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.

1 MODEL EVALUATION

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

1. 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.

SSReg(k variables)≤ SSReg(k+1 variables)

Proof: modeling - Is MSE decreasing with increasing number of explanatory variables? - Cross Validated (stackexchange.com)

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).

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

1. LIMITATION OF R-SQUARE

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.

2. Adjusted R-square

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

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

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

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

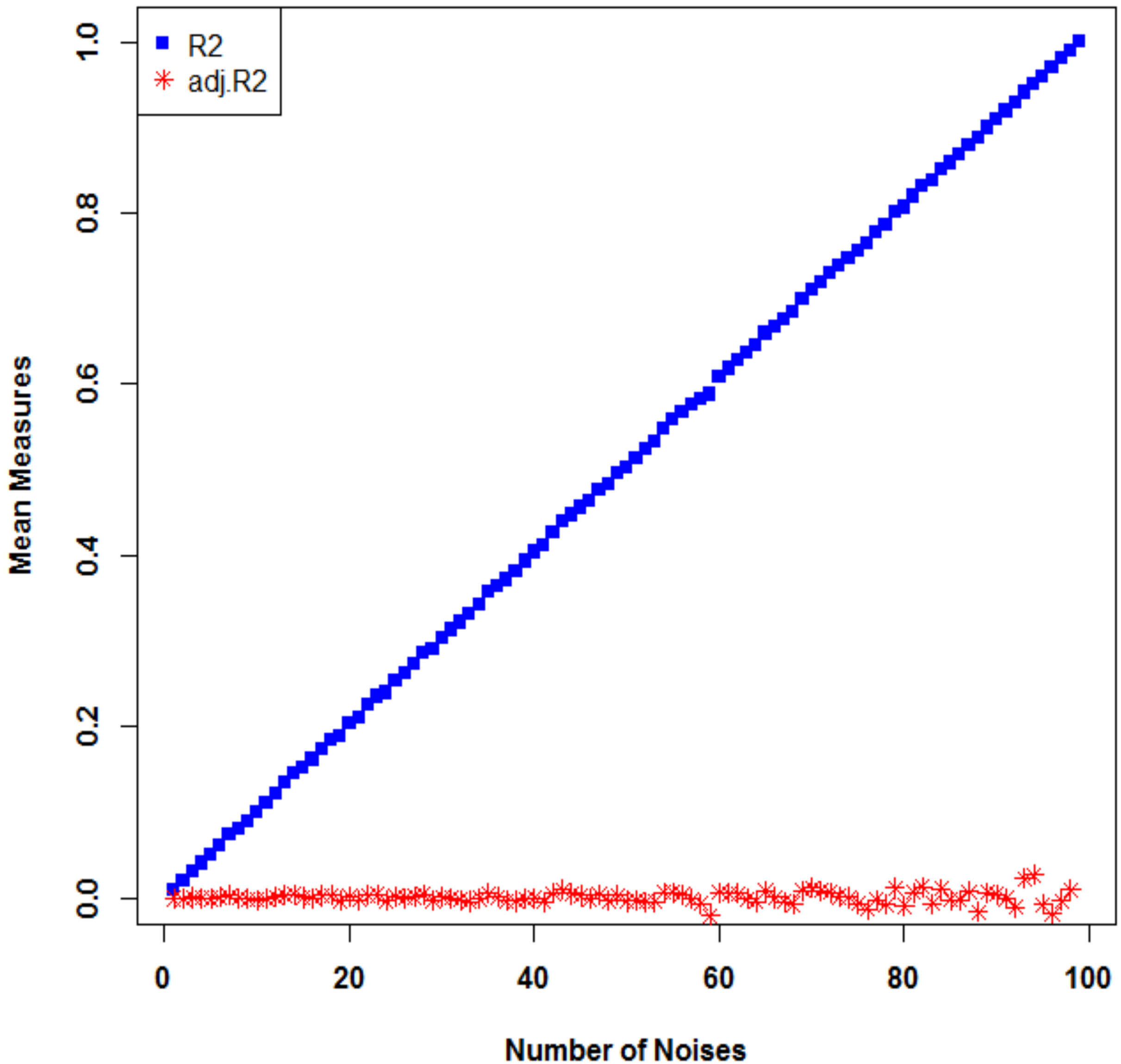


Figure 1:

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).

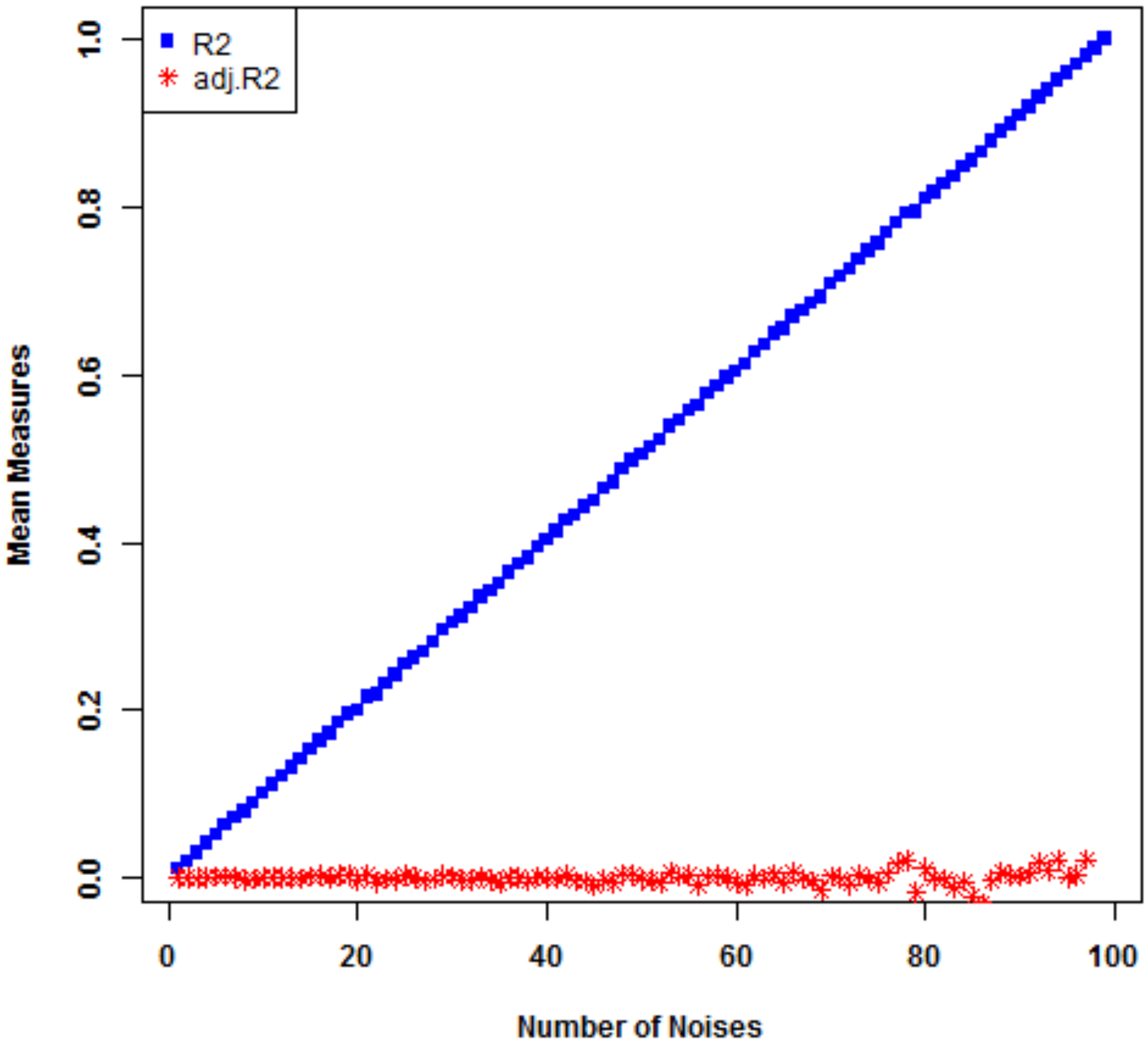


Figure 2:

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

3. Mallows' Cp-statistic

$$C_p = k + 1 + \frac{(MSE_k - MSE_{all})(n - k - 1)}{MSE_{all}}$$

MSE_all, the mean squared error obtained from fitting the model containing *all* of the candidate predictors.
MSE_k, the mean squared error obtained from fitting the model containing *k* candidate predictors.

For the largest model containing all of the candidate predictors, $C_p = k+1$ (always).
When more than one model has a small value of C_p value near k+*1*, in general, choose the simpler model or the model that meets your research needs.

4. Akaike Information Criterion

$$AIC = 2k + n \, Log(ResSS/n),$$

ResSS is the Residual Sum of Squares and K is the number of model parameters

$$AIC = -2\ln(L)+2k$$

with ln(L) the maximum log-likelihood of the model and k the number of free parameters.

Lower AIC values indicate a better-fit model, and a model with a delta-AIC (the difference between the two AIC values being compared) of more than -2 is considered significantly better than the model it is being compared to.

5. Bayesian Information Criterion

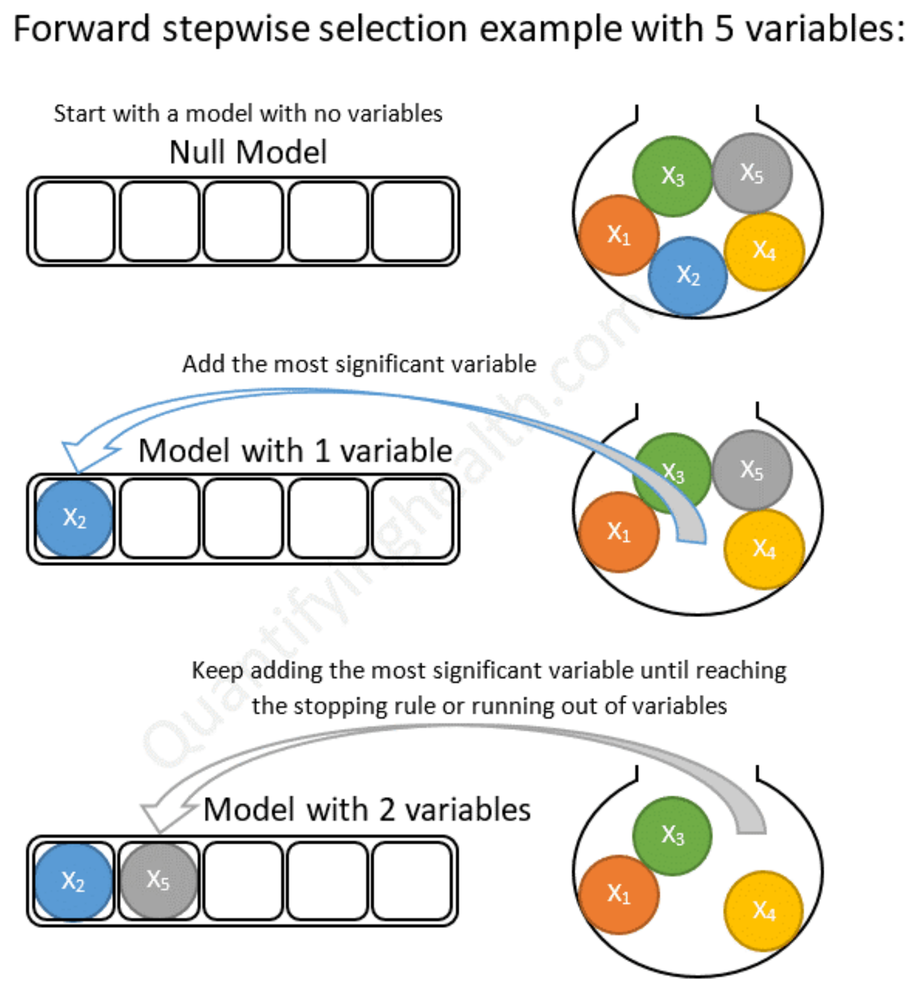
BIC = -2*ln(L)+k*ln(n)

Here n is the number of observations in the model, and k-1 is the number of predictors. That is, k is the number of total parameters (also the total number of coefficients, including the intercept) in the model, ln(L) is the maximum log-likelihood of the model and k the number of free parameters.

Lower BIC values indicate a better-fit model, BIC tends to select models with fewer parameters

MODEL BUILDING

Forward selection
(forward stepwise selection)



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)

```r
##model selection##############################################
#based on fev data, to see the forward selection process
#define intercept-only model
intercept_only <- lm(fev ~ 1, data=x)
#define model with all predictors
all <- lm(fev ~ ., data=x)
#perform forward stepwise regression
forward <- step(intercept_only, direction='forward', scope=formula(all), trace=1)
#view results of forward stepwise regression
forward$anova
```

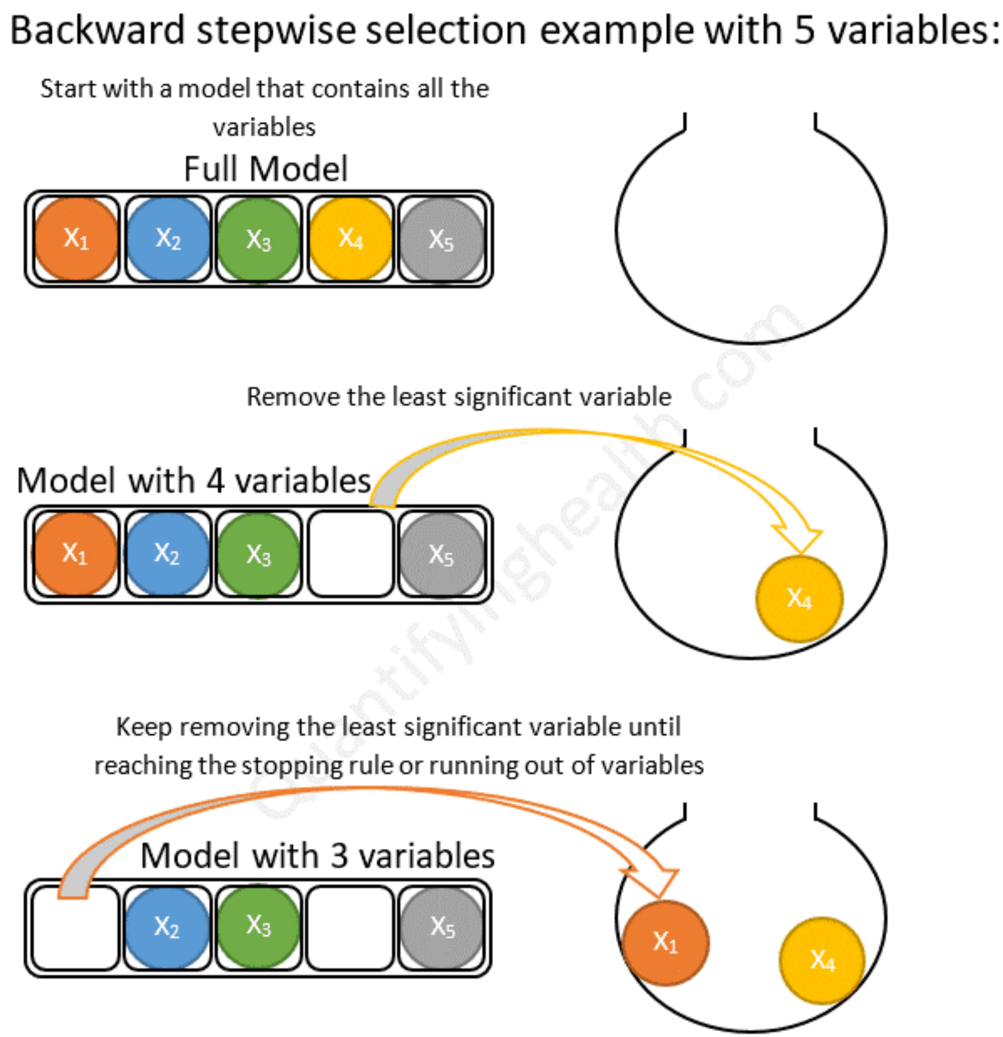
```
Start:  AIC=-185.58
fev ~ 1

         Df Sum of Sq    RSS     AIC
+ hgt     1    369.99 120.93 -1099.86
+ age     1    280.92 210.00  -738.94
+ smoker  1     29.57 461.35  -224.21
+ sex     1     21.32 469.60  -212.63
<none>                490.92  -185.58
```

```
Step:  AIC=-1099.86
fev ~ hgt

         Df Sum of Sq    RSS     AIC
+ age     1    6.2591 114.67 -1132.6
+ sex     1    2.4931 118.44 -1111.5
<none>                120.93 -1099.9
+ smoker  1    0.0022 120.93 -1097.9
```

```
Step:  AIC=-1132.62
fev ~ hgt + age

         Df Sum of Sq    RSS     AIC
+ sex     1    4.0269 110.65 -1154.0
+ smoker  1    0.5921 114.08 -1134.0
<none>                114.67 -1132.6
```

```
Step:  AIC=-1154
fev ~ hgt + age + sex

         Df Sum of Sq    RSS     AIC
+ smoker  1    0.3684 110.28 -1154.2
<none>                110.65 -1154.0

Step:  AIC=-1154.18
fev ~ hgt + age + sex + smoker
```

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)

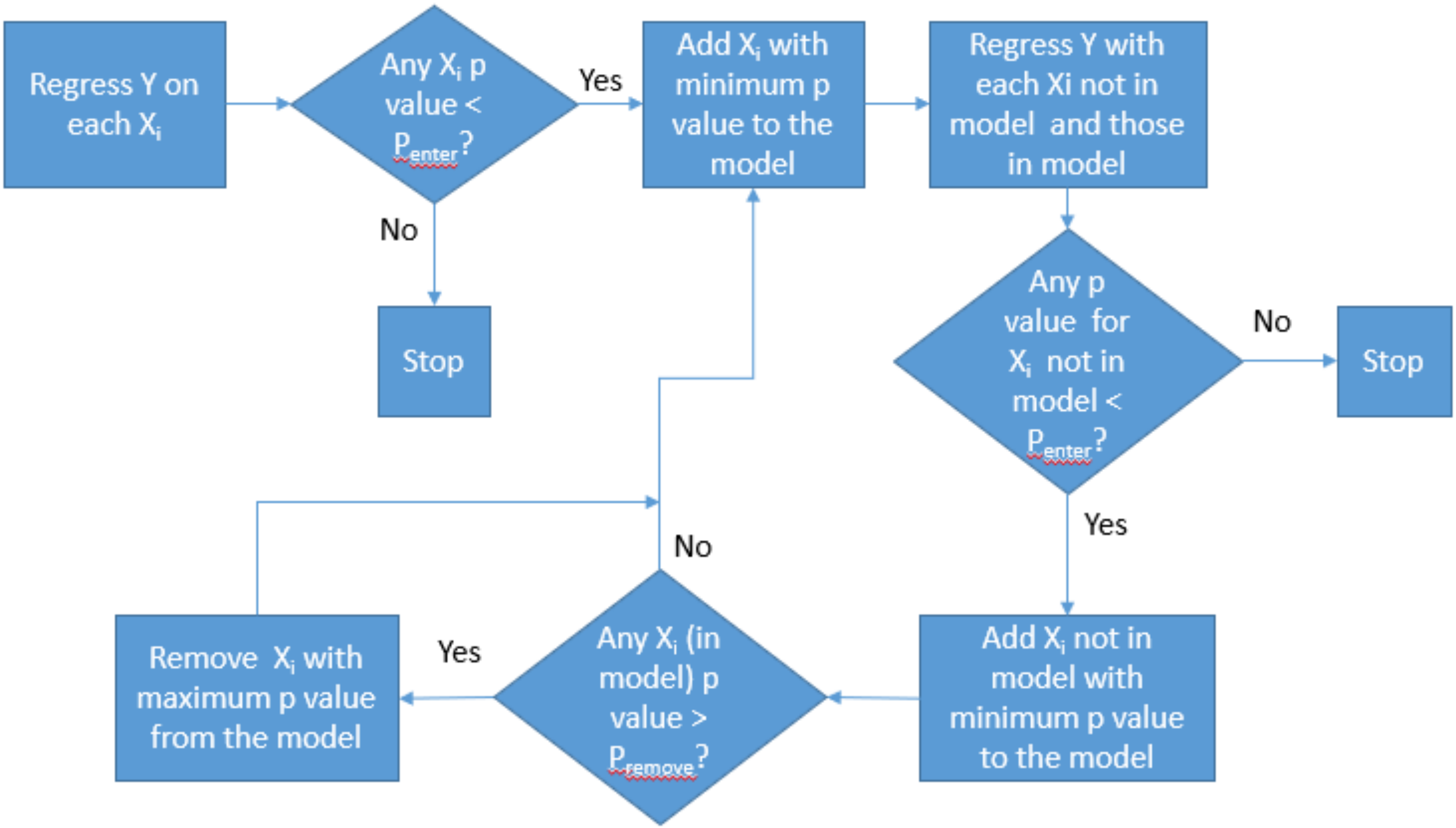
③ Determined by BIC (Bayesian information criterion)

```
###Backward Stepwise Selection
#perform backward stepwise regression
backward <- step(all, direction='backward', scope=formula(all), trace=1)
#view results of backward stepwise regression
backward$anova
```

```
> backward <- step(all, direction='backward', scope=formula(all), trace=1)
Start:  AIC=-1154.18
fev ~ age + hgt + sex + smoker

         Df Sum of Sq    RSS      AIC
<none>                110.28 -1154.18
- smoker  1     0.368 110.65 -1154.00
- sex     1     3.803 114.08 -1134.00
- age     1     8.099 118.38 -1109.83
- hgt     1    81.505 191.78  -794.28
> #view results of backward stepwise regression
> backward$anova
  Step Df Deviance Resid. Df Resid. Dev       AIC
1      NA       NA       649   110.2796 -1154.178
```

Stepwise selection

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



```
> stepwise <- step(intercept_only, direction='both', scope=formula(all), trace=1)
Start:  AIC=-185.58
fev ~ 1

         Df Sum of Sq    RSS      AIC
+ hgt     1    369.99 120.93 -1099.86
+ age     1    280.92 210.00  -738.94
+ smoker  1     29.57 461.35  -224.21
+ sex     1     21.32 469.60  -212.63
<none>                490.92  -185.58

Step:  AIC=-1099.86
fev ~ hgt

         Df Sum of Sq    RSS      AIC
+ age     1      6.26 114.67 -1132.62
+ sex     1      2.49 118.44 -1111.49
<none>                120.93 -1099.86
+ smoker  1      0.00 120.93 -1097.87
- hgt     1    369.99 490.92  -185.58

Step:  AIC=-1132.62
fev ~ hgt + age

         Df Sum of Sq    RSS      AIC
+ sex     1     4.027 110.65 -1154.00
+ smoker  1     0.592 114.08 -1134.00
<none>                114.67 -1132.62
- age     1     6.259 120.93 -1099.86
- hgt     1    95.326 210.00  -738.94

Step:  AIC=-1154
fev ~ hgt + age + sex

         Df Sum of Sq    RSS      AIC
+ smoker  1     0.368 110.28 -1154.18
<none>                110.65 -1154.00
- sex     1     4.027 114.67 -1132.62
- age     1     7.793 118.44 -1111.49
- hgt     1    82.287 192.94  -792.37

Step:  AIC=-1154.18
fev ~ hgt + age + sex + smoker
```