Bootstrap and Linear Regression

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

You should have downloaded studio12.zip and unzipped it into your 18.05 working directory.

Review: Computing a bootstrap confidence interval

Starting with data $x_1,\ x_2,\ \ldots x_n$ and test statistic $\hat{\theta}$.

1. Generate a resample of the data of size `n`.
2. Compute the test statistic θ^*.
3. Compute and store the bootstrap difference $\theta^* - \hat{\theta}$.
4. Repeat steps 1-3 `nboot` times.
5. The bootstrap confidence interval is

$$\left[\hat{\theta} - \delta^*_{1-\alpha/2},\ \hat{\theta} - \delta^*_{\alpha/2}\right]$$

where $\delta^*_{\alpha/2}$ is a *quantile*.

Board question: two independent samples

Suppose

$$x_1, x_2, \ldots, x_n \qquad \text{and} \qquad y_1, y_2, \ldots, y_m$$

are independent samples drawn from distributions with means μ_x and μ_y respectively.

Describe in detail the steps for computing an empirical bootstrap 95% confidence interval for $\mu_x - \mu_y$

Solution

The steps are almost identical to the ones outlined above. The only difference is that you have to generate two independent bootstrap resamples at each step:
The test statistic is $\hat{\theta} = \overline{x} - \overline{y}$.

1. Resample:

$$x_1^*, x_2^*, \ldots, x_n^* \text{ resampled from } x_1, \ldots, x_n$$
$$y_1^*, y_2^*, \ldots, y_m^* \text{ resampled from } y_1, \ldots, y_m$$

2. Compute $\theta^* = \overline{x}^* - \overline{y}^*$
3. Compute and store the bootstrap difference $\theta^* - \hat{\theta}$
4. Repeat steps 1-3 `nboot` times.
5. The bootstrap confidence interval is

$$\hat{\theta} - \delta^*_{1-\alpha/2}, \ \hat{\theta} - \delta^*_{\alpha/2}$$

where $\delta^*_{\alpha/2}$ is a *quantile*.

R Problem 1: Bootstrapping

The data file `salaries.csv` contains two columns of data: `Salaries.1` and `Salaries.2`. Using R compute a 95% bootstrap confidence interval for the difference of the two means.

- The file `studio12.r` has code that will show you how load the data in `salaries.csv`
- `studio12.r` also has sample code for computing a one-sample bootstrap confidence interval.

answer: *Code for the solution is in studio12-sol.r*

Linear regression using R

We will use R to analyze financial data using simple linear regression.

- We'll use publically available data on the price of several stocks over a 14 year period from 2000 to 2014.
- We'll use a simplified version of the *Capital Asset Pricing Model* or CAPM.
- The file `studio12.r` contains code for loading the data and fitting a line to two of the variables.

Exploring the original data

• The original data is in the file `studio12financialOriginal.csv`
The next two sides show a quick exporation of this data.

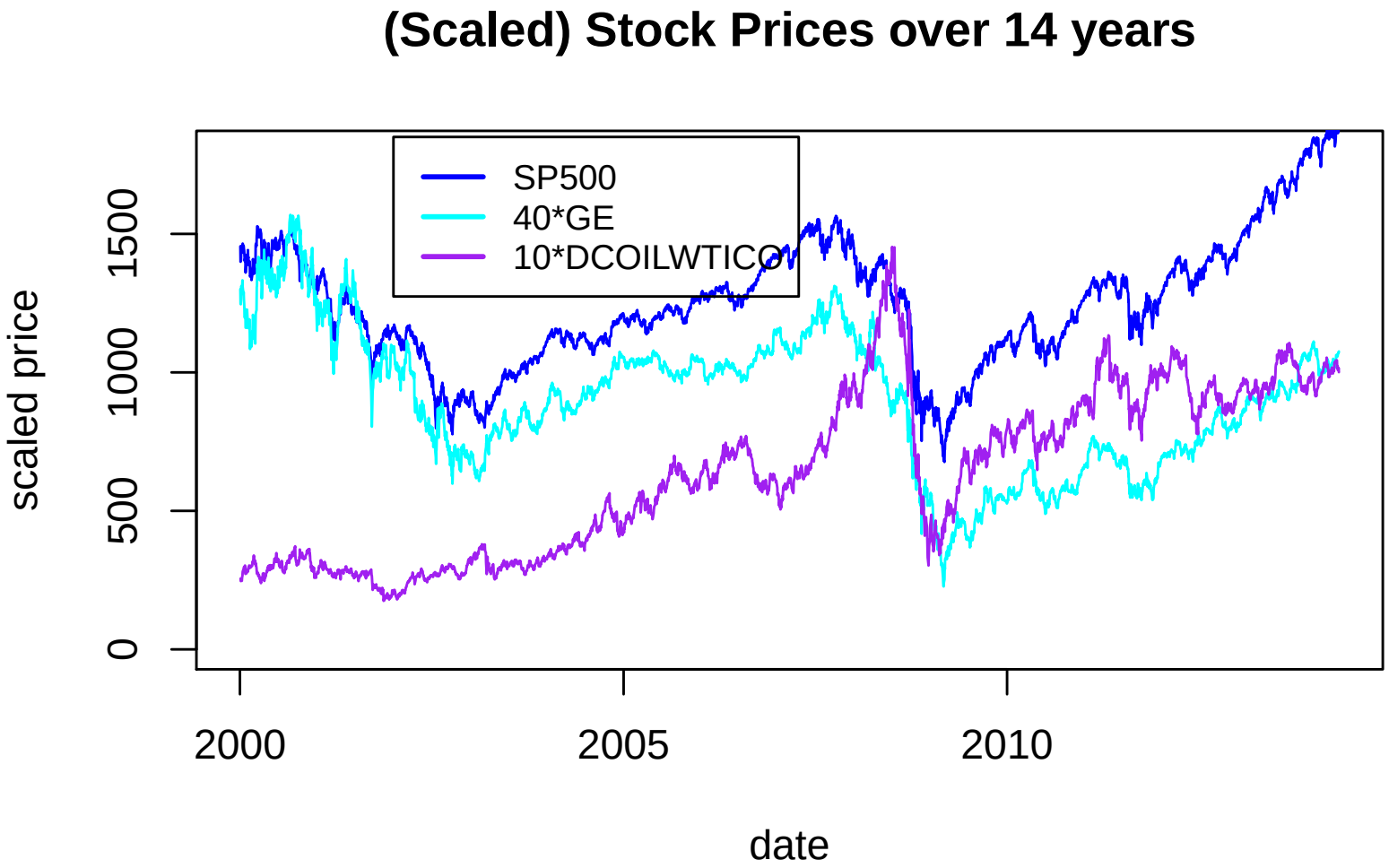
• The file contains daily prices or interest rates, over 14 years, for several stocks, bonds and a barrel of oil.

We will focus on:

- SP500: Standard & Poors stock index, the cost of a certain basket of stocks.
- GE: General Electric.
- DCOICWTICO: the price of a barrel of the benchmark West Texas crude oil.

Price plots

• The prices are scaled so they would all fit nicely on the same axes. For example, by plotting 40*GE vs. date it is at roughly the same scale as the others.



Daily rate of return

For this project we'll look at the daily rate of return for the financial variables.

Our goal is to fit the data with linear models of the form

```
DCOICWTICO.daily = a + b*SP500.daily
```

- The daily rate of return was precomputed using `studio12-prep.r`
- The data is stored in `studio12financialDaily.csv`
- The R file `studio12.r` has code to load this data and fit a linear model

Let's look at `studio12.r` now!

Fitting a linear model using R

Linear model: `DCOICWTICO.daily = a + b*SP500.daily`

R code:

```
lmfit = lm(DCOILWTICO.daily ~ SP500.daily)
summary(lmfit):
```

```
Coefficients:
              Estimate Std.       Error  t value  Pr(>|t|)
(Intercept)     0.0003469  0.0003546    0.978  0.328
SP500.daily     0.3325257  0.0311009   10.692  <2e-16 ***
```

- This estimates the model coefficients as

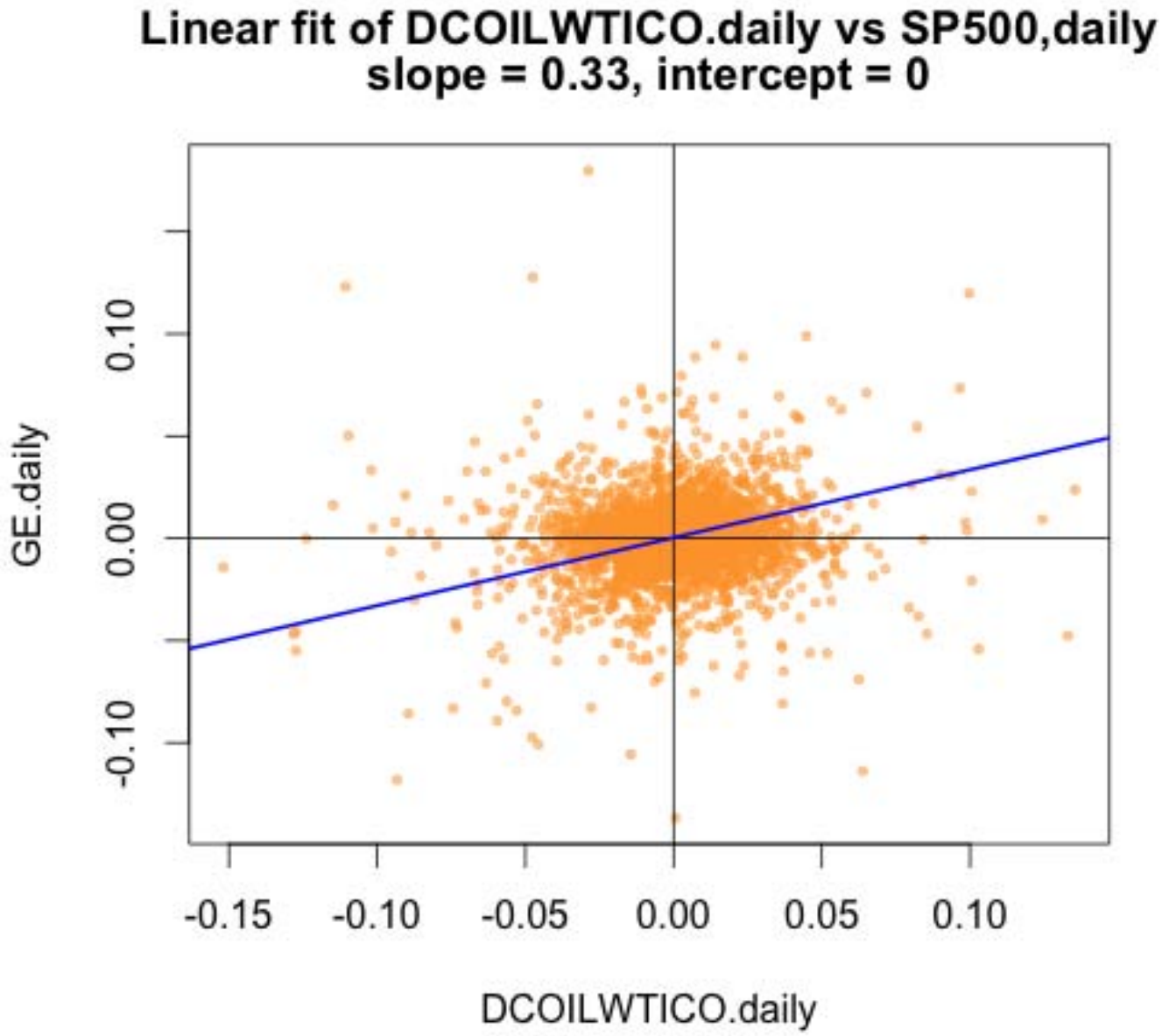
$$\hat{a} = 0.0003469 \quad \text{and} \quad \hat{b} = 0.3325257.$$

- `Pr(>|t|)` are p-values for a NHST with H_0 that the given coefficient is 0.

  $\hat{b}$ is significantly different from 0, with p-value $< 2 \cdot 10^{-16}$.

  $\hat{a}$ (the constant term) is not significantly different from 0.

Linear fit


Linear fit of DCOILWTICO.daily vs SP500,daily
slope = 0.33, intercept = 0
GE.daily
0.10
0.00
-0.10
-0.15
-0.10
-0.05
0.00
0.05
0.10
DCOILWTICO.daily

R Problem 2: Linear regression

1. Load the data from `studio12financialDaily.csv`
2. Do a linear fit of `GE.daily` vs. `SP500.daily`
3. Interpret the results.
4. Run the multiple linear regression at the end of studio12.r and interpret the results.

answer: The answers to 1 and 2 are in `studio12-sol.r`

3. The coefficient of SP500.daily is positive and significantly different from 0. This indicates that SP500.daily and GE.daily are positively correlated. They tend to rise and fall together.

Continued on next slide.

Solution continued

4. The coefficients of SP500.daily and DCOILWTICO.daily are both significantly different from 0. Interestingly the coefficient of DCOILWTICO.daily is negative.

When we compared GE and DCOILWTICO without including SP500 the coefficient of DCOILWTICO was positive. So each pair of GE, SP500 and DCOILWTICO are positively correlated: they all tend to rise and fall together.

With both SP500.daily and COILWTICO.daily as predictor variables for GE, we see that if the SP500 is flat, then a rise in the price of oil predicts a fall in the price of GE.

One possible explanation: higher energy costs that aren't rooted in broader economic growth raise GE's costs without increasing its profits.

Note: The `lm` function returns a lot of other information bearing on the predictive power of the predictor variables.

18.05 Introduction to Probability and Statistics

Spring 2014