PubH 7405: REGRESSION ANALYSIS

MLR: INFERENCES, Part II

TESTING HYPOTHESES

- Once we have fitted a multiple linear regression model and obtained estimates for the various parameters of interest, we want to **answer questions about the contributions of factor or factors** to the prediction of the dependent variable Y. There are three types of tests:

(1) An **overall** test

(2) Test for the value of a **single factor**

(3) Test for contribution of a **group of factors**

MARGINAL CONTRIBUTION

- The change, for example, from the model containing only X_1 to the model containing all three variables X_1, X_2 & X_3 represent "marginal contribution" by the **addition of X_2 & X_3**.

- The marginal contribution represent the part (of SSE) that is further explained by X_2 and X_3 (in addition to what already explained by X_1):

- $$SSR(X_2, X_3 | X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$$

A TYPICAL STRATEGY

- Two models: a larger one & a smaller one; larger model has more terms and larger SSR

- The difference in SSR is accountable for by extra terms in the regression model; the group of terms under investigation.

- **Decompose the SSR and the df(R); then calculating the MR and the F ratio**

- **Numerator of F is the MS due to the extra terms; denominator is MSE of larger model.**

- Use the F ratio to test difference of 2 models.

COEFFICIENT OF
PARTIAL DETERMINATION #a

- Suppose we have a multiple regression model with 2 independent variables (the Full Model) and suppose we are interested in the marginal contribution of X_2:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

- The **coefficient of partial determination**, between Y and X_2 measures the marginal reduction in the variation of Y associated with the addition of X_2:

$$R_{Y2|1}^2 = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)}$$

COEFFICIENTS OF PARTIAL CORRELATION

The square root of a coefficient of partial determination is called a **coefficient of partial correlation**. It is given the **same sign as that of the corresponding regression coefficient** in the fitted regression model. **Coefficients of partial correlation are sometimes used in practice, although they do not have a clear meaning** as coefficients of partial determination nor the (single) coefficient of correlation.

PARTIAL CORRELATION & PARTIAL DETERMINATION

Let both the response variable Y and the predictor under investigation (say, X_1) be both regressed against the other predictor variables already in the regression model and the residuals are obtained for each. **These two sets of residuals reflect the part of each (Y and X_1) that is not linearly associated with the other predictor variables.**

Result: The correlation coefficient **between the above two sets of residuals** is equal to the Coefficient of Partial Correlation between Y & X1, which is the square root of the Coefficient of Partial Determination (of X_1 on Y).

In the decomposition of the sums of squares; the "extra sums of squares" are very useful for **testing/measuring for the marginal contribution of individual variable or group of variables.**

However, they are only useful when we want to test **whether several regression coefficients are equal zero; for the case of one factor, we don't need the test but we use Partial Determination Coefficient**

In addition, the role/meaning of the Coefficient of Partial Correlation is now clearer: the correlation coefficient between the above two sets of residuals, the part of each (Y and X1) that is not linearly associated with the other predictor variables. This opens up to the use nonparametric correlations.

There are times when the Null Hypothesis does not fit the "**usual pattern**". For example:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

We may be interested in :

$$H_0 : \beta_1 = \beta_2$$

What should we do? – Similar strategy!

Full Model :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

We are interested in $H_0 : \beta_1 = \beta_2$

Reduced Model :

$$Y = \beta_0 + \beta_c (x_1 + x_2) + \beta_3 x_3 + \varepsilon$$

$$= \beta_0 + \beta_c x_c + \beta_3 x_3 + \varepsilon$$

The Reduced Model has 2 independent variables but they are not a subset of the original 3; X_c is a **"new" variable**.

GENERAL LINEAR TEST APPROACH

- The general linear test approach can be used for highly complex tests of linear statistical models, as well as for simple tests. **The basic steps, which are similar but not identical to the use of extra sums of squares:**

(1) Fit the Full Model to obtain SSE(F)

(2) Fit the Reduced Model to obtain SSE(R)

(3) Use the following Test Statistic & Rule

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_E(R) - df_E(F)}}{\dfrac{SSE(F)}{df_E(F)}}$$

Refer to percentiles of the F distribution with $(df_E(R) - df_E(F); df_E(F))$ degrees of freedom. Note that {SSE(R) - SSE(F)} is similar to the reduction in SSE when a variable is added.

In the recent example:

Full Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Reduced Model:

$$Y = \beta_0 + \beta_c (x_1 + x_2) + \beta_3 x_3 + \varepsilon$$

COVARIATES

- In biomedical research, the independent variables or covariates represent patients' characteristics and, in many cases of clinical research, one of them represents the treatment.

- Do we need to impose any assumptions on these predictor variables?

- First, unlike the response variable, we treat the covariates as "fixed" – without assuming any kind of distributions (it seems like a language problem with "variable" – as used in "predictor variable" or "independent variable"!)

MEASUREMENT SCALE

- There are no restriction on measurement scale for Independent Variables.

- Examples of **binary covariates** include Gender, and presence or absence of certain co-morbidity. Polytomous or **categorical covariates** include race, and different grades of symptoms. **Continuous covariates** include patient age, blood pressure, etc...

- Dependent Variable must be on the continuous scale; the **Model** imposes a **normal distribution** for the dependent variable; we will briefly look at the case of a Binary Dependent Variable on the very last lecture.

We can investigate binary covariates, we can investigate categorical covariates, and we can investigate continuous covariates. However, **when we have a <u>combination the predictor variables having substantially different magnitudes</u> (say, one is 0/1 while another varies from 0 to 50,000), we may have serious problems**. The **inversion of the matrix X'X** might run into serious rounding-off errors (even with computer programs using double-precision calculations), which maybe magnified in calculating estimated regression coefficients and other subsequent statistics.

EXAMPLE: TWO COVARIATES

$$\mathbf{X'X} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{bmatrix}$$

The key step is **to bring the entries of this matrix to similar magnitudes**; the solution is called : "Correlation Transformation" (which precedes data analysis)

STANDARDIZATION

- Standardizing a variable X involves centering & scaling X
- **Centering** involves taking the difference between each observation and the mean of all observations for the X
- **Scaling** involves expressing the centered observations in units of the standard deviation of the observations for X

$$x_{st} = \frac{x - \bar{x}}{s_x}$$

CORRELATION TRANSFORMATION

- The **Correlation Transformation** of X is a simple modification of its usual standardization.

- **The result**: it makes all entries in the **X'X** matrix for the transformed variables falling between -1 and 1 (will prove soon).

$$\mathbf{x}^* = \frac{1}{\sqrt{\mathbf{n}-1}} \left(\frac{\mathbf{x} - \bar{\mathbf{x}}}{\mathbf{s_x}} \right)$$

STANDARDIZED REGRESSION MODEL

- With correlation-transformed variables Y*'s <u>and</u> all X*'s, the result is called the **Standardized Regression Model**

- It has no intercept because, with correlation transformed values of data, the least squares method always would lead to an estimated **intercept term of zero**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

&

$$\mathbf{Y}^* = \boldsymbol{\beta}_1^* \mathbf{x}_1^* + \boldsymbol{\beta}_2^* \mathbf{x}_2^* + \cdots \boldsymbol{\beta}_k^* \mathbf{x}_k^* + \boldsymbol{\varepsilon}^*$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \varepsilon$$

&

$$Y^* = \beta_1^* x_1^* + \beta_2^* x_2^* + \cdots \beta_k^* x_k^* + \varepsilon^*$$

$$\beta_i = \left(\frac{s_Y}{s_{Xi}} \right) \beta_i^*$$

The **standardized regression coefficients <u>and</u> the original regression coefficients are related by simple scaling factors** involving ratios of standard deviations.

NOTE: For Regression Model without intercept, the "Design Matrix" **X** of correlation-transformed data becomes:

$$\mathbf{X}^*_{nxk} = \begin{bmatrix} x^*_{11} & x^*_{12} & \cdot & x^*_{1k} \\ x^*_{21} & x^*_{22} & \cdot & x^*_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x^*_{n1} & x^*_{n2} & \cdot & x^*_{nk} \end{bmatrix}$$

No first column with all 1's

$$\sum (x_{i1}^*)^2 = \sum (\frac{x_{i1} - \bar{x}_1}{\sqrt{n-1}s_1})^2$$

$$= 1$$

$$\sum (x_{i1}^*)(x_{i2}^*) = \sum (\frac{x_{i1} - \bar{x}_1}{\sqrt{n-1}s_1})(\frac{x_{i2} - \bar{x}_2}{\sqrt{n-1}s_2})$$

$$= r_{12}$$

Result : All entries in $\mathbf{X'X}$ are between -1 and 1

$$\mathbf{X^{*'}X^*} = \mathbf{r_{XX}}$$

$$= \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

$$= \text{Correlation Matrix of } \mathbf{X's}$$

$$\mathbf{X}^{*'}\mathbf{Y}^{*} = \mathbf{r}_{YX}$$

$$= \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Yk} \end{bmatrix}$$

$$= \text{Correlation between Y \& X's}$$

$$\mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$$

$$= \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$$

EXAMPLE: k=2

$$b_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2}$$

$$b_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}$$

Very easy calculations!

We start to see a **new potential problem** here: **what if r_{12} is near ±1?**

$$b_1^* = \frac{r_{Y1} - r_{12} r_{Y2}}{1 - r_{12}^2}$$

$$b_2^* = \frac{r_{Y2} - r_{12} r_{Y1}}{1 - r_{12}^2}$$

The use of the Standardized Regression Model not only help to reduce/avoid rounding-off errors but also permit comparison of the estimated regression coefficients in common units. That allows us to "rank" the level of importance of predictors – just like using the coefficients of partial determination.

After all the results for the Standardized Regression Model are obtained, the return to the original model is proceeded as follows:

$$b_i = \left(\frac{s_Y}{s_i} \right) b_i^*$$

$$b_0 = \bar{y} - \sum b_i \bar{x_i}$$

We treat the covariates as fixed; no assumptions on their distributions and even no restrictions on their measurement scale (standardized regression model would help here). However, as briefly see, **their relationships with each other** **may present serious problems** for data analysis.

SOME QUESTIONS OF INTEREST

(1) What is the magnitude of the effect of a given predictor variable on the response variable?

(2) What is the relative importance of the effects of different predictor variables?

(3) Can any predictor variable be dropped from the model because it has little or no effect on the response variable?

(4) Should any predictors variables not yet included in the model be considered for inclusion?

If the predictor variables in the model are:
    (1) uncorrelated among themselves and
    (2) uncorrelated with any other predictor variables not yet in the model but related to the response, then relatively simple answers can be given to some the above questions. For example:

For question #3, Can any predictor variable be dropped from the model because it has little or no effect on the response variable?, the answer is yes. However, if predictor variables are related, dropping one from the model would **change the relationships between the response variable and other remaining predictors.**

If the predictor variables in the model are:
    (1) uncorrelated among themselves and
    (2) uncorrelated with any other predictor variables not yet in the model but related to the response, then relatively simple answers can be given to some the above questions. For example:

For question #4, Should any predictors variables not yet included in the model be considered for inclusion? The answer is yes. <u>However</u>, if a "potential predictor variable" are related with predictor variables in the model, **its "value" may be <u>reduced</u> because of redundancy.**

In many non-experimental situations in biomedical research – or any other field – the **predictor variables tend to be correlated among themselves and with other variables related to the response but not yet in the regression model.** For example, we investigated possible effects of Age, Weight, and Height on FEV1, a lung health measure. The problem is **Height** is correlated – to different degrees – to **Age**, to **Weight**, **and** to **Gender** which is not yet considered in the model.

UNCORRELATED PREDICTORS

- The text gives an example (page 279) to illustrate that when 2 predictors are uncorrelated:

(1) Regression coefficient for one variable is the same whether or not the model includes the other.

(2) The marginal contribution of a variable is the same as its (simple regression) contribution:

$$SSR(X_2 \mid X_1) = SSR(X_2)$$

Recall that we have, in Simple Linear Regression of Y on X_1:

& on Multiple Regression
of Y on X1 and X2:

$$b_1 = \frac{s^2_{x_1 y}}{s^2_{x_1}}$$

$$= \mathbf{r_{Y1}} \frac{\mathbf{s_y}}{\mathbf{s_1}}$$

$$b_1^* = \frac{r_{Y1} - r_{12} r_{Y2}}{1 - r_{12}^2}$$

$$= r_{Y1} \text{ (because } r_{12} = 0)$$

$$= \frac{s_1}{s_Y} b_1$$

$$b_1 = \mathbf{r_{Y1}} \frac{\mathbf{s_Y}}{\mathbf{s_1}}$$

PERFECTLY CORRELATED PREDICTORS

- The text gives another example (page 281) to show that, when two predictors are perfectly correlated:

(1) It is **still possible** to obtain estimated regression coefficients, but

(2) The **solutions are not unique** for the same good fit; therefore, one of these solutions cannot provide meaningful interpretation.

In the real life/practice, we seldom have predictor variables that are perfectly related, nor uncorrelated.

When the predictor variables are correlated among themselves, "**intercorrelation**" or "**multicollinearity**" among them are said to exist. (However, sometimes the latter term, multicollinearity, is reserved only for those instances when the correlation among the predictor variables is very high).

We are exploring brief the problems created by the phenomenon of multicollinearity.

(1)

$$b_1 = \left(\frac{s_Y}{s_1} \right) b_1^*$$

$$= \left(\frac{s_Y}{s_1} \right) \left(\frac{r_{Y1} - r_{12} r_{Y2}}{1 - r_{12}^2} \right)$$

$$b_2 = \left(\frac{s_Y}{s_2} \right) b_2^*$$

$$= \left(\frac{s_Y}{s_2} \right) \left(\frac{r_{Y2} - r_{12} r_{Y1}}{1 - r_{12}^2} \right)$$

High correlation among predictors does not inhibit our ability to obtain good fit but solutions may not be unique. You can see from the formula that, when r_{12} is near 1 or (-1), the **estimated coefficients become increasingly undetermined** (because the denominators approach zero)

(2)

The common interpretation of a coefficient, in the context of multiple regression, as measuring the change in the mean response when the given predictor variable is increased by one unit **while** all other predictor variables are held constant. It is not conceptually feasible to think of varying one predictor variable and holding others constant if they are highly correlated. In other words, **usual interpretation is no longer meaningful.**

When predictor variables are highly correlated among themselves, the estimated regression coefficients tend to have large sampling variability, i.e. **large standard errors**. That is, the estimated regression coefficients very widely from one sample to the next; that values obtained in one samples become meaningless.

Let think in terms of transforming the Original Model into the Standardized Regression Model:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

&

$$E(\mathbf{Y}^*) = \boldsymbol{\beta}_1^* \mathbf{x}_1^* + \boldsymbol{\beta}_2^* \mathbf{x}_2^*$$

And recall of the following results:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\sigma^2(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{s}^2(\mathbf{b}) = MSE(\mathbf{X}'\mathbf{X})^{-1}$$

Using data after the correlation transformation:

$$\boldsymbol{\sigma}^2(\mathbf{b}^*) = \sigma^{*2}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma^{*2}\mathbf{r}_{\mathbf{XX}}^{-1}$$

$$= \frac{\sigma^{*2}}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

$$\sigma^2(b_1^*) = \sigma^2(b_2^*)$$

$$= \frac{\sigma^{*2}}{1-r_{12}^2}$$

$$\sigma^2(b_1^*) = \sigma^2(b_2^*)$$

$$= \frac{\sigma^{*2}}{1 - r_{12}^2}$$

$$s(b_1^*) = \sqrt{\frac{MSE^*}{1 - r_{12}^2}}$$

When predictor variables are highly correlated among themselves, correlation coefficients, such as r_{12}, are high leading to large variances and standard errors (near zero denominator). That, in turns, leads to same effects on estimated regression coefficients of the Original Model.

Multicollinearity is a very serious problem without easy solutions. It might even go undetected; serious multicollinearity even exists **without being disclosed by the pairwise correlation coefficient.** More powerful diagnostics are desirable. Maybe several pairwise correlation coefficients at medium level could, together, cause serious problems.

Readings & Exercises

- Readings: A thorough reading of the text's section2.8 (pages 72-73) and sections 7.5-7.6 (pp.271-289) is recommended.

- Exercises: The following exercises are good for practice, all from chapter 7 of text: 7.18-7.19, 7.24-7.25

Due As Homework

#17.1 Refer to dataset "Cigarettes", let Y = log(NNAL) and consider a model with three independent variables, X_1 = CPD, X_2 = Age, & $X_3 = X_2^2$.

a) Fit the model with all 3 predictors and draw your conclusion – especially with respect to the marginal contribution of the quadratic term. Is this reasonable to interpret β_2 as indicating the effect of X_2 when the other two variables are held constant?

b) Transform the variables by means of the correlation transformation and fit the standardized regression model. Are the conclusions different from those in part (a)?

c) Transform the estimated standardized regression coefficients obtained in part (b) back to the ones for the regular model using original variables. Verify that we get the sane results as in part (a).

#17.2 Answer the 3 questions of Exercise 17.1 using dataset "Infants" with Y = Birth Weight, X_1 = Gestational Weeks, X_2 = Mother's Age, & $X_3 = X_2^2$.