

Chapter 3 Examining Data

This chapter discusses methods of displaying *quantitative* data with the objective of understanding the distribution of the data.

Example During childhood and adolescence, bone mineral density increases until peak bone mass is reached. Peak bone mass and subsequent bone loss are important determinants of osteoporosis later in life. To investigate the process of bone mineral acquisition, measurements of the bone mineral density and age of 261 North American adolescents was collected.¹

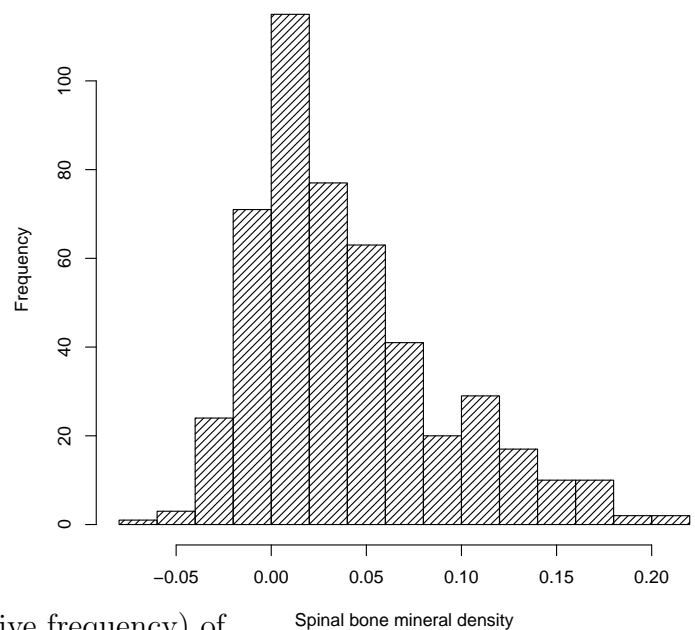
Ignoring age and gender, the distribution of values can be visually represented several ways. A histogram is shown to the right.

A histogram is constructed by forming a set of contiguous intervals, or bins and counting the number of observations belong to each. A bar drawn above the bin has height proportional to the count (and relative frequency) of observations in the interval.

Typically, 10 to 20 bins are used depending on the number of observations, more observations allow for a finer division of the variables range and more detail in the histogram. A bin is an interval a plot which breaks data values of a variable into

intervals and displays the frequency (or relative frequency) of the observations that fall into each interval.

The distribution is unimodal and somewhat right-skewed. It's unknown why there are negative values.



Density plots are smoothed versions of histograms. A density plot is constructed by estimating the density of the underlying probability distribution that generated the observations. For example, the underlying probability distribution might be approximately normal, and the greatest density occurs in a neighborhood of the mean.

¹Bachrach LK, Hastie T, Wang M-C, Narasimhan B, Marcus R. Bone Mineral Acquisition in Healthy Asian, Hispanic, Black and Caucasian Youth. A Longitudinal Study. J Clin Endocrinol Metab(1999) 84, 4702-12.

The figure to the right shows the probability distribution, or density function of a standard normal random variable.

Given a set of data, a density function is constructed from a large set of points spanning the range of the data. At each point, the average number of observations near the point determine the height of the estimated density function. The average is a weighted average where the distance of the observation to the point are (typically) inversely proportional to the distance to the point.

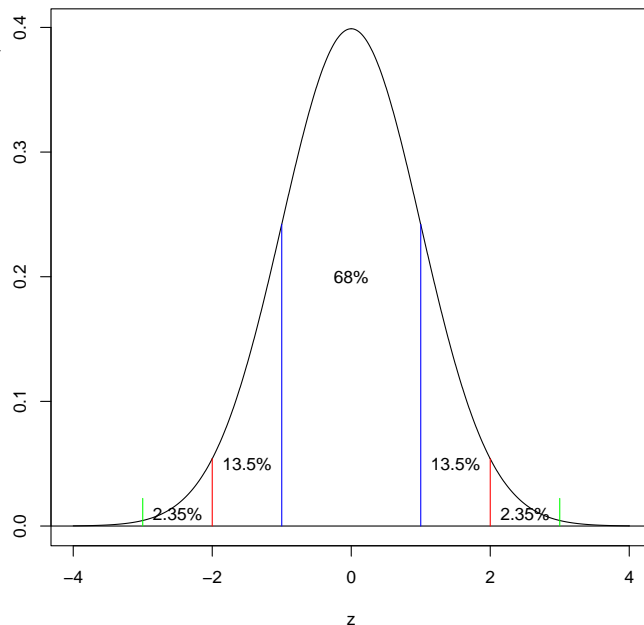
A distant point contributes little or no weight towards the estimate of the density at the point. A common a set of weights are drawn from the normal probability density function. For example, range of the bone density data is from -0.06 to $.220$ and so a 1000 points x_1, \dots, x_{1000} beginning at -0.06 and increasing by $.000284$ might be used. The estimated density at the k th point is

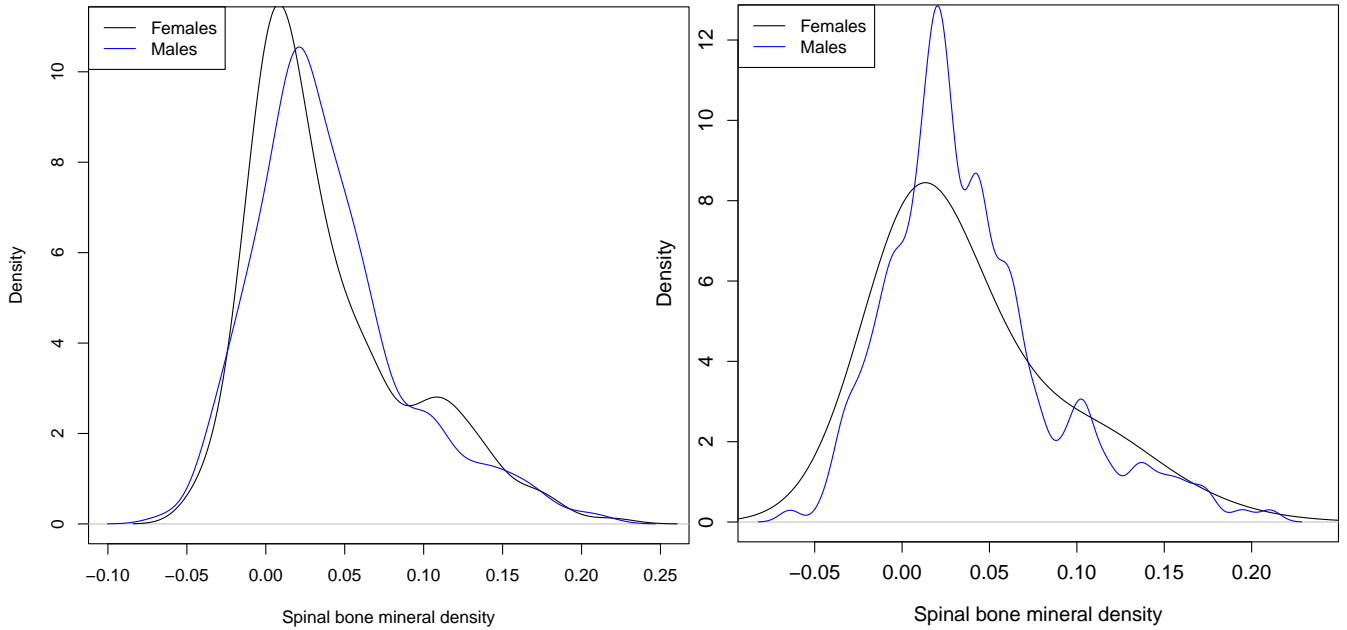
$$\hat{p}(x_k) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_k - x_i}{h}\right).$$

The n data values are x_1, \dots, x_n , the kernel function K determines a weight assigned to the data points depending on the difference $x_k - x_i$, and h is a scaling term (often the sample standard deviation). An example of a kernel function K is the standard normal density function

$$K(x_k) = \frac{1}{\sqrt{2\pi}} \exp[-(x_k - x_i)^2].$$

The figure below and left show density plots for males and females using the default smooth value; the figure below and right shows density plots after adjusting the degree of smoothing.





Boxplots are alternatives to histograms and density plots that reveal less of the distribution but provide quantitative information regarding quantiles; specifically, a boxplot is a graphical display of a 5-number summary with one modification: outliers are identified.

5-number summary: While the median and IQR are a useful two-number summary of center and spread, a more complete summary of the distribution is given by the 5-number summary:

$$(\text{Min}, Q_1, \text{Median}, Q_3, \text{Max}).$$

The *quartiles* are the 25th, 50th and 75th percentiles²; and so they partition the data set into four sets of approximately equal numbers of observations.

$$\begin{aligned} Q_1 &= 25^{\text{th}} \text{ percentile} = 1^{\text{st}} \text{ quartile} \\ M &= 50^{\text{th}} \text{ percentile} = 2^{\text{nd}} \text{ quartile, or the median} \\ Q_3 &= 75^{\text{th}} \text{ percentile} = 3^{\text{rd}} \text{ quartile} \end{aligned}$$

The *interquartile range* is the distance between the 25th and 75th percentiles, i.e.,

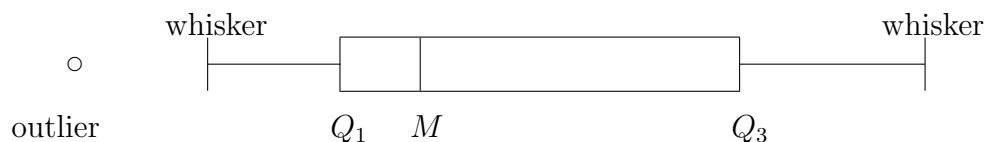
$$\text{IQR} = Q_3 - Q_1.$$

The interquartile range is a robust measure of spread or variance since Q_1 , Q_2 and Q_3 are resistant to outliers.

Again, a *boxplot* is a graphical display of a 5-number summary showing outliers.

²Recall that the p^{th} percentile of a distribution is that value such that $p\%$ of the data values fall below it. If your SAT math percentile was 80%, then your score was larger than 80% of all scores.

The components are shown:



- The central *box* shows Q_1 , the median, and Q_3 .
- The *whiskers* extend to the most extreme values that are within the fences.
- The *lower fence* is $Q_1 - 1.5 \times \text{IQR}$ and the *upper fence* is $Q_3 + 1.5 \times \text{IQR}$.
- The fences are not shown; instead *whiskers* are drawn at largest datum less than the upper fence and the smallest datum greater than the lower fence.
- Any points outside the fences are *outliers* and are plotted individually.

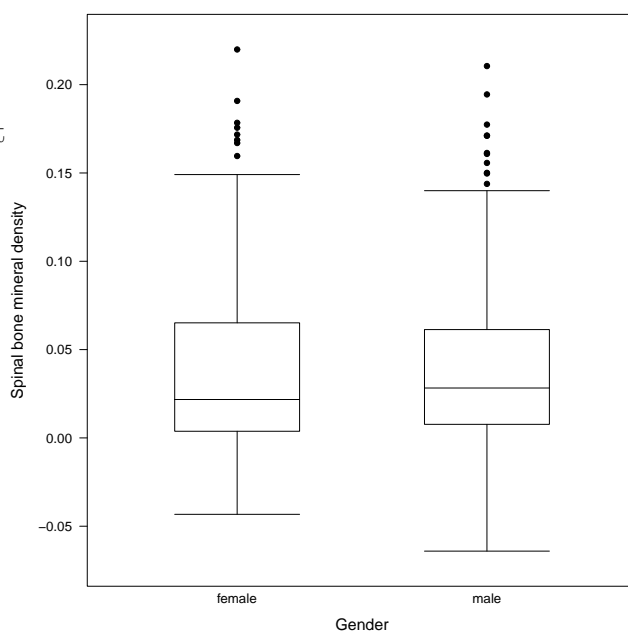
Outlier Detection using the IQR: A common rule for detecting outlying values is called the *1.5 IQR rule*: values at least $1.5 \times \text{IQR}$ greater than Q_3 or at least $1.5 \times \text{IQR}$ less than Q_1 are outliers. Hence, outliers lie outside the interval:

$$[Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}]$$

Example: To the right are two boxplots comparing the spinal bone density between genders. The boxplots show that there is little difference between the sample quantiles, and little evidence of systematic differences among genders.

Both sample distributions contain large outliers.

The R code for constructing the boxplots is



```

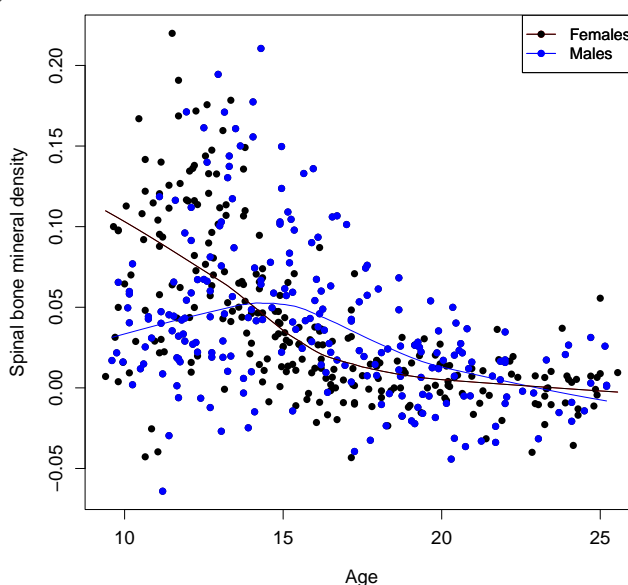
library(ElemStatLearn)
library(lattice)
data(bone)
head(bone)
dim(bone)
bwplot(bone$spnbmd~bone$gender,xlab="Gender",cex.axis=1.2,cex.lab=1.2,
  pch = "|",ylab="Spinal bone mineral density",
  par.settings = list(strip.background = list(col = "transparent"),
  box.rectangle = list(col = "black",lty = 1),
  box.umbrella = list(col = "black",lty = 1),
  plot.symbol = list(alpha = 1,col = "black",cex = 1,pch = 20),
  superpose.symbol = list(cex = rep(0.7, 7),col = "black", pch = rep(20,7))))

```

Most of the code is used to produce a nice looking boxplot. The last function call can be replaced with `bwplot(bone$spnbmd~bone$gender)`.

Males and females reach puberty at different ages³, and so it's possible that, if bone density changes over time, then the timing of the changes may differ among genders.

A *scatterplot* plotting bone density against age, with females and males identified by color shows that there is overall little differences. However, over the age range 10-15 years, there are differences: female bone density decreases with age more rapidly (on average) than male bone density. The lowest smooths clearly show the difference in mean bone density. The R code for constructing the boxplots is



```

plot(bone$age,bone$spnbmd,pch=16,ylab="Spinal bone mineral density",col="black",
  xlab="Age",cex.lab=1.2,cex.axis=1.2)
points(bone$age[females==FALSE],bone$spnbmd[females==FALSE],col="blue",pch=16)
lines(lowess(bone$age[females==FALSE],bone$spnbmd[females==FALSE]),col="blue")
lines(lowess(bone$age[females==T],bone$spnbmd[females==T]),col="black")
legend(x="topright",legend=c("Females","Males"),col=c("red","blue"),pch=16,lty=1)

```

Consider a linear regression model of the response variable spinal bone density (Y) on age

³Females at age 10 or 11, and males at age 12 or 13, on average.

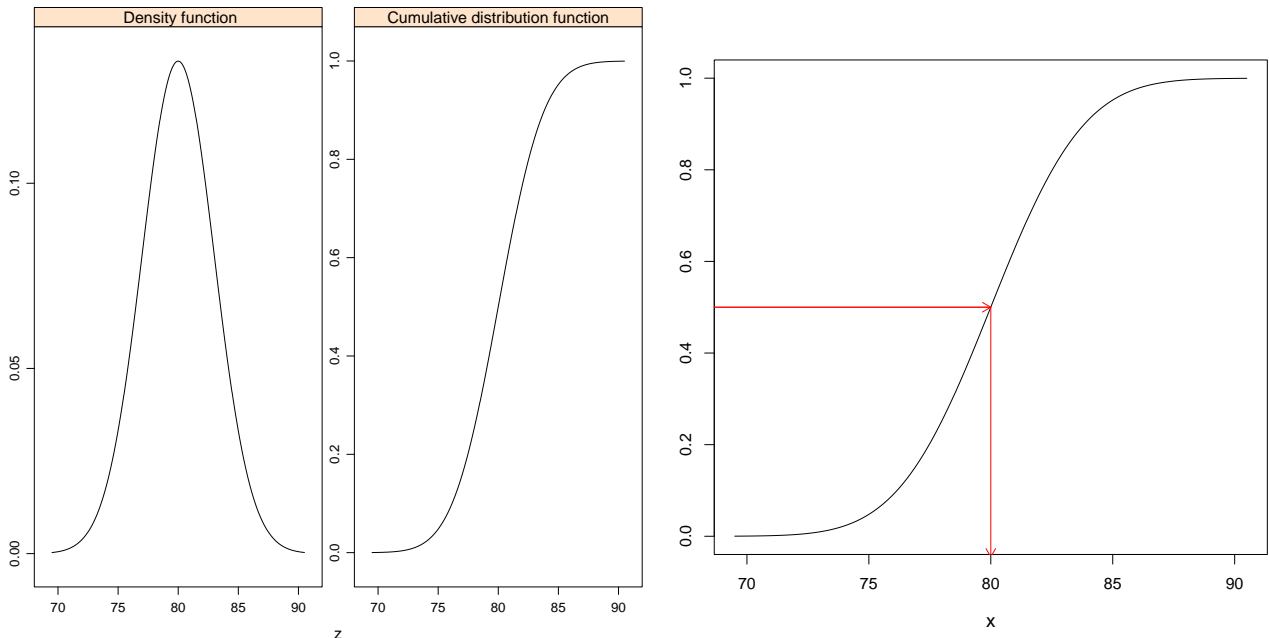
(x):

$$\begin{aligned} Y &= E(Y|x) + \varepsilon, \\ E(Y|x) &= \beta_0 + \beta_1 x \\ \varepsilon &\sim N(0, \sigma_\varepsilon^2). \end{aligned}$$

The model is untenable because the lowest smooths indicate that the relationship is not linear (at least for the males if not the females), and depends on gender, and also because the constant variance assumption does not hold. The variance of the residuals (ε 's) decreases with age. The third condition, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ may be tenable.

Quantile-comparison plots

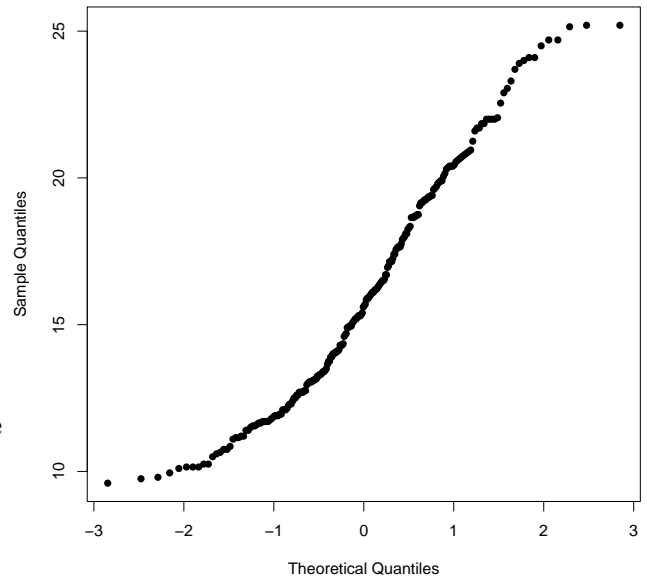
Distributional assumptions are investigated graphically using a quantile comparison or similar plot. The principal is to compare the expected values of the ordered data to the observed data. If the data are to have originated from a normal distribution with mean μ equal to the sample mean \bar{x} and standard deviation σ equal to the sample standard deviation, then the expected value of the sample median is μ . Other expected values are more difficult to determine, but the figures below illustrate the idea. The figures on the left show the density and cumulative density function of the $N(80, \sigma = 3)$ distribution. The cumulative distribution function shows the portion of the distribution less than a particular choice of x . For example, the function call `pnorm(82, mean=80, sd = 3)` returns the value .747; hence, 75.5% of the distribution is less than 83.



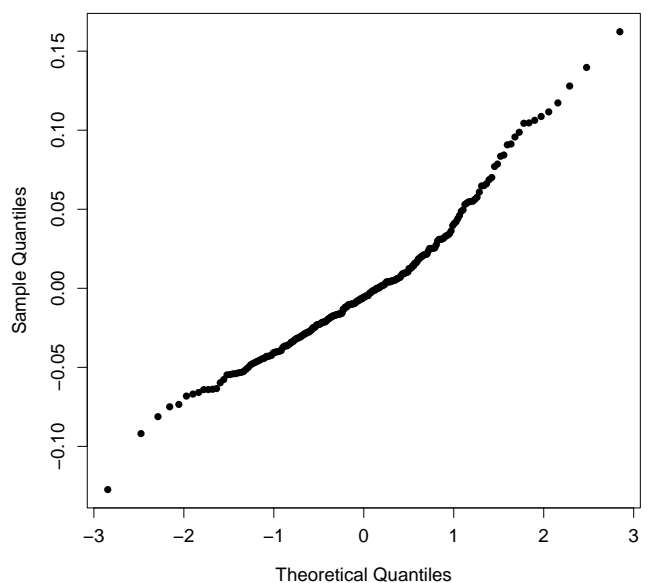
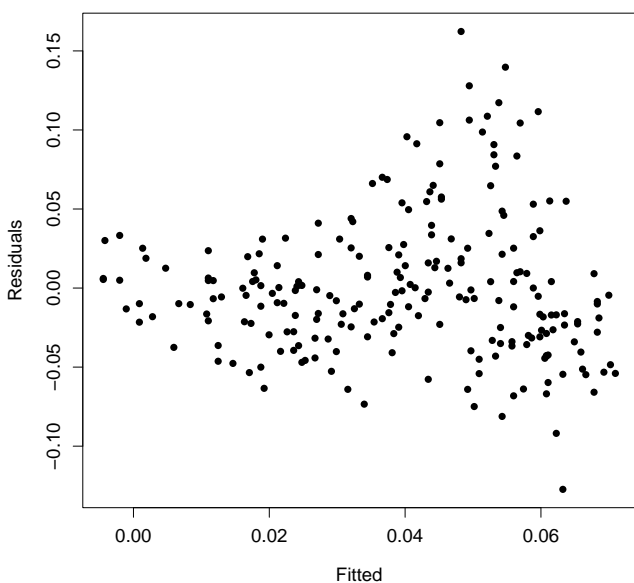
The figure on the right shows how the theoretical quantiles are determined. Given a proportion (take $p = .5$), the x with 50% of the distribution less than x is $x = 80$. If

$n = 100$, then the first quantile has (roughly) 1% of the distribution to its left (that value is 73.02). If the data are approximately normal, then the smallest data value ought to be approximately 73.02, and the second-smallest ought to be approximately equal to the 2nd quantile, and so on. A plot of the theoretical quantiles and the ordered data ought to appear to have slope 1 and follow a straight line. If not, then the normal distribution is not tenable.

The function call `qqnorm(bone$age[females==FALSE], pch=16, main="")` yields a quantile plot for the male sample. The normal distribution assumption is in doubt since the tails of the sample distribution are too short. This may be because of variation associated with age is not accounted for. Consequently, the sample standard deviation overestimates the population standard deviation. If the population standard deviation were equal to the sample standard deviation, then there would be more extreme values (large and small) in the sample data.



Since age is apparently associated with bone density, any analysis of bone density ought to account for age differences (and possibly gender differences). A first attempt at analysis is to conduct a simple linear regression analysis for each gender (separately), in which case, the distributional conditions presented above ought to hold for the regression residuals.



To examine the distributional conditions, a simple linear regression model is fit using bone density as the response and age as the explanatory variable. The sample residuals $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1, 2, \dots, n$ are plotted against the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \dots, n$ above and left, and a quantile-quantile plot is above and right constructed from the residuals.

The *residual plot*⁴ (left and above) reveals that the constant variance condition is not satisfied. Also, there is evidence that the linear model assumption is not satisfied since the residual display a curvilinear trend. The quantile-quantile plot indicates that the residual distribution deviates slightly from normality.⁵ The R code is

```
lm.obj <- lm(bone$spnbmd[females==FALSE] ~ bone$age[females==FALSE])
summary(lm.obj)
Residuals <- lm.obj$resid
Fitted <- lm.obj$fitted
plot(Fitted,Residuals,pch=16,cex.lab=1.2,cex.axis=1.2)
qqnorm(Residuals,pch=16,main="",cex.lab=1.2,cex.axis=1.2)
```

Multivariate data arises from observing more than 2 variables on observational unit. Bivariate data is plotted with one variable per axis. More than 2 quantitative variables require more than 2 axes, and so it is difficult to visualize (simultaneously) more than two quantitative variables. Three dimensional plots are feasible and sometimes insightful, but often difficult to produce. There are two widely used approaches - pair-wise plotting of variables and coplots.

Example Gross domestic product (GDP) refers to the market value of all final goods and services produced in a country in a given period. GDP per capita is often considered an indicator of a country's standard of living. Treasury bills are short-term loans to the federal government, and are nearly risk-free investment. Presumably, GDP and inflation are associated, though there are undoubtedly other covariates affecting both variables; Treasury bills may also be related to these variables. The figures below provide some information on the apparently complex relationship. The data are quarterly observations from 1950 through 1996 from Canada.⁶

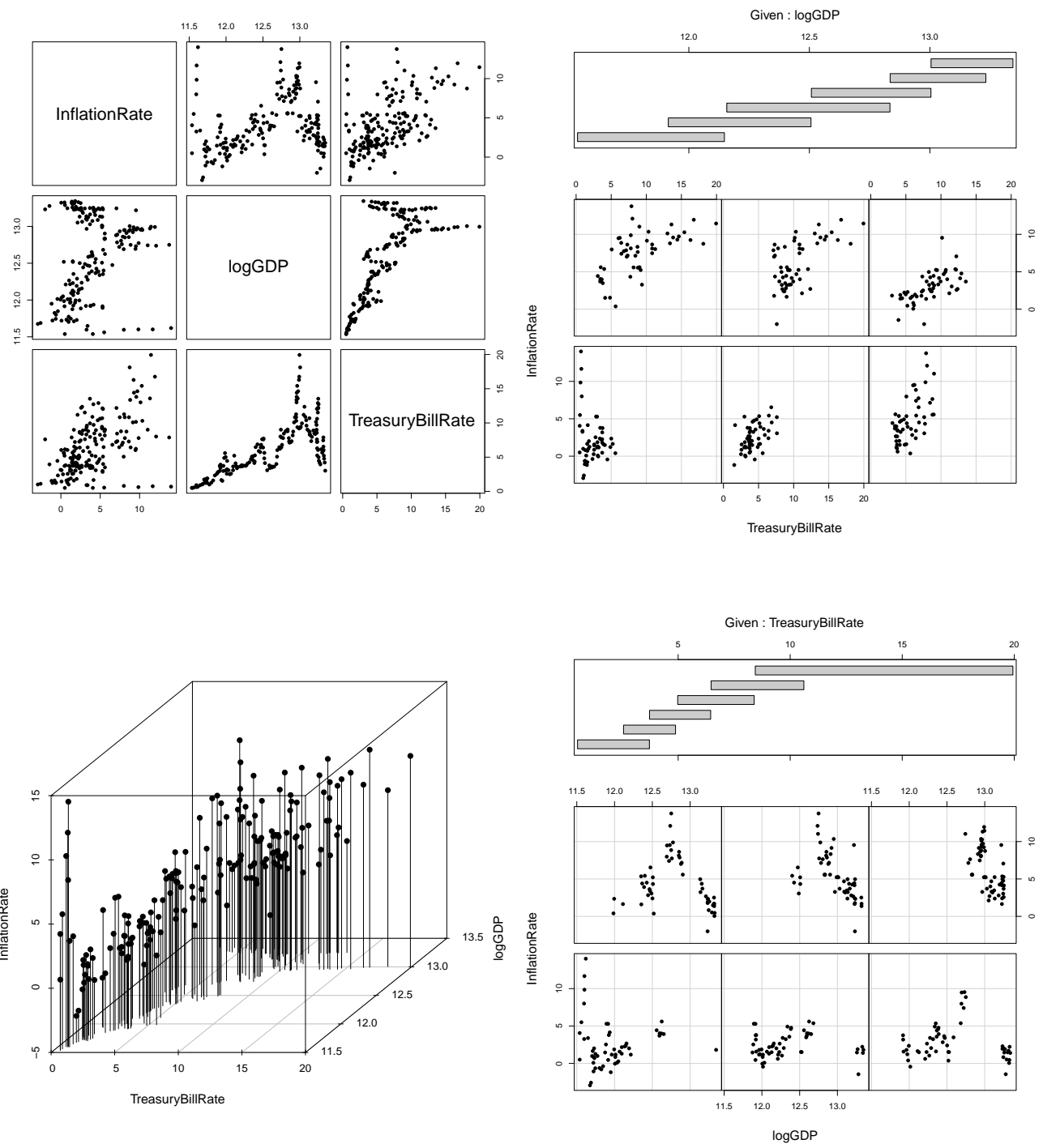
The plot on the upper left is a set of pair-wise scatterplots. It can be obtained using a data frame using the function call `plot(dataframe)` where `dataframe` is the name of the data frame. The upper left plot is a covariate plot (coplot) wherein the range of one variable–

⁴Residual plots almost always plot the residuals on the vertical axis and the fitted values on the horizontal. If the linear regression model is correct (for the sample population), then the residual distribution should be random, and without trend or pattern besides random.

⁵The amount of deviation is not of concern for conducting inference (hypothesis testing, confidence intervals).

⁶Available in the R package `Ecdat`, topic: `Tbrate`.

$\log(\text{GNP})$ is divided into 6 overlapping intervals and the observations with $\log(\text{GNP})$ in a particular level are plotted.⁷ The figure in the lower right is a 3-dimensional scatterplot constructed using the function call `scatterplot3d` available in the library `scatterplot3d`.

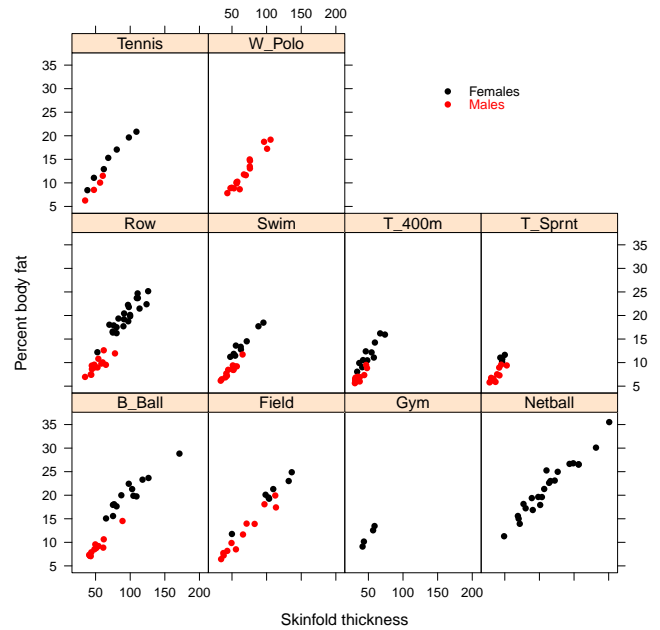


It appears that both T-bill and inflation rate are associated with $\log(\text{GNP})$ in a non-linear manner; the association between T-bill and inflation rate is weak but positive.

⁷The function call was `coplot(InflationRate~logGDP|TreasuryBillRate,pch=16)`.

The lattice panel functions in R are often useful if there are a few of categorical variables and a few quantitative variables. The Figure to the right shows differences among athletes (by sport) and gender in the relationship between percent body fat and skinfold thickness.

The R code is below.



```
xyplot(ais$pcBfat~ais$ssf|ais$sport,groups=ais$sex,col=c("black","red"),pch=16,
xlab="Skinfold thickness",ylab="Percent body fat",
key = list(corner=c(.8,.9),cex=.8,cex.title=1.05,points=T,
           pch=16,col=c("black","red"),border=F,
           text = list(lab = c("Females","Males"),
                       columns = 1)))
```