Memilih Model Terbaik

Sevi Nurafni
Data Sains, Fakultas Sains dan Teknologi
Universitas Koperasi Indonesia
slideshare.net/sevinurafni

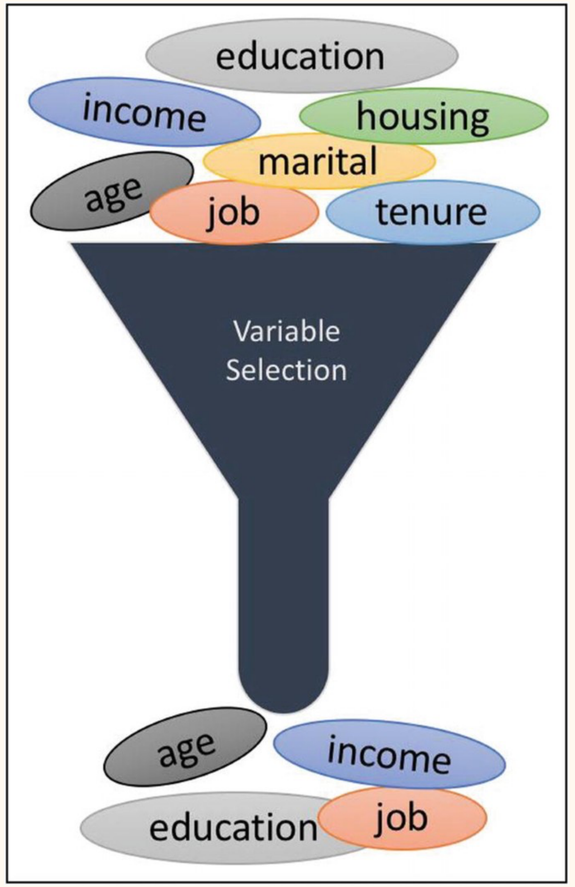
Contents:

- Problem
- All Possible Regression
- Best Subset Regression

Intro

Problem

Masalah utama dalam analisis regresi adalah untuk memutuskan variabel yang dimasukan dalam model regresi tersebut.



https://link.springer.com/chapter/10.1007/978-1-4842-6500-0_4

https://blog.minitab.com/en/how-to-choose-the-best-regression-model

https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html

Choosing a subset variables

1. Untuk membuat model serealistis mungkin, analis dapat menyertakan sebanyak mungkin variabel penjelas.

2. Untuk membuat model sesederhana mungkin, salah satu caranya adalah dengan memasukkan lebih sedikit variabel penjelas.

Why Bother?

1. Kita ingin menjelaskan data dengan cara yang paling sederhana - prediktor yang berlebihan harus dibuang.

2. Prediktor yang tidak perlu akan menambah noise pada estimasi kuantitas lain yang kita minati. Derajat kebebasan akan terbuang sia-sia.

3. Kolinieritas disebabkan oleh terlalu banyak variabel yang mencoba melakukan pekerjaan yang sama.

4. Biaya: jika model akan digunakan untuk prediksi, kita dapat menghemat waktu dan/atau uang dengan tidak mengukur prediktor yang berlebihan.

Methods

- All-Possible
- Best-Subsets Regression

} Setiap subset dari variabel independen diEVALUASI

- Backward Elimination
- Stepwise Regression
- Principle Component Regression

} Satu variabel independen pada suatu waktu ditambahkan atau dihilangkan berdasarkan F-test

- Ridge Regression
- Latent Root Regression
- Stagewise Regression

} Mengatasi multikolinearitas sebagai dasar centering dan scalling

All Possible Regression

Identify all of the possible regression models

Menentukan kombinasi persamaan menggunakan 2^r

r : banyaknya variabel bebas

Misal: terdapat 3 buah variabel bebas, maka regresi yang mungkin adalah

1. $Y = \beta_0$
2. $Y = \beta_0 + \beta_1 X_1$
3. $Y = \beta_0 + \beta_2 X_2$
4. $Y = \beta_0 + \beta_3 X_3$
5. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
6. $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$
7. $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$
8. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

All Possible Regression, R^2

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$$

Y_i nilai aktual dari variabel dependen

$\hat{Y}_i$ nilai prediksi dari model regresi

$\bar{Y}_i$ rata-rata dari nilai aktual Y

All Possible Regression, S^2

$$S^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - p}$$

Y_i nilai aktual dari variabel dependen

$\hat{Y}_i$ nilai prediksi dari model regresi

n jumlah observasi

p parameter model regresi (termasuk intercept)

All Possible Regression, C_p

$$C_p = \frac{\sum (Y_i - \hat{Y}_i)^2}{S^2} - (n - 2p)$$

Y_i nilai aktual dari variabel dependen

$\hat{Y}_i$ nilai prediksi dari model regresi

n jumlah observasi

p parameter model regresi (termasuk intercept)

https://en.wikipedia.org/wiki/Mallows's_Cp

Conclusion

1. All Possible Regression, R^2 —> Persamaan yang dipilih adalah yang nilai R^2 terbesar
2. All Possible Regression, S^2 —> Persamaan yang dipilih adalah yang nilai S^2 terkecil
3. All Possible Regression, C_p —> Persamaan yang dipilih adalah yang nilai C_p terkecil

Best Subset Regression

Identify all of the possible regression models

Menentukan kombinasi persamaan menggunakan 2^r

r : banyaknya variabel bebas

Misal: terdapat 3 buah variabel bebas, maka regresi yang mungkin adalah

1. $Y = \beta_0$
2. $Y = \beta_0 + \beta_1 X_1$
3. $Y = \beta_0 + \beta_2 X_2$
4. $Y = \beta_0 + \beta_3 X_3$
5. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
6. $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$
7. $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3$
8. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Best Subset Regression, R^2

$$R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y}_i)^2}$$

Y_i nilai aktual dari variabel dependen

$\hat{Y}_i$ nilai prediksi dari model regresi

$\bar{Y}_i$ rata-rata dari nilai aktual Y

Best Subset Regression, R_p^2 dengan parameter

$$R_p^2 = 1 - [(1 - R^2)(\frac{n-1}{n-p})]$$

Y_i nilai aktual dari variabel dependen

$\hat{Y}_i$ nilai prediksi dari model regresi

$\bar{Y}_i$ rata-rata dari nilai aktual Y

Referensi

https://search.r-project.org/CRAN/refmans/olsrr/html/ols_step_all_possible.html

https://rpubs.com/Baalgainti/PembentukanModelTerbaik

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/

http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html