Chapter 6: Multivariate Regression

Jingyuan Liu
Department of Statistics, School of Economics
Wang Yanan Institute for Studies in Economics
Xiamen University

Outline

Introduction

In this chapter, we consider the linear relationship between one or more y's (dependent or response variables) and one or more x's (independent or predictor variables).

1. **Simple linear regression:** one y and one x. E.g. predict college GPA based on an applicant's high school GPA.
2. **Multiple linear regression:** one y and several x's. E.g. improve our prediction of college GPA by using high school GPA, standardized test scores (such as SAT), and rating of high school.
3. **Multivariate (multiple) linear regression:** several y's and several x's. E.g. we may also wish to predict the number of years of college the person will complete or other performances in college.

Outline

This section is based on the Regression course. Refer to any regression book, e.g. "*Applied Linear Statistical Models*" by Kutner, Nachtsheim, Neter and Li. Therefore, we here only briefly review the key points without going to much details.

Multiple Linear Regression Model

To study the linear relationship between y and multiple x's, we build up the following linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q + \varepsilon.$$

Or with the sample $\{y_i, x_{i1}, \ldots, x_{iq}, i = 1, \ldots, n\}$, the sample regression model is

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_q x_{nq} + \varepsilon_n. \end{aligned}$$

Model Assumption

To fit the model, the following assumptions about the random error ε_i are imposed:

1. $E(\varepsilon_i) = 0$ for all $i = 1, 2, \ldots, n$.
2. $\text{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$.
3. $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

Or, from the perspective of y:

1. $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq}, i = 1, 2, \ldots, n$.
2. $\text{var}(y_i) = \sigma^2, i = 1, 2, \ldots, n$.
3. $\text{cov}(y_i, y_j) = 0$, for all $i \neq j$.

Matrix Presentation

Using matrix notation, the model can be represented by $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Thus the above assumptions can be rewritten as

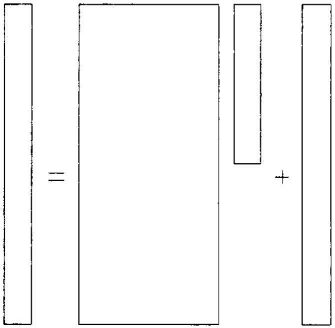
$$1.E(\varepsilon) = \mathbf{0}; \;\; 2.COV(\varepsilon) = \sigma^2\mathbf{I}$$

or in terms of y:

$$1.E(\mathbf{y}) = \mathbf{X}\beta; \;\; 2.COV(\mathbf{y}) = \sigma^2\mathbf{I}$$

Outline

For estimation and testing purposes we need to have $n > q$. Therefore, the matrix expression should have the following typical pattern:



Least Squares Estimation of $\boldsymbol{\beta}$

To estimate the coefficient $\boldsymbol{\beta}$, the **least squares estimates** can be computed to minimize the sum of squares of deviations

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_q x_{iq}$. The resulting estimator is

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

providing that $\mathbf{X}'\mathbf{X}$ is nonsingular (which ordinarily holds if $n > q + 1$ and no multicollinearity exists).

Properties of $\hat{\beta}$

- $\hat{\beta}$ is an unbiased estimator for β, i.e. $E(\hat{\beta}) = \beta$.
- $COV(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- **Gauss-Markov Theorem:** The least squares estimator $\hat{\beta}$ is the "best linear unbiased estimator" (BLUE) of β, i.e. it has minimum variance among all linear unbiased estimators.

An Estimator for σ^2

It can be shown that $E(SSE) = \sigma^2(n - q - 1)$, thus we can obtain an unbiased estimator of σ^2 as

$$s^2 = \frac{SSE}{n - q - 1} = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n - q - 1}$$

Decomposition of Sum of Squares

The SSE can be considered as "the variation in y due to random error" which cannot be captured by the regression model. So correspondingly, we would also have SSR to be "the variation in y that can be explained by the model" and " the total variation in y" SST. Mathematically,

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ with } n - q - 1 \text{ degrees of freedom}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \text{ with } q \text{ degrees of freedom}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ with } n - 1 \text{ degrees of freedom}$$

And it is easy to check that $SST = SSR + SSE$.

Outline

Test for Overall Regression

- In order to conduct the tests, we need distributional assumption. Specifically, assume $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.
- The first hypothesis is to investigate the overall linear relationship between y and x's:

$$H_0: \ \beta_1 = \ldots = \beta_q = 0 \text{ vs. } H_0: \ \text{at least one } \beta_j \text{ is not } 0.$$

- The test is based on the F statistic:

$$F = \frac{SSR/q}{SSE/(n-q-1)} = \frac{MSR}{MSE} \sim F(q, n{-}q{-}1) \text{ under } H_0.$$

Thus reject H_0 if $F > F_\alpha(q, n-q-1)$.

Lack-of-fit Test on a Subset of $\boldsymbol{\beta}$

For ease of presentation, $\boldsymbol{\beta}$ is partitioned into $\boldsymbol{\beta} = (\boldsymbol{\beta}_r', \boldsymbol{\beta}_d')'$. Then we may be interested in

$$H_0: \ \boldsymbol{\beta}_d = \mathbf{0} \text{ vs. } H_1: \ \boldsymbol{\beta}_d \neq \mathbf{0}.$$

That is to compare two models:

$$\text{Full model (f): } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\text{Reduced model (r): } \mathbf{y} = \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$$

where $\mathbf{X}_r$ is the reduced design matrix obtained by extracting the columns of $\mathbf{X}$ corresponding to $\boldsymbol{\beta}_r$.

The test statistic for the "lack of fit" (LOF) of the reduced model is

$$F = \frac{(SSR_f - SSR_r)/h}{SSE_f/(n-q-1)} = \frac{MSR_{LOF}}{MSE_f} \sim F(h, n-q-1) \text{ under } H_0,$$

where h designates the number of parameters in $\boldsymbol{\beta}_d$, or equivalently, the difference in the degrees of freedom between SSR_f and SSR_r. If $F > F_\alpha(h, n-q-1)$, reject H_0 and claim the effect of $\mathbf{X}_d$ is still significant in the presence of $\mathbf{X}_r$.

Test on a Single β_j

Sometimes we may want to examine the importance of X_j with the other x's are in the model. Specifically,

$$H_0: \ \beta_j = 0 \text{ vs. } H_1: \ \beta_j \neq 0.$$

Essentially, this is a special case of the lack-of-fit test, with $\boldsymbol{\beta}_d = \beta_j$. But by the equivalence between the squared $t(\nu)$ distribution and $F(1, \nu)$, we could apply the t test statistic

$$t = \frac{\hat{\beta}_j}{sd(\hat{\beta}_j)} \sim t(n-q-1) \text{ under } H_0,$$

where $sd(\hat{\beta}_j) = \sqrt{MSE_f(\mathbf{X}'\mathbf{X}^{-1})_{jj}}$ is the standard deviation of $\hat{\beta}_j$, and $(\mathbf{X}'\mathbf{X}^{-1})_{jj}$ is the jth diagonal element of $\mathbf{X}'\mathbf{X}^{-1}$.

Coefficient of Multiple Determination R^2

To check the model fit, one way is to utilize the **coefficient of (multiple) determination**, or more commonly referred as **squared multiple correlation**:

$$R^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{SSR}{SST}.$$

The **multiple correlation** is defined as $R = \sqrt{R^2}$.

Remarks:

- The F statsitic for the overall regression can be expressed in terms of R^2:

$$F = \frac{n-q-1}{q} \frac{R^2}{1-R^2}.$$

- The lack-of-fit F statistic can also be expressed as

$$F = \frac{(R_f^2 - R_r^2)/h}{(1-R_f^2)/(n-q-1)},$$

  where R_f^2 and R_r^2 are R^2 of the full and reduced model.
- R^2 never decreases when more variables are included, thus it can not be used to compare models with different model sizes.

Outline

Introduction to Variable Selection

In practice, one often has more x's than needed for predicting y - some x's could be discarded for the ease of interpretation and increased precision of model fit. To choose the "best" model, we need

1. A group formulation to select the subset of the full model
2. A criterion to evaluate the current model

Group Formulation

- **Subset selection:** (R: "*regsubsets*" in library "*leap*")
  Evaluate all subsets of the full model.
- **Stepwise selection:** (R: "*step*")
  - **Forward selection:**
    Start from the null model, add one variable at a time with the most improvement, stop when no improvement any more.
  - **Backward elimination:**
    Start from the full model with all candidate x's, delete one variable at a time that is most insignificant, stop when all remaining x's are needed.
  - **Stepwise regression:**
    Combination of forward and backward.

Evaluation Criteria

Criteria to evaluate the current model with model size k:

- ***p*-Value** of the individual t test for β_j
- **Adjusted R^2:**

$$R_a^2 = 1 - \frac{n-1}{n-k}(1 - R^2),$$

where $R^2 = SSR/SST$ is the coefficient of determination. Note that R^2 cannot be used for variable selection.

- **Mallow's C_p:**

$$C_p = \frac{SSE_k}{SSE_f/(n-q-1)} + 2k - n,$$

where SSE_k and SSE_f are the SSE for the current model and the full model with all x's, respectively.

- **Prediction sum of squares $PRESS_p$:**

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{ip})^2,$$

  where $\hat{y}_{ip}$ is the predicted value for y_i using the current candidate model, with the ith observation delected. This is realized by cross validation.

- **Generalized cross validation GCV:**

$$GCV = \frac{SSE_k}{(1 - k/n)^2}.$$

  GCV is an approximation to $PRESS_p$ when n is large.

- **Akaike information criterion AIC:**

$$AIC = \frac{SSE_k}{SSE_f/(n-q-1)} + 2k.$$

AIC is equivalent to Mallow's C_p in linear models, and tends to over-fit the model (conservative).

- **Bayesian information criterion BIC:**

$$BIC = \frac{SSE_k}{SSE_f/(n-q-1)} + k\log(n).$$

BIC is consistent with the true model when q is fixed and n goes to infinity.

- **Generalized information criterion GIC:**

$$GIC = \frac{SSE_k}{SSE_f/(n-q-1)} + k\tau_n,$$

where τ_n is a multiplier determined by n. AIC and BIC are the special cases of GIC.

Multiple Regression: Example

Example: The manager of a company wanted to study the relation between the employees' current salary (Y) and their starting salary (X_1), the number of working months for the current job (X_2) and that for the previous work experiences (X_3), and the number of years of education (X_4). 36 of the employees were randomly selected.

Part of the data are shown below.

```
> salary<-read.table("/Users/jingyuan/快盘/Teaching/Multivariate
Analysis/R code/Chap6/salary.csv",sep=",",header=T)
> salary[1:15,]
        y    x1 x2  x3 x4
1   79220 14010 98 115 15
2   79670 13260 98  26  8
3  186320 81240 96 199 19
4  161945 46260 96 120 19
5   74570 15510 95  46 12
6   86120 15810 93   8 16
7   91520 20760 92 168 17
8   82820 20010 90 205 12
9   75620 16260 90 191 15
10  82220 16260 88 252 12
11  78020 14760 88  38 12
12  76370 14010 87 123 16
13  78020 14760 86 367 12
14 120570 43740 85 134 20
15  83270 16260 85 438  8
```

We could fit the following multiple linear model.

```
> lm.salary<-lm(y~x1+x2+x3+x4,data=salary)  #build the multiple regression model
> summary(lm.salary)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = salary)

Residuals:
     Min       1Q   Median       3Q      Max
-12924.2  -4588.1   -269.6   1756.2  25215.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 48386.0620 11237.2882   4.306 0.000155 ***
x1              1.6831     0.1302  12.929 5.01e-14 ***
x2            -34.5520   130.2602  -0.265 0.792570
x3            -13.0004    13.7882  -0.943 0.353043
x4            808.3223   547.8017   1.476 0.150144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7858 on 31 degrees of freedom
Multiple R-squared:  0.919,    Adjusted R-squared:  0.9086
F-statistic: 87.95 on 4 and 31 DF,  p-value: < 2.2e-16
```

- How to check the model fit?
- What is the p-value for overall test? What does it mean?
- How to interpret the estimated coefficients?
- How to explain the p-values for each x-variable?
- What should we do next?

We can conduct the stepwise regression with AIC criterion.

```
> lm.step<-step(lm.salary,direction="both")
Start:  AIC=650.41
y ~ x1 + x2 + x3 + x4

       Df  Sum of Sq        RSS    AIC
- x2    1 4.3448e+06 1.9186e+09 648.49
- x3    1 5.4896e+07 1.9692e+09 649.43
<none>               1.9143e+09 650.41
- x4    1 1.3445e+08 2.0487e+09 650.85
- x1    1 1.0323e+10 1.2237e+10 715.19

Step:  AIC=648.49
y ~ x1 + x3 + x4

       Df  Sum of Sq        RSS    AIC
- x3    1 6.2078e+07 1.9807e+09 647.64
<none>               1.9186e+09 648.49
- x4    1 1.3011e+08 2.0487e+09 648.85
+ x2    1 4.3448e+06 1.9143e+09 650.41
- x1    1 1.0341e+10 1.2259e+10 713.26

Step:  AIC=647.64
y ~ x1 + x4

       Df  Sum of Sq        RSS    AIC
<none>               1.9807e+09 647.64
+ x3    1 6.2078e+07 1.9186e+09 648.49
+ x2    1 1.1527e+07 1.9692e+09 649.43
- x4    1 2.9640e+08 2.2771e+09 650.66
- x1    1 1.1654e+10 1.3635e+10 715.09
```

The result from stepwise regression is summarized as follows.

```
> summary(lm.step)

Call:
lm(formula = y ~ x1 + x4, data = salary)

Residuals:
   Min     1Q Median     3Q    Max
-13632  -4759   -615   1761  25076

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42097.165   5265.218   7.995 3.18e-09 ***
x1              1.631      0.117  13.934 2.22e-15 ***
x4           1039.260    467.671   2.222   0.0332 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7747 on 33 degrees of freedom
Multiple R-squared:  0.9162,  Adjusted R-squared:  0.9111
F-statistic: 180.4 on 2 and 33 DF,  p-value: < 2.2e-16
```

Multiple Regression: Remarks

Practically, after we fit the model, we should check for

- heteroscedasticity (equal variance) by the residual plots;
- multicollinearity by VIF;
- influential points / significant outliers by the residual plots or Cook's distance;
- normality by QQ plots or normality tests.

If one or more assumption is violated, data transformation or other remedy procedures are necessary.

Outline

We turn now to the **multivariate multiple linear regression model**, where "multivariate" refers to the dependent variables $\mathbf{y} = (Y_1, \ldots, Y_p)'$ and "multiple" pertains to the independent variables $\mathbf{x} = (X_1, \ldots, X_q)'$. Sometimes "multiple" is omitted.

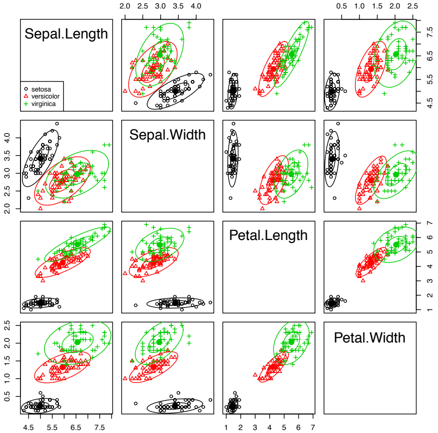
Multivariate Example: Iris Data

Recall the iris data from Chapter 1:

```
> library(car)
> some(iris)
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
12           4.8         3.4          1.6         0.2     setosa
14           4.3         3.0          1.1         0.1     setosa
16           5.7         4.4          1.5         0.4     setosa
50           5.0         3.3          1.4         0.2     setosa
71           5.9         3.2          4.8         1.8 versicolor
73           6.3         2.5          4.9         1.5 versicolor
114          5.7         2.5          5.0         2.0  virginica
117          6.5         3.0          5.5         1.8  virginica
129          6.4         2.8          5.6         2.1  virginica
147          6.3         2.5          5.0         1.9  virginica
```

```
> scatterplotMatrix(~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width | Species,
+ data=iris, smooth=FALSE, reg.line=FALSE, ellipse=TRUE,by.groups=TRUE, diagonal="none",
+ legend.pos="bottomleft")
```



The bivariate scatter plots reveal the relations between all pairs of variables (within each group). What if, however, our target is the relation between the sepal measurements (length and width) and the petal measurements?

Multivariate Data Structure

The n observed values of $\mathbf{y}$ can be listed as rows in the matrix:

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}.$$

The n values of $\mathbf{x}$ can be placed in the following $\mathbf{X}$ matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1q} \\ 1 & x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix}.$$

Multivariate (Multiple) Linear Model

- Since each column of $\mathbf{Y}$ will need different coefficient β's, we should have a **coefficient matrix** $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$.
- Therefore, the multivariate model is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{\Xi},$$

where $\mathbf{Y}$ and $\mathbf{\Xi}$ are $n \times p$ matrices, $\mathbf{X}$ is $n \times (q+1)$, $\mathbf{B}$ is $(q+1) \times p$.

Multivariate Model: Example

We illustrate the multivariate model with $p = 2$ and $q = 3$:

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{pmatrix}$$

The model for the first column of $\mathbf{Y}$ is

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \\ \beta_{31} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{pmatrix}$$

The model for the second column of $\mathbf{Y}$ is

$$\begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{02} \\ \beta_{12} \\ \beta_{22} \\ \beta_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n2} \end{pmatrix}$$

Multivariate Model: Assumptions

By analogy with the univariate case, the assumptions needed for estimation are as follows:

1. $E(\mathbf{Y}) = \mathbf{XB}$, or $E(\mathbf{\Xi}) = \mathbf{O}$.
2. $COV(\mathbf{y}_i) = \mathbf{\Sigma}$ for all $i = 1, \ldots, n$.
3. $COV(\mathbf{y}_i, \mathbf{y}_k) = \mathbf{O}$ for all $i \neq k$.

Note that the elements within each row of $\mathbf{Y}$ are correlated, with covariance matrix $\mathbf{\Sigma}$, but are uncorrelated with elements from other rows.

Outline

Multivariate Least Squares Estimation

Analogous to the univariate case, we estimate $\mathbf{B}$ by

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

We call $\hat{\mathbf{B}}$ the **least squares estimator** of $\mathbf{B}$, since it "minimizes" the matrix $\mathbf{E}$ analogous to SSE:

$$\mathbf{E} = \hat{\boldsymbol{\Xi}}'\hat{\boldsymbol{\Xi}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}$$

in the sense of $tr(\mathbf{E})$ and $|\mathbf{E}|$.

Properties of $\hat{\mathbf{B}}$

(1) The jth column of $\hat{\mathbf{B}}$ is the usual least squares estimate β for the jth dependent variable Y_j, $j = 1, \ldots, p$. That is, denote the p columns of $\mathbf{Y}$ by $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(p)}$, then

$$\begin{aligned}\hat{\mathbf{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(p)}) \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(1)}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(2)}, \ldots, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{(p)}] \\ &= [\hat{\boldsymbol{\beta}}_{(1)}, \hat{\boldsymbol{\beta}}_{(2)}, \ldots, \hat{\boldsymbol{\beta}}_{(p)}].\end{aligned}$$

(2) All $\hat{\beta}_{jk}$'s in $\hat{\mathbf{B}}$ are correlated with each other - it is the reason that we need multivariate tests for hypotheses about $\mathbf{B}$ instead of separately univariate tests.

(3) $\hat{\mathbf{B}}$ is unbiased, i.e. $E(\hat{\mathbf{B}}) = E(\mathbf{B})$. Furthermore, it is BLUE for $\mathbf{B}$.

(4) The covariance matrix between the columns of $\hat{\mathbf{B}}$ is

$$COV(\hat{\boldsymbol{\beta}}_{(j)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{jk}(\mathbf{X}'\mathbf{X})^{-1},$$

where σ_{jk} is the covariance between Y_j and Y_k.

An Estimator for $\boldsymbol{\Sigma}$

By analogy with the univariate case, an unbiased estimator of $\boldsymbol{\Sigma} = COV(\mathbf{y}_i)$ is given by

$$\mathbf{S}_e = \frac{\mathbf{E}}{n-q-1} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})}{n-q-1} = \frac{\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}}{n-q-1}.$$

Remark:
We could also decompose the total sum of squares into those due to regression and those due to random error:

$$\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}' = (\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}) + (\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}') \equiv \mathbf{E} + \mathbf{H},$$

corresponding to $SST = SSE + SSR$ in the univariate model.

Estimate of **B**: Iris Data

```
> # fit the multivariate model
> p<-q<-2
> n<-dim(iris)[1]
> mod.iris<-lm(cbind(Sepal.Length, Sepal.Width)~cbind(Petal.Length,
+              Petal.Width),data=iris)
> (B.hat<-mod.iris$coeff)   # estimate coefficient matrix B
                                          Sepal.Length Sepal.Width
(Intercept)                                  4.1905824   3.5870492
cbind(Petal.Length, Petal.Width)Petal.Length  0.5417772  -0.2571378
cbind(Petal.Length, Petal.Width)Petal.Width  -0.3195506   0.3640421
```

Estimate of $\boldsymbol{\Sigma}$: Iris Data

```
> summary(Manova(mod.iris))

Type II MANOVA Tests:

Sum of squares and products for error:
             Sepal.Length Sepal.Width
Sepal.Length     23.88069    14.49716
Sepal.Width      14.49716    22.27463

------------------------------------------

Term: cbind(Petal.Length, Petal.Width)

Sum of squares and products for the hypothesis:
             Sepal.Length Sepal.Width
Sepal.Length     78.28764  -20.819824
Sepal.Width     -20.81982    6.032303

Multivariate Tests: cbind(Petal.Length, Petal.Width)
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            2  0.900783  60.2316      4    294 < 2.22e-16 ***
Wilks             2  0.112817 144.3376      4    292 < 2.22e-16 ***
Hotelling-Lawley  2  7.743329 280.6957      4    290 < 2.22e-16 ***
Roy               2  7.727729 567.9881      2    147 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

E

H

The estimator $\mathbf{S}_e$ of $\boldsymbol{\Sigma}$ can be computed directly from $\mathbf{E}$:

```
> names(summary(Manova(mod.iris)))
[1] "type"                "repeated"            "multivariate.tests"
[4] "univariate.tests"    "pval.adjustments"    "sphericity.tests"
[7] "SSPE"
> (E<-summary(Manova(mod.iris))$SSPE)
             Sepal.Length Sepal.Width
Sepal.Length     23.88069    14.49716
Sepal.Width      14.49716    22.27463
> (Se<-E/(n-q-1))
             Sepal.Length Sepal.Width
Sepal.Length   0.16245370  0.09862012
Sepal.Width    0.09862012  0.15152810
```

E, and hence $\mathbf{S}_e$, can also be computed using the formulas:

```
> attach(iris)
> Y<-cbind(Sepal.Length, Sepal.Width)
> X<-cbind(rep(1,n),Petal.Length, Petal.Width)
> (E<-t(Y)%*%Y-t(B.hat)%*%t(X)%*%Y)
             Sepal.Length Sepal.Width
Sepal.Length     23.88069    14.49716
Sepal.Width      14.49716    22.27463
> (Se<-E/(n-q-1))
             Sepal.Length Sepal.Width
Sepal.Length   0.16245370  0.09862012
Sepal.Width    0.09862012  0.15152810
```

Outline

Test for Overal Regression

As in the univariate case, we first consider the hypothesis

$$H_0: \ \mathbf{B}_1 = \mathbf{O} \text{ vs, } H_1: \ \mathbf{B}_1 \neq \mathbf{O}$$

where $\mathbf{B}_1$ includes all rows of $\mathbf{B}$ except the first row:

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_0' \\ \mathbf{B}_1 \end{pmatrix} = \left(\begin{array}{cccc} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \hline \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{array}\right).$$

H_0 means that none of the x's predicts any of the y's. The idea of the tests are comparing $\mathbf{E}$ and $\mathbf{H}$.

Wilks' Lambda Test

- Test statistic:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \frac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'|}$$

- Alternative expressions of Λ:
  - Let $\lambda_1, \ldots, \lambda_s$ be the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ and $s = \min(p, q)$. Then Λ can be also expressed as

$$\Lambda = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}.$$

  - Let $\mathbf{S}$, $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ be the sample covariance matrices of $(\mathbf{Y}, \mathbf{X})$, $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then

$$\Lambda = \frac{|\mathbf{S}|}{|\mathbf{S}_{xx}||\mathbf{S}_{yy}|}.$$

- Null distribution:
  Λ is distributed as the **Wilks' lambda** distribution $\Lambda(p, q, n-q-1)$ when H_0 is true.
- Rejection region:

$$\Lambda < \Lambda_{\alpha}(p, q, n-q-1).$$

Note that we here reject H_0 for small Λ value. The quantiles of the Wilks' lambda distribution can be found in Table A.9 in the book.

Part of Wilks' Lambda Table

Table A.9. Lower Critical Values of Wilks Λ, $\alpha = .05$

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_s$ are eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Reject H_0 if $\Lambda \leq$ table value. [a] Multiply entry by 10^{-3}.

	ν_H											
ν_E	1	2	3	4	5	6	7	8	9	10	11	12
						$p = 1$						
1	6.16[a]	2.50[a]	1.54[a]	1.11[a]	.868[a]	.712[a]	.603[a]	.523[a]	.462[a]	.413[a]	.374[a]	.341[a]
2	.098	.050	.034	.025	.020	.017	.015	.013	.011	.010	9.28[a]	8.51[a]
3	.229	.136	.097	.076	.062	.053	.046	.041	.036	.033	.030	.028
4	.342	.224	.168	.135	.113	.098	.086	.076	.069	.063	.058	.053
5	.431	.302	.236	.194	.165	.144	.128	.115	.104	.096	.088	.082
6	.501	.368	.296	.249	.215	.189	.169	.153	.140	.129	.119	.111
7	.556	.425	.349	.298	.261	.232	.209	.190	.175	.161	.150	.140
8	.601	.473	.396	.343	.303	.271	.246	.225	.208	.193	.180	.169
9	.638	.514	.437	.382	.341	.308	.281	.258	.239	.223	.209	.196

Properties of Wilks' Lambda

- $\Lambda(p, m, n)$ can be approximated with a χ^2 distribution:

$$\left(\frac{p-n+1}{2} - m\right) \log \Lambda(p, m, n) \sim \chi^2(n, p) \text{ approximately.}$$

- Wilks' lambda can be related to the F distribution:

$$\frac{1-\Lambda(p, m, 1)}{\Lambda(p, m, 1)} \sim \frac{p}{m-p+1} F(p, m-p+1),$$
$$\frac{1-\sqrt{\Lambda(p, m, 2)}}{\sqrt{\Lambda(p, m, 2)}} \sim \frac{p}{m-p+1} F(2p, 2(m-p+1))$$

- Symmetry of λ: $\Lambda(p, q, n-q-1) \sim \Lambda(q, p, n-p-1)$.

Roy's Test

- Intuition:

$$\lambda_1 = \max_{\mathbf{a}} \frac{\mathbf{a}'\mathbf{H}\mathbf{a}}{\mathbf{a}'\mathbf{E}\mathbf{a}}$$

  where the maximizer $\mathbf{a} = \mathbf{e}_1$ is the eigenvector associated with λ_1.

- Test statistics:

$$\theta = \frac{\lambda_1}{1+\lambda_1},$$

  where λ_1 is the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$.

- Null distribution:
  The critical values for θ are given in Table A.10 in the book, with the parameters

$$s = \min(p, q), \ m = \frac{1}{2}(|q-p|-1), \ N = \frac{1}{2}(n-q-p-2).$$

- Rejection region:

$$\theta > \theta_\alpha(s, m, N).$$

Part of Roy's Upper Quantile Table

	m								
N	0	1	2	3	4	5	7	10	15
					$s = 2$				
5	.565	.651	.706	.746	.776	.799	.834	.868	.901
10	.374	.455	.514	.561	.598	.629	.679	.732	.789
15	.278	.348	.402	.446	.483	.515	.567	.627	.696
20	.221	.281	.329	.369	.404	.434	.486	.546	.620
25	.184	.236	.278	.314	.346	.375	.424	.484	.558
30	.157	.203	.241	.274	.303	.330	.376	.433	.507
40	.122	.159	.190	.218	.243	.266	.306	.359	.428
50	.099	.130	.157	.180	.202	.222	.259	.306	.370
60	.084	.110	.133	.154	.173	.191	.223	.266	.326
80	.064	.085	.103	.119	.135	.149	.176	.211	.263
120	.043	.058	.070	.082	.093	.104	.123	.150	.190
240	.022	.030	.036	.042	.048	.054	.065	.080	.103
					$s = 3$				
5	.669	.729	.770	.800	.822	.840	.867	.894	.920
10	.472	.537	.586	.625	.656	.683	.725	.770	.819
15	.362	.422	.469	.508	.541	.569	.616	.669	.730
20	.293	.346	.390	.427	.458	.486	.533	.589	.656

Pillai's Test

- Test statistics:

$$V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i},$$

  where λ_i's are the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

- Null distribution:
  The critical values for $V^{(s)}$ are given in Table A.11 in the book, with the same parameters as in the Roy's test.

- Rejection region:

$$V^{(s)} > V_{\alpha}^{(s)}(s, m, N).$$

Part of Pillai's Upper Quantile Table

Table A.11. Upper Critical Values of Pillai's Statistic $V^{(s)}$, $\alpha = .05$

$$V^{(s)} = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_s$ are eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. Reject H_0 if $V^{(s)}$ exceeds table value. The parameters s, m, and N are defined in Table A.10.

	N													
m	0	1	2	3	4	5	6	7	8	9	10	15	20	25
							$s = 2$							
0	1.536	1.232	1.031	.890	.782	.698	.629	.573	.526	.485	.451	.333	.263	.218
1	1.706	1.452	1.258	1.109	.991	.896	.817	.751	.694	.646	.604	.455	.364	.304
2	1.784	1.573	1.397	1.254	1.137	1.039	.956	.886	.825	.772	.725	.556	.451	.379
3	1.829	1.649	1.492	1.358	1.245	1.149	1.065	.993	.930	.875	.825	.643	.526	.445
4	1.859	1.703	1.560	1.436	1.329	1.235	1.153	1.081	1.018	.961	.910	.719	1.594	.506
5	1.880	1.742	1.613	1.497	1.395	1.305	1.226	1.155	1.091	1.034	.983	.786	.655	.561
6	1.895	1.772	1.654	1.546	1.450	1.364	1.286	1.217	1.154	1.098	1.046	.846	.710	.612
7	1.907	1.796	1.687	1.586	1.495	1.413	1.338	1.270	1.209	1.153	1.102	.901	.761	.658
8	1.917	1.815	1.714	1.620	1.534	1.455	1.383	1.317	1.257	1.202	1.151	.950	.808	.702
9	1.924	1.831	1.737	1.649	1.567	1.491	1.422	1.358	1.299	1.245	1.195	.995	.851	.743
10	1.931	1.844	1.757	1.673	1.595	1.523	1.456	1.394	1.337	1.284	1.235	1.036	.891	.781
15	1.951	1.888	1.822	1.758	1.695	1.636	1.580	1.527	1.477	1.430	1.386			
20	1.963	1.913	1.860	1.807	1.756	1.706	1.658	1.612	1.568	1.527	1.487			
25	1.969	1.929	1.885	1.840	1.796	1.753	1.711	1.671	1.632	1.595	1.559			

Lawley-Hotelling Test

- Test statistics:

$$U^{(s)} = \sum_{i=1}^{s} \lambda_i,$$

  where λ_i's are the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

- Null distribution:
  The upper critical values for $\nu_E U^{(s)}/\nu_H$ are given in Table A.12 in the book, where $\nu_E = n - q - 1$ and $\nu_H = q$.

- Rejection region:

$$\nu_E U^{(s)}/\nu_H > \text{table value}.$$

Part of Lawley–Hotelling Table

Table A.12. Upper Critical Values for the Lawley–Hotelling Test Statistic, $\alpha = .05$

The test statistic is $\nu_E U^{(s)}/\nu_H$, where $U^{(s)}$ is the Lawley–Hotelling statistic. Reject H_0 if $\nu_E U^{(s)}/\nu_H >$ table value.

	ν_H										
ν_E	2	3	4	5	6	8	10	12	15	20	25
						$p = 2$					
2[a]	9.8591	10.659	11.098	11.373	11.562	11.952	11.804	12.052	12.153	12.254	12.316
3	58.428	58.915	59.161	59.308	59.407	59.531	59.606	59.655	59.705	59.755	59.785
4	23.999	23.312	22.918	22.663	22.484	22.250	22.104	22.003	21.901	21.797	21.733
5	15.639	14.864	14.422	14.135	13.934	13.670	13.504	13.391	13.275	13.156	13.083
6	12.175	11.411	10.975	10.691	10.491	10.228	10.063	9.9489	9.8320	9.7118	9.6381
7	10.334	9.5937	9.1694	8.8927	8.6975	8.4396	8.2765	8.16399	8.0480	7.9285	7.8549
8	9.2069	8.4881	8.0752	7.8054	7.6145	7.3614	7.2008	7.0896	6.9748	6.8560	6.7826
10	7.9095	7.2243	6.8294	6.5702	6.3860	6.1405	5.9837	5.8745	5.7612	5.6433	5.5701
12	7.1902	6.5284	6.1461	5.8942	5.7147	5.4744	5.3200	5.2122	5.0997	4.9820	4.9085
14	6.7350	6.0902	4.7168	5.4703	5.2941	5.0574	4.9048	4.7977	4.6856	4.5678	4.4939
16	6.4217	5.7895	5.4230	5.1804	5.0067	4.7727	4.6213	4.5147	4.4028	4.2846	4.2102
18	6.1932	5.5708	5.2095	4.9700	4.7982	4.5663	4.4157	4.3094	4.1976	4.0791	4.0042
20	6.0192	5.4046	5.0475	4.8105	4.6402	4.4099	4.2600	4.1539	4.0420	3.9231	3.8477
25	5.7244	5.1237	4.7741	2.5415	4.3740	4.1465	3.9977	3.8919	3.7798	3.6598	3.5832
30	5.5401	4.9487	4.6040	4.3743	4.2086	3.9829	3.8347	3.7291	3.6166	3.4957	3.4181

Comparison of the Test Statistics

- All four tests have the same probability of type I error α.
- When H_0 is false, the power ranking of the tests depends on the configuration of the population eigenvalues, which are estimated by the sample eigenvalues $\lambda_1, \ldots, \lambda_s$ from $\mathbf{E}^{-1}\mathbf{H}$.
    - If the population eigenvalues are equal or nearly equal, the power ranking is $V^{(s)} \geq \Lambda \geq U^{(s)} \geq \theta$.
    - If only one population eigenvalues is nonzero, the powers are reversed: $V^{(s)} \leq \Lambda \leq U^{(s)} \leq \theta$.

Overall Test: Iris Data

Back to the iris data example, the overall significance of petal length and width can be assessed by

```
> # overall significance tests
> summary(Manova(mod.iris))$multivariate.test
$`cbind(Petal.Length, Petal.Width)`

Sum of squares and products for the hypothesis:
             Sepal.Length Sepal.Width
Sepal.Length     78.28764  -20.819824
Sepal.Width     -20.81982    6.032303

Sum of squares and products for error:
             Sepal.Length Sepal.Width
Sepal.Length     23.88069    14.49716
Sepal.Width      14.49716    22.27463

Multivariate Tests: cbind(Petal.Length, Petal.Width)
                 Df test stat approx F num Df den Df     Pr(>F)    
Pillai            2  0.900783  60.2316      4    294 < 2.22e-16 ***
Wilks             2  0.112817 144.3376      4    292 < 2.22e-16 ***
Hotelling-Lawley  2  7.743329 280.6957      4    290 < 2.22e-16 ***
Roy               2  7.727729 567.9881      2    147 < 2.22e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lack-of-fit Test on a Subset of $\mathbf{B}$

For ease of presentation, $\mathbf{B}$ is partitioned into $\mathbf{B} = (\mathbf{B}_r', \mathbf{B}_d')'$. Then we may be interested in

$$H_0: \ \mathbf{B}_d = \mathbf{O} \text{ vs. } H_1: \ \mathbf{B}_d \neq \mathbf{O}.$$

That is to compare two models:

$$\begin{aligned} &\text{Full model (f): } \mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{\Xi} \\ &\text{Reduced model (r): } \mathbf{Y} = \mathbf{X}_r\mathbf{B}_r + \mathbf{\Xi} \end{aligned}$$

where $\mathbf{X}_r$ is the reduced design matrix obtained by extracting the columns of $\mathbf{X}$ corresponding to $\mathbf{B}_r$.

- Lack-of-fit sum of squares matrix:

$$\mathbf{H}_{LOF} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y},$$

  where $\mathbf{H}_{LOF}$ is the difference in the "regression sum of squares matrix" between the full and reduced model.

- Wilks' test statistic:

$$\Lambda(x_{q-h+1}, \dots, x_q | x_1, \dots, x_{q-h}) = \frac{|\mathbf{E}|}{|\mathbf{E}_r|} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_{LOF}|} = \frac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}|},$$

  where h is the number of rows in $\mathbf{B}_d$, $\mathbf{E}$ and $\mathbf{E}_r$ are the "residual sum of squares matrix" of the full and reduced model, respectively.

- Another expression of the test statistic:

$$
\begin{aligned}
\Lambda(x_{q-h+1}, \ldots, x_q | x_1, \ldots, x_{q-h}) &= \frac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}|} \\
&= \frac{\dfrac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'|}}{\dfrac{|\mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}_r'\mathbf{X}_r'\mathbf{Y}|}{|\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'|}} \\
&= \frac{\Lambda_f}{\Lambda_r},
\end{aligned}
$$

where Λ_f and Λ_r are the overall Wilks' test statistics for the full and reduced model, respectively.

- Null distribution:

$$\Lambda(x_{q-h+1}, \ldots, x_q | x_1, \ldots, x_{q-h}) \sim \Lambda(p, h, n-q-1) \text{ under } H_0.$$

- Rejection region:

$$\Lambda(x_{q-h+1}, \ldots, x_q | x_1, \ldots, x_{q-h}) < \Lambda_\alpha(p, h, n - q - 1)$$

In Table A.9, $s = \min(p, h)$, $\nu_H = h$ and $\nu_E = n - q - 1$.

Lack-of-fit Test: Iris Data

For predicting the sepal length and sepal width, do we actually need both petal length and petal width?

```
> # lack of fit test
> mod2.iris<-lm(cbind(Sepal.Length, Sepal.Width)~Petal.Length)
> anova(mod2.iris,mod.iris,test="Wilks")
Analysis of Variance Table

Model 1: cbind(Sepal.Length, Sepal.Width) ~ Petal.Length
Model 2: cbind(Sepal.Length, Sepal.Width) ~ cbind(Petal.Length, Petal.Width)
  Res.Df Df Gen.var.   Wilks approx F num Df den Df    Pr(>F)
1    148     0.13126
2    147 -1  0.12203 0.85265   12.616      2    146 8.836e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Outline

Prediction

For a new observation $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0q})'$, we can predict the response by $\hat{\mathbf{y}}_0 = \hat{\mathbf{B}}'\mathbf{x}_0$. Furthermore, we can provide the interval predictions, which are also natural extension of the univariate ones. We only give the results here, including

- confidence interval for $E(\mathbf{y}_0)$
- prediction interval for the future observation $\mathbf{y}_0$

Confidence Interval for $E(\mathbf{y}_0)$

- Confidence interval for the jth component of $E(\mathbf{y}_0)$:

$$\hat{\beta}'_{(j)}\mathbf{x}_0 \pm t_{\alpha/2}(n-q-1)\sqrt{s_{jj}[\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]}$$

where $\hat{\beta}_{(j)}$ is the jth column of $\hat{\mathbf{B}}$ and s_{jj} is the jth diagonal element of $\mathbf{E}/(n-q-1)$.

- Simultaneous confidence intervals for all the p components of $E(\mathbf{y}_0)$:

$$\hat{\beta}'_{(j)}\mathbf{x}_0 \pm \sqrt{T^2_{\alpha}(n-q-1)s_{jj}[\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]}$$

Prediction Interval for $\mathbf{y}_0$

- Prediction interval for the jth component of $\mathbf{y}_0$:

$$\hat{\beta}'_{(j)}\mathbf{x}_0 \pm t_{\alpha/2}(n-q-1)\sqrt{s_{jj}[1+\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]}$$

  where $\hat{\beta}_{(j)}$ is the jth column of $\hat{\mathbf{B}}$ and s_{jj} is the jth diagonal element of $\mathbf{E}/(n-q-1)$.

- Simultaneous prediction intervals for all the p components of $\mathbf{y}_0$:

$$\hat{\beta}'_{(j)}\mathbf{x}_0 \pm \sqrt{T^2_\alpha(n-q-1)s_{jj}[1+\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0]}$$

Prediction: Remarks

Remarks:

- The prediction intervals are always wider than the confidence intervals due to the randomness of individual observations.
- The point estimate of $E(\mathbf{y}_0)$ can be obtained directly from the function "*predict*" in R. While the interval estimates cannot be directly obtained as in the univariate case.

Outline

2 Multivariate Multiple Regression
- Multivariate Multiple Regression Model
- Estimations
- Hypothesis Tests
- Prediction
- Variable Selection

Variable Selection

In the multivariate regression, two issues may occur:

- Some x's may be redundant in the presence of other x's
- Some y's may be deleted if they are not well predicted by any of the x's.

The group formulations of the univariate models can also be adopted to choose the "best" x's as well as y's:

- Forward selection
- Backward elimination
- Stepwise regression
- All subset selection (less used)

The criterion used is the **partial Wilks' Λ-statistic**.

- For selecting x's: After m x variables, denoted by $x_1, \dots, x_m$ have been selected, the next step studies

$$\Lambda(X_j | x_1, \dots, x_m) = \frac{\Lambda(x_1, x_2, \dots, x_m, X_j)}{\Lambda(x_1, x_2, \dots, x_m)}$$

- For selecting y's: After s y variables, denoted by $y_1, \dots, y_s$ have been selected, the next step studies

$$\Lambda(Y_j | y_1, \dots, y_s) = \frac{\Lambda(y_1, y_2, \dots, y_m, Y_j)}{\Lambda(y_1, y_2, \dots, y_s)}$$

Summary and Take-home Messages

- What is the difference between multivariate regression and multiple regression?
- How to matrix-represent the multivariate model?
- How to estimate the coefficient matrix and the covariance matrix of the error?
- How to conduct the overall test and the lack-of-fit test (the four matrix-based tests)?
- For a new subject $\mathbf{y}_0$, how to obtain the confidence interval of the expected value of $\mathbf{y}_0$ and the prediction interval of $\mathbf{y}_0$?