CPH 636 – Spring 2009 – Dr. Charnigo

Lecture 5

Diagnostics in multiple linear regression

Augmented partial residual plots. For each explanatory variable in a multiple linear regression model, we may create an augmented partial residual plot. For concreteness I describe how this is done for X_1, but the same approach may be used for $X_2, X_3, \ldots, X_k$.

First, fit an auxiliary multiple linear regression model that includes the quadratic term X_1^2,

$$Y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \beta_{k+1} x_{1,i}^2 + \epsilon_i. \tag{1}$$

The augmented partial residual for subject i is defined by

$$APR_i := Y_i - \hat{\alpha} - \hat{\beta}_2 x_{2,i} - \cdots - \hat{\beta}_k x_{k,i}. \tag{2}$$

Next, fit a second auxiliary multiple linear regression model in which APR is the response variable and X_1, X_1^2 are the explanatory variables,

$$APR_i = \gamma_0 + \gamma_1 x_{1,i} + \gamma_2 x_{1,i}^2 + \delta_i. \tag{3}$$

Let $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$ denote the estimates of $\gamma_0, \gamma_1, \gamma_2$ in model (3).

Finally, plot the points $(x_{1,i}, APR_i)$ along with the parabola

$$\hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_1^2. \tag{4}$$

Examples of augmented partial residual plots are found on pages 2, 5, 8, and 11 of {RCLib.fevtrain27.rtf}. Let's look at page 2 specifically. The points marked "O" represent values of FEV plotted against values of AGE. The green line is obtained from a simple linear regression model with FEV as the response variable and AGE as the explanatory variable. The points marked “R” represent values of augmented partial residuals plotted against values of AGE. The red curve is the parabola identified in expression (4). Thus, the green line is the best-fitting line to the points marked “O”, while the red curve is the best-fitting parabola to the points marked “R”.

What should we look for in augmented partial residual plots?

- Is the red curve much different from a straight line? If so, and if this impression is not created by just one or two points marked “R”, then model (1) with the quadratic term X_1^2 may be more appropriate than the original multiple linear regression model that did not have any quadratic terms. The p-value following “quadratic” at the top of the page may be considered as well, subject to the caveat that we may get a misleadingly small p-value if the sample size is large. In any case, note that having a quadratic term does not make sense if the explanatory variable is dichotomous. **Discussion Question**: Why not?

- Do we get a qualitatively different impression from the red curve than from the green line? If so, then one of the other explanatory variables in the model is confounding the relationship between the response variable and the present explanatory variable. On page 11, the red curve is a straight line that is nearly horizontal, whereas the green line is clearly positively sloped. Here, we know that age is confounding the relationship between FEV and smoking. In particular, whatever positive association there is between FEV and smoking (green line) all but disappears after we control for age and the other variables in the multiple linear regression model (red line). **Discussion Question**: Why is the red curve on page 11 a straight line instead of a parabola?

Partial leverage plots. For each explanatory variable in a multiple linear regression model, we may create a partial leverage plot. For concreteness I describe how this is done for X_1, but the same approach may be used for $X_2, X_3, \ldots, X_k$.

First, fit an auxiliary multiple linear regression model from which X_1 is absent,

$$Y_i = \lambda_0 + \lambda_2 x_{2,i} + \cdots + \lambda_k x_{k,i} + \zeta_i, \tag{5}$$

and a second auxiliary multiple linear regression model in which X_1 is treated as the response variable,

$$X_{1,i} = \theta_0 + \theta_2 x_{2,i} + \cdots + \theta_k x_{k,i} + \eta_i. \tag{6}$$

Let $\hat{\zeta}_i$ and $\hat{\eta}_i$ denote the residuals from models (5) and (6). Also, for future reference, let R_1^2 denote the value of R^2 based on model (6).

Next, define the "partial regression" estimate $PR_{y,i} := \hat{\zeta}_i + \bar{y}$ and the "partial leverage" estimate $PL_{x_1,i} := \hat{\eta}_i + \bar{x}_1$.

Finally, fit a simple linear regression model with PR_y as the response variable and PL_{x_1} as the explanatory variable. Plot the points $(PL_{x_1,i}, PR_{y,i})$, the best-fitting line through these points (based on the aforementioned simple linear regression model), a 95% confidence band around the best-fitting line, and a horizontal line at height $\bar{y}$.

Examples of partial leverage plots are found on pages 3, 6, 9, and 12 of {RCLib.fevtrain27.rtf}. Let's look at page 3 specifically. The points marked "E" represent values of PR_{FEV} plotted against values of PL_{AGE}. The blue line is the best-fitting line through these points, and the 95% confidence band around the best-fitting line does not contain the horizontal line at height $\overline{FEV}$. This plot illustrates the nontrivial linear relationship between FEV and AGE after we control for the other variables in the multiple linear regression model. **Discussion Question**: What would the plot have looked like if there were no relationship between FEV and AGE?

Variance inflation factors and collinearity. Recall that R_1^2 was defined as the value of R^2 based on model (6). We refer to $1/(1 - R_1^2)$ as the variance inflation factor (VIF) for X_1. We can define VIFs analogously for $X_2, X_3, \ldots, X_k$. If the VIF for an explanatory variable exceeds 10 (and especially if it exceeds 100), we say that we have a "collinearity" (or a "multicollinearity") problem. Speaking roughly, collinearity means that some of the explanatory variables are so strongly correlated that precise estimation of partial slope coefficients is difficult. **Discussion Question**: Intuitively, why should strong correlations among $X_1, X_2, \ldots, X_k$ impede precise estimation of partial slope coefficients?

[[Speaking technically, a high VIF is really a diagnostic for rather than a definition of collinearity. A more proper definition of collinearity is that $X_1, X_2, \ldots, X_k$ move together in such a way that

$$c_1X_1 + c_2X_2 + \cdots + c_kX_k \approx c$$

for some constants $c_1, c_2, \ldots, c_k$ and c (other than $c_1 = \cdots = c_k = c = 0$). Moreover, if the VIF for X_1 equals 10, this means that the variance of the estimator of β_1 is 10 times larger than it would have been had $X_1, X_2, \ldots, X_k$ been mutually uncorrelated.]]

Options to address collinearity include removing the offending variable(s), reducing the original collection of explanatory variables to a smaller collection of variables without a collinearity problem, or performing ridge regression. [[One technique for "dimension reduction" is principal components, which we will visit in Lecture 11 this semester. Regarding ridge regression, see pages 7-11 of {www.richardcharnigo.net/CPH931F08/L1931F08.ps}.]] Note that VIFs are reported in the "Parameter Estimates" box on page 25 and that, in our example, there are no large VIFs.

VIF plots. For each explanatory variable in a multiple linear regression model, we may create a VIF plot. This is done by superimposing the "R" points from the augmented partial residual plot and the "E" points from the partial leverage plot.

Examples are found on pages 4, 7, 10, and 13 of {RCLib.fevtrain27.rtf}. If the "E" points are compressed (horizontally) relative to the "R" points, then there may be a collinearity problem. However, VIF plots are only useful for continuous explanatory variables. With a dichotomous variable, the "R" points are completely compressed because they fall on vertical lines determined by the possible values of the dichotomous variable.

Interaction. In an ordinary multiple linear regression model, the partial slope coefficient β_1 describes what happens to the mean response when x_1 increases by one unit while $x_2, \ldots, x_k$ are fixed. In particular, β_1 is a(n unknown) constant that does not depend on $x_2, \ldots, x_k$. Yet, there may be situations in which the effect of changing x_1 <u>should</u> depend on $x_2, \ldots, x_k$.

For instance, suppose that mean FEV is 0.3 liters lower for smokers aged 18 than for otherwise similar nonsmokers, while mean FEV is 0.1 liters lower for smokers aged 14 than for otherwise similar nonsmokers. In other words, suppose that the effect of smoking becomes more pronounced with increasing age (because older smokers will have been smoking for a longer time and thus will have harmed themselves more than younger smokers). An ordinary multiple linear regression model cannot capture this phenomenon, as the partial slope coefficient for SMOKE cannot equal both -0.3 and -0.1.

We can address this problem by including an <u>interaction term</u> that is the product of SMOKE and AGE. Then the effect of smoking is given by

$$\beta_{SMOKE} + \beta_{SMOKE \times AGE}\ AGE.$$

As you can verify, with $\beta_{SMOKE} = 0.60$ and $\beta_{SMOKE \times AGE} = -0.05$, the effect of smoking can equal -0.1 at age 14 and -0.3 at age 18.

Plots for detecting interaction are exemplified on pages 14 through 18 of {RCLib.fevtrain27.rtf}. In each such plot, a flat surface at height $\bar{y}$ suggests no interaction, while a highly irregular surface suggests important interaction. The p-values may be considered as well, subject to the caveat that we may get misleadingly small p-values if the sample size is large.

Other diagnostic output. Pages 19 and 20 may be disregarded. We will return to pages 21 and 22 later. Page 23 may be disregarded.

Continuing past the portion of the output reviewed in Lecture 4, we are now at page 31. The plot on page 31 is like the plot on page 27 except that the axes have been switched. Instead of plotting the observed responses y_i against the predicted values $\hat{y}_i$, we now plot the prediced values $\hat{y}_i$ against the observed responses y_i. This plot is useful for assessing the assumption of constant error variance. If the points become more (or less) spread out horizontally as we move along the line, then the assumption of constant error variance may be questioned. If the error variance seems to be an increasing function of the mean response, we may do well to analyze our data using a logarithmically transformed version of the response variable. [[**Discussion Question**: Can you make a convincing case for the logarithmic transformation, assuming that $\epsilon_i = E[Y|x_{1,i}, x_{2,i}, \ldots, x_{k,i}]\delta_i$, where the δ_i are independent and identically distributed normal random variables with mean 0 and variance $\tau^2 << 1$?]]

The plot on page 32 depicts the residuals, and the test referred to at the top of the page is a test for independence of the errors in the multiple linear regression model (or, more precisely, a test for zero correlation between successive residuals, which are proxies for successive errors). If the observations pertain to distinct and unrelated subjects, then the assumption of independent errors can be taken for granted without consultation of this plot.

The table on page 33 lists observations that are unusual and/or influential, as judged by studentized residuals or hat matrix values. The studentized residuals are rescaled versions of the ordinary residuals. The rescaling allows us to compare the studentized residuals to a standard normal distribution and, in particular, to declare as unusual those observations with studentized residuals greater than 2.5 or less than -2.5. The hat matrix values measure how influential the various observations are.

[[The hat matrix values are the diagonal elements of $\mathbf{H} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$, where $\mathbf{X}$ is the $n \times (k+1)$ matrix whose columns contain the values of the explanatory variables for the subjects in our sample. Since predicted and actual values of the response are related by the matrix equation $\hat{\mathbf{Y}} = \mathbf{HY}$, the hat matrix values quantify how much each y_i contributes to its own predicted value $\hat{y}_i$. In fact, the hat matrix values also enter into the computation of the studentized residuals, which are defined by $e_i/\sqrt{\text{Res MS }(1-h_i)}$.]]

Observations whose hat matrix values are more than twice the average hat matrix value are flagged. The hat matrix values are determined solely by the values of the explanatory variables (i.e., are unaffected by the values of the response variable).

The table on page 34 lists results for tests involving the residuals. If the errors are normally distributed, then the residuals will be approximately normally distributed as well. Hence, markedly nonnormal residuals suggest that the errors are not normally distributed. If the residuals have a strong right-skewed distribution, we may do well to analyze our data using a logarithmically transformed version of the response variable.

The plots on page 35 further examine the residuals. The first plot is a histogram of the residuals with a superimposed normal density curve; if the histogram does not align well with the normal density curve, then the assumption of normally distributed errors may be questioned. The second plot shows the residuals versus the fitted values; this plot is repetitive of the plot on page 27, in effect replacing a diagonal reference line with a horizontal reference line. The third plot is a normal probability plot of the studentized residuals; if there is a marked departure from a straight-line pattern, especially at the ends of the plot, then the assumption of normally distributed errors may be questioned. The fourth plot shows the values of the studentized residuals and the hat matrix values (after division by the average hat matrix value), helping us to visualize which observations are the most unusual and/or influential.

Criteria for model selection from the training subset

Page 21. The REGDIAG macro has identified the two one-variable models with the highest R^2. The first such model includes only HGT and has $R^2 = 0.7744$. The second such model includes only AGE and has $R^2 = 0.6340$. The REGDIAG macro has also identified the two two-variable models with the highest R^2 and the two three-variable models with the highest R^2.

For all of these models, as well as for the model with all four explanatory variables, the REGDIAG macro lists values of adjusted R^2, $C(p)$, AIC (Akaike Information Criterion), $RMSE$ (root mean square error), and SBC (Schwarz-Bayesian Information Criterion). I have already described adjusted R^2 and $RMSE$ in Lecture 4. The quantities $C(p)$, AIC, and SBC are referred to as model selection criteria because, even though they are computed from the training subset, they compensate for the optimism of the residual sum of squares and hence can be used to judge between competing models.

[[Formally, we define $C(p)$ as the quotient of the residual sum of squares for the present model divided by the residual mean square for the most complicated model under consideration, minus the sample size, plus twice the number of parameters in the present model. We define AIC as negative twice the log likelihood for the present model, plus twice the number of parameters in the present model. We define SBC similarly to AIC, except the multiplier of two is replaced by the natural logarithm of the sample size.]]

Some people use $C(p)$ to look for a preferred model by asking which model has the smallest $C(p)$, but many statisticians feel that a better approach is to find the simplest model for which $C(p)$ is less than or equal to the number of parameters (including the intercept). Note that $C(p)$ is defined relative to the variables supplied on the fifth and sixth lines of the REGDIAG macro, so you cannot compare $C(p)$ values across two separate runs of the REGDIAG macro.

Most people use AIC and SBC to look for a preferred model by asking which model has the smallest AIC and SBC. If AIC and SBC are negative numbers, this means that you are looking for the biggest negative numbers.

In our example, all three model selection criteria prefer the model with AGE, HGT, and SEX. However, the model selection criteria will not always agree. The SBC has the best statistical justification, but the justification is "asymptotic" (i.e., based on having a large sample size). For small samples (e.g., those with fewer than 200 observations), the AIC and $C(p)$ may be more credible. Another consideration is the strength of preference. For instance, if the AIC prefers one model over another by 30 points while the SBC prefers the opposite model by only 2 points, then perhaps the AIC should prevail.

Page 22. This is a plot of $C(p)/p$, where p is the number of parameters (including the intercept), against the number of explanatory variables for the models listed on page 21. The bubbles are sized in relation to $RMSE$. We see that only two models have $C(p)/p$ less than 1, of which the simpler model is the one with AGE, HGT, and SEX.

Model selection and evaluation from the validation and test subsets

Model selection. If we have a validation subset, we do not have to make a final decision about which model to choose based on the model selection criteria computed from the training subset. We can identify a number of promising candidate models using the model selection criteria and then employ the validation subset to make a final decision.

For a given candidate model, let $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denote the parameter estimates based on the training subset. Then, for subject i in the validation subset, define

$$\hat{y}_i := \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i},$$

where $x_{1,i}, \ldots, x_{k,i}$ are the values of $X_1, \ldots, X_k$ for subject i in the validation subset. Letting y_i denote the response for subject i in the validation subset, and letting $\bar{y}$ denote the mean response in the validation subset, we can then calculate

$$R^2_{valid} := 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y})^2}.$$

Our final decision about which model to choose is that we will choose whichever candidate model produces the highest R^2_{valid}. Note that R^2_{valid} has the same basic structure as R^2 calculated from the training subset, but now we have avoided the problem of using the same data twice (for model fitting and model selection).

Unfortunately, Fernandez tampered with his RSCORE macro between February 2006 and the present day, so the RSCORE macro can no longer be used for calculating R^2_{valid}. Thus, I have written code in {SASMacroInstr4.txt} for this purpose.

Model evaluation. In the end, whatever model we have chosen is formally evaluated as follows.

Once again, let $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denote the parameter estimates <u>based on the training subset</u>. Then, for subject i in the <u>test subset</u>, define

$$\hat{y}_i := \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_k x_{k,i},$$

where $x_{1,i}, \ldots, x_{k,i}$ are the values of $X_1, \ldots, X_k$ for subject i in the test subset. Letting y_i denote the response for subject i in the test subset, and letting $\bar{y}$ denote the mean response in the test subset, we can then calculate

$$R^2_{test} := 1 - \frac{\Sigma_i (y_i - \hat{y}_i)^2}{\Sigma_i (y_i - \bar{y})^2}.$$

We regard R^2_{test} as a formal evaluation of the chosen model's predictive capabilities. Indeed, R^2_{test} estimates the fraction of variability in the response that can be accounted for by the chosen model in the larger population from which the sample was drawn.