1.4: Data Exploration

Dr. Bean - Stat 5100

1 Why Data Exploration

Data Modeling is a lot like:

In order to avoid disaster, you need to **look** before you **jump**.

Example: Consider four scenarios where we use to create a model that uses values of x to predict values of y. We make the assumption in each case that the data can be modeled as

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i \tag{1}$$

This assumption means that we assume that X and Y share a linear relationship. That is, as X increases, Y will increase proportionally. We will explore this further in Handout 2.1.
I estimate the values of β_0 and β_1 using SAS for all four scenarios. The estimated models all have identical form, with identical measures of model goodness (which we will learn about in Handouts 2.2 and beyond).

$$\hat{Y} = 3 + 0.5X$$

(Groups) Using the results of Figure 1, which models are appropriate, and which are inappropriate? Why?

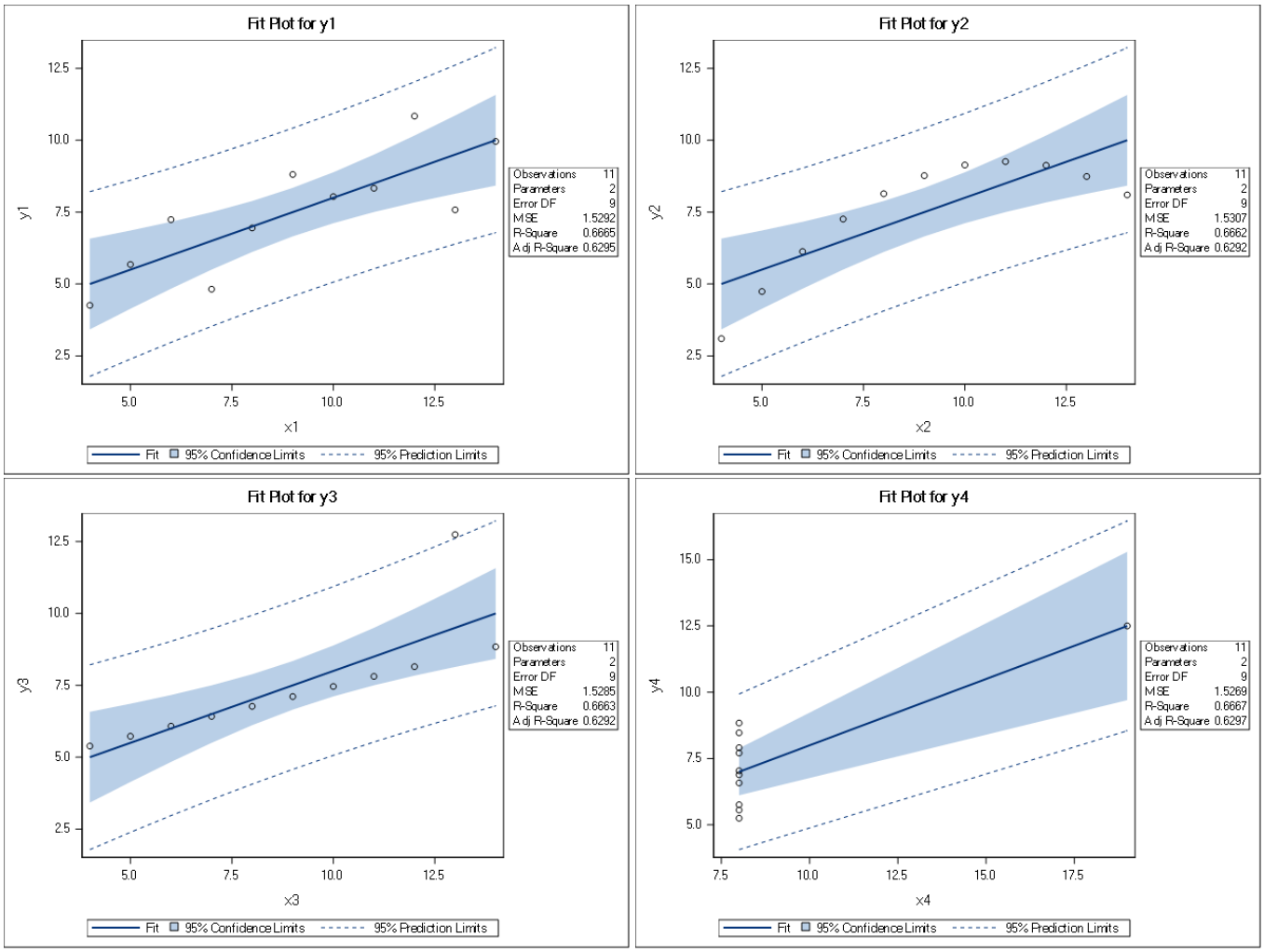

Figure 1: Plots of X vs Y, along with the estimated regression line, for Models 1-4.

Model 1 is the only appropriate model as the rest have outlier points or a non-linear relationship between X and Y.

Data Explorations BEFORE modeling will help us to detect:

- Skewed distributions
- Outlier points
- Non-linear trends

Often, we can use **variable transformations** to get data that are normal, or at least symmetric, in distribution.

Why symmetric data? Consider the "door hinge" problem.

Common Exploratory Plots

- **Boxplots:**: Show the five quartlies of the data (min, 25th percentile, median, 75th percentile, and maximum).
  - Values that are farther than 1.5*IQR (Interquartile Range, which is the 75th percentile minus the 25 percentile) above the 75th percentile or below the 25th percentile are typically plotted as "outlier" points.
  - Great way to quickly summarize the range of values.

```
proc sgplot data=concord1;
vbox Water81;
run;
```

Add option "/ datalabel =" to identify potential outlier points.

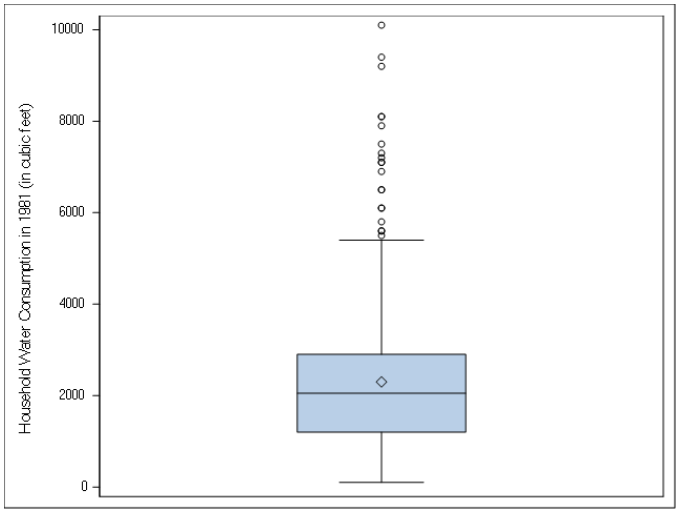

Figure 2: Sample boxplot.

- **Histograms:** Use bins to show the number of observations in a range.
  - Help us to visualize the distribution of the data by imagining a smooth curve running along the top of the bins.
  - Word of caution: the choice of bin width can drastically change the shape of a histogram.

```
PROC UNIVARIATE DATA = concord1 noprint;
HISTOGRAM Water81;
run;
```

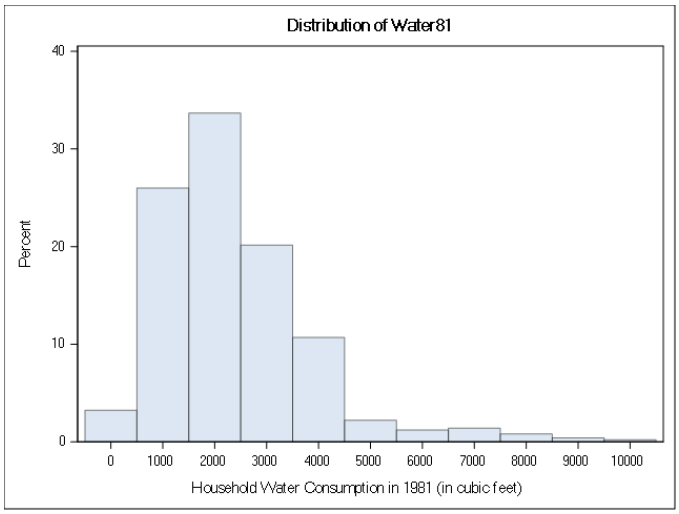

Figure 3: Sample histogram.

- **QQ Plot:** "Quantile Comparison" plots help to easily compare the observed distribution of points to a theoretical (typically normal) distribution.
  - Plots the data quantiles against the theoretical quantiles of similar observations that are normal in distribution.
  - Points that closely follow the diagonal line indicate that the observed data follow the theoretical distribution.
  - While they don't help to visualize shape, qqplots are superior to histograms as a visual check for normality.

```
PROC UNIVARIATE DATA = concord1 noprint;
qqplot Water81 / NORMAL(mu=est sigma=est);
run;
```

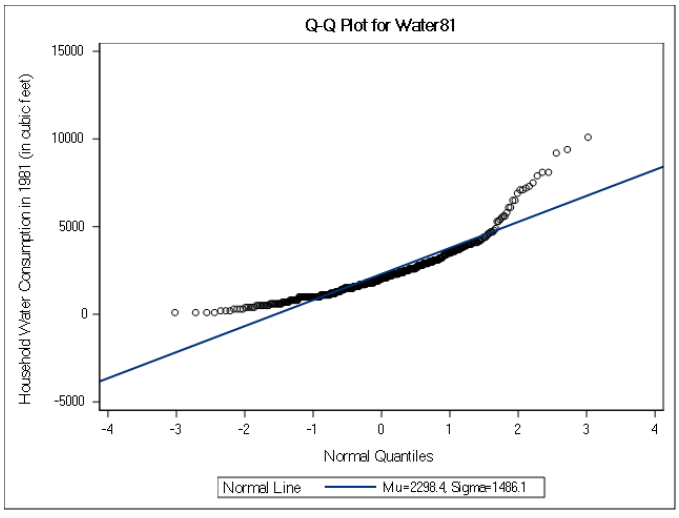

Figure 4: Sample quantile comparison plot for a normal probability distribution.

- **Scatterplots:** Plots paired observations from two variables as points on a two-dimensional plot.

- Excellent way to determine if two variables share a relationship.
- Can combine in a **scatterplot matrix** when looking at relationships between more than two variables.
- Subject to **overplotting** when you have thousands of observations that you are trying to plot at the same time.

```
proc sgscatter data=concord1;
matrix Water81 Water80 Water79;
run;
```

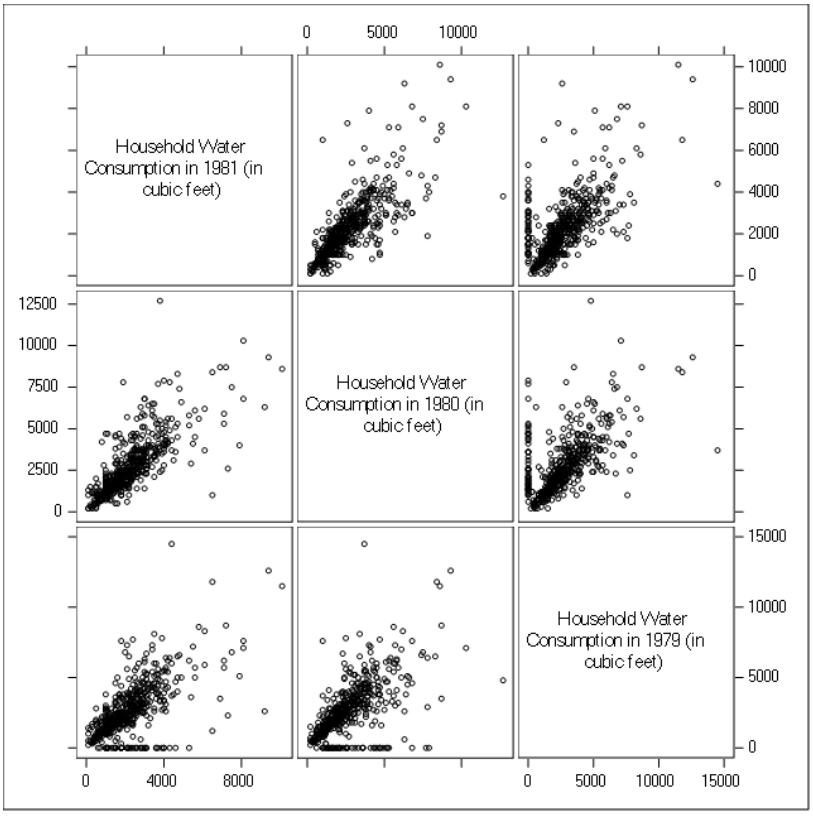

Figure 5: Sample scatterplot matrix.

See **Handout 1.4.2** for an extended example in SAS of data explorations.