Linear Regression and Correlation

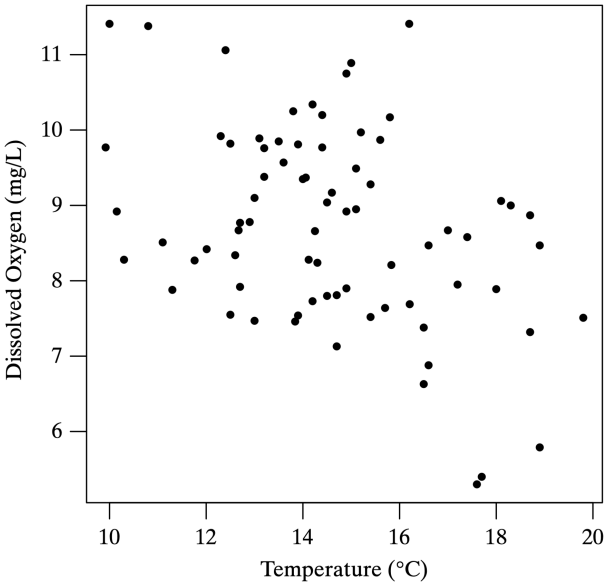
In this chapter we discuss some methods for analyzing the relationship between two quantitative variables, X and Y. Linear regression and correlation analysis are techniques based on fitting a straight line to the data.

Example: dissolved oxygen

The level of dissolved oxygen in a river is one measure of the overall health of the river. Researchers recorded water temperature (°C) and level of dissolved oxygen (mg/L) for 75 days at Dairy Creek in California. The figure below shows a scatterplot of the data, with

$$Y = \text{level of dissolved oxygen (mg/L) plotted against } X = \text{water temperature } (^\circ\text{C}).$$

The scatterplot suggests that higher water temperatures (X) are associated with lower levels of dissolved oxygen (Y).



The Correlation Coefficient

Suppose we have a sample of n pairs for which each pair represents the measurements of two variables, X and Y. If a scatterplot of Y versus X shows a general linear trend, then it is natural to try to describe the strength of the linear association. We will learn how to measure the strength of linear association using the correlation coefficient.

Example: length and weight of snakes

In a study of a free-living population of the snake Vipera bertis, researchers caught and measured nine adult females. Their body lengths and weights are shown and displayed as a

scatterplot in the following table and figure, separately. The number of observations is $n = 9$.

Table 12.2.1

	Length X (cm)	Weight Y (g)
	60	136
	69	198
	66	194
	64	140
	54	93
	67	172
	59	116
	65	174
	63	145
Mean	63	152
SD	4.6	35.3

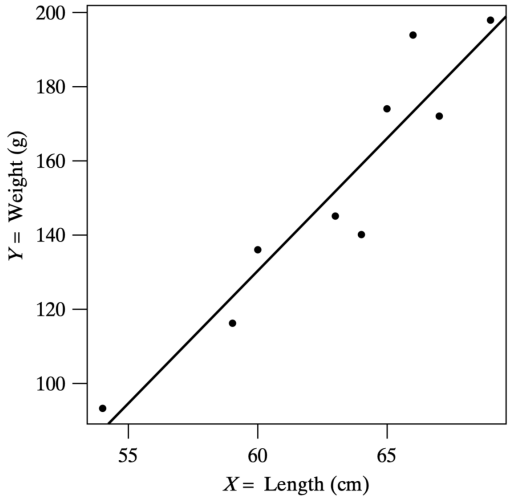


Figure 12.2.1 Body length and weight of nine snakes with fitted regression line

The scatterplot shown in the preceding figure shows a clear upward trend. We say that weight shows a **positive association** with length, indicating that greater lengths are associated with greater weights. Thus, snakes that are longer than the average length of $\bar{X} = 63$ tend to be heavier than the average weight of $\bar{Y} = 152$. The line superimposed on the plot is called the **fitted regression line** or **least-squares line** of Y on X. We will learn how to compute and interpret the regression line later.

Measuring strength of linear association

How strong is the linear relationship between snake length and weight? Are the data points tightly clustered around the regression line, or is the scatter loose? To answer these questions we will compute the correlation coefficient, a **scale-invariant** numeric measure of the strength of linear association between two quantitative variables.

To understand how the correlation coefficient works, consider again the snake length and weight example. Rather than plotting the original data, the figure and table below show the standardized data (Z scores); note that the figure looks identical to our original figure except now our scales are unit-less.

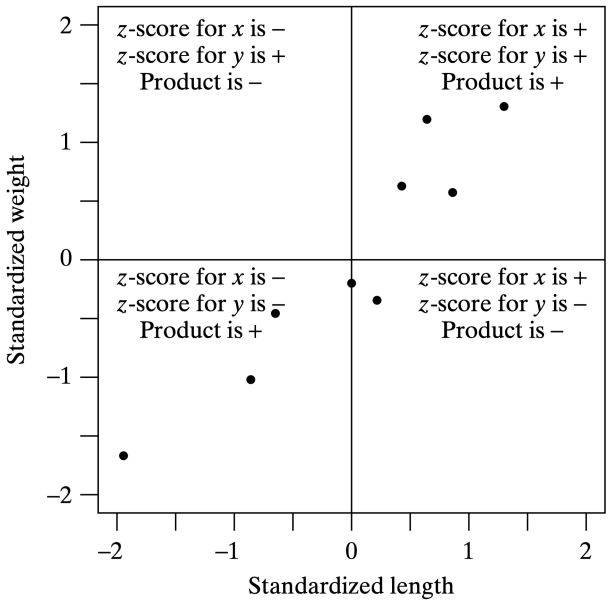


Table 12.2.2 Standardized snake weights, lengths, and their products

	Weight	Length	Standardized weight	Standardized length	Product of standardized values
	X	Y	$z_x = \frac{x - \bar{x}}{s_x}$	$z_y = \frac{y - \bar{y}}{s_y}$	$z_x z_y$
	60	136	−0.65 . . .	−0.45 . . .	0.29 . . .
	69	198	1.29 . . .	1.30 . . .	1.68 . . .
	66	194	0.65 . . .	1.19 . . .	0.77 . . .
	64	140	0.22 . . .	−0.34 . . .	−0.07 . . .
	54	93	−1.94 . . .	−1.67 . . .	3.24 . . .
	67	172	0.86 . . .	0.57 . . .	0.49 . . .
	59	116	−0.86 . . .	−1.02 . . .	0.88 . . .
	65	174	0.43 . . .	0.62 . . .	0.27 . . .
	63	145	0.00 . . .	−0.20 . . .	0.00 . . .
Sum	567	1368	0.00	0.00	7.5494
Mean	63.000	152.000	0.00	0.00	
SD	4.637	35.338	1.00	1.00	

Values in the table are truncated for ease of reading. Because the summary values will be used in subsequent calculations, they include more digits than one would typically report when following our rounding conventions.

- Dividing the plot into quadrants based on the sign of the standardized score.
- Most points fall into the upper-right and lower-left quadrants, indicating positive products of standardized scores.
- Upper-left and lower-right quadrants have points with negative products of standardized scores.
- The sum of these products provides a numeric measure of dominant quadrants.
- Positive association between length and weight leads to a positive sum of standardized score products. A negative relationship would yield a negative sum, and no linear relationship would result in a balanced sum of zero.

The **correlation coefficient** is based on this sum. It is computed as the average product of standardized scores (using $n-1$ rather than n to compute the average):

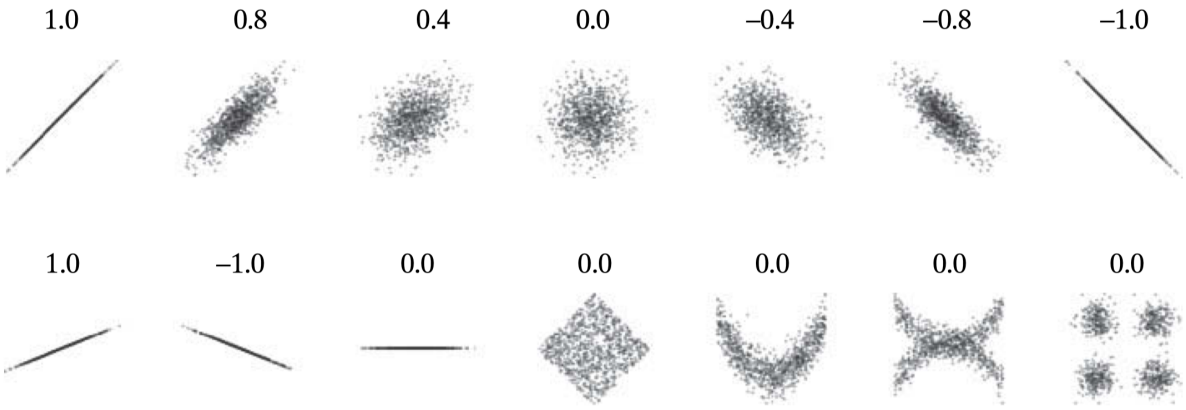
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right).$$

From this formula it is clear that X and Y enter r symmetrically; therefore, if we were to interchange the labels X and Y of our variables, r would remain unchanged.

Interpreting the correlation coefficient

- The correlation coefficient is unit-free and ranges between -1 and 1.
- The sign of the correlation indicates the sign of the relationship (positive or negative) and matches the sign of the slope of the regression line.
- The closer the correlation is to -1 or 1, the stronger the **linear relationship** between X and Y.
- A correlation of -1 or 1 indicates a **perfect linear relationship**, while a correlation of zero means **no linear relationship** between X and Y, but there might still be a **non-linear relationship**.

The figure below displays several examples with a variety of correlation coefficient values.



For the data in length and wight of snakes example, we showed that for the snake data the sum of the products of the standardized scores is 7.5494. Thus, the correlation coefficient for the lengths and weights of our sample of nine snakes is about 0.94.

$$r = \frac{1}{9-1} \times 7.5494 = 0.94.$$

In this example we may also refer to the value 0.94 as the sample correlation, since the lengths and weights of these nine snakes comprise a sample from a larger population. The sample correlation is an estimate of the population correlation (often denoted by the Greek letter "rho", ρ).

Inference concerning correlation

In some investigations it is not a foregone conclusion that there is any relationship between X and Y. It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

$$H_0 : X \text{ and } Y \text{ are uncorrelated in the population}$$

or, equivalently

$$H_0 : \text{ There is no linear relationship between } X \text{ and } Y$$

or symbolically as

$$H_0 : \rho = 0 \text{ v.s. } H_A : \rho \neq 0.$$

A traditional approach to investigate the null hypothesis is to use a t test that is based on the test statistic

$$T = r\sqrt{\frac{n-2}{1-r^2}}.$$

The null distribution of the test statistic is t_{n-2}, i.e.,

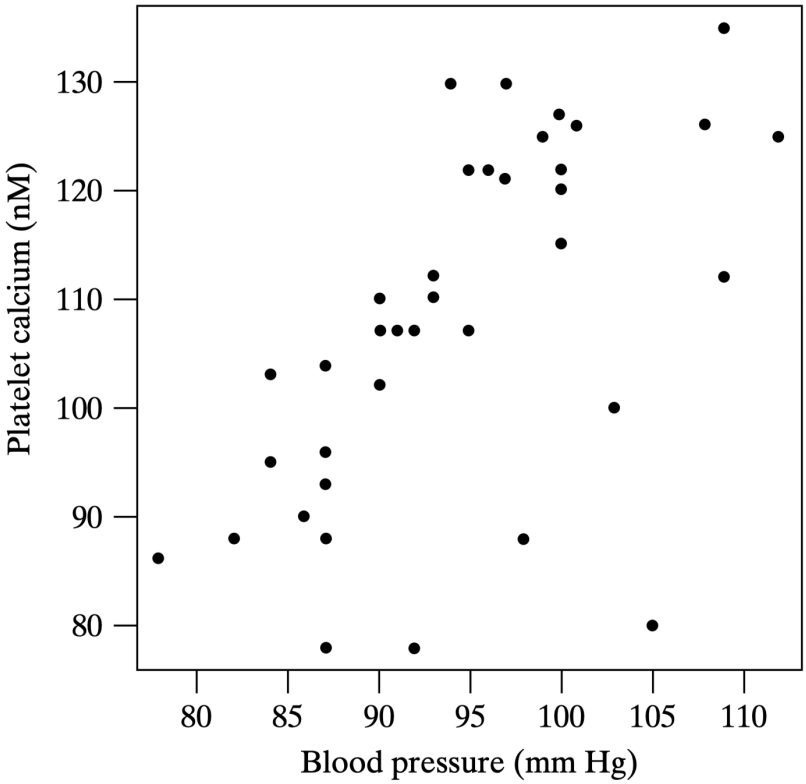
$$T \overset{H_0}{\sim} t_{n-2}.$$

Therefore, H_0 is rejected at the α level of significance if

$$p\text{-value } = 2 \times P(t_{n-2} > |T|) < \alpha \text{ or } |T| > t_{n-2}(\alpha/2).$$

Example: blood pressure and platelet calcium

It is suspected that calcium in blood platelets may be related to blood pressure. As part of a study of this relationship, researchers recruited 38 subjects whose blood pressure was normal (i.e., not abnormally elevated). For each subject two measurements were made: pressure (average of systolic and diastolic measurements) and calcium concentration in the blood platelets. The data are shown in the figure below. The sample size is $n = 38$, and the sample correlation is $r = 0.5832$.



We wish to test the null hypothesis that there is no linear relationship between blood pressure and blood platelet calcium. Let us choose $\alpha = 0.05$. The test statistic is

$$T = 0.5832\sqrt{\frac{38-2}{1-0.5832^2}} = 4.308.$$

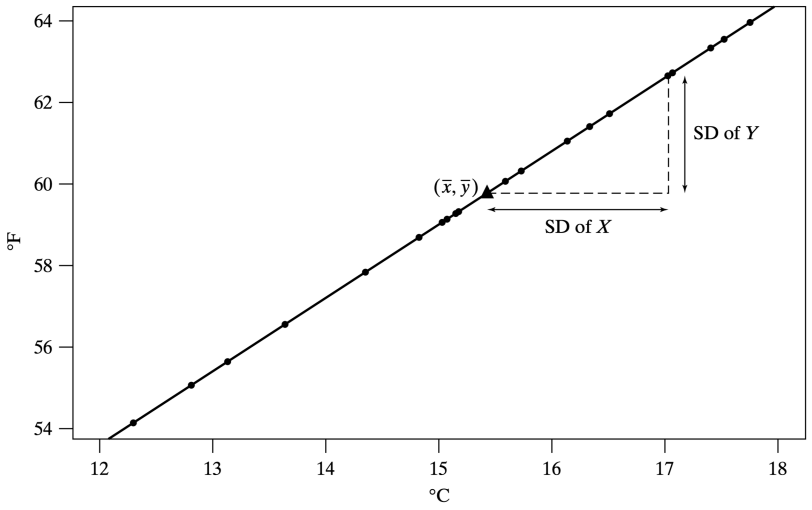
From t Table with $\mathrm{df} = n - 2 = 36$, we find $t_{30}(0.0005) = 3.646$. Thus, we find p-value $< 2 \times 0.0005 = 0.001$ (two-sided), and we reject H_0. The data provide strong evidence that platelet calcium is linearly related with blood pressure.

The Fitted Regression Line

We learned how the correlation coefficient describes the strength of linear association between two numeric variables, X and Y. In this section we will learn how to find and interpret the line that best summarizes their linear relationship.

Example: ocean temperature

Consider a data set for which there is a perfect linear relationship between X and Y, for example, temperature measured in $X =$ Celsius and $Y =$ Fahrenheit. The following figure displays 20 weekly ocean temperatures (in both °C and °F) for a coastal California city along with a line that perfectly describes the relationship: $Y = 32 + 1.8X$.



A summary of the data appears in the following table.

Table 12.3.1 Summary of water temperature data

	X = temperature (°C)	Y = temperature (°F)
Mean	15.43	59.77
SD	1.60	2.88

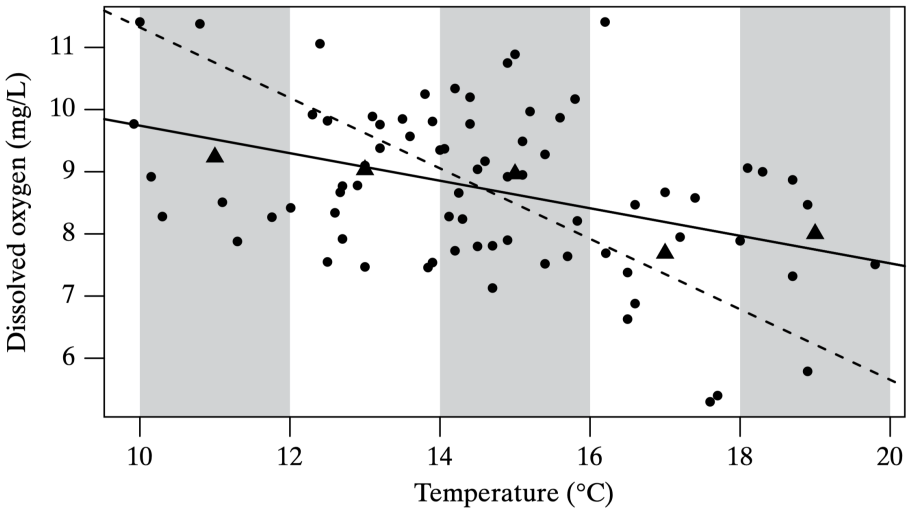
Because X and Y are measuring the same variable (temperature), it stands to reason that a water specimen that is 1 SD above average in °C ($s_X = 1.60$) will also be 1 SD above average in °F ($s_Y = 2.88$). Combined, these values can describe the slope of the line that fits these data exactly:

$$\frac{\text{rise}}{\text{run}} = \frac{s_Y}{s_X} = \frac{2.88}{1.60} = 1.8.$$

In this example we also happen to know the equation of the line that describes the Celsius to Fahrenheit conversion. The slope of this line is 1.80, the same value we found previously.

- In perfect linear relationships (i.e., when $r = \pm 1$) the line that fits the data exactly will have slope $\pm s_Y / s_X$ (the sign of the slope matches the sign of the correlation coefficient) and passes through the point $(\bar{X}, \bar{Y})$.
- This line is sometimes referred to as the **SD line**. Our previous temperature example displays this property.
- But what about situations in which r is not exactly ± 1, that is, when the relationship between X and Y is less than perfectly linear?

In the dissolved oxygen example, we observed a scatterplot indicating that the amount of dissolved oxygen in a river and water temperature appear to be linearly related ($r = -0.391$). The following figure displays a scatterplot of these data along with the SD line (dashed line) and fitted regression line (solid line). Each solid triangle indicates the mean dissolved oxygen level for a range of temperatures specified by the shading.



The dissolved oxygen example shows that the SD line tends to overestimate the mean value of Y for below average X values and underestimate the mean value of Y for above average X values.

- Our examples illustrate that if the relationship is not perfect, the relationship between the mean Y values and X values has a **flatter** slope.
- Mathematically, it can be shown that the line that is best suited to predicting Y (the so called least-squares or **fitted regression line**) has a slope equal to $r(s_Y/s_X)$ and passes through the point $(\bar{X}, \bar{Y})$.
- That is, for X values one standard deviation above average, the mean Y value will only be r standard deviations above average (assuming that r is positive; if r is negative, then for X values one standard deviation above average, the mean Y value will be r standard deviations below average).

For the dissolved oxygen data, the slope of the fitted regression line is

$$r\frac{s_Y}{s_X} = -0.391 \times \frac{1.30}{1.20} = -0.22,$$

meaning that each additional 1 °C increase in water temperature is associated with a 0.22 mg/L decrease in dissolved oxygen level, **on average**.

Table 12.3.2 Summary of dissolve oxygen data

	X = temperature (°C)	Y = dissolved oxygen (mg/L)
Mean	14.58	8.73
SD	2.30	1.30
	$r = -0.391$	

Equation of the fitted regression line.

- The equation of a straight line can be written as

$$Y = b_0 + b_1 X,$$

where b_0 is the y-intercept and b_1 is the slope of the line. The slope b_1 is the rate of change of Y with respect to X.

- Y is the dependent/response variable while X is the independent/explanatory variable.
- The fitted regression line of Y on X is written $\hat{Y} = b_0 + b_1 X$. We write $\hat{Y}$ (read "Y-hat") in place of Y to remind us that this line is providing only estimated or predicted Y values; unless the correlation is ± 1, we don't expect the data values to fall exactly on the line.
- The fitted regression line estimates the **mean value** of Y for any given value of X.
- For the fitted regression line, one has

$$b_1 = r\frac{s_Y}{s_X}, \quad b_0 = \bar{Y} - b_1 \bar{X}.$$

- The formula for the intercept indicates that the fitted regression line passes through the joint mean $(\bar{X}, \bar{Y})$ of our data.

For the dissolved oxygen data, we found that the slope of the fitted regression line to be $b_1 = -0.22$. Using this value we find the intercept,

$$b_0 = 8.73 - (-0.22) \times 14.58 = 11.94.$$

Thus, our fitted regression line is $\hat{Y} = 11.94 - 0.22X$.

The residual sum of squares

We now consider a statistic that describes the scatter of the points about the fitted regression line. The equation of the fitted line is $\bar{Y} = b_0 + b_1 X$. Thus, for each observed X_i in our data there is a predicted Y value of

$$\hat{Y}_i = b_0 + b_1 X_i.$$

Also associated with each observed pair (X_i, Y_i) is a quantity called a **residual**, defined as

$$e_i = Y_i - \hat{Y}_i.$$

A summary measure of the distances of the data points from the regression line is the **error sum of squares**, or SSE, which is defined as follows:

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

For the dissolved oxygen data, the table below indicates how SSE would be calculated from its definition. The values displayed are abbreviated to improve readability.

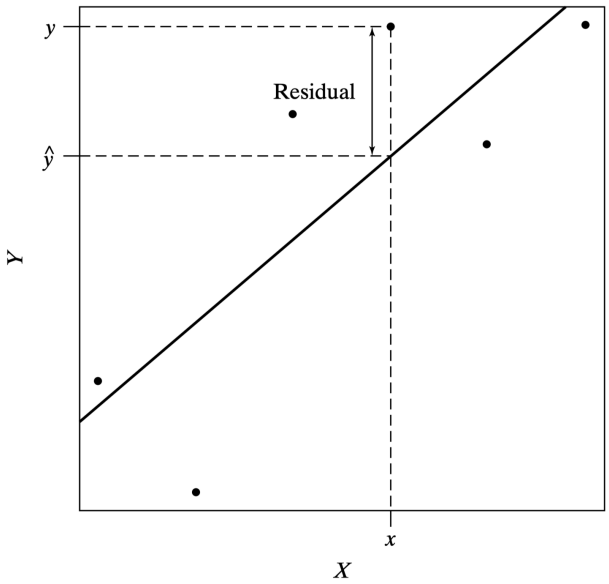


Table 12.3.3 Calculation of SS(resid) for a portion of the dissolved oxygen data

Obs #	x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	15.0	10.89	8.64 ...	2.25 ...	5.06 ...
2	15.7	7.64	8.48 ...	−0.84 ...	0.71 ...
3	17.7	5.40	8.04 ...	−2.64 ...	7.00 ...
4	12.7	7.92	9.14 ...	−1.22 ...	1.50 ...
5	17.2	7.95	8.15 ...	−0.20 ...	0.04 ...
6	18.9	5.79	7.78 ...	−1.99 ...	3.96 ...
7	14.2	10.34	8.81 ...	1.52 ...	2.32 ...
8	16.5	6.63	8.31 ...	−1.68 ...	2.82 ...
9	14.9	7.90	8.66 ...	−0.76 ...	0.58 ...
⋮	⋮	⋮	⋮	⋮	⋮
70	14.0	9.35	8.86 ...	0.49 ...	0.24 ...
71	15.1	9.49	8.61 ...	0.87 ...	0.76 ...
72	16.6	6.88	8.28 ...	–1.40 ...	1.98 ...
73	13.2	9.38	9.03 ...	0.34 ...	0.11 ...
74	12.0	8.42	9.29 ...	–0.87 ...	0.77 ...
75	18.1	9.06	7.95 ...	1.10 ...	1.21 ...
Sum				0.0	106.14 = SS(resid)

Interpreting the fitted regression line

- *Slope*: The average/expected/typical change in Y when X increases by 1 unit.
- *Intercept*: The average/expected/typical value of Y when $X = 0$. The intercept could have no practical meaning, for example, $X = 0$ does not make sense for the snake data.
- *Predict a value of Y based on a value of X*: Substitute values into X to predict a Y ($\hat{Y}$).
- *Calculate the error from the fitted regression line*: For a specific pair of observations (X_i, Y_i), the error for Y_i is $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$.

Several facts:

- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

- $\sum_{i=1}^{n} e_i = 0$
- $\sum_{i=1}^{n} X_i e_i = 0$
- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

The least-squares criterion

Many different criteria can be proposed to define the straight line that “best” fits a set of data points. The classical criterion is the least-squares criterion:

- The "best" straight line is the one that minimizes the error sum of squares (SSE).

The formulas given for b_0 and b_1 were derived from the least-squares criterion by applying calculus to solve the minimization problem. The fitted regression line is also called the "least-squares line".

The residual standard deviation

A measure derived from the error sum of squares (SSE) and easier to interpret is the residual standard deviation,

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}.$$

The **residual standard deviation** tells how far above or below the regression line points tend to be. Thus, the residual standard deviation specifies how far off predictions made using the regression model tend to be.

For the dissolved oxygen data, the residual standard deviation is

$$s_e = \sqrt{\frac{106.14}{75-2}} = \sqrt{1.454} = 1.21.$$

Thus, predictions for the levels of dissolved oxygen based on the regression model tend to deviate by about 1.21 mg/L on average.

The coefficient of determination

We have said that the magnitude of r describes the tightness of the linear relationship between X and Y and have seen how its value is related to the slope of the regression line. When squared, it also provides an additional and very interpretable summary of the regression relationship. The **coefficient of determination**, r^2, describes *the proportion of the variance in Y that is explained by the linear relationship between Y and X.*

$$r^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\text{SSE}}{(n-1)s_Y^2}.$$

For the dissolved oxygen data, we found $r = -0.391$, so $r^2 = 0.153$. Thus, 15.3% of the variance in dissolved oxygen level is explained by the linear relationship between dissolved oxygen level and water temperature.

Parametric Interpretation of Regression: The Linear Model

One use of regression analysis is simply to provide a concise description of the data. The quantities b_0 and b_1 locate the regression line, and s_e describes the scatter of the points about the line. For many purposes, however, data description is not enough. In this section we consider inference from the data to a larger population.

Conditional populations and conditional distributions

A conditional population of Y values is a population of Y values associated with a fixed, or given, value of X. Within a conditional population we may speak of the conditional distribution of Y. The mean and standard deviation of a conditional population distribution are denoted as

$$E(Y|X) = \mu_{Y|X} = \text{ Population mean } Y \text{ value for a given } X$$

$$\text{Var}(Y|X) = \sigma^2_{Y|X} = \text{ Population variance of } Y \text{ values for a given } X$$

Consider the variables $X =$ Height and $Y =$ Weight for a population of young men. The conditional means and standard deviations are

$$\mu_{Y|X} = \text{ Mean weight of men who are } X \text{ inches tall}$$

$$\sigma_{Y|X} = \text{ SD of weights of men who are } X \text{ inches tall}$$

Thus, $\mu_{Y|X}$ and $\sigma_{Y|X}$ are the mean and standard deviation of weight in the subpopulation of men whose height is X. Of course, there is a different subpopulation for each value of X.

The linear model

When we conduct a linear regression analysis, we think of Y as having a distribution that depends on X. The analysis can be given a parametric interpretation if two conditions are met.

- **Linearity**: $Y = \mu_{Y|X} + \varepsilon$ where $\mu_{Y|X}$ is a linear function of X; that is $\mu_{Y|X} = \beta_0 + \beta_1 X$. Thus

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

- **Constancy of standard deviation**: $\sigma_{Y|X}$ does not depend on X. We denote this constant value as σ_ε.

In the linear model $Y = \beta_0 + \beta_1 X + \varepsilon$, the ε term represents random error. We include this term in the model to reflect the fact that Y varies, even when X is fixed.

Estimation in the linear model

Consider now the analysis of a set of (X, Y) data. Suppose we assume that the linear model is an adequate description of the true relationship of Y and X. Suppose further that we are willing to adopt the following random subsampling model:

- For each observed X, the corresponding observed Y is viewed as randomly chosen from the conditional population distribution of Y values for that X.

Within the framework of the linear model and the random subsampling model, the quantities b_0, b_1, and s_e calculated from a regression analysis can be interpreted as estimates of population parameters:

- b_0 is an estimate of β_0.
- b_1 is an estimate of β_1.
- s_e is an estimate of σ_ε.

From the summaries of the snake data, we can compute the following regression coefficients $b_0 = -301, b_1 = 7.19$, and $s_e = 12.5$ (computing these yourself from the provided summaries would be a good exercise). Thus,

- 301 is our estimate of β_0.
- 7.19 is our estimate of β_1.
- 12.5 is our estimate of σ_ε.

Statistical inference concerning β_1

The linear model provides interpretations of b_0, b_1, and s_e that take them beyond data description into the domain of statistical inference. In this section we consider inference about the true slope b_1 of the regression line. The methods are based on the condition that the conditional population distribution of Y for each value of X is a **normal distribution**. This is equivalent to stating that in the linear model of $Y = \beta_0 + \beta_1 X + \varepsilon$, the ε values come from a normal distribution.

The standard error of b_1

Within the context of the linear model, b_1 is an estimate of β_1. Like all estimates calculated from data, b_1 is subject to sampling error. The standard error of b_1 is

$$\mathrm{SE}_{b_1} = \frac{s_e}{s_X\sqrt{n-1}}.$$

For the snake data, we found that $n = 9, s_X = 4.637$ and $s_e = 12.5$. The standard error of b_1 is

$$\mathrm{SE}_{b_1} = \frac{12.5}{4.637\sqrt{9-1}} = 0.9531.$$

Confidence interval for β_1

In many studies the quantity β_1 is a biologically meaningful parameter and a primary aim of the data analysis is to estimate β_1. A confidence interval for β_1 can be constructed by the familiar method based on the SE and Student's t distribution.

A $1-\alpha$ confidence interval for β_1 is constructed as

$$b_1 \pm t_{n-2}(\alpha/2) \times \mathrm{SE}_{b_1}.$$

For the snake data, we found that $b_1 = 7.19, \mathrm{SE}_{b_1} = 0.9531$. There are $n = 9$ observations; we refer to t Table with $\mathrm{df} = 9 - 2 = 7$, and obtain $t_7(0.025) = 2.365$. The 95% confidence interval is

$$7.19 \pm 2.365 \times 0.9531$$

or $(4.94, 9.45)$. We are 95% confident that the true slope of the regression of weight on length for this snake population is between 4.94 gm/cm and 9.45 gm/cm; this is a rather wide interval because the sample size is not very large.

Testing the hypothesis: $H_0 : \beta_1 = 0$

In some investigations it is not a foregone conclusion that there is any linear relationship between X and Y. It then may be relevant to consider the possibility that any apparent trend in the data is illusory and reflects only sampling variability. In this situation it is natural to formulate the null hypothesis

$$H_0 : \mu_{Y|X} \text{ does not depend on } X.$$

Within the linear model, this hypothesis can be translated as

$$H_0 : \beta_1 = 0.$$

A t test of H_0 is based on the test statistic

$$T = \frac{b_1 - 0}{\mathrm{SE}_{b_1}}.$$

The null distribution of the test statistic is t_{n-2}. Specifically,

$$T \overset{H_0}{\sim} t_{n-2}.$$

H_0 is rejected at the α level of significance if

$$p\text{-value } = 2 \times P(t_{n-2} > |T|) < \alpha \text{ or } |T| > t_{n-2}(\alpha/2).$$

While the forms of the test statistic are quite different, testing $H_0 : \beta_1 = 0$ is equivalent to testing $H_0 : \rho = 0$. Recall that a population correlation of zero indicates that there is no linear relationship between X and Y. In this case, the slope that best summarizes "no linear relationship" is a slope of zero.

Note that

$$b_1 = r\frac{s_Y}{s_X}, \quad r^2 = 1 - \frac{\text{SSE}}{(n-1)s_Y^2}, \quad s_e = \sqrt{\frac{\text{SSE}}{n-2}}, \quad \text{SE}_{b_1} = \frac{s_e}{s_X\sqrt{n-1}}.$$

One can verify that the test statistic for $H_0 : \beta_1 = 0$ is equal to the test statistic for $H_0 : \rho = 0$, i.e.,

$$\frac{b_1 - 0}{\text{SE}_{b_1}} = r\sqrt{\frac{n-2}{1-r^2}}.$$

For the snake data, we found that $b_1 = 7.19, \text{SE}_{b_1} = 0.9531$. The test statistic is

$$T = \frac{7.19 - 0}{0.9531} = 7.54.$$

There are $n = 9$ observations; we refer to t Table with $\text{df} = 9 - 2 = 7$, and obtain $t_7(0.0005) = 5.408$. Thus, we find that p-value < 0.001 and we reject H_0. The data provide sufficient (and very strong) evidence to conclude that the true slope of the regression of snake body weight on body length in this population is nonzero.

Note that the test on β_1 does not ask whether the relationship between $\mu_{Y|X}$ and X is linear. Rather, the test asks whether, assuming that the linear model holds, we can conclude that **the slope is nonzero**. It is therefore necessary to be careful in phrasing the conclusion from this test. For instance, the statement "There is a significant linear trend" could easily be misunderstood.

Conditions for inference

The quantities b_0, b_1, s_e, and r can be used to describe a scatterplot that shows a linear trend. However, statistical inference based on these quantities depends on certain conditions concerning the design of the study, the parameters, and the conditional population distributions.

- Design conditions.
  - Random subsampling model: For each observed X, the corresponding observed Y is viewed as randomly chosen from the conditional population distribution of Y values for that X.
- Conditions concerning parameters. The linear model states that

    - $\mu_{Y|X} = \beta_0 + \beta_1 X$.
    - σ_ε does not depend on X.
- Condition concerning population distributions.
    - The confidence interval and t test are based on the conditional population distribution of Y for each fixed X having a **normal distribution**.