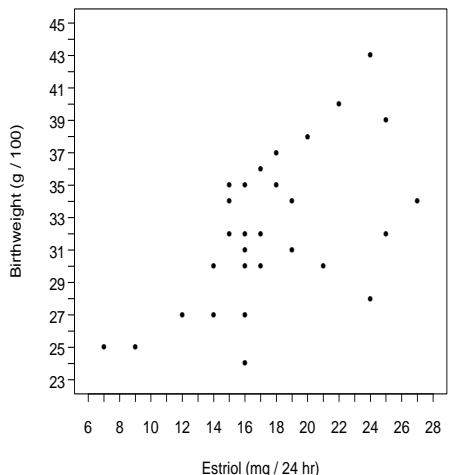


Linear Regression (Review)

Recall the study to relate birthweight to the estriol level of pregnant women.



Linear Regression - Summary of Assumptions:

- Values of X are fixed (at x).
- The outcomes of Y are normally distributed (independent) random variables with mean $\mu_{Y|x}$ and variance $\sigma^2_{Y|x}$.
- $\sigma^2_{Y|x}$ is the same for all x . This assumption of constant variability across all x values is known as homoscedasticity. ($\sigma^2_{Y|x} = \sigma^2$)
- The relationship between $\mu_{Y|x}$ and x is described by the straight line

$$\mu_{Y|X=x} = \alpha + \beta x$$

Linear Regression - Notation:

We have been using capital X or Y to denote a random variable and x or y to denote the values that the respective random variables could assume. In linear regression, we assume that values of X are fixed (not random). The notation above is consistent with this idea. Many books follow this notation. At the same time, many books (including your text) do not. We will follow the notation in your book.

The full linear regression model takes the following form

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

where ϵ is a normally distributed random variable with mean 0 and variance σ^2 .

(x_i, y_i) denote data values. There are 31 (x_i, y_i) pairs shown in the scatterplot.

How do we get estimates of β and α ? We would like the line to be as close to the data as possible.

The line is given in general by $\alpha + \beta x$. For a fixed x_i , the corresponding point on the line would be defined as

$$\alpha + \beta x_i$$

Consider measuring the distance from the data point y_i to the line.

$$y_i - \alpha - \beta x_i$$

The distance from the data point y_i to the value of the line for a given x_i is

$$y_i - \alpha - \beta x_i$$

We could sum up all of the differences. As always, it's a good idea to square the differences so that they don't cancel out.

S denotes the sum of squared differences (distances) between data points and the line

$$S = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

The least squares line, or estimated regression line, is the line that minimizes S.

Finding the least squares line means finding the values of α and β that minimize S.

With a little calculus, you can show that these values are

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

When we estimate α and β based on data $((x_i, y_i)$ pairs), the estimates, $\hat{\alpha}$ and $\hat{\beta}$, are called **estimated regression coefficients** or just **regression coefficients**.

Once estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β have been computed, the predicted value of y_i given x_i is obtained from the estimated regression line.

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i,$$

where \hat{y}_i is the **prediction** of the true value of y_i , for observation i , $i = 1 \dots n$. In our example, $n = 31$, and

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0.608$$

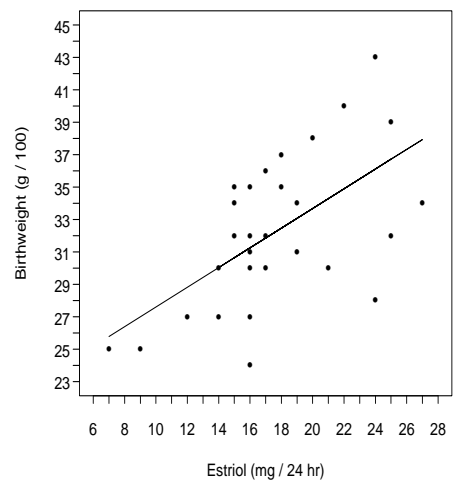
and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 21.52$$

The estimated regression line is

$$\hat{y}_i = 21.52 + 0.608x_i$$

The estimated regression line ($\hat{y}_i = 21.52 + 0.608x_i$) is shown along with the data below.



Linear regression inference

We would like to be able to use the least-squares regression line

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

to make inference about the population regression line.

$$\mu_{y|x} = \alpha + \beta x$$

We begin by noting that $\hat{\alpha}$ and $\hat{\beta}$ are point estimates of the population intercept and slope, respectively. As always, if we obtain different data values, the estimates will change.

Suppose we are interested in β and want to infer something about β using the information in $\hat{\beta}$. In particular, we want to know if there is evidence to reject the null $H_0 : \beta = 0$.

We also talked about the correlation coefficient. Note that R^2 denotes a quantity called the coefficient of determination and $R^2 = r^2$.

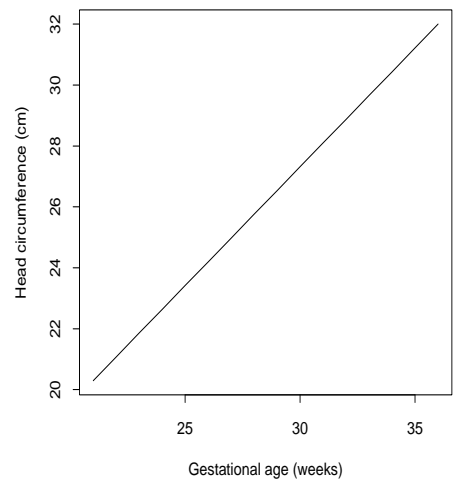
The statistic r has the following properties:

- $-1 \leq r \leq 1$.
- $r = 1$ iff $y = \hat{\alpha} + \hat{\beta}x$ for some $\hat{\beta} > 0$.
- $r = -1$ iff $y = \hat{\alpha} + \hat{\beta}x$ for some $\hat{\beta} < 0$.
- r remains the same under location-scale transforms of x and y .
- r measures the extent of linear association.
- r tends to be close to zero if there is no linear association between x and y .

If the linear regression model holds and we were to select repeated samples of size n from the underlying population of paired outcomes (x, y) and calculate a least squares line for each set of observations, the estimated values of α and β would vary from sample to sample. To test any hypotheses regarding the underlying population regression coefficients, we need to know (or estimate) the standard error of the estimates (we showed formulas for these standard errors last time).

We conducted a test of $H_0 : \beta = 0$ against $H_A : \beta \neq 0$ last time.

Another simple linear regression example: Consider a study that investigates the relationship between head circumference and gestational age for 100 infants. Using techniques from linear regression, a regression line was fit to the data. The line is shown below.



For that data, $\hat{\alpha} = 3.9143$, $\hat{\beta} = 0.7801$, and so

$$\hat{y} = 3.9143 + 0.7801x.$$

In the context of this example, we know that the prediction of the mean head circumference at age 29 weeks is given by \hat{y} .

Question: How accurate is this prediction ?

Answer: It depends on whether we are making predictions for the mean value of all infants that are 29 gestational weeks or one specific infant at age 29 weeks.

The first answer might be useful to a researcher interested in the relationship between head circumference and age at 29 weeks over a large population of infants; the second answer might be useful to an MD interested in assessing the head circumference of a particular infant.

We would like to construct a confidence interval around the true mean for a fixed value of x ($E[y|x] = \mu_{y|x}$) and a confidence interval (most commonly called a prediction interval) for a future value (new value) of y , again for a fixed x .

A $100(1 - \alpha^*)$ confidence interval for the mean value of y for a fixed x ($\mu_{y|x}$) is:

$$\left(\hat{y} - t_{n-2, \alpha^*/2} \hat{s}e(\hat{y}), \hat{y} + t_{n-2, \alpha^*/2} \hat{s}e(\hat{y}) \right)$$

where \hat{y} is the predicted mean of the normally distributed outcomes. The standard error of \hat{y} is given by:

A $100(1 - \alpha^*)$ confidence interval for a (predicted) future value of y (denoted by \tilde{y}) for a given x is:

$$\left(\tilde{y} - t_{n-2, \alpha^*/2} \hat{s}e(\tilde{y}), \tilde{y} + t_{n-2, \alpha^*/2} \hat{s}e(\tilde{y}) \right)$$

The standard error of \tilde{y} is always bigger than the standard error of \hat{y} since an extra source of variability is being considered.

Using these formulas, we could show that for gestational age fixed at 29 weeks, a 95 % confidence interval for the mean value of y is

$$(26.23, 26.85)$$

Similarly we could show that for gestational age fixed at 29 weeks, a 95 % prediction interval for an individual value of y is

$$(23.38, 29.70)$$

Using this interval, we can identify individuals that look unusual.

Nenana Ice Classic - Alaska's coolest lottery

See <http://www.nenanaaiceclassic.com>

