Criteria for model comparison

- In the last video, during each step of the forward, backward, and stepwise selection, we are using an *F*-test to choose between two models
- We can use other criteria to compare between models
- There are many proposed approach in the literature
- They have their own advantages and limitations
- But there is no 'standard' or 'best' criterion to use
- We'll look at some popular model selection criteria

Criteria for comparing models

There are a number of commonly used criteria for comparing models:

- R^2
- Adjusted R^2
- Mallow's C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- PRESS statistics

R^2

The proportion of variation in Y (about $\hat{Y}$) explained by the predictors in the model is known as the coefficient of multiple determination, denoted by R^2,

SSE$_D$ SSE$_F$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

SSE$_R$

- larger R^2 is preferred
- Useful for comparing models with same number of predictors

In this case, we are effectively choosing the model with the smallest SST

Adjusted R^2

The adjusted R^2 is the R^2 'corrected' for degrees of freedom:

$$\tilde{R}^2 = 1 - \frac{\overset{MSE_F}{MSE}}{\underset{MSE_R}{MST}} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2).$$

- larger $\tilde{R}^2$ is preferred
- can compare models of different size

$$\tilde{R}^2 = 1 - \frac{MSE}{MST} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SST}\right) = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Mallow's C_p

The Mallow's C_p statistic compares the predictive ability of subset models to that of the full model:

smaller model

$$C_p = \frac{SSE}{\hat{\sigma}^2} - (n - 2p),$$

MSE of full model $= S_e^2$

where $\hat{\sigma}^2$ is an estimate of σ^2 based on the full model.

- the model with the smallest p for which $C_p \approx p$

$E(SSE) = (n-p)\sigma^2$ if the smaller model is adequate

$E(SSE) > (n-p)\sigma^2$ if the smaller model is not adequate

$$E[C_p] = E\left(\frac{SSE}{\hat{\sigma}^2}\right) - (n-2p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) = (n-p) - (n-2p) = p$$

Akaike information criterion (AIC)

The Akaike information criterion (AIC) is defined as

$$AIC = 2p - 2\ln(\hat{L}),$$

where p is the number of parameter in the model, and $\ln(\hat{L})$ is the log likelihood evaluated at the maximum likelihood estimates.

- the model with the lowest AIC is preferred

For linear regression, $\ln(\hat{L}) = \text{constant} - \frac{n}{2}\ln(SSE)$

$AIC = \text{constant} + 2p + n\ln(SSE)$

Akaike information criterion corrected (AICc)

To adjust for small sample sizes, the AICc is used. It is defined as

penalty term

$$AICc = AIC + \boxed{\frac{2p(p+1)}{n-p-1}},$$

where n is the sample size.

As $n \to \infty$, penalty term $\to 0$, AICc $\to$ AIC

Bayesian information criterion (BIC)

A more stringent criterion with respect to the number of parameters is the Bayesian information criterion (BIC). It is defined as

$$BIC = \boxed{\ln(n)}p - 2\ln(\hat{L})\,.$$

- the model with the lowest BIC is preferred

For linear regression, $BIC = p\ln(n) + n\ln(\hat{L})$

In R, the step() function uses the formula

$$IC = \boxed{k}p - 2\ln(\hat{L})$$

It uses AIC by default.

For AIC, set $k=2$.

For BIC, set $k=\ln(n)$.

If we want to use P-value with an arbitrary significance level α, set $k=\chi^2_{1,\alpha}$.

Example 4.4

Consider again the marks data in Example 4.1.

Fit a multiple linear regression to the data using AIC and

(a) Forward selection

(b) Backward selection

(c) Stepwise selection

Example 4.4 Solutions

If using BIC: , k = log(n)

```
fAIC <- step(null, scope=scope, direction="forward")
```

backward selection: "backward"

stepwise selection: "both"

```
## Start:  AIC=-938.43
## E ~ 1
##
##         Df Sum of Sq    RSS      AIC
## + A6    1    9.3016 11.853 -1132.80
## + A4    1    7.1621 13.993 -1076.55
## + OQ    1    7.1258 14.029 -1075.67
## + A5    1    6.9001 14.255 -1070.26
## + A3    1    4.2472 16.908 -1012.40
## + A2    1    1.7407 19.414  -965.54
## + A1    1    1.0852 20.070  -954.28
## <none>                21.155  -938.43
##
## Step:  AIC=-1132.8
## E ~ A6
##
##         Df Sum of Sq    RSS     AIC
## + OQ    1   1.12811 10.725 -1164.7
## + A3    1   0.56043 11.293 -1147.2
## + A4    1   0.55393 11.299 -1147.0
## + A5    1   0.32088 11.532 -1140.1
## + A2    1   0.09538 11.758 -1133.5
## + A1    1   0.08001 11.773 -1133.1
## <none>               11.853 -1132.8
```

add A6

Find the lowest AIC value.
Add the corresponding variable to our model.

add OQ

Example 4.4 Solutions (cont.)

```
##
## Step:  AIC=-1164.71
## E ~ A6 + OQ
##
##        Df Sum of Sq    RSS    AIC
## + A3    1   0.33372 10.391 -1173.4
## + A4    1   0.18839 10.537 -1168.7
## + A5    1   0.09645 10.629 -1165.8
## <none>              10.725 -1164.7
## + A1    1   0.03104 10.694 -1163.7
## + A2    1   0.02419 10.701 -1163.5
##
## Step:  AIC=-1173.43
## E ~ A6 + OQ + A3
##
##        Df Sum of Sq    RSS    AIC
## + A4    1  0.070882 10.320 -1173.8
## <none>              10.391 -1173.4
## + A5    1  0.039884 10.351 -1172.7
## + A2    1  0.006228 10.385 -1171.6
## + A1    1  0.000108 10.391 -1171.4
```

add A3

add A4

Example 4.4 Solutions (cont.)

```
##
## Step:  AIC=-1173.75
## E ~ A6 + OQ + A3 + A4
##
##         Df Sum of Sq    RSS     AIC
## <none>                10.320 -1173.8
## + A5     1 0.0235297 10.297 -1172.5
## + A2     1 0.0136998 10.307 -1172.2
## + A1     1 0.0003186 10.320 -1171.8
```

current model

The current model <none> has the lowest AIC. Hence we can stop here.

Example 4.4 Solutions (cont.)

```
summary(fAIC)
```

```
##
## Call:
## lm(formula = E ~ A6 + OQ + A3 + A4, data = stats_marks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81971 -0.06460  0.02923  0.09030  0.61607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12306    0.03321   3.705 0.000247 ***
## A6           0.32917    0.04666   7.054 1.00e-11 ***
## OQ           0.18475    0.03868   4.777 2.67e-06 ***
## A3           0.12196    0.04611   2.645 0.008549 **
## A4           0.08014    0.05291   1.515 0.130826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1758 on 334 degrees of freedom
## Multiple R-squared:  0.5121, Adjusted R-squared:  0.5063
## F-statistic: 87.66 on 4 and 334 DF,  p-value: < 2.2e-16
```

PRESS statistic

The prediction sum of squares (PRESS) is used to assess a model's predictive ability, and is given by

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2,$$

where $\hat{y}_{(i)}$ is the predicted value of y_i using the model fitted with ith observation removed.

- smaller PRESS is preferred

Model validation

Q: Is our model generalizable beyond our sample data?

Ideally, we want to get new observations to validate our model:

(1) Refit the model with the new observations included. Check to see if the parameter of the new model is significantly different from our previous model.

(2) Use our original model to predict the response of the new observations. Calculate the prediction errors for the new observations.

In practice, it may be difficult to get new observations. An alternative is to split our data into a training set and a validation/testing set. This is called cross-validation (CV).

(1) Fit our model using the training set

(2) Calculate prediction error using the testing set.

Cross-validation

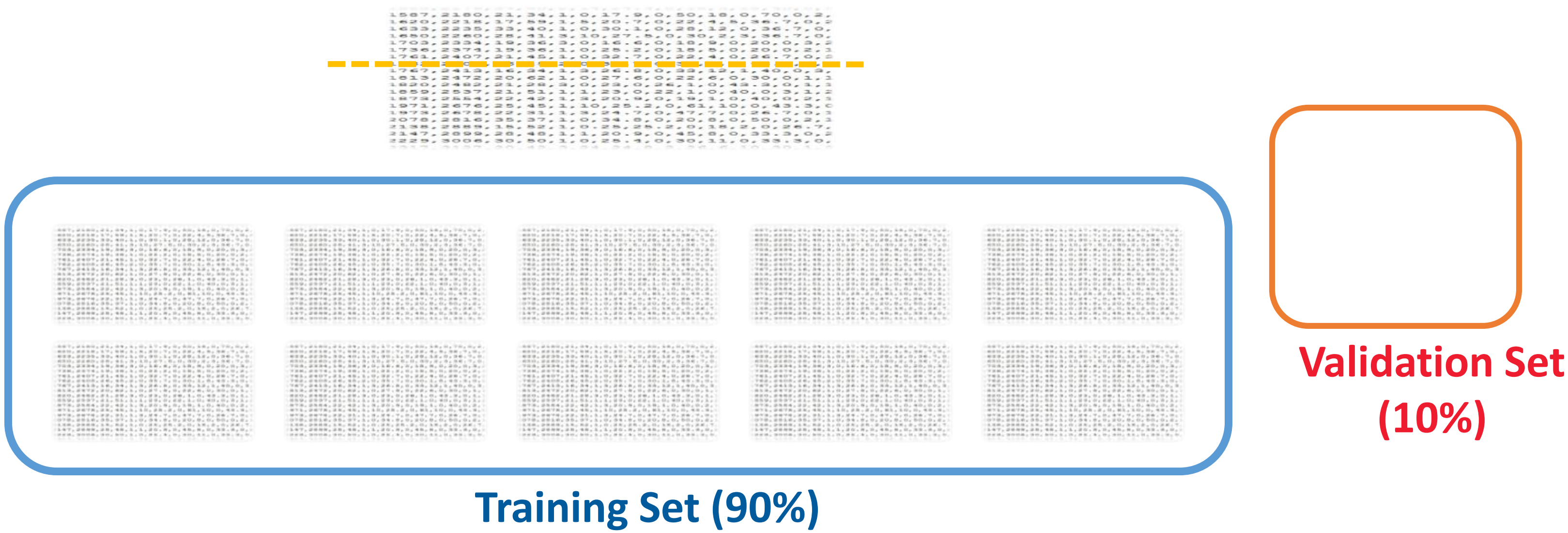
Another useful method to access how good a model is for prediction is cross-validation.

For k-fold cross-validation, you split the data into k parts.

In each step, train the model on $k-1$ parts and test for kth part.

Cross-validation

- 10-fold cross-validation



- Training set: for model development
- Validation set: to assess the accuracy on new data
- Repeat 10 times: estimate prediction error

Notation

Label each part

$$C_1, C_2, \dots, C_K$$

Let the number of observations in C_k be n_k $(k = 1, 2, \dots, K)$ so that

$$n = \sum_{k=1}^{K} n_k \,,$$

i.e. n is the total number of observations.

Prediction error

The cross validation estimate of the prediction error is

$$CV_{(k)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k,$$

weighted sum of the MSE of the k folds

where

$$\boxed{MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k},}$$

where $\hat{y}_i$ is the fitted value for observation i for the model with part k removed.

If $k = n$, we have leave-one-out cross validation (LOOCV). In this case, $CV_{(n)} = PRESS$.