STA 111: Probability & Statistical Inference

Lecture Eighteen – Simple Linear Regression and Least Squares
D.S. Sections 11.1 & 11.2

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Correlation
- Method of Least Squares
- Regression
- Examples
- Recap

Introduction

- In the last lecture we talked about testing for independence between two categorical variables using the Chi-squared test.
- Today we will talk about predicting one variable from another when observations are observed in pairs, taking us back to a formal discussion of correlation.
- We will also learn about fitting a line to such data using least squares and regression.

Correlation

Recall that we defined correlation between two random variables x and y as:

$$Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

Formally, **Correlation** is a measure of the strength of the linear association between two variables.

Consider an early example which studied the relationship between the height of fathers and the height of sons. Clearly, tall fathers tend to have tall sons, and short fathers tend to have short sons. If the father's height were a perfect predictor of the son's height, then all father-son pairs would lie on a straight line in a scatterplot.

We will review two methods (least squares and regression) of fitting a line to the points in a scatterplot in a bit.

Correlation

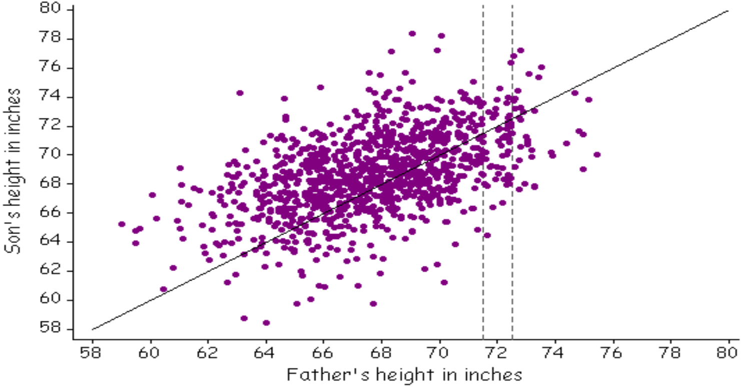
The sample correlation coefficient r estimates the true correlation and measures the strength of the linear association between x and y values in a **scatterplot**. If the absolute value of the correlation is near 1, then knowing one variable determines the other variable almost perfectly (if the relationship is linear).

- r lies between -1 and 1, inclusive.
- r equals 1 iff all points lie on a line with positive slope.
- r equals -1 iff all points lie on a line with negative slope.
- non-zero r does not imply a causal relationship; that is, we can't say for sure that x causes y or y causes x.

The square of the correlation is sometimes called the **coefficient of determination**. It is the proportion of the variation in y that is explained by knowledge of x. This is usually used in the simple case of two variables.

Correlation

In our height example, we clearly have a strong positive correlation between the height of fathers and the height of sons.



Correlation

To estimate the true correlation coefficient, define

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

Note: observe that if divided by $n-1$, these are the sample versions of the variances and the covariance. So there's no need to memorize.

Then the sample correlation is

$$r = \frac{\hat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{SS_{xy}\,/\,n-1}{\sqrt{(SS_{xx}\,/\,n-1)(SS_{yy}\,/\,n-1)}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}.$$

Correlation

One can show that the coefficient of determination r^2 is the proportion of the variance in y that is explained by knowledge of x.

Correlations are often high when some factor affects both x and y.

- GPA and SAT scores are both affected by IQ.
- number of hours spent listening to Taylor Swift and GPA are both affected by lifestyle.

It is hard to argue that correlation implies causation. GPA does not cause SAT, and Taylor Swift does not hurt GPA. Sometimes, the link might be causal. Hours of study are probably correlated with GPA, and the link is likely causal.

Method of Least Squares

Suppose we are interested in describing the relationship between two variables x and y, we might simply fit a straight line to the data. This is relatively easy and straightforward to do.

Recall that the algebraic equation for a line is

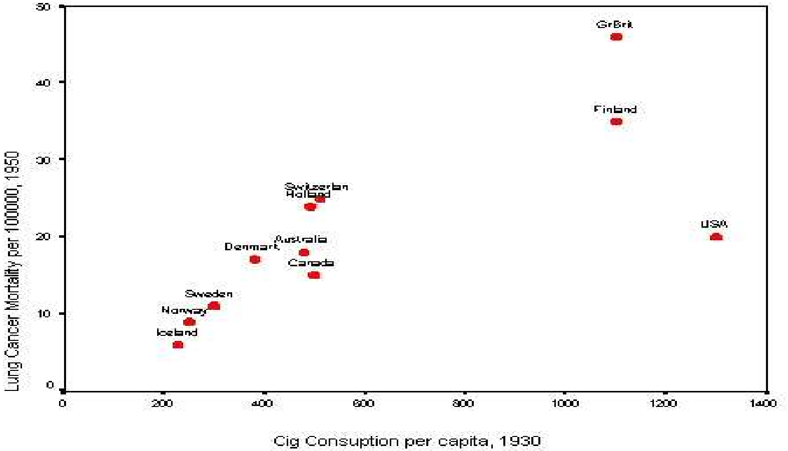
$$y = \beta_0 + \beta_1 x$$

where β_o is the intercept and β_1 is the slope. If we can estimate the values of β_0 and β_1, then predicting a value for y given a new value for x is again easy by simply plugging in the values.

But how exactly should we fit the line if we have data that don't already lie on a straight line?

Method of Least Squares

Example 1: The scatterplot below shows lung cancer rate against the proportion of smokers for 11 different countries (Doll, 1955).

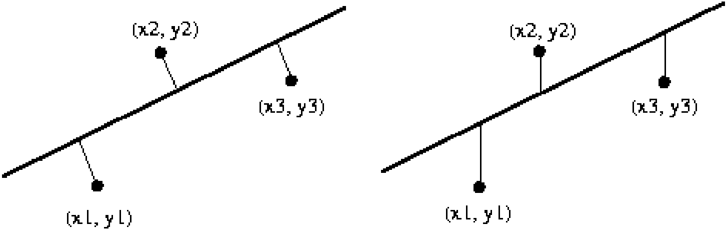


Method of Least Squares

In this example, it is not quite possible to fit a line that accommodates or fits all points reasonably well.

One method of constructing a straight line to fit the data is known as **method of least squares**. The method of least squares tries to fit the line that minimizes the sum of the squared deviations from each point to the line, where deviation is measured in the **vertical** direction.

Note: This does **not** measure deviation as the perpendicular distance from the point to the line.



Method of Least Squares

Thus, to find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the coefficients in the equation of a line, we need to get the values that minimize the sum of the squared vertical distances. The sum of the squared vertical distances is

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

So take the derivative of $f(\beta_0, \beta_1)$ with respect to β_0 and β_1, set these equal to zero, and solve. One finds that:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \textbf{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

Thus, the line defined by $y = \hat{\beta}_0 + \hat{\beta}_1 x$ is the least squares line.

Regression

Regression terminology:

- The **response** variable is labeled y, the variable we are often interested in explaining or predicting. This is sometimes called the **dependent** variable.
- The **explanatory** variable is labeled x, the variable we think/hope explains y. This is sometimes called the **independent** variable, or the **covariate**.

The regression model assumes that the observed response is :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the ϵ_i's are random error (due to genetics, measurement error, etc.). We assume that these errors are independent and normal with mean 0 and unknown sd σ_ϵ. We assume the x_i's are measured without error.

Regression

In regression, we assume that the y values are observed values of a collection of random variables. In this case, it turns out that assuming a normal distribution makes the maximum likelihood estimates of β_0 and β_1 the same as what we had for the method of least squares.

In fact, we have that $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2\right)$. By MLE, it turns out that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \qquad \hat{\beta}_1 = SS_{xy} / SS_{xx}$$

just like we had before using the method of least squares. **Can we find $\mathbb{V}[\hat{\beta}_0]$ and $\mathbb{V}[\hat{\beta}_1]$?**

By the way, how should we interpret β_0 and β_1? β_0 is usually the average of y when x is zero, thus it is only meaningful when x can be zero (think about when x represents the height of a person). For β_1, every unit increase in x corresponds to an increase in y by β_1.

Regression

Regression predicts the average value of y for a specific value of x.

This is not the same as saying that an individual value lies on the line. In fact, an individual is often likely to be far from the line.

Example 2: Suppose we regress exam grade against number of hours of study per week. Assume the regression line is $y = 20 + (7x)$. *(Is this reasonable? When would it break down?)*

- If you are advising a class on how to study, you tell them that the regression model says they should work for 10 hours a week on the material if they want to score a 90.
- John complains that he studied conscientiously for 10 hours each week, but his exam grade was only 40.

Is the result for John unexpected? We can calculate how likely the event is since we have assumed a normal distribution.

Regression

To decide this, we first need to estimate σ_ϵ, which says how far from the line an observation is likely to be. To do this, we look at the sample standard deviation of the **residuals**.

The residuals are our estimates of the random errors, so $\{\hat{\epsilon}_i = y_i - \hat{y}_i\}$, where $\hat{y}_i$ is the value predicted by the regression line. The difference is the estimated error for the ith observation.

Thus an unbiased estimate is

$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

and we divide by $n-2$ because we have estimated two parameters β_0 and β_1, in calculating the residuals, and thus we "used up" a quantity of information equivalent to two observations.

Regression

Suppose the sample standard deviation of the residuals around the line that predicts score from the number of hours of study is 38. What is the probability that John would get a grade of 40 or below?

The predicted value for John is $Y = 20 + (7 \times 10) = 90$. The standard deviation around that is 38.

Assuming that the deviations from the line are normal, then the chance of John getting a 40 or less is the area under the curve and to the left of a normal distribution with mean 90 and sd 38.

So the z-transformation $z = (40 - 90)/38 = -1.315$. From the table, he has a 9.68% chance of this low a grade.

We are able to do this because under the regression assumptions, an individual with explanatory variable x_i has response value y_i where $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_\epsilon)$. So we can estimate the probability of particular outcomes using $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_\epsilon$.

Regression

Be aware that regressing y as a function of x gives a different regression line than regressing x against y. If your best estimate of the weight of a man who is 5'10" is 170 pounds, that **does not** mean that the best estimate of the height of a man who weighs 170 pounds is 5'10".

The mathematical model for regression assumes that:

1. Each point (x_i, y_i) in the scatterplot satisfies:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

   where the ϵ_i's have a normal distribution with mean zero and (usually) unknown standard deviation.
2. The errors ϵ_i have nothing to do with one another; they are independent.
3. The X_i values are measured without error. Thus all the error occurs in the y, and we do not need to minimize perpendicular distance to the line.

Recap

Today we learned about fitting lines to observed data when we only have two variables using either regression or method os least squares.

In the next lecture, we will extend the regression idea to multiple explanatory variables.