Model Selection

In a multiple regression, which of the predictor variables are important?

What is the best model (choice of predictor variables) to use?

Looking at *t*-statistics for various choices can lead to contradictory results.
(Eg: In housing price data set, age was significant when it was the only explanatory variable, but became insignificant when the other two variables were also included.)

Trying to maximize the R^2 is not helpful, since R^2 always goes up when a new variable is included.

The *F*-statistic (next handout) cannot be used to select a model since it will tend to be significant even if *just one* of the variables has a nonzero true coefficient.

A better way: use *information theory*.

Evaluate the Corrected Akaike Information Criterion, AIC_C for each model.

Choose the model that gives the smallest value (or the most negative value).

For a model with *k* predictors, define

$$AIC_C = \log(SSE) + \frac{2(k+2)}{n-k-3}$$ ("log" = Natural Log).

Zagat 2006 data: y = Cost, n = 299.

Model	*k*	SSE	AIC_C
Food	1	49781	10.836
Décor	1	31268	10.371
Service	1	26159	10.192
Food, Décor	2	27259	10.240
Food, Service	2	26110	10.197
Service, Décor	2	21189	9.988
Food, Décor, Service	3	21189	9.995

Selected Model: Service, Décor. Minimum AIC_C=9.988.