First Example Set

Matt Taddy – Chicago Booth
A Three Day Course in Applied Regression Analysis

1 Rent Data Exploratory Analysis

The `rent.csv` data (available on the course page) contains information on $Rent$ (in \$), number of $Rooms$ and $Bathrooms$, the year of construction ($YearBuilt$), square footage (100 $SqFt$), and existence of Air Conditioning (AC) or $Parking$ for $n = 696$ apartments in Chicago.

You are going to explore how these variables affect rent.

(i) Decide from looking at the data (give your reasons) which of the 2 or 3 level variables (AC, $Bathrooms$, and $Parking$) are most influential. Produce scatterplots for the other variables ($Rooms$, $YearBuilt$, and $SqFt$) and comment on what you see.

(ii) Produce a boxplot for the marginal distribution of $Rent$, and compare this to boxplots for conditional distributions for $Rent$ given each level of your chosen factor in (i), and for $Rent$ given the different numbers of $Rooms$. What can you say about the effect of these variables on rent?

(iii) Now investigate the effect of $SqFt$ on rent. First, do you see any problems in the scatterplot? Could you get a better model by ignoring some observations? Fit the line $Rent = b_0 + b_1 SqFt$.

(iv) What does b_1 tell you about the influence of $SqFt$? What would you say if asked to predict at $SqFt = 0$? Finally, what is your forecast $rent$ for my 1480 sq.ft. flat?

2 Newspaper Circulation

In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of 34 newspapers concerning their daily and Sunday circulations (in thousands). The data is available int the course website, file `newspaper.csv`.

(i) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggests a linear relationship between the variables? Do you think this is a plausible relationship?

(ii) Fit a regression line predicting Sunday circulation from daily circulation, and obtain the 95% confidence intervals for β_0 and β_1

(iii) Is there a significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Indicate what hypothesis you are testing and your conclusion.

(iv) Suppose that you are proposing to add a Sunday edition of a newspaper with weekday circulation of 200,000 copies. What would you tell advertisers is the expected Sunday circulation? What is the standard deviation of this expectation? What would you say when they ask you to predict a likely range of possible Sunday circulation numbers?

3 A Classic Height Regression

The first regression was carried out in 1885 by Sir Francis Galton who wanted to see the effect of parents' height on the height of their children. We will reconsider this problem and take a look at a sample of heights for Chicago MBA students and their parents. The data is in file `MBA-hgt.csv`, and the variables are: `Female` - One if female, zero if male, `SHGT` - Student's height in inches, `MHGT` - Mother's height in inches, and `FHGT` - Father's height in inches.

(i) Find the correlations for each possible input/output combination.

(ii) Create plots of student's height vs each parent's height. Would it be better to fit different lines for male and female students?

(iii) Use your results from (ii) to choose the two best models (one to predict male heights and one to predict female heights). What are the least squares lines for each model?

(iv) What is R^2 for each regression? What are the residuals for prediction of your personal height(s) using the regression?

4 Crime Statistics

In this question we consider crime-related and demographic statistics for 47 US states in 1960 (available as `crime.csv`). The data were collected from the FBI's Uniform Crime Report and other gov-

ernment agencies to determine how the Crime Rate (CR, offenses per million population) depends on socio-economic variables: residents' average years of education (EDU), unemployment rate (UE), median income (INC), and the number of male residents per 1000 female residents ($PMALE$). The data also include the per capita state expenditure on policing ($POLICE$, in dollars).

(i) Present a visual summary of the data. How does the crime rate relate to our other variables?

(ii) What is the correlation between police expenditure and the crime rate? Do you think that police budget is a useful variable for understanding what *causes* crime?

(iii) Consider the regression of crime rate onto state median income. Is there a significant relationship between INC and CR? How do you interpret this result? Do you find it surprising?

(iv) Plot the residuals from your regression in (iii), and use this plot to comment on the model fit.

(v) A continental US state not in our sample had a median income of \$2800 in 1960, but the crime rate recordings were not considered accurate enough for inclusion. What is a 90% confidence interval for the unknown crime rate in this state? Is there anything disturbing about this interval?

5 Market Model Example

The Capital Asset Pricing Model (CAPM) for a given asset assumes the rate of return $R^s = (V_t^s - V_{t-1}^s)/V_{t-1}^s$ on a generic stock is linearly related to the rate of return (R^m) on the overall market as:

$$R_i^s = \alpha + \beta R_i^m + \epsilon_i$$

where the error term ϵ follow the assumptions of the SLR Model. The slope coefficient measures the sensitivity of the stock's rate of return to changes in the level of the overall market, and the intercept is market independent income.

For this problem, use the file `mktmodel.csv` from the course website. The dataset contains 60 monthly returns (from 1992 to 1996) of the S&P500 and 30 individual US stocks (labelled by ticker).

- Calculate the market correlation for each stock. Based on this information alone, which CAPM fit would yield the highest R^2? Can you give a practical reasoning for this?

- Estimate α and β for each stock and plot them against each other. Briefly describe the results.