Chapter 10
Inference for Bivariate Data

10.1 Inference for Regression

In Chapter 9 we used the least squares regression line to summarize the linear relationship between a response variable Y and an explanatory variable X. We will now consider a more formal (inferential) approach to this problem based on a model for the population distribution of Y. The model we will use is the simple linear regression model which assumes that the population mean response (the population mean of Y) is a linear function of the explanatory variable X.

We will use a simple example to motivate the simple linear regression model and to develop the associated methods of inference. Throughout this discussion we will provide formulae and computations to clarify definitions. You will not need to perform most of these computations, since they can be performed using a suitable calculator or computer statistics program.

Example. Arsenic concentrations. Bencko and Symon (*Env. Res.* 1977) considered the effects of air pollution from a power plant burning coal with a high arsenic content on the health of persons living near the plant. Groups of ten year old boys, each group consisting of 20–27 boys, were selected from ten communities southwest (downwind) of the plant. For each group the response variable Y = average concentration of arsenic in the hair (in parts per million, ppm) was measured and the explanatory variable X = distance of the community from the plant (in kilometers, km) was recorded. The data are given in Table 1.

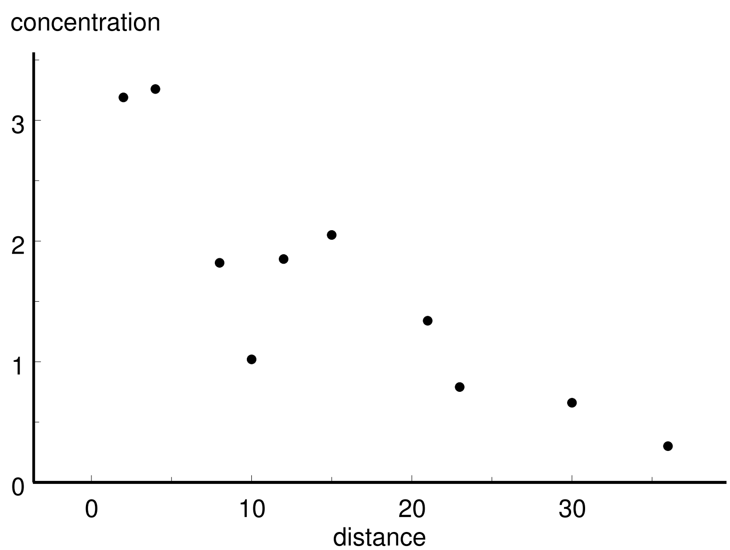
Table 1. Arsenic Data.

distance	2	4	8	10	12	15	21	23	30	36
arsenic conc.	3.19	3.26	1.82	1.02	1.85	2.05	1.34	0.79	0.66	0.30

First consider a model for the distribution of the response variable Y = arsenic concentration at a particular community located at a distance $X = x$ kilometers downwind from the plant. We will model the observed responses ($Y's$) using normal distributions. More specifically, we will assume that the distribution of the arsenic concentration Y corresponding to a community at a distance of $X = x$ is a normal distribution with population mean $\mu(x)$ and population variance σ^2 (population standard deviation σ). The notation $\mu(x)$ indicates that the population mean response depends on the distance $X = x$ of the community from the plant. The population variance is assumed to be constant so that the variance of Y is the same regardless of the distance $X = x$.

In the context of this example we would expect the distribution of the arsenic concentration Y to depend on the distance X of the corresponding community from the power plant. In general, we would expect to observe smaller values of Y for communities which are farther away from the plant. The tendency to observe lower arsenic concentrations at communities farther from the plant is supported by the plot of arsenic concentration versus distance in Figure 1.

Figure 1. Plot of arsenic concentration versus distance.



The simple linear regression model assumes that the population mean response $\mu(x)$ is a linear function of the corresponding value $X = x$ of the explanatory variable. The **population simple linear regression line** can be parameterized in the intercept and slope form

$$\mu(x) = \alpha + \beta x,$$

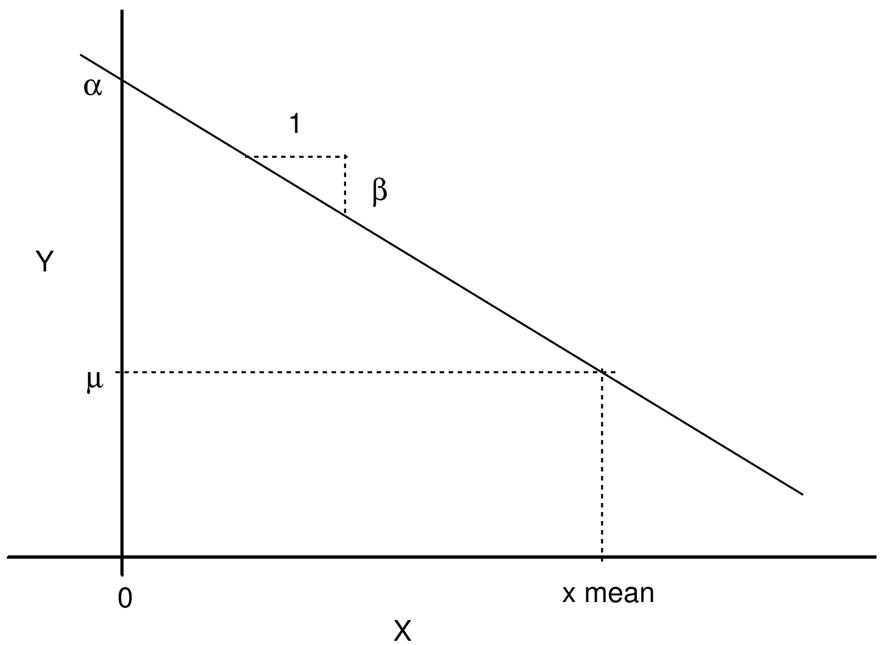
where α is the population intercept and β is the population slope; or, letting $\overline{x}$ denote the mean of the observed values of X, in the mean and slope form

$$\mu(x) = \mu + \beta(x - \overline{x}),$$

where $\mu = \mu(\overline{x})$ denotes the population mean response corresponding to $X = \overline{x}$. The simple linear regression model assumes that there is a constant rate of change, β, in the population mean response, $\mu(x)$, as a function of the explanatory variable x. In many applications this assumption of a constant rate of change will not be appropriate for all possible values of X; however, it may be reasonable if we restrict our attention to a suitable range of X

values. A population regression line, drawn with negative slope, with the parameters α, β, and μ indicated is provided in Figure 2.

Figure 2. A population regression line (drawn with negative slope).



Based on the plot of Figure 1, the assumption of a constant rate of change in the population mean arsenic concentration seems reasonable for distances between 2 and 36 km from the plant. However, we would not necessarily expect this constant rate of change to hold for distances less than 2 km or greater than 36 km. In particular, as the distance from the plant gets very large we would expect the population mean arsenic concentration to decrease more slowly and to eventually stabilize at some background level. Therefore, in this example and in general, we must be careful about making inferences which correspond to extrapolations beyond the range of X values for which we have data.

The estimated or fitted regression line is the least squares regression line introduced in Section 9.3. This fitted regression line passes through the point $(\overline{x}, \overline{Y})$ and has slope b, where

$$b = \frac{\sum(x-\overline{x})(Y-\overline{Y})}{\sum(x-\overline{x})^2} \ .$$

This fitted regression line, which can be expressed as

$$\hat{Y}(x) = a + bx = \overline{Y} + b(x - \overline{x}),$$

is now viewed as an estimate (the least squares estimate) of the population regression line

$$\mu(x) = \alpha + \beta x = \mu + \beta(x - \overline{x})$$

defined above. The least squares estimates of the population slope β, the population mean response $\mu = \mu(\overline{x})$ (corresponding to $X = \overline{x}$), and the population intercept α are the estimated slope $\hat{\beta} = b$, the sample mean response $\hat{\mu} = \overline{Y}$, and the estimated intercept $\hat{\alpha} = a = \overline{Y} - b\overline{x}$, respectively.

The fitted regression line for the arsenic example (graphed in Figure 3) has slope $b = -.07815$ ppm per km which indicates that if the distance of a community from the plant was increased by one km, we would estimate that the population mean arsenic concentration would decrease by .07815 ppm. The fitted line passes through the point $(\overline{x}, \overline{Y}) = (16.1, 1.628)$ which indicates that the estimated mean response for a distance of $X = \overline{x} = 16.1$ km is equal to $\hat{Y}(16.1) = \overline{Y} = 1.628$ ppm. The intercept for this fitted line is $a = 2.8862$ ppm. We can use the residuals and a residual plot, as discussed in Section 9.3, to determine whether this fitted regression line supports the simple linear regression model as an appropriate model for the data at hand. The observed values of the arsenic concentrations Y, the fitted values $\hat{Y}(x)$, and the residual values $Y - \hat{Y}(x)$ are given in Table 2 and the residual plot (a plot of the residuals $Y - \hat{Y}(x)$ versus the distances x) is given in Figure 4.

Table 2. Arsenic data, fitted values, and residuals.

distance X	concentration Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
2	3.19	2.7299	0.4601
4	3.26	2.5736	0.6864
8	1.82	2.2610	-0.4410
10	1.02	2.1047	-1.0847
12	1.85	1.9484	-0.0984
15	2.05	1.7140	0.3360
21	1.34	1.2451	0.0949
23	0.79	1.0888	-0.2988
30	0.66	0.5417	0.1183
36	0.30	0.0728	0.2272

The residual plot appears reasonable overall, especially for such a small data set, with little if any evidence that a straight line (constant rate of change) model is not appropriate. There is one residual, -1.0847 for the community 10 km from the plant, which is somewhat large in magnitude indicating that the observed concentration at a distance of 10 kilometers is somewhat smaller than that predicted by the fitted regression line. The magnitude of this residual is not large enough to cause much concern about the simple linear regression model. We might argue that there is some (slight) evidence of curvature in the residual plot suggesting that the relationship between arsenic concentration and distance is nonlinear; but, again there is not enough evidence to cause much concern. Based on these observations it seems reasonable to use the simple linear regression model for the arsenic example.

Figure 3. Arsenic data with fitted line.

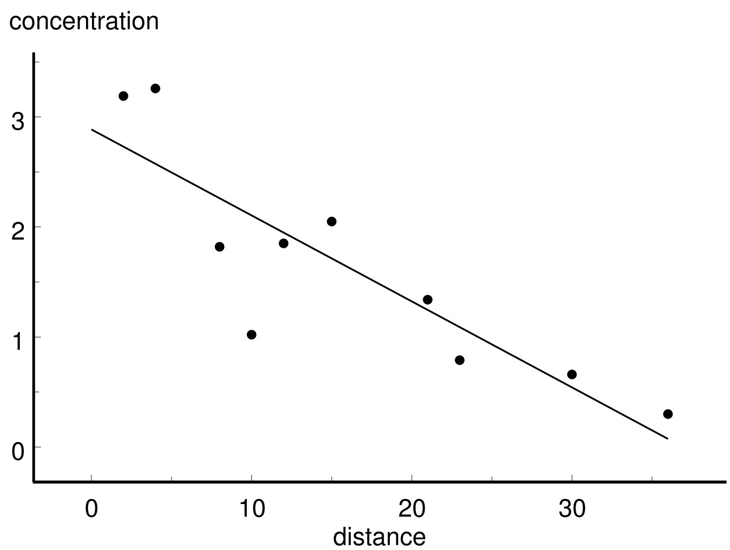
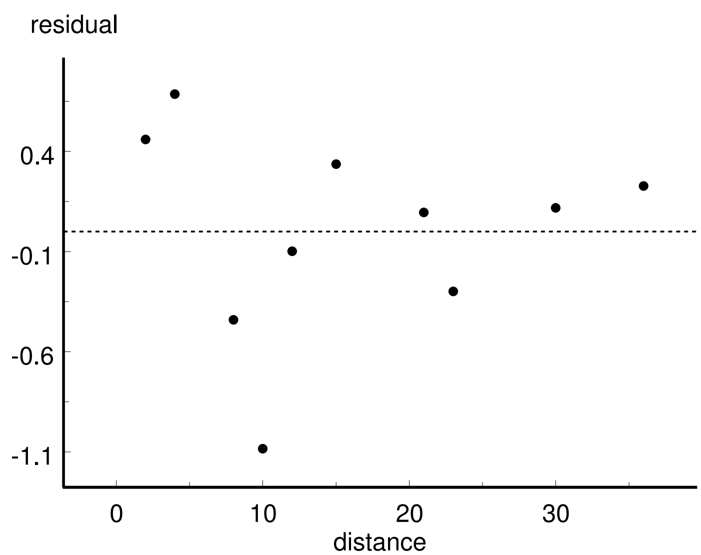


Figure 4. Arsenic example residual plot.



The next step in developing inferential methods for the simple linear regression model is to determine a suitable estimator of the common variance σ^2. We will use a pooled estimator of the common variance based on the residuals. This pooled variance estimator is analogous to the pooled variance estimator of the two sample problem of Chapter 8. In the regression context the model allows a different mean for each distinct value of X and the fitted values, the $\hat{Y}'s$, provide estimates of these means. Thus the pooled variance estimator for the regression problem is the "average" of the squared residuals

$$S_p^2 = \frac{\sum (Y - \hat{Y})^2}{n-2},$$

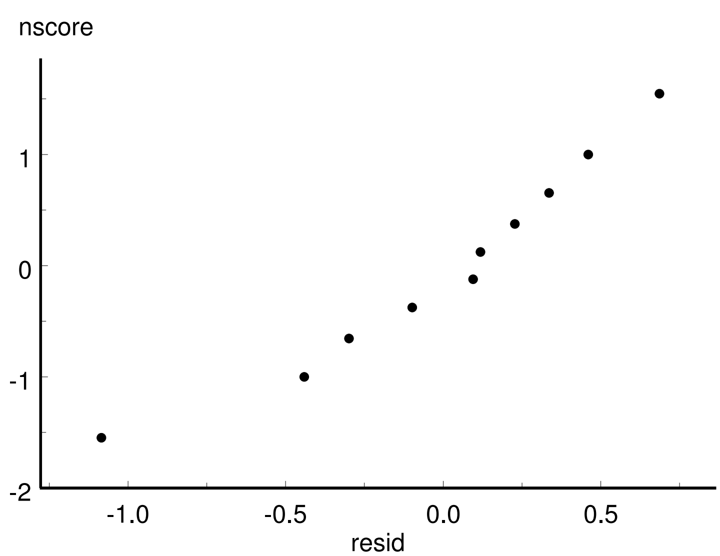
where the sum is over all n observations. The divisor in this pooled variance estimator is $n-2$, since we need two degrees of freedom to estimate the line which determines the fitted values. For the arsenic example the pooled variance estimate is $S_p^2 = .29255$ ($S_p = .54088$) with $n - 2 = 8$ degrees of freedom.

Figure 5. Stem and leaf histogram for arsenic residuals.

In this stem and leaf histogram the stem represents ones and the leaf represents tenths. (ppm)

-1	1
-0	
-0	431
0	1123
0	57

Figure 6. Arsenic residuals normal probability plot.



The confidence interval estimates and hypothesis tests we will now develop are based on the pooled variance estimator S_p^2 from above and the Student's t distribution with $n-2$ degrees of freedom. We can examine the residuals to verify that the normality assumption required for these inferential methods is reasonable. The only disturbing feature of the stem and leaf histogram in Figure 5 for the residuals from the arsenic example is the one large (negative) residual corresponding to $X = 10$ we mentioned above; otherwise, this histogram is consistent with a sample of size ten from a normal distribution. The normal probability plot for the residuals given in Figure 6 also indicates that the normality assumption is reasonable here.

First consider inference for the population slope β. The estimated standard error of the estimated slope b, based on the pooled variance estimator S_p^2, is

$$\widehat{\text{S.E.}}(b) = \frac{S_p}{\sqrt{\sum (x - \overline{x})^2}},$$

and the quantity

$$T = \frac{b - \beta}{\widehat{\text{S.E.}}(b)}$$

follows a Student's t distribution with $n-2$ degrees of freedom. Therefore, we can use the Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for β or to test a hypothesis about β.

For the arsenic example we postulated that the population slope β (the rate of change in arsenic concentration as a function of distance) should be negative, since we expect the arsenic concentration to decrease as the distance from the plant increases. We can address this contention by testing the null hypothesis $H_0 : \beta \geq 0$ versus the research hypothesis $H_1 : \beta < 0$. Under the null hypothesis, with $\beta = 0$, the quantity

$$T = \frac{b - 0}{\widehat{\text{S.E.}}(b)}$$

follows the Student's t distribution with $n - 2 = 8$ degrees of freedom and we can use the Student's t test statistic

$$T_{calc} = \frac{b}{\widehat{\text{S.E.}}(b)}$$

to test $H_0 : \beta \geq 0$ versus $H_1 : \beta < 0$ by rejecting $H_0 : \beta \geq 0$ if T_{calc} is sufficiently far below zero. The P–value for this test is $P(T \leq T_{calc})$, where T denotes a Student's t variable with 8 degrees of freedom. In this example we have $b = -.0782$, $\widehat{\text{S.E.}}(b) = .0161$ and $T_{calc} = -4.85$, which gives a P–value of $P(T \leq -4.85) = .0006$. This P–value is very small providing strong evidence that the population slope β is negative. The 97.5 percentile of the Student's t distribution with 8 degrees of freedom is 2.306 which gives a 95% margin of error for b of $\text{M.E.}(b) = 2.306(.0161) = .0371$ and a 95% confidence interval from $-.0782 - .0371 = -.1153$ to $-.0782 + .0371 = -.0411$. Thus we are 95% confident that the population slope β is at least $-.1153$ ppm/km and at most $-.0411$ ppm/km indicating a decrease in the population mean arsenic concentration of at least .0411 ppm and at most .1153 ppm for each increase of one kilometer in distance from the plant. Note that this interpretation of the slope of the regression line should only be used in the range of distances for which we have data, since we would not necessarily expect this simple linear regression model to hold beyond this range.

The slope provides an estimate of how the population mean response changes as a function of X. We might also want an estimate of the vertical location of the regression line. The population mean response $\mu = \mu(\overline{x})$ (the population mean of the response variable Y at the mean $\overline{x}$ of the explanatory variable values) can be used to indicate the location the population regression line. The sample mean response $\overline{Y}$ provides our estimate of μ. The estimated standard error of the estimated mean response $\overline{Y}$, based on the pooled variance estimator S_p^2, is

$$\widehat{\text{S.E.}}(\overline{Y}) = \frac{S_p}{\sqrt{n}}$$

and the quantity

$$T = \frac{\overline{Y} - \mu}{\widehat{\text{S.E.}}(\overline{Y})}$$

follows a Student's t distribution with $n-2$ degrees of freedom. Therefore, we can use the Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for μ or to test a hypothesis about μ.

To get a feel for the overall population mean arsenic concentration at distances between 2 and 36 km from the plant we can estimate the population mean concentration for a distance of $X = \overline{x} = 16.1$ km, *i.e.*, we can estimate $\mu = \mu(16.1)$. The estimate of the population mean concentration at 16.1 km is $\overline{Y} = 1.628$ and the estimated standard error of $\overline{Y}$ is

$$\widehat{\text{S.E.}}(\overline{Y}) = \frac{S_p}{\sqrt{n}} = .1710.$$

Since there are $n-2 = 8$ degrees of freedom associated with S_p, we know that the quantity

$$T = \frac{\overline{Y} - \mu}{\widehat{\text{S.E.}}(\overline{Y})}$$

follows a Student's t distribution with 8 degrees of freedom. Therefore, the 95% margin of error of $\overline{Y}$ is $\text{M.E.}(\overline{Y}) = 2.306(.1710) = .3943$ and the interval from $1.628 - .3943 = 1.2337$ ppm to $1.628 + .3943 = 2.0223$ ppm is a 95% confidence interval for μ, the population mean arsenic concentration at 16.1 kilometers from the plant. Hence, we are 95% confident that the population mean arsenic concentration for a community 16.1 km from the plant is between 1.2337 and 2.0223 ppm.

The Student's t test statistic obtained from the quantity T above by replacing μ by a specific hypothesized concentration μ_0 could be used to conduct a hypothesis test for comparing μ with μ_0. Since a relevant μ_0 value is not available, we will not consider such a hypothesis test for the arsenic example.

You may have wondered why we used the population mean response $\mu = \mu(\overline{x})$ instead of the population intercept α to quantify the vertical location of the population regression line. Since the population intercept is the population mean response for $X = 0$ and since, as in the arsenic example, $X = 0$ is often not within the range of the values of the explanatory variable we are interested in, there is often little interest in the value of α except as part of the equation for the fitted regression line. Therefore, it is usually more appropriate to consider inference for μ instead of α.

For a specified value x^* of the explanatory variable (note that this x^* should be in the range of the explanatory variable values for which we have data) we can estimate the corresponding population mean response $\mu(x^*)$ as

$$\hat{Y}(x^*) = a + bx^*$$

or as

$$\hat{Y}(x^*) = \overline{Y} + b(x^* - \overline{x}).$$

The first expression, giving $\hat{Y}(x^*)$ in terms of a and b, is more convenient for computation while the second expression, giving $\hat{Y}(x^*)$ in terms of $\overline{Y}$ and b, allows us to more easily find the estimated standard error of $\hat{Y}(x^*)$ and see how it depends on the location of x^* relative to $\overline{x}$. It can be shown that the estimators $\overline{Y}$ and b are statistically independent and that, because of this independence, we can express the estimated standard error of $\hat{Y}(x^*)$ in terms of the estimated standard errors of $\overline{Y}$ and b. For ease of notation let

$$\widehat{\text{var}}(\overline{Y}) = (\widehat{\text{S.E.}}(\overline{Y}))^2 \text{ and } \widehat{\text{var}}(b) = (\widehat{\text{S.E.}}(b))^2$$

denote the estimated variances of $\overline{Y}$ and b. The estimated standard error of $\hat{Y}(x^*)$ is

$$\widehat{\text{S.E.}}(\hat{Y}(x^*)) = \sqrt{\widehat{\text{var}}(\overline{Y}) + (x^* - \overline{x})^2 \widehat{\text{var}}(b)}.$$

Notice that the $(x^* - \overline{x})^2$ term in this standard error causes the standard error of $\hat{Y}(x^*)$ to increase as the distance between x^* and $\overline{x}$ increases. That is, the variability in $\hat{Y}(x^*)$ as an estimator of $\mu(x^*)$ is smaller for values of x^* close to $\overline{x}$ than it is for values of x^* farther from $\overline{x}$. Some calculators and computer programs will provide the standard errors of $\overline{Y}$ and b but will not provide the standard error of $\hat{Y}(x^*)$, if this is true for your calculator or computer program, you can use the expression above to find $\text{S.E.}(\hat{Y}(x^*))$.

We can use the fact that the quantity

$$T = \frac{\hat{Y}(x^*) - \mu(x^*)}{\widehat{\text{S.E.}}(\hat{Y}(x^*))}$$

follows a Student's t distribution with $n-2$ degrees of freedom to form a confidence interval for $\mu(x^*)$ or to test a hypothesis about $\mu(x^*)$. Notice that $\alpha = \mu(0)$ and we can make inferences about the population intercept using the present approach with $x^* = 0$.

Consider the problem of estimating the population mean arsenic concentration $\mu(20)$ for a hypothetical community located 20 km from the power plant. Our estimate of $\mu(20)$ is

$$\hat{Y}(20) = 2.8862 - .07815(20) = 1.3232.$$

In this example

$$\widehat{\text{var}}(\overline{Y}) = .02925 \text{ and } \widehat{\text{var}}(b) = .0002595$$

so that the estimated standard error of $\hat{Y}(20)$ is

$$\widehat{\text{S.E.}}(\hat{Y}(20)) = \sqrt{.02925 + (20 - 16.1)^2(.0002595)} = .1822,$$

which gives a margin of error of M.E.$(\hat{Y}(20)) = (2.306)(.1822) = .4202$. Therefore, we can be 95% confident that the population mean response $\mu(20)$ for a distance of 20 km is between 1.3232 - .4202 = .9030 ppm and 1.3232 + .4202 = 1.7434 ppm.

In some situations instead of estimating the population mean response $\mu(x^*)$ for $X = x^*$ we might wish to predict the actual response $Y(x^*)$ which would be observed if we were to measure Y when $X = x^*$. We can model the actual response value corresponding to $X = x^*$ as $Y(x^*) = \mu(x^*) + \epsilon$, where $\mu(x^*)$ is the corresponding population mean response and ϵ represents a random, normally distributed quantity with mean zero and standard deviation σ. The fitted value $\hat{Y}(x^*)$ which served as our estimate of $\mu(x^*)$ provides a suitable prediction (estimate) of the actual response value $Y(x^*)$ as well, since $\mu(x^*)$ is the mean of the distribution of $Y(x^*)$. However, there is more variability in $\hat{Y}(x^*)$ when it is viewed as a predictor of $Y(x^*)$ than there is when it is viewed as an estimator of $\mu(x^*)$. We can use the standard error of prediction

$$\text{S.E.P.}(\hat{Y}(x^*)) = \sqrt{\widehat{\text{var}}(\overline{Y}) + (x^* - \overline{x})^2\widehat{\text{var}}(b) + S_p^2}$$

to quantify the variability in $\hat{Y}(x^*)$ as a predictor of $Y(x^*)$, and in particular, we can use this standard error of prediction to form an interval estimate of $Y(x^*)$. Notice that the standard error of prediction differs from the standard error for estimating $\mu(x^*)$ by the addition of the term S_p^2 under the square root sign. This added term accounts for the variability in the ϵ of the expression for $Y(x^*)$ given above.

We estimated the population mean arsenic concentration $\mu(20)$ for a hypothetical community located 20 km from the power plant above. Now consider the prediction of the actual response we would have observed if there was a community 20 km from the plant. Since $S_p^2 = 0.29255$, the standard error for prediction for a distance of 20 km is

$$\text{S.E.P.}(\hat{Y}(20)) = \sqrt{.02925 + (20 - 16.1)^2(.0002595) + .29255} = .5707$$

and the 95% prediction interval for the actual response at 20 km is the interval from 1.3232 - (2.306)(.5707) = 1.3232 - 1.3161 = .0071 ppm to 1.3232 + (2.306)(.5707) = 1.3232 + 1.3161 = 2.6393 ppm. Therefore, with 95% confidence we can predict that if we were to measure the actual arsenic concentration at a community 20 km from the plant we would get a value between .0071 ppm and 2.6393 ppm. Notice that this prediction interval is a good bit longer than the 95% confidence interval (.9030, 1.7434) for the population mean arsenic concentration for a community 20 km from the plant, since the prediction interval takes the variability of the measurement process into consideration.

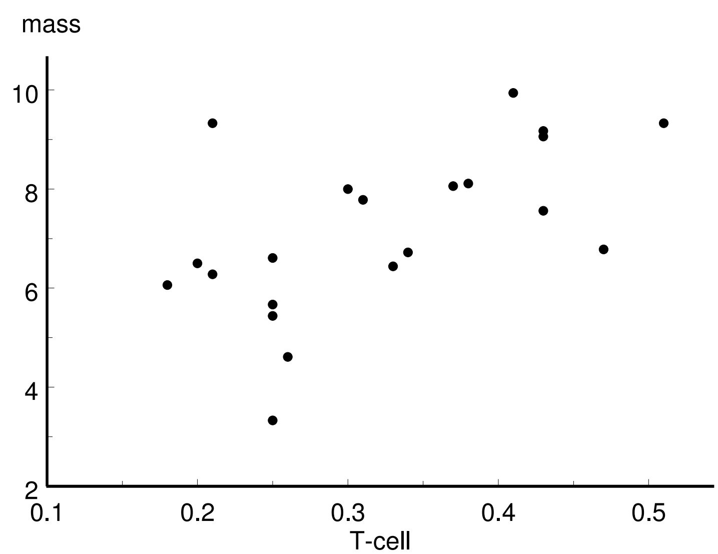
Example. Wheatear weight lifting and health status example. The black wheatear is a small passerine (perching) bird that is resident in Spain and Morroco. The male black wheatear demonstrates an exaggerated sexual display by collecting stones from the ground and placing them in cavities in cliffs, caves, or buildings while the female mate is present. Soler, Martín–Vivaldi, Marín, and Møller, *Behav. Ecol.* **10**, 281–286, (1999) investigated the relationship between such weight lifting and health status for black wheatears. The data in the first two columns of Table 3 (which were read from Figure 1 of this paper) correspond to a sample of $n = 21$ male black wheatears. The two variables are: the bird's T–cell response (in mm) which is a measure of the strength of the bird's immune system; and stone mass (in g) which is the average weight of the stones moved by the bird. The T–cell response is essentially the increase in the thickness of the patagium (wing web) in response to the injection of a lectin. A larger T–cell value indicates a stronger immune system response.

These authors conjectured that male black wheatears signal their current health status to their partners by carrying heavy stones. In particular, they conjectured that birds with stronger immune systems would be expected to carry heavier stones. The plot of stone mass versus T–cell response in Figure 7 shows a reasonably strong linear relationship between stone mass and T–cell response. The authors argued that the T–cell response was very precisely measured; thus, it is reasonable to treat T–cell response as the explanatory variable in a simple linear regression model for stone mass.

Table 3. Wheatear data, fitted values, and residuals.

T-cell response (mm) X	stone mass (g) Y	fitted value $\hat{Y}$	residual $Y - \hat{Y}$
.18	6.06	5.7537	0.3063
.20	6.50	5.9540	0.5460
.21	6.28	6.0542	0.2258
.21	9.33	6.0542	3.2758
.25	3.33	6.4549	-3.1249
.25	5.44	6.4549	-1.0149
.25	5.67	6.4549	-0.7849
.25	6.61	6.4549	0.1551
.26	4.61	6.5551	-1.9451
.30	8.00	6.9558	1.0442
.31	7.78	7.0560	0.7240
.33	6.44	7.2563	-0.8163
.34	6.72	7.3565	-0.6365
.37	8.06	7.6570	0.4030
.38	8.11	7.7572	0.3528
.41	9.94	8.0577	1.8823
.43	7.56	8.2581	-0.6981
.43	9.06	8.2581	0.8019
.43	9.17	8.2581	0.9119
.47	6.78	8.6588	-1.8788
.51	9.33	9.0595	0.2705

Figure 7. Plot of stone mass versus T-cell response.



The equation for the fitted regression line for this example (see Figure 8) is

$$\hat{Y} = 3.9505 + 10.0177X,$$

where Y denotes the stone mass in grams and X denotes the T–cell response in mm.

Figure 8. Wheatear data with fitted line.

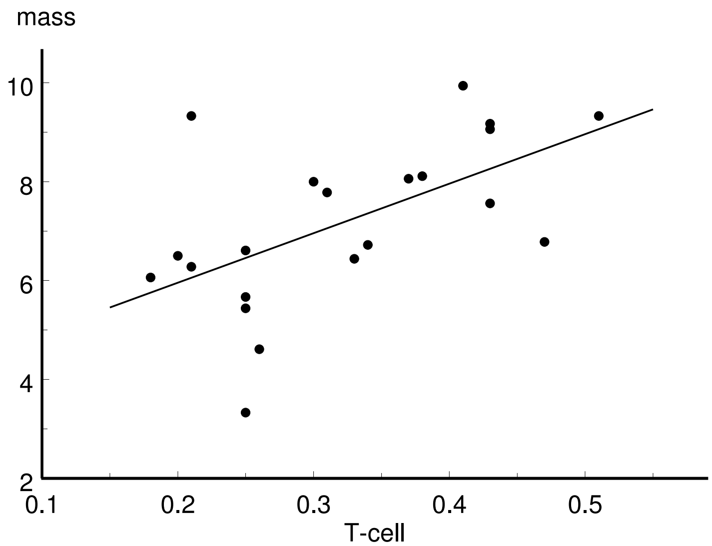
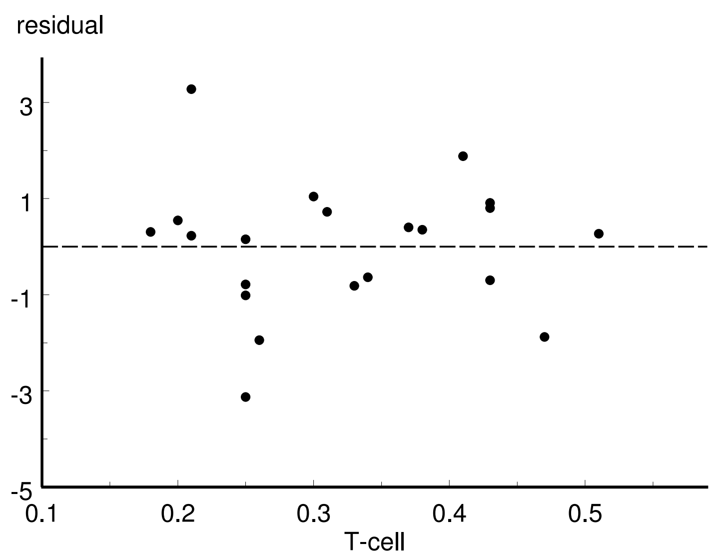


Figure 9. Wheatear example residual plot.



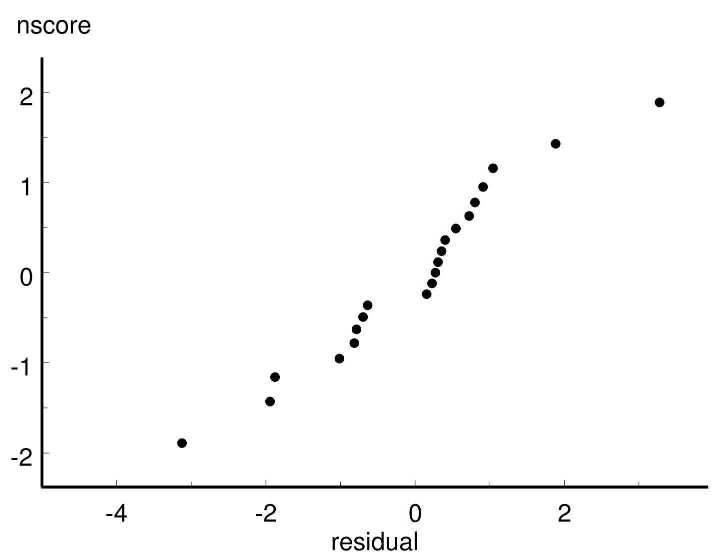
The plot in Figure 8 and the residual plot in Figure 9 show that there are two points which are relatively far away from the fitted line. The observation with $X = .21$ and $Y = 9.33$ has a residual of 3.2758 (see Table 3) and the observation with $X = .25$ and $Y = 3.33$ has a residual of -3.1249. All of the other residuals have magnitudes which are less than two. Because of these two mild outliers the coefficient of determination $R^2 = .3330$ is not very large. Notice that even with these two unusual points T–cell response alone still explains 33.3% of the variability in stone mass. If we had data for some other relevant explanatory variables we could fit a more complex regression model which would account for more of the variability in stone mass.

Figure 10. Stem and leaf histogram for wheatear residuals.

In this stem and leaf histogram the stem represents ones and the leaf represents tenths. (grams)

-3	1
-2	
-1	980
-0	8766
0	1223345789
1	08
2	
3	2

Figure 11. Wheatear residuals normal probability plot.



The two mild outliers are apparent in the stem and leaf histogram of Figure 10 and in the normal probability plot of Figure 11. Even with these points there does not seem to be any problem in treating these stone mass data as forming a random sample of size 21 from a normal distribution.

Table 4. Wheatear regression summary statistics.

$n =$	21	error d.f. $=$	19
$\overline{X} =$	.3223	$S_p^2 =$	2.0249
$\overline{Y} =$	7.1800	$\widehat{SE}(\overline{Y}) =$	.3105
y–intercept, $a =$	3.9505	$\widehat{SE}(a) =$	1.0936
slope, $b =$	10.0177	$\widehat{SE}(b) =$	3.2528
$R^2 =$	.3330		

The slope, $b = 10.0177$ g per mm, of the fitted line indicates that if the T–cell response was increased by 1 mm, then we would estimate that the population mean stone mass would increase by 10.0177 grams. Since the T–cell response cannot increase by 1 mm and stay within the range of the data, we might rephrase this by saying that if the T–cell response increased by .1 mm, then we would estimate that the population mean stone mass would increase by 1.00177 grams. We can state, with 95% confidence, that the population slope β is between 3.2096 and 16.8258 g per mm. We can use a test of $H_0 : \beta \leq 0$ versus $H_1 : \beta > 0$ to quantify the evidence in favor of the conjecture that the population mean stone mass is an increasing function of the T–cell response. For this test we have $T_{calc} = 3.08$ giving a P–value of .0031. This provides strong evidence that the population slope is positive so that a stronger T–cell response yields a higher population mean stone mass.

There is some interest here in considering the population mean stone mass for a low T–cell response value (say $X = .25$) and for a high T–cell response value (say $X = .45$). The estimate of the population mean stone mass for $X = .25$ is $\hat{Y}(.25) = 6.4549$ with standard error .3897; and a 95% confidence interval for $\mu(.25)$ goes from 5.6393 to 7.2705 grams. The estimate of the population mean stone mass for $X = .45$ is $\hat{Y}(.45) = 8.4584$ with standard error .5184; and a 95% confidence interval for $\mu(.45)$ goes from 7.3734 to 9.5435 grams.