Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine
qing.li2@uclagary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents
Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

One Health at UCalgary

UNIVERSITY OF CALGARY

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance assumptions.
  - Lecture 7: Multiple regression diagnostics: verify normality assumptions and identify multicollinearity and outliers.
  - Lecture 8: Multiple regression diagnostics: data transformation
- Topic 5: Transfer learning
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

Learning Outcomes: At the end of the course, participants will be able to

1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize correlations between a response variable (Y) and all explanatory variables (X) by graphs (scatter plot, residual plot) to assess model validity.
2. Predict the response variable at a certain level of the explanatory variables once the fit model exists.
3. Implement R-software and analyze statistical results for biomedical and other data.

- **Evaluations**
  1. Assignments will be posted on Slack (our communication tool with students).
  2. Students must attend 70% (6/9) of the sessions in order to receive the certificate and are encouraged to work on the assignments progressively throughout the course as the relevant material is covered.

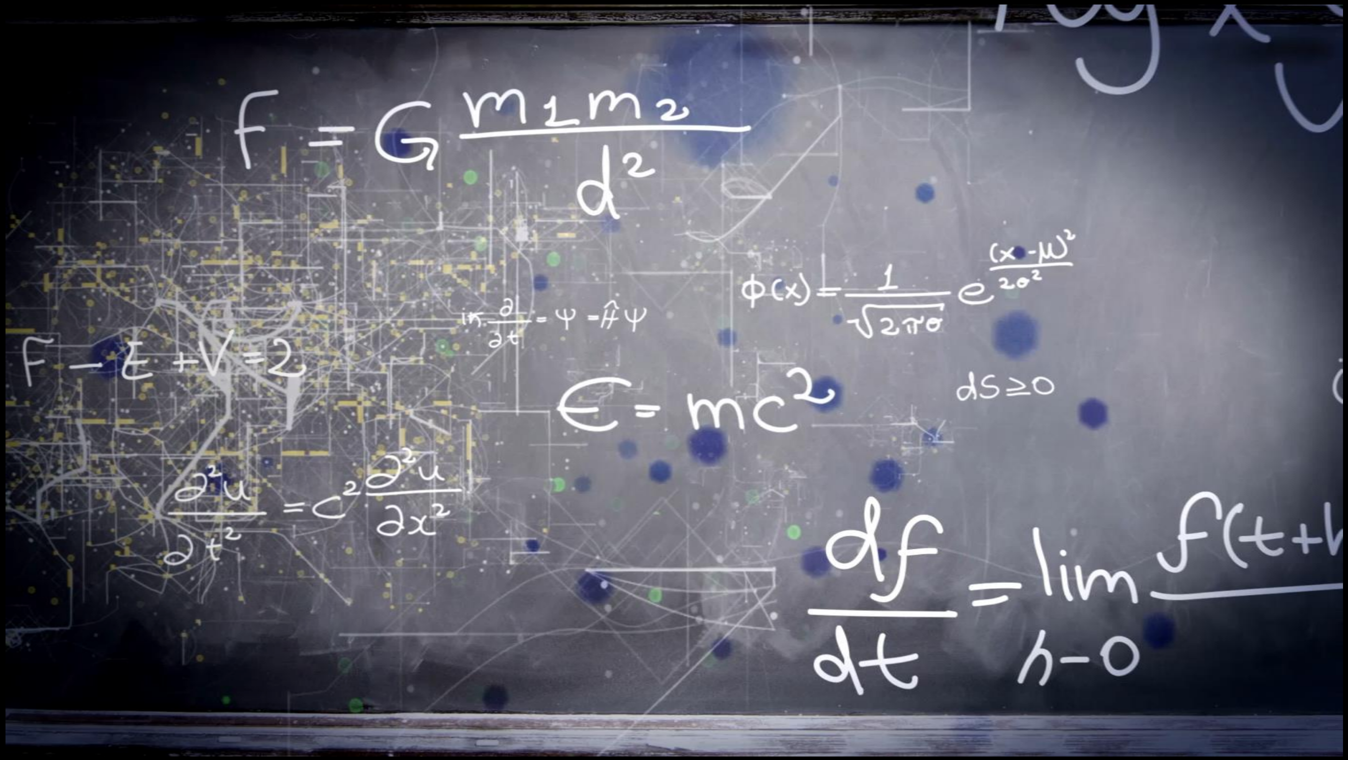
One Health at UCalgary

UNIVERSITY OF CALGARY

Statistical Modelling with Data

- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
  - Recoding videos
  - Exercises
  - Course-documents

Lecture 4: Model Selection



Quick recap of lecture 3

- Statistics:
  - Interaction Effect in Multiple Regression with both Quantitative and Qualitative (Dummy) Variable models
  - Two different approaches but result in the same optimal model
  - A Quadratic (Second Order) Model with Quantitative predictors
  - Cross validation to avoid overfitting and underfitting model
- Code:
  - `lm(y ~ x1+x2+(x1+x2)^2 + I(X1^2)+I(X2^2))`
  - `CVlm()`

Model Selection

- Stepwise model selection methods
  - Backwards Elimination
  - Forward Selection
  - Stepwise regression
- Criteria for model selection
  - p-values
  - adjusted R^2
  - RMSE
  - Mallows’s Cp Criterion
  - AIC (Akaike’s information criterion)

Backward Elimination Procedure

- Step 1: The Backward procedure initially fits a model containing terms for all potential independent variables

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- Step 2: The variable with the smallest t (or F) test statistic (largest p-value) for testing $H0{:}\beta i{=}0$ is identified and dropped from the model if the t-value is less than some specified critical value or p-value more than a cut-off (common: 0.3)
- The model with the remaining $(p-1)$ independent variables is fit in step 2, and again, the variable associated with the smallest nonsignificant t-value (largest p-value) is dropped. This process is repeated until no further nonsignificant independent variables can be found.

Backward Elimination Procedure

Dr. Thuntida Ngamkham's approach

1. Build an additive model
2. Determine significant predictors
3. Build an interaction model with significant predictors
4. Remove non-significant interactions
5. Rerun model to ensure all predictors are significant
6. Iterate at step 5 until done

Leah's approach:

1. Start with an interaction model with all predictors
2. Remove non-significant interactions
3. Rerun model to ensure all predictors are significant
4. Iterate step 3 until done.

Similar to backward elimination. But backward elimination is more rigorous as it only remove one predictor at a time

Forward selection

- Do a similar thing like a backward selection but the other way around.

$$null\ model:\ Y = \beta_0$$

- Step 1: The software program fits all possible one-variable models of the form. The independent variable that produces the largest (absolute) t -value (smallest p-value) is then declared the best one-variable predictor of Y. Call this independent variable X_1.

$$Y = \beta_0 + \beta_1 X_1$$

Forward selection

- Step 2: The stepwise program now begins to search through the remaining (p−1) independent variables for the best two-variable model of the form

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_i$$

- This is done by fitting all two-variable models containing X_1 and each of the other (p−1) options for the second variable X_i. The t-values for the test H_0:β_2=0 are computed for each of the p−1 models (corresponding to the remaining independent variables, X_i,i=2,3,...,p−1), and the variable having the largest t is retained. Call this variable X_2.
- Repeat this step 2 until the model go through all predictors

Stepwise Regression

- Similar to forward selection, but add one more step after adding additional predictor
- Null model
- Add 1st predictor
- When 1 more predictor on top of the 1st predictor : After step 2 of forward selection, the stepwise routine will go back and check the t-value of $\widehat{\beta_1}$ after $\widehat{\beta_2} X_2$ has been added to the model (check step). If the t-value has become nonsignificant at some specified α level (say α=0.3), the variable X_1 is removed and a search is made for the independent variable with a β parameter that will yield the most significant t-value in the presence of $\widehat{\beta_2} X_2$.

Comparisons of three approaches

To test whether a predictor can be added/removed from the model:

	STEP1	STEP2	STEP3
Backward elimination	The optimal model (Start from Full model)	Identify the predictor with largest p-value and remove it predictor if its p-value is larger than a threshold in anova (f-test).	NA
Forward selection	The optimal model (Start from Y ~ 1)	Identify the predictor with smallest p-value its p-value is smaller than a threshold in t-test.	NA
Stepwise regression	The optimal model (Start from Y ~ 1)	Identify the predictor with smallest p-value its p-value is smaller than a threshold in t-test.	Check the p-value for predictors before adding the new predictor, replace it if its p-value is no longer significant after adding the new predictor.

One Health at UCalgary

UNIVERSITY OF CALGARY

In class Practice Problem 10

Use the Executive Salary (EXECSAL2.csv) data.

(a) Use backward elimination to identify potentially important predictor variables (make sure to specify which are qualitative).

(b) Use forward selection to identify potentially important predictor variables. Are they the same as backward elimination?

(c) Use stepwise selection to identify potentially important predictor variables. Are they the same as previous two methods?

Lecture_4.R

```
> backmodel

                              Elimination Summary
-----------------------------------------------------------------------------
        Variable                       Adj.
Step     Removed      R-Square      R-Square     C(p)         AIC        RMSE
-----------------------------------------------------------------------------
   1    X10             0.9228        0.9151    9.1091    -220.9265    0.0757
   2    X7              0.9227        0.9159    7.1956    -222.8295    0.0753
   3    X8              0.9225        0.9166    5.4499    -224.5448    0.0750
   4    factor(X6)       0.922         0.917    4.0235    -225.9056    0.0749
-----------------------------------------------------------------------------
```

```
> forwardmodel

                              Selection Summary
-----------------------------------------------------------------------------
        Variable                      Adj.
Step     Entered      R-Square      R-Square     C(p)          AIC        RMSE
-----------------------------------------------------------------------------
   1    X1              0.6190        0.6151    343.8566     -77.2678    0.1612
   2    factor(X3)      0.7492        0.7440    195.5192    -117.0905    0.1314
   3    X4              0.8391        0.8341     93.7538    -159.4705    0.1058
   4    X2              0.9075        0.9036     16.8128    -212.8048    0.0807
   5    X5              0.9206        0.9164      3.6279    -226.1391    0.0751
   6    factor(X9)      0.9220        0.9170      4.0235    -225.9056    0.0749
-----------------------------------------------------------------------------
```

```
> ols_step_both_p(full_model, pent=0.3, prem = 0.3, details = FALSE)

                             Stepwise Selection Summary
-------------------------------------------------------------------------------------
                        Added/                     Adj.
Step    Variable       Removed     R-Square     R-Square     C(p)        AIC        RMSE
-------------------------------------------------------------------------------------
   1       X1          addition       0.619        0.615    343.8570    -77.2678    0.1612
   2    factor(X3)     addition       0.749        0.744    195.5190   -117.0905    0.1314
   3       X4          addition       0.839        0.834     93.7540   -159.4705    0.1058
   4       X2          addition       0.907        0.904     16.8130   -212.8048    0.0807
   5       X5          addition       0.921        0.916      3.6280   -226.1391    0.0751
   6    factor(X9)     addition       0.922        0.917      4.0240   -225.9056    0.0749
-------------------------------------------------------------------------------------
```

One Health at UCalgary

In class Practice Problem 10

Save you day with "olsrr" package!
olsrr: Tools for Building OLS Regression Models

>backmodel=ols_step_backward_p(fullmodel, prem = 0.3, details=TRUE)
>forwardmodel=ols_step_forward_p(fullmodel, penter = 0.3, details=TRUE)
>stepwisemodel=ols_step_both_p(fullmodel, pent = 0.3, prem=0.3, details=TRUE)

Lecture_4.R

In class Practice Problem 11

Use the CREDIT.CSV data.

(a) Use **backward elimination** to identify potentially important predictor variables (make sure to specify which are qualitative).
(b) Use **forward selection** to identify potentially important predictor variables.
(c) Are the results the same?

Lecture_4.R

In class Practice Problem 11

```
> backward_model
```

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Education	0.955	0.9539	8.4729	4821.8576	98.7238
2	factor(Ethnicity)	0.9549	0.954	7.9305	4819.3557	98.6556
3	factor(Married)	0.9548	0.954	6.4620	4817.9006	98.5968

backward_model=lm(Balance ~ Income+Limit+Rating+Cards+Age+factor(Gender)+factor(Student), data=credit)

```
> forward_model = ols_step_forward_p(full_model,penter=0.3)
> forward_model
```

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Rating	0.7458	0.7452	1800.3084	5496.7815	232.0713
2	Income	0.8751	0.8745	685.1965	5214.5571	162.8813
3	factor(Student)	0.9499	0.9495	41.1339	4851.3870	103.3189
4	Limit	0.9522	0.9517	23.1825	4834.5240	101.0389
5	Cards	0.9542	0.9536	8.1316	4819.6668	99.0576
6	Age	0.9547	0.9540	5.5749	4817.0390	98.6115
7	factor(Gender)	0.9548	0.9540	6.4620	4817.9006	98.5968

forward_model=lm(Balance ~ Rating+Income+Factor(student)+Limit+Cards+Age+factor(Gender), data=credit)

Lecture_4.R

In class Practice Problem 12

Use the CREDIT.CSV data.

a) Use **stepwise regression** to find the potentially important independent variables for predicting credit card balance. (make sure to specify which are qualitative).

b) Is the result the same as the previous two approach?

Lecture_4.R

In class Practice Problem 12

```
> credit_stepwise
```

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Rating	addition	0.746	0.745	1800.3080	5496.7815	232.0713
2	Income	addition	0.875	0.874	685.1970	5214.5571	162.8813
3	factor(Student)	addition	0.950	0.949	41.1340	4851.3870	103.3189
4	Limit	addition	0.952	0.952	23.1820	4834.5240	101.0389
5	Cards	addition	0.954	0.954	8.1320	4819.6668	99.0576
6	Age	addition	0.955	0.954	5.5750	4817.0390	98.6115
7	factor(Gender)	addition	0.955	0.954	6.4620	4817.9006	98.5968

stepwise_model=lm(Balance ~ Rating+Income+factor(student)+Limit+Cards+Age+factor(Gender), data=credit)

forward_model=lm(Balance ~ Rating+Income+factor(student)+Limit+Cards+Age+factor(Gender), data=credit)

Lecture_4.R

Coffee break

All Possible Regressions Selection Criteria

1. R^2 Criterion the multiple coefficient of determination

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

will increase when independent variables are added to the model. Therefore, the model that includes all p independent variables $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$ will yield the largest R^2.

2. Adjusted R^2 or RMSE Criterion

We can use the adjusted R^2 instead of R^2. It is easy to show that R^2_{adj} is related to MSE as follows:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

$$R^2_{adj} = 1 - (n-1)\frac{MSE}{SST}$$

$$s = RMSE = \sqrt{\frac{1}{n-p-1}SSE}$$

Adjusted based on the number of predictors

Note that R^2_{adj} increases only if RMSE decreases [since SST remains constant for all models]. Thus, an equivalent procedure is to search for the model with the minimum, or near minimum, RMSE.

3. Mallows's Cp Criterion

The Cp criterion, named for Colin Lingwood Mallow, selects as the best subset model with

1. a small value of Cp (i.e., a small total mean square error), means that the model is relatively precise.
2. a value of Cp near p + 1, a property that indicates that slight (or no) bias exists in the subset regression model.

Thus, the Cp criterion focuses on minimizing total mean square error and the regression bias. If we are mainly concerned with minimizing total mean square error, we will want to choose the model with the smallest Cp value, as long as the bias is not large. On the other hand, we may prefer a model that yields a Cp value slightly larger than the minimum but that has slight (or no) bias.

4. AIC (Akaike's information criterion)

When using the model to predict Y, some information will be lost. Akaike's information criterion estimates the relative information lost by a given model. It is defined as

$$AIC = n\ln(\frac{SSE}{n}) + 2(p+1)$$

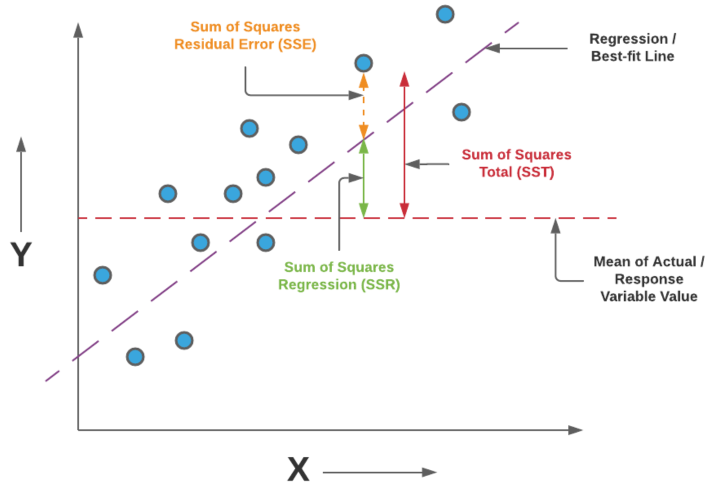
The formula is formulated by the statistician **Hirotugu Akaike**. Models with smaller values of AIC are preferred.

Hen, these are outdated! We have a powerful tool that is CV MSE! DAAG, CVlm()
That's the point of view of machine learning engineers and data analysts.
But statisticians sometimes still use these criteria as they are more focused on attributes of parameters
Besides, cross validation works better with a lot of samples. Sometimes, we only have a few samples (n<500).

The fans knew that no truck could beat Motozilla. They figured it would be an easy win for the monster—like always. But the challenger wasn't *one* truck. It was a *team of trucks*.

Criteria for model selection

- R^2 Criterion the multiple coefficient of determination



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

- R^2 will increase when independent variables are added to the model. Therefore, the model that includes all p independent variables $E(Y)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p$ will yield the largest R^2.

RMSE or Adjusted R^2

- Adjusted R^2 or RMSE Criterion

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

n: samples size, p: # of predictors

$$R^2_{adj} = 1 - (n-1)\frac{MSE}{SST}$$

$$R^2_a = R^2 - \left(\frac{p-1}{n-p}\right)(1-R^2)$$

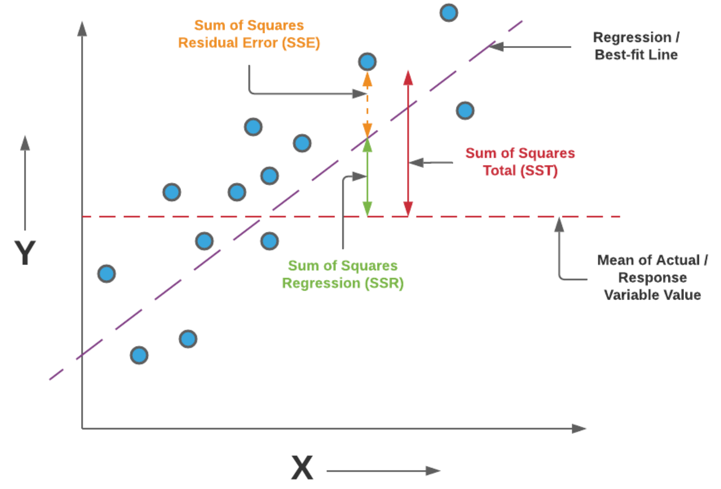
$$s = RMSE = \sqrt{\frac{1}{n-p-1}SSE}$$

AIC (Akaike's information criterion)

- Akaike's information criterion estimates the relative information lost by a given model.

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

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



- Models with smaller values of AIC are preferred.

BIC (Bayesian information criterion)

- Akaike's information criterion estimates the relative information lost by a given model.

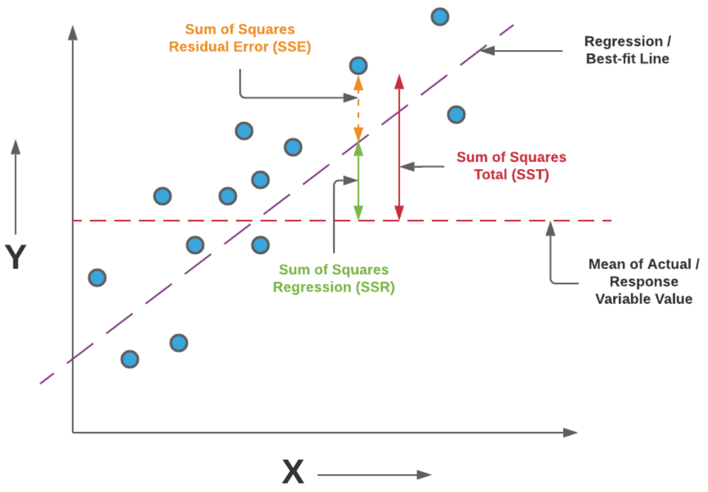
$$BIC = -2\ln\left(\hat{L} \approx \frac{SSE}{N}\right) + 2\ln(n)\,p$$

- Models with small values of BIC are preferred.

- Akaike's information criterion estimates the relative information lost by a given model.

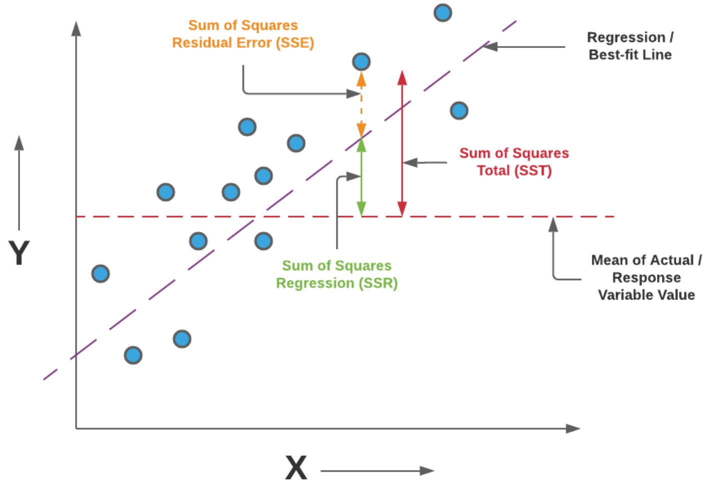
$$BIC = -2\ln\left(\hat{L} \approx \frac{SSE}{N}\right) + 2\ln(n)\,p$$

- Models with small values of BIC are preferred.



Mallows's C_p Criterion

- A measure for evaluating and comparing models
- Similar in usage to other model evaluation techniques:
  - Adjusted R2
  - Standard error of the regression / RMSE /s
  - AIC and BIC
- Estimated the magnitude of bias (prediction error) present in predicted val when important features are missing; large residuals.
- Here, p full model parameter numbers, k is the subset (reduced) with k adjustable model parameters, n is the sample size
- Full model: $C_p = \mathrm{p} + 1$
- For reduced mode: $C_p = \frac{SSE_p}{\hat{\sigma}_\varepsilon^2} - n + 2(p+1) = \frac{SSE_p}{MSE_k} - n + 2(p+1)$
- If the subset model is correct, the C_p should be about equal to p (smaller is better)
- In best subsets regression, we are looking for the reduced model that is closest to $C_p = \mathrm{k} + 1$ approaching from above.
- Only useful when n >>k



$$SSE_p = \sum^N (Y_i - Y_{pi})^2$$

$$\text{MSEk}=\frac{1}{N-K}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Combos 5

- #Construct a table for combos 5
- library(olsrr)
- salary=read.csv("EXECSAL2.csv", header = TRUE)
- firstordermodel<-lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data= salary)
- *#Select the subset of predictors that do the best at meeting some well-defined objective criterion, such*
- ks=ols_step_backward_p(firstordermodel, details=TRUE)
- *# for the output interpretation*
- R2<-c(ks$rsquare)
- AdjustedR2<-c(ks$adjr)
- Cp<-c(ks$cp)
- AIC<-c(ks$aic)
- RMSE<-c(ks$rmse)
- cbind(R2, AdjustedR2,Cp,AIC, RMSE)

```
> forwardmodel

                              Selection Summary
-----------------------------------------------------------------------------------
        Variable                     Adj.
Step     Entered        R-Square    R-Square     C(p)          AIC         RMSE
-----------------------------------------------------------------------------------
   1    X1                0.6190      0.6151    343.8566     -77.2678    0.1612
   2    factor(X3)        0.7492      0.7440    195.5192    -117.0905    0.1314
   3    X4                0.8391      0.8341     93.7538    -159.4705    0.1058
   4    X2                0.9075      0.9036     16.8128    -212.8048    0.0807
   5    X5                0.9206      0.9164      3.6279    -226.1391    0.0751
   6    factor(X9)        0.9220      0.9170      4.0235    -225.9056    0.0749
-----------------------------------------------------------------------------------

> stepwisemodel

                             Stepwise Selection Summary
---------------------------------------------------------------------------------------
                        Added/                  Adj.
Step     Variable      Removed     R-Square    R-Square     C(p)        AIC        RMSE
---------------------------------------------------------------------------------------
   1        X1         addition       0.619       0.615    343.8570    -77.2678    0.1612
   2    factor(X3)     addition       0.749       0.744    195.5190   -117.0905    0.1314
   3        X4         addition       0.839       0.834     93.7540   -159.4705    0.1058
   4        X2         addition       0.907       0.904     16.8130   -212.8048    0.0807
   5        X5         addition       0.921       0.916      3.6280   -226.1391    0.0751
---------------------------------------------------------------------------------------
```

Combos 5	R2 (+)	Adjusted R2 (+)	Cp (k+1)	AIC/BIC (-)	RMSE (-)
Model1					
…					
Model5					

In class Practice Problem 13

Use the CREDIT.CSV data.
From the credit card example, using the **All Possible Regressions Selection Criteria** (combo5) to analyze which independent predictors should be used in the model.

Hints: construct combos 5 tables for the best models you get from problem 11, 12 and make a final decision.

Lecture_4.R

In class Practice Problem 13 Answers

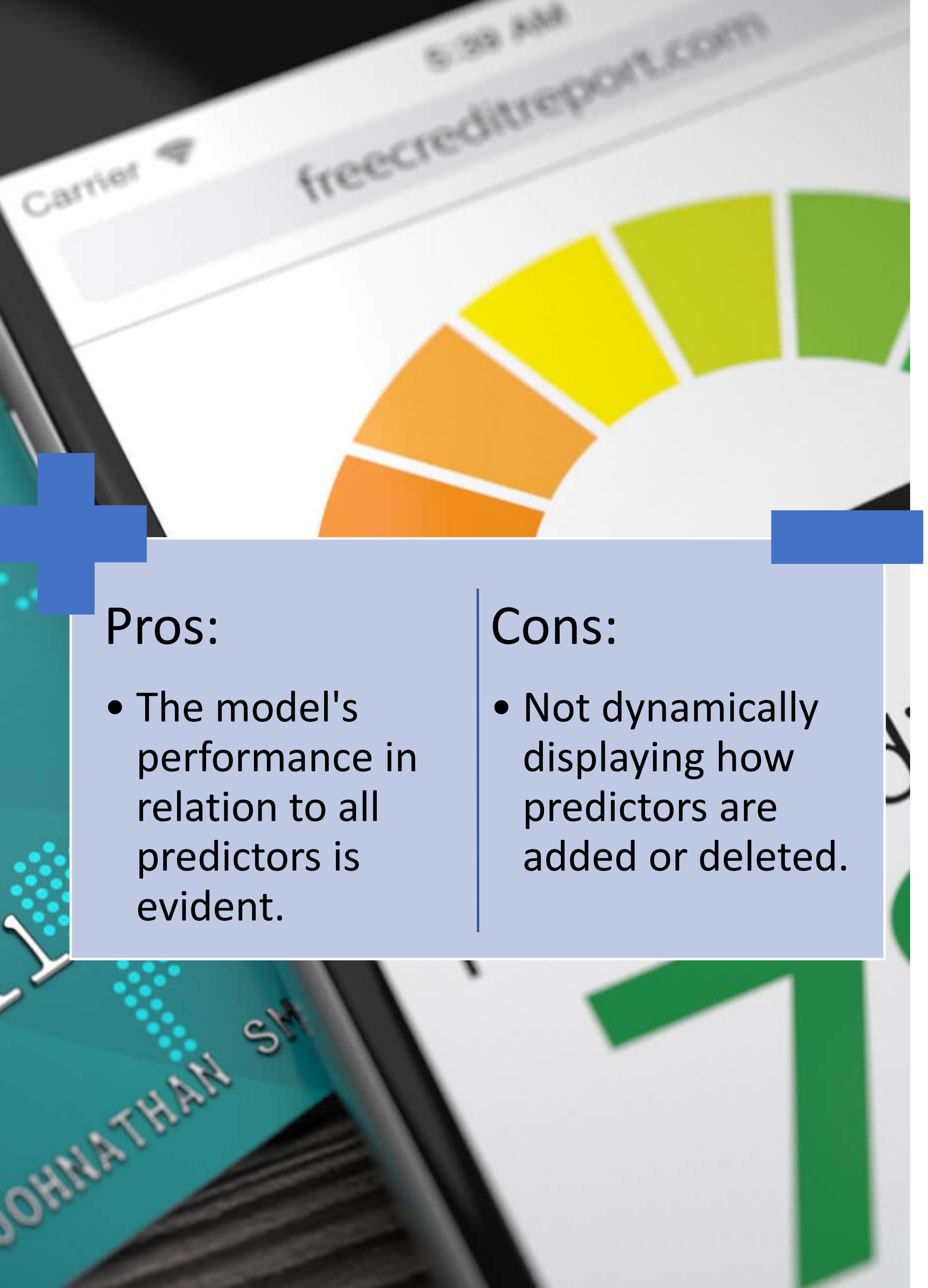
```
> forward_model

                                Selection Summary
--------------------------------------------------------------------------------------
        Variable                            Adj.
Step       Entered          R-Square      R-Square       C(p)          AIC         RMSE
--------------------------------------------------------------------------------------
   1    Rating               0.7458        0.7452      1800.3084    5496.7815    232.0713
   2    Income               0.8751        0.8745       685.1965    5214.5571    162.8813
   3    factor(Student)      0.9499        0.9495        41.1339    4851.3870    103.3189
   4    Limit                0.9522        0.9517        23.1825    4834.5240    101.0389
   5    Cards                0.9542        0.9536         8.1316    4819.6668     99.0576
   6    Age                  0.9547        0.9540         5.5749    4817.0390     98.6115
   7    factor(Gender)       0.9548        0.9540         6.4620    4817.9006     98.5968
--------------------------------------------------------------------------------------
```

```
> stepwise_model

                              Stepwise Selection Summary
--------------------------------------------------------------------------------------
                               Added/                   Adj.
Step        Variable          Removed     R-Square    R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------------------
   1         Rating           addition       0.746       0.745    1800.3080    5496.7815    232.0713
   2         Income           addition       0.875       0.874     685.1970    5214.5571    162.8813
   3     factor(Student)      addition       0.950       0.949      41.1340    4851.3870    103.3189
   4          Limit           addition       0.952       0.952      23.1820    4834.5240    101.0389
   5          Cards           addition       0.954       0.954       8.1320    4819.6668     99.0576
   6           Age            addition       0.955       0.954       5.5750    4817.0390     98.6115
   7     factor(Gender)       addition       0.955       0.954       6.4620    4817.9006     98.5968
--------------------------------------------------------------------------------------
```

Lecture_4.R



Pros:

- The model's performance in relation to all predictors is evident.

Cons:

- Not dynamically displaying how predictors are added or deleted.

In class Practice Problem 13 Answers

model	Variables	R2	AdjR2	Cp	AIC	RMSE
forward	Income	0.745848418	0.74521	1800.308	5496.781548	232.0713
forward	Limit	0.875117948	0.874489	685.1965	5214.557085	162.8813
forward	Rating	0.94987878	0.949499	41.13387	4851.386992	103.3189
forward	Cards	0.952187504	0.951703	23.1825	4834.524008	101.0389
forward	Age	0.954160597	0.953579	8.131573	4819.66682	99.05763
forward	Education	0.954687886	0.953996	5.574883	4817.038963	98.61148
forward	factor(Gender)	0.954816662	0.95401	6.462042	4817.90056	98.59677
forward	factor(Ethnicity)	0.745848418	0.74521	1800.308	5496.781548	232.0713
forward	factor(Married)	0.875117948	0.874489	685.1965	5214.557085	162.8813
forward	factor(Student)	0.94987878	0.949499	41.13387	4851.386992	103.3189
stepwise	Income	0.745848418	0.74521	1800.308	5496.781548	232.0713
stepwise	Limit	0.875117948	0.874489	685.1965	5214.557085	162.8813
stepwise	Rating	0.94987878	0.949499	41.13387	4851.386992	103.3189
stepwise	Cards	0.952187504	0.951703	23.1825	4834.524008	101.0389
stepwise	Age	0.954160597	0.953579	8.131573	4819.66682	99.05763
stepwise	Education	0.954687886	0.953996	5.574883	4817.038963	98.61148
stepwise	factor(Gender)	0.745848418	0.74521	1800.308	5496.781548	232.0713
stepwise	factor(Ethnicity)	0.875117948	0.874489	685.1965	5214.557085	162.8813
stepwise	factor(Married)	0.94987878	0.949499	41.13387	4851.386992	103.3189
stepwise	factor(Student)	0.952187504	0.951703	23.1825	4834.524008	101.0389

Lecture_4.R

Take away messages

- Statistics:
  - Backward elimination
  - Forward selection
  - Stepwise regression
  - All Possible Regressions Selection Procedure (makes your own selection criteria combo)
- Code:
  - lm(); anova();
  - ols_step_backward_p
  - ols_step_forward_p
  - ols_step_both_p
  - rbind(); cbind(); colnames(); write.csv(); getwd()

Thank you

- Questions OR Comments?

- Slack channel: section2-course-documents
- Email: qing.li2@uclagary.ca