Simple linear regression

Patrick Breheny

January 18

Introduction

- Today's lecture/lab is about fitting a regression line to a scatter plot of data, also known as *simple linear regression*
- This is interesting both by itself and as a precursor to multiple linear regression

Pearson's height data

- Statisticians in Victorian England were fascinated by the idea of quantifying hereditary influences
- Two of the pioneers of modern statistics, the Victorian Englishmen Francis Galton and Karl Pearson were quite passionate about this topic
- In pursuit of this goal, they measured the heights of 1,078 fathers and their (fully grown) sons
- Introducing standard regression notation, we have $n = 1,078$ pairs of observations $\{x_i, y_i\}$, in which x_i is the height of the father in family i, and y_i is the height of the son

Importing the data

- All the data sets for this class will be provided in a tab-delimited format
- In R, such files can be read in via

```
pearson <- read.delim("pearson.txt")
```

- In SAS, you can import the data through **File** $\rightarrow$ **Import Data**; when it asks you for the data source, select "Tab Delimited File (.txt)" from the drop-down menu

Plotting the data

- In R, the data can be plotted with either

```
plot(pearson$Father,pearson$Son)
```

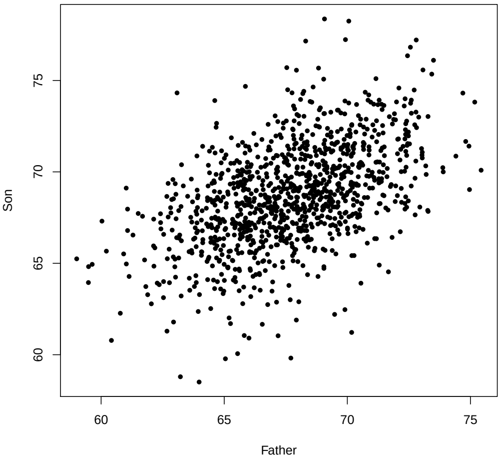
  or

```
attach(pearson)
plot(Father,Son)
```

- In SAS, the data can be plotted via

```
PROC SGPLOT DATA=Pearson;
  SCATTER X=Father Y=Son;
RUN;
```

Scatter plot of Pearson's height data



Observations about the scatter plot

- Taller fathers tend to have taller sons
- The scatter plot shows how strong this association is – there is a tendency, but there are plenty of exceptions

Simple linear regression

- Simple linear regression aims to draw a line through those points which
  - Approximates the average height of the sons, given the heights of their fathers
  - Can be used to predict a son's height, given the height of his father
  - Can be used to draw conclusions about the heredity of height
- The regression line, like all lines, has an equation of the form

$$y = \alpha + \beta x$$

Fitting the regression line

- However, the heights of fathers and sons clearly do not fall exactly on a line; there are *random errors*:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- Note that x_i and y_i are observed, while α, β, and $\{\epsilon_i\}$ are not
- The parameters of interest are α and β; *i.e.*, we are interested in obtaining the estimates $\hat{\alpha}$ and $\hat{\beta}$, which in turn determine the regression line

Fitted values and residuals

- Suppose we use the regression line to predict y_i
- The resulting prediction is called the *fitted value*:

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$$

(this quantity is also called the "predicted value", though this is potentially a little misleading, as you're not really "predicting" y_i, since you've already observed it)

- The amount by which each fitted value differs from the observed value y_i is called the *residual*:

$$r_i = y_i - \hat{\mu}_i$$

The method of least squares

- We want a regression line that fits the data well (*i.e.* does a good job of passing through the average y for a given x)
- Regression lines are fit by minimizing the *residual sum of squares*:

$$RSS = \sum_i r_i^2$$

(we will discuss the justifications for this in a moment)

- **Proposition:** The values $\{\hat{\alpha}, \hat{\beta}\}$ which minimize the residual sum of squares are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

Obtaining least squares estimates in R/SAS

- These estimates can be obtained via
  - In R:

    ```
    lm(Son~Father)
    ```
  - In SAS:

    ```
    PROC REG;
      MODEL Son = Father;
    RUN;
    ```
- They both yield the estimates $\hat{\alpha} = 33.9$, $\hat{\beta} = 0.514$

Reproducing these estimates manually

These estimates can also be obtained manually using the solution that we derived earlier:

```
x <- pearson[,1]
y <- pearson[,2]

xx <- x - mean(x)
yy <- y - mean(y)

beta <- sum(xx*yy)/sum(xx^2)
alpha <- mean(y) - beta*mean(x)
```

Adding the regression line to the plot

Let's add the regression line to the plot:

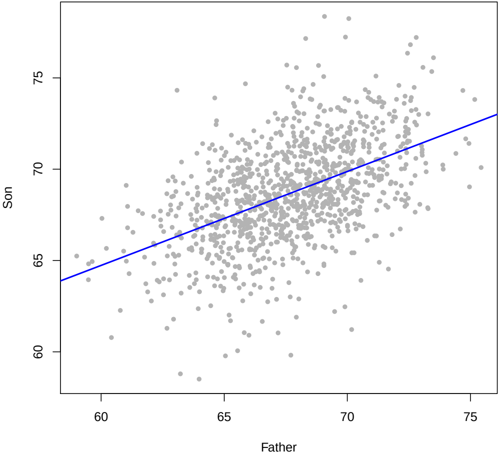
- In R:

```
abline(alpha,beta)
```

- In SAS:

```
PROC SGPLOT;
  REG X=Father Y=Son;
RUN;
```

The regression line for Pearson's height data



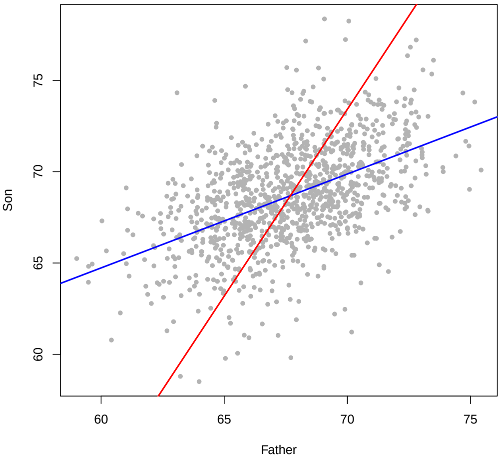
There are two regression lines

- When we regress y on x, we are predicting y based on x – not the other way around
- This matters, because the outcome and explanatory variables are not interchangeable with respect to estimation:

$$\frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \neq \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(y_i - \bar{y})^2}$$

- We obtain different lines, and different predictions, depending on which variable is chosen as the outcome

The two regression lines



Justifications for least squares

- The original justification for least squares was that it was convenient to work with: $\sum r_i^2$ is differentiable, whereas, for example, $\sum_i |r_i|$ is not
- An additional justification is that if y_i follows a normal distribution, $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimates:

$$l(\alpha, \beta) \propto -\sum (y_i - \alpha - \beta x_i)^2$$

- A further justification, which we discuss in more detail later in the course, is that the method of least squares produces the best (*i.e.* minimum variance) linear unbiased estimator of α and β

Regression and correlation

- There is an intimate connection between regression and correlation
- Given the regression line, you can calculate the correlation, and vice versa
- Letting r denote the correlation coefficient and $s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$, we have

$$\hat{\beta} = r\frac{s_y}{s_x}$$

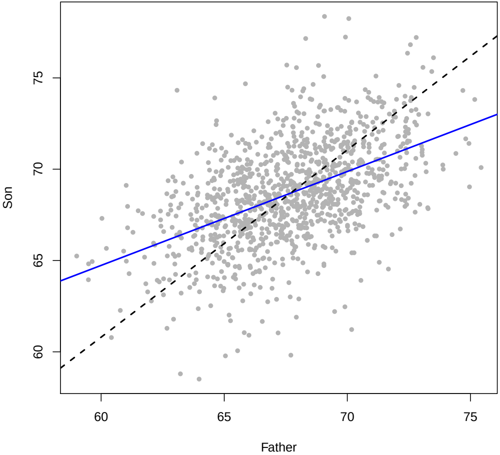
Regression and correlation (cont'd)

- Furthermore, substituting this expression into $y = \alpha + \beta x$, we have

$$\frac{y - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}$$

- This neat little equation summarizes quite nicely the interplay between regression, correlation, and standardized variables
- Also note that because $r \in [-1, 1]$, this equation places a bound on the slope of the regression line

The regression and SD lines



Simple vs. multiple regression

- A "simple" regression equation has on the right hand side an intercept, a single explanatory variable, and single slope
- A *multiple regression* equation has several explanatory variables, each with its own slope
- Before we study multiple regression, we will need to develop some matrix algebra tools, which is what we will do in our next lecture