Testing a subset of regression coefficients

Situation: two models, one contains a subset of the regression predictor variables

"Complete" model contains β_0, β_1, β_2, ..., β_k

"Reduced" model contains β_0, β_1, ..., β_g $(g < k)$

H_0: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$ (reduced model)

H_a: $\neq$ (complete model)

ex. $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

vs

$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

vs

$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$

Test statistic (based on LR):

- fit H_0, get SS(regression, reduced)
- fit H_a, get SS(regression, complete) and SS(residual, complete)

$$f = \frac{[\text{SS(regression, complete)} - \text{SS(regression, reduced)}]/(k-g)}{\text{SS(residual, complete)}/(n-k-1)}$$

Rejection region: reject H_0 if $f \geq f_\alpha$ where the F distribution has $k - g$ and $n - k - 1$ df.

Overall F test: special case of this test is the "overall F test" that a given multiple regression model offers any added predictive value over just using the grand mean of the y_i's.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (model with just β_0; $g = 0$)

H_a: $\neq$ (complete model)

$$f = \frac{[\text{SS(regression, complete)} - \text{SS(regression, reduced)}]/(k-0)}{\text{SS(residual, complete)}/(n-k-1)}$$

$$= \frac{\text{SS(regression, complete)}/k}{\text{SS(residual, complete)}/(n-k-1)}$$

Overall test uses $\text{F}(k, n - k - 1)$ distribution.

Inferences for the individual $\boldsymbol{\beta_j}$'s

The estimated matrix of variances and covariances for the parameter estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, ..., $\widehat{\beta}_k$ is given by

$$\mathbf{W} = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

The estimated variances of $\widehat{\beta}_0$, $\widehat{\beta}_1$, ..., $\widehat{\beta}_k$ are the corresponding elements w_{00}, w_{11}, ..., w_{kk} on the diagonal of **W** ($(j+1)$th row, $(j+1)$th column)

$100(1-\alpha)$% CI for β_j:

$$\widehat{\beta}_j \pm t_{\alpha/2}\sqrt{w_{jj}}$$

where the Student's t distribution has $n-k-1$ df.

Hypothesis test:

$$\mathrm{H_0}:\ \beta_j = 0$$

$$\mathrm{H_a}:\ \beta_j \left\{ \begin{array}{c} > \\ < \\ \neq \end{array} \right\} 0$$

Test statistic:

$$t = \frac{\widehat{\beta}_j - 0}{\sqrt{w_{jj}}}$$

Rejection region: reject H_0 if

$$\left\{ \begin{array}{c} t \geq t_\alpha \\ t \leq -t_\alpha \\ |t| \geq t_{\alpha/2} \end{array} \right\}$$

where the Student's t distribution has $n-k-1$ df.

Confidence interval for $\mathbf{E}(Y)$ at $x_{h1}, x_{h2}, ..., x_{hk}$

$$\mathrm{E}(Y_h) = \beta_0 + \beta_1 x_{h1} + \cdots + \beta_k x_{hk}$$

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 x_{h1} + \cdots + \widehat{\beta}_k x_{hk}$$

$$\text{Let } \mathbf{X}_h' = [x_{h1}, x_{h2}, ..., x_{hk}]$$

$100(1-\alpha)\%$ CI:

$$\widehat{Y}_h \pm t_{\alpha/2} \sqrt{\widehat{\sigma}^2 \left[\mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h\right]}$$

Student's t distribution has $n-k-1$ df.

Prediction interval for a new observation Y_h at $x_{h1}, x_{h2}, ..., x_{hk}$

$100(1-\alpha)\%$ PI:

$$\widehat{Y}_h \pm t_{\alpha/2} \sqrt{\widehat{\sigma}^2 \left[1 + \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h\right]}$$

SAS program for multiple regression

ex. $$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 {x_1}^2 + \beta_5 {x_2}^2$$

```
DATA;
  INPUT Y X1 X2;
  X1SQ=X1*X1;
  X2SQ=X2*X2;
  CARDS;
      ⋮
  ;
PROC REG;
  MODEL Y=X1 X2 X1*X2 X1SQ X2SQ;
  QUAD:  TEST X1SQ, X2SQ;
  QUAD&INTER:  TEST X1SQ, X2SQ, X1*X2;
  MODEL Y=X1 X2 X1*X2;
  INTER:  TEST X1*X2;
RUN;
```

Collinearity (multicollinearity)

The predictor variables are frequently correlated, especially in observational studies. Sometimes one variable can act essentially as a surrogate for another.

ex. predictors x_1, x_2, x_3, with x_2 & x_3 strongly correlated; adding x_3 into the model *after* x_2 might not contribute much to predicting y. Likewise, adding x_2 *after* x_3 might not contribute to prediction. Typical outcomes:

test (x_1) vs (x_1&x_2): reject H_0
test (x_1) vs (x_1&x_3): reject H_0
test (x_1&x_2) vs (x_1&x_2&x_3): do not reject H_0
test (x_1&x_3) vs (x_1&x_2&x_3): do not reject H_0

Interestingly, adding any predictor variables, even ones totally unrelated to Y, make a model fit better!

The problem with: (a) adding predictors strongly correlated with other predictors, or (b) adding predictors that contribute little or nothing to prediction, is that confidence intervals for parameters (& functions of parameters) and prediction intervals become *wider* when such useless variables are included in the model.

Model *goodness-of-fit* is not necessarily model *quality*. Model *goodness-of-fit* can be made arbitrarily perfect by simply adding parameters (i.e. predictor terms) to the model. But, estimating parameters is "expensive"! The price one pays for unnecessary parameters is the reduced usefulness of the model as a prediction tool.