STA 674

Regression Analysis And Design Of Experiments

Comparing and Selecting Models – Lecture 1

STA 674, RADOE:

Comparing and Selecting Models

- Where does it fit in?
- What is it?
- Where next?

STA 674, RADOE:

Comparing and Selecting Models

Variable Selection

- Parsimony: extreme unwillingness to spend money or use resources; principle (or law) of parsimony is the scientific principle that things are usually connected or behave in the simplest or most economical way. – Oxford Online Dictionary (2019)
- "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances." – Sir Isaac Newton

Objective

- Identify the model that best describes the data with the least number of predictors.

STA 674, RADOE:

Comparing and Selecting Models

Variable Selection

1. All Possible Regressions

- Fit separate models with all possible combinations of predictors and choose the one that maximizes or minimizes some criteria.
- Variable Selection Criteria

1. Residual error variance, s_e^2 (minimize)
2. Proportion of variance explained, R^2 (maximize)
3. Adjusted R^2 (maximize) NOT proportion of variance explained...use for model selection only
4. Mallow's C_p (small or close to $p = K + 1$) measures total mean square error, including bias
5. PRESS (minimize) prediction error compared to predicted value with rest of data?

… and many more.

STA 674, RADOE:

Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

- Response

  y = log(Full Expiratory Volume)

- Predictor Variables

  x_1 = height

  x_2 =smoking (0=no,1=yes) indicator variable

  x_3 =gender (0=female,1=male) indicator variable

  all pairwise interactions:

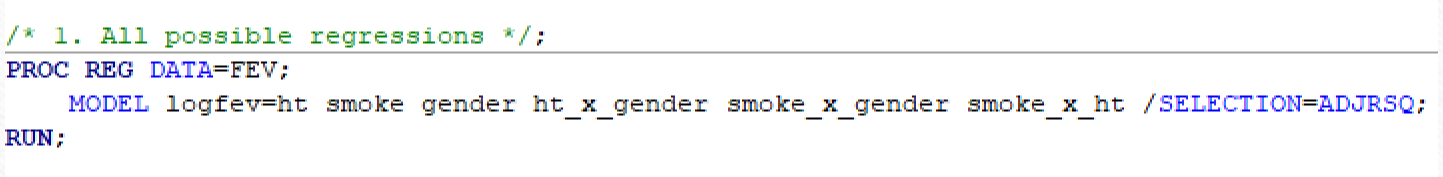
  - height and smoking,
  - height and gender,
  - smoking and gender

STA 674, RADOE:

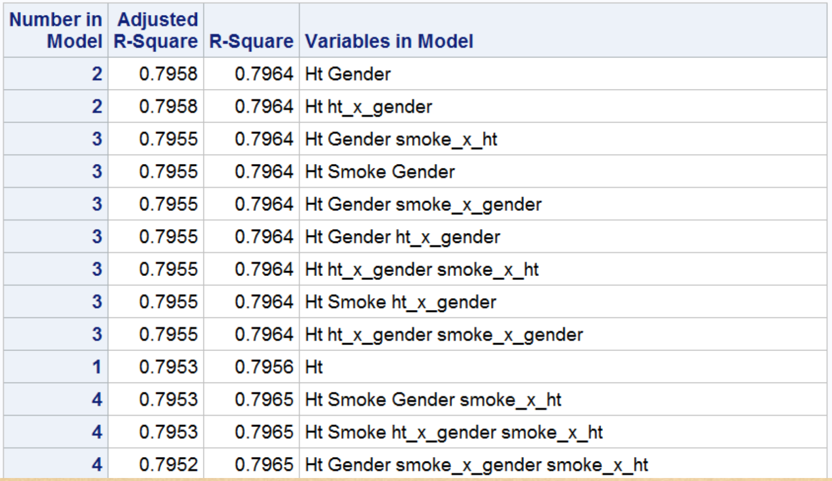
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

- All possible regressions with adjusted R^2 criterion:



```
/* 1. All possible regressions */;
PROC REG DATA=FEV;
    MODEL logfev=ht smoke gender ht_x_gender smoke_x_gender smoke_x_ht /SELECTION=ADJRSQ;
RUN;
```



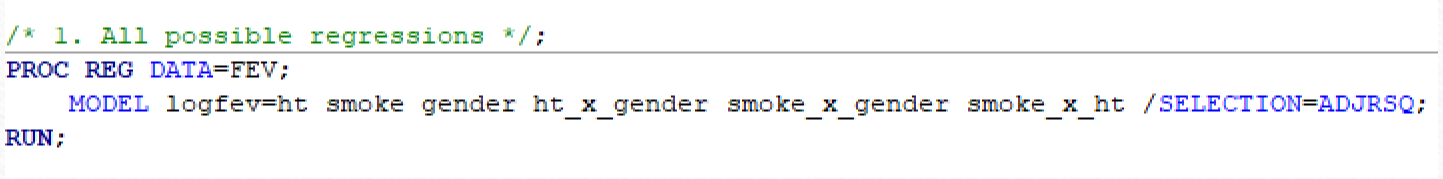
Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.7958	0.7964	Ht Gender
2	0.7958	0.7964	Ht ht_x_gender
3	0.7955	0.7964	Ht Gender smoke_x_ht
3	0.7955	0.7964	Ht Smoke Gender
3	0.7955	0.7964	Ht Gender smoke_x_gender
3	0.7955	0.7964	Ht Gender ht_x_gender
3	0.7955	0.7964	Ht ht_x_gender smoke_x_ht
3	0.7955	0.7964	Ht Smoke ht_x_gender
3	0.7955	0.7964	Ht ht_x_gender smoke_x_gender
1	0.7953	0.7956	Ht
4	0.7953	0.7965	Ht Smoke Gender smoke_x_ht
4	0.7953	0.7965	Ht Smoke ht_x_gender smoke_x_ht
4	0.7952	0.7965	Ht Gender smoke_x_gender smoke_x_ht

STA 674, RADOE:

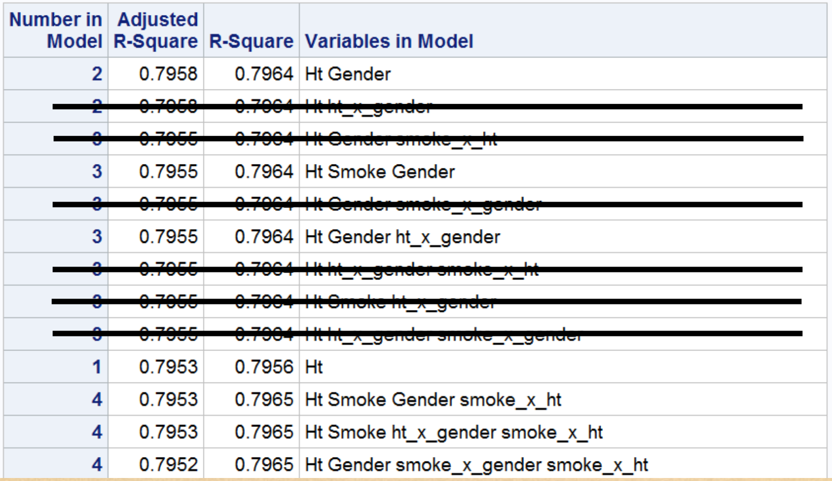
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

- All possible regressions with adjusted R^2 criterion:



```
/* 1. All possible regressions */;
PROC REG DATA=FEV;
    MODEL logfev=ht smoke gender ht_x_gender smoke_x_gender smoke_x_ht /SELECTION=ADJRSQ;
RUN;
```



Number in Model	Adjusted R-Square	R-Square	Variables in Model
2	0.7958	0.7964	Ht Gender
~~2~~	~~0.7958~~	~~0.7964~~	~~Ht ht_x_gender~~
~~3~~	~~0.7955~~	~~0.7964~~	~~Ht Gender smoke_x_ht~~
3	0.7955	0.7964	Ht Smoke Gender
~~3~~	~~0.7955~~	~~0.7964~~	~~Ht Gender smoke_x_gender~~
3	0.7955	0.7964	Ht Gender ht_x_gender
~~3~~	~~0.7955~~	~~0.7964~~	~~Ht ht_x_gender smoke_x_ht~~
~~3~~	~~0.7955~~	~~0.7964~~	~~Ht Smoke ht_x_gender~~
~~3~~	~~0.7955~~	~~0.7964~~	~~Ht ht_x_gender smoke_x_gender~~
1	0.7953	0.7956	Ht
4	0.7953	0.7965	Ht Smoke Gender smoke_x_ht
4	0.7953	0.7965	Ht Smoke ht_x_gender smoke_x_ht
4	0.7952	0.7965	Ht Gender smoke_x_gender smoke_x_ht

STA 674, RADOE:

Comparing and Selecting Models

Variable Selection

1. All Possible Regressions

- Advantages
  - Clear definition of "best" model.
  - Objective (once you have chosen a criterion).
  - All models considered.
- Disadvantages
  - Different criteria select different models.
  - Are small differences in criteria really meaningful?
  - The number of possible models grows exponentially.