CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 2

Nonnormal errors and robust regression

Scenario. Let us assume that the mean of a continuous response variable Y is linear in the values assumed by explanatory variables $X_1, \ldots, X_k$,

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

and that

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

The error terms $\epsilon_1, \ldots, \epsilon_n$ represent the differences between actual and expected responses. Let us also assume 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. Consequently, the probability of Y_i assuming a numerical value "close" to y_i is roughly proportional to

$$f(y_i|\mathbf{x}_i) := (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta_1 x_{1,i} - \cdots - \beta_k x_{k,i})^2\right]. \tag{3}$$

Because ϵ_1 through ϵ_n are independent, so are Y_i through Y_n. Thus, the probability of Y_1 through Y_n assuming numerical values "close" to y_1 through y_n is roughly proportional to

$$\prod_{i=1}^{n} f(y_i|\mathbf{x}_i) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta_1 x_{1,i} - \cdots - \beta_k x_{k,i})^2\right]. \tag{4}$$

Let us define the likelihood function

$$L(r_0, r_1, \ldots, r_k) := (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - r_0 - r_1x_{1,i} - \cdots - r_kx_{k,i})^2\right]. \tag{5}$$

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

$$SS(r_0, r_1, \ldots, r_k) := \sum_{i=1}^{n}(y_i - r_0 - r_1x_{1,i} - \cdots - r_kx_{k,i})^2. \tag{6}$$

Substituting (6) into (5) shows that

$$L(r_0, r_1, \ldots, r_k) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}SS(r_0, r_1, \ldots, r_k)\right]. \tag{7}$$

From (7) 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 redefine $f(y_i|\mathbf{x}_i)$ to accord with the distribution of ϵ_1 through ϵ_n, calculate $\Pi_{i=1}^n f(y_i|\mathbf{x_i})$, and use the expression for $\Pi_{i=1}^n f(y_i|\mathbf{x_i})$ to redefine the likelihood function $L(r_0, r_1, \ldots, r_k)$. Of course relation (7) 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. Other partial slope coefficient estimates may be interpreted similarly.

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

$$CRIT(r_0, r_1, \ldots, r_k) := \sum_{i=1}^{n} \rho\left[(y_i - r_0 - r_1 x_{1,i} - \cdots - r_k x_{k,i})/\sigma\right] \qquad (8)$$

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

$$\rho(t) := t^2 - t^4/c^2 + t^6/(3c^4) \quad \text{for} \quad |t| < c \qquad (9)$$

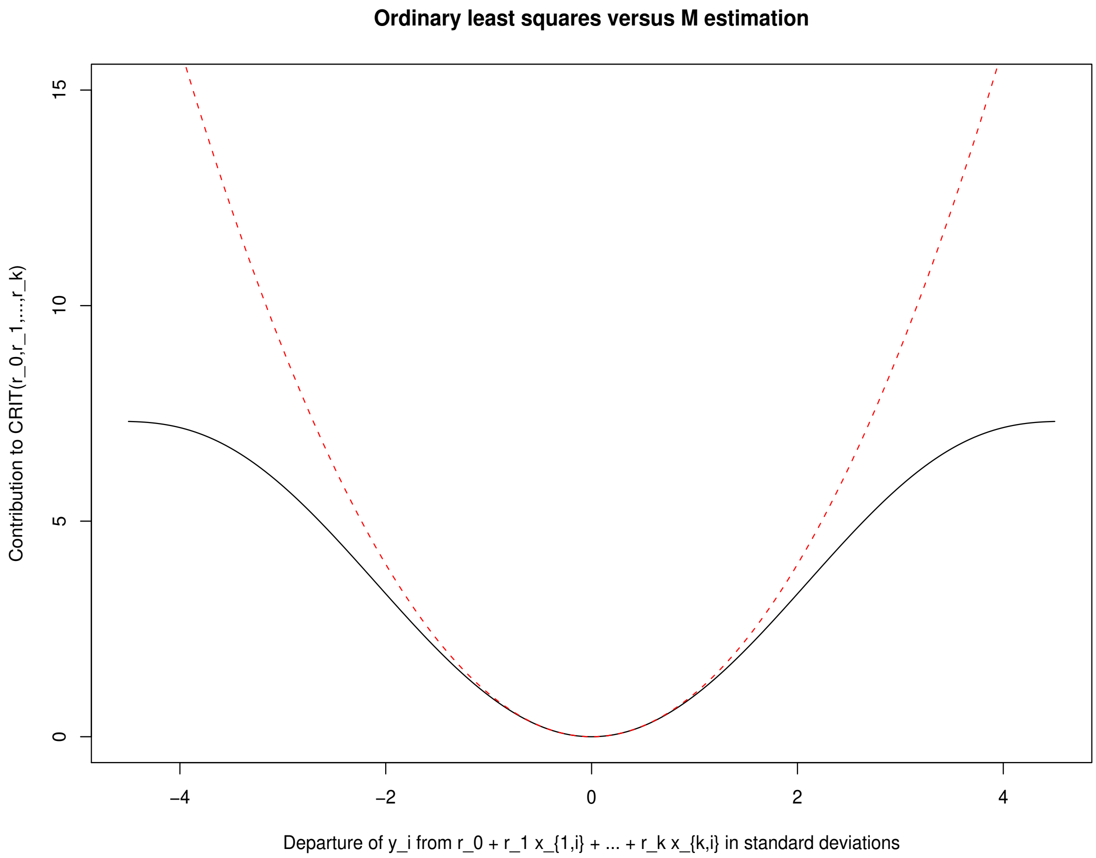
and

$$\rho(t) := c^2 - c^4/c^2 + c^6/(3c^4) = c^2/3 \quad \text{for} \quad |t| \geq c, \qquad (10)$$

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. Thus, ρ permits each observation to make only a limited contribution to $CRIT(r_0, r_1, \ldots, r_k)$, so that 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 $\alpha, \beta_1, \ldots, \beta_k$.

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. Refer to page 12 of {pulmresults.rtf}. The "Summary Statistics" box provides the 25^{th}, 50^{th}, and 75^{th} percentiles for the response and explanatory variables. Also provided are the mean, standard deviation, and "median absolute deviation". The median absolute deviation for data $z_1, \ldots, z_n$ is a measure of dispersion defined as the median of $|z_i - median(z_1, \ldots, z_n)|$. Note the resemblance to the definition of the variance, which is in essence the mean of $[z_i - mean(z_1, \ldots, z_n)]^2$.

The "Parameter Estimates" box 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.

Some other output of interest in {pulmresults.rtf} is on page 13. The $R - Square$ in M estimation is analogous to the R^2 from ordinary least squares and may be used as a summary of model goodness, with two caveats. First, the mathematical definition of $R - Square$ on page 13 is different from the mathematical definition of R^2 from ordinary least squares. Thus, comparing $R - Square = 0.6518$ to $R^2 = 0.7458$ on page 1 would be misleading. Second, as is true for R^2 in ordinary least squares, the $R - Square$ in M estimation is susceptible to inflation through the addition of extraneous explanatory variables. To choose among competing models, we are better off using analogues to the AIC and BIC called the AICR and BICR.

Nonlinear mean response and nonparametric regression

Flexibility in modeling the mean response. Suppose that we have a single continuous explanatory variable X. The linear regression model says that

$$m(x) = \alpha + \beta x \tag{11}$$

and

$$Y_i = m(x_i) + \epsilon_i. \tag{12}$$

In practice the mean response is almost never exactly linear in x. We usually live with the linear specification in (11) because we believe that it is not a serious departure from the truth.

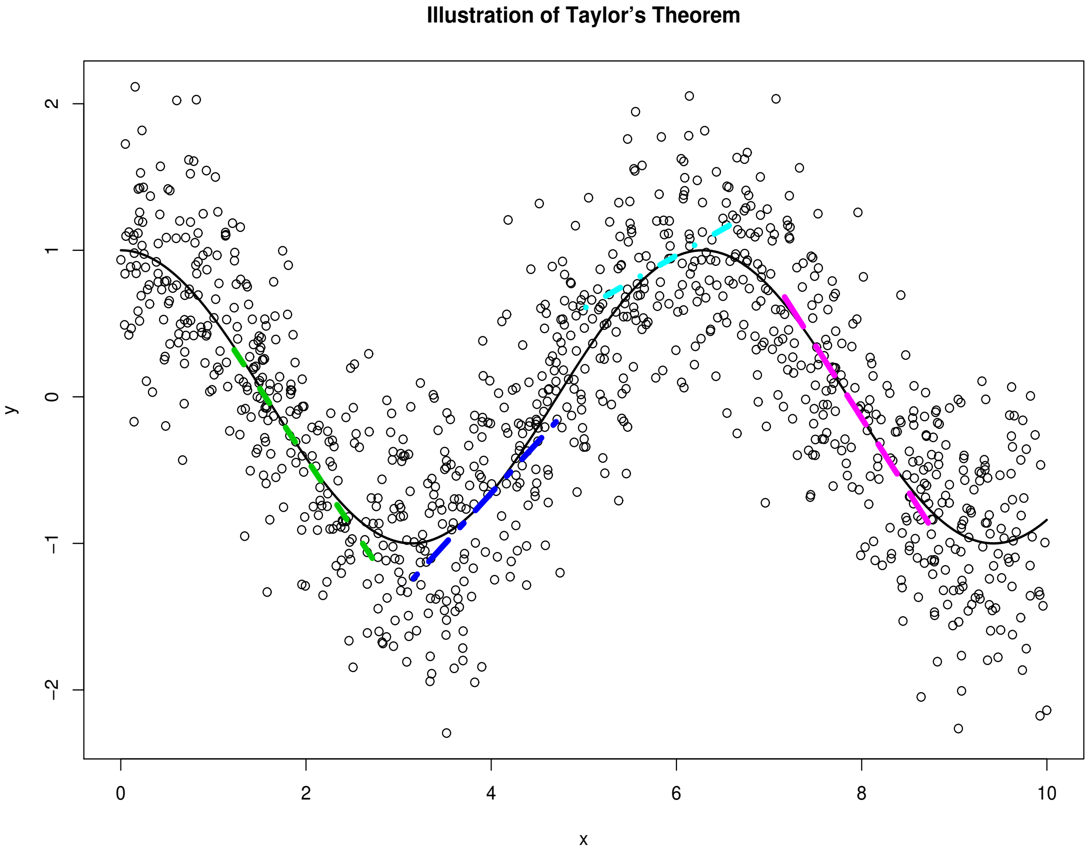
From your calculus course you may recall Taylor's Theorem, which says that any smooth function is well approximated by a straight line if you confine attention to a small enough interval. This idea is illustrated in the figure on the next page.

The linear specification is convenient for testing the straw-man null hypothesis of no association between X and Y (all we have to do is see whether $\widehat{\beta}$ differs significantly from zero) and is highly amenable to interpretation (increasing X by one unit increases the expected value of Y by β units).

However, there may be some occasions on which we desire more flexibility in modeling the mean response.

Example. The data set {BWeight.xls} provides birthweights (grams) and gestational ages (weeks) for 700 infants. I used ordinary least squares to estimate the expected value of birthweight as a linear function of gestational age.

The results of the ordinary least squares analysis are summarized on page 1 of {infresults.rtf}. The slope coefficient estimate of 125.2 tells us that, for



every one-week increase in gestational age, the expected birthweight is estimated to increase by 125.2 grams. Is this reasonable?

Page 2 of {infresults.rtf} shows a scatterplot of birthweights against gestational ages with the fitted regression line superimposed. The fitted regression line appears too high both for unusually low and for unusually high gestational ages. However, this is not a failure of ordinary least squares *per se*. There is no way to draw a straight line through these points without being way off for some gestational ages; whatever approach that we might employ to estimate α and β would fail. The real problem is the linear specification.

Nonparametric regression and LOESS smoothing. There exist several nonparametric regression methods. They are so called because they do not require us to express the mean response in terms of finitely many parameters like α and β. The specific method that we will consider today is called LOESS smoothing and is available in SAS through the LOESS procedure.

The basic idea is as follows. Let x_0 denote a specific value of X. For x near x_0, we have the approximate linear relationship

$$m(x) \approx \alpha_{x_0} + \beta_{x_0} x. \tag{13}$$

The x_0 subscripts on α_{x_0} and β_{x_0} remind us that the approximating linear relationship should depend on x_0. The figure on the preceding page illustrates: for x near $x_0 = 2$ we have (green line) $m(x) \approx 1.40 - 0.91x$, for x near $x_0 = 4$ we have (blue line) $m(x) \approx -3.68 + 0.76x$, for x near $x_0 = 6$ we have (turquoise line) $m(x) \approx -0.76 + 0.28x$, and for x near $x_0 = 8$ we have (purple line) $m(x) \approx 7.77 - 0.99x$.

So, given a specific value x_0, how can we estimate α_{x_0} and β_{x_0}? The answer is that we will employ weighted least squares; we will estimate α_{x_0} and β_{x_0} by minimizing

$$\sum_{i=1}^{n} w_{i;x_0}(y_i - r_{0;x_0} - r_{1;x_0} x_i)^2. \tag{14}$$

However, the weights $w_{i;x_0}$ will <u>not</u> be chosen in relation to error variances; although we have discarded assumption (11), we are still assuming (12) with errors that have a common variance. Rather, the weights $w_{i;x_0}$ will be chosen to place more emphasis on the observations with x_i close to x_0. Thus, observations with x_i close to 2 will do most of the work in estimating the green line on the preceding page, while observations with x_i close to 4 will do most of the work in estimating the blue line.

More specifically, we choose (or rely on SAS to choose for us) a number $q << n$ such that nonzero weights are assigned only to the q observations for which x_i is closest to x_0; all other observations receive zero weights. The nonzero weights are assigned as follows. Let $d_1 \leq d_2 \leq \cdots \leq d_q$ denote the distances from x_i to x_0 for the q observations. The observation corresponding to the distance d_j receives a weight of $[1-(d_j/d_q)^3]^3$. Hence, observations corresponding to smaller distances receive greater weights. The estimates of α_{x_0} and β_{x_0} are then determined using these weights.

The estimate of $m(x_0)$ resulting from this procedure is $\hat{\alpha}_{x_0} + \hat{\beta}_{x_0} x_0$, where $\hat{\alpha}_{x_0}$ and $\hat{\beta}_{x_0}$ denote the estimates of α_{x_0} and β_{x_0}. By varying x_0 through the entire range of X and repeating the above computations for each x_0, we can estimate the mean response over the entire range of X.

Example, continued. Refer to page 5 of {infresults.rtf}. There are many pieces of information here, and we will not be concerned with all of them. However, I will call your attention to a few items. The entry of 175 for "Points in Local Neighborhood" tells us what SAS used for q. The "Smoothing Parameter" equals 0.25071 and is basically q/n, apart from minor discrepancies due to rounding. Hence, in this example only 25% of the observations were used to estimate α_{x_0} and β_{x_0} for each x_0. The "AICC" is a measure that SAS uses to describe the quality of the fit associated with a given value for the "Smoothing Parameter"; lower is better.

Page 6 of {infresults.rtf} provides a visual display of the results. Although there is some "jiggling" (partly because gestational age is measured only to the nearest week), the smooth curve shown here is a more credible estimate of mean birthweight as a function of gestational age than the straight line shown on page 2 (or any other straight line).

Remark on LOESS smoothing. Although LOESS smoothing can be generalized to accommodate multiple explanatory variables, making sense of the results is difficult when there are more than two explanatory variables. This is because, in the absence of parameters, we must visualize the results to make sense of them; human beings can only see cross sections when there are more than three spatial dimensions.

Discussion questions

1. 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.)

2. If we were uncomfortable assuming that $m(x) = \alpha + \beta x$, what other option besides nonparametric regression might we consider? What would be the advantages and disadvantages of this option?

3. Nonparametric regression is, in my opinion, underutilized in medicine and public health. Other than lack of familiarity, why do you suppose that some researchers are reluctant to employ nonparametric regression?