Unit 1: Linear models: Estimation

Eugene Katsevich

September 13, 2021

1 Types of predictors; interpreting linear model coefficients (Agresti 1.2)

The types of predictors x_j (e.g. binary or continuous) has less of an effect on the regression than the type of response, but it is still important to pay attention to the former.

Intercepts. It is common to include an *intercept* in a linear regression model, a predictor x_0 such that $x_{i0} = 1$ for all i. When an intercept is present, we index it as the 0th predictor. The simplest kind of linear model is the *intercept-only model* or the *one-sample model*:

$$y = \beta_0 + \epsilon. \tag{1}$$

The parameter β_0 is the mean of the response.

Binary predictors. In addition to an intercept, suppose we have a binary predictor $x_1 \in \{0, 1\}$ (e.g. $x_1 = 1$ for patients who took blood pressure medication and $x_1 = 0$ for those who didn't). This leads to the following linear model:

$$y = \beta_0 + \beta_1 x_1 + \epsilon. \tag{2}$$

Here, β_0 is the mean response (say blood pressure) for observations with $x_1 = 0$ and $\beta_0 + \beta_1$ is the mean response for observations with $x_1 = 1$. Therefore, the parameter β_1 is the difference in mean response between observations with $x_1 = 1$ and $x_1 = 0$. This parameter is sometimes called the *effect* or *effect size* of x_1, though a causal relationship might or might not be present. The model (2) is sometimes called the *two-sample model*, because the response data can be split into two "samples": those corresponding to $x_1 = 0$ and those corresponding to $x_1 = 1$.

Categorical predictors. A binary predictor is a special case of a categorical predictor: A predictor taking two or more discrete values. Suppose we have a predictor $w \in \{w_0, w_1, \dots, w_{C-1}\}$, where $C \geq 2$ is the number of categories and $w_0, \dots, w_{C-1}$ are the *levels* of w. E.g. suppose $\{w_0, \dots, w_{C-1}\}$ is the collection of U.S. states, so that $C = 50$. If we want to regress a response on the categorical predictor w, we cannot simply set $x_1 = w$ in the context of the linear regression (2). Indeed, w does not necessarily take numerical values. Instead, we need to add a predictor x_j for each of the levels of w. In particular, define $x_j \equiv \mathbb{1}(w = w_j)$ for $j = 1, \dots, C-1$ and consider the regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{C-1} x_{C-1} + \epsilon. \tag{3}$$

Here, category 0 is the *base category*, and β_0 represents the mean response in the base category. The coefficient β_j represents the difference in mean response between the jth category and the base category.

Quantitative predictors. A quantitative predictor is one that can take on any real value. For example, suppose that $x_1 \in \mathbb{R}$, and consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \epsilon. \tag{4}$$

Now, the interpretation of β_1 is that an increase in x_1 by 1 is associated with an increase in y by β_1. We must be careful to avoid saying "an increase in x_1 by 1 *causes* y to increase by β_1" unless we make additional causal assumptions. Note that the units of x_1 matter. If x_1 is the height of a person, then the value and the interpretation of β_1 changes depending on whether that height is measured in feet or in meters.

Ordinal predictors. There is an awkward category of predictor in between categorical and continuous called *ordinal*. An ordinal predictor is one that takes a discrete number of values, but these values have an intrinsic ordering, e.g. $x_1 \in \{\texttt{small}, \texttt{medium}, \texttt{large}\}$. It can be treated as categorical at the cost of losing the ordering information, or as continuous if one is willing to assign quantatitive values to each category.

Multiple predictors. A linear regression need not contain just one predictor (aside from an intercept). For example, let's say x_1 and x_2 are two predictors. Then, a linear model with both predictors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon. \tag{5}$$

When there are multiple predictors, the interpretation of coefficients must be revised somewhat. For example, β_1 in the above regression is the effect of an increase in x_1 by 1 *while holding* x_2 *constant* or *while adjusting for* x_2 or *while controlling for* x_2. If y is blood pressure, x_1 is a binary predictor indicating blood pressure medication taken and x_2 is sex, then β_1 is the effect of the medication on blood pressure while controlling for sex. In general, the coefficient of a predictor depends on what other predictors are in the model. As an extreme case, suppose the medication has no actual effect, but that men generally have higher blood pressure and higher rates of taking the medication. Then, the coefficient β_1 in the single regression model (2) would be nonzero but the coefficient in the multiple regression model (5) would be equal to zero. In this case, sex acts as a *confounder*.

Interactions. Note that the multiple regression model (5) has the built-in assumption that the effect of x_1 on y is the same for any fixed value of x_2 (and vice versa). In some cases, the effect of one variable on the response may depend on the value of another variable. In this case, it's appropriate to add another predictor called an *interaction*. Suppose x_2 is quantitative (e.g. years of job experience) and x_2 is binary (e.g. sex, with $x_2 = 1$ meaning male). Then, we can define a third predictor x_3 as the product of the first two, i.e. $x_3 = x_1 x_2$. This gives the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \tag{6}$$

Now, the effect of adding another year of job experience is β_1 for females and $\beta_1 + \beta_3$ for males. The coefficient β_3 is the difference in the effect of job experience between males and females.

2 Model matrices, model vectors spaces, and identifiability (Agresti 1.3-1.4)

The matrix $\boldsymbol{X}$ is called the *model matrix* or the *design matrix*. Concatenating the linear model equations across observations give us an equivalent formulation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0},\ \mathrm{Var}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n$$

or

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X\beta} = \boldsymbol{\eta}.$$

As $\boldsymbol{\beta}$ varies in $\mathbb{R}^p$, the set of possible vectors $\boldsymbol{\eta} \in \mathbb{R}^n$ is defined

$$C(\boldsymbol{X}) \equiv \{\boldsymbol{\eta} = \boldsymbol{X\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

$C(\boldsymbol{X})$, called the *model vector space*, is a subspace of $\mathbb{R}^n$: $C(\boldsymbol{X}) \subseteq \mathbb{R}^n$. Since

$$\boldsymbol{X\beta} = \beta_1 \boldsymbol{x}_{*1} + \cdots + \beta_p \boldsymbol{x}_{*p},$$

the model vector space is the column space of the matrix $\boldsymbol{X}$.

The *dimension* of $C(\boldsymbol{X})$ is the rank of $\boldsymbol{X}$, i.e. the number of linearly independent columns of $\boldsymbol{X}$. If $\text{rank}(\boldsymbol{X}) < p$, this means that there are two different vectors $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ such that $\boldsymbol{X\beta} = \boldsymbol{X\beta}'$. Therefore, we have two values of the parameter vector that give the same model for $\boldsymbol{y}$. This makes $\boldsymbol{\beta}$ *not identifiable*, and makes it impossible to reliably determine $\boldsymbol{\beta}$ based on the data. For this reason, we will generally assume that $\boldsymbol{\beta}$ is *identifiable*, i.e. $\boldsymbol{X\beta} \neq \boldsymbol{X\beta}'$ if $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$. This is equivalent to the assumption that $\text{rank}(\boldsymbol{X}) = p$. Note that this cannot hold when $p > n$, so for the majority of the course we will assume that $p \leq n$. In this case, $\text{rank}(\boldsymbol{X}) = p$ if and only if $\boldsymbol{X}$ has *full-rank*.

As an example when $p \leq n$ but when $\boldsymbol{\beta}$ is still not identifiable, consider the case of a categorical predictor. Suppose the categories of w were $\{w_1, \dots, w_{C-1}\}$, i.e. the baseline category w_0 did not exist. In this case, the model (3) would not be identifiable because $x_0 = 1 = x_1 + \cdots + x_{C-1}$ and thus $x_{*0} = 1 = x_{*1} + \cdots + x_{*,C-1}$. Indeed, this means that one of the predictors can be expressed as a linear combination of the others, so $\boldsymbol{X}$ cannot have full rank. A simpler way of phrasing the problem is that we are describing $C-1$ intrinsic parameters (the means in each of the $C-1$ groups) with C model parameters. There must therefore be some redundancy. For this reason, if we include an intercept term in the model then we must designate one of our categories as the baseline and exclude its indicator from the model.

3 Least squares estimation (Agresti 2.1.1, 2.7.1)

Now, suppose that we are given a dataset $(\boldsymbol{X}, \boldsymbol{y})$. How do we go about estimating $\boldsymbol{\beta}$ based on this data? The canonical approach is the *method of least squares*:

$$\widehat{\boldsymbol{\beta}} \equiv \underset{\boldsymbol{\beta}}{\arg\min}\ \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2. \tag{7}$$

The quantity

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{y} - \widehat{\boldsymbol{\mu}}\|^2 = \sum_{i=1}^n (y_i - \widehat{\mu}_i)^2 \tag{8}$$

is called the *residual sum of squares (RSS)*, and it measures the lack of fit of the linear regression model. We therefore want to choose $\widehat{\boldsymbol{\beta}}$ to minimize this lack of fit. Note that if $\boldsymbol{\epsilon}$ is assumed to be $N(0, \sigma^2 \boldsymbol{I}_n)$, then the least squares solution would also be the maximum likelihood solution. Indeed, for $y_i \sim N(\mu_i, \sigma^2)$, the log-likelihood is

$$\log\left[\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)\right] = \text{constant} - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu_i)^2.$$

Letting $L(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$, we can do some calculus to derive that

$$\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}) = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{9}$$

Setting this vector of partial derivatives equal to zero, we arrive at the *normal equations*:

$$-\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0 \quad \Longleftrightarrow \quad \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}. \tag{10}$$

If $\boldsymbol{X}$ is full rank, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible and we can therefore conclude that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{11}$$

Now that we have derived the least squares estimator, we can compute its bias and variance. To obtain the bias, we first calculate that

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}[\boldsymbol{y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Therefore, the least squares estimator is unbiased. To obtain the variance, we compute

$$\begin{aligned}
\text{Var}[\widehat{\boldsymbol{\beta}}] &= \text{Var}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Var}[\boldsymbol{y}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.
\end{aligned} \tag{12}$$

According to the Gauss-Markov theorm, this covariance matrix computed above is the smallest (in the sense of positive semidefinite matrices) among all linear unbiased estimates of $\boldsymbol{\beta}$.

4 Least squares estimation R demo (Agresti 2.6)

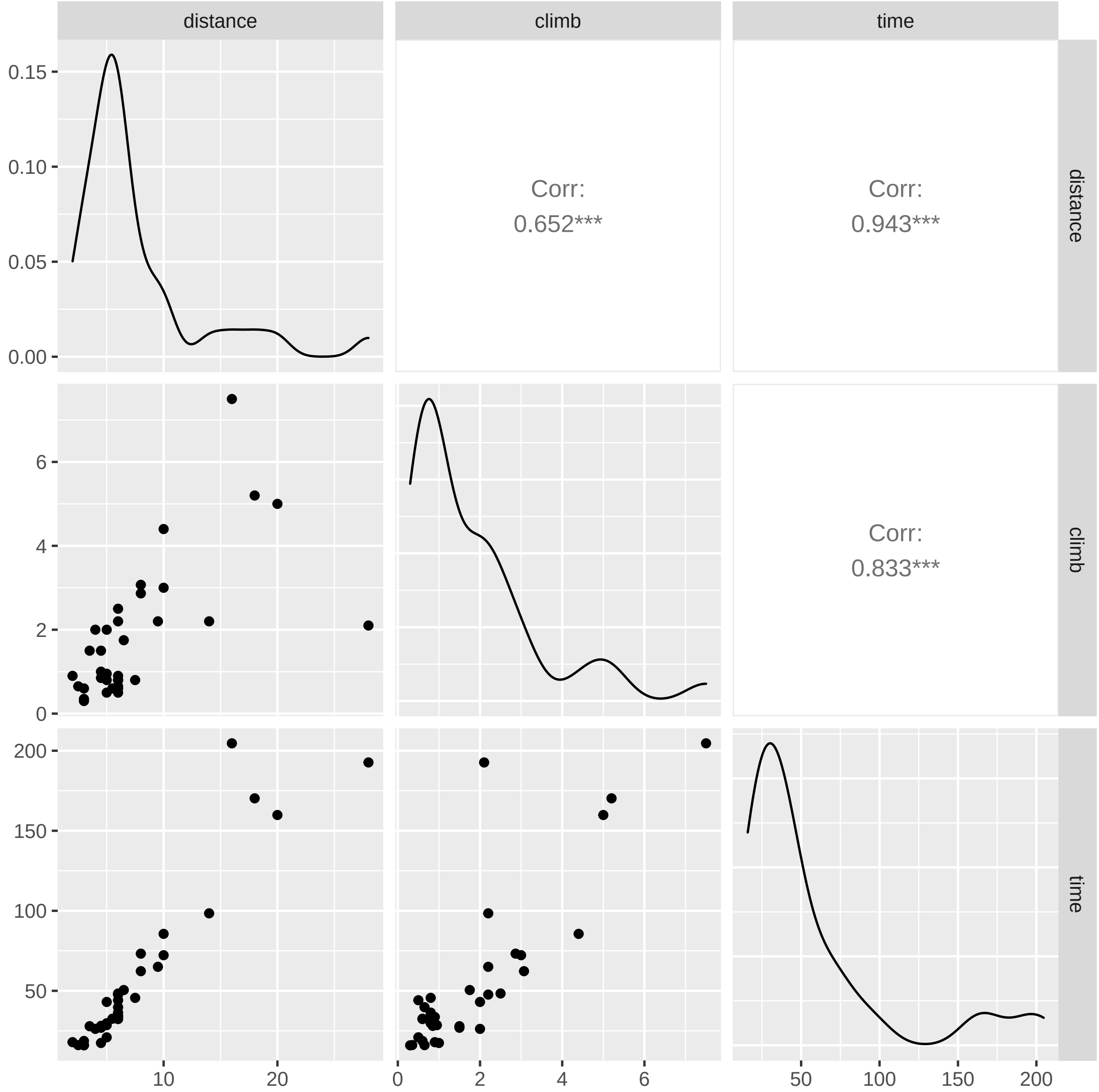
The R demo will be based on the `ScotsRaces` data from the textbook. Data description (quoted from the textbook):

> “Each year the Scottish Hill Runners Association publishes a list of hill races in Scotland for the year. The table below shows data on the record time for some of the races (in minutes). Explanatory variables listed are the distance of the race (in miles) and the cumulative climb (in thousands of feet).”

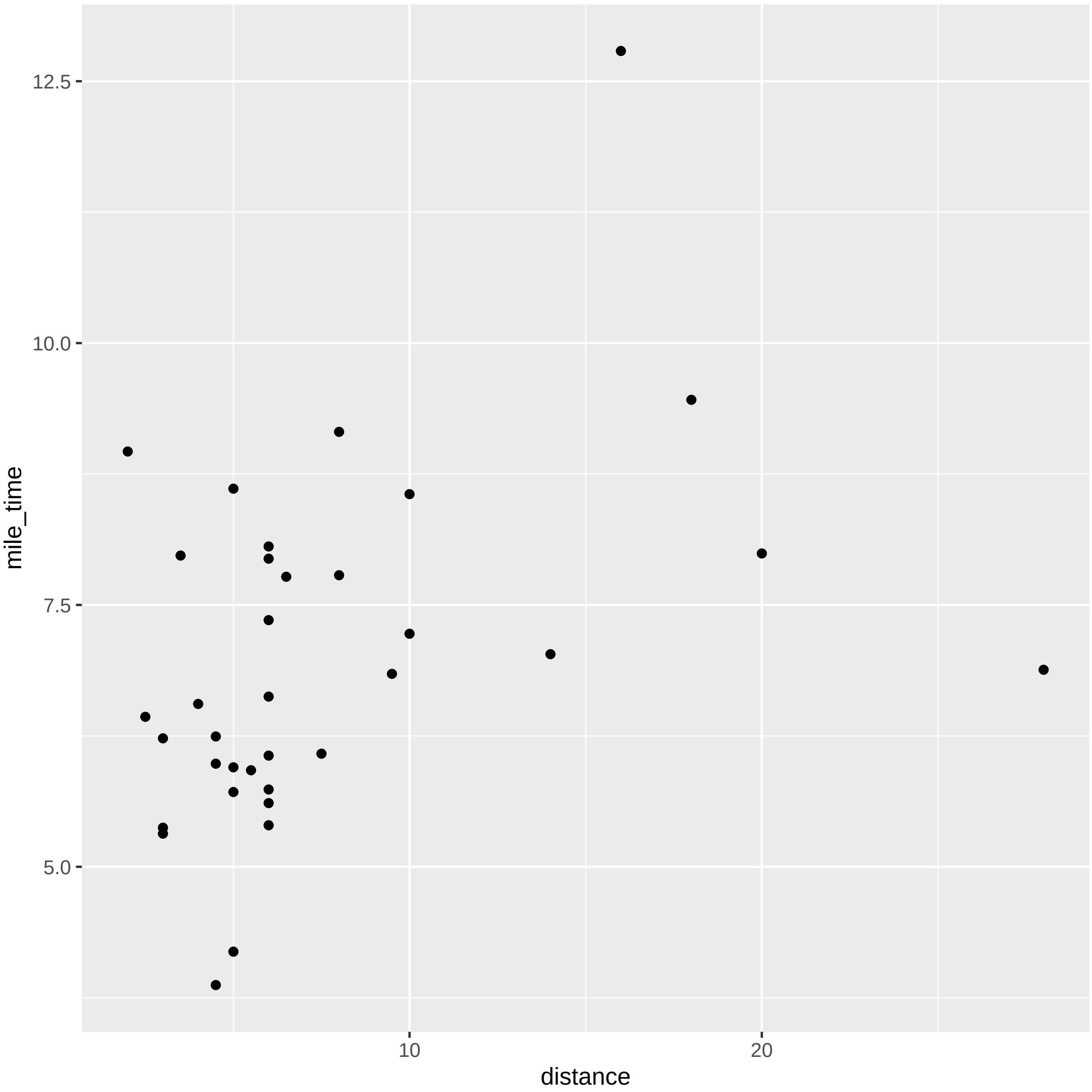
```r
library(tidyverse)
```

```r
# read the data into R
scots_races = read_tsv("../data/ScotsRaces.dat", col_types = "cddd")
scots_races

## # A tibble: 35 x 4
##    race                     distance climb  time
##    <chr>                       <dbl> <dbl> <dbl>
##  1 GreenmantleNewYearDash        2.5  0.65  16.1
##  2 Carnethy5HillRace             6    2.5   48.4
##  3 CraigDunainHillRace           6    0.9   33.6
##  4 BenRhaHillRace                7.5  0.8   45.6
##  5 BenLomondHillRace             8    3.07  62.3
##  6 GoatfellHillRace              8    2.87  73.2
##  7 BensofJuraFellRace         16    7.5  205.
##  8 CairnpappleHillRace         6    0.8   36.4
##  9 ScoltyHillRace              5    0.8   29.8
## 10 TraprainLawRace             6    0.65  39.8
## # ... with 25 more rows
```

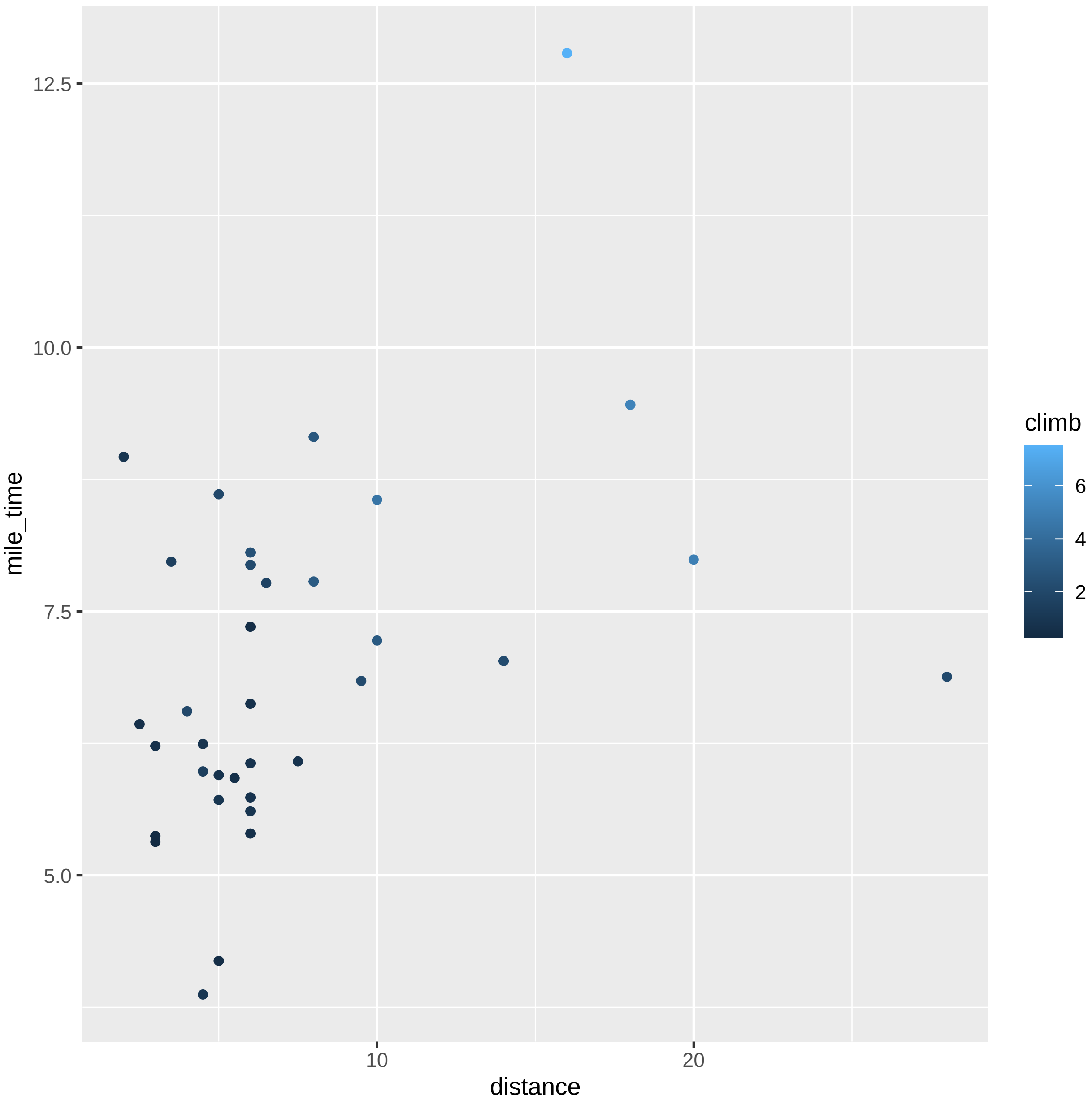
```
# Exploration

# pairs plot
GGally::ggpairs(scots_races %>% select(-race))

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```



```
# Q: What are the typical ranges of the variables?
# Q: What are the relationships among the variables?


# mile time versus distance

scots_races %>%
  mutate(mile_time = time/distance) %>%
  ggplot(aes(x = distance, y = mile_time)) +
  geom_point()
```



```
scots_races %>%
  mutate(mile_time = time/distance) %>%
  ggplot(aes(x = distance, y = mile_time, label = race,
             color = climb)) +
  geom_point()
```



```
# Q: How does mile time vary with distance?
# Q: What races deviate from this trend?
# Q: How does climb play into it?
```

```
# Linear model

# Q: What is the effect of an extra mile of distance on time?
lm_fit = lm(time ~ distance + climb, data = scots_races)
summary(lm_fit)

##
## Call:
```

```
## lm(formula = time ~ distance + climb, data = scots_races)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.654  -4.842   1.110   4.667  27.762
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.1086     2.5608  -5.119 1.41e-05 ***
## distance      6.3510     0.3578  17.751  < 2e-16 ***
## climb        11.7801     1.2206   9.651 5.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.734 on 32 degrees of freedom
## Multiple R-squared:  0.9717,Adjusted R-squared:   0.97
## F-statistic: 549.9 on 2 and 32 DF,  p-value: < 2.2e-16
```

```
# Linear model with interaction

# Q: What is the effect of an extra mile of distance on time
#  for a run with low climb?

# Q: What is the effect of an extra mile of distance on time
#  for a run with high climb?

lm_fit_int = lm(time ~ distance*climb, data = scots_races)
summary(lm_fit_int)

##
## Call:
## lm(formula = time ~ distance * climb, data = scots_races)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.197  -2.797   0.628   2.243  18.963
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.7672     3.9058  -0.196  0.84556
## distance         4.9623     0.4742  10.464 1.07e-11 ***
## climb            3.7133     2.3647   1.570  0.12650
## distance:climb   0.6598     0.1743   3.786  0.00066 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.338 on 31 degrees of freedom
## Multiple R-squared:  0.9807,Adjusted R-squared:  0.9788
## F-statistic: 524.1 on 3 and 31 DF,  p-value: < 2.2e-16
```

5 Linear regression as orthogonal projection (Agresti 2.2, 2.3, 2.4.2, 2.4.3, 2.4.4)

Let's think about the mapping $\boldsymbol{y} \mapsto \widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \in C(\boldsymbol{X})$. We claim that this mapping is an *orthogonal projection*. Geometrically it makes sense, since we define $\widehat{\boldsymbol{\beta}}$ so that $\widehat{\boldsymbol{\mu}} \in C(\boldsymbol{X})$ is as close to $\boldsymbol{y}$ as possible. The shortest path between a point and a plane is the perpendicular. One way of seeing this is to show that $\boldsymbol{v}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$ for each $\boldsymbol{v} \in C(\boldsymbol{X})$. Since the columns $\{\boldsymbol{x}_{*1}, \dots, \boldsymbol{x}_{*p}\}$ of $\boldsymbol{X}$ form a basis for $C(\boldsymbol{X})$, it suffices to show that $\boldsymbol{x}_{*j}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$ for each $j = 1, \dots, p$. This is a consequence of the normal equations $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = 0$ derived in (10).

To derive the projection matrix corresponding to this orthogonal projection, we write

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{13}$$

where

$$\boldsymbol{H} \equiv \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \tag{14}$$

is called the *hat matrix*. This is the orthogonal projection matrix onto $C(\boldsymbol{X})$. Recall that a matrix $\boldsymbol{P}$ is an orthogonal projection onto a subspace $\boldsymbol{W}$ if for all $\boldsymbol{v} \in \boldsymbol{W}$ we have $\boldsymbol{P}\boldsymbol{v} = \boldsymbol{v}$ and for all $\boldsymbol{v} \in \boldsymbol{W}^{\perp}$ we have $\boldsymbol{P}\boldsymbol{v} = 0$. We can check for example the first of these conditions by noting that if $\boldsymbol{v} \in C(\boldsymbol{X})$, then $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$. Therefore, we have

$$\boldsymbol{H}\boldsymbol{v} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{v}.$$

A simple example of $\boldsymbol{H}$ can be obtained by considering the intercept-only regression.

One consequence of this observation is that the fitted values $\widehat{\boldsymbol{\mu}}$ depend on $\boldsymbol{X}$ only through $C(\boldsymbol{X})$. As we will see in Homework 1, there are many different model matrices $\boldsymbol{X}$ leading to the same model space. Essentially, this reflects the fact that there are many different bases for the same vector space. Consider for example changing the units on the columns of $\boldsymbol{X}$. It can be verified that not just the fitted values $\widehat{\boldsymbol{\mu}}$ but also the predictions on a new set of features remain invariant to reparametrization (this follows from parts (a) and (b) of Homework 1 Problem 1). Therefore, while reparametrization can have a huge impact on the fitted coefficients, it has no impact on the predictions of linear regression.

The orthogonality property of least squares, together with the Pythagorean theorem, leads to the following fundamental relationship. Let's say that $S \subset \{0, 1, \dots, p\}$ is a subset of the predictors. First regress $\boldsymbol{y}$ on $\boldsymbol{X}$ to get $\widehat{\boldsymbol{\beta}}$ as usual. Then, we consider the *partial model matrix* $\boldsymbol{X}_{*S}$ obtained by selecting only the columns in S. Regression $\boldsymbol{y}$ on $\boldsymbol{X}_{*S}$ results in $\widehat{\boldsymbol{\beta}}_S$ (note: $\widehat{\boldsymbol{\beta}}_S$ is not necessarily obtained from $\widehat{\boldsymbol{\beta}}$ by extracting the coefficients corresponding to S). Now, consider the three points $\boldsymbol{y}, \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S \in \mathbb{R}^n$. Since $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S$ are both in $C(\boldsymbol{X})$, it follows by the orthogonal projection properpty that $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is orthogonal to $\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S$. In other words, these three points form a right triangle. By the Pythagorean theorem, we conclude that

$$\|\boldsymbol{y} - \boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S\|^2 = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S\|^2 + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2. \tag{15}$$

We will rely on this fundamental relationship throughout this course.

For now, we can extract a few consequences of the relationship (15). As a starting point, consider the case when $S = \{0\}$, i.e. the partial model is the intercept-only model. In this case, $\boldsymbol{X}_{*S} = \mathbf{1}_n$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}} = \bar{y}$. Therefore, equation (15) implies that

$$\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}_n\|^2 + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2. \tag{16}$$

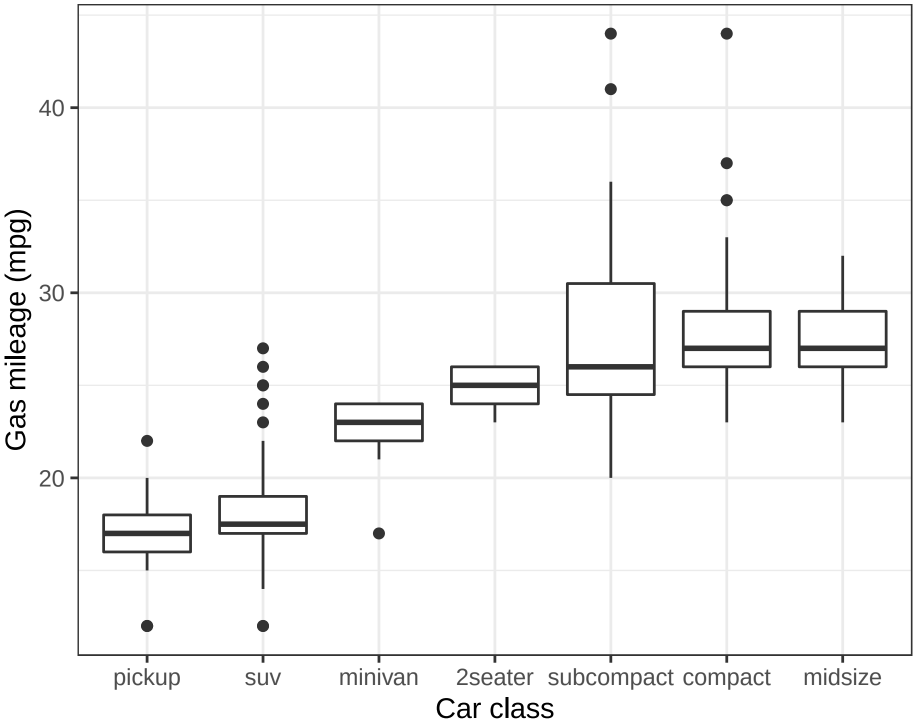
Equivalently, we can rewrite this equation as follows:

$$\text{SST} \equiv \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\widehat{\mu}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \widehat{\mu}_i)^2 \equiv \text{SSR} + \text{SSE}. \tag{17}$$

6 Correlation, multiple correlation, and R^2 (Agresti 2.1.3, 2.4.6)

ANOVA decomposition for C groups model. Let's consider the special case of the ANOVA decomposition (17) when the model matrix $\boldsymbol{X}$ represents a single categorical predictor w. In this case, each observation i is associated to one of the C classes of w, which we denote $c(i) \in \{1, \dots, C\}$. Let's consider the C groups of observations $\{i : c(i) = c\}$ for $c \in \{1, \dots, C\}$. For example, w may be the type of a car (compact, midsize, minivan, etc.) and y might be its fuel efficiency in miles per gallon.

```r
mpg %>%
  ggplot() +
  geom_boxplot(aes(x = fct_reorder(class, hwy), y = hwy)) +
  labs(x = "Car class", y = "Gas mileage (mpg)") +
  theme_bw()
```



It is easy to check that the least squares fitted values $\widehat{\mu}_i$ are simply the means of the corresponding groups:

$$\widehat{\mu}_i = \bar{y}_{c(i)}, \quad \text{where } \bar{y}_{c(i)} \equiv \frac{\sum_{i:c(i)=c} y_i}{|\{i : c(i) = c\}|}. \tag{18}$$

Therefore, we have

$$\text{SSR} = \sum_{i=1}^{n} (\widehat{\mu}_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y}_{c(i)} - \bar{y})^2 \equiv \text{between-groups sum of squares (SSB)} \tag{19}$$

and

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y}_{c(i)})^2 \equiv \text{within-groups sum of squares (SSW)}. \tag{20}$$

We therefore obtain the following corollary of the ANOVA decomposition (17):

$$\text{SST} = \text{SSB} + \text{SSW}. \tag{21}$$

R^2 definition and (multiple) correlation. The ANOVA decompositions (17) and (21) of the variation in $\boldsymbol{y}$ into that explained by the linear regression model (SSR) and that left over (SSE) leads naturally to the definition of R^2 as the fraction of variation in $\boldsymbol{y}$ explained by the linear regression model:

$$R^2 \equiv \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n} (\widehat{\mu}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}_n\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2}. \tag{22}$$

By the decomposition (17), we have $R^2 \in [0, 1]$. The closer R^2 is to 1, the closely the data follow the fitted linear regression model. There is a connection between R^2 and correlation. To see this, let us first consider the case of the simple linear regression model with one predictor

$$y = \beta_0 + \beta_1 x_1 + \epsilon. \tag{23}$$

In this simple case, one can directly derive a formula for the fitted coefficients:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}; \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{24}$$

Therefore,

$$\widehat{\boldsymbol{\mu}} - \bar{y}\mathbf{1}_n = \widehat{\beta}_0 \mathbf{1}_n + \widehat{\beta}_1 \boldsymbol{x}_{*1} - \bar{y}\mathbf{1}_n = \widehat{\beta}_1 (\boldsymbol{x}_{*1} - \bar{x}\mathbf{1}_n)$$

and thus

$$R^2 = \frac{\|\widehat{\boldsymbol{\mu}} - \bar{y}\mathbf{1}_n\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2} = \frac{\widehat{\beta}_1^2 \|\boldsymbol{x}_{*1} - \bar{x}\mathbf{1}_n\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2} = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{1/2} \left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)^{1/2}} \right)^2 \equiv \rho_{xy}^2, \tag{25}$$

where ρ_{xy} is the sample correlation between x_1 and y. Therefore, in a simple linear regression, R^2 is the squared sample correlation between x_1 and y. For general regressions, one can derive that R^2 is the squared sample correlation between $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{y}$. For this reason, R^2 is sometimes called the *multiple correlation coefficient*.

Regression to the mean. Let's go back to the simple regression model (23), and let's take a closer look at $\widehat{\beta}_1$ in (24). Denoting by ρ_x is the sample standard deviation of x_1 and ρ_y is the sample standard deviation of y, we can rewrite $\widehat{\beta}_1$ as

$$\widehat{\beta}_1 = \frac{\rho_y}{\rho_x} \cdot \rho_{xy}. \tag{26}$$

Assuming that $\boldsymbol{x}_{*1}$ and $\boldsymbol{y}$ have been normalized to have the same sample standard deviation $\rho_x = \rho_y$, we find that the least squares coefficient $\widehat{\beta}_1$ is equal to the sample correlation ρ_{xy} between x and y. Since $|\rho_{xy}| < 1$ unless $\boldsymbol{x}_{*1}$ and $\boldsymbol{y}$ are perfectly correlated (by the Cauchy-Schwarz inequality), this means that

$$|\widehat{\mu}_i - \bar{y}| < |x_i - \bar{x}| \quad \text{for each } i. \tag{27}$$

Therefore, we expect y_i to be closer to its mean than x_i is to its mean. This phenomenon is called *regression to the mean* (and is in fact the origin of the term "regression"). Many mistakenly attribute a causal mechanism to this phenomenon, when in reality it is simply a statistical artifact. For example, suppose x_i is the number of games a sports team won last season and y_i is the number of games it won this season. It is widely observed that teams with exceptional performance in a given season suffer a "winner's curse", performing worse in the next season. The reason for the winner's curse is simple: teams perform exceptionally well due to a combination of skill and luck. While skill stays roughly constant from year to year, the team which performed exceptionally well in a given season is unlikely to get as lucky as it did next season.

R^2 increases as predictors are added. The R^2 is an *in-sample* measure, i.e. it uses the same data to fit the model and to assess the quality of the fit. Therefore, it is generally an optimistic measure of the (out-of-sample) prediction error. One manifestation of this is that the R^2 increases if any predictors are added to the model (even if these predictors are "junk"). To see this, it suffices to show that SSE decreases as we add predictors. Without loss of generality, suppose that we start with a model including predictors $S \subset \{0, 1, \dots, p\}$ and compare it to the model including all the predictors $\{0, 1, \dots, p\}$. We can read off from the Pythagorean theorem (15) that

$$\text{SSE}(\boldsymbol{X}_{*S}, \boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{X}_{*S}\widehat{\boldsymbol{\beta}}_S\|^2 \geq \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 = \text{SSE}(\boldsymbol{X}, \boldsymbol{y}).$$

Adding many junk predictors will have the effect of degrading predictive performance but will nevertheless increase R^2.

7 Collinearity, adjustment, and partial correlation (Agresti 2.2.4, 2.5.6, 2.5.7, 4.6.5)

An important part of linear regression analysis is the dependence of the least squares coefficient for a predictor on what other predictors are in the model. This relationship is dictated by the extent to which the given predictor is correlated with the other predictors. In this section, we'll use some additional notation. Let $S \subset \{1, \dots, p\}$ be a group of predictors (we can assume without loss of generality that $S = \{1, \dots, s\}$ for some $1 \leq s < p$). Then, denote $\text{-}S \equiv \{1, \dots, p\} \setminus S$. Let $\widehat{\boldsymbol{\beta}}_S$ denote the least squares coefficients when regressing $\boldsymbol{y}$ on $\boldsymbol{X}_{*S}$ and let $\widehat{\boldsymbol{\beta}}_{S|\text{-}S}$ denote the least squares coefficients corresponding to S when regressing $\boldsymbol{y}$ on $\boldsymbol{X} = (\boldsymbol{X}_{*S}, \boldsymbol{X}_{*,\text{-}S})$.

Least squares estimates in the orthogonal case. The simplest case to analyze is when a groups of predictors $\boldsymbol{X}_{*S}$ is orthogonal to the rest of the predictors $\boldsymbol{X}_{*,\text{-}S}$ in the sense that

$$\boldsymbol{X}_{*S}^T \boldsymbol{X}_{*,\text{-}S} = \boldsymbol{0}. \tag{28}$$

In this case, we can derive the least squares coefficient vector $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{S|\text{-}S}, \widehat{\boldsymbol{\beta}}_{\text{-}S|S})$ from the normal equations:

$$\begin{aligned}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{S|\text{-}S} \\ \widehat{\boldsymbol{\beta}}_{\text{-}S|S}\end{pmatrix} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \begin{pmatrix}\boldsymbol{X}_S^T\boldsymbol{X}_S & 0 \\ 0 & \boldsymbol{X}_{\text{-}S}^T\boldsymbol{X}_{\text{-}S}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{X}_S^T \\ \boldsymbol{X}_{\text{-}S}\end{pmatrix}\boldsymbol{y} \\ &= \begin{pmatrix}(\boldsymbol{X}_S^T\boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^T\boldsymbol{y} \\ (\boldsymbol{X}_{\text{-}S}^T\boldsymbol{X}_{\text{-}S})^{-1}\boldsymbol{X}_{\text{-}S}^T\boldsymbol{y}\end{pmatrix} \\ &= \begin{pmatrix}\widehat{\boldsymbol{\beta}}_S \\ \widehat{\boldsymbol{\beta}}_{\text{-}S}\end{pmatrix}.\end{aligned} \tag{29}$$

Therefore, the least squares coefficients when regressing $\boldsymbol{y}$ on $(\boldsymbol{X}_S, \boldsymbol{X}_{\text{-}S})$ are the same as those obtained from regressing $\boldsymbol{y}$ separately on $\boldsymbol{X}_S$ and $\boldsymbol{X}_{\text{-}S}$, i.e.

$$\widehat{\boldsymbol{\beta}}_{S|\text{-}S} = \widehat{\boldsymbol{\beta}}_S. \tag{30}$$

Least squares estimates via orthogonalization. Let's now focus our attention on a single predictor x_j. If this predictor is orthogonal to the remaining predictors, then the result (30) states that $\widehat{\beta}_{j|\text{-}j}$ can be obtained from simply regressing y on x_j. However, this is usually not the case. Usually, $\boldsymbol{x}_{*j}$ has a nonzero projection $\boldsymbol{X}_{*,\text{-}j}\widehat{\gamma}$ onto $C(\boldsymbol{X}_{*,\text{-}j})$:

$$\boldsymbol{x}_{*j} = \boldsymbol{X}_{*,\text{-}j}\widehat{\gamma} + \boldsymbol{x}_{*j}^{\perp}, \tag{31}$$

where $\boldsymbol{x}_{*j}^{\perp}$ is the residual from regressing $\boldsymbol{x}_{*j}$ onto $\boldsymbol{X}_{*,\text{-}j}$ and is therefore orthogonal to $C(\boldsymbol{X}_{*,\text{-}j})$. In other words, $\boldsymbol{x}_{*j}^{\perp}$ is the projection of $\boldsymbol{x}_{*j}$ onto the orthogonal complement of $C(\boldsymbol{X}_{*,\text{-}j})$.

With this decomposition, let us change basis from $(\boldsymbol{x}_{*j}, \boldsymbol{X}_{*,\text{-}j})$ to $(\boldsymbol{x}_{*j}^{\perp}, \boldsymbol{X}_{*,\text{-}j})$ by the process explored in Homework 1 Question 1. Let us write

$$\begin{aligned}\boldsymbol{y} = \boldsymbol{x}_{*j}\beta_{j|\text{-}j} + \boldsymbol{X}_{*,\text{-}j}\boldsymbol{\beta}_{\text{-}j|j} + \boldsymbol{\epsilon} \iff &\ \boldsymbol{y} = (\boldsymbol{X}_{*,\text{-}j}\widehat{\gamma} + \boldsymbol{x}_{*j}^{\perp})\beta_{j|\text{-}j} + \boldsymbol{X}_{*,\text{-}j}\boldsymbol{\beta}_{\text{-}j|j} + \boldsymbol{\epsilon} \\ \iff &\ \boldsymbol{y} = \boldsymbol{x}_{*j}^{\perp}\beta_{j|\text{-}j} + \boldsymbol{X}_{*,\text{-}j}\boldsymbol{\beta}'_{\text{-}j|j} + \boldsymbol{\epsilon}.\end{aligned}$$

What this means is that $\widehat{\beta}_{j|\text{-}j}$, the least squares coefficient of $\boldsymbol{x}_{*j}$ in the regression of $\boldsymbol{y}$ on $(\boldsymbol{x}_{*j}, \boldsymbol{X}_{*,\text{-}j})$ is also the least squares coefficient of $\boldsymbol{x}_{*j}^{\perp}$ in the regression of $\boldsymbol{y}$ on $(\boldsymbol{x}_{*j}^{\perp}, \boldsymbol{X}_{*,\text{-}j})$. However, since $\boldsymbol{x}_{*j}^{\perp}$ is orthogonal to $\boldsymbol{X}_{*,\text{-}j}$ by construction, we can use the result (29) to conclude that

$\widehat{\beta}_{j|\text{-}j}$ is the least squares coefficient of $\boldsymbol{x}_{*j}^{\perp}$ in the *univariate* regression of $\boldsymbol{y}$ on $\boldsymbol{x}_{*j}^{\perp}$ (without intercept).

We can solve this univariate regression explicitly to obtain

$$\widehat{\beta}_{j|\text{-}j} = \frac{(\boldsymbol{x}_{*j}^{\perp})^T\boldsymbol{y}}{\|\boldsymbol{x}_{*j}^{\perp}\|^2}. \tag{32}$$

Adjustment and partial correlation. Equivalently, letting $\widehat{\boldsymbol{\beta}}_{\text{-}j}$ be the least squares estimate in the regression of $\boldsymbol{y}$ on $\boldsymbol{X}_{*,\text{-}j}$ (note that this is *not* the same as $\widehat{\boldsymbol{\beta}}_{\text{-}j|j}$), we can write

$$\widehat{\beta}_{j|\text{-}j} = \frac{(\boldsymbol{x}_{*j}^{\perp})^T(\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\beta}}_{\text{-}\boldsymbol{j}})}{\|\boldsymbol{x}_{*j}^{\perp}\|^2} = \frac{(\boldsymbol{x}_{*j} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\gamma}})^T(\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\beta}}_{\text{-}\boldsymbol{j}})}{\|\boldsymbol{x}_{*j} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\gamma}}\|^2}. \tag{33}$$

We can interpret this result as follows: The linear regression coefficient $\widehat{\beta}_{j|\text{-}j}$ results from first adjusting $\boldsymbol{y}$ and $\boldsymbol{x}_{*j}$ for the effects of all other variables, and then regressing the residuals from $\boldsymbol{y}$ onto the residuals from $\boldsymbol{x}_{*j}$. In this sense, *the least squares coefficient for a predictor in a multiple linear regression reflects the effect of the predictor on the response after controlling for the effects of all other predictors.* A related quantity is the *partial correlation* between $\boldsymbol{x}_{*j}$ and $\boldsymbol{y}$ after controlling for $\boldsymbol{X}_{*,\text{-}j}$, defined as the correlation between $\boldsymbol{x}_{*j} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\gamma}}$ and $\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\beta}}_{\text{-}\boldsymbol{j}}$. We can then connect the least squares coefficient $\widehat{\beta}_j$ to this partial correlation in a similar spirit to equation (26).

Aside: Average treatment effect estimation in causal inference. Suppose we'd like to study the effect of an exposure or treatment on a response y. Letting y_1 and y_0 denote the responses under treatment and control (Neyman-Rubin causal model), the most basic goal is to estimate the *average treatment effect* $\tau \equiv \mathbb{E}[y_1 - y_0]$. Usually in observational studies we have *confounding variables* $z_1, \dots, z_p$: variables that influence both the treatment assignment and the response. It is important to control for these confounders in order to get an unbiased estimate of the treatment effect. Suppose all confounders are measured (i.e. $(y_1, y_0) \perp\!\!\!\perp t \mid z_1, \dots, z_p$), the treatment effect is constant, and the response is a linear function of the treatment and confounders:

$$y = \beta t + \gamma_1 z_1 + \cdots + \gamma_p z_p + \epsilon. \tag{34}$$

Then, the average treatment effect τ is identified as the coefficient β in the above regression, i.e. $\tau = \beta$. Therefore, the least squares estimate $\widehat{\beta}_{t|z}$ is an unbiased estimate of the average treatment effect. (Causal inference is beyond the scope of STAT 961; see STAT 921 instead.)

Effects of collinearity. Collinearity between a predictor x_j and the other predictors tends to make the estimate $\widehat{\beta}_{j|\text{-}j}$ unstable. Intuitively, this makes sense because it becomes harder to distinguish between the effects of predictor x_j and those of the other predictors on the response. To find the variance of $\widehat{\beta}_{j|\text{-}j}$ for a model matrix $\boldsymbol{X}$, we could in principle use the formula (12). However, this formula involves the inverse of the matrix $\boldsymbol{X}^T\boldsymbol{X}$, which is hard to reason about. Instead, we can employ the formula (32) to calculate directly that

$$\text{Var}[\widehat{\beta}_{j|\text{-}j}] = \frac{\sigma^2}{\|\boldsymbol{x}_{*j}^{\perp}\|^2}. \tag{35}$$

We see that the variance of $\widehat{\beta}_{j|\text{-}j}$ is inversely proportional to $\|\boldsymbol{x}_{*j}^{\perp}\|^2$. This means that the greater the collinearity, the less of $\boldsymbol{x}_{*j}$ is left over after adjusting for $\boldsymbol{X}_{*,\text{-}j}$, and the greater the variance of $\widehat{\beta}_{j|\text{-}j}$. To quantify the effect of this adjustment, suppose there were no other predictors other than the intercept term. Then, we would have

$$\text{Var}[\widehat{\beta}_j] = \frac{\sigma^2}{\|\boldsymbol{x}_{*j} - \bar{x}_j\boldsymbol{1}_n\|^2}. \tag{36}$$

Therefore, we can rewrite the variance (35) as

$$\text{Var}[\widehat{\beta}_{j|\text{-}j}] = \frac{\|\boldsymbol{x}_{*j} - \bar{x}_j\boldsymbol{1}_n\|^2}{\|\boldsymbol{x}_{*j} - \boldsymbol{X}_{*,\text{-}j}\widehat{\boldsymbol{\gamma}}\|^2} \cdot \text{Var}[\widehat{\beta}_j] = \frac{1}{1 - R_j^2} \cdot \text{Var}[\widehat{\beta}_j] \equiv \text{VIF}_j \cdot \text{Var}[\widehat{\beta}_j], \tag{37}$$

where R_j^2 is the R^2 value when regressing $\boldsymbol{x}_{*j}$ on $\boldsymbol{X}_{*,\text{-}j}$ and VIF stands for *variance inflation factor*. The higher R_j^2, the more of the variance in $\boldsymbol{x}_{*j}$ is explained by other predictors, the higher the variance in $\widehat{\beta}_{j|\text{-}j}$.

8 R demo

Let's go back to the `ScotsRaces` data from before:

```
scots_races = read_tsv("../data/ScotsRaces.dat", col_types = "cddd")
scots_races

## # A tibble: 35 x 4
##    race                   distance climb  time
##    <chr>                     <dbl> <dbl> <dbl>
##  1 GreenmantleNewYearDash      2.5  0.65  16.1
##  2 Carnethy5HillRace           6    2.5   48.4
##  3 CraigDunainHillRace         6    0.9   33.6
##  4 BenRhaHillRace              7.5  0.8   45.6
##  5 BenLomondHillRace           8    3.07  62.3
##  6 GoatfellHillRace            8    2.87  73.2
##  7 BensofJuraFellRace         16    7.5  205. 
##  8 CairnpappleHillRace         6    0.8   36.4
##  9 ScoltyHillRace              5    0.8   29.8
## 10 TraprainLawRace             6    0.65  39.8
## # ... with 25 more rows
```

Recall we had run a linear regression of `time` on `distance` and `climb`:

```
lm_fit = lm(time ~ distance + climb, data = scots_races)
summary(lm_fit)

## 
## Call:
## lm(formula = time ~ distance + climb, data = scots_races)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -16.654  -4.842   1.110   4.667  27.762 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -13.1086     2.5608  -5.119 1.41e-05 ***
## distance      6.3510     0.3578  17.751  < 2e-16 ***
## climb        11.7801     1.2206   9.651 5.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.734 on 32 degrees of freedom
## Multiple R-squared:  0.9717,Adjusted R-squared:   0.97
## F-statistic: 549.9 on 2 and 32 DF,  p-value: < 2.2e-16
```

We get a coefficient of 6.35 with standard error 0.347 for `distance`, where the standard error is an estimate of the quantity (35).

R^2 and sum-of-squared decompositions. We can extract the R^2 from this fit by reading it off from the bottom of the summary, or by typing

```
summary(lm_fit)$r.squared

## [1] 0.971725
```

We can construct sum-of-squares decompositions (15) using the `anova` function. This function takes as arguments the partial model and the full model. For example, consider the partial model `time distance`.

```
lm_fit_partial = lm(time ~ distance, data = scots_races)
anova(lm_fit_partial, lm_fit)

## Analysis of Variance Table
##
## Model 1: time ~ distance
## Model 2: time ~ distance + climb
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1     33 9546.9
## 2     32 2441.3  1    7105.6 93.14 5.369e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that adding the predictor `climb` reduces the RSS by 7106, from 9547 to 2441. As another example, we can compute the R^2 by comparing the full model with the null model:

```
lm_fit_null = lm(time ~ 1, data = scots_races)
anova(lm_fit_null, lm_fit)

## Analysis of Variance Table
##
## Model 1: time ~ 1
## Model 2: time ~ distance + climb
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     34 86340
## 2     32  2441  2     83899 549.87 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, the R^2 is $83899/86340 = 0.972$, consistent with the above regression summary.

Adjustment and collinearity. We can also test the adjustment formula (32) numerically. Let's consider the coefficient of `distance` in the regression `time ~ distance + climb`. We can obtain this coefficient by first regressing `climb` out of `distance` and `time`:

```
lm_dist_on_climb = lm(distance ~ climb, data = scots_races)
lm_time_on_climb = lm(time ~ climb, data = scots_races)

scots_races_resid =
  bind_cols(scots_races,
            dist_residuals = lm_dist_on_climb$residuals,
            time_residuals = lm_time_on_climb$residuals)

lm_adjusted = lm(time_residuals ~ dist_residuals-1,
                 data = scots_races_resid)
summary(lm_adjusted)

##
## Call:
## lm(formula = time_residuals ~ dist_residuals - 1, data = scots_races_resid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.654  -4.842   1.110   4.667  27.762
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## dist_residuals   6.3510     0.3471    18.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.474 on 34 degrees of freedom
## Multiple R-squared:  0.9078,Adjusted R-squared:  0.9051
## F-statistic: 334.8 on 1 and 34 DF,  p-value: < 2.2e-16
```

We find a coefficient of 6.35 with standard error 0.347, which matches that obtained in the original regression. Suppose we had not regressed `climb` out of `time`:

```
lm_dist_adjusted = lm(time ~ dist_residuals-1,
                 data = scots_races_resid)
summary(lm_dist_adjusted)

##
## Call:
## lm(formula = time ~ dist_residuals - 1, data = scots_races_resid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  18.42  30.93  40.10  64.74 231.17
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## dist_residuals    6.351      2.917   2.177   0.0365 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.21 on 34 degrees of freedom
## Multiple R-squared:  0.1224,Adjusted R-squared:  0.09655
## F-statistic:  4.74 on 1 and 34 DF,  p-value: 0.0365
```

The estimate stays the same but the standard error increases. Why is this the case? To obtain the variance inflation factors defined in equation (37), we can use the `vif` function from the `car` package:

```
car::vif(lm_fit)

## distance    climb
## 1.740812 1.740812
```

Why are these two VIF values the same?