رگرسیون خطی چندگانه

به‌همراه مثال‌های کاربردی، توضیحات آماری و کد پایتون

نگارنده: آرمان موجودی

برای دانلود و یا اجرای آنلاین کد پایتون این آموزش به آدرس گیتهاب آموزشها مراجعه کنید.

https://github.com/ML-OilGas/ML_Algorithms/Multiple_Regression

ارتباط با نگارنده و ارائه‌ی نظرات و پیشنهادات:

https://www.linkedin.com/in/arman-mojoodi

مقدمه

در رگرسیون خطی ساده فرض می‌کنیم که تنها یک متغیر بر خروجی مؤثر است. سؤال اینجاست که اگر بخواهیم اثر چند متغیر را به‌صورت همزمان بر خروجی بررسی کنیم، چه باید بکنیم؟

به‌عنوان مثال، اگر اثر اضافه‌وزن بر افزایش خطر سکته‌ی قلبی بررسی شده باشد، چگونه می‌توان اثر فشار خون بالا را نیز در نظر گرفت؟

یک راه ساده این است که دو مدل رگرسیون خطی ساده در نظر بگیریم؛ یکی برای اثر اضافه‌وزن و دیگری برای اثر فشار خون بالا. اما این رویکرد دو اشکال اساسی دارد. اول اینکه اثر همزمان اضافه‌وزن و فشار خون بالا بر خطر سکته مشخص نمی‌شود. دوم اینکه در رگرسیون خطی ساده، در واقع فرض می‌کنیم هیچ عامل دیگری به‌جز متغیر x بر خروجی y تأثیر ندارد. بنابراین زمانی که چند عامل بر خروجی مؤثر هستند و برای هرکدام به‌صورت جداگانه یک رگرسیون ساده اجرا می‌کنیم، اثر سایر متغیرها نادیده گرفته می‌شود؛ موضوعی که می‌تواند منجر به برآوردهای نادرست و گمراه کننده از میزان اثر متغیرها شود.

راه‌حل مناسب، تعمیم مدل رگرسیون خطی ساده به مدلی است که تمام متغیرها را به‌طور همزمان در نظر بگیرد تا بتوان اثر مشترک آن‌ها را بر خروجی بررسی کرد. در این چارچوب، اثر هر متغیر در حضور و با کنترل سایر متغیرها تفسیر می‌شود.

برای سادگی، فرض می‌کنیم دو متغیر x_1 و x_2 داریم. در این حالت مدل به‌صورت زیر نوشته می‌شود که به آن **رگرسیون خطی چندگانه** گفته می‌شود:

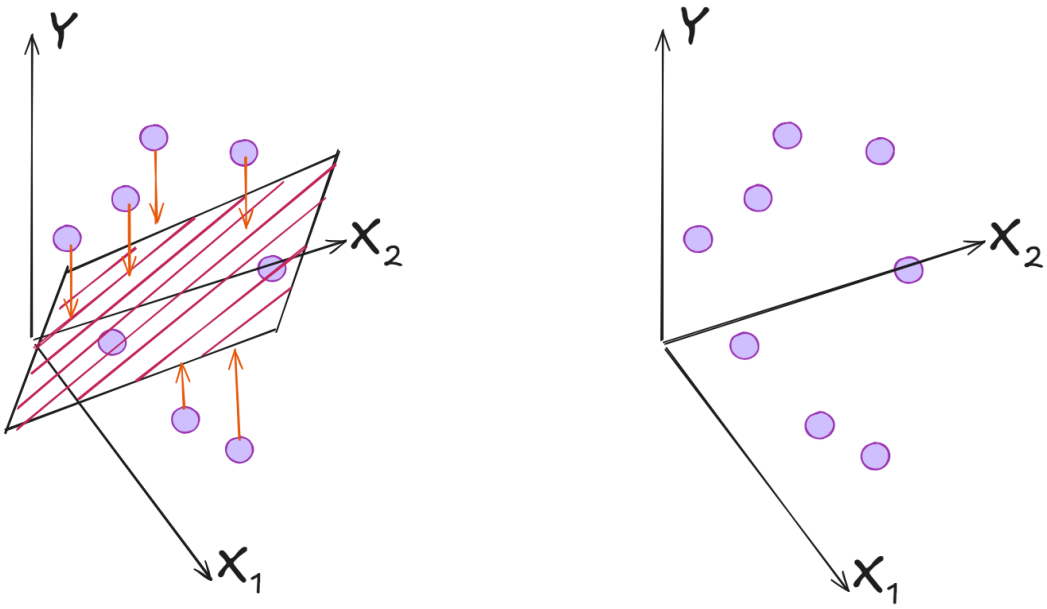
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad (۱)$$

در این رابطه، ϵ بیانگر نویز، خطای اندازه‌گیری و اثر عوامل مشاهده‌نشده است که حتی با بهترین مدل خطی نیز به‌طور کامل قابل حذف نیست.

ضرایب β_i (برای $i = 1,2$) به این معنا هستند که با افزایش یک واحد در x_i، در حالی که **سایر متغیرها ثابت نگه داشته شده‌اند**، مقدار میانگین خروجی y چه تغییری می‌کند. این نکته نشان‌دهنده‌ی تفاوت معنای ضرایب در رگرسیون خطی ساده و چندگانه است. در حالت ساده، β_1 میزان تغییرات خروجی به ازای افزایش یک واحد x_1 را

نشان می‌دهد، به شرط آنکه عامل مؤثر دیگری وجود نداشته باشد؛ در حالی که در رگرسیون چندگانه، اثر هر متغیر به‌صورت مستقل و با کنترل سایر متغیرها تفسیر می‌شود.

مشابه رگرسیون خطی ساده، ضرایب این مدل نیز با کمینه کردن مجموع مربعات باقیمانده‌ها (RSS) به‌دست می‌آیند، با این تفاوت که شکل روابط پیچیده‌تر می‌شود. در حالت دو متغیره، به‌جای یک خط رگرسیون، با یک صفحه‌ی رگرسیون مواجه هستیم که در آن مقدار RSS حداقل می‌شود (شکل ۱).



شکل ۱ : در حالت دو متغیره، خط رگرسیون تبدیل به صفحه‌ی رگرسیون می‌شود (شکل سمت چپ)

در بخش بعد، برای یک دیتاست نفتی واقعی و با استفاده از پایتون، مدل رگرسیون خطی چندگانه محاسبه و نتایج آن تحلیل می‌شود.

اجرای عملی

به‌منظور نمایش نحوه‌ی اجرای رگرسیون خطی چندگانه، در این بخش از داده‌های واقعی مربوط به میدان **Volve** در دریای شمال استفاده می‌کنیم. این داده‌ها به‌صورت منبع باز در دسترس عموم قرار دارند.

داده‌های خام ارائه‌شده در منبع اصلی نیازمند انجام مراحل **پاکسازی** و **تحلیل اکتشافی داده**[1] هستند. از آنجا که این مراحل خارج از اهداف این فایل آموزشی است، در اینجا مورد بررسی قرار نمی‌گیرند. داده‌های مورد استفاده در این بخش پس از انجام پاکسازی و تحلیل اکتشافی انتخاب شده‌اند. توضیحات مربوط به این مراحل به‌صورت جداگانه در بخش ویکی گیت‌هاب ارائه خواهد شد.

فایل داده‌ی مورد استفاده شامل متغیرهای زیر است:

- BVW (V/V): حجم کل آب موجود در سنگ (نسبت حجمی)
- CARB_FLAG (Unitless): شاخص وجود سازند کربناته
- COAL_FLAG (Unitless): شاخص وجود لایه‌های زغالی
- KLOGH (mD): تراوایی افقی سنگ (**متغیر پاسخ**)
- PHIF (V/V): تخلخل شکستگی
- RHOFL (g/cm³): چگالی سیال
- RHOMA (g/cm³): چگالی ماتریس سنگ
- RW (ohm·m): مقاومت الکتریکی آب سازند
- SAND_FLAG (Unitless): شاخص وجود ماسه‌سنگ
- SW (V/V): اشباع آب
- TEMP (°C): دمای مخزن
- VSH (V/V): حجم شیل
- LITHOTYPE: نوع لیتولوژی سنگ

[1] Exploratory Data Analysis (Eda)

هدف، بررسی رابطه‌ی بین تراوایی افقی سنگ (KLOGH) و مجموعه‌ای از پارامترهای پتروفیزیکی است. در این راستا، مدل رگرسیون خطی چندگانه به‌صورت زیر در نظر گرفته می‌شود:

$$KLOGH = \beta_0 + \beta_1 BVW + \beta_2 CARB_{FLAG} + \beta_3 COAL_{FLAG} + \beta_4 PHIF + \beta_5 RHOFL + \beta_6 RHOMA + \beta_7 RW + \beta_8 SAND_{FLAG} + \beta_9 SW + \beta_{10} TEMP + \beta_{11} VSH + \epsilon \quad (۲)$$

تعدادی از ردیف‌های این دیتاست در شکل ۲ نشان داده شده است.

	Unnamed: 0	DEPTH (M)	BVW (V/V)	CARB_FLAG (UNITLESS)	COAL_FLAG (UNITLESS)	PHIF (V/V)	RHOFL (G/CM3)	RHOMA (G/CM3)	RW (OHMM)	SAND_FLAG (UNITLESS)	SW (V/V)	TEMP (DEGC)	VSH (V/V)	KLOGH (MD)
0	0	3666.5916	0.111705	0.0	0.0	0.199843	0.8	2.66	0.021643	0.0	0.5590	112.7249	0.598800	0.0003
1	1	3666.7440	0.112236	0.0	0.0	0.218743	0.8	2.66	0.021642	0.0	0.5131	112.7285	0.599555	0.0000
2	2	3666.8964	0.114825	0.0	0.0	0.242380	0.8	2.66	0.021642	0.0	0.4737	112.7321	0.600310	0.0000
3	3	3667.0488	0.114419	0.0	0.0	0.226408	0.8	2.66	0.021641	0.0	0.5054	112.7358	0.601065	0.0000
4	4	3667.2012	0.115867	0.0	0.0	0.216169	0.8	2.66	0.021640	0.0	0.5360	112.7394	0.601820	0.0000

شکل ۲ : بخشی از دیتاست مورد بررسی

برای به دست آوردن مدل رگرسیون از کتابخانه‌ی statsmodels و تابع OLS استفاده می‌کنیم که علاوه بر ضرایب، خروجی‌های آماری موردنیاز برای ارزیابی مدل را نیز محاسبه می‌کند. در ادامه، ضرایب مدل برآورد شده و نتایج آماری حاصل از برازش مدل مورد بررسی قرار می‌گیرند.

OLS Regression Results

Dep. Variable:	KLOGH (MD)	R-squared:	0.237
Model:	OLS	Adj. R-squared:	0.235
Method:	Least Squares	F-statistic:	135.8
Date:	Tue, 23 Dec 2025	Prob (F-statistic):	2.07e-272
Time:	12:00:32	Log-Likelihood:	-37568.
No. Observations:	4818	AIC:	7.516e+04
Df Residuals:	4806	BIC:	7.524e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.529e+05	4.22e+04	3.621	0.000	7.01e+04	2.36e+05
BVW (V/V)	-9694.8618	642.750	-15.083	0.000	-1.1e+04	-8434.778
CARB_FLAG (UNITLESS)	85.9356	73.525	1.169	0.243	-58.207	230.078
COAL_FLAG (UNITLESS)	-2.1787	139.581	-0.016	0.988	-275.822	271.465
PHIF (V/V)	9683.5161	585.995	16.525	0.000	8534.697	1.08e+04
RHOFL (G/CM3)	1361.5109	285.535	4.768	0.000	801.731	1921.290
RHOMA (G/CM3)	-2736.5151	1143.659	-2.393	0.017	-4978.609	-494.421
RW (OHMM)	-3.867e+06	1.1e+06	-3.531	0.000	-6.01e+06	-1.72e+06
SAND_FLAG (UNITLESS)	248.8619	36.961	6.733	0.000	176.402	321.322
SW (V/V)	1387.9500	128.846	10.772	0.000	1135.353	1640.547
TEMP (DEGC)	-572.3689	161.885	-3.536	0.000	-889.737	-255.000
VSH (V/V)	-163.8915	20.146	-8.135	0.000	-203.388	-124.395

شکل ۳ : خروجی آماری ضرایب مدل رگرسیون چندگانه

ارتباط با یادگیری ماشین:

در این فایل آموزشی، هدف اصلی بررسی **معناداری آماری متغیرها و تفسیر ضرایب مدل** است؛ بنابراین محاسبه‌ی ضرایب، آزمون‌های آماری (t, F, p-value) و شاخص‌هایی مانند RSE بر اساس **کل دیتاست** انجام می‌شود تا برآوردها از کمترین واریانس ممکن برخوردار باشند.

در مقابل، در یادگیری ماشین معمولاً داده‌ها به مجموعه‌های آموزش و آزمون تقسیم می‌شوند تا عملکرد پیش‌بینی مدل روی داده‌های دیده‌نشده ارزیابی شود. این رویکرد بیشتر بر دقت پیش‌بینی تمرکز دارد تا تفسیر آماری ضرایب. این موضوع به‌صورت جداگانه در کدهای موجود در مخزن گیت‌هاب مورد بررسی قرار گرفته است.

بحث در نتایج

۱- آیا حداقل یکی از متغیرها بر خروجی موثر است؟

مشابه آنچه در رگرسیون خطی ساده انجام دادیم، در اینجا نیز از آزمون فرضیه‌ی صفر استفاده می‌کنیم که به‌صورت زیر تعریف می‌شود:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \qquad (۳)$$

رد شدن این فرضیه نشان می‌دهد که حداقل یکی از ضرایب غیرصفر است و در نتیجه حداقل یکی از متغیرها بر خروجی مؤثر است.

در رگرسیون خطی چندگانه، علاوه بر محاسبه‌ی آماره‌ی (t) و مقدار (p) برای هر ضریب به‌صورت جداگانه، یک آماره‌ی کلی به نام **آماره‌ی (F)** نیز محاسبه می‌شود که به‌صورت زیر تعریف می‌گردد:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \qquad (۴)$$

که در آن (p) تعداد متغیرهای توضیحی و (n) تعداد مشاهدات است. برای این آماره نیز مقدار (p)-value محاسبه می‌شود. به‌طور شهودی، هرچه مقدار (F) بزرگ‌تر باشد، شواهد قوی‌تری علیه فرضیه‌ی صفر وجود دارد. به‌صورت دقیق‌تر، فرضیه‌ی صفر زمانی رد می‌شود که مقدار F از مقدار بحرانی متناظر با توزیع F فراتر رود.

سؤال مهمی که در اینجا مطرح می‌شود این است که با وجود محاسبه‌ی آماره‌ی t و مقدار p برای هر ضریب به‌صورت جداگانه، چه نیازی به استفاده از آماره‌ی F وجود دارد؟ آیا بررسی جداگانه‌ی ضرایب کافی نیست؟

پاسخ این سؤال در این نکته نهفته است که وقتی تعداد پارامترها زیاد باشد (برای مثال ۱۰۰ پارامتر)، به‌صورت تصادفی انتظار داریم حدود ۵٪ از ضرایب دارای مقدار (p) کمتر از ۰.۰۵ شوند، حتی اگر در واقع هیچ‌کدام از متغیرها اثری بر خروجی نداشته باشند. بنابراین معنادار شدن چند ضریب به‌تنهایی لزوماً نشان‌دهنده‌ی اعتبار کلی مدل نیست. آماره‌ی (F) با در نظر گرفتن همزمان تعداد پارامترها و میزان کاهش RSS، از بروز چنین خطایی جلوگیری می‌کند.

اگر مقدار (F) نزدیک به یک باشد، نقش تعداد داده‌ها (n) اهمیت بیشتری پیدا می‌کند. برای مقادیر بزرگ n، حتی مقادیر نزدیک به یک نیز می‌توانند منجر به رد شدن فرضیه‌ی صفر شوند، در حالی که برای nهای کوچک باید با احتیاط بیشتری نتیجه‌گیری کرد.

در برخی موارد، لازم است اثر یک **زیرمجموعه از متغیرها** بررسی شود. اگر بخواهیم بررسی کنیم که (q) پارامتر مشخص بر خروجی مؤثر هستند یا خیر، فرضیه‌ی صفر به‌صورت زیر تعریف می‌شود:

$$H_0: \beta_{\{p-q+1\}} = \beta_{\{p-q+2\}} = \cdots = \beta_p = 0 \qquad (۵)$$

در این حالت، این (q) پارامتر از مدل حذف شده و مدل جدید تنها با پارامترهای باقی‌مانده برازش داده می‌شود. مقدار RSS این مدل را با (RSS_0) نشان می‌دهیم. سپس آماره‌ی (F) از رابطه‌ی زیر محاسبه می‌شود:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \qquad (۶)$$

این روش در واقع مقایسه‌ی دو مدل **تو در تو**[۲] است. توجه شود که آماره‌ی (t) برای هر ضریب خاص، برابر با جذر آماره‌ی (F) متناظر با حذف همان پارامتر است که در این حالت $(q=1)$ خواهد بود.

نتایج مربوط به دیتاست مورد بررسی

برای دیتاست مورداستفاده، مقادیر محاسبه‌شده مطابق شکل ۳ هستند. مشاهده می‌شود که:

- مقدار $F=135.8$ عددی بزرگ است.
- مقدار p-value متناظر با آن ((Prob (F-statistic) بسیار کوچک است.

بنابراین، H_0 **رد می‌شود** و نتیجه می‌گیریم که **حداقل یکی از متغیرهای توضیحی تاثیر معناداری بر متغیر پاسخ دارد**.

[2] Nested Models

۲- آیا تمام پارامترها بر خروجی مؤثر هستند یا تنها بخشی از آن‌ها ؟

با اســتفاده از آزمون آماره‌ی F می‌توان نتیجه گرفت که حداقل یکی از پارامترها بر خروجی مؤثر اســت. با این حال، سؤال مهم‌تری مطرح می‌شود: **کدام پارامتر یا پارامترها واقعاً نقش مؤثری در مدل دارند؟**

این موضوع در مبحثی با عنوان انتخاب متغیرها[3] بررسی می‌شود که در آن معیارهایی مانند *Mallow's Cp*، AIC[4] ، BIC[5] و $Adjusted\ R^2$ معرفی و مقایسه می‌شوند. در اینجا، به‌اختصار به ایده‌ی کلی این روش‌ها اشاره می‌کنیم.

در حالت ایده‌آل، باید تمام مدل‌های ممکن ساخته شده و با یکدیگر مقایسه شوند تا بهترین زیرمجموعه‌ی متغیرها انتخاب شود. برای مثال، اگر تنها دو متغیر در نظر گرفته شود، چهار مدل قابل تعریف است: مدل بدون متغیر، مدل با متغیر اول، مدل با متغیر دوم و مدل با هر دو متغیر. این مدل‌ها می‌توانند بر اســاس معیارهایی مانند RSE یا $Adjusted\ R^2$ با یکدیگر مقایسه شوند. اما با افزایش تعداد متغیرها، این روش عملاً غیرقابل اجرا می‌شود، زیرا تعداد مدل‌های ممکن برابر با 2^p خواهد بود که p تعداد متغیرهای توضیحی است.

به همین دلیل، معمولاً از روش‌های گام‌به‌گام استفاده می‌شود که مهم‌ترین آن‌ها عبارتند از:

۱- انتخاب جلوسو[6]:

در این روش، ابتدا p مدل رگرسیون ساده ساخته می‌شود و متغیری انتخاب می‌شود که بهترین بهبود را طبق یک معیار مشخص (مانند AIC، BIC یا $Adjusted\ R^2$) ایجاد کند. سپس متغیرها به‌صورت مرحله‌ای به مدل چندگانه اضافه می‌شوند و این روند تا زمانی که شرط توقف از پیش تعیین‌شده برقرار شود ادامه می‌یابد.

۲- انتخاب عقب‌گرد[7]:

در این روش، از مدل کامل شامل تمام متغیرها شروع می‌شود و در هر مرحله، متغیری که کمترین شواهد آماری برای مؤثر بودن دارد (مثلاً بزرگ‌ترین مقدار p-value) حذف می‌شود. این فرآیند تا زمانی ادامه می‌یابد که معیار انتخاب (مانند حد آستانه‌ی p-value یا یک معیار اطلاعاتی) برقرار شود.

---

[3] Variable Selection

[4] *Akaike Information Criterion*

[5] *Bayesian Information Criterion*

[6] Forward Selection

[7] Backward Selection

۳- انتخاب ترکیبی[8]:

این روش ترکیبی از دو روش قبل است. ابتدا مدل بدون متغیر در نظر گرفته می‌شود و متغیرها به‌صورت جلوسو اضافه می‌شوند. در هر مرحله، اگر مقدار p-value یکی از متغیرهای موجود در مدل از حد مشخصی فراتر رود، آن متغیر به‌صورت عقب‌گرد حذف می‌شود. این روند تا زمانی ادامه می‌یابد که تمام متغیرهای موجود در مدل دارای p-value کوچک و تمام متغیرهای خارج از مدل، در صورت اضافه شدن، p-value بزرگی داشته باشند.

توجه

همانطور که در رگرسیون خطی ساده بحث شد، R^2 معیاری است که نشان می‌دهد مدل چه مقدار از تغییرات خروجی را توضیح می‌دهد:

$$R^2 = 1 - \frac{RSS}{TSS}$$

در مدل چندگانه، برای مقایسه کیفیت مدل‌هایی با تعداد متفاوتی از متغیرها، Adjusted R^2 تعریف می‌شود که واریانس توضیح داده شده را نسبت به تعداد پارامترها و تعداد نمونه‌ها تصحیح می‌کند:

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \quad (۷)$$

تفاوت اصلی:

- R^2 همواره با اضافه کردن متغیر جدید **افزایش می‌یابد** حتی اگر آن متغیر اثر واقعی نداشته باشد.
- Adjusted R^2 اضافه شدن متغیرهای بی‌اثر را جریمه می‌کند و معیار مناسبتری برای مقایسه‌ی مدل‌های چندگانه است.

براساس نتایج مدل کامل (شکل ۳)، متغیرهای *CARB_FLAG* و *COAL_FLAG* دارای مقدار p بالایی بوده و در از نظر آماری معنادار نیستند. به‌منظور بررسی اثر حذف این متغیرها، یک مدل ساده‌شده بدون آنها برازش داده شد که نتایج آن در شکل ۴ ارائه شده است.

[8] Mixed Selection

OLS Regression Results

Dep. Variable:	KLOGH (MD)	R-squared:	0.228
Model:	OLS	Adj. R-squared:	0.227
Method:	Least Squares	F-statistic:	160.2
Date:	Tue, 23 Dec 2025	Prob (F-statistic):	1.79e-266
Time:	11:56:49	Log-Likelihood:	-38351.
No. Observations:	4891	AIC:	7.672e+04
Df Residuals:	4881	BIC:	7.679e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.552e+05	4.19e+04	3.706	0.000	7.31e+04	2.37e+05
BVW (V/V)	-1.022e+04	652.304	-15.673	0.000	-1.15e+04	-8944.585
PHIF (V/V)	1.045e+04	602.298	17.350	0.000	9268.932	1.16e+04
RHOFL (G/CM3)	1155.5981	288.533	4.005	0.000	589.944	1721.253
RHOMA (G/CM3)	-1736.7630	833.829	-2.083	0.037	-3371.442	-102.084
RW (OHMM)	-4.027e+06	1.09e+06	-3.707	0.000	-6.16e+06	-1.9e+06
SAND_FLAG (UNITLESS)	188.1023	37.511	5.015	0.000	114.564	261.641
SW (V/V)	1449.4188	129.287	11.211	0.000	1195.959	1702.879
TEMP (DEGC)	-584.7684	159.995	-3.655	0.000	-898.430	-271.106
VSH (V/V)	-160.1278	20.918	-7.655	0.000	-201.137	-119.119

شکل ۴ : خروجی آماری ضرایب مدل ساده شده با حذف پارامترهای CARB_FLAG و COAL_FLAG

مقایسه‌ی نتایج نشان می‌دهد که با وجود کاهش جزئی در مقدار R^2 و Adjusted R^2، این کاهش بسیار ناچیز بوده و در مقابل، مدل ساده‌تر شده و آماره‌ی F افزایش یافته است. این موضوع نشان می‌دهد که این متغیرها اطلاعات افزوده‌ی معناداری فراتر از سایر متغیرهای مدل فراهم نمی‌کنند.

بر این اساس، مدل ساده‌شده به‌عنوان گزینه‌ای مناسب‌تر از نظر توازن بین سادگی و دقت آماری انتخاب می‌شود.

توجه:

عدم معناداری آماری این متغیرها در مدل خطی به معنی بی‌تأثیر بودن آن‌ها در مدل‌های غیرخطی یا روش‌های یادگیری ماشین نیست، زیرا این روش‌ها قادر به مدل‌سازی روابط غیرخطی و اثرات متقابل پیچیده هستند.

۳- تفاوت تفسیر معناداری پارامترها در رگرسیون خطی ساده و چندگانه

در برخی موارد، یک متغیر در رگرسیون خطی ساده بسیار معنادار به‌نظر می‌رسد، اما با افزودن سایر متغیرهای مرتبط در یک مدل چندگانه، میزان معناداری آماری آن کاهش می‌یابد یا حتی از بین می‌رود. این مسئله معمولاً ناشی از همبستگی آن متغیر با سایر متغیرهای ورودی (مسئله‌ی هم‌خطی[9]) است.

به عنوان مثال، پارامتر RHOMA که بیانگر چگالی ماتریس سنگ است، در یک مدل رگرسیون خطی ساده رابطه‌ی آماری بسیار معناداری با تراوایی افقی سنگ (KLOGH) نشان می‌دهد (باتوجه به مقادیر t و p این پارامتر در شکل ۵).

OLS Regression Results

Dep. Variable:	KLOGH (MD)	R-squared:	0.038
Model:	OLS	Adj. R-squared:	0.038
Method:	Least Squares	F-statistic:	195.1
Date:	Tue, 23 Dec 2025	Prob (F-statistic):	1.69e-43
Time:	11:56:49	Log-Likelihood:	-38888.
No. Observations:	4891	AIC:	7.778e+04
Df Residuals:	4889	BIC:	7.779e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.987e+04	2124.123	14.061	0.000	2.57e+04	3.4e+04
RHOMA (G/CM3)	-1.117e+04	799.445	-13.967	0.000	-1.27e+04	-9598.485

شکل ۵ : خروجی آماری مدل رگرسیون خطی ساده با پارامتر RHOMA

اما در مدل رگرسیون چندگانه، با وارد شدن متغیرهایی مانند تخلخل و حجم شیل، بخشی از اثری که در مدل ساده به RHOMA نسبت داده شده بود، توسط این متغیرها توضیح داده می‌شود (به مقادیر t و p این پارامتر در شکل ۳ و شکل ۴ توجه کنید).

به این ترتیب، ضرایب در مدل چندگانه اثر **مستقل** هر متغیر را در حضور سایر متغیرها نشان می‌دهند، نه اثر ترکیبی آن با پارامترهای همبسته. به بیان دیگر، معناداری یک متغیر در مدل ساده لزوماً به معنای اثرگذاری مستقیم

[9] Collinearity

و مستقل آن بر خروجی نیست. این تفاوت یکی از دلایل اصلی استفاده از رگرسیون چندگانه به‌جای چند رگرسیون ساده‌ی جداگانه است.

٤- مدل چندگانه چقدر در توضیح داده‌ها موفق بوده است؟

پاسخ به این پرسش مشابه رگرسیون خطی ساده و با استفاده از معیارهایی مانند RSE داده می‌شود. در مدل رگرسیون خطی چندگانه، مقدار RSE به‌صورت زیر تعریف می‌شود:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\,\mathrm{RSS}} \qquad (۸)$$

که در آن n تعداد داده‌ها و p تعداد پارامترهای مدل است.

RSE تخمینی از انحراف معیار خطای مدل است و نشان می‌دهد پیش‌بینی‌های مدل، به‌طور متوسط، چه فاصله‌ای از مقادیر واقعی پاسخ دارند. هرچه مقدار RSE کوچکتر باشد، مدل داده‌ها را با دقت بیشتری توضیح می‌دهد.

نکته‌ی مهم در مدل چندگانه این است که با افزایش تعداد پارامترها، مقدار RSS همواره کاهش می‌یابد؛ اما RSE با در نظر گرفتن درجه‌ی آزادی $(n-p-1)$ از کاهش ظاهری خطا که صرفا ناشی از افزایش تعداد پارامترهاست جلوگیری می‌کند. به همین دلیل، RSE معیار مناسب‌تری برای ارزیابی کیفیت برازش مدل نسبت به RSS خام محسوب می‌شود.

نتایج مربوط به دیتاست مورد بررسی

در این بخش، مشابه توضیحات ارائه شده در رگرسیون خطی ساده، مقدار RSE و نسبت آن به میانگین و انحراف معیار متغیر پاسخ برای دو مدل کامل و ساده‌شده (شکل ۴) گزارش شده است.

	RSE	RSE / mean(y)	RSE / std(y)
مدل کامل	589.821	296.38 %	87.45 %
مدل ساده شده	615.985	308.22 %	87.94 %

نسبت RSE به میانگین مقدار پاسخ، در این داده‌ها عدد بزرگی است که ناشی از توزیع بسیار چولگی‌دار تراوایی است. در چنین شرایطی، میانگین، معیار مناسبی برای مقیاس داده محسوب نمی‌شود و تفسیر این نسبت می‌تواند گمراه‌کننده باشد.

در مقابل، نسبت RSE به انحراف معیار پاسخ معیار مناسب‌تری برای ارزیابی عملکرد مدل است. مقدار به‌دست‌آمده نشان می‌دهد که خطای مدل رگرسیونی هنوز بخش قابل‌توجهی از پراکندگی داده‌ها را در بر می‌گیرد، اما در عین حال نسبت به مدل مرجعِ «پیش‌بینی مقدار ثابت برابر با میانگین KLOGH»، بخشی از پراکندگی داده‌ها کاهش یافته است. به‌طور مشخص، اگر این نسبت حدود ٪۸۷ باشد، به این معناست که حدود ٪۸۷ از پراکندگی داده‌ها همچنان توسط خطای مدل باقی مانده و مدل توانسته است حدود ٪۱۳ از پراکندگی داده‌ها را نسبت به مدل مبتنی بر میانگین توضیح دهد.

این برداشت کاملاً با مقادیر R^2 و R^2 Adjusted گزارش‌شده برای مدل‌های کامل و ساده‌شده سازگار است و نشان می‌دهد که اگرچه داده‌ها ذاتاً نویزی هستند و پیش‌بینی‌پذیری کامل آن‌ها محدود است، مدل رگرسیون خطی چندگانه همچنان اطلاعات معناداری از روابط بین متغیرها استخراج کرده است.

جمع‌بندی و نکات تکمیلی

در این فایل، چارچوب رگرسیون خطی چندگانه به‌عنوان یک ابزار آماری برای تحلیل همزمان اثر چند متغیر بر یک خروجی بررسی شد. تمرکز اصلی بر تفسیر ضرایب، معناداری آماری، و محدودیت‌های ذاتی مدل‌های خطی بود. مثال عملی از داده‌های واقعی مخزن نشان داد که اگرچه داده‌های مهندسی معمولاً نویزی هستند، رگرسیون چندگانه همچنان می‌تواند روابط معناداری بین متغیرها استخراج کند. این نگاه آماری، پایه‌ای ضروری برای درک صحیح‌تر روش‌های پیشرفته‌تر یادگیری ماشین فراهم می‌کند.

برای مقادیر داده شده‌ی x_i ، پیش‌بینی y با چه میزان عدم‌قطعیتی همراه است؟

برای هر مقدار مشخص از متغیرهای مستقل x_i، می‌توان مقدار خروجی y پیش‌بینی شده را به همراه عدم قطعیت پیش‌بینی بیان کرد. در مدل رگرسیون چندگانه، بازه پیش‌بینی[10] شامل دو مولفه اصلی است:

۱. خطای برآورد ضرایب β که ناشی از محدود بودن حجم نمونه است.

۲. خطای تصادفی ε که بخشی از تغییرات y را تشکیل می‌دهد و با هیچ مدلی قابل حذف نیست.

فرمول بازه پیش‌بینی برای یک مشاهده جدید x_0 به صورت زیر است:

$$\hat{y}_0 \pm t_{\left(\frac{\alpha}{2}, n-p-1\right)} \times s\sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

- $\hat{y}_0$ مقدار پیش‌بینی شده برای x_0 است.
- $s^2 = \frac{\text{RSS}}{n-p-1}$ برآورد واریانس خطا (مربع RSE) است.
- عبارت $x_0^T (X^T X)^{-1} x_0$ میزان فاصله x_0 از داده‌های مشاهده‌شده را نشان می‌دهد که بر بزرگی بازه‌ی پیش‌بینی اثر می‌گذارد.

[10] Prediction Interval

توجه: به دلیل وجود جمله‌ی ۱+ زیر رادیکال، بازه‌ی پیش‌بینی همواره بزرگتر از بازه اطمینان میانگین پاسخ[11] است. هر مشـاهده‌ی جدید دارای بازه‌ی پیش‌بینی مختص به خود اسـت؛ برخلاف RSE که یک معیار کلی برای کل مدل محسوب می‌شود.

در عمل، این بازه‌ها می‌توانند به‌صورت نمودار همراه با مقادیر پیش‌بینی‌شده ترسیم شوند تا هم دقت پیش‌بینی و هم عدم‌قطعیت هر نقطه به‌صورت تصویری مشخص شود (کد و نمودارهای مربوطه در مخزن گیت‌هاب ارائه شده‌اند).

اثر متقابل پارامترها

در مدل رگرسیون خطی چندگانه فرض می‌شود که هر متغیر به‌صورت مستقل و جمع‌پذیر بر خروجی تأثیر می‌گذارد؛ به این معنا که اثر هر متغیر بر خروجی مسـتقل از مقدار سایر متغیرها در نظر گرفته می‌شـود. این فرض را فرض جمع‌پذیری[12] می‌نامند.

با این حال، در بسیاری از مسائل واقعی ممکن است اثر یک متغیر بر خروجی به مقدار متغیر دیگری وابسته باشد. در چنین شرایطی، می‌توان اثر متقابل[13] متغیرها را مستقیما در مدل لحاظ کرد.

برای مثال، در یک مدل دو متغیره می‌توان جمله‌ی تعاملی x_1x_2 را به مدل اضافه کرد:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

برای روشن‌تر شدن تفسیر، این مدل را می‌توان به‌صورت زیر بازنویسی کرد:

$$Y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + \epsilon$$

در این حالت مشاهده می‌شود که اثر x_1 بر خروجی دیگر ثابت نیست، بلکه به مقدار x_2 وابسته است. به‌طور مشخص، افزایش یک واحدی x_1 باعث تغییر خروجی به اندازه‌ی $(\beta_1 + \beta_3 x_2)$ می‌شود. به این ترتیب، وابستگی اثر متغیرها به یکدیگر در مدل گنجانده می‌شود، در حالی که مدل همچنان نسبت به ضرایب خطی باقی می‌ماند.

---

[11] Confidence Interval

[12] Additive Assumption

[13] Interaction

اثرات غیرخطی پارامترها

تا اینجا فرض شد که اثر هر پارامتر بر خروجی به‌صورت خطی است؛ به این معنا که فارغ از مقدار خود x_i، تغییر یک واحدی در x_i همواره باعث تغییر ثابتی در خروجی y می‌شود. با این حال، در بسیاری از مسائل واقعی این فرض برقرار نیست و ممکن است رابطه‌ی یک پارامتر با خروجی غیرخطی باشد. به‌عنوان مثال، ممکن است خروجی به توان دوم یک پارامتر، مانند x_i^2، وابسته باشد.

در چنین شرایطی نیز می‌توان از چارچوب رگرسیون خطی استفاده کرد. کافی است عبارت غیرخطی به‌عنوان یک متغیر جدید وارد مدل شود. برای مثال، برای در نظر گرفتن اثر درجه‌ی دوم یک پارامتر، مدل را می‌توان به‌صورت زیر نوشت:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

توجه شود که با وجود غیرخطی بودن رابطه‌ی y و x، این مدل همچنان یک مدل خطی محسوب می‌شود، زیرا **خطی بودن در رگرسیون به رابطه‌ی خطی نسبت به ضرایب $\boldsymbol{\beta}$ اشاره دارد**، نه به شکل متغیرها. به این نوع مدل‌ها رگرسیون چندجمله‌ای[۱۴] گفته می‌شود.

به‌طور مشابه، می‌توان تبدیل‌هایی مانند $\log(x)$، $\sqrt{x}$ یا سایر توابع را نیز به‌عنوان متغیرهای جدید وارد مدل کرد. این کار انعطاف‌پذیری مدل را افزایش می‌دهد، اما همزمان می‌تواند منجر به افزایش پیچیدگی مدل و خطر بیش‌برازش شود؛ بنابراین استفاده از این عبارات باید همراه با معیارهای مناسب انتخاب مدل انجام گیرد.

نکته‌ی مهم

نتایج رگرسیون تنها در صورتی معتبر هستند که فروض اصلی مدل تقریباً برقرار باشند. در عمل، چالش‌های زیر می‌توانند منجر به تفسیر نادرست ضرایب و آزمون‌های آماری شوند:

- غیرخطی بودن رابطه‌ی متغیرها
- ناهمسانی واریانس خطا
- وجود داده‌های پرت یا نقاط با اهرم بالا
- هم‌خطی بین متغیرهای توضیحی

بررسی این موارد پیش از نتیجه‌گیری نهایی ضروری است. توضیحات تکمیلی و کدهای مربوطه در بخش ویکی مخزن گیت‌هاب ارائه شده‌اند.

---

[14] Polynomial Regression