

LINEAR REGRESSION

Systematic variation in the true value

Up to now, we have been thinking about measurement as sampling of values from an ensemble of all possible outcomes in order to estimate the true value (which would, according to our previous discussion, be well approximated by the mean of a very large sample). Given a sample of outcomes, we have sometimes checked the hypothesis that it is a random sample from some ensemble of outcomes, by plotting the data points against some other variable, such as ordinal position. Under the hypothesis of random selection, no clear trend should appear. However, the contrary case, where one finds a clear trend, is very important. A clear trend can be a discovery, rather than a nuisance! Whether it is a discovery or a nuisance (or both) depends on what one finds out about the reasons underlying the trend. In either case one must be prepared to deal with trends in analyzing data.

Figure 2.1 (a) shows a plot of (hypothetical) data in which there is a very clear trend. The y axis scales concentration of coliform bacteria sampled from rivers in various regions (units are colonies per liter). The x axis is a hypothetical index of regional urbanization, ranging from 1 to 10. The hypothetical data consist of 6 different measurements at each level of urbanization. The mean of each set of 6 measurements gives a rough estimate of the true value for coliform bacteria concentration for rivers in a region with that urbanization level. The jagged dark line drawn on the graph connects these estimates of true value and makes the trend quite clear: more extensive urbanization is associated with higher true values of bacteria concentration. The straight dashed line shows that much of this trend can be approximated by a linear relationship. The equation of the dashed line is $y = 5.3 + 8.0x$, i.e., the true value of bacteria concentration is estimated to increase by about 8 colonies per liter for each unit increase in the urbanization index.

The linear relationship does not fit the data too well at the lowest and especially at the highest end of urbanization. A linear trend represents a useful approximation, but there may be better ways to represent the actual trend. One device that is often very useful is to replace concentration by its logarithm. Figure 2.1 (b) shows the analogous plot, using base-10 logarithms of bacterial concentration for the y-axis scale. Now the linear relationship fits very well, and indeed, one can view the bumps in the jagged dark line as plausibly due to noise. The equation of the dashed straight line in 2.1(b) is $y = 1.25 + 0.071x$. Each increase of one unit urbanization adds 0.071 to the log concentration, i.e., multiplies the concentration itself by a factor of $10^{0.071} \approx 1.18$. This trend can be summarized by saying that the true value of bacteria concentration is estimated to increase by about 18% for each unit increase in urbanization index.

Fitting lines (or other models) by least-squares

The lines in Figure 2.1 (a,b) could be drawn freehand, or by eye, so as to come as close as possible to the various plotted points. A more systematic approach, introduced by the great mathematician Gauss around 1800, is to draw a line such that the **sum of squared vertical deviations** of the points from the line is as small as possible. This used to be called the method of "least squares." Nowadays, it is called "ordinary least squares," often abbreviated as OLS, because some situations turn out to be handled better by methods that also use a least-squares principle but in a more complex fashion. (These latter methods are beyond the scope of what we will do here.)

The vertical deviations from a fitted line or curve are called **residuals**. The least-squares criterion minimizes the sum of squared residuals; when this has been done, the (unsquared) residuals, which can be positive (points above the curve) or negative (below) always add up to 0.

Figure 2.2 demonstrates what is meant with a very small set of paired numbers (X, Y). Four points are plotted, and the least-squares line is drawn. One can see that its slope is over 2 (actually about 2.44). At the lower left point, there is a small positive residual, 0.36, at the next point, a very small negative residual, -0.08 , and the last two points have larger residuals, one negative and one positive. The sum of the four residuals is 0: this will always be the case, apart from small rounding errors. When one squares these four residuals, only the two large ones matter very much, giving a sum of squares of around 4. Note also that the point 'm' whose coordinates are the means of X and Y falls exactly on the line. This will always be true: the least-squares line goes through the point defined by the empirical means, (\bar{X}, \bar{Y}) .

One can imagine different lines drawn on this plot. Each line would produce a different set of residuals, but the sum of their squares would always be larger than for the line depicted.

The ordinary sample average is actually a least-squares fit: the sum of squared deviations of a set of observations from their average is smaller than the sum of squared deviations from any other value. For example, the average of 1 and 5 is 3; the sum of squared deviations of 1 and 5 from 3 is $2^2 + 2^2 = 8$. The sum of squared deviations from a value different from the average would be larger than 8; for example, if one replaces 3 by 3.1, the sum is $2.1^2 + 1.9^2 = 8.02$, slightly larger; etc.

As another example, consider Figure 2.2 again. The horizontal dashed line is drawn at the level of the Y mean, and so one can visualize the residuals about the mean value, as the vertical distances of the 4 points from this horizontal. One of these is actually much smaller than the residual from the least-squares line, but the other 3 are all much larger. Because this horizontal line goes through \bar{Y} , it has a smaller sum of squared residuals than any other horizontal line, but a much larger sum of squared residuals than the sloping least-squared line. This sum of squared residuals about the horizontal line at the level of the mean will be discussed further below, in the section on Analysis of Variance tables.

Each sample mean in Figure 2.1 (a,b) is actually a least-squares fit to the set of 6 measurements for the given region; and so the empirical jagged line connecting the points in each of these two figures, which goes through all the sample means, can be considered as the best "piecewise-linear" fit, according to the least-squares criterion. The dashed lines in these figures are the best fully linear fits, according the least-squares criterion. Naturally, the dashed lines do not go through all the sample means. This would be impossible, unless the 10 sample means lined up on a perfectly straight line. For the dashed lines, there is a deviation-tradeoff among all sixty points in the plot: each line has a smaller sum of squared residuals, summing across all 60 points, than any other straight line that could be drawn on its plot. If there were one very outlying point, the line would move farther away from many or most of the other 59 points in order not produce a huge deviation from the outlier. This might or might not be a good thing, depending on the reliability and validity of that distinctive measurement.

Four additional comments should be noted.

(i) First, the deviations that are minimized are **vertical** deviations, along the y axis only, **NOT distances** in the plane. This is because we are still using a simple measurement-error model, in which there is error in the y measurements; the variable along the x axis is taken as known more or less exactly. (Where there are important measurement errors in both x and y variables, other methods come in to play, including, at times, minimization of planar distance rather than vertical distance.)

(ii) Second, the deviations minimized are different for the two parts of Figure 2.1. In 2.1 (a), deviations are measured in concentration units; in 2.1 (b), they are deviations in log (base 10) of concentration. In this case, the logarithms are much smaller numbers, and the deviations comparably smaller, but there is no straightforward way to compare deviations in Figure 2.1 (a) from deviations in 2.1 (b), in absolute magnitude. One can see qualitatively that the logarithmic approach produces a simpler model (good fit by a straight line), but it makes no sense to compare the two sums of deviations.

(iii) Third, there is the matter of using **squared** deviations. This was introduced by Gauss because it turns out to be much simpler, mathematically and computationally, to minimize squared deviations than other possible measures (such as absolute values of deviations, or square roots or 4th powers of deviations). Computational simplicity is still quite important, but less so nowadays than 200 years ago, given the speed of modern computers. Mathematical simplicity remains highly desirable. Keep in mind the Pythagorean rule for combining standard errors for independent samples, which is possible only because one is dealing with squared deviations in calculating the variance. Similar considerations will arise in the present context. Apart from computational and mathematical simplicity, squaring deviations has plusses and minusses, as suggested in the earlier discussion about the effect of a single very deviant point. If one thinks that one or more severely deviant points may represent some sort of major error, and should not be taken so seriously, then the least-squares criterion should be abandoned, in favor of newer computational techniques, such as different varieties of "robust regression" found in many statistical packages and programming languages. (These techniques would be covered in the a full-semester course in regression methods.)

(iv) Finally, there is the peculiar word "regression," which has gone unexplained thus far. For now, it should be taken as an odd historical accident. Modern statisticians often speak of linear modelling, or nonlinear modelling, and avoid this odd word. A somewhat interesting explanation can be given for the term, but this is not the place for it.

In a situation where the trend is clearly curvilinear, so that a straight line fit is really not satisfactory, one can often get a better representation with a family of functions more complicated than straight lines. For example, the data in Figure 2.1 (a) are better fit by the quadratic polynomial,

$$y = 26.2 - 2.47x + 0.95x^2$$

than by the linear equation $y = 5.3 + 8.0x$. In fact, the total sum of squared deviations of the 60 points from the dashed line in Figure 2.1 (a) is a little over 5900, but the sum of squared residuals from the quadratic goes down to a little over 3000. We do not graph this, because we think that the logarithmic fit in Figure 2.1 (b) is in this instance a more appropriate way to handle the deviation from a straight line. However, quadratic (or cubic) fits do often come up in data analysis. The point we want to make is that the selection of a best-fitting curve is again done by least squares. This gives us three examples of least squares: a horizontal line at the level of \bar{Y} minimizes the sum of squared residuals, among all possible horizontal lines; the regression line minimizes this sum among all possible lines, horizontal or otherwise; and the function shown above minimizes this sum among all possible functions of the form $y = a + bx + cx^2$.

The output of regression functions

A fairly standard set of summary calculations is reported by regression or linear modelling functions¹ in most data-analysis packages; including the regression function in Excel. We give a list of the summary calculations that we consider most important, and then subsequently comment on each one in detail.

1. **Coefficients.** For a purely linear trend, these are the **intercept** and the **slope** of the best straight line, i.e., 5.3 and 8.0 in the fit of Figure 2.1 (a). When trend is represented by the quadratic family, $y = a + bx + cx^2$, there would be three coefficients, the estimates of a , b , and c that minimize the sum of squared residuals.

2. **Standard errors of coefficients.** In thinking about a sample average, we took the standard error as an indicator of how much that sample average might deviate from the true value that one is trying to estimate from the sample. This idea carries over to the present situation. We think of "true value" now as the true straight line that represents the relationship between y and x ; thus there is a true intercept a and a true slope b . The standard error tells us how far the estimated values of a and b are likely to be from their true values.

Just as before, one can form 95% confidence intervals for the true value of the slope or intercept by centering the confidence interval around the estimated value and taking \pm the standard error multiplied by the appropriate t percentile.

To illustrate, the Excel regression output for the linear fit in Figure 2.1 (a) would look as follows:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.27	2.81	1.87	0.066	-0.35	10.89
Urbanization	8.00	0.45	17.65	0.000	7.10	8.90

We focus here on the coefficients, standard errors, and the lower and upper limits for the 95%-confidence intervals for the coefficients. (The t statistic is just the ratio between the estimated coefficient and its estimated standard error; the P-values in tables like this can be very tricky to interpret and are better ignored entirely.)

1. The terms "regression" and "linear modelling" mean the same thing and they include both linear trends, i.e., functions of the form $y = a + bx$, and many nonlinear trends, such as the quadratic discussed above. The key requirement is that the predicted variable y has to be related to the variable parameters (for example, a , b , and c , in the quadratic) in a linear way. To take a contrary example, $y = \sqrt{a + bx}$ would not be included under linear models, because y is related to a and to b via the encompassing $\sqrt{\quad}$ function. In the case of Figure 2.1 (b), we used linear regression with the logarithm of y taken as the basic observable. If we used the same relationship for a least-squares fit to y itself, e.g., $y = 10^{a + bx}$, minimizing deviations on the y rather than the $\log y$ scale, this would fall outside the scope of linear modelling.

In this case, the t-multiplier is almost exactly 2 (using 58 degrees of freedom, as explained below) and so the confidence limits go two standard errors above and below the estimated coefficient. The slope estimate for the rate of growth of coliform bacteria with Urbanization index could be written compactly as 8.00 ± 0.90 colonies per liter per unit urbanization. The intercept is not very meaningful: 5.3 ± 5.6 colonies per liter at Urbanization = 0; but the latter is entirely off the hypothetical urbanization scale used here.

3. Multiple R and R squared. The degree to which the fit approximates the data is indicated by these two closely related statistics. For a simple linear trend, as in Figures 2.1 (a,b), the R value is called the correlation coefficient (more technical terms are "Pearson" or "product-moment" correlation coefficient), and R squared is just its square, i.e., R^2 . The correlation coefficient R can be positive or negative (depending on whether the best fitting line has positive or negative slope), and it can vary between -1 and 1 . R^2 is of course positive, between 0 and 1. Values of R^2 near 1 indicate a near-perfect fit, while values near 0 indicate that the sum of squared vertical deviations is barely smaller than the corresponding sum from a horizontal line that goes through the mean \bar{y} (x has no relation at all to y).

To illustrate, the correlation in Figure 2.1 (a) is $+0.92$, giving $R^2 = 0.84$; while this is a very high value, that in Figure 2.1 (b) is even higher, $+0.97$ ($R^2 = 0.94$).

Further explanation of R and R^2 is best deferred for a little while.

4. Residual standard deviation. This is almost, but not quite, the standard deviation of the residuals. To calculate the latter, one would divide the sum of squared deviations by $N - 1$, the total degrees of freedom in the N observations of y; but now we divide by $N - 2$ (for a linear trend) or by $N - 3$ (for a quadratic trend) or more generally, by $N - k$ when k different parameters have been estimated from the data. In fact, $N - k$ is called the residual degrees of freedom.

Here is where the idea of degrees of freedom (df) really comes into its own. A mean can be estimated with only 1 observation, but that leaves no measure of intrinsic variability, and so there are 0 residual df. Since any two points on a scatterplot can be joined exactly by a unique straight line, one cannot measure the intrinsic variability about a straight line from just two points. If one fits a straight line to 3 points, there are three residuals, but they contain only 1 item of information about intrinsic variability, hence 1 residual df. In Figure 2.2, the four residuals shown have 2 residual df. Similarly, the empirical (solid) line through all the sample means (in Figure 2.1 a or 2.1 b) requires that we estimate all 10 means from the data, leaving $60 - 10 = 50$ df. This can also be obtained by reflecting that each sample of 6 measurements at a single level of urbanization has 5 df for estimating intrinsic variability; combining all 10 samples, we have $10 \times 5 = 50$ residual df. (One could also find a 9th-degree polynomial that would go exactly through all 10 sample means in the figure, and thus would produce exactly the same residuals as the empirical solid line. A 9th-degree polynomial has 10 coefficients, this again gives us $60 - 10 = 50$ df.)

To reiterate, then, the residual standard deviation is obtained by calculating the

(minimized) sum of squared residuals, dividing by $N - k$, and then taking the square root. For a linear trend, $k = 2$ (slope and intercept are estimated), so one divides by $N - 2$ before taking the square root.

There is a very sad fact about the residual standard deviation. We hate to tell you this, it is so sad. Most statistical software packages (including Excel) continue to mislabel it, even in the year 2002, calling it the "residual standard error" or sometimes just the "standard error" and thereby confusing a lot of people. How would a student feel if all her professors not only called her by a wrong name but actually misled a lot of her friends? In these notes, we are going to deal with such a sad fact by pretending henceforth that it just is not so; we use the correct term, "residual standard deviation." Perhaps by 2102 things will be labelled correctly.

Analysis of variance table. This table is built around the idea of sums of squared residuals for a nested sequence of fitted models. In the simplest case, there are only two models in the sequence. One, called the **null model**, assumes no effect of x on y . That is, in the equation $y = a + bx$, the slope b is zero, and so the model is just $y = a$. The least-squares value of the parameter a is just the sample mean \bar{y} of the N values of y . In Figure 2.2, this null model is shown by the horizontal line. The sum of squared residuals from the null model is just the numerator of the usual estimate of the variance of N numbers. It is often called the **total sum of squares**. The second model in the sequence is the **linear regression model**, where the intercept a and the slope b are both obtained by least squares. The sum of squared residuals for this model is necessarily less than for the null model (since $b = 0$ and $a = \bar{y}$ is a possible set of parameter estimates, and the actual estimates are chosen so as to do at least slightly better than that). The difference between these two sums of squares is the **sum of squares attributed to regression**. The analysis of variance (ANOVA) table reports these three quantities (two residual sums of squares and their difference) and the corresponding residual **df** (and their difference).

This is how the analysis of variance (ANOVA) table would look in Excel, for the fitted linear trend in Figure 2.1 (a).

	df	SS	MS	F	Significance F
Regression	1	31700	31700	311.4	0
Residual	58	5904	102		
Total	59	37604			

Pay attention only to the first two columns, labelled **df** and **SS** (sum of squares), where there is an entry in each row. The bottom row (**Total**) shows the **df** and sum of squared residuals for the null model, the next row up (**Residual**) shows the corresponding values for the linear regression model, and the top row (**Regression**) shows the differences in **df** and in sum of squares between the null and the regression models.

In this case, clearly, the linear regression model fits much better than the null model. That is what we already knew as soon as we looked at Figure 2.1 (a) and saw a clear upward trend that was well approximated by a simple straight line, over much of the

range. The proportional reduction in residual sum of squares is $31700/37604$, or 84.3%. And that, in turn is R^2 , exactly. (Earlier we rounded to 84%.) So now you know more exactly how R^2 measures the goodness-of-fit of the fitted model to the data: it looks at the reduction in squared residuals for y , by using the particular regression model, as compared with a null model that attributes all the variation to random error, rather than a systematic trend involving x . This definition of R^2 applies equally well to quadratic trends or to many other models one might try.

The last 3 columns of the analysis of variance table are best ignored for the present. They will not be needed. Just to make one interconnection: the F value (should be called "F Stat") is just the square of the t Stat for the Urbanization slope in the preceding table ($17.65^2 = 311.5$, which differs only through rounding error).

Examining Residuals. The final output from most regression functions is some sort of examination of the residuals. In the simple case of a sample of measurements, one checks for randomness by looking for an alternative: one plots the data against ordinal position or against some other paired variable to look for a trend. If one does not find any clearly systematic pattern, one continues to assume that observations can be viewed as a "random sample" from some ensemble. In regression, one has already found a trend; now one similarly looks for randomness in the residuals from that trend.

Figure 2.3 examines the residuals from the linear trends in Figure 2.1, by simply plotting the residuals against Urbanization. We already noted, in Figure 2.1 (a), that the linear trend was far from a perfect description. The corresponding residual plot, Figure 2.3 (a), dramatizes the departure. In addition to showing the systematic trend in the mean residuals, this plot also makes obvious something that was partly hidden in Figure 2.1 (a): there is also a trend in the standard deviation of the residuals, with small variability at the left changing to large variability at the right-hand side of the plot. This in turn means that the residual standard deviation is not a good representation of variability—it is an average, but does not do justice to either the left or the right side of the plot.

Figure 2.3 (b) does the same job for the linear trend in Figure 2.1 (b). Now one sees two reasons for strongly preferring the logarithmic model. The trend in mean residuals is abolished, and so is the trend in standard deviation of the residuals. The residuals in 2.3 (b) still do not look quite random, since there are clusters above and below zero at different Urbanization levels, but at least there are no trends, and so the linear fit is giving quite a good approximation to the data and the measurement errors look fairly homogeneous on the logarithmic scale.

Figure 2.1 (a)
bacterial concentration versus urbanization
arithmetic plot, showing empirical and linear trend

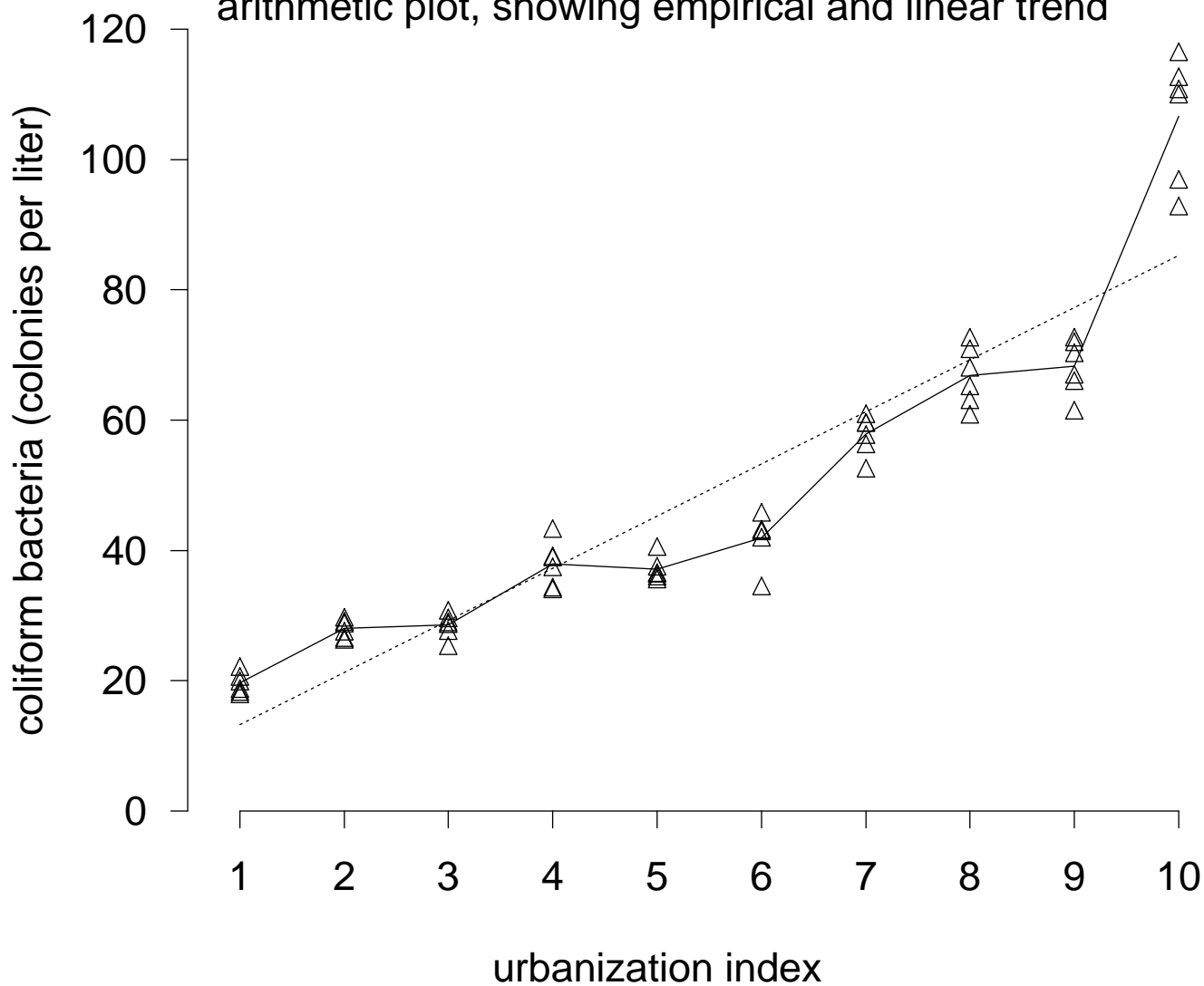


Figure 2.1 (b)
bacterial concentration versus urbanization
logarithmic plot, showing empirical and linear trend

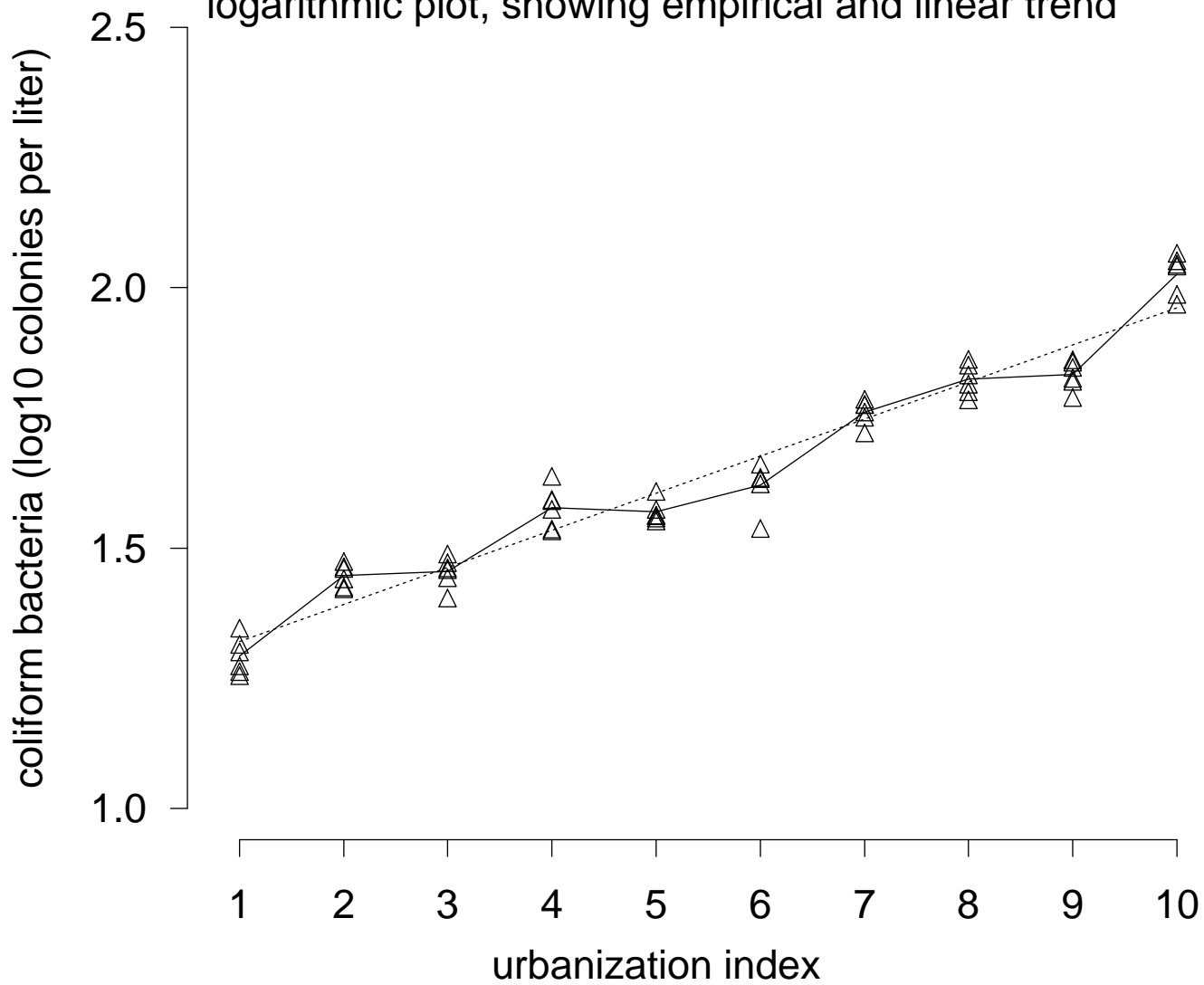


Figure 2.2
linear regression demonstration
showing data points, regression line, and residuals
(mean(X), mean(Y)) plotted as 'm'

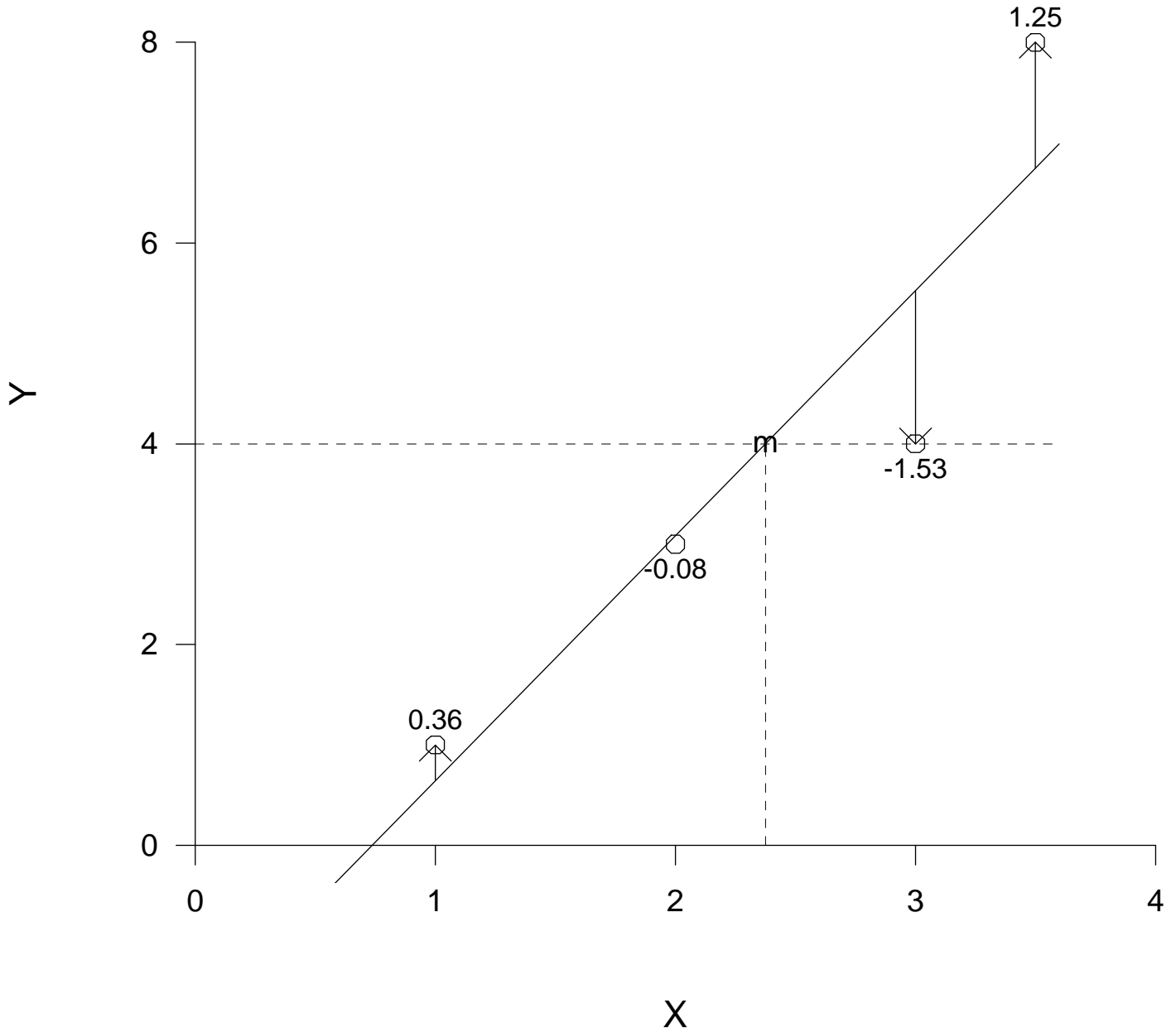


Figure 2.3 (a)
residuals from linear fit in Figure 2.1 (a)

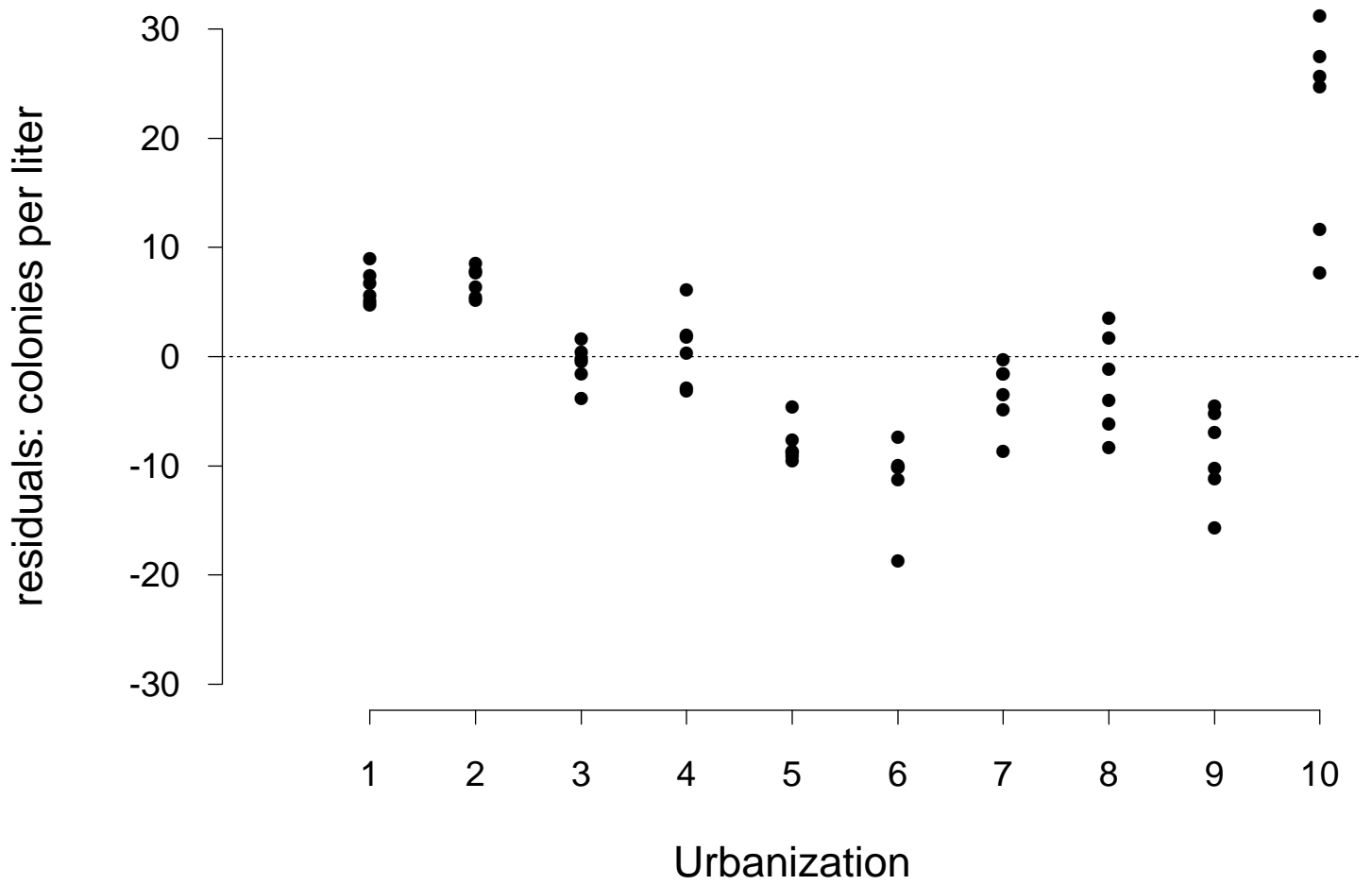


Figure 2.3 (b)
residuals from linear fit in Figure 2.1 (b)

