This material was produced for instructors using Llaudet, Elena and Kosuke Imai.
Data Analysis for Social Science: A Friendly and Practical Introduction. (Princeton University Press)
and should not be shared beyond those who are enrolled in this class.

Predicting Elections Using Betting Markets

In this exercise, we aim to predict electoral outcomes based on data from the betting markets. For this purpose, we will analyze state-level data on the 2008 U.S. presidential election where the two main candidates were Obama (the Democratic candidate) and McCain (the Republican candidate) and model the relationship between (i) the closing price the day before the election of contracts from the online betting company Intrade and (ii) the real electoral outcomes. At Intrade, people trade contracts such as 'Obama to win the electoral college votes of Florida.' The market price of each contract fluctuates based on its sales. The price of each nominee's contract can be interpreted as the vote share that the market expects that nominee to receive in that state.

The dataset we will use is in the *intrade.csv* file. Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is states (plus DC).

variable	description
state	state's two letter abbreviation
D_expected	closing price or expected vote share of the Democratic candidate's contract in that state (in percentages)
R_expected	closing price or expected vote share of the Republican candidate's contract in that state (in percentages)
D_real	real vote share received by the Democratic candidate in that state (in percentages)
R_real	real vote share received by the Republican candidate in that state (in percentages)

Table 1: Variables in "intrade.csv"

In this problem set, we practice creating new variables, creating histograms, computing correlations, creating scatter plots, fitting linear models, making predictions with the fitted line, and computing R^2.

As always, let's start by loading and exploring the data:

```
## load and look at the data
intrade <- read.csv("intrade.csv") # reads and stores data
head(intrade) # shows first observations
##   state D_expected R_expected D_real R_real
## 1    AK        6.0       94.0     38     59
## 2    AL        3.3       95.0     39     60
## 3    AR        3.9       90.0     39     59
## 4    AZ       18.9       80.2     45     54
## 5    CA       97.9        2.1     61     37
## 6    CO       91.5       12.0     54     45
```

And, to confirm that we have data on all 50 states plus DC, we can run:

```
dim(intrade) # provides dimensions: rows, columns
## [1] 51  5
```

1. First, let's create our Y and X variables, figure out their unit of measurement, and explore whether they are moderately to strongly linearly associated with each other.

   a. Create our X variable, that is our predictor (i.e., what we will use as the basis of our prediction): the expected democratic margin based on the betting markets, defined as the closing price of the Democratic candidate's contract minus the closing price of the Republican candidate's contract. Call this variable *expected_margin* and make sure to store it inside the dataframe *intrade*. (R code only) (2.5 points)

   b. Create our Y variable, that is our outcome variable (i.e., what we want to predict): the real democratic margin, defined as the real vote share received by the Democratic candidate minus the real vote share received by the Republican candidate. Call this variable *real_margin* and make sure to store it inside the dataframe *intrade*. (R code only) (2.5 points)

   c. Use the function head() to look at the first few observations of *intrade* again to ensure that the two new variables were created correctly. In what unit of measurement are the two new variables, *expected_margin* and *real_margin*? (5 points)

   d. Do the betting markets do a good job in predicting the real democratic margin? To answer this question, create the histogram of the differences between *real_margin* and *expected_margin*. If the betting markets do a good job in predicting the real democratic margin, most of the observations in the histogram will be around 0 because the value of *real_margin* will equal the value of *expected_margin*. Is that what you observe in the histogram? A yes/no answer will suffice. (5 points)

   e. Is the relationship between *expected_margin* and *real_margin* moderate to strongly linear? If so, we could model the relationship between *expected_margin* and *real_margin* with a line and then use that line to predict *real_margin* based on the value of *expected_margin*. Please answer this question, by computing the correlation coefficient between *expected_margin* and *real_margin*. A yes/no answer will suffice. (5 points)

2. Second, let's fit the linear model that we will use to make predictions.

   a. Use the function lm() to fit a linear model to summarize the relationship between *expected_margin* and *real_margin* and store the output in an object called *fit*. Then, ask R to provide the contents of *fit* by running its name. (R code only.) (5 points)

   b. What is the fitted line? In other words, provide the formula $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute $\widehat{\alpha}$ for the estimated value of the intercept coefficient, substitute $\widehat{\beta}$ for the estimated value of the slope coefficient, and substitute X for the name of the predictor.) (5 points)

   c. Create a visualization of the relationship between *expected_margin* and *real_margin* and add the fitted line to the graph using the function abline(). (R code only.). (5 points)

3. Now, let's use the fitted line to make some predictions.

   a. Computing $\widehat{Y}$ based on X: Suppose the day before the next U.S. presidential election, you notice that the difference between the closing price of the Democratic candidate's contract and the closing price of the Republican candidate's contract (i.e., the expected democratic margin) in a particular state equals 20 percentage points. By how much would you predict that the Democratic candidate will win that state? (5 points)

   b. Computing $\triangle\widehat{Y}$ based on $\triangle X$: What is the predicted change in the real democratic margin associated with an increase in the expected democratic margin of 10 percentage points? (5 points)

c. To explore how good the model is at making predictions, compute the R^2 of the fitted model and interpret its value. (5 points)