Lecture 6
Multiple Linear Regression, cont.

BIOST 515

January 22, 2004

Testing general linear hypotheses

Suppose we are interested in testing linear combinations of the regression coefficients. For example, we might be interested in testing whether two regression coefficients are equal

$$H_0 : \beta_i = \beta_j.$$

Equivalently,

$$H_0 : \beta_i - \beta_j = 0.$$

Such hypotheses can be expressed as $H_0 : T\beta = 0$, where T is an $m \times p$ matrix of constants, such that only r of the m equations in $T\beta = 0$ are independent.

For example, consider the model

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \epsilon_i$$

and testing the hypothesis

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

This hypothesis is equivalent to

$$H_0 : \left(\begin{array}{cccc} 0 & 1 & -1 & 0 \end{array} \right) \beta = 0.$$

We may also consider the hypothesis

$$H_0 : \beta_1 - \beta_2 = 0, \beta_3 = 0$$

which is equivalent to

$$H_0 : T\beta = 0$$

where

$$T = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We can use sums of squares to test general linear hypotheses. The **full model** is

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

with residual sum of squares

$$SSE(FM) = y'y - \hat{\beta}'X'y \ (n-p \text{ degrees of freedom}).$$

Obtain the **reduced model** by solving $T\beta = 0$ for r of the regression coefficients in the full model in terms of the remaining $p+1-r$ regression coefficients. Substitutin these values into the full model will yield the reduced model,

$$y = Z\gamma + \epsilon,$$

where Z is an $n \times (p+1-r)$ matrix and γ is a $(p+1-r) \times 1$ vector of unknown regression coefficients. The residual sum of squares for the reduced model is

$$SSE(RM) = y'y - \hat{\gamma}Z'y \text{ } (n - p + r \text{ degrees of freedom})$$

$SSE(RM) - SSE(FM)$ is called the **sum of squares due to the hypothesis** $T\beta = 0$. We can test this hypthesis using

$$F_0 = \frac{(SSE(RM) - SSE(FM))/r}{MSE} \sim F_{r,n-p-1}.$$

CHS smoking example

Recall the example where smoking status was recoded to

$$smoke_{1i} = \begin{cases} 1, & never\ smoked \\ 0, & otherwise \end{cases}$$

and

$$smoke_{2i} = \begin{cases} 1, & former\ smoker \\ 0, & otherwise \end{cases} ,$$

and we fit the model

$$BP_i = \beta_0 + \beta_1 smoke_{1i} + \beta_2 smoke_{2i} + \epsilon_i.$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	69.2963	1.6176	42.84	0.0000
smoke1	2.9860	1.7629	1.69	0.0909
smoke2	2.6239	1.8162	1.44	0.1492

We may be interested in testing $H_0 : \beta_1 = \beta_2$ which is equivalent to testing $H_0 : \begin{pmatrix} 0 & 1 & -1 \end{pmatrix}\beta$ The full model is

$$BP_i = \beta_0 + \beta_1 smoke_{1i} + \beta_2 smoke_{2i} + \epsilon_i,$$

andthe reduced model is

$$\begin{aligned} BP_i &= \beta_0 + \beta_1 smoke_{1i} + \beta_1 smoke_{2i} + \epsilon_i \\ &= \beta_0 + \beta_1(smoke_{1i} + smoke_{2i}) + \epsilon_i \\ &= \gamma_0 + \gamma_1 z_i + \epsilon_i \end{aligned}$$

The reduced model is equivalent to the model we fit with current smokers vs. former and never smokers.

Full model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
smoke1	1	101.65	101.65	0.79	0.3737
smoke2	1	267.61	267.61	2.09	0.1492
Residuals	495	63465.82	128.21		

Reduced model:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
smoker	1	354.93	354.93	2.77	0.0965
Residuals	496	63480.15	127.98		

$$F_0 = \frac{(63480.15 - 63465.82)/1}{128.21} = 0.11 < 3.86.$$

Therefore we fail to reject the null hypothesis.

We could also test this hypothesis using the t statistic

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\hat{se}(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\hat{\sigma^2}(C_{11} + C_{22} - 2C_{12})}}$$

where

$$C = \begin{pmatrix} 0.0204 & -0.0204 & -0.0204 \\ -0.0204 & 0.0242 & 0.0204 \\ -0.0204 & 0.0204 & 0.0257 \end{pmatrix}.$$

Therefore

$$t_0 = \frac{(2.986 - 2.624)}{\sqrt{128.21 \times (.0242 + .0257 - 2 \times .0204)}} = .335 < t_{n-p-1,.975}$$

Consider the model

$$BP_i = \beta_0 + \beta_1 smoke_{1i} + \beta_2 smoke_{2i} + \beta_3 age_i + \epsilon_i.$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
smoke1	1	101.65	101.65	0.83	0.3638
smoke2	1	267.61	267.61	2.18	0.1409
AGE	1	2687.39	2687.39	21.84	0.0000
Residuals	494	60778.42	123.03		

Suppose we want to test

$$H_0 : \beta_1 = \beta_2, \beta_3 = 0$$

which is equivalent to

$$H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \beta = 0.$$

The reduced model is

$$\begin{aligned} BP_i &= \beta_0 + \beta_1(smoke_{1i} + smoke_{2i}) + \epsilon_i \\ &= \gamma_0 + \gamma_1 z_i + \epsilon_i \end{aligned}$$

$$F_0 = \frac{(63480.15 - 60778.42)/2}{123.03} = 10.98 > F_{2,494,.95} = 3.01.$$

We reject the null hypothesis.

Confidence intervals in multiple linear regression

- Confidence interval for a single coefficient
- Confidence interval for a fitted value
- Simultaneous confidence intervals on multiple coefficients

Confidence interval for a single coefficient

We can construct a confidence interval for β_j as follows. Given that

$$\frac{\hat{\beta}_j - \beta_j}{\hat{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma^2} C_{jj}}} \sim t_{n-p-1},$$

we can define a $100(1-\alpha)$ confidence interval for β_j as

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma^2} C_{jj}}.$$

Confidence interval for a fitted value

We can construct a confidence interval for the fitted response for a set of predictor values, $x_{01}, x_{02}, ..., x_{0p}$. Define the vector x_0 as

$$x_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0p} \end{pmatrix}.$$

The fitted value at this point is

$$\hat{y_0} = x_0' \hat{\beta}.$$

$\hat{y_0}$ is an unbiased estimator of $E(y|x_0)$, and the variance of $\hat{y_0}$ is

$$\text{var}(\hat{y_0}) = \sigma^2 x_0'(X'X)^{-1}x_0.$$

Therefore, the $100(1-\alpha)\%$ confidence interval for the fitted response at $x_{01}, x_{02}, ..., x_{0p}$ is

$$\hat{y_0} \pm t_{n-p-1,\alpha/2}\sqrt{\hat{\sigma^2} x_0'(X'X)^{-1}x_0}.$$

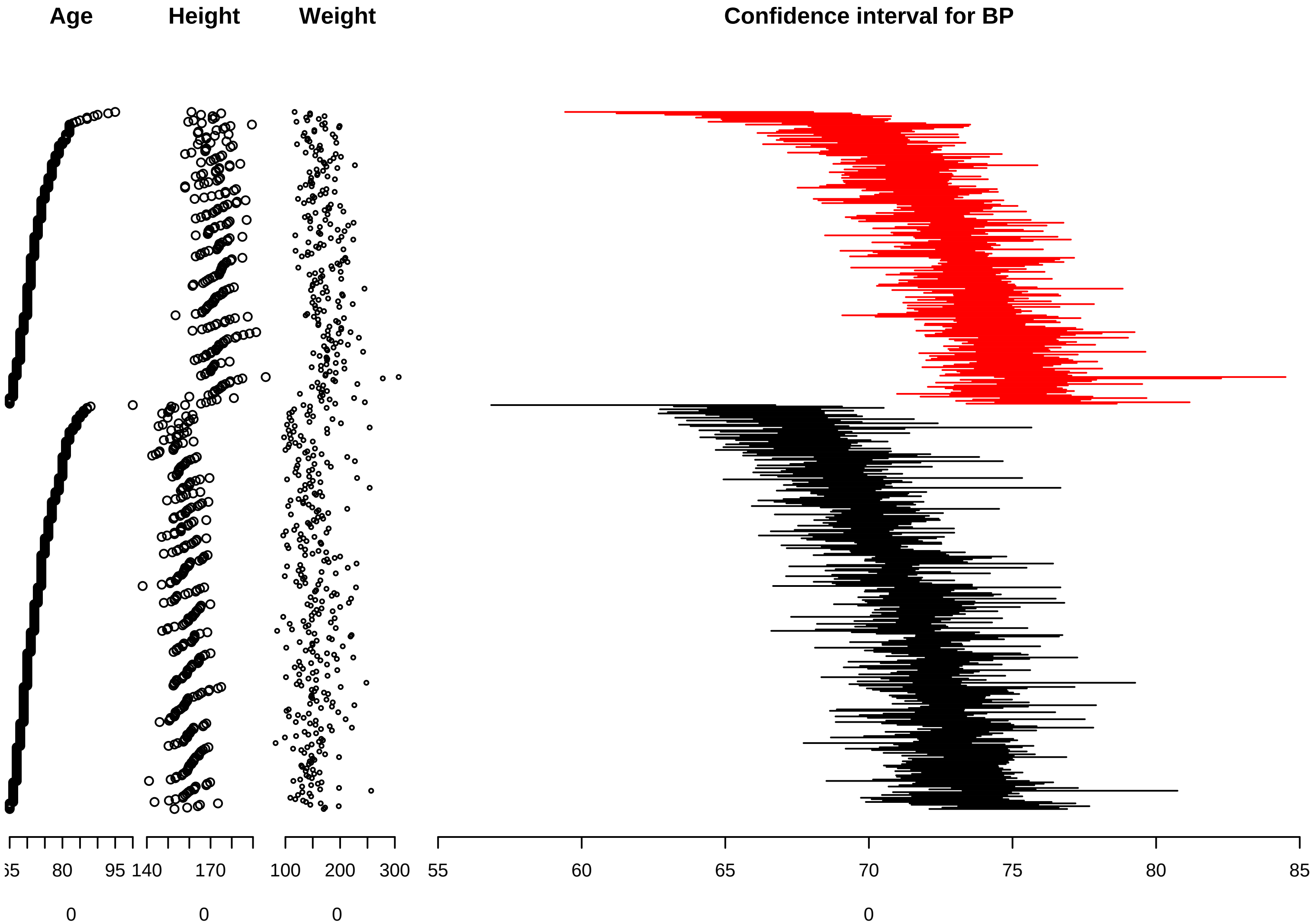
Example from CHS

In the last lecture, we fit the model

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

Let's calculate the confidence interval for the fitted value for the 100th subject who has the covariate vector $(\ 194.8 \quad 159.2 \quad 70.0 \quad 0.0\ )$. The fitted value for BP is 73.67 and $x_0'(X'X)^{-1}x_0 = 0.007972541$ and the 95% confidence interval is

$$73.67 \pm 1.96 \times 11.11 \times \sqrt{0.007972541} = (71.72, 75.62).$$


Age
Height
Weight
Confidence interval for BP
65 80 95
140 170
100 200 300
55 60 65 70 75 80 85
0
0
0
0

Simultaneous confidence intervals

Sometimes we may be interested in specifying a $(1-\alpha)100\%$ confidence interval (or region) for the entire set or a subset of the coefficients.

$$\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{(p+1)MSE} \sim F_{p+1,n-p-1}$$

Therefore, we can define a $(1-\alpha)100\%$ **joint confidence region** for all the parameters in β as

$$\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{(p+1)MSE} \leq F_{p+1,n-p-1}$$

Bonferroni intervals

Another general pproach for obtaining simultaneous confidence intervals is

$$\hat{\beta}_j \pm \Delta \hat{se}(\hat{\beta}_j), \; j = 0, 1, \ldots, p. \tag{1}$$

Using the **Bonferroni method**, we set $\Delta = t_{n-p-1,\alpha/(2(p+1))}$ leading to a **Bonferroni confidence interval** of

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/(2(p+1))} \hat{se}(\hat{\beta}_j).$$

Bonferroni intervals CHS example

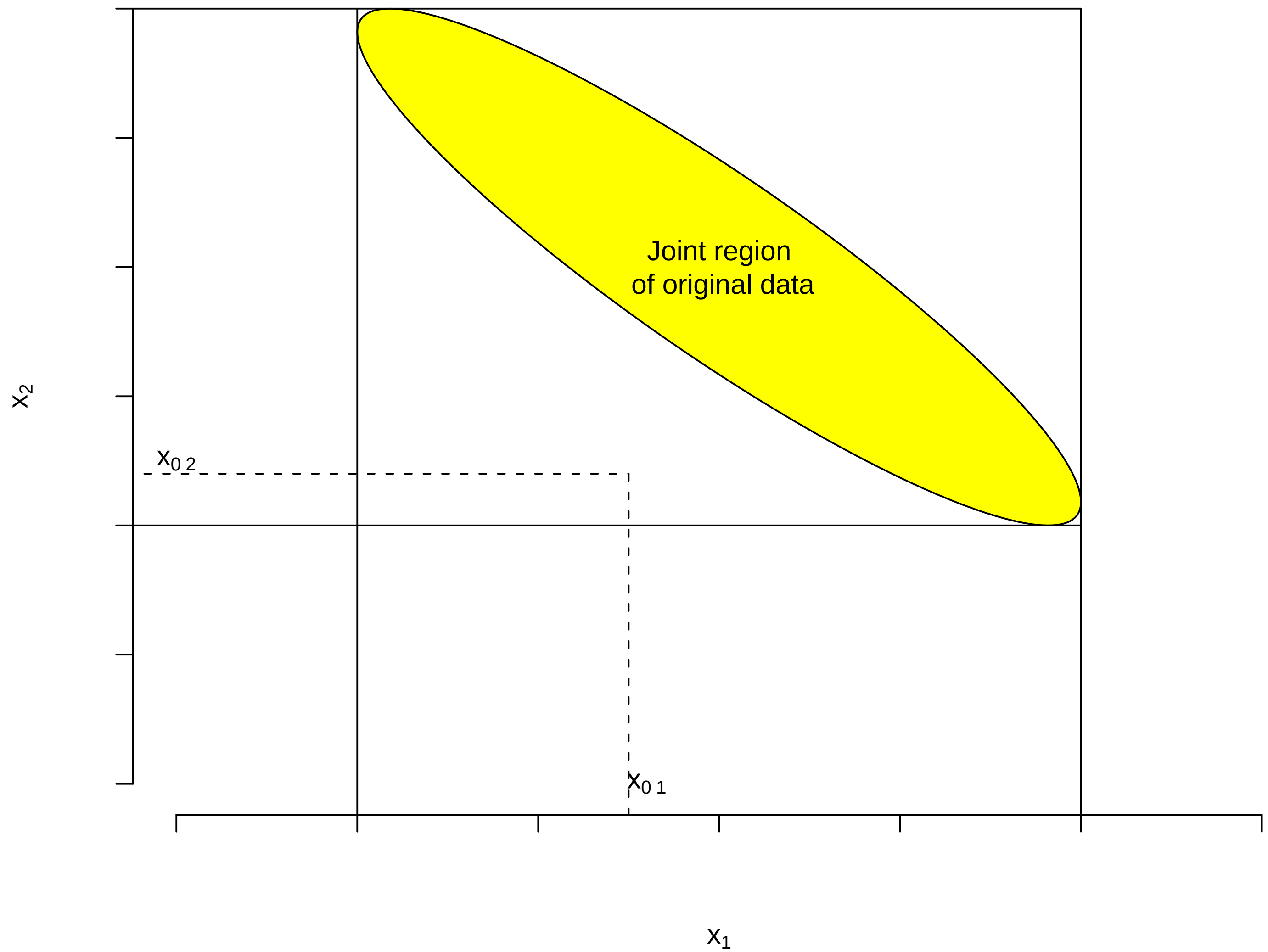
$$BP_i = \beta_0 + weight_i\beta_1 + height_i\beta_2 + age_i\beta_3 + gender_i\beta_4 + \epsilon.$$

The Bonferroni intervals are

$$\hat{\beta}_j \pm t_{493,.005}\hat{se}(\hat{\beta}_j)$$

	Lower	Upper
(Intercept)	49.25	131.64
WEIGHT	−0.01	0.08
HEIGHT	−0.22	0.22
AGE	−0.57	−0.09
GENDER	−3.11	4.78

Hidden extrapolation in multiple regression



R^2 and adjusted R^2

As in simple linear regression

$$R^2 = 1 - \frac{SSE}{SSTO}.$$

In general, R^2 increases whenever new terms are added to the model.

Therefore, for model comparison, we may prefer to use an R^2 that is adjusted for the number of predictors in the model. This is the adjusted R^2 and is equivalent to

$$R^2_{adj} = 1 - \frac{MSE}{SSTO/(n-1)}$$

Predictors	R^2	R^2_{adj}
weight, height, age, gender	0.0464	0.0386
smoke1, smoke2	0.0058	0.0018
weight, height	0.0221	0.0181
smoke1, smoke2, age	0.048	0.042