Evaluasi Model Regresi Lanjutan

Pada Hari 17 dan 19, kita telah membahas tentang analisis regresi linear, asumsi-asumsi yang mendasarinya, dan cara melakukan validasi silang untuk mengevaluasi kinerja model. Hari ini, kita akan membahas beberapa teknik evaluasi model regresi lanjutan.

# Analisis Residu Lanjutan

Selain plot scatter antara prediksi dan residu, serta plot residual vs. fitted values yang sudah kita bahas sebelumnya, ada beberapa plot diagnostik lain yang dapat membantu Anda menganalisis residu lebih dalam:

* **Partai Regression Plot:** Menunjukkan hubungan antara variabel independen dan dependen setelah efek dari variabel independen lainnya dihilangkan. Berguna untuk mengidentifikasi hubungan non-linear atau pengaruh outlier pada masing-masing variabel independen.
* **CCPR Plot (Component-Component plus Residual Plot):** Mirip dengan partikel regression plot, tetapi juga menunjukkan residu pada sumbu y. Berguna untuk melihat apakah ada pola non-linear yang tidak tertangkap oleh model.
* **Scale-Location Plot:** Menunjukkan akar kuadrat dari residu standar terhadap nilai prediksi. Berguna untuk memeriksa asumsi homoskedastisitas (varians error konstan).

# Uji Statistik

Selain analisis visual, Anda juga dapat menggunakan uji statistik untuk memeriksa asumsi-asumsi regresi linear:

* **Uji Breusch-Pagan:** Uji statistik untuk memeriksa homoskendastistik.
* **Uji Shapiro-Wilk:** Uji statistik untuk memeriksa normalitas residu.

# Pemilihan Model dengan Informasi Kriteria

Selain menggunakan cross-validation dan rata-rata MAE, Anda juga dapat menggunakan kriteria informasi seperti AIC (Akaike Information Criterion) atau BIC (Bayesian Information Criterion) untuk memilih model terbaik. Kriteria ini memperhitungkan kebaikan fit model (seberapa baik model sesuai dengan data) dan kompleksitas model (jumlah parameter). Model dengan nilai AIC atau BIC yang lebih rendah biasanya lebih disukai.

# 

# 

# **Tugas HAri 20:**

Lanjutkan dengan dataset “Advertising” dan model regresi linear yang telah Anda buat.

**Tugas**

1. **Analisa Residu Lanjutan:**

* Buat partial regression plot dan CCPR plot untuk masing-masing variabel independen.
* Buat scale-location plot.
* Analisis plot-plot tersebut. Apakah ada indikasi pelanggaran asumsi regresi linear?

1. **Uji Statistik (Opsional):**

* Lakukan uji Breusch-Pagan dan uji Shapiro-Wilk pada residu model.
* Tampilkan hasil uji dan interpretasikannya.

**Contoh kode (Statsmodels)**

import statsmodels.api as sm

# ... (kode untuk membaca, mengolah data, membagi data, dan membuat model)

# Fit model dengan statsmodels

X\_train\_const = sm.add\_constant(X\_train) # Menambahkan intercept

model\_sm = sm.OLS(y\_train, X\_train\_const).fit()

# Partial Regression Plot

fig = sm.graphics.plot\_partregress\_grid(model\_sm)

fig.tight\_layout(pad=1.0)

plt.show()

# CCPR Plot

fig = sm.graphics.plot\_ccpr\_grid(model\_sm)

fig.tight\_layout(pad=1.0)

plt.show()

# Scale-Location Plot

fig = sm.graphics.plot\_regress\_exog(model\_sm, exog\_idx=1) # Ganti exog\_idx sesuai variabel

plt.show()

# Uji Breusch-Pagan (opsional)

from statsmodels.stats.diagnostic import het\_breuschpagan

\_, pvalue, \_\_, fvalue = het\_breuschpagan(model\_sm.resid, model\_sm.model.exog)

print('p-value uji Breusch-Pagan:', pvalue)

# Uji Shapiro-Wilk (opsional)

from scipy.stats import shapiro

\_, pvalue = shapiro(model\_sm.resid)

print('p-value uji Shapiro-Wilk:', pvalue)

**Catatan:**

* Tugas nomor 2 (uji statistik) bersifat opsional karena membutuhkan pemaham yang lebih mendalam tentang statistik. Boleh di lewati saja.

**Selamat Mengerjakan Tugas! 🙂**

**Tugas:**

1. **Analisis Residu Lanjutan:**

**# Fit model dengan statsmodels**

**X\_train\_const = sm.add\_constant(X\_train) # Menambahkan intercept**

**model\_sm = sm.OLS(y\_train, X\_train\_const).fit()**

**# Partial Regression Plot**

**fig = sm.graphics.plot\_partregress\_grid(model\_sm)**

**fig.tight\_layout(pad=1.0)**

**plt.show()**

**# CCPR Plot**

**fig = sm.graphics.plot\_ccpr\_grid(model\_sm)**

**fig.tight\_layout(pad=1.0)**

**plt.show()**

**# Scale-Location Plot**

**fig = sm.graphics.plot\_regress\_exog(model\_sm, exog\_idx=1) # Ganti exog\_idx sesuai variabel**

**plt.show()**

1. **Uji Statistik (Opsional):**

**# Uji Breusch-Pagan (opsional)**

**from statsmodels.stats.diagnostic import het\_breuschpagan**

**\_, pvalue, \_\_, fvalue = het\_breuschpagan(model\_sm.resid, model\_sm.model.exog)**

**print('p-value uji Breusch-Pagan:', pvalue)**

**# Uji Shapiro-Wilk (opsional)**

**from scipy.stats import shapiro**

**\_, pvalue = shapiro(model\_sm.resid)**

**print('p-value uji Shapiro-Wilk:', pvalue)**

**Interpretasikan hasilnya:**

**1. Analisis Residu Lanjutan**

* **Partial Regression Plot:**
  + **Plot ini menunjukkan hubungan antara setiap variabel prediktor (TV, Radio, Social Media, dan Influencer\_Nano, Influencer\_Mega) dengan variabel target (Sales) setelah mengontrol efek dari variabel prediktor lainnya.**
  + **Jika titik-titik data tersebar secara acak di sekitar garis regresi, maka asumsi linearitas terpenuhi untuk variabel tersebut.**
  + **Dari plot yang dihasilkan, terlihat bahwa hubungan antara masing-masing variabel prediktor dengan Sales cenderung linear, menunjukkan bahwa asumsi linearitas terpenuhi dengan cukup baik.**
* **CCPR Plot:**
  + **Plot ini mirip dengan partial regression plot, tetapi juga menunjukkan residu pada sumbu y.**
  + **Jika titik-titik data tersebar secara acak di sekitar garis horizontal nol, maka asumsi linearitas terpenuhi.**
  + **Dari plot yang dihasilkan, terlihat bahwa sebagian besar titik data tersebar secara acak, tetapi ada beberapa titik yang mungkin outlier.**
* **Scale-Location Plot:**
  + **Plot ini menunjukkan akar kuadrat dari residu standar terhadap nilai prediksi.**
  + **Jika titik-titik data tersebar secara acak dan tidak membentuk pola tertentu (seperti kerucut melebar atau menyempit), maka asumsi homoskedastisitas (varians error konstan) terpenuhi.**
  + **Dari plot yang dihasilkan, terlihat bahwa titik-titik data cenderung menyebar secara acak, menunjukkan bahwa asumsi homoskedastisitas mungkin terpenuhi.**

**2. Uji Statistik**

* **Uji Breusch-Pagan:**
  + **Uji ini digunakan untuk menguji asumsi homoskedastisitas.**
  + **Hasil p-value yang besar (biasanya lebih besar dari 0.05) menunjukkan bahwa kita tidak dapat menolak hipotesis nol tentang homoskedastisitas (artinya, kita tidak memiliki bukti yang cukup untuk mengatakan bahwa varians error tidak konstan).**
  + **Pada hasil Anda, p-value uji Breusch-Pagan sangat kecil, yang menunjukkan adanya indikasi heteroskedastisitas (varians error tidak konstan).**
* **Uji Shapiro-Wilk:**
  + **Uji ini digunakan untuk menguji asumsi normalitas residu.**
  + **Hasil p-value yang besar (biasanya lebih besar dari 0.05) menunjukkan bahwa kita tidak dapat menolak hipotesis nol tentang normalitas residu (artinya, kita tidak memiliki bukti yang cukup untuk mengatakan bahwa residu tidak terdistribusi normal).**
  + **Pada hasil Anda, p-value uji Shapiro-Wilk sangat kecil, yang mengindikasikan bahwa residu mungkin tidak terdistribusi normal.**

**Kesimpulan:**

* **Analisis visual (plot) menunjukkan bahwa asumsi linearitas cukup terpenuhi, tetapi ada sedikit indikasi adanya outlier.**
* **Uji statistik menunjukkan adanya indikasi pelanggaran asumsi homoskedastisitas dan normalitas residu.**

**Rekomendasi:**

1. **Investigasi Outlier: Coba identifikasi dan analisis lebih lanjut outlier yang terdeteksi pada plot CCPR. Jika memungkinkan, perbaiki atau hapus outlier tersebut.**
2. **Transformasi Data: Karena ada indikasi heteroskedastisitas, Anda bisa mencoba melakukan transformasi data pada variabel dependen (Sales) atau variabel independen. Misalnya, Anda bisa mencoba menggunakan transformasi logaritmik atau akar kuadrat.**
3. **Pertimbangkan Model Lain: Jika asumsi homoskedastisitas dan normalitas tidak terpenuhi dengan baik setelah transformasi data, Anda bisa mencoba menggunakan model regresi yang lebih robust terhadap pelanggaran asumsi ini, seperti regresi quantile atau robust regression.**

**Catatan:**

**Ingatlah bahwa analisis diagnostik regresi dan pemilihan model adalah proses iteratif. Anda mungkin perlu mencoba beberapa pendekatan berbeda untuk menemukan model yang paling sesuai dengan data Anda.**

**Jangan ragu untuk bertanya jika Anda memiliki pertanyaan lain.**