MTH 4230 Spring 2021
Module 5 Notes and Exercises

Simple Linear Regression Diagnostics

Recall that the simple linear regression *model equation* for the value of a response variable Y as a function of an independent variable x is given by

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is a normally-distributed random variable with mean 0 and variance σ^2.

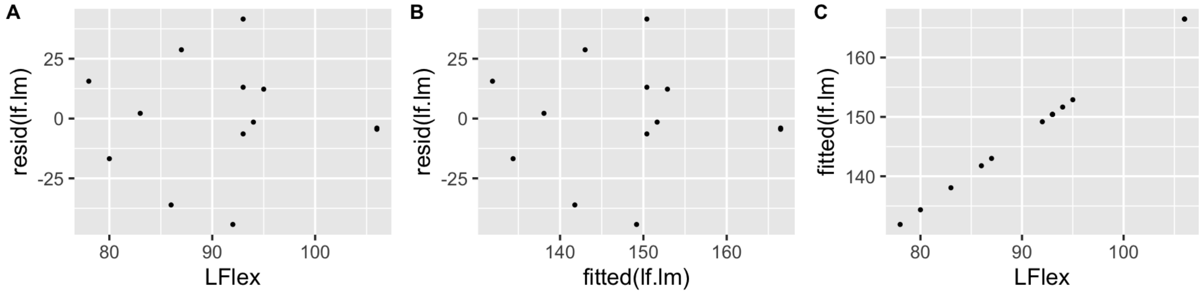
Error Term Diagnostics. First we consider diagnostic methods for verifying the following model assumptions associated with the error terms, ϵ_i, for $i = 1, 2, \ldots, n$:

(1) The error terms are independent random variables
(2) The error terms are sampled from a normal distribution with mean zero
(3) The variance of each error term is σ^2

More specifically, the last assumption is that the error terms have constant variance that is not dependent on predictor levels.

Each of these three specific model assumptions can be tested using diagnostic plots or inferential methods developed for diagnostics. Graphical analysis requires intuition but is wider in scope in terms of evaluating model assumptions. Large data sets in particular lend themselves to graphical analysis. Hypothesis testing is more useful for decisions that are narrow in scope and based on relatively small random samples.

Residuals plotted against fitted or predictor values. We use the "punting " data set from textbook to exemplify some of the graphs covered here. In the case of a single predictor variable, note that the residuals plotted against the predictor variable (Plot A) is simply a rescaled version of the residuals plotted against the fitted values (Plot B) since there is a one-to-one linear dependence between the predictors and the fitted values (Plot C).



Here 'lf.lm' is the linear model in R (see code).

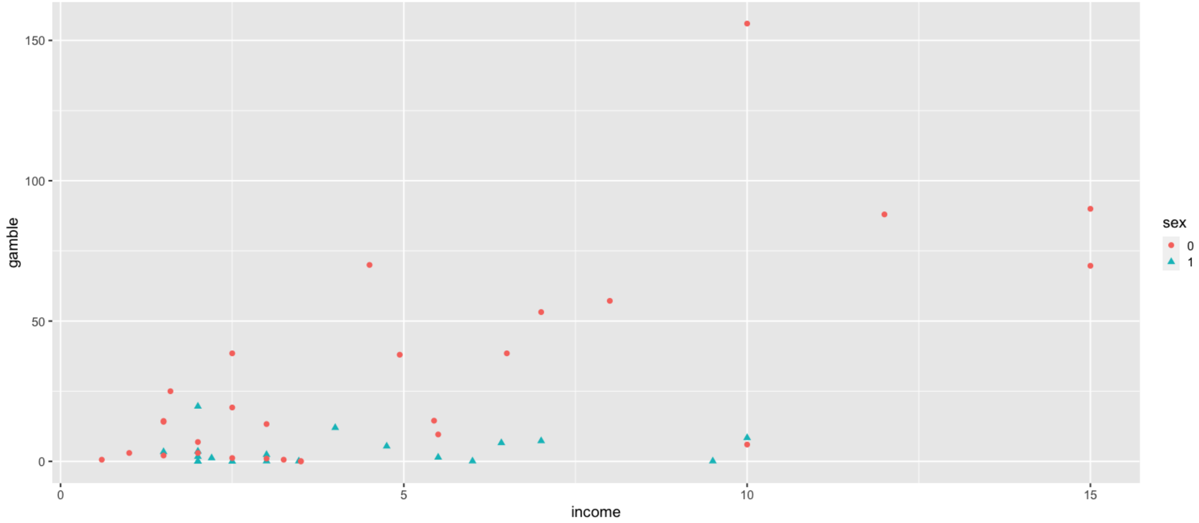
Both plots of residuals versus fitted values and residuals versus predictor values are used crucially in assessing model assumptions (1) and (3) from above. When patterns are observed in the residuals, this is an indication that error terms are not independent. When the spread of the error appears more pronounced for certain subsets of predictors, this is an indication that the constant variance assumption may be incorrect.

Patterns in the residual error plots may also indicate that the linear structure of the relationship is incorrect in the first place, and the model is not linear. This is an assumption violation that is usually just as easily recognized in the scatterplot of the data, before one proceeds to residual error analysis.

When multiple predictor are incorporated into the model, residuals plotted against the fitted values incorporate information from all of the predictors used, so they are more appropriate for assessing the combined effects of all predictors in the linear models. Plots of residuals against each predictor and so-called *added variable* or *partial regression* plots covered in Module 9 are also useful for multiple regression diagnostics covered later in the course.

We address further residual analysis for multiple regression in subsequent modules.

Residuals plots for possible omissions of important variables. If other variable measurements are available but left out of a linear model with one predictor, correlations in the error can be revealed by visualizing the values of another, usually binary, predictor using different colors or shapes. For example, if we compare income and amount gambled in the 'teengamb' data set, we see a stark contrast with respect to the additional variable 'teengamb$sex' as can be seen in the figure below.



Histograms and normal $q-q$ plots of the residual error. These methods have already been covered in the Module 3 worksheet and assessing normality is something covered in prerequisite courses.

Hypothesis testing methods for goodness-of-fit of a data set to a particular distribution are covered in prerequisite courses, but these tests are usually too rigid for assessing the normality of the residual error.

In general, detecting violations of the normal distribution assumption is more difficult than detecting violations of the other assumptions.

Project 3 Exercises (due 2/19/21)

Exercise 1 (4 points): This is essentially Exercise 8 from Chapter 6 of *Linear Models in R*. This Exercise uses the 'divusa' data from the 'faraway' package in *R*.

(1) Construct scatterplots of the response variable 'divorce' against all other predictor variables except for 'year.' Use these scatterplots to assess the criterion for a regression line in each case.
(2) Construct plots of the residual errors against the fitted value of linear models using response variable 'divorce' and each predictor variable except for 'year.' Use these plots to assess the assumption of independent error terms and the constant variance assumption.
(3) Indicate which predictor variables are most suspect in terms of not satisfying the assumptions of the simple linear regression model linking that prediction variable to 'divorce.'

Exercise 2 (8 points): This is similar to Exercise 2 from Chapter 6 of *Linear Models in R*. This Exercise uses the 'teengamb' data from the 'faraway' package in *R*.

(1) Fit a model with 'gamble' as the response and the 'income ' as the predictor. Complete each of the following parts:
    (a) Check the constant variance assumption for the errors.
    (b) Check the normality assumption.
    (c) Check for outliers.
    (d) Check the structure of the relationship between the predictors and the response.
(2) Complete the previous part, but only for males in the original data set.
(3) Complete the previous part, but only for females in the original data set.

Exercise 3 (8 points): Duplicate Exercise 2 and include a residual plot for the omitted variable for **one** of the following cases:

- For the 'swiss' data, fit a model with 'Fertility' and another non-binary variable as the predictor. Use another binary variable as the omitted variable.
- For the 'happy' data, fit a model with 'happy' and another non-binary variable as the predictor. Use another binary variable as the omitted variable.
- For the 'tvdoctor' data, fit a model with 'life' and another non-binary variable as the predictor. Use another binary variable as the omitted variable.

References

[1] Cornillon, Pierre-Andre. *R for Statistics, 1rst ed.*. Chapman and Hall, (2012).

[2] Draper, N.R., Smith, H. *Applied regression analysis 3rd ed.* Wiley (1998)

[3] Faraway, J. *Linear Models with R, 2nd ed.*. Chapman and Hall, (2014).

[4] Faraway, J. (April 27, 2018) *Linear Models with R* translated to Python. Blog post. https://julianfaraway.github.io/post/linear-models-with-r-translated-to-python/

[5] Fahrmeir, Kneib, Lang, Marx , *Regression.* Springer-Verlag Berlin Heidelberg (2013).

[6] Kutner, Nachtsheim, Neter, Li *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin (2005).