LECTURE 10.4

Model evaluation and selection

(MODEL BUILDING)

SHOULD WE INCLUDE VARIABLES AS MANY AS POSSIBLE?

NO!

- First, any correlation among predictors will increase the standard error of the estimated regression coefficients.
- Second, having more slope parameters in our model will reduce interpretability and cause problems with multiple testing.
- Third, the model may suffer from overfitting. As the number of predictors approaches the sample size, we begin fitting the model to the noise.

$$R^2 = \frac{\text{SSReg}}{SS\ total} = 1 - \frac{\text{SSerror}}{SS\ total}$$

LIMITATIONS OF R-SQUARE

In least-squares regression, R^2 is a statistic to reflect the strength of linear relationship between outcome and *a given set* of predictors. However, it does not indicate whether:

- the correct regression was used;
- omitted-variable bias exists;
- the most appropriate set of predictors has been chosen;
- the model might be improved by using transformed predictors.

In particular, R^2 has an *undesired property*: It increases when more variables enter the linear regression as predictors, even they are irrelevant to the outcome (Figures 1 and 2).

Adjusted R-square

To take account of this drawback of R^2, we use an adjusted R^2_{adj}, which is defined by:

$$\boldsymbol{R^2_{adj} = 1 - \left(1 - R_k^2\right)\frac{n-1}{n-k-1}}$$

$$\boldsymbol{= R_k^2 - \left(1 - R_k^2\right)\frac{k}{n-k-1}}$$

Note : R^2_{adj} is always be less than or equal to that of R^2 and could be could be negative. The R^2_{adj} measure *penalizes* the inclusion of a new predictor and thus it increases only if the contribution of the k^{th} predictor is large enough.

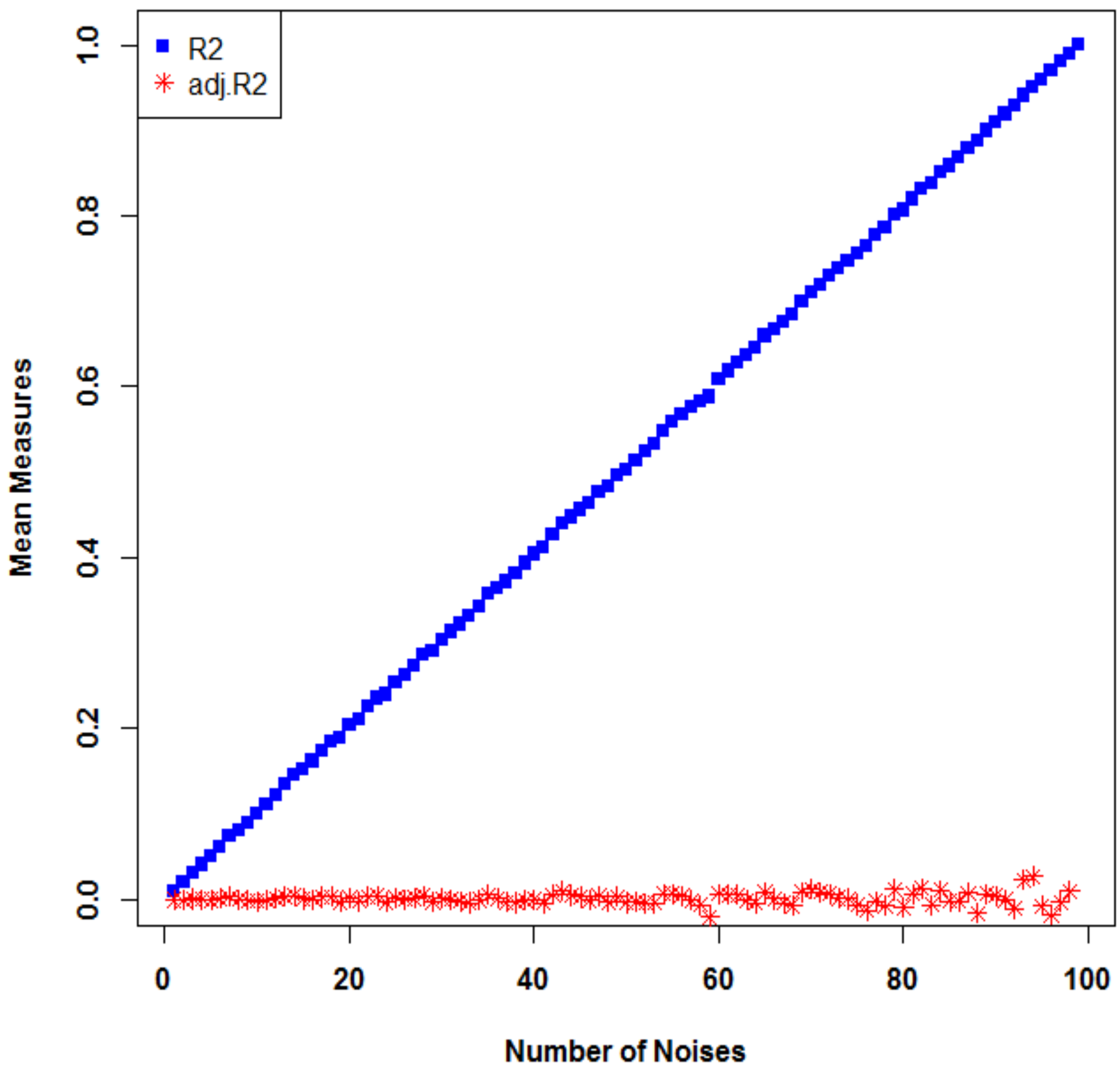


Figure 1:

Mean R^2 and R^2_{adj} over 1000 least-squares of y_i 's on noises (x_{i1}, ..., x_{ik})'s. In each fit, sample size n= 100, and all y_i's and all the x_{ik}'s were iid N(0,1).

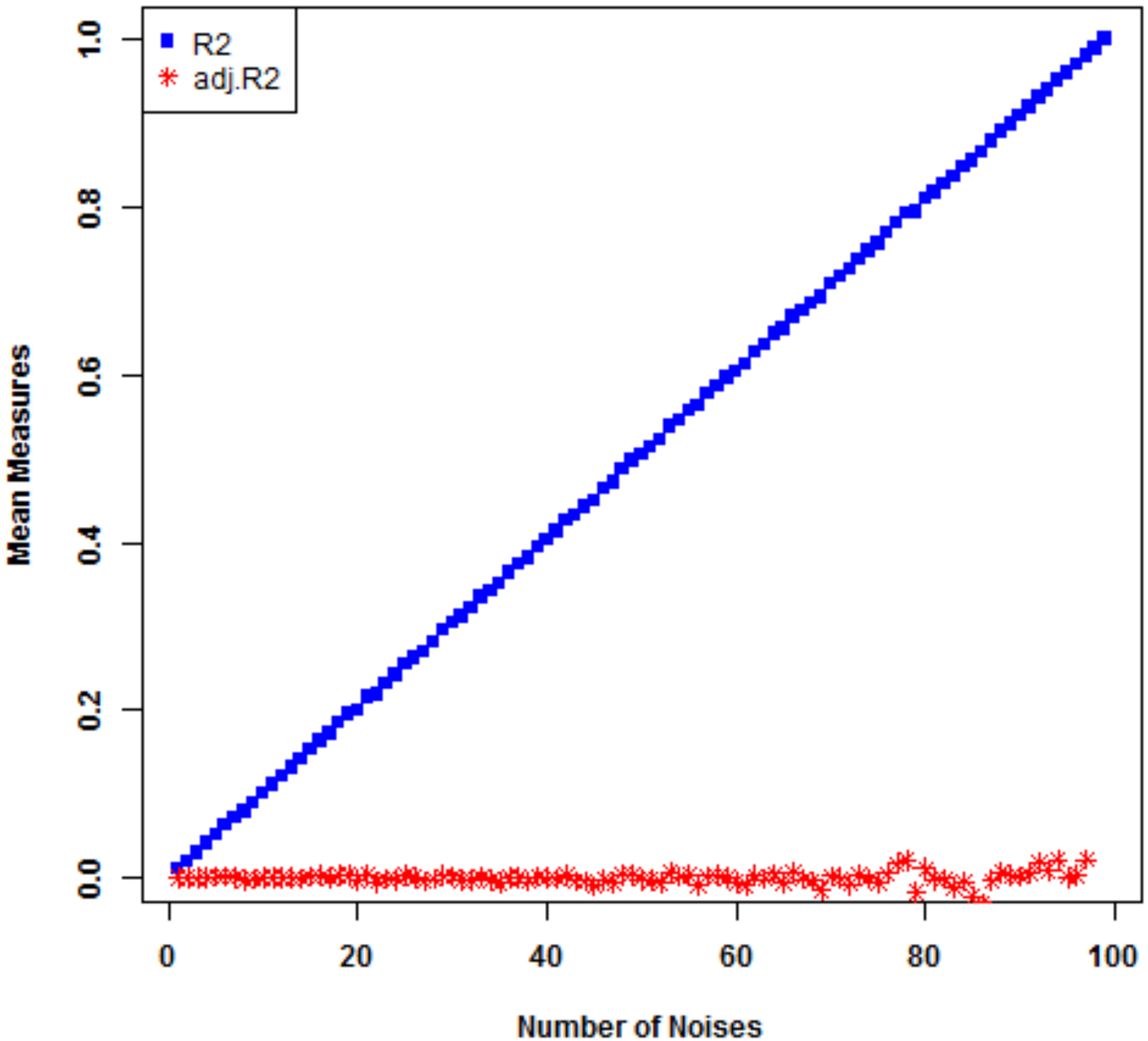


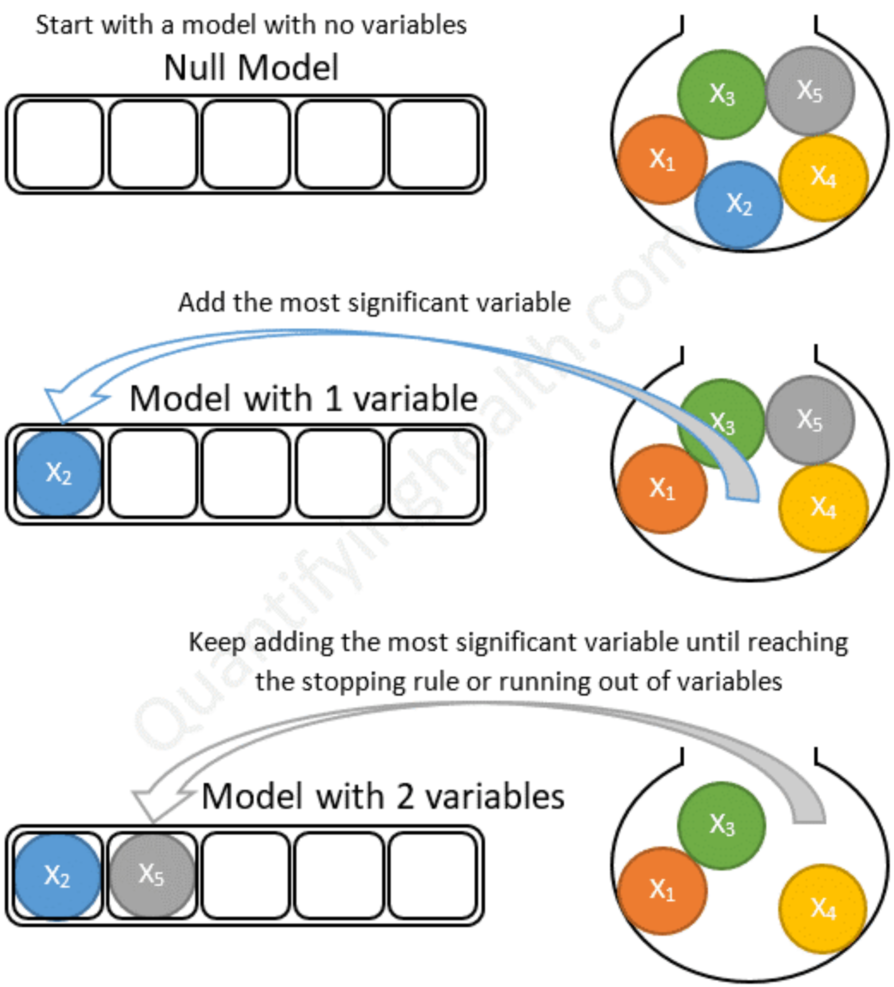
Figure 2:

Mean R^2 and R^2_{adj} over 1000 least-squares of y_i 's on noises $(x_{i1}$, ..., $x_{ik})$'s. In each fit, sample size $n =$ 100, and all y_i's and all the x_{ik}'s were independent, $y_i \sim \mathrm{N}(0,1)$, and $x_{ik} \sim B(10, 0.3)$.

MODEL SELECTION

Forward selection
(forward stepwise selection)

Forward stepwise selection example with 5 variables:



PROCEDURES

- **1. Determine the most significant variable to add at each step**
- The most significant variable can be chosen so that, when added to the model:
  - It has the smallest p-value, or
  - It provides the highest increase in R^2, or
  - It provides the highest drop in model RSS (Residuals Sum of Squares) compared to other predictors under consideration.

PROCEDURES

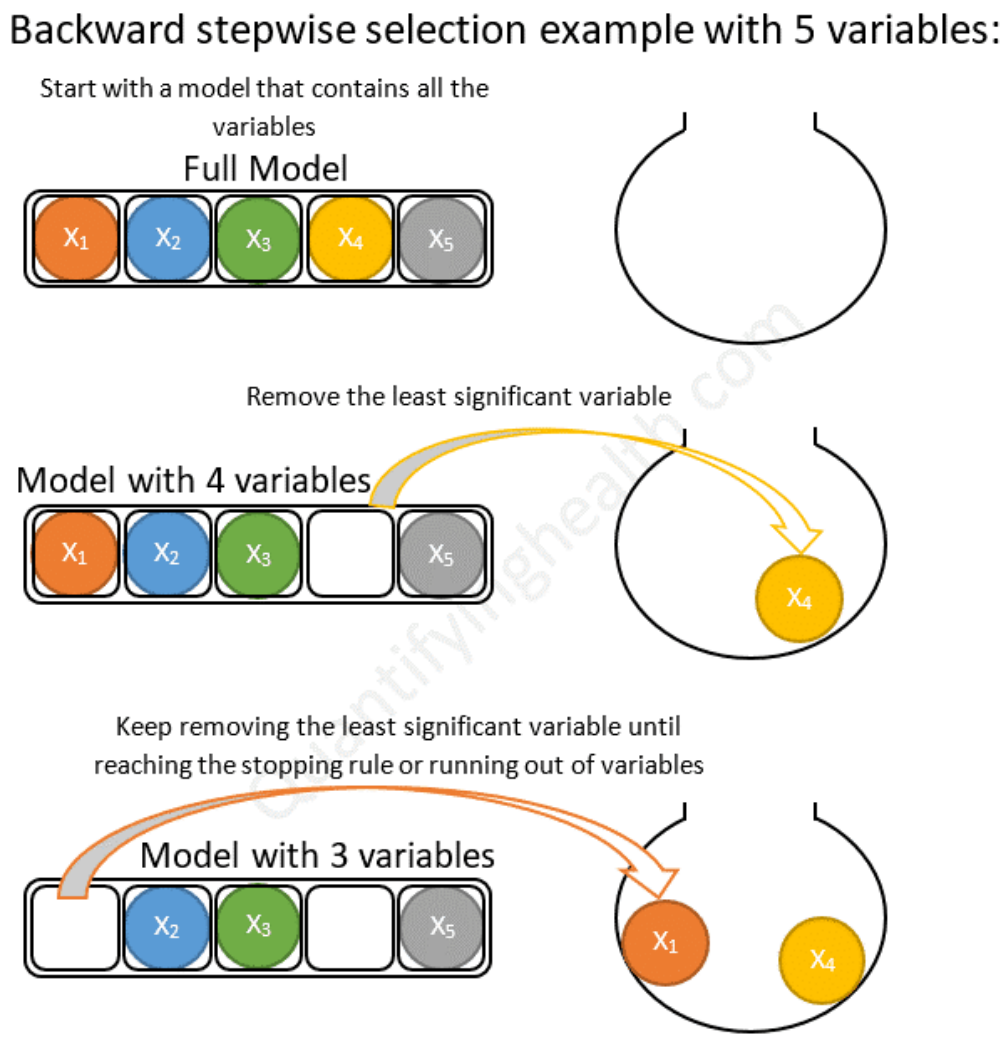
- **2. Choose a stopping rule**
- The stopping rule is satisfied when all remaining variables to consider have a p-value larger than some specified threshold, if added to the model. When we reach this state, forward selection will terminate and return a model that only contains variables with p-values < threshold.
- The threshold can be:

① A fixed value (for instance: 0.05 or 0.2 or 0.5)

② Determined by AIC (Akaike Information Criterion)

③ Determined by BIC (Bayesian information criterion)

Backward selection
(backward stepwise selection)



PROCEDURES

- **1. Determine the least significant variable to remove at each step**
- The least significant variable is a variable that:

① Has the highest p-value in the model, or

② Its elimination from the model causes the lowest drop in R^2, or

③ Its elimination from the model causes the lowest increase in RSS (Residuals Sum of Squares) compared to other predictors

PROCEDURES

- **1. Choose a stopping rule**
- The stopping rule is satisfied when all remaining variables in the model have a p-value smaller than some pre-specified threshold:

① A fixed value (for instance: 0.05 or 0.2 or 0.5)

② Determined by AIC (Akaike Information Criterion)

③ Determined by BIC (Bayesian information criterion)

Stepwise selection

p value to enter = P_{enter} = 0.15, p value to remove = P_{remove} = 0.15