17.0 Linear Regression

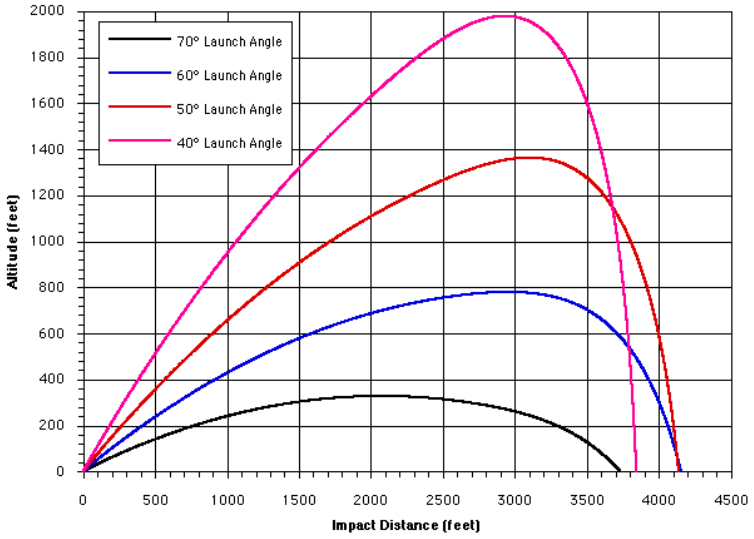
- Answer Questions
- Lines
- Correlation
- Regression

17.1 Lines

The algebraic equation for a line is

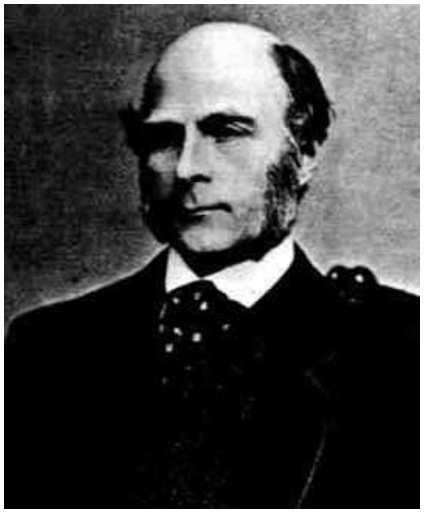
$$Y = \beta_0 + \beta_1 X$$

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.

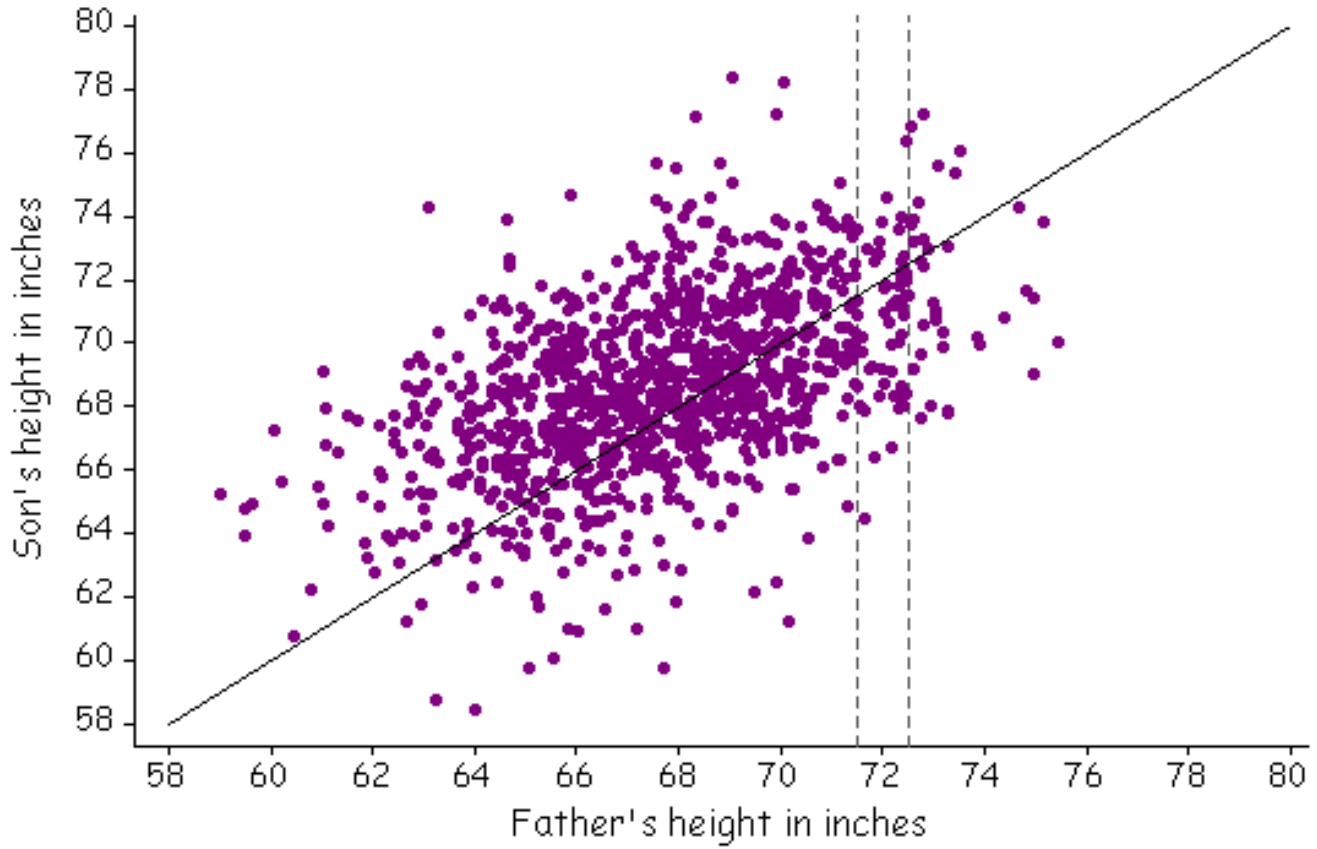
Clearly, tall fathers tend to have tall sons, and short fathers tend to have short sons. If the father's height were a perfect predictor of the son's height, then all father-son pairs would lie on a straight line in a scatterplot.

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.

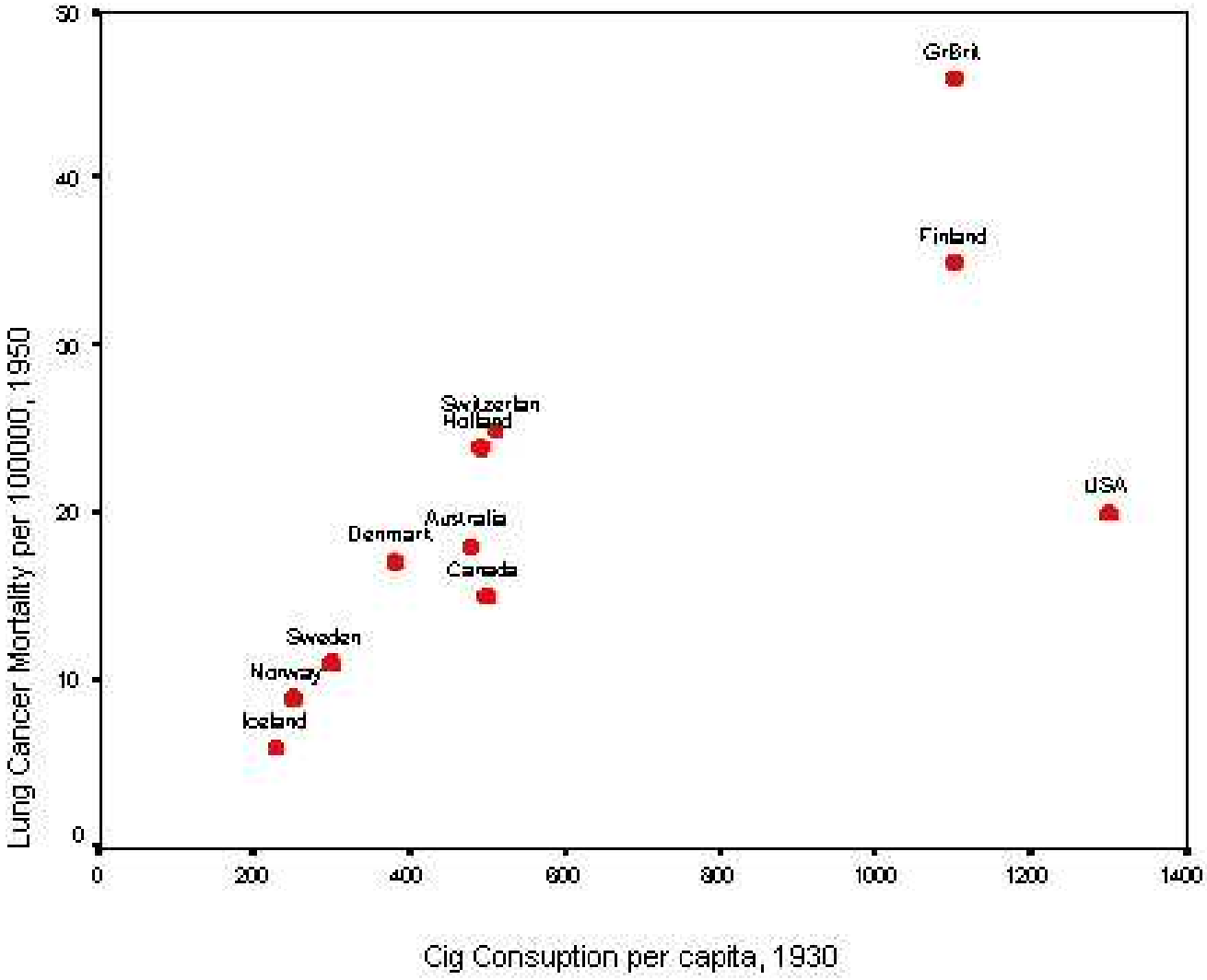
Correlations are often high when some factor affects both X and Y.

- 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.


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

17.3 Regression

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.

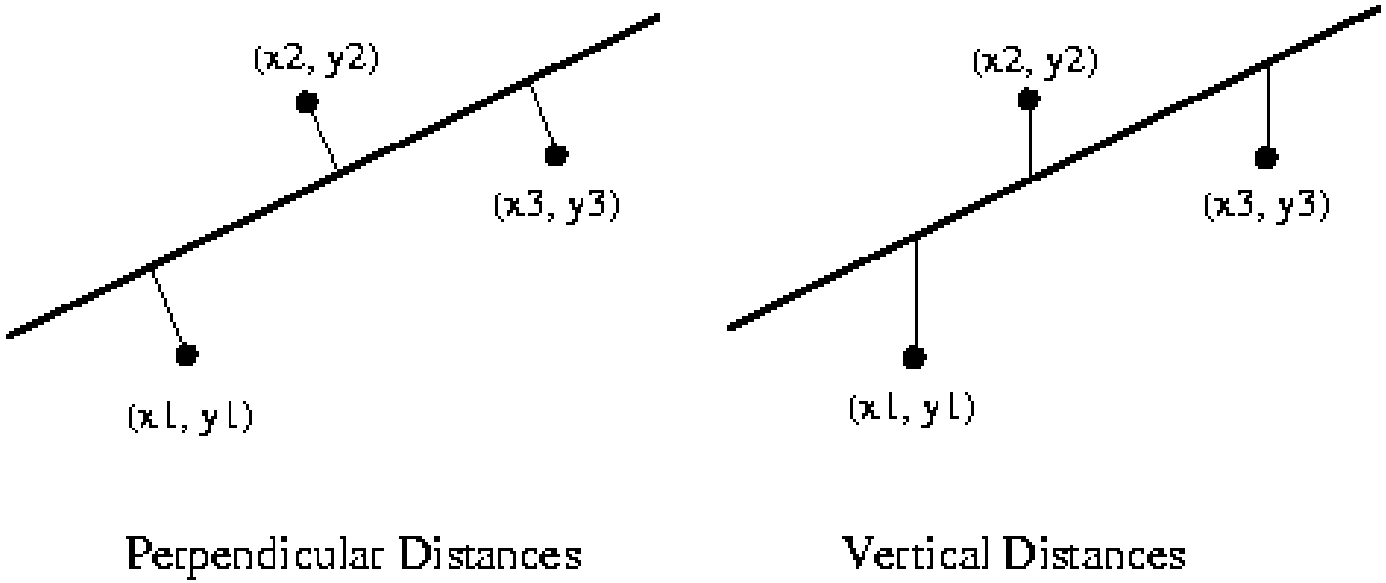
2. The errors ϵ_i are independent.

3. The X_i values are measured without error. (Thus all error occurs in the vertical direction.)

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**.

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:

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

The regression line predicts the average value of Y for a specific value of X. This is not the same as saying that an individual's value lies on the line. An individual is likely to be far from the line.

Under our assumptions, the distance of an individual from the regression line is normally distributed with mean 0 and standard deviation σ_ϵ. We do not know the true σ_ϵ, but we can estimate it from 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. This difference is the estimated error $\hat{\epsilon}_i$ for the ith observation. Then

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

Why do we divide by $n-2$?

Recall that $SS_x = \sum_{i=1}^{n} (X_i - \bar{X})^2$. Then a two-sided $100(1-\alpha)\%$ confidence interval on the location of the true regression line at x is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{SS_x}} t_{n-2,\alpha/2}.$$

A two-sided $100(1-\alpha)\%$ **prediction interval** on the location of an individual whose value of the explanatory variable is x is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{X})^2}{SS_x}} t_{n-2,\alpha/2}.$$

One-sided intervals are formed in the obvious way. If the sample size is large, you use the z-table instead of the t_{n-2}-table. And if you sample without replacement, you can use the FPCF to multiply $\hat{\sigma}_\epsilon$.

The $\hat{\sigma}_\epsilon$ is sometimes called root mean squared error or rmse.

Example 1.a: The DUS want to set 95% confidence intervals on the starting salaries (in thousands) of Duke statistics majors as a function of their GPA. Based on the 15 people who majored last year, we find $\beta_0 = 20$ and $\hat{\beta}_1 = 10$. The rmse was 4, $SS_x = 4$, $\bar{X} = 3.2$. What is the average starting salary for people who have GPAs of 3.5?

$$\begin{aligned} \hat{\beta}_0 &+ \hat{\beta}_1 x \pm \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{SS_x}} t_{n-2,\alpha/2} \\ &= 20 + 10 * 3.5 \pm 4 * \sqrt{\frac{1}{15} + \frac{(3.5-3.2)^2}{4}} * 2.160. \end{aligned}$$

The DUS is 95% confident that the mean starting salary is between $L =$ \$52.42K and $U =$ \$57.58K.

Example 1.b: Poindexter has a GPA of 3.5 and asks the DUS to set a 95% prediction interval on his starting salary.

$$\begin{aligned} \hat{\beta}_0 &+ \hat{\beta}_1 x \pm \hat{\sigma}_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{SS_x}} t_{n-2,\alpha/2} \\ &= 20 + 10 * 3.5 \pm 4 * \sqrt{1 + \frac{1}{15} + \frac{(3.5 - 3.2)^2}{4}} * 2.160. \end{aligned}$$

The DUS is 95% confident that his starting salary will be between $L =$ \$45.98K and $U =$ \$64.02K.

Note that for both intervals, the uncertainty increases as one tries to set intervals for x-values that are far from $\bar{X}$. This is reasonable—if there is a certain amount of "wigggle error" in the fitted regression line, the magnitude of the error increases with distance from $\bar{X}$.

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?