The Correlation Coefficient and Regression

Ryan Miller

Introduction

- Lately we've been working with *regression models*, a general statistical approach that uses one or more explanatory variables to model a numeric outcome

Introduction

- Lately we've been working with *regression models*, a general statistical approach that uses one or more explanatory variables to model a numeric outcome
- *Simple linear regression* describes the special case involving a single numeric explanatory variable
    - You may recall the *correlation coefficient* is used in this same scenario (numeric explanatory variable, numeric response variable)
    - This lecture will cover the relationship between these two methods

Correlation (review)

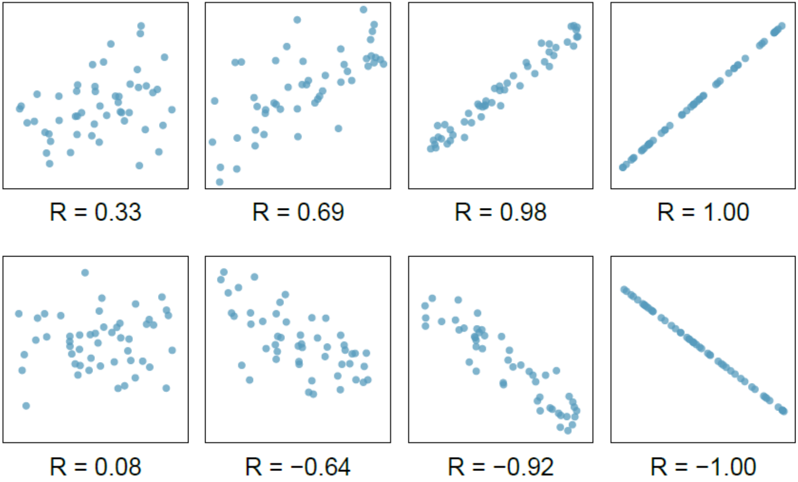
The correlation coefficient measures the strength of linear association between two numeric variables, X and Y:

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

- It is useful to understand this calculation as the *average product of Z-scores* across the two variables
  - For example, if cases tend to be either above average in both X and Y or below average in both X and Y, the correlation coefficient will be positive

Correlation (review)

Below are some examples of different correlation coefficients:



Correlation (review)

As a reminder, you should always graph the data before blindly interpreting a correlation coefficient:

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

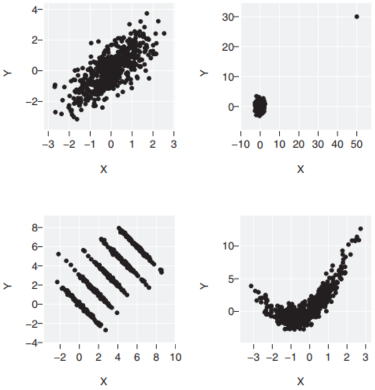


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Example - Pearson's Height Data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying heritable traits

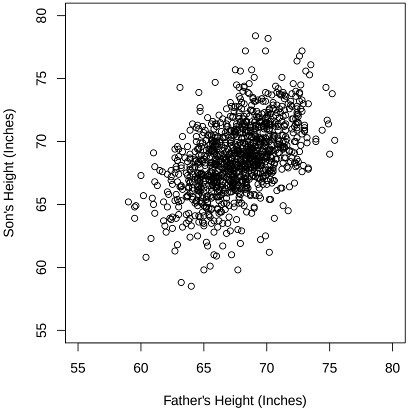
Example - Pearson's Height Data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying heritable traits
- Wondering if height is heritable, they measured the heights of 1,078 fathers and their (fully grown) sons:

Father	Son
65	59.8
63.3	63.2
65	63.3
65.8	62.8
...	...

Example - Pearson's Height Data

Does height appear heritable? To what degree?



Hypothesis Testing and Correlation

- While tall fathers tend to have tall sons, there are plenty of exceptions, so do these data provide convincing evidence of an association?
  - What hypothesis might we want to test?

Hypothesis Testing and Correlation

- While tall fathers tend to have tall sons, there are plenty of exceptions, so do these data provide convincing evidence of an association?
  - What hypothesis might we want to test?

```
pearson <- read.delim("https://remiller1450.github.io/data/Pearson.txt", sep = "\t")
cor.test(x = pearson$Father, y = pearson$Son)
```

```
##
##  Pearson's product-moment correlation
##
## data:  pearson$Father and pearson$Son
## t = 18.997, df = 1076, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4550726 0.5445746
## sample estimates:
##       cor
## 0.5011627
```

- There is overwhelming statistical evidence of a non-zero correlation between father and son heights

Hypothesis Testing and Correlation (a second example)

- Do larger colleges tend to have a higher proportion of female students? What do you make of these results?

```r
colleges19 <- read.csv("https://remiller1450.github.io/data/Colleges2019.csv")
cor.test(colleges19$PercentFemale, colleges19$Enrollment)
```

```
##
##  Pearson's product-moment correlation
##
## data:  colleges19$PercentFemale and colleges19$Enrollment
## t = -2.1114, df = 1606, p-value = 0.03489
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.101237512 -0.003740064
## sample estimates:
##         cor
## -0.05261417
```

Using Correlation to make Predictions

- Suppose we want to use Pearson and Galton's height data to make predictions

```
summary(pearson)
```

```
##      Father            Son       
##  Min.   :59.00   Min.   :58.50  
##  1st Qu.:65.80   1st Qu.:66.90  
##  Median :67.80   Median :68.60  
##  Mean   :67.69   Mean   :68.68  
##  3rd Qu.:69.60   3rd Qu.:70.50  
##  Max.   :75.40   Max.   :78.40  
```

- What height would predict for a future son of a father who is 67.69 inches tall?

Using Correlation to make Predictions

- Suppose we want to use Pearson and Galton's height data to make predictions

```r
summary(pearson)
```

```
##      Father            Son       
##  Min.   :59.00   Min.   :58.50  
##  1st Qu.:65.80   1st Qu.:66.90  
##  Median :67.80   Median :68.60  
##  Mean   :67.69   Mean   :68.68  
##  3rd Qu.:69.60   3rd Qu.:70.50  
##  Max.   :75.40   Max.   :78.40
```

- What height would predict for a future son of a father who is 67.69 inches tall?
  - This father is exactly average height, so the logical prediction is that the son is also average height, or 68.68 inches

Using Correlation to make Predictions

- How would you predict the son's height if the father were 65.0 inches, or 1 standard deviation below the average?
  - You'd be wise to predict a below average height for the son, but by how much?

Using Correlation to make Predictions

- How would you predict the son's height if the father were 65.0 inches, or 1 standard deviation below the average?
  - You'd be wise to predict a below average height for the son, but by how much?
- Part of the answer is *standardization*
  - 65.0 inches is 1 standard deviation below the average for father's height
- But we know that father's height and son's height aren't perfectly correlated, so we shouldn't expect the son to be *exactly* 1 standard deviation below average

Using Correlation to make Predictions

- How would you predict the son's height if the father were 65.0 inches, or 1 standard deviation below the average?
  - You'd be wise to predict a below average height for the son, but by how much?
- Part of the answer is *standardization*
  - 65.0 inches is 1 standard deviation below the average for father's height
- But we know that father's height and son's height aren't perfectly correlated, so we shouldn't expect the son to be *exactly* 1 standard deviation below average
  - Since the correlation is 0.50 between father/son heights, the "best" prediction we can make is that the son will be $0.5 * 1$ standard deviations between below average height

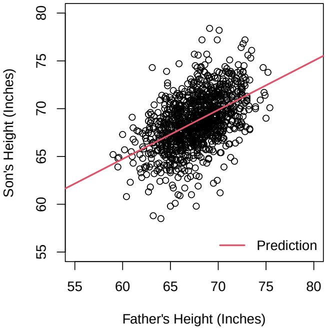
Using Correlation to make Predictions

The logic on the previous slide can be proceduralized:

1. Standardize the explanatory variable (In the previous example, $z_x = -1$)
2. Use the correlation coefficient to make a prediction (ie: $z_y = z_x * r_{xy} = -1 * .5$)
3. Un-standardize the prediction to get an answer in the original units (ie: predicted son's height $= \bar{y} + z_y * s_y$)

Using Correlation to make Predictions

The approach can be used to make predictions for *any* father's height:



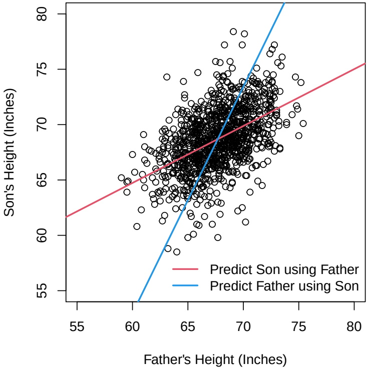
This line is the simple linear regression model!

How Regression got its Name

- The correlation coefficient is always less than 1 (in absolute value)
  - So being 1 standard deviation above/below average in the explanatory variable *always* leads to the prediction that a case is less than 1 standard deviation above/below average in the response variable
  - Galton described this phenomenon as: "regression to mediocrity"

Regression is Asymmetric

- Correlation is a **symmetric** statistical method: $r_{x,y} = r_{y,x}$
- Regression is an **asymmetric** statistical method: the choice of explanatory and response variables matters



The Madden Curse

Article Link: "Is the 'Madden' cover curse still a thing? A look back at 20 years of NFL stars offers a verdict"

- Madden is an iconic video game whose cover features a different NFL player each year, usually a player who performed exceptionally well in the previous season
- Frequently, the player featured on the Madden cover suffers from a decline in play or sustains an injury in their next season (see the article)
- Is the "Madden Curse" real? What might be a more statistically sound explanation?

Regression to Mediocrity

- Each player featured on the Madden cover was selected because they had exceptional season
- Performance in the subsequent season is correlated with that of the prior season, but the correlation is no where near 1
- The best prediction is for these players to regress
- The NFL is such that seasons near the league's statistical averages are not generally regarded as "good"
  - In 2017, the 16th rated passer was Tyrod Taylor, with 2799 yds, 14 tds, 4 ints
  - The 16th rusher was Lamar Miller with 888 yds, 3 tds

The Coefficient of Variation (R^2)

- Correlation shares another connection with regression, the **coefficient of variation**, more commonly known as R^2
  - R^2 describes the *proportion of total variability explained by the explanatory variable*

The Coefficient of Variation (R^2)

- Correlation shares another connection with regression, the **coefficient of variation**, more commonly known as R^2
  - R^2 describes the *proportion of total variability explained by the explanatory variable*
  - We can express proportion this using sums of squares (as we defined in ANOVA):

$$R^2 = \frac{SST - SSE}{SST}$$

- Recall SST is the *sum of squares total* (the maximum possible amount of modeling error), and SSE is the *sum of squared error* (how much error remains in our model)

Example in `R` (directions)

1) Fit a simple linear regression model that uses father's height to predict son's height
2) Use the `anova()` function to view an ANOVA table summarizing this model
3) Calculate the model's R^2 based upon the information in the ANOVA table, then compare your answer with the R^2 given by the `summary()` function
4) Take the square-root of the model's R^2, then compare this value with the correlation coefficient found using `cor()`

Example in R(solution)

```r
## Fit the simple linear regression
pearson <- read.delim("https://remiller1450.github.io/data/Pearson.txt", sep = "\t")
model <- lm(Son ~ Father, data = pearson)

## Get ANOVA table and extract SSE/SST
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Son
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Father       1 2145.4 2145.35   360.9 < 2.2e-16 ***
## Residuals 1076 6396.3    5.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSE <- 6396.3
SST <- SSE + 2145.4

## Cacluate R^2
(SST - SSE)/SST
```

```
## [1] 0.2511678
```

Example in R(solution)

Notice the summary() function will also calculate our model's R^2 value

```
##
## Call:
## lm(formula = Son ~ Father, data = pearson)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8910 -1.5361 -0.0092  1.6359  8.9894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.89280    1.83289   18.49   <2e-16 ***
## Father       0.51401    0.02706   19.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.438 on 1076 degrees of freedom
## Multiple R-squared:  0.2512, Adjusted R-squared:  0.2505
## F-statistic: 360.9 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Example in R(solution)

```r
## Square-root R^2
R2 <- (SST - SSE)/SST
sqrt(R2)
```

```
## [1] 0.5011664
```

```r
## Compare with cor()
cor(x = pearson$Father, y = pearson$Son)
```

```
## [1] 0.5011627
```

Comments on R^2

- Much like the correlation coefficient, R^2 can be thought of as summarizing the strength of linear association between explanatory and response variables

Comments on R^2

- Much like the correlation coefficient, R^2 can be thought of as summarizing the strength of linear association between explanatory and response variables
- Because one-way ANOVA is also a regression model, R^2 can be used to summarize the association between a categorical *explanatory variable* and a *numeric response variable*
  - This is useful because the correlation coefficient can only be calculated when both variables are numeric

Conclusion

We've now covered analysis approaches for every possible comparison of two variables:

- Two numeric variables - scatterplot, simple linear regression, correlation coefficient, R^2

Conclusion

We've now covered analysis approaches for every possible comparison of two variables:

- Two numeric variables - scatterplot, simple linear regression, correlation coefficient, R^2
- One categorical and one numeric variable - side-by-side boxplots, two-sample t-test (binary), ANOVA (nominal), Tukey's HSD (post-hoc tests)

Conclusion

We've now covered analysis approaches for every possible comparison of two variables:

- Two numeric variables - scatterplot, simple linear regression, correlation coefficient, R^2
- One categorical and one numeric variable - side-by-side boxplots, two-sample t-test (binary), ANOVA (nominal), Tukey's HSD (post-hoc tests)
- Two categorical variables - stacked bar charts, two-sample z-test (both binary), Chi-square test of independence or Fisher's exact test (either nominal)

Conclusion

We've now covered analysis approaches for every possible comparison of two variables:

- Two numeric variables - scatterplot, simple linear regression, correlation coefficient, R^2
- One categorical and one numeric variable - side-by-side boxplots, two-sample t-test (binary), ANOVA (nominal), Tukey's HSD (post-hoc tests)
- Two categorical variables - stacked bar charts, two-sample z-test (both binary), Chi-square test of independence or Fisher's exact test (either nominal)

We've also covered methods for univariate analyses:

- One numeric variable - boxplot/histogram, t-test
- One categorical variable - bar charts or pie charts, exact binomial test (binary), Chi-square goodness of fit test (nominal)