

Linear and Generalized Linear Models (4433LGLM6Y)

Model selection

Meeting 8

Vahe Avagyan

Biometris, Wageningen University and Research



Model selection (Fox: chapter 22, Faraway PRA: chapter 10)

Model selection and criteria

Model validation

Collinearity

Model Selection: Caution!

- Suppose we have many predictor variables, not all necessarily related to the response variable.
- We want to select the "best" subset of predictors.
- The following problems may appear (according to Fox):
 - Problem of simultaneous inference.
 - What does “failing to reject a null hypothesis” mean?
 - Impact of large samples on hypothesis tests: trivially small effects become statistically significant if dataset is large.
 - Exaggerated precision.

General strategies

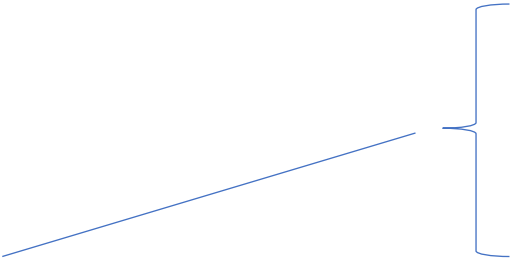
- Addressing these concerns (according to Fox):
 - Use alternative model-selection criteria instead of statistical significance.
 - Compensate for simultaneous inference, e.g., by Bonferroni adjustments,
 - Validate a statistical model selected by another approach.
 - Model averaging
 - Avoid model selection. Specify maximally complex and flexible model without trying to simplify it. Issues here?

Model Selection

- Reasons for selecting "best" subset of regressors:
 - We can explain the data in simplest way, removing redundant predictors.
 - **Principle of Occam's Razor (parsimony)** : among several plausible explanations for phenomenon, simplest is best.
 - Unnecessary predictors will add noise to the estimation of other quantities that are of interest,
 - degrees of freedom are wasted.
 - Collinearity is caused by having too many variables doing same job.
 - Remove excess predictors.

Types of variable selection

- Two main types of variable selection:
 - **Stepwise approach**, comparing successive models
 - Stepwise approach may use hypothesis testing to select the next step, but other criteria may be used too.
 - **Criterion approach**, finding a model that optimizes some measure of goodness of fit.
- **Marginality principle**: keep lower order terms in the model, if higher order term is important.
- Model selection is conceptually simplest if the goal is **prediction**
 - Example: develop regression model that will predict new data as accurately as possible.

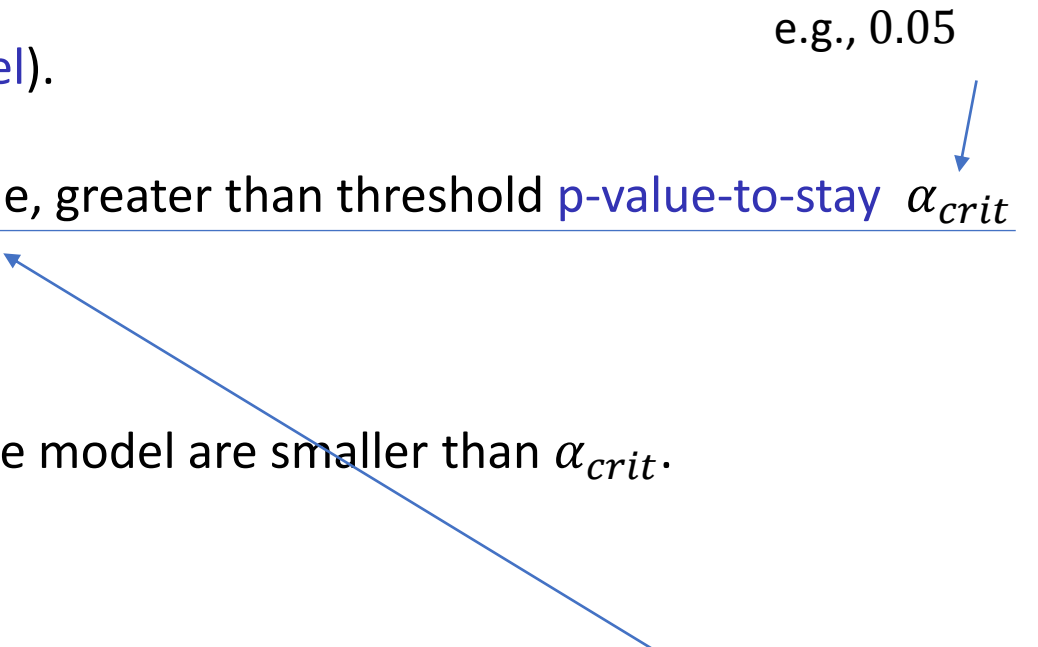
- 
- Backwards
 - Forward
 - Stepwise (mixed)

Stepwise procedures

- Backward elimination vs Forward elimination
- Both procedures can be implemented using `step()` command in R.
- Stepwise selection (mixed)
- Stepwise procedures may also be used in combination with other criteria, e.g., AIC (see `stepAIC()`).

Stepwise procedures: Backward Elimination

- Backward Elimination

1. Start with all predictors in model (**full model**).
 2. Remove a predictor with the highest p-value, greater than threshold **p-value-to-stay** α_{crit}
 3. Refit the model and repeat the step 2.
 4. Stop if all p-values of terms remaining in the model are smaller than α_{crit} .
- 

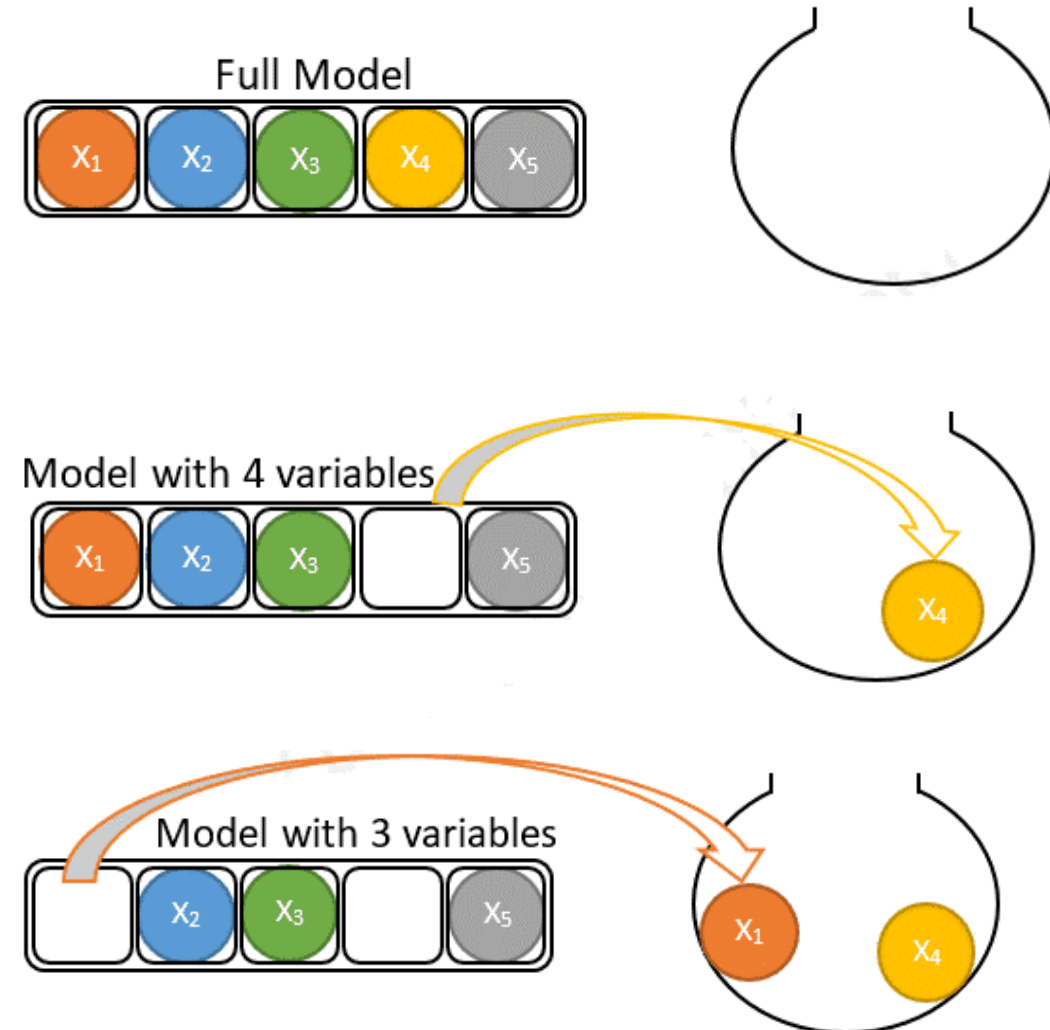
- What is the main drawback of this approach?

Other criteria may be used here as well,
e.g., *AIC*.

Stepwise procedures: Backward Elimination with 5 predictors

- Backward Elimination

1. Start with all 5 predictors in model (full model).
2. Remove a predictor with the least significant predictor (e.g., X_4)
3. Refit the model and repeat the step 2.
4. Keep removing the least significant predictors until all of them are significant (or running out of predictors).



Example: Life expectancy dataset

- Examine the relationship between life expectancy and other socio-economic variables for the U.S. states.

```
> statedata <- read.csv("statedata.csv", header = TRUE)
> head(statedata)
```

	life.exp	population	income	illiteracy	murder	highSchoolGrad	frost	area
1	69.05	3615	3624	2.1	15.1	41.3	20	50708
2	69.31	365	6315	1.5	11.3	66.7	152	566432
3	70.55	2212	4530	1.8	7.8	58.1	15	113417
4	70.66	2110	3378	1.9	10.1	39.9	65	51945
5	71.71	21198	5114	1.1	10.3	62.6	20	156361
6	72.06	2541	4884	0.7	6.8	63.9	166	103766

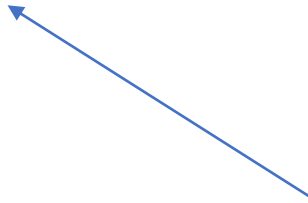
- There is an easier way to do this in **R**, but let's try manually first.
- See using `step()` or `stepAIC()` commands.

Example: Life expectancy dataset

```
> statreg <- lm(life.exp ~ ., data = statedata)
> coef(summary(statreg))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.094322e+01	1.747975e+00	40.58594017	2.510609e-35
population	5.180036e-05	2.918703e-05	1.77477309	8.318351e-02
income	-2.180424e-05	2.444256e-04	-0.08920603	9.293422e-01
illiteracy	3.382032e-02	3.662799e-01	0.09233464	9.268712e-01
murder	-3.011232e-01	4.662073e-02	-6.45899735	8.679582e-08
highSchoolGrad	4.892948e-02	2.332328e-02	2.09788176	4.197175e-02
frost	-5.735001e-03	3.143230e-03	-1.82455682	7.518682e-02
area	-7.383166e-08	1.668163e-06	-0.04425927	9.649075e-01

```
> summary(statreg)$r.squared
[1] 0.7361563
```



“area” shows the highest p-value above the threshold (e.g., 0.05),
i.e., the least significant predictor.

Example: Life expectancy dataset

```
> statreg <- update(statreg, ~ . -area, data = statedata)
> coef(summary(statreg))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	70.9893185176	1.387454e+00	51.16515405	3.694989e-40
population	0.0000518827	2.878768e-05	1.80225346	7.851808e-02
income	-0.0000244403	2.342908e-04	-0.10431609	9.174036e-01
illiteracy	0.0284588124	3.416329e-01	0.08330231	9.339978e-01
murder	-0.3018231392	4.334432e-02	-6.96338357	1.453868e-08
highSchoolGrad	0.0484723220	2.066727e-02	2.34536620	2.369166e-02
frost	-0.0057757582	2.970228e-03	-1.94455035	5.838883e-02

```
> summary(statreg)$r.squared
[1] 0.736144
```

- Next, “illiteracy” shows the highest p-value above the threshold.

```
> statreg <- update(statreg, ~ . -illiteracy, data = statedata)
> coef(summary(statreg))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.106575e+01	1.0289414512	69.0668559	1.659665e-46
population	5.114856e-05	0.0000270945	1.8877836	6.566097e-02
income	-2.477076e-05	0.0002315986	-0.1069555	9.153104e-01
murder	-3.000077e-01	0.0370418231	-8.0991613	2.907482e-10
highSchoolGrad	4.775797e-02	0.0185907897	2.5689048	1.367027e-02
frost	-5.909864e-03	0.0024677801	-2.3948100	2.095338e-02

```
> summary(statreg)$r.squared
[1] 0.7361014
```

- “income” shows the highest p-value above the threshold.

Example: Life expectancy dataset

```
> statreg <- update(statreg, ~ . -income, data = statedata)
```

```
> coef(summary(statreg))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.102713e+01	9.528530e-01	74.541541	8.612596e-49
population	5.013998e-05	2.512002e-05	1.996017	5.200514e-02
murder	-3.001488e-01	3.660946e-02	-8.198669	1.774520e-10
highSchoolGrad	4.658225e-02	1.482706e-02	3.141704	2.968091e-03
frost	-5.943290e-03	2.420875e-03	-2.455017	1.801778e-02

```
> summary(statreg)$r.squared
```

```
[1] 0.7360328
```

- Next, “population” shows the highest p-value above the threshold.

```
> statreg <- update(statreg, ~ . -population, data = statedata)
```

```
> coef(summary(statreg))
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	71.036378813	0.983262169	72.245614	5.253889e-49
murder	-0.283065168	0.036731323	-7.706370	8.039156e-10
highSchoolGrad	0.049948704	0.015201124	3.285856	1.950415e-03
frost	-0.006911735	0.002447477	-2.824025	6.987727e-03

```
> summary(statreg)$r.squared
```

```
[1] 0.7126624
```

- The procedure stops here.

Stepwise procedures: Forward Selection with 5 predictors

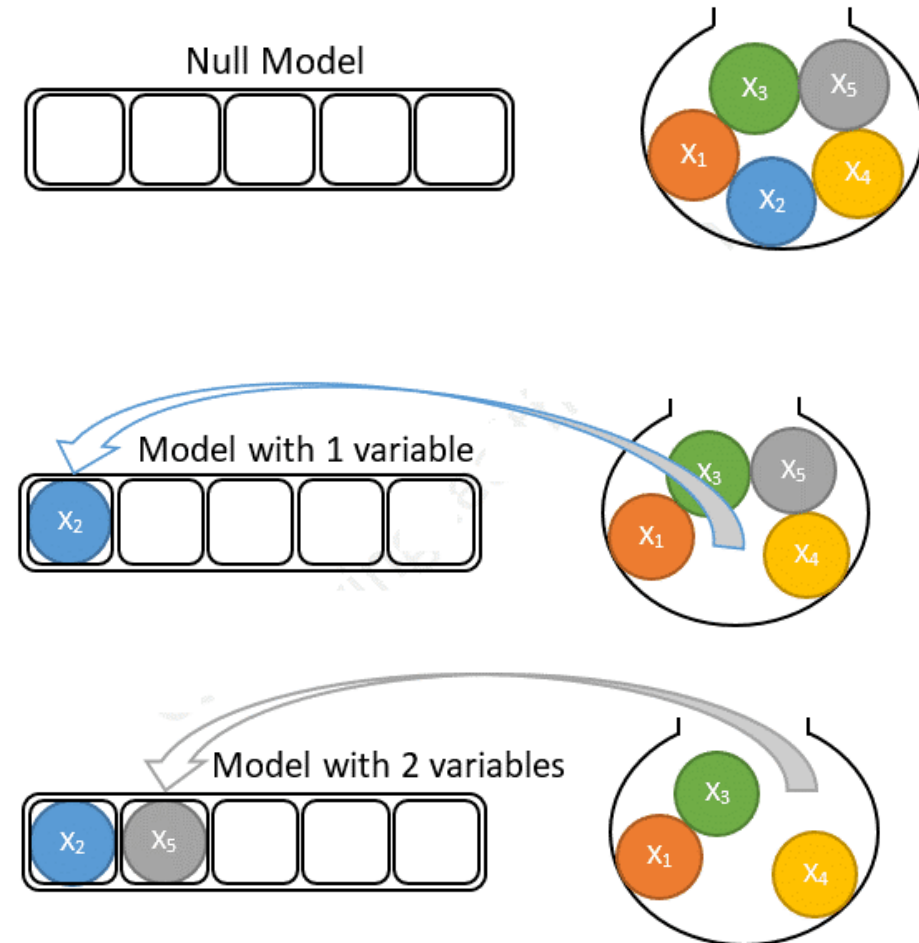
- Forward Selection

1. Start with no predictor in model (**null model**).
2. Enter a predictor with the smallest p-value, smaller than threshold **p-value-to-enter** α_{crit} .
3. Refit the model and repeat the step 2.
4. Stop if all p-values of terms not in the model are higher than α_{crit} .

Stepwise procedures: Forward Selection

- Forward Selection

1. Start with no predictor in model (**null model**).
2. Enter the most significant predictor (e.g., X_2).
3. Refit the model and repeat the step 2.
4. Keep adding the most significant predictors until those not in the model are not significant.



Stepwise procedures: Stepwise (or mixed) selection

- This is a combination of backward elimination and forward selection. After entering variable, all variables in model are candidate for removal.
- Thresholds **p-value-to-enter** and **p-value-to-stay** need to be specified.
- Drawback related to earlier mentioned caveats:
 - “Optimal” model may be missed due to adding / dropping of single variables.
 - Stepwise selection tends to pick models smaller than desirable for prediction purposes.
 - If using p-values: don’t treat p-values literally! (**recall multiple testing problem**)
 - Procedures are not linked to final objective of prediction or explanation.

Criterion-Based Procedures

- Criterion-based procedures typically compare all possible models (i.e., all possible “subsets regression”)
- A model with k regressors has 2^k possible sub-models! (why?)
- Different criteria may be used, e.g.:
 - R_{adj}^2 (R^2 - adjusted)
 - Mallows's C_p
 - Predicted residual error sum of squares (PRESS) and Cross-Validation
 - Akaike information criterion (AIC)
 - Bayesian (sometimes Schwarz's Bayesian) information criterion (BIC)

Criterion-Based Procedures: R^2_{adj}

- Recall

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

- Why can't we use R^2 as a model selection criterion?

- Instead, we use the adjusted R^2 :


$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2_{Model}}{\hat{\sigma}^2_{Null}}$$

$\hat{\sigma}^2_{Null}$ is the estimate of error variance based on the “empty” model (intercept only).

- R^2_{adj} will only increase by changing a model, if the estimate of error variance based on new model $\hat{\sigma}^2_{Model}$ decreases. It will only decrease, if the “change” in RSS is compensated by change in residual df.

Criterion-Based Procedure

- Good model should predict well, so total prediction MSE (on population level) should be small.
- The scaled (normalized) MSE is:

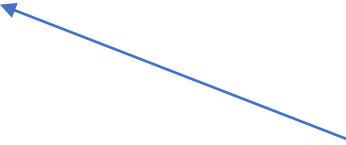
$$\frac{1}{\sigma_{\epsilon}^2} \sum_{i=1}^n E(\hat{Y}_i - E(Y_i))^2 = \frac{1}{\sigma_{\epsilon}^2} \sum_{i=1}^n \left\{ V(\hat{Y}_i) + \left(E(\hat{Y}_i) - E(Y_i) \right)^2 \right\}.$$


- There are two components: variance $V(\hat{Y}_i)$ and squared bias $\left(E(\hat{Y}_i) - E(Y_i) \right)^2$.
- **Bias-variance trade-off:** by removing a variable, the decrease in variance offsets any increase in bias

Criterion-Based Procedure: Mallow's C_p

- Prediction MSE is estimated by Mallow's C_p :

$$C_p = \frac{RSS_p}{\hat{\sigma}_\epsilon^2} + 2p - n$$



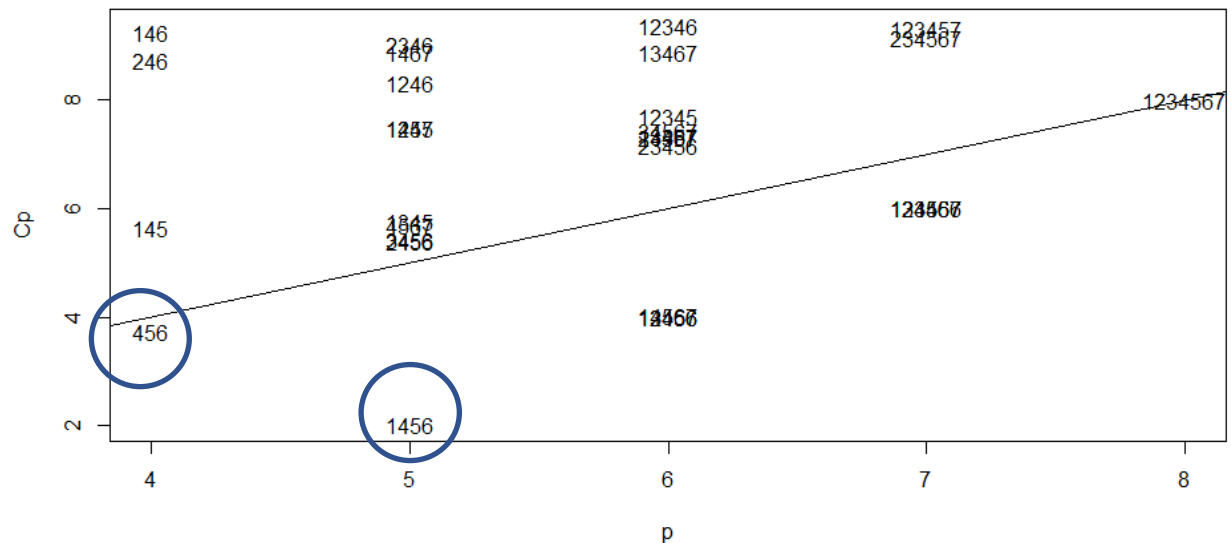
$\hat{\sigma}_\epsilon^2$ is from the full model, and RSS_p from current model (with p parameters).

- Good model should have C_p close to or below p . Model with a bad fit has C_p much bigger than p .
- For full model, we have $C_p = p$ (why?).

Mallow's C_p : Example

- In practice, we plot C_p against p and look for models with small p and with C_p around or less than p .

```
> library(leaps)
> statreg <- lm(life.exp ~ ., data = statedata)
> x <- model.matrix(statreg)[,-1]
> y <- statedata$life.exp
> statregCP <- leaps(x,y)
> Cpplot(statregCP)
```



- We have $k = 7$ predictors
- How many models are there, in total?
- Good options are the models “456” and “1456”. Smaller model is more parsimonious, but larger models fits slightly better.

Criterion-Based Procedure: R_{adj}^2

- Now, let's check the model selection with R_{adj}^2 .

```
> adjr <- leaps(x, y, method = "adjr2")
> maxadjr(adjr, 8)
```

1,4,5,6	1,2,4,5,6	1,3,4,5,6	1,4,5,6,7	1,2,3,4,5,6	1,3,4,5,6,7
0.713	0.706	0.706	0.706	0.699	0.699
1,2,4,5,6,7	4,5,6				
0.699	0.694				

- Model with largest R_{adj}^2 is "1456".
- What about the best 3-predictor model ?
- Variable selection methods are sensitive to outliers, influential points, and transformations.

Criterion-Based Procedure: Cross-Validation

- *PRESS* = Predicted REsidual Sum of Squares (Leave-one-out cross-validated residuals):

$$PRESS = \sum_{i=1}^n (\hat{Y}_{-i} - Y_i)^2.$$

- \hat{Y}_{-i} is the prediction for i -th observation, using a model fitted without the i -th observation.
- Cross-validation criterion estimates the mean-squared error of prediction as:

$$CV \equiv \frac{\sum_{i=1}^n (\hat{Y}_{-i} - Y_i)^2}{n} = \frac{PRESS}{n}$$

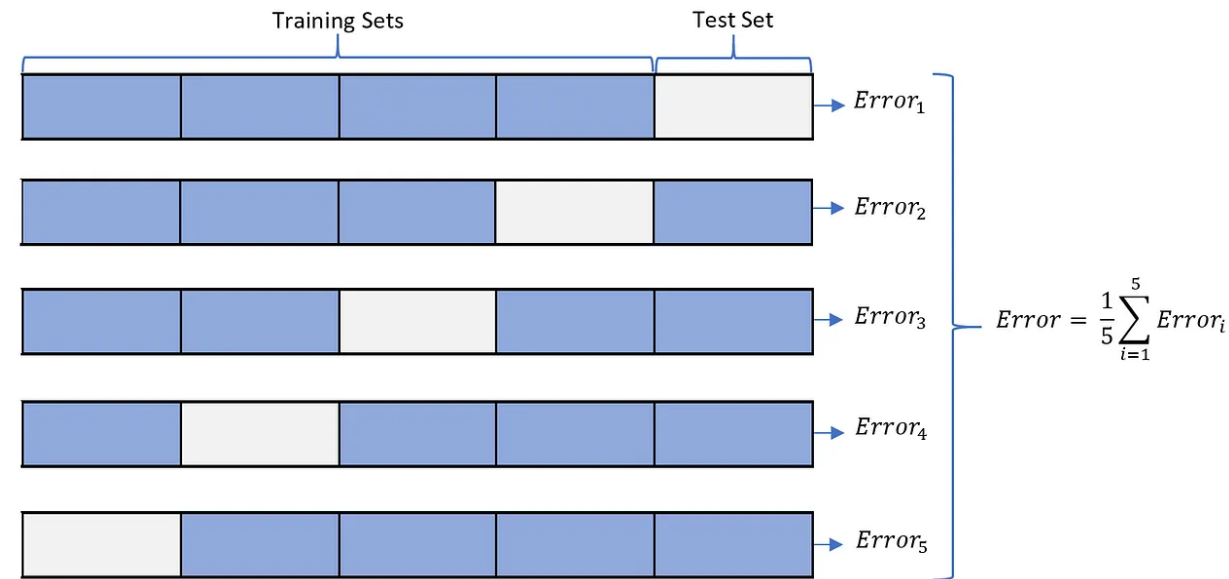
- We prefer the model with the smallest value of CV or PRESS.

What is the drawback of LOOCV?

Criterion-Based Procedure: PRESS and Cross-Validation

- Alternative is k -fold cross validation:

1. Divide the data into small number of subsets or **folds** (e.g., 5 or 10) of roughly equal size
2. Fit a model omitting each subset in turn (i.e., use only the training data)
3. Obtain the fitted values for all observations in omitted subset (i.e., the test data).



Criterion-Based Procedure: *AIC* and *BIC*

- Penalized model-fit statistics:

$$-2 \log L(\hat{\theta}) + \textit{penalty}$$

- θ is the vector of parameters of the model, including the regression coefficients and the error variance. Here, $\hat{\theta}$ is the [m.l.e.](#)
 - $L(\hat{\theta})$ is the maximized likelihood under current model.
 - $\textit{penalty} \equiv c \times p$ (c is a scaling parameter).
-
- The magnitude of a criterion is not interpretable, but differences are.
 - Model with the smallest value of information criterion is preferred.

Criterion-Based Procedure: *AIC* and *BIC*

- Most popular criteria are

$$AIC = -2 \log L(\hat{\theta}) + 2p, \text{ thus } c = 2$$

$$BIC = -2 \log L(\hat{\theta}) + p \log(n), \text{ thus } c = \log(n)$$

- When the sample size n is small, there is a high chance that AIC will select models that have too many parameters (i.e., AIC will overfit). In this case, the **corrected AIC** can be used:

$$AIC_c = AIC + \frac{2p^2 + 2k}{n - k - 1}$$

Example: Life expectancy dataset

```
> statreg <- lm(life.exp ~ ., data = statedata)
```

```
> stepres <- step(statreg, steps = 2)
```

Start: AIC=-22.18

life.exp ~ population + income + illiteracy + murder + highSchoolGrad +
frost + area

	Df	Sum of Sq	RSS	AIC
- area	1	0.0011	23.298	-24.182
- income	1	0.0044	23.302	-24.175
- illiteracy	1	0.0047	23.302	-24.174
<none>			23.297	-22.185
- population	1	1.7472	25.044	-20.569
- frost	1	1.8466	25.144	-20.371
- highSchoolGrad	1	2.4413	25.738	-19.202
- murder	1	23.1411	46.438	10.305

What would be the AIC value, if the predictor is removed.

The variables are removed step-by-step, and AIC is checked at each step.

Step: AIC=-24.18

life.exp ~ population + income + illiteracy + murder + highSchoolGrad +
frost

	Df	Sum of Sq	RSS	AIC
- illiteracy	1	0.0038	23.302	-26.174
- income	1	0.0059	23.304	-26.170
<none>			23.298	-24.182
- population	1	1.7599	25.058	-22.541
- frost	1	2.0488	25.347	-21.968
- highSchoolGrad	1	2.9804	26.279	-20.163
- murder	1	26.2721	49.570	11.569

Step: AIC=-26.17

life.exp ~ population + income + murder + highSchoolGrad + frost

Summary model selection

- Stepwise procedures:
 - Search through space of potential models.
 - Testing-based procedures use dubious hypothesis testing.
- Criterion-based procedures:
 - search through a wider space of models ("all possible subsets regression")
 - compare the models using a particular criterion.
- Criterion based procedures are usually preferred.

Summary model selection

- The aim of variable selection is to construct a model that predicts well or explains relationships in data well.
- It is part of process of model building, like identification of outliers and influential points, and variable transformation.
- Automatic selections are not guaranteed to be consistent. Use methods as guide only.
- Accept possibility that several models are suggested which fit equally well. Then consider:
 - Do models have similar qualitative consequences?
 - Do they make similar predictions?
 - What is cost of measuring predictors?
 - Which has best diagnostics?

Model validation

- **Model validation**: split dataset into **training subsample** and **validation subsample**:
 - Training subsample is used to specify the statistical model.
 - Validation subsample is used to evaluate the fitted model.
- **Cross-validation** is an application of this idea where the roles of training and validation subsamples are interchanged.
- **Statistical modeling**: iterative sequence of data exploration, model fitting, model criticism, model re-specification.
- Variables may be dropped, interactions may be incorporated or deleted, variables may be transformed, unusual data may be corrected, removed, or otherwise accommodated.

Example: Model validation



```
> library(caret)
> set.seed(123)
> # creating training data as 80% of the dataset
> random_sample <- createDataPartition(statedata$life.exp, p = 0.8, list = FALSE)
> # training dataset from the random_sample
> training_dataset <- statedata[random_sample, ]
> # testing dataset from rows which are not included in random_sample
> testing_dataset <- statedata[-random_sample, ]

> # Building the model on the training data
> statereg <- lm(life.exp ~., data = training_dataset)
> # predicting the target variable
> predictions <- predict(statereg, testing_dataset)
```

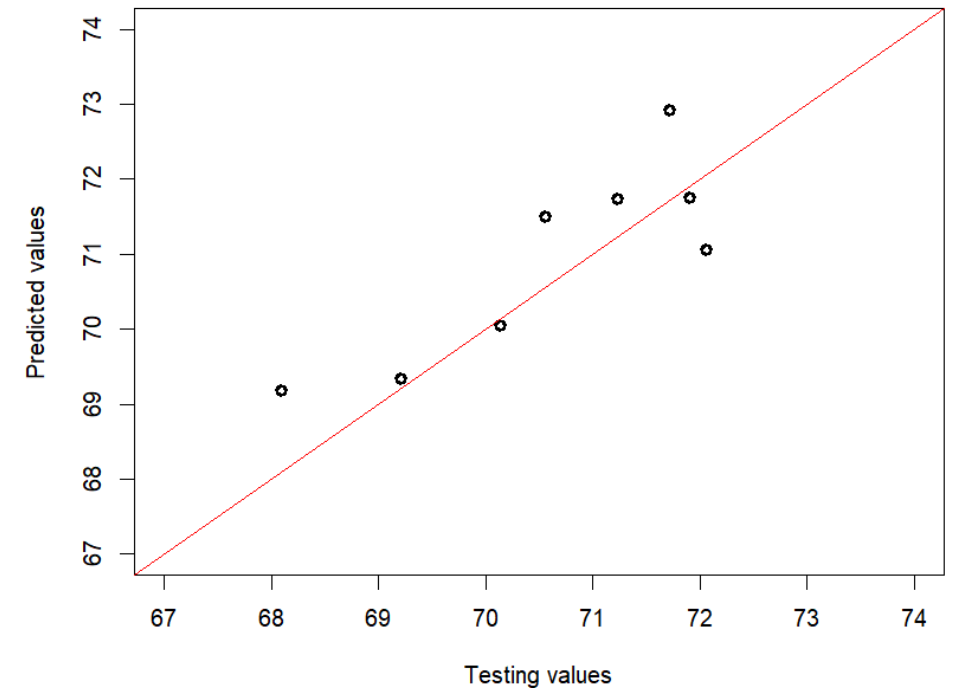
Example: Model validation

```
> plot(predictions ~ testing_dataset$life.exp,  
+       xlab="Testing values", ylab = "Predicted values",  
+       lwd= 3, xlim = c(67, 74), ylim=c(67,74))  
> abline(a=0, b=1, col="red")
```

- Evaluation is usually done using Pearson's correlation or Root mean squared error (RMSE) metrics.

```
> # computing model performance metrics  
> (R2 <- R2(predictions, testing_dataset$life.exp))  
[1] 0.720348  
> (RMSE <- RMSE(predictions, testing_dataset$life.exp))  
[1] 0.7794504
```

- Usually, several random partition is used.



Model validation

- Resulting model should accurately reflect the principal characteristics of your data.
- Danger: **overfitting and overstating strength of results.**
- **Ideal solution:** collect new data with which to validate model (often not possible).
- Model validation simulates the collection of new data by randomly dividing data into two parts:
 - First, for exploration and model formulation,
 - second for checking adequacy of model, formal estimation, and testing.

Collinearity

- If a perfect linear relationship among regressors exist, least-squares coefficient are no longer uniquely defined.
- Strong, but less-than-perfect linear relationship among X 's causes least-squares coefficients to be unstable:
 - large standard errors of coefficients,
 - broad confidence intervals,
 - hypothesis tests with low power.

Collinearity and Remedies

- Small changes in data can greatly change the coefficients.
- Large changes in coefficients coincide with only very small changes in residual sum of squares.
- This problem is known as **collinearity or multicollinearity**.
 - Collinearity is a relatively rare problem in social-science applications of linear models.
 - Methods employed as remedies for collinearity **may be worse than the disease**.
 - Usually, it is **impossible to redesign study** to decrease correlations between X 's.

Detecting Collinearity

- Suppose a perfect linear relationship exists between X 's:

$$c_1X_{i1} + c_2X_{i2} + \dots + c_kX_{ik} = c_0.$$

- Then, the matrix $\mathbf{X}'\mathbf{X}$ is **singular** (why?),
- Therefore.
 - least squares normal equations do not have unique solution
 - sampling variances of regression coefficients are infinite.
- Perfect collinearity is often a product of some error in formulating linear model
 - e.g., too many dummies.

Detecting Collinearity

- The sampling variance of the slope B_j is

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{(n - 1)S_j^2}$$

Here, R_j^2 is the R^2 for regression of X_j on other X 's .

S_j^2 is the sample variance of X_j .

- The first term is called **Variance Inflation Factor**:

$$VIF = \frac{1}{1 - R_j^2}$$

- VIF indicates directly the impact of collinearity on precision of B_j .
- VIF is a basic diagnostic for collinearity.
- A rule of thumb**: VIF is greater than 10 (very strong) or 5 (strong) (why these values?).

Detecting Collinearity (Optional)

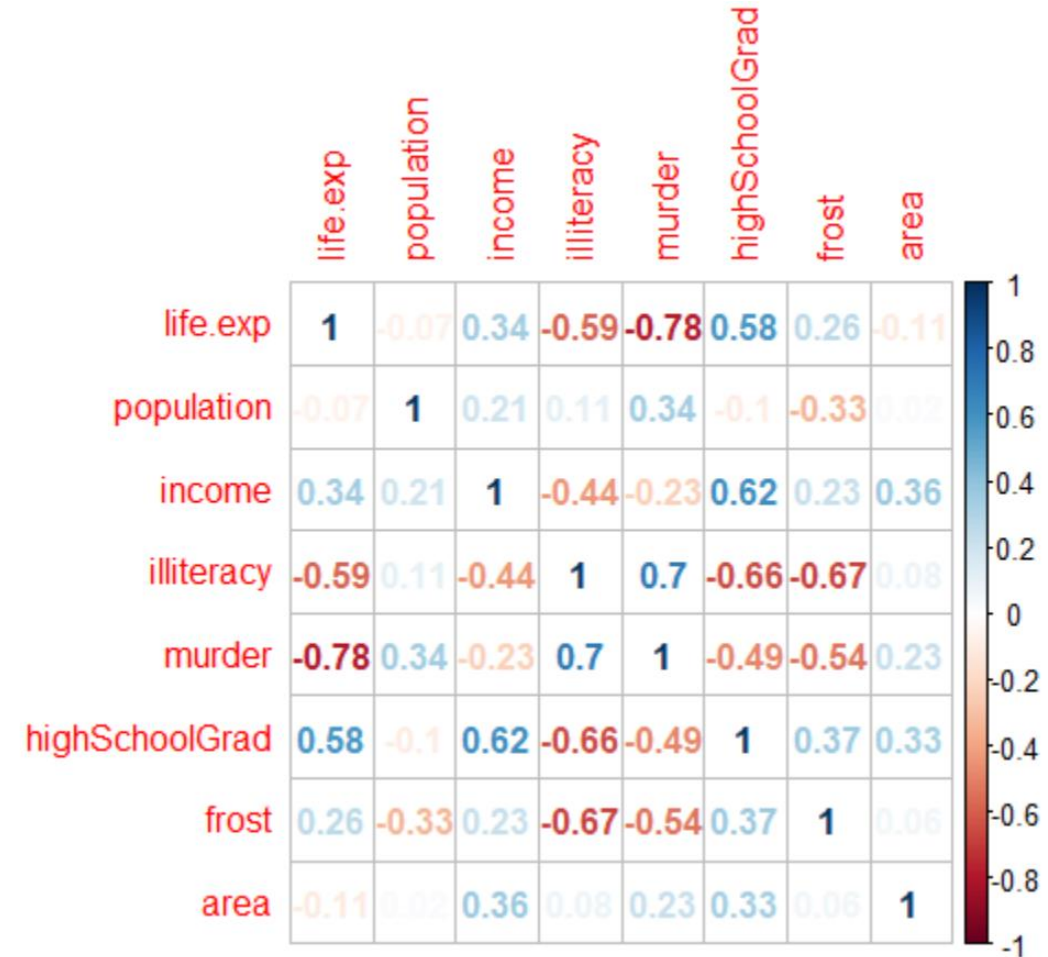
- Ways for detecting collinearity besides looking at the VIF's.
 - Examination of the correlation matrix of predictors:
 - Regress X_i on all other X 's and repeat this for all predictors. R_i^2 close to one indicates the problem.
 - Examine the eigenvalues λ_i of $\mathbf{X}'\mathbf{X}$: small eigenvalues indicate problem.
 - Large **condition numbers** $\kappa(\mathbf{X}'\mathbf{X}) = \sqrt{\lambda_1/\lambda_p}$ ($\kappa > 30$ is considered large).
 - Also check the values of **condition index** $\sqrt{\lambda_1/\lambda_i}$

Detecting Collinearity: Example

```
> statreg <- lm(life.exp ~ ., data = statedata)
> vif(statreg)
```

population	income	illiteracy
1.499915	1.992680	4.403151
murder	highSchoolGrad	frost
2.616472	3.134887	2.358206
area		
1.789764		

- What is your opinion?



Collinearity: No Quick Fix

- Collinearity leads to
 - imprecise estimates of β ; even the signs of coefficients may be misleading.
 - t-tests fail to reveal significant factors.
- Coping With Collinearity
 - Model re-specification.
 - Variable Selection.
 - Biased Estimation: e.g., Ridge Regression..
 - Prior Info About Regression Coefficients: e.g., Bayesian approaches.

Geometric interpretation of collinearity (Optional)

- Imagine a table: as two diagonally opposite legs are moved closer together, the table becomes increasing
- no collinearity (a), complete collinearity (b) and strong collinearity

