Residual Analysis and Model Refinement

To determine the adequacy of a regression model with respect to regression assumptions 1-4, check that the residuals have properties I-IV **in order**. In addition, check V and, to the extent possible, assess VI (regression assumption 5).

I. $\mu_{e_i} = 0$. (Residuals have mean 0.) If the residuals satisfy this property, then the regression model is adequately approximating the mean function

$$\mu_{Y|x_1,\ldots,x_k} = f(x_1, \ldots, x_k)$$

To check that the residuals satisfy this property, construct the following residual plots:

1. Plot the residuals vs. every predictor of interest, and
2. If you have a multiple regression model (model with more than one predictor variable), plot the residuals vs. the predicted values $\hat{Y}_i$.

Check that there are no systematic departures of the residuals from 0. If this assumption is met, you have adequately modeled the deterministic relationship between y and the predictor variables.

II. $\sigma_e^2 \approx \sigma^2$. (Homoscedasticity) To check this property, use the following graphs:

1. Use the plots described in I.
2. IF the residuals satisfy I, redo the plots in (1) using the absolute values $|e_i|$ instead of the raw residuals, e_i.
3. IF additional assumptions are met, formal procedures can be used to test the equality of the variances.

Check that there are no systematic trends in the scatter/variability of the residuals.

III. $\text{corr}(e_i, e_j) = 0,\ i \neq j$. (Residuals uncorrelated.) IF the residuals satisfy I, check for correlation using

1. Plot e_j vs. e_{j-1}, i.e., create a *lag plot*, after sorting the residuals with respect to the order in which the data was generated or with respect to some other order of interest, e.g., location. Once the residuals are in proper order, you can use the `lagplot` macro to generate the plot. The plot points should be randomly scattered in the x-y plane.
2. IF the residual satisfy II, you can use the Durbin-Watson test and other formal procedures not discussed in our text.

IV. (Residuals are normally distributed.) IF the residuals satisfy conditions I, II, and III, you can check this requirement using a good test of normality. Minitab has two good tests, the Ryan-Joiner procedure and the Anderson-Darling test; see `Stat -> Basic Statistics -> Normality Test`. Do not use the Kolmogorov-Smirnov Test; it is inferior to the aforementioned tests.

V. (Outliers.) If a *deleted t* residual (*studentized deleted residual* has a magnitude exceeding 3, then it may be an outlier. If it exceeds 4, then it most certainly is an outlier. Deleted t residuals can be requested from Minitab using the same regression menu as for regular residuals.

VI. (Predictor variables measured with negligible error.) This assumption cannot usually be checked using the regression data alone; additional data/information is typically needed to determine if this assumption is met or is reasonable given the data collection/generating process.

Model Refinement and Transformations

In the following we describe how to refine the model so that the residuals satisfy properties I-IV.

I. Suppose $\mu_{e_i} \neq 0$, i.e., the residuals e_i exhibit systematic trends about 0. This means that the deterministic component of the regression model is inadequate. Recall that regression analysis assumes that each y-value is the sum of a deterministic mean function of the predictors $x_1 - x_k$, $f(x_1, \ldots, x_k)$, plus zero mean random error, e. Our goal is to find a linear function $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ which adequately approximates the mean function $f(x_1, \ldots, x_k)$:

$$f(x_1, \ldots, x_k) \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

There are two basic ways to improve this approximation:

i. Add higher order terms, i.e., use a higher order polynomial model. Be sure to center and/or standardize your predictor variables to minimize variance inflation.

ii. Transform variables:

1. More complex transformations of the predictor variables, e.g., $\log(x_1)$, $\sin(x_2)$, etc. Sometimes you can achieve a simple mean function through the use of transformations, particularly if you only have one predictor variable. See pages 129-132 of your book.
2. Transformation of the response variable, e.g., $\log(y)$.

Developing an adequate model for f is an iterative and potentially tedious process.

II. Perhaps the most difficult requirement to achieve is homoscedasticity. If the residuals for your current model don't satisfy this condition, there are two possible remedies:

i. Use a more complex model for the mean function, $f(x_1, \ldots, x_k)$. Occasionally, by explaining more of the variation in the response variable, you can get to a point where the residual variances are approximately constant. However, you usually need the following approach.

ii. Transform y. There are various possible transformations, including the Box-Cox family of transformations; see pages 132-137. The Minitab macro `BCtrans` can help you determine the best Box-Cox transformation (λ value) to use. The convention when using the Box-Cox transformation is to use the best integer multiple of 1/2. Note that $\lambda = 0$ indicates you should log transform your data. If your model for the mean function $f(x_1, \ldots, x_k)$ was adequate prior to transforming y, you will probably need to modify it after transforming y, hence the phrase, "transform both sides."

III. If your residuals are not independent, there are two remedies:

i. In some cases, adding additional predictor variables will eliminate correlation among the residuals.

ii. Typically, you can't eliminate correlation among the residuals so you have to use special procedures/models - time series models and procedures - which explicitly incorporate residual correlation into the model. These methods are beyond the scope of this course.

IV. Nonnormality of the residuals indicates that the random errors are not normally distributed. Depending on the goal(s) of your regression analysis, this may or may not be a serious problem. If you need to correct lack of normality - for example, so you can construct prediction intervals - there is only one remedy: transform the response variable y. For example, if your reponse variable does not take on negative values, you can use a Box-Cox transformation. Use the Minitab macro `BCtrans` to determine initial estimates of λ.

Macros for Model Assessment and Refinement

`lagplot:` This macro takes a column of ordered residuals and creates the corresponding lag plot of e_j vs. e_{j-1}. This plot is used to assess regression assumption 3: independent and/or uncorrelated random errors. Here is the usage:

```
%lagplot C
```

where `C` is the column containing the residuals.

`BCtrans:` See next page

`BCtrans:` This macro takes the column containing the response variable (`C0`) and the columns containing the predictor variables (`C1-CK`) and determines the best power λ to use when Box-Cox transforming the response variable. Here is the usage:

```
%BCtrans C0 C1-CK.
```

There are several subcommands which may prove useful in the course:

range K K Use this subcommand to specify the range for λ in the PRESS plot. Specifying a range for λ overrides the default range that covers the 95% confidence interval.

influence Use this subcommand to display an index plot showing the influence of individual cases on the likelihood estimate of λ.

bcstore C C Use this subcommand to store the log-likelihood and corresponding λ values used in creating the log-likelihood plot. Two columns must be specified. The first column will contain the log-likelihood values and the second column will contain the λ values.

infstore C C Use this subcommand to store the approximate likelihood distances and case numbers used in creating the index plot. Two columns must be specified. The first column will contain the PRESS statistic values and the second column will contain the λ values.

presstore C C Use this subcommand to store the PRESS statistic and corresponding λ values used in creating the PRESS plot. Two columns must be specified. The first column will contain the PRESS statistic values and the second will contain the λ values.