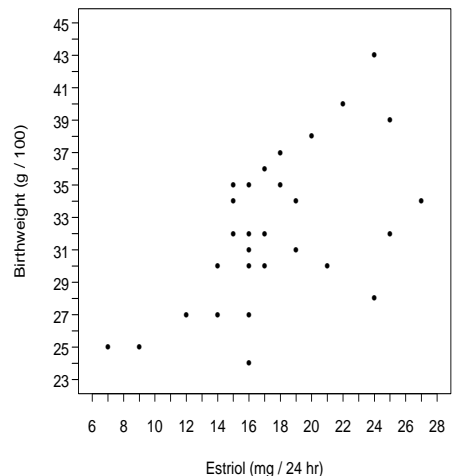


Linear Regression

Consider a study to relate birthweight to the estriol level of pregnant women. The data is below.

i	Estriol (mg / 24 hr)	Weight (g / 100)	i	Estriol (mg / 24 hr)	Weight (g / 100)
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Scatter plot of data from the study.



Important features of the scatter plot:

- Two variables are associated with each unit.
- There appears to be a rough trend or relationship between the two variables.
- This relationship is not exactly precise in that there exists substantial variation or scatter.

Important questions concerning the scatter plot:

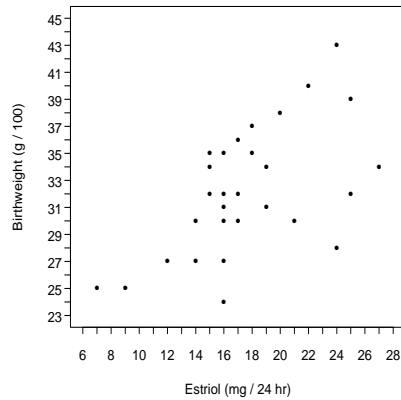
- Can we summarize the relationship with a linear equation (line) ?
- Is the apparent trend statistically significant ?
- Can we characterize the variability about a summary line ?

Linear regression is a statistical method used to address these questions.

The simple linear regression model

A simple linear regression model is a summary of the relationship between a **dependent variable** (or **response variable**) Y and an **independent variable** (or **covariate variable**) X .

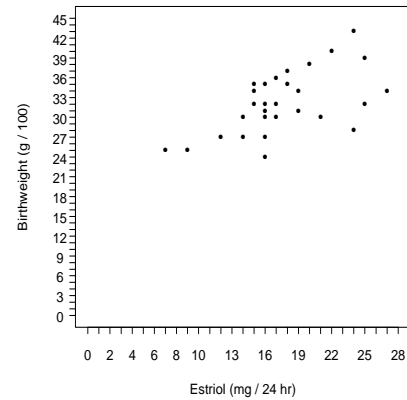
Y is assumed to be a random variable while, even if X is a random variable, we condition on it (assume it is fixed). Essentially, we are interested in knowing the behavior of Y given we know $X = x$.



Let $E[Y|x]$ represent the expected value of Y given $X = x$. We see that (on average) Y is increasing with X . To describe this relationship, I could define

$$E[Y|x] = \mu_{Y|x} = \beta x$$

where $\beta > 0$.



I could define

$$E[Y|x] = \mu_{Y|x} = \beta x$$

where $\beta > 0$; but note that when $x = 0$, $E[Y|X = 0] = 0$, which is not what we see in the data. To account for this (non-zero intercept), I could define

$$E[Y|x] = \mu_{Y|x} = \beta x + \alpha$$

where $\beta > 0$ and $\alpha > 0$.

$$E[Y|x] = \mu_{Y|x} = \beta x + \alpha$$

This line is the **population regression line**.

α and β are the **population regression coefficients** (coefficients).

α is the y-intercept of the line ($E[Y|X = 0]$). β is the slope of the line. It gives the change in the mean value of Y that corresponds to a one unit increase in X . If $\beta > 0$, the mean increases as X increases; if $\beta < 0$, the mean decreases as X increases.

Note that even if the population regression line accurately describes the relationship between the two variables, individual measurements of these variables will not necessarily fall on that line.

Note that even if the population regression line accurately describes the relationship between the two variables, individual measurements of these variables will not necessarily fall on that line.

To account for this, an error term (ϵ) which represents the variance of responses Y conditional on x is introduced.

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 .

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 (capital Y , lowercase x) 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.

You have 31 (x_i, y_i) pairs. Recall these questions:

- Can we summarize the relationship with a linear equation (line) ?
- Is the apparent trend statistically significant ?
- Can we characterize the variability about some line ?

Summarizing the relationship with a linear equation.

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

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

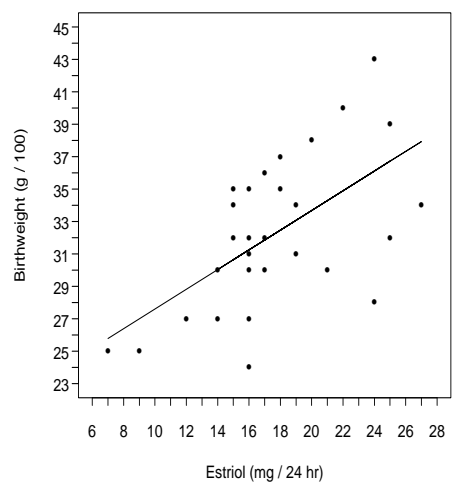
and

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

The estimated regression line is

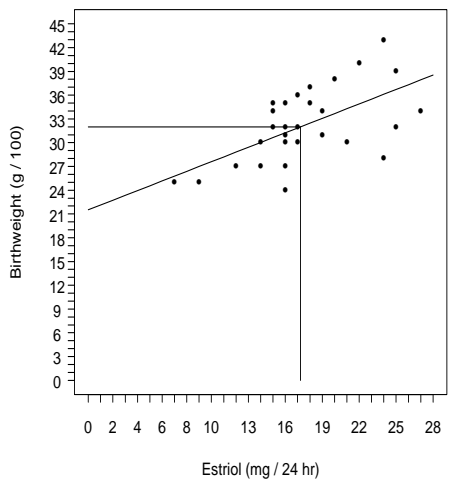
$$\hat{y}_i = 21.52 + 0.608x_i$$

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



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

Note that $\frac{1}{n} \sum_{i=1}^n x_i = 17.22$ and $\frac{1}{n} \sum_{i=1}^n y_i = 32$. Also note that when $x_i = 0$, $\hat{y}_i = 21.52$.



You have 31 (x_i, y_i) pairs. Recall these questions:

- Can we summarize the relationship with a linear equation (line) ?
- Is the apparent trend statistically significant ?
- Can we characterize the variability about some line ?

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 $H_0 : \beta = 0$.

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.

It can be shown that

$$se(\hat{\beta}) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

and

$$se(\hat{\alpha}) = \sigma_{y|x} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Note that the standard errors of the estimated coefficients $\hat{\alpha}$ and $\hat{\beta}$ depend on $\sigma_{y|x}$, which is usually unknown. We can estimate it using

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

Note that the difference between the true y_i and the predicted value of y_i , denoted by \hat{y}_i , is

$$y_i - \hat{y}_i$$

This difference is called the **residual value**.

The estimate of the standard deviation $s_{y|x}$ is

$$s_{y|x} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}}$$

Correlation Coefficient

In simple linear regression analysis, we saw that the estimate slope of the regression line $\hat{\beta}$, is a measure of the linear association between a dependent variable y and an independent variable x .

However, the value $\hat{\beta}$ depends on which variable we declare the independent variable, and which variable we declare the dependent variable. Also, $\hat{\beta}$ depends on the dispersion of the variables.

A measure of the linear association between two variables X and Y which is independent of these concerns is the *Pearson correlation coefficient* r .

It turns out that r is related to the regression coefficient $\hat{\beta}$ by,

$$r = \hat{\beta}(s_x/s_y) \text{ or } \hat{\beta} = r(s_y/s_x)$$

The correlation coefficient r is computed from the sample and estimates the population correlation coefficient ρ .

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of the x and y values, respectively.

We can now test $H_0 : \beta = 0$ against $H_A : \beta \neq 0$.

Note that if the null is true, then

$$t = \frac{\hat{\beta}}{se(\hat{\beta})}$$

follows a t distribution with $n - 2$ degrees of freedom.

In our example,

$$se(\hat{\beta}) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = 0.147$$

and

$$t = \frac{0.608}{0.147} = 4.14$$

Note that from table 5, $P(T_{29} \geq 3.659) = 0.0005$

We reject the null (at $\alpha = 0.01$ or $\alpha = 0.05$).

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 .