Displays for Statistics 5401/8401

Lecture 16

October 12, 2005

Christopher Bingham, Instructor

612-625-1024, kb@umn.edu
372 Ford Hall

Class Web Page

http://www.stat.umn.edu/~kb/classes/5401

Review

We are looking at univariate models of the form

y = (predictable part) + {unpredictable part}

where the predictable (fixed) part depends linearly on one or more parameters. The unpredictable (random) part has 0 mean.

I introduced three types of models for the predictable part:

Regression: $\beta_0 + Z_1\beta_1 + Z_2\beta_2 + \ldots Z_k\beta_k$

ANOVA: The predictable part is made up of a sum of subscripted parameters, typified by the one way case"

$$y_{ij} = \mu_j + \varepsilon_{ij} \text{ or } y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

ANACOVA: The predictable part is a combination of regression and ANOVA forms like

$$y_{ij} = \mu + \alpha_j + Z_{ij}\beta$$

For each of these linear model types you can always express any given model in at more than one *different* way.

Example: *Regression* with k = 2 predictors Z_1 and Z_2:

$$y = (\beta_0 + Z_1\beta_1 + Z_2\beta_2) + \{\varepsilon_i\}$$

Define new predictor variables and slopes

$$\tilde{Z}_1 \equiv (Z_1 + Z_2)/2,\ \tilde{Z}_2 \equiv (Z_1 - Z_2)/2$$

$$\tilde{\beta}_1 \equiv \beta_1 + \beta_2,\ \tilde{\beta}_2 \equiv \beta_1 - \beta_2$$

Then,

$$\beta_0 + Z_1\beta_1 + Z_2\beta_2 = \beta_0 + \tilde{Z}_1\tilde{\beta}_1 + \tilde{Z}_2\tilde{\beta}_2$$

so

$$y_i = (\beta_0 + \tilde{Z}_1\tilde{\beta}_1 + \tilde{Z}_2\tilde{\beta}_2) + \{\varepsilon_i\},$$

is a linear model with *different* coefficients and *different* predictor variables but equally well describing y. This is typical. Different parametrizations have different predictor variables.

Example: *One-way ANOVA*

Define $\mu_i = \mu + \alpha_i$, the group i mean.

Then

$$y_{ij} = (\mu_i) + \{\varepsilon_{ij}\},\ i = 1,\ldots,g,\ j = 1,\ldots,n_i$$

is the same model as

$$y_{ij} = (\mu + \alpha_i) + \{\varepsilon_{ij}\},\ i = 1,\ldots,g,\ j = 1,\ldots,n_i$$

but involves different parameters.

Changing restrictions on the α's changes the meaning of the α's, but not the actual model.

Example: If $\{\alpha_i\}$ satisfy $\sum_i \alpha_i = 0$ define

$\tilde{\mu} \equiv \mu_k$ and $\tilde{\alpha}_i \equiv \alpha_i - \alpha_k$, for some k

$$\tilde{\mu} + \tilde{\alpha}_i = (\mu + \alpha_k) + (\alpha_i - \alpha_k) = \mu + \alpha_i = \mu_i$$

so $y_{ij} = (\tilde{\mu} + \tilde{\alpha}_i) + \{\varepsilon_{ij}\}$ is the same model.

The $\tilde{\alpha}_i$'s satisfy the restriction $\tilde{\alpha}_k = 0$.

- Some programs assume $\tilde{\alpha}_1 = 0$ (k = 1)
- Others assume $\tilde{\alpha}_g = 0$ (k = g)
- MacAnova assumes $\sum_i \alpha_i = 0$

I want to collect all the linear parameters in a single vector. So define $\mathbf{b}$ to be the vector of all coefficients in a particular parametrization of the linear model.

- Multiple regression:

$$\mathbf{b} = [\beta_0, \beta_1, \ldots, \beta_k]'$$

- One-way ANOVA:

$\mathbf{b} = [\mu, \alpha_1, \ldots, \alpha_g]'$ or $\mathbf{b} = [\mu_1, \mu_2, \ldots, \mu_g]'$

- Two factor ANOVA with interaction:

$\mathbf{b} =$

$[\mu, \alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots, (\alpha\beta)_{11}, (\alpha\beta)_{21}, \ldots]'$

- One-way ANACOVA with g-groups and k predictors:

$$\mathbf{b} = [\mu, \beta_1, \ldots, \beta_k, \alpha_1, \alpha_2, \ldots, \alpha_g]'$$

ANOVA allows you to test one or more "linear" hypotheses about $\mathbf{b}$:

Null hypothesis

H_0: For $m \geq 1$ specific vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_m$, and m specific $\delta_1, \ldots, \delta_m$,

H_0: $\boldsymbol{\ell}_j'\mathbf{b} = \delta_j$, $1 \leq j \leq m$.

Usually $\delta_j = 0$ so H_0: $\boldsymbol{\ell}_j'\mathbf{b} = 0$, $1 \leq j \leq m$

Alternative hypothesis

H_1: For at least one j, $\boldsymbol{\ell}_j'\mathbf{b} \neq \delta_j$ $(\neq 0)$.

Note that $\boldsymbol{\ell}_j'\mathbf{b}$ is a linear combination of the parameters in $\mathbf{b}$

Examples with $\mathbf{b} = [\beta_0, \beta_1, \beta_2]'$,

H_0: $\beta_1 = \beta_2 = 0$ is the same as H_0: $\boldsymbol{\ell}_j'\mathbf{b} = 0$, $j=1, 2$ with $\boldsymbol{\ell}_1 = [0,1,0]'$, $\boldsymbol{\ell}_2 = [0,0,1]'$ and $\delta_1 = \delta_2 = 0$

H_0: $\beta_1 = \beta_2$ is the same as H_0: $\boldsymbol{\ell}_1'\mathbf{b}$, with $\boldsymbol{\ell}_1 = [0, 1, -1]'$ and $\delta_1 = 0$

Let $f_h \leq m$ be the number of <u>linearly independent</u> $\boldsymbol{\ell}_j$'s.

Vocabulary

f_h is the *hypothesis degrees of freedom*.

When $\boldsymbol{\beta} = [\mu_1, \mu_2, \ldots, \mu_g]'$, the hypothesis

H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$

can be expressed as

$\boldsymbol{\ell}_{jk}'\boldsymbol{\beta} = \mu_j - \mu_k = 0$, $\boldsymbol{\ell}_{jk} = [0 \ldots 1 \ldots -1 \ldots 0]'$ $1 \leq j < k \leq g$, where $\boldsymbol{\ell}_{jk}$ has with 1 in position j and -1 in position k.

Here $m = k(k-1)/2$. But all you need are the $g-1$ vectors $\boldsymbol{\ell}_{12}, \boldsymbol{\ell}_{13}, \ldots, \boldsymbol{\ell}_{1g}$ defining the contrasts $\mu_1 - \mu_2, \mu_1 - \mu_3, \ldots, \mu_1 - \mu_g$. These are linearly independent so $f_h = g - 1$.

This same definition will apply to multivariate models.

Principle of change of residual SS

Let $\hat{\boldsymbol{\beta}}^0$ be a least squares estimate of $\boldsymbol{\beta}$ when you assume H_0 to be true and $\hat{\boldsymbol{\beta}}^1$ be an estimate when you assume H_1 is true.

Example:

Regression with H_0: $\beta_{k-1} = \beta_k = 0$ ($f_h = 2$):

$$\hat{\boldsymbol{\beta}}^0 = [\hat{\beta}_0^0 \; \hat{\beta}_1^0 \; \ldots \; \hat{\beta}_{k-2}^0 \; 0 \; 0]'$$

$$\hat{\boldsymbol{\beta}}^1 = [\hat{\beta}_0^1 \; \hat{\beta}_1^1 \; \ldots \; \hat{\beta}_{k-2}^1 \; \hat{\beta}_{k-1}^1 \; \hat{\beta}_k^1]'$$

where

- $\hat{\beta}_j^0$, $j = 0, \ldots, k-2$ are the least squares coefficients in regression on $Z_1, \ldots, Z_{k-2}$, perhaps, when $k = 4$, from `regress("y=z1+z2")`
- $\hat{\beta}_j^1$, $j = 0, \ldots, k$ are LS coefficients in the full regression on $Z_1, \ldots, Z_k$, perhaps from `regress("y=z1+z2+z3+z4")`

Notation

$$RSS(H_0) = RSS(\hat{\beta}^0) = \sum_i \{y_i - \hat{y}_i(\hat{\beta}^0)\}^2$$
$$RSS(H_1) = RSS(\hat{\beta}^1) = \sum_i \{y_i - \hat{y}_i(\hat{\beta}^1)\}^2$$

$\hat{y}(\hat{\beta}^0)$ and $\hat{y}(\hat{\beta}^1)$ are the fitted values using estimates $\hat{\beta}^0$ and $\hat{\beta}^1$, that is the estimated predictable parts when you substitute $\hat{\beta}$ for the true parameter vector β.

$RSS(H_0)$ is the residual SS when you estimate β by $\hat{\beta}^0$ (assuming H_0 is true).

$RSS(H_i)$ is the residual SS when you estimate β by $\hat{\beta}^1$ (assuming H_1 is true).

Neither $RSS(H_0)$ or $RSS(H_1)$ depends on the parametrization used.

Therefore, you can use the most convenient parametrization. It may be different in computing $RSS(H_0)$ and $RSS(H_1)$.

Always $RSS(H_0) \geq RSS(H_1)$.

When $RSS(H_1)$ is a lot smaller than $RSS(H_0)$ it suggests that a model satisfying H_0 is inadequate.

This idea leads to a fundamental inference principle for linear hypotheses:

> The statistical significance of evidence *against* H_0 is determined from the *relative increase* in RSS when you assume the null hypothesis is true as compared to the RSS when you don't.

That is, significance depends on the ratio

$$\{RSS(H_0) - RSS(H_1)\}/RSS(H_1) = SS_h/SS_e$$
$$SS_h \equiv RSS(H_0) - RSS(H_1), \quad SS_e \equiv RSS(H_1)$$

When SS_h/SS_e is "large enough", H_0 is significantly worse than H_1 and is rejected.

In the case of normal errors, this principle comes from the *likelihood ratio statistic* $\lambda = \Lambda^{N/2}$, where

$$\Lambda = RSS(H_1)/RSS(H_0) = SS_e/(SS_h+SS_e)$$
$$= 1/(1 + SS_h/SS_e) = 1/(1 + (f_h/f_e)F)$$

You reject H_0 for "small" λ or Λ, which corresponds to "large" SS_h/SS_e or F.

Even when the errors are not normal, this has an intuitive appeal, since SS_h/SS_e is a scale free index of how much worse the H_0 fit is compared to the H_1 fit.

Looking ahead: In MANOVA

- SS_h becomes a p by p hypothesis matrix **H**
- SS_e becomes a p by p error matrix **E**
- Inference is based on comparing **H** with **E**.

To decide on when SS_h/SS_e is "large", you need its distribution when H_0 is true.

In the case of normal ε

$$F = (SS_h/f_h)/(SS_e/f_e) = (f_e/f_h)SS_h/SS_e$$

has an F_{f_h,f_e} distribution.

You reject H_0 for

- large F, that is, $F > F_{f_h,f_e}(\alpha)$

or

- small $\lambda = 1/(1 + (f_h/f_e)F)^{N/2}$, that is,

$$\lambda < 1/(1 + (f_h/f_e)F_{f_h,f_e}(\alpha))^{N/2}$$

or

$$\Lambda < 1/(1 + (f_h/f_e)F_{f_h,f_e}(\alpha))$$

The F-test is fairly robust against non-normality so it can be use fairly safely as long as the error distribution is not too far from normal.

LR theory says that, when H_0 is true

$$-2\log \lambda = N\log(1 + (f_h/f_e)F) \tilde{=} \chi^2_m,$$

where m = the number of linearly independent linear combinations of elements of **b** being tested. In this univariate case, $m = f_h$.

You can improve this large sample result by replacing N by a well chosen multiplier m(N) such that $m(N)/N \to 1$ as $N \to \infty$.

The best such m(N) for this problem is

$$m(N) = m_1 \equiv f_e + f_h/2 - 1.$$

That is the adjusted LR test statistic is

$$(f_e + f_h/2 - 1)\log(1 + (f_h/f_e)F) \tilde{=} \chi^2_{f_h}$$

You can use this to get approximate critical values for F from critical values for χ^2 without need of F-tables or **invF()**.

When f_e is large, this gives a very usable approximation for the F-distribution:

$$F \tilde{=} (f_e/f_h)(\exp(\chi^2_{f_h}/(f_e + f_h/2 - 1)) - 1)$$

Here are numerical comparisons with exact F-critical values with the approximate values assuming

$$(f_e + f_h/2 - 1)\log(1 + (f_h/f_e)F) \text{ is } \chi^2_{f_h}$$

$\alpha = .05$, $f_e = 30$			$\alpha = .05$, $f_e = 100$		
f_h	F	From χ^2	f_h	F	From χ^2
1	4.171	4.172	1	3.936	3.936
2	3.316	3.316	2	3.087	3.087
3	2.922	2.920	3	2.696	2.695
4	2.690	2.685	4	2.463	2.462
5	2.534	2.527	5	2.305	2.305
10	2.165	2.140	10	1.927	1.925
15	2.015	1.967	15	1.768	1.764
20	1.932	1.856	20	1.676	1.670
25	1.878	1.773	25	1.616	1.607

The approximation is better for larger f_e and smaller f_h.

To compute F you need to find residual sums of squares RSS(H_0) and RSS(H_1).

MacAnova **anova()** and **regress()** allow you to do this in a "black box" way.

Regression with k = 2 predictors:

Test H_0: $\beta_1 = \beta_2 = 0$.

ℓ_1=[0,1,0]', ℓ_2 = [0,0,1]', f_h = 2

Example using data from Table 4.3 of Draper and smith.

```
Cmd> y <- vector(10.98,11.13,12.51,8.4,9.27,8.73,6.36,\
       8.5,7.82,9.14,8.24,12.19,11.88,9.57,10.94,\
       9.58,10.09,8.11,6.83,8.88,7.68,8.47,8.86,10.36,11.08)

Cmd> x6 <- vector(20,20,23,20,21,22,11,23,21,20,\
       20,21,21,19,23,20,22,22,11,23,20,21,20,20,22)

Cmd> x8 <- vector(35.3,29.7,30.8,58.8,61.4,71.3,74.4,\
       76.7,70.7,57.5,46.4,28.9,28.1,39.1,46.8,48.5,59.3,\
       70,70,74.5,72.1,58.1,44.6,33.4,28.6)

Cmd> regress("y=1") # null hypothesis model y = beta_0
Model used is y=1
               Coef      StdErr           t
CONSTANT      9.424     0.32613      28.897

N: 25, MSE: 2.659, DF: 24, R^2: 0.00000
Regression F(0,24): undefined, Durbin-Watson: 1.1415
To see the ANOVA table type 'anova()'

Cmd> ss0 <- sum(RESIDUALS^2); ss0 # RSS(H_0)
(1)      63.816

Cmd> SS#it's also the last element in SS, computed by regress()
    CONSTANT      ERROR1
      2220.3      63.816

Cmd> ss0 <-  reverse(SS)[1]; ss0#this works with any size model
(1)      63.816
```



```
Cmd> regress("y=x6+x8") # alternative hypothesis model
Model used is y=x6+x8
               Coef      StdErr           t
CONSTANT      9.1269      1.1028      8.2761
x6           0.20282    0.045768      4.4314
x8         -0.072393   0.0079994     -9.0498

N: 25, MSE: 0.43767, DF: 22, R^2: 0.84912
Regression F(2,22): 61.904, Durbin-Watson: 2.1955
To see the ANOVA table type 'anova()'
```

Note: **regress()**, **anova()**, **manova()** and other linear and generalized linear model fitting commands create variables SS and DF.

```
Cmd> SS # SS for an ANOVA
    CONSTANT          x6          x8      ERROR1
      2220.3      18.342      35.845      9.6287

Cmd> DF # degrees of freedom for an ANOVA
    CONSTANT          x6          x8      ERROR1
           1           1           1          22

Cmd> ss1 <- sum(RESIDUALS^2); ss1 # RSS(H_1)
(1)      9.6287

Cmd> ss1 <- reverse(SS)[1]; ss1 # alternate
(1)      9.6287

Cmd> n <- 25; fh <- 2; fe <- reverse(DF)[1]

Cmd> fe
(1)          22
```

```
Cmd> ssh <- ss0 - ss1; ssh # hypothesis sum of squares
(1)      54.187

Cmd> sse <- ss1 # error sum of squares

Cmd> fstat <- (ssh/fh)/(sse/fe); fstat # F-statistic
(1)      61.904
```

This F-statistic is the same as the **Regression F(2,22)** printed by **regress()**:

```
Regression F(2,22): 61.904
```

Use **cumF()** to find a P-value

```
Cmd> cumF(fstat,fh,fe,upper:T)# P-value (very small)
(1)  9.2265e-10  Very strong evidence against H_0
```

Now a less standard null hypothesis:

Test H_0: $\beta_1 = \beta_2$, that is H_0: $\beta_1 - \beta_2 = 0$.

Re-parametrize with variables $\tilde{Z}_1$ & $\tilde{Z}_2$ with coefficients $\tilde{\beta}_1 = \beta_1+\beta_2$, $\tilde{\beta}_2 = \beta_1-\beta_2$ so H_0 becomes $\tilde{\beta}_2 = 0$.

```
Cmd> z1 <- (x6 + x8)/2

Cmd> z2 <- (x6 - x8)/2

Cmd> regress("y=z1") # restricted (Null) model
Model used is y=z1
                Coef      StdErr           t
CONSTANT      14.825     0.96266        15.4
z1           -0.1483    0.025775     -5.7537

N: 25, MSE: 1.1374, DF: 23, R^2: 0.59005
Regression F(1,23): 33.105, Durbin-Watson: 2.7833
To see the ANOVA table type 'anova()'

Cmd> ss0 <- reverse(SS)[1]; ss0 # or sum(RESIDUALS^2): RSS(H_0)
(1)      26.161
```


Now fit the full model.

```
Cmd> regress("y=z1 + z2",pval:T) # full model
Model used is y=z1 + z2
                Coef      StdErr           t     P-Value
CONSTANT      9.1269      1.1028      8.2761  3.3456e-08
z1           0.13042    0.048086      2.7123    0.012723
z2           0.27521    0.044778      6.1461   3.471e-06

N: 25, MSE: 0.43767, DF: 22, R^2: 0.84912
Regression F(2,22): 61.904, P-value: < 1e-08, Durbin-Watson:
2.1955
To see the ANOVA table type 'anova()'

Cmd> ss1 <- revese(SS)[1]; ss1# = RSS(H_1) = sum(RESIDUALS^2)
(1)      9.6287  Note ss1 is same as for regress("y=x1+x2")

Cmd> ssh <- ss0 - ss1; sse <- ss1 # hypothesis SS

Cmd> fh <- 1; fe <- reverse(DF)[1]; fe
(1)          22

Cmd> fstat <- (ssh/fh)/(sse/fe); fstat # F-statistic
(1)      37.774

Cmd> cumF(fstat,fh,fe,upper:T) # compute P-value
(1)   3.471e-06      Strong evidence against H_0
```

Note the P-value is the same as the P-value for z2 in the regression output. This is because the F-statistic is in fact t^2:

```
Cmd> 6.1461^2
(1)      37.775
```

Multivariate Linear Models

Just as for *univariate* linear models, multivariate linear models have the form

$\mathbf{y}$ = (predictable part) + {unpredictable part}

where the predictable part depends linearly on one or more parameters. The unpredictable (random) part has 0 mean. $\mathbf{y}$ and both the parts or p-dimensional vectors.

As before, ther are three forms of *multivariate* linear models (p > 1 response variables):

- *Multivariate* ***regression***
- *Multivariate* ***analysis of variance*** (MANOVA)
- *Multivariate* ***analysis of covariance*** (MANACOVA).

In all three, the response being modeled is a vector $\mathbf{y}$.

Multivariate regression

$\mathbf{Y} = \mathbf{Z}_0\boldsymbol{\beta}_0' + \mathbf{Z}_1\boldsymbol{\beta}_1' + \mathbf{Z}_2\boldsymbol{\beta}_2' + \ldots + + \mathbf{Z}_k\boldsymbol{\beta}_k' + \boldsymbol{\varepsilon}$

$\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_p]$ and $\boldsymbol{\varepsilon}$ with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ are n by p matrices, each row corresonding to a case

Each $\mathbf{Z}_j$ is n by 1, $\mathbf{Z}_0 = \mathbf{1}_N$

Each $\boldsymbol{\beta}_j = [\beta_{1j}, \beta_{2j}, \ldots \beta_{pj}]'$, is p by 1 so $\boldsymbol{\beta}_j'$ is 1 by p..

This is equivalent to p *univariate* multiple regressions, *each with the same independent variables.*

$\mathbf{Y}_\ell = \beta_{\ell 0} + \mathbf{Z}_1\beta_{\ell 1} + \mathbf{Z}_2\beta_{\ell 2} + \ldots + \mathbf{Z}_k\beta_{\ell k} + \boldsymbol{\varepsilon}_\ell$

$= \mathbf{Z}\mathbf{b}_\ell + \boldsymbol{\varepsilon}_\ell$, $\mathbf{b}_\ell = [\beta_{\ell 0}, \beta_{\ell 1}, \beta_{\ell 2}, \ldots, \beta_{\ell k}]'$

Don't confuse $\mathbf{b}_\ell$, the vector of coefficients for variable ℓ with $\boldsymbol{\beta}_j$, the vector of coefficients of $\mathbf{Z}_j$, one for each response.

Using matrices this is,

$$\mathbf{Y} = \mathbf{ZB} + \boldsymbol{\varepsilon}, \qquad n \text{ by } p$$
$$\mathbf{Z} = [\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_k], \qquad n \text{ by } k+1$$
$$\mathbf{B} = [\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_k]', \qquad k+1 \text{ by } p$$
$$= [\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_p], \text{ each } \mathbf{b}_j \; k+1 \text{ by } 1$$

- $\boldsymbol{\beta}_j'$ is a 1 by p *row* of $\mathbf{B}$, j = 0, ..., k
- $\boldsymbol{\beta}_j$ is the p-vector of coefficients of predictor $\mathbf{Z}_j$ for all response variables.
- $\mathbf{b}_\ell$ is a (k+1)-vector, column ℓ of $\mathbf{B}$
- $\mathbf{b}_\ell = [\beta_{\ell 0}, \beta_{\ell 1}, \beta_{\ell 2}, \ldots, \beta_{\ell k}]'$ are the coefficients for response variable $\mathbf{Y}_\ell$.
- A *column* $\mathbf{b}_\ell$ of $\mathbf{B}$ has all coefficients for a single *response* variable
- A *row* $\boldsymbol{\beta}_j'$ of $\mathbf{B}$ has coefficients of one *predictor* for all responses.

In all, there are (k+1)×p coefficients in $\mathbf{B}$ and kp when you omit the intercepts $\boldsymbol{\beta}_0$.