Data Analysis
Spring Semester, 2023
March 28, 2023
Lecture 16

Second Midterm on Thursday, March 30. It will focus on Chapters 11 and 12. I will not hold Zoom hours on Wednesday, March 29. Please use a TA Zoom hour. Their hours are listed in the announcement section of the Blackboard.

Chapter 12
Multiple Regression and the General Linear Model

Specifically, the model for Chapter 12 is $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_{Y\bullet 12} Z_i$. The parameters $(\beta_0, \beta_1, \beta_2)$ are fixed but unknown. The parameter $\sigma_{Y\bullet 12}$ is the unknown conditional standard deviation of Y_i controlling for x_{1i} and x_{2i}, $i = 1, \ldots, n$. The standard deviation of Y_i is assumed to be equal for each observation. The random errors Z_i are assumed to be independent. The independence of the random errors (and hence independence of Y_i) is important. The assumption of a linear regression function (that is, $E(Y_i \mid x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$) is also important. As in Chapter 11, this is equivalent to the joint distribution of the dependent variable values being $NID(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2_{Y\bullet 12})$.

Estimating the Linear Model Parameters

A linear model with arbitrary arguments $b_0 + b_1 x_1 + b_2 x_2$ is used as a *fit* for the dependent variable values. The method uses the *residual* $y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}$. OLS minimizes the sum of squares function $SS(b_0, b_1, b_2) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$. The OLS method is to find the arguments $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ that make $SS(b_0, b_1, b_2)$ as small as possible. This minimization is a standard calculus problem. One finds the arguments $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ that make the three partial derivatives simultaneously zero. The resulting equations are still called the *normal equations*:

$$\sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = 0,$$

$$\sum_{i=1}^{n} (-2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) x_{1i} = 0, \text{ and}$$

$$\sum_{i=1}^{n}(-2)(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-\hat{\beta}_2x_{2i})x_{2i}=0,.$$

These equations still have a very important mathematical interpretation. Let $r_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-\hat{\beta}_2x_{2i},\ i=1,\ldots,n$. The first normal equation is equivalent to $\sum_{i=1}^{n}r_i=0$; the second is $\sum_{i=1}^{n}r_ix_{1i}=0$; and the third is $\sum_{i=1}^{n}r_ix_{2i}=0$ That is, there are three constraints on the *n* residuals. The OLS residuals must sum to zero, and the OLS residuals are orthogonal to the two independent variable values. The *n* residuals then have $n-3$ degrees of freedom.

Next, one solves this three linear equation system in three unknowns. There is a more general approach to solving systems like this. The first equation is $\sum_{i=1}^{n}(-2)(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-\hat{\beta}_2x_{2i})=0,$ which can be written $\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i})$ and $\sum_{i=1}^{n}(1\times y_i)=[\sum_{i=1}^{n}(1\times 1)]\hat{\beta}_0+[\sum_{i=1}^{n}(1\times x_{1i})]\hat{\beta}_1+[\sum_{i=1}^{n}(1\times x_{2i})]\hat{\beta}_2$. Similarly, the second normal equation can be written $\sum_{i=1}^{n}(x_{1i}\times y_i)=[\sum_{i=1}^{n}(1\times x_{1i})]\hat{\beta}_0+[\sum_{i=1}^{n}(x_{1i}\times x_{1i})]\hat{\beta}_1+[\sum_{i=1}^{n}(x_{1i}\times x_{2i})]\hat{\beta}_2$; and the third

$$\sum_{i=1}^{n}(x_{2i}\times y_i)=[\sum_{i=1}^{n}(1\times x_{2i})]\hat{\beta}_0+[\sum_{i=1}^{n}(x_{2i}\times x_{1i})]\hat{\beta}_1+[\sum_{i=1}^{n}(x_{2i}\times x_{2i})]\hat{\beta}_2 .$$

While these look like complicated equations, matrix algebra leads to a simpler expression. Let $y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}$ be the $n\times 1$ column vector of dependent variable values, and let $X=[\begin{matrix}1 & x_{11} & x_{21}\\1 & x_{12} & x_{22}\\\vdots & \vdots & \vdots\\1 & x_{1n} & x_{2n}\end{matrix}]$ be the $n\times 3$ matrix of coefficients of the parameters

$(\beta_0, \beta_1, \beta_2)$. From matrix algebra, $X^T X = \begin{bmatrix} \sum_{i=1}^n (1\times 1) & \sum_{i=1}^n (1\times x_{1i}) & \sum_{i=1}^n (1\times x_{2i}) \\ \sum_{i=1}^n (1\times x_{1i}) & \sum_{i=1}^n (x_{1i}\times x_{1i}) & \sum_{i=1}^n (x_{1i}\times x_{2i}) \\ \sum_{i=1}^n (1\times x_{2i}) & \sum_{i=1}^n (x_{1i}\times x_{21}) & \sum_{i=1}^n (x_{2i}\times x_{2i}) \end{bmatrix}$, and

$X^T Y = \begin{bmatrix} \sum_{i=1}^n (1\times y_i) \\ \sum_{i=1}^n (x_{1i}\times y_i) \\ \sum_{i=1}^n (x_{2i}\times y_i) \end{bmatrix}$.

Recall the three normal equations above:

$$\sum_{i=1}^n (1\times y_i) = [\sum_{i=1}^n (1\times 1)]\hat{\beta}_0 + [\sum_{i=1}^n (1\times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^n (1\times x_{2i})]\hat{\beta}_2$$

$$\sum_{i=1}^n (x_{1i}\times y_i) = [\sum_{i=1}^n (1\times x_{1i})]\hat{\beta}_0 + [\sum_{i=1}^n (x_{1i}\times x_{1i})]\hat{\beta}_1 + [\sum_{i=1}^n (x_{1i}\times x_{2i})]\hat{\beta}_2$$

$$\sum_{i=1}^n (x_{2i}\times y_i) = [\sum_{i=1}^n (1\times x_{2i})]\hat{\beta}_0 + [\sum_{i=1}^n (x_{1i}\times x_{2i})]\hat{\beta}_1 + [\sum_{i=1}^n (x_{2i}\times x_{2i})]\hat{\beta}_2$$

The left-hand side terms are the same as the terms of $X^T Y$, and the coefficients of the OLS estimators match with the terms of $X^T X$. For this problem, then, the normal equations can be written in matrix form as

$$(X^T X)\hat{\beta} = X^T Y .$$

This result also holds for three or more independent variables. The proof is the same as for the two independent variable case.

If $(X^T X)^{-1}$ exists, then $\hat{\beta} = (X^T X)^{-1} X^T Y$. The existence of $(X^T X)^{-1}$ is the usual case in observational studies using multiple regression. If $(X^T X)^{-1}$ does not exist, then the OLS estimators exist but are not unique.

Distribution of $\hat{\beta} = (X^T X)^{-1} X^T Y$

Let $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ be the vector of the random outcome variables. That is, the data will be collected in the future as opposed to having the data in hand as we assumed in our OLS estimator derivation. The probabilistic model for the data can be written in matrix form $Y = X\beta + \sigma_{Y\bullet 12}Z$, where Z is the column vector of random errors Z_i that are assumed to be independent.

The model is that

$$E(Y) = E(X\beta + \sigma_{Y\bullet 12}Z) = E(X\beta) + E(\sigma_{Y\bullet 12}Z) = X\beta + \sigma_{Y\bullet 12}E(Z) = X\beta,$$ and

$\text{vcv}(Y) = \sigma^2_{Y\bullet 12}I_{n\times n}$. An equivalent description is to say that Y is multivariate normal with dimension n; that is, Y has the distribution $MVN_n(X\beta, \sigma^2_{Y\bullet 12}I_{n\times n})$.

For this model with three parameters, when the matrix X has rank 3, $(X^TX)^{-1}$ exists. Then the vector of OLS estimators is
$\hat{\beta} = (X^TX)^{-1}X^TY$. The expected value is given by

$$E(\hat{\beta}) = E[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^TE(Y) = (X^TX)^{-1}X^T(X\beta) = [(X^TX)^{-1}(X^TX)]\beta = I_{p\times p}\beta = \beta.$$

The variance-covariance matrix of $\hat{\beta} = (X^TX)^{-1}X^TY$ is calculated by
$\text{vcv}(\hat{\beta}) = \text{vcv}[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T\text{vcv}(Y)[(X^TX)^{-1}X^T]^T$.

Recall that the transpose of the transpose of a matrix is just the matrix so that $(X^T)^T = X$. Further, a matrix is symmetric if its transpose is the matrix itself. That is, $(X^TX)^T = X^T(X^T)^T = X^TX$. The inverse of a symmetric matrix is symmetric so that $[(X^TX)^{-1}]^T = (X^TX)^{-1}$. Using these results in

$$\text{vcv}(\hat{\beta}) = (X^TX)^{-1}X^T\text{vcv}(Y)[(X^TX)^{-1}X^T]^T = (X^TX)^{-1}X^T\sigma^2_{\text{Y}\bullet\text{x}}I_{n\times n}X(X^TX)^{-1},$$
$$\text{vcv}(\hat{\beta}) = (X^TX)^{-1}X^T\sigma^2_{\text{Y}\bullet\text{x}}I_{n\times n}X(X^TX)^{-1} = \sigma^2_{\text{Y}\bullet\text{x}}\{(X^TX)^{-1}[X^TX]\}(X^TX)^{-1} = \sigma^2_{\text{Y}\bullet\text{x}}\{I_{p\times p}\}(X^TX)^{-1} = \sigma^2_{\text{Y}\bullet\text{x}}(X^TX)^{-1}.$$

The distribution of $\hat{\beta} = (X^TX)^{-1}X^TY$ is $MVN_p(\beta, \sigma^2_{\text{Y}\bullet\text{x}}(X^TX)^{-1})$.

Fisher's Decomposition of the (Uncorrected) Total Sum of Squares

The uncorrected total sum of squares of the dependent variable is defined to be Y^TY with n degrees of freedom. In Chapter 11, the (corrected) total sum of squares was used. This is $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 = Y^TY - n\bar{Y}_n^2$ with $n-1$ degrees of freedom. First, Fisher's decomposition of the uncorrected total sum of squares follows from

$$\begin{aligned} Y^TY &= (Y - X\hat{\beta} + X\hat{\beta})^T(Y - X\hat{\beta} + X\hat{\beta}) \\ &= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) + (X\hat{\beta})^T(Y - X\hat{\beta}) + (Y - X\hat{\beta})^TX\hat{\beta} + (X\hat{\beta})^T(X\hat{\beta}). \end{aligned}$$

This result can be simplified using

$$(Y - X\hat{\beta})^T X\hat{\beta} = (Y - X(X^TX)^{-1}X^TY)^T X(X^TX)^{-1}X^TY = Y^T(I_{n\times n} - X(X^TX)^{-1}X^T)^T X(X^TX)^{-1}X^TY$$
$$= Y^T\{(I_{n\times n})^T - [X(X^TX)^{-1}X^T]^T\}X(X^TX)^{-1}X^TY$$
$$= Y^T[I_{n\times n} - X(X^TX)^{-1}X^T]X(X^TX)^{-1}X^TY$$
$$= Y^T[X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T]Y$$
$$= Y^T[X(X^TX)^{-1}X^T - X(X^TX)^{-1}\{(X^TX)(X^TX)^{-1}\}X^T]Y$$
$$= Y^T[X(X^TX)^{-1}X^T - X(X^TX)^{-1}\{I_{p\times p}\}X^T]Y = 0.$$

Of course, $(X\hat{\beta})^T(Y - X\hat{\beta}) = 0.$

Then
$$Y^TY = (Y - X\hat{\beta} + X\hat{\beta})^T(Y - X\hat{\beta} + X\hat{\beta})$$
$$= (Y - X\hat{\beta})^T(Y - X\hat{\beta}) + (X\hat{\beta})^T(X\hat{\beta}).$$

The residuals R are defined to be the $n \times 1$ vector $R = Y - X\hat{\beta}$ on $n - p$ degrees of freedom, and the fitted values $\hat{Y} = X\hat{\beta}$ on p degrees of freedom. Then the uncorrected total sum of squares is $Y^TY = R^TR + \hat{Y}^T\hat{Y}$. The error sum of squares is defined to be R^TR with $n - p$ degrees of freedom. The uncorrected sum of squares due to regression is defined to be $\hat{Y}^T\hat{Y}$ with p degrees of freedom. Statistical computing programs subtract the correction $n\overline{Y}_n^2$ with 1 degree of freedom from both the uncorrected total sum of squares and uncorrected regression sum of squares. That is, the programs display the corrected total sum of squares $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 = Y^TY - n\overline{Y}_n^2$ and the corrected regression sum of squares in the Analysis of Variance Table as below:

Analysis of Variance Table

$p-1$ Predictor Multiple Linear Regression

Source	DF	Sum of Squares	Mean Square	F
Regression	$p-1$	$(X\hat{\beta})^T(X\hat{\beta}) - n\overline{Y}_n^2$	$\frac{(X\hat{\beta})^T(X\hat{\beta}) - n\overline{Y}_n^2}{p-1}$	$\frac{MS\ REG}{MSE}$
Error	$n-p$	R^TR	$\frac{R^TR}{(n-p)}$	
Total	$n-1$	$TSS = (n-1)s_{DV}^2$		

Inferences

With *p-1* independent variables, the probabilistic model for the data is $Y = X\beta + \sigma_{Y\bullet1\ldots(p-1)}Z$. The outcome or dependent (random) variables $Y_i, i = 1,\ldots,n$ are each assumed to be the sum of the linear regression expected value $\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1}x_{(p-1)i}$ and a random error term $\sigma_{Y\bullet1\ldots(p-1)}Z_i$. The random variables $Z_i, i = 1,\ldots,n$ are assumed to be independent standard normal random variables. The parameter β_0 is the intercept parameter and is fixed but unknown. The parameters $\beta_1,\ldots,\beta_{p-1}$ are partial regression coefficient parameters and are also fixed but unknown. These parameters are the focus of the statistical analysis. The parameter $\sigma_{Y\bullet1\ldots(p-1)}$ is also fixed but unknown. Another description of this model is that $Y_i, i = 1,\ldots,n$ are independent normally distributed random variables with $Y_{n\times1}$ having the distribution $MVN(X\beta, \sigma^2_{Y\bullet1\cdots(p-1)}I_{n\times n})$.

Again, there are four assumptions. The two important assumptions are that the outcome variables $Y_i, i = 1,\ldots,n$ are independent and that the regression function is $\beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1}x_{(p-1)i}$ $i = 1,\ldots,n$. Homoscedasticity is less important. The assumption that $Y_i, i = 1,\ldots,n$ are normally distributed random variables is least important.

Testing null hypotheses about the partial regression coefficients (Not in text)

The mathematical analysis of the general problem is complicated. The analysis for two independent variables, however, is more manageable—particularly the problem of sequential tests. As before, the model for the data is $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_{Y\bullet12}Z_i.$

The research problem is to consider a sequence of models. The first model is that $Y_i = \beta_0 + \beta_1 x_{1i} + \sigma_{Y\bullet1}Z_i,$ with null hypothesis $H_0 : \beta_1 = 0.$ This is a Chapter 11 problem. The second model is that $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_{Y\bullet12}Z_i,$ with null hypothesis $H_0 : \beta_2 = 0.$ This is an example of a sequential test. That is, the second hypothesis is tested after the first one. These tests require the definition of the partial correlation coefficient.

Partial correlation coefficient

Let the correlation matrix of (Y, x_1, x_2) be

$$\begin{pmatrix} 1 & \rho(y,x_1)=\rho_{y1} & \rho(y,x_2)=\rho_{y2} \\ \rho(y,x_1)=\rho_{y1} & 1 & \rho(x_1,x_2)=\rho_{12} \\ \rho(y,x_2)=\rho_{y2} & \rho(x_1,x_2)=\rho_{12} & 1 \end{pmatrix}$$

The *partial correlation* between Y and x_2 controlling for x_1 is defined to be $\rho_{y2.1}=\frac{\rho_{y2}-\rho_{y1}\rho_{12}}{\sqrt{(1-\rho_{y1}^2)(1-\rho_{12}^2)}}$. Analogous definitions hold for the Pearson product moment correlations. That is, $r_{y2.1}=\frac{r_{y2}-r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}$

Analysis of variance table for a sequential test

The (corrected) total sum of squares is always $TSS=(n-1)s_{DV}^2$. The first model is that $Y_i=\gamma_0+\gamma_1 x_{1i}+\sigma_{Y\bullet 1}Z_i$, with null hypothesis $H_0:\gamma_1=0.$ The sum of squares due to the regression on x_1 is then $[r(x_1,y)]^2 TSS$ on 1 degree of freedom.

The regression on x_1 has error sum of squares $\{1-[r(x_1,y)]^2\}TSS$ on $n-2$ degrees of freedom. The new aspect of multiple regression is that there is a second independent (predictor) variable. The regression on x_2 after x_1 has been entered explains an additional $r_{y2\bullet 1}^2$ of the $\{1-[r(x_1,y)]^2\}TSS$ that was not explained by x_1. That is, the sum of squares due to the regression on $x_2\mid x_1$ is $[r_{y2\bullet 1}^2\{1-[r(x_1,y)]^2\}TSS$ with one degree of freedom. The error sum of squares is obtained by subtracting both the sum of squares due to the regression on x_1 and the sum of squares due to the regression on $x_2\mid x_1$. The analysis of variance table below summarizes these results.

Analysis of variance table
Multiple regression of Y on x_1 and $x_2 \mid x_1$

Source	DF	Sum of Squares	Mean Square	
Reg on x_1	1	$[r(x_1, y)]^2 TSS$	$[r(x_1, y)]^2 TSS$	
Reg on $x_2 \mid x_1$	1	$[r^2_{y2\bullet 1}\{1-[r(x_1, y)]^2\} TSS$	$[r^2_{y2\bullet 1}\{1-[r(x_1, y)]^2\} TSS$	
Error	$n-3$	Subtraction	MSE	
Total (corrected)	$n-1$	$TSS = (n-1)s^2_{DV}$		

The test of $H_0 : \beta_2 = 0$ against the alternative hypothesis that $H_1 : \beta_2 \neq 0$ uses the test statistic $F_{2\bullet 1} = \dfrac{[r^2_{y2\bullet 1}\{1-[r(x_1, y)]^2\} TSS}{MSE}$, which has 1 numerator and $n-3$ denominator degrees of freedom.

This presentation of the models has disguised the complexity of the coefficients. The first model was $Y_i = \gamma_0 + \gamma_1 x_{1i} + \sigma_{Y\bullet 1} Z_i$. The specification of the γ_1 parameter requires taking expectation to get $E(Y \mid x_1) = \gamma_0 + \gamma_1 x_1$. Then, $\gamma_1 = \dfrac{\partial}{\partial x_1} E(Y_i \mid x_1)$; that is, γ_1 is the expected increase in the value of the dependent variable associated with a unit increase in x_1. The extended model was $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sigma_{Y\bullet 12} Z_i$, so that $E(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Then $\beta_2 = \dfrac{\partial}{\partial x_2}\{E(Y \mid x_1, x_2)\}\mid_{x_1 \text{ fixed}}$. The coefficient $\beta_2 = \dfrac{\partial}{\partial x_2}\{E(Y \mid x_1, x_2)\}\mid_{x_1 \text{ fixed}}$ is the expected increase in the value of the dependent variable associated with a unit increase in x_2, controlling for x_1 being held constant (sometimes called *ceteris paribus*). These coefficients are partial derivatives of the conditional expectation of the dependent variable.

Complete Mediation and Complete Explanation Causal Models

In analyzing research data from engineering or physical sciences studies, the independent variables typically operate at the same time. Given this, the fact that a partial regression coefficient is an estimate of a partial derivative strongly indicates to the user that caution is warranted in the interpretation of a partial regression coefficient. In social science and epidemiological research, however, the independent variables may operate at different points of time. For example, x_1 may describe a variable measured when the participant was between ages 5 and 6, and

x_2 may describe a variable measured when the participant was between the ages of 8 and 9. The time-ordering of the independent variables is a crucial consideration in the interpretation of partial regression coefficients.

For example, often one sees that ρ_{y2} appears significant (that is, x_2 has a significant *F* statistic in a multiple regression analysis or the r_{y2}, the Pearson product moment correlation, is significant) but that $\rho_{y2.1}$ does not appear significant. That is, in multiple regression analysis, the variable x_2 does not have a significant F-to-enter once x_1 is in the regression equation. There is a fundamental paper (Simon, 1954, available on JSTOR and on the Blackboard site) that you should download and read it.

Simon points out that when one has a common cause model (or *explanation*), the independent variable x_1 precedes both x_2 and *y* with regard to operation impact. Then if x_1 "causes" x_2 and if x_1 "causes" *y*, then there will be a "spurious" correlation ρ_{y2} (this correlation will be non-zero even though x_2 has no causal relation to *y*) and $\rho_{y2.1}$ will be zero. For example, consider G. B. Shaw's correlation between the number of suicides in England in a given year and the number of churches of England in the same year.

In a causal chain model, the independent variable x_2 operates before and causes x_1, and x_1 operates before *y* and causes *y*. Simon also points out that, when the model is a causal chain (or *mediation*), one also observes that ρ_{y2} will be non-zero and $\rho_{y2.1}$ will be zero (even though x_2 causes *y* through the mediation of x_1). Both causal modeling situations have the same empirical fact that a partial correlation is near 0. Deciding which interpretation is valid requires clarifying the sequence of operation of the variables. In practice, the relevant partial correlation may not be essentially 0. In this event, researchers speak of partial explanation and partial mediation.

Second Computer Project

Your second project expands on the concept in a paper by Caspi et al. that is posted on the class blackboard. The model considered in that paper is $Y_i = \beta_0 + \beta_1 G_{1i} + \beta_2 E_{1i} + \beta_3 (E_{1i} \times G_{1i}) + \sigma Z_i$, where E_{1i} is an "environmental" variable (in the paper "stressful life events between ages 21 and 26"), G_{1i} is a genetic variable (in the paper the number of copies of an allele that puts the participant at risk), and

Y_i is “depression outcomes at age 26.” The model contains a gene-environment interaction term (namely, $E_{1i} \times G_{1i}$). Caspi et al. reported that the sequential test of entering the gene-environment interaction after the gene and environment variables was significant.

The model that you are given for your second project is inspired by this paper. Each file contains one dependent variable and nineteen independent variables. The values of the dependent variable are in the DV column. The values of the nineteen independent variables are in the columns with names of E1 to E4 and G1 to G15. The variables E1 to E4 are continuous and positive and simulate “environmental” variables. The variables G1 to G20 are indicator variables. That is, the values are either 0 or 1. The value 1 for GJ_i indicates that the participant *i* is “at risk” based on the participant’s genotype on the *J*th gene. The value 0 indicated that the participant is not at risk.
The model that generates the data for your group may contain a number of significant variables. Researchers typically have a collection of environmental variables that are strongly associated with any outcome variable. Typically, one would expect to find one or more of E1 to E4 to be significant. The association may be non-linear; for example, E1**2 may be associated with *Y*. Genetic associations are more problematic. There may well not be an association between *Y* and any of the genetic variables in your model. Typically, one or more genes in a study are associated with *Y*. There may also be a gene environment interaction; for example, E2xG3 may be significantly associated with *Y*. There may also be gene-gene interactions; for example, G4xG5 may be significantly associated with *Y*. Your group’s task is to find the model that generated your data.

First steps in analyzing the data

- Get summary statistics on each variable.
- Get histograms of *Y* and each environmental variable
- Get correlation matrix of all variables. Identify all of the variables with relatively strong correlations with *Y*.
- Get scatterplot of *Y* versus each environmental variable.
- Do a stepwise regression analysis of *Y* using environmental and genetic variables. Examine the residual vs. predicted plot. See if you can identify transformations of the environmental variables and/or Y that would make the assumptions of multiple regression better satisfied.
- Generate the interactions that you think are plausible and run another stepwise selection regression.

- Keep working until you have a satisfactory residual plot. The R-squared values will vary among groups. There is no correct R-squared value.
- There are many models with essentially equal R-squared values. The variables associated with the outcome are the same in these models.
- The p-value for terms in the model that you report should be small, on the order of 0.005 or smaller. Term with p-values around 0.02 to 0.04 are likely to be false positives (i.e., Type I errors).
- Avoid overly complex models. One rule of thumb is that you should have at least 5 (better 20) observations per parameter.
- Challenge your results. Make sure that the associations that you are ready to report are really there.

Chapter Eight
Inferences about More than Two Population Central Values

Context

The procedures in this chapter generalize the test of the equality of means of two independent populations. This generalization is often called the one-way layout. While this design has somewhat limited value in practice, the material in this chapter is fundamental for further generalizations. The key ideas that are first developed in the one-way analysis of variance are: the generalization of the t-test, the expected mean square calculation (which is described in Chapter 14 and is crucial for power calculations), and the introduction to multiple testing of hypotheses in Chapter 9.

The Model of Observations in a Completely Randomized Design

The usual "effects" model is $Y_{ij} = \mu + \alpha_i + \sigma_{1W} Z_{ij}$, for $i = 1, \ldots, I$ (where *I* is the number of treatment settings), $j = 1, \ldots, J_i$, and $\sum_{i=1}^{I} J_i \alpha_i = 0$. The use of Z_{ij} in this model is the assumption that the dependent variable data is normally distributed and independent. The use of the multiplier σ_{1W} is the assumption that the variances within groups are homogeneous. The important assumption is independence of the error terms. This is guaranteed when there is a random assignment of experimental unit to treatments. Sometimes researchers apply these techniques to data not generated by a randomized experiment. In that event, checking the assumption of independence is crucial. The $\{\alpha_i\}$ parameters are called the treatment effects. Under the effects model, $E(Y_{ij}) = \mu + \alpha_i$, and the distribution of Y_{ij} is $NID(\mu + \alpha_i, \sigma_{1W}^2)$.

OLS Estimates

A model that is equivalent to the effects model is called the means model and is $Y_{ij} = \mu_i + \sigma_{1W} Z_{ij}$, where $\mu_i = \mu + \alpha_i$. The sum of squares function is then $SS(m_1, \ldots, m_I) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - m_i)^2$. We seek values of the arguments that make the *SS* function as small as possible. As before, we take the partial derivatives and solve the normal equations.

Partial derivatives

One must calculate in turn $\frac{\partial}{\partial m_1} SS(m_1,\ldots,m_I),\ldots,\frac{\partial}{\partial m_I} SS(m_1,\ldots,m_I)$. First, focus on $\frac{\partial}{\partial m_1} SS(m_1,\ldots,m_I)$:

$$\frac{\partial}{\partial m_1} SS(m_1,\ldots,m_I) = \frac{\partial}{\partial m_1}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-m_i)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\frac{\partial}{\partial m_1}(y_{ij}-m_i)^2$$. Now,

$$\frac{\partial}{\partial m_1} SS(m_1,\ldots,m_I) = \sum_{j=1}^{J_1}\frac{\partial}{\partial m_1}(y_{1j}-m_1)^2 + \sum_{i=2}^{I}\sum_{j=1}^{J_i}\frac{\partial}{\partial m_1}(y_{ij}-m_i)^2 .$$

One must be careful with the partial derivative calculations. For observations from the first treatment,

$\frac{\partial}{\partial m_1}(y_{1j}-m_1)^2 = 2(y_{1j}-m_1)(\frac{\partial}{\partial m_1}(y_{1j}-m_1)) = 2(y_{1j}-m_1)(-1)$. For observations from the second and other treatments,

$\frac{\partial}{\partial m_1}(y_{2j}-m_2)^2 = 2(y_{2j}-m_2)(\frac{\partial}{\partial m_1}(y_{2j}-m_2)) = 2(y_{2j}-m_2)(0) = 0$. That is,

$$\frac{\partial}{\partial m_1} SS(m_1,\ldots,m_I) = \sum_{j=1}^{J_1}[-2(y_{1j}-m_1)] + \sum_{i=2}^{I}\sum_{j=1}^{J_i}0 = -2\sum_{j=1}^{J_1}(y_{1j}-m_1) = -2[\sum_{j=1}^{J_1}y_{1j} - J_1 m_1].$$

In general, $\frac{\partial}{\partial m_i} SS(m_1,\ldots,m_I) = -2[\sum_{j=1}^{J_i}y_{ij} - J_i m_i],\ i=1,\ldots,I.$

Normal Equations

Let $(\hat{\mu}_1,\ldots\hat{\mu}_I)$ be one of the solutions to the normal equations. Then, the first normal equation is

$\frac{\partial}{\partial m_1} SS(\hat{\mu}_1,\ldots,\hat{\mu}_I) = -2[\sum_{j=1}^{J_1}y_{1j} - J_1\hat{\mu}_1] = 0$. This can easily be solved to obtain

$\sum_{j=1}^{J_1} y_{1j} - J_1\hat{\mu}_1 = 0$ or $\hat{\mu}_1 = \frac{\sum_{j=1}^{J_1} y_{1j}}{J_1} = \bar{y}_1 = y_{1\bullet}$. The same analysis holds for the other treatment settings so that $\hat{\mu}_i = \frac{\sum_{j=1}^{J_i} y_{ij}}{J_i} = \bar{y}_i = y_{i\bullet}$, $i = 1,\ldots,I$.

The treatment model $Y_{ij} = \mu + \alpha_i + \sigma_{1W} Z_{ij}$, $j = 1,\ldots,J_i$, and $\sum_{i=1}^{I} J_i\alpha_i = 0$ has $I+1$ parameters (namely $\mu, \alpha_1, \ldots, \alpha_I$). The constraint on the treatment effects that $\sum_{i=1}^{I} J_i\alpha_i = 0$ is needed to make the parameters of the model and hence the OLS estimates unique. The OLS estimates are $\hat{\mu} = \frac{\sum_{i=1}^{I} J_i\hat{\mu}_i}{\sum_{i=1}^{I} J_i} = \frac{\sum_{i=1}^{I} J_i y_{i\bullet}}{\sum_{i=1}^{I} J_i} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i} y_{ij}}{\sum_{i=1}^{I} J_i} = y_{\bullet\bullet}$, where $\hat{\mu} = y_{\bullet\bullet}$ is called the grand mean (or overall mean) of the observations. Then, $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = y_{i\bullet} - y_{\bullet\bullet}$, $i = 1,\ldots,I$.

Sum of Squared Errors

As in Chapters 11 and 12, the minimized value of the *SS* function is the sum of squared error and is crucial for our analysis. Now,

$\min[SS(m_1,\ldots,m_I)] = SS(\hat{\mu}_1,\ldots,\hat{\mu}_I) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - y_{i\bullet})^2 = \sum_{i=1}^{I}(J_i - 1)s_i^2$, where $s_i^2 = \frac{\sum_{j=1}^{J_i}(y_{ij} - y_{i\bullet})^2}{J_i - 1}$ is the usual sample variance estimator applied to the J_i observations from the *i*th setting of the treatment. Then the sum of squared error *SSE* is given by $\min[SS(m_1,\ldots,m_I)] = SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - y_{i\bullet})^2 = \sum_{i=1}^{I}(J_i - 1)s_i^2$, with $\sum_{i=1}^{I}(J_i - 1) = n - I$ degrees of freedom, where *n* is the total number of observations in the study.

Fisher's decomposition of the total sum of squares

I will now shift the discussion from a realized experiment to a planned experiment. That is, I will use the random variable notation. The total sum of squares is always $SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{\bullet\bullet})^2$, with $n - 1 = \sum_{i=1}^{I} J_i - 1$ degrees of freedom. Then

$$SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet} + Y_{i\bullet} - Y_{\bullet\bullet})^2 \text{ and}$$

$$SS_{Total} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{i\bullet} - Y_{\bullet\bullet})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}).$$

Recall that $SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet})^2$. Further,

$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{i\bullet} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})^2\sum_{j=1}^{J_i}1 = \sum_{i=1}^{I}J_i(Y_{i\bullet} - Y_{\bullet\bullet})^2 = SS_{Treatment}$. The sum of squares due to treatment settings is defined to be

$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{i\bullet} - Y_{\bullet\bullet})^2 = \sum_{i=1}^{I}J_i(Y_{i\bullet} - Y_{\bullet\bullet})^2 = SS_{Treatment}$ and has $I-1$ degrees of freedom.

Finally, $2\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet})(Y_{i\bullet} - Y_{\bullet\bullet}) = 2\sum_{i=1}^{I}(Y_{i\bullet} - Y_{\bullet\bullet})\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet}).$

Since $\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet}) = 0,$ the cross-product term is 0. This proves that $SS_{Total} = SSE + SS_{Treatment}$.

Analysis of Variance Table

These results are conventionally displayed in an analysis of variance table

Analysis of Variance Table
Complete Randomized Experiment

Source	Degrees of Freedom	Sum of Squares	Mean Square	F
Treatment	$I-1$	$\sum_{i=1}^{I}J_i(Y_{i\bullet} - Y_{\bullet\bullet})^2$	$SS_{Treatment}/(I-1)$	$\frac{MS_{Treatment}}{MSE}$
Error	$n-I$	$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{i\bullet})^2 = \sum_{i=1}^{I}(J_i - 1)S_i^2$	$SSE/(n-I)$	
Total	$n-1$	$\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - Y_{\bullet\bullet})^2$		

As in Chapters 11 and 12, the statistical estimate of the variance parameter in the model is the mean squared error. The model is $Y_{ij} = \mu + \alpha_i + \sigma_{1W} Z_{ij}$, for $i = 1, \ldots, I$ (where I is the number of treatment settings), $j = 1, \ldots, J_i$, and $\sum_{i=1}^{I} J_i \alpha_i = 0$. Then $\hat{\sigma}_{1W}^2 = MSE$.