

Scatterplots and Correlation

Ryan Miller

Introduction

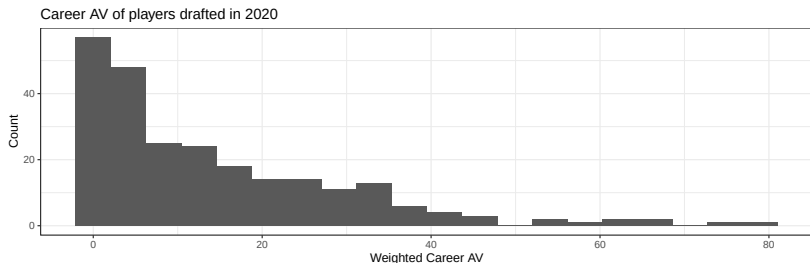
So far we've focused on *one-sample* and *two-sample* data:

1. One-sample data - is our sample consistent with a hypothesized population mean or proportion?
2. Two-sample data - does our sample show evidence of a difference between two populations (groups)?

These analyses have involved various combinations of *binary*, *categorical* and *quantitative* variables. Which combination haven't we covered?

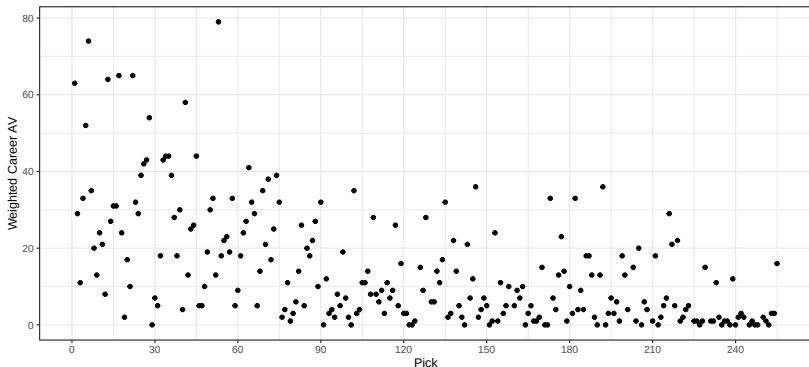
NFL Player Values

- ▶ Approximate value (AV) is an attempt to quantify the value of an NFL player based upon their individual statistics, team performance, and professional awards.
 - ▶ A typical starter has AV between 10-20, while an elite player might have AV of 30+



NFL Player Values

Is a player's career AV related to how high they were picked in the NFL draft?



Describing Relationships (Qualitatively)

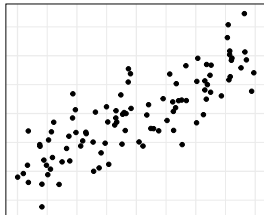
There are *three* important aspects of a relationship between two quantitative variables that we should consider:

1. **Direction** - positive, negative, or unclear
2. **Form** - linear or non-linear
3. **Strength** - weak, moderate, strong, etc.

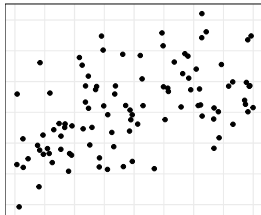
For example: “the resting heart rate and age have a weak, positive, linear association”

Examples

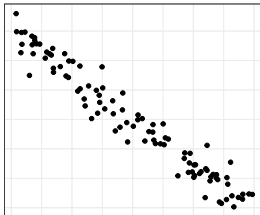
Moderate Positive Linear



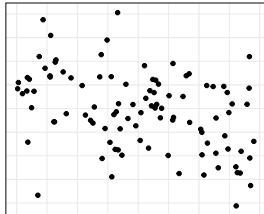
Weak Positive Linear



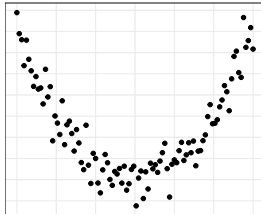
Strong Negative Linear



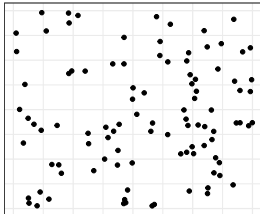
Weak Negative Linear



Strong Nonlinear



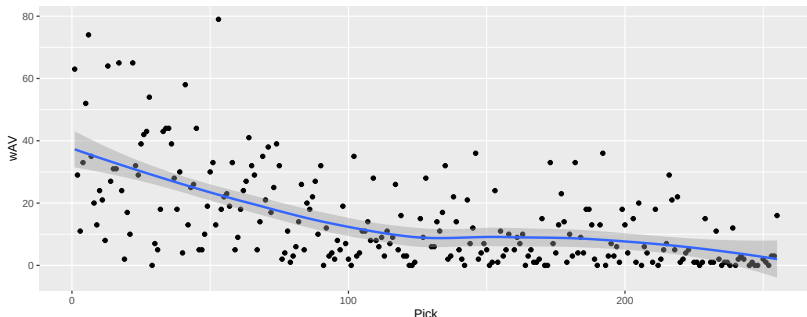
No Clear Relationship



LOESS

LOESS (locally estimated scatterplot smoothing) is a tool to help us judge form:

```
ggplot(data = draft_data, aes(x = Pick, y = wAV)) +  
  geom_point() + geom_smooth(method = "loess")
```



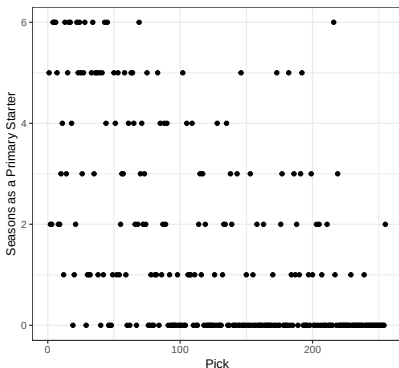
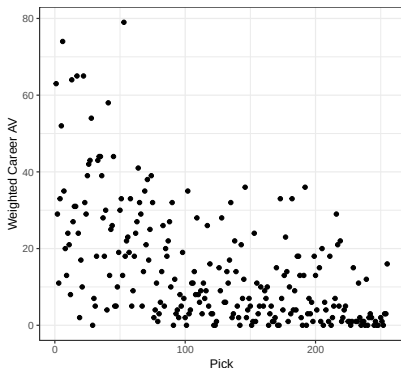
Practice

1. Load data into R using code below
2. Use the code on the previous slide as a template to make a scatterplot with Loess smoother
3. Describe the direction, form, and strength of the relationship between Ret_Rate (percentage of 1st years who return for a 2nd year) and Med_Grad_Earnings_4Y (median annual earnings 4 years after graduation)

```
grinnell_peers = read.csv("https://remiller1450.github.io/data/grinnell_peers.csv")
```

Quantifying Strength

Is an NFL player's draft pick *more strongly* related to their career AV or the number of seasons they play as a starter?



Quantifying Strength

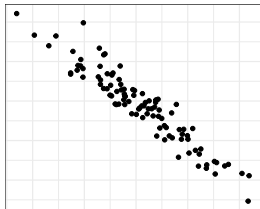
Karl Pearson developed **Pearson's correlation coefficient** to quantify the strength of *linear association*:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

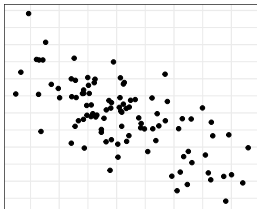
- ▶ Notice r is essentially the average product of Z-scores across observations, so it has no units
- ▶ r can range from -1 (perfect negative association) to +1 (perfect positive association) with 0 reflecting no linear association between variables

Correlation Examples

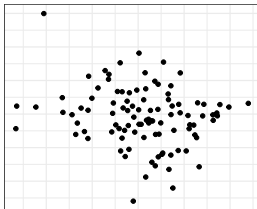
$r = -0.95$



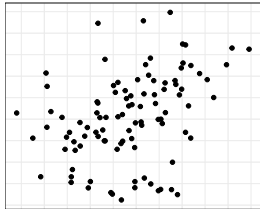
$r = -0.69$



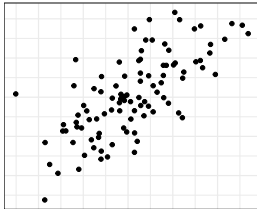
$r = -0.19$



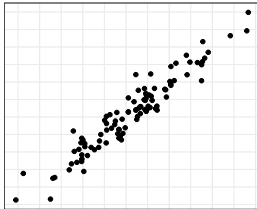
$r = 0.36$



$r = 0.7$



$r = 0.95$



Correlation Examples

Whether a correlation is considered “strong”, “moderate”, or “weak” depends upon the field:

Correlation Coefficient		Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	-1	Perfect	Perfect	Perfect
+0.9	-0.9	Strong	Very Strong	Very Strong
+0.8	-0.8	Strong	Very Strong	Very Strong
+0.7	-0.7	Strong	Very Strong	Moderate
+0.6	-0.6	Moderate	Strong	Moderate
+0.5	-0.5	Moderate	Strong	Fair
+0.4	-0.4	Moderate	Strong	Fair
+0.3	-0.3	Weak	Moderate	Fair
+0.2	-0.2	Weak	Weak	Poor
+0.1	-0.1	Weak	Negligible	Poor
0	0	Zero	None	None

Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969/>

Returning to the NFL Player Data

Is an NFL player's draft pick more strongly related to their career AV or the number of seasons they play as a starter?

```
cor(x=draft_data$Pick, y=draft_data$wAV, use = "complete.obs")
```

```
## [1] -0.5617705
```

```
cor(x=draft_data$Pick, y=draft_data$St, use = "complete.obs")
```

```
## [1] -0.5391188
```

Practice

1. Use the “Grinnell Peers” data set from earlier
2. Is Ret_Rate more strongly correlated with Med_Grad_Earnings_4Y or with Enrollment (total number of enrolled students)?

Problematic Correlation Examples

From Cook & Swayne's *Interactive and Dynamic Graphics for Data Analysis*:

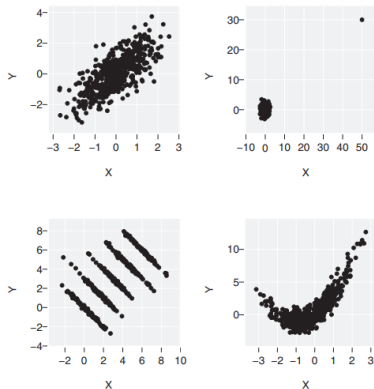


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Quantifying Strength for Non-linear Associations

A popular alternative to Pearson's correlation is **Spearman's correlation**, which replaces each observation with its *rank* from smallest to largest and computes the Pearson correlation of those ranks. As an example:

$$X = \{2, 4, 6, 9, 8\}$$

$$Y = \{7, 4, 1, 5, 3\}$$

Replace each value with its rank:

$$X_{\text{rank}} = \{1, 2, 3, 5, 4\}$$

$$Y_{\text{rank}} = \{5, 3, 1, 4, 2\}$$

Then correlate the ranks: $\text{cor}(X_{\text{rank}}, Y_{\text{rank}}) = -0.3$

Quantifying Strength for Non-linear Associations

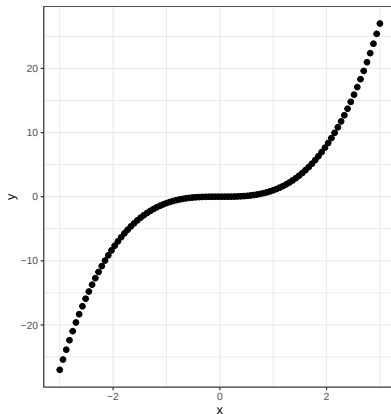
- ▶ **Pearson's correlation** measures the strength and direction of a *linear relationship*.
- ▶ **Spearman's correlation** measures the strength and direction of a *monotonic relationship*.

Note: A relationship is *monotonic* if larger values of one variable tend to be associated with larger (or smaller) values of the other, even if the relationship is not linear.

Spearman's Correlation

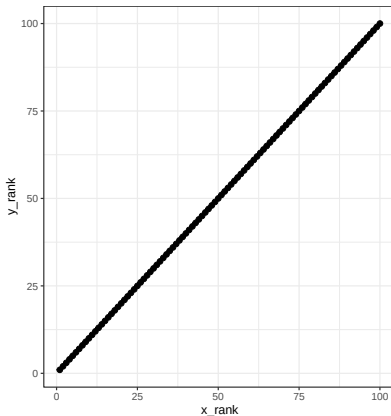
Original Data (n = 100)

Pearson: 0.917, Spearman: 1



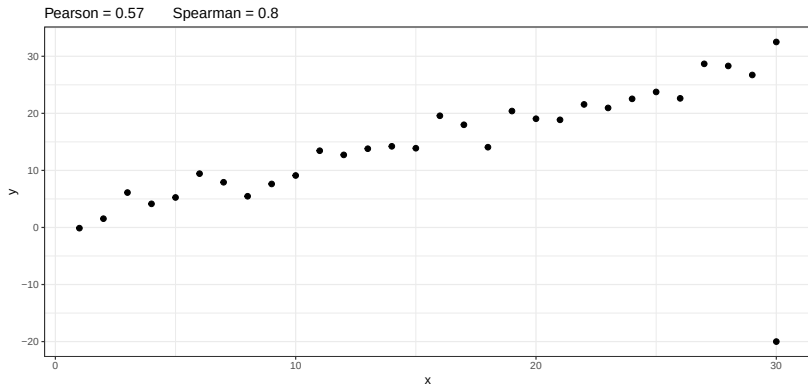
After Ranking

Perfect Linear Relationship



Spearman's Correlation

Spearman's correlation is more robust to outliers than Pearson's correlation:



Practice

1. Spearman's correlation can be found using the argument `method = "spearman"` in the `cor()` function (see example below). Find Spearman's correlation between the variables `Adm_Rate` and `med_ACT` in the Grinnell peers data set
2. Pearson's correlation between these two variables is -0.824, create a scatterplot of the data and use it to explain why Pearson's and Spearman's correlations show noticeably different degrees of strength.

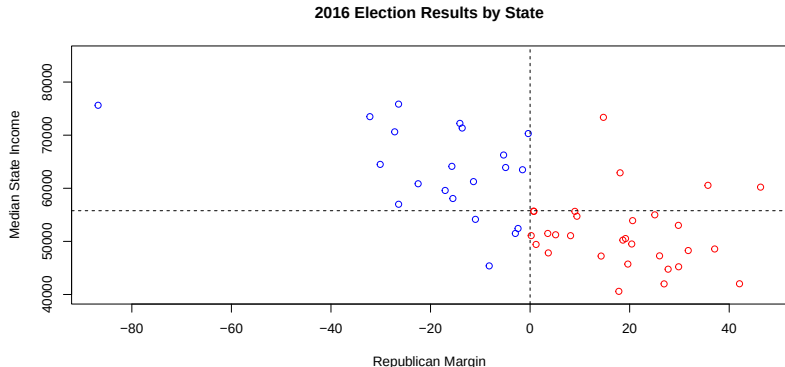
```
## Example
```

```
cor(x=draft_data$Pick, y=draft_data$St, method="spearman")
```

```
## [1] -0.5515649
```

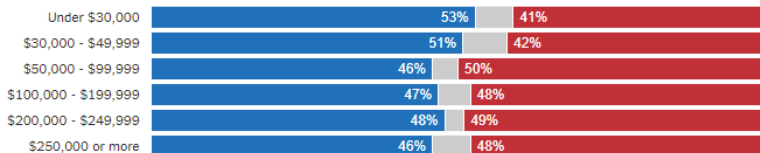
Other Correlation Misconceptions

Below are state-level data for vote margin in the 2016 presidential election and median household income. Note that $r = -0.63$, so do Republican voters tend to have lower incomes than Democrat voters?



Ecological Fallacy

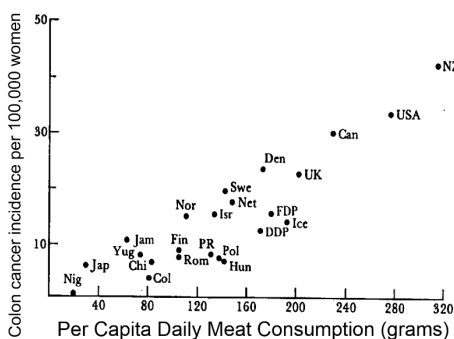
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



The **ecological fallacy** occurs when conclusions about *individuals* are made using aggregated (ecological) data. We cannot make reliable conclusions about individuals using aggregate data.

Practice

- 1) Describe the association (form, strength, and direction) and estimate the correlation coefficient
- 2) Explain how the ecological fallacy might impact the conclusion most people are tempted to draw from this graph



Conclusion

- ▶ We use scatterplots to visualize and qualitatively understand relationships between two quantitative variables
- ▶ Correlation is used to quantify the strength of relationship
 - ▶ Pearson's correlation for linear relationships
 - ▶ Spearman's correlation for non-linear (but monotonic) relationships
- ▶ The ecological fallacy frequently arises in correlation analysis, so you should be careful when analyzing aggregate data (ie: state-level, country-level, etc.)
- ▶ Another common mistake in correlation analysis is inferring that a strong correlation implies a causal relationship (we'll talk more about causation soon)