# Statistical Inference for Correlation and Regression

Ryan Miller

# Outline

- Video #1
  - Inference on the correlation coefficient
- Video #2
  - Inference for simple linear regression

# Introduction

So far, we've covered statistical methods for evaluating associations between following combinations of variables:

- Two categorical variables - difference in proportions (two-sample) z-test
- One quantitative and one categorical variable - difference in means (two-sample) t-test

This week our focus will be on the remaining combination: two quantitative variables

# Review of the Correlation Coefficient

- The correlation coefficient, $r$, is a standardized measure of the strength of *linear association* between two variables, $X$ and $Y$:

$$r_{xy} = \frac{1}{n-1} \sum_i \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

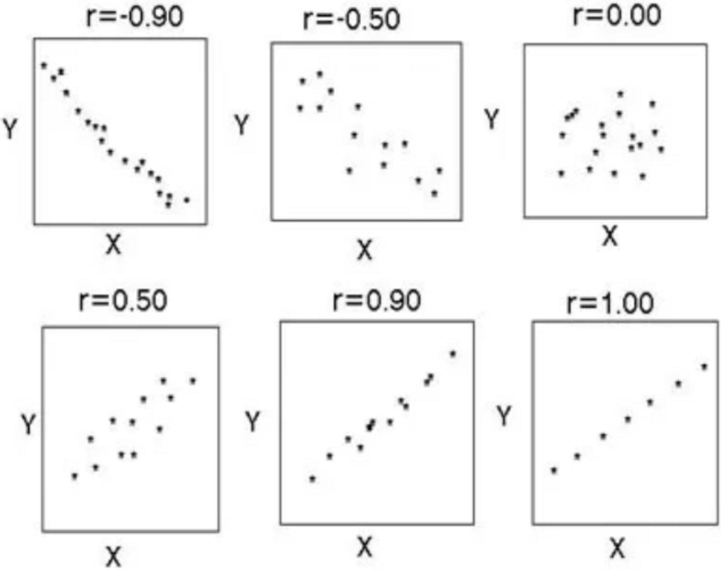
# Review of the Correlation Coefficient

- The correlation coefficient, $r$, is a standardized measure of the strength of *linear association* between two variables, $X$ and $Y$:

$$r_{xy} = \frac{1}{n-1} \sum_i \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

- Correlations near 1.0 indicate a *strong, positive relationship* (ie: higher values of $X$ correspond with higher values $Y$)
- Correlations near -1.0 indicate a *strong, negative relationship* (ie: higher values of $X$ correspond with *lower* values $Y$)
- Correlations near 0 indicate no *linear* association

# Correlation Examples



# Correlation and Statistical Inference

- Like any descriptive measure, the correlation coefficient observed in a sample is unlikely to perfectly reflect the correlation of the entire population

# Correlation and Statistical Inference

- Like any descriptive measure, the correlation coefficient observed in a sample is unlikely to perfectly reflect the correlation of the entire population
  - As was the case with other descriptive measures, we can use a *probability model* to describe the statistical uncertainty in the correlation coefficient observed in sample data:

$$r \sim N\left(\rho, \sqrt{\tfrac{1-\rho^2}{n-2}}\right)$$

- Note: The Central Limit theorem is not the basis of this probability model; additionally, there are accurate (though more complex) models for the sampling distribution of the correlation coefficient (which we will not cover)

# The t-distribution

Notice the sample correlation coefficient requires us to use *two different* sample standard deviations:

$$r_{xy} = \frac{1}{n-1}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Similar to what we've seen in other scenarios involving quantitative data, this will create extra uncertainty in our inferences about the population
  - To account for this additional uncertainty, we must use a t-distribution (this time with $n-2$ degrees of freedom, as we are estimating two additional parameters using the sample data)

# Correlation and Statistical Inference

The aforementioned probability model can serve as the basis for confidence intervals:

$$r \pm t^* \sqrt{\frac{1-r^2}{n-2}}$$

It can also be used as the basis for a T-test (usually of the null hypothesis $H_0 : \rho = 0$):

$$T = \frac{r - 0}{\sqrt{\frac{1-r^2}{n-2}}}$$

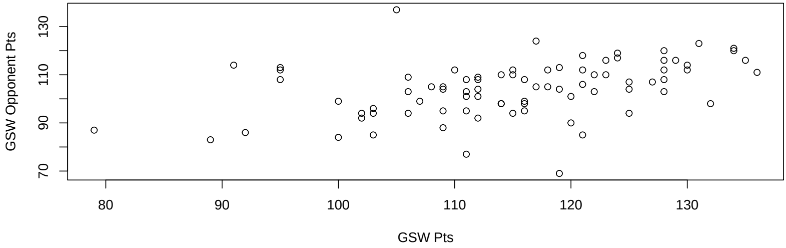
# Example

Do basketball teams tend to play to style and ability of their opponents?

- If this is the case, the points scored by each team should be correlated
- If it's not the case, the correlation between each team's point total should be zero (with some degree of sampling variability)

# Example

The scatterplot below displays game results from the Golden State Warrior's historical 2014-15 NBA season:




Let's use StatKey to calculate the correlation coefficient and perform a hypothesis test on it. The data are available here: https://remiller1450.github.io/data/GSWarriors.csv

# Example (continued)

- The sample correlation is $r = 0.412$
  - We want to test $H_0 : \rho = 0$

# Example (continued)

- The sample correlation is $r = 0.412$
  - We want to test $H_0 : \rho = 0$

To perform this test, we can calculate a T-value of
$T = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.412-0}{\sqrt{\frac{1-0.412^2}{82-2}}} = 4.04$

# Example (continued)

- The sample correlation is $r = 0.412$
  - We want to test $H_0 : \rho = 0$

To perform this test, we can calculate a T-value of
$T = \frac{r-0}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.412-0}{\sqrt{\frac{1-0.412^2}{82-2}}} = 4.04$

- Comparing this T-value with a t-distribution (having $df = 80$), the two-sided p-value is 0.0001
  - So, we can conclude there is a correlation between the Warrior's and their opponents points scored

# Guidelines for Interpretting Clinical Significance

| 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/

# Review of Simple Linear Regression

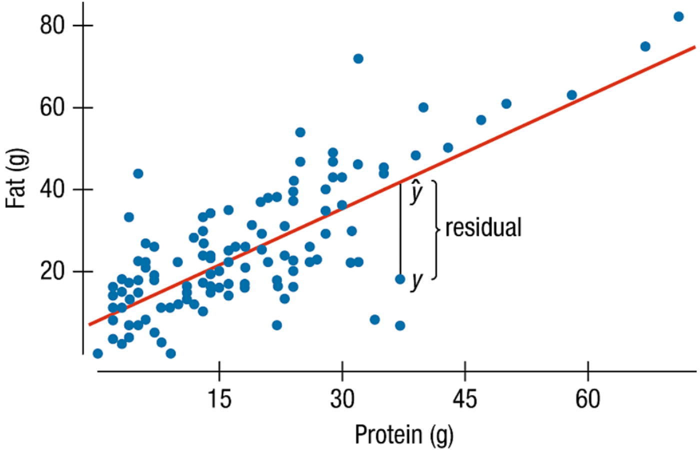
- Simple linear regression uses a straight-line (ie: a slope and an intercept) to model the relationship between an explanatory and a response variable
- The *population-level* model is stated below:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- $\beta_0$ is the model's y-intercept
- $\beta_1$ is the model's slope
- $\epsilon$ is a *random error* component that allows individual data-points to deviate from the line

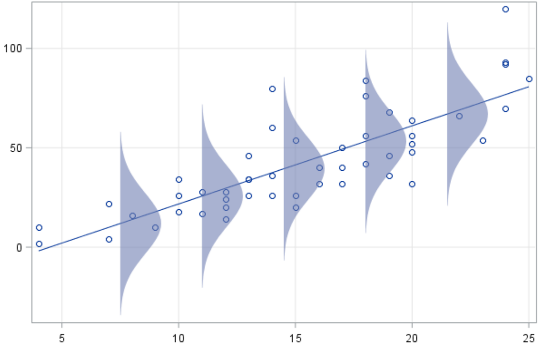
# Population vs. Fitted Models

To make use of the simple linear regression model, the unknown population-level parameters need to estimated from sample data via *least squares estimation*:



# Population vs. Fitted Models

- In addition to estimating the slope and intercept, the *variance* of the model's random errors is also estimated
  - The error variance is important for statistical inference, as it describes how much uncertainty exists in the sample data



# Inference on the Model's Slope

Typically, the most interesting statistical test that can be performed on a simple linear regression model is whether the population-level slope could be zero:

$$H_0 : \beta_1 = 0$$

This can be evaluated using a t-test:

$$T = \frac{b_1 - 0}{SE}$$

- For similar linear regression, this T-value is expected to follow a t-distribution with $n - 2$ degrees of freedom
- The standard error of $b_1$ is complicated and is typically found using statistical software

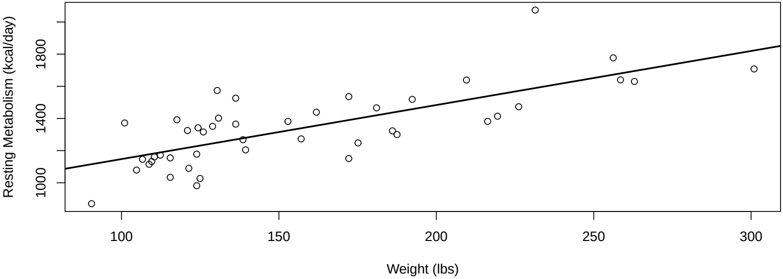
# Inference on the Model's Slope

Similarly, confidence interval estimates for the population-level slope are also based upon a t-distribution with $n - 2$ degrees of freedom:

$$\text{Estimate} \pm MOE$$
$$b_1 \pm t^* SE$$

# Example

Shown below are data on $n = 44$ adult women, along with a simple linear regression model that uses the variable "weight" (in lbs) to predicting resting metabolism (in kcal/day):




```
##
## Call:
## lm(formula = rate_kcal ~ weight_lbs, data = rmr)
##
```

# Example (continued)

- The estimated slope and intercept of this model are $b_0 = 811.2$ and $b_1 = 3.36$ respectively
  - The standard error of the slope is 0.466
- Based upon this information, do these data providing compelling statistical evidence that weight is associated with resting metabolism?

# Example (continued)

- The estimated slope and intercept of this model are $b_0 = 811.2$ and $b_1 = 3.36$ respectively
  - The standard error of the slope is 0.466
- Based upon this information, do these data providing compelling statistical evidence that weight is associated with resting metabolism?
  - We can answer this question via a t-test of $H_0 : \beta_1 = 0$

# Example (continued)

1) $H_0 : \beta_1 = 0$
2) $T = \frac{b_1 - 0}{SE} = \frac{3.36 - 0}{0.466} = 7.2$
3) Comparing this T-value to a t-distribution with $n - 2 = 42$ degrees of freedom, the p-value is nearly zero
4) These data provide overwhelming statistical evidence of a linear association between weight and resting metabolism in adult women

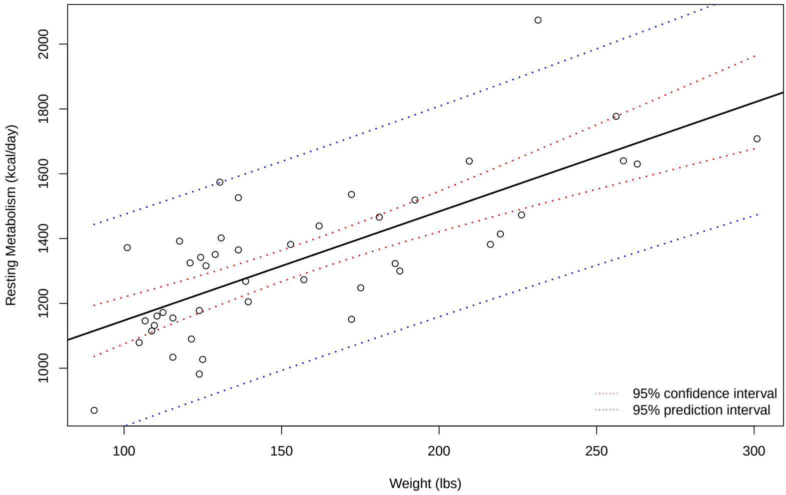
# Confidence vs. Prediction Intervals

Regression is somewhat unique statistical method in that you'll often encounter two different types of intervals:

- **Confidence Intervals** - describe the uncertainty in the *expected value* of the outcome, $Y$, given the explanatory variable, $X$
- **Prediction Intervals** - describe the uncertainty in *individual values* of the outcome, $Y$, given the explanatory variable, $X$

Prediction intervals are *always* wider than than confidence intervals, as there's less uncertainty in an expected value (ie: an average) than there is in individual data-points

# Confidence vs. Prediction Intervals



# Closing Remarks

- This presentation is aimed at providing a brief introduction to statistical inference in situations containing two quantitative variables
  - You could spent an entire course learning about the statistical details of regression modeling

# Closing Remarks

- This presentation is aimed at providing a brief introduction to statistical inference in situations containing two quantitative variables
  - You could spent an entire course learning about the statistical details of regression modeling
- For now, you should recognize the following:
  - Correlation is a symmetric method of describing the strength of linear association
  - Regression is asymmetric, meaning the choice of explanatory and response variables matter
  - Statistical inference is possible for both methods