Chapter 9

Simple Linear Regression

An analysis appropriate for a quantitative outcome and a single quantitative explanatory variable.

9.1   The model behind linear regression

When we are examining the relationship between a quantitative outcome and a single quantitative explanatory variable, simple linear regression is the most commonly considered analysis method. (The "simple" part tells us we are only considering a single explanatory variable.) In linear regression we usually have many different values of the explanatory variable, and we usually assume that values between the observed values of the explanatory variables are also possible values of the explanatory variables. We postulate a linear relationship between the population mean of the outcome and the value of the explanatory variable. If we let Y be some outcome, and x be some explanatory variable, then we can express the structural model using the equation

$$E(Y|x) = \beta_0 + \beta_1 x$$

where E(), which is read "expected value of", indicates a population mean; $Y|x$, which is read "Y given x", indicates that we are looking at the possible values of Y when x is restricted to some single value; β_0, read "beta zero", is the intercept parameter; and β_1, read "beta one". is the slope parameter. A common term for any parameter or parameter estimate used in an equation for predicting Y from

x is **coefficient**. Often the "1" subscript in β_1 is replaced by the name of the explanatory variable or some abbreviation of it.

So the structural model says that for each value of x the population mean of Y (over all of the subjects who have that particular value "x" for their explanatory variable) can be calculated using the simple linear expression $\beta_0 + \beta_1 x$. Of course we cannot make the calculation exactly, in practice, because the two parameters are unknown "secrets of nature". In practice, we make estimates of the parameters and substitute the estimates into the equation.

In real life we know that although the equation makes a prediction of the true mean of the outcome for any fixed value of the explanatory variable, it would be unwise to use **extrapolation** to make predictions *outside* of the range of x values that we have available for study. On the other hand it *is* reasonable to **interpolate**, i.e., to make predictions for unobserved x values in between the observed x values. The structural model is essentially the assumption of "linearity", at least within the range of the observed explanatory data.

It is important to realize that the "linear" in "linear regression" does *not* imply that only linear relationships can be studied. Technically it only says that the beta's must not be in a transformed form. It is OK to transform x or Y, and that allows many non-linear relationships to be represented on a new scale that makes the relationship linear.

The structural model underlying a linear regression analysis is that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is $\beta_0 + \beta_1 x$.

The error model that we use is that for each particular x, if we have or could collect many subjects with that x value, their distribution around the population mean is Gaussian with a spread, say σ^2, that is the same value for each value of x (and corresponding population mean of y). Of course, the value of σ^2 is an unknown parameter, and we can make an estimate of it from the data. The error model described so far includes not only the assumptions of "Normality" and "equal variance", but also the assumption of "fixed-x". The "fixed-x" assumption is that the explanatory variable is measured without error. Sometimes this is possible, e.g., if it is a count, such as the number of legs on an insect, but usually there is some error in the measurement of the explanatory variable. In practice,

we need to be sure that the size of the error in measuring x is small compared to the variability of Y at any given x value. For more on this topic, see the section on robustness, below.

The error model underlying a linear regression analysis includes the assumptions of fixed-x, Normality, equal spread, and independent errors.

In addition to the three error model assumptions just discussed, we also assume "independent errors". This assumption comes down to the idea that the **error** (deviation of the true outcome value from the population mean of the outcome for a given x value) for one observational unit (usually a subject) is not predictable from knowledge of the error for another observational unit. For example, in predicting time to complete a task from the dose of a drug suspected to affect that time, knowing that the first subject took 3 seconds longer than the mean of all possible subjects with the same dose should not tell us anything about how far the next subject's time should be above or below the mean for their dose. This assumption can be trivially violated if we happen to have a set of identical twins in the study, in which case it seems likely that if one twin has an outcome that is below the mean for their assigned dose, then the other twin will also have an outcome that is below the mean for their assigned dose (whether the doses are the same or different).

A more interesting cause of correlated errors is when subjects are trained in groups, and the different trainers have important individual differences that affect the trainees performance. Then knowing that a particular subject does better than average gives us reason to believe that most of the other subjects in the same group will probably perform better than average because the trainer was probably better than average.

Another important example of non-independent errors is **serial correlation** in which the errors of adjacent observations are similar. This includes adjacency in both time and space. For example, if we are studying the effects of fertilizer on plant growth, then similar soil, water, and lighting conditions would tend to make the errors of adjacent plants more similar. In many task-oriented experiments, if we allow each subject to observe the previous subject perform the task which is measured as the outcome, this is likely to induce serial correlation. And worst of all, if you use the same subject for every observation, just changing the explanatory

variable each time, serial correlation is extremely likely. Breaking the assumption of independent errors does not indicate that no analysis is possible, only that linear regression is an inappropriate analysis. Other methods such as time series methods or mixed models are appropriate when errors are correlated.

> **The worst case of breaking the independent errors assumption in regression is when the observations are repeated measurement on the same experimental unit (subject).**

Before going into the details of linear regression, it is worth thinking about the variable types for the explanatory and outcome variables and the relationship of ANOVA to linear regression. For both ANOVA and linear regression we assume a Normal distribution of the outcome for each value of the explanatory variable. (It is equivalent to say that all of the errors are Normally distributed.) Implicitly this indicates that the outcome should be a continuous quantitative variable. Practically speaking, real measurements are rounded and therefore some of their continuous nature is not available to us. If we round too much, the variable is essentially discrete and, with too much rounding, can no longer be approximated by the smooth Gaussian curve. Fortunately regression and ANOVA are both quite robust to deviations from the Normality assumption, and it is OK to use discrete or continuous outcomes that have at least a moderate number of different values, e.g., 10 or more. It can even be reasonable in some circumstances to use regression or ANOVA when the outcome is ordinal with a fairly small number of levels.

The explanatory variable in ANOVA is categorical and nominal. Imagine we are studying the effects of a drug on some outcome and we first do an experiment comparing control (no drug) vs. drug (at a particular concentration). Regression and ANOVA would give equivalent conclusions about the effect of drug on the outcome, but regression seems inappropriate. Two related reasons are that there is no way to check the appropriateness of the linearity assumption, and that after a regression analysis it is appropriate to interpolate between the x (dose) values, and that is inappropriate here.

Now consider another experiment with 0, 50 and 100 mg of drug. Now ANOVA and regression give different answers because ANOVA makes no assumptions about the relationships of the three population means, but regression assumes a linear relationship. If the truth is linearity, the regression will have a bit more power
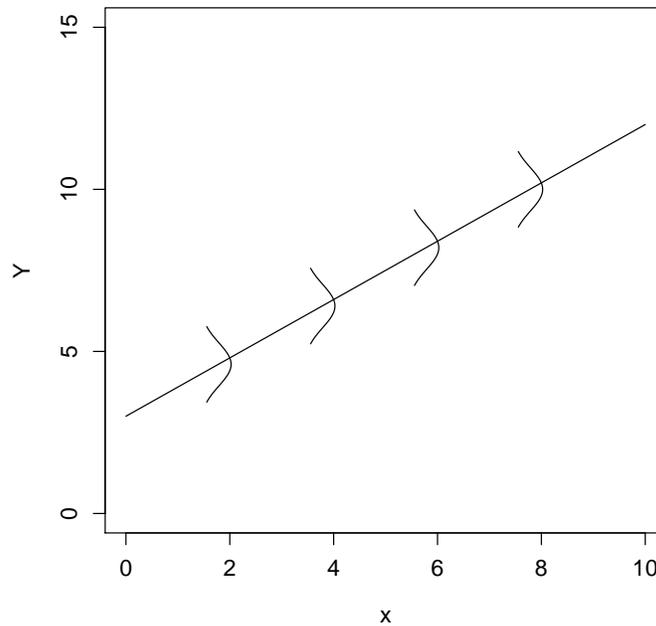
Figure 9.1: Mnemonic for the simple regression model.

than ANOVA. If the truth is non-linearity, regression will make inappropriate predictions, but at least regression will have a chance to detect the non-linearity. ANOVA also loses some power because it incorrectly treats the doses as nominal when they are at least ordinal. As the number of doses increases, it is more and more appropriate to use regression instead of ANOVA, and we will be able to better detect any non-linearity and correct for it, e.g., with a data transformation.

Figure 9.1 shows a way to think about and remember most of the regression model assumptions. The four little Normal curves represent the Normally distributed outcomes (Y values) at each of four fixed x values. The fact that the four Normal curves have the same spreads represents the equal variance assumption. And the fact that the four means of the Normal curves fall along a straight line represents the linearity assumption. Only the fifth assumption of independent errors is not shown on this mnemonic plot.

9.2   Statistical hypotheses

For simple linear regression, the chief null hypothesis is $H_0 : \beta_1 = 0$, and the corresponding alternative hypothesis is $H_1 : \beta_1 \neq 0$. If this null hypothesis is true, then, from $E(Y) = \beta_0 + \beta_1 x$ we can see that the population mean of Y is β_0 for *every* x value, which tells us that x has no effect on Y. The alternative is that changes in x are associated with changes in Y (or changes in x cause changes in Y in a randomized experiment).

Sometimes it is reasonable to choose a different null hypothesis for β_1. For example, if x is some **gold standard** for a particular measurement, i.e., a best-quality measurement often involving great expense, and y is some cheaper substitute, then the obvious null hypothesis is $\beta_1 = 1$ with alternative $\beta_1 \neq 1$. For example, if x is percent body fat measured using the cumbersome whole body immersion method, and Y is percent body fat measured using a formula based on a couple of skin fold thickness measurements, then we expect either a slope of 1, indicating equivalence of measurements (on average) or we expect a different slope indicating that the skin fold method proportionally over- or under-estimates body fat.

Sometimes it also makes sense to construct a null hypothesis for β_0, usually $H_0 : \beta_0 = 0$. This should only be done if each of the following is true. There are data that span $x = 0$, or at least there are data points near $x = 0$. The statement "the population mean of Y equals zero when $x = 0$" both makes scientific sense and the difference between equaling zero and not equaling zero is scientifically interesting. See the section on interpretation below for more information.

The usual regression null hypothesis is $H_0 : \beta_1 = 0$. Sometimes it is also meaningful to test $H_0 : \beta_0 = 0$ or $H_0 : \beta_1 = 1$.

9.3   Simple linear regression example

As a (simulated) example, consider an experiment in which corn plants are grown in pots of soil for 30 days after the addition of different amounts of nitrogen fertilizer. The data are in corn.dat, which is a space delimited text file with column headers. Corn plant final weight is in grams, and amount of nitrogen added per pot is in
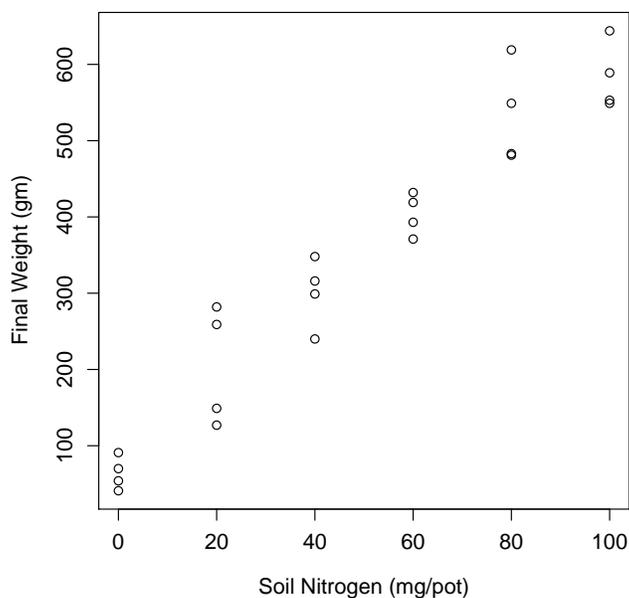
Figure 9.2: Scatterplot of corn data.

mg.

EDA, in the form of a scatterplot is shown in figure 9.2.

We want to use EDA to check that the assumptions are reasonable before trying a regression analysis. We can see that the assumptions of linearity seems plausible because we can imagine a straight line from bottom left to top right going through the center of the points. Also the assumption of equal spread is plausible because for any narrow range of nitrogen values (horizontally), the spread of weight values (vertically) is fairly similar. These assumptions should only be doubted at this stage if they are drastically broken. The assumption of Normality is not something that human beings can test by looking at a scatterplot. But if we noticed, for instance, that there were only two possible outcomes in the whole experiment, we could reject the idea that the distribution of weights is Normal at each nitrogen level.

The assumption of fixed-x cannot be seen in the data. Usually we just think about the way the explanatory variable is measured and judge whether or not it is measured precisely (with small spread). Here, it is not too hard to measure the amount of nitrogen fertilizer added to each pot, so we accept the assumption of

fixed-x. In some cases, we can actually perform repeated measurements of x on the same case to see the spread of x and then do the same thing for y at each of a few values, then reject the fixed-x assumption if the ratio of x to y variance is larger than, e.g., around 0.1.

The assumption of independent error is usually not visible in the data and must be judged by the way the experiment was run. But if serial correlation is suspected, there are tests such as the Durbin-Watson test that can be used to detect such correlation.

Once we make an initial judgement that linear regression is not a stupid thing to do for our data, based on plausibility of the model after examining our EDA, we perform the linear regression analysis, then further verify the model assumptions with residual checking.

9.4   Regression calculations

The basic regression analysis uses fairly simple formulas to get estimates of the parameters β_0, β_1, and σ^2. These estimates can be derived from either of two basic approaches which lead to identical results. We will not discuss the more complicated maximum likelihood approach here. The least squares approach is fairly straightforward. It says that we should choose as the best-fit line, that line which minimizes the sum of the squared residuals, where the **residuals** are the vertical distances from individual points to the best-fit "regression" line.

The principle is shown in figure 9.3. The plot shows a simple example with four data points. The diagonal line shown in black is close to, but not equal to the "best-fit" line.

Any line can be characterized by its intercept and slope. The intercept is the y value when x equals zero, which is 1.0 in the example. *Be sure to look carefully at the x-axis scale; if it does not start at zero, you might read off the intercept incorrectly.* The slope is the change in y for a one-unit change in x. Because the line is straight, you can read this off anywhere. Also, an equivalent definition is the change in y divided by the change in x for *any* segment of the line. In the figure, a segment of the line is marked with a small right triangle. The vertical change is 2 units and the horizontal change is 1 unit, therefore the slope is 2/1=2. Using b_0 for the intercept and b_1 for the slope, the equation of the line is $y = b_0 + b_1x$.
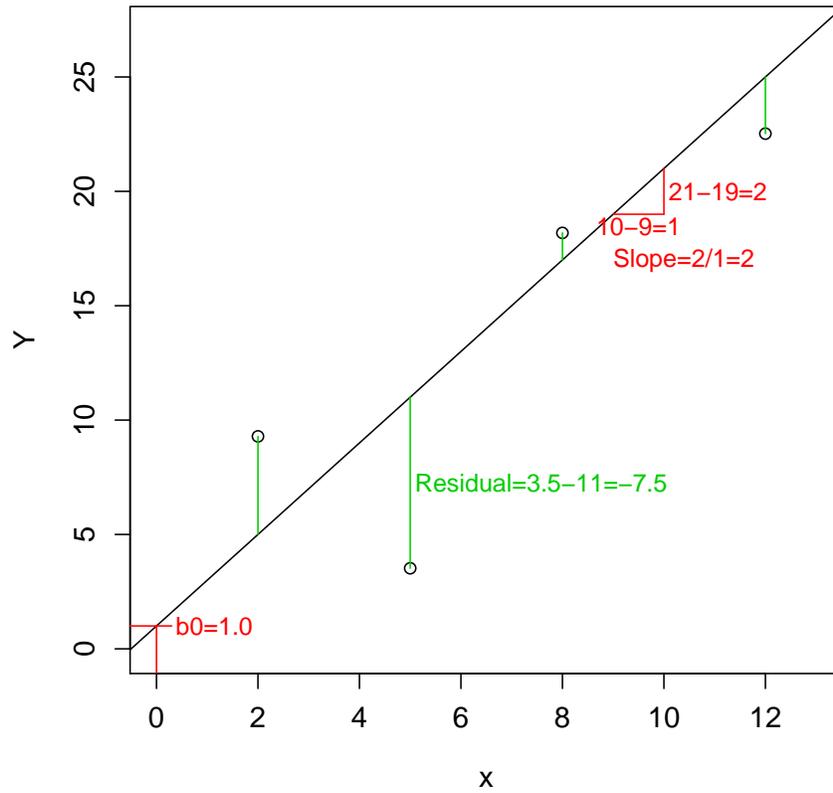
Figure 9.3: Least square principle.

By plugging different values for x into this equation we can find the corresponding y values that are on the line drawn. For any given b_0 and b_1 we get a potential best-fit line, and the vertical distances of the points from the line are called the **residuals**. We can use the symbol $\hat{y}_i$, pronounced "y hat sub i", where "sub" means subscript, to indicate the fitted or predicted value of outcome y for subject i. (Some people also use the y_i' "y-prime sub i".) For subject i, who has explanatory variable x_i, the prediction is $\hat{y}_i = b_0 + b_1 x_i$ and the residual is $y_i - \hat{y}_i$. The least square principle says that the best-fit line is the one with the smallest sum of squared residuals. It is interesting to note that the sum of the residuals (not squared) is zero for the least-squares best-fit line.

In practice, we don't really try every possible line. Instead we use calculus to find the values of b_0 and b_1 that give the minimum sum of squared residuals. You don't need to memorize or use these equations, but here they are in case you are interested.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Also, the best estimate of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

Whenever we ask a computer to perform simple linear regression, it uses these equations to find the best fit line, then shows us the parameter estimates. Sometimes the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$ are used instead of b_0 and b_1. Even though these symbols have Greek letters in them, the "hat" over the beta tells us that we are dealing with statistics, not parameters.

Here are the derivations of the coefficient estimates. SSR indicates sum of squared residuals, the quantity to minimize.

$$SSR = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{9.1}$$

$$= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2\right) \tag{9.2}$$

$$\frac{\partial SSR}{\partial \beta_0} = \sum_{i=1}^{n} (-2y_i + 2\beta_0 + 2\beta_1 x_i) \tag{9.3}$$

$$0 = \sum_{i=1}^{n} \left(-y_i + \hat{\beta}_0 + \hat{\beta}_1 x_i\right) \tag{9.4}$$

$$0 = -n\bar{y} + n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x} \tag{9.5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9.6}$$

$$\frac{\partial SSR}{\partial \beta_1} = \sum_{i=1}^{n} \left(-2x_i y_i + 2\beta_0 x_i + 2\beta_1 x_i^2\right) \tag{9.7}$$

$$0 = -\sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{9.8}$$

$$0 = -\sum_{i=1}^{n} x_i y_i + (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{9.9}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \tag{9.10}$$

A little algebra shows that this formula for $\hat{\beta}_1$ is equivalent to the one shown above because $c \sum_{i=1}^{n}(z_i - \bar{z}) = c \cdot 0 = 0$ for any constant c and variable z.

In multiple regression, the matrix formula for the coefficient estimates is $(X'X)^{-1}X'y$, where X is the matrix with all ones in the first column (for the intercept) and the values of the explanatory variables in subsequent columns.

Because the intercept and slope estimates are statistics, they have sampling distributions, and these are determined by the true values of β_0, β_1, and σ^2, as well as the positions of the x values and the number of subjects at each x value. If the model assumptions are correct, the sampling distributions of the intercept and slope estimates both have means equal to the true values, β_0 and β_1, and are Normally distributed with variances that can be calculated according to fairly simple formulas which involve the x values and σ^2.

In practice, we have to estimate σ^2 with s^2. This has two consequences. First we talk about the standard errors of the sampling distributions of each of the betas

instead of the standard deviations, because, by definition, SE's are estimates of s.d.'s of sampling distributions. Second, the sampling distribution of $b_j - \beta_j$ (for j=0 or 1) is now the t-distribution with $n-2$ df (see section 3.9.5), where n is the total number of subjects. (Loosely we say that we lose two degrees of freedom because they are used up in the estimation of the two beta parameters.) Using the null hypothesis of $\beta_j = 0$ this reduces to the null sampling distribution $b_j \sim t_{n-2}$.

The computer will calculate the standard errors of the betas, the t-statistic values, and the corresponding p-values (for the usual two-sided alternative hypothesis). We then compare these p-values to our pre-chosen alpha (usually $\alpha = 0.05$) to make the decisions whether to retain or reject the null hypotheses.

> The formulas for the standard errors come from the formula for the variance covariance matrix of the joint sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ which is $\sigma^2 (X'X)^{-1}$, where X is the matrix with all ones in the first column (for the intercept) and the values of the explanatory variable in the second column. This formula also works in multiple regression where there is a column for each explanatory variable. The standard errors of the coefficients are obtained by substituting s^2 for the unknown σ^2 and taking the square roots of the diagonal elements.
>
> For simple regression this reduces to
>
> $$SE(b_0) = s\sqrt{\frac{\sum x^2}{n\sum(x^2) - (\sum x)^2}}$$
>
> and
>
> $$SE(b_1) = s\sqrt{\frac{n}{n\sum(x^2) - (\sum x)^2}}.$$

The basic regression output is shown in table 9.1 in a form similar to that produced by SPSS, but somewhat abbreviated. Specifically, "standardized coefficients" are not included.

In this table we see the number 94.821 to the right of the "(Constant)" label and under the labels "Unstandardized Coefficients" and "B". This is called the intercept estimate, estimated intercept coefficient, or estimated constant, and can

	Unstandardized Coefficients				95% Confidence Interval for B	
	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
(Constant)	94.821	18.116	4.682	.000	47.251	122.391
Nitrogen added	5.269	.299	17.610	.000	4.684	5.889

Table 9.1: Regression results for the corn experiment.

be written as b_0, $\hat{\beta}_0$ or rarely B_0, but β_0 is incorrect, because the parameter value β_0 is a fixed, unknown "secret of nature". (Usually we should just say that b_0 equals 94.8 because the original data and most experimental data has at most 3 significant figures.)

The number 5.269 is the slope estimate, estimated slope coefficient, slope estimate for nitrogen added, or coefficient estimate for nitrogen added, and can be written as b_1, $\hat{\beta}_1$ or rarely B_1, but β_1 is incorrect. Sometimes symbols such as β_{nitrogen} or β_N for the parameter and b_{nitrogen} or b_N for the estimates will be used as better, more meaningful names, especially when dealing with multiple explanatory variables in multiple (as opposed to simple) regression.

To the right of the intercept and slope coefficients you will find their standard errors. As usual, standard errors are estimated standard deviations of the corresponding sampling distributions. For example, the SE of 0.299 for B_N gives an idea of the scale of the variability of the estimate B_N, which is 5.269 here but will vary with a standard deviation of approximately 0.299 around the true, unknown value of β_N if we repeat the whole experiment many times. The two t-statistics are calculated by all computer programs using the default null hypotheses of $H_0 : \beta_j = 0$ according to the general t-statistic formula

$$t_j = \frac{b_j - \text{hypothesized value of } \beta_j}{\text{SE}(b_j)}.$$

Then the computer uses the null sampling distributions of the t-statistics, i.e., the t-distribution with $n-2$ df, to compute the 2-sided p-values as the areas under the null sampling distribution more extreme (farther from zero) than the coefficient estimates for this experiment. SPSS reports this as "Sig.", and as usual gives the misleading output ".000" when the p-value is really "< 0.0005".

> **In simple regression the p-value for the null hypothesis $H_0 : \beta_1 = 0$ comes from the t-test for b_1. If applicable, a similar test is made for β_0.**

SPSS also gives **Standardized Coefficients** (not shown here). These are the coefficient estimates obtained when both the explanatory and outcome variables are converted to so-called **Z-scores** by subtracting their means then dividing by their standard deviations. Under these conditions the intercept estimate is zero, so it is not shown. The main use of standardized coefficients is to allow comparison of the importance of different explanatory variables in multiple regression by showing the comparative effects of changing the explanatory variables by one standard deviation instead of by one unit of measurement. I rarely use standardized coefficients.

The output above also shows the "95% Confidence Interval for B" which is generated in SPSS by clicking "Confidence Intervals" under the "Statistics" button. In the given example we can say "we are 95% confident that β_N is between 4.68 and 5.89." More exactly, we know that using the method of construction of coefficient estimates and confidence intervals detailed above, and if the assumptions of regression are met, then each time we perform an experiment in this setting we will get a different confidence interval (center and width), and out of many confidence intervals 95% of them will contain β_N and 5% of them will not.

> **The confidence interval for β_1 gives a meaningful measure of the location of the parameter and our uncertainty about that location, regardless of whether or not the null hypothesis is true. This also applies to β_0.**

9.5 Interpreting regression coefficients

It is very important that you learn to correctly and completely interpret the coefficient estimates. From $E(Y|x) = \beta_0 + \beta_1 x$ we can see that b_0 represents our estimate of the mean outcome when $x = 0$. Before making an interpretation of b_0,

first check the range of x values covered by the experimental data. If there is no x data near zero, then the intercept is still needed for calculating $\hat{y}$ and residual values, but it should not be interpreted because it is an extrapolated value.

If there are x values near zero, then to interpret the intercept you must express it in terms of the actual meanings of the outcome and explanatory variables. For the example of this chapter, we would say that b_0 (94.8) is the estimated corn plant weight (in grams) when no nitrogen is added to the pots (which is the meaning of $x = 0$). This point estimate is of limited value, because it does not express the degree of uncertainty associated with it. So often it is better to use the CI for b_0. In this case we say that we are 95% confident that the mean weight for corn plants with no added nitrogen is between 47 and 122 gm, which is quite a wide range. (It would be quite misleading to report the mean no-nitrogen plant weight as 94.821 gm because it gives a false impression of high precision.)

After interpreting the *estimate* of b_0 and it's CI, you should consider whether the *null hypothesis*, $\beta_0 = 0$ makes scientific sense. For the corn example, the null hypothesis is that the mean plant weight equals zero when no nitrogen is added. Because it is unreasonable for plants to weigh nothing, we should stop here and not interpret the p-value for the intercept. For another example, consider a regression of weight gain in rats over a 6 week period as it relates to dose of an anabolic steroid. Because we might be unsure whether the rats were initially at a stable weight, it might make sense to test $H_0 : \beta_0 = 0$. If the null hypothesis is rejected then we conclude that it is not true that the weight gain is zero when the dose is zero (control group), so the initial weight was not a stable baseline weight.

Interpret the estimate, b_0, only if there are data near zero and setting the explanatory variable to zero makes scientific sense. The meaning of b_0 is the estimate of the mean outcome when $x = 0$, and should always be stated in terms of the actual variables of the study. The p-value for the intercept should be interpreted (with respect to retaining or rejecting $H_0 : \beta_0 = 0$) only if both the equality and the inequality of the mean outcome to zero when the explanatory variable is zero are scientifically plausible.

For interpretation of a slope coefficient, this section will assume that the setting is a randomized experiment, and conclusions will be expressed in terms of causa-

tion. Be sure to substitute association if you are looking at an observational study. The general meaning of a slope coefficient is the change in Y caused by a one-unit increase in x. It is very important to know in what units x are measured, so that the meaning of a one-unit increase can be clearly expressed. For the corn experiment, the slope is the change in mean corn plant weight (in grams) caused by a one mg increase in nitrogen added per pot. If a one-unit change is not substantively meaningful, the effect of a larger change should be used in the interpretation. For the corn example we could say the a 10 mg increase in nitrogen added causes a 52.7 gram increase in plant weight on average. We can also interpret the CI for β_1 in the corn experiment by saying that we are 95% confident that the change in mean plant weight caused by a 10 mg increase in nitrogen is 46.8 to 58.9 gm.

Be sure to pay attention to the sign of b_1. If it is positive then b_1 represents the increase in outcome caused by each one-unit increase in the explanatory variable. If b_1 is negative, then each one-unit increase in the explanatory variable is associated with a *fall* in outcome of magnitude equal to the absolute value of b_1.

A significant p-value indicates that we should reject the null hypothesis that $\beta_1 = 0$. We can express this as evidence that plant weight is affected by changes in nitrogen added. If the null hypothesis is retained, we should express this as having no good evidence that nitrogen added affects plant weight. Particularly in the case of when we retain the null hypothesis, the interpretation of the CI for β_1 is better than simply relying on the general meaning of retain.

The interpretation of b_1 is the change (increase or decrease depending on the sign) in the average outcome when the explanatory variable increases by one unit. This should always be stated in terms of the actual variables of the study. Retention of the null hypothesis $H_0 : \beta_1 = 0$ indicates no evidence that a change in x is associated with (or causes for a randomized experiment) a change in y. Rejection indicates that changes in x *cause* changes in y (assuming a randomized experiment).

9.6 Residual checking

Every regression analysis should include a residual analysis as a further check on the adequacy of the chosen regression model. Remember that there is a residual value for each data point, and that it is computed as the (signed) difference $y_i - \hat{y}_i$. A positive residual indicates a data point higher than expected, and a negative residual indicates a point lower than expected.

> **A residual is the deviation of an outcome from the predicated mean value for all subjects with the same value for the explanatory variable.**

A plot of all residuals on the y-axis vs. the predicted values on the x-axis, called a **residual vs. fit plot**, is a good way to check the linearity and equal variance assumptions. A quantile-normal plot of all of the residuals is a good way to check the Normality assumption. As mentioned above, the fixed-x assumption cannot be checked with residual analysis (or any other data analysis). Serial correlation can be checked with special residual analyses, but is not visible on the two standard residual plots. The other types of correlated errors are not detected by standard residual analyses.

To analyze a residual vs. fit plot, such as any of the examples shown in figure 9.4, you should mentally divide it up into about 5 to 10 vertical stripes. Then each stripe represents all of the residuals for a number of subjects who have a similar predicted values. For simple regression, when there is only a single explanatory variable, similar predicted values is equivalent to similar values of the explanatory variable. But be careful, if the slope is negative, low x values are on the right. (Note that sometimes the x-axis is set to be the values of the explanatory variable, in which case each stripe directly represents subjects with similar x values.)

To check the linearity assumption, consider that for each x value, if the mean of Y falls on a straight line, then the residuals have a mean of zero. If we incorrectly fit a straight line to a curve, then some or most of the predicted means are incorrect, and this causes the residuals for at least specific ranges of x (or the predicated Y) to be non-zero on average. Specifically if the data follow a simple curve, we will tend to have either a pattern of high then low then high residuals or the reverse. So the technique used to detect non-linearity in a residual vs. fit plot is to find the
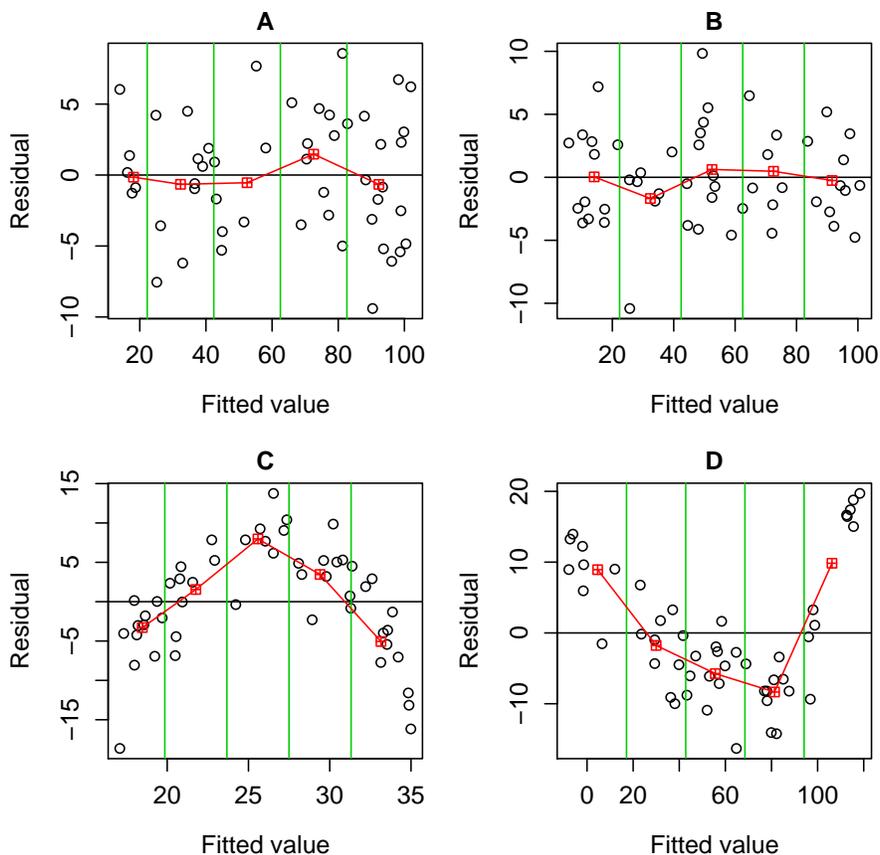
Figure 9.4: Sample residual vs. fit plots for testing linearity.

(vertical) mean of the residuals for each vertical stripe, then actually or mentally connect those means, either with straight line segments, or possibly with a smooth curve. If the resultant connected segments or curve is close to a horizontal line at 0 on the y-axis, then we have no reason to doubt the linearity assumption. If there is a clear curve, most commonly a "smile" or "frown" shape, then we suspect non-linearity.

Four examples are shown in figure 9.4. In each band the mean residual is marked, and lines segments connect these. Plots A and B show no obvious pattern away from a horizontal line other that the small amount of expected "noise". Plots C and D show clear deviations from normality, because the lines connecting the mean residuals of the vertical bands show a clear frown (C) and smile (D) pattern, rather than a flat line. Untransformed linear regression is inappropriate for the
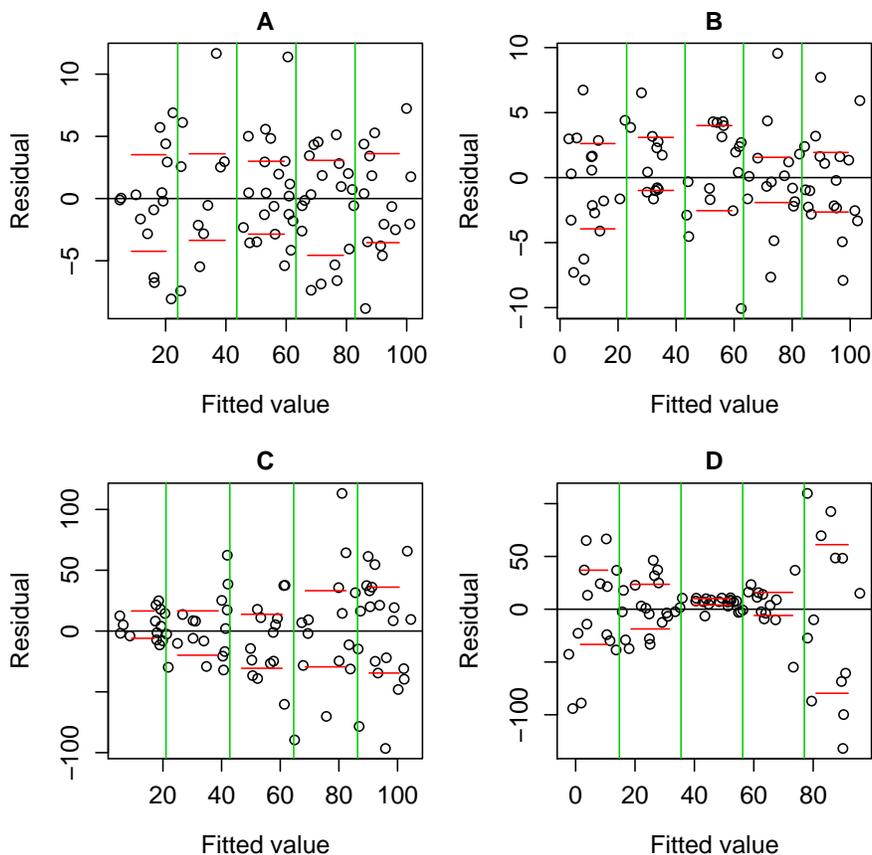
Figure 9.5: Sample residual vs. fit plots for testing equal variance.

data that produced plots C and D. With practice you will get better at reading these plots.

To detect unequal spread, we use the vertical bands in a different way. Ideally the vertical spread of residual values is equal in each vertical band. This takes practice to judge in light of the expected variability of individual points, especially when there are few points per band. The main idea is to realize that the minimum and maximum residual in any set of data is not very robust, and tends to vary a lot from sample to sample. We need to estimate a more robust measure of spread such as the IQR. This can be done by eyeballing the middle 50% of the data. Eyeballing the middle 60 or 80% of the data is also a reasonable way to test the equal variance assumption.

Figure 9.5 shows four residual vs. fit plots, each of which shows good linearity. The red horizontal lines mark the central 60% of the residuals. Plots A and B show no evidence of unequal variance; the red lines are a similar distance apart in each band. In plot C you can see that the red lines increase in distance apart as you move from left to right. This indicates unequal variance, with greater variance at high predicted values (high x values if the slope is positive). Plot D show a pattern with unequal variance in which the smallest variance is in the middle of the range of predicted values, with larger variance at both ends. Again, this takes practice, but you should at least recognize obvious patterns like those shown in plots C and D. And you should avoid over-reading the slight variations seen in plots A and B.

> **The residual vs. fit plot can be used to detect non-linearity and/or unequal variance.**

The check of normality can be done with a quantile normal plot as seen in figure 9.6. Plot A shows no problem with Normality of the residuals because the points show a random scatter around the reference line (see section 4.3.4). Plot B is also consistent with Normality, perhaps showing slight skew to the left. Plot C shows definite skew to the right, because at both ends we see that several points are higher than expected. Plot D shows a severe low outlier as well as heavy tails (positive kurtosis) because the low values are too low and the high values are too high.

> **A quantile normal plot of the residuals of a regression analysis can be used to detect non-Normality.**

9.7 Robustness of simple linear regression

No model perfectly represents the real world. It is worth learning how far we can "bend" the assumptions without breaking the value of a regression analysis.

If the linearity assumption is violated more than a fairly small amount, the regression loses its meaning. The most obvious way this happens is in the interpretation of b_1. We interpret b_1 as the change in the mean of Y for a one-unit
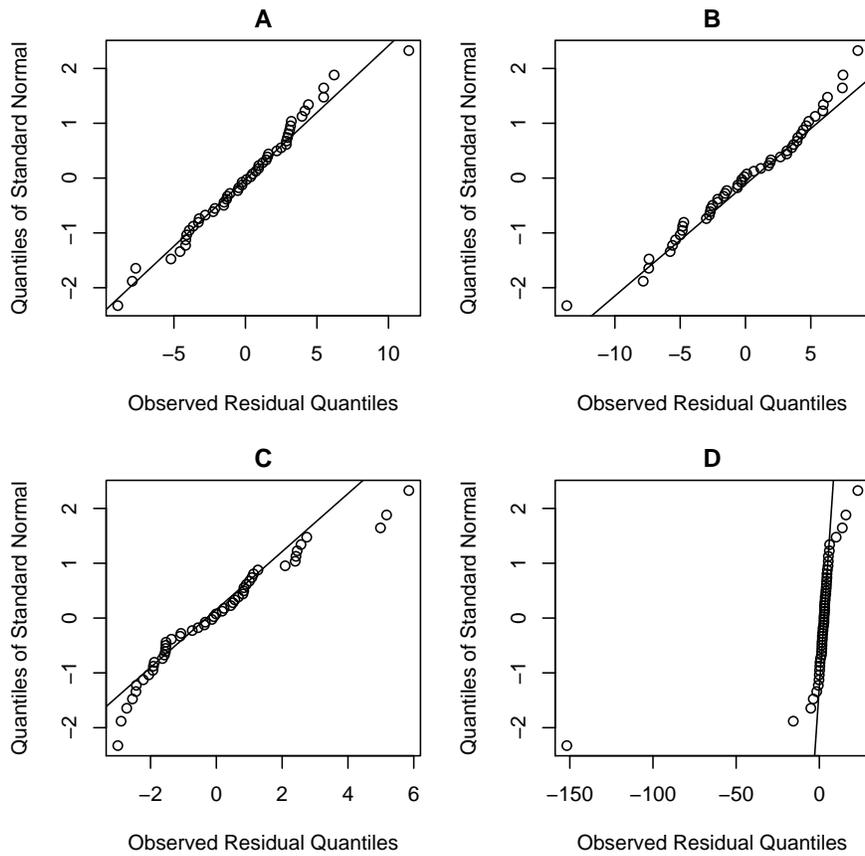
Figure 9.6: Sample QN plots of regression residuals.

increase in x. If the relationship between x and Y is curved, then the change in Y for a one-unit increase in x *varies* at different parts of the curve, invalidating the interpretation. Luckily it is fairly easy to detect non-linearity through EDA (scatterplots) and/or residual analysis. If non-linearity is detected, you should try to fix it by transforming the x and/or y variables. Common transformations are log and square root. Alternatively it is common to *add* additional new explanatory variables in the form of a square, cube, etc. of the original x variable one at a time until the residual vs. fit plot shows linearity of the residuals. For data that can only lie between 0 and 1, it is worth knowing (but not memorizing) that the square root of the arcsine of y is often a good transformation.

You should not feel that transformations are "cheating". The original way the data is measured usually has some degree of arbitrariness. Also, common measurements like pH for acidity, decibels for sound, and the Richter earthquake scale are all log scales. Often transformed values are transformed back to the original scale when results are reported (but the fact that the analysis was on a transformed scale must also be reported).

Regression is reasonably robust to the equal variance assumption. Moderate degrees of violation, e.g., the band with the widest variation is up to twice as wide as the band with the smallest variation, tend to cause minimal problems. For more severe violations, the p-values are incorrect in the sense that their null hypotheses tend to be rejected more that $100\alpha\%$ of the time when the null hypothesis is true. The confidence intervals (and the SE's they are based on) are also incorrect. For worrisome violations of the equal variance assumption, try transformations of the y variable (because the assumption applies at each x value, transformation of x will be ineffective).

Regression is quite robust to the Normality assumption. You only need to worry about severe violations. For markedly skewed or kurtotic residual distributions, we need to worry that the p-values and confidence intervals are incorrect. In that case try transforming the y variable. Also, in the case of data with less than a handful of different y values or with severe truncation of the data (values piling up at the ends of a limited width scale), regression may be inappropriate due to non-Normality.

The fixed-x assumption is actually quite important for regression. If the variability of the x measurement is of similar or larger magnitude to the variability of the y measurement, then regression is inappropriate. Regression will tend to give smaller than correct slopes under these conditions, and the null hypothesis on the

slope will be retained far too often. Alternate techniques are required if the fixed-x assumption is broken, including so-called Type 2 regression or "errors in variables regression".

The independent errors assumption is also critically important to regression. A slight violation, such as a few twins in the study doesn't matter, but other mild to moderate violations destroy the validity of the p-value and confidence intervals. In that case, use alternate techniques such as the paired t-test, repeated measures analysis, mixed models, or time series analysis, all of which model correlated errors rather than assume zero correlation.

> **Regression analysis is not very robust to violations of the linearity, fixed-x, and independent errors assumptions. It is somewhat robust to violation of equal variance, and moderately robust to violation of the Normality assumption.**

9.8 Additional interpretation of regression output

Regression output usually includes a few additional components beyond the slope and intercept estimates and their t and p-values.

Additional regression output is shown in table 9.2 which has what SPSS labels "Residual Statistics" on top and what it labels "Model Summary" on the bottom. The Residual Statistics summarize the predicted (fit) and residual values, as well as "standardized" values of these. The standardized values are transformed to Z-scores. You can use this table to detect possible outliers. If you know a lot about the outcome variable, use the unstandardized residual information to see if the minimum, maximum or standard deviation of the residuals is more extreme than you expected. If you are less familiar, standardized residuals bigger than about 3 in absolute value suggest that those points may be outliers.

The "Standard Error of the Estimate", s, is the best estimate of σ from our model (on the standard deviation scale). So it represents how far data will fall from the regression predictions on the scale of the outcome measurements. For the corn analysis, only about 5% of the data falls more than 2(49)=98 gm away from

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	84.8	611.7	348.2	183.8	24
Residual	-63.2	112.7	0.0	49.0	24
Std. Predicted Value	-1.43	1.43	0.00	1.00	24
Std. Residual	-1.26	2.25	0.00	0.978	24

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.966	0.934	0.931	50.061

Table 9.2: Additional regression results for the corn experiment.

the prediction line. Some programs report the mean squared error (MSE), which is the estimate of σ^2.

In simple regression, how close the simple correlation of x and y is to 1 or -1 is a measure of the strength of the association. Because there are many correlations in multiple regression (one for each x) we need a different way to measure the overall strength of the regression. We use a quantitiy call the R^2 value or **multiple correlation coefficient**. Note that when applied to simple regression, R^2 is equal to the *square* of the simple correlation.

R^2 can be interpreted as the fraction (or percent if multiplied by 100) of the total variation in the outcome that is "accounted for" by regressing the outcome on the explanatory variables.

A little math helps here. The total variation used in the case of R^2, is the sum of squared deviations of each y value from the mean of y, (SS$_{tot}$). Note that this quantity ignores x. Since the mean of y is the best guess of the outcome for any subject if the values of the explanatory variables are unknown, we can think of total variation as measuring how well we can predict y without knowing x.

If we perform regression and then focus on the residuals, we can square and then sum these residual to get SS_{res}. The better x helps to predict y, the smaller the residuals will be, and therefore, the smaller SS$_{res}$ will be. We can interpret SS_{res} as a measure of "unexplained variablility".

If we compute total minus residual variability (SS$_{tot}$ − SS$_{res}$) we can call the result "explained variability". It represents the amount of variability in y that is

explained away by regressing on x. Then we can compute R^2 as

$$R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{\text{SS}_{tot} - \text{SS}_{res}}{\text{SS}_{tot}}.$$

So R^2 is the portion of the total variation in Y that is explained away by using the x information in a regression. R^2 is always between 0 and 1. An R^2 of 0 means that x provides no information about y. An R^2 of 1 means that use of x information allows perfect prediction of y with every point of the scatterplot exactly on the regression line. Anything in between represents different levels of closeness of the scattered points around the regression line.

So for the corn problem we can say the 93.4% of the total variation in plant weight can be explained by regressing on the amount of nitrogen added. Unfortunately, there is no clear general interpretation of the values of R^2. While $R^2 = 0.6$ might indicate a great finding in social sciences, it might indicate a very poor finding in a chemistry experiment.

R^2 **is a measure of the fraction of the total variation in the outcome that can be explained by the explanatory variable. It runs from 0 to 1, with 1 indicating perfect prediction of** y **from** x**.**

9.9 Using transformations

If you find a problem with the equal variance or Normality assumptions, you will probably want to see if the problem goes away if you use $\log(y)$ or y^2 or $\sqrt{y}$ or $1/y$ instead of y for the outcome. (It never matters whether you choose natural vs. common log.) For non-linearity problems, you can try transformation of x, y, or both. If regression on the transformed scale appears to meet the assumptions of linear regression, then go with the transformations. In most cases, when reporting your results, you will want to back transform point estimates and the ends of confidence intervals for better interpretability. By "back transform" I mean do the inverse of the transformation to return to the original scale. The inverse of common log of y is 10^y; the inverse of natural log of y is e^y; the inverse of y^2 is $\sqrt{y}$; the inverse of $\sqrt{y}$ is y^2; and the inverse of $1/y$ is $1/y$ again. *Do not transform a p-value – the p-value remains unchanged.*

Here are a couple of examples of transformation and how the interpretations of the coefficients are modified. If the explanatory variable is dose of a drug and the outcome is log of time to complete a task, and $b_0 = 2$ and $b_1 = 1.5$, then we can say the best estimate of the log of the task time when no drug is given is 2 or that the the best estimate of the time is $10^2 = 100$ or $e^2 = 7.39$ depending on which log was used. We also say that for each 1 unit increase in drug, the log of task time increases by 1.5 (additively). On the original scale this is a *multiplicative* increase of $10^{1.5} = 31.6$ or $e^{1.5} = 4.48$. Assuming natural log, this says every time the dose goes up by another 1 unit, the mean task time get multiplied by 4.48.

If the explanatory variable is common log of dose and the outcome is blood sugar level, and $b_0 = 85$ and $b_1 = 18$ then we can say that when log(dose)=0, blood sugar is 85. Using $10^0 = 1$, this tells us that blood sugar is 85 when dose equals 1. For every 1 unit increase in log dose, the glucose goes up by 18. But a one unit increase in log dose is a ten fold increase in dose (e.g., dose from 10 to 100 is log dose from 1 to 2). So we can say that every time the dose increases 10-fold the glucose goes up by 18.

Transformations of x or y to a different scale are very useful for fixing broken assumptions.

9.10   How to perform simple linear regression in SPSS

To perform simple linear regression in SPSS, select Analyze/Regression/Linear... from the menu. You will see the "Linear Regression" dialog box as shown in figure 9.7. Put the outcome in the "Dependent" box and the explanatory variable in the "Independent(s)" box. I recommend checking the "Confidence intervals" box for "Regression Coefficients" under the "Statistics..." button. Also click the "Plots..." button to get the "Linear Regression: Plots" dialog box shown in figure 9.8. From here under "Scatter" put "*ZRESID" into the "Y" box and "*ZPRED" into the "X" box to produce the residual vs. fit plot. Also check the "Normal probability plot" box.
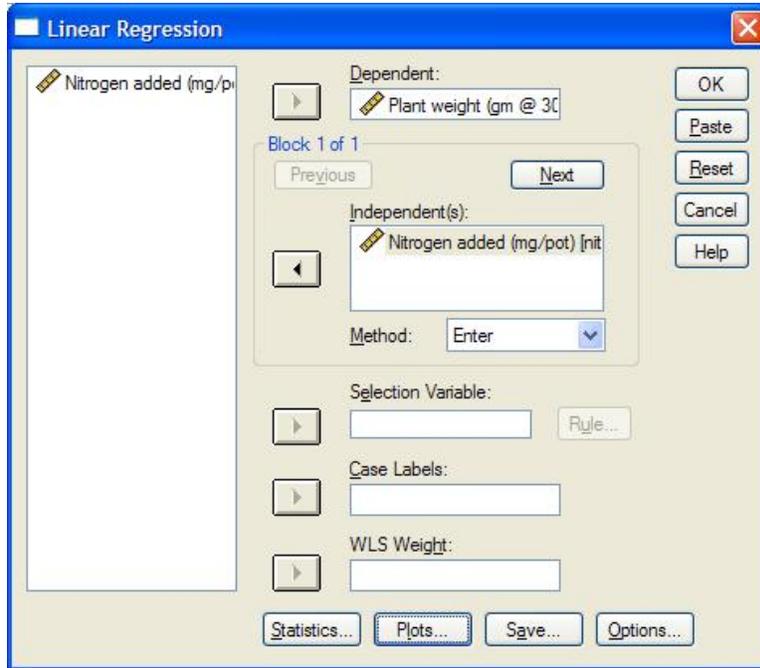
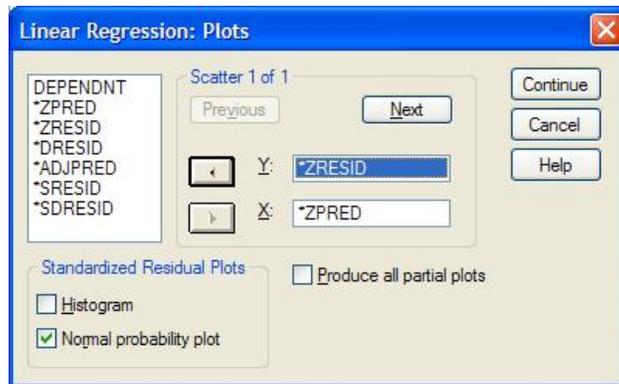Figure 9.7: Linear regression dialog box.

Figure 9.8: Linear regression plots dialog box.

In a nutshell: Simple linear regression is used to explore the relationship between a quantitative outcome and a quantitative explanatory variable. The p-value for the slope, b_1, is a test of whether or not changes in the explanatory variable really are associated with changes in the outcome. The interpretation of the confidence interval for β_1 is usually the best way to convey what has been learned from a study. Occasionally there is also interest in the intercept. No interpretations should be given if the assumptions are violated, as determined by thinking about the fixed-x and independent errors assumptions, and checking the residual vs. fit and residual QN plots for the other three assumptions.