Broadening Your Statistical Horizons

Generalized Linear Models and Multilevel Models

Julie Legler and Paul Roback

2019-01-25

Contents

Preface

Broadening Your Statistical Horizons (BYSH): Generalized Linear Models and Multilevel Models is intended to be accessible to undergraduate students who have successfully completed a regression course through, for example, a textbook like *Stat2* (Cannon et al. 2019). We started teaching this course at St. Olaf in 2003 so students would be able to deal with the non-normal, correlated world we live in. It has been offered at St. Olaf every year since; in fact, it is required for all statistics concentrators. Even though there is no mathematical prerequisite, we still introduce fairly sophisticated topics such as likelihood theory, zero-inflated Poisson, and parametric bootstrapping in an intuitive and applied manner. We believe strongly in case studies featuring real data and real research questions; thus, most of the data in the textbook and available at our GitHub repo arises from collaborative research conducted by the authors and their students, or from student projects. Our goal is that, after working through this material, students will not necessarily be expert in these methods and associated theory, but that they will develop an expanded toolkit and a greater appreciation for the wider world of data and statistical modeling.

Acknowledgements. We would like to thank students of Stat 316 at St. Olaf College since 2010 for their patience as this book has taken shape with their feedback. We would especially like to thank these St. Olaf students for their summer research efforts which significantly improved aspects of this book: Cecilia Noecker, Anna Johanson, Nicole Bettes, Kiegan Rice, Anna Wall, Jack Wolf, and Josh Pelayo. Early editions of this book also benefitted greatly from feedback from instructors who used these materials in their classes, including Matt Beckman, Laura Boehm Vock, Beth Chance, Laura Chihara, Mine Dogucu, and Katie Ziegler-Graham. Finally, we have appreciated the support of two NSF grants (#DMS-1045015 and #DMS-0354308) and of our colleagues in Mathematics, Statistics, and Computer Science at St. Olaf.

Chapter 1

Review of Multiple Linear Regression

1.1 Learning Objectives

After finishing this chapter, you should be able to:

- Identify cases where ordinary least squares (OLS) assumptions are violated.
- Generate exploratory data analysis plots and summary statistics.
- Use residual diagnostics to examine OLS assumptions.
- Interpret parameters and associated tests and intervals from multiple regression models.
- Understand the basic ideas behind bootstrapped confidence intervals.

1.2 Introduction to Broadening Your Statistical Horizons

Ecologists count species, criminologists count arrests, and cancer specialists count cases. Political scientists seek to explain who is a Democrat, pre-med students are curious about who gets in to medical school, and sociologists study which people get tattoos. In the first case, ecologists, criminologists and cancer specialists are concerned about outcomes which are counts. The political scientists', pre-med students' and sociologists' interest centers on binary responses: Democrat or not, accepted or not, and tattooed or not. We can model these non-Gaussian (non-normal) responses in a more natural way by fitting **generalized linear models (GLMs)** as opposed to using to **ordinary least squares (OLS)** models.

When models are fit to data using OLS, inferences are possible using traditional statistical theory under certain conditions: if we can assume that there is a linear relationship between the response (Y) and an explanatory variable (X), the observations are independent of one another, the responses are approximately normal for each level of the X, and the variation in the responses is the same for each level of X. If we intend to make inferences using GLMs, necessary assumptions are different. First, we will not be constrained by the normality assumption. When conditions are met, GLMs can accommodate non-normal responses such as the counts and binary data in our preceding examples. While the observations must still be independent of one another, the variance in Y at each level of X need not be equal nor does the assumption of linearity between Y and X need to be plausible.

However GLMs cannot be used for models in the following circumstances: medical researchers collect data on patients in clinical trials weekly for 6 months; rat dams are injected with teratogenic substances and their offspring are monitored for defects; and, musicians' performance anxiety is recorded for several performances. Each of these examples involves correlated data: the same patient's outcomes are more likely to be similar from week-to-week than outcomes from different patients; litter mates are more likely to suffer defects at similar rates in contrast to unrelated rat pups; and, a musician's anxiety is more similar from performance to

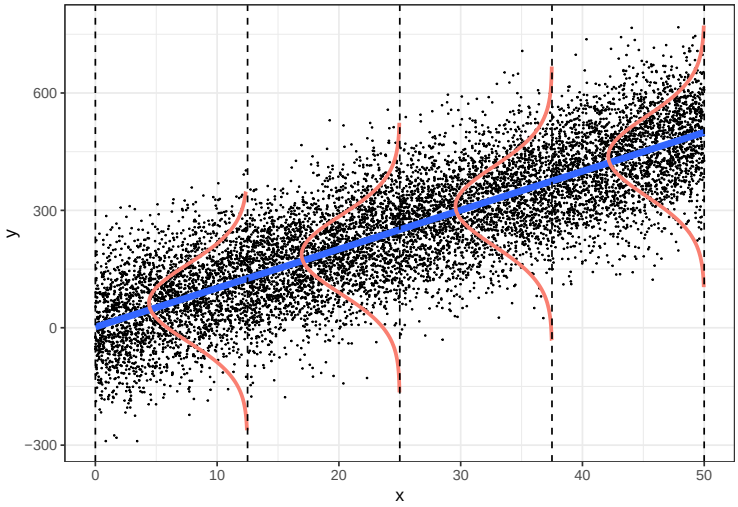

Figure 1.1: Ordinary least squares assumptions

performance than it is with other musicians. Each of these examples violate the independence assumption of simpler linear models for OLS or GLM inference.

The **Generalized Linear Models** in the book's title extends OLS methods you may have seen in linear regression to handle responses that are non-normal. The **Multilevel Methods** will allow us to create models for situations where the observations are not independent of one another. Overall, these approaches will permit us to get much more out of data and may be more faithful to the actual data structure than models based on ordinary least squares. These models will *broaden your statistical horizons.*

In order to understand the motivation for handling violations of assumptions, it is helpful to be able to recognize the model assumptions for inference with OLS in the context of different studies. While linearity is sufficient for fitting an OLS model, in order to make inferences and predictions the observations must also be independent, the responses should be approximately normal at each level of the predictors, and the standard deviation of the responses at each level of the predictors should be approximately equal. After examining circumstances where inference with OLS modeling is appropriate, we will look for violations of these assumptions in other sets of circumstances. These are settings where we may be able to use the methods of this text. We've kept the examples in the exposition simple to fix ideas. There are exercises which describe more realistic and complex studies.

1.3 Ordinary Least Squares (OLS) Assumptions

Recall that making inferences or predictions with models fit using ordinary least squares (OLS) requires that the following assumptions be tenable. The acronym LINE can be used to recall the assumptions required for making inferences and predictions with models based on OLS. If we consider a simple linear regression with just a single predictor X, then:

- **L:** there is a linear relationship between the mean response (Y) and the explanatory variable (X),
- **I:** the errors are independent—there's no connection between how far any two points lie from the regression line,
- **N:** the responses are normally distributed at each level of X, and
- **E:** the variance or, equivalently, the standard deviation of the responses is equal for all levels of X.

These assumptions are depicted in Figure 1.1.

- **L:** The mean value for Y at each level of X falls on the regression line.

- **I:** We'll need to check the design of the study to determine if the errors (vertical distances from the line) are independent of one another.
- **N:** At each level of X, the values for Y are normally distributed.
- **E:** The spread in the Y's for each level of X is the same.

1.3.1 Cases that do not violate the OLS assumptions for inference

It can be argued that the following studies do not violate the OLS assumptions for inference. We begin by identifying the response and the explanatory variables followed by describing each of the LINE assumptions in the context of the study, commenting on possible problems with the assumptions.

1) **Reaction times and car radios.** A researcher suspects that loud music can affect how quickly drivers react. She randomly selects drivers to drive the same stretch of road with varying levels of music volume. Stopping distances for each driver are measured along with the decibel level of the music on their car radio.

   - *Response variable:* Reaction time
   - *Explanatory variable:* Decibel level of music

   The OLS assumptions for inference would apply if:

   - **L:** The mean reaction time is linearly related to decibel level of the music.
   - **I:** Stopping distances are independent. The random selection of drivers should assure independence.
   - **N:** The stopping distances for a given decibel level of music vary and are normally distributed.
   - **E:** The variation in stopping distances should be approximately the same for each decibel level of music.

   There are potential problems with the linearity and equal standard deviation assumptions. For example, if there is threshold for the volume of music where the effect on reaction times remains the same, mean reaction times would not be a linear function of music. Another problem may occur if a few subjects at each decibel level took a really long time to react. In this case, reaction times would be right skewed and the normality assumption would be violated. Often we can think of circumstances where the OLS assumptions may be suspect. Later in this chapter we will describe plots which can help diagnose issues with OLS assumptions.

2) **Crop yield and rainfall.** The yield of wheat per acre for the month of July is thought to be related to the rainfall. A researcher randomly selects acres of wheat and records the rainfall and bushels of wheat per acre.

   - *Response variable:* Yield of wheat measured in bushels per acre for July
   - *Explanatory variable:* Rainfall measured in inches for July
   - **L:** The mean yield per acre is linearly related to rainfall.
   - **I:** Fields' yields are independent; knowing one (X, Y) pair does not provide information about another.
   - **N:** The yields for a given amount of rainfall are normally distributed.
   - **E:** The standard deviation of yields is approximately the same for each rainfall level.

   Again we may encounter problems with the linearity assumption if mean yields increase initially as the amount of rainfall increases after which excess rainfall begins to ruin crop yield. The random selection of fields should assure independence if fields are not close to one another.

3) **Heights of sons and fathers.** Sir Francis Galton suspected that a son's height could be predicted using the father's height. He collected observations on heights of fathers and their firstborn sons (Stigler 2002).

   - *Response variable:* Height of the firstborn son
   - *Explanatory variable:* Height of the father
   - **L:** The mean height of firstborn sons is linearly related to heights of fathers.
   - **I:** The height of one firstborn son is independent of the heights of other firstborn sons in the study. This would be the case if firstborn sons were randomly selected.
   - **N:** The heights of firstborn sons for a given fathers' height are normally distributed.
   - **E:** The standard deviation of firstborn sons' heights at a given father's height are the same.

   Heights and other similar measurements are often normally distributed. There would be a problem with the independence assumption if multiple sons from the same family were selected. Or, there would be a problem with equal variance if sons of tall fathers had much more variety in their heights than sons of shorter fathers.

1.3.2 Cases where the OLS assumptions for inference are violated

1) **Grades and studying.** Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

   - *Response variable:* Exam outcome (Pass or Fail)
   - *Explanatory variable:* Time spent studying (in hours)

   Here the response is a binary outcome which violates the OLS assumption of a normally distributed response at each level of X. In Chapter **??**, we will see logistic regression which is more suitable for models with binary responses.

2) **Income and family size.** Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

   - *Response variable:* Family size, number of children
   - *Explanatory variable:* Annual income, in dollars

   Family size is a count taking on integer values from 0 to (technically) no upper bound. The normality assumption may be problematic again because the distribution of family size is likely to be skewed, with more families having one or two children and only a few with a much larger number of children. Both of these concerns lead us to question the validity of the normality assumption. Study design should also specify that families are done adding children to their family.

3) **Exercise, weight, and sex.** Investigators collected the weight, sex, and amount of exercise for a random sample of college students.

   - *Response variable:* Weight
   - *Explanatory variables:* Sex and hours spent exercising in a typical week

   With two predictors, the assumptions now apply to the combination of sex and exercise. For example, the linearity assumption implies that there is a linear relationship in mean weight and amount of exercise for males and, similarly, a linear relationship in mean weight and amount of exercise for females. This data may not be appropriate for OLS modeling because the standard deviation in weight for students

who do not exercise for each sex is likely to be considerably greater than the standard deviation in weight for students who follow an exercise regime. We can assess this potential problem by plotting weight by amount of exercise for males and females separately. There may also be a problem with the independence assumption because there is no indication that the subjects were randomly selected. There may be subgroups of subjects likely to be more similar, e.g. selecting students at a gym and others in a TV lounge.

4) **Surgery Outcome and Patient Age.** Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of one to ten.

   - *Response variable:* Surgery outcome, scale 1-10
   - *Explanatory variable:* Patient age, in years

   Outcomes for patients operated on by the same surgeon are more likely to be similar and have similar results. For example, if surgeons' skills differ or if their criteria for selecting patients for surgery vary, individual surgeons may tend to have better or worse outcomes, and patient outcomes will be dependent on surgeon. Furthermore, outcomes at one hospital may be more similar possibly due to factors associated with different patient populations. The very structure of this data suggests that the independence assumption will be violated. Multilevel models will explicitly take this structure into account for a proper analysis of this study's results.

While we identified possible violations of OLS assumptions for inference for each of the examples in this section, there may be violations of the other assumptions that we have not pointed out. Prior to reading this book, you have presumably learned some ways to handle these violations such as applying variance stabilizing transformations or logging responses, but you will discover other models in this text that may be more appropriate for the violations we have presented.

1.4 Review of Multiple Linear Regression

1.4.1 Case Study: Kentucky Derby

Before diving into generalized linear models and multilevel modeling, we review key ideas from multiple linear regression using an example from horse racing. The Kentucky Derby is a 1.25 mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky. Our data set `derbyplus.csv` contains the `year` of the race, the winning horse (`winner`), the `condition` of the track, the average `speed` (in feet per second) of the winner, and the number of `starters` (field size, or horses who raced) for the years 1896-2017 (Wikipedia contributors 2018). The track `condition` has been grouped into three categories: fast, good (which includes the official designations "good" and "dusty"), and slow (which includes the designations "slow", "heavy", "muddy", and "sloppy"). We would like to use OLS linear regression techniques to model the speed of the winning horse as a function of track condition, field size, and trends over time.

1.5 Initial Exploratory Analyses

1.5.1 Data Organization

The first five and last five rows from our data set are illustrated in Table 1.5.1. Note that, in certain cases, we created new variables from existing ones:

- `fast` is an **indicator variable**, taking the value 1 for races run on fast tracks, and 0 for races run under other conditions,

- `good` is another indicator variable, taking the value 1 for races run under good conditions, and 0 for races run under other conditions,
- `yearnew` is a **centered variable**, where we measure the number of years since 1896, and
- `fastfactor` replaces `fast` = 0 with the description "not fast", and `fast` = 1 with the description "fast". Changing a numeric categorical variable to descriptive phrases can make plot legends more meaningful.

The first five and the last five observations from the Kentucky Derby case study.

year

winner

condition

speed

starters

fast

good

yearnew

fastfactor

1896

Ben Brush

good

51.66

8

0

1

0

not fast

1897

Typhoon II

slow

49.81

6

0

0

1

not fast

1898

Plaudit

good

51.16

4

0

1

2

not fast

1899

Manuel

fast

50.00

5

1

0

3

fast

1900

Lieut. Gibson

fast

52.28

7

1

0

4

fast

2013

Orb

slow

53.71

19

0

0

117

not fast

2014

California Chrome

fast

53.37

19

1

0

118

fast

2015

American Pharoah

fast

53.65

18

1

0

119

fast

2016

Nyquist

fast

54.41

20

1

0

120

fast

2017

Always Dreaming

fast

53.40

20

1

0

121

fast

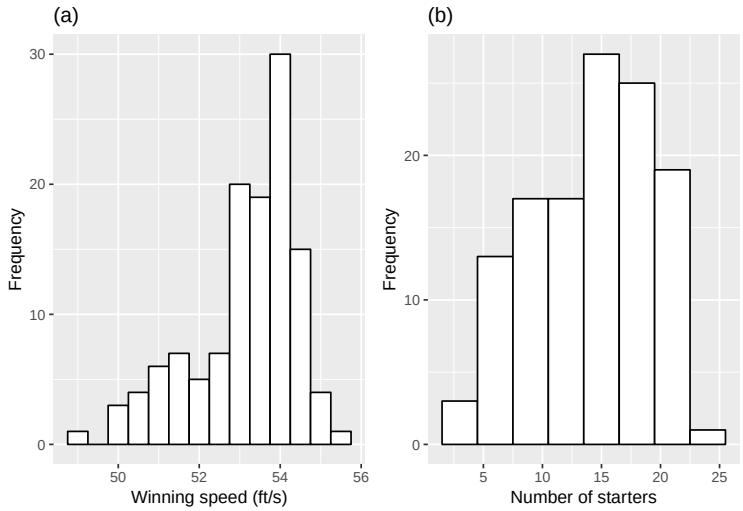

Figure 1.2: Histograms of key continuous variables. Plot (a) shows winning speeds, while plot (b) shows the number of starters.

1.5.2 Univariate Summaries

With any statistical analysis, our first task is to explore the data, examining distributions of individual responses and predictors using graphical and numerical summaries, and beginning to discover relationships between variables. This should *always* be done *before* any model fitting! We must understand our data thoroughly before doing anything else.

First, we will examine each response variable and potential covariate individually. Continuous variables can be summarized using histograms and statistics indicating center and spread; categorical variables can be summarized with tables and possibly bar charts.

In Figure 1.2(a), we see that the primary response, winning speed, follows a distribution with a slight left skew, with a large number of horses winning with speeds between 53-55 feet per second. Plot (b) shows that the number of starters is mainly distributed between 5 and 20, with the largest number of races having between 15 and 20 starters.

The primary categorical explanatory variable is track condition, where 88 (72%) of the 122 races were run under fast conditions, 10 (8%) under good conditions, and 24 (20%) under slow conditions.

1.5.3 Bivariate Summaries

The next step in an initial exploratory analysis is the examination of numerical and graphical summaries of relationships between model covariates and responses. Figure 1.3 is densely packed with illustrations of bivariate relationships. The relationship between two continuous variables is depicted with scatterplots below the diagonal and correlation coefficients above the diagonal. Here, we see that higher winning speeds are associated with more recent years, while the relationship between winning speed and number of starters is less clear cut. We also see a somewhat strong correlation between year and number of starters—we should be aware of highly correlated explanatory variables whose contributions might overlap too much.

Relationships between categorical variables like track condition and continuous variables can be illustrated with side-by-side boxplots as in the top row, or with stacked histograms as in the first column. As expected, we see evidence of higher speeds on fast tracks and also a tendency for recent years to have more fast conditions. These observed trends can be supported with summary statistics generated by subgroup. For instance, the mean speed under fast conditions is 53.6 feet per second, compared to 52.7 ft/s under good

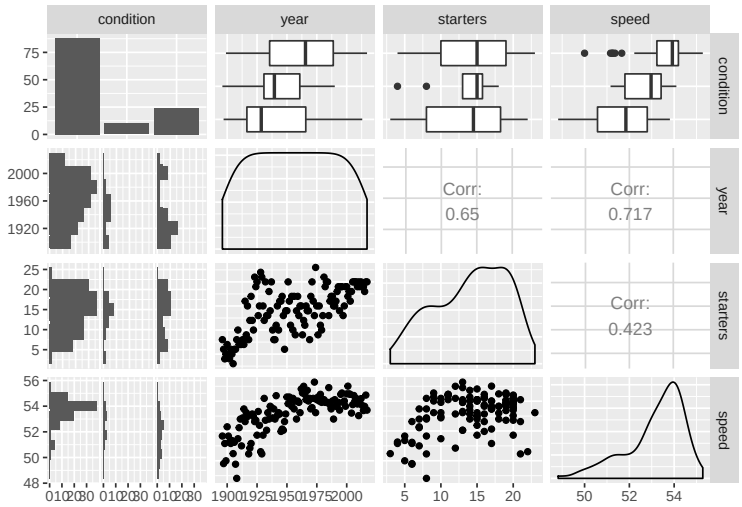

Figure 1.3: Relationships between pairs of variables in the Kentucky Derby data set

conditions and 51.7 ft/s under slow conditions. Variability in winning speeds, however, is greatest under slow conditions (SD = 1.36 ft/s) and least under fast conditions (0.94 ft/s).

Finally, notice that the diagonal illustrates the distribution of individual variables, using density curves for continuous variables and a bar chart for categorical variables. Trends observed in the last two diagonal entries match trends observed in Figure 1.2.

By using shape or color or other attributes, we can incorporate the effect of a third or even fourth variable into the scatterplots of Figure 1.3. For example, in the **coded scatterplot** of Figure 1.4 we see that speeds are generally faster under fast conditions, but the rate of increasing speed over time is greater under good or slow conditions.

Of course, any graphical analysis is exploratory, and any notable trends at this stage should be checked through formal modeling. At this point, a statistician begins to ask familiar questions such as:

- are winning speeds increasing in a linear fashion?
- does the rate of increase in winning speed depend on track condition or number of starters?
- after accounting for other explanatory variables, is greater field size (number of starters) associated with faster winning speeds (because more horses in the field means a greater chance one horse will run a very fast time) or slower winning speeds (because horses are more likely to bump into each other or crowd each others' attempts to run at full gait)?
- are any of these associations statistically significant?
- how well can we predict the winning speed in the Kentucky Derby?

As you might expect, answers to these questions will arise from proper consideration of variability and properly identified statistical models.

1.6 Multiple linear regression modeling

1.6.1 Simple linear regression with a continuous predictor

We will begin by modeling the winning speed as a function of time; for example, have winning speeds increased at a constant rate since 1896? For this initial model, let Y_i be the speed of the winning horse in year i. Then, we might consider Model 1:

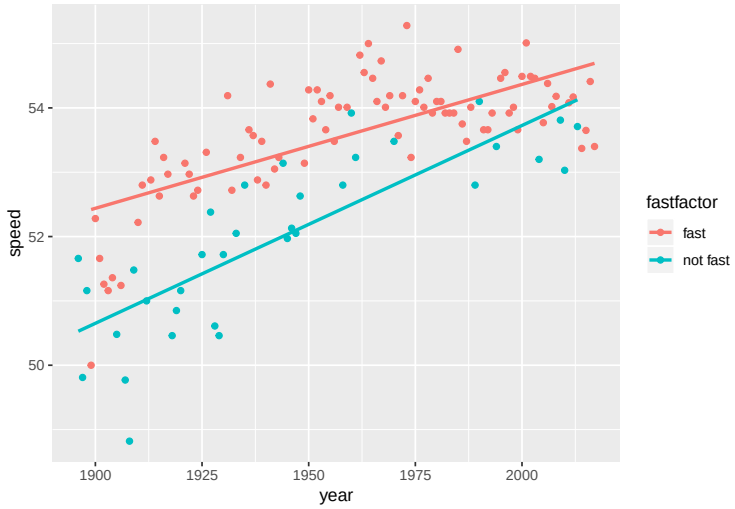

Figure 1.4: Linear trends in winning speeds over time, presented separately for fast conditions vs. good or slow conditions

$$Y_i = \beta_0 + \beta_1(\text{Year}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{1.1}$$

In this case, β_0 represents the true intercept—the expected winning speed during Year 0. β_1 represents the true slope—the expected increase in winning speed from one year to the next, assuming the rate of increase is linear (i.e., constant with each successive year since 1896). Finally, the **error** (ϵ_i) terms represent the deviations of the actual winning speed in Year i (Y_i) from the expected scores under this model ($\beta_0 + \beta_1(\text{Year}_i)$)—the part of a horse's winning speed that is not explained by a linear trend over time. The variability in these deviations from the regression model is denoted by σ^2.

The parameters in this model (β_0, β_1, and σ^2) can be estimated through OLS methods; we will use hats to denote estimates of population parameters based on empirical data. Values for $\hat{\beta}_0$ and $\hat{\beta}_1$ are selected to minimize the sum of squared residuals, where a **residual** is simply the observed prediction error—the actual winning speed for a given year minus the winning speed predicted by the model. In the notation of this section,

- Predicted speed: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(\text{Year}_i)$
- Residual (estimated error): $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- Estimated variance of points around the line: $\hat{\sigma}^2 = \sum \hat{\epsilon}_i^2/(n-2)$

Using Kentucky Derby data, we estimate $\hat{\beta}_0 = 2.05$, $\hat{\beta}_1 = 0.026$, and $\hat{\sigma} = 0.90$. Thus, according to our simple linear regression model, winning horses of the Kentucky Derby have an estimated winning speed of 2.05 ft/s in Year 0 (more than 2000 years ago!), and the winning speed improves by an estimated 0.026 ft/s every year. With an R^2 of 0.513, the regression model explains a moderate amount (51.3%) of the year-to-year variability in winning speeds, and the trend toward a linear rate of improvement each year is statistically significant at the 0.05 level ($t(120) = 11.251$, $p < .001$).

```
lm(formula = speed ~ year, data = derby.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.053473   4.543754   0.452    0.652
year        0.026126   0.002322  11.251   <2e-16 ***
---
Residual standard error: 0.9032 on 120 degrees of freedom
```

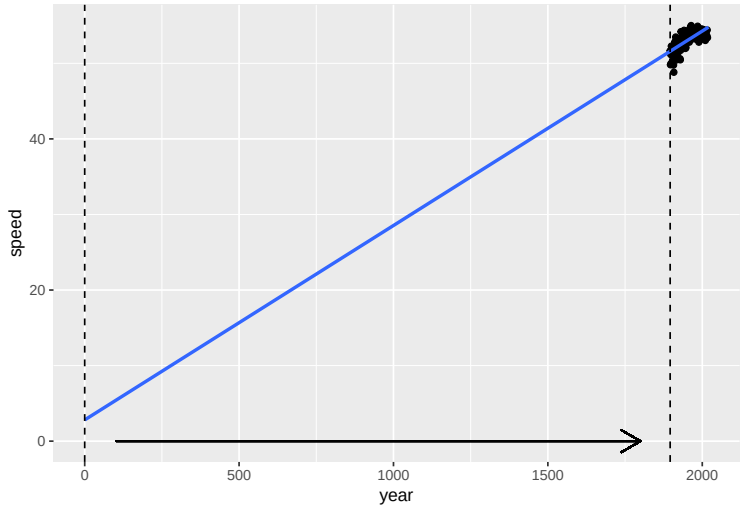

Figure 1.5: Compare Model 1 (with intercept at 0) to Model 2 (with intercept at 1896)

```
Multiple R-squared:  0.5134,    Adjusted R-squared:  0.5093
```

You may have noticed in Model 1 that the intercept has little meaning in context, since it estimates a winning speed in Year 0, when the first Kentucky Derby run at the current distance (1.25 miles) was in 1896. One way to create more meaningful parameters is through **centering**. In this case, we could create a centered year variable by subtracting 1896 from each year for Model 2:

$$Y_i = \beta_0 + \beta_1(\text{Yearnew}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2) \quad \text{and} \quad \text{Yearnew} = \text{Year} - 1896. \tag{1.2}$$

Note that the only thing that changes from Model 1 to Model 2 is the estimated intercept; $\hat{\beta}_1$, R^2, and $\hat{\sigma}$ all remain exactly the same. Now $\hat{\beta}_0$ tells us that the estimated winning speed in 1896 is 51.59 ft/s, but estimates of the linear rate of improvement or the variability explained by the model remain the same. As Figure 1.5 shows, centering year has the effect of shifting the y-axis from year 0 to year 1896, but nothing else changes.

```
lm(formula = speed ~ yearnew, data = derby.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.588393   0.162549  317.37   <2e-16 ***
yearnew      0.026126   0.002322   11.25   <2e-16 ***
---
Residual standard error: 0.9032 on 120 degrees of freedom
Multiple R-squared:  0.5134,    Adjusted R-squared:  0.5093
```

We should also attempt to verify that our LINE linear regression model assumptions fit for Model 2 if we want to make inferential statements (hypothesis tests or confidence intervals) about parameters or predictions. Most of these assumptions can be checked graphically using a set of residual plots as in Figure 1.6:

- The upper left plot, Residuals vs. Fitted, can be used to check the Linearity assumption. Residuals should be patternless around Y = 0; if not, there is a pattern in the data that is currently unaccounted for.
- The upper right plot, Normal Q-Q, can be used to check the Normality assumption. Deviations from a straight line indicate that the distribution of residuals does not conform to a theoretical normal curve.

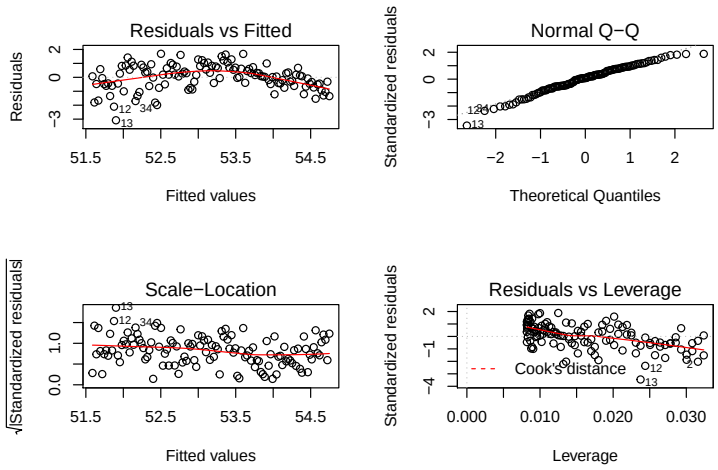

Figure 1.6: Residual plots for Model 2

- The lower left plot, Scale-Location, can be used to check the Equal Variance assumption. Positive or negative trends across the fitted values indicate variability that is not constant.
- The lower right plot, Residuals vs. Leverage, can be used to check for influential points. Points with high leverage (having unusual values of the predictors) and/or high absolute residuals can have an undue influence on estimates of model parameters.

In this case, the Residuals vs. Fitted plot indicates that a quadratic fit might be better than the linear fit of Model 2; other assumptions look reasonable. Influential points would be denoted by high values of Cook's Distance; they would fall outside cutoff lines in the northeast or southeast section of the Residuals vs. Leverage plot. Since no cutoff lines are even noticeable, there are no potential influential points of concern.

We recommend relying on graphical evidence for identifying regression model assumption violations, looking for highly obvious violations of assumptions before trying corrective actions. While some numerical tests have been devised for issues such as normality and influence, most of these tests are not very reliable, highly influenced by sample size and other factors. There is typically no residual plot, however, to evaluate the Independence assumption; evidence for lack of independence comes from knowing about the study design and methods of data collection. In this case, with a new field of horses each year, the assumption of independence is pretty reasonable.

Based on residual diagnostics, we should test Model 2Q, in which a quadratic term is added to the linear term in Model 2.

$$Y_i = \beta_0 + \beta_1(\text{Yearnew}_i) + \beta_2(\text{Yearnew}_i^2) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{1.3}$$

This model could suggest, for example, that the rate of increase in winning speeds is slowing down over time. In fact, there is evidence that the quadratic model improves upon the linear model (see Figure 1.7). R^2, the proportion of year-to-year variability in winning speeds explained by the model, has increased from 51.3% to 64.1%, and the pattern in the Residuals vs. Fitted plot of Figure 1.6 has disappeared in Figure 1.8, although normality is a little sketchier in the left tail, and the larger mass of points with fitted values near 54 appears to have slightly lower variability. The significantly negative coefficient for β_2 suggests that the rate of increase is indeed slowing in more recent years.

```
lm(formula = speed ~ yearnew + yearnew2, data = derby.df)

Coefficients:
```

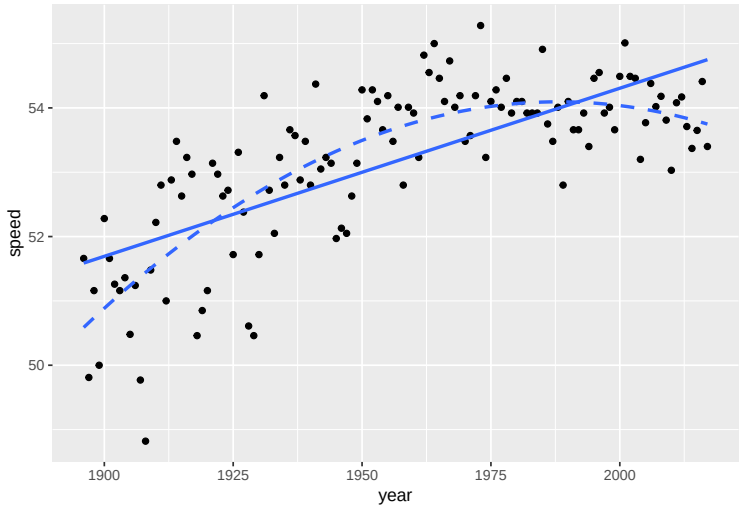

Figure 1.7: Linear vs. quadratic fit

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.059e+01  2.082e-01 243.010  < 2e-16 ***
yearnew      7.617e-02  7.950e-03   9.581  < 2e-16 ***
yearnew2    -4.136e-04  6.359e-05  -6.505 1.92e-09 ***
---
Residual standard error: 0.779 on 119 degrees of freedom
Multiple R-squared:  0.641, Adjusted R-squared:  0.635
```

1.6.2 Simple linear regression with a binary predictor

We also may want to include track condition as an explanatory variable. We could start by using `fast` as the lone predictor: Do winning speeds differ for fast and non-fast conditions? `fast` is considered an **indicator variable**—it takes on only the values 0 and 1, where 1 indicates presence of a certain attribute (like fast racing conditions). Since `fast` is numeric, we can use simple linear regression techniques to fit Model 3:

$$Y_i = \beta_0 + \beta_1(\text{Fast}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{1.4}$$

Here, it's easy to see the meaning of our slope and intercept by writing out separate equations for the two conditions:

- Good or slow conditions (`fast` $= 0$)

$$Y_i = \beta_0 + \epsilon_i \tag{1.5}$$

- Fast conditions (`fast` $= 1$)

$$Y_i = (\beta_0 + \beta_1) + \epsilon_i \tag{1.6}$$

β_0 is the expected winning speed under good or slow conditions, while β_1 is the difference between expected winning speeds under fast conditions vs. non-fast conditions. According to our fitted Model 3, the estimated winning speed under non-fast conditions is 52.0 ft/s, while mean winning speeds under fast conditions are estimated to be 1.6 ft/s higher.

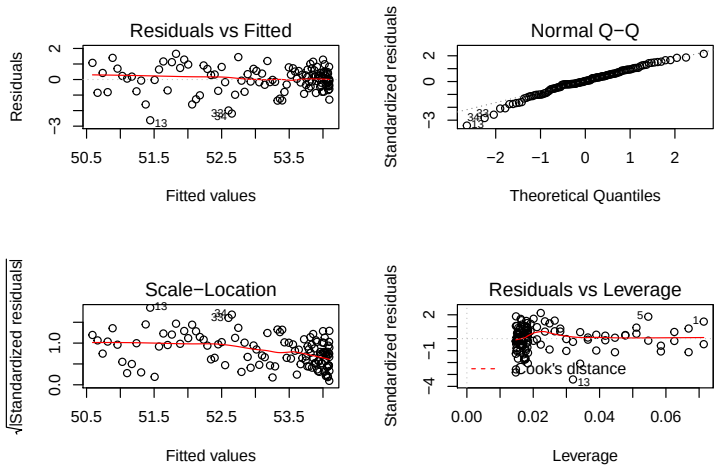

Figure 1.8: Residual plots for Model 2Q

```
lm(formula = speed ~ fast, data = derby.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.9938     0.1826 284.698  < 2e-16 ***
fast          1.6292     0.2150   7.577 8.17e-12 ***
---
Residual standard error: 1.065 on 120 degrees of freedom
Multiple R-squared:  0.3236,    Adjusted R-squared:  0.318
```

You might be asking at this point: If we simply wanted to compare mean winning speeds under fast and non-fast conditions, why didn't we just run a two-sample t-test? The answer is: we did! The t-test corresponding to β_1 is equivalent to an independent-samples t-test with under equal variances. Convince yourself that this is true, and that the equal variance assumption is needed.

1.6.3 Multiple linear regression with two predictors

The beauty of the linear regression framework is that we can add additional explanatory variables in order to explain more variability in our response, obtain better and more precise predictions, and control for certain covariates while evaluating the effect of others. For example, we could consider adding `yearnew` to Model 3, which has the indicator variable `fast` as its only predictor. In this way, we would estimate the difference between winning speeds under fast and non-fast conditions *after accounting for the effect of time*. As we observed in Figure 1.3, recent years have tended to have more races under fast conditions, so Model 3 might overstate the effect of fast conditions because winning speeds have also increased over time. A model with terms for both year and track condition will estimate the difference between winning speeds under fast and non-fast conditions *for a fixed year*; for example, if it had rained in 2016 and turned the track muddy, how much would we have expected the winning speed to decrease?

Our new model (Model 4) can be written:

$$Y_i = \beta_0 + \beta_1(\text{Yearnew}_i) + \beta_2(\text{Fast}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \tag{1.7}$$

and OLS provides the following parameter estimates:

```
lm(formula = speed ~ yearnew + fast, data = derby.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.917822   0.154602  329.35  < 2e-16 ***
yearnew      0.022583   0.001919   11.77  < 2e-16 ***
fast         1.226846   0.150721    8.14 4.39e-13 ***
---
Residual standard error: 0.7269 on 119 degrees of freedom
Multiple R-squared:  0.6874,    Adjusted R-squared:  0.6822
```

Our new model estimates that winning speeds are, on average, 1.23 ft/s faster under fast conditions after accounting for time trends, which is down from an estimated 1.63 ft/s without accounting for time. It appears our original model (Model 3) may have overestimated the effect of fast conditions by conflating it with improvements over time. Through our new model, we also estimate that winning speeds increase by 0.023 ft/s per year, after accounting for track condition. This yearly effect is also smaller than the 0.026 ft/s per year we estimated in Model 1, without adjusting for track condition. Based on the R^2 value, Model 4 explains 68.7% of the year-to-year variability in winning speeds, a noticeable increase over using either explanatory variable alone.

1.6.4 Inference in multiple linear regression: normal theory

So far we have been using linear regression for descriptive purposes, which is an important task. We are often interested in issues of statistical inference as well—determining if effects are statistically significant, quantifying uncertainty in effect size estimates with confidence intervals, and quantifying uncertainty in model predictions with prediction intervals. Under LINE assumptions, all of these inferential tasks can be completed with the help of the t-distribution and estimated standard errors.

Here are examples of inferential statements based on Model 4:

- We can be 95% confident that average winning speeds under fast conditions are between 0.93 and 1.53 ft/s higher than under non-fast conditions, after accounting for the effect of year.
- Fast conditions lead to significantly faster winning speeds than non-fast conditions (t = 8.14 on 119 df, p < .001), holding year constant.
- Based on our model, we can be 95% confident that the winning speed in 2017 under fast conditions will be between 53.4 and 56.3 ft/s. Note that Always Dreaming's actual winning speed barely fit within this interval—the 2017 winning speed was a borderline outlier on the slow side.

```
confint(model4)
```

```
                2.5 %      97.5 %
(Intercept) 50.61169473 51.22394836
yearnew      0.01878324  0.02638227
fast         0.92840273  1.52528902
```

```
new.data <- data.frame(yearnew = 2017 - 1896, fast = 1)
predict(model4, new = new.data, interval = "prediction")
```

```
       fit     lwr      upr
1 54.87718 53.4143 56.34006
```

1.6.5 Inference in multiple linear regression: bootstrapping

Remember that you must check LINE assumptions using the same residual plots as in Figure 1.6 to ensure that the inferential statements in the previous section are valid. In cases when model assumptions are shaky, one

alternative approach to statistical inference is **bootstrapping**; in fact, bootstrapping is a robust approach to statistical inference that we will use frequently throughout this book because of its power and flexibility. In bootstrapping, we use only the data we've collected and computing power to estimate the uncertainty surrounding our parameter estimates. Our primary assumption is that our original sample represents the larger population, and then we can learn about uncertainty in our parameter estimates through repeated samples (with replacement) from our original sample.

If we wish to use bootstrapping to obtain confidence intervals for our coefficients in Model 4, we could follow these steps:

- take a (bootstrap) sample of 122 years of Derby data with replacement, so that some years will get sampled several times and others not at all. This is **case resampling**, so that all information from a given year (winning speed, track condition, number of starters) remains together.
- fit Model 4 to the bootstrap sample, saving $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.
- repeat the two steps above a large number of times (say 1000)
- the 1000 bootstrap estimates for each parameter can be plotted to show the **bootstrap distribution** (see Figure 1.9)
- a 95% confidence interval for each parameter can be found by taking the middle 95% of each bootstrap distribution—i.e., by picking off the 2.5 and 97.5 percentiles. This is called the **percentile method**.

```
# updated code from tobiasgerstenberg on github
library(rsample)
library(purrr)
bootreg = derby.df %>%
  bootstraps(1000) %>%
  pull(splits) %>%
  map_dfr(~lm(speed ~ yearnew + fast, data = .) %>%
            tidy())
summarize = dplyr::summarize
bootreg %>%
  group_by(term) %>%
  summarize(low=quantile(estimate, .025),
            high=quantile(estimate, .975))
```

```
# A tibble: 3 x 3
  term            low    high
  <chr>         <dbl>   <dbl>
1 (Intercept) 50.6    51.3
2 fast         0.929   1.56
3 yearnew      0.0183  0.0268
```

In this case, we see that 95% bootstrap confidence intervals for β_0, β_1, and β_2 are very similar to the normal-theory confidence intervals we found earlier. For example, the normal-theory confidence interval for the effect of fast tracks is 0.93 to 1.53 ft/s, while the analogous bootstrap confidence interval is 0.89 to 1.55 ft/s.

There are many variations on this bootstrap procedure. For example, you could sample residuals rather than cases, or you could conduct a parametric bootstrap in which error terms are randomly chosen from a normal distribution. In addition, researchers have devised other ways of calculating confidence intervals besides the percentile method, including normality, studentized, and bias-corrected and accelerated (Hesterberg (2015); Efron and Tibshirani (1993); Davison and Hinkley (1997)). We will focus on case resampling and percentile confidence intervals for now for their understandability and wide applicability.

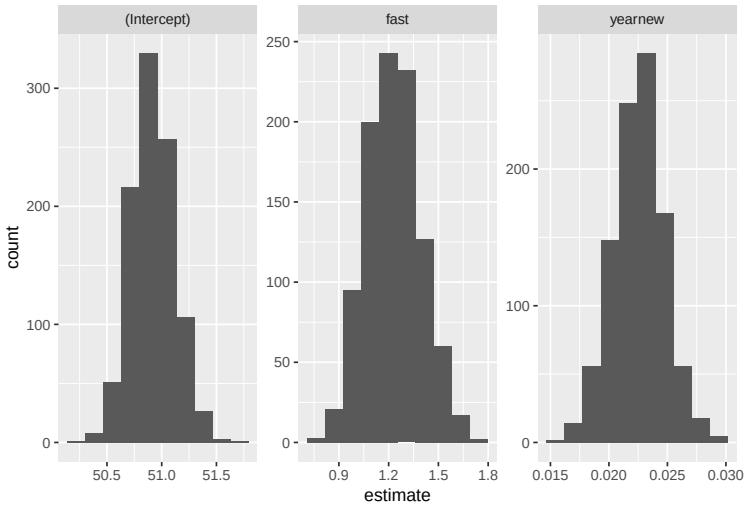

Figure 1.9: Bootstrapped distributions for Model 4 coefficients

1.6.6 Multiple linear regression with an interaction term

Adding terms to form a multiple linear regression model as we did in Model 4 is a very powerful modeling tool, allowing us to account for multiple sources of uncertainty and to obtain more precise estimates of effect sizes after accounting for the effect of important covariates. One limitation of Model 4, however, is that we must assume that the effect of track condition has been the same for 122 years, or conversely that the yearly improvements in winning speeds are identical for all track conditions. To expand our modeling capabilities to allow the effect of one predictor to change depending on levels of a second predictor, we need to consider **interaction terms**. Amazingly, if we create a new variable by taking the product of `yearnew` and `fast` (i.e., the **interaction** between `yearnew` and `fast`), adding that variable into our model will have the desired effect.

Thus, consider Model 5:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(\text{Yearnew}_i) + \beta_2(\text{Fast}_i) \\ &\quad + \beta_3(\text{Yearnew} \times \text{Fast}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \end{aligned}$$

where OLS provides the following parameter estimates:

```
lm(formula = speed ~ yearnew + fast + yearnew:fast, data = derby.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.528629   0.205072 246.394  < 2e-16 ***
yearnew       0.030751   0.003471   8.859 9.84e-15 ***
fast          1.833523   0.262175   6.994 1.73e-10 ***
yearnew:fast -0.011490   0.004117  -2.791  0.00613 **
---
Residual standard error: 0.7071 on 118 degrees of freedom
Multiple R-squared:  0.7068,    Adjusted R-squared:  0.6993
```

According to OLS estimates, estimated winning speeds can be found by:

$$\hat{Y}_i = 50.53 + 0.031(\text{Yearnew}_i) + 1.83(\text{Fast}_i) - 0.011(\text{Yearnew} \times \text{Fast}_i). \tag{1.8}$$

Interpretations of model coefficients are most easily seen by writing out separate equations for fast and non-fast track conditions:

$$\begin{aligned}
&\text{Fast} = 0: \\
&\qquad \hat{Y}_i = 50.53 + 0.031(\text{Yearnew}_i) \\
&\text{Fast} = 1: \\
&\qquad \hat{Y}_i = (50.53 + 1.83) + (0.031 - 0.011)(\text{Yearnew}_i)
\end{aligned}$$

leading to the following interpretations for estimated model coefficients:

- $\hat{\beta}_0 = 50.53$. The expected winning speed in 1896 under non-fast conditions was 50.53 ft/s.
- $\hat{\beta}_1 = 0.031$. The expected yearly increase in winning speeds under non-fast conditions is 0.031 ft/s.
- $\hat{\beta}_2 = 1.83$. The winning speed in 1896 was expected to be 1.83 ft/s faster under fast conditions compared to non-fast conditions.
- $\hat{\beta}_3 = -0.011$. The expected yearly increase in winning speeds under fast conditions is 0.020 ft/s, which is 0.011 ft/s less than the expected annual increase under non-fast conditions.

In fact, using interaction allows us to model the relationships we noticed in Figure 1.4, where both the intercept and slope describing the relationships between `speed` and `year` differ depending on whether track conditions were fast or not. Note that we interpret the coefficient for the interaction term by comparing slopes under fast and non-fast conditions; this produces a much more understandable interpretation for a reader than attempting to interpret the -0.011 directly.

1.6.7 Building a multiple linear regression model

We now begin iterating toward a "final model" for these data, on which we will base conclusions. Typical features of a "final multiple linear regression model" include:

- explanatory variables allow one to address primary research questions
- explanatory variables control for important covariates
- potential interactions have been investigated
- variables are centered where interpretations can be enhanced
- unnecessary terms have been removed
- LINE assumptions and the presence of influential points have both been checked using residual plots
- the model tells a "persuasive story parsimoniously"

Although the process of reporting and writing up research results often demands the selection of a sensible final model, it's important to realize that (a) statisticians typically will examine and consider an entire taxonomy of models when formulating conclusions, and (b) different statisticians sometimes select different models as their "final model" for the same set of data. Choice of a "final model" depends on many factors, such as primary research questions, purpose of modeling, tradeoff between parsimony and quality of fitted model, underlying assumptions, etc. Modeling decisions should never be automated or made completely on the basis of statistical tests; subject area knowledge should always play a role in the modeling process. You should be able to defend any final model you select, but you should not feel pressured to find the one and only "correct model", although most good models will lead to similar conclusions.

Several tests and measures of model performance can be used when comparing different models for model building:

- R^2 we have seen; it measures the variability in the response variable explained by the model. One problem is that R^2 always increases with extra predictors, even if the predictors add very little information.

- adjusted R^2. Adds a penalty for model complexity to R^2 so that any increase in performance must outweigh the cost of additional complexity. We should ideally favor any model with higher adjusted R^2, regardless of size, but the penalty for model complexity (additional terms) is fairly ad-hoc.
- AIC (Akaike Information Criterion). Again attempts to balance model performance with model complexity, with smaller AIC levels being preferable, regardless of model size. The BIC (Bayesian Information Criterion) is similar to the AIC, but with a greater penalty for additional model terms.
- extra sum of squares F test. This is a generalization of the t-test for individual model coefficients which can be used to perform significance tests on **nested models**, where one model is a reduced version of the other. For example, we could test whether our final model (below) really needs to adjust for track condition, which is comprised of indicators for both fast condition and good condition (leaving slow condition as the reference level). Our null hypothesis is then $\beta_3 = \beta_4 = 0$. We have statistically significant evidence (F = 57.2 on 2 and 116 df, p < .001) that track condition is associated with winning speeds, after accounting for quadratic time trends and number of starters.

One potential final model for predicting winning speeds of Kentucky Derby races is:

$$\begin{aligned} Y_i = \beta_0 + \beta_1(\text{Yearnew}_i) + \beta_2(\text{Yearnew}_i^2) + \beta_3(\text{Fast}_i) \\ + \beta_4(\text{Good}_i) + \beta_5(\text{Starters}_i) + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \text{N}(0, \sigma^2). \end{aligned} \tag{1.9}$$

and OLS provides the following parameter estimates:

```
lm(formula = speed ~ yearnew + yearnew2 + fast + good + starters,
    data = derby.df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.002e+01  1.946e-01 256.980  < 2e-16 ***
yearnew      7.003e-02  6.130e-03  11.424  < 2e-16 ***
yearnew2    -3.697e-04  4.598e-05  -8.041 8.44e-13 ***
fast         1.393e+00  1.305e-01  10.670  < 2e-16 ***
good         9.157e-01  2.077e-01   4.409 2.33e-05 ***
starters    -2.528e-02  1.360e-02  -1.859   0.0656 .
---
Residual standard error: 0.5483 on 116 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.8192

Analysis of Variance Table

Model 1: speed ~ yearnew + yearnew2 + starters
Model 2: speed ~ yearnew + yearnew2 + fast + good + starters
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    118 69.257
2    116 34.870  2    34.386 57.196 < 2.2e-16 ***
```

This model accounts for the slowing annual increases in winning speed with a negative quadratic term, adjusts for baseline differences stemming from track conditions, and suggests that, for a fixed year and track condition, a larger field is associated with slower winning times (unlike the positive relationship we saw between speed and number of starters in our exploratory analyses). The model explains 82.7% of the year-to-year variability in winning speeds, and residual plots show no serious issues with LINE assumptions. We tested interaction terms for different effects of time or number of starters based on track condition, but we found no significant evidence of interactions.

1.7 Preview

Having reviewed key ideas from multiple linear regression, you are now ready to extend those ideas, especially to handle non-normal responses and lack of independence. This section provides a preview of the type of problems you will encounter in the book. For each journal article cited, we provide an abstract in the authors' words, a description of the type of response and, when applicable, the structure of the data. Each of these examples appears later as an exercise, where you can play with the actual data or evaluate the analyses detailed in the articles.

1.7.1 Soccer

Roskes et al. (2011) . The right side? Under time pressure, approach motivation leads to right-oriented bias. *Psychological Science* [Online] **22(11)**:1403-7. DOI: 10.1177/0956797611418677, October 2011.

> **Abstract:** Approach motivation, a focus on achieving positive outcomes, is related to relative left-hemispheric brain activation, which translates to a variety of right-oriented behavioral biases. [...] In our analysis of all Federation Internationale de Football Association (FIFA) World Cup penalty shoot-outs, we found that goalkeepers were two times more likely to dive to the right than to the left when their team was behind, a situation that we conjecture induces approach motivation. Because penalty takers shot toward the two sides of the goal equally often, the goalkeepers' right-oriented bias was dysfunctional, allowing more goals to be scored.

The response for this analysis is the direction of the goalkeeper dive, a binary variable. For example, you could let Y=1 if the dive is to the right and Y=0 if the dive is to the left. This response is clearly not normally distributed. One approach to the analysis is logistic regression as described in Chapter **??**. A binomial random variable could also be created for this application by summing the binary variables for each game so that Y= the number of dives right out of the number of dives the goalkeeper makes during a game. [Thought question: Do you buy the last line of the abstract?]

1.7.2 Elephant Mating

Poole (1989) . Mate guarding, reproductive success and female choice in African elephants. *Animal Behavior* **37**:842-49.

> **Abstract:** Male guarding of females, male mating success and female choice were studied for 8 years among a population of African elephants, Loxodonta africana. Males were not able to compete successfully for access to oestrous female until approximately 25 years of age. Males between 25 and 35 years of age obtained matings during early and late oestrus, but rarely in mid-oestrus. Large musth males over 35 years old guarded females in mid-oestrus. Larger, older males ranked above younger, smaller males and the number of females guarded by males increased rapidly late in life. Body size and longevity are considered important factors in determining the lifetime reproductive success of male elephants...

Poole and her colleagues recorded, for each male elephant, his age (in years) and the number of matings for a given year. The researchers were interested in how age affects the males' mating patterns. Specifically, questions concern whether there is a steady increase in mating success as an elephant ages or if there is an optimal age after which the number of matings decline. Because the responses of interest are counts (number of matings for each elephant for a given year), we will consider a Poisson regression (see Chapter **??**). The general form for Poisson responses is the number of events for a specified time, volume, or space.

1.7.3 Parenting and Gang Activity

Walker-Barnes and Mason (2001) . Ethnic differences in the effect of parenting on gang involvement and gang delinquency: a longitudinal, hierarchical linear modeling perspective. *Child Development* **72(6)**:1814-31.

> **Abstract:** This study examined the relative influence of peer and parenting behavior on changes in adolescent gang involvement and gang-related delinquency. An ethnically diverse sample of 300 ninth-grade students was recruited and assessed on eight occasions during the school year. Analyses were conducted using hierarchical linear modeling. Results indicated that, in general, adolescents decreased their level of gang involvement over the course of the school year, whereas the average level of gang delinquency remained constant over time. As predicted, adolescent gang involvement and gang-related delinquency were most strongly predicted by peer gang involvement and peer gang delinquency, respectively. Nevertheless, parenting behavior continued to significantly predict change in both gang involvement and gang delinquency, even after controlling for peer behavior. A significant interaction between parenting and ethnic and cultural heritage found the effect of parenting to be particularly salient for Black students, for whom higher levels of behavioral control and lower levels of lax parental control were related to better behavioral outcomes over time, whereas higher levels of psychological control predicted worse behavioral outcomes.

The response for this study is a gang activity measure which ranges from 1 to 100. While it may be reasonable to assume this measure is approximately normal, the structure of this data implies that it is not a simple regression problem. Individual students have measurements made at 8 different points in time. We cannot assume that we have 2400 independent observations because the same measurements on one individual are more likely to be similar than a measurement of another student. Multilevel modeling as discussed in Chapter **??** can often be used in these situations.

1.7.4 Crime

Gelman, Fagan, and Kiss (2007) . An analysis of the NYPD's stop-and-frisk policy in the context of claims of racial bias. *Journal of the American Statistical Association* **102(479)**:813-823.

> **Abstract:** Recent studies by police departments and researchers confirm that police stop racial and ethnic minority citizens more often than whites, relative to their proportions in the population. However, it has been argued stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas such as neighborhoods or precincts. Most of the research on stop rates and police-citizen interactions has focused on traffic stops, and analyses of pedestrian stops are rare. In this paper, we analyze data from 175,000 pedestrian stops by the New York Police Department over a fifteen-month period. We disaggregate stops by police precinct, and compare stop rates by racial and ethnic group controlling for previous race-specific arrest rates. We use hierarchical multilevel models to adjust for precinct-level variability, thus directly addressing the question of geographic heterogeneity that arises in the analysis of pedestrian stops. We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.

This application involves both non-normal data (number of stops by ethnic group can be modeled as a Poisson response) and multilevel data (number of stops within precincts will likely be correlated due to characteristics of the precinct population). This type of analysis will be the last type you encounter, generalized linear multilevel modeling, as addressed in Chapter **??**.

1.8 Exercises

1.8.1 Conceptual Exercises

1. **Applications that do not violate the OLS assumptions for inference**. Identify the response and explanatory variable(s) for each problem. Write the OLS assumptions for inference in the context of each study.
    a. **Cricket Chirps.** Researchers record the number of cricket chirps per minute and temperature during that time to investigate whether the number of chirps varies with the temperature.
    b. **Women's Heights.** A random selection of women aged 20-24 years are selected and their shoe size is used to predict their height

2. **Applications that do violate the OLS assumptions for inference**. All of the examples in this section have at least one violation of the OLS assumptions for inference. Begin by identifying the response and explanatory variables. Then, identify which OLS assumption(s) are violated.
    a. **Low Birthweights.** Researchers are attempting to see if socioeconomic status and parental stability are predictive of low birthweight. They classify a child as having a low birthweight if their birthweight is less than 2,500 grams.
    b. **Clinical Trial I.** A Phase II clinical trial is designed to compare the number of patients getting relief at different dose levels. 100 patients get dose A, 100 get dose B, and 100 get dose C.
    c. **Canoes and zip codes.** For each of over 27,000 overnight permits for the Boundary Water Canoe area, the zip code for the group leader has been translated to the distance traveled and socioeconomic data. This data is used to create a model for the number of trips made per zip code.
    d. **Clinical Trial II.** A randomized clinical trial investigated postnatal depression and the use of an estrogen patch. Patients were randomly assigned to either use the patch or not. Depression scores were recorded on 6 different visits.
    e. **Elephant mating.** Researchers are interested in how elephant age affects mating patterns among males. In particular, do older elephants have greater mating success, and is there an optimal age for mating among males? Data collected includes, for each elephant, age and number of matings in a given year.

3. **Kentucky Derby.** The next set of questions is related to the Kentucky Derby case study from this chapter.
    a. Discuss the pros and cons of using side-by-side boxplots vs. stacked histograms to illustrate the relationships between year and track condition in Figure 1.3.
    b. Why is a scatterplot more informative than a correlation coefficient to describe the relationship between speed of the winning horse and year in Figure 1.3.
    c. How might you incorporate a fourth variable, say number of starters, into Figure 1.4?
    d. Explain why ϵ_i in Equation (1.1) measures the vertical distance from a data point to the regression line.
    e. In the first t-test in Section 1.6.1 (t = 11.251 for $H_0 : \beta_1 = 0$), notice that $t = \frac{\hat{\beta}_1}{SE(\beta_1)} = \frac{.026}{.0023} = 11.251$. Why is the t-test based on the ratio of the estimated slope to its standard error?
    f. In Equation (1.4), explain why the t-test corresponding to β_1 is equivalent to an independent-samples t-test under equal variances. Why is the equal variance assumption needed?
    g. When interpreting β_2 in Equation (1.7), why do we have to be careful to say *for a fixed year* or *after adjusting for year*? Is it wrong to leave a qualifier like that off?
    h. Interpret in context a 95% confidence interval for β_0 in Model 4.
    i. State (in context) the result of a t-test for β_1 in Model 4.
    j. Why is there no ϵ_i term in Equation (1.8)?
    k. If you considered the interaction between two continuous variables (like `yearnew` and `starters`), how would you provide an interpretation for that coefficient in context?

l. Interpret (in context) the OLS estimates for β_3 and β_5 in Equation (1.9).

4. **Moneyball.** In a 2011 article in *The Sport Journal*, Farrar and Bruggink attempt to show that Major League Baseball general managers did not immediately embrace the findings of Michael Lewis's 2003 *Moneyball* book (Lewis 2003). They contend that players' on-base percentage remained relatively undercompensated compared to slugging percentage three years after the book came out. Two regression models are described: a Team Run Production Model and a Player Salary Model. (Farrar and Bruggink 2011)

   a. Discuss potential concerns (if any) with the LINE assumptions for linear regression in each model.
   b. In Table 3, the authors contend that Model 1 is better than Model 3. Could you argue that Model 3 is actually better? How could you run a formal hypothesis test comparing Model 1 to Model 3?
   c. If authors had chosen Model 3 in Table 3 with the two interaction terms, how would that affect their final analysis, in which they compare coefficients of slugging and on-base percentage? (Hint: write out interpretations for the two interaction coefficients—the first one should be NL:OBP and the second one should be NL:SLG)
   d. The authors write that "It should also be noted that the runs scored equation fit is better than the one Hakes and Sauer have for their winning equation." What do you think they mean by this statement? Why might this comparison not be relevant?
   e. In Table 4, Model 1 has a higher adjusted R^2 than Model 2, yet the extra term in Model 1 (an indicator value for the National League) is not significant at the 5% level. Explain how this is possible.
   f. What limits does this paper have on providing guidance to baseball decision makers?

1.8.2 Guided Exercises

1. **Gender discrimination in bank salaries**. In the 1970's, Harris Trust was sued for gender discrimination in the salaries it paid its employees. One approach to addressing this issue was to examine the starting salaries of all skilled, entry-level clerical workers between 1965 and 1975. The following variables, which can be found in `banksalary.csv`, were collected for each worker (Ramsey and Schafer 2002):
   - `bsal` = beginning salary (annual salary at time of hire)
   - `sal77` = annual salary in 1977
   - `sex` = MALE or FEMALE
   - `senior` = months since hired
   - `age` = age in months
   - `educ` = years of education
   - `exper` = months of prior work experience

   Creating an indicator variable based on `sex` could be helpful.

   a. Identify observational units, the response variable, and explanatory variables.
   b. The mean starting salary of male workers ($5957) was 16% higher than the mean starting salary of female workers ($5139). Confirm these mean salaries. Is this enough evidence to conclude gender discrimination exists? If not, what further evidence would you need?
   c. How would you expect age, experience, and education to be related to starting salary? Generate appropriate exploratory plots; are the relationships as you expected? What implications does this have for modeling?
   d. Why might it be important to control for seniority (number of years with the bank) if we are only concerned with the salary when the worker started?
   e. By referring to exploratory plots and summary statistics, are any explanatory variables (including sex) closely related to each other? What implications does this have for modeling?
   f. Fit a simple linear regression model with starting salary as the response and experience as the sole explanatory variable (Model 1). Interpret the intercept and slope of this model; also inter-

pret the R-squared value. Is there a significant relationship between experience and starting salary?

g. Does Model 1 meet all linear regression assumptions? List each assumption and how you decided if it was met or not.
h. Is a model with all 4 confounding variables (Model 2, with `senior`, `educ`, `exper`, and `age`) better than a model with just experience (Model 1)? Justify with an appropriate significance test in addition to summary statistics of model performance.
i. You should have noticed that the term for age was not significant in Model 2. What does this imply about age and about future modeling steps?
j. Generate an appropriate coded scatterplot to examine a potential age-by-experience interaction. How would you describe the nature of this interaction?
k. A potential final model (Model 3) would contain terms for seniority, education, and experience in addition to sex. Does this model meet all regression assumptions? State a 95% confidence interval for sex and interpret this interval carefully in the context of the problem.
l. Based on Model 3, what conclusions can be drawn about gender discrimination at Harris Trust? Do these conclusions have to be qualified at all, or are they pretty clear cut?
m. Often salary data is logged before analysis. Would you recommend logging starting salary in this study? Support your decision analytically.
n. Regardless of your answer to the previous question, provide an interpretation for the coefficient for the male coefficient in a modified Model 3 after logging starting salary.
o. Build your own final model for this study and justify the selection of your final model. You might consider interactions with gender, since those terms could show that discrimination is stronger among certain workers. Based on your final model, do you find evidence of gender discrimination at Harris Trust?

2. **Sitting and MTL Thickness.** Siddarth et al. (2018) researched relations between time spent sitting (sedentary behavior) and the thickness of participant's medial temporal lobe (MTL) in their paper, Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults. MTL volume is negatively associated with Alzheimer's disease and memory impairment. Their data on 35 adults can be found in `sitting.csv`. Key variables include:
   - `MTL` = Medial temporal lobe thickness in mm
   - `sitting` = Reported hours/day spent sitting
   - `MET` = Reported metabolic equivalent unit minutes per week
   - `age` = Age in years
   - `sex` = Sex (`M` = Male, `F` = Female)
   - `education` = Years of education completed

   a. In their article's introduction, Siddarth et al. (2018) differentiate their analysis on sedentary behavior from analysis on active behavior by citing evidence supporting the claim that "one can be highly active yet still be sedentary for most of the day." Fit your own linear model with `MET` and `sitting` as your explanatory and response variables, respectively. Using R^2, how much of the subject to subject variability in hours/day spent sitting can be explained by MET minutes per week? Does this support the claim that sedentary behaviors may be independent from physical activity?
   b. In the paper's section, "Statistical analysis", the authors report that "Due to the skewed distribution of physical activity levels, we used log-transformed values in all analyses using continuous physical activity measures." Generate both a histogram of `MET` values and log–transformed `MET` values. Do you agree with the paper's decision to use a log-transformation here?
   c. Fit a preliminary model with `MTL` as the response and `sitting` as the sole explanatory variable. Are OLS conditions satisfied?
   d. Expand on your previous model by including a centered version of `age` as a covariate. Interpret all three coefficients in this model.
   e. One model fit in Siddarth et al. (2018) includes `sitting`, log–transformed `MET`, and `age` as explanatory variables. They report an estimate $\widehat{\beta_1} = -0.02$ with confidence interval $(-0.04, -0.002)$ for the coefficient corresponding to `sitting`, and $\widehat{\beta_2} = 0.007$ with confidence interval $(-0.07, 0.08)$

for the coefficient corresponding to `MET`. Verify these intervals and estimates on your own.

f. Based on your confidence intervals from the previous part, do you support the paper's claim that "it is possible that sedentary behavior is a more significant predictor of brain structure, specifically MTL thickness [than physical activity]"? Why or why not?
g. A New York Times Article was published discussing Siddarth et al. (2018) with the title "Standing Up at Your Desk Could Make You Smarter" (Friedman 2018). Do you agree with this headline choice? Why or why not?

3. **Housing Prices and log Transformations.** The dataset `kingCountyHouses.csv` contains data on over 20,000 houses sold in King County, Washington ("House Sales in King County, Usa" 2018). The dataset includes the following variables:
   - `price` = selling price of the house
   - `date` = date house was sold, measured in days since January 1, 2014
   - `bedrooms` = number of bedrooms
   - `bathrooms` = number of bathrooms
   - `sqft` = interior square footage
   - `floors` = number of floors
   - `waterfront` = 1 if the house has a view of the waterfront, 0 otherwise
   - `yr_built` = year the house was built
   - `yr_renovated` = 0 if the house was never renovated, the year the house was renovated if else

   We wish to create a linear model to predict a house's selling price.
   a. Generate appropriate graphs and summary statistics detailing both `price` and `sqft` individually and then together. What do you notice?
   b. Fit a simple linear regression model with `price` as the response variable and `sqft` as the explanatory variable (Model 1). Interpret the slope coefficient β_1. Are all conditions met for linear regression?
   c. Create a new variable, `logprice`, the natural log of `price`. Fit Model 2, where `logprice` is now the response variable and `sqft` is still the explanatory variable. Write out the regression line equation.
   d. How does `logprice` change when `sqft` increases by 1.
   e. Recall that $\log(a) - \log(b) = \log\left(\frac{a}{b}\right)$, and use this to derive how `price` changes as `sqft` increases by 1.
   f. Are OLS assumptions satisfied in Model 2? Why or why not?
   g. Create a new variable, `logsqft`, the natural log of `sqft`. Fit Model 3 where `price` and `logsqft` are the explanatory and response variables, respectively. Write out the regression line equation.
   h. How does predicted `price` change as `logsqft` increases by 1 in Model 3?
   i. How does predicted `price` change as `sqft` increases by 1%? As a hint, this is the same as multiplying `sqft` by 1.01.
   j. Are OLS assumptions satisfied in Model 3? Why or why not?
   k. Fit Model 4, with `logsqft` and `logprice` as the response and explanatory variables, respectively. Write out the regression line equation.
   l. In Model 4, what is the effect on `price` corresponding to a 1% increase in `sqft`?
   m. Are OLS assumptions satisfied in Model 4? Why or why not?
   n. Find another explanatory variable which can be added to Model 4 to create a model with a higher adjusted R^2 value. Interpret the coefficient of this added variable.

1.8.3 Open-ended Exercises

1. **The Bechdel Test.** In April, 2014, website FiveThirtyEight published the article The Dollar-And-Cents Case Against Hollywood's Exclusion of Women (Hickey 2014). There, they analyze returns on investment for 1,615 films released between 1990 and 2013 based on the Bechdel test. The test, developed by cartoonist Alison Bechdel, measures gender bias in films by checking if a film meets three criteria:

- there are at least two named women in the picture
- they have a conversation with each other at some point
- that conversation isn't about a male character

While the test is not a perfect metric of gender bias, data from it does allow for statistical analysis. In the FiveThirtyEight article, they find that "passing the Bechdel test had no effect on the film's return on investment." Their data can be found in `bechdel.csv`. Key variables include:

- `year` = the year the film premiered
- `pass` = 1 if the film passes the Bechdel test, 0 otherwise
- `budget` = budget in 2013 U.S. dollars
- `totalGross` = total gross earnings in 2013 U.S. dollars
- `domGross` = domestic gross earnings in 2013 U.S. dollars
- `intGross` = international gross earnings in 2013 U.S. dollars
- `totalROI` = total return on investment (total gross divided by budget)
- `domROI` = domestic return on investment
- `intROI` = international return on investment

With this in mind, carry out your own analysis. Does passing the Bechdel test have any effect on a film's return on investment?

2. **Waitress Tips.** A student collected data from a restaurant where she was a waitress (Dahlquist and Dong 2011). The student was interested in learning under what conditions a waitress can expect the largest tips—for example: At dinner time or late at night? From younger or older patrons? From patrons receiving free meals? From patrons drinking alcohol? From patrons tipping with cash or credit? And should tip amount be measured as total dollar amount or as a percentage? Data can be found in `TipData.csv`. Here is a quick description of the variables collected:

   - `Day` = day of the week
   - `Meal` = time of day (Lunch, Dinner, Late Night)
   - `Payment` = how bill was paid (Credit, Cash, Credit with Cash tip)
   - `Party` = number of people in the party
   - `Age` = age category of person paying the bill (Yadult, Middle, SenCit)
   - `GiftCard` = was gift card used?
   - `Comps` = was part of the meal complimentary?
   - `Alcohol` = was alcohol purchased?
   - `Bday` = was a free birthday meal or treat given?
   - `Bill` = total size of the bill
   - `W.tip` = total amount paid (bill plus tip)
   - `Tip` = amount of the tip
   - `Tip.Percentage` = proportion of the bill represented by the tip

Cannon, Ann, George Cobb, Brad Hartlaub, Julie Legler, Robin Lock, Tom Moore, Allan Rossman, and Jeff Witmer. 2019. *Stat2: Modeling with Regression and Anova.* Macmillan.

Dahlquist, Samantha, and Jin Dong. 2011. "The Effects of Credit Cards on Tipping." Project for Statistics 212-Statistics for the Sciences, St. Olaf College.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application.* Cambridge University Press.

Efron, Bradley, and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

Farrar, Anthony, and Thomas H. Bruggink. 2011. "A New Test of the Moneyball Hypothesis." *The Sport Journal*, May. http://thesportjournal.org/article/a-new-test-of-the-moneyball-hypothesis/.

Friedman, Richard A. 2018. "Standing up at Your Desk Could Make You Smarter." *The New York Times*, April.

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. 2007. "An Analysis of the Nypd's Stop-and-Frisk Policy in

the Context of Claims of Racial Bias" 102 (September): 813–23.

Hesterberg, Tim C. 2015. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." *The American Statistician* 69 (4). Taylor & Francis: 371–86. doi:10.1080/00031305.2015.1089789.

Hickey, Walt. 2014. "The Dollar-and-Cents Case Against Hollywood's Exclusion of Women." FiveThirtyEight. https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.

"House Sales in King County, Usa." 2018. Accessed June 29. https://www.kaggle.com/harlfoxem/housesalesprediction/home.

Lewis, Michael M. 2003. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton & Company.

Poole, Joyce H. 1989. "Mate Guarding, Reproductive Success and Female Choice in African Elephants." *Animal Behaviour* 37: 842–49. doi:https://doi.org/10.1016/0003-3472(89)90068-7.

Ramsey, Fred, and Daniel Schafer. 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Second Edition. Boston, Massachusetts: Brooks/ Cole Cengage Learning.

Roskes, Marieke, Daniel Sligte, Shaul Shalvi, and Carsten K. W. De Dreu. 2011. "The Right Side? Under Time Pressure, Approach Motivation Leads to Right-Oriented Bias." *Psychology Science* 22 (11): 1403–7. doi:10.1177/0956797611418677.

Siddarth, Prabha, Alison C. Burggren, Harris A. Eyre, Gary W. Small, and David A. Merrill. 2018. "Sedentary Behavior Associated with Reduced Medial Temporal Lobe Thickness in Middle-Aged and Older Adults." *PLOS ONE* 13 (4). Public Library of Science: 1–13. doi:10.1371/journal.pone.0195549.

Stigler, Stephen M. 2002. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press.

Walker-Barnes, Chanequa J., and Craig A. Mason. 2001. "Ethnic Differences in the Effect of Parenting on Gang Involvement and Gang Delinquency: A Longitudinal, Hierarchical Linear Modeling Perspective." *Child Development* 72 (6): 1814–31. doi:10.1111/1467-8624.00380.

Wikipedia contributors. 2018. "Kentucky Derby — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Kentucky_Derby&oldid=846316018.