11 Correlation and Regression

SW, Chapter 12.

Suppose we select $n = 10$ persons from the population of college seniors who plan to take the MCAT exam. Each takes the test, is coached, and then retakes the exam. Let X_i be the pre-coaching score and let Y_i be the post-coaching score for the i^{th} individual, $i = 1, 2, \cdots, n$. There are several questions of potential interest here, for example: Are Y and X related (associated), and how? Does coaching improve your MCAT score? Can we use the data to develop a mathematical model (formula) for predicting post-coaching scores from the pre-coaching scores? These questions can be addressed using **correlation** and **regression** models.

The **correlation coefficient** is a standard measure of **association** or relationship between two features Y and X. Most scientists equate Y and X being correlated to mean that Y and X are associated, related, or **dependent** upon each other. However, correlation is only a measure of the strength of a **linear relationship.** For later reference, let ρ be the correlation between Y and X in the population and let r be the sample correlation. I define r below. The population correlation is defined analogously from population data.

Suppose each of n sampled individuals is measured on two quantitative characteristics called Y and X. The data are pairs of observations $(X_1,\ Y_1)$, $(X_2,\ Y_2)$, $\cdots (X_n,\ Y_n)$, where $(X_i,\ Y_i)$ is the $(X,\ Y)$ pair for the i^{th} individual in the sample. The sample correlation between Y and X, also called the **Pearson product moment correlation coefficient**, is

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}},$$

where

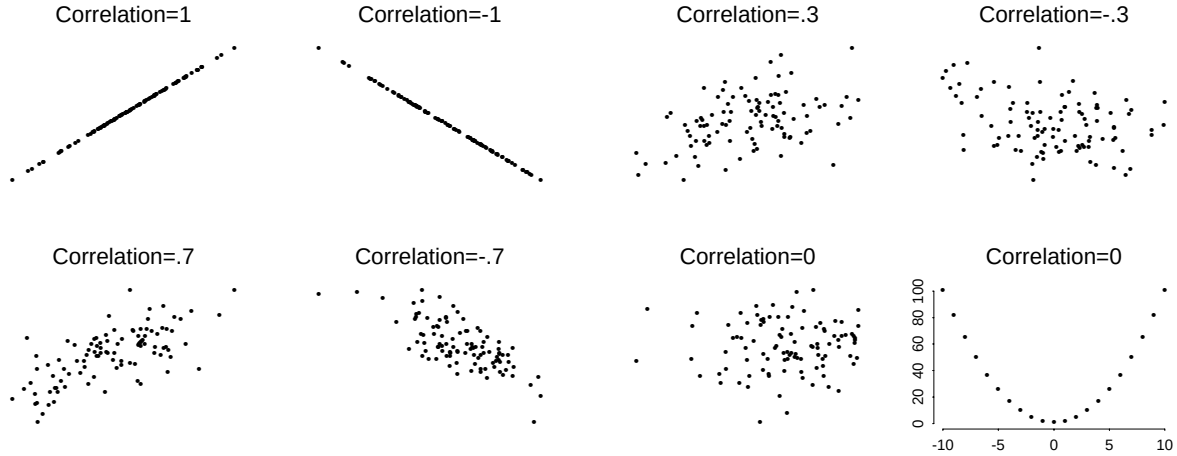
$$S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

is the **sample covariance** between Y and X, and $S_Y = \sqrt{\sum_i (Y_i - \overline{Y})^2/(n-1)}$ and $S_X = \sqrt{\sum_i (X_i - \overline{X})^2/(n-1)}$ are the standard deviations for the Y and X samples. Here are eight important properties of r:

1. $-1 \leq r \leq 1$.
2. If Y_i tends to increase linearly with X_i then $r > 0$.
3. If Y_i tends to decrease linearly with X_i then $r < 0$.
4. If there is a perfect linear relationship between Y_i and X_i with a positive slope then $r = +1$.
5. If there is a perfect linear relationship between Y_i and X_i with a negative slope then $r = -1$.
6. The closer the points (X_i, Y_i) come to forming a straight line, the closer r is to ± 1.
7. The magnitude of r is unchanged if either the X or Y sample is transformed linearly (i.e. feet to inches, pounds to kilograms, Celsius to Fahrenheit).
8. The correlation does not depend on which variable is called Y and which is called X.

If r is near ± 1, then there is a strong linear relationship between Y and X in the sample. This suggests we might be able to accurately predict Y from X with a linear equation (i.e. linear regression). If r is near 0, there is a weak linear relationship between Y and X, which suggests that a linear equation provides little help for predicting Y from X. The pictures below should help you develop a sense about the size of r.

Note that $r = 0$ does not imply that Y and X are not related in the sample. It only implies they are not linearly related. For example, in the last plot $r = 0$ yet $Y_i = X_i^2$.



Testing that $\rho = 0$

Suppose you want to test $H_0 : \rho = 0$ against $H_A : \rho \neq 0$, where ρ is the population correlation between Y and X. This test is usually interpreted as a test of no association, or relationship, between Y and X in the population. Keep in mind, however, that ρ measures the strength of a linear relationship.

The standard test of $H_0 : \rho = 0$ is based on the magnitude of r. If we let

$$t_s = r\sqrt{\frac{n-2}{1-r^2}},$$

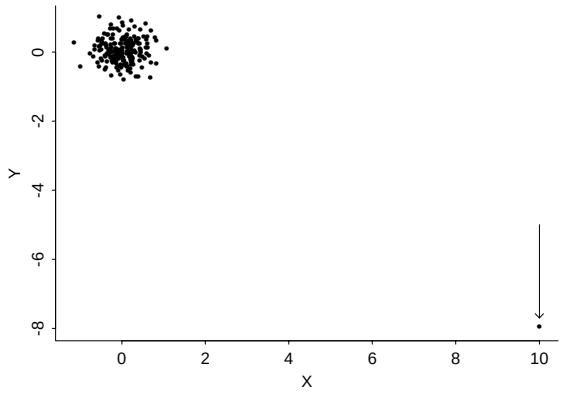
then the test rejects H_0 in favor of H_A if $|t_s| \geq t_{crit}$, where t_{crit} is the two-sided test critical value from a t-distribution with $df = n - 2$. The p-value for the test is the area under the t-curve outside $\pm t_s$ (i.e. two-tailed test p-value).

This test assumes that the data are a random sample from a **bivariate normal population** for $(X,\ Y)$. This assumption implies that all linear combinations of X and Y, say $aX + bY$, are normal. In particular, the (marginal) population frequency curves for X and Y are normal. At a minimum, you should make boxplots of the X and Y samples to check marginal normality. For large-sized samples, a plot of Y against X should be roughly an elliptical cloud, with the density of the points decreasing as the points move away from the center of the cloud.

The Spearman Correlation Coefficient

The Pearson correlation r can be highly influenced by outliers in one or both samples. For example, $r \approx -1$ in the plot below. If you delete the one extreme case with the largest X and smallest Y

value then $r \approx 0$. The two analyses are contradictory. The first analysis (ignoring the plot) suggests a strong linear relationship, whereas the second suggests the lack of a linear relationship. I will not strongly argue that you should (must?) delete the extreme case, but I am concerned about any conclusion that depends heavily on the presence of a single observation in the data set.



Spearman's rank correlation coefficient r_S is a sensible alternative to r when normality is unreasonable or outliers are present. Most books give a computational formula for r_S. I will verbally describe how to compute r_S. First, order the X_is and assign them ranks. Then do the same for the Y_is and replace the original data pairs by the pairs of ranked values. The Spearman rank correlation is the Pearson correlation computed from the pairs of ranks.

The Spearman correlation r_S estimates the **population rank correlation coefficient**, which is a measure of the strength of linear relationship between population ranks. The Spearman correlation, as with other rank based methods, is not sensitive to the presence of outliers in the data. In the plot above, $r_S \approx 0$ whether the unusual point is included or excluded from the analysis. In samples without unusual observations and a linear trend, you often find that $r_S \approx r$.

An important point to note is that the magnitude of the Spearman correlation does not change if either X or Y or both are transformed (monotonically). Thus, if r_S is noticeably greater than r, a transformation of the data might provide a stronger linear relationship.

Example

Eight patients underwent a thyroid operation. Three variables were measured on each patient: weight in kg, time of operation in minutes, and blood loss in ml. The scientists were interested in the factors that influence blood loss. Minitab output for this data set is a separate document.

weight	time	blood loss
44.3	105	503
40.6	80	490
69.0	86	471
43.7	112	505
50.3	109	482
50.2	100	490
35.4	96	513
52.2	120	464

Comments:

1. (Pearson correlations). Blood loss tends to decrease linearly as weight increases, so r should be negative. The output gives $r = -.77$. There is not much of a linear relationship between blood loss and time, so r should be close to 0. The output gives $r = -.11$. Similarly, weight and time have a weak negative correlation, $r = -.07$.

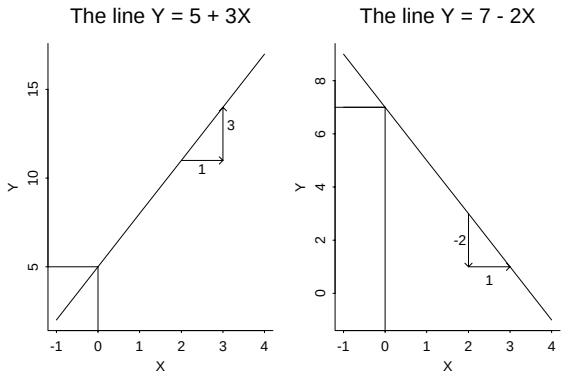
2. The Pearson and Spearman correlations are fairly consistent here. Only the correlation between blood loss and weight is significant at the $\alpha = 0.05$ level (the p-values are given below the correlations).

Simple Linear Regression

In linear regression, we are interested in developing a linear equation that best summarizes the relationship in a sample between the **response variable** Y and the **predictor variable** (or **independent variable**) X. The equation is also used to predict Y from X. The variables are not treated symmetrically in regression, but the appropriate choice for the response and predictor is usually apparent.

Linear Equation

If there is a perfect linear relationship between Y and X then $Y = \beta_0 + \beta_1 X$ for some β_0 and β_1, where β_0 is the Y-intercept and β_1 is the slope of the line. Two plots of linear relationships are given below. The left plot has $\beta_0 = 5$ and $\beta_1 = 3$. The slope is positive, which indicates that Y increases linearly when X increases. The right plot has $\beta_0 = 7$ and $\beta_1 = -2$. The slope is negative, which indicates that Y decreases linearly when X increases.



Least Squares

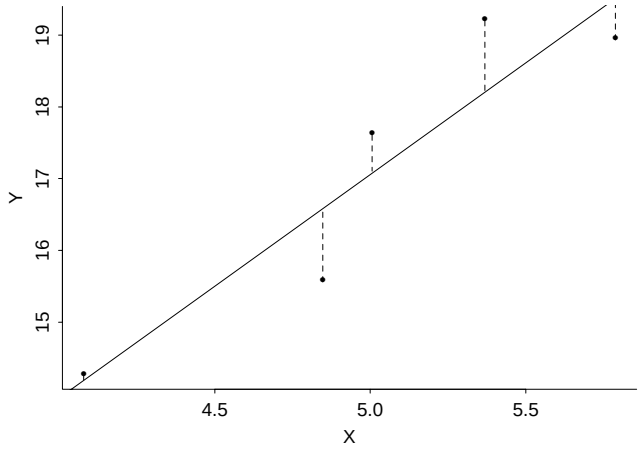
Data rarely, if ever, fall on a straight line. However, a straight line will often describe the **trend** for a set of data. Given a data set (X_i, Y_i), $i = 1, ..., n$ with a **linear trend**, what linear equation "best" summarizes the observed relationship between Y and X? There is no universally accepted

definition of "best", but many researchers accept the **Least Squares** line (LS line) as a reasonable summary.

Mathematically, the LS line chooses the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all possible choices of β_0 and β_1. These values can be obtained using calculus. Rather than worry about this calculation, note that the LS line makes the sum of squared deviations between the responses Y_i and the line as small as possible, over all possible lines. The LS line typically goes through "the heart" of the data, and is often closely approximated by an eye-ball fit to the data.



The equation of the LS line is

$$\hat{y} = b_0 + b_1 X$$

where the intercept b_0 satisfies

$$b_0 = \overline{Y} - b_1\overline{X}$$

and the slope is

$$b_1 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} = r\frac{S_Y}{S_X}.$$

As before, r is the Pearson correlation between Y and X, whereas S_Y and S_X are the sample standard deviations for the Y and X samples, respectively. The **sign of the slope** and the **sign of the correlation** are **identical** (i.e. + correlation implies + slope).

Special symbols b_0 and b_1 identify the LS intercept and slope to distinguish the LS line from the generic line $Y = \beta_0 + \beta_1 X$. You should think of $\hat{Y}$ as the **fitted value** at X, or the value of the LS line at X.

Minitab Implementation

The separate document shows Minitab output from a least squares fit.

For the **thyroid operation data** with Y = Blood loss in ml and X = Weight in kg, the LS line is $\hat{Y} = 552.44 - 1.30X$, or Predicted Blood Loss = 552.44 − 1.30 Weight. For an $86kg$ individual, the Predicted Blood Loss = $552.44 - 1.30 * 86 = 440.64ml$.

The LS regression coefficients for this model are interpreted as follows. The intercept b_0 is the predicted blood loss for a 0 kg individual. The intercept has no meaning here. The slope b_1 is the predicted increase in blood loss for each additional kg of weight. The slope is -1.30, so the predicted *decrease* in blood loss is 1.30 ml for each increase of 1 kg in weight.

Any fitted linear relationship holds only approximately and does not necessarily extend outside the range of the data. In particular, nonsensical predicted blood losses of less than zero are obtained at very large weights outside the range of data.

ANOVA Table for Regression

The LS line minimizes

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all choices for β_0 and β_1. Inserting the LS estimates b_0 and b_1 into this expression gives

$$\text{Residual Sums of Squares} = \sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\}^2.$$

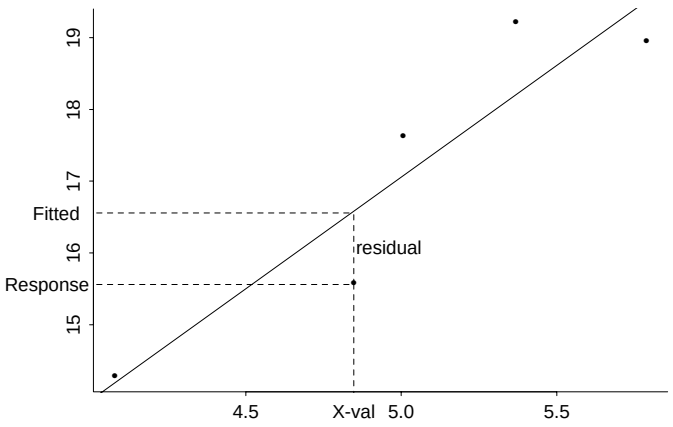
Several bits of notation are needed. Let

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

be the **predicted** or fitted $Y-$value for an $X-$value of X_i and let $e_i = Y_i - \hat{Y}_i$. The fitted value $\hat{Y}_i$ is the value of the LS line at X_i whereas the **residual** e_i is the distance that the observed response Y_i is from the LS line. Given this notation,

$$\text{Residual Sums of Squares} = \text{Res SS } = \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Here is a picture to clarify matters:



The Residual SS, or sum of squared residuals, is *small* if each $\hat{Y}_i$ is *close to* Y_i (i.e. the line closely fits the data). It can be shown that

$$\text{Total SS in Y} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \geq \text{Res SS} \geq 0.$$

Also define

$$\text{Regression SS} = \text{Reg SS} = \text{Total SS} \ - \ \text{Res SS} = b_1 \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}).$$

The Total SS measures the variability in the $Y-$sample. Note that

$$0 \leq \text{Regression SS} \leq \text{Total SS}.$$

The percentage of the variability in the $Y-$ sample that is **explained by the linear relationship** between Y and X is

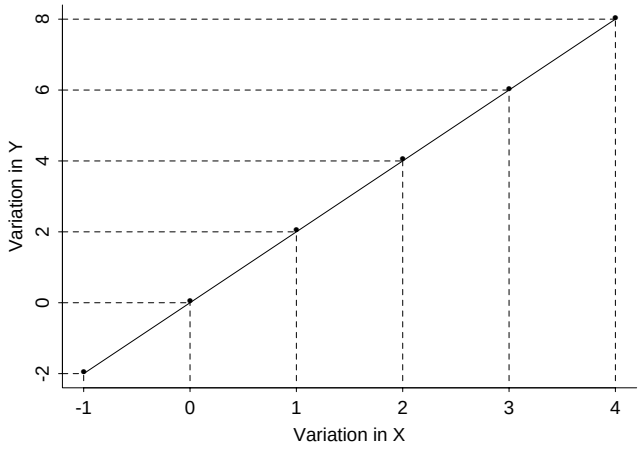
$$R^2 = \text{coefficient of determination} = \frac{\text{Reg SS}}{\text{Total SS}}.$$

Given the definitions of the Sums of Squares, we can show $0 \leq R^2 \leq 1$ and

$$R^2 = \text{square of Pearson correlation coefficient } = r^2.$$

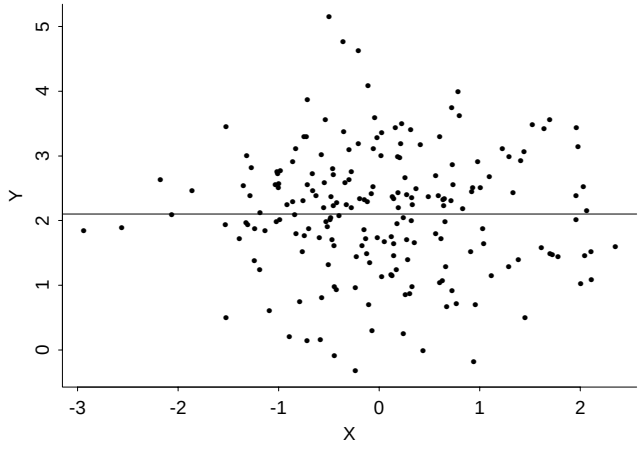
To understand the interpretation of R^2, at least in two extreme cases, note that

$$\begin{aligned}
\text{Reg SS} = \text{Total SS} \quad &\Leftrightarrow \quad \text{Res SS} = 0 \\
&\Leftrightarrow \quad \text{all the data points fall on a straight line} \\
&\Leftrightarrow \quad \text{all the variability in } Y \text{ is explained by the linear relationship with } X \\
&\qquad\quad \text{(which has variation)} \\
&\Leftrightarrow \quad R^2 = 1. \quad \text{(see the picture below)}
\end{aligned}$$



Furthermore,

$$\begin{aligned}
\text{Reg SS} = 0 &\Leftrightarrow \text{Total SS} = \text{Res SS} \\
&\Leftrightarrow b_1 = 0 \\
&\Leftrightarrow \text{LS line is } \hat{Y} = \overline{Y} \\
&\Leftrightarrow \text{none of the variability in } Y \text{ is explained by a linear relationship} \\
&\Leftrightarrow R^2 = 0.
\end{aligned}$$



Each Sum of Squares has a corresponding *df* (degrees of freedom). The Sums of Squares and *df* are arranged in an analysis of variance (ANOVA) table:

Source	*df*	SS	MS
Regression	1		
Residual	$n-2$		
Total	$n-1$		

The Total *df* is $n-1$. The Residual *df* is n minus the number of parameters (2) estimated by the LS line. The Regression *df* is the number of predictor variables (1) in the model. A Mean Square is always equal to the Sum of Squares divided by the *df*. SW use the following notation for the Residual MS: $s^2_{Y|X} = Resid(SS)/(n-2)$.

Brief Discussion of Minitab Output for Blood Loss Problem

1. Identify fitted line: Blood Loss = 552.44 - 1.30 Weight (i.e. $b_0 = 552.44$ and $b_1 = -1.30$).
2. Locate Analysis of Variance Table. More on this later.
3. Locate Parameter Estimates Table. More on this later.
4. Note that $R^2 = .5967 = (-.77247)^2 = r^2$.

The regression model

The following statistical model is assumed as a means to provide error estimates for the LS line, regression coefficients, and predictions. Assume that the data (X_i, Y_i), $i = 1, ..., n$ are a sample of (X, Y) values from the population of interest, and

1. The mean in the population of all responses Y at a given X value (called $\mu_{Y|X}$ by SW) falls on a straight line, $\beta_0 + \beta_1 X$, called the population regression line.

2. The variation among responses Y at a given X value is the same for each X, and is denoted by $\sigma^2_{Y|X}$.

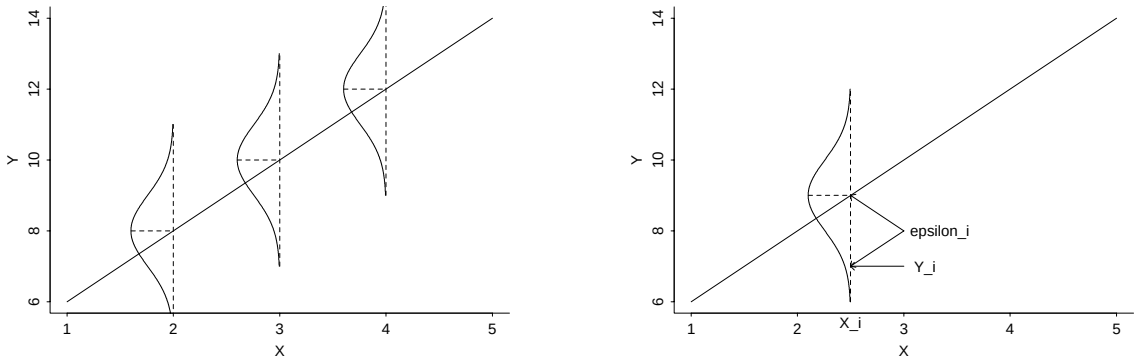
3. The population of responses Y at a given X is normally distributed.

4. The pairs (X_i, Y_i) are a random sample from the population. Alternatively, we can think that the X_is were fixed by the experimenter, and that the Y_i are random responses at the selected predictor values.

The model is usually written in the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

(i.e. Response = Mean Response + Residual), where the ϵ_is are, by virtue of assumptions 2, 3 and 4, independent normal random variables with mean 0 and variance $\sigma^2_{Y|X}$. The following picture might help see this. Note that the population regression line is unknown, and is estimated from the data using the LS line.



Back to the Data

There are three unknown population parameters in the model: β_0, β_1 and $\sigma^2_{Y|X}$. Given the data, the LS line

$$\hat{Y} = b_0 + b_1 X$$

estimates the population regression line $\beta_0 + \beta_1 X$. The LS line is our best guess about the unknown population regression line. Here b_0 estimates the intercept β_0 of the population regression line and b_1 estimates the slope β_1 of the population regression line.

The i^{th} **observed residual** $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 X_i$ is the i^{th} **fitted value**, estimates the **unobservable residual** ϵ_i. (ϵ_i is unobservable because β_0 and β_1 are unknown.) See the picture on page 10 to refresh your memory on the notation. The Residual MS from the ANOVA table is used to estimate $\sigma^2_{Y|X}$:

$$s^2_{Y|X} = \text{Res MS} = \frac{\text{Res SS}}{\text{Res df}} = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}.$$

CI and tests for β_1

A CI for β_1 is given $b_1 \pm t_{crit} SE_{b_1}$, where the standard error of b_1 under the model is

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum_i (X_i - \overline{X})^2}},$$

and where t_{crit} is the appropriate critical value for the desired CI level from a $t-$distribution with df =Res df.

To test $H_0 : \beta_1 = \beta_{1,0}$ (a given value) against $H_A : \beta_1 \neq \beta_{1,0}$, reject H_0 if $|t_s| \geq t_{crit}$, where

$$t_s = \frac{b_1 - \beta_{1,0}}{SE_{b_1}},$$

and t_{crit} is the $t-$critical value for a two-sided test, with the desired size and df =Res df. Alternatively, you can evaluate a p-value in the usual manner to make a decision about H_0.

The parameter estimates table in Minitab gives the standard error, $t-$statistic, and p-value for testing $H_0 : \beta_1 = 0$. Analogous summaries are given for the intercept, but these are typically of less interest.

Testing $\beta_1 = 0$

Assuming the mean relationship is linear, consider testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. This test can be conducted using a t-statistic, as outlined above, or with an ANOVA $F-$test, as outlined below.

For the analysis of variance (ANOVA) F-test, compute

$$F_s = \frac{\text{Reg MS}}{\text{Res MS}}$$

and reject H_0 when F_s exceeds the critical value (for the desired size test) from an $F-$table with numerator $df = 1$ and denominator $df = n - 2$; see SW, page 654. The hypothesis of zero slope (or no relationship) is rejected when F_s is large, which happens when a significant portion of the variation in Y is explained by the linear relationship with X. Minitab gives the $F-$statistic and p-value with the ANOVA table output.

The p-values from the $t-$test and the $F-$test are always equal. Furthermore this p-value is equal to the p-value for testing no correlation between Y and X, using the $t-$test described earlier. Is this important, obvious, or disconcerting?

A CI for the population regression line

I can not overemphasize the **power** of the regression model. The model allows you to estimate the mean response at any X value in the range for which the model is reasonable, even if little or no data is observed at that location.

We estimate the mean population response among individuals with $X = X_p$

$$\mu_p = \beta_0 + \beta_1 X_p,$$

with the fitted value, or the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

X_p is not necessarily one of the observed X_is in the data. To get a CI for μ_p, use $\hat{Y}_p \pm t_{crit} SE(\hat{Y}_p)$, where the standard error of $\hat{Y}_p$ is

$$SE(\hat{Y}_p) = s_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}}.$$

The $t-$critical value is identical to that used in the subsection on CI for β_1.

CI for predictions

Suppose a future individual (i.e. someone not used to compute the LS line) has $X = X_p$. The best prediction for the response Y of this individual is the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

To get a CI (prediction interval) for an individual response, use $\hat{Y}_p \pm t_{crit} SE_{pred}(\hat{Y}_p)$, where

$$SE_{pred}(\hat{Y}_p) = s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}},$$

and t_{crit} is identical to the critical value used for a CI on β_1.

For example, in the blood loss problem you may want to estimates the blood loss for an $50kg$ individual, and to get a CI for this prediction. This problem is different from computing a CI for the mean blood loss of all $50kg$ individuals!

Comments

1. The prediction interval is wider than the CI for the mean response. This is reasonable because you are less confident in predicting an individual response than the mean response for all individuals.

2. The CI for the mean response and the prediction interval for an individual response become wider as X_p moves away from $\overline{X}$. That is, you get a more sensitive CI and prediction interval for X_ps near the center of the data.

3. In `Stat > Regression > Fitted Line Plot` Minitab will plot a band of 95% confidence intervals and a band of 95% prediction intervals on the data plot, along with the fitted LS line.

A further look at the blood loss data (Minitab Output)

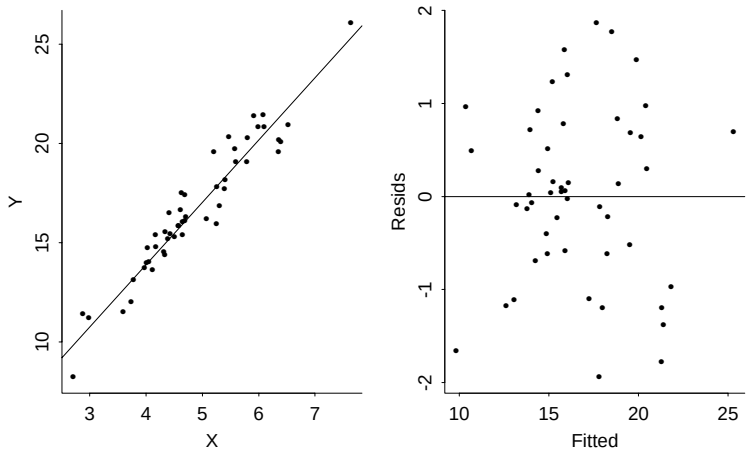
- The LS line is: Predicted Blood Loss = 552.442 - 1.30 Weight.
- The R^2 is .597 (i.e. 59.7%).
- The $F-$statistic for testing $H_0 : \beta_1 = 0$ is $F_{obs} = 8.88$ with a $p - value = .025$. The Error MS is $s^2_{Y|X} = 136.0$; see ANOVA table.
- The Parameter Estimates table gives b_0 and b_1, their standard errors, and $t-$statistics and p-values for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The $t-$test and $F-$ test p-values for testing that the slope is zero are identical. We could calculate a 95% CI for β_0 and β_1. If we did so (using the t critical value) we find we are 95% confident that the slope of the population regression line is between -2.37 and -.23.
- Suppose we are interested in estimating the average blood loss among all 50kg individuals. The estimated mean blood loss is $552.442 - 1.30033 * 50 = 487.43$. Reading off the plot, we are 95% confident that the mean blood loss of all 50kg individuals is between (approximately) 477 and 498 ml. A 95% prediction interval for the blood loss of a single 50 kg person is less precise (about 457 to 518 ml).

As a summary we might say that weight is important for explaining the variation in blood loss. In particular, the estimated slope of the least squares line (Predicted Blood loss = 552.442 - 1.30 Weight) is significantly different from zero (p-value = .0247), with weight explaining approximately 60% (59.7%) of the variation in blood loss for this sample of 8 thyroid operation patients.

Checking the regression model

A regression analysis is never complete until the assumptions of the model have been checked. In addition, you need to evaluate whether individual observations, or groups of observations, are unduly influencing the analysis. A first step in any analysis is to plot the data. The plot provides information on the linearity and constant variance assumption. For example, the data plot below shows a linear relationship with roughly constant variance.

In addition to plotting the data, a variety of methods for assessing model adequacy are based on plots of the residuals, $e_i = Y_i - \hat{Y}_i$ (i.e. Observed $-$ Fitted values). For example, an option in Minitab is to plot the e_i against the fitted values $\hat{Y}_i$, as given below. This residual plot should exhibit no systematic dependence of the sign or the magnitude of the residuals on the fitted values.

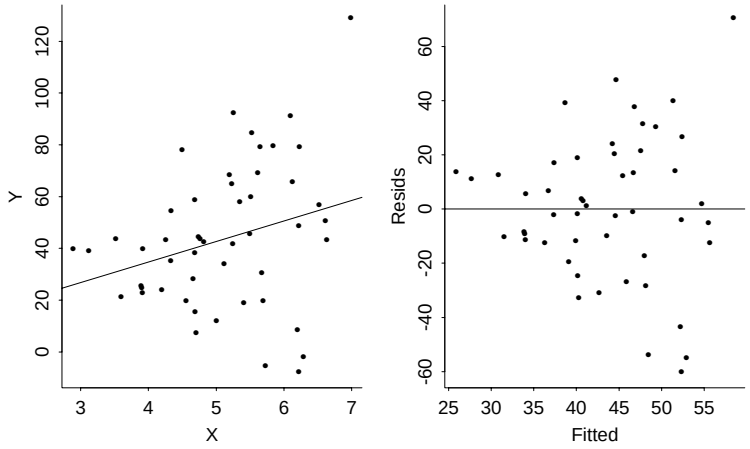


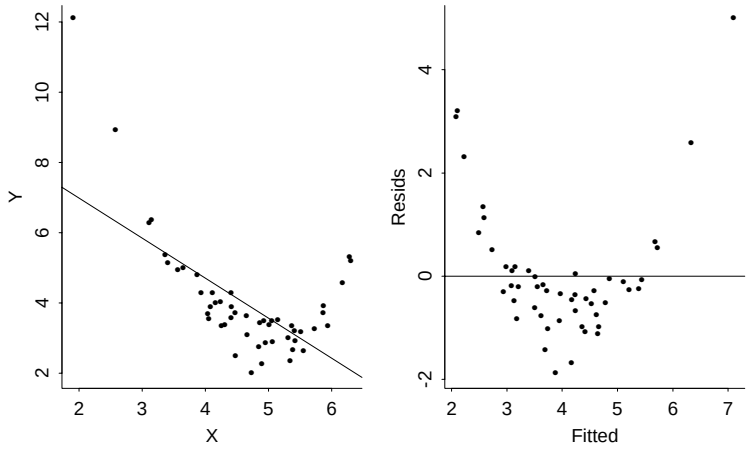
The real power of this plot is with multiple predictor problems (multiple regression). For simple linear regression, the information in this plot is similar to the information in the original data plot, except that the residual plot eliminates the effect of the trend on your perceptions of model adequacy.

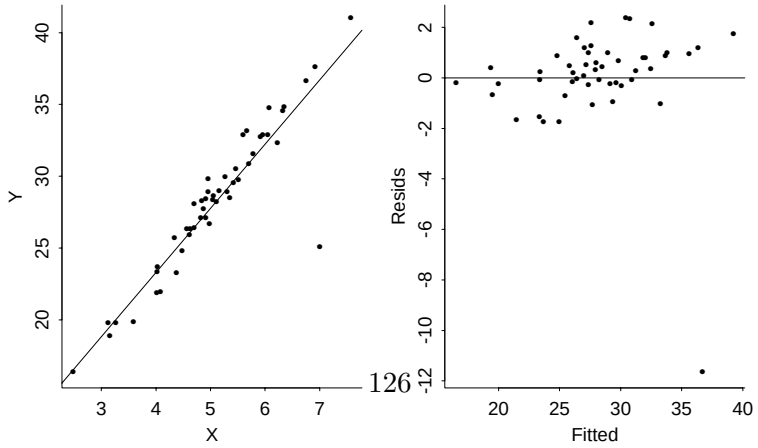
The following plots show how inadequacies in the data plot appear in a residual plot. The first plot shows a roughly linear relationship between Y and X with non-constant variance. The residual plot shows a megaphone shape rather than the ideal horizontal band. A possible remedy is a **weighted least squares** analysis to handle the non-constant variance, or to transform Y to stabilize the variance. Transforming the data may destroy the linearity.

The second plot shows a nonlinear relationship between Y and X. The residual plot shows a systematic dependence of the sign of the residual on the fitted value. A possible remedy is to transform the data.

The last plot shows an outlier. This point has a large residual. A sensible approach is to refit the model after deleting the case and see if any conclusions change.


Y
X
Resids
Fitted


Y
X
Resids
Fitted


Y
X
Resids
Fitted

Checking normality

The normality assumption can be evaluated with a boxplot or a normal quantile plot of the residuals. A formal test of normality using the residuals can be computed as discussed earlier this semester.

Checking independence

Diagnosing dependence among observations usually requires some understanding of the mechanism that generated the data. There are a variety of graphical and inferential tools for checking independence for data collected over time (called a time series). The easiest thing to do is plot the r_i against time index and look for any suggestive patterns.

Outliers

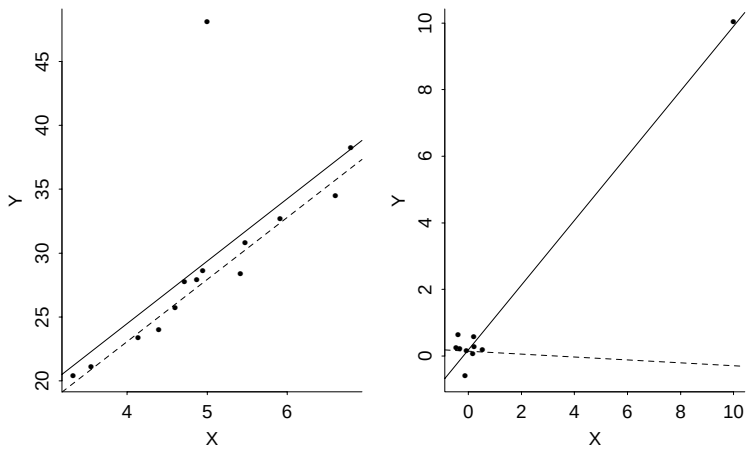
Outliers are observations that are poorly fitted by the regression model. The response for an outlier is far from the fitted line, so outliers have large positive or negative values of the residual e_i.

What do you do with outliers? Outliers may be due to incorrect recordings of the data or failure of the measuring device, or indications or a change in the mean or variance structure for one or more cases. Incorrect recordings should be fixed if possible, but otherwise deleted from the analysis.

Routine deletion of outliers from the analysis is not recommended. This practice can have a dramatic effect on the fit of the model and the perceived precision of parameter estimates and predictions. Analysts who routinely omit outliers without cause tend to overstate the significance of their findings and get a false sense of precision in their estimates and predictions. At the very least, a data analyst should repeat the analysis with and without the outliers to see whether any substantive conclusions are changed.

Influential observations

Certain data points can play a very important role in determining the position of the LS line. These data points may or may not be outliers. For example, the observation with $Y > 45$ in the first plot below is an outlier relative to the LS fit. The extreme observation in the second plot has a very small e_i. Both points are highly **influential observations** - the LS line changes dramatically when these observations are deleted. The influential observation in the second plot is not an outlier because its presence in the analysis determines that the LS line will essentially pass through it! In these plots the solid line is the LS line from the full data set, whereas the dashed line is the LS line after omitting the unusual point.



There are well defined measures of the influence that individual cases have on the LS line, and they are available in Minitab. On the separate output I calculated Cook's D (labelled COOK1) – large values indicate influential values. Which observations are most influential according to this measure? For simple linear regression most influential cases can be easily spotted by carefully looking at the data plot. If you identify cases that you suspect might be influential, you should hold them out (individually) and see if any important conclusions change. If so, you need to think hard about whether the cases should be included or excluded from the analysis.