

Chapter 11

Multiple Linear Regression

The 100-year flood is to be estimated for locations without streamflow gages using basin characteristics at those locations. A regression equation is first developed relating the 100-year flood to several basin characteristics at sites which have a streamflow gage. Each characteristic used is known to influence the magnitude of the 100-year flood, has already been used in adjoining states, and so will be included in the equation regardless of whether it is significant for any individual data set. Values for the basin characteristics at each ungaged site are then input to the multiple regression equation to produce the 100-year flood estimate for that site.

Residuals from a simple linear regression of concentration versus streamflow show a consistent pattern of seasonal variation. To make better predictions of concentration from streamflow, additional explanatory variables are added to the regression equation, modeling the pattern seen in the data.

As an exploratory tool in understanding possible causative factors of groundwater contamination, data on numerous potential explanatory variables are collected. Each variable is plausible as an influence on nitrate concentrations in the shallowest aquifer. Stepwise or similar procedures are performed to select the "most important" variables, and the subsequent regression equation is then used to predict concentrations. The analyst does not realize that this regression model is calibrated, but not verified.

Multiple linear regression (MLR) is the extension of simple linear regression (SLR) to the case of multiple explanatory variables. The goal of this relationship is to explain as much as possible of the variation observed in the response (y) variable, leaving as little variation as possible to unexplained "noise". In this chapter methods for developing a good multiple regression model are explained, as are the common pitfalls such as multi-collinearity and relying on R^2 . The mathematics of multiple regression, best handled by matrix notation, will not be extensively covered here. See Draper and Smith (1981) or Montgomery and Peck (1982) for this.

11.1 Why Use MLR?

When are multiple explanatory variables required? The most common situation is when scientific knowledge and experience tells us they are likely to be useful. For example, average runoff from a variety of mountainous basins is likely to be a function both of average rainfall and of altitude; average dissolved solids yields are likely to be a function of average rainfall, percent of basin in certain rock types, and perhaps basin population. Concentrations of contaminants in shallow groundwater are likely to be functions of both source terms (application rates of fertilizers or pesticides) and subsurface conditions (soil permeability, depth to groundwater, etc.).

The use of MLR might also be indicated by the residuals from a simple linear SLR. Residuals may indicate there is a temporal trend (suggesting time as an additional explanatory variable), a spatial trend (suggesting spatial coordinates as explanatory variables), or seasonality (suggesting variables which indicate which season the data point was collected in). Analysis of a residuals plot may also show that patterns of residuals occur as a function of some categorical grouping representing a special condition such as: on the rising limb of a hydrograph, at cultivating time, during or after frontal storms, in wells with PVC casing, measurements taken before 10:00 a.m., etc. These special cases will only be revealed by plotting residuals versus a variety of variables -- in a scatterplot if the variable is continuous, in grouped boxplots if the variable is categorical. Seeing these relationships should lead to definition of an appropriate explanatory variable and its inclusion in the model if it significantly improves the fit.

11.2 MLR Model

The MLR model will be denoted:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

where y is the response variable

β_0 is the intercept

β_1 is the slope coefficient for the first explanatory variable

β_2 is the slope coefficient for the second explanatory variable

β_k is the slope coefficient for the k th explanatory variable, and

ε is the remaining unexplained noise in the data (the error).

To simplify notation the subscript i , referring to the $i=1,2,\dots,n$ observations, has been omitted from the above. There are k explanatory variables, some of which may be related or correlated to each other (such as the previous 5-day's rainfall and the the previous 1-day rainfall). It is therefore best to avoid calling these "independent" variables. They may or may not be independent of each other. Calling them explanatory variables describes their purpose: to explain the variation in the response variable.

11.3 Hypothesis Tests for Multiple Regression

11.3.1 Nested F Tests

The single most important hypothesis test for MLR is the F test for comparing any two nested models. Let model "s" be the "simpler" MLR model

$$y_s = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon_s.$$

It has $k+1$ parameters including the intercept, with degrees of freedom (df_s) of $n-(k+1)$. Again, the degrees of freedom equals the number of observation minus the number of parameters estimated, as in SLR. Its sum of squared errors is SSE_s .

Let model "c" be the more complex regression model

$$y_c = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \beta_{k+1} x_{k+1} + \dots + \beta_m x_m + \epsilon_c.$$

It has $m+1$ parameters and residual degrees of freedom (df_c) of $n-(m+1)$. Its sum of squared errors is SSE_c .

The test of interest is whether the more complex model provides a sufficiently better explanation of the variation in y than does the simpler model. In other words, do the extra explanatory variables x_{k+1} to x_m add any new explanatory power to the equation? The models are "nested" because all of the k explanatory variables in the simpler model are also present in the complex model, and thus the simpler model is nested within the more complex model. The null hypothesis is

$$H_0: \beta_{k+1} = \beta_{k+2} = \dots = \beta_m = 0 \text{ versus the alternative}$$

$$H_1: \text{at least one of these } m-k \text{ coefficients is not equal to zero.}$$

If the slope coefficients for the additional explanatory variables are all not significantly different from zero, the variables are not adding any explanatory power in comparison to the cost of adding them to the model. This cost is measured by the loss in the degrees of freedom = $m-k$, the number of additional variables in the more complex equation.

The test statistic is

$$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{(SSE_c / df_c)} \quad \text{where } (df_s - df_c) = m-k.$$

If F exceeds the tabulated value of the F distribution with $(df_s - df_c)$ and df_c degrees of freedom for the selected α (say $\alpha=0.05$), then H_0 is rejected. Rejection indicates that the more complex model should be chosen in preference to the simpler model. If F is small, the additional variables are adding little to the model, and the simpler model would be chosen over the more complex.

Note that rejection of H_0 does not mean that all of the $K+1$ to m variables have coefficients significantly different from zero. It merely states that some of the coefficients in the more complex model are significant, making that model better than the simpler model tested. Other simpler models having different subsets of variables may need to be compared to the more complex model before choosing it as the "best".

11.3.2 Overall F Test

There are two special cases of the nested F test. The first is of limited use, and is called the overall F test. In this case, the simpler model is

$$y_s = \beta_0 + \epsilon_s, \text{ where } \beta_0 = \bar{y}.$$

The rules for a nested F test still apply: the $df_s = n-1$ and SSE_s equals $(n-1)$ times the sample variance of y . Many computer packages give the results of this F-test. It is not very useful because it tests only whether the complex regression equation is better than no regression at all. Of much greater interest is which of several regression models is best.

11.3.3 Partial F Tests

The second special case of nested F tests is the partial F test, which is called a Type III test by SAS. Here the complex model has only 1 additional explanatory variable over the simpler model, so that $m=k+1$. The partial F test evaluates whether the m th variable adds any new explanatory power to the equation, and so ought to be in the regression model, given that all the other variables are already present. Note that the F statistics on a coefficient will change depending on what other variables are in the model. Thus the simple question "does variable m belong in the model?" cannot be answered. What can be answered is whether m belongs in the model in the presence of the other variables.

With only one additional explanatory variable, the partial F test is identical in results to a t-test on the coefficient for that variable. In fact, $t^2 = F$, where both are the statistics computed for the same coefficient for the partial test. Some computer packages report the F statistic, and some the t-test, but the p-values for the two tests are identical. The partial t-test can be easily performed by comparing the t statistic for the slope coefficient to a student's t-distribution with $n-(m+1)$ degrees of freedom. H_0 is rejected if $|t| > t_{1-(\alpha/2)}$. For a two-sided test with $\alpha = 0.05$ and sample sizes n of 20 or more, the critical value of t is $|t| \cong 2$. Larger t-statistics (in absolute value) for a slope coefficient indicate significance. Squaring this, the critical partial F value is near 4.

Partial tests guide the evaluation of which variables to include in a regression model, but are not sufficient for every decision. If every $|t| > 2$ for each coefficient, then it is clear that every explanatory variable is accounting for a significant amount of variation, and all should be present. When one or more of the coefficients has a $|t| < 2$, however, some of the variables should be removed from the equation, but **the t values are not a certain guide as to which**

ones to remove. These partial t or F tests are precisely the tests used to make automatic decisions for removal or inclusion in "stepwise" procedures: forward, backward, and stepwise multiple regression. These procedures do not guarantee that some "best" model is obtained, as discussed later. Better procedures are available for doing so.

11.4 Confidence Intervals

Confidence intervals can be computed for the regression slope coefficients β_k , and for the mean response \hat{y} at a given value for all explanatory variables. Prediction intervals can be similarly computed around an individual estimate of y . These are entirely analogous to the SLR situation, but require matrix manipulations for computation. A brief discussion of them follows. More complete treatment can be found in many statistics textbooks, such as Montgomery and Peck (1982), Draper and Smith (1981), and Walpole and Myers (1985), among others.

11.4.1 Variance-Covariance Matrix

In MLR, the values of the k explanatory variables for each of the n observations, along with a vector of 1s for the intercept term, can be combined into a matrix X :

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdot & \cdot & x_{1k} \\ 1 & x_{12} & x_{22} & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} & \cdot & \cdot & x_{nk} \end{bmatrix}$$

X is used in MLR to compute the variance-covariance matrix $\sigma^2 \cdot (X'X)^{-1}$, where $(X'X)^{-1}$ is the "X prime X inverse" matrix. Elements of $(X'X)^{-1}$ for three explanatory variables are as follows:

$$(X'X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} & C_{03} \\ C_{10} & C_{11} & C_{12} & C_{13} \\ C_{20} & C_{21} & C_{22} & C_{23} \\ C_{30} & C_{31} & C_{32} & C_{33} \end{bmatrix} \quad [11.1]$$

When multiplied by the error variance σ^2 (estimated by the variance of the residuals, s^2), the diagonal elements of the matrix C_{00} through C_{33} become the variances of the regression coefficients, while the off-diagonal elements become the covariances between the coefficients. Both $(X'X)^{-1}$ and s^2 can be output from MLR software.

11.4.2 Confidence Intervals for Slope Coefficients

Interval estimates for the regression coefficients β_0 through β_k are often printed by MLR software. If not, the statistics necessary to compute them are. As with SLR it must be assumed

that the residuals are normally distributed with variance σ^2 . A 100•(1- α)% confidence interval on β_j is

$$\hat{b}_j - t_{(\alpha/2, n-p)} \sqrt{s^2 C_{jj}} \leq \beta_j \leq \hat{b}_j + t_{(\alpha/2, n-p)} \sqrt{s^2 C_{jj}} \quad [11.2]$$

where C_{jj} is the diagonal element of $(X'X)^{-1}$ corresponding to the j th explanatory variable. Often printed is the standard error of the regression coefficient:

$$se(\hat{b}_j) = \sqrt{s^2 C_{jj}}. \quad [11.3]$$

Note that C_{jj} is a function of the other explanatory variables as well as the j th. Therefore the interval estimate, like \hat{b}_j and its partial test, will change as explanatory variables are added to or deleted from the model.

11.4.3 Confidence Intervals for the Mean Response

A 100•(1- α)% confidence interval for the expected mean response $\mu(y_0)$ for a given point in multidimensional space x_0 is symmetric around the regression estimate \hat{y}_0 . These intervals also require the assumption of normality of residuals.

$$\hat{y}_0 - t_{(\alpha/2, n-p)} \sqrt{s^2 x_0'(X'X)^{-1}x_0} \leq \mu(y_0) \leq \hat{y}_0 + t_{(\alpha/2, n-p)} \sqrt{s^2 x_0'(X'X)^{-1}x_0} \quad [11.4]$$

The variance of the mean is the term under the square root sign. It changes with x_0 , increasing as x_0 moves away from the multidimensional center of the data. In fact, the term $x_0'(X'X)^{-1}x_0$ is the leverage statistic h_i , expressing the distance that x_0 is from the center of the data.

11.4.4 Prediction Intervals for an Individual y

A 100•(1- α)% prediction interval for a single response y_0 , given a point in multidimensional space x_0 , is symmetric around the regression estimate \hat{y}_0 . It requires the assumption of normality of residuals.

$$\hat{y}_0 - t_{(\alpha/2, n-p)} \sqrt{s^2 \langle 1 + x_0'(X'X)^{-1}x_0 \rangle} \leq y_0 \leq \hat{y}_0 + t_{(\alpha/2, n-p)} \sqrt{s^2 \langle 1 + x_0'(X'X)^{-1}x_0 \rangle} \quad [11.5]$$

11.5 Regression Diagnostics

As was the case with SLR, it is important to use graphical tools to diagnose deficiencies in MLR. The following residuals plots are very important: normal probability plots of residuals, residuals versus predicted (to identify curvature or heteroscedasticity), residuals versus time sequence or location (to identify trends), and residuals versus any candidate explanatory variables not in the model (to identify variables, or appropriate transformations of them, which may be used to improve the model fit).

11.5.1 Partial Residual Plots

As with SLR, curvature in a plot of residuals versus an explanatory variable included in the model indicates that a transformation of that explanatory variable is required. Their relationship should be linear. To see this relation, however, residuals should not be plotted directly against explanatory variables; the other explanatory variables will influence these plots. For example, curvature in the relationship between e and x_1 may show up in the plot of e versus x_2 , erroneously indicating that a transformation of x_2 is required. To avoid such effects, partial residuals plots (also called adjusted variable plots) should be constructed.

The partial residual is

$$e_j^* = y - \hat{y}_{(j)}$$

where $\hat{y}_{(j)}$ is the predicted value of y from a regression equation where x_j is left out of the model. All other candidate explanatory variables are present.

This partial residual is then plotted versus an adjusted explanatory variable

$$x_j^* = x - \hat{x}_{(j)}$$

where $\hat{x}_{(j)}$ is the x_j predicted from a regression against all other explanatory variables. So x_j is treated as a response variable in order to compute its adjusted value. The partial plot (e_j^* versus x_j^*) describes the relationship between y and the j th explanatory variable after all effects of the other explanatory variables have been removed. Only the partial plot accurately indicates whether a transformation of x_j is necessary.

11.5.2 Leverage and Influence

The regression diagnostics of Chapter 9 are much more important in MLR than in SLR. It is very difficult when performing multiple regression to recognize points of high leverage or high influence from any set of plots. This is because the explanatory variables are multidimensional. One observation may not be exceptional in terms of each of its explanatory variables taken one at a time, but viewed in combination it can be very exceptional. Numerical diagnostics can accurately detect such anomalies.

The leverage statistic $h_i = x_0'(X'X)^{-1}x_0$ expresses the distance of a given point x_0 from the center of the sample observations (see also section 11.4.3). It has two important uses in MLR. The first is the direct extension of its use in SLR -- to identify points unusual in value of the explanatory variables. Such points warrant further checking as possible errors, or may indicate a poor model (transformation required, relationships not linear, etc.).

The second use of h_i is when making predictions. The leverage value for a prediction should not exceed the largest h_i in the original data set. Otherwise an extrapolation beyond the envelope surrounding the original data is being attempted. The regression model may not fit well in that region. It is sometimes difficult to recognize that a given x_0 for which a predicted \hat{y} is attempted is outside the boundaries of the original data. This is because the point may not be

beyond the bounds of any of its individual explanatory variables. Checking the leverage statistic guards against an extrapolation that is difficult to detect from a plot of the data.

Example 1

Variations in chemical concentrations within a steeply dipping aquifer are to be described by location and depth. The data are concentrations (C) plus three coordinates: distance east (DE), distance north (DN), and well depth (D). Data were generated using $C = 30 + 0.5 D + \epsilon$. Any acceptable regression model should closely reproduce this true model, and should find C to be independent of DE and DN. Three pairwise plots of explanatory variables (figure 11.1) do not reveal any "outliers" in the data set. Yet compared to the critical leverage statistic $h_i = 3p/n = 0.6$, and critical influence statistic $DFFITs = 2\sqrt{p/n} = 0.9$, the 16th observation is found to be a point of high leverage and high influence (table 11.1). In figure 11.2 the axes have been rotated, showing observation 16 to be lying outside the plane of occurrence of the rest of the data, even though its individual values for the three explanatory variables are not unusual.

Obs. #	DE	DN	D	C	h_i	DFFITs
1	1	1	4.2122	30.9812	0.289433	-0.30866
2	2	1	8.0671	33.1540	0.160670	-0.01365
3	3	1	10.7503	37.1772	0.164776	0.63801
4	4	1	11.9187	35.3864	0.241083	-0.04715
5	1	2	11.2197	35.9388	0.170226	0.42264
6	2	2	12.3710	31.9702	0.086198	-0.51043
7	3	2	12.9976	34.9144	0.087354	-0.19810
8	4	2	15.0709	36.5436	0.165040	-0.19591
9	1	3	12.9886	38.3574	0.147528	0.53418
10	2	3	18.3469	39.8291	0.117550	0.45879
11	3	3	20.0328	40.0678	0.121758	0.28961
12	4	3	20.5083	37.4143	0.163195	-0.47616
13	1	4	17.6537	35.3238	0.165025	-0.59508
14	2	4	17.5484	34.7647	0.105025	-0.77690
15	3	4	23.7468	40.7207	0.151517	0.06278
16	4	4	13.1110	42.3420	0.805951	4.58558
17	1	5	20.5215	41.0219	0.243468	0.38314
18	2	5	23.6314	40.6483	0.165337	-0.08027
19	3	5	24.1979	42.8845	0.160233	0.17958
20	4	5	28.5071	43.7115	0.288632	0.09397

Table 11.1 Data and diagnostics for Example 1

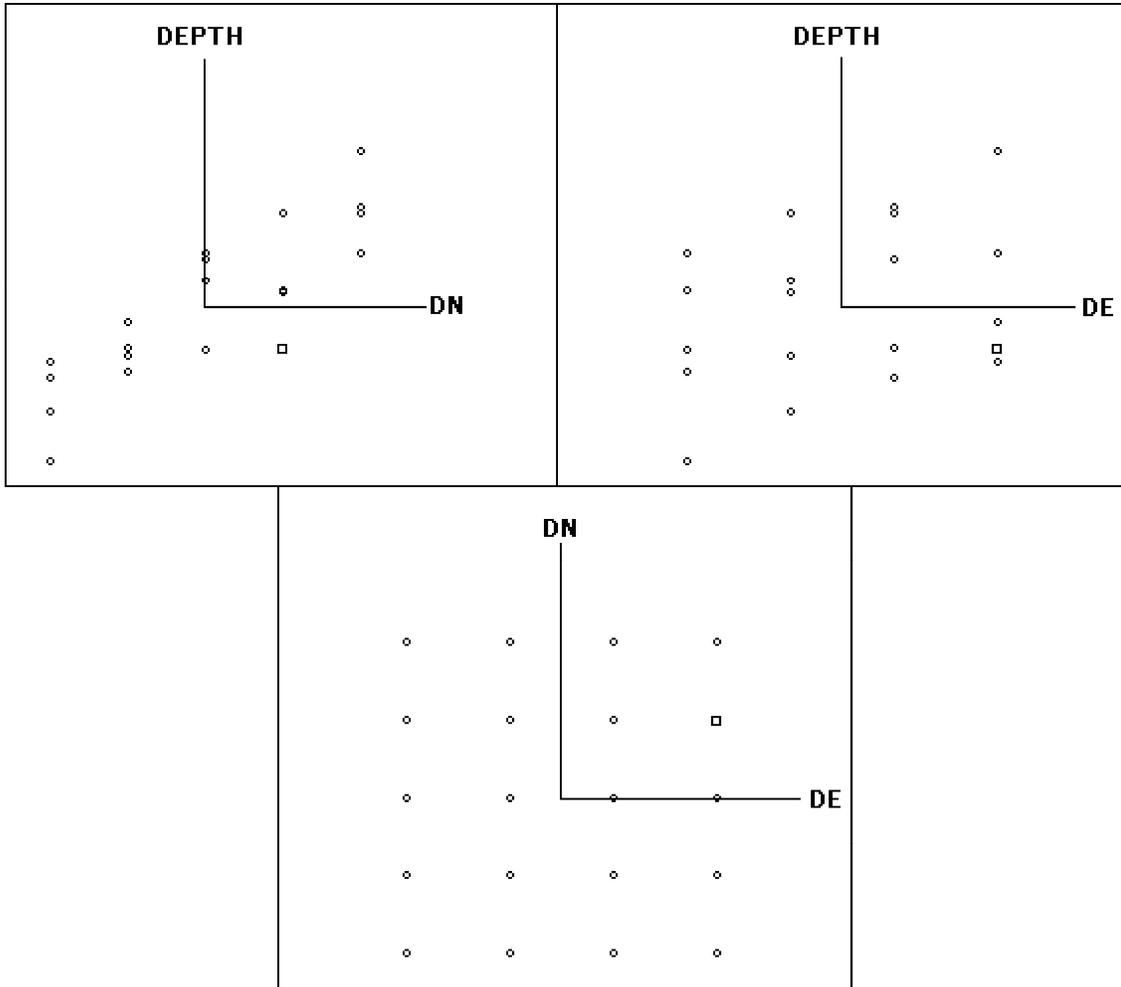


Figure 11.1 Scatterplot matrix for the 3 explanatory variables (obs. 16 is shown as a square)

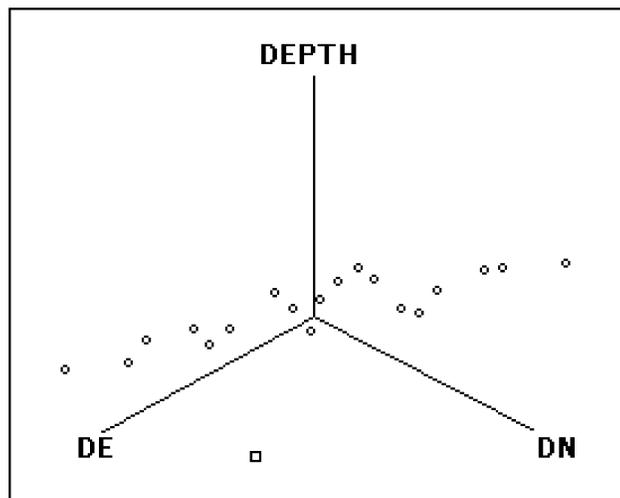


Figure 11.2 Rotated scatterplot showing the position of the high leverage point (obs. 16 is shown as a square)

The depth value for observation 16 was set as a "typographical error", and should be 23.111 instead of 13.111. What does this error and resulting high leverage point do to a regression of concentration versus the three explanatory variables? From the t-ratios of table 11.2 it is seen that DN and perhaps DE appear to be significantly related to Conc, but that depth (D) is not. This is exactly opposite of what is known to be true.

Conc = 28.9 + 0.991 DE + 1.60 DN + 0.091 D				
n = 20	s = 2.14	R ² = 0.71		
<u>Parameter</u>	<u>Estimate</u>	<u>Std.Err(β)</u>	<u>t-ratio</u>	<u>p</u>
Intercept β ₀	28.909	1.582	18.28	0.000
Slopes β _k				
DE	0.991	0.520	1.90	0.075
DN	1.596	0.751	2.13	0.049
D	0.091	0.186	0.49	0.632

Table 11.2 Regression statistics for Example 1

One outlier has had a severe detrimental effect on the regression coefficients and model structure. Points of high leverage and influence should always be examined before accepting a regression model, to determine if they represent errors. Suppose that the "typographical error" was detected and corrected. Table 11.3 shows that the resulting regression relationship is drastically changed:

C = 29.2 - 0.419 DE - 0.82 DN + 0.710 D				
n = 20	s = 1.91	R ² = 0.77		
<u>Parameter</u>	<u>Estimate</u>	<u>Std.Err(β)</u>	<u>t-ratio</u>	<u>p</u>
Intercept β ₀	29.168	1.387	21.03	0.000
Slopes β _k				
DE	-0.419	0.833	-0.50	0.622
DN	-0.816	1.340	-0.61	0.551
D	0.710	0.339	2.10	0.052

Table 11.3 Regression statistics for the corrected Example 1 data

Based on the t-statistics, DE and DN are not significantly related to C, while depth is related. The intercept of 29 is close to the true value of 30, and the slope for depth (0.7) is not far from the true value of 0.5. For observation 16, $h_i = 0.19$ and $DFFITS = 0.48$, both well below their critical values. Thus no observations have undue influence on the regression equation. Since

DE and DN do not appear to belong in the regression model, dropping them produces the equation of table 11.4, with values very close to the true values from which the data were generated. Thus by using regression diagnostics to inspect observations deemed unusual, a poor regression model was turned into an acceptable one.

Conc = 29.0 + 0.511 D				
n = 20	s = 1.83	$R^2 = 0.77$		
<u>Parameter</u>	<u>Estimate</u>	<u>Std.Err(β)</u>	<u>t-ratio</u>	<u>p</u>
Intercept β_0	29.036	1.198	24.23	0.000
Slope D	0.511	0.067	7.65	0.000
Table 11.4 Final regression model for the corrected Example 1 data				

11.5.3 Multi-Collinearity

It is very important that practitioners of MLR understand the causes and consequences of multi-collinearity, and can diagnose its presence. Multi-collinearity is the condition where at least one explanatory variable is closely related to one or more other explanatory variables. It results in several undesirable consequences for the regression equation, including:

- 1) Equations acceptable in terms of overall F-tests may have slope coefficients with magnitudes which are unrealistically large, and whose partial F or t-tests are found to be insignificant.
- 2) Coefficients may be unrealistic in sign (a negative slope for a regression of streamflow vs. precipitation, (etc). Usually this occurs when two variables describing approximately the same thing are counter-balancing each other in the equation, having opposite signs.
- 3) Slope coefficients are unstable. A small change in one or a few data values could cause a large change in the coefficients.
- 4) Automatic procedures such as stepwise, forwards and backwards methods produce different models judged to be "best".

Concern over multi-collinearity should be strongest when the purpose is to make inferences about coefficients. Concern can be somewhat less when only predictions are of interest, provided that these predictions are for cases within the observed range of the x data.

An excellent diagnostic for measuring multi-collinearity is the variance inflation factor (VIF) presented by Marquardt (1970). For variable j the VIF is

$$\text{VIF}_j = 1/(1-R_j^2) \quad [11.6]$$

where R_j^2 is the R^2 from a regression of the jth explanatory variable on all of the other explanatory variables -- the equation used for adjustment of x_j in partial plots. The ideal is VIF_j

$\cong 1$, corresponding to $R_j^2 \cong 0$. Serious problems are indicated when $VIF_j > 10$ ($R_j^2 > 0.9$). A useful interpretation of VIF is that multi-collinearity "inflates" the width of the confidence interval for the j th regression coefficient by the amount $\sqrt{VIF_j}$ compared to what it would be with a perfectly independent set of explanatory variables.

11.5.3.1 Solutions for multi-collinearity

There are four options for working with a regression equation having one or more high VIF values.

- 1) **Center the data.** A simple solution which works in some specific cases is to center the data. Multi-collinearity can arise when some of the explanatory variables are functions of other explanatory variables, such as for a polynomial regression of y against x and x^2 . When x is always of one sign, there may be a strong relationship between it and its square. Centering redefines the explanatory variables by subtracting a constant from the original variable, and then recomputing the derived variables. This constant should be one which produces about as many positive values as negative values, such as the mean or median. When all of the derived explanatory variables are recomputed as functions (squares, products, etc.) of these centered variables, their multi-collinearity will be reduced.

Centering is a mathematical solution to a mathematical problem. It will not reduce multi-collinearity between two variables which are not mathematically derived one from another. It is particularly useful when the original explanatory variable has been defined with respect to some arbitrary datum (time, distance, temperature) and is easily fixed by resetting the datum to roughly the middle of the data. In some cases the multi-collinearity can be so severe that the numerical methods used by the statistical software fail to perform the necessary matrix computations correctly. Such numerical problems occur frequently when doing trend surface analysis (e.g., fitting a high order polynomial of distances north of the equator and west of Greenwich) or trend analysis (e.g., values are a polynomial of years). This will be demonstrated in Example 2.

- 2) **Eliminate variables.** In some cases prior judgment suggests the use of several different variables which describe related but not identical attributes. Examples of this might be: air temperature and dew point temperature, the maximum 1-hour rainfall, and the maximum 2-hour rainfall, river basin population and area in urban land use, basin area forested and basin area above 6,000 feet elevation, and so on. Such variables may be strongly related as shown by their VIFs, and one of them must be eliminated on judgmental grounds, or on the basis of comparisons of models fit with one eliminated versus the other eliminated, in order to lower the VIF.

- 3) **Collect additional data.** Multi-collinearity can sometimes be solved with only a few additional but strategically selected observations. Suppose some attributes of river basins are being studied, and small basins tend to be heavily forested while large basins tend to be less heavily forested. Discerning the relative importance of size versus the importance of forest cover will prove to be difficult. Strong multi-collinearity will result from including both variables in the regression equation. To solve this and allow the effects of each variable to be judged separately, collect additional samples from a few small less forested basins and a few large but heavily-forested basins. This produces a much more reliable model. Similar problems arise in ground-water quality studies, where rural wells are shallow and urban wells are deeper. Depth and population density may show strong multi-collinearity, requiring some shallow urban and deeper rural wells to be sampled.
- 4) **Perform ridge regression.** Ridge regression was proposed by Hoerl and Kennard (1970). Montgomery and Peck (1982) give a good brief discussion of it. It is based on the idea that the variance of the slope estimates can be greatly reduced by introducing some bias into them. It is a controversial but useful method in multiple regression.

Example 2 -- centering

The natural log of concentration of some contaminant in a shallow groundwater plume is to be related to distance east and distance north of a city. The city was arbitrarily chosen as a geographic datum. The data are presented in table 11.5.

Since the square of distance east (DE²) must be strongly related to DE, and similarly DN² and DN, and DE•DN with both DE and DN, multi-collinearity between these variables will be detected by their VIFs. Using the rule that any VIF above 10 indicates a strong dependence between variables, table 11.6 shows that all variables have high VIFs. Therefore all of the slope coefficients are unstable, and no conclusions can be drawn from the value of 10.5 for DE, or 15.1 for DN, etc. This cannot be considered a good regression model, even though the R² is large.

Obs. #	C	ln(C)	DE	DN	DESQ	DNSQ	DE•DN
1	14	2.63906	17	48	289	2304	816
2	88	4.47734	19	48	361	2304	912
3	249	5.51745	21	48	441	2304	1008
4	14	2.63906	23	48	529	2304	1104
5	29	3.36730	17	49	289	2401	833
6	147	4.99043	19	49	361	2401	931
7	195	5.27300	21	49	441	2401	1029
8	28	3.33220	23	49	529	2401	1127
9	21	3.04452	17	50	289	2500	850
10	276	5.62040	19	50	361	2500	950
11	219	5.38907	21	50	441	2500	1050
12	40	3.68888	23	50	529	2500	1150
13	22	3.09104	17	51	289	2601	867
14	234	5.45532	19	51	361	2601	969
15	203	5.31320	21	51	441	2601	1071
16	35	3.55535	23	51	529	2601	1173
17	15	2.70805	17	52	289	2704	884
18	115	4.74493	19	52	361	2704	988
19	180	5.19296	21	52	441	2704	1092
20	16	2.77259	23	52	529	2704	1196

Table 11.5 Data for Example 2

DE and DN are centered by subtracting their medians. Following this, the three derived variables DESQ, DNSQ and DEDN are recomputed, and the regression rerun. Table 11.7 give the results, showing that all multi-collinearity is completely removed. The coefficients for DE and DN are now more reasonable in size, while the coefficients for the derived variables are exactly the same. The t-statistics for DE and DN have changed because their uncentered values were unstable and t-tests unreliable. Note that the s and R^2 are also unchanged. In fact, this is exactly the same model as the uncentered equation, but only in a different and centered coordinate system.

$\ln(C) = -479 + 10.5 \text{ DE} + 15.1 \text{ DN} - 0.264 \text{ DESQ} - 0.151 \text{ DNSQ} + 0.0014 \text{ DEDN}$					
$n = 20$	$s = 0.27$	$R^2 = 0.96$			
Parameter	Estimate	Std.Err(β)	t-ratio	p	VIF
Intercept β_0	-479.03	91.66	-5.23	0.000	
Slopes β_k					
DE	10.55	1.12	9.40	0.000	1751.0
DN	15.14	3.60	4.20	0.001	7223.9
DESQ	-0.26	0.015	-17.63	0.000	501.0
DNSQ	-0.15	0.04	-4.23	0.001	7143.9
DEDN	0.001	0.02	0.07	0.943	1331.0

Table 11.6 Regression statistics and VIFs for Example 2

$\ln(C) = 5.76 + 0.048 \text{ DE} + 0.019 \text{ DN} - 0.264 \text{ DESQ} - 0.151 \text{ DNSQ} + 0.001 \text{ DNDE}$					
$n = 20$	$s = 0.27$	$R^2 = 0.96$			
Parameter	Estimate	Std.Err(β)	t-ratio	p	VIF
Intercept β_0	5.76	0.120	48.15	0.000	
Slopes β_k					
DE	0.048	0.027	1.80	0.094	1.0
DN	0.019	0.042	0.44	0.668	1.0
DESQ	-0.264	0.015	-17.63	0.000	1.0
DNSQ	-0.151	0.036	-4.23	0.001	1.0
DEDN	0.001	0.019	0.07	0.943	1.0

Table 11.7 Regression statistics and VIFs for centered Example 2 data

11.6 Choosing the Best MLR Model

One of the major issues in multiple regression is the appropriate approach to variable selection. The benefit of adding additional variables to a multiple regression model is to account for or explain more of the variance of the response variable. The cost of adding additional variables is that the degrees of freedom decreases, making it more difficult to find significance in hypothesis tests and increasing the width of confidence intervals. A good model will explain as much of the variance of y as possible with a small number of explanatory variables.

The first step is to consider only explanatory variables which ought to have some effect on the dependent variable. There must be plausible theory behind why a variable might be expected to influence the magnitude of y . Simply minimizing the SSE or maximizing R^2 are not sufficient criteria. In fact, any explanatory variable will reduce the SSE and increase the R^2 by some small amount, even those irrelevant to the situation (or even random numbers!). The benefit of

adding these unrelated variables, however, is small compared to the cost of a degree of freedom. Therefore the choice of whether to add a variable is based on a "cost-benefit analysis", and variables enter the model only if they make a significant improvement in the model. There are at least two types of approaches for evaluating whether a new variable sufficiently improves the model. The first approach uses partial F or t-tests, and when automated is often called a "stepwise" procedure. The second approach uses some overall measure of model quality. The latter has many advantages.

11.6.1 Stepwise Procedures

Stepwise procedures are automated model selection methods in which the computer algorithm determines which model is preferred. There are three versions, usually called forwards, backwards, and stepwise. These procedures use a sequence of partial F or t-tests to evaluate the significance of a variable. The three versions do not always agree on a "best" model, especially when multi-collinearity is present. They also do not evaluate all possible models, and so cannot guarantee that the "best" model is even tested. They were developed prior to modern computer technology, taking shortcuts to avoid running all possible regression equations for comparison. Such shortcuts are no longer necessary.

Forward selection starts with only an intercept and adds variables to the equation one at a time. Once in, each variable stays in the model. All variables not in the model are evaluated with partial F or t statistics in comparison to the existing model. The variable with the highest significant partial F or t statistic is included, and the process repeats until either all available variables are included or no new variables are significant. One drawback to this method is that the resulting model may have coefficients which are not significantly different from zero; they must only be significant when they enter. A second drawback is that two variables which each individually provide little explanation of y may never enter, but together the variables would explain a great deal. Forward selection is unable to capitalize on this situation.

Backward elimination starts with all explanatory variables in the model and eliminates the one with the lowest partial-F statistic (lowest $|t|$). It stops when all remaining variables are significant. The backwards algorithm does ensure that the final model has only significant variables, but does not ensure a "best" model because it also cannot consider the combined significance of groups of variables.

Stepwise regression combines the ideas of forward and backward. It alternates between adding and removing variables, checking significance of individual variables within and outside the model. Variables significant when entering the model will be eliminated if later they test as insignificant. Even so, stepwise does not test all possible regression models.

Example 3:

Haan (1977) attempted to relate the mean annual runoff of several streams (ROFF) with 9 other variables: the precipitation falling at the gage (PCIP), the drainage area of the basin (AREA), the average slope of the basin (SLOPE), the length of the drainage basin (LEN), the perimeter of the basin (PERIM), the diameter of the largest circle which could be inscribed within the drainage basin (DI), the "shape factor" of the basin (Rs), the stream frequency -- the ratio of the number of streams in the basin to the basin area (FREQ), and the relief ratio for the basin (Rr). The data are found in Appendix C14. Haan chose to select a 3-variable model (using PCIP, PERIM and Rr) based on a levelling off of the incremental increase in R^2 as more variables were added to the equation (see figure 11.3).

What models would be selected if the stepwise or overall methods are applied to this data? If a forwards routine is performed, no single variables are found significant at $\alpha = 0.05$, so an intercept-only model is declared "best". Relaxing the entry criteria to a larger α , AREA is first entered into the equation. Then Rr, PCIP, and PERIM are entered in that order. Note that AREA has relatively low significance once the other three variables are added to the model (Model 4).

Forwards		Model 1	Model 2	Model 3	Model 4
AREA	β	0.43	0.81	0.83	-0.62
	t	1.77	4.36	4.97	-1.68
Rr	β		0.013	0.011	0.009
	t		3.95	3.49	4.89
PCIP	β			0.26	0.54
	t			1.91	5.05
PERIM	β				1.02
	t				4.09

The backwards model begins with all variables in the model. It checks all partial t or F statistics, throwing away the variable which is least significant. Here the least significant single variable is AREA. So while forwards made AREA the first variable to bring in, backwards discarded AREA first of all! Then other variables were also removed, resulting in a model with Rr, PCIP, PERIM, DI and FREQ remaining in the model. Multi-collinearity between measures of drainage basin size, as well as between other variables, has produced models from backwards and forwards procedures which are quite different from each other. The slope coefficient for DI is also negative, suggesting that runoff decreases as basin size increases. Obviously DI is counteracting another size variable in the model (PERIM) whose coefficient is large.

11.6.2 Overall Measures of Quality.

Three newer statistics can be used to evaluate each of the 2^k regressions equations possible from k candidate explanatory variables. These are Mallows's C_p , the PRESS statistic, and the adjusted R^2 .

Mallows's C_p , is designed to achieve a good compromise between the desire to explain as much variance in y as possible (minimize bias) by including all relevant variables, and to minimize the variance of the resulting estimates (minimize the standard error) by keeping the number of coefficients small. The C_p statistic is

$$C_p = p + \frac{(n-p) \cdot (s_p^2 - \hat{\sigma}^2)}{\hat{\sigma}^2} \quad [11.7]$$

where n is the number of observations, p is the number of coefficients (number of explanatory variables plus 1), s_p^2 is the mean square error (MSE) of this p coefficient model, and $\hat{\sigma}^2$ is the best estimate of the true error, which is usually taken to be the minimum MSE among the 2^k possible models. The best model is the one with the lowest C_p value. When several models have nearly equal C_p values, they may be compared in terms of reasonableness, multi-collinearity, importance of high influence points, and cost in order to select the model with the best overall properties.

The second overall measure is the PRESS statistic. PRESS was defined in Chapter 9 as the sum of the squared prediction errors $e_{(i)}$. By minimizing PRESS, the model with the least error in the prediction of future observations is selected. PRESS and C_p generally agree as to which model is "best", even though their criteria for selection are not identical.

A third overall measure is the adjusted R^2 (R_a^2). This is an R^2 value adjusted for the number of explanatory variables (or equivalently, the degrees of freedom) in the model. The model with the highest R_a^2 is identical to the one with the smallest standard error (s) or its square, the mean squared error (MSE). To see this, in Chapter 9 R^2 was defined as a function of the total (SS_y) and error (SSE) sum of squares for the regression model:

$$R^2 = 1 - (SSE / SS_y) \quad [11.8]$$

The weakness of R^2 is that it must increase, and the SSE decrease, when any additional variable is added to the regression. This happens no matter how little explanatory power that variable has. R_a^2 is adjusted to offset the loss in degrees of freedom by including as a weight the ratio of total to error degrees of freedom:

$$R_a^2 = 1 - \frac{(n-1)}{(n-p)} \frac{SSE}{SS_y} = 1 - \frac{MSE}{(SS_y/(n-1))} \quad [11.9]$$

As $(SS_Y/(n-1))$ is constant for a given data set, R^2_a increases as MSE decreases. Either maximize R^2_a or minimize MSE as an overall measure of quality. However, when p is considerably smaller than n , R^2_a is a less sensitive measure than either PRESS or C_p . PRESS has the additional advantage of being a validation criteria.

Overall methods use the computer to perform large computations (such as C_p and PRESS for many models), letting the scientist judge which model to use. This allows flexibility in choosing between models. For example, two "best" models may be nearly identical in terms of their C_p , R^2_a and/or PRESS statistics, yet one involves variables that are much less expensive to measure than the other. The less expensive model can be selected with confidence. In contrast, stepwise procedures ask the computer to judge which model is best. Their combination of inflexible criteria and inability to test all models often results in the selection of something much less than the best model.

Example 3, continued

Instead of the stepwise procedures run on Haan's data, models are evaluated using the overall statistics C_p and PRESS. Smaller values of C_p and PRESS are associated with better models. Computing PRESS and C_p for the $2^9 = 512$ possible regression models can be done with modern statistical software. A list of these statistics for the two best k -variable models, where best is defined as the highest R^2 , is given in table 11.8.

Based on C_p , the best model would be the 5 variable model having PCIP, PERIM, DI, FREQ and Rr as explanatory variables -- the same model as selected by stepwise and forwards. Remember that there is no guarantee that stepwise procedures regularly select the lowest C_p or PRESS models. The advantage of using an overall statistic like C_p is that options are given to the scientist to select what is best. If the modest multi-collinearity ($VIF=5.1$) between PERIM and DI is of concern, with its resultant negative slope for DI, the model with the next lowest C_p that does not contain both these variables (a four-variable model with $C_p= 3.6$) could be selected. If the scientist decided AREA must be in the model, the lowest CP model containing AREA (the same four-variable model) could be selected. C_p and PRESS allow model choice to be based on multiple criteria such as prediction quality (PRESS), low VIF, cost, etc..

linearity. Considerable help can be obtained from statistics such as R^2 (maximize it), or SSE or PRESS (minimize it). Many transformations can be rapidly checked with such statistics, but a residuals plot should always be inspected prior to making any final decision.

3) **Which of several models, each with the same y and with the same number of explanatory variables, is preferable?** Use of R^2 , SSE, or PRESS is appropriate here, but back it up with a residuals plot.

4) **Which of several nested models, each with the same y, is preferable?** Use the partial F test between any pair of nested models to find which is best. One may also select the model based on minimum C_p or minimum PRESS.

5) **Which of several models is preferable when each uses the same y variable but are not necessarily nested?** C_p or PRESS must be used in this situation.

11.8 Analysis of Covariance

Often there are factors which influence the dependent variable which are not appropriately expressed as a continuous variable. Examples of such grouped or qualitative variables include location (stations, aquifers, positions in a cross section), or time (day & night; winter & summer; before & after some event such as a flood, a drought, operation of a sewage treatment plant or reservoir). These factors are perfectly valid explanatory variables in a multiple regression context. They can be incorporated by the use of binary or "dummy" variables, essentially blending regression and analysis of variance into an analysis of covariance.

11.8.1 Use of One Binary Variable

To the simple one-variable regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad [11.10]$$

(again with subscripts i assumed), an additional factor is believed to have an important influence on Y for any given value of X . Perhaps this factor is a seasonal one: cold season versus warm season -- where some precise definition exists to classify all observations as either cold or warm.

A second variable, a binary variable Z , is added to the equation where

$$Z_i = \begin{cases} 0 & \text{if } i \text{ is from cold season} \\ 1 & \text{if } i \text{ is from warm season} \end{cases}$$

to produce the model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon. \quad [11.11]$$

When the slope coefficient β_2 is significant, model 11.11 would be preferred to the SLR model 11.10. This also says that the relationship between Y and X is affected by season.

Consider $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$. The null hypothesis is tested using a student's t-test with $(n-3)$ degrees of freedom. There are $(n-3)$ because there are 3 betas being estimated. If the partial $|t| \geq t_{\alpha/2}$, H_0 is rejected, inferring that there are two models:

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X && \text{for the cold season } (Z = 0), \text{ and} \\ \hat{Y} &= b_0 + b_1 X + b_2 && \text{for the warm season } (Z = 1), \text{ or} \\ &= (b_0 + b_2) + b_1 X.\end{aligned}$$

Therefore the regression lines differ for the two seasons. Both seasons have the same slope, but different intercepts, and will plot as two parallel lines (figure 11.4).

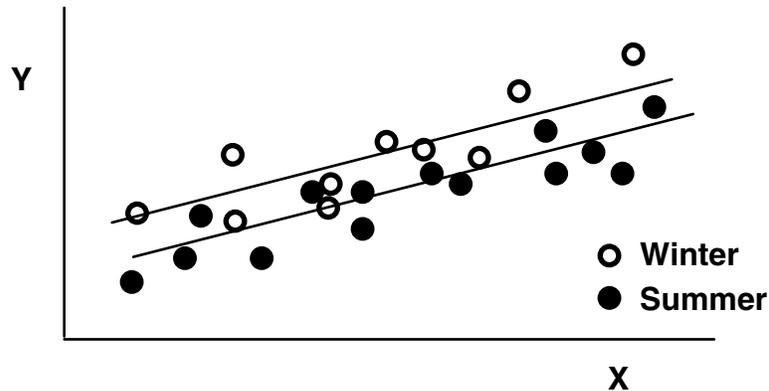


Figure 11.4 Regression lines for data differing in intercept between two seasons

Suppose that the relationship between X and Y for the two seasons is suspected not only to differ in intercept, but in slope as well. Such a model is written as:

$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 Z X + \varepsilon && [11.12] \\ \text{or } Y &= (\beta_0 + \beta_2 Z) + (\beta_1 + \beta_3 Z) \cdot X + \varepsilon\end{aligned}$$

The intercept equals β_0 for the cold season and $\beta_0 + \beta_2$ for the warm season; the slope equals β_1 for the cold season and $\beta_1 + \beta_3$ for the warm season. This model is referred to as an "interaction model" because of the use of the explanatory variable $Z X$, the interaction (product) of the original predictor X and the binary variable Z .

To determine whether the simple regression model with no Z terms can be improved upon by model 11.12, the following hypotheses are tested:

$$H_0: \beta_2 = \beta_3 = 0 \text{ versus } H_1: \beta_2 \text{ and/or } \beta_3 \neq 0.$$

A nested F statistic is computed
$$F = \frac{(SSE_s - SSE_c) / (df_s - df_c)}{(SSE_c / df_c)}$$

where s refers to the simpler (no Z terms) model, and c refers to the more complex model. For the two nested models 11.10 and 11.12 this becomes

$$F = \frac{(SSE_{11.10} - SSE_{11.12}) / 2}{MSE_{11.12}}$$

where $MSE_{11.12} = SSE_{11.12} / (n-4)$, rejecting H_0 if $F > F_{\alpha, 2, n-4}$.

If H_0 is rejected, model 11.12 should also be compared to model 11.11 (the shift in intercept only model) to determine whether there is a change in slope in addition to the change in intercept, or whether the rejection of model 11.10 in favor of 11.12 was due only to a shift in intercept. The null hypothesis $H_0': \beta_3 = 0$ is compared to $H_1': \beta_3 \neq 0$ using the test statistic

$$F = \frac{(SSE_{11.11} - SSE_{11.12}) / 1}{MSE_{11.12}}$$

rejecting H_0' if $F > F_{\alpha, 1, n-4}$.

Assuming H_0 and H_0' are both rejected, the model can be expressed as the two separate equations (see figure 11.5):

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X && \text{cold season} \\ \hat{Y} &= (b_0 + b_2) + (b_1 + b_3) X && \text{warm season} \end{aligned}$$

Furthermore, the coefficient values in these two equations will be exactly those computed if the two regressions were estimated by separating the data, and computing two separate regression equations. By using analysis of covariance, however, the significance of the difference between those two equations has been established.

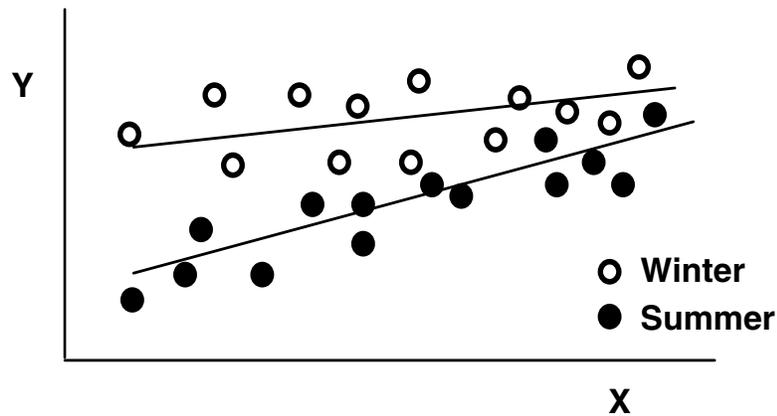


Figure 11.5 Regression lines differing in slope and intercept for data from two seasons

11.8.2 Multiple Binary Variables

In some cases, the factor of interest must be expressed as more than two categories: 4 seasons, 12 months, 5 stations, 3 flow conditions (rising limb, falling limb, base flow), etc. To illustrate, assume there are precise definitions of 3 flow conditions so that all discharge (X_i) and

concentration (Y_i) pairs are classified as either rising, falling, or base flow. Two binary variables are required to express these three categories -- there is always one less binary variable required than the number of categories.

$$\text{Let } R_i = \begin{cases} 1 & \text{if } i \text{ is a rising limb observation} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Let } D_i = \begin{cases} 1 & \text{if } i \text{ is falling limb observation} \\ 0 & \text{otherwise} \end{cases}$$

so that	category	value of R	value of D
	rising	1	0
	falling	0	1
	base flow	0	0

The following model results:

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \varepsilon \quad [11.13]$$

To test $H_0: \beta_2 = \beta_3 = 0$ versus $H_1: \beta_2$ and/or $\beta_3 \neq 0$, F tests comparing simpler and more complex models are again performed. To compare model 11.13 versus the SLR model 11.10 with no rising or falling terms,

$$F = \frac{(\text{SSE}_{11.10} - \text{SSE}_{11.13}) / 2}{\text{MSE}_{11.13}} \quad \text{where } \text{MSE}_{11.13} = \text{SSE}_{11.13} / (n-4),$$

rejecting H_0 if $F > F_{2, n-4, \alpha}$.

To test for differences between each pair of categories:

1. Is rising different from base flow? This is tested using the t-statistic on the coefficient β_2 .
If $|t| > t_{\alpha/2}$ on $n-4$ degrees of freedom, reject H_0 where $H_0: \beta_2 = 0$.
2. Is falling different from base flow? This is tested using the t-statistic on the coefficient β_3 .
If $|t| > t_{\alpha/2}$ with $n-4$ degrees of freedom, reject H_0 where $H_0: \beta_3 = 0$.
3. Is rising different from falling? There are two ways to determine this.
 - (a) the standard error of the difference ($b_2 - b_3$) must be known. The null hypothesis is $H_0: (\beta_2 - \beta_3) = 0$. The estimated variance of $b_2 - b_3$,

$$\text{Var}(b_2 - b_3) = \text{Var}(b_2) + \text{Var}(b_3) - 2\text{Cov}(b_2, b_3)$$
 where Cov is the covariance between b_2 and b_3 . To determine these terms, the matrix $(X'X)^{-1}$ and s^2 (s^2 is the mean square error) are required. Then

$\widehat{\text{Var}}(b_2) = C_{22} \cdot s^2$, $\widehat{\text{Var}}(b_3) = C_{33} \cdot s^2$, and $\widehat{\text{Cov}}(b_2, b_3) = C_{23} \cdot s^2$.
 The test statistic is $t = (b_2 - b_3) / \sqrt{\widehat{\text{Var}}(b_2 - b_3)}$. If $|t| > t_{\alpha/2}$ with $n-4$ degrees of freedom, reject H_0 .

(b) The binary variables can be re-defined so that a direct contrast between rising and falling is possible. This occurs when either is set as the (0,0) "default" case. This will give answers identical to (a).

Ever greater complexity can be added to these kinds of models, using multiple binary variables and interaction terms such as

$$Y = \beta_0 + \beta_1 X + \beta_2 R + \beta_3 D + \beta_4 R X + \beta_5 D X + \varepsilon. \quad [11.14]$$

The procedures for selecting models follow the pattern described above. The significance of an individual β coefficient, given all the other β s, can be determined from the t statistic. The comparison of two models, where the set of explanatory variables for one model is a subset of those used in the other model, is computed by a nested F test. The determination of whether two coefficients in a given model differ significantly from each other is computed either by re-defining the variables, or by using a t test after estimating the variance of the difference between the coefficients based on the elements of the $(X'X)^{-1}$ matrix and s^2 .

Model #	Explanatory variables	SSE	df(error)
1	X, X ²	69.89	124
2	X, X ² , S	65.80	123
3	X, X ² , S, SX	65.18	122
4	X, X ² , S, SX, SX ²	64.84	121
5	X, X ² , W	63.75	123
6	X, X ² , W, WX	63.53	122
7	X, X ² , W, WX, WX ²	63.46	121
8	X, X ² , S, W	63.03	122
9	X, X ² , S, W, SX, WX	62.54	120
10	X, X ² , S, W, SX, WX, SX ² , WX ²	61.45	118

11.3 The Ogallala aquifer was investigated to determine relationships between uranium and other concentrations in its waters. Construct a regression model to relate uranium to total dissolved solids and bicarbonate, using the data in Appendix C16. What is the significance of these predictor variables?

11.4 You are asked to estimate uranium concentrations in irrigation waters from the Ogallala aquifer for a local area. Four supply wells pump waters with the characteristics given below. The relative amounts of water pumped by each well are also given below. Using this and the regression equation of Exercise 11.3, estimate the mean concentration of uranium in the water applied to this area.

<u>Well #</u>	<u>Relative volume of water used</u>	<u>TDS</u>	<u>Bicarbonate</u>
1	2	500	≤ 50%
2	1	900	≤ 50%
3	1	400	> 50%
4	2	600	> 50%