Multivariate Analysis

Project 1

Jeremy Morris

February 20, 2006

1 Generating bivariate normal data

Definition 2.2 from our text states that we can transform a sample from a standard normal random variable ($\boldsymbol{Z}$) into a multivariate random variable with the distribution $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the equation

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} \tag{1}$$

The matrix $\boldsymbol{A}$ is defined as $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A'}$ [2]. Theorems A.4.3 and A.3.3 in the text suggest that, if $\boldsymbol{\Sigma}$ is positive definite, we use the spectral decomposition $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T'}$ and set $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}$. It follows that we get the correct decomposition of $\boldsymbol{\Sigma}$. The function `mvrnorm` in the MASS library of `R` uses this method.

2 Kernel Density Estimators

The bivariate kernel density estimator with kernel K and bandwidth h is defined by

$$f(\boldsymbol{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left\{\frac{\boldsymbol{x} - X_i}{h}\right\} \tag{2}$$

If we assume the data to be bivariate normal, the kernel function $K(x)$ will be defined as

$$K(\boldsymbol{x}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\boldsymbol{x}'\boldsymbol{x}\right) \tag{3}$$

The kernel density estimator can be calculated using the function `bivden` provided by Everitt [4].

The parameter h is known as the bandwidth of the estimator. There is no closed form solution for an optimal bandwidth, but there are some suggestions. Everitt's function `bivden` uses $h = an^{-0.2}$, where a is some constant provided by the user. This choice for h may not be optimal since it does not use any characteristics of the data set. Härdle suggests that if we can assume the data to be normally distributed, Silverman's rule of thumb can be used to choose the bandwidth [1]. Silverman's rule of thumb suggests taking an estimate from the sample based on the sample variance so that we have the estimate

$$\hat{h}_{rot} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06\hat{\sigma}n^{-1/5} \tag{4}$$

Härdel further notes that this definition of $\hat{h}_{rot}$ is sensitive to outliers, so it is suggested to modify the definition of $\hat{h}_{rot}$ and take into account the 25% and 75% quantiles. Then if we define

$$R = X_{[0.75n]} - X_{[0.25n]} \tag{5}$$

where the X_i are normally distributed, we can use

$$\hat{h}_{rot} = 1.06\frac{R}{1.34}n^{-1/5} \approx 0.79\widehat{R}n^{-1/5} \tag{6}$$

The two estimates in 4 and 6 can be combined so that we have the final version of Silverman's rule of thumb.

$$\hat{h}_{rot} = 1.06 \min\left\{\hat{\sigma}, \frac{R}{1.34}\right\} n^{-1/5} \tag{7}$$

The function `kde2d` in the `MASS` library of `R` uses the $\hat{h}_{rot}$ by default, whereas the `bivden` function assumes that the user will provide a constant. This small difference could change the smoothness in the plots if the proper constant is not provided to `bivden`. For this reason, the function `kde2d` will be used so that no unwieldy calculations have to be done.

Figure 1 shows contour plots for four different values of ρ. These values are $\rho = (0, 0.25, 0.75, 1)$. The plot shows that as the correlation between the data samples goes up, the contour plots get more elongated along the line $y = x$. This makes sense because a bivariate sample that is completely correlated will have the same observations in both elements.

3 Testing for Multivariate Normality

Johnson and Wichern suggest that a good test for multivariate normality is to look at Q-Q plots of the components of the sample [3]. This is done by plotting the sample quantiles $x_{(j)}$ against the quantiles for the standard normal distribution $q_{(j)}$. Sample quantiles can be thought of as the order statistics from the sample. Equation 8

$$P\{Z \leq q_{(j)}\} = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = p_{(j)} = \frac{j - \frac{1}{2}}{n} \tag{8}$$

can be used to find the standard normal quantiles, $q_{(j)}$, by using $p_{(j)}$ as calculated from the ordered observations. If the points $(x_{(j)}, q_{(j)})$ are approximately linear, we can be fairly certain that the observations do not violate the assumption of normality.

Figure 2 shows Normal Q-Q plot for two bivariate data sets, with the least squares line plotted in red. The first sample, $\boldsymbol{X}$, was generated using the method described in Section 1. The Q-Q plots show that this data is normally distributed. The second bivariate sample $\boldsymbol{Y}$ was generated using `R`'s exponential random number generator. The Q-Q plots show this data to be far from normality.

4 Visualizing Trivariate Data

Everitt suggests using a scatterplot matrix or conditioning plots to visualize trivariate data.

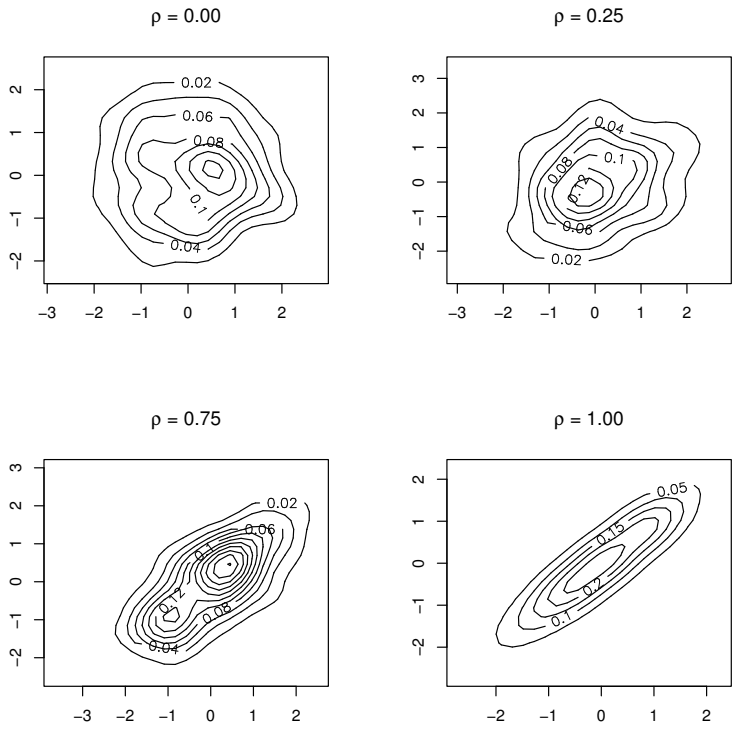

Figure 1: Contour Plots of the Kernel Density Estimator

4.1 Scatterplot Matrix

A scatterplot matrix is a square matrix of bivariate plots for each pair of components in the data set. In the case of trivariate data, six plots will be created. Everitt also suggests placing least square lines and lowess lines in order to better visualize the relationships between the different variables.

Figure 3 shows a scatterplot matrix for data on population, income level and murder rates in all fifty states during the 1970's. Also plotted are the least squares fit and lowess fit for each plot. It is apparent that the scatterplot, in this case, does not fully capture the relationships between the different variables in this data set.

4.2 Conditioning Plots

Conditioning plots, or coplots, can provide more information than scatterplots. To construct a coplot for trivariate data, we need to specify a conditioning variable. We will use the same data set as in the previous section and choose to condition on population. The `coplot` function is used to display six bivariate scatterplots, for the variables representing income level and murder rate, where the data are separated by population density.

In Figure 4 we can see the top panel which gives six bars representing the density of

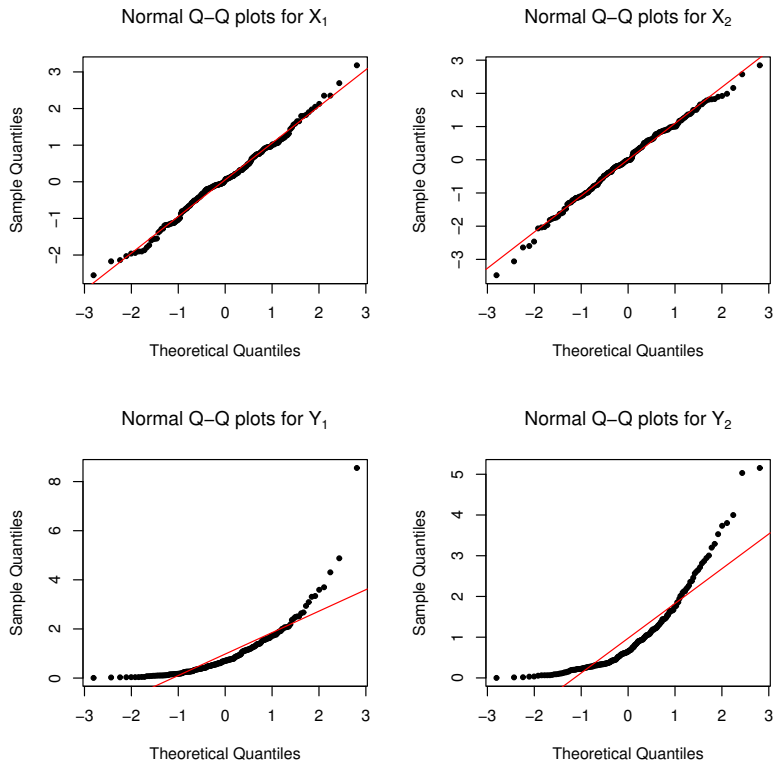

Figure 2: Q-Q plots to test for normality

the population variable. Then each bivariate plot, from left to right, starting in the bottom row plots income level against murder rate for each of the six bars in the density plot. The coplot gives a better picture of the relationship between murder rates and income level. Notice that in all but one plot (bottom left) there appears to be a downward trend, suggesting that as income level rises, murder rates decline. This relationship is not as clear when looking at the scatterplot in Figure 3.

5 Application to a Bivariate Data Set

From the data in Section 4 we will look at a bivariate data set consisting of the income levels and murder rates from the 1970's states data.

Figure 5 shows Q-Q plots for both components of the data set. The income level data appears to be approximately normal. Data for murder rates appears to be normally distributed, but the fit is not great. This could be because the murder rates are integer values, and are thus not normally distributed. A transformation could be used to improve the normality of the data.

Figure 6 shows a contour plot for the kernel density estimator of this data. The contour plots shows some interesting clusters of data. There is one large cluster around the income

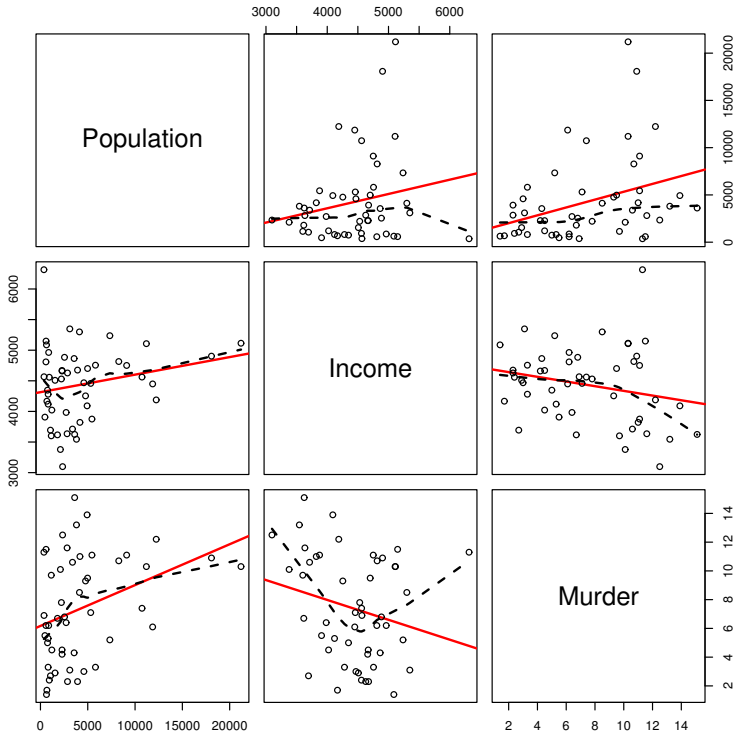

Figure 3: Scatterplot Matrix

levels of 4500-5000 suggesting that there is a low murder rate in middle income areas. There is another smaller cluster in the income range of 3500-4000 suggesting high murder rates in low income levels. There is another small cluster around the 6000+ income level suggesting a high murder rate in a few upper income areas. These observations would suggest that there may be at least one other factor that should be included in any analysis attempting to explain murder rates.

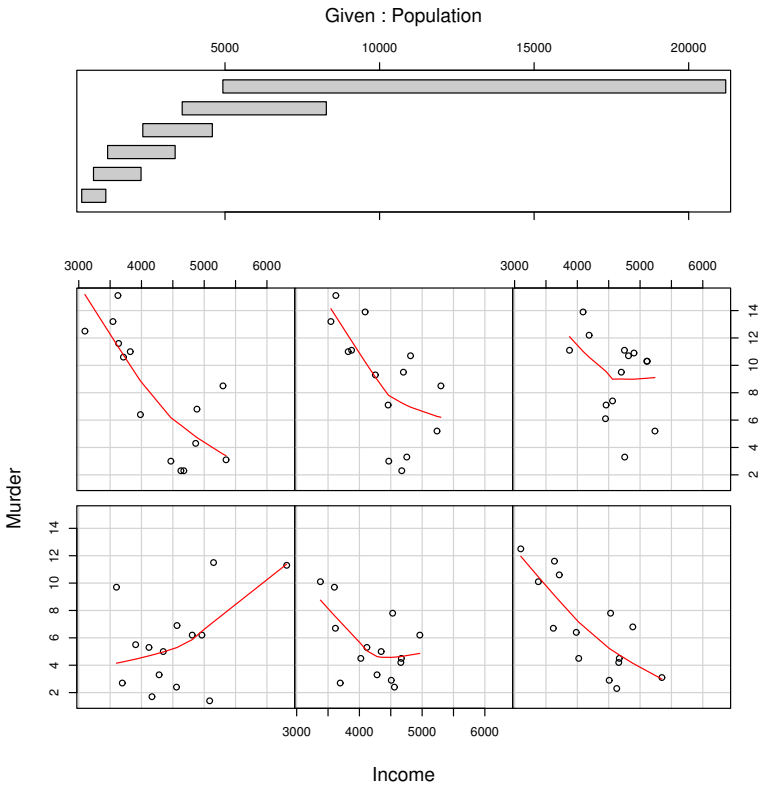

Figure 4: Conditioning Plot

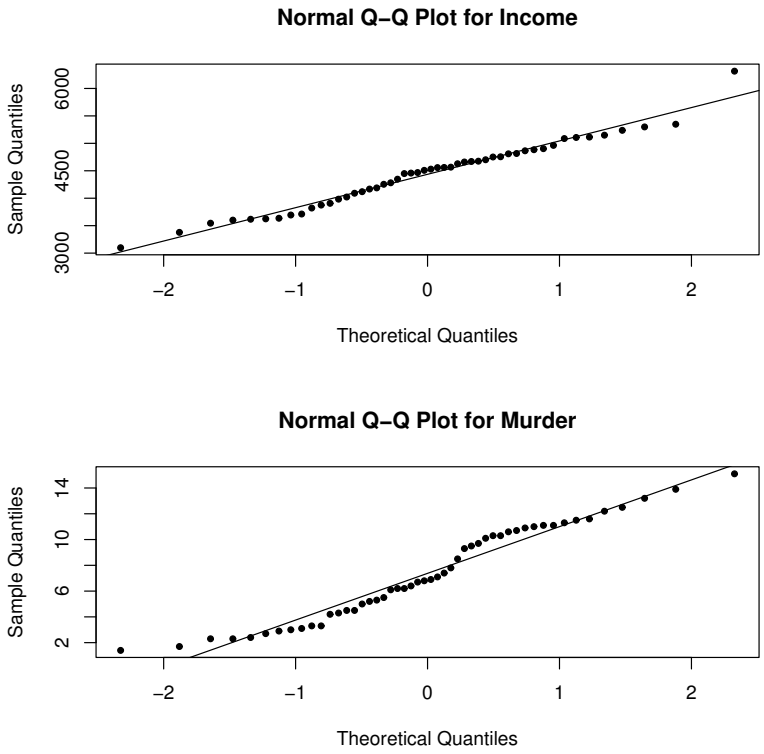

Figure 5: Testing for Normality

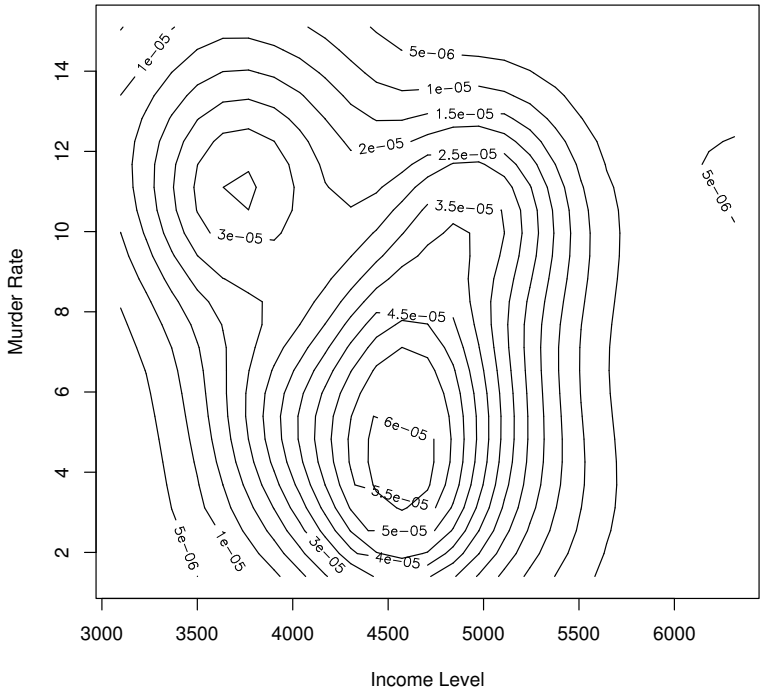

Figure 6: Contour Plots of the Kernel Density Estimator

References

[1] Härdle, Wolfgang, Marlene Müller, Stefan Sperlich, and Axel Werwatz. Nonparametric and Semiparametric Models. Berlin: Springer, 2004.

[2] Seber, George, and Alan Lee. Linear Regression Analysis. 2nd ed. Hoboken: John Wiley & Sons, 2003.

[3] Johnson, Richard, and Dean Wichern. Applied Multivariate Statistical Analysis. 5th ed. Upper Saddle River: Pearson Education, 2002.

[4] Everitt, Brian. An R and S-Plus© Companion to Multivariate Analysis. London: Springer, 2005.