Eksplorasi dan Visualisasi Data

Pertemuan 11:

Influential Observation

Anomali

- Deteksi anomali, yaitu memahami apakah suatu pengamatan "tidak biasa (*unusual*)".
- Sebuah observasi disebut tidak biasa karena:
  1. memiliki karakteristik atau perilaku yang tidak biasa
  2. mungkin tidak cocok dengan model
  3. mungkin sangat berpengaruh dalam melatih model
- Dalam regresi linier, karakteristik terakhir adalah hasil sampingan dari dua yang pertama.
- Menjadi tidak biasa belum tentu buruk, bahkan sering kali membawa lebih banyak informasi.

Alasan:
- dibangkitkan oleh proses pembangkitan-data yang berbeda
- kesalahan pengukuran atau penipuan (*fraud*), dll

https://towardsdatascience.com/outliers-leverage-residuals-and-influential-observations-df3065a0388e

Leverage

- Metrik pertama yang akan kita gunakan untuk mengevaluasi pengamatan "tidak biasa" adalah leverage.
- Tujuan dari leverage adalah untuk menangkap seberapa banyak satu titik berbeda sehubungan dengan titik data lainnya. Titik-titik data ini sering disebut *outlier* dan terdapat jumlah algoritma dan aturan praktis yang hampir tak terbatas untuk menandainya.
- Salah satu interpretasi dari leverage adalah sebagai ukuran jarak, di mana pengamatan individu dibandingkan dengan rata-rata semua pengamatan.
- Interpretasi lain dari leverage adalah sebagai pengaruh hasil pengamatan ke-i, y_i, pada nilai $\hat{y}$.

https://towardsdatascience.com/outliers-leverage-residuals-and-influential-observations-df3065a0388e

Residual

- Residual regresi mengukur perilaku yang tidak biasa.
- Residu regresi adalah perbedaan antara nilai hasil yang diprediksi dan nilai hasil yang diamati.
- Dalam arti tertentu, mereka menangkap apa yang tidak dapat dijelaskan oleh model, yaitu semakin tinggi residual dari satu pengamatan, semakin tidak biasa dalam arti bahwa model tidak dapat menjelaskannya.

Influence

- Ide umumnya adalah untuk mendefinisikan observasi sebagai berpengaruh jika menghapusnya secara signifikan mengubah model yang diestimasi.
- Ada hubungan yang erat antara leverage dan residual e_i: pengaruh meningkat pada keduanya. Pengamatan dengan leverage tinggi adalah pengamatan yang bersifat outlier dan memiliki residual yang tinggi. Tak satu pun dari dua kondisi saja yang cukup untuk observasi memiliki pengaruh pada model.
- Tidak satu pun dari dua kondisi saja yang cukup untuk sebuah observasi menjadi berpengaruh dan mendistorsi model (membuuat kemiringan model ke arah dirinya sendiri).

Influential Observation

- Influential Observation adalah pengamatan dalam kumpulan data yang ketika dihapus, secara dramatis mengubah estimasi koefisien model regresi.
- Cara paling umum untuk mengukur influential observation adalah dengan menggunakan jarak Cook (*Cook's distance*), yang mengkuantifikasi berapa banyak semua nilai yang dipasang dalam model regresi berubah ketika pengamatan ke-i dihapus,
- a rule of thumb: pengamatan apa pun dengan jarak Cook lebih besar dari 1 dianggap sebagai pengamatan dengan pengaruh tinggi. Aturan ini harus diterapkan dengan bijaksana dan tidak sembarangan.

https://www.statology.org/influential-observation/

Cook's Distance

- Jarak Cook, sering dilambangkan Di, digunakan dalam analisis regresi untuk mengidentifikasi influential observation(s) yang dapat berpengaruh negatif terhadap model regresi. Jarak Cook adalah jarak antara koefisien yang dihitung dengan dan tanpa pengamatan ke-i.
- Rumus jarak Cook adalah:

$$D_i = \frac{e_i^2}{ps^2}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{(\boldsymbol{b}-\boldsymbol{b}_{(i)})'\mathbf{X}'\mathbf{X}(\boldsymbol{b}-\boldsymbol{b}_{(i)})}{ps^2}$$

Keterangan:

e_i	i^{th} residual
h_i	i^{th} diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
p	number of model parameters, including the constant
s^2	mean square error
b	coefficient vector
$\boldsymbol{b}_{(i)}$	coefficient vector calculated after deleting the i^{th} observation
X	design matrix

https://www.statology.org/how-to-identify-influential-data-points-using-cooks-distance/
https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-regression-model/methods-and-formulas/diagnostic-measures/#cook-s-distance

Leverage (H)

- Leverage suatu observasi didasarkan pada seberapa besar perbedaan nilai observasi pada variabel prediktor dengan mean dari variabel prediktor. Semakin besar pengaruh suatu pengamatan, semakin besar potensinya untuk menjadi suatu pengamatan yang berpengaruh.
- Misalnya, pengamatan dengan nilai yang sama dengan mean pada variabel prediktor tidak memiliki pengaruh terhadap kemiringan garis regresi terlepas dari nilainya pada variabel kriteria. Di sisi lain, pengamatan yang ekstrim pada variabel prediktor berpotensi sangat mempengaruhi kemiringan.
- Rumus:

$$H = X(X'X)^{-1}X'$$

dengan

X	design matrix
h_i	i^{th} diagonal element of the hat matrix
p	number of terms in the model, including the constant
n	number of observations

https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-regression-model/methods-and-formulas/diagnostic-measures/#cook-s-distance

DFITS

- DFITS adalah ukuran yang menggabungkan nilai leverage dan studentized residual (t residual yang dihapus) menjadi satu ukuran keseluruhan tentang seberapa tidak biasa suatu pengamatan.
- DFITS mengukur pengaruh setiap pengamatan pada nilai-nilai yang dipasang dalam model regresi dan ANOVA.
- Pengamatan dengan nilai DFITS besar mungkin merupakan outlier.
- DFITS mewakili secara kasar jumlah deviasi standar yang nilai *fit*-nya berubah ketika setiap pengamatan dihapus dari kumpulan data dan modelnya diperbaiki.

$$DFITS = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_i}} = e_i \sqrt{\frac{n-p-1}{SSE(1-h_i) - e_i^2}} \left[\sqrt{\frac{h_i}{1-h_i}} \right]$$

Keterangan:

e_i	i^{th} residual
h_i	i^{th} diagonal element of $\mathbf{X(X'X)^{-1}X'}$
X	design matrix
$\hat{y}_i$	i^{th} fitted response
$\hat{y}_{i(i)}$	fitted value calculated without the i^{th} observation
$MSE_{(i)}$	mean square error calculated without the i^{th} observation
n	number of observations
p	number of model parameters

https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-regression-model/methods-and-formulas/diagnostic-measures/#cook-s-distance

Deteksi Influential Observation

Misalkan kita memiliki dataset berikut dengan 14 nilai:

x	y	COOK	DFIT	HI
1	23	0,0138	0,1599	0,1663
2	24	0,0056	0,1016	0,1425
3	23	0,0005	0,0292	0,1221
4	19	0,0022	-0,0641	0,1052
5	34	0,0016	0,0544	0,0916
7	35	0,0002	-0,0170	0,0749
3	36	0,0196	0,1921	0,1221
2	36	0,0384	0,2705	0,1425
12	34	0,0319	-0,2483	0,0934
11	32	0,0229	-0,2096	0,0828
15	38	0,1026	-0,4573	0,1457
14	41	0,0502	-0,3128	0,1248
17	42	0,2022	-0,6553	0,1978
22	180	3,6933	15,0420	0,3882

Cook's distance > 1

Dengan sotware **Minitab**, klik Stat -> Regression -> Regression -> Fit Regression Model
Responses : y, Continuous predictors: x, Storage: centang Leverages, Cook's distance, DFITS

Regression Analysis: y versus x

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	9556,0	9555,98	10,04	0,008
x	1	9556,0	9555,98	10,04	0,008
Error	12	11423,2	951,94		
Lack-of-Fit	10	11266,7	1126,67	14,40	0,067
Pure Error	2	156,5	78,25		
Total	13	20979,2			

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8,5	13,6	0,62	0,544	
x	4,05	1,28	3,17	0,008	1,00

Regression Equation

y = 8,5 + 4,05 x

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
30,8535	45,55%	41,01%	0,00%

Fits and Diagnostics for Unusual Observations

Obs	y	Fit	Resid	Std Resid	
14	180,0	97,7	82,3	3,41	R

R Large residual

https://www.statology.org/influential-observation/

Deteksi Influential Observation

Misalkan kita menghapus observasi ke-14, output analisis regresinya adalah

Regression Analysis: y_new versus x_new

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	319,1	319,08	10,27	0,008
x_new	1	319,1	319,08	10,27	0,008
Error	11	341,8	31,08		
Lack-of-Fit	9	185,3	20,59	0,26	0,936
Pure Error	2	156,5	78,25		
Total	12	660,9			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
5,57467	48,28%	43,58%	31,69%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	25,34	2,61	9,71	0,000	
x_new	0,913	0,285	3,20	0,008	1,00

Regression Equation

y_new = 25,34 + 0,913 x_new

Tidak ada unusual obs.

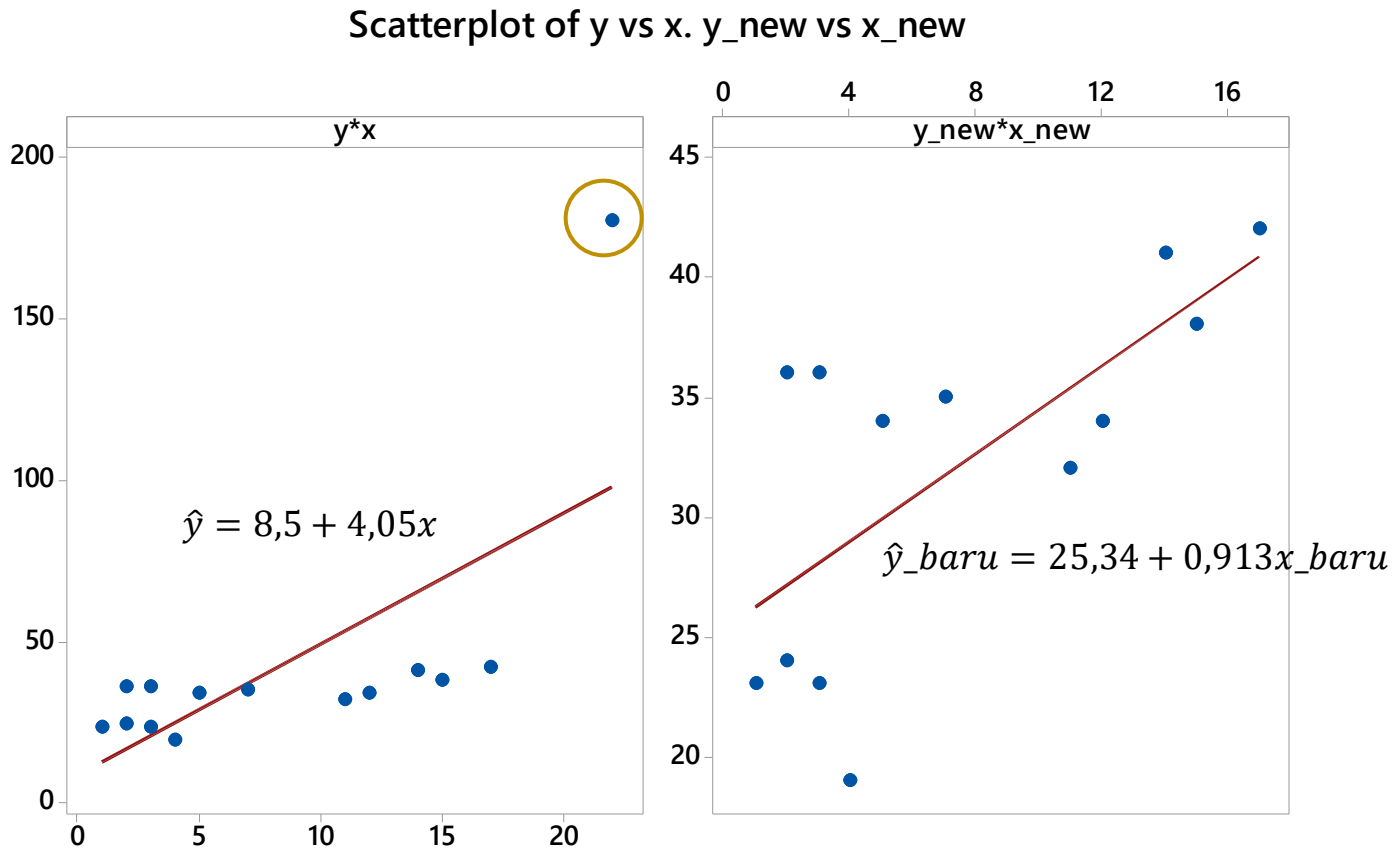
Koefisien regresi untuk intersep dan x **keduanya berubah secara dramatis**, Ini memberi tahu kita bahwa **menghapus influential observation** dari kumpulan data sepenuhnya (asal) mengubah model regresi.

x_new	y_new	COOK	DFIT	HI
1	23	0.0467	-0.2971	0.1833
2	24	0.0342	-0.2538	0.1526
3	23	0.0691	-0.3709	0.1271
4	19	0.2149	-0.7619	0.1068
5	34	0.0301	0.2405	0.0918
7	35	0.0157	0.1717	0.0773
3	36	0.1686	0.6230	0.1271
2	36	0.2671	0.8153	0.1526
12	34	0.0148	-0.1658	0.1325
11	32	0.0258	-0.2207	0.1110
15	38	0.0065	-0.1090	0.2283
14	41	0.0391	0.2709	0.1912
17	42	0.0145	0.1629	0.3183

Cook's distance < 1

https://www.statology.org/influential-observation/

Deteksi Influential Observation



dengan obs. 14

tanpa obs. 14

Perhatikan seberapa besar satu *influential observation* mengubah garis regresi. Dengan menghapus pengamatan ini, ditemukan garis regresi yang lebih cocok dengan data.

https://www.statology.org/influential-observation/

Menangani Influential Observation(s)

- Jarak Cook digunakan sebagai cara untuk mengidentifikasi pengamatan yang berpotensi berpengaruh. Namun, hanya karena pengamatan berpengaruh tidak berarti pengamatan itu harus dihapus dari kumpulan data.
- Pertama, Anda harus memverifikasi bahwa pengamatan tersebut bukan hasil dari kesalahan entri data atau kejadian aneh lainnya. Jika ternyata menjadi nilai yang sah, Anda kemudian dapat memutuskan untuk menanganinya dengan salah satu cara berikut:
  - Hapus dari kumpulan data.
  - Biarkan dalam kumpulan data.
  - Ganti dengan nilai alternatif seperti mean atau median.

https://www.statology.org/influential-observation/

Eksplorasi dan Visualisasi Data

Pertemuan 11:

Influential Observation