Lecture 23

Multiple Linear Regression II

The following are some more sketchy notes about multiple linear regression.

R^2 versus $R^2_{\text{adj.}}$

The meaning of R^2 is always the proportion of the variation explained by the model. It's value is given as

$$R^2 = \frac{\text{Sums of Squares due to the Model}}{\text{Sums of Squares Total}} = 1 - \frac{\text{Sums of Squares due to Error}}{\text{Sums of Squares Total}}$$

The value of R^2 naturally lies between 0 and 100%. However, as the number of variables (parameters) used to predict the response Y increases, the better Y will be predicted. For instance, if you use as many (independent) variables as you have data points, you can predict the responses perfectly. But your result would only be expected to work with your data set, not with any new one. The modified statistic, $R^2_{\text{adj.}}$, adjusts R^2 for the fact that you are using more predictors in your model. The formula is given as

$$R^2_{\text{adj.}} = 1 - \frac{\text{Mean Sums of Squares due to Error}}{\text{Mean Sums of Squares Total}} = 1 - \frac{\text{Sums of Squares due to Error}}{\text{Sums of Squares Total}} \frac{n-1}{n-k-1}$$

For simple linear regression when $k = 1$, the adjustment is minor and distracting because the adjusted parameter does not have the same, simple interpretation as the proportion of variance explained. However, for multiple predictors and for comparing models with varying numbers of predictors, $R^2_{\text{adj.}}$ becomes more useful. Adding predictors always mean increasing R^2 but does not always mean increasing $R^2_{\text{adj.}}$.

Diagnostics: Cook's Distances

The single more useful measure of problem data points is Cook's Distances. You should always store these and look for ones with value greater than 1 or those which are outliers to the other Cook's Distances. The formulation is the same as before:

Cook's Distance - USE THIS - this is a single measure that combines leverages and residuals to determine whether a point exerted its leverage and whether the residual was large enough to significantly influence the regression line. The formula is

$$D_i = \frac{1}{k+1}(\text{studentized residual}_i)^2 \left(\frac{h_i}{1-h_i}\right) = \sum_{j=1}^{n} \frac{(\hat{Y}_{j(i)} - \hat{Y}_j)^2}{(k+1)\hat{\sigma}^2}$$

where $\hat{Y}_{j(i)}$ is the fit for the j^{th} value if observation i is removed from the data set, $\hat{Y}_j$ is the corresponding fit when all the data is used, $k+1$ is the number of parameters in the model (2 for simple linear regression), and $\hat{\sigma}$ is the estimate for the residual standard deviation. Flag any point with $D_i > 1$ and also examine any point with an outlying Cook's distance value.

Diagnostics: Pure Error and Data Subsetting

Minitab will perform the same goodness-of-fit tests with multiple regression as it does with simple linear regression. If combinations of the predictors have repeats, then it can perform a pure error test where each combination of predictors could have its own mean (versus that mean being given by the best fitting linear model.) The set up is similar to but more complicated than the set up for simple linear regression. For two predictor variables, you can think of it in the following way:

Repeats by X_1 and X_2:	$X_{1,1}, X_{2,1}$	$X_{1,2}, X_{2,2}$	$\cdots$	$X_{1,I}, X_{2,I}$
Full Model:	μ_1	μ_2	$\cdots$	μ_I
Reduced Model:	$\beta_0 + \beta_1 X_{1,1} + \beta_2 X_{2,1}$	$\beta_0 + \beta_1 X_{1,2} + \beta_2 X_{2,2}$	$\cdots$	$\beta_0 + \beta_1 X_{1,I} + \beta_2 X_{2,I}$

To compare nested models, one constructs a nested-models F-test. The (mean) difference in the sums of squares explained by the two models goes in the numerator, and the (mean) sums of squares that is the best/smallest (from the full separate means model, of course) goes in the denominator:

$$\begin{aligned} F &= \frac{\text{SSE}_{\text{Extra}}/\text{d.f.}_{\text{Extra}}}{\text{SSE}_{\text{Full}}/\text{d.f.}_{\text{Full}}} \\ &= \frac{\left(\text{SSE}_{\text{Reduced}} - \text{SSE}_{\text{Full}}\right) / \left(\text{difference in the number of parameters between the models}\right)}{\text{SSE}_{\text{Full}}/\text{d.f.}_{\text{Full}}} \end{aligned}$$

For data subsetting, Minitab does multiple tests and does a Bonferroni correction on their p-values (multiplying the raw p-values by the number of tests conducted.) For each predictor variable, Minitab looks for whether 2 lines, one on each side of the predictor variable mean, rather than one line fits the data better. It also looks for curvature in the outer values for each variable by fitting a line to the middle portion and seeing if that line is a good fit for the outer regions. It also looks for possible interactions between variables (something we will discuss later). Minitab then tells you which tests were problematic in order to help you fix the model.

Diagnostics: Serial Correlation in the Residuals

Where relevant (when data are known to be collected in order, especially time-order, or when calibration of equipment might be an issue and the order is known), Minitab will report the Durbin-Watson statistic. The original paper must be consulted to determine the significance. In their tables, across the top, k is the number of predictor variables in the model.

The statistic is:

$$d = \frac{\sum_{t=2}^{n} \left(\hat{\epsilon}_t - \hat{\epsilon}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}$$

The expected value of d when there is no correlation amongst the residuals is 2. If d is substantially lower than 2, that is a sign of positive serial correlation - the most likely possibility and the one with the most serious consequences. If you do not correct for positive serial correlation, you are stating more confidence in your results than you really have. This is called anti-conservatism in statistics and statisticians regard it as a serious error. If d is substantially higher than 2, that is a sign of negative serial correlation, but this is a rarer and less of a problem.

Diagnostics: Variance Inflation Factors

When one predictor variable is itself predicted well from the other predictor variables, the matrix on which the regression estimates rely, $(\mathbf{X}^T\mathbf{X})$ becomes singular (non-invertible). This causes the estimates of the coefficients to become unstable. Minitab reports this problem as a "Variance Inflation Factor." Minitab determines how well any one predictor variables, say X_1, is predicted from the other predictor variables, X_2, X_3, ..., X_k and records the $R^2(X_1)$ value from this regression. In general, Minitab records a value, $R^2(X_i)$ for each predictor variable X_i. The Variance Inflation Factor (VIF) for each variable is then reported as

$$\text{VIF for variable } X_i = \frac{1}{1 - R^2(X_i)}$$

Values of VIF over 10 can be causes of concern. They mean if that variable suddenly differs from its expected value from the other predictor variables, your estimate of the response may be way off (we'll talk about balancing a plane on a line to demonstrate this issue in class.)

In some cases, this is NOT a problem. For instance, X and X^2 can be highly correlation (for instance, as X varies from 100 to 200). However, X^2 will never vary from the square of X and X will never vary from the square root of X^2, so there is no problem with these two variables being correlated.

Example

The data we will examine in class comes originally from Brofeldt and was contributed to the Journal of Statistical Education Data Archive by Juha Puranen. The data is the weight, length, width, and height of various fish caught in a particular lake in Finland. For now, we will concentrate on the data from Perch (Perca fluviatilis). We will try to predict the weight of the fish from its other measurements. Running a multiple regression with no diagnostics gives no hint of trouble (strong R^2) but looking at the data and using our diagnostic tools indicates that there are problems with a simple linear model. We will find problem points, make good transformations, and try to find the "right model" in class.

References and Readings

[1] Pekka Brofeldt. Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi. T.H.Jaervi: Finlands Fiskeriet Band 4, Meddelanden utgivna av fiskerifoereningen i Finland. Helsingfors. 1917.

Exercises for Lecture 23

1. –

2. –