

WRITING ASSIGNMENT (HAND IN SEPARATELY)

COLLEGES (20 POINTS)

This question explores information about selected colleges and universities taken from *US News and World Report*. The data include the following variables: the name of the school, number of undergraduates, student/faculty ratio, average SAT score (out of 1600), percent of freshman in top 10% of high school class, percent of accepted applicants, percent of alumni who donate to school, and expenditures per student in dollars. I have categorized the schools into three categories: liberal arts colleges, private universities, and public universities.

Your assignment is to explore this dataset using four different visualization methods—Chernoff faces, a scatterplot matrix, a parallel axis plot, and a data image—focusing particularly on clustering in the data. You may also use any other methods, if you wish. For instance, what similarities and differences are there between the three types of schools? Are there certain schools that are more similar to schools in other categories than to those in their own, and if so, which schools are they? Are there other outliers? You don't need to describe every single relationship you find; just point out some "interesting" features. **This part should be limited to a maximum of one page of text.**

In the Data Desk file, the points are already colored according to the type of school. If you are pasting the data in as text, you will need to add color manually. To do this, select the icon for *type* and select **Modify ♦ Colors ♦ Add ♦ By Group**.

Also compare the four visualization methodologies. Which plots did you find easier to work with for which questions, or in general? What were the strengths and weaknesses of each type of plot? **This part should be limited to a maximum of two pages of text.** Hand in a printout of your parallel axis plot.

The dataset is on my web site in the usual place. Please type your answer. The entire assignment should be a maximum of four pages total, including graphs. (Note that the maximums given above are for text only.) As always, this number is meant to give you a rough idea of the scope of the assignment. Again, don't try to cover every possible aspect of the data; point out the noteworthy features and discuss how the plots helped you discover these features.

Data Desk does not create Chernoff faces or Data Image plots. I have created these plots using the program *R* and put a copy on my web site (*R* is a program commonly used by academic statisticians; you may download a free copy at <http://www.r-project.org>). In the data image plot, high values are plotted in red and low values in cyan. You don't need to hand in a printout of these plots.

TO MAKE A PARALLEL AXIS PLOT IN DATA DESK:

- 1) Download the data, and copy and paste it into DD.
- 2) Select the numeric variables, and then choose **Manip ♦ Transform ♦ Misc ♦ Zscores(y)**. This will create standardized versions of all the variables, which you may want to rename.
- 3) Choose **Plot ♦ Dotplot** side by side.
- 4) Click on the dotplot window, and then choose **Modify ♦ Lines ♦ Show lines**.
- 5) To add color, select the "type" icon, then choose **Modify ♦ Colors ♦ Add ♦ By group**.
- 6) To identify individual points, open the "college" icon. Open the palettes (**Modify ♦ Palettes**) if you haven't already done so, and select the question mark tool from the palette. Then click on the point you wish to identify.
- 7) It may be useful to reverse the sign on some of the variables. To do this, open the standardized variable icon, and insert a negative sign in the formula.

PROBLEMS (HAND IN SEPARATELY; PLEASE TYPE WHERE POSSIBLE)

1. PREDICTION INTERVALS (3 POINTS)

I mentioned in class that prediction intervals are often very wide, sometimes so much so that they are not informative. In this problem we will think about the margin of error for a prediction interval.

In the Colleges dataset we've seen in class, we used $Y = \text{gradrate}$ (the college's graduation rate) and $X = \text{top25}$ (the percentage of the college's students who were in the top 25% of their high school graduating class). Suppose we are looking at the University of Colorado and we know $\text{top25} = 66\%$, but we don't know its graduation rate. To predict its graduation rate, you would find the regression equation for gradrate vs top25 , calculate the conditional mean of gradrate at $\text{top25} = 66\%$, and then add and subtract the margin of error (MOE) to your estimate.

(Note that we would use a prediction interval rather than a confidence interval because we want to predict the gradrate for just the University of Colorado, a *single* college having $\text{top25} = 66\%$ — not the mean gradrate of *all* colleges having $\text{top25} = 66\%$.)

At the review session for the midterm, it was asked why the MOE for a prediction interval was not simply equal to s , the conditional SD of Y . At that time, I did not have a good answer for this question. I've now figured out how to answer it, but instead of just telling you the answer, I thought I'd make it into a homework question. :-)

The conditional SD of Y measures one source of variability or uncertainty: the spread of points around the line for a given x value. In addition to what is measured by the conditional SD of Y , what other source of variability or uncertainty is reflected in the MOE for a prediction interval?

In other words, the conditional SD of Y accounts for the spread of points around the conditional mean of Y . The MOE for a prediction interval has to account for not only this source of variability, but what other source as well?

2. TRANSFORMATIONS (3 POINTS)

In the regular linear regression model

$$Y = \beta_0 + \beta_1 x$$

the slope β_1 means that a 1-unit increase in x is associated with an additive increase in Y of β_1 units. This can be shown as follows:

At the original value of x , we have

$$\text{original } Y = \beta_0 + \beta_1 x$$

If we increase x by 1 unit, then

$$\begin{aligned} \text{new } Y &= \beta_0 + \beta_1(x + 1) \\ &= \beta_0 + \beta_1 x + \beta_1 \\ &= (\text{original } Y) + \beta_1 \end{aligned}$$

Using a similar argument, show that in a regression with a log-transformed Y variable


$$\log(Y) = \beta_0 + \beta_1 x$$

a 1-unit increase in x is associated with a *multiplicative* increase in Y by a factor of e^{β_1} .

3. HOME PRICES (12 POINTS)

Download the dataset on home prices from my web site. These data give the selling prices (in thousands of dollars) of randomly sampled homes in Albuquerque, New Mexico, in Spring 1993, together with various features of the home: square footage, the age of the home, the number of features possessed out of a list of 11 (e.g., dishwasher, refrigerator, washer and dryer, etc.), whether it is located in the Northeast section of the city, whether it was custom-built, whether it is on a corner lot, and the amount of real estate taxes assessed. Realtors use data such as these to determine the proper selling price for a home.

(a) The size of a home, measured in square feet, should be an important determinant of the home's selling price. The web site <http://lib.stat.cmu.edu/DASL/Stories/homeprice.html> also suggests that homes in the Northeast section ("the largest residential area... more Anglo and more Republican than the rest of the city") may be more expensive, and that the effect of square footage may differ in the Northeast section — in other words, that square footage and being in the Northeast have an interaction.

First, make a scatterplot of price vs square footage (sqft). Color the points in the Northeast section ($NE = 1$) blue and the other points yellow. (To do this, make a bar chart of NE by selecting its icon and choosing **Plot** \blacktriangleright **Bar Charts**. Then open the palettes (**Modify** \blacktriangleright **Palettes**) and select the  icon. Click on the $NE = 1$ bar and select blue from the color palette. Then click on the $NE = 0$ bar and select yellow from the color palette. Now go to the HyperView menu of the scatterplot and choose **Add Color Regression Lines**. If sqft and NE had an interaction, what would be true about the regression lines? Do you see evidence for such an interaction? (Hand in the scatterplot.)

(b) Now fit a regression model for price based on sqft and NE, and include an interaction term. (To create a variable representing the product of sqft and NE, click on the sqft and NE icons, then choose **Manip** \blacktriangleright **Transform** \blacktriangleright **Arithmetic** \blacktriangleright $y * x$.) Do you see evidence for a statistically significant interaction? (Hand in the regression output.)

(c) Are there outliers or influential points? Calculate the leverages and Cook's distances by clicking on the HyperView menu of the regression output and choosing **Compute** \blacktriangleright **Leverages** and **Compute** \blacktriangleright **Cook**. You will get two new icons; click on each one and choose **Plot** \blacktriangleright **Dotplot Side by Side** (make two separate dotplots, one for the leverages and one for the Cook's distances). Indicate which point(s) may be outliers or influential points in your scatterplot from (a).

(d) Repeat steps (a) and (b) with any outliers or influential points deleted. (A quick way to delete an outlier is to type an "x" in front of its Y value, which makes the value unreadable by Data Desk; then select Turn on automatic update in each window's HyperView menu.) For the purposes of this assignment, don't delete more than one point. Do your answers change? (Hand in the new scatterplot and regression output.)