Relationships Between Two Variables: Scatterplots and Correlation

- **Example: Consider the population of cars manufactured in the U.S. What is the relationship (1) between engine size and horsepower? (2) Between engine size and gas mileage?**
- **Do cars with large engines tend to have high horsepower?**
- **Do cars with large engines tend to have high gas mileage?**
- **We will see both graphical and numerical ways to summarize this information about the *relationship between two variables*.**

Scatterplots

- **A *scatterplot* is a graph that shows the relationship between two *quantitative* variables.**
- **Each individual in the data set has *two variables* measured on it.**
- **For each individual, the values of one variable are plotted on the horizontal axis, with the values of the other variable on the vertical axis.**
- **On the plot, there is a dot for each observation in the data set.**
- **See example:**

Scatterplot of Horsepower and Vehicle Weight for 32 cars.

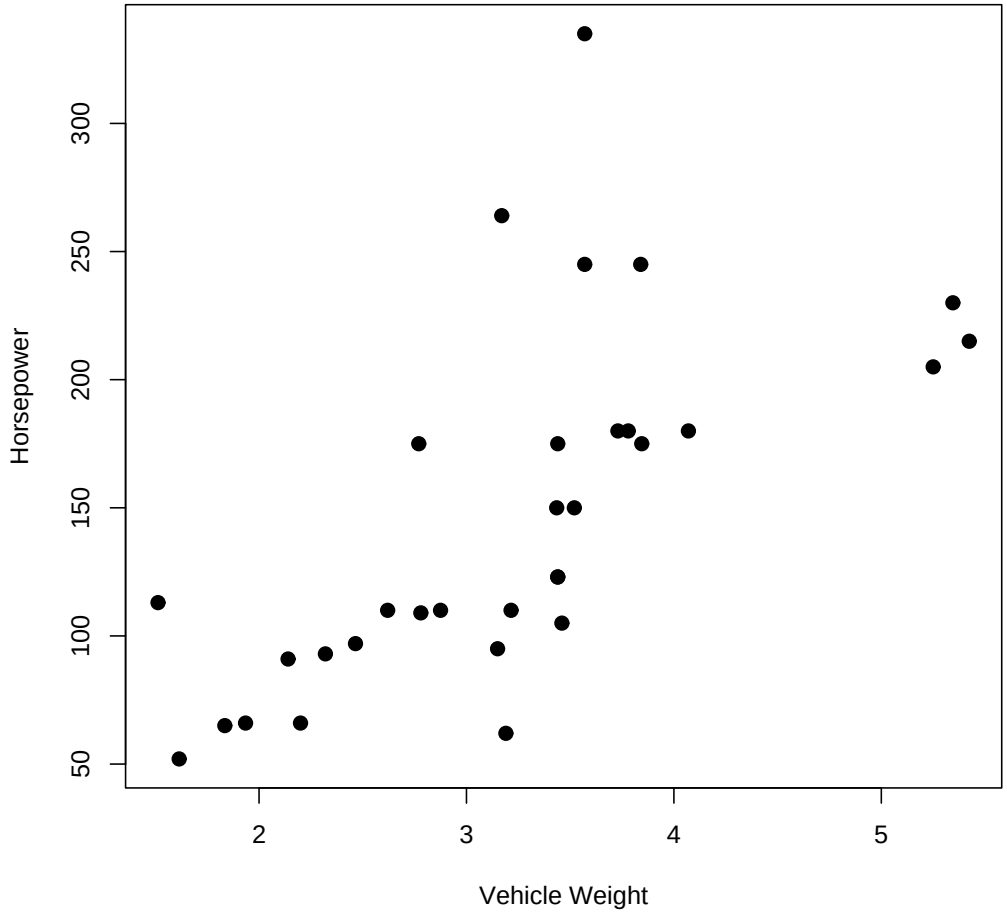

Figure 1: Vehicle weight (in 1000s of pounds) on horizontal axis, horsepower on vertical axis.

Explanatory and Response Variables

- **Sometimes one variable (called the *explanatory variable*) may naturally *explain* or *predict* the value of the other variable (called the *response variable*).**
- **If so, the explanatory variable is denoted *X* and plotted on the X-axis. The response variable is denoted *Y* and plotted on the Y-axis.**
- **In the engine size / gas mileage example, which is naturally the response variable?**
- **Other times, there is no natural explanatory-response relationship between the two variables – either one can go on the horizontal axis. (Example: Height/Weight)**

Interpreting Scatterplots

- **Same process as interpreting other graphs**
- **Look for *overall pattern* in the plot and look for *deviations* from that pattern.**
- **Describe general pattern (not counting outliers)**
- **Then look closely at individual outlying values to determine their cause.**

Positive and Negative Associations

- **A scatterplot can show us the *direction* of the relationship between two variables.**
- **Two variables have a *positive association* if observations having large values for one variable also tend to have large values for the other variable.**
- **Also, when variables are positively associated, observations having small values for one variable also tend to have small values for the other variable.**
- **The scatterplot for such positively associated variables has a pattern that slopes upward from left to right. (Example: Horsepower and Vehicle Weight)**

Positive and Negative Associations (Continued)

- **Two variables have a *negative association* if observations having large values for one variable tend to have *small* values for the other variable.**
- **The scatterplot for such negatively associated variables has a pattern that slopes downward from left to right.**

Scatterplot of Gas Mileage and Vehicle Weight for 32 cars.

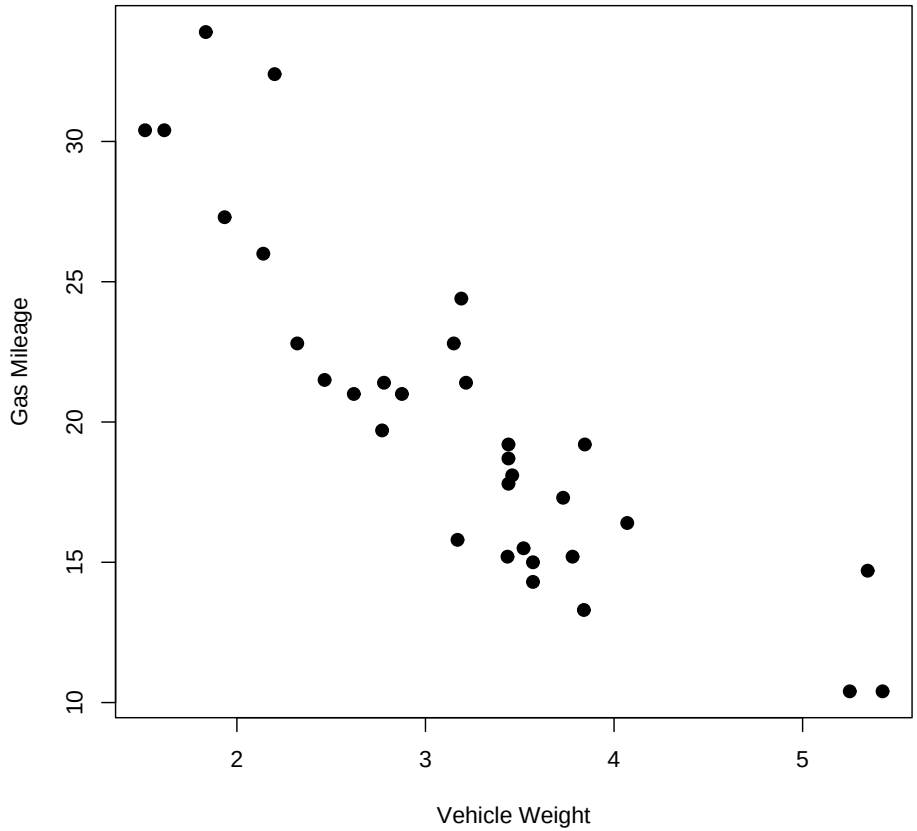

Figure 2: Vehicle weight (in 1000s of pounds) on horizontal axis, mileage (in miles per gallon) on vertical axis.

Form and Strength of Association

- **The *form* of the relationship between two variables may be approximately *linear* or *curved*.**
- **Once we identify the *form* of the relationship, we can characterize the *strength* of the relationship between the two variables.**
- **The relationship is *strong* if most of the data values *closely follow* the major trend in the plot.**
- **The relationship is *weak* if the data values show a great deal of *random scatter* around the major trend in the plot.**

Form and Strength of Association (Continued)

- **In the previous two scatterplots shown, what are the *forms* of the relationships?**
- **Which of the two scatterplots shows a *stronger* relationship between the two variables plotted?**
- **Are there any notable outliers in the scatterplots?**

Clicker Quiz 1

What is the best description of the relationship between vehicle weight and gas mileage, based on the scatterplot?

A. There is a fairly strong, roughly linear, positive association between vehicle weight and gas mileage.

B. There is a very weak, roughly linear, negative association between vehicle weight and gas mileage.

C. There is a very weak, curved, positive association between vehicle weight and gas mileage.

D. There is a fairly strong, roughly linear, negative association between vehicle weight and gas mileage.

Correlation

- **Straight-line relationships between two variables are often of interest in data analyses.**
- ***Correlation* is a numerical measure of the *strength* and *direction* of the *linear* relationship between two quantitative variables.**
- **This could give us a more precise measure of the association than a scatterplot.**
- **Correlation coefficient (denoted *r*) is a number between -1 and 1.**
- **A positive value of *r* indicates the two variables are *positively* linearly associated.**
- **A negative value of *r* indicates the two variables are *negatively* linearly associated.**

More on Correlation

- **The correlation also tells us how *strong* the linear relationship is.**
- **A value of *r* near -1 or near 1 indicates the two variables are *strongly* linearly associated.**
- **A value of *r* near 0 indicates the two variables are *weakly* linearly associated.**
- **See example pictures:**

Clicker Quiz 2

If two variables are strongly negatively linearly associated, what might be a value of *r* for a sample of data on these two variables?

A. *r* = 0.78

B. *r* = 0.13

C. *r* = -0.95

D. *r* = -0.04

Cautions about Correlation

- **The correlation does not change if we change the units of measurement for the variables (example: measuring vehicle weight in tons).**
- **Correlation ignores any distinction between explanatory and response variables.**
- **Correlation only describes the *linear* association between two variables, not any curved relationship!**
- **Correlation may be strongly affected (either increased or decreased) by a few outlying observations.**

Clicker Quiz 3

If two variables are very strongly associated, what might be a value of *r* for a sample of data on these two variables? Choose the best answer:

A. *r* = 0.98

B. *r* = 0.03

C. Impossible to say for sure, but definitely near 1 or near -1.

D. Impossible to say for sure.

The Effect of Outliers on Correlation

- **Recall the scatterplot of Horsepower and Vehicle Weight.**
- **The Maserati Bora is somewhat outlying with respect to the overall pattern.**
- **The correlation for the entire data set is 0.659.**
- **If we delete the Maserati Bora observation, will the correlation be larger or smaller?**

Scatterplot: Horsepower and Weight for 32 cars (Outlier shown)

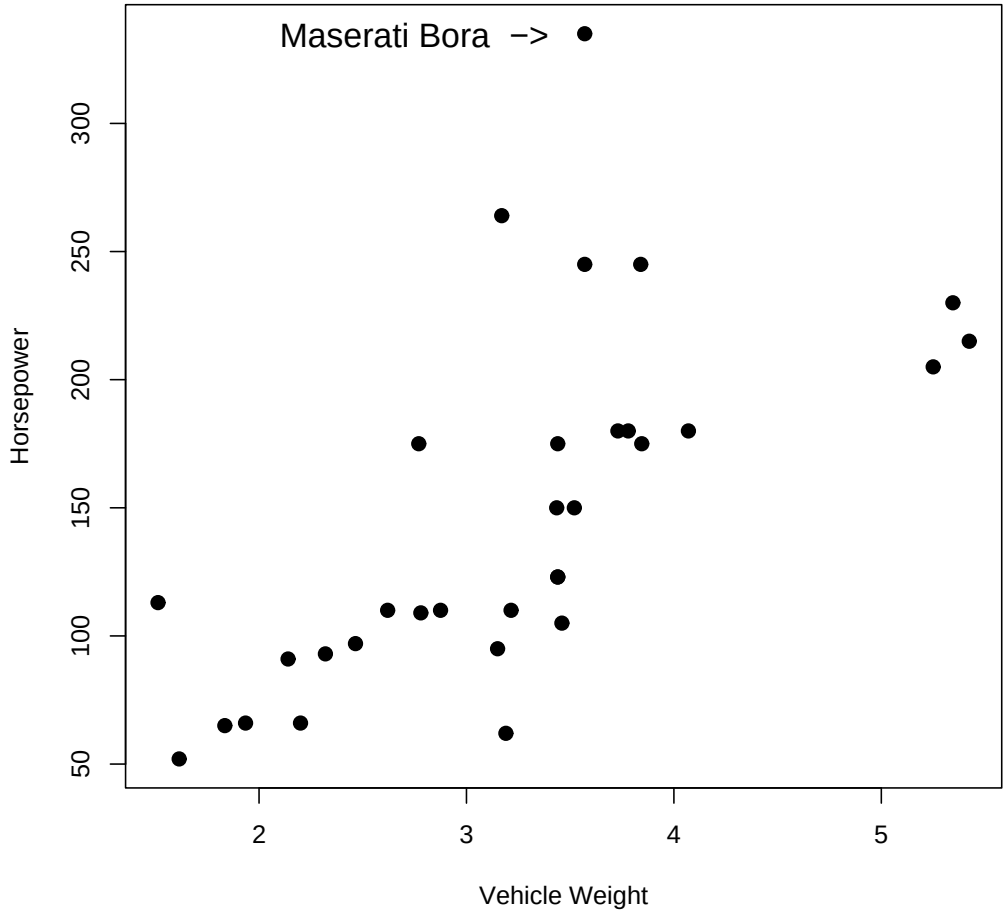

Figure 3: Vehicle weight (in 1000s of pounds) on horizontal axis, horsepower on vertical axis.

The Effect of Outliers on Correlation

- **The correlation for the entire data set is 0.659.**
- **If we delete the Maserati Bora observation, will the correlation be stronger or weaker?**
- **If we delete the Maserati Bora observation, the correlation becomes 0.725.**
- **The Maserati Bora observation had dampened the strength of the linear relationship.**
- **In other cases, an outlier could *increase* the correlation between the variables (see "Thought Question 3" picture)**

Clicker Quiz 4

Consider the *left graph* in the "Thought Question 3" slide. Suppose the correlation between the two variables is 0.57. What might be the correlation *if we deleted* the solitary outlier?

A. 0.47

B. 0.02

C. -0.83

D. 0.74

Clicker Quiz 5

Consider the *right graph* in the "Thought Question 3" slide. Suppose the correlation between the two variables is 0.83. What might be the correlation *if we deleted* the outlying value?

A. 0.72

B. -0.12

C. -0.64

D. 0.95

More about Correlation

- **The ordinary correlation coefficient can only be calculated when both variables are quantitative.**
- **It doesn’t make sense to talk about a “high correlation between gender and preferred TV network.”**
- **There may be a *relationship* between two categorical variables, but it’s not measured by correlation.**
- **There are other statistics that measure the association between categorical variables, or between a categorical variable and a quantitative variable (we won’t cover these).**
- ***Also note:* To be complete, we should report means and standard deviations for each variable, along with the correlation.**