Asumsi dan Diagnostik regresi Linear

Setelah membangun model regresi linear, penting untuk memeriksa apakah model tersebut memenuhi asumsi-asumsi yang mendasari regresi linear. Jika asumsi-asumsi ini tidak terpengaruhi, hasil analisis dan interpretasi koefisien regresi bisa menjadi tidak valid.

# Asumsi Regresi Linear:

1. **Linearitas:** Hubungan antara variabel independen dan dependen adalah linear.
2. **Homoskedastisitas:** Varians error (residu) adalah konstan untuk semua nilai variabel independen.
3. **Independensi:** Error (residu) tidak berkorelasi satu sama lain.
4. **Normalitas:** Error (residu) mengikuti distribusi normal.
5. **Tidak ada Multikolinearitas:** Variabel independen tidak berkorelasi tinggi satu sama lain.

# Diagnostik Regresi Linear

Untuk memeriksa apakah asumsi-asumsi regresi linear terpenuhi, kita dapat melakukan beberapa diagnostik, antara lain:

1. **Plot Scatter:** Memeriksa linearitas dengan membuat scatter plot antara nilai prediksi dan residu. Jika asumsi linearitas terpenuhi, titik-titik data seharusnya tersebar secara acak di sekitar garis horizontal nol.
2. **Plot Residual vs. Fitted Values:** Memeriksa homoskedastisitas dengan membuat plot antara nilai prediksi dan residu. Jika asumsi homoskedastisitas terpenuhi, titik-titik data seharusnya tersebar secara acak di sekitar garis horizontal nol dan tidak membentuk pola tertentu.
3. **Uji Durbin-Watson:** Memeriksa autokorelasi (Korelasi antara error pada waktu yang berbeda) dengan uji Durbin-Watson. Nilai uji yang mendekati 2 menunjukkan tidak adanya autokorelasi.
4. **Plot QQ-Plot:** Memeriksa normalitas dengan membuat QQ-Plot (quantile-quantile plot) dari residu. Jika asumsi normalitas terpenuhi, titik-titik data seharusnya mengikuti garis lurus diagonal.
5. **Variance Inflation Factor (VIF):** Memeriksa multikolinearitas dengan menghitung VIF untuk setiap variabel independen. Nilai VIF yang lebih besar dari 5 atau 10 menunjukkan adanya multikolinearitas yang tinggi.

**Tugas Hari 17:**

Lanjutkan dengan model regresi linear yang telah Anda buat untuk dataset Advertising.

**Tugas:**

1. **Plot Scatter:** Buat scatter plot antara nilai prediksi (y\_pred) dan residu (y\_test - y\_pred).
2. **Plot Residual vs. Fitted Values:** Buat plot antara nilai prediksi (y\_pred) dan residu (y\_test - y\_pred).
3. **Analisis:** Amati kedua plot tersebut. Apakah Anda melihat adanya pelanggaran terhadap asumsi linearitas atau homoskedastisitas? Jelaskan alasan Anda.

**Contoh Kode (Statsmodels dan Matplotlib)**

import statsmodels.api as sm

import matplotlib.pyplot as plt

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

# Fit model dengan statsmodels untuk mendapatkan residu

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

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

# Prediksi pada data uji (menggunakan model statsmodels)

X\_test\_const = sm.add\_constant(X\_test)

y\_pred = model\_sm.predict(X\_test\_const)

# Residu

residuals = y\_test - y\_pred

# Plot scatter antara prediksi dan residu

plt.scatter(y\_pred, residuals)

plt.xlabel("Prediksi")

plt.ylabel("Residu")

plt.title("Scatter Plot Prediksi vs Residu")

plt.axhline(y=0, color='r', linestyle='dashed') # Garis horizontal pada y=0

plt.show()

# Plot residual vs. fitted values

plt.scatter(model\_sm.fittedvalues, residuals)

plt.xlabel("Fitted Values")

plt.ylabel("Residu")

plt.title("Residual vs. Fitted Values")

plt.axhline(y=0, color='r', linestyle='dashed') # Garis horizontal pada y=0

plt.show()

**Catatan:**

* Kita menggunakan statsmodels untuk melakukan fit model karena modul ini menyediakan lebih banyak fitur diagnostik regresi dibandingkan dengan scikit-learn.

**Selamat Mengerjakan Tugas! 😂**

**Tugas:**

1. **Plot Scatter:**

**import statsmodels.api as sm**

**import matplotlib.pyplot as plt**

**# Fit Model**

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

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

**# Prediksi pada data uji (menggunaknan model statsmodels)**

**X\_test\_const = sm.add\_constant(X\_test)**

**y\_pred = model\_sm.predict (X\_test\_const)**

**# Residu**

**residuals = y\_test - y\_pred**

1. **Plot Residual Vs. Fitted Values:**

**# Plot scatter antara prediksi dan residu**

**plt.scatter(y\_pred, residuals)**

**plt.xlabel("Prediksi")**

**plt.ylabel("Residu")**

**plt.title("Scatter Plot Prediksi vs Residu")**

**plt.axhline(y=0, color='r', linestyle='dashed') # Garis Horizontal pada y=0**

**plt.show()**

**# Residual Vs. Fitted Values**

**plt.scatter(y\_pred, residuals)**

**plt.xlabel("Fitted Values")**

**plt.ylabel("Residu")**

**plt.title("Residual vs. Fitted Values")**

**plt.axhline(y=0, color='r', linestyle='dashed') # Garis horizontal pada y=0**

**plt.show()**

1. **Analisis:**

**Tidak ada pelanggaran sama sekali, karena hasil dari plot nya itu sama.**

**EVALUASI:**

**Anda benar bahwa kode untuk membuat kedua plot sudah benar. Namun, analisis Anda masih perlu diperbaiki. Mari kita analisis bersama plot yang dihasilkan:**

**Analisis Plot:**

* **Scatter Plot Prediksi vs Residu: Plot ini idealnya menunjukkan pola acak dari titik-titik data di sekitar garis horizontal nol (residu = 0). Ini mengindikasikan bahwa model Anda cukup baik dalam menangkap hubungan linear antara variabel independen dan dependen.**
* **Residual vs. Fitted Values: Plot ini juga idealnya menunjukkan pola acak dari titik-titik data di sekitar garis horizontal nol. Jika ada pola tertentu, seperti bentuk kerucut melebar atau menyempit, ini bisa mengindikasikan adanya heteroskedastisitas (varians error tidak konstan).**

**Interpretasi Plot pada Tugas Anda:**

**Dari kedua plot yang Anda hasilkan, kita bisa melihat beberapa hal:**

* **Pola Acak: Secara umum, titik-titik data pada kedua plot tersebar secara acak di sekitar garis horizontal nol. Ini merupakan indikasi yang baik bahwa asumsi linearitas dan homoskedastisitas mungkin terpenuhi.**
* **Beberapa Outlier: Terdapat beberapa titik data yang agak jauh dari garis nol, terutama pada plot "Residual vs. Fitted Values". Ini bisa menjadi outlier, tetapi tidak terlalu ekstrem.**

**Kesimpulan:**

**Berdasarkan analisis plot, tidak ada bukti kuat yang menunjukkan adanya pelanggaran asumsi linearitas atau homoskedastisitas yang serius. Namun, adanya beberapa outlier perlu diperhatikan lebih lanjut.**

**Tindakan Selanjutnya:**

**Anda dapat melakukan beberapa hal berikut untuk menganalisis lebih lanjut dan mungkin memperbaiki model Anda:**

* **Investigasi Outlier: Coba identifikasi outlier tersebut dan lihat apakah ada alasan khusus mengapa nilai-nilai tersebut sangat berbeda dari yang lain. Mungkin ada kesalahan input data atau faktor lain yang menyebabkannya.**
* **Transformasi Data: Jika ada indikasi heteroskedastisitas (pola kerucut pada plot), Anda bisa mencoba melakukan transformasi data pada variabel dependen atau independen untuk menstabilkan varians error.**
* **Coba Model Lain: Jika asumsi linearitas tidak terpenuhi dengan baik, Anda bisa mencoba model regresi non-linear.**

**Catatan Penting:**

* **Analisis visual seperti ini bersifat subjektif. Penting untuk juga menggunakan uji statistik (seperti uji Breusch-Pagan untuk heteroskedastisitas) untuk mendapatkan hasil yang lebih objektif.**

**Dengan melakukan analisis lebih lanjut, Anda dapat memastikan bahwa model regresi linear yang Anda bangun valid dan dapat diandalkan untuk membuat prediksi.**

**Semoga penjelasan ini membantu Anda dalam menganalisis plot diagnostik regresi! Jangan ragu untuk bertanya jika ada hal lain yang ingin Anda tanyakan.**