# Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine

qing.li2@ucalgary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents

Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

# 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**

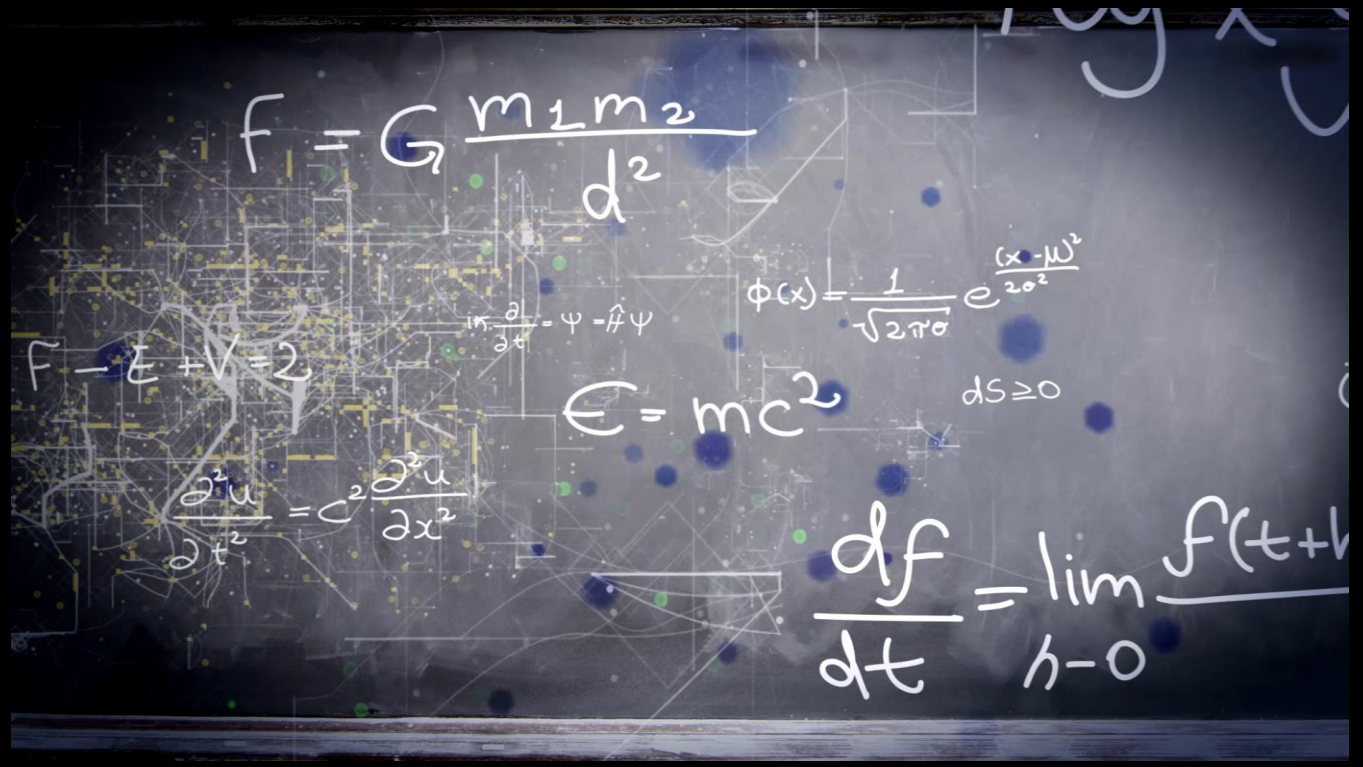
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize 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.


# Statistical Modelling with Data

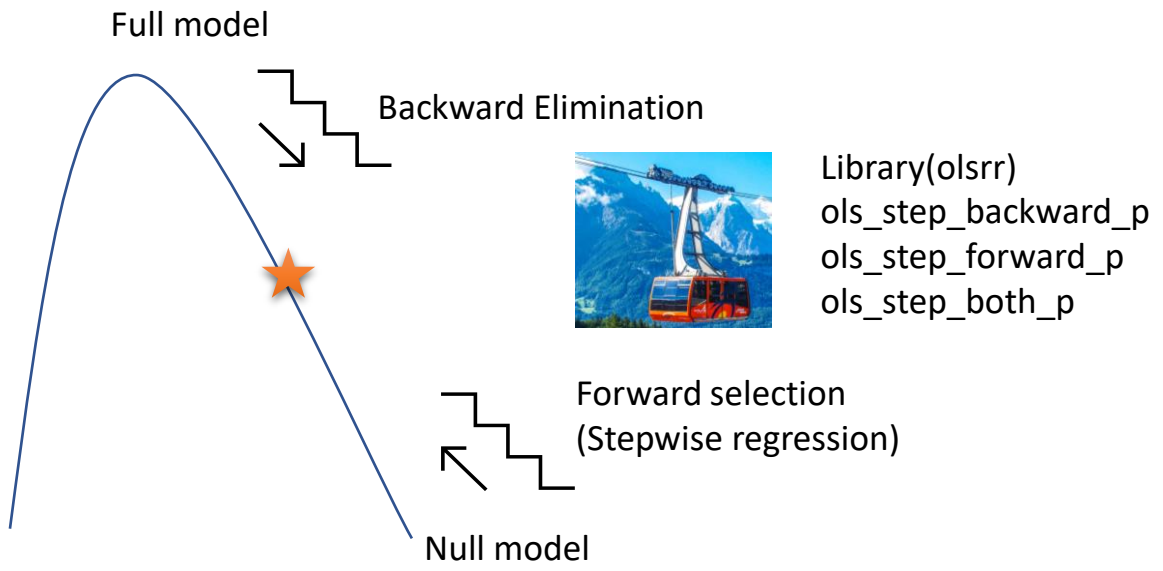
- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
    - Recoding videos
    - Exercises
    - Course-documents

# Lecture 5: Model Selection, evaluate the reliability of the model chosen



# Quick recap of lecture 4

- Statistics:
  - Backward elimination
  - Forward selection
  - Stepwise regression
  - All Possible Regressions Selection Criteria (makes your own combos of selection criteria)
    - R square
    - adjusted $R^2$
    - Mallows's Cp Criterion
    - AIC (Akaike's information criterion) or BIC (Bayesian information criterion)
    - RMSE
- Code:
  - lm(); anova();
  - ols_step_backward_p
  - ols_step_forward_p
  - ols_step_both_p
  - rbind(); cbind(); colnames(); write.csv(); getwd()

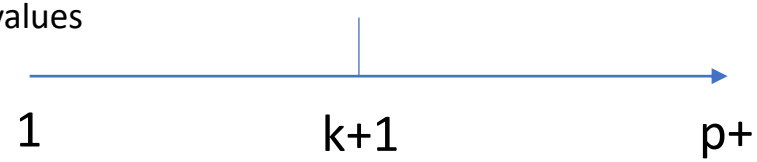



| Criteria Combos 5 | R2 (+) | Adjusted R2 (+) | Cp (k+1) | AIC/BIC (-) | RMSE (-) |
|---|---|---|---|---|---|
| Model1 | | | | | |
| … | | | | | |
| Model5 | | | | | |

Cp < k +1, sampling errors, no bias between predicted values and actual values

Cp near k +1, the optimal model

Cp > k +1, substantial bias between predicted values and actual values

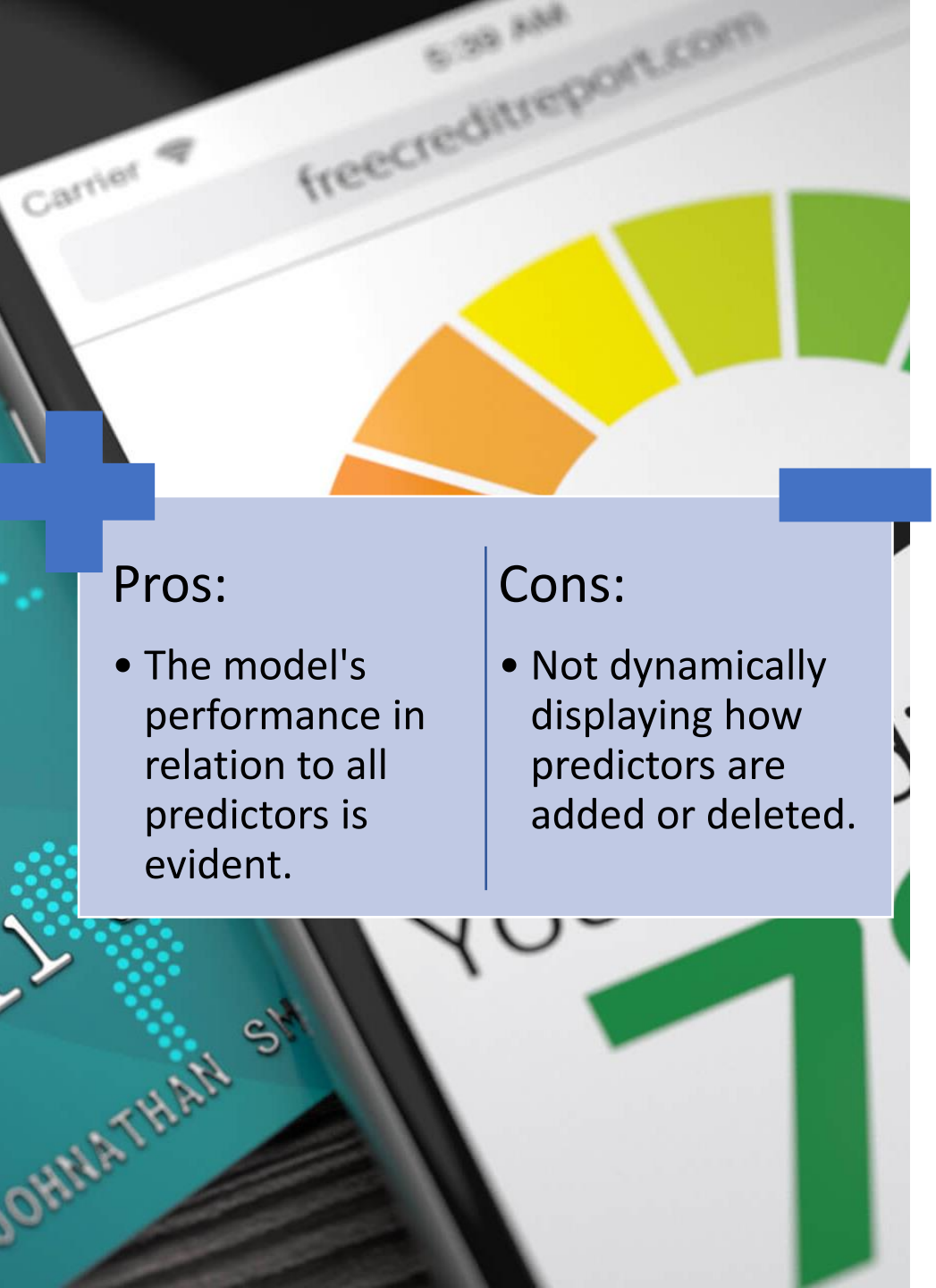



# In class Practice Problem 13

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

Lecture_4.R




One Health at UCalgary | UNIVERSITY OF CALGARY

# In class Practice Problem 13

| 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 |

**Pros:**

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

**Cons:**

- Not dynamically displaying how predictors are added or deleted.

Lecture_4.R

# Combos 5

- Model selection functions, such as ols_step_backward_p, ols_step_forward_p, ols_step_both_p, include or exclude predictor on their coefficients' p-values. Additionally, these functions also give values for criteria combos 5.
- Does the optimal model constructed based on p-values matches the one uses criteria combos 5?
- To answer this question: we use the function named *"ols_step_best_subset"*

# Comparison of the optimal models

The optimal model based on combos5 criterion

| Predictors | # of Predictors | R2 | AdjR2 | Cp | | AIC | RMSE |
|---|---|---|---|---|---|---|---|
| Rating | 1 | 0.745848418 | 0.745209846 | 1800.308406 | 1798.308406 | 5496.781548 | 21542838 |
| Income Rating | 2 | 0.875117948 | 0.874488819 | 685.1965138 | 682.1965138 | 5214.557085 | 10612200 |
| Income Rating factor(Student) | 3 | 0.94987878 | 0.949499073 | 41.13386746 | 37.13386746 | 4851.386992 | 4269973 |
| Income Limit Cards factor(Student) | 4 | 0.953580003 | 0.953109927 | 11.14890997 | 6.148909969 | 4822.701337 | 3964692 |
| Income Limit Rating Cards factor(Student) | 5 | 0.954160597 | 0.953578879 | 8.131573242 | 2.131573242 | 4819.66682 | 3925066 |
| Income Limit Rating Cards Age factor(Student) | 6 | 0.954687886 | 0.953996098 | 5.574883125 | 1.425116875 | 4817.038963 | 3889814 |
| Income Limit Rating Cards Age factor(Ethnicity) factor(Student) | 7 | 0.954824445 | 0.953900137 | 6.394781129 | 1.605218871 | 4819.83165 | 3888010 |
| Income Limit Rating Cards Age factor(Gender) factor(Ethnicity) factor(Student) | 8 | 0.954959282 | 0.95391988 | 7.229560218 | 1.770439782 | 4820.635975 | 3886345 |
| Income Limit Rating Cards Age factor(Gender) factor(Ethnicity) factor(Married) factor(Student | 9 | 0.955046842 | 0.953891234 | 8.472883277 | 1.527116723 | 4821.857603 | 3888761 |
| Income Limit Rating Cards Age Education factor(Gender) factor(Ethnicity) factor(Married) fact | 10 | 0.955101563 | 0.95382867 | 10 | 1 | 4823.370391 | 3894037 |

---

The optimal model using forward selection procedure

```
> 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 |

These two results do not match with each other. Which one should we choose?

>> Based on your circumstances. Maybe more information the better.

# Evaluate the reliability of the model chosen

After using model selection by automatic methods or all possible regression methods, we might not have the best fit model yet, as we consider only main effects on independent variables. After eliminating some variables that are not important out of the model, we consider interaction terms and/or high order multiple regression model to improve the model.


# In class Practice Problem 14

Use the CREDIT.CSV data.
Include interaction terms and higher order terms to improve the optimal first-order model we got.

The Credit data set records balance (average credit card debt for a number of individuals) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). In addition to these quantitative variables, we also have four qualitative variables: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian).

Lecture_4.R


# In class Practice Problem 14 Answers

```
> ols_step_both_p(inter_full_model, pent=0.05, prem=0.05, details = FALSE)
```

Stepwise Selection Summary

| Step | Variable | Added/ Removed | R-Square | Adj. R-Square | C(p) | AIC | RMSE |
|---|---|---|---|---|---|---|---|
| 1 | Rating:factor(Student) | addition | 0.808 | 0.807 | 3686.6670 | 5386.1303 | 201.8427 |
| 2 | Income | addition | 0.947 | 0.946 | 738.5400 | 4875.2153 | 106.4426 |
| 3 | factor(Student) | addition | 0.951 | 0.951 | 640.6990 | 4840.3003 | 101.7711 |
| 4 | Age | addition | 0.952 | 0.951 | 631.4390 | 4837.9153 | 101.3432 |
| 5 | Cards | addition | 0.952 | 0.951 | 630.4880 | 4838.7578 | 101.3253 |
| 6 | Age | removal | 0.952 | 0.951 | 640.3350 | 4841.3837 | 101.7835 |
| 7 | Limit:Rating | addition | 0.965 | 0.965 | 350.7520 | 4710.3128 | 86.2957 |
| 8 | factor(Student) | removal | 0.963 | 0.962 | 403.5540 | 4736.9367 | 89.3257 |
| 9 | Income:Rating | addition | 0.977 | 0.976 | 110.0590 | 4552.6157 | 70.8566 |
| 10 | Rating | addition | 0.977 | 0.976 | 112.0590 | 4552.6157 | 70.8566 |
| 11 | Limit | addition | 0.978 | 0.978 | 84.2300 | 4529.9117 | 68.7896 |
| 12 | Income:factor(Student) | addition | 0.980 | 0.979 | 52.7520 | 4502.2385 | 66.3695 |
| 13 | Age:factor(Student) | addition | 0.981 | 0.981 | 24.1550 | 4477.0478 | 64.1558 |
| 14 | Rating:Age | addition | 0.981 | 0.981 | 21.3440 | 4474.2571 | 63.8548 |

Lecture_5.R


# In class Practice Problem 14 Answers

```
> summary(best_inter_model_stepwise)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    factor(Student) + Rating:factor(Student) + Limit:Rating +
    Income:Rating + Income:factor(Student) + Age:factor(Student) +
    Rating:Age, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-207.563  -42.088    7.096   38.556  158.181

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.425e+02  3.154e+01  -7.689 1.24e-13 ***
Income                    -1.891e+00  5.276e-01  -3.584 0.000381 ***
Limit                      1.146e-01  2.163e-02   5.300 1.95e-07 ***
Rating                    -2.636e-01  3.281e-01  -0.803 0.422199
Cards                      1.807e+01  2.802e+00   6.449 3.36e-10 ***
Age                        1.798e-01  4.663e-01   0.386 0.699957
factor(Student)Yes         9.240e+01  5.016e+01   1.842 0.066234 .
Rating:factor(Student)Yes  1.136e+00  1.164e-01   9.757  < 2e-16 ***
Limit:Rating               3.376e-04  1.720e-05  19.633  < 2e-16 ***
Income:Rating             -1.682e-02  1.196e-03 -14.069  < 2e-16 ***
Income:factor(Student)Yes -1.757e+00  4.772e-01  -3.682 0.000264 ***
Age:factor(Student)Yes     5.430e-01  7.400e-01   0.734 0.463498
Rating:Age                -2.682e-03  1.160e-03  -2.311 0.021368 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.66 on 387 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9808
F-statistic:  1702 on 12 and 387 DF,  p-value: < 2.2e-16
```

Lecture_5.R


# In class Practice Problem 14 Answers

```
> best_inter_model_final1 <-lm(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                              +Rating:factor(Student)+Limit:Rating+Income:Rating+Income:factor(Student)
+                              +Rating:Age
+                               ,data=credit)
> summary(best_inter_model_final1)

Call:
lm(formula = Balance ~ Income + Limit + Rating + Cards + Age +
    factor(Student) + Rating:factor(Student) + Limit:Rating +
    Income:Rating + Income:factor(Student) + Rating:Age, data = credit)

Residuals:
     Min       1Q   Median       3Q      Max
-211.690  -42.756    7.535   38.314  159.449

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -2.455e+02  3.126e+01  -7.852 4.03e-14 ***
Income                    -1.931e+00  5.244e-01  -3.683 0.000263 ***
Limit                      1.156e-01  2.158e-02   5.357 1.45e-07 ***
Rating                    -2.666e-01  3.279e-01  -0.813 0.416576
Cards                      1.795e+01  2.795e+00   6.421 3.96e-10 ***
Age                        2.406e-01  4.586e-01   0.525 0.600073
factor(Student)Yes         1.221e+02  2.965e+01   4.118 4.68e-05 ***
Rating:factor(Student)Yes  1.120e+00  1.143e-01   9.798  < 2e-16 ***
Limit:Rating               3.365e-04  1.712e-05  19.658  < 2e-16 ***
Income:Rating             -1.674e-02  1.189e-03 -14.073  < 2e-16 ***
Income:factor(Student)Yes -1.643e+00  4.506e-01  -3.645 0.000303 ***
Rating:Age                -2.739e-03  1.157e-03  -2.367 0.018416 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.62 on 388 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.9809
F-statistic:  1859 on 11 and 388 DF,  p-value: < 2.2e-16
```

Lecture_5.R


# In class Practice Problem 14 Answers

It takes too much time to run ols_step_best_subset and construct the Combos 5 table. Let's quit.

Lecture_5.R


# In class Practice Problem 14 Answers

Add high order terms [Income, Limit, Rating, Cards, Age].

Lecture_5.R


# Coffee break

# Stepwise regression is cumbersome

```
> full_model=lm(Balance ~ Income+Limit+Rating+Cards+Age+Education+factor(Gender)
+               +factor(Ethnicity)+factor(Married)+factor(Student), data=credit)
> best_inter_model_final1_high_order <-lm(Balance ~ Income+Limit+Rating+Cards+Age+factor(Student)
+                                         +Rating:factor(Student)+Limit:Rating+Income:Rating+Income:factor(Student)
+                                         +Rating:Age+I(Income^2)+I(Limit^2)+I(Rating^2)+I(Cards^2)+I(Age^2)
+                                         ,data=credit)
>
> ols_step_both_p(best_inter_model_final1_high_order,pent=0.05, prem=0.05, details = FALSE)
```

Stepwise Selection Summary

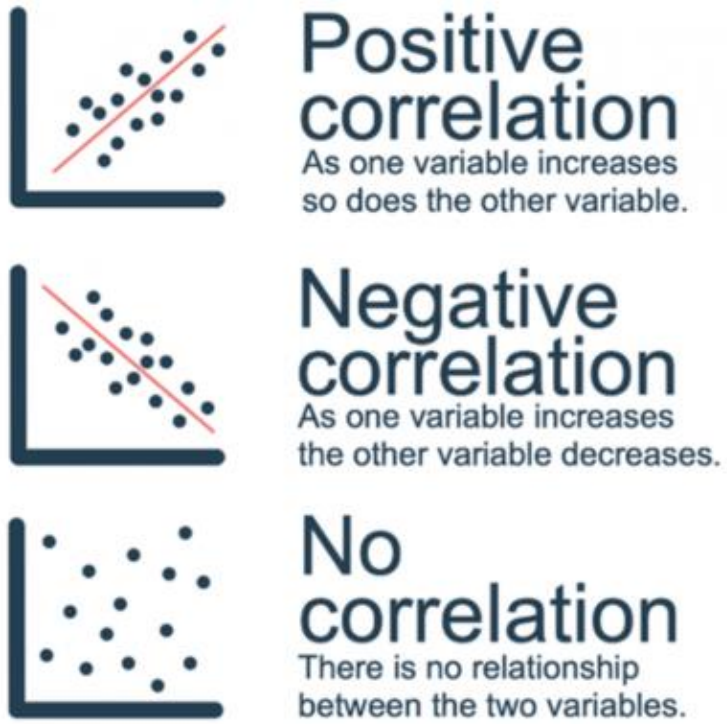
| Step | Variable | Added/ Removed | R-Square | Adj. R-Square | C(p) | AIC | RMSE |
|---|---|---|---|---|---|---|---|
| 1 | Rating:factor(Student) | addition | 0.808 | 0.807 | 4890.2550 | 5386.1303 | 201.8427 |
| 2 | Income | addition | 0.947 | 0.946 | 1072.4180 | 4875.2153 | 106.4426 |
| 3 | factor(Student) | addition | 0.951 | 0.951 | 945.1430 | 4840.3003 | 101.7711 |
| 4 | I(Income^2) | addition | 0.952 | 0.951 | 935.0100 | 4838.6543 | 101.4368 |
| 5 | factor(Student) | removal | 0.947 | 0.947 | 1061.2610 | 4873.6102 | 106.0980 |
| 6 | Age | addition | 0.948 | 0.947 | 1047.8970 | 4871.3589 | 105.6696 |
| 7 | Income | removal | 0.912 | 0.911 | 2040.1500 | 5079.5965 | 137.2558 |
| 8 | I(Age^2) | addition | 0.912 | 0.911 | 2039.7200 | 5081.1968 | 137.3612 |
| 9 | I(Income^2) | removal | 0.816 | 0.814 | 4678.5110 | 5373.4629 | 198.1808 |
| 10 | Cards | addition | 0.818 | 0.816 | 4623.7010 | 5370.9560 | 197.3173 |
| 11 | Rating:factor(Student) | removal | 0.008 | 0.001 | 26946.2300 | 6045.4007 | 459.5913 |
| 12 | Limit | addition | 0.759 | 0.756 | 6262.5020 | 5482.0835 | 227.0012 |
| 13 | Age | removal | 0.758 | 0.756 | 6291.3090 | 5481.9316 | 227.2387 |
| 14 | Income:factor(Student) | addition | 0.926 | 0.925 | 1643.7420 | 5010.0372 | 125.6707 |
| 15 | I(Limit^2) | addition | 0.943 | 0.942 | 1188.0880 | 4910.0645 | 110.7715 |
| 16 | I(Cards^2) | addition | 0.943 | 0.942 | 1190.0860 | 4912.0642 | 110.9127 |
| 17 | I(Limit^2) | removal | 0.926 | 0.925 | 1645.7100 | 5012.0310 | 125.8295 |
| 18 | I(Rating^2) | addition | 0.942 | 0.941 | 1220.5580 | 4919.7239 | 111.9797 |
| 19 | Income:factor(Student) | removal | 0.762 | 0.759 | 6184.9390 | 5479.2707 | 225.9258 |
| 20 | Income:Rating | addition | 0.891 | 0.889 | 2618.6040 | 5168.1446 | 152.9437 |
| 21 | Cards | removal | 0.891 | 0.890 | 2616.7380 | 5166.1625 | 152.7529 |
| 22 | Rating | addition | 0.891 | 0.890 | 2607.4040 | 5166.6507 | 152.6583 |
| 23 | I(Cards^2) | removal | 0.891 | 0.889 | 2625.4030 | 5167.3142 | 152.9729 |

# Correlation

response, coefficients of correlation, predictors

$$g(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, g(y_i) = y_i$$

- Correlations is an easy way to tell which predictors are mostly likely to have high orders
- Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate in relation to each other. A positive correlation (+) indicates the extent to which those variables increase or decrease in parallel; a negative correlation (-) indicates the extent to which one variable increases as the other decreases.
- A correlation coefficient is a statistical measure, of the degree to which changes to the value of one variable predict change to the value of another.
- Note, correlation does not imply causation.



# Correlation

- Pearson correlation: Pearson correlation evaluates the linear relationship between two quantitative (continuous) variables. It's also known as a parametric correlation test because it depends to the distribution of the data. It can be used only when x and y are from normal distribution.
- Spearman (rank) correlation: Spearman correlation evaluates the monotonic relationship, suitable for both quantitative and qualitative variables. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.
- Kendall's Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 to 1, where: 0 is no relationship, 1 is a perfect relationship.
- A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. Examples of monotonic and non-monotonic relationships are presented in the diagram below

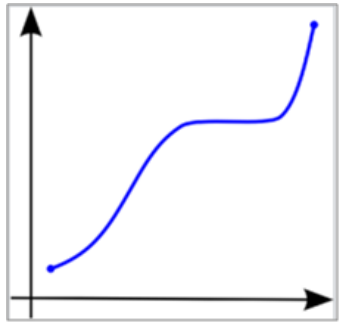


Figure 1 - A monotonically increasing function

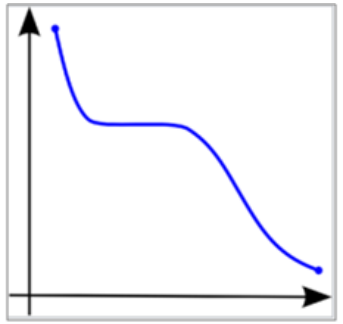


Figure 2 - A monotonically decreasing function

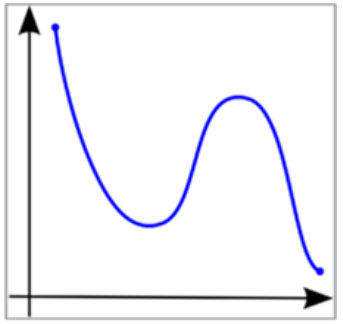


Figure 3 - A function that is not monotonic

# Pearson correlation

- Pearson correlation evaluates the linear relationship between two quantitative (continuous) variables.
- Requirements:
  - Scale of measurement should be interval or ratio
  - Variables should be approximately normally distributed
  - The association should be linear
  - There should be no outliers in the data

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# Spearman (rank) correlation

| | Marks | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 56 | 75 | 45 | 71 | 62 | 64 | 58 | 80 | 76 | 61 |
| Maths | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

No tied observations where d = difference between ranks and $d^2$ = difference squared.

Create a table of ranks, example of no tied observations

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) | d | $d^2$ |
|---|---|---|---|---|---|
| 56 | 66 | 9 | 4 | 5 | 25 |
| 75 | 70 | 3 | 2 | 1 | 1 |
| 45 | 40 | 10 | 10 | 0 | 0 |
| 71 | 60 | 4 | 7 | 3 | 9 |
| 62 | 65 | 6 | 5 | 1 | 1 |
| 64 | 56 | 5 | 9 | 4 | 16 |
| 58 | 59 | 8 | 8 | 0 | 0 |
| 80 | 77 | 1 | 1 | 0 | 0 |
| 76 | 67 | 2 | 3 | 1 | 1 |
| 61 | 63 | 7 | 6 | 1 | 1 |

$$\rho(df) = \rho\ coefficient, P = P\ value$$

$$\rho(8) = 0.67, P = 0.033$$

where $df = N - 2$, where $N$ = number of pairwise cases.

$H_0$: There is no [monotonic] association between the two variables [in the population].

P value of spearman correlation < 0.05 means there is a monotonic association between two variables.



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

# Spearman (rank) correlation

| | Marks | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| English | 56 | 75 | 45 | 71 | 61 | 64 | 58 | 80 | 76 | 61 |
| Maths | 66 | 70 | 40 | 60 | 65 | 56 | 59 | 77 | 67 | 63 |

Create a table of ranks, example of tied observations

| English (mark) | Maths (mark) | Rank (English) | Rank (maths) |
|---|---|---|---|
| 56 | 66 | 9 | 4 |
| 75 | 70 | 3 | 2 |
| 45 | 40 | 10 | 10 |
| 71 | 60 | 4 | 7 |
| **61** | 65 | 6.5 | 5 |
| 64 | 56 | 5 | 9 |
| 58 | 59 | 8 | 8 |
| 80 | 77 | 1 | 1 |
| 76 | 67 | 2 | 3 |
| **61** | 63 | 6.5 | 6 |

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Tied observation, where *i* = paired score.

It is important to realize that statistical significance does not indicate the strength of Spearman's correlation. In fact, the statistical significance testing of the Spearman correlation does not provide you with *any* information about the strength of the relationship.

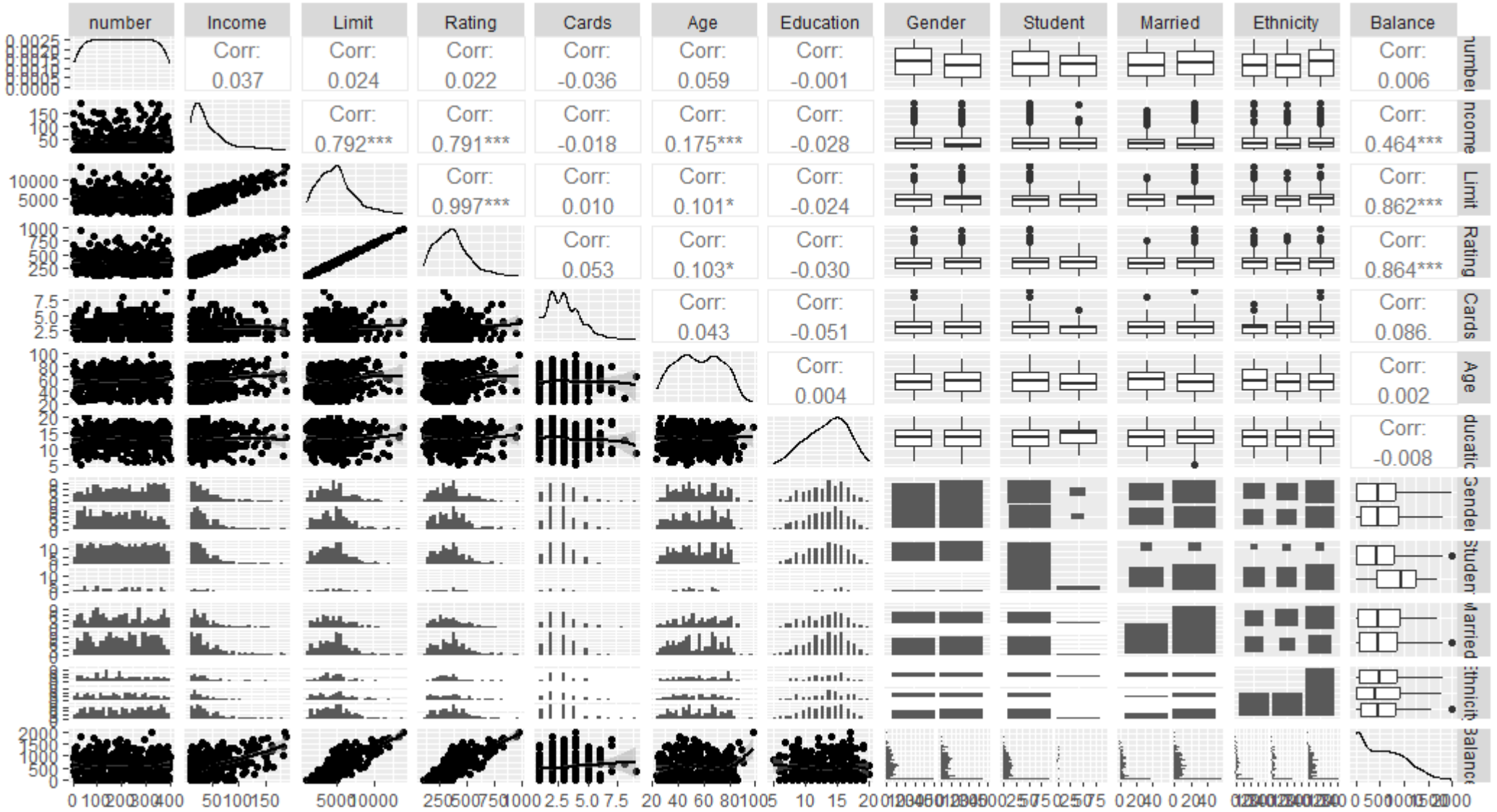
R code: cor(x, y, method="pearson")



https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

# Correlation matrix

Library(GGally)
ggpairs(credit,lower = list(continuous = "smooth_loess",
combo ="facethist", discrete = "facetbar", na = "na"))




# In class Practice Problem 15

Clerical staff work hours. In any production process in which one or more workers are engaged in a variety of tasks, the total time spent in production varies as a function of the size of the work pool and the level of output of the various activities.

For example, in a large metropolitan department store, the number of hours worked (Y) per day by the clerical staff may depend on the following variables:

X1 = Number of pieces of mail processed (open, sort, etc.)
X2 = Number of money orders and gift certificates sold,
X3 = Number of window payments (customer charge accounts) transacted ,
X4 = Number of change order transactions processed ,
X5 = Number of checks cashed ,
X6 =Number of pieces of miscellaneous mail processed on an ‘’as available” basis , and
X7 =Number of bus tickets sold

The data are provided in **CLERICAL.csv** file count for these activities on each of 52 working days. Conduct a Stepwise Regression Procedure and All-Possible-Regressions procedure of the data using R software package.

Lecture_5.R

# Take away messages

- Statistics:
  - Model selection based on p-values of predictor's coefficients OR combos 5 criterion
  - Correlation measures the extent to which two or more variables fluctuate in relation to each other.
    - Pearson correlation
    - Spearman rank correlation
- Code:
  - ols_step_best_subset
  - cor(x,y)
  - ggpairs(data, lower = list(continuous = "smooth_loess", combo ="facethist", discrete = "facetbar", na = "na")))

# Thank you

- Questions OR Comments?

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