CPH 931 — Fall 2009 — Dr. Charnigo

Lecture 3

Going beyond linear regression via ordinary least squares

Recalling the linear regression model. Suppose that the mean of a continuous response variable Y depends on the numerical values taken on by several explanatory variables $X_1, \ldots, X_k$. Let $m(x_1, \ldots, x_k)$ denote the mean of Y when the variables $X_1, \ldots, X_k$ have taken on the numerical values $x_1, \ldots, x_k$.

A linear regression model asserts that

$$m(x_1, \ldots, x_k) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \quad \text{and} \tag{1}$$

$$Y_i = m(x_{1,i}, \ldots, x_{k,i}) + \epsilon_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \epsilon_i, \tag{2}$$

where the ϵ_i are error terms representing the differences between actual and expected responses.

Assumptions of the linear regression model and ordinary least squares. In addition to assuming that equation (1) is correct (i.e., no important explanatory variables have been omitted, and the mean of Y really is linear in $x_1, \ldots, x_k$), we usually assume that the ϵ_i are independent normal random variables with mean 0 and (unknown but) constant variance $\sigma^2 > 0$.

Under these assumptions, the method of ordinary least squares is employed to estimate the regression coefficients $\alpha, \beta_1, \ldots, \beta_k$. That is, for a given data set the numbers $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are taken as estimates of $\alpha, \beta_1, \ldots, \beta_k$ because the sum of squares

$$\sum_{i=1}^{n} (y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})^2 \tag{3}$$

is minimized when $r_0 = \hat{\alpha}, r_1 = \hat{\beta}_1, \ldots, r_k = \hat{\beta}_k$.

Implications of violating the assumptions. If the usual assumptions are not satisfied or if there are other problems (e.g., multicollinearity or missing data), then the inferences obtained through ordinary least squares estimation (e.g., decisions to accept or reject null hypotheses based on t-tests and f-tests) may not be credible. In Lectures 3 and 4 we will discuss alternative estimation approaches that can accommodate departures from the usual assumptions and other problems. (In Lecture 2 we have already discussed an alternative modeling strategy.)

Heteroscedasticity and weighted least squares

Heteroscedasticity. Suppose that the usual assumptions are satisfied <u>except</u> that the ϵ_i do not have constant variance. That is,

$$Var[\epsilon_i] = \sigma^2(x_{1,i}, \ldots, x_{k,i}), \tag{4}$$

where $\sigma^2(x_{1,i}, \ldots, x_{k,i})$ is not constant as a function of $x_{1,i}, \ldots, x_{k,i}$. Then we say that heteroscedasticity is present.

Weighted least squares. If $Var[\epsilon_i]$ is large, then Y_i may be far away from its expected value $m(x_{1,i}, \ldots, x_{k,i}) = \alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$. In this case Y_i will not provide good information about the regression coefficients $\alpha, \beta_1, \ldots, \beta_k$. On the other hand, if $Var[\epsilon_i]$ is small, then Y_i will be close to its expected value and will provide good information about the regression coefficients. Thus, rather than minimize

$$\sum_{i=1}^{n} (y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})^2 \tag{5}$$

to obtain estimates, we should minimize

$$\sum_{i=1}^{n} w_i \; (y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})^2. \tag{6}$$

Above, the weights w_i are positive numbers that are "large" for y_i providing good information and "small" for y_i not providing good information. This estimation approach is called weighted least squares.

The maximum likelihood weights. The terms "large" and "small" are, of course, rather vague. Fortunately, the statistical principle of maximum likelihood provides a more concrete answer. The w_i should be chosen inversely proportional to $Var[\epsilon_i]$. That is, we should set

$$w_i = C/Var[\epsilon_i] \tag{7}$$

for some positive constant C that does not depend on i. The specific value of C does not matter since it can be factored out of expression (6).

Remark on the maximum likelihood weights. If the ϵ_i did have constant variance σ^2, then by taking $C = \sigma^2$ we would attain $w_i = 1$ in expression (7), thereby reducing expression (6) to expression (5). In other words, ordinary least squares is consistent with the statistical principle of maximum likelihood when the usual assumptions are met.

Example: prostate data set. Refer to "prostresults.rtf". Page 1 shows the results of fitting a linear regression model by ordinary least squares. The response variable is the level of prostate specific antigen (after square root transformation). The explanatory variables are cancer volume (after logarithmic transformation) and prostate weight (after logarithmic transformation). Page 2 shows a plot of the ordinary residuals against the predicted/fitted values of prostate specific antigen, while page 3 shows a plot of the internally studentized residuals against the predicted/fitted values of prostate specific antigen.

In CPH 930 you learned that a fan-shaped pattern in a plot of residuals against predicted/fitted values is a hallmark of nonconstant error variance. In this example the error variances seem to increase (as evidenced by the greater spread of the residuals) with larger expected responses (for which predicted/fitted values are proxies). That is, values $x_{1,i}, \ldots, x_{k,i}$ that make $m(x_{1,i}, \ldots, x_{k,i})$ large also make $\sigma^2(x_{1,i}, \ldots, x_{k,i})$ large.

Since we have already transformed the data, we will employ weighted least squares to address the heteroscedasticity. However, there is an obstacle: we don't know what the error variances are, so we can't evaluate expression (7).

A first (unsuccessful) attempt at a practical solution. Although we don't know what the error variances are, we can try to estimate them from the ordinary residuals obtained via ordinary least squares.

Suppose that equation (4) applies but that we have employed ordinary least squares. Probability theory tells us that

$$E|\epsilon_i| = \sqrt{2/\pi} \times \sigma(x_{1,i}, \ldots, x_{k,i}), \tag{8}$$

where E denotes expected value. Since an ordinary residual e_i is a proxy for the corresponding error ϵ_i, one can argue that (up to a multiplicative factor of $\sqrt{2/\pi}$) the absolute value of the ordinary residual $|e_i|$ is a proxy for the error standard deviation $\sigma(x_{1,i}, \ldots, x_{k,i}) = \sqrt{Var[\epsilon_i]}$.

These considerations suggest choosing

$$w_i = 1/|e_i|^2, \tag{9}$$

where e_i is as determined by ordinary least squares.

A second (successful) attempt at a practical solution. Equation (9) does not provide a practical solution because an ordinary residual can equal 0, in which case equation (9) mandates an infinite weight. Even if none of the ordinary residuals equals 0 exactly, equation (9) can still provide unreasonably large weights for some observations.

The issue here is that $|e_i|$ is an imperfect proxy for $E|\epsilon_i|$. Sometimes $|e_i|$ will not be close to $|\epsilon_i|$. Moreover, $|\epsilon_i|$ may not be close to its own expected value. This begs the question, is there a way to replace $|e_i|$ by a quantity closer to $E|\epsilon_i|$?

Well, there is a statistical technique for replacing a datum by a quantity closer to its expected value: that technique is linear regression! Thus, a practical solution entails fitting an auxiliary regression model in which the absolute values of the ordinary residuals are regressed against the corresponding predicted/fitted values from the original model in equations (1) and (2). The form of the auxiliary regression model is

$$|e_i| = \alpha^* + \beta^* \hat{y}_i + \epsilon_i^*, \tag{10}$$

where the $\hat{y}_i$ represent the predicted/fitted values and the asterisks distinguish the auxiliary regression model in equation (10) from the original model in equations (1) and (2).

Now let $\widehat{|e_i|}$ denote the predicted/fitted values for the auxiliary regression model in equation (10). The $\widehat{|e_i|}$ are closer to the $E|\epsilon_i|$ than are the $|e_i|$, and so the $\widehat{|e_i|}$ can replace the $|e_i|$ in equation (9). That is, we can take

$$w_i = 1/\widehat{|e_i|}^2. \tag{11}$$

To summarize, we end up fitting three linear regression models. First, we fit the original model in equations (1) and (2) via ordinary least squares; the purpose is to extract the $|e_i|$ and $\hat{y}_i$ for equation (10). Second, we fit the auxiliary regression model in equation (10) via ordinary least squares; the purpose is to extract the $\widehat{|e_i|}$ for equation (11). Third, we fit the original model in equations (1) and (2) via weighted least squares using equation (11); this is the model from which we actually make our inferences.

Example, continued: prostate data set. Page 5 of "prostresults.rtf" shows the results of regressing the absolute values of the ordinary residuals against the predicted/fitted values from the model on page 4 of "prostresults.rtf". Note that the estimate of β^* is positive and significantly different from zero, confirming what we already knew from the plots on pages 2 and 3: the errors have greater variances as the expected responses become larger.

The predicted/fitted values from the model on page 5 are used, via equation (11), to perform the weighted least squares analysis on page 6. The standard errors for the regression coefficient estimates have been drastically reduced: from 1.53915 to 0.84712 for α, from 0.17923 to 0.08762 for β_1, and from 0.42535 to 0.24549 for β_2. The last reduction is partially responsible for the qualitatively different conclusion about prostate weight: the p-value is less than 0.0001 with weighted least squares, compared to 0.1203 with ordinary least squares.

The most important point to take from this example is that the estimation of regression coefficients will be needlessly imprecise (possibly leading to Type II errors) if one disregards heteroscedasticity and applies ordinary least squares when weighted least squares is appropriate.

Multicollinearity and ridge regression

Multicollinearity. Recall from CPH 930 that explanatory variables $X_1, X_2, \ldots, X_k$ exhibit multicollinearity if

$$c_1 X_1 + c_2 X_2 + \cdots + c_k X_k$$

is approximately constant for some numbers $c_1, c_2, \ldots, c_k$ that are not all 0. The problem with multicollinearity is that, when it is present, ordinary least squares will yield highly imprecise estimates of some regression coefficients.

Example: cholesterol data. Refer to "cholresults.rtf". Page 1 shows the results of fitting a linear regression model by ordinary least squares. The response variable is total serum cholesterol. The explanatory variables are total fat intake, saturated fat intake, vegetable fat intake, polyunsaturated fat intake, and animal fat intake.

Just from the names of the explanatory variables we have good reason to anticipate multicollinearity, and this is confirmed by the variance inflation factors. A common rule of thumb is that variance inflation factors greater than 10 may warrant attention, but here there are three variance inflation factors in excess of $100,000$. Moreover, the corresponding standard errors are in excess of 100.

A key insight. Although the regression coefficients are unknown (otherwise why collect data and fit a linear regression model?), we often have some prior knowledge/beliefs about them. Since the problem with multicollinearity is that ordinary least squares yields highly imprecise estimates, one strategy for addressing multicollinearity is to employ an alternative estimation framework that exploits our prior knowledge/beliefs about the regression coefficients. The alternative estimation framework that we will consider is called ridge regression.

Ridge regression. For the following description I assume that the response and explanatory variables have been standardized. This is handled automatically by SAS when you request ridge regression; you yourself do not need to preprocess the data. Moreover, SAS presents results that apply to the response and explanatory variables as they were originally scaled; you yourself do not need to convert results for the standardized variables to results for the variables on their original scales.

Instead of minimizing

$$\sum_{i=1}^{n}(y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})^2, \tag{12}$$

we minimize

$$\sum_{i=1}^{n}(y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})^2 + \lambda(r_0^2 + r_1^2 + \cdots + r_k^2), \tag{13}$$

where λ is a nonnegative quantity called the “ridge parameter”.

If λ exceeds 0, then the effect of the “penalty term” $\lambda(r_0^2 + r_1^2 + \cdots + r_k^2)$ is that extremely large (positive or negative) values for $r_0, r_1, \ldots, r_k$ cannot minimize expression (13). Thus, $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ cannot be extremely large. The penalty term reflects our prior knowledge/beliefs that $\alpha, \beta_1, \ldots, \beta_k$ are not too large; the ridge parameter determines how strongly we assert our prior knowledge/beliefs.

If λ equals 0, then ridge regression reduces to ordinary least squares.

Remarks on ridge regression. The downside of imposing prior knowledge/beliefs about $\alpha, \beta_1, \ldots, \beta_k$ in ridge regression is that $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ will be biased toward 0. Thus, the average value of $\hat{\beta}_1$ over repeated sampling will not equal

β_1 but will be some number that is smaller in magnitude than β_1. On the other hand, if X_1 is a primary cause of the multicollinearity, then the variance of $\hat{\beta}_1$ over repeated sampling will be much smaller with ridge regression than with ordinary least squares. Similar statements apply to the other regression coefficients.

Although many statistical methods reflect a desire to avoid bias (recall from your first statistics course the reason for dividing by $n-1$ instead of by n when computing a sample variance!), there are some situations in which we gladly accept a small amount of bias to reduce variability. Multicollinearity is one such situation.

The ridge trace. Page 2 of "cholresults.rtf" illustrates a "ridge trace". This is a plot of coefficient estimates obtained via ridge regression using several values for the ridge parameter. When λ equals 0, we have the ordinary least squares estimates. As λ moves away from 0, the estimates shift abruptly but then appear to stabilize. If we let λ increase indefinitely, then the estimates would very gradually dissipate to 0.

Remark on the ridge trace. If there were no multicollinearity, then there would be no abrupt shift near the beginning of the ridge trace.

Choosing a value for the ridge parameter. A ridge trace shows multiple sets of coefficient estimates, but in the end we need to be able to report a single set of coefficient estimates. Thus, we must choose a single value for the ridge parameter. A rule of thumb is to take the smallest value at which the coefficient estimates appear to have stabilized.

Example, continued: cholesterol data. The ridge trace on page 2 of "cholresults.rtf" shows that the coefficient estimates appear to have stabilized at $\lambda = 0.002$. Choosing this value for the ridge parameter, we obtain the results on page 4 of "cholresults.rtf".

The coefficient estimate (standard error) for total fat intake changed from -3.361 (108.530) with ordinary least squares to 0.070 (0.247) with ridge regression.

The coefficient estimate (standard error) for vegetable fat intake changed from 2.860 (108.589) to -0.574 (0.375).

The coefficient estimate (standard error) for animal fat intake changed from 4.316 (108.428) to 0.870 (0.629).

The ridge regression estimates for total fat intake, vegetable fat intake, and animal fat intake are thus much closer to 0 than the ordinary least squares estimates. Even more dramatically, the corresponding standard errors have been reduced by at least 99%.

On the other hand, the ridge regression estimates (and standard errors) for saturated fat intake and polyunsaturated fat intake are not appreciably different from the ordinary least squares estimates (and standard errors).

Nonnormal errors and robust regression

Scenario. Suppose that $\epsilon_1, \ldots, \epsilon_n$ are independent and that they have identical probability distributions — not necessarily a normal distribution — with mean 0 and (unknown constant) variance $\sigma^2 > 0$.

Normality of the errors, maximum likelihood, and ordinary least squares. If ϵ_i is normally distributed, then Y_i is normally distributed with mean $\alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$ and variance σ^2. In this case, the likelihood function is defined by

$$L(r_0, r_1, \ldots, r_k) := (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - r_0 - r_1x_{1,i} - \cdots - r_kx_{k,i})^2\right]. \tag{14}$$

Roughly speaking, the likelihood function represents how probable the observed responses would be if the intercept were r_0 and the partial slope coefficients were $r_1, \ldots, r_k$. Since the observed responses should be more probable for $r_0, r_1, \ldots, r_k$ near to $\alpha, \beta_1, \ldots, \beta_k$ (i.e., for $r_0, r_1, \ldots, r_k$ consistent with the mechanism actually generating the observed responses) than for $r_0, r_1, \ldots, r_k$ far away from $\alpha, \beta_1, \ldots, \beta_k$ (i.e, for $r_0, r_1, \ldots, r_k$ inconsistent with the mechanism actually generating the observed responses), finding $r_0, r_1, \ldots, r_k$ that maximize the likelihood function is anticipated to yield good estimates of $\alpha, \beta_1, \ldots, \beta_k$.

Now consider the sum of squares

$$SS(r_0, r_1, \ldots, r_k) := \sum_{i=1}^{n}(y_i - r_0 - r_1x_{1,i} - \cdots - r_kx_{k,i})^2. \tag{15}$$

Substituting (15) into (14) shows that

$$L(r_0, r_1, \ldots, r_k) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}SS(r_0, r_1, \ldots, r_k)\right]. \tag{16}$$

From (16) we see that the likelihood is big when the sum of squares is small. In particular, the likelihood is maximized when the sum of squares is minimized. Thus, if the errors are normally distributed, ordinary least squares is the same as maximum likelihood and is anticipated to yield good estimates.

Nonnormality of the errors. Suppose that we have fitted a linear regression model by ordinary least squares. To determine whether the errors are normally distributed, we can create a normal probability plot of studentized residuals. If this plot exhibits a severe departure from a straight-line pattern, then we must conclude that the errors are not normally distributed or that there are outliers in the data set. If transforming the response variable does not fix the problem, and if we do not want to simply "throw away" outliers, how shall we estimate $\alpha, \beta_1, \ldots, \beta_k$ once we know that ϵ_1 through ϵ_n are not normally distributed?

One option, if we knew (or were willing to guess) the distribution of ϵ_1 through ϵ_n, would be to employ maximum likelihood estimation; we would appropriately redefine the likelihood function $L(r_0, r_1, \ldots, r_k)$. Of course, relation (16) would no longer be valid, meaning that maximum likelihood estimation with nonnormal errors is not equivalent to ordinary least squares.

A second option, which we will pursue today, replaces $SS(r_0, r_1, \ldots, r_k)$ by an alternative "badness" criterion to be minimized. This option can be used if we do not know (and are unwilling to guess) the distribution of ϵ_1 through ϵ_n. Before providing details, I will introduce a motivating example.

Example. The data set {FEV.xls} provides information on forced expiratory volume, age, height, gender, and smoking status for 145 children and adolescents. I used ordinary least squares to estimate the presumed-linear relationship between forced expiratory volume (after logarithmic transformation) and age, height, gender, smoking status.

The results of the ordinary least squares analysis are summarized on page 1 of {pulmresults.rtf}. Corresponding to each one-year increase in age (holding fixed the other explanatory variables) is an estimated average increase of 0.047 points in forced expiratory volume on the logarithmic scale. Initially this seems difficult to interpret, but a 0.047-point increase on the logarithmic scale implies a *multiplication* by $\exp[0.047] = 1.048$ on the original scale. So, roughly speaking, each one-year increase in age (holding fixed the other explanatory variables) yields an estimated average increase of 4.8% in forced expiratory volume.

However, if we continue to pages 7 through 11 of {pulmresults.rtf}, we see that there are four outlying observations. On page 7, I have plotted the externally studentized residuals against the predicted/fitted values of log forced expiratory volume; on pages 8 through 11, I have plotted the externally studentized residuals against the values of age, height, gender, and smoking status. A normal probability plot of studentized residuals (not shown) would also call attention to these four observations.

Pages 2 through 6 of {pulmresults.rtf} tabulate ordinary residuals and externally studentized residuals along with DFBETAs and other measures of influence. The two "low" outlying observations correspond to subjects 21 and 78, while the two "high" outlying observations correspond to subjects 135 and 145. All four of the observations have positive DFBETAs for age. This suggests that the estimate of 0.047 may be overstated. We can understand this intuitively by noting that subjects 21 and 78 were younger and had rather low forced expiratory volumes, while subjects 135 and 145 were older and had rather high forced expiratory volumes. Hence, the observations from these four subjects may exaggerate the strength of the relationship between forced expiratory volume and age.

Robust regression and M estimation. A method allowing us to make inferences about $\alpha, \beta_1, \ldots, \beta_k$ in the absence of a normality assumption is generically referred to as a robust regression method. The specific robust regression method we will consider today is called M estimation and is the default for the ROBUSTREG procedure in SAS.

Suppose for the sake of discussion that σ^2 is known. The idea of M estimation is to replace $SS(r_0, r_1, \ldots, r_k)$ by another "badness" criterion

$$CRIT(r_0, r_1, \ldots, r_k) := \sum_{i=1}^{n} \rho\left[(y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})/\sigma\right] \tag{17}$$

that does not "overreact" when y_i is far away from $r_0 + r_1 x_{1,i} + \cdots + r_k x_{k,i}$.

Of course there are many ways in which we could choose ρ, and not all of them would be good. If we defined ρ by $\rho(t) := t^2$, then $CRIT(r_0, r_1, \ldots, r_k)$ would be equivalent to $SS(r_0, r_1, \ldots, r_k)$ and nothing would change. The default choice of ρ in SAS is

$$\rho(t) := t^2 - t^4/c^2 + t^6/(3c^4) \quad \text{for } |t| < c \quad \text{and} \tag{18}$$

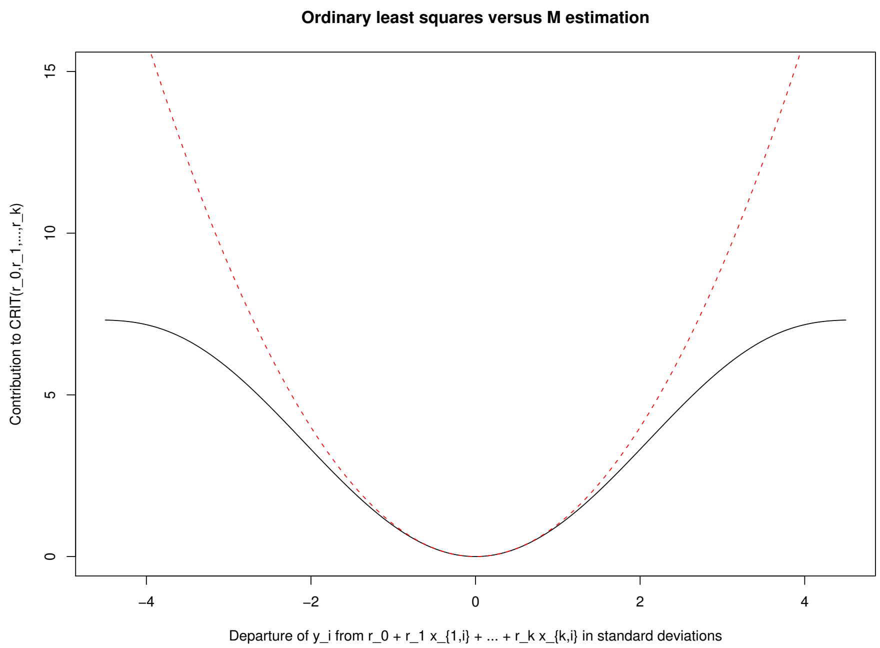
$$\rho(t) := c^2 - c^4/c^2 + c^6/(3c^4) = c^2/3 \quad \text{for } |t| \geq c, \tag{19}$$

where $c := 4.685$.

The graphic on the next page shows a plot of $\rho(t)$ against t (black solid curve) and, for comparative purposes, a plot of t^2 against t (red dashed curve). Clearly $\rho(t)$ is similar to t^2 when y_i departs from $r_0 + r_1 x_{1,i} + \cdots + r_k x_{k,i}$ by less than two standard deviations; however, $\rho(t)$ is much smaller than t^2 when y_i departs from $r_0 + r_1 x_{1,i} + \cdots + r_k x_{k,i}$ by more than two standard deviations. In essence ρ permits each observation to make only a limited contribution to $CRIT(r_0, r_1, \ldots, r_k)$, so that <u>each observation — no matter how bad an outlier it might be or how severely nonnormal the error distribution might be — can have only a limited impact on the estimation of $\alpha, \beta_1, \ldots, \beta_k$</u>.

In practice σ^2 is not known. Fortunately the ROBUSTREG procedure in SAS can adapt what has been described above to simultaneously estimate σ^2 along with $\alpha, \beta_1, \ldots, \beta_k$.

Remark on M estimation. Although M estimation is useful when there are outlying observations in the sense that y_i is not close to $\alpha + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$,



M estimation does nothing to help us when there are outlying observations in the sense that one or more of $x_{1,i}, \ldots, x_{k,i}$ are extreme. The latter outlying observations are referred to as "high leverage points" and must be handled another way (e.g., transformations of the affected explanatory variables, weighted least squares with reduced weights for high leverage points, or a different robust regression method).

Example, continued. The "Parameter Estimates" box on page 12 provides inferences about $\alpha, \beta_1, \ldots, \beta_k$ along with an estimate of σ. In this example we are particularly interested in seeing what happened to the coefficient estimate for age. Ordinary least squares yielded a coefficient estimate of 0.047, which we thought might have been overstated. We see that M estimation yields a coefficient estimate of 0.037, supporting the idea that the 0.047 might have been overstated.

Discussion questions

1. Suppose that you applied ordinary least squares and obtained a plot of ordinary residuals against predicted/fitted values in which the residuals had a great deal of spread for extremely small or large predicted/fitted values but rather little spread for moderate predicted/fitted values. How should the approach presented in equation (10) be modified?

2. Why might we want to use ridge regression to address multicollinearity, rather than simply eliminating one of the offending explanatory variables?

3. Other than a normal distribution, identify a continuous probability distribution that has zero mean and positive variance. (Hence, that ϵ_1 through ϵ_n are normally distributed is a stronger assumption than that they have zero mean and positive variance.)