Statistical Methods for the Social Sciences

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter 14

MODEL BUILDING WITH MULTIPLE REGRESSION

CHAPTER OUTLINE

This chapter introduces tools for building regression models and evaluating the effects on their fit of unusual observations or highly correlated predictors. It also shows ways of modeling variables that badly violate the assumptions of straight-line relationships with a normal response variable.

We first discuss criteria for *selecting a regression model* by deciding which of a possibly large collection of variables to include in the model. We then introduce methods for *checking regression assumptions* and evaluating the influence of individual observations. We also discuss effects of *multicollinearity*—such strong "overlap" among the explanatory variables that no one of them seems useful when the others are also in the model.

Section 14.4 introduces a ***generalized linear model*** that can handle response variables having distributions other than the normal. For example, the *gamma distribution* is useful for positive variables that exhibit skew to the right and have variability that grows with the mean. We also introduce models for nonlinear relationships, such as *exponential* increase or decrease. The final section introduces alternative regression methods with weaker assumptions, such as not assuming a functional form for the relationship or common response variability.

14.1 Model Selection Procedures

Social research studies usually have several explanatory variables. For example, for modeling mental impairment, potential predictors include income, educational attainment, an index of life events, social and environmental stress, marital status, age, self-assessment of health, number of jobs held in previous five years, number of relatives who live nearby, number of close friends, membership in social organizations, and frequency of church attendance.

Usually, the regression model for a study includes some explanatory variables for theoretical reasons, such as to analyze whether a predicted effect truly occurs under certain controls. Other explanatory variables may be included to see if they mediate the predicted effects. Others may be included for exploratory purposes, to check whether they explain other variability in the response variable. The model might also include terms to allow for interactions. In such situations, it is not simple to decide which variables to include and which to exclude from a final model.

SELECTING EXPLANATORY VARIABLES FOR A MODEL

A strategy that you might first consider is to include every potentially useful explanatory variable and then delete those terms not making statistically significant partial contributions at some preassigned α-level. Unfortunately, this usually is inadequate.

Because of correlations among the explanatory variables, any one variable may have little unique predictive power, especially when the number of predictors is large. It is conceivable that few, if any, explanatory variables would make significant *partial* contributions, given that all of the other explanatory variables are in the model.

Here are three general guidelines for selecting explanatory variables:

1. Include the relevant variables to make the model useful for theoretical purposes, so you can address hypotheses posed by the study, with sensible control and mediating variables.
2. Include enough variables to obtain good predictive power.
3. Keep the model simple.

Goal 3 is a counterbalance to goal 2. Having a large number of explanatory variables in a model has disadvantages. The correlations among them can result in inflated standard errors of the parameter estimates, and may make it impossible to assess the partial contributions of variables that are important theoretically. To avoid multicollinearity, it is helpful for the explanatory variables to be correlated with the response variable but not highly correlated among themselves.

Goal 2 of obtaining good predictive power might suggest "Maximize R^2" as a criterion for selecting a model. Because R^2 cannot decrease as you add variables to a model, however, this approach would lead you to the most complex model in the set being considered. Related to the goal 3 of simplicity, don't try to build a complex model if the data set is small. If you have only 25 observations, you won't be able to untangle the complexity of effects among 10 explanatory variables. With small to moderate sample sizes (say, 100 or less), it is safer to use relatively few predictors.

Keeping these thoughts in mind, no unique or optimal approach exists for selecting explanatory variables. For p potential predictors, since each can be either included or omitted (two possibilities for each variable), there are 2^p potential subsets. For $p = 2$, for example, there are $2^p = 2^2 = 4$ possible models: one with both x_1 and x_2, one with x_1 alone, one with x_2 alone, and one with neither variable. The set of potential models is too large to evaluate practically if p is even moderate; with $p = 7$, there are $2^7 = 128$ potential models.

Statistical software has automated variable selection procedures that scan the explanatory variables to construct a model. These routines sequentially enter or remove variables, one at a time according to some criterion. For any particular sample and set of variables, however, different procedures may select different subsets of variables, with no guarantee of selecting a sensible model. The most popular automated variable selection methods are *backward elimination*, *forward selection*, and *stepwise regression*.

BACKWARD ELIMINATION

Backward elimination begins by placing all of the explanatory variables under consideration in the model. It deletes one at a time until reaching a point where the remaining variables all make significant partial contributions to predicting y. The variable deleted at each stage is the one that is the least significant, having the largest P-value in the significance test for its effect.

Here is the sequence of steps for backward elimination: The initial model contains all potential explanatory variables. If all variables make significant partial contributions at some fixed α-level, according to the usual t test or F test, then that model is the final one. Otherwise, the explanatory variable having the largest P-value, controlling for the other variables in the model, is removed. Next, the model is refitted with that variable removed, and the partial contributions of the variables remaining in the model are reassessed, controlling for the other variables still in the model. If they are all significant, that model is the final model. Otherwise, the variable having the largest P-value is removed. The process continues until each remaining predictor explains a significant partial amount of the variability in y.

Example 14.1

Selecting Explanatory Variables for House Selling Price Example 9.10 (page 265) introduced a data set consisting of 100 observations on house selling prices with several explanatory variables. The data are in the `Houses` data file at the text website. We use y = selling price of home, with explanatory variables size of home (denoted SIZE), annual taxes (TAXES), number of bedrooms (BEDS), number of bathrooms (BATHS), and a dummy variable for whether the home is new (NEW). We use backward elimination with these variables as potential explanatory variables but without interaction terms, requiring a variable to reach significance at the $\alpha = 0.05$ level for inclusion in the model.

Table 14.1 shows the first stage of the process, fitting the model containing all the explanatory variables. The variable making the least partial contribution to the model is BATHS. Its P-value ($P = 0.85$) is the largest. Although BATHS is moderately correlated with the selling price ($r = 0.56$), the other explanatory variables together explain most of the same variability in selling price. Once those variables are in the model, BATHS is essentially redundant.

TABLE 14.1: Model Fit at Initial Stage of Backward Elimination for Predicting House Selling Price

Variable	Coef.	Std. Error	t	Sig
(Constant)	4525.75	24474.05		
SIZE	68.35	13.94	4.90	.000
NEW	41711.43	16887.20	2.47	.015
TAXES	38.13	6.81	5.60	.000
BATHS	-2114.37	11465.11	-.18	.854
BEDS	-11259.10	9115.00	-1.23	.220

When we refit the model after dropping BATHS, the only nonsignificant variable is BEDS, having a t statistic of -1.31 and P-value $= 0.19$. Table 14.2 shows the third stage, refitting the model after dropping BATHS and BEDS as explanatory variables. Each variable now makes a significant contribution, controlling for the others in the model. Thus, this is the final model. Backward elimination provides the prediction equation

$$\hat{y} = -21{,}353.8 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

Other things being equal, an extra thousand square feet of size increases the predicted selling price by about 62 thousand dollars, and having a new home increases

TABLE 14.2: Model Fit at Third Stage of Backward Elimination for Predicting House Selling Price

Variable	Coef.	Std. Error	Std. Coeff	t	Sig
(Constant)	-21353.8	13311.49			
SIZE	61.70	12.50	0.406	4.94	.000
NEW	46373.70	16459.02	0.144	2.82	.006
TAXES	37.23	6.74	0.466	5.53	.000

it by about 46 thousand dollars. With standardized variables, the equation is

$$\hat{z}_y = 0.406z_S + 0.144z_N + 0.464z_T.$$

SIZE and TAXES have similar partial effects.

If we had included interactions in the original model, we would have ended up with a different final model. However, the model given here has the advantage of simplicity, and it has good predictive power ($R^2 = 0.790$, compared to 0.793 with all the explanatory variables). ■

FORWARD SELECTION AND STEPWISE REGRESSION PROCEDURES

Whereas backward elimination begins with *all* the potential explanatory variables in the model, ***forward selection*** begins with *none* of them. It adds one variable at a time to the model until no remaining variable not yet in the model makes a significant partial contribution to predicting y. At each step, the variable added is the one that is most significant, having the smallest P-value. For quantitative explanatory variables, this is the variable having the largest t test statistic, or equivalently the one providing the greatest increase in R^2.

For the data on selling prices of homes, Table 14.3 depicts the process. The variable most highly correlated with selling price is TAXES, so it is added first. Once TAXES is in the model, SIZE provides the greatest boost to R^2, and it is significant ($P = 0.000$), so it is the second variable added. Once both TAXES and SIZE are in the model, NEW provides the greatest boost to R^2 and it is significant ($P = 0.006$), so it is added next. At this stage, BEDS gives the greatest boost to R^2 (from 0.790 to 0.793), but it does not make a significant contribution ($P = 0.194$), so the final model does not include it. In this case, forward selection reaches the same final model as backward elimination.

TABLE 14.3: Steps of Forward Selection for Predicting House Selling Price. The model chosen has predictors TAXES, SIZE, and NEW.

Step	Variables in Model	P-Value for New Term	R^2
0	None	—	0.000
1	TAXES	0.000	0.709
2	TAXES, SIZE	0.000	0.772
3	TAXES, SIZE, NEW	0.006	0.790
4	TAXES, SIZE, NEW, BEDS	0.194	0.793

Once forward selection provides a final model, not all the explanatory variables appearing in it are necessarily significantly related to y. The variability in y explained by a variable entered at an early stage may overlap with the variability explained by variables added later, so it may no longer be significant. Figure 14.1 illustrates this. The figure portrays the portion of the total variability in y explained by each of three explanatory variables. Variable x_1 explains a similar amount of variability, by itself, as x_2 or x_3. However, x_2 and x_3 between them explain much of the same variation that x_1 does. Once x_2 and x_3 are in the model, the unique variability explained by x_1 is minor.

Stepwise regression is a modification of forward selection that drops variables from the model if they lose their significance as other variables are added. The approach is the same as forward selection except that at each step, after entering the new variable, the procedure drops from the model any variables that no longer make

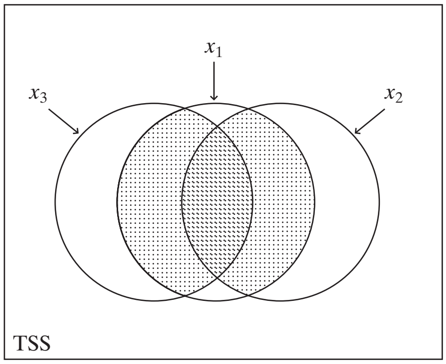


FIGURE 14.1: Variability in y Explained by x_1, x_2, and x_3. The shaded portion is the amount explained by x_1 that is also explained by x_2 and x_3.

significant partial contributions. A variable entered into the model at some stage may eventually be eliminated because of its overlap with variables entered at later stages.

For the home sales data, stepwise regression behaves the same way as forward selection. At each stage, each variable in the model makes a significant contribution, so no variables are dropped. For these variables, backward elimination, forward selection, and backward elimination all agree. This need not happen.

LIMITATIONS AND ABUSES OF AUTOMATIC SELECTION PROCEDURES

It may seem appealing to select explanatory variables automatically according to established criteria. But any variable selection method should be used with caution and should not substitute for theory and careful thought. There is no guarantee that the final model chosen will be sensible.

For instance, suppose we specify all the pairwise interactions as well as the main effects as the potential explanatory variables. In this case, it is inappropriate to remove a main effect from a model that contains an interaction composed of that variable. Yet, most software does not have this safeguard. To illustrate, we used forward selection with the home sales data, including the 5 explanatory variables as well as their 10 cross-product interaction terms. The final model has $R^2 = 0.866$, using four interaction terms (SIZE×TAXES, SIZE ×NEW, TAXES×NEW, BATHS×NEW) and the TAXES main effect. It is inappropriate, however, to use these interactions as explanatory variables without the SIZE, NEW, and BATHS main effects.

Also, a variable selection procedure may exclude an important explanatory variable that really should be in the model according to other criteria. For instance, using backward elimination with the five explanatory variables of home selling price and their interactions, TAXES was removed. At a certain stage, TAXES explained an insignificant part of the variation in selling price. Nevertheless, it is the best single predictor of selling price, having $r^2 = 0.709$ by itself. (Refer to step 1 of the forward selection process in Table 14.3.) Since TAXES is such an important determinant of selling price, it seems sensible that any final model should include it as an explanatory variable.

Although P-values provide a guide for making decisions about adding or dropping variables in selection procedures, they are not the *true* P-values for the tests conducted. We add or drop a variable at each stage according to a minimum or maximum P-value, but the sampling distribution of the maximum or minimum of a set of t or F statistics differs from the sampling distribution for the statistic for an a priori chosen test. For instance, suppose we add variables in forward selection according to whether the P-value is less than 0.05. Even if none of the potential explanatory variables truly affect y, the probability is considerably larger than 0.05 that at least one of the separate test statistics provides a P-value below 0.05 (Exercise 14.48). At least one variable that is not really important may look impressive merely due to chance.

Similarly, for the final model suggested by a particular selection procedure, any inferences conducted with it are highly approximate. In particular, P-values are likely to appear smaller than they should be and confidence intervals are likely to be too narrow, because the model was chosen that most closely reflects the data, in some sense. The inferences are more believable if performed for that model with a new set of data. (See the related discussion about *cross-validation* on page 425.)

EXPLORATORY VERSUS EXPLANATORY (THEORY-DRIVEN) RESEARCH

There is a basic difference between *explanatory* and *exploratory* modes of model selection. ***Explanatory research*** has a theoretical model to test using multiple regression. We might test whether a hypothesized spurious association disappears when other variables are controlled, for example. In such research, automated selection procedures are usually not appropriate, because theory dictates which variables are in the model.

Exploratory research, by contrast, has the goal not of examining theoretically specified relationships but merely finding a good set of explanatory variables. This approach searches for explanatory variables that give a large R^2, without concern about theoretical explanations. Thus, educational researchers might use a variable selection procedure to search for a set of test scores and other factors that predict well how students perform in college. They should be cautious about giving causal interpretations to the effects. For example, possibly the best predictor of students' success in college is whether their parents use the Internet for voice communication (with a program such as Skype).

In summary, automated variable selection procedures are no substitute for careful thought in formulating models. For most scientific research, they are not appropriate.

INDICES FOR SELECTING A MODEL: ADJUSTED R^2, PRESS, AND AIC

Instead of using an automated algorithm to choose a model, we could specify a set of potentially adequate models, and then use some established criterion to select among them. We next present some possible criteria.

Recall that maximizing R^2 is not a sensible criterion, because the most complicated model will have the largest R^2-value. This reflects the upward bias that R^2 has as an estimator of the population value of R^2. This bias can be considerable with small n or with many explanatory variables. In comparing predictive power of different models, it is more helpful to use *adjusted* R^2 instead of R^2. This is

$$R^2_{\text{adj}} = \frac{s_y^2 - s^2}{s_y^2} = 1 - \frac{s^2}{s_y^2},$$

where $s^2 = \sum(y - \hat{y})^2/[n - (p+1)]$ is the estimated conditional variance (i.e., the residual mean square) and $s_y^2 = \sum(y - \bar{y})^2/(n-1)$ is the sample variance of y. This is a less biased estimator of the population R^2. Unlike ordinary R^2, if we add a term to a model that is not especially useful, then R^2_{adj} may even *decrease*. This happens when the new model has poorer predictive power, in the sense of a larger value of s^2. A possible criterion for selecting a model is to choose the one having the greatest value of R^2_{adj}. This is, equivalently, the model with smallest residual MS.

Most other criteria for selecting a model attempt to find the model for which the predicted values tend to be closest to the true expected values. One type of method for doing this uses ***cross-validation***. For a given model, you fit the model using some of the data and then analyze how well its prediction equation predicts the rest of the data. In one version, you use all observations except one to fit the model, and then check how well it predicts the remaining observation. Suppose we fit a model using all the data except observation 1. Using the prediction equation we get, let $\hat{y}_{(1)}$ denote the predicted selling price for observation 1. That is, we find a prediction equation using the data for observations $2, 3, \ldots, n$, and then we substitute the values of the explanatory variables for observation 1 into that prediction equation to get $\hat{y}_{(1)}$. Likewise, let $\hat{y}_{(2)}$ denote the prediction for observation 2 when we fit the model to observations $1, 3, 4, \ldots, n$, leaving out observation 2. In general, for observation i, we leave it out in fitting the model and then use the resulting prediction equation to get $\hat{y}_{(i)}$. Then, $(y_i - \hat{y}_{(i)})$ is a type of residual, measuring how far observation i falls from the value predicted for it using the prediction equation generated by the other $(n - 1)$ observations.

In summary, for a model for n observations, this version of cross-validation fits the model n times, each time leaving out one observation and using the prediction equation to predict that observation. We then get n predicted values and corresponding prediction residuals. The ***predicted residual sum of squares***, denoted by PRESS, is

$$\text{PRESS} = \sum (y_i - \hat{y}_{(i)})^2.$$

The smaller the value of PRESS, the better the predictions tend to be, in a summary sense. According to this criterion, the best-fitting model is the one with the smallest value of PRESS.

The ***AIC***, short for ***Akaike information criterion***, attempts to find a model for which the $\{\hat{y}_i\}$ tend to be closest to $\{E(y_i)\}$ in an average sense. The AIC is also scaled in such a way that the lower the value, the better the model. The best model is the one with the smallest AIC. We do not show its formula here, but it is sufficient to know that for ordinary regression models, minimizing the AIC corresponds to minimizing

$$n \log(\text{SSE}) + 2p,$$

where p is the number of model parameters. So, this criterion penalizes a model for having more parameters than are useful for getting good predictions. An advantage of the AIC is that its general definition also makes it applicable for models that assume nonnormal distributions for y, in which case a sum of squared errors is often not a useful summary.

Example 14.2

Using Indices to Select a Model for House Selling Price Table 14.4 shows the model selection indices for five models for the house selling price data. The table shows the models in the order built by forward selection (reverse order for backward elimination).

TABLE 14.4: Model Selection Criteria for Models for House Selling Price

Variables in Model	R^2	R^2_{adj}	PRESS	AIC
TAXES	0.709	0.706	3.17	2470.5
TAXES, SIZE	0.772	0.767	2.73	2448.0
TAXES, SIZE, NEW	0.790	0.783	2.67	2442.0
TAXES, SIZE, NEW, BEDS	0.793	0.785	2.85	2442.2
TAXES, SIZE, NEW, BEDS, BATHS	0.793	0.782	2.91	2444.2

Note: Actual PRESS equals value reported times 10^{11}.

According to the criterion of minimizing adjusted R^2, the selected model has all the explanatory variables except BATHS. It has $R^2_{\text{adj}} = 0.785$. To illustrate that R^2_{adj} can decrease when variables are added, note that this model as well as the model with SIZE, NEW, and TAXES predictors have R^2_{adj} values that are higher than $R^2_{\text{adj}} = 0.782$ for the full model with all the explanatory variables.

According to the criterion of minimizing the predicted residual sum of squares, the selected model has explanatory variables TAXES, SIZE, and NEW. It has the minimum PRESS = 2.67. (The y values were in dollars, so squared residuals tended to be huge numbers, and the actual PRESS values are the numbers reported multiplied by 10^{11}.) This was also the model selected by backward elimination and by forward selection.

According to the criterion of minimizing AIC, the selected model is also the one with explanatory variables TAXES, SIZE, and NEW. It has the minimum AIC = 2442.0. The model also containing NEW fits essentially as well. ■

14.2 Regression Diagnostics

Once we have selected the explanatory variables for a model, how do we know that model fits the data adequately? This section introduces diagnostics that indicate (1) when model assumptions are grossly violated and (2) when certain observations are highly influential in affecting the model fit or inference about model parameters.

Recall that inference about parameters in a regression model has these assumptions:

- The true regression function has the form used in the model (e.g., linear).
- The conditional distribution of y is normal.
- The conditional distribution of y has constant standard deviation throughout the range of values of the explanatory variables. This condition is called ***homoscedasticity***.
- The sample is randomly selected.

In practice, the assumptions are never perfectly fulfilled, but the regression model can still be useful. It is adequate to check that no assumption is grossly violated.

EXAMINE THE RESIDUALS

Several checks of assumptions use the residuals, $y - \hat{y}$. One check concerns the normality assumption. If the observations are normally distributed about the true regression equation with constant conditional standard deviation σ, then the residuals should be approximately normally distributed. To check this, plot the residuals about their mean value 0, using a histogram. They should have approximately a bell shape about 0.

A standardized version of the residual equals the residual divided by its standard error, which describes how much residuals vary because of ordinary sampling variability. In regression, this is called[1] a ***studentized residual***. Under the normality assumption, a histogram of these residuals should have the appearance of a standard normal distribution (bell shaped with mean of 0 and standard deviation of 1).

[1] Some software reports also a *standardized residual*, which divides $y - \hat{y}$ by s, which is slightly larger than the standard error of the residual.

If a studentized residual is larger than about 3 in absolute value, the observation is a potential outlier and should be checked. If an outlier represents a measurement error, it could cause a major bias in the prediction equation. Even if it is not an error, it should be investigated. It represents an observation that is not typical of the sample data, and it may have too much impact on the model fit. Consider whether there is some reason for the peculiarity. Sometimes the outliers differ from the other observations on some variable not included in the model, and once that variable is added, they cease to be outliers.

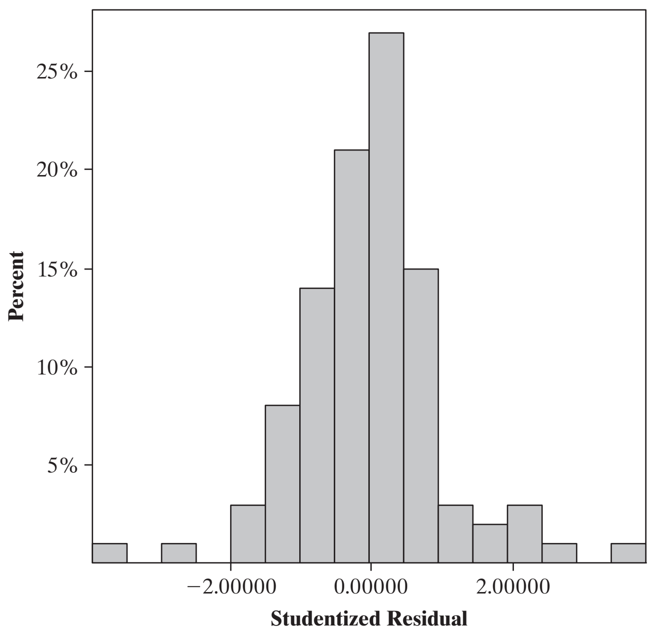
Example 14.3

Residuals for Modeling Home Selling Price For the `Houses` data file, with $y =$ selling price, variable selection procedures in Example 14.1 (page 421) and the AIC and PRESS indices in Example 14.2 suggested the model with prediction equation

$$\hat{y} = -21{,}353.8 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

Figure 14.2 is a histogram of the studentized residuals for this fit. No severe nonnormality seems to be indicated, since they are roughly bell shaped about 0. However, the plot indicates that two observations have relatively large residuals. On further inspection, we find that observation 6 had a selling price of \$499,900, which was \$168,747 higher than the predicted selling price for a new home of 3153 square feet with a tax bill of \$2997. The residual of \$168,747 has a studentized value of 3.88. Observation 64 had a selling price of \$225,000, which was \$165,501 lower than the predicted selling price for a non-new home of 4050 square feet with a tax bill of \$4350. Its residual of $-$\$165,501 has a studentized value of -3.93.

FIGURE 14.2: Histogram of Studentized Residuals for Multiple Regression Model Fitted to House Selling Prices, with Explanatory Variables Size, Taxes, and New



A severe outlier on y can substantially affect the fit, especially when the values of the explanatory variables are not near their means. So, we refitted the model without these two observations. The R^2-value changes from 0.79 to 0.83, and the prediction equation changes to

$$\hat{y} = -32{,}226 + 68.9(\text{SIZE}) + 20{,}436(\text{NEW}) + 38.3(\text{TAXES}).$$

The parameter estimates are similar for SIZE and TAXES, but the estimated effect of NEW drops from 46,374 to 20,436. Moreover, the effect of NEW is no longer significant, having a P-value of 0.17. Because the estimated effect of NEW is affected substantially by these two observations, we should be cautious in making conclusions about its effect. Of the 100 homes in the sample, only 11 were new. It is difficult to make precise estimates about the NEW effect with so few new homes, and results are highly affected by a couple of unusual observations. ■

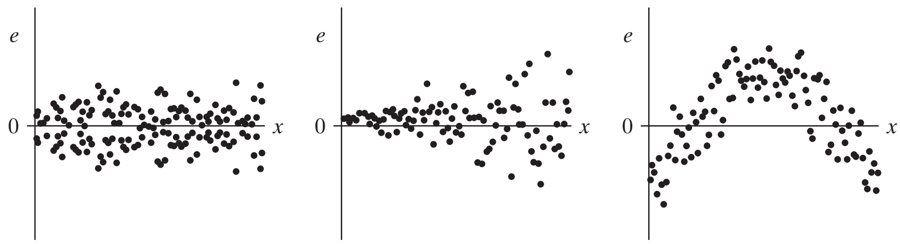
PLOTTING RESIDUALS AGAINST EXPLANATORY VARIABLES

The normality assumption is not as important as the assumption that the model provides a good approximation for the true relationship between the explanatory variables and the mean of y. If the model assumes a linear effect but the effect is actually strongly nonlinear, some conclusions may be faulty.

For bivariate models, the scatterplot provides a simple check on the form of the relationship. For multiple regression, it is also useful to construct a scatterplot of each explanatory variable against the response variable. This displays only the *bivariate* relationships, however, whereas the model refers to the *partial* effect of each explanatory variable, with the others held constant. The *partial regression plot* introduced on page 314 provides some information about this.

For multiple regression models, plots of the residuals (or studentized residuals) against the predicted values $\hat{y}$ or against each explanatory variable also help us check for potential problems. If the residuals appear to fluctuate randomly about 0 with no obvious trend or change in variation as the values of a particular x_i increase, then no violation of assumptions is indicated. The pattern should be roughly like Figure 14.3a.

FIGURE 14.3: Possible Patterns for Residuals (e), Plotted against an Explanatory Variable x



(a) Assumptions satisfied (b) Nonconstant standard deviation (c) Nonlinear term needed

In practice, most response variables can take only nonnegative values. For such responses, a fairly common occurrence is that the variability increases as the mean increases. For example, suppose we model y = annual income (in dollars) using several explanatory variables. For those subjects having $E(Y) = \$10,000$, the standard deviation of income is probably much less than for those subjects having $E(Y) = \$200,000$. Plausible standard deviations might be \$4000 and \$80,000. When this happens, the conditional standard deviation of y is not constant, whereas ordinary regression assumes that it is. An indication that this is happening is when the residuals are more spread out as the y_i-values increase. If we plot the residuals against a predictor that has a positive partial association with y, such as number of years of education, the residuals are then more spread out for larger values of the predictor, as in Figure 14.3b.

Figure 14.3c shows another possible abnormality, in which y tends to be below $\hat{y}$ for very small and very large x_i-values (giving negative residuals) and above $\hat{y}$ for medium-sized x_i-values (giving positive residuals). Such a scattering of residuals suggests that y is actually nonlinearly related to x_i. Sections 14.5 and 14.6 show how to address nonlinearity.

For the model relating selling price of home to size, taxes, and whether new for all 100 observations, Figure 14.4 plots the residuals against size. There is some suggestion of more variability at the higher size values. It does seem sensible that selling prices would vary more for very large homes than for very small homes. A similar picture occurs when we plot the residuals against taxes.

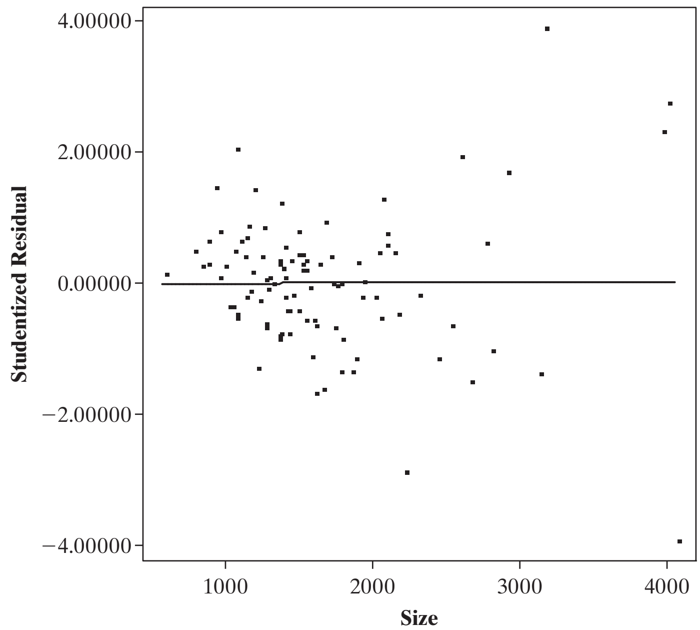


FIGURE 14.4: Scatterplot of Studentized Residuals of Home Selling Price Plotted against Size of Home, for Model with Explanatory Variables Size, Taxes, and Whether Home Is New

If the change in variability is severe, then a method other than ordinary least squares provides better estimates with more valid standard errors. Section 14.4 presents a generalized regression model that allows the variability to be greater when the mean is greater.

In practice, residual patterns are rarely as neat as the ones in Figure 14.3. Don't let a few outliers or ordinary sampling variability influence too strongly your interpretation of a plot. Also, the plots described here just scratch the surface of the graphical tools now available for diagnosing potential problems. Fox (2015, Section III) described a variety of modern graphical displays and diagnostic tools.

TIME SERIES DATA AND LONGITUDINAL STUDIES

Some social research studies collect observations sequentially over time. For economic variables such as a stock index or the unemployment rate, for example, the observations often occur daily or monthly. The observations are then recorded in sequence, rather than randomly sampled. Sampling subjects randomly from some population ensures that one observation is not statistically dependent on another, and this simplifies derivations of sampling distributions and their standard errors. However, neighboring observations from a time sequence are usually correlated rather than independent. For example, if the unemployment rate is relatively low in January 2018, it will probably also be relatively low in February 2018.

A plot of the residuals against the time of making the observation checks for this type of dependence. Ideally, the residuals should fluctuate in a random pattern about

0 over time, rather than showing a trend or periodic cycle. The methods presented in this text are based on independent observations and are inappropriate when time effects occur. For example, when observations next to each other tend to be positively correlated, the standard error of the sample mean is larger than the $\sigma/\sqrt{n}$ formula that applies for independent observations.

The term *time series* refers to relatively long sequences of observations over time. Books specializing in econometrics, such as Kennedy (2008), present methods for time series data. The term *longitudinal data* refers to studies, common in the social sciences and public health, that observe subjects over a relatively small number of times. For analyzing such data, see the linear mixed model in Section 13.5 and books by Fitzmaurice et al. (2011) and Hedeker and Gibbons (2006).

DETECTING INFLUENTIAL OBSERVATIONS: RESIDUAL AND LEVERAGE

Least squares estimates of parameters in regression models can be strongly influenced by an outlier, especially when n is small. A variety of statistics summarize the influence each observation has. These statistics refer to how much the predicted values $\hat{y}$ or the model parameter estimates change when we remove an observation from the data set. An observation's influence depends on two factors: (1) how far its y-value falls from the overall trend in the data and (2) how far the values of the explanatory variables fall from their means.

The first factor on influence (how far y falls from the overall trend) is measured by the observation's residual, $y - \hat{y}$. The larger the residual, the farther the observation falls from the overall trend. We can search for observations with large studentized residuals (say, larger than about 3 in absolute value) to find observations that may be influential.

The second factor on influence (how far the explanatory variables fall from their means) is summarized by the ***leverage*** of the observation. The leverage is a nonnegative statistic such that the larger its value, the greater weight that observation receives in determining the $\hat{y}$-values (hence, it also is sometimes called a *hat value*). The formula for the leverage in multiple regression is complex. For the bivariate model, the leverage for observation i simplifies to

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})^2}.$$

So, the leverage gets larger as the x-value x_i for observation i gets farther from the mean. It gets smaller as the sample size increases. When calculated for each observation in a sample, the average leverage equals the number p of parameters in the model divided by n.

DETECTING INFLUENTIAL OBSERVATIONS: DFFIT AND DFBETA

For an observation to be influential, it must have both a relatively large residual and a relatively large leverage. Statistical software reports diagnostics that depend on the residuals and the leverages. Two popular ones are called ***DFFIT*** and ***DFBETA***.

For an observation, DFBETA summarizes the *effect on the model parameter estimates* of removing the observation from the data set. For the effect β_j of x_j, DFBETA equals the change in the estimate $\hat{\beta}_j$ due to deleting the observation. The larger the absolute value of DFBETA, the greater the influence of the observation on that parameter estimate. Each observation has a DFBETA value for each parameter in the model.

DFFIT summarizes the *effect on the fit* of deleting the observation. For observation i, DFFIT equals the change in the predicted value due to deleting that observation (i.e., $\hat{y}_i - \hat{y}_{(i)}$). The DFFIT value has the same sign as the residual. ***Cook's distance*** is an alternative measure with the same purpose. Cook's distance and DFFIT are based on the effect that observation i has on *all* the parameter estimates. They summarize more broadly the influence of an observation, as each observation has a single DFFIT value and a single Cook's distance, whereas it has a separate DFBETA for each parameter. The larger their absolute values, the greater the influence that observation has on the fitted values.

Some software reports *standardized* versions of the DFBETA and DFFIT measures, often denoted by DFBETAS and DFFITS. The standardized DFBETA divides the change in the estimate $\hat{\beta}_j$ due to deleting the observation by the standard error of $\hat{\beta}_j$ for the adjusted data set. For observation i, the standardized DFFIT equals the change in the predicted value due to deleting that observation, divided by the standard error of $\hat{y}$ for the adjusted data set.

In practice, scan or plot these diagnostic measures to see if some observations stand out from the rest, having relatively large values. Each measure has approximate cutoff points for noteworthy observations. For example, a Cook's distance larger than about $4/n$ indicates a potentially large influence. A standardized DFBETA larger than 1 suggests a substantial influence on that parameter estimate. However, Cook's distance, DFBETA, and DFFIT tend to decrease as n increases, so normally it is a good idea to examine observations having extreme values relative to the others. Individual data points have less influence for larger sample sizes.

Example 14.4

DFBETA and DFFIT for an Influential Observation Example 14.3 (page 427) showed that observations 6 and 64 were influential on the equation for predicting home selling price using size of home, taxes, and whether the house is new. The prediction equation for all 100 observations is

$$\hat{y} = -21{,}354 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

For observation 6, the DFBETA values are 12.5 for size, 16,318.5 for new, and -5.7 for taxes. This means, for example, that if this observation is deleted from the data set, the effect of NEW changes from 46,373.7 to $46{,}373.7 - 16{,}318.5 = 30{,}055.2$. Observation 6 had a predicted selling price of $\hat{y} = 331{,}152.8$. Its DFFIT value is 29,417.0. This means that if observation 6 is deleted from the data set, then $\hat{y}$ at the explanatory variable values for observation 6 changes to $331{,}152.8 - 29{,}417.0 = 301{,}735.8$. This analysis shows that this observation is quite influential. ■

Example 14.5

Influence Diagnostics for Crime Data Table 9.1 (page 248) listed y = murder rate for the 50 states and the District of Columbia (D.C.), with explanatory variables x_1 = percentage of families below the poverty level and x_2 = percentage of single-parent families. The data are in the `Crime2` data file at the text website. The least squares fit of the multiple regression model is

$$\hat{y} = -40.7 + 0.32x_1 + 3.96x_2.$$

Table 14.5 shows the influence diagnostics for the model fit, including the standardized versions of DFBETA and DFFIT. The studentized residuals fall in a reasonable range except the one for the last observation (D.C.), which equals 14.2. The observed murder rate of 78.5 for D.C. falls far above the predicted value of 55.3, causing a large positive residual. This is an extreme outlier. In addition, the leverage for D.C. is 0.54, more than three times as large as any other leverage and nine times the average

TABLE 14.5: Influence Diagnostics for Model Using Poverty Rate and Single-Parent Percentage to Predict Murder Rate for 50 U.S. States and District of Columbia

Obs	Dep Var MURDER	Predict Value	Residual	Student Resid	Leverage h	Dffits	POVERTY Dfbeta	SINGLE Dfbeta
AK	9.0	18.88	−9.88	−2.04	0.162	−0.895	0.714	−0.761
AL	11.6	10.41	1.18	0.22	0.031	0.039	0.024	−0.011
AR	10.2	8.07	2.13	0.40	0.079	0.117	0.100	−0.069
AZ	8.6	12.16	−3.55	−0.65	0.022	−0.099	−0.005	−0.025
CA	13.1	14.63	−1.53	−0.28	0.034	−0.053	−0.027	−0.004
CO	5.8	10.41	−4.61	−0.87	0.060	−0.220	0.174	−0.134
CT	6.3	2.04	4.25	0.79	0.051	0.185	−0.130	0.015
DE	5.0	7.73	−2.73	−0.50	0.043	−0.107	0.079	−0.045
FL	8.9	6.97	1.92	0.35	0.048	0.080	0.059	−0.047
GA	11.4	15.12	−3.72	−0.69	0.042	−0.145	0.071	−0.105
HI	3.8	−2.07	5.87	1.11	0.059	0.279	−0.153	−0.058
IA	2.3	−1.74	4.04	0.75	0.045	0.164	−0.034	−0.081
ID	2.9	1.12	1.77	0.32	0.035	0.063	0.012	−0.040
IL	11.4	9.21	2.18	0.40	0.020	0.058	−0.013	0.011
IN	7.5	5.99	1.50	0.27	0.023	0.043	−0.014	−0.000
KS	6.4	2.71	3.68	0.68	0.029	0.117	0.013	−0.062
KY	6.6	7.79	−1.19	−0.22	0.088	−0.070	−0.061	0.043
LA	20.3	26.74	−6.44	−1.29	0.161	−0.568	−0.412	−0.055
MA	3.9	5.91	−2.01	−0.37	0.033	−0.068	0.042	−0.014
MD	12.7	9.95	2.74	0.51	0.060	0.130	−0.104	0.077
ME	1.6	4.72	−3.12	−0.57	0.031	−0.104	0.058	−0.008
MI	9.8	15.72	−5.92	−1.10	0.033	−0.204	0.035	−0.124
MN	3.4	2.23	1.16	0.21	0.029	0.037	−0.007	−0.013
MO	11.3	7.62	3.67	0.67	0.027	0.115	0.059	−0.049
MS	13.5	25.40	−11.90	−2.45	0.126	−0.933	−0.623	−0.151
MT	3.0	6.84	−3.84	−0.70	0.023	−0.108	−0.033	0.039
NC	11.3	7.87	3.42	0.62	0.020	0.090	0.009	−0.013
ND	1.7	−3.83	5.53	1.04	0.057	0.259	0.016	−0.184
NE	3.9	−0.15	4.05	0.75	0.039	0.153	−0.047	−0.056
NH	2.0	−1.07	3.07	0.57	0.044	0.123	−0.039	−0.047
NJ	5.3	0.82	4.47	0.83	0.035	0.158	−0.041	−0.058
NM	8.0	19.53	−11.53	−2.25	0.046	−0.499	−0.017	−0.308
NV	10.4	11.57	−1.17	−0.22	0.069	−0.060	0.048	−0.040
NY	13.3	14.85	−1.55	−0.28	0.028	−0.048	−0.005	-0.019
OH	6.0	8.62	−2.62	−0.48	0.022	−0.072	0.024	−0.015
OK	8.4	9.62	−1.22	−0.22	0.067	−0.061	−0.051	0.031
OR	4.6	7.84	−3.24	−0.59	0.027	−0.101	0.054	−0.029
PA	6.8	1.55	5.24	0.97	0.034	0.183	0.036	−0.115
RI	3.9	5.67	−1.77	−0.32	0.028	−0.056	0.029	−0.006
SC	10.3	13.99	−3.69	−0.68	0.038	−0.137	−0.084	0.008
SD	3.4	1.07	2.32	0.43	0.042	0.091	0.036	−0.067
TN	10.2	9.92	0.27	0.05	0.060	0.013	0.010	−0.006
TX	11.9	11.60	0.29	0.05	0.029	0.009	0.005	−0.001
UT	3.1	2.34	0.75	0.13	0.032	0.025	−0.010	−0.004
VA	8.3	3.21	5.08	0.94	0.039	0.192	−0.119	0.010
VT	3.6	6.08	−2.48	−0.46	0.040	−0.094	0.067	−0.028
WA	5.2	9.52	−4.32	−0.80	0.029	−0.139	0.078	−0.059
WI	4.4	4.53	−0.13	−0.02	0.023	−0.003	0.000	0.001
WV	6.9	3.60	3.29	0.66	0.178	0.307	0.274	−0.229
WY	3.4	6.34	−2.94	−0.54	0.021	−0.079	0.006	0.012
DC	78.5	55.28	23.22	14.20	0.536	15.271	−0.485	12.792

leverage of $p/n = 3/51 = 0.06$. Since D.C. has both a large studentized residual and a large leverage, it has considerable influence on the model fit.

Not surprisingly, DFFIT for D.C. is much larger than for the other observations. This suggests that the predicted values change considerably if we refit the model after removing this observation. The DFBETA value for the single-family explanatory variable x_2 is much larger for D.C. than for the other observations. This suggests that the effect of x_2 could change substantially with the removal of D.C. By contrast, DFBETA for poverty is not so large.

These diagnostics suggest that the D.C. observation has a large influence, particularly on the coefficient of x_2 and on the fitted values. The prediction equation for the model fitted without the D.C. observation is

$$\hat{y} = -14.6 + 0.36x_1 + 1.52x_2.$$

Not surprisingly, the estimated effect of x_1 did not change much, but the coefficient of x_2 is now less than half as large. The standard error of the coefficient of x_2 also changes dramatically, decreasing from 0.44 to 0.26. ■

An observation with a large studentized residual does not have a major influence if its values on the explanatory variables do not fall far from their means. Recall that the leverage summarizes how far the explanatory variables fall from their means. For instance, New Mexico has a relatively large negative studentized residual (−2.25) but a relatively small leverage (0.046), so it does not have large values of DFFIT or DFBETA. Similarly, an observation far from the mean on the explanatory variables (i.e., with a large leverage) need not have a major influence if it falls close to the prediction equation and has a small studentized residual. For instance, West Virginia has a relatively large poverty rate and its leverage of 0.178 is triple the average. However, its studentized residual is small (0.66), so it has little influence on the fit.

14.3 Effects of Multicollinearity

In many social science studies using multiple regression, the explanatory variables "overlap" considerably. A variable may be nearly redundant, in the sense that it can be predicted well using the others. If we regress an explanatory variable on the others and get a large R^2-value, this suggests that it may not be needed in the model once the others are there. This condition is called *multicollinearity*, or sometimes simply *collinearity*. This section describes the effects of multicollinearity and ways to diagnose it.

VIF: MULTICOLLINEARITY CAUSES VARIANCE INFLATION

Multicollinearity causes inflated standard errors for estimates of regression parameters. The standard error of the estimator of the coefficient β_j of x_j in the multiple regression model can be expressed as

$$se = \frac{1}{\sqrt{1 - R_j^2}}\left[\frac{s}{\sqrt{n-1}s_{x_j}}\right],$$

where s is the square root of the residual mean square and s_{x_j} denotes the sample standard deviation of x_j values. Let R_j^2 denote R^2 from the regression of x_j on the other explanatory variables from the model. So, when x_j overlaps a lot with the other explanatory variables, in the sense that R_j^2 is large for predicting x_j using the other explanatory variables, this *se* is relatively large. Then, the confidence interval for β_j

is wide, and the test of H_0: $\beta_j = 0$ has a large P-value unless the sample size is very large or the effect is very strong.

In this *se* formula for the estimate of β_j, the quantity

$$\text{VIF} = 1/(1 - R_j^2)$$

is called a ***variance inflation factor*** (VIF). It represents the multiplicative increase in the variance (squared standard error) of the estimator due to x_j being correlated with the other explanatory variables. When any R_j^2-value from regressing each explanatory variable on the other explanatory variables in the model is close to 1, say above 0.90, severe multicollinearity exists.

For example, if $R_j^2 > 0.90$, then VIF > 10 for the effect of that explanatory variable. That is, the variance of the estimate of β_j inflates by a factor of more than 10. The standard error inflates by a factor of more than $\sqrt{10} = 3.2$, compared to the standard error for uncorrelated explanatory variables. When an explanatory variable is in the model primarily as a control variable, and we do not need precise estimates of its effect on the response variable, it is not crucial to worry about its VIF value.

For the model selected in Section 14.1 that predicts house selling price using taxes, size, and whether the house is new, software reports the VIF values

```
          VIF
TAXES    3.082
SIZE     3.092
NEW      1.192
```

The standard error for whether the house is new is not affected much by correlation with the other explanatory variables, but the other two standard errors multiply by a factor of roughly $\sqrt{3.1} = 1.76$.

OTHER INDICATORS OF MULTICOLLINEARITY

Even without checking VIFs, various types of behavior in a regression analysis can indicate potential problems due to multicollinearity. A warning sign occurs when the estimated coefficient for a predictor already in the model changes substantially when another variable is introduced. For example, perhaps the estimated coefficient of x_1 is 2.4 for the bivariate model, but when x_2 is added to the model, the coefficient of x_1 changes to 25.9.

Another indicator of multicollinearity is when a highly significant R^2 exists between y and the explanatory variables, but individually each partial regression coefficient is not significant. In other words, H_0: $\beta_1 = \cdots = \beta_k = 0$ has a small P-value in the overall F test, but H_0: $\beta_1 = 0$, H_0: $\beta_2 = 0$, and so forth do not have small P-values in the separate t tests. Thus, it is difficult to assess individual partial effects when severe multicollinearity exists. Other indicators of multicollinearity are surprisingly large standard errors, or standardized regression coefficients that are larger than 1 in absolute value.

When multicollinearity exists, it is rather artificial to interpret a regression coefficient as the effect of an explanatory variable when other variables are held constant. For instance, when $|r_{x_1x_2}|$ is high, then as x_1 changes, x_2 also tends to change in a linear manner, and it is artificial to envision x_1 or x_2 as being held constant.

REMEDIAL ACTIONS WHEN MULTICOLLINEARITY EXISTS

Remedial measures can help to reduce the effects of multicollinearity. One solution is to choose a subset of the explanatory variables, removing those variables that explain a small portion of the remaining unexplained variation in y. If x_4 and x_5 have a correlation of 0.96, it is only necessary to include one of them in the model.

When several explanatory variables are highly correlated and are indicators of a common feature, you could construct a summary index by combining responses on those variables. For example, suppose that a model for predicting y = opinion about president's performance in office uses 12 explanatory variables, of which three refer to the subject's opinion about whether a woman should be able to obtain an abortion (1) when she cannot financially afford another child, (2) when she is unmarried, and (3) anytime in the first three months. Each of these items is scaled from 1 to 5, with a 5 being the most conservative response. They are likely to be highly positively correlated, contributing to multicollinearity. A possible summary measure for opinion about abortion averages (or sums) the responses to these items. Higher values on that summary index represent more conservative responses. If the items were measured on different scales, we could first standardize the scores before averaging them. Socioeconomic status is a variable of this type, summarizing the joint effects of education, income, and occupational prestige.

Often multicollinearity occurs when the explanatory variables include interaction terms. Since cross-product terms are composed of other explanatory variables in the model, it is not surprising that they tend to be highly correlated with the other terms. The effects of this are diminished if we center the explanatory variables by subtracting their sample means before entering them in the interaction model (see page 328).

Other procedures, beyond the scope of this chapter, can handle multicollinearity. For example, *factor analysis* (introduced in Chapter 16) is a method for creating artificial variables from the original ones in such a way that the new variables can be uncorrelated. In most applications, though, it is more advisable to use a subset of the variables or create some new variables directly, as just explained.

Multicollinearity does not adversely affect all aspects of regression. Although multicollinearity makes it difficult to assess *partial* effects of explanatory variables, it does not hinder the assessment of their *joint* effects. If newly added explanatory variables overlap substantially with ones already in the model, then R and R^2 will not increase much, but the fit will not be poorer. So, the presence of multicollinearity does not diminish the predictive power of the equation. For further discussion of the effects of multicollinearity and methods for dealing with it, see DeMaris (2004), Fox (2015, Chapter 13), and Kutner et al. (2008).

14.4 Generalized Linear Models

The models presented in this book are special cases of ***generalized linear models***. This broad class of models includes ordinary regression models for response variables assumed to have a normal distribution, alternative models for continuous variables that do not assume normality, and models for discrete response variables including categorical variables. This section introduces generalized linear models. We use the acronym *GLM*.

NONNORMAL DISTRIBUTIONS FOR A RESPONSE VARIABLE

As in other regression models, a GLM identifies a response variable y and a set of explanatory variables. The regression models presented in Chapters 9–14 are GLMs that assume that y has a normal distribution.

In many applications, the potential outcomes for y are binary rather than continuous. Each observation might be labeled as a *success* or *failure*, as in the methods for proportions presented in Sections 5.2, 6.3, and 7.2. For instance, in a study of factors that influence votes in U.S. presidential elections, the response variable indicates the preferred candidate in the previous presidential election—the Democratic or the Republican candidate. In this case, models usually assume a *binomial* distribution for y.

In some applications, each observation is a count. For example, in a study of factors associated with family size, the response variable is the number of children in a family. GLMs for count data most often use two distributions for y not presented in this text, called the *Poisson* and the *negative binomial*.

Binary outcomes and counts are examples of discrete variables. Regression models that assume normal distributions are not optimal for models with discrete responses. Even when the response variable is continuous, the normal distribution may not be optimal. When each observation must take a positive value, for instance, the distribution is often skewed to the right with greater variability when the mean is greater. In that case, a GLM can assume a *gamma* distribution for y, as discussed later in this section.

THE LINK FUNCTION FOR A GLM

In a GLM, as in ordinary regression models, $\mu = E(y)$ varies according to values of explanatory variables, which enter linearly as predictors on the right-hand side of the model equation. However, a GLM allows a function $g(\mu)$ of the mean rather than just the mean μ itself on the left-hand side. The GLM formula states that

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The function $g(\mu)$ is called the ***link function***, because it links the mean of the response variable to the explanatory variables.

For instance, the link function $g(\mu) = \log(\mu)$ models the log of the mean. The log function applies to positive numbers, so this ***log link*** is appropriate when μ cannot be negative, such as with count data. GLMs that use the log link,

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

are often called ***loglinear models***. The final section of this chapter shows an example.

For binary data, the most common link function is $g(\mu) = \log[\mu/(1-\mu)]$. This is called the ***logit link***. It is appropriate when μ falls between 0 and 1, such as a probability, in which case $\mu/(1-\mu)$ is the *odds*. When y is binary, this link is used in models for the probability of a particular outcome, for instance, to model the probability that a subject votes for the Republican candidate. A GLM using the logit link, called a ***logistic regression model***, is presented in the next chapter.

The simplest possible link function is $g(\mu) = \mu$. This models the mean directly and is called the ***identity link***. It specifies a linear model for the mean response,

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

When employed with a normal assumption for y, this is the ordinary regression model.

A GLM generalizes ordinary regression in two ways: First, y can have a distribution other than the normal. Second, it can model a function of the mean. Both generalizations are important, especially for discrete responses.

GLMS VERSUS ORDINARY REGRESSION FOR TRANSFORMED DATA

Before GLMs were developed in the 1970s, the traditional way of analyzing “nonnormal” data was to transform the y-values. The goal of this approach is to find a function

g such that $g(y)$ has an approximately normal distribution, with constant standard deviation at all levels of the explanatory variables. Square root or log transforms are often applied to do this. If the variability is then more nearly constant, least squares works well with the transformed data. In practice, this may not work well. Simple linear models for the explanatory variables may fit poorly on that scale. If the original relationship is linear, it is no longer linear after applying the transformation. If we fit a straight line and then transform back to the original scale, the fit is no longer linear. Also, technical problems can occur, such as taking logarithms of 0. Moreover, conclusions that refer to the mean response on the scale of the transformed variable are less relevant.

With the GLM approach, it is not necessary to transform data and use normal methods. This is because the GLM fitting process utilizes the ***maximum likelihood*** estimation method (page 126), for which the choice of distribution for y is not restricted to normality. Maximum likelihood employs a generalization of least squares called ***weighted least squares***. It gives more weight to observations over regions that show less variability. In addition, in GLMs the choice of link function is separate from the choice of distribution for y. If a certain link function makes sense for a particular type of data, it is not necessary that it also stabilize variation or produce normality.

The family of GLMs unifies a wide variety of statistical methods. Ordinary regression models as well as models for discrete data (Chapter 15) are special cases of one highly general model. In fact, the same fitting method yields parameter estimates for all GLMs. Using GLM software, there is tremendous flexibility and power in the model-building process. You pick a probability distribution that is most appropriate for y. For instance, you might select the normal option for a continuous response or the binomial option for a binary response. You specify the variables that are the explanatory variables. Finally, you pick the link function, determining which function of the mean to model. Software then fits the model and provides the maximum likelihood estimates of model parameters. For further details about GLMs, see Fox (2015), Gill (2000), and King (1989).

GLMS FOR A RESPONSE ASSUMING A GAMMA DISTRIBUTION

The residual analysis in Example 14.3 for modeling selling prices of homes showed a tendency for greater variability of selling prices at higher house size values. (See Figure 14.4 on page 429.) Small homes show little variability in selling price, whereas large homes show high variability. Large homes are the ones that tend to have higher selling prices, so variability in y increases as its mean increases.

This phenomenon often happens for positive-valued response variables. When the mean response is near 0, less variation occurs than when the mean response is high. For such data, least squares is not optimal. Least squares is identical to maximum likelihood for a GLM in which y is assumed to have a normal distribution with *identical* standard deviation σ at all values of explanatory variables.

An alternative approach for data of this form assumes a distribution for y for which the standard deviation increases as the mean increases (i.e., that permits *heteroscedasticity*). The family of ***gamma distributions*** has this property. When y has a gamma distribution with mean μ, then y has

$$\text{Variance} = \phi\mu^2, \quad \text{Standard deviation} = \sqrt{\phi}\mu,$$

where ϕ is called a *scale parameter*. The standard deviation increases proportionally to the mean: When the mean doubles, the standard deviation doubles. The gamma distribution falls on the positive part of the line. It exhibits skewness to the right, like the chi-squared distribution, which is a special case of the gamma.

The scale parameter, or an equivalent *shape parameter* that is the reciprocal of the scale parameter, determines the shape of the distribution. The gamma distribution becomes more bell shaped as ϕ decreases, being quite bell shaped when $\phi < 0.1$. It becomes more skewed as ϕ increases, being so highly skewed when $\phi \geq 1$ that the mode is 0.

With GLMs, you can fit a regression model assuming a gamma distribution for y instead of a normal distribution. Even if the data are close to normal, this alternative fit is more appropriate than the least squares fit when the standard deviation increases proportionally to the mean. Just as ordinary regression models assume that the variance is constant for all values of the explanatory variables, software for gamma GLMs assumes a constant scale parameter and estimates it as part of the model-fitting process.[2]

When the relationship is closer to linear on a log scale for $E(y)$, it is preferable to apply the log as a link function with a gamma GLM. The log link is also used when a linear model for the mean would give *negative* values at some explanatory variable values, because negative values are not permitted with a gamma distribution.

Example 14.6 **Gamma GLM for House Selling Price** The least squares fit of the model to the data on y = selling price using explanatory variables size of home, taxes, and whether new, discussed in Example 14.1 (page 421), is

$$\hat{y} = -21{,}353.8 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

However, Example 14.3 (page 427) showed that two outlying observations had a substantial effect on the estimated effect of NEW. Figure 14.4 showed that the variability in selling prices seems to increase as its mean does. This suggests that a model assuming a gamma distribution may be more appropriate, because the gamma permits the standard deviation to increase as the mean does.

We can use software, as explained in Appendix A, to fit GLMs that assume a gamma distribution for y. For these data, we obtain

$$\hat{y} = -940.9 + 48.7(\text{SIZE}) + 32{,}868.0(\text{NEW}) + 37.9(\text{TAXES}).$$

The estimated effect of TAXES is similar as with least squares, but the estimated effect of SIZE is weaker and the estimated effect of NEW is much weaker. Moreover, the effect of NEW is no longer significant, as the ratio of the estimate to the standard error is 1.5. This result is similar to what we obtained in Example 14.4 (page 431) after deleting observation 6, an outlier corresponding to a large new house with an unusually high selling price. The outliers are not as influential for the gamma fit, because that model expects more variability in the data when the mean is larger.

The estimate of the scale parameter is $\hat{\phi} = 0.07$. The estimated standard deviation $\hat{\sigma}$ of the conditional distribution of y relates to the estimated conditional mean $\hat{\mu}$ by

$$\hat{\sigma} = \sqrt{\hat{\phi}}\hat{\mu} = \sqrt{0.07}\hat{\mu} = 0.27\hat{\mu}.$$

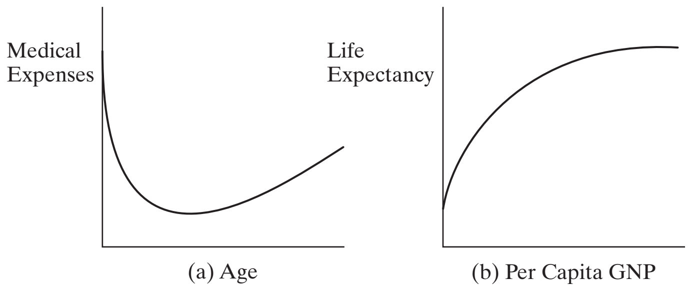
For example, at explanatory variable values such that the estimated mean selling price is $\hat{\mu}$ = \$100,000, the estimated standard deviation of selling prices is $\hat{\sigma}$ = 0.27(\$100,000) = \$27,000. By contrast, at explanatory variable values such that $\hat{\mu}$ = \$400,000, $\hat{\sigma}$ = 0.27(\$400,000) = \$108,000, four times as large. ■

[2] R, SPSS, and Stata estimate the scale parameter, whereas SAS estimates the shape parameter.

14.5 Nonlinear Relationships: Polynomial Regression

The ordinary multiple regression model assumes that the relationship between the mean of y and each quantitative explanatory variable is linear, controlling for other explanatory variables. Although social science relationships are not *exactly* linear, the degree of nonlinearity is often so minor that they can be reasonably well approximated with linearity. Occasionally, though, such a model is inadequate, even for approximation. A scatterplot may reveal a highly nonlinear relationship. Or, you might expect a nonlinear relationship because of the nature of the variables. For example, you might expect $y =$ medical expenses to have a curvilinear relationship with $x =$ age, being relatively high for the very young and the very old but lower for older children and young adults (Figure 14.5a). The relationship between $x =$ per capita income and $y =$ life expectancy for a sample of countries might be approximately a linearly increasing one, up to a certain point. However, beyond a certain level, additional income would probably result in little, if any, improvement in life expectancy (Figure 14.5b).

FIGURE 14.5: Two Nonlinear Relationships



If we use straight-line regression to describe a curvilinear relationship, what can go wrong? Measures of association designed for linearity, such as the correlation, may underestimate the true association. Estimates of the mean of y at various x-values may be badly biased, since the prediction line may poorly approximate the true regression curve. Two approaches are common to deal with nonlinearity. The first approach, presented in this section, uses a *polynomial* regression function. The class of polynomial functions includes a diverse set of functional patterns, including straight lines. The second approach, presented in Section 14.6, uses a generalized linear model with a link function such as the logarithm. For example, for certain curvilinear relationships, the logarithm of the mean of the response variable is linearly related to the explanatory variables.

QUADRATIC REGRESSION MODELS

A ***polynomial regression function*** for a response variable y and single explanatory variable x has the form

$$E(y) = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p.$$

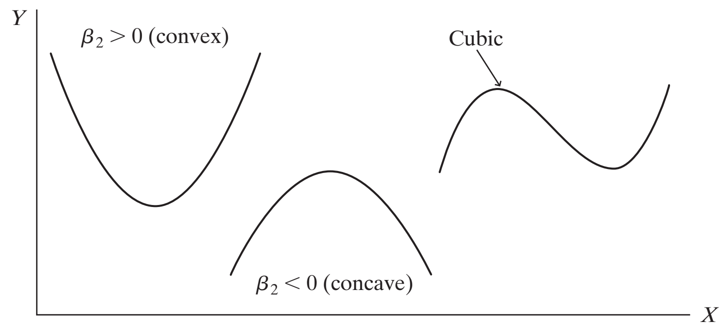
In this model, x occurs in powers from the first ($x = x^1$) to some integer p. For $p = 1$, this is the straight line $E(y) = \alpha + \beta_1 x$. The index p, the highest power in the polynomial equation, is called the ***degree*** of the polynomial function.

The polynomial function most commonly used for nonlinear relationships is the *second-degree polynomial*

$$E(y) = \alpha + \beta_1 x + \beta_2 x^2.$$

This is called a ***quadratic regression model***. The graph of this function is parabolic, as Figure 14.6 portrays. This shape is limited in scope for applications, because it is symmetric about a vertical axis. That is, its appearance when increasing is a mirror image of its appearance when decreasing.

FIGURE 14.6: Graphs of Two Second-Degree Polynomials (Quadratic Functions) and a Third-Degree Polynomial (Cubic Function)

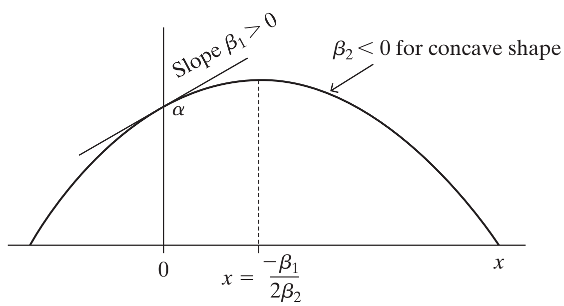


If a scatterplot reveals a pattern of points with one bend, then a second-degree polynomial usually improves upon the straight-line fit. A third-degree polynomial $E(y) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, called a ***cubic function***, is a curvilinear function having *two* bends. See Figure 14.6. But it is rarely necessary to use higher than a second-degree polynomial to describe the trend.

INTERPRETING AND FITTING QUADRATIC REGRESSION MODELS

The quadratic regression model $E(y) = \alpha + \beta_1 x + \beta_2 x^2$, plotted for the possible values of α, β_1, and β_2, describes the possible parabolic shapes. Unlike straight lines, for which the slope remains constant over all x-values, the mean change in y for a one-unit increase in x *depends on the value of* x. For example, as the value of x increases, a straight line drawn tangent to the parabola in Figure 14.7 first has positive slope, then zero slope where the parabola achieves its maximum value, and then negative slope. The rate of change of the line varies to produce a curve having a smooth bend.

FIGURE 14.7: Interpretation of Parameters of the Quadratic Model $E(y) = \alpha + \beta_1 x + \beta_2 x^2$



The sign of the coefficient β_2 of the x^2 term determines whether the function is bowl shaped (opens up) relative to the x-axis or mound shaped (opens down). Bowl-shaped functions, also called *convex* functions, have $\beta_2 > 0$. Mound-shaped functions, also called *concave* functions, have $\beta_2 < 0$. See Figure 14.6.

As usual, the coefficient α is the y-intercept. The coefficient β_1 of x is the slope of the line that is tangent to the parabola as it crosses the y-axis. If $\beta_1 > 0$, for example, then the parabola is sloping upward at $x = 0$ (as Figure 14.7 shows). At the point at which the slope is zero, the relationship changes direction from positive to negative or from negative to positive. This happens at $x = -\beta_1/2\beta_2$. This is the x-value at which the mean of y takes its maximum if the parabola is mound shaped and its minimum if it is bowl shaped.

To fit quadratic regression models, we treat them as a special case of the multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 = \alpha + \beta_1 x + \beta_2 x^2$$

with two explanatory variables. We identify x_1 with the explanatory variable x and x_2 with its square, x^2. The data for the model fit consist of the y-values for the subjects in the sample, the x-values (called x_1), and an artificial variable (x_2) consisting of the squares of the x-values. Software can create these squared values for us. It then uses least squares to find the best-fitting function out of the class of all second-degree polynomials.

Example 14.7

Fertility Predicted Using Gross Domestic Product (GDP) Table 14.6 shows values reported by the United Nations for several nations on y = fertility rate (the mean number of children per adult woman) and x = per capita gross domestic product (GDP, in tens of thousands of dollars). Fertility tends to decrease as GDP increases. However, a straight-line model may be inadequate, since it might predict negative fertility for sufficiently high GDP. In addition, some demographers predict that after GDP passes a certain level, fertility rate may increase, since the nation's wealth makes it easier for a parent to stay home and take care of children rather than work.

TABLE 14.6: Data on Fertility Rate and Per Capita Gross Domestic Product GDP (`Fertility`GDP Data File at the Text Website)

Nation	GDP	Fertility Rate	Nation	GDP	Fertility Rate	Nation	GDP	Fertility Rate
Algeria	0.21	2.5	Germany	2.91	1.3	Pakistan	0.06	4.3
Argentina	0.35	2.4	Greece	1.56	1.3	Philippines	0.10	3.2
Australia	2.63	1.7	India	0.06	3.1	Russia	0.30	1.3
Austria	3.13	1.4	Iran	0.21	2.1	S. Africa	0.35	2.8
Belgium	2.91	1.7	Ireland	3.85	1.9	Saudi Ar.	0.95	4.1
Brazil	0.28	2.3	Israel	1.65	2.9	Spain	2.04	1.3
Canada	2.71	1.5	Japan	3.37	1.3	Sweden	3.37	1.6
Chile	0.46	2.0	Malaysia	0.42	2.9	Switzerland	4.36	1.4
China	0.11	1.7	Mexico	0.61	2.4	Turkey	0.34	2.5
Denmark	3.93	1.8	Netherlands	3.15	1.7	UK	3.03	1.7
Egypt	0.12	3.3	New Zealand	1.98	2.0	United States	3.76	2.0
Finland	3.11	1.7	Nigeria	0.04	5.8	Vietnam	0.05	2.3
France	2.94	1.9	Norway	4.84	1.8	Yemen	0.06	6.2

Source: `hdr.undp.org/en/data`.

Figure 14.8, a scatterplot for the 39 observations, shows a clear decreasing trend. The linear prediction equation is $\hat{y} = 3.04 - 0.415x$, and the correlation equals -0.56. This prediction equation gives absurd predictions for very large x-values; $\hat{y}$ is negative for $x > 7.3$ (i.e., \$73,000). However, the predicted values are positive over the range of x-values for this sample. To allow for potential nonlinearity and for the possibility that fertility rate may increase for sufficiently large GDP, we could fit a quadratic regression model to these data. ■

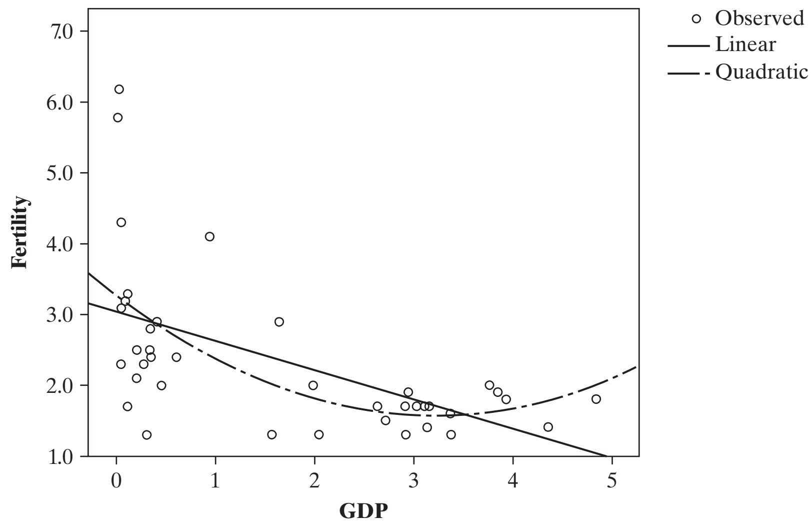


FIGURE 14.8: Scatterplot and Best-Fitting Straight-Line Model and Quadratic Model for Data on Fertility Rate and Per Capita GDP from the `FertilityGDP` Data File

Table 14.7 shows some output for the quadratic regression of y = fertility rate on x = GDP. Here, GDP2 denotes an artificial variable constructed as the square of GDP. The prediction equation is

$$\hat{y} = 3.28 - 1.054x + 0.163x^2.$$

Figure 14.8 plots the linear and quadratic prediction equations in the scatter diagram.

TABLE 14.7: Some Output for Quadratic Regression Model for y = Fertility Rate and x = GDP from the `FertilityGDP` Data File

Variable	Coef.	Std. Error	t	Sig
INTERCEP	3.278	.257	12.750	.000
GDP	-1.054	0.366	-2.880	.007
GDP2	.163	0.090	1.810	.079
R-square	0.375			

A bowl-shaped quadratic equation takes its minimum at $x = -\beta_1/2\beta_2$. For these data, we estimate this point to be $x = 1.054/2(0.163) = 3.23$. The predicted fertility rate increases as GDP increases above this point (i.e., \$32,300).

DESCRIPTION AND INFERENCE ABOUT THE NONLINEAR EFFECT

For a polynomial model, R^2 for multiple regression describes the strength of the association. In this context, it describes the proportional reduction in error obtained from using the quadratic model, instead of $\bar{y}$, to predict y. Comparing this measure to r^2 for the straight-line model indicates how much better a fit the curvilinear model provides. Since a polynomial model has additional terms besides x, R^2 always is at least as large as r^2. The difference $R^2 - r^2$ measures the additional reduction in prediction error obtained by using the polynomial instead of the straight line.

For Table 14.6, the best-fitting straight-line prediction equation has $r^2 = 0.318$. From Table 14.7 for the quadratic model, $R^2 = 0.375$. The best quadratic equation explains about 6% more variability in y than does the best-fitting straight-line equation.

If $\beta_2 = 0$, the quadratic regression equation $E(y) = \alpha + \beta_1 x + \beta_2 x^2$ simplifies to the linear regression equation $E(y) = \alpha + \beta_1 x$. Therefore, to test the null hypothesis that the relationship is linear against the alternative that it is quadratic, we test H_0: $\beta_2 = 0$. The usual t test for a regression coefficient does this, dividing the estimate of β_2 by its standard error. The assumptions for applying inference are the same as for ordinary regression: randomization for gathering the data, a conditional distribution of y-values that is normal about the mean, with constant standard deviation σ at all x-values. The set of nations in Table 14.6 is not a random sample of nations, so inference is not relevant for those data.

CAUTIONS IN USING POLYNOMIAL MODELS

Some cautions are in order before you take the conclusions in this example too seriously. The scatterplot (Figure 14.8) suggests that the variability in fertility rates is considerably higher for nations with low GDPs than it is for nations with high GDPs. The fertility rates show much greater variability when their mean is higher. A GLM that permits nonconstant standard deviation by assuming a gamma distribution for y (see page 437) provides somewhat different results, including stronger evidence of nonlinearity (Exercise 14.14).

In fact, before we conclude that fertility rate increases above a certain value, we should realize that other models for which this does not happen are also consistent with these data. For instance, Figure 14.8 suggests that a "piecewise linear" model that has a linear decrease until GDP is about $25,000 and then a separate, nearly horizontal, line beyond that point fits quite well. A more satisfactory model for these data is one discussed in the next section of this chapter for *exponential regression*. Unless a data set is very large, several models may be consistent with the data.

In examining scatterplots, be cautious not to read too much into the data. Don't let one or two outliers suggest a curve in the trend. Good model building follows the principle of ***parsimony***: Models should have no more parameters than necessary to represent the relationship adequately. One reason is that simple models are easier to understand and interpret than complex ones. Another reason is that when a model contains unnecessary variables, the standard errors of the estimates of the regression coefficients tend to inflate, hindering efforts at making precise inferences. Estimates of the conditional mean of y also tend to be poorer than those obtained with well-fitting simple models.

When a polynomial regression model is valid, the regression coefficients do not have the partial slope interpretation usual for coefficients of multiple regression models. It does not make sense to refer to the change in the mean of y when x^2 is increased one unit and x is held constant. Similarly, it does not make sense to interpret the partial correlations $r_{yx^2 \cdot x}$ or $r_{yx \cdot x^2}$ as measures of association, controlling for x or x^2. However, the coefficient $r^2_{yx^2 \cdot x}$ does measure the proportion of the variation in y unaccounted for by the straight-line model that is explained by the quadratic model. In Example 14.7, applying the formula for $r^2_{yx_2 \cdot x_1}$ from page 332 yields

$$r^2_{yx^2 \cdot x} = \frac{R^2 - r^2_{yx}}{1 - r^2_{yx}} = \frac{0.375 - 0.318}{1 - 0.318} = 0.08.$$

Of the variation in y unexplained by the linear model, about 8% is explained by the introduction of the quadratic term.

With multiple explanatory variables, we may find that the fit improves by permitting one or more of them to have quadratic effects. For example, the model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

allows nonlinearity in x_2. For fixed x_1, the mean of y is a quadratic function of x_2. For fixed x_2, the mean of y is a linear function of x_1 with slope β_1. This model is a special case of multiple regression with three explanatory variables, in which x_3 is the square of x_2. Models allowing both nonlinearity and interaction are also possible.

14.6 Exponential Regression and Log Transforms*

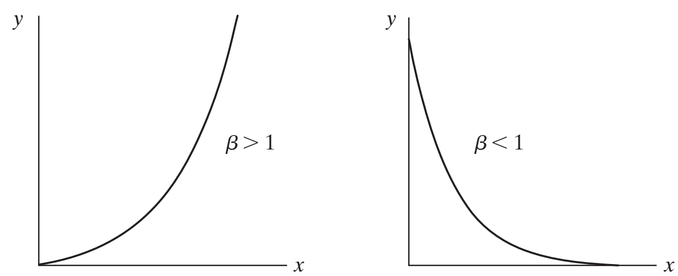
Although polynomials provide a diverse collection of functions for modeling nonlinearity, other mathematical functions are often more appropriate. The most important case is when the mean of the response variable is an *exponential* function of the explanatory variable.

Exponential Regression Function

An ***exponential regression*** function has the form $E(y) = \alpha\beta^x$.

In this equation, the explanatory variable appears as the exponent of a parameter. Unlike a quadratic function, an exponential function can take only positive values, and it continually increases (if $\beta > 1$) or continually decreases (if $\beta < 1$). In either case, it has a convex shape, as Figure 14.9 shows. We provide interpretations for the model parameters later in this section.

FIGURE 14.9: The Exponential Regression Function $E(y) = \alpha\beta^x$



For the exponential regression function, the *logarithm* of the mean is linearly related to the explanatory variable. That is, if $\mu = E(y) = \alpha\beta^x$, then

$$\log(\mu) = \log\alpha + (\log\beta)x.$$

The right-hand side of this equation has the straight-line form $\alpha' + \beta' x$ with intercept $\alpha' = \log(\alpha)$, the log of the α parameter, and slope $\beta' = \log(\beta)$, the log of the β parameter for the exponential regression function. This model form is the special case of a generalized linear model (GLM) using the log link function. If the model holds, a plot of the log of the y-values should show approximately a linear relation with x. (Don't worry if you have forgotten your high school math about logarithms. You will not need to know this in order to understand how to fit or interpret the exponential regression model.)

You can use GLM software to estimate the parameters in the model $\log[E(y)] = \alpha' + \beta' x$. The antilogs of these estimates are the estimates for the parameters in the exponential regression model $E(y) = \alpha\beta^x$, as shown below.

Example 14.8

Exponential Population Growth Exponential regression models well the growth of some populations over time. If the rate of growth remains constant, in percentage terms, then the size of that population grows exponentially fast. Suppose that the population size at some fixed time is α and the growth rate is 2% per year. After one year, the population is 2% larger than that at the beginning of the year. This means that the population size grows by a multiplicative factor of 1.02 each year. The population size after one year is $\alpha(1.02)$. Similarly, the population size after two years is

$$(\text{Population size at the end of one year})(1.02) = [\alpha(1.02)]1.02 = \alpha(1.02)^2.$$

After three years, the population size is $\alpha(1.02)^3$. After x years, the population size is $\alpha(1.02)^x$. The population size after x years follows an exponential function $\alpha\beta^x$ with parameters given by the initial population size α and the rate of growth factor, $\beta = 1.02$, corresponding to 2% growth.

Table 14.8 shows the U.S. population size (in millions) at 10-year intervals beginning in 1890. Figure 14.10 plots these values over time. Table 14.8 also shows the natural logarithm of the population sizes. (This uses the base e, where $e = 2.718\ldots$ is an irrational number that appears often in mathematics. The model makes sense with logs to any base, but software fits the GLM using natural logs, often denoted by $\log_e$ or LN.)

TABLE 14.8: U.S. Population Sizes and Log Population Sizes by Decade from 1890 to 2010, with Predicted Values for Exponential Regression Model

Year	No. Decades Since 1890 x	Population Size y	$\log_e(y)$	$\hat{y}$
1890	0	62.95	4.14	73.2
1900	1	75.99	4.33	82.7
1910	2	91.97	4.52	93.5
1920	3	105.71	4.66	105.6
1930	4	122.78	4.81	119.4
1940	5	131.67	4.88	134.9
1950	6	151.33	5.02	152.4
1960	7	179.32	5.19	172.3
1970	8	203.30	5.31	194.7
1980	9	226.54	5.42	220.0
1990	10	248.71	5.52	248.7
2000	11	281.42	5.64	281.0
2010	12	308.75	5.73	317.5

Source: U.S. Census Bureau; data file `Population` at the text website.

Figure 14.11 plots these log of population size values over time. The log population sizes appear to grow approximately linearly. This suggests that population growth over this time period was approximately exponential, with a constant rate of growth. We now estimate the regression curve, treating time as the explanatory variable x. For convenience, we identify the time points $1890, 1900, \ldots, 2010$ as times $0, 1, \ldots, 12$; that is, x represents the number of decades since 1890.

We use software to estimate the generalized linear model $\log(\mu) = \alpha' + \beta' x$, assuming a normal distribution for y. The prediction equation is

$$\log_e(\hat{\mu}) = 4.29285 + 0.12233x.$$

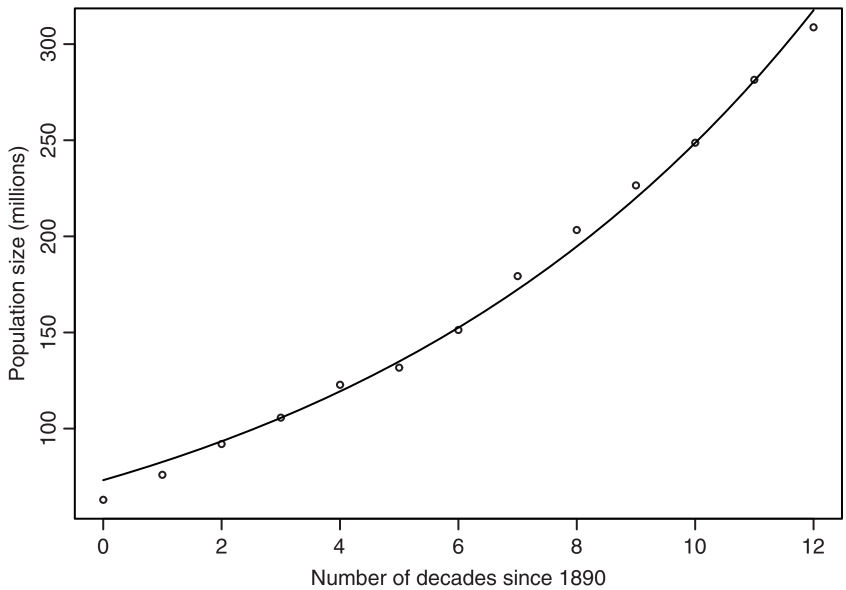


FIGURE 14.10: U.S. Population Size Since 1890. The fitted curve is the exponential regression, $\hat{y} = 73.2(1.1301)^x$.

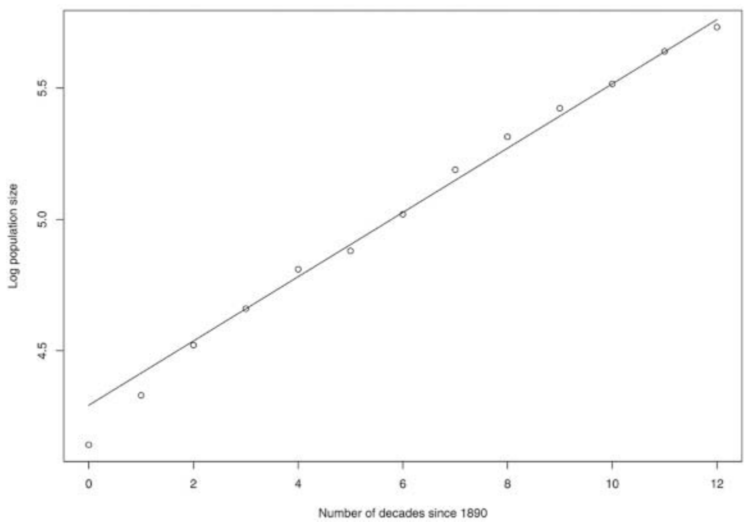


FIGURE 14.11: Log Population Sizes Since 1890. The prediction equation is $\widehat{\log y} = 4.29 + 0.122x$.

Antilogs of these estimates are the parameter estimates for the exponential regression model. For natural logs, the antilog function is the exponential function e^x. That is, antilog(4.29285) $= e^{4.29285} = 73.175$, and antilog(0.12233) $= e^{0.12233} = 1.1301$. Thus, for the exponential regression model $E(y) = \alpha\beta^x$, the estimates are $\hat{\alpha} = 73.175$ and $\hat{\beta} = 1.1301$. The prediction equation is

$$\hat{y} = \hat{\alpha}\hat{\beta}^x = 73.175(1.1301)^x.$$

The predicted initial population size (in 1890) is $\hat{\alpha} = 73.2$ million. The predicted population size x decades after 1890 is $\hat{y} = 73.175(1.1301)^x$. For 2010, for instance, $x = 12$, and the predicted population size is $\hat{y} = 73.175(1.1301)^{12} = 317.5$ million. Table 14.8 shows the predicted values for each decade. Figure 14.10 plots the exponential prediction equation.

The predictions are quite good, except for the first couple of observations. The total sum of squares of population size values about their mean is TSS = 76,791, whereas the sum of squared errors about the prediction equation is SSE = 419. The proportional reduction in error is (76,791 − 419)/76,791 = 0.995. The ordinary linear model $E(y) = \alpha + \beta x$ also fits quite well, having $r^2 = 0.980$. ■

A caution: The fit of the model $\log[E(y)] = \alpha' + \beta' x$ that you get with GLM software will *not* be the same as you get by taking logarithms of all the y-values and then fitting a straight-line model using least squares. The latter approach[3] gives the fit for the model $E[\log(y)] = \alpha' + \beta' x$. For that model, taking antilogs does not take you back to $E(y)$, because $E[\log(y)]$ is not equivalent to $\log[E(y)]$. So, in software it is preferable to use a generalized linear modeling option rather than an ordinary regression option.

INTERPRETING EXPONENTIAL REGRESSION MODELS

Now let's take a closer look at how to interpret parameters in the exponential regression model, $E(y) = \alpha\beta^x$. The parameter α represents the mean of y when $x = 0$. The parameter β represents the ***multiplicative*** change in the mean of y for a one-unit increase in x. The mean of y at $x = 12$ equals β *multiplied* by the mean of y at $x = 11$. For instance, for the equation $\hat{y} = 73.175(1.1301)^x$, the predicted population size at a particular date equals 1.1301 times the predicted population size a decade earlier.

By contrast, the parameter β in the *linear* model $E(y) = \alpha + \beta x$ represents the ***additive*** change in the mean of y for a one-unit increase in x. In the linear model, the mean of y at $x = 12$ is β *plus* the mean of y at $x = 11$. The prediction equation for the linear model (i.e., identity link) fitted to Table 14.8 is $\hat{y} = 46.51 + 20.33x$. This model predicts that the population size increases by 20.33 million people every decade.

In summary, for the linear model, $E(y)$ changes by the same *quantity* for each one-unit increase in x, whereas for the exponential model, $E(y)$ changes by the same *percentage* for each one-unit increase. For the exponential regression model with Table 14.8, the multiplicative effect of 1.1301 for each decade corresponds to a predicted 13.01% growth per decade.

Suppose the growth rate is 15% per decade, to choose a rounder number. This corresponds to a multiplicative factor of 1.15. After five decades, the population grows by a factor of $(1.15)^5 = 2.0$. That is, after five decades, the population size doubles. If the rate of growth remained constant at 15% per decade, the population would double every 50 years. After 100 years, the population size would be quadruple the original size, after 150 years it would be 8 times as large, after 200 years it would be 16 times its original size, and so forth.

The exponential function with $\beta > 1$ has the property that its doubling time is a constant. As can be seen from the sequence of population sizes at 50-year intervals, this is an extremely fast increase even though the annual rate of growth (1.4% annually for a decade increase of 15%) seems small. In fact, this has been the approximate growth of the world population in the past century. (See Exercise 14.22.)

Example 14.9 **Exponential Regression for Fertility Data** When $\beta < 1$ in the exponential regression model, $\beta' = \log(\beta) < 0$ in the log-transformed GLM. In this case, the mean of y *decreases* exponentially fast as x increases. The curve then looks like the second curve in Figure 14.9.

[3] For example, as SPSS gives by selecting *Regression* in the *Analyze* menu, followed by the choice of *Curve Estimation* with the *Exponential* option.

In Example 14.7 with Table 14.6 (page 441), we modeled y = fertility rate for several countries, with x = per capita GDP. If we expect $E(y)$ to continually decrease as x increases, an exponentially decreasing curve may be more appropriate. In fact, the exponential regression model provides a good fit for those data. Using the GLM with log link for y = fertility rate and x = per capita GDP and assuming a normal distribution for y, we get the prediction equation

$$\log_e(\hat{\mu}) = 1.148 - 0.206x.$$

Taking antilogs yields the exponential prediction equation

$$\hat{y} = \hat{\alpha}\hat{\beta}^x = e^{1.148}(e^{-0.206})^x = 3.15(0.81)^x.$$

The predicted fertility rate at GDP value $x + 1$ equals 81% of the predicted fertility rate at GDP value x; that is, it decreases by 19% for a \$10,000 increase in per capita GDP.

With this fit, the correlation between the observed and predicted fertility rates equals 0.59, nearly as high as the value of 0.61 achieved with the quadratic model, which has an extra parameter. If we expect fertility rate to decrease continuously as GDP increases, the exponential regression model is a more realistic model than the quadratic regression model, which predicted increasing fertility above a certain GDP level. Also, unlike the straight-line model, the exponential regression model cannot yield negative predicted fertility rates.

Since the scatterplot in Figure 14.8 suggests greater variability when the mean fertility rate is higher, it may be even better to assume a gamma distribution for y with this exponential regression model. The prediction equation is then

$$\log_e(\hat{\mu}) = 1.112 - 0.177x, \quad \text{for which} \quad \hat{y} = e^{1.112}(e^{-0.177})^x = 3.04(0.84)^x.$$

This gives a slightly shallower rate of decrease than the fit $3.15(0.81)^x$ for the normal response model. ■

TRANSFORMING THE EXPLANATORY VARIABLE TO ACHIEVE LINEARITY

Other transformations of the response mean or of explanatory variables are useful in some situations. For example, suppose y tends to increase or decrease over a certain range of x-values, but once a certain x-value has been reached, further increases in x have less effect on y, as in Figure 14.5b. For this concave increasing type of trend, x behaves like an exponential function of y. Taking the logarithms of the x-values often linearizes the relationship. Another possible transform for this case uses $1/x$ as the explanatory variable.

14.7 Robust Variances and Nonparametric Regression*

Recent years have seen yet other ways developed to generalize regression to handle violations of assumptions for the ordinary linear model. Detailed explanations of such generalizations are beyond the scope of this book, but in this section we briefly introduce two popular ones.

ROBUST VARIANCE ESTIMATES

The ordinary regression model assumes a normal distribution for y with constant variability at all settings of the explanatory variables. Section 14.4 introduced the *generalized linear model*, which permits alternative distributions that have nonconstant variability, such as the gamma distribution. An alternative approach uses the least squares estimates but does not assume constant variance in finding standard errors. Instead, it adjusts ordinary standard error formulas to reflect the empirical variability displayed by the sample data.

This alternative standard error estimate is sometimes called the ***sandwich estimate***, because of how its formula sandwiches the empirical variability between two terms from the ordinary formula. It is also referred to as a ***robust standard error estimate***, because it is more valid than the ordinary *se* when the true response variability is not constant. Some software[4] now makes this available. If you use it and find standard errors quite different from those given in an ordinary regression analysis, basic assumptions are likely violated and you should treat its results skeptically.

To illustrate, we found robust standard errors for the house selling price data analyzed on page 438 with least squares and with a gamma GLM. For the effects of (size, new, taxes), the robust *se* values are (22.4, 26245, 9.3), compared to (12.5, 16459, 6.7) for the ordinary *se* values. Such highly different results make us wary of the ordinary *se* values. As explained previously, the clear increase in the variability of $y =$ selling price as its mean increases made us skeptical of the ordinary regression results.

This robust variance approach also extends to handle violations of the assumption of independent observations, such as those that occur with clustered data and longitudinal studies. This approach incorporates the empirical variability and correlation within clusters to generate standard errors that are more reliable than ones that treat observations within clusters as independent. This method for clustered correlated data uses ***generalized estimating equations*** (GEEs) that resemble equations used to obtain maximum likelihood estimates, but without a parametric probability distribution incorporating correlations.

This way of handling clustered data is an alternative to the *linear mixed model* introduced in Section 13.5. The linear mixed model assumes normality for the response variable and adds *random effects* to an ordinary model. Likewise, we can add random effects to a generalized linear model to obtain a *generalized linear mixed model* to handle clustering with nonnormal responses such as the binomial. The robust variance approach has the advantage of not requiring an assumption about the distribution of y or the correlation structure within clusters. However, it has the disadvantage that (because of the lack of a distribution assumption) likelihood-based methods such as maximum likelihood estimates and likelihood ratio tests are not available.

NONPARAMETRIC REGRESSION

Recent advances make it possible to fit models to data without assuming particular functional forms, such as straight lines or parabolas, for the relationship. These approaches are *nonparametric*, in terms of having fewer (if any) assumptions about the functional form and the distribution of y. It is helpful to look at a plot of a fitted nonparametric regression model to learn about trends in the data.

One nonparametric regression method, called ***generalized additive modeling***, is a further generalization of the generalized linear model. It has the form

$$g(\mu) = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where $f_1, \ldots, f_p$ are unspecified and potentially highly complex functions. The GLM is the special case in which each of these functions is linear. The estimated functional

[4] For example, Stata with the *robust* option for its *regress* command.

form of the relationship for each explanatory variable is determined by a computer algorithm, using the sample data. As in GLMs, with this model you can select a particular link function g and also a distribution for y. This model is useful for smoothing data to reveal overall trends.

Other nonparametric smoothing methods do not even require selecting a link function or a distribution for y. Popular smoothers are ***LOESS*** and ***kernel*** methods that get the prediction at a particular point by smoothly averaging nearby values. The smoothed value is found by fitting a low-degree polynomial while giving more weight to observations near the point and less weight to observations further away. You can achieve greater smoothing by choosing a larger *bandwidth*, essentially by letting the weights die out more gradually as you move away from each given point.

Figure 14.12 shows two plots of nonparametric regression fits for the fertility rate data of Table 14.6. The first plot employs greater smoothing and has a curved, decreasing trend. It is evident that the response may not eventually increase, as a quadratic model predicts. This fit suggests that the exponential regression model is more satisfactory than the quadratic model for these data.

To learn more about robust regression and nonparametric regression, see Fox (2015, Chapters 18 and 19).

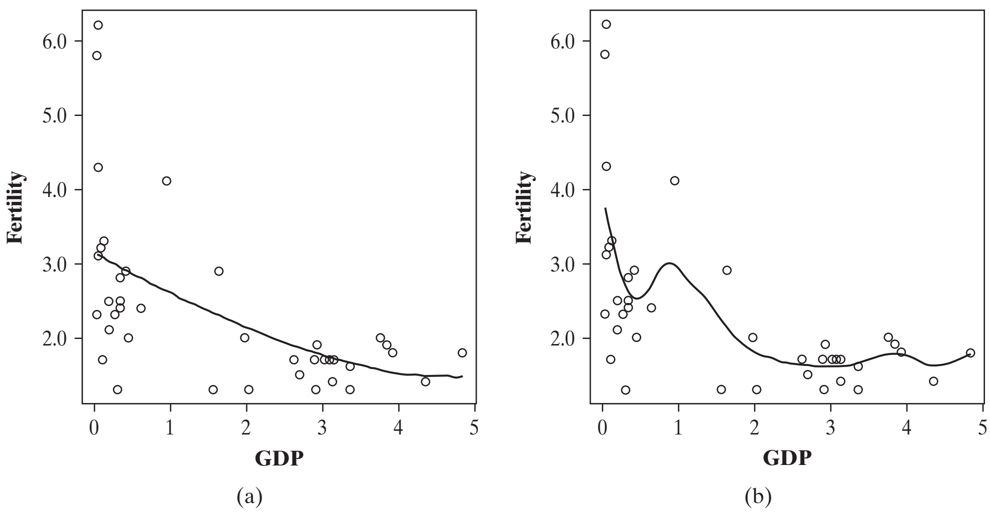


FIGURE 14.12: Fits of Nonparametric Regression Model (Using SPSS) to Smooth the Fertility Rate Data of Table 14.6. Fit (a) employs greater smoothing (bandwidth = 5) than fit (b) (bandwidth = 1).

14.8 Chapter Summary

This chapter discussed issues about building regression models and showed how to check assumptions and how to ease some restrictions of the basic linear model.

- With a large number of potential explanatory variables for a model, the ***backward elimination*** and ***forward selection*** procedures use a sequential algorithm to select variables. These are exploratory in purpose and should be used with caution. Fit indices such as adjusted R^2, PRESS, and AIC also provide criteria for model selection.

- Plots of the ***residuals*** check whether the model is adequate and whether the assumptions for inferences are reasonable. Observations having a large leverage and large studentized residual have a strong influence on the model fit. Diagnostics such as DFBETA and DFFIT describe which observations have a strong influence on the parameter estimates and the model fit.