

Regression analysis

In this handout we take a look at regression models. These are undoubtedly the most important, and most widely-used statistical methods around.

The most common uses of statistical methods are modeling and prediction: investigating and establishing a relationship between, or among, a set of variables.

- (a) Sir Francis Galton (1885): the relationship between the heights of parents and the heights of their children. The term originates from Galton’s observation of “regression to the mean,” which he observed in the context of the relationship between heights of parents and heights of their children. Regression to the mean comes from the natural variability in the population in (virtually) any relationship. For example, consider two children of different parents. Even if the heights of the parents are identical, we don’t expect the children to have exactly the same height, due to random nongenetic (environmental) determinants of height. Unusually tall parents (say 78.7 inches, or 2 meters) might be either taller or shorter than what their genetics would have determined, but the former is more likely, since there are many more people whose heights based on genetics are predicted to be less than 2 meters than there are people whose heights are predicted to be more than 2 meters. Thus, while tall parents are likely to be taller than average genetically, the observed parental heights overestimate that genetic predisposition handed down to children, and a child is likely to be shorter than the parent (although probably still taller than the average in the entire population). Similarly, unusually short parents are more likely to have had random environmental effects that “shortened” them, implying taller children than themselves.

- (1) Retesting in standardized tests.
- (2) Mutual funds.
- (3) Legal and social interventions.
- (4) *Sports Illustrated* cover jinx.

Does regression to the mean imply that heights are converging to a central value, or that eventually everyone would get the same score on tests? Of course not. The same argument that says that people who score well on the first test will score less well on the second test can be reversed: people who score unusually well on the second test probably scored less well on the first test, since they were more likely to have been a bit lucky on the second test. That is, there is also “regression away from the mean” — people who are unusually close to average on the first test are likely to be further

away from the mean on the second test.

The mistaken impression that regression to the mean implies lessening inherent variability over time is an easy trap to fall into. Nobel laureate W.F. Sharpe, in his 1985 text *Investments, 3rd. ed.*, studied the profits of the 20% of firms with highest profits in 1966 and the 20% of firms with lowest profits in 1966. In 1980 the profits of both groups were closer to the mean, leading to his conclusion that “ultimately, economic forces will force the convergence of the profitability and growth rates of different firms” (p. 430). This is not correct; rather, firms that were lucky (or unlucky) enough to do particularly well (or poorly) are not likely to stay as lucky (or unlucky) in the future. By the same token, firms that had profit unusually close to the mean profit in 1966 were no doubt further from the mean profit in 1980, again just due to random fluctuation.

- (b) Economics: the relationship between the supply or demand of a product and its price.
- (c) Finance: predicting the prepayment of mortgages from loan information.
- (d) Environmental science: the EPA projects 60,000 deaths from microscopic particles of soot and dust. How?

Regression analyses have three basic purposes:

- (1) to **predict** the values of a particular variable from one or several other variables
- (2) to try to establish the relationship for explanatory (**model building**) purposes
- (3) to determine if any apparent relationship is just because of random chance or if it reflects a genuine relationship (**testing**).

We will restrict ourselves to *linear relationships*:

$$y = 10 + 3x_1 - 5x_2.$$

They’re easier to understand, are easier to handle mathematically, and (most importantly) do a surprisingly good job in many different applications.

The multiple regression model

The functional model that we will be considering has the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i. \tag{1}$$

The variable y is the **dependent** or **target** or **response** variable, while the x variables are the **independent** or **predictor** variables. Here i indexes the observation number,

with y_i the i th observation of y , x_{1i} the i th observation of x_1 , x_{2i} the i th observation of x_2 , and so on; $\beta_0, \beta_1, \dots, \beta_p$ are unknown parameters; and ε_i is the i th error value. *Simple regression* corresponds to the special case where there is only one predictor,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (2)$$

In that case the model is consistent with imagining a straight line going through a cloud of points; in the multiple regression case it corresponds to a plane or a hyperplane.

Say we find that there is a tendency for y to vary as \mathbf{x} varies, and vice versa. The model (1) seems to imply that the x 's **cause** y , but we cannot assume that. Remember — **correlation does not imply causation!** As George Box has said, “To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it).” That is, we cannot infer causality statistically from observational data. We should also understand that “causality” in this context is not referring to a proposed deterministic causal link. This is not physics, where pushing an object definitely *causes* it to move; this is not chemistry, where adding a catalyst to a reaction definitely *causes* the reaction to speed up. For random processes, causation doesn't mean that if A occurs, B **must** occur; rather, causation means that if A occurs, that *causes* a change in the **probability** that B occurs.

We would like to estimate $\beta_0, \beta_1, \dots, \beta_p$ (that is, choose a reasonable line, plane, or hyperplane to represent the relationship between \mathbf{x} and y). In order to do that effectively, we need to make some assumptions about the error term ε_i . If the assumptions hold, our estimates of $\beta_0, \beta_1, \dots, \beta_p$ should be accurate and as precise as they can be; if they do not hold, we need to do something else to estimate them. This is the same idea as assuming that a population is reasonably Gaussian when we use the sample mean as a location estimate or construct a confidence or prediction interval — if we have long tails or outliers, we need to do something different, such as use the median or use transformations judiciously.

It is important to realize that these are **not** merely theoretical issues. If you go to a regression package and blindly obtain estimates of $\beta_0, \beta_1, \dots, \beta_p$ without worrying about (and checking) assumptions, the estimates are likely to be **inappropriate** and **incorrect**.

The table on the next page summarizes the assumptions, and problems associated with their violation.

Assumption	What does it really mean?	When is it likely to be violated?	Why is it a problem?
$E(\varepsilon_i) = 0$ for all i	It cannot be the case that some members of the population have y value that is systematically below the regression line, while others have y value systematically above it.	Well-defined subgroups in the data can cause this problem. For example, if $x \equiv$ Years on the job, and $y \equiv$ Salary, and women are systematically underpaid, they will have $E(\varepsilon_i) < 0$, while the men have $E(\varepsilon_i) > 0$.	Estimates of β_0 will be inappropriate. More importantly, a part of the signal is being mistakenly treated as noise.
$V(\varepsilon_i) = \sigma^2$ for all i (homoscedasticity)	It cannot be the case that the x/y relationship is stronger for some members of the population, and weaker for others (heteroscedasticity).	Well-defined subgroups in the data can cause this problem. For example, it could be the case that the salaries of women vary less around their typical values than those of men. Another possible cause is if the data vary over a wide range. Say, e.g., that $y \equiv$ Revenues of a firm, while $x \equiv$ the Advertising budget. It is reasonable to expect that it would be possible to predict revenues more accurately for smaller firms than for larger ones.	Estimates of the parameters will be less precise than they could be. More importantly, assessments of predictive power will be incorrect.
ε_i and ε_j are not related to each other for $i \neq j$.	It cannot be the case that knowing that the value of y for the i^{th} case is, e.g., below its expected value tells us anything about whether the value of y for another case is above or below its expected value.	This occurs most often for time series data. It is quite likely that if, e.g., sales of a product are higher than expected in July, they will also be higher than expected in June and August.	Measures of the strength of the relationship between x and y can be very misleading.
$\varepsilon_i \sim N(0, \sigma^2)$	The errors are normally distributed.	Can happen any time.	Confidence and prediction intervals, and hypothesis tests, can be misleading.

Note, by the way, that the examples of when these assumptions could be violated given in the table are not exhaustive; model violations can arise in other situations as well. For example, if the different response values are themselves summary statistics from samples, those different summary statistics could have different variances if they are based on different sample sizes (because of the Law of Large Numbers). Nonindependence of errors could occur in a non-time series situation if different observations are based on the same underlying object, for example if response values for sets of different observations are actually repeated blood pressure measurements for the same patient.

Least squares estimation

This describes our model, but it doesn't directly answer the question of how to estimate the parameters. Any choice of estimated values ($\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, respectively) corresponds to a specific straight line, plane, or hyperplane; how do we decide if it is a good one or a bad one?

We need a criterion that corresponds to some measure of closeness. We need to compare the observed target values, y_i , to the so-called *fitted values*, the best guesses for y_i based on the chosen equation, which equal $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}$. It seems clear that the criterion should be based on the difference between these two, the *residual*, $y_i - \hat{y}_i$.

- $\sum_{i=1}^n (y_i - \hat{y}_i)$?
- $\sum_{i=1}^n |y_i - \hat{y}_i|$?
- $\sum_{i=1}^n (y_i - \hat{y}_i)^2$?

The estimates that minimize the last-named criterion are called the **least squares regression estimates** (Gauss, 1795?; Legendre, 1805). The actual minimizers can be determined using calculus (this is actually a straightforward application of multivariate calculus). For simple regression the estimates have the following form:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{(n - 1)s_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

We won't actually use these formulas; we'll use the computer to calculate the least squares estimates (the formulas used by the computer are more computationally efficient and numerically stable than these formulas). For multiple regression there aren't any simple formulas like this, but it is still straightforward for the computer to determine the least squares estimates.

What do these estimates actually mean? Each means something very specific, and it is crucial not to get it wrong:

$\hat{\beta}_0$ is the estimated expected value of the response variable y when all of the predicting variables x_j equal zero. Note that in many situations $x_j = 0$ might be meaningless for some predictor(s), or you might have no data where the observed x_j values are anywhere near 0; in that situation, β_0 has no physical meaning, and there is no reason to spend any trying to interpret $\hat{\beta}_0$ (that doesn't mean that you don't need it in the model; only that it has no physical interpretation).

$\hat{\beta}_j$ is the estimated expected change in y associated with a one unit change in x_j holding all else in the model fixed. Consider the following example. Say we take a sample of college students and determine their College grade point average (GPA), High school GPA, and SAT score. We then build a model of College GPA as a function of High school GPA and SAT:

$$\text{College GPA} = 1.3 + .7 \times \text{High school GPA} - .0003 \times \text{SAT}.$$

It is tempting to say (and many people do) that the coefficient for SAT has the “wrong sign,” because it says that higher values of SAT are associated with lower values of College GPA. **This is absolutely incorrect!** What it says is that higher values of SAT are associated with lower values of College GPA, holding High school GPA fixed. High school GPA and SAT are no doubt correlated with each other, so changing SAT by one unit holding High school GPA fixed may not ever happen! **The coefficients of a multiple regression are conditional, given everything else in the model, and must not be interpreted marginally!** If you really are interested in the relationship between College GPA and just SAT, you should simply do a regression of College GPA on only SAT. **Note also the use of the words “associated with”; this reflects the fact that regression only uncovers associations, and does not prove causation.** You should always use the words “associated with” when describing the interpretation of a slope coefficient; you should never say “a change in x leads to a change in y ,” “results in a change in y ,” or any other statement that could be interpreted causally.

Say we performed a regression relating the monthly percentage stock price change (i.e., the return) of a particular stock (y) to the percentage change in the Standard & Poor's 500 Index (x) for 4 years of data, and got the following fitted model:

$$\hat{y} = -.492 + 1.253x$$

(this is an example of the *Capital Asset Pricing Model*, CAPM). What does this model say? The constant term (called “alpha” in the CAPM application) says that the estimated expected percentage change in the stock price in a flat market ($x = 0$) is $-.492$, or a drop of about one-half of one percent, which is obviously pretty close to zero itself. The slope term (called “beta” in the CAPM application) says that a one unit (that is, one percentage point) increase in the return of the market as a whole is associated with an estimated expected increase in the stock’s price of 1.253 percentage points, or (equivalently), a month with S&P return one percentage point higher than another month’s S&P return is estimated to have an expected return for this stock 1.253 percentage points higher (thus, this stock is more volatile than the market itself). Note that you should always report the interpretations of coefficients in meaningful terms when summarizing the results of a regression (i.e., “1 percentage point” or “\$1” as opposed to “one unit”).

If we added other variables to this model, the regression coefficients would be conditional, given everything else in the model. For example, the “three factor” model of Eugene Fama and Kenneth French adds two additional predictors, which refer to the difference in returns between small capitalization and large capitalization portfolios and the difference in return between value stock portfolios (securities that appear to be underpriced based on fundamental analysis) and growth stock portfolios (securities that historically achieve high returns on equity). In that formulation, the estimated slope of S&P return represents the estimated expected change in the stock’s return given a one percentage point change in the market return holding the two additional factors fixed.

One of the most useful aspects of multiple regression is its ability to statistically represent a conditioning action that would otherwise be impossible. In experimental situations, it is common practice to change the setting of one experimental condition while holding others fixed, thereby isolating its effect, but this is not possible with observational data. Multiple regression provides a statistical version of this practice. This is the reasoning behind the use of “control variables” in multiple regression — variables that are not necessarily of direct interest, but ones that the researcher wants to “correct for” in the analysis, such as demographic variables.

Example. *Dinner prices in Manhattan*

Estimating σ

The estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ tell us something about the location of the regression relationship, but say nothing about the variability of the relationship. Is the relationship

between the x 's and y a strong one or a weak one? This can be measured in our model by σ^2 , the variance of the error term, which is estimated by the **residual mean square**,

$$\text{Residual MS} = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}$$

(we divide by $n - p - 1$ because we are estimating $p + 1$ parameters, $\beta_0, \beta_1, \dots, \beta_p$). **The square root of the residual mean square estimates σ , and is called the standard error of the estimate** (Minitab calls it s in the output).

The standard error of the estimate helps us to determine if the observed relationship is of any practical importance. If the errors are roughly normally distributed, we know that roughly 95% of the population values are within $\pm 2\sigma$ of the regression relationship; that is, ± 2 times the standard error of the estimate is a rough 95% prediction interval for where we think observations are likely to be off the regression relationship. If this is relatively small, then \mathbf{x} is doing a good job of predicting y .

Proportion of variability accounted for by the regression

The following relationship is a tautology:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

For linear least squares regression, it turns out that

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Total SS = Residual SS + Regression SS.

An effective regression would have small Residual SS compared to Total SS; that is,

$$\frac{\text{Residual SS}}{\text{Total SS}}$$

would be close to zero, or equivalently

$$R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}} = \frac{\text{Regression SS}}{\text{Total SS}}$$

would be close to 1. R^2 is called the **coefficient of determination**, and is an **estimate of the proportion of variability in the target variable accounted for by the regression**. Note that R^2 is an estimate of a population value ρ^2 , and for simple regression $R^2 = r^2$ (where r is the correlation between x and y). R^2 satisfies $0 \leq R^2 \leq 1$,

where $R^2 = 0$ corresponds to no observed linear relationship between \mathbf{x} and y and $R^2 = 1$ corresponds to perfect prediction of the response from the predictors based on a linear relationship. It turns out the R^2 is biased upwards slightly as an estimate of ρ^2 , so we sometimes use the **adjusted R^2** ,

$$R_a^2 = R^2 - \frac{p}{n - p - 1}(1 - R^2).$$

What is a “large” value of R^2 ? This depends on the context of the data. In the physical sciences, for example, modeled relationships must be extremely accurate before they will even be considered, so R^2 values must be very close to 1. In the social sciences, on the other hand, much lower R^2 values can be indicative of important relationships.

Another way to see how the R^2 is a measure of how well the model fits the model is to recognize that the following is true:

$$R^2 = [\text{corr}(\mathbf{y}, \hat{\mathbf{y}})]^2;$$

that is, the R^2 is just the correlation between the observed and fitted response values, squared. Thus, in a very real sense, the R^2 is telling you how well the predictions from the model track the actual responses (at least in the sample on which the model is being fit). This suggests that a plot of y_i versus \hat{y}_i provides a graphical summary that corresponds to the numerical one given by the R^2 value, with a stronger relationship implying better predictive power for the model.

Example. Salaries and performance measures: baseball players and CEOs.

Inference in multiple regression

Inferential questions arise naturally in the regression context. The first question is concerned with a very basic question — is anything actually going on? Is an observed pattern of \mathbf{x} and y values reflecting a genuine relationship, or is it simply due to random chance? This is a hypothesis testing question. Looking again at the basic regression equation,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i,$$

we can see that all of the slope terms equaling zero corresponds to no relationship between \mathbf{x} and y , so **a test of the hypotheses**

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

versus

$$H_a : \text{some } \beta_j \neq 0, \quad j = 1, \dots, p$$

is a test of whether there is any relationship between x and y . The test of these hypotheses is the **F-test**:

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{\text{Regression SS}/p}{\text{Residual SS}/(n - p - 1)}. \quad (3)$$

This is compared to a critical value for an F-distribution on $(p, n - p - 1)$ degrees of freedom.

A different question is whether a particular variable x_j adds any predictive power to the model given the other predictors in the model. If we look at the regression model, we can see that a slope term equaling zero corresponds to no relationship between that predictor and y given the other predictors in the model, so **a test of the hypotheses**

$$H_0 : \beta_j = 0, \quad j = 1, \dots, p$$

versus

$$H_a : \beta_j \neq 0$$

is a test of whether there is a relationship between a predictor and the response given the other predictors in the model. This is tested using a **t-test**:

$$t_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}, \quad (4)$$

which is compared to a critical value for a t -distribution on $n - p - 1$ degrees of freedom. A t -test for whether the intercept β_0 equals 0 also can be constructed in this way, but as was noted earlier, this might not have any physical meaning, and hence not be of any practical interest. More generally, other values of β_j can be specified in the null hypothesis (say β_{j0}), with the t -statistic becoming

$$t_j = \frac{\hat{\beta}_j - \beta_{j0}}{\text{s.e.}(\hat{\beta}_j)}.$$

A confidence interval for β_j can help indicate if a relationship is meaningful:

$$\hat{\beta}_j \pm t_{\alpha/2}^{(n-2)} \text{s.e.}(\hat{\beta}_j).$$

Note that in the case of simple regression, the hypotheses

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

and

$$H_0 : \beta_1 = 0$$

are identical to each other. This means that the overall F -test (3) and the t -test for the slope (4) are testing the same thing, and unsurprisingly, they are guaranteed to give the same result. It is in the multiple regression case that the distinction between the two becomes important.

The usual caveats that apply to hypothesis tests apply here. For very large sample sizes, a very weak relationship can still be statistically significant (that is, **statistical significance and practical importance are not the same thing**).

Using 0/1 variables as predictors

It should be remembered that 0/1 (indicator) variables can be used as predictors in a multiple regression. Consider, for example, doing a regression of Math SAT score on High school GPA and a 0/1 variable defining gender (Gender = 0 for male, 1 for female), and getting the following result:

$$\text{Math SAT} = 400 + 73 \times \text{High school GPA} - 20 \times \text{Gender}.$$

The coefficient estimate of -20 has the following appealing interpretation: given High school GPA, the estimated expected difference in Math SAT between men and women is -20 (that is, holding High school GPA fixed, women average an estimated 20 points lower). This is called a “constant shift” model, since it models the GPA/SAT relationship as two parallel lines shifted by a constant (one for men and one for women):

$$\text{Math SAT} = 400 + 73 \times \text{High school GPA}$$

for men, and

$$\text{Math SAT} = 380 + 73 \times \text{High school GPA}$$

for women. More complex relationships also can be fit using transformations based on indicator variables. By convention, the “0” in an indicator variable indicates absence of a characteristic, while the “1” indicates presence; thus, for example, a variable called Male

would be 1 for males and 0 for females, while a variable called Female would be 1 for females and 0 for males. This is one of the ways that people try to establish if there is statistical support for discrimination in civil suits, by the way.

A special case of this is when the *only* predictor in the regression is a 0/1 indicator variable. This is the *two-sample* problem (sometimes called the independent samples problem), where the goal is to assess whether the mean values of the target variable in two independent samples are significantly different from each other. The slope coefficient in this case is an estimate of the difference in the average value of the target between the groups (that is, the two samples, which are indexed by the 0/1 variable). In this situation the usual regression assumptions are equivalent to assuming that the target variable is normally distributed within each group, with possibly different means but the same variance.

Note that an indicator variable is just another predictor in your regression model, and as such you should look at the marginal relationship between the response and it. You can use a scatter plot, of course, but a plot that is easier to interpret would be a set of side-by-side boxplots of the response separated by the different groups. Detailed description of this plot, along with the `Minitab` commands to construct it, can be found in the “Data presentation and summary” handout.

A natural generalization to consider is to allow group effects in a regression model when there are more than two groups. That is beyond the scope of this course, but discussion of how to use `Minitab` to fit such models is given as an appendix to this document, in case you are faced with data of this type.

Confidence intervals for expected responses and prediction intervals

Consider a hypothetical regression between heights of parents and heights of children. Say the fitted regression had the following form:

$$\widehat{\text{Height of child}} = 21.5 + 0.685 \times \text{Midheight of parents}.$$

There are two kinds of questions that we might want to answer using this fitted model:

- (1) What is our best guess for the *average* height of children for *all* children with parents with midheight 70 inches?
- (2) Given a *particular* set of parents whose midheight is 70 inches, what is our best guess for the height of their *particular* child?

In either case, our best guess is obtained by substituting into the regression equation:

$$\widehat{\text{Height of child}} = 21.5 + 0.685 \times 70 = 69.45 \text{ inches}.$$

But, there is a different level of accuracy with the two answers: the former is a confidence interval statement (the *average* $y|x$), while the latter is a prediction interval statement (a *particular* $y|x$). Just as was true for univariate data, a prediction interval has an extra source of variation (σ^2). If we knew β_0 and β_1 exactly, we would know the answer to question (1) exactly (the confidence interval would have zero width), but we wouldn't know the answer to question (2) with any more accuracy than the actual regression relationship allows. The “ ± 2 times the standard error of the estimate” prediction interval is a rough approximation to the exact interval, which depends on the actual value of the predictor.

Here is **Minitab** output giving a fitted value, confidence and prediction interval (these are not based on the actual Galton data):

```

Fit   StDev Fit      95.0% CI          95.0% PI
69.4524  0.18481 ( 69.0983, 69.8265) ( 65.5056, 73.4192)

```

This output is obtained by giving **Minitab** the value(s) of (all of) the predicting variable(s) as input. The coverage is 95% by default, but can be changed. The value under **Fit** is the fitted (or predicted) value, \hat{y} . **StDev Fit** is the estimated standard error of the fitted value, which is used in the confidence interval given under **95.0% CI**. The estimated standard error of the predicted value, which is used to construct the prediction interval given under **95.0% PI**, is not given, but equals $\sqrt{\text{StDev Fit}^2 + \mathbf{s}^2}$, where **s** is the standard error of the estimate.

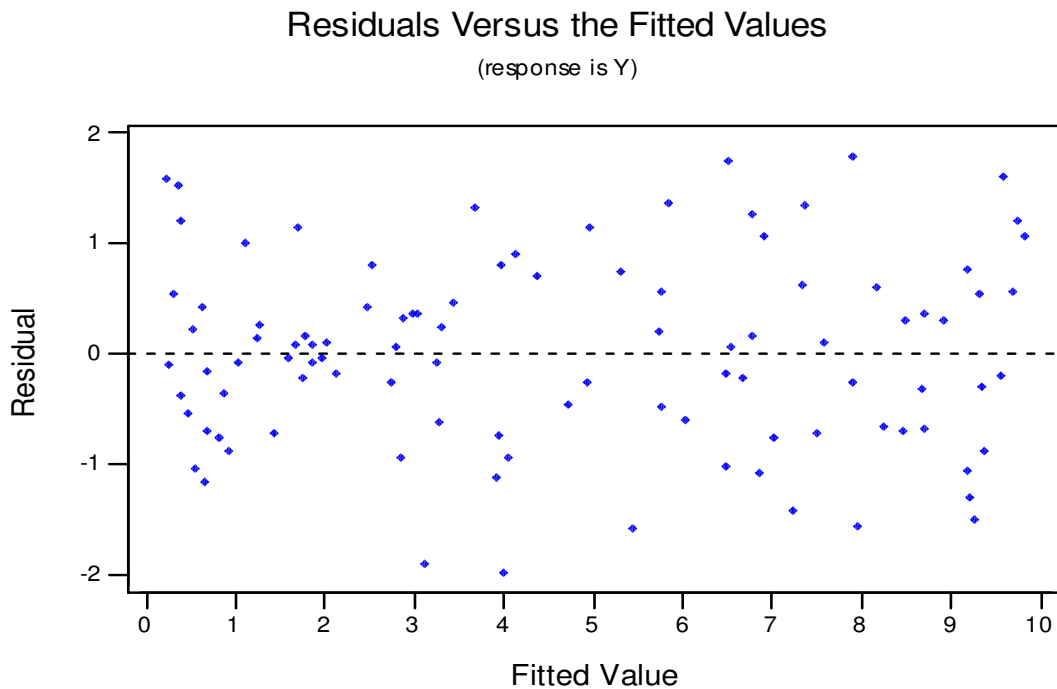
Note, by the way, that while these intervals are confidence and prediction intervals in the same sense as that we used for univariate data, they are **not** constructed in the same way. There is a formula for these intervals for simple regression, but you needn't worry about it; there is no closed-form representation for these intervals when there is more than one predicting variable.

Residual plots and checking assumptions

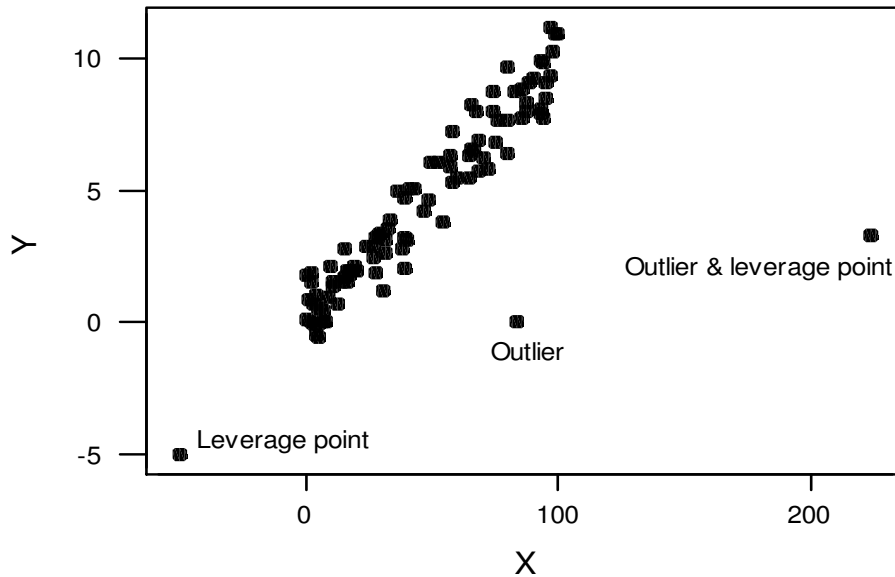
Plots should be a routine part of any regression analysis. Before the analysis is even done, scatter plots should be constructed of y versus each of the x 's to see what the marginal relationships look like, and if there are any noticeable unusual patterns in those relationships. Once the model is fit, plots of the residuals can help to identify unusual observations and violations of regression assumptions. **One plot that should**

be constructed routinely is a normal plot of the residuals. What we're looking for, of course, is (roughly) a straight line, indicating rough normality of the residuals (and hence, hopefully, the errors).

The other residual plot that should always be constructed routinely is a plot of the residuals $(y_i - \hat{y}_i)$ versus the fitted values (\hat{y}_i) . This residual plot should exhibit no apparent patterns — just a cloud of points on the page (the reason for this is that the regression assumptions are stating that there is no structure in the errors, so we would like to see the lack of a pattern in the residuals). Here is an example of the kind of plot we'd like to see:



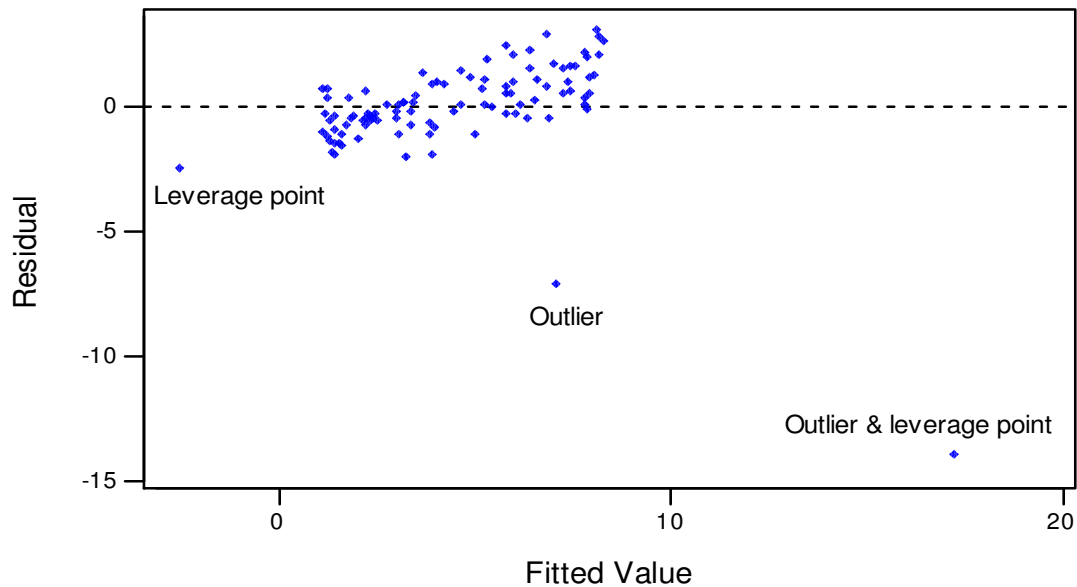
There are several ways that the residual plot can indicate a problem. Consider the following scatter plot, for example:



There is a nice linear relationship between the target and the predictor variables here, except that there are three unusual observations. An *outlier* is an observation where the target variable value is unusual given the predicting variable value, and it shows up as being unusually above or below the bulk of the points in the plot (in this case below them). A *leverage point* is an observation that has an unusual predicting variable value. Its target variable value might fall roughly along the same regression line as the other points (marked “Leverage point” above), or it might not (marked “Leverage point & outlier” above). The points are apparent in a scatter plot corresponding to a simple regression, but might be harder to see in scatter plots related to a multiple regression. They might still be seen in a plot of the residuals versus the fitted values, however. The plot of residuals versus fitted values below demonstrates how outliers show up at the top or bottom of the plot, while leverage points show up to the left or right.

Residuals Versus the Fitted Values

(response is Y)



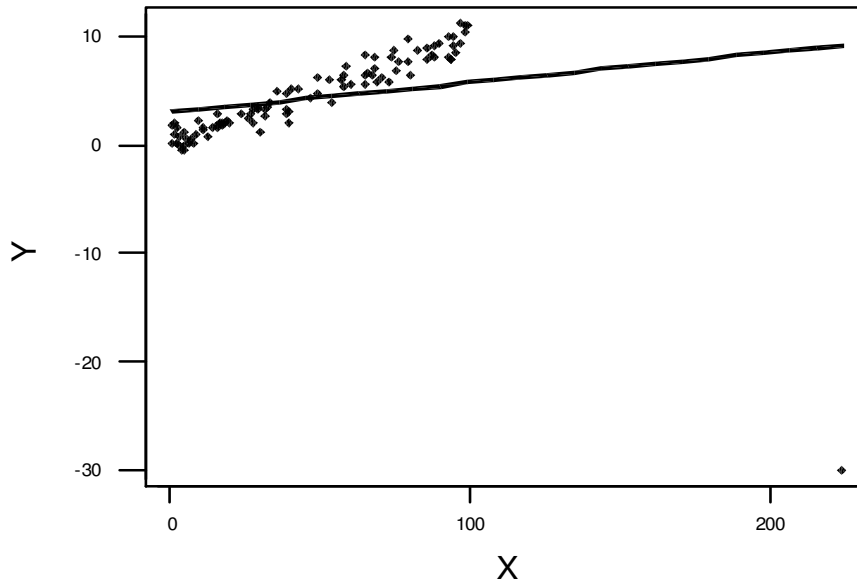
Outliers also show up at the bottom or top of a normal plot of the residuals.

Why is it important to identify these observations? There are many reasons. First, outliers don't follow the general pattern supported by the rest of the data, so it's not reasonable to act as if they do by ignoring them. Second, outliers and leverage points can have a large effect on a fitted regression. An unusual observation can draw the regression line away from the relationship exhibited by the bulk of the points towards itself, especially if it is both a leverage point and outlier:

Regression Plot

$$Y = 2.93335 + 2.78E-02X$$

$$R\text{-Sq} = 4.6\%$$



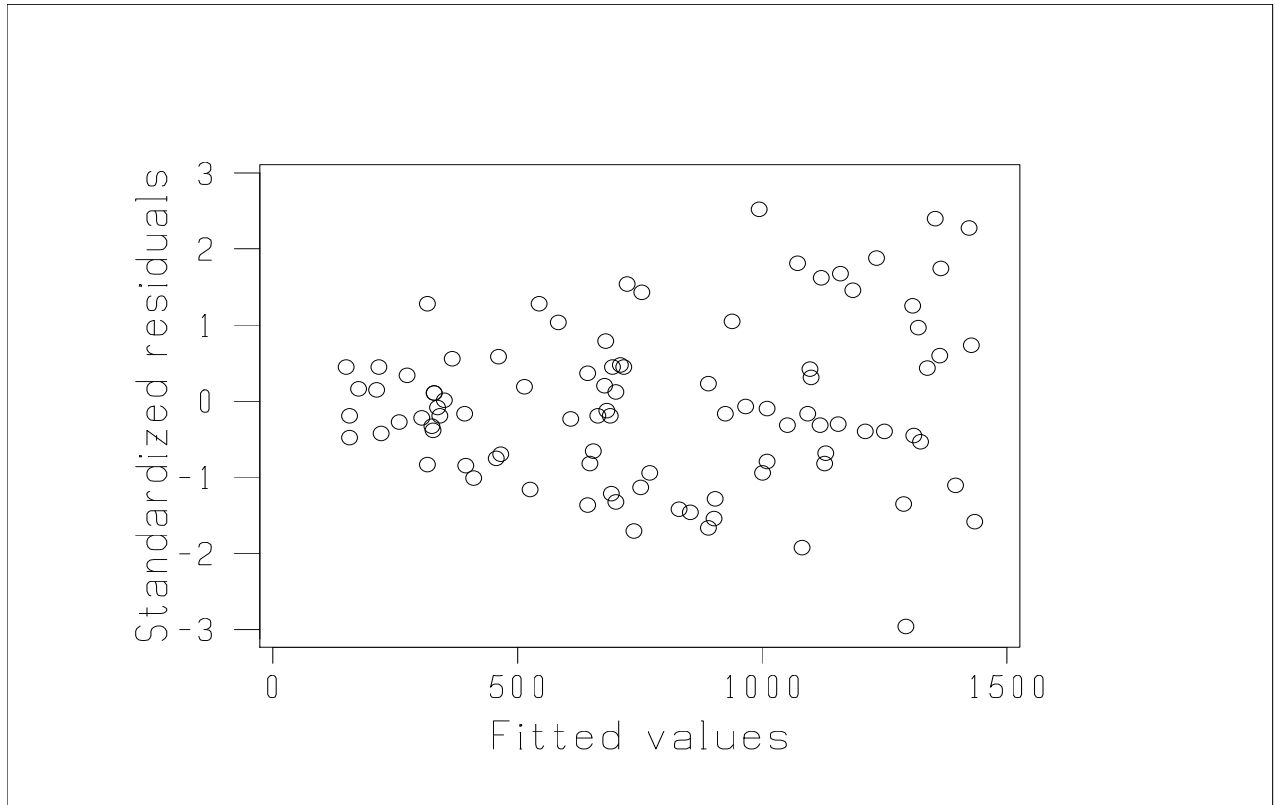
Even if the regression line hasn't been moved by the point, the point can still affect other aspects of the regression, such as R^2 , F -, and t -statistics. It is **not** correct to keep a point in a data set if removing it makes the apparent relationship weaker (e.g., lowers the R^2), since the stronger relationship using the original relationship was misleading. **The goal of an analysis is to learn the truth, not to maximize the R^2 .** Finally, unusual observations are often the most interesting ones in the data, since they can tell you when your model doesn't work (and hence, what you've missed).

This is a good time to debunk an argument that you might hear regarding unusual observations and statistical modeling. You might hear people say that they are not going to omit unusual observations from their data, because all of the observations in the data are "legitimate" (that is, the observations don't correspond to coding errors, observations that shouldn't have been in the original sample, or other obvious mistakes), and they want to keep the data "as they really are." This is a fundamentally incorrect attitude, as it ignores the key goal of any statistical model, which is to describe as accurately as possible what is going on in the data. Say in the plot above all of these data are "legitimate," and an analyst fits a regression model based on all of the data. The resultant fitted regression line is obviously an extremely poor representation of what is going on in the data — it does not in any way describe the data "as they really are," because of the deficiencies

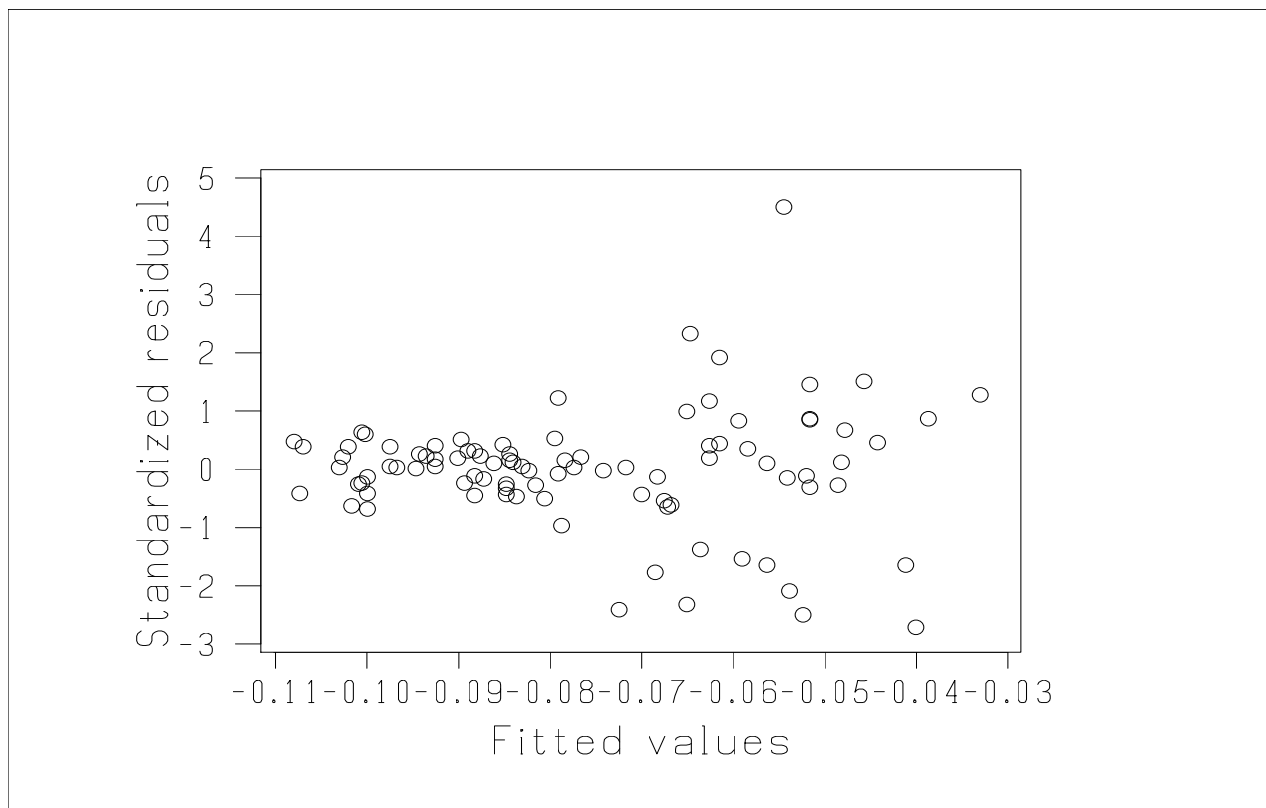
of least squares regression modeling (and its sensitivity to unusual observations). Since a different regression method that was insensitive to unusual observations (a *robust* regression technique) would lead to a completely different fitted regression (one not affected by the unusual observation), the issue is not whether the data are legitimate, but rather whether the description provided by the statistical model is legitimate. It is apparent that a much better description of what is going on in these data would be to report and discuss the one unusual observation, and then fit the model to the remaining data points, providing a good description of what is going on in the bulk of the data. **We remove unusual observations from a regression analysis not because there is something “wrong” with the data, but rather because there is something wrong with least squares regression (it is sensitive to unusual observations).**

Minitab sometimes provides in its regression output a list of observations that it considers “unusual.” Cases labeled with an R are potential outliers, while those labeled X are potential leverage points. While there’s nothing wrong with looking at those points, you should know that I **don’t** recommend using this output to decide which points actually are unusual. The reason is that Minitab uses certain regression diagnostics to label these points (I discuss these later in this handout) using cutoffs that I don’t like. In particular, I think they label too many points as outliers, and not enough as leverage points. This display can be disabled in Minitab, but it’s just as easy to just not pay much attention to it.

Two other particular kinds of patterns in residual plots are worth knowing about. One possibility is that the cloud of points seems to be narrower or wider at different parts of the plot (*standardized* residuals are residuals that are scaled to have unit standard deviation; I say more about them in a later handout). This indicates a violation of the constant variance (homoscedasticity) assumption.

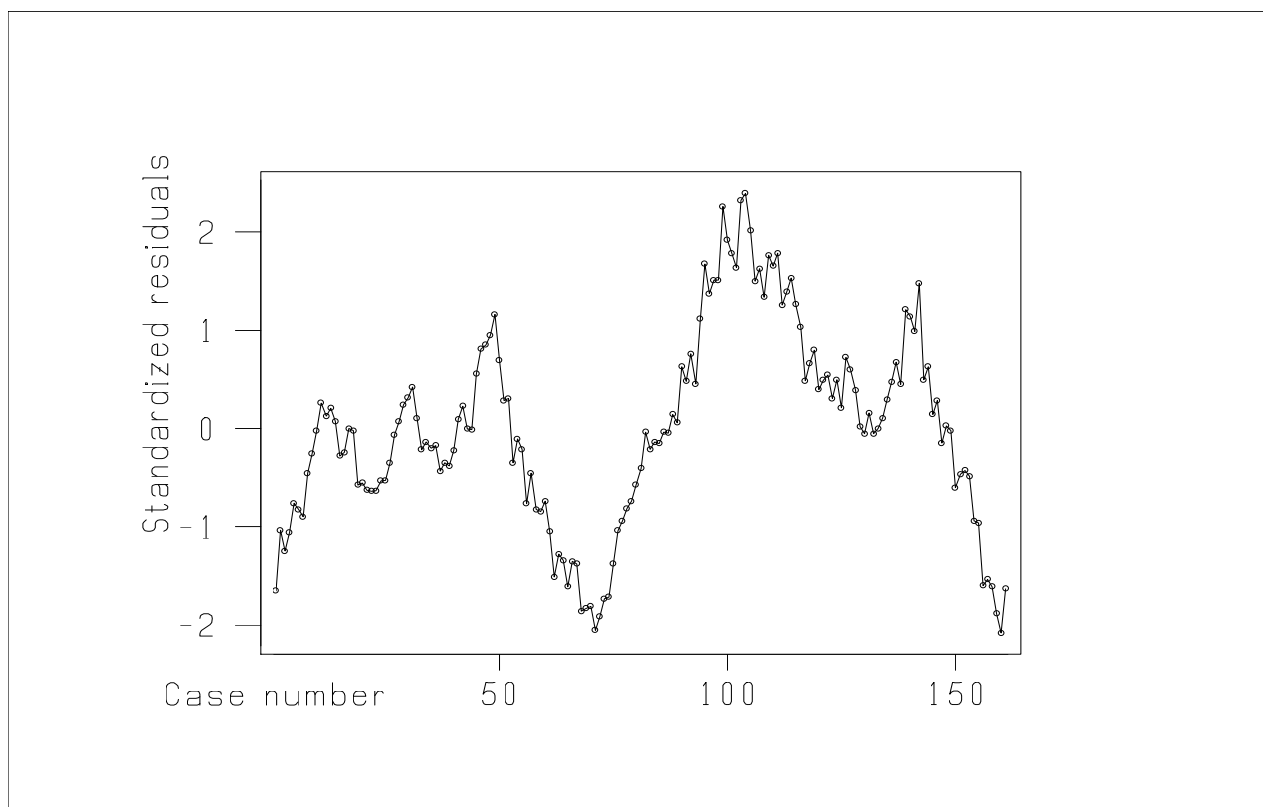


The residual plot above exhibits a gradual widening as the fitted values increase. This is often indicative of a multiplicative, rather than additive, relationship. Thus, a picture like this indicates that using the logarithms of the variables could very well be useful (we'll say more about this later).



The picture on the previous page is a little different. Rather than a gradual widening of the residual cloud, there appears to be two subgroups in the data with different amounts of variability off the regression line. This is also non-constant variance, but the proper course of action here is to use *weighted least squares*. This is a technique whereby the observations that are closer to the regression line (where the relationship is stronger) are weighted more heavily than observations that are farther away from the regression line.

Let's say your data had a natural time ordering. This is a situation where the assumption that the errors are uncorrelated with each other might be violated. A time series plot of the residuals (just a plot of the residuals versus time) can indicate possible problems:



Once again, we would like to see a shapeless cloud, with no apparent patterns. This is not the case in the plot above. There is a clear cyclical effect, of the residuals alternately increasing and decreasing together. This indicates a correlation structure in the residuals related to time, and is called *autocorrelation*. The solution to this problem is “time series stuff” — there are diagnostics to help detect it (e.g., runs test, Durbin–Watson test, autocorrelation function plots), and methods to try to address it (e.g., Cochrane–Orcutt procedure, Box–Jenkins methods, spectral analysis).

Transformations

Everything we’ve done so far assumes a linear relationship between the x ’s and y . What if that’s not true? Then none of this analysis makes any sense. What are the possibilities? We can determine these from either examination of scatter plots or from our understanding of the underlying process itself.

- (1) In some contexts the relationship between x and y is inherently nonlinear. Consider, for example, the science of *pharmacokinetics*, which is the study of the way drugs spread through the body after being administered to a patient. A standard pharmacokinetic model is the so-called two-compartment model, which says that C_t , the

concentration of a drug in the bloodstream t minutes after the drug has been administered into a patient's arm, satisfies

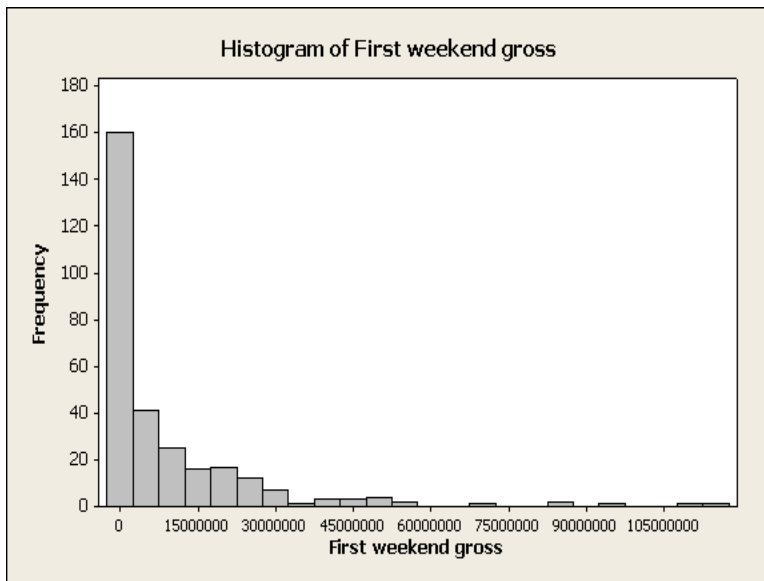
$$C_t = \theta_1 e^{-\theta_1 t} + \theta_2 e^{-\theta_2 t} + \epsilon_t,$$

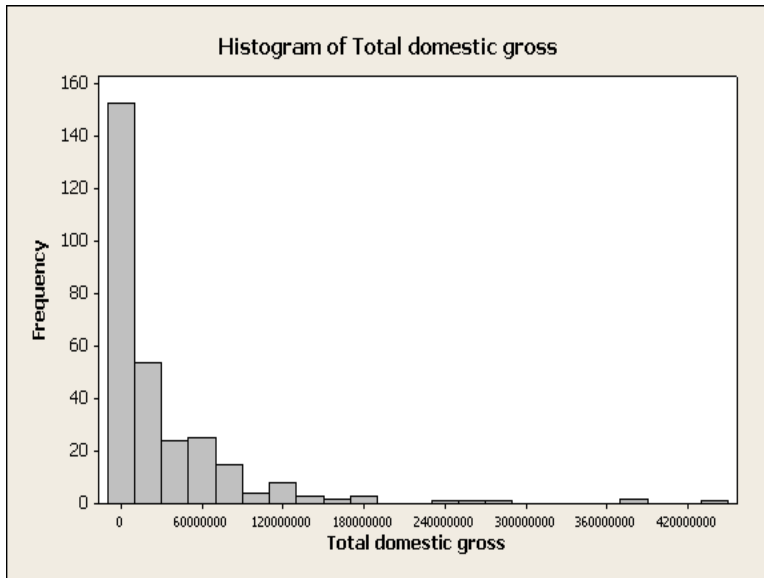
where θ_1 and θ_2 are parameters that determine the rate of absorption of the drug (the motivation for this model is that the body can be thought of as consisting of two compartments: the vascular system, including the blood, liver, and kidneys, where the drug is distributed throughout the body quickly, and poorly perfused tissues, such as muscle, lean tissue, and fat, where the drug is eliminated more slowly. The only way that this model can fit to observed data is by using *nonlinear regression* methods. Several statistical packages include such routines (Minitab is not one of them, however), but nonlinear regression estimation is a tricky business. All of the nice properties of linear least squares regression that we take for granted no longer hold for nonlinear regression (e.g., R^2 measures can be negative, t - and F -statistics don't follow t - and F -distributions, estimates may be difficult to calculate, the usual confidence and prediction intervals might not be appropriate, etc.). Still, in these circumstances, there is no alternative to the use of nonlinear regression methods. For a discussion of several ways to address some of the difficulties in using two-compartment models, see D. Niedzwiecki and J.S. Simonoff, "Estimation and inference in pharmacokinetic models: the effectiveness of model reformulation and resampling methods for functions of parameters," *Journal of Pharmacokinetics and Biopharmaceutics*, **18**, 361–377 (1991).

- (2) We might notice a parabolic (quadratic) relationship between x and y . This just suggests enriching our model to include both linear and quadratic terms. That is, we should fit a model with two predictors related to x , x and x^2 . In theory, this could be extended to cubics, quartics, quintics, and so on, but while this would lead to a more flexible shape for a fitted curve, it would also lead to complete loss of interpretability of the implications of the model.
- (3) It is often the case that using the logarithm of a variable, rather than the variable itself, makes a relationship look more like one consistent with the assumptions of least squares regression. We've talked about this already, of course, but let's quickly review a bit. Consider the following possibilities:
 - (a) Variables that are long right-tailed often benefit from using logs.
 - (b) Relationships that look exponential rather than linear benefit from using logs.

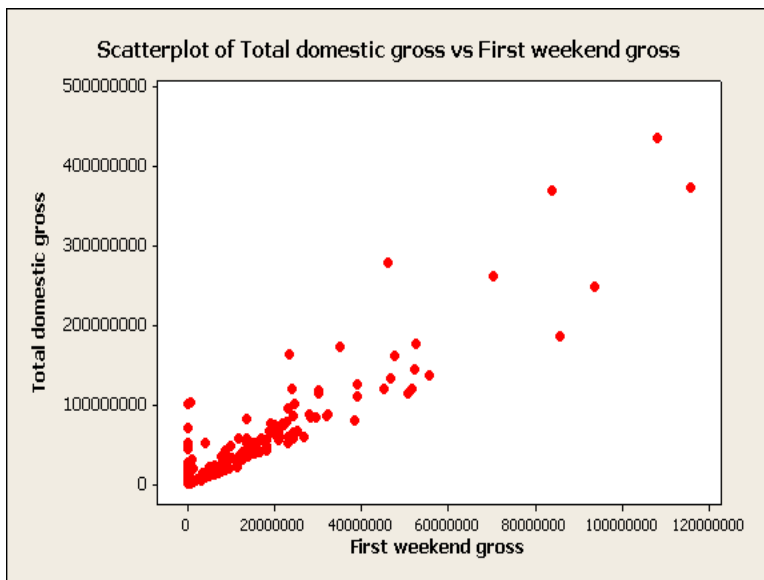
- (c) Nonconstant variance is often cured by using logged variables.
- (d) Money data, which usually operate multiplicatively, rather than additively, are often reasonably modeled using logged variables.

Consider the following graphs. These data refer to the first weekend gross and total domestic (U.S.) gross in millions of dollars for movies released during 2004. The histograms show that each of these variables is right-tailed, and of course, both are money variables:

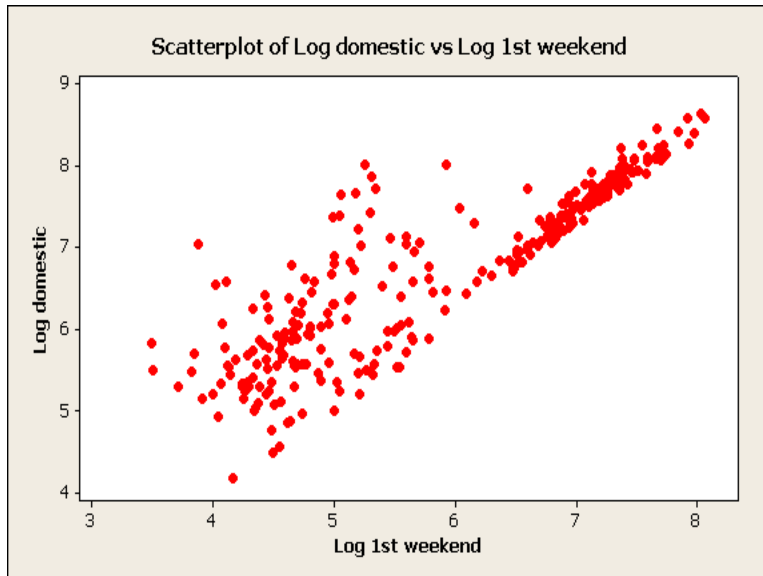




Further, a scatter plot shows that the relationship here is apparently nonlinear, and there is very obvious nonconstant variance:



All of these facts suggest that logging **both** variables will lead to a relationship that looks more linear, with better variance properties, and that is in fact the case:



It is clear, of course, that there is still something unusual with these data, in that the movies with lower revenues have much higher variability than those with higher revenues. There is enormous variability in the way small-release movies are screened and marketed. For the most part, the kinds of movies that receive an initial and/or full release of less than ten screens are niche market pictures, such as foreign films, small independent productions and documentaries, and often get released in specific art-house theaters in major markets only. Additionally, the length of time a given film may play can vary widely. That time frame is dependent on factors as diverse as competition for screens, film festival awards, word of mouth, and reviews. Overall, these movies generally rely most heavily on word of mouth and reviews to keep them in theaters for any appreciable length of time. This is very different from the standard sort of approach with wide-release films, which involves many newspaper and television ads and release on many screens. In the output below, we'll focus on only films that opened with at least \$2 million in receipts in the first weekend.

How do we interpret the output from a regression using logged variables? There are several situations that need to be considered:

- (1) **The target variable is in logged form.** Here is the regression output for the regression of logged domestic gross on logged first weekend gross:

Regression Analysis

The regression equation is

$$\text{Log domestic} = -0.164 + 1.09 \text{ Log 1st weekend}$$

Predictor	Coef	SE Coef	T	P
Constant	-0.1637	0.2009	-0.81	0.417
Log 1st weekend	1.08829	0.02821	38.58	0.000

$$S = 0.126895 \quad R\text{-Sq} = 91.6\% \quad R\text{-Sq}(\text{adj}) = 91.5\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	23.962	23.962	1488.08	0.000
Residual Error	137	2.206	0.016		
Total	138	26.168			

Note that the standard error of the estimate $s = .13$. We know that a rough prediction interval is $\pm(2)(.13) \approx .26$, but what does that mean in this context? The key is to remember that additive changes for logs are equivalent to multiplicative changes for the original variables. An additive increase in logged domestic gross of .26 corresponds to *multiplying* domestic gross by $10^{.26}$, or 1.82; similarly, an additive decrease in logged domestic gross of .26 corresponds to *multiplying* domestic gross by $10^{-.26}$, or .55. Thus, the interval $\pm.26$ in the logged scale corresponds to us saying that knowing first week gross allows us to predict total domestic gross to within a multiplicative factor of roughly 1.8. So, for example, if I predict total domestic gross to be \$20 million, I wouldn't be surprised (with 95% probability) if the actual total domestic gross was as little as \$11 million, and as much as \$36 million; if I predict it to be \$100 million, I'm not surprised if it ranges from \$55 million to as much as \$182 million. That is of course a pretty wide range, which reinforces how unpredictable the movie business is.

- (2) **Both the target and predicting variable are in logged form.** The situation where both the target and predicting variables are in logged form is particularly appealing because the slope coefficient has a very nice interpretation. Consider the

functional relationship

$$y = \alpha x^\beta.$$

This is a **multiplicative** relationship; it is consistent with **proportional** changes in x being associated with **proportional** changes in y . For example, if $\beta = 1.58496$, doubling x is associated with tripling y (since $2^{1.58496} = 3$). This functional form is *linearizable*, since if we take the logarithm of both sides of the equation we obtain

$$\log y = \log \alpha + \beta \log x.$$

That is, the model is linear after logging both x and y . This multiplicative relationship is called the *log–log model*. The log–log model is important in the construction and estimation of demand functions. Let y represent demand for a product, and x be the price. The price elasticity is defined as the proportional change in demand for a proportional change in price; that is,

$$\frac{dy/y}{dx/x} = \frac{dy/dx}{y/x},$$

where dy/dx is the derivative of y with respect to x . Some calculus shows that for the log–log model, the elasticity is a constant β , and the log–log model is therefore sometimes called the *constant elasticity model*. Thus, if it is assumed that elasticities are constant, they can be estimated using the slope coefficient for price in a log–log regression model fit. So, in the output above, the slope estimate is 1.088; what this says is that a 1% change in first weekend gross is associated with a 1.088% change in total domestic gross (in a multiple regression, this would be holding all else fixed). [Technically, elasticities are only valid for *small* proportional changes in the predictor; thus while a 1% change in x goes with a (roughly) 1.088% change in y , it isn't true that a 50% change in x goes with a $(50)(1.088) = 54.4\%$ change in y .] Note that a proportional relationship corresponds to the slope equaling one.

- (3) **The target variable is in logged form, but the predicting variable is not.** Consider the functional relationship

$$y = \alpha \beta^x.$$

This is a mixed **additive / multiplicative** relationship; it is consistent with **additive** changes in x being associated with **proportional** changes in y . For example, if $\beta = 2$,

adding two units to x is associated with multiplying y by 4 (that is, multiplying y by two twice). This functional form is also linearizable, since if we take the logarithm of both sides of the equation we obtain

$$\log y = \log \alpha + \log \beta \times x.$$

That is, the model is linear after logging y but not x . This model is particularly appropriate for modeling the growth of objects over time; for example, the total amount of money in an investment as a function of time, or the number of people suffering from a disease as a function of time. Since the coefficient in a semilog model is actually $\log \beta$, a little care must be taken in its interpretation, since it depends on what base is used for the logarithm. Say logs base 10 are used. Then, a slope coefficient of 1.5, say, says that adding 1 to x is associated with multiplying y by $10^{1.5} = 31.62$. If natural logs (base e) had been used instead, the slope coefficient would have been 3.454, since $e^{3.454} = 31.62$, implying the same additive/multiplicative relationship. The coefficient in this case is called a *semielasticity*.

(4) **The predicting variable is in logged form, but the target variable is not.**

This is obviously possible, but the functional relationship it implies between y and x is a little strange:

$$10^y = \alpha x^\beta$$

(I've used 10 as the base here, assuming that the logs being taken are to that base). Logging both sides gives the relationship

$$y = \log \alpha + \beta \log x.$$

Interpretation of the slope β comes from the usual interpretation, except that adding one to $\log x$ corresponds to multiplying x by 10. That is, the model implies that multiplying x by 10 is associated with an expected increase of β in y (in a multiple regression, holding all else fixed). When might such a relationship make sense? It seems most appropriate in the situation where the target variable is a “pure” number of some sort, such as a return or a score of some sort, and the predictor is long right-tailed.

Note that for all of these models, you should always report coefficients and the standard error of the estimate in terms corresponding to the original scale, not the transformed

(log) scale. So, for example, we report the rough prediction interval $\pm .26$ above as corresponding to a multiplicative factor of 1.8 (which corresponds to dollars), not an additive factor of .26 in the logged scale. Similarly, elasticities and semielasticities are in the original scale of the variables, and that is how they should be reported, not in terms of the underlying logged variables.

There is one additional important point to consider if the response variable in the logged scale. If that is the case, confidence intervals for average response cannot be easily converted to the original scale (which is, of course, what we would be interested in); the reason for that is that the log of the expected value of a random variable is not equal to the expected value of the log of a random variable. Prediction intervals, on the other hand, can be converted directly to the original scale by antilogging the two ends (just as was true in the univariate case), providing an interval estimate for the value in the original scale of a future observation.

Variable selection

An important question is how to decide what predicting variables to consider for your regression model. You do **not** want to throw in every variable that you can think of (and then let the computer decide what to keep), since this adds unnecessary random noise to the model. Also, just by random chance, predictors that don't add anything to the fit can appear useful, inflating the apparent usefulness of the model.

The first step in applying a multiple regression model is to choose a reasonable set of possible predictors based on your knowledge of the problem, weeding out obviously irrelevant or redundant variables *before* fitting the regression. Only then would you run your regression. Even at this point you still want to decide which variables you want to keep, and which are not needed. The adjusted R^2 can be a useful tool when initially fitting a regression model with several predictors, since it takes the number of predictors into account (unlike R^2). A model with high R^2 but low R_a^2 means that (at least some of) the predictors are probably just picking up random fluctuation, and nothing much is going on. Another way that R_a^2 can be informative is that removing unimportant variables from a multiple regression fit should result in R_a^2 remaining virtually unchanged, or even increasing.

An additional problem when many possible models are possible is that the act of looking for models makes the resultant models look more useful than they really are (prediction intervals are too narrow, compared with the accuracy you would actually achieve on new

data). This *variable selection* question is a difficult one, and has been the focus of a good deal of research. A tool that can be useful in ordering different multiple regression models is *best subsets regression*. There are important issues in using this tool indiscriminately, but the output from a best subsets regression can help to see how complex a model needs to get to account for most of the available predictive power in the predictors.

Note that while marginal relationships between the target and each predictor, while interesting, do not reflect the way that predictors work together to model the response variable, and hence cannot be used to determine which predictors will be useful in a multiple regression, or which variables will be most important or most useful in a multiple regression model. The individual regression coefficients also can be misleading if used for this purpose; remember, different variables are on different scales and have different amounts of variability, so the importance of a variable in a regression as reflected by its coefficient must be viewed in context (one centimeter is not the same as one year, and one dollar might be a lot or a little, depending on the context).

Multicollinearity

We know from our earlier discussion of multiple regression that, generally speaking, the presence of one predictor in a regression model affects the slope coefficients of other variables, since the coefficients only represent estimates of the expected change in the target given that the other variables are held fixed. A related issue is that of multicollinearity. When predicting (x) variables are highly correlated with each other, this can lead to instability in the regression coefficients, and the t -statistics for the variables can be deflated. From a practical point of view, this can lead to two problems:

- (1) If one value of one of the x -variables is changed only slightly, the fitted regression coefficients can change dramatically.
- (2) It can happen that the overall F -statistic is significant, yet each of the individual t -statistics are not significant (multicollinearity generally has little effect on overall measures of fit, but can have a serious effect on measures of the importance of individual variables given the others in the model). Another indication of this problem is that the p -value for the F test is considerably smaller than those of any of the individual coefficient t -tests.

Another problem with multicollinearity comes from attempting to use the regression model for prediction. In general, simple models tend to forecast better than more complex ones, since they make fewer assumptions about what the future must look like. That is, if

a model exhibiting collinearity is used for prediction in the future, the implicit assumption is that the relationships among the predicting variables, as well as their relationship with the target variable, remain the same in the future. This is less likely to be true if the predicting variables are collinear.

How can we diagnose multicollinearity? A useful diagnostic is the *variance inflation factor* (*VIF*) for each predicting variable, which is defined as

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the R^2 of the regression of the variable x_j on the other predicting variables. The *VIF* gives the proportional increase in the variance of $\hat{\beta}_j$ compared to what it would have been if the predicting variables had been completely uncorrelated. Values of *VIF* greater than $\max(10, 1/(1 - R_{model}^2))$, where R_{model}^2 is the usual R^2 from the regression output, are worthy of concern (Minitab supplies *VIF* values).

What can we do about multicollinearity? The simplest solution is to simply drop out any collinear variables; so, if High school GPA and SAT are highly correlated, you don't need to have to both in the model, so use only one. Note, however, that this advice is **only a general guideline** — sometimes two (or more) collinear predictors are needed in order to adequately model the target variable.

Regression diagnostics

As is true of all statistical methodologies, linear regression analysis can be a very effective way to model data, *as long as the assumptions being made are true*. For the regression model, these assumptions include that all of the data follow the hypothesized linear model; that is, there aren't any cases far off the regression line (*outliers*). In addition, cases that are isolated in X-space (*leverage points*) are also problematic, as they can have a strong effect on estimated regression parameters, measures of fit, and so on. Once you've identified such a case, it's very important to identify what it is that makes the case unusual (e.g., y is surprisingly large for the given X values, the observed values for two X-variables don't typically occur together, etc.). Further, you need to try to determine what might have happened in the random process under study that would result in such a case.

As discussed above, residual plots are very useful to detect outliers and leverage points. In a residuals versus fitted values plot, points by themselves on the top or bottom are outliers; points by themselves on the left or right are leverage points. In a normal plot

of the residuals, outliers show up as distinct at the bottom left (negative outliers) or top right (positive outliers).

Still, it is sometimes the case (particularly for multiple regression data sets) that these plots don't identify these cases very well. For this reason, several diagnostics have been developed to help identify unusual cases. Three of them seem to adequately cover the many possibilities that have been suggested:

Standardized residuals

By definition, an outlier is a point off the regression line. Thus, its residual should be large (in absolute value). The standardized residual is the residual divided by the standard deviation of the residual; that is, it is a residual standardized to have standard deviation 1. Recalling that the (unknown) errors are assumed to be normally distributed, we can see that the standardized residuals can be expected to be (roughly) standard normal. For example, we would expect about 95% of them to be within ± 2 . A good guideline for standardized residuals is that a case with a standardized residual larger than about ± 2.5 should be investigated as a potential outlier.

Leverage values

Looking at residuals doesn't help in the detection of leverage points, since they don't necessarily fall off the line (and can, in fact, draw the line towards them, thereby *reducing* their residuals). What is needed is a measure of how far a case is from a "typical" value. This is provided by the so-called *leverage value*. (It is sometimes referred to as the "hat" value, or the "diagonal element of the hat matrix", or the "diagonal element of the prediction matrix". There are good reasons for that, but we won't go into them here.) The leverage value is simply a measure of how far a particular case is (based on only the X-values) from the average of all cases, with distance being measured in such a way that the correlations between the X-variables is taken into account.

It can be shown that the sum of the N leverage values must equal $p + 1$, where p is the number of predicting variables in the regression. That is, the average leverage value is $\frac{p+1}{N}$. A good guideline for what constitutes a large leverage value is $2.5\left(\frac{p+1}{N}\right)$; cases with values greater than that should be investigated as leverage points.

Cook's distances

A different way to look at the unusual case problem is to focus on the effect a case has on the regression. A case that, if it were removed, would result in a large change in the regression is an *influential point*, and obviously dangerous to leave in. A common measure

of influence (although not the only one) is *Cook's distance*, which measures the change in the fitted regression coefficients if a case were dropped from the regression, relative to the inherent variability of the coefficient estimates themselves. A value of Cook's D over 1 or so is flagging a point that should probably be studied further.

If an observation or observations are ultimately omitted from the data, it is important to remember that you have effectively created a new data set. That means that you have to re-examine again from the beginning what model you should fit, the properties of that model, check assumptions, etc.

Minitab (and all other *good* statistical packages) provides these diagnostics as a standard option from the regression. These values should **always** be determined and looked at. That can mean simply printing them out to look at; using univariate pictures of them, like histograms, stem-and-leaf displays or boxplots, can also be useful, as that might make unusually high or low values stand out more.

It is worth noting what these diagnostics are *not* so good at. Specifically, they are all sensitive to the so-called *masking effect*. This occurs when several unusual cases are all in the same region. When this happens, the diagnostics, which all focus on changes in the regression when a *single* point is deleted, fail, since the presence of the other outliers means that the regression line changes very little. The problem of multiple outliers in regression is one of the hardest problems in statistics, and is a topic of ongoing research. See, for example, A.S. Hadi and J.S. Simonoff, "Procedures for the identification of multiple outliers in linear models," *Journal of the American Statistical Association*, **88**, 1264–1272 (1993).

Minitab commands

To fit a linear least squares regression line, click on **Stat** → **Regression** → **Regression**. Enter the target variable under **Response:** and the predicting variable under **Predictors:**. To perform a multiple regression, just enter the variables desired as predictors under **Predictors:**. When fitting a multiple regression, variance inflation factors are supplied under **Options** for a multiple regression fit.

The correlation between any two variables is obtained by clicking on **Stat** → **Basic Statistics** → **Correlation**. Enter the variables in the box under **Variables:**. Entering more than two variables gives a correlation matrix.

A prediction interval and a confidence interval for average y for a given (set of) predictor value(s) are obtained as a subcommand of the regression fit. Click on **Options**. Enter the value(s) of the predicting variable(s) under **Prediction intervals for new observations:**.

To obtain regression residual plots, while in the regression dialog box click on **Graphs**. Under **Residual Plots** click **Normal plot of residuals**, **Residuals versus fits**, and **Residuals versus order** (if you have data with a natural ordering, like time series data). To get a four-in-one residual plot, click on **Graphs** while fitting a regression, and click the radio button next to **Four in one**. To get plots of residuals versus different variables, just enter the names of the variables you want under **Residuals versus the variables:**.

Although it is possible to omit observations in a sample by simply highlighting them in the data worksheet and pressing the delete key, this is generally not advisable, since then the observation cannot be recovered without reopening the original file (and if you save the data before doing that, the observation is gone completely). A better approach is to create a subset of the worksheet that has the observations you want; this will create a new worksheet that can be analyzed, but the original worksheet will still be there as well. Click on **Data → Subset Worksheet**. You can give the new worksheet an identifying name if you like under **Name:**. Click the radio button next to **Specify which rows to exclude**, click the radio button next to **Specify rows:**, and enter the row numbers of the outliers in the associated box. Note that there is a good deal of flexibility in the subsetting; you can identify rows to include or exclude, identify them by some condition (for example, observations with values of a predictor greater than 10), or brush them on a scatter plot and identify them that way, in addition to specifying them by row number(s).

To create a regression plot with pointwise confidence and prediction intervals superimposed, click on **Stat → Regression → Fitted Line Plot**. Enter the target variable under **Response (Y):** and the predicting variable under **Predictor**. Click on **Options**, and click on **Display confidence bands** and **Display prediction bands** under **Display Options**.

To get plots of standardized residuals when constructing graphs in a regression, click on the radio button next to **Standardized**.

To save regression diagnostics when performing a regression, click on **Storage**, and then click on **Standardized residuals**, **Hi (leverages)**, and **Cook's distance**. To display the values, click on **Manip → Data Display**. Enter the variable names, such as *H11 SRES1 COOK1* under **Columns, constants, and matrices to display:**.

Best subsets regression is performed by clicking on **Stat** → **Regression** → **Best Subsets**. Enter the target variable under **Response:** and the predicting variable(s) under **Free predictors:**. If there are any variables that you want to be in all regression models, enter them instead under **Predictors in all models:**. The output will list the two models of each size (one predictor, two predictors, etc.) with highest R^2 , which allows you to see how many predictors it takes to account for most (or all) of the potential predictive power in the predictors.

Appendix: Allowing for group effects with more than two groups (categorical predictors)

All of the predicting variables that we have discussed so far fall into two types: continuous (or effectively continuous) variables, and 0/1 (indicator) variables. A third possibility is that of categorical predictors — variables that represent group membership, when there are multiple groups to consider. For example, a person’s political party might be a potential predictor in a regression model, where party could be registered Democrat, registered Republican, registered Independent, and unregistered. This is just a generalization of the use of indicator variables, allowing for constant shift effects for more than two groups.

A full discussion of models of this type (called *analysis of covariance*, or ANCOVA, models) is beyond the scope of this course, but it’s worth knowing how to get Minitab to fit such models, and how to interpret the output that results. The facility within Minitab that fits ANCOVA models is the General Linear Model, not the Regression facility we’ve been using. Here is a description of how to fit such models. Of course, the first step when using a categorical predictor in a regression model is to look at the relationship between the response and the predictor, which would be done using side-by-side boxplots of the response variable separated by the different groups; as was noted earlier in the discussion of indicator variables, this display is discussed in detail in the “Data presentation and summary” handout.

- (1) To do a regression that includes predictors that are categorical, click on **Stat** → **ANOVA** → **General Linear Model**.
- (2) Put your target variable under **Responses**. Put all of your predictors (continuous or categorical) under **Model**. Click on the button **Covariates** and put your continuous predictors in under **Covariates**.
- (3) Click on the button labeled **Results**, and click on the radio button next to **In addition, coefficients for all terms**. Also, enter the names of any categorical predictors under **Display least squares means corresponding to the terms**.
- (4) Click on the button labeled **Graphs**, and ask for the usual residual plots.
- (5) Click on the button labeled **Storage**, and ask for the usual regression diagnostics.

The output from General Linear Model looks a little different from what we have seen so far. Note that the discussion below includes some material on regression inference (points 2 and 3) that we haven’t discussed yet in the usual (numerical predictors) context, but will later.

- (1) The bottom part of the output will give regression coefficients and t -statistics for the

covariates, as usual. It will also give coefficients for the terms that correspond to the categorical predictor(s). There is one coefficient for each level of the categorical variable, except for the last one. The coefficient for the last one is determined so that all of the coefficients for that categorical predictor sum to zero; that is, to get the last coefficient, add up all of the others for that predictor, and change the sign from positive to negative or vice versa. You would use the coefficients to make predictions in the usual way, by plugging in x -values for your continuous predictors, and then adding the coefficient value(s) based on the level of the categorical predictor(s).

Below the coefficients will be a list of the levels of each of your categorical predictors under the heading **Least Squares Means**. These summarize for you the effect of the categorical variable. Say you have a categorical variable that has a level called **Group 1**, and the entry under **Least Squares Means** next to **Group 1** is 10.0. This number is an estimate of the expected value of the target variable for an observation in **Group 1**, given that all numerical predictors are at their respective sample means, and all other categorical variable levels are unknown (if there are any other categorical variables in the model). If the corresponding entry next to another level of the categorical variable (**Group 2**, say) was 15.0, that tells you that part of the effect of the categorical variable is that members of **Group 2** are estimated to be 5 units higher than members of **Group 1**, given everything else in the model.

- (2) In the top part of the output, you'll find some F -statistics. The ones that refer to the continuous predictors are equivalent to the t -tests given in the lower part of the output ($F = t^2$, with identical p -values). The one(s) that refer to categorical predictor(s) are tests as to whether there is any difference in the average of your target variable for different levels of the categorical predictor, given the other predictors. If that test has a high p -value, that's saying that the categorical predictor doesn't add anything to the fit.
- (3) There's no overall F statistic in the output. You can get it, however, from the output that is there. You need to determine the regression sum of squares, which is split into parts corresponding to the predictors. Take all of the values under "Seq SS" on the lines that correspond to predictors (all but the last two lines, which correspond to "Error" and "Total"), and add them up. Divide that number by the total of all of the values under "DF" on the lines that correspond to the predictors. This is the regression mean square. Now, divide that by the residual mean square, the value under "Adj MS" on the line labeled "Error". This is the overall F -statistic. It can

be compared to an F -distribution on (df_1, df_2) degrees of freedom, where df_1 is the sum of the DF values for the predictors, and $df_2 = n - df_1 - 1$.

- (4) If you're fitting a model with a categorical predictor that is coded with words (rather than numbers), there's no obvious way to get a plot of the residuals versus that variable. In order to do this, first save the residuals by clicking on the *Storage* button, and then checking the box next to **Residuals**. After you run your model, a variable named `RESI1` will be created (if you do it again, a new variable named `RESI2` is formed, and so on). You can then construct side-by-side boxplots of `RESI1`, with the categorical predictor being the grouping variable (see the "Data presentation and summary" handout).

MYTHS ABOUT DATA ANALYSIS

1. *The results of a data analysis hinge on the statistical significance of hypothesis tests.*

Hypothesis tests are a useful tool to help determine what is going on in a data set, but they have no inherent superiority over other tools, such as graphical methods. Hypothesis tests can give misleading results when samples are small, when samples are very large, and when assumptions being made do not hold. **Don't fall in love with the number .05 — it is not a magic number!**

2. *There is a single correct way to analyze a given data set.*

There are many different ways to analyze a typical data set, each with their own strengths and weaknesses. Usually any reasonable analyses will end up with similar results and implications. **There is more than one path to the summit!**

3. *When you come to a point in your analysis where you have to make a decision, you only can choose one possibility and follow it until you're done.*

Good data analysis is a process of following up leads that often reach dead ends. If you're not sure what path to take at a given point, try both paths and see what happens — the only thing you lose is a little time. The answer to the question “I'm not sure if this will help; what should I do?” is always “Try it and see.” **Any choices you make that you can justify are okay, as long as you tell people what you are doing.**

4. *The goal of an analysis is to ultimately come up with a model that has the strongest measures of fit possible.*

There is only one goal in any data analysis — to uncover what is actually going on in the data. All data analytic decisions should be driven by that concern, **not** by whether they make the R^2 (or F , or t) larger. Don't succumb to “ R^2 envy” (“Ha ha! Mine is bigger than yours!”). **Good data analysis is very much like good detective work — its goal is not to verify our own beliefs, but rather to search for the truth.**

“Out of the clutter find simplicity.
Out of discord make harmony.
Out of difficulty find opportunity.”

— Albert Einstein

“Embrace your data, not your models.”

— John Tukey

“What can be done with fewer assumptions is done in vain with more.”

— William of Ockham (“Ockham’s Razor”)

“In matters of science, it is always a good idea to remember and use the KISS method; that is, Keep It Simple, Stupid.”

— William Dawes

“Make things as simple as possible, but no simpler.”

— Albert Einstein

“When forecasting, remember — the future ain’t what it used to be.”

— Edward L. Leamer

“*Post hoc, ergo propter hoc.* (After this, therefore because of this.)”

— Latin proverb

“An approximate answer to the right question is far better than an exact answer to the wrong question.”

— John Tukey

“All models are wrong, but some are useful.”

— George E.P. Box

“The only relevant test of validity of a hypothesis is comparison of prediction with experience.”

— Milton Friedman

“There is no result in nature without a cause; understand the cause and you will have no need of the experiment.”

— Leonardo da Vinci

“Management is prediction.”

— W. Edwards Deming

“It is very hard to predict, especially the future.”

— Neils Bohr

“*Qui bene conjiciet, hunc vatem.* (He who guesses right is the prophet.)”

— Greek proverb

“The only useful function of a statistician is to make predictions, and thus to provide a basis for action.”

— W. Edwards Deming

- “Most economists think of God as working great multiple regressions in the sky.”
- “The moment you forecast you know you’re going to be wrong, you just don’t know when and in which direction.”
- “The herd instinct among forecasters makes sheep look like independent thinkers.”
- “When you know absolutely nothing about the topic, make your forecast by asking a carefully selected probability sample of 300 others who don’t know the answer either.”
- “If you have to forecast, forecast often.”

— Edgar R. Fiedler

*The Three Rs of Economic Forecasting —
Irrational, Irrelevant, and Irreverent*