MTH 4230 Spring 2021
Module 6 Notes and Exercises

Multiple Linear Regression with Normal Error

In Module 2, we introduced the matrix form of the generalized linear model; namely,

$$\mathbf{Y} = \mathbf{X}\,\beta + \epsilon$$

where $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{pmatrix}$

is a dimension $n \times p$ matrix of predictor variable observations,

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ is a vector of parameters and,

$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$ is a dimension $n \times 1$ vector of independent $N(0, \sigma^2)$ random variables.

The response variable $\mathbf{Y}$ on the left hand side of the equation is, consequently, a $n \times 1$ vector of normally distributed random variables with means given by,

$$\mu_i := E[Y_i] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

We can write this in the concise matrix from $E\left[\mathbf{Y}\right] = \mathbf{X}\beta$.

The capitol X_{ij} notation and capitol Y notation above is used to emphasize that these are, in general, random variables in the conceptual model formalized *before experimental or observational data is collected*. The response, Y, is aways a random variable because the error term, ϵ, is random. The predictors are also usually random in real world experiments except in the context of controlled experiments where researchers have complete control of all of the predictor variable values used in the experiment.

Note that, in a real world application, the rows of the matrix $\mathbf{X}$ (the parts proceeding the '1' values) can be interpreted as individual predictor observations, $(x_{j1}, \ldots, x_{j,p-1})$, associated with real world data. We use $i = 1, \ldots, n$ to iterate the real world sample and we use $j = 0, \ldots, p-1$ to iterate the parameter values β_j. The textbook *Linear Models in R* uses the lower case sample notation (e.g. Section 2.2 Matrix Representation).

Model Parameters. The model equation is a $p+1$-parameter family of statistical models; this includes $p-1$ *partial regression coefficients* denoted by β_j for $j = 1, \ldots, p-1$, the intercept β_0, and the variance σ^2. Each choice of $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and σ^2, corresponds to a specific member of the family. The model is a *general linear regression model with normal error.* Parameters can be interpreted as follows.

- The partial regression coefficient, β_j, for $j = 1, \ldots, p-1$, of the least squares regression line for the bivariate population: this parameter can be interpreted as the average change in the response with respect to the jth predictor variable x_j when all other predictors remain constant. If there is no way to vary the jth predictor while keeping the other predictor values constant (correlated predictors), this value may not yield practical interpretation.
- The y-intercept is the intercept with the line $\mathbf{x} = (x_1, \ldots, x_{p-1}) = (0, \ldots, 0) = \mathbf{0}$ in this context. We interpret β_0 as such. This parameter usually does not have direct practical interpretation, as the predictor value $\mathbf{0}$ is usually not within the range of the experimental data.
- The variance, σ^2, of the residual errors associated with the least squares regression line for the bivariate population: this parameter yields the practical interpretation that 68/95/99% of possible values of Y given that $\mathbf{x}^* = \left(x_1^*, \ldots, x_{p-1}^*\right)$ should be within one/two/three standard deviations, σ, of the number $\mu_{Y|\mathbf{x}^*} = \mathbf{x}^*\ \beta$.

Sample Statistics. The multivariate simple random sample is denoted by

$$\{(x_{i1}, \ldots, x_{i,p-1}, y_i)\}_{i=1}^n$$

For observational studies (with a conceptually finite population), the population notation is the same except that N replaces n. Here we use the following notations for the finite sequence $(a_1, a_2, \ldots, a_n)$:

$$\{a_j\}_{j=1}^n = \{a_1, a_2, \ldots, a_n\} \quad \text{and} \quad (a_j)_{j=1}^n = (a_1, a_2, \ldots, a_n)$$

The first, with set brackets, can be used when order is not important. The latter, with parenthetical vector brackets, should be used if order is important.

In observational studies, researchers may not be able to control predictor values, and measurement error of predictors is a serious concern. In such cases, analytic methods for determining the sensitivity of the model to measurement error have been developed.

In controlled experiments, predictor variable values can be adjusted to the researchers specifications. Typically, predictor variable values equally partitioned over the domain of interest are preferable, but this depends on the goals of the analysis and correlations between the predictor variables.

For the sample statistic matrix computations given below, we use the sample notation,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix}, \text{ and } \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

where $e_i = \hat{y}_i - y_i$ for $i = 1, \ldots n$.

Given an estimate $\mathbf{b} = (b_0, \dots b_{p-1})$ of β, the fitted values associated with the estimate are computed using the matrix multiplication given by,

$$\hat{\mathbf{Y}} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{pmatrix} = \mathbf{Xb}$$

This allows us to compute residual errors associated with the observations as follows,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

We define $\mathbf{I}$ to be the $n \times n$ identity matrix and $\mathbf{J}$ to be the $n \times n$ matrix with all entries equal to 1. Using this notation the following sample statistics can be defined. These statistics are used to calculate estimates and apply inferential methods. Here, $\mathbf{X'}$ represents the transpose of the matrix $\mathbf{X}$ and an inverse power, $(\ )^{-1}$, of a matrix is the matrix inverse.

$$\begin{aligned}
\mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
\hat{\mathbf{Y}} &= \mathbf{HY} \\
\mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\
SSR &= \mathbf{Y}'(\mathbf{H} - \frac{1}{n}\mathbf{J})\mathbf{Y} \\
SSE &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \\
SST &= \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y} = SSR + SSE \\
MSR &= \frac{SSR}{p-1} \qquad MSE = \frac{SSE}{n-p} \\
\mathbf{s^2} &= MSE(\mathbf{I} - \mathbf{H})
\end{aligned}$$

The last item, $\mathbf{s}$, is called the ‘variance-covariance’ matrix because it estimates the variance and covariance between error terms. Model assumptions imply that the population variance-covariance matrix is $\boldsymbol{\sigma^2} = \sigma^2\mathbf{I}$.

The *coefficient of multiple determination* measures the proportion of total variance in the response variable explained by the regression model. The formula for the coefficient of multiple determination is given by

$$R^2 = 1 - \frac{SSE}{SST}$$

The square root of R^2 is called the *coefficient of multiple correlation.*

Inference: Point Estimates. A **point estimate** of a population parameter is, in general, any statistic calculated from a sample and used to estimate a population parameter; this is discussed in the Module 3 notes.

In our case, the population distribution assumes a parametrized form. We use estimates, called **maximum likelihood estimators**, that maximize the likelihood of the observed data. In this case, the likelihood function is written in terms of the parameters β and σ as follows:

$$L(\beta, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \cdots - \beta_{p-1} X_{i,p-1})^2 \right)$$

The parameter values that maximize this function are the maximum likelihood estimators and can be shown to also be minimum variance unbiased estimators of the population parameters. Point estimates of the parameters are,

$$\begin{aligned} \hat{\beta} &= \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\boldsymbol{\sigma^2}} &= \mathbf{s^2} = MSE(\mathbf{I} - \mathbf{H}) \end{aligned}$$

The real number σ^2 (not the bold matrix $\boldsymbol{\sigma^2}$) is estimated as $\hat{\sigma^2} = MSE$. It can also be shown that $E(MSR) \geq \sigma^2$ with equality holding if and only if β is identically zero (equal to a zero vector).

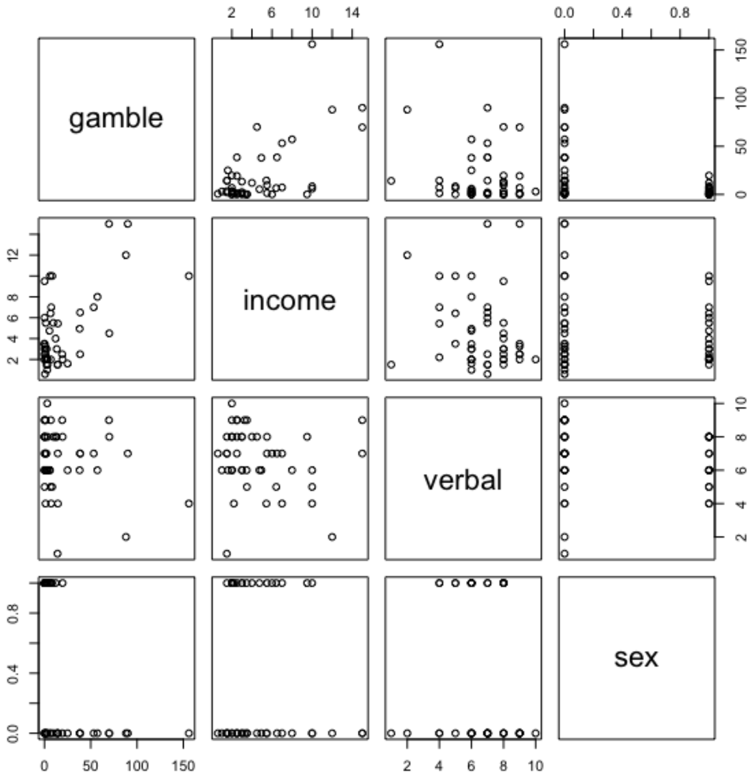
Point estimates for mean response variables given specific predictor values can be denoted $\mu_{Y|\mathbf{x}=\mathbf{x}^*}$ (here $\mathbf{x}$ is a vector of length $p-1$).

Model Diagnostics. When using a mathematical or statistical model, it is important to perform diagnostic tests to determine if the model is appropriate. The objective of **statistical model diagnostics** is to determine if the data under consideration fits the model assumptions. For the simple linear regression model with multitiple predictors and normally distributed error, the following diagnostic techniques are routine.

The *Real World Considerations* subsection of the Module 3 notes also applies here, and most of the techniques used to diagnose simple linear regression with one predictor variable assumptions apply or can be generalized to accommodate multiple predictors.

Scatterplot and Correlation Matrix. A matrix of scatterplots and a correlation coefficient matrix that give scatterplots between any two of the predictor variables and/or response variable are commonly used to both diagnose model assumptions and understand the extent to which variables are correlated with each other. Correlated predictor variables do not violate model assumptions and not specific structure is assumed that relates the predictor variables. The response variable should satisfy the criterion for a linear regression when plotted against any individual response variable.

The *R* function ‘pairs’ constructs the desired scatterplot matrix; for example, the variables ‘gamble’, ‘income’, ‘verbal’, and ‘sex’, from the ‘teengamb’ data set are plotted below.



Residual Error Normality Plots. The residual errors should have a normal distribution, so histograms and normal probability plots are applied as before to assess normality.

Residual Error Plotted Against Fitted Values and Predictors. Residual errors are assumed to be independent random variables. By plotting predictors or the fitted values against residual error values, (x_i, ϵ_i), we can assess this independence assumption by looking for any patterns that indicate a correlation of residual errors, which violates model assumptions.

These plots are also used to assess nonlinearity of the regression functions. If residual error is scattered but appears to follow a trend that is not centered about the predictor axis, this can be a sign of a nonlinear relationship.

Further Residual Error Analysis. Residual errors provide a rich resource for model assumption analysis. In addition to the preceding two items, absolute (or squared) residuals are often plotted against the predictor to assess constancy of error variance. Other techniques included investigating omitted factors using residual analysis and plots of residual errors vs. fitted values, time, or another sequence.

Hypothesis Tests for Comparing Models. The preceding diagnostic techniques listed above all involve some subjectivity. Model assumptions can also be tested using objective, programmable inferential hypothesis testing techniques. When multiple predictors are potentially included in a model, different models and combinations of predictors can be tested against each other use hypothesis testing procedures. These procedures are covered in Module 10.

Project 5 Exercises

Exercise 1 (4 points): Complete Textbook Exercise 4 from Chapter 2 (page 30), stated as follows: The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2. Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.

Exercise 2 (6 points): Find five different sized vessels such as a cup or a pot or a vase in your house. Measure the weight (or width) of each vessel, the height of each vessel, and the volume of water each vessel holds as accurately as possible. Organize this data in a data frame in R, identifying volume as the response and the other two measurements as predictors.

(1) Identify $\mathbf{X}, \mathbf{Y}, \hat{\mathbf{Y}}$, and $\mathbf{e}$ in this context; include the R code you use.
(2) Identify $\mathbf{H}, SSR, SSE, SST, MSR, MSE, \mathbf{s^2}$ and R^2 in this context; include the R code you use.
(3) Interpret R^2 in terms of the real world variables.

Exercise 3 (10 points): Complete Textbook Exercise 7 from Chapter 2 (page 31), stated as follows: An experiment was conducted to determine the effect of four factors on the resistivity of a semiconductor wafer. The data is found in wafer where each of the four factors is coded as $-$ or $+$ depending on whether the low or the high setting for that factor was used. Fit the linear model resist $\sim x1 + x2 + x3 + x4$.

(1) Extract the $\mathbf{X}$ matrix using the model.matrix function. Examine this to deter-mine how the low and high levels have been coded in the model.
(2) Compute the correlation in the $\mathbf{X}$ matrix. Why are there some missing values in the matrix?
(3) What difference in resistance is expected when moving from the low to the high level of $x1$?
(4) Refit the model without $x4$ and examine the regression coefficients and standard errors? What stayed the the same as the original fit and what changed?
(5) Explain how the change in the regression coefficients is related to the correlation matrix of $\mathbf{X}$.

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).