```

```
summary(model.final)

Call:
lm(formula = density ~ abdomen + hip + wrist + thigh + ankle,
    data = bfat)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0233941 -0.0076618  0.0008563  0.0065818  0.0269610

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept)  1.1015316  0.0161119  68.368  < 2e-16 ***
abdomen     -0.0020339  0.0001435 -14.170  < 2e-16 ***
hip          0.0011987  0.0003120   3.842 0.000171 ***
wrist        0.0025030  0.0011511   2.174 0.031028 *
thigh       -0.0007388  0.0003391  -2.179 0.030700 *
ankle        0.0009059  0.0006401   1.415 0.158805
Residual standard error: 0.009961 on 174 degrees of freedom
Multiple R-squared:  0.7315,    Adjusted R-squared:  0.7237
F-statistic: 94.79 on 5 and 174 DF,  p-value: < 2.2e-16
```

4.3 Analysis of 1960 US Crime Data

A well-described dataset with "crime rate" in 47 states of the USA for 1960 as the response variable and 15 predictor variables. This may seem an odd choice as more recent data with richer sets of predictor features are certainly available. The choice of a limited dataset with a limited number of features, some of which are linearly correlated, will allow us to more easily interpret differences between the regression methods as well as evaluate the usefulness of principal component analysis in combination with the three different regression methods[1].

4.3.1 R-code for LASSO and Model Selection

```
options("repos" = c(CRAN = "https://cran.rstudio.com"))
library(MASS)
library(ggplot2)
library(gridExtra)
if(!require(glmnet)){install.packages("glmnet")}
if(!require(caret)){install.packages("caret")}

#Load Crime Data
dfC <- read.csv("http://www.statsci.org/data/general/uscrime.txt", sep="\t")
#print summary
```

[1]This material is found in Stepwise, Lasso, and Elastic Net by Michael Boerrigter in July 06, 2017

```r
summary(dfC)

#Normalize Data
#The variable So is really a factor variable indicating
#whether a state is considered Southern or not and,
#therefore, should not be scaled.
colNames <- colnames(dfC[,-2])[1:14]

#Define a function to normalize data
normalize <- function(df, cols) {
  result <- df # make a copy of the input data frame

  for (j in cols) { # each specified col
    m <- mean(df[,j]) # column mean
    std <- sd(df[,j]) # column (sample) sd

    result[,j] <- sapply(result[,j], function(x) (x - m) / std)
  }
  return(result)
}

#normalize predictors except 'So'
dfC.norm <- normalize(dfC, colNames)

#Perform a backwards stepwise regression with cross-validation.
#The output of all the iterations is suppressed.
#The results of the final step are given in the
#commented section of the next code block.
ctrl <- trainControl(method = "repeatedcv", number = 5,
                     repeats = 5)

lmFit_Step <- train(Crime ~ ., data = dfC.norm, "lmStepAIC",
              scope = list(lower = Crime~1, upper = Crime~.),
              direction = "backward", trControl=ctrl)
##
##Step:  AIC=503.93
##.outcome ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob

##       Df Sum of Sq     RSS    AIC
##<none>               1453068 503.93
##- M.F   1    103159 1556227 505.16
##- U1    1    127044 1580112 505.87
##- Prob  1    247978 1701046 509.34
##- U2    1    255443 1708511 509.55
##- M     1    296790 1749858 510.67
##- Ed    1    445788 1898855 514.51
##- Ineq  1    738244 2191312 521.24
##- Po1   1   1672038 3125105 537.93

#Develop a new model with the eight variables found
#with stepwise regression.
#This yields an adjusted R-squared of 0.7444.

mod_Step = lm(Crime ~ M.F+U1+Prob+U2+M+Ed+Ineq+Po1, data = dfC.norm)
```

```
summary(mod_Step)
Call:
lm(formula = Crime ~ M.F + U1 + Prob + U2 + M + Ed + Ineq + Po1,
    data = dfC.norm)

Residuals:
    Min      1Q  Median      3Q     Max
-444.70 -111.07    3.03  122.15  483.30

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   905.09      28.52  31.731  < 2e-16 ***
M.F            65.83      40.08   1.642  0.10874
U1           -109.73      60.20  -1.823  0.07622 .
Prob          -86.31      33.89  -2.547  0.01505 *
U2            158.22      61.22   2.585  0.01371 *
M             117.28      42.10   2.786  0.00828 **
Ed            201.50      59.02   3.414  0.00153 **
Ineq          244.70      55.69   4.394 8.63e-05 ***
Po1           305.07      46.14   6.613 8.26e-08 ***

Residual standard error: 195.5 on 38 degrees of freedom
Multiple R-squared:  0.7888,    Adjusted R-squared:  0.7444
F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10

#Next use cross-validation to see how good this model
#really is. Because we only have 47 data points,
# lets  use 47-fold cross-validation (equivalent to
#leave-one-out cross-validation).
SStot <- sum((dfC.norm$Crime - mean(dfC.norm$Crime))^2)
totsse <- 0

for(i in 1:nrow(dfC.norm)) {
  mod_Step_i = lm(Crime ~ M.F+U1+Prob+U2+M+Ed+Ineq+Po1, data = dfC.norm[-i,])
  pred_i <- predict(mod_Step_i, newdata=dfC.norm[i,])
  totsse <- totsse + ((pred_i - dfC.norm[i,16])^2)
}

R2_mod <- 1 - totsse/SStot
R2_mod
       1
0.667621

#Note that in the previous model, the p-value for M.F is higher than
#0.1.  Lets  see what happens if it is removed.

mod_Step = lm(Crime ~ U1+Prob+U2+M+Ed+Ineq+Po1, data = dfC.norm)
summary(mod_Step)

Call:
lm(formula = Crime ~ U1 + Prob + U2 + M + Ed + Ineq + Po1, data = dfC.norm)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-520.76 -105.67    9.53  136.28  519.37

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   905.09      29.14  31.062  < 2e-16 ***
U1            -63.86      54.48  -1.172   0.2482
Prob          -84.83      34.61  -2.451   0.0188 *
U2            134.13      60.71   2.209   0.0331 *
M             134.20      41.70   3.218   0.0026 **
Ed            244.38      54.07   4.520 5.62e-05 ***
Ineq          264.65      55.52   4.767 2.61e-05 ***
Po1           314.89      46.73   6.738 4.91e-08 ***
Residual standard error: 199.8 on 39 degrees of freedom
Multiple R-squared:  0.7738,    Adjusted R-squared:  0.7332
F-statistic: 19.06 on 7 and 39 DF,  p-value: 8.805e-11

#Now it appears that U1 is not all that significant.
# Lets  remove it as well and re-run the model,

mod_Step = lm(Crime ~ Prob+U2+M+Ed+Ineq+Po1, data = dfC.norm)
summary(mod_Step)

#Next, use cross-validation to evaluate this latest model.
#The results show that the simpler model has an
#R-squared of 0.666 which is almost the same compared to
#the model which included M.F and U1. This would suggest
#that these features may be ommitted from the final model.

SStot <- sum((dfC.norm$Crime - mean(dfC.norm$Crime))^2)
totsse <- 0

for(i in 1:nrow(dfC.norm)) {
  mod_Step_i = lm(Crime ~ Prob+U2+M+Ed+Ineq+Po1, data = dfC.norm[-i,])
  pred_i <- predict(mod_Step_i,newdata=dfC.norm[i,])
  totsse <- totsse + ((pred_i - dfC.norm[i,16])^2)
}
R2_mod <- 1 - totsse/SStot
R2_mod

#The same result can be obtained with  step  as shown
#below (output of code not shown).

model <- lm(Crime ~ ., data = dfC.norm)
step(model, direction = "backward")

mod3 =lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
   data = dfC.norm)
summary(mod3)

Call:
lm(formula = Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob,
    data = dfC.norm)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-444.70 -111.07    3.03  122.15  483.30

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   905.09      28.52  31.731  < 2e-16 ***
M             117.28      42.10   2.786  0.00828 **
Ed            201.50      59.02   3.414  0.00153 **
Po1           305.07      46.14   6.613 8.26e-08 ***
M.F            65.83      40.08   1.642  0.10874
U1           -109.73      60.20  -1.823  0.07622 .
U2            158.22      61.22   2.585  0.01371 *
Ineq          244.70      55.69   4.394 8.63e-05 ***
Prob          -86.31      33.89  -2.547  0.01505 *


Residual standard error: 195.5 on 38 degrees of freedom
Multiple R-squared:  0.7888,    Adjusted R-squared:  0.7444
F-statistic: 17.74 on 8 and 38 DF,  p-value: 1.159e-10


##
##LASSO Regression
##
#Prepare Data for Lasso
#building lasso

XP=data.matrix(dfC.norm[,-16])
YP=data.matrix(dfC.norm$Crime)
lasso=cv.glmnet(x=as.matrix(dfC.norm[,-16]),
                y=as.matrix(dfC.norm$Crime), alpha=1,
                nfolds = 5, type.measure="mse",
                family="gaussian")
coef(lasso, s=lasso$lambda.min)

##16 x 1 sparse Matrix of class "dgCMatrix"
##                1
##(Intercept) 890.529769
##M            93.625851
##So           42.756304
##Ed          147.747936
##Po1         300.403501
##Po2           .
##LF            .
##M.F          56.647483
##Pop          -0.351305
##NW            7.822136
##U1          -46.220197
##U2           83.637150
##Wealth       16.416623
##Ineq        204.642645
##Prob        -84.355911
##Time          .
```

```
#Fit a new model with the  remaining  ten variables.
#The adjusted R-squared is slightly lower as compared
#to that obtained with stepwise regression.

mod_lasso = lm(Crime ~So+M+Ed+Po1+LF+M.F+NW+U2+Ineq+Prob,
            data = dfC.norm)
summary(mod_lasso)
Call:
lm(formula = Crime ~ So + M + Ed + Po1 + LF + M.F + NW + U2 +
    Ineq + Prob, data = dfC.norm)

Residuals:
    Min      1Q  Median      3Q     Max
-410.55 -121.42    5.76  110.54  550.24

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)  863.284     53.510  16.133  < 2e-16 ***
So           122.792    129.896   0.945  0.35080
M            113.614     49.962   2.274  0.02903 *
Ed           188.755     64.750   2.915  0.00608 **
Po1          333.470     49.353   6.757 6.86e-08 ***
LF            25.162     49.588   0.507  0.61496
M.F           31.513     43.179   0.730  0.47021
NW            -2.883     59.595  -0.048  0.96169
U2            72.881     41.795   1.744  0.08973 .
Ineq         229.121     68.845   3.328  0.00203 **
Prob         -99.145     40.243  -2.464  0.01867 *


Residual standard error: 206.6 on 36 degrees of freedom
Multiple R-squared:  0.7767,    Adjusted R-squared:  0.7147
F-statistic: 12.53 on 10 and 36 DF,  p-value: 5.374e-09

#Use cross-validation to evaluate the model. Note that
#four of the variables, namely So, LF, M.F., and NW do
#not appear to be significant. In fact, without
#these variables, the model is the same as that obtained
#with stepwise regression.

SStot <- sum((dfC.norm$Crime - mean(dfC.norm$Crime))^2)
totsse <- 0

for(i in 1:nrow(dfC.norm)) {
  mod_lasso_i = lm(Crime ~ So+M+Ed+Po1+LF+M.F+NW+U2+Ineq+Prob, data = dfC.norm[-i,])
  pred_i <- predict(mod_lasso_i,newdata=dfC.norm[i,])
  totsse <- totsse + ((pred_i - dfC.norm[i,16])^2)
}

R2_mod <- 1 - totsse/SStot
R2_mod
##         1
## 0.5840486
```

```
##
##Elastic Net
##

R2=c()

for (i in 0:10) {
  mod_elastic = cv.glmnet(x=as.matrix(dfC.norm[,-16]),y=as.matrix(dfC.norm$Crime),
              alpha=i/10,nfolds = 5,type.measure="mse",family="gaussian")


  #The deviance(dev.ratio ) shows the percentage of deviance explained,
  #(equivalent to r squared in case of regression)

  R2 = cbind(R2, mod_elastic$glmnet.fit$dev.ratio[which(mod_elastic$glmnet.fit$lambda == mod_elastic$lambda.min)])

}


R2

          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.7570353 0.7733668 0.7593456 0.7068747 0.7357399
          [,6]      [,7]      [,8]      [,9]     [,10]
[1,] 0.7608016 0.7851519 0.7582144 0.7854984 0.7569563
         [,11]
[1,] 0.7922808

alpha_best = (which.max(R2)-1)/10
alpha_best
## [1] 1

#The alpha values are varied in steps of 0.1,
#from 0 to 1, and subsequently the resultant
#R-Squared values are calculated.


#Use the   best   alpha value to build the model.

Elastic_net=cv.glmnet(x=as.matrix(dfC.norm[,-16]),
y=as.matrix(dfC.norm$Crime),
alpha=alpha_best,nfolds=5,
type.measure="mse",family="gaussian")

#Output the coefficients of the variables selected by Elastic Net

coef(Elastic_net, s=Elastic_net$lambda.min)

## 16 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) 888.760971
## M            82.888559
## So           47.952148
## Ed          116.058564
```

```
## Po1          310.189409
## Po2            .
## LF             2.228622
## M.F           49.818594
## Pop            .
## NW             3.623195
## U1           -13.837819
## U2            46.698552
## Wealth         .
## Ineq         175.547425
## Prob         -80.415268
## Time           .

#The Elastic Net selects 11 variables compared to 10 in
#Lasso and 8 in stepwise. Next, compare how this new model
#performs compared to the Lasso and Stepwise models.

mod_Elastic_net = lm(Crime ~So+M+Ed+Po1+M.F+Po2+NW+U2+Ineq+Prob, data = dfC.norm)
summary(mod_Elastic_net)

Call:
lm(formula = Crime ~ So + M + Ed + Po1 + M.F + Po2 + NW + U2 +
    Ineq + Prob, data = dfC.norm)

Residuals:
    Min      1Q  Median      3Q     Max
-396.35 -106.93   -7.52   99.99  575.49

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   867.57      51.04  16.999  < 2e-16 ***
So            110.20     121.33   0.908  0.36980
M             107.28      49.29   2.176  0.03617 *
Ed            204.10      62.97   3.241  0.00256 **
Po1           551.82     280.24   1.969  0.05667 .
M.F            36.89      39.44   0.935  0.35588
Po2          -228.85     289.14  -0.792  0.43384
NW             10.27      58.22   0.176  0.86095
U2             61.61      37.16   1.658  0.10604
Ineq          226.18      67.98   3.327  0.00203 **
Prob         -103.57      39.72  -2.607  0.01319 *

Residual standard error: 205.5 on 36 degrees of freedom
Multiple R-squared:  0.779,     Adjusted R-squared:  0.7176
F-statistic: 12.69 on 10 and 36 DF,  p-value: 4.527e-09

##

#The R-squared appears to be similar to that obtained
#with Lasso and Stepwise regression. Lets
#use cross-validation to evaluate the model.

SStot <- sum((dfC.norm$Crime - mean(dfC.norm$Crime))^2)
totsse <- 0
```

```
for(i in 1:nrow(dfC.norm)) {
  mod_lasso_i = lm(Crime ~ So+M+Ed+Po1+Po2+M.F+LF+Pop+NW+U1+U2+Wealth+Ineq+Prob+Time, data = dfC.norm[-i,])
  pred_i <- predict(mod_lasso_i,newdata=dfC.norm[i,])
  totsse <- totsse + ((pred_i - dfC.norm[i,16])^2)
}
R2_mod <- 1 - totsse/SStot
R2_mod
##        1
## 0.485607

#The resultant R- squared value is a much worse
#cross-validated R-squared estimate because for
#most of the variables, the p-values seemed to
#indicate they were not significant. If the
#insignificant variables would be remove, one would
#end up with a model akin to that obtained with
#Lasso or Stepwise regression.


#write data to SAS
tosas = data.frame(dfC.norm)
library(foreign)
write.dbf(tosas,"tosas.dbf")
```

4.3.2 R code for Collinearity

The R-code[2] for identifying collinearity issues is

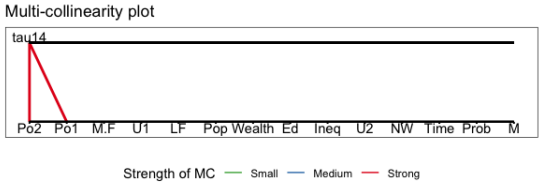
```
# Source of function is
# http://highstat.com/Books/BGS/GAMM/RCodeP2/HighstatLibV6.R
#
#############################################################################
#VIF FUNCTION.
#To use:  corvif(YourDataFile)
corvif <- function(dataz) {
  dataz <- as.data.frame(dataz)
  #correlation part
  #cat("Correlations of the variables\n\n")
  #tmp_cor <- cor(dataz,use="complete.obs")
  #print(tmp_cor)

  #vif part
  form    <- formula(paste("fooy ~ ",paste(strsplit(names(dataz)," "),
                           collapse=" + ")))
  dataz   <- data.frame(fooy=1,dataz)
  lm_mod  <- lm(form,dataz)

  cat("\n\nVariance inflation factors\n\n")
  print(myvif(lm_mod))
}
corvif(dfC.norm)
```

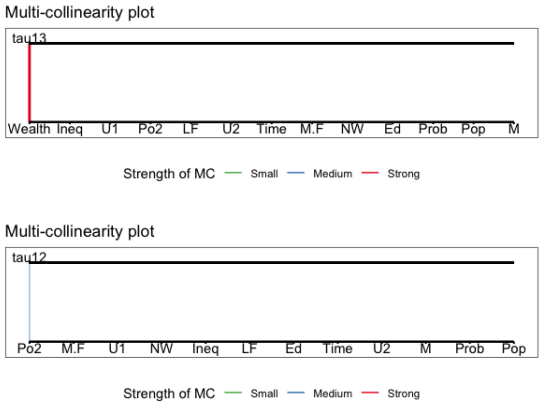
[2]The mcvis code came from https://github.com/leaffur/mcvis

```
#Variance inflation factors

#GVIF
#M        3.306095
#So       5.342896
#Ed       6.584317
#Po1    115.805131
#Po2    116.737306
#LF       3.737122
#M.F      3.875256
#Pop      2.563147
#NW       4.737546
#U1       6.438623
#U2       5.770691
#Wealth  10.822692
#Ineq    11.343460
#Prob     3.223199
#Time     2.734422
#Crime    5.078379

## An easier method is
library(usdm)
install.packages("usdm")
library("usdm", lib.loc="~/Library/R/4.0/library")
library(usdm)
vif(dfC.norm)

#Variables        VIF
#1          M   3.306095
#2         So   5.342896
#3         Ed   6.584317
#4        Po1 115.805131
#5        Po2 116.737306
#6         LF   3.737122
#7        M.F   3.875256
#8        Pop   2.563147
#9         NW   4.737546
#10        U1   6.438623
#11        U2   5.770691
#12    Wealth  10.822692
#13      Ineq  11.343460
#14      Prob   3.223199
#15      Time   2.734422
#16     Crime   5.078379

#####################################
## Install mcvis package
#####################################
devtools::install_github("kevinwang09/mcvis")
#####################################
#Use Crime Data
plot(mcvis_result)
library(usdm)
```

```
vif(dfC.norm)
M=dfC.norm$M
Ed=dfC.norm$Ed
Po1=dfC.norm$Po1
Po2=dfC.norm$Po2
LF=dfC.norm$LF
M.F=dfC.norm$M.F
Pop=dfC.norm$Pop
NW=dfC.norm$NW
U1=dfC.norm$U1
U2=dfC.norm$U2
Wealth=dfC.norm$Wealth
Ineq=dfC.norm$Ineq
Prob=dfC.norm$Prob
Time=dfC.norm$Time
summary(dfC.norm)
X = cbind(M,Ed,Po1,Po2,LF,M.F,Pop,NW,U1,
          U2,Wealth,Ineq,Prob,Time)
X
mcvis_result = mcvis(X = X)
mcvis_result
plot(mcvis_result)
#
# Remove Po1
X = cbind(M,Ed,Po2,LF,M.F,Pop,NW,U1,
          U2,Wealth,Ineq,Prob,Time)
X
mcvis_result = mcvis(X = X)
mcvis_result
plot(mcvis_result)

# Remove Wealth
X = cbind(M,Ed,Po2,LF,M.F,Pop,NW,U1,
          U2,Ineq,Prob,Time)
X
mcvis_result = mcvis(X = X)
mcvis_result
plot(mcvis_result)
```

Some of the mcvis plots are





4.3.3 SAS Code for Model Selection and collinearity Issues

```
libname ldata "/folders/myfolders/Large Data Sets/SAS Data Sets";
options nodate nonumber ps=200 ls=80 formdlim=' ';

   title 'US Crime in 1960';
   title2 'Original Data';
   title3 ' ';

data crime; set ldata.crime_1960; run;
proc contents short; run;

title3 'Plots for the Crime Variable';
proc sgplot data=crime;
histogram crime;
density crime;
density crime/ type=kernel;
run;

title3 'Linear Regression';
proc sgplot data=crime;
scatter y = crime x = Po1;
reg y = crime x = Po1;
run;

proc reg data=crime plots=diagnostics;
model crime = Po1;
run;

title3 'Scatterplot for Crime vs some other variables';
proc sgscatter data=crime;
  matrix crime Ed Ineq LF M M_F
  NW Po1 Po2 Pop /diagonal=(histogram normal);run;

title3 'Scatterplot for Crime vs some other variables';
proc sgscatter data=crime;
  matrix crime Prob Shape Time U1 U2 Wealth /diagonal=(histogram normal);run;
```

```
title3 'Correlations';
proc corr data=crime pearson nosimple noprob;
var crime Ed Ineq LF M M_F NW Po1
    Po2 ;
run;

title3 'Correlations';
proc corr data=crime pearson nosimple noprob;
var crime  Prob Shape
    Time U1 U2 Wealth;
run;

title3 'Multiple Regression';
proc reg data=crime  plots=none;
model crime = Ed Ineq LF M M_F NW Po1 Po2 Pop Prob Shape
              Time U1 U2 Wealth/ss1 ss2;
run;

title3 'Multiple Regression with Ridge Plots';
  proc reg data=crime plots(only)=ridge(unpack VIFaxis=log)
 outest=b ridge=0 to 0.2 by .02;
;*   plots = (diagnostics partial);
model crime = Ed Ineq LF M M_F NW Po1 Po2 Pop Prob Shape
              Time U1 U2 Wealth/  vif tol  collinoint;
run;

title3 'Stepwise Regression';
  proc reg data=crime plots=none;*   plots = (diagnostics partial);
model crime = Ed Ineq LF M M_F NW Po1 Po2 Pop Prob Shape
              Time U1 U2 Wealth/ selection=stepwise details=summary;
run;

title3 '5 Best Stepwise Regression with Mallows Po2 removed';
  proc reg data=crime plots=none;*   plots = (diagnostics partial);
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=cp best=5 details=summary;
run;

title3 'Lasso Regression Po2 removed';
  proc glmselect data=crime   plots = all;
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=lasso(choose = cp steps=5)
                                 hierarchy=single stb details=summary;
run;

title3 'ElasticNet Regression Po2 removed';
  proc glmselect data=crime   plots = all;
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=elasticnet(choose = cp steps=5)
                                 hierarchy=single stb details=summary;
run;

title3 'Stepwise Regression Po2 removed';
```

```
  proc glmselect data=crime   plots = all;
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=stepwise(choose = validate
                                 select = sl)
                                 hierarchy=single stb details=summary;
 partition fraction(validate=0.3 test=0.2);
run;

title3 'Lasso Regression Po2 removed';
  proc glmselect data=crime   plots = all;
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=lasso(choose = validate steps=5)
                                 hierarchy=single stb details=summary;
 partition fraction(validate=0.3 test=0.2);
run;

title3 'ElasticNet Regression Po2 removed';
  proc glmselect data=crime   plots = all;
model crime = Ed Ineq LF M M_F NW Po1 Pop Prob Shape
              Time U1 U2 Wealth/ selection=elasticnet(choose = validate steps=5)
                                 hierarchy=single stb details=summary;
 partition fraction(validate=0.3 test=0.2);
run;

quit;
```

4.4 Quantile Regression

In this section have provided two examples. One with the certification dat sand the other with the Texas Childhood Obesity data.

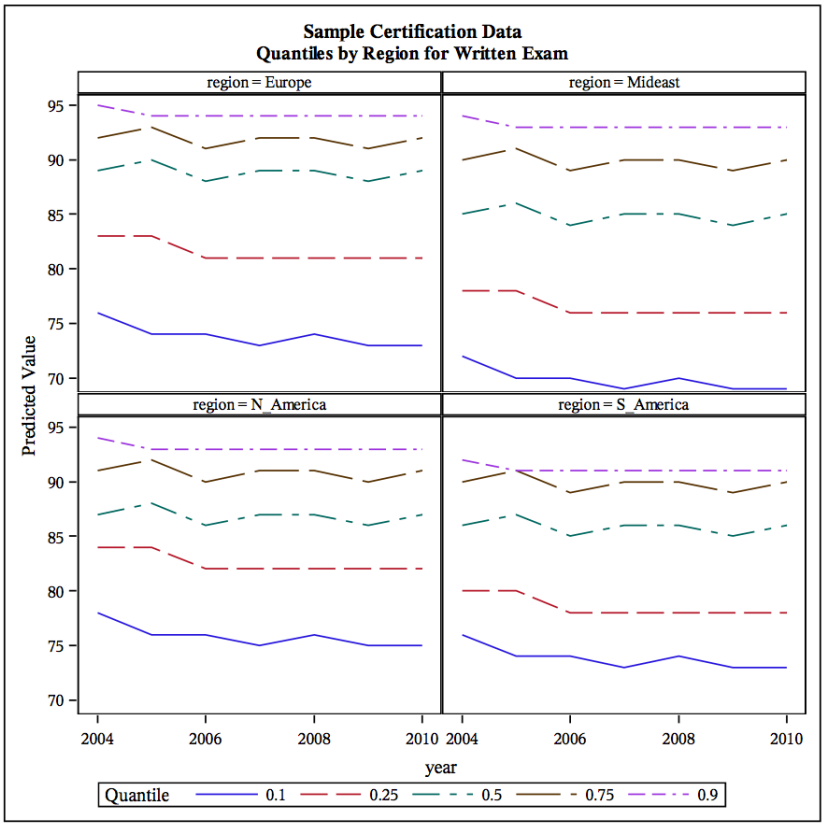
4.4.1 Certification Example

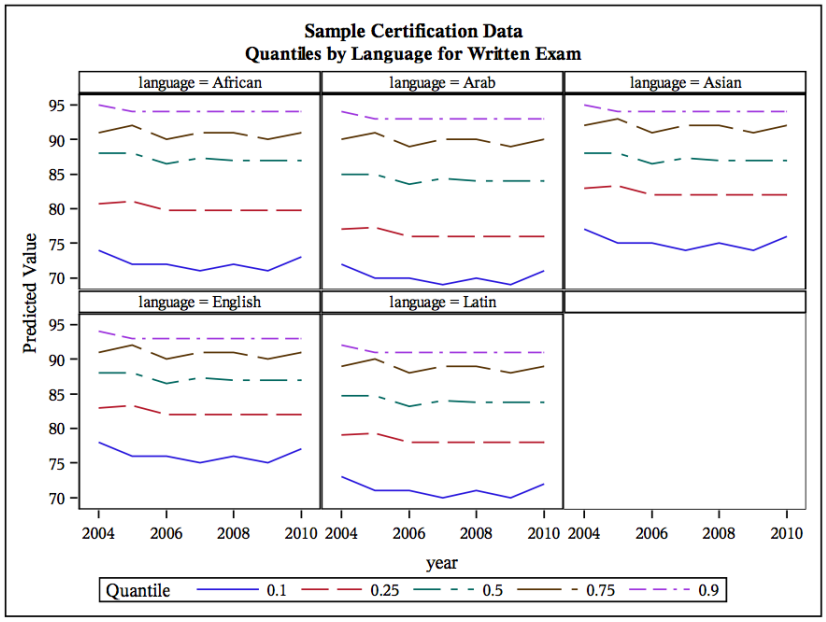
In this example I am finding the quartile regression for the *written exam scores* as a function of year and either *language* or *region*. The SAS code is given in

`certification-quantreg.sas`

The SAS output is

Sample Certification Data
Quantiles by Region for Written Exam





4.4.2 Obesity Example

In this example I am finding the quartile regression for the *bmi* as a function of age, gender and county. Since the CDC is mainly interested BMI values at the .85, .95, and .975 quantiles these are the ones that are presented in this analysis. The SAS code is given in

`texas-obesity-quantreg.sas`

```
options nodate nonumber ps=200 ls=80 formdlim=' ';

   title 'Obesity - All Counties';
/*
   There are two data sets
       1. CDC BMI age/gender adjusted by CDC 2000
       2. Combined Texas preschool data
*/

   * define work data for cdc data files;
   * the permanent dataset is sasuser.cdc_bmi;
data cdc; set sasuser.cdc_bmi;
if agemos=24 then agemos=23.9;
age=floor(agemos);
if age < 13 then N_AGE = 0;
if age > 12 and age < 25 then N_AGE = 1;
if age > 24 and age < 37 then N_AGE = 2;
if age > 36 and age < 49 then N_AGE = 3;
if age > 48 and age < 61 then N_AGE = 4;
if age > 60 then N_AGE = 5;
if sex=1 then Gender= 'M';
if sex=2 then Gender= 'F'; drop sex;
```

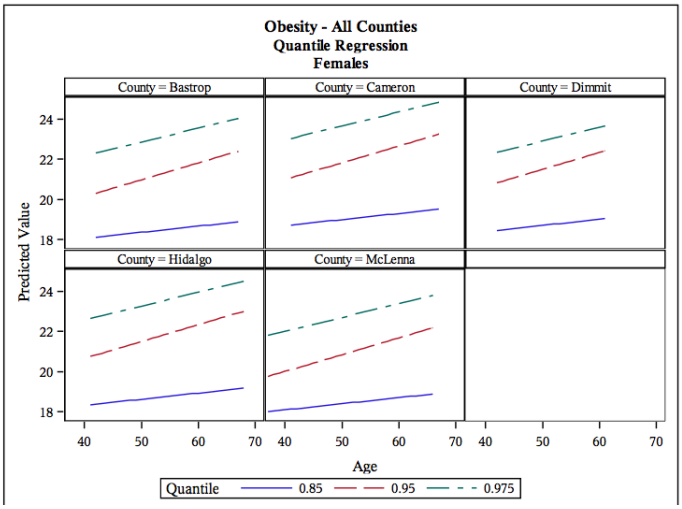
```
run;
*proc contents data=cdc; run;

* define work data for texas_obesity data files;
* the permanent dataset is sasuser.new_combined_clean;
* year 2002 is removed due to the limited number of observations for that year;

data texas_obesity; set sasuser.new_combined_clean;
if year > 2002;
if gender ne '.';
run;


*Quantile Regression;
title2 "Quantile Regression";
title3 'Model 1 ';

proc quantreg data=texas_obesity;
class county gender;
where year > 2002 and year < 2009 and N_age > 2;
model bmi = county gender age
    / nosummary quantile=0.85,0.95,0.975;
output out=outp pred=p/columnwise;
   run;
proc sort data=outp; by quantile county gender age; run;
proc means data=outp noprint; var p;output out=aout; by quantile county gender age;run;
data ab; set aout; keep county gender age p quantile; if _STAT_='MEAN'; run;
*proc print data=ab;run;

data new_ab; set ab;* ab_85 ab_95 ab_975;* by quantile; run;

title3 'Females';
proc sgpanel data=new_ab; where gender='F';
panelby county;
   series  x=age y=p/group=QUANTILE;
run;

proc sgpanel data=new_ab;where gender='F';
panelby quantile;
   series  x=age y=p/group=county;
run;

title3 'Males';
proc sgpanel data=new_ab; where gender='M';
panelby county;
   series  x=age y=p/group=QUANTILE;
run;

proc sgpanel data=new_ab;where gender='M';
panelby quantile;
   series  x=age y=p/group=county;
run;
quit;
```

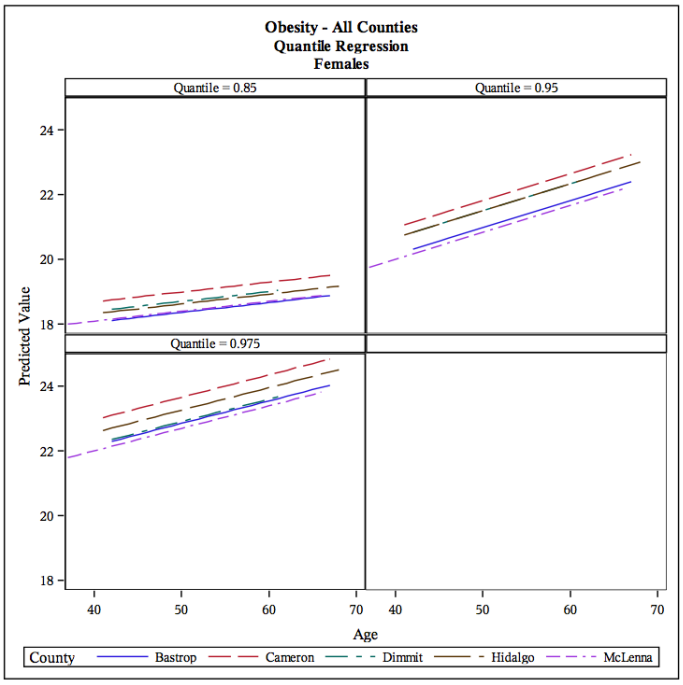
The SAS output is

Quantile and Objective Function	
Quantile	0.85
Objective Function	12495.8852
Predicted Value at Mean	18.4920

Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept		1	16.9371	0.3260	16.2981	17.5760	51.96	<.0001
county	Bastrop	1	-0.0387	0.2056	-0.4417	0.3644	-0.19	0.8508
county	Cameron	1	0.5924	0.1385	0.3210	0.8639	4.28	<.0001
county	Dimmit	1	0.3033	0.1407	0.0275	0.5791	2.16	0.0311
county	Hidalgo	1	0.2242	0.1007	0.0269	0.4215	2.23	0.0259
county	McLenna	0	0.0000	0.0000	0.0000	0.0000	.	.
Gender	F	1	-0.0675	0.0710	-0.2067	0.0717	-0.95	0.3417
Gender	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Age		1	0.0305	0.0066	0.0175	0.0435	4.61	<.0001




Obesity - All Counties
Quantile Regression
Females
Quantile = 0.85
Quantile = 0.95
Quantile = 0.975
Predicted Value
24
22
20
18
40
50
60
70
Age
County
Bastrop
Cameron
Dimmit
Hidalgo
McLenna

Chapter 5

K Population Methods

The methods presented in this chapter are for the one-way ANOVA type problems when the data are parametric (normal) and non-parametric.

I will use the certification data to illustrate these methods. In the original program I create two new variables; *language* and *region* in which the countries are classified according to the native language groups and regions of the world. In each case $k > 2$ groups have been created for the purpose of using these data to illustrate the methods found in this chapter.

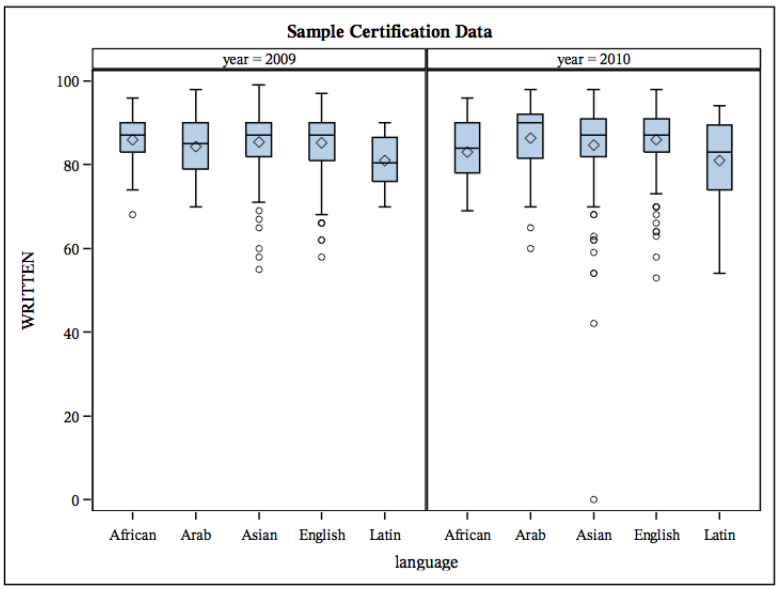
5.1 Certification Data for K Population Methods

The SAS code is

```
libname ldata  '/folders/myfolders/Large Data Sets/SAS Data Sets/' ;

options nodate nonumber ps=200 ls=80 formdlim=' ';
title 'Sample Certification Data';
data cert; set ldata.certification;
english = (language = 'English');
asian = (language = 'Asian');
proc sort data=cert; by year;
run;

ods graphics on;

proc sgpanel data=cert; where year > 2008;
panelby  year;
vbox written/category=language;run;

ods graphics off;
proc anova data=cert; where year > 2008;
class language;
model written = language;
means language / tukey;
means language / waller;
means language / sidak;
means language / dunnett ('English');
run;
```

```
ods graphics on;
proc glm data=cert; where year > 2008;
class language;
model written = language;
lsmeans language / adjust=tukey;
lsmeans language / adjust=sidak;
lsmeans language / adjust=dunnett pdiff=control('English');
run;

ods graphics off;
proc npar1way data=cert wilcoxon savage edf;where year > 2008;
class language;
var written;
run;

*/
proc export data=cert
outfile = "/folders/myfolders/Large Data Sets/SAS Data Sets/certification.dbf"
replace dbms=dbf;

quit;
```

The SAS output is



The ANOVA Procedure

Dependent Variable: WRITTEN WRITTEN

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	744.73374	186.18343	2.65	0.0322
Error	855	60091.93952	70.28297		
Corrected Total	859	60836.67326			

R-Square	Coeff Var	Root MSE	WRITTEN Mean
0.012242	9.847980	8.383494	85.12907

Source	DF	Anova SS	Mean Square	F Value	Pr > F
language	4	744.7337353	186.1834338	2.65	0.0322

Waller-Duncan K-ratio t Test for WRITTEN

Note: This test minimizes the Bayes risk under additive loss and certain other assumptions.

Kratio	100
Error Degrees of Freedom	855
Error Mean Square	70.28297
F Value	2.65
Critical Value of t	2.20311
Minimum Significant Difference	3.3307
Harmonic Mean of Cell Sizes	61.50085

Note: Cell sizes are not equal.

Means with the same letter are not significantly different.				
Waller Grouping		**Mean**	**N**	**language**
	A	85.608	424	English
	A			
	A	85.367	60	Arab
	A			
	A	84.997	308	Asian
	A			
B	A	84.250	32	African
B				
B		81.000	36	Latin

Dunnett's t Tests for WRITTEN

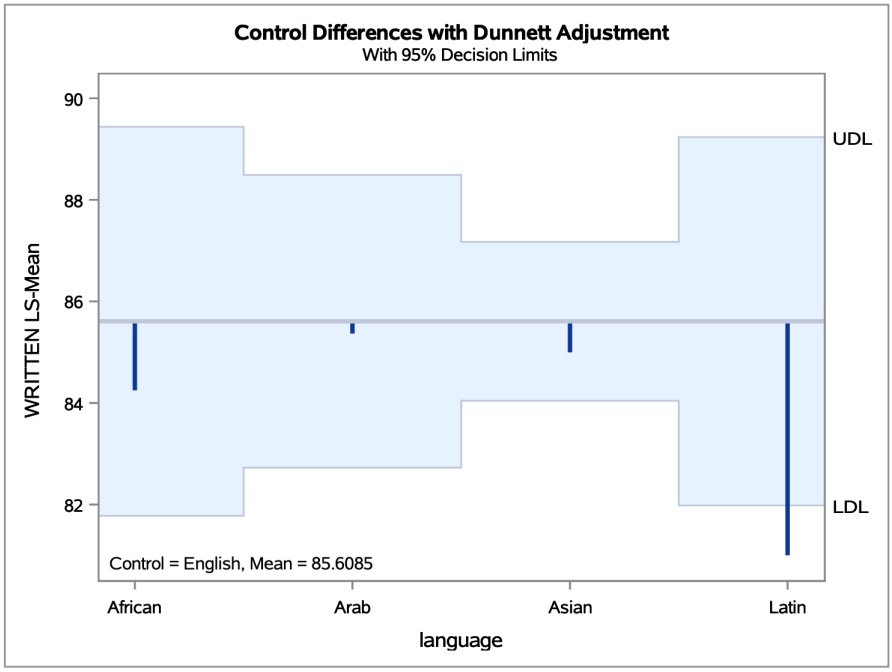
Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	855
Error Mean Square	70.28297
Critical Value of Dunnett's t	2.49194

Comparisons significant at the 0.05 level are indicated by *.**

language Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Arab - English	-0.2418	-3.1234	2.6397	
Asian - English	-0.6117	-2.1758	0.9523	
African - English	-1.3585	-5.1884	2.4714	
Latin - English	-4.6085	-8.2352	-0.9818	***

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett



The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

language	WRITTEN LSMEAN	LSMEAN Number
African	84.2500000	1
Arab	85.3666667	2
Asian	84.9967532	3
English	85.6084906	4
Latin	81.0000000	5

Least Squares Means for effect language
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: WRITTEN

i/j	1	2	3	4	5
1		0.9738	0.9892	0.9029	0.5006
2	0.9738		0.9979	0.9996	0.0983
3	0.9892	0.9979		0.8666	0.0538
4	0.9029	0.9996	0.8666		0.0138
5	0.5006	0.0983	0.0538	0.0138	

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Sidak

language	WRITTEN LSMEAN	LSMEAN Number
African	84.2500000	1
Arab	85.3666667	2
Asian	84.9967532	3
English	85.6084906	4
Latin	81.0000000	5

Least Squares Means for effect language
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: WRITTEN

i/j	1	2	3	4	5
1		0.9996	1.0000	0.9912	0.6915
2	0.9996		1.0000	1.0000	0.1287
3	1.0000	1.0000		0.9818	0.0672
4	0.9912	1.0000	0.9818		0.0159
5	0.6915	0.1287	0.0672	0.0159	

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable WRITTEN Classified by Variable language					
language	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
African	32	12498.0	13776.0	1376.49978	390.562500
Asian	308	134865.0	132594.0	3486.83090	437.873377
English	424	185804.0	182532.0	3635.90194	438.216981
Latin	36	10503.0	15498.0	1456.46766	291.750000
Arab	60	26560.0	25830.0	1852.70643	442.666667
Average scores were used for ties.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
12.9265	4	0.0116

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable WRITTEN Classified by Variable language			
language	N	EDF at Maximum	Deviation from Mean at Maximum
African	32	0.281250	0.466197
Asian	308	0.194805	-0.070762
English	424	0.162736	-0.743373
Latin	36	0.444444	1.473643
Arab	60	0.283333	0.654504
Total	860	0.198837	
Maximum Deviation Occurred at Observation 405			
Value of WRITTEN at Maximum = 79.0			

Kolmogorov-Smirnov Statistics (Asymptotic)			
KS	0.062645	KSa	1.837104

The R-code is

```
library(foreign)
cert= read.dbf("certification.dbf")

#Build boxplots
cert09 = cert[year==2009,]
cert10 = cert[year==2010,]
par(mfrow=c(1,2))
boxplot(cert09$WRITTEN~cert09$language, ds=cert09)
boxplot(cert10$WRITTEN~cert10$language, ds=cert10)

#Buils ANOVA
#Using the entire data set (all years)
cert = transform(cert, lang.f=as.factor(language))
mod1 = aov(WRITTEN ~ lang.f, data=cert)
```

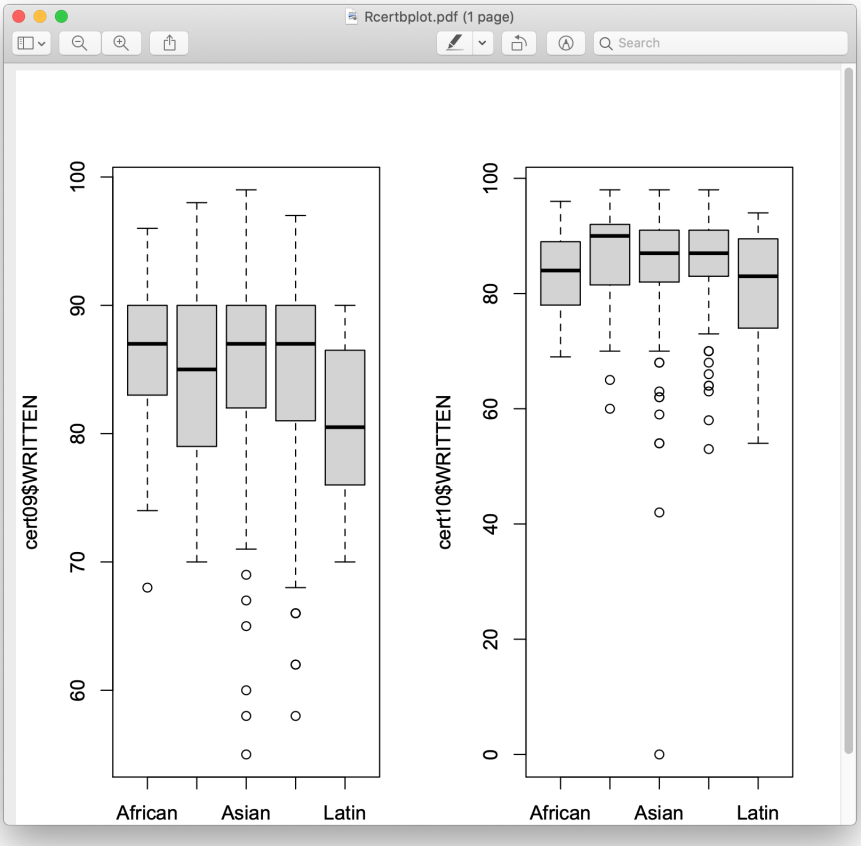
```
summary(mod1)
             Df Sum Sq Mean Sq F value   Pr(>F)
lang.f        4   2454   613.4    8.69 5.94e-07 ***
Residuals  1995 140825    70.6
anova(mod1)
Analysis of Variance Table

Response: WRITTEN
            Df Sum Sq Mean Sq F value    Pr(>F)
lang.f       4   2454  613.45  8.6904 5.942e-07 ***
Residuals 1995 140825   70.59

#Using year > 2008
cert.new=cert[year>2008,]
summary(cert.new)
       region         year         STATUS       WRITTEN         PRACTICA
 N_America:368   Min.   :2009   Failed: 72   Min.   : 0.00   Min.   :  0.00
 Asia     :308   1st Qu.:2009   Passed:788   1st Qu.:82.00   1st Qu.: 80.00
 Mideast  : 60   Median :2010                Median :87.00   Median : 87.00
 Austrilia: 38   Mean   :2010                Mean   :85.13   Mean   : 84.41
 Africa   : 32   3rd Qu.:2010                3rd Qu.:91.00   3rd Qu.: 92.00
 S_America: 22   Max.   :2010                Max.   :99.00   Max.   :100.00
 (Other)  : 32
     COUNTRY       language    year.f        lang.f
 USA      :330   African: 32   2004:  0   African: 32
 India    :125   Arab   : 60   2005:  0   Arab   : 60
 China    : 97   Asian  :308   2006:  0   Asian  :308
 Japan    : 41   English:424   2007:  0   English:424
 Australia: 38   Latin  : 36   2008:  0   Latin  : 36
 Canada   : 38                 2009:384
 (Other)  :191                 2010:476

 #
 #Using another method for ds=cert.new
 #
 library(tidytext)
install.packages("dplyr")
set.seed(1234)
dplyr::sample_n(cert.new,10)
levels(cert.new$lang.f)
library(dplyr)
group_by(cert.new, lang.f) %>%
  summarise(
    count = n(),
    mean = mean(written, na.rm = TRUE),
    sd = sd(written, na.rm = TRUE)
  )

install.packages("ggpubr")
# Box plots
# ++++++++++++++++++++
# Plot weight by group and color by group
library("ggpubr")
ggboxplot(cert.new, x = "lang.f", y = "WRITTEN",
```

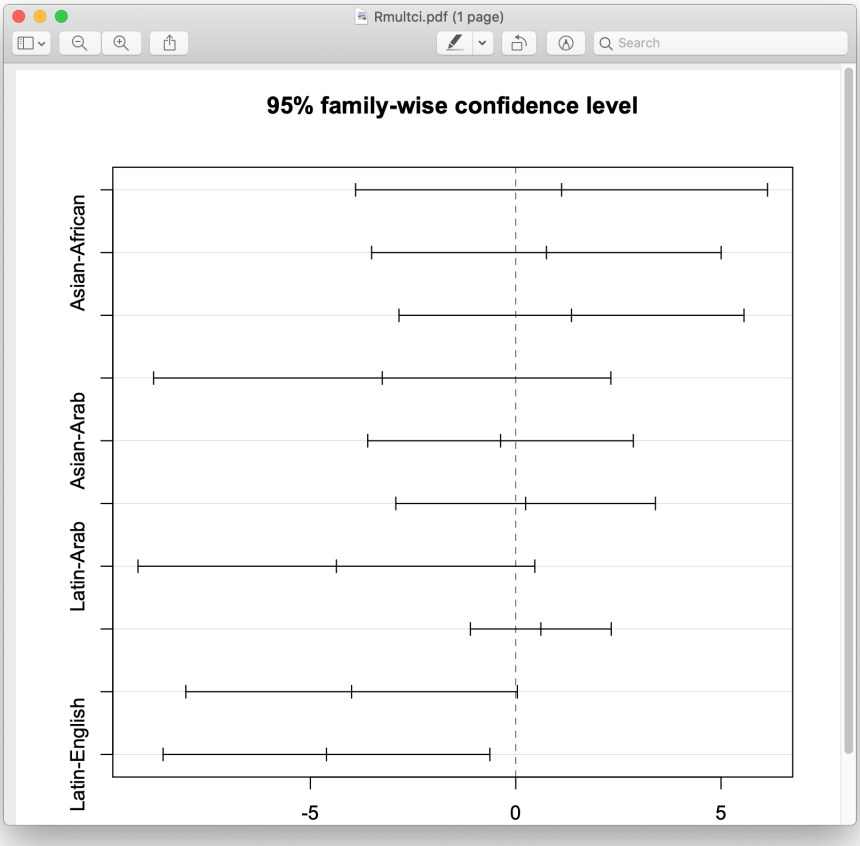
```
          color = "lang.f", palette = c("#00AFBB", "#E7B800", "#FC4E07","#00AFBB","#E7B800"),
          order = c("African","Arab","Asian","English","Latin"),
          ylab = "WRITTEN", xlab = "Language")


# Mean plots
# ++++++++++++++++++++
# Plot weight by group
# Add error bars: mean_se
# (other values include: mean_sd, mean_ci, median_iqr, ....)
library("ggpubr")
ggline(cert.new, x = "lang.f", y = "WRITTEN",
       add = c("mean_se", "jitter"),
       order = c("African","Arab","Asian","English","Latin"),
       ylab = "WRITTEN", xlab = "Language")

# Box plot
boxplot(WRITTEN ~ lang.f, data = cert.new,
        xlab = "Treatment", ylab = "WRITTEN",
        frame = FALSE, col = c("#00AFBB", "#E7B800", "#FC4E07","#00AFBB","#E7B800"))
# plotmeans
library("gplots")
plotmeans(WRITTEN ~ lang.f, data = cert.new, frame = FALSE,
          xlab = "Treatment", ylab = "WRITTEN",
          main="Mean Plot with 95% CI")

# Compute the analysis of variance
res.aov <- aov(WRITTEN ~ lang.f, data = cert.new)
# Summary of the analysis
summary(res.aov)
             Df Sum Sq Mean Sq F value Pr(>F)
lang.f        4    745  186.18   2.649 0.0322 *
Residuals   855  60092   70.28

TukeyHSD(res.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = WRITTEN ~ lang.f, data = cert.new)

$lang.f
                     diff       lwr         upr
Arab-African     1.1166667 -3.899876  6.13320891
Asian-African    0.7467532 -3.509724  5.00323020
English-African  1.3584906 -2.842829  5.55980976
Latin-African   -3.2500000 -8.817874  2.31787391
Asian-Arab      -0.3699134 -3.603869  2.86404253
English-Arab     0.2418239 -2.919182  3.40282977
Latin-Arab      -4.3666667 -9.198032  0.46469908
English-Asian    0.6117373 -1.104030  2.32750479
Latin-Asian     -3.9967532 -8.033335  0.03982883
Latin-English   -4.6084906 -8.586867 -0.63011378
                    p adj
Arab-African    0.9737655
```

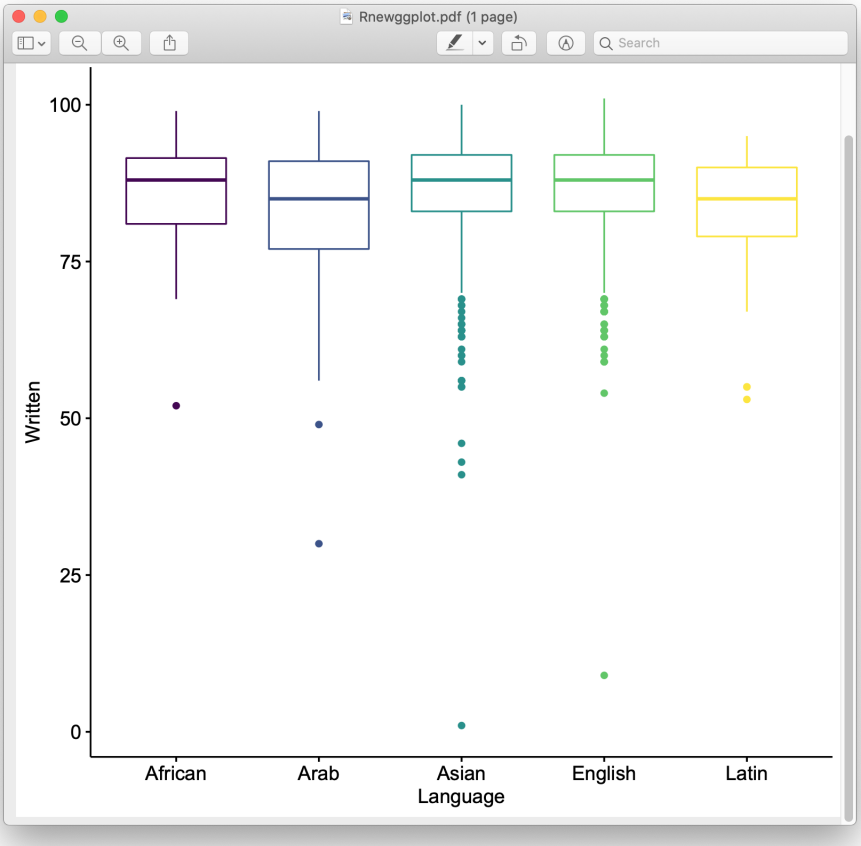
```
Asian-African   0.9892050
English-African 0.9028614
Latin-African   0.5005736
Asian-Arab      0.9979262
English-Arab    0.9995745
Latin-Arab      0.0983112
English-Asian   0.8666419
Latin-Asian     0.0537933
Latin-English   0.0137969

library(multcomp)
summary(glht(res.aov, linfct = mcp(lang.f = "Tukey")))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = WRITTEN ~ lang.f, data = cert.new)

Linear Hypotheses:
                       Estimate Std. Error t value
Arab - African == 0      1.1167     1.8351   0.608
Asian - African == 0     0.7468     1.5571   0.480
English - African == 0   1.3585     1.5369   0.884
Latin - African == 0    -3.2500     2.0368  -1.596
Asian - Arab == 0       -0.3699     1.1830  -0.313
English - Arab == 0      0.2418     1.1563   0.209
Latin - Arab == 0       -4.3667     1.7674  -2.471
English - Asian == 0     0.6117     0.6277   0.975
Latin - Asian == 0      -3.9968     1.4767  -2.707
Latin - English == 0    -4.6085     1.4554  -3.167
                       Pr(>t)
Arab - African == 0      0.9702
Asian - African == 0     0.9876
English - African == 0   0.8913
Latin - African == 0     0.4702
Asian - Arab == 0        0.9976
English - Arab == 0      0.9995
Latin - Arab == 0        0.0865 .
English - Asian == 0     0.8518
Latin - Asian == 0       0.0469 *
Latin - English == 0     0.0119 *

(Adjusted p values reported -- single-step method)


pairwise.t.test(cert.new$WRITTEN, cert.new$lang.f,
                p.adjust.method = "BH")

        Pairwise comparisons using t tests with pooled SD

data:  cert.new$WRITTEN and cert.new$lang.f
```

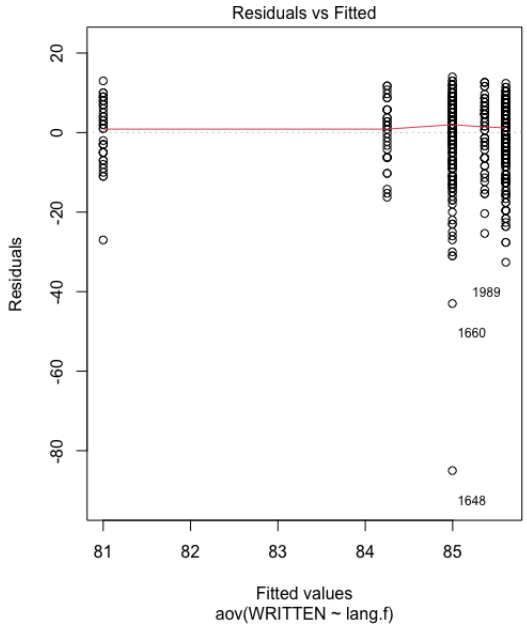
```
        African Arab  Asian English
Arab    0.776   -     -     -
Asian   0.790   0.834 -     -
English 0.628   0.834 0.628 -
Latin   0.277   0.046 0.035 0.016

# 1. Homogeneity of variances
plot(res.aov, 1)

library(car)
leveneTest(WRITTEN ~ lang.f, data = cert.new)
Levene's Test for Homogeneity of Variance' (center = median)
       Df F value Pr(>F)
group   4  1.4626 0.2116
      855

# 2. Normality
plot(res.aov, 2)

# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test[Note Not a very good test,
sure there are others in R but I dont know them at this time]
shapiro.test(x = aov_residuals )

        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.86086, p-value < 2.2e-16

#Run Kruskal Wallis Nonparametric test
kruskal.test(WRITTEN ~ lang.f, data = cert.new)

        Kruskal-Wallis rank sum test

data:  WRITTEN by lang.f
Kruskal-Wallis chi-squared = 12.927, df = 4,
p-value = 0.01164
```

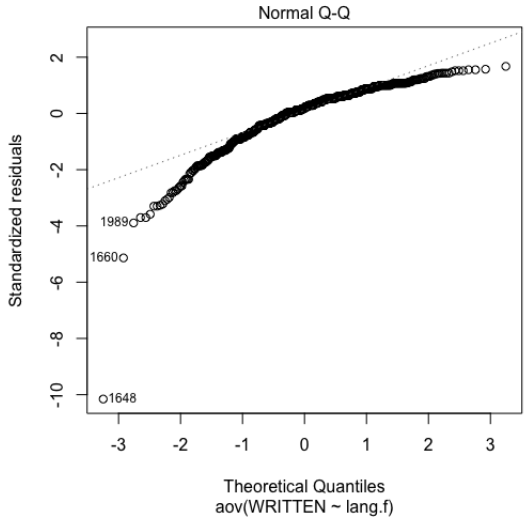
Some of the R output is


Rcertbplot.pdf (1 page)
Search
cert09$WRITTEN
100
90
80
70
60
African
Asian
Latin
cert10$WRITTEN
100
80
60
40
20
0
African
Asian
Latin


Rmultci.pdf (1 page)
Search
95% family-wise confidence level
Asian-African
Asian-Arab
Latin-Arab
Latin-English
-5
0
5


Rnewggplot.pdf (1 page)
Search
100
75
50
25
0
Written
African
Arab
Asian
English
Latin
Language





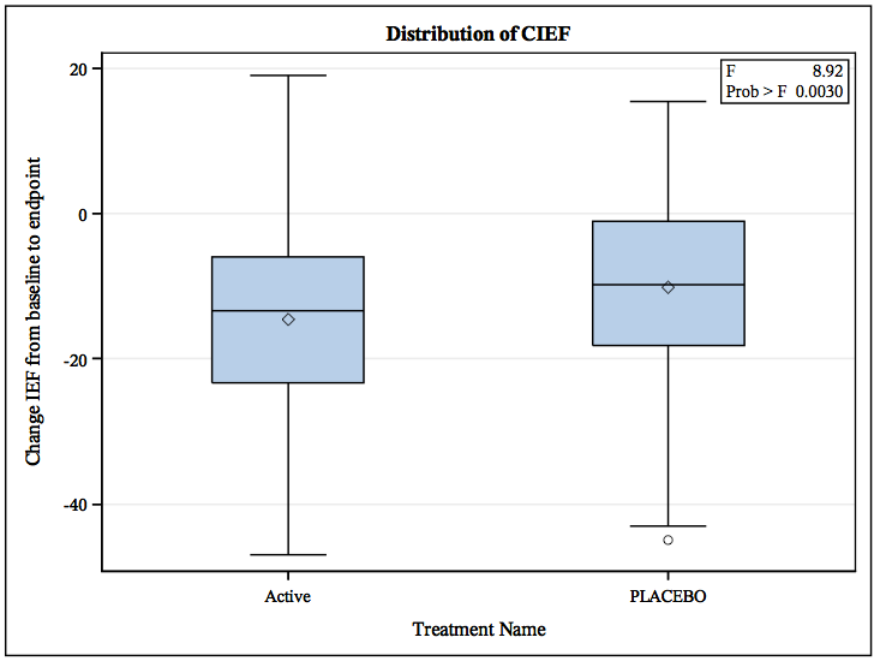
5.2 Analysis of Covariance

Few of the large data sets lend themselves to this model. I have included one example using the Urinary Incontinence Trial. The response variable is the change in incontinence episodes (*cief*) as a function of the baseline number of incidents (*bief*) and the treatments (*therapy*). The analysis was done when the Strata level = 4. The SAS code is given in

`urinary-canova.sas`

```
options center nodate pagesize=100 ls=80;

title1 'Urinary Incontinence Data';

data urinary; set sasuser.urinary; if  -50 < cief < 20;run;

proc sgpanel data=urinary; where strata=4;
panelby therapy;
  histogram cief;
  density cief / type=normal;
  density cief / type=kernel;
run;

* ANOVA model;
proc glm data = urinary plots=all; where strata=4;
class therapy;* horm50;
  model cief =   therapy/ solution;
run;

* Analysis of Covariance;
proc glm data = urinary plots=all; where strata=4;
class therapy;* horm50;
  model cief =  bief  therapy/ solution;run;proc glm data = urinary plots=all; where strata=4;class therapy;*
horm50;model cief = bief therapy/ solution;run;quit;
```

The SAS(some) output is

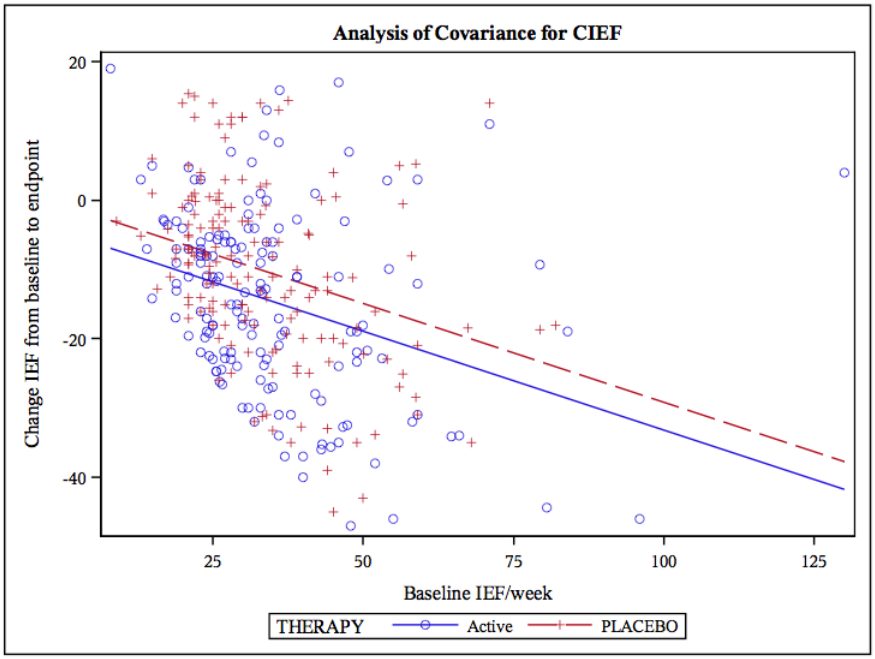


Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7025.91426	3512.95713	22.16	<.0001
Error	312	49470.06071	158.55789		
Corrected Total	314	56495.97497			

R-Square	Coeff Var	Root MSE	CIEF Mean
0.124361	-102.0724	12.59198	-12.33632

Source	DF	Type I SS	Mean Square	F Value	Pr > F
BIEF	1	5761.793365	5761.793365	36.34	<.0001
THERAPY	1	1264.120896	1264.120896	7.97	0.0051

Source	DF	Type III SS	Mean Square	F Value	Pr > F
BIEF	1	5460.569073	5460.569073	34.44	<.0001
THERAPY	1	1264.120896	1264.120896	7.97	0.0051



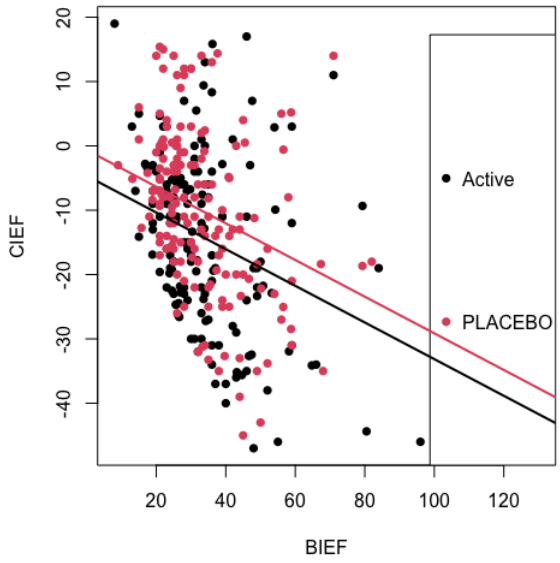
The R code is (using RStudio)

```
library(foreign)
urinary = read.dbf("new_urinary.dbf")
bief = urinary$BIEF
cief = urinary$CIEF
therapy = urinary$THERAPY
```

```r
bmi = urinary$BMI
urinary = transform(urinary, therapy.f=as.factor(therapy))
summary(urinary)
# Simple Plot
plot(x   = bief,
     y   = cief,
     col = therapy,
     pch = 16,
     xlab = "BIEF",
     ylab = "CIEF")

legend('bottomright',
       legend = levels(therapy),
       col = 1:2,
       cex = 1,
       pch = 16)

#ANCOVA Model
mod1 = lm(cief ~ bief + therapy.f + bief:therapy, data=urinary)
summary.aov(mod1)
              Df Sum Sq Mean Sq F value   Pr(>F)
bief           1   5762    5762  36.506 4.34e-09 ***
therapy.f      1   1264    1264   8.009  0.00496 **
bief:therapy   1    385     385   2.439  0.11940
Residuals    311  49085     158

mod2 = lm(cief ~ bief + therapy.f, data=urinary)
summary.aov(mod2)
             Df Sum Sq Mean Sq F value   Pr(>F)
bief          1   5762    5762  36.339 4.68e-09 ***
therapy.f     1   1264    1264   7.973  0.00505 **
Residuals   312  49470     159

#Analysis of the models
options(contrasts = c("contr.treatment", "contr.poly"))
library(car)

Anova(mod1, type="II")
Anova Table (Type II tests)

Response: cief
             Sum Sq  Df F value    Pr(>F)
bief           5461   1 34.5978 1.047e-08 ***
therapy.f      1018   1  6.4523   0.01157 *
bief:therapy    385   1  2.4386   0.11940
Residuals     49085 311
Anova(mod2, type="II")
summary(mod2)
Anova Table (Type II tests)

Response: cief
          Sum Sq  Df F value    Pr(>F)
bief        5461   1 34.4390 1.123e-08 ***
therapy.f   1264   1  7.9726  0.005054 **
```

```
Residuals  49470 312

contrasts(therapy)
        PLACEBO
Active        0
PLACEBO       1

# Simple Plot with fitted equation
I.nought = -4.65314
I1 = I.nought + 0
I2 = I.nought + 4.01279
B  = -0.28492

plot(x   = bief,
     y   = cief,
     col = therapy,
     pch = 16,
     xlab = "BIEF",
     ylab = "CIEF")

legend('bottomright',
       legend = levels(therapy),
       col = 1:2,
       cex = 1,
       pch = 16)

abline(I1, B,
       lty=1, lwd=2, col = 1)

abline(I2, B,
       lty=1, lwd=2, col = 2)
```

Some of the R output is


CIEF
BIEF
Active
PLACEBO
20
10
0
-10
-20
-30
-40
20
40
60
80
100
120