17.0 Linear Regression

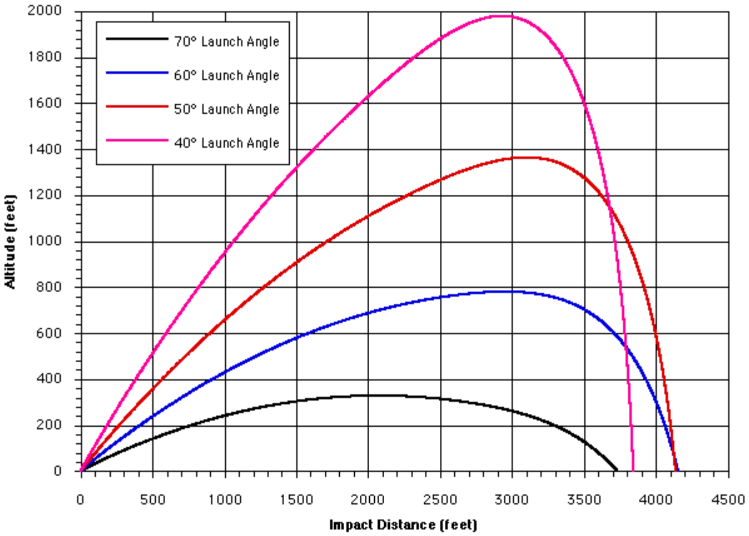
- Answer Questions
- Lines
- Correlation
- Regression

17.1 Lines

The algebraic equation for a line is

$$Y = \beta_0 + \beta_1 X$$

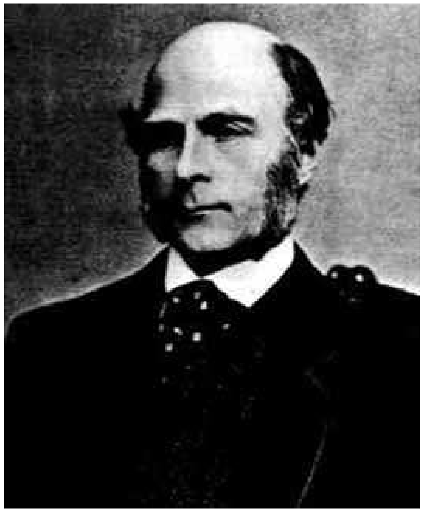
The use of coordinate axes to show functional relationships was invented by René Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.



17.2 Correlation

Sir Francis Galton explored Africa, invented eugenics, studied whether ships that carried missionaries were less likely to be lost at sea, pioneered birth-and-death models and meteorology, and was Charles Darwin's cousin.

He also was the first to conceive of linear regression (although he did not have the mathematical skill to develop the formulae, and got a friend of his at Cambridge to do the derivations).



Correlation is a measure of the strength of the linear association between two continuous variables. An early example studied the relationship between the height of fathers and the height of sons.

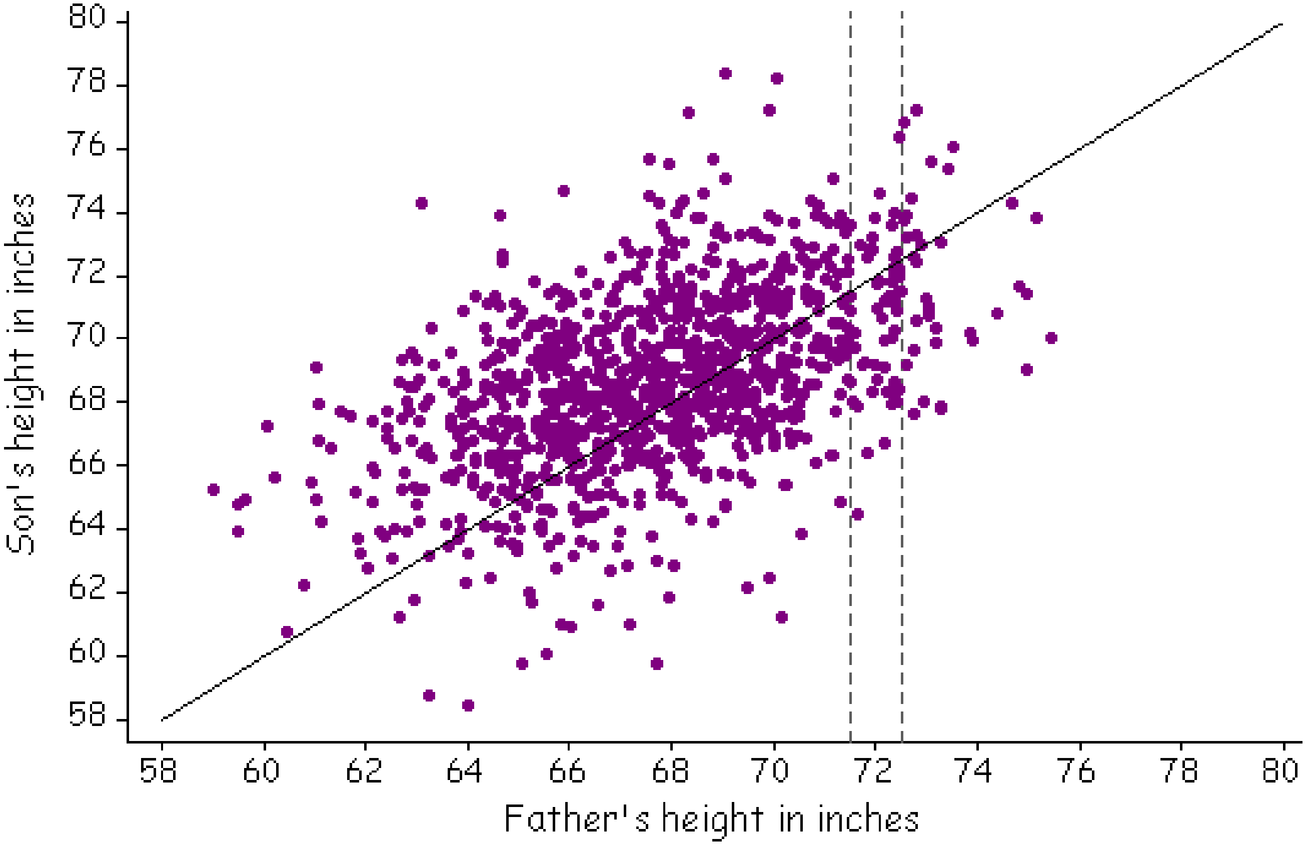
Clearly, tall fathers tend to have tall sons, and short fathers tend to have short sons. If the father's height were a perfect predictor of the son's height, then all father-son pairs would lie on a straight line in a scatterplot.

Regression fits a line to the points in a scatterplot. The term comes from the father-son example. An exceptionally tall father tends to have sons that are shorter than himself; an exceptionally short father tends to have sons that are taller than himself. Thus the sons' height tend to "regress towards the mean".

The sample correlation coefficient r 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.

The square of the correlation is called the **coefficient of determination**. It is the proportion of the variation in Y that is explained by knowledge of X.


Son's height in inches
80
78
76
74
72
70
68
66
64
62
60
58
58
60
62
64
66
68
70
72
74
76
78
80
Father's height in inches

To estimate the true correlation coefficient, define

$$\begin{aligned} SS_{xx} &= \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 \\ SS_{yy} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 \\ SS_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}. \end{aligned}$$

Note: 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

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}.$$

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.

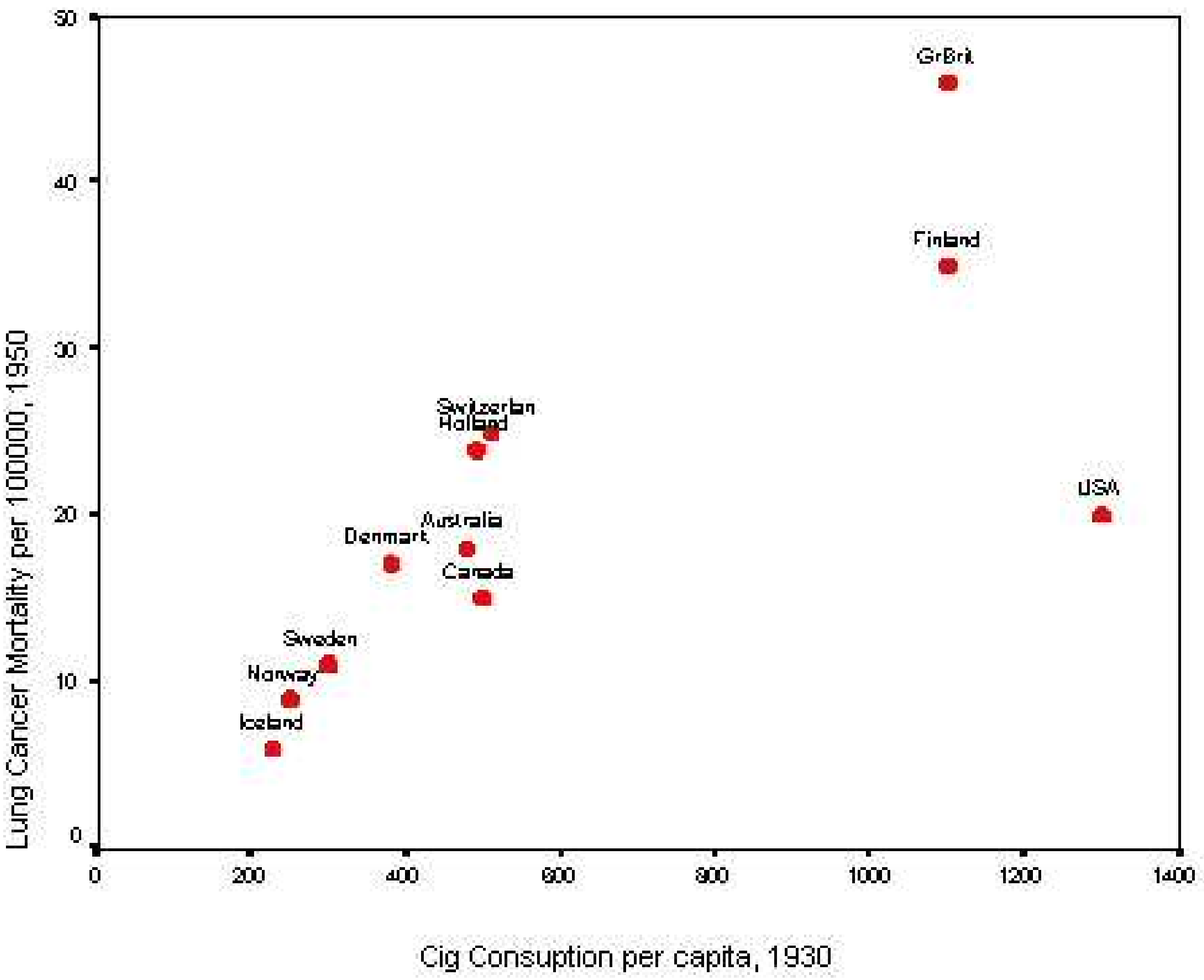
- GPA and SAT scores are both affected by IQ.
- number of hours spent listening to Rob Zombie and GPA are both affected by lifestyle.

It is hard to argue that correlation implies causation. GPA does not cause SAT, and Rob Zombie does not hurt GPA. But sometimes, there might be a causal link. Hours of study are probably correlated with GPA, and it seems likely to be causal.

∞

Ecological correlations occur when X or Y or both is an average, proportion, or a percentage for a group. Here causation is especially difficult to show.

The original link between smoking and lung cancer was an ecological correlation (Doll, 1955). The scatterplot showed the lung cancer rate against the proportion of smokers for 11 different countries.


GrBrit
Finland
Switzerlan
Holland
USA
Australia
Denmark
Canada
Sweden
Norway
Iceland
50
40
30
20
10
0
0
200
400
600
800
1000
1200
1400
Lung Cancer Mortality per 100000, 1950
Cig Consuption per capita, 1930

17.3 Regression

Regression terminology:

- The response variable is labeled Y. This is sometimes called the **dependent** variable.
- The explanatory variable is labeled X. This is sometimes called the **independent** variable, or the **covariate**.

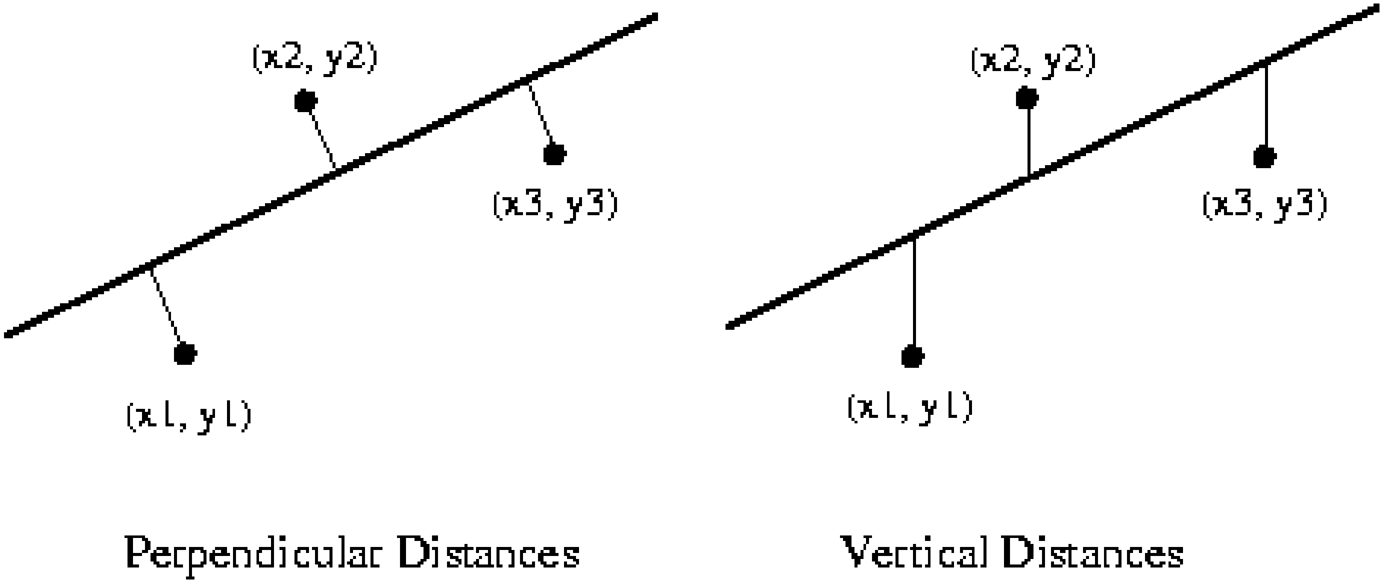
The regression model assumes that the observed response is :

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the ϵ_i is 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 are measured without error.

Regression tries to fit the "best" straight line to the data. Specifically, it fits 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.



Perpendicular Distances

Vertical Distances

How does one find the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the coefficients in the regression equation? We need to get the values that minimize the sum of the squared vertical distances. (Gauss, of course.)

The sum of the squared vertical distances is

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2.$$

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:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}; \qquad \hat{\beta}_1 = SS_{xy}/SS_{xx}.$$

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.

For example, suppose we regress exam grade against number of hours of study per week. Assume the regression line is $Y = 20 + 7 * X$. (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.
- Beavis complains that he studied conscientiously for 10 hours each week, but his exam grade was only 40.

Is the result for Beavis unexpected?

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 the $\{\hat{\epsilon}_i = y_i - \hat{y}_i\}$, where $\hat{y}_i$ is the value predicted by the regression line. The difference is the estimated error for the ith observation.

Thus

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

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

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 Beavis would get a grade of 40 or below?

The predicted value for Beavis is $Y = 20 * 7 * 10 = 90$. The standard deviation around that is 38.

Assuming that the deviations from the line are normal, then the chance of Beavis 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 $z = (40 - 90)/38 = -1.315$. From the table, he has a 9.68% chance of this low a grade.

Under the regression assumptions, an individual with explanatory variable x_i has response value Y_i where Y_i is $N(\beta_0 + \beta_1 x_i, \sigma_\epsilon)$. So we can estimate the probability of particular outcomes using $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_\epsilon$.

Be aware that regressing weight as a function of height gives a different regression line than regressing height against weight. 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 **regression fallacy** mistakenly argues that there is some effect or force that causes sons to be more average than their fathers. In fact, this is only the natural operation of random chance. Consider scores on a first and second exam, and also the father-son height example.

What can you say about the performance of baseball players in the first and second halves of the season? Or stock-traders, or new employees?

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 have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors ϵ_i have nothing to do with one another. A large error on the first observation does not tend to be followed by another large error on the second observation, for example.

3. The X_i values are measured without error. (Thus all the error occurs in the vertical direction, and we do not need to minimize perpendicular distance to the line.)