UNIVERSITAS SULTAN AGENG TIRTAYASA

UNTIRTA Jawara

Jujur Adil Wibawa Amanah Religius Akuntabel

Analisis Regresi

#12-13 Meeting

Model Selection

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 199005202024061001

Seleksi Model Terbaik

Dalam praktik analisis data nyata, kita sering kali dihadapkan pada dataset yang memiliki banyak sekali kandidat variabel independen (X).

Misalkan kita ingin memprediksi Gaji Seseorang (Y). Kita punya data:

- X1: Usia
- X2: Pendidikan
- X3: Pengalaman Kerja
- X4: IQ
- X5: Tinggi Badan
- X6: Jarak rumah ke kantor
- ... hingga X20 misalnya.

Pertanyaan besarnya: Apakah kita harus memasukkan SEMUA variabel tersebut ke dalam model regresi?
Jawabannya: Hampir pasti TIDAK.
Memasukkan variabel yang tidak relevan (sampah) ke dalam model bukan hanya sia-sia, tetapi juga merusak kualitas model (memperbesar varians estimasi).

Seleksi Model Terbaik

Tujuan utama seleksi model adalah mencari model yang Parsimoni.
Prinsip Parsimoni: Model terbaik adalah model yang paling sederhana (sedikit variabel) namun memiliki kemampuan penjelasan (prediksi) yang paling tinggi.

Kita menginginkan model yang:

1. Goodness of Fit Tinggi: Mampu menjelaskan variasi data dengan baik.
2. Kompleksitas Rendah: Menggunakan sesedikit mungkin variabel.

Seleksi Model Terbaik

Dalam memilih model, kita selalu menyeimbangkan dua risiko ini:

1. Underfitting (Model Terlalu Sederhana):
   - Misal: Hanya pakai 1 variabel padahal yang berpengaruh ada 5.
   - Akibat: Model memiliki Bias Tinggi. Model gagal menangkap pola data yang sebenarnya. Prediksi tidak akurat.

2. Overfitting (Model Terlalu Kompleks):
   - Misal: Memasukkan 20 variabel (termasuk yang tidak penting).
   - Akibat: Model memiliki Varians Tinggi. Model tidak hanya menangkap pola, tapi juga "menghafal" noise (acak) dalam data. Model ini bagus di data latih, tapi hancur saat dipakai memprediksi data baru.

Tujuan Seleksi Model: Menemukan Titik Tengah di antara keduanya.

Seleksi Model Terbaik

Kita sering tergoda untuk memilih model dengan R^2 tertinggi. Ini berbahaya.

- Fakta: Setiap kali kita menambah variabel X baru (walaupun variabel itu acak/sampah), nilai R^2 PASTI NAIK (atau minimal tetap). Tidak akan pernah turun.
- Jika kita hanya mengejar R^2, kita akan terjebak memilih model dengan semua variabel (X1 ... X20), yang kemungkinan besar mengalami Overfitting dan Multikolinearitas.

Oleh karena itu, kita membutuhkan:

1. Kriteria Informasi (Metrics): Ukuran yang mengevaluasi model jika terlalu kompleks (seperti Adj-R^2, AIC, BIC).
2. Algoritma Seleksi: Cara sistematis untuk memilih kombinasi variabel (Forward, Backward, Stepwise).

Ukuran dalam Seleksi Model

Dalam memilih model regresi terbaik, kita sering dihadapkan pada banyak pilihan variabel. Prinsip utamanya adalah menyeimbangkan kebaikan suai (seberapa dekat data dengan garis) dengan kompleksitas model (jumlah variabel).

Kita akan menggunakan tiga ukuran utama:

1. Koefisien Determinasi Terkoreksi (Adjusted R^2)
2. AIC (Akaike Information Criterion)
3. BIC (Bayesian Information Criterion)

Ukuran dalam Seleksi Model

Mahasiswa (i)	Nilai (Y)	Jam (X1)	Hadir (X2)	IQ (X3)
1	60	2	10	105
2	65	3	11	100
3	70	4	12	110
4	75	5	12	105
5	72	4	10	115
6	80	6	13	110
7	85	7	14	112
8	82	6	12	108
9	90	8	15	120
10	95	9	15	118

Disajikan data sampel sebanyak n=10$ mahasiswa. Variabel respon adalah Nilai Ujian (Y), dengan tiga kandidat prediktor:

1. Jam Belajar (X1)
2. Kehadiran (X2)
3. IQ (X3)

Ukuran dalam Seleksi Model

Skenario Perbandingan:

Kita akan membandingkan tiga model dengan tingkat kompleksitas berbeda:

1. Model A (1 Variabel): Y = b0 + b1 X1
2. Model B (2 Variabel): Y = b0 + b1 X1 + b2 X2
3. Model C (3 Variabel): Y = b0 + b1 X1 + b2 X2 + b3 X3

Informasi Hasil Estimasi OLS:

Diketahui Jumlah Kuadrat Total (JKT) = 1.122,5.

Berikut adalah Jumlah Kuadrat Galat (JKG) yang didapat dari perhitungan software R:

- Model A (X1): k=2, JKG = 85,4
- Model B (X1, X2): k=3, JKG = 78,2 (Error turun)
- Model C (X1, X2, X3): k=4, JKG = 75,0 (Error paling kecil)

Catatan: Semakin banyak variabel, JKG pasti semakin kecil. Namun, apakah penurunan itu sebanding dengan biaya penambahan variabel? Mari kita uji.

Ukuran dalam Seleksi Model

Koefisien Determinasi Terkoreksi (R^2_{adj})

$1-(\frac{JKG}{JKT}\times\frac{n-1}{n-k})$. Nilai **makin besar** makin baik.

Model A ($k=2$):

$$1-(\frac{85,4}{1.122,5}\times\frac{9}{8})=1-(0,076\times1,125)=1-0,085=\mathbf{0,915}$$

Model B ($k=3$):

$$1-(\frac{78,2}{1.122,5}\times\frac{9}{7})=1-(0,069\times1,285)=1-0,089=\mathbf{0,911}$$

Model C ($k=4$):

$$1-(\frac{75,0}{1.122,5}\times\frac{9}{6})=1-(0,067\times1,500)=1-0,100=\mathbf{0,900}$$

Model A memiliki R^2_{adj} tertinggi. Menambah variabel justru menurunkan kualitas model.

Akaike Information Criterion (AIC)

$n\ln(\frac{JKG}{n})+2k$. Nilai **makin kecil** makin baik.

Model A ($k=2$):

$$10\ln(8,54)+2(2)=10(2,145)+4=21,45+4=\mathbf{25,45}$$

Model B ($k=3$):

$$10\ln(7,82)+2(3)=10(2,057)+6=20,57+6=\mathbf{26,57}$$

Model C ($k=4$):

$$10\ln(7,50)+2(4)=10(2,015)+8=20,15+8=\mathbf{28,15}$$

Model A memiliki AIC terendah.

Ukuran dalam Seleksi Model

Bayesian Information Criterion (BIC)

$n\ln(\frac{JKG}{n}) + k\ln(n)$. Nilai **makin kecil** makin baik.

Model A ($k = 2$):

$21,45 + 2\ln(10) = 21,45 + 2(2,30) = 21,45 + 4,6 = \mathbf{26,05}$

Model B ($k = 3$):

$20,57 + 3\ln(10) = 20,57 + 3(2,30) = 20,57 + 6,9 = \mathbf{27,47}$

Model C ($k = 4$):

$20,15 + 4\ln(10) = 20,15 + 4(2,30) = 20,15 + 9,2 = \mathbf{29,35}$

Model A memiliki BIC terendah.

Ukuran dalam Seleksi Model

Kriteria	Aturan	Model A (X1)	Model B (X1,X2)	Model C (X1,X2,X3)	Pemenang
JKG	Min	854	782	**750**	Model C
R2 adj	Max	**915**	911	900	**Model A**
AIC	Min	**2,545**	2,657	2,815	**Model A**
BIC	Min	**2,605**	2,747	2,935	**Model A**

Seleksi Model : Metode Forward

Metode Seleksi Maju adalah algoritma pencarian model terbaik yang dimulai dari model yang paling sederhana (Model Kosong), kemudian menambahkan variabel satu per satu.
Variabel yang ditambahkan adalah variabel yang memberikan perbaikan model paling signifikan (penurunan AIC terbesar). Algoritma ini akan berhenti ketika tidak ada lagi variabel tersisa yang mampu memperbaiki model.

Seleksi Model : Metode Forward

Algoritma :

1. Langkah Awal (Step 0):
2. Mulai dengan "Model Kosong" (Null Model) yang hanya berisi intersep (Y = b0). Hitung AIC-nya.
3. Langkah Pertama (Step 1):
   - Lakukan regresi Y terhadap masing-masing variabel X secara terpisah (Y dengan X1, Y dengan X2, Y dengan X3).
   - Pilih variabel yang menghasilkan AIC terendah.
   - Jika model ini lebih baik dari Model Kosong (AIC turun), masukkan variabel tersebut ke dalam model.
4. Langkah Kedua (Step 2):
   - Dengan satu variabel yang sudah terpilih (misal X1), coba kombinasikan dengan variabel tersisa satu per satu.
   - Pilih kombinasi yang menghasilkan AIC terendah.
   - Keputusan:
     - Jika AIC turun dibandingkan langkah sebelumnya → Masukkan variabel tersebut.
     - Jika AIC naik (memburuk) → Tolak variabel tersebut dan Berhenti.

Seleksi Model : Metode Forward

Kita menggunakan data yang sama ($n = 10$, Y =Nilai Ujian). Kandidat: X_1 (Jam Belajar), X_2 (Kehadiran), X_3 (IQ). Kriteria: **AIC** (Makin kecil makin baik).

Langkah 0: Model Kosong (Intersep Saja)

Model: $Y = \beta_0$ Pada model ini, prediksi Y hanyalah rata-rata data. JKG sama dengan JKT ($1.122, 5$).

- **AIC Model Kosong: 49,20** (Status: Model Terbaik Sementara).

Seleksi Model : Metode Forward

Langkah 1: Mencari Variabel Pertama Masuk

Kita menguji tiga model regresi linier sederhana secara terpisah untuk mencari "kandidat terbaik".

1. **Model 1a ($Y \sim X_1$):**
   - Diketahui JKG turun drastis menjadi 85,4.
   - **AIC = 25,45**
2. **Model 1b ($Y \sim X_2$):**
   - Korelasi lemah, JKG masih tinggi (misal 500,0).
   - **AIC = 43,12**
3. **Model 1c ($Y \sim X_3$):**
   - Korelasi sangat lemah, JKG tinggi (misal 600,0).
   - **AIC = 44,94**

Keputusan Langkah 1:

- Bandingkan ketiga kandidat: 25,45 vs 43,12 vs 44,94.
- Pemenangnya adalah X_1 **(Jam Belajar)** dengan AIC **25,45**.
- Bandingkan dengan langkah sebelumnya (49,20): **AIC Turun Drastis (Membaik)**.
- **Tindakan:** Masukkan X_1 ke dalam model. (Model Aktif: $Y \sim X_1$).

Seleksi Model : Metode Forward

UNIVERSITAS SULTAN AGENG TIRTAYASA

UNTIRTA Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

Langkah 2: Mencari Variabel Kedua Masuk

Sekarang model kita memuat X_1. Variabel tersisa adalah X_2 dan X_3. Kita coba tambahkan satu per satu.

1. **Model 2a ($Y \sim X_1 + X_2$):**
   - Kita uji jika X_2 masuk. JKG turun sedikit menjadi 78,2.
   - **AIC = 26,57**
2. **Model 2b ($Y \sim X_1 + X_3$):**
   - Kita uji jika X_3 masuk. JKG turun sedikit menjadi 80,0.
   - **AIC = 26,79**

Keputusan Langkah 2:

- Di antara kandidat, Model 2a ($X_1 + X_2$) memiliki AIC terendah yaitu **26,57**.
- Sekarang, bandingkan nilai ini dengan Model Terbaik dari Langkah 1 (AIC = **25,45**).
- **Perbandingan:** AIC Baru (26,57) **LEBIH BESAR** dari AIC Lama (25,45).
- Artinya, model justru **memburuk** jika kita menambah variabel lagi. Penalti kompleksitasnya lebih besar daripada perbaikan error-nya.
- **Tindakan: STOP (BERHENTI).** Jangan masukkan X_2 maupun X_3.
- Kembali ke model terbaik terakhir.

Seleksi Model : Metode Forward

Langkah	Model Terbaik di Langkah Ini	AIC	Keputusan
0	Y = Intersep	49.20	Mulai
1	Tambah X1 (Jam Belajar)	**25.45**	**Lanjut** (AIC Turun)
2	Coba Tambah X2	26.57	**STOP** (AIC Naik)

Seleksi Model : Metode Backward

Kebalikan dari metode Forward, metode ini memiliki pendekatan "Buang yang Kurang Penting".
Jika Forward bersifat hati-hati (mulai dari nol), Backward bersifat agresif (memasukkan semuanya dulu, baru seleksi).

UNIVERSITAS SULTAN AGENG TIRTAYASA

Seleksi Model : Metode Backward

UNTIRTA Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

Algoritma :

1. Langkah Awal (Step 0):
2. Mulai dengan "Model Penuh" (Full Model). Masukkan SEMUA kandidat variabel prediktor ke dalam model. Hitung AIC-nya.
3. Langkah Eliminasi (Step 1, 2, dst):
   - Coba keluarkan satu per satu variabel dari model saat ini.
   - Hitung AIC untuk setiap model yang dikurangi tersebut.
   - Bandingkan: Cari variabel mana yang jika dibuang, menghasilkan penurunan AIC paling besar (membuat model jadi lebih baik).
   - Keputusan:
     - Jika membuang variabel membuat AIC turun → Buang variabel tersebut permanen. Lanjut ke langkah berikutnya.
     - Jika membuang variabel manapun justru membuat AIC naik (model jadi lebih buruk) → Berhenti. Pertahankan model terakhir.

Seleksi Model : Metode Backward

Kita gunakan data yang sama ($n = 10$, Y=Nilai Ujian). Kandidat: X_1 (Jam Belajar), X_2 (Kehadiran), X_3 (IQ).

Langkah 0: Model Penuh (Start)

Kita mulai dengan memasukkan semua variabel.

- **Model:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- Dari perhitungan sebelumnya (di materi Ukuran Seleksi), kita tahu:
  - **AIC Full Model: 28,15**

Seleksi Model : Metode Backward

Langkah 1: Uji Eliminasi Pertama

Kita coba buang satu variabel dari Model Penuh, lalu cek apakah AIC-nya membaik (turun bawah 28,15).

1. **Coba Buang X_1 (Sisa X_2, X_3):**
   - Model ini buruk karena X_1 adalah prediktor terkuat. JKG akan melonjak tinggi.
   - Misal AIC jadi **45,00** (Naik drastis). $\rightarrow$ *Jangan buang* X_1.
2. **Coba Buang X_2 (Sisa X_1, X_3):**
   - Model tinggal X_1 dan X_3.
   - Misal AIC turun menjadi **26,79**. (Lebih baik dari 28,15).
3. **Coba Buang X_3 (Sisa X_1, X_2):**
   - Model tinggal X_1 dan X_2.
   - Dari data sebelumnya, kita tahu AIC model ini adalah **26,57**.

Keputusan Langkah 1:

- Membuang X_3 memberikan AIC terendah (**26,57**).
- Bandingkan dengan Langkah 0 (28,15): **AIC Turun (Membaik)**.
- **Tindakan: Buang X_3 (IQ)** dari model secara permanen.

UNIVERSITAS SULTAN AGENG TIRTAYASA

Seleksi Model : Metode Backward

UNTIRTA Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

Langkah 2: Uji Eliminasi Kedua

Model kita sekarang: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (AIC saat ini = 26,57). Variabel tersisa di dalam model: X_1, X_2. Kita coba buang lagi satu per satu.

1. **Coba Buang X_1 (Sisa hanya X_2):**
   - Model hanya Kehadiran. Ini prediktor lemah.
   - Misal AIC jadi **43,12**. (Memburuk drastis).
2. **Coba Buang X_2 (Sisa hanya X_1):**
   - Model hanya Jam Belajar.
   - Dari data sebelumnya, kita tahu AIC model ini adalah **25,45**.

Keputusan Langkah 2:

- Membuang X_2 memberikan AIC terendah (**25,45**).
- Bandingkan dengan AIC saat ini (26,57): **AIC Turun (Membaik)**.
- **Tindakan: Buang X_2 (Kehadiran)** dari model.

Seleksi Model : Metode Backward

Langkah 3: Uji Eliminasi Ketiga

Model kita sekarang: $Y = \beta_0 + \beta_1 X_1$ (AIC saat ini = 25,45). Variabel tersisa: X_1.

1. **Coba Buang X_1 (Sisa Model Kosong / Intersep saja):**
   - Jika X_1 dibuang, kita kembali ke model rata-rata.
   - Dari data sebelumnya, AIC Model Kosong = **49,20**.

Keputusan Langkah 3:

- Jika kita membuang X_1, AIC melonjak dari **25,45** ke **49,20** (Sangat buruk).
- **Tindakan: Jangan buang X_1. Berhenti di sini.**

Seleksi Model : Metode Backward

Langkah	Model Aktif	AIC	Keputusan
0	X1 + X2 + X3 (Full)	28.15	Mulai
1	Buang X3 (IQ)	26.57	Lanjut (AIC Turun)
2	Buang X2 (Kehadiran)	**25.45**	Lanjut (AIC Turun)
3	Coba Buang X1	49.20	**STOP** (AIC Naik)

Seleksi Model : Metode Stepwise

Metode Stepwise adalah algoritma hibrida (campuran) yang menggabungkan kelebihan dari Forward Selection dan Backward Elimination.

- Kelemahan Forward: Sekali variabel masuk, ia tidak bisa keluar lagi (meskipun di langkah selanjutnya variabel tersebut menjadi tidak penting karena adanya variabel baru).
- Kelemahan Backward: Sekali variabel dibuang, ia tidak bisa masuk lagi.

Metode Stepwise memperbaiki ini dengan cara: Memasukkan variabel seperti Forward, tetapi di setiap langkah, ia melakukan "evaluasi ulang" (seperti Backward) untuk membuang variabel yang sudah tidak lagi signifikan.

Seleksi Model : Metode Stepwise

Algoritma :

1. Langkah Awal (Step 0): Mulai dengan Model Kosong (hanya Intersep).
2. Langkah Maju (Forward Step): Sama seperti Forward Selection, cari variabel kandidat di luar model yang memiliki AIC terendah (atau p-value terkecil). Masukkan ke dalam model.
3. Langkah Evaluasi Mundur (Backward Step) - INI KUNCINYA: Setelah variabel baru masuk, cek kembali statistik (AIC/p-value) dari semua variabel yang sudah ada di dalam model.
    - Apakah ada variabel lama yang mendadak menjadi tidak signifikan (AIC-nya memburuk) setelah variabel baru masuk?
    - Jika YA: Buang variabel lama tersebut.
    - Jika TIDAK: Lanjut ke langkah berikutnya.
4. Ulangi: Ulangi proses maju dan mundur ini sampai tidak ada lagi variabel yang bisa masuk dan tidak ada lagi variabel yang perlu dibuang.

Seleksi Model : Metode Stepwise

Skenario Data:

- **Respon (Y):** Volume Penjualan.
- **Prediktor (X):** Iklan TV (X_1), Iklan Radio (X_2), Iklan Digital (X_3).
- **Karakteristik:** X_1 dan X_2 memiliki multikolinearitas tinggi. Secara individual X_1 terbaik, namun kombinasi X_2 dan X_3 adalah yang paling efisien.

1. Inisialisasi

Model dimulai dari **Model Kosong** (Hanya Intersep β_0).

- **Status Awal:** AIC = **120,0**

Seleksi Model : Metode Stepwise

UNIVERSITAS SULTAN AGENG TIRTAYASA

UNTIRTA Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

2. Iterasi Pertama

A. Fase Forward (Memilih Variabel Pertama) Algoritma menguji pengaruh masing-masing variabel secara individual:

- Model $\{X_1\} \rightarrow$ AIC = **100,0** (Terendah)
- Model $\{X_2\} \rightarrow$ AIC = 102,0
- Model $\{X_3\} \rightarrow$ AIC = 110,0

Keputusan: Masukkan X_1. **Model Aktif:** $\{X_1\}$ (AIC = 100,0)

B. Fase Backward (Evaluasi Variabel)

- Cek hapus X_1: Kembali ke Model Kosong (AIC 120,0).
- **Hasil:** $120,0 > 100,0$ (AIC Memburuk).
- **Keputusan:** Pertahankan X_1.

Seleksi Model : Metode Stepwise

3. Iterasi Kedua

A. Fase Forward (Memilih Variabel Kedua) Algoritma mencari variabel terbaik untuk ditambahkan ke model $\{X_1\}$:

- Model $\{X_1, X_2\} \rightarrow$ AIC = 98,0
- Model $\{X_1, X_3\} \rightarrow$ AIC = **90,0** (Terendah)

Keputusan: Masukkan X_3. **Model Aktif:** $\{X_1, X_3\}$ (AIC = 90,0)

B. Fase Backward (Evaluasi Variabel Lama) Algoritma mengecek apakah variabel lama (X_1) masih relevan setelah X_3 masuk.

- Cek hapus X_1 (Sisa $\{X_3\}$): AIC model $\{X_3\}$ adalah 110,0.
- **Perbandingan:** $110,0 > 90,0$ (AIC Memburuk).
- **Keputusan:** Pertahankan X_1. Kedua variabel tetap di dalam model.

Seleksi Model : Metode Stepwise

4. Iterasi Ketiga

A. Fase Forward (Memilih Variabel Ketiga) Algoritma mencoba menambahkan variabel terakhir (X_2) ke model $\{X_1, X_3\}$.

- Model Saat Ini $\{X_1, X_3\}$: AIC = 90,0.
- Model Baru $\{X_1, X_3, X_2\}$: AIC = **88,0**.

Analisis: Penambahan X_2 menurunkan AIC. **Keputusan:** Masukkan X_2. **Model Aktif (Sementara):** $\{X_1, X_3, X_2\}$ (AIC = 88,0)

Seleksi Model : Metode Stepwise

B. Fase Backward (Evaluasi Menyeluruh) Model sekarang berisi 3 variabel. Algoritma melakukan simulasi penghapusan satu per satu untuk mencari redundansi.

- **Opsi 1: Hapus** X_3 (Sisa $\{X_1, X_2\}$)
  - AIC menjadi **98,0**.
  - Bandingkan dengan Model Aktif (88,0) $\rightarrow$ **Memburuk**.
  - *Kesimpulan:* X_3 Penting.
- **Opsi 2: Hapus** X_2 (Sisa $\{X_1, X_3\}$)
  - AIC menjadi **90,0**.
  - Bandingkan dengan Model Aktif (88,0) $\rightarrow$ **Memburuk**.
  - *Kesimpulan:* X_2 Penting.
- **Opsi 3: Hapus** X_1 (Sisa $\{X_2, X_3\}$)
  - Karena informasi X_1 sudah terwakili oleh X_2, penghapusan X_1 mengurangi kompleksitas (k) tanpa menambah error secara signifikan.
  - Hasil perhitungan AIC model $\{X_2, X_3\}$ adalah **85,0**.
  - Bandingkan dengan Model Aktif (88,0) $\rightarrow$ **MEMBAIK (AIC Turun)**.

Keputusan Akhir Fase Backward: Karena model tanpa X_1 memiliki AIC lebih rendah (**85,0**) dibandingkan model dengan X_1 (**88,0**), maka X_1 **dikeluarkan** dari model.

Algoritma berlanjut dengan model aktif $\{X_2, X_3\}$. Pada iterasi berikutnya, algoritma akan mencoba memasukkan kembali X_1, namun karena akan menghasilkan AIC 88,0 (lebih tinggi dari 85,0), maka proses berhenti.

Model Final: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

UNIVERSITAS SULTAN AGENG TIRTAYASA

UNTIRTA Jawara
Jujur Adil Wibawa Amanah Religius Akuntabel

SEE YOU NEXT WEEK !

Ferdian Bangkit Wijaya, S.Stat., M.Si
NIP. 199005202024061001
ferdian.bangkit@untirta.ac.id