Lecture 5
Hypothesis Testing in Multiple Linear Regression

BIOST 515

January 20, 2004

Types of tests

- Overall test
- Test for addition of a single variable
- Test for addition of a group of variables

Overall test

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \epsilon_i$$

Does the $entire$ set of independent variables contribute significantly to the prediction of y?

Test for an addition of a single variable

Does the addition of one particular variable of interest add significantly to the prediction of y acheived by the other independent variables already in the model?

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \epsilon_i$$

Test for addition of a group of variables

Does the addition of some *group* of independent variables of interest add significantly to the prediction of y obtained through other independent variables already in the model?

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{i,p-1}\beta_{p-1} + x_{ip}\beta_p + \epsilon_i$$

The ANOVA table

Source of variation	Sums of squares	Degrees of freedom	Mean square	E[Mean square]
Regression	$SSR = \hat{\beta}'X'y - n\bar{y}^2$	p	$\frac{SSR}{p}$	$p\sigma^2 + \beta_R' X_C' X_C \beta_R$
Error	$SSE = y'y - \hat{\beta}'X'y$	$n-(p+1)$	$\frac{SSE}{n-(p+1)}$	σ^2
Total	$SSTO = y'y - n\bar{y}^2$	$n-1$		

X_C is the matrix of centered predictors:

$$X_C = \begin{pmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{pmatrix}$$

and $\beta_R = (\beta_1, \cdots, \beta_p)'$.

The ANOVA table for

$$y_i = \beta_0 + x_{i1}\beta 1 + x_{i2}\beta 2 + \cdots + x_{ip}\beta_p + \epsilon_i$$

is often provided in the output from statistical software as

Source of variation	Sums of squares	Degrees of freedom	F
Regression	x_1	1	
	$x_2 \vert x_1$	1	
	$\vdots$		
	$x_p \vert x_{p-1}, x_{p-2}, \cdots, x_1$	1	
Error	SSE	$n-(p+1)$	
Total	$SSTO$	$n-1$	

where $SSR =$
$SSR(x_1) + SSR(x_2|x_1) + \cdots + SSR(x_p|x_{p-1}, x_{p-2}, \ldots, x_1)$
and has p degrees of freedom.

Overall test

$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$
$H_1: \ \beta_j \neq 0$ for at least one $j, \ j = 1, \ldots, p$

Rejection of H_0 implies that at least one of the regressors, $x_1, x_2, \ldots, x_p$, contributes significantly to the model.

We will use a generalization of the F-test in simple linear regression to test this hypothesis.

Under the null hypothesis, $SSR/\sigma^2 \sim \chi^2_p$ and $SSE/\sigma^2 \sim \chi^2_{n-(p+1)}$ are independent. Therefore, we have

$$F_0 = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \sim F_{p,n-p-1}$$

Note: as in simple linear regression, we are assuming that $\epsilon_i \sim N(0, \sigma^2)$ or relying on large sample theory.

CHS example, cont.

$$y_i = \beta_0 + weight_i\beta_1 + height_i\beta_2 + \epsilon_i$$

```
> anova(lmwtht)
Analysis of Variance Table

Response: DIABP
           Df Sum Sq Mean Sq F value   Pr(>F)
WEIGHT      1   1289    1289 10.2240 0.001475 **
HEIGHT      1    120     120  0.9498 0.330249
Residuals 495  62426     126
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$F_0 = \frac{(1289 + 120)/2}{62426/495} = 5.59 > F_{2,495,.95} = 3.01$$

We reject the null hypothesis at $\alpha = .05$ and conclude that at least one of β_1 or β_2 is not equal to 0.

The overall F statistic is also available from the output of `summary()`.

```
> summary(lmwtht)

Call:
lm(formula = DIABP ~ WEIGHT + HEIGHT, data = chs)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.65777    8.91267   6.245 9.14e-10 ***
WEIGHT       0.04140    0.01723   2.403   0.0166 *
HEIGHT       0.05820    0.05972   0.975   0.3302
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.23 on 495 degrees of freedom
Multiple R-Squared: 0.02208,    Adjusted R-squared: 0.01812

F-statistic: 5.587 on 2 and 495 DF,  p-value: 0.003987
```

Tests on individual regression coefficients

Once we have determined that at least one of the regressors is important, a natural next question might be which one(s)?

Important considerations:

- Is the increase in the regression sums of squares sufficient to warrant an additional predictor in the model?
- Additional predictors will increase the variance of $\hat{y}$ - include only predictors that explain the response (note: we may not know this through hypothesis testing as confounders may not test significant but would still be necessary in the regression model).
- Adding an unimportant predictor may increase the residual mean square thereby reducing the usefulness of the model.

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ij}\beta_j + \cdots + x_{ip}\beta_p + \epsilon_i$$
$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

As in simple linear regression, under the null hypothesis

$$t_0 = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \sim t_{n-p-1}.$$

We reject H_0 if $|t_0| > t_{n-p-1,1-\alpha/2}$.

This is a **partial test** because $\hat{\beta}_j$ depends on all of the other predictors $x_i,\ i \neq j$ that are in the model. Thus, this is a test of the contribution of x_j given the other predictors in the model.

CHS example, cont.

$$y_i = \beta_0 + weight_i \beta_1 + height_i \beta_2 + \epsilon_i$$

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$, given that $weight$ is in the model.

From the ANOVA table, $\hat{\sigma^2} = 126.11$.

$$C = (X'X)^{-1} = \begin{pmatrix} 0.6299 & 2.329 \times 10^{-4} & -4.05 \times 10^{-3} \\ 2.329 \times 10^{-4} & 2.353 \times 10^{-6} & -3.714 \times 10^{-6} \\ -4.050 \times 10^{-3} & -3.714 \times 10^{-6} & 2.828 \times 10^{-5} \end{pmatrix}$$

$$t_0 = 0.05820/\sqrt{126.11 \times 2.828 \times 10^{-5}} = 0.975 < t_{495,.975} = 1.96$$

Therefore, we fail to reject the null hypothesis.

Tests for groups of predictors

Often it is of interest to determine whether a group of predictors contribute to predicting y given another predictor or group of predictors are in the model.

- In CHS example, we may want to know if age, height and sex are important predictors given weight is in the model when predicting blood pressure.
- We may want to know if additional powers of some predictor are important in the model given the linear term is already in the model.
- Given a predictor of interest, are interactions with other confounders of interest as well?

Using sums of squares to test for groups of predictors

Determine the contribution of a predictor or group of predictors to SSR given that the other regressors are in the model using the **extra-sums-of-squares** method.

Consider the regression model with p predictors

$$y = X\beta + \epsilon.$$

We would like to determine if some subset of $r < p$ predictors contributes significantly to the regression model.

Partition the vector of regression coefficients as

$$\beta = \left[\frac{\beta^1}{\beta^2}\right]$$

where β^1 is $(p+1-r) \times 1$ and β^2 is $r \times 1$. We want to test the hypothesis

$$H_0 : \beta^2 = 0$$
$$H_1 : \beta^2 \neq 0$$

Rewrite the model as

$$y = X\beta + \epsilon = X^1\beta^1 + X^2\beta^2 + \epsilon, \tag{1}$$

where $X = [X^1|X^2]$.

Equation (1) is the **full model** with SSR expressed as

$$SSR(X) = \hat{\beta}' X' y \text{ (p+1 degrees of freedom)}$$

and

$$MSE = \frac{y'y - \hat{\beta}' X' y}{n - p - 1}.$$

To find the contribution of the predictors in X^2, fit the model assuming H_0 is true. This **reduced model** is

$$y = X^1 \beta^1 + \epsilon,$$

where

$$\hat{\beta}^1 = (X^{1'} X^1)^{(-1)} X^{1'} y$$

and

$$SSR(X^1) = \hat{\beta}^1 {X^1}' y \text{ (p+1-r degrees of freedom).}$$

The regression sums of squares due to X^2 when X^1 is already in the model is

$$SSR(X^2|X^1) = SSR(X) - SSR(X^1)$$

with r degrees of freedom. This is also known as the **extra sum of squares due to** X^2.

$SSR(X^2|X^1)$ is independent of MSE. We can test $H_0 : \beta^2 = 0$ with the statistic

$$F_0 = \frac{SSR(X^2|X^1)/r}{MSE} \sim F_{r,n-p-1}.$$

CHS example, cont.

Full model: $y_i = \beta_0 + weight_i\beta_1 + height_i\beta_2$
$H_0 : \beta_2 = 0$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WEIGHT	1	1289.38	1289.38	10.22	0.0015
HEIGHT	1	119.78	119.78	0.95	0.3302
Residuals	495	62425.91	126.11		

$$F_0 = 119.78/126.11 = 0.95 < F_{1,495,0.95} = 3.86$$

This should look very similar to the t-test for H_0.

$$BP_i = \beta_0 + weight_i\beta_1 + height_i\beta_2 + age_i\beta_3 + gender_i\beta_4 + \epsilon$$

```
> summary(lm(DIABP~WEIGHT+HEIGHT+AGE+GENDER,data=chs))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.4481265 15.9317114   5.677 2.34e-08 ***
WEIGHT       0.0326655  0.0172310   1.896 0.058579 .
HEIGHT      -0.0009921  0.0852395  -0.012 0.990718
AGE         -0.3283816  0.0926922  -3.543 0.000434 ***
GENDER       0.8348105  1.5264106   0.547 0.584687
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.11 on 493 degrees of freedom
Multiple R-Squared: 0.04636,    Adjusted R-squared: 0.03862
F-statistic: 5.991 on 4 and 493 DF,  p-value: 0.0001031
```

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ vs $H_1 : \beta_j \neq, \; j = 2, 3, 4$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WEIGHT	1	1289.38	1289.38	10.44	0.0013
HEIGHT	1	119.78	119.78	0.97	0.3252
AGE	1	1513.06	1513.06	12.25	0.0005
GENDER	1	36.93	36.93	0.30	0.5847
Residuals	493	60875.92	123.48		

SSR(intercept, weight, height, age, gender) $=$

$$2571019 + 1289.38 + 119.89 + 1513.06 + 36.93 = 2573978$$

SSR(intercept, weight) $= 257019 + 1289.38 = 2572308$

SSR(height, age, gender| intercept, weight) $= 2573978 - 2572308 = 1670$

Notice we can also get this from the ANOVA table above

SSR(height, age, gender| intercept,weight) $= 119.78{+}1513.06{+}36.93 = 1670$

The observed F statistic is

$$F_0 = 1670/3/123.48 = 13.5 > F_{3,493,.95} = 2.62,$$

and we reject the null hypothesis, concluding that at least one of β_2, β_3 or β_4 is not equal to 0.

This should look very similar to the overall F test if we considered the intercept to be a predictor and all the covariates to be the additional variables under consideration.

What if we had put the predictors in the model in a different order?

$$diabp_i = \beta_0 + height_i\beta_2 + age_i\beta_3 + weight_i\beta_1 + gender_i\beta_4 + \epsilon$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HEIGHT	1	680.76	680.76	5.51	0.0193
AGE	1	1798.91	1798.91	14.57	0.0002
WEIGHT	1	442.55	442.55	3.58	0.0589
GENDER	1	36.93	36.93	0.30	0.5847
Residuals	493	60875.92	123.48		

Could we use this table to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$?

What if we had the ANOVA table for the reduced model?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WEIGHT	1	1289.38	1289.38	10.23	0.0015
Residuals	496	62545.69	126.10		

Given that

$$SSR = SSR(x_2) + SSR(x_3|x_2) + SSR(x_1|x_2, x_3) + SSR(x_4|x_3, x_2, x_1)$$

and

$$SSR(x_2, x_3, x_4|x_1) = SSR - SSR(x_1)$$

then

$$SSR(x_2, x_3, x_4|x_1) = 680.76 + 1798.91 + 442.55 + 36.93 - 1289.38 = 1680.$$

One other question we might be interested in asking is if there are any significant interactions in the model?

```
lm(DIABP~WEIGHT*HEIGHT*AGE*GENDER,data=chs)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−1479.5964	1219.6693	−1.21	0.2257
WEIGHT	12.8828	8.3636	1.54	0.1241
HEIGHT	9.9984	7.7695	1.29	0.1988
AGE	20.7270	16.4946	1.26	0.2095
GENDER	−1429.3377	1638.6646	−0.87	0.3835
WEIGHT:HEIGHT	−0.0816	0.0530	−1.54	0.1244
WEIGHT:AGE	−0.1713	0.1135	−1.51	0.1319
HEIGHT:AGE	−0.1342	0.1052	−1.28	0.2025
WEIGHT:GENDER	8.9610	10.7075	0.84	0.4031
HEIGHT:GENDER	7.2497	10.0955	0.72	0.4730
AGE:GENDER	22.2077	22.8169	0.97	0.3309
WEIGHT:HEIGHT:AGE	0.0011	0.0007	1.51	0.1312
WEIGHT:HEIGHT:GENDER	−0.0436	0.0658	−0.66	0.5084
WEIGHT:AGE:GENDER	−0.1449	0.1498	−0.97	0.3339
HEIGHT:AGE:GENDER	−0.1146	0.1404	−0.82	0.4148
WEIGHT:HEIGHT:AGE:GENDER	0.0007	0.0009	0.79	0.4298

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WEIGHT	1	1289.38	1289.38	10.65	0.0012
HEIGHT	1	119.78	119.78	0.99	0.3204
AGE	1	1513.06	1513.06	12.50	0.0004
GENDER	1	36.93	36.93	0.31	0.5810
WEIGHT:HEIGHT	1	19.88	19.88	0.16	0.6855
WEIGHT:AGE	1	4.44	4.44	0.04	0.8483
HEIGHT:AGE	1	73.22	73.22	0.60	0.4371
WEIGHT:GENDER	1	21.53	21.53	0.18	0.6734
HEIGHT:GENDER	1	597.64	597.64	4.94	0.0268
AGE:GENDER	1	214.78	214.78	1.77	0.1835
WEIGHT:HEIGHT:AGE	1	298.24	298.24	2.46	0.1172
WEIGHT:HEIGHT:GENDER	1	167.07	167.07	1.38	0.2407
WEIGHT:AGE:GENDER	1	1051.41	1051.41	8.69	0.0034
HEIGHT:AGE:GENDER	1	5.07	5.07	0.04	0.8379
WEIGHT:HEIGHT:AGE:GENDER	1	75.58	75.58	0.62	0.4298
Residuals	482	58347.07	121.05		

We can simplify the ANOVA table to

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Main effects	4	2959.15	739.79		
Interactions	11	2528.86	229.8964		
Residuals	482	58347.07	121.05		

How do we fill in the rest of this table?