# 2 · *Simple statistical models*

**Due Monday, February 2, 2015**

*(1) Group-wise models*

The data set oxford.csv contains data on the agreed closing price of 1,134 "flats" (i.e. apartments) sold in Oxfordshire, England between 2000 and 2011. As you might guess from the name, Oxfordshire is the county containing the ancient and beautiful university town of Oxford.

This is a subset of a much larger database of home prices across the United Kingdom over that period, but for now we'll consider only two variables: the sales price in British pounds sterling, and the year in which the flat was sold. For your convenience, the data set depicts this second piece of information in two different ways: the calendar year (2000 through 2011), and the number of years that had elapsed since 2000 when the flat was sold (0 through 11). When importing the data, remember to tell RStudio that there is a header row in the file.

Fit a simple group-wise model for price versus year, treating year as a categorical predictor. An important thing to remember here is that year is given as a number in the data set, and so R treat it by default as a numerical variable. If you want to override this behavior and get R to treat a numerical variable as a categorical variable, enclose the name of the variable in the `factor` command. For example, compare the default plots from the following two commands.

```
plot(Price~Year, data=oxford)
plot(Price~factor(Year), data=oxford)
```

In the first case, you should get a scatter plot, because R is treating Year as a quantitative predictor. In the second, you should get a boxplot, because R is treating year as a categorical predictor. You'll need to use the factor() command if you want to compute the mean Price by Year.

Describe the overall trend in the data, marshalling appropriate visual and numerical evidence (for example: a simple table of the groupwise means). Do you perceive any shortcomings of a simple group-wise model here? In particular, could you use the groupwise model to forecast what might happen next year? If so, how? If not, why not?

*(2) Austin food critics*

The data in "afc.csv" contains the following information about 104 restaurants scattered around central Austin:
*Name:* the name of the restaurant
*Neighborhood:* where the restaurant is located
*Type:* the type of food served at the restaurant
*FoodScore:* a numerical rating of the food (0–10) assigned by food critics from the Austin newspaper, with 10 being best.
*FeelScore:* a numerical rating of the atmosphere (0–10) assigned assigned by those same food critics, with 10 being best.
*Price:* average price of a meal at the restaurant, including tax, tips, and drinks

(A) Which are the two most expensive neighorhoods, on average? Which are the two cheapest? How did you judge this?

(B) Which seems to predict the price of a meal better: the food quality, or the atmosphere? How did you judge this?

(C) Now, using tools you've learned, compute a "food-adjusted value" measure for each restaurant: that is, a measure of the price of a meal at a restaurant that adjusts for the quality of the food one finds there, as judged by Austin food critics. Which are the two "best-value" neighborhoods, on average? Which are the two worst? Describe your procedure concisely, and present visual and/or quantitative evidence in support of your judgments.

*(3) The CAPM and a stock's "beta"*

If you've taken a finance class that covered anything about asset-pricing theory, you may recall that the Capital Asset Pricing Model (CAPM) assumes that the rate of return on an individual stock is linearly related to the rate of return on the overall stock market. Roughly speaking, this means that each stock's returns follows a linear regression model! That is,

$$Y_t^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} X_t + e_t^{(k)} ,$$

where $Y_t^{(k)}$ is the return of an individual stock ($k$) in some given time period $t$; $X_t$ is the return of the entire stock market in that same time period; and $e_t^{(k)}$ is the residual for stock $k$ in that time period. The superscript $(k)$'s here are simply denoting the different stocks, while the subscript $t$'s are denoting the different time periods. The market ($X_t$) is a predictor common to all stocks.

The data set "marketmodel.csv" contains information on the daily percentage returns for the S&P 500 stock index, along with 6 individual stocks: Apple, Google, Merck, Johnson and Johnson, Wal-Mart, and Target. The data are from February 1, 2011 through January 30, 2012. The entries are interpretable as percentages—for example, looking at the first row, we see that the S&P 500 gained 1.595% in value on February 1, 2011, while Target lost 0.335% in value on that same day.

(A) Regress the returns for each of the 6 stocks individually on the return of S&P 500 (which is like $X_t$, the market return, in the equation above). Make a table that shows the ticker symbol, intercept, slope, and residual standard deviation for each of the 6 regressions. Which stock seems to be the most tightly coupled to the movements of the wider market?

(B) What do you notice about the intercepts? Are they mostly small, or mostly large? Interpret these intercepts in terms of whether any of the individual stocks appear to be outperforming the market, on average.[1]

(C) Does your estimate of the slope for Wal-Mart versus the S&P 500 agree (roughly) with the "beta" reported by Yahoo Finance?[2] If you notice a discrepancy, offer a possible explanation or two.

(D) Assess the evidence in the data for the following claim: "*Even after adjusting for their shared dependence on the broader market*, we should expect Wal-Mart's stock market returns to be most closely related to Target's returns than with any of the other four firms, because they are both large retailers."

[1] Note: these intercepts are often referred to as "alpha" rather than $\beta_0$ in the finance community, e.g.`http://en.wikipedia.org/wiki/Alpha_(investment)`.

[2] Check for yourself on: `http://finance.yahoo.com/q/ks?s=WMT`. The beta is in the right-most column, under "Trading Information." If you don't quite know how to interpret a stock's beta, do a bit of background reading, e.g. `http://en.wikipedia.org/wiki/Beta_(finance)`.

*Note to those who might enjoy playing around with stock-market data:* if you install the R package "fImport" and use the command `yahooSeries`, you can get data on any stock, for any time period, that you want! For example, I got this data using the following commands:

```
library(fImport)
mytickers=c("SPY", "AAPL", "GOOG", "MRK", "JNJ", "WMT", "TGT")
X = yahooSeries(mytickers, from = "2011-01-31", to = "2012-01-31", frequency="daily")
```

I then constructed the percentage returns from each "Adjusted Closing Price" column.