رگرسیون خطی ساده

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.

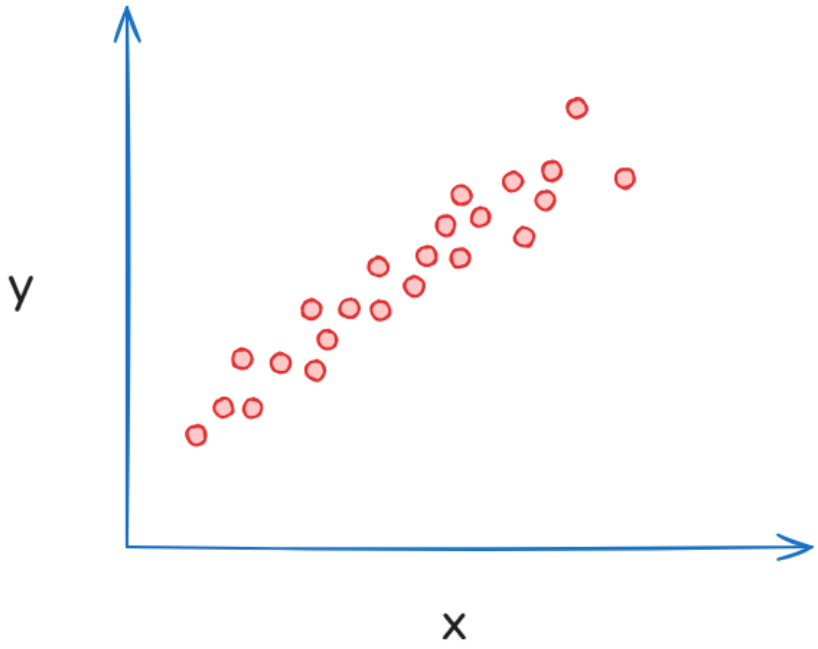
https://github.com/ML-OilGas/ML_Algorithms/Simple_Regression

ارتباط با نگارنده و ارائه‌ی نظرات و پیشنهادات:

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

مقدمه

پس از انجام آزمایش‌هایی، مجموعه‌ای از داده‌ها مطابق شکل ۱ به دست آمده است. هدف، یافتن رابطه‌ای میان x و y است تا بتوان برای مقادیر دلخواه x، مقدار متناظر y را بدون نیاز به انجام آزمایش مجدد تخمین زد.



شکل ۱ : داده‌های بدست آمده از آزمایش

$$y = f(x) + \varepsilon \qquad (۱)$$

همان‌طور که در رابطه‌ی (۱) نشان داده شده است، هدف تعیین تابع f یا به‌دست‌آوردن تقریبی از آن است. توجه کنید که در واقعیت، به دلیل وجود خطای ε، هیچ‌گاه نمی‌توان به رابطه‌ای رسید که مقادیر واقعی را دقیقاً بازتولید کند. به این خطا، خطای غیرقابل‌کاهش[1] گفته می‌شود. این خطا بخشی از تغییرات y است که حتی بهترین مدل نیز قادر به پیش‌بینی آن نیست؛ زیرا ناشی از عواملی مانند خطای اندازه‌گیری، تغییرات تصادفی محیط، یا متغیرهایی است که در مدل در نظر گرفته نشده‌اند.

بنابراین، تمام تلاش ما در یافتن تقریب مناسبی از تابع f، کاهش خطای مدل‌سازی و نزدیک‌تر کردن $\hat{f}$ (تابع تخمینی) به f واقعی است. با این حال، حتی اگر $f = \hat{f}$ باشد، باز هم مقادیر پیش‌بینی‌شده‌ی $\hat{y}$ دقیقاً برابر با مقادیر واقعی y نخواهد بود، زیرا وجود ε اجتناب‌ناپذیر است.

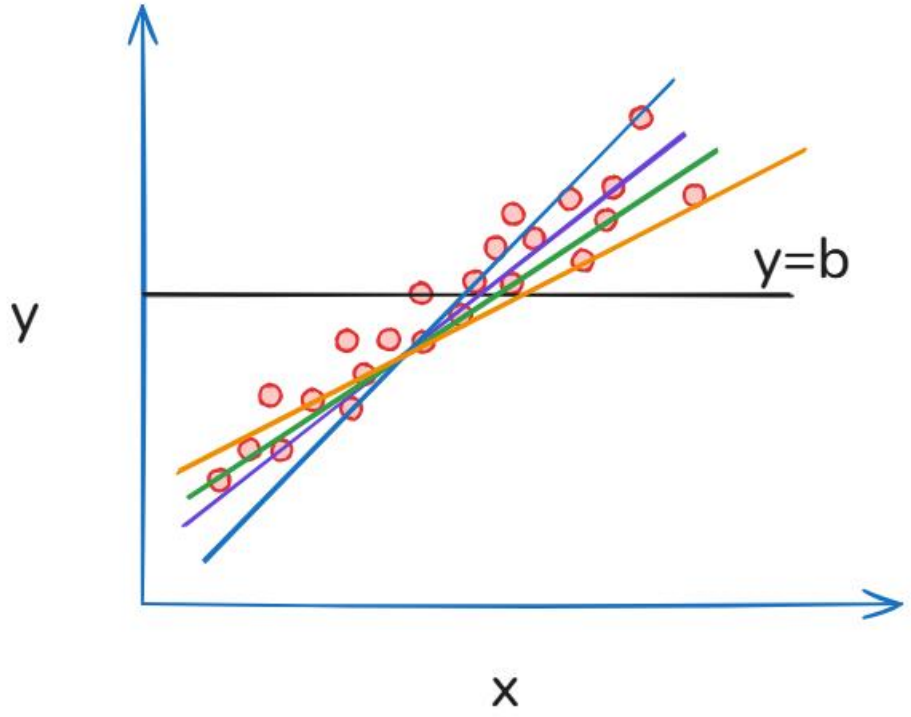
---

[1] Irreducible Error

برای یافتن تابع f، ساده‌ترین راه تقریب‌زدن آن با یک خط است. با توجه به اینکه در نمودار ارائه‌شده می‌توان بی‌نهایت خط از میان داده‌ها رسم کرد، لازم است معیاری برای انتخاب «بهترین» خط داشته باشیم. به‌عنوان مثال، اگر x فشار و y دبی تولیدی باشد، هدف یافتن بهترین خطی است که تغییرات دبی را بر حسب فشار توضیح دهد.

از میان تمام خط‌های ممکن، ساده‌ترین آن‌ها **خط میانگین** است که به صورت خط افقی $y = b$ در شکل ۲ نشان داده شده است. همچنین چند خط دیگر نیز به‌عنوان نمونه رسم شده‌اند. معادله‌ی کلی این خط‌ها را می‌توان به‌صورت زیر نوشت:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (۲)$$



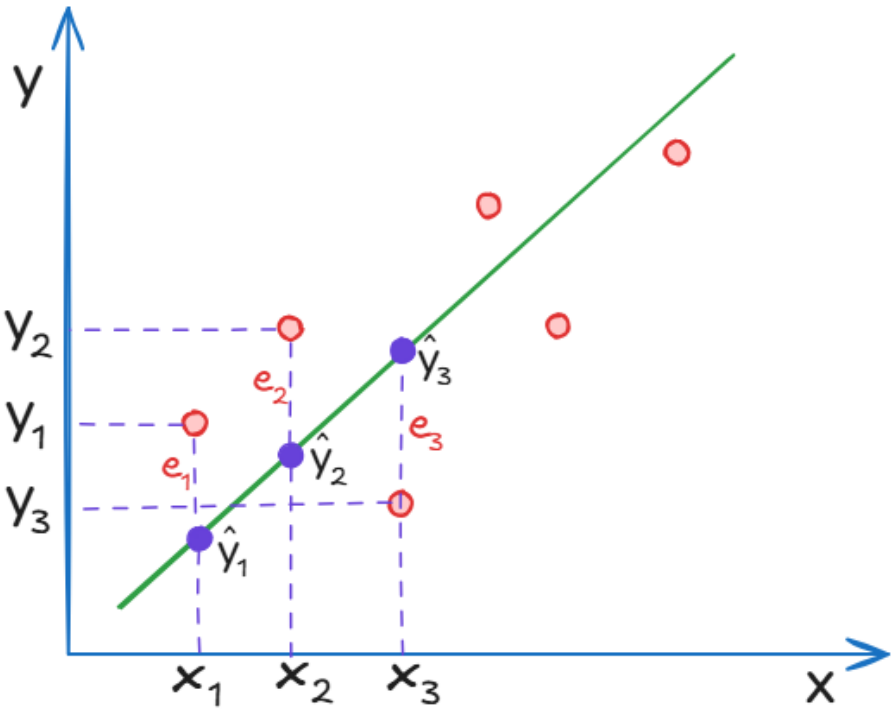
شکل ۲ : خط میانگین و خطوط فرضی

اکنون لازم است معیاری برای سنجش «خوبی» یا «بدی» یک خط داشته باشیم. برای این منظور **خطا** را تعریف می‌کنیم. مطابق شکل ۳، از نقطه‌ی x_1 خطی عمودی رسم می‌کنیم تا خط انتخاب‌شده را قطع کند. محل تلاقی، مقدار تخمینی برای نقطه‌ی x_1 است که در شکل با دایره‌ی بنفش نشان داده شده و مقدار آن $\hat{y}_1$ است. سپس مقدار واقعی و مقدار تخمینی را از یکدیگر کم می‌کنیم. به این ترتیب، **خطای تخمین** در نقطه‌ی x_1 به‌دست می‌آید که با e_1 نشان داده شده است.

$$e_1 = y_1 - \hat{y}_1 \quad (۳)$$

همین کار را برای نقطه‌ی x_2 و سایر نقاط انجام می‌دهیم و خطاها را با یکدیگر جمع می‌کنیم. از آن‌جایی که برخی خطاها مثبت و برخی منفی هستند، برای جلوگیری از خنثی‌شدن آن‌ها، از توان دوم خطاها استفاده می‌شود.

در نهایت، رابطه‌ای به‌دست می‌آید که در آن توان دوم خطای تخمین برای تمام نقاط جمع شده است. به این رابطه **مجموع مربعات باقیمانده** یا RSS[2] گفته می‌شود.



شکل ۳ : نحوه‌ی محاسبه‌ی خطای نقاط با خط رگرسیون

$$RSS = {e_1}^2 + {e_2}^2 + {e_3}^2 + \cdots = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + \cdots$$

$$RSS = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (۴)$$

در یادگیری ماشین و مسائل رگرسیون به اختلاف بین مقدار واقعی و مقدار تخمینی، **باقیمانده**[3] گفته می‌شود.

---

[2] Residual Sum Of Squares (RSS)

[3] Residual

اگر مقدار RSS را برای خط‌های مختلف محاسبه کنیم، بهترین خط آن است که **کمترین RSS** را داشته باشد که به آن **خط رگرسیون**[۴] گفته می‌شود. برای یافتن این خط به صورت تحلیلی، کافی است معادله‌ی کلی خط را در رابطه‌ی RSS جای‌گذاری کرده و با مشتق‌گیری نسبت به ضرایب β_0 و β_1، مقادیری را بیابیم که RSS را کمینه می‌کنند. نتیجه‌ی این فرایند، ضرایب خطی‌ای هستند که داده‌ها را به بهترین شکل توصیف می‌کنند. (برای جزئیات ریاضی، به پیوست مراجعه کنید).

با توجه به این‌که ضرایب خط رگرسیون با کمینه‌کردن مجموع مربعات خطا (RSS) به‌دست می‌آیند، این روش با نام **روش حداقل مربعات**[۵] نیز شناخته می‌شود.

در بخش بعد، برای یک دیتاست نفتی و با استفاده از پایتون، این خط را رسم می‌کنیم.

---

[4] Regression Line

5 Least Squares Method

اجرای عملی

در این تمرین از مجموعه‌داده‌ای استفاده می‌کنیم که شامل چند ویژگی زمین‌شناسی مهم در ارزیابی مخازن نفت و گاز است. این ویژگی‌ها عبارت‌اند از:

- Porosity (%): درصد تخلخل سنگ، یعنی نسبت حجم فضای خالی به حجم کل سنگ. هرچه تخلخل بیشتر باشد، ظرفیت ذخیره‌ی سیال (نفت، گاز یا آب) بیشتر است.
- Matrix permeability (nd): تراوایی ماتریکس سنگ، که بیانگر توانایی سنگ برای عبور دادن سیال از درون خود است.
- Acoustic impedance ($kg/m^2 \cdot s \times 10^6$): امپدانس آکوستیکی، حاصل‌ضرب چگالی در سرعت موج صوتی است و معمولاً برای تفسیر داده‌های لرزه‌ای و شناسایی لایه‌های زمین‌شناسی به کار می‌رود.
- Brittleness ratio: نسبت شکنندگی سنگ، که نشان می‌دهد سنگ تا چه حد مستعد شکستگی است (ویژگی مهم در تحلیل شکست هیدرولیکی).
- TOC (%): درصد کربن آلی کل، معیاری از غنای سنگ از مواد آلی قابل تبدیل به هیدروکربن.
- Vitrinite reflectance (%): بازتاب ویتیرنیت، شاخصی از بلوغ حرارتی سنگ مخزن و میزان تبدیل مواد آلی به نفت و گاز.

ویژگی موردنظر برای بررسی و تحلیل، $A\sqrt{K}$ است؛ پارامتری که از تحلیل RTA[6] به‌دست می‌آید و در واقع معیاری از بهره‌وری یا توان تولیدی چاه محسوب می‌شود. در مخازن غیرمتعارف، $A\sqrt{K}$ نقش مشابه پارامتر kh (تراوایی ضرب‌در ضخامت) را در مخازن متعارف دارد. از نظر فیزیکی، چون $A\sqrt{K}$ با ریشه‌ی دوم تراوایی متناسب است، افزایش تخلخل یا شکنندگی معمولاً باعث رشد این شاخص می‌شود. بنابراین انتظار می‌رود میان $A\sqrt{K}$ و برخی از متغیرهای سنگی مانند تخلخل یا تراوایی نوعی رابطه‌ی همبستگی مثبت دیده شود.

در ادامه، قصد داریم رفتار $A\sqrt{K}$ را نسبت به تخلخل بررسی کنیم و ببینیم چه رابطه‌ای بین این دو برقرار است. تعدادی از ردیف‌های این دیتاست در شکل ۴ نشان داده شده است.

[6] Rate Transient Analysis

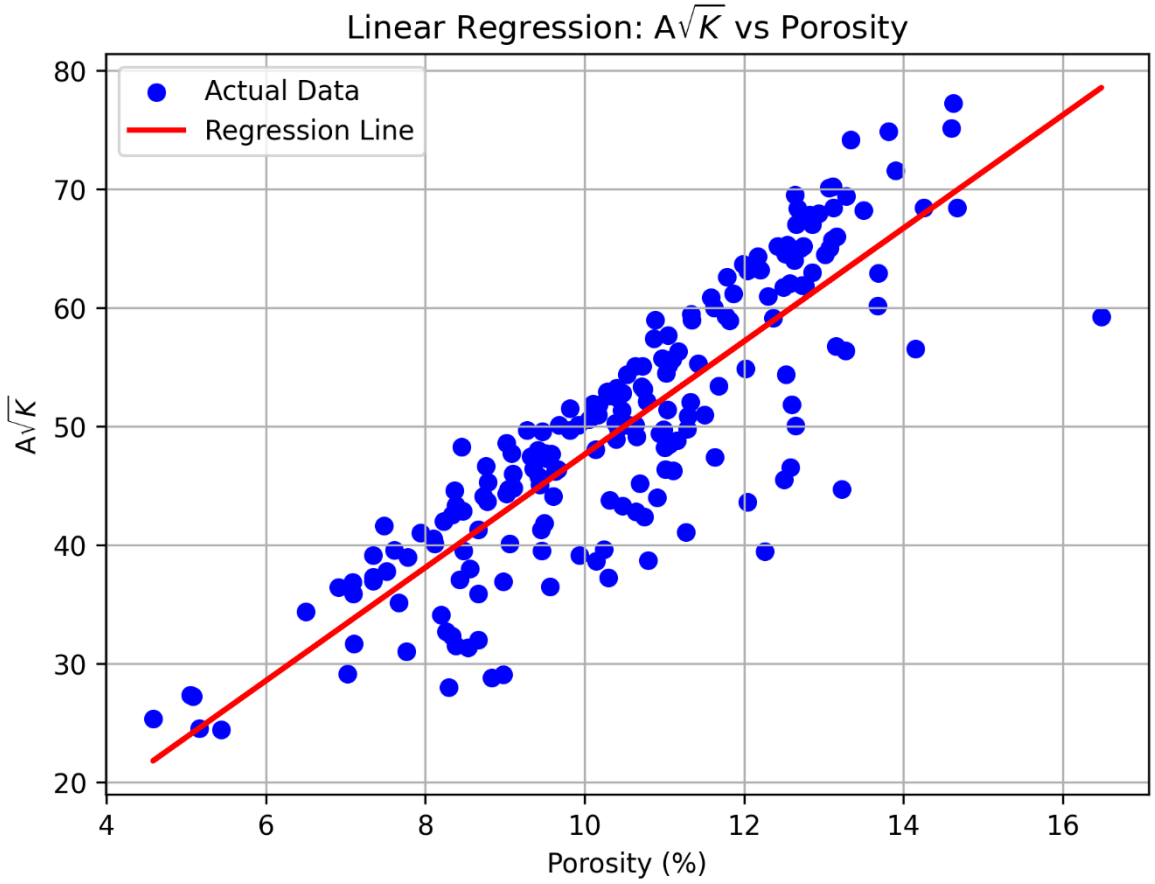
	Porosity (%)	Matrix Perm (nd)	Acoustic impedance (kg/m2s*10^6)	Brittleness Ratio	TOC (%)	Vitrinite Reflectance (%)	Aroot(K)
0	8.456	292	3.080	97.680	4.64	1.848	48.306469
1	8.666	353	3.542	55.404	3.56	1.504	41.300912
2	9.814	259	4.411	87.360	3.56	2.176	49.688356
3	12.369	675	2.893	47.772	4.32	1.504	59.132694
4	12.264	457	3.498	13.128	6.04	1.520	39.503121

شکل ۴ : بخشی از دیتاست مورد بررسی

برای رسم خط رگرسیون از کتابخانه‌ی sklearn و تابع LinearRegression استفاده می‌کنیم. خروجی در شکل ۵ نشان داده شده است. معادله‌ی خط رگرسیون بدست آمده برابر است با:

$$y = -0.034 + 4.768\,x \tag{5}$$

که در آن $x \equiv Porosity$ و $y \equiv A\sqrt{K}$ است. بر اساس این معادله، می‌توان گفت که **به‌طور متوسط، هر ۱ درصد افزایش در تخلخل باعث افزایش حدود ۴.۷۶۸ واحدی در** $A\sqrt{K}$ می‌شود.



شکل ۵ : داده‌های اصلی (نقاط آبی) و خط رگرسیون (خط قرمز)

ارتباط با یادگیری ماشین:

در این فایل، رگرسیون خطی ساده عمدتاً با رویکرد آماری و با هدف تفسیر ضرایب و بررسی معناداری متغیرها بررسی شده است. در کاربردهای یادگیری ماشین، تمرکز اصلی بر قدرت پیش‌بینی مدل روی داده‌های دیده‌نشده است که مستلزم تقسیم داده به مجموعه‌های آموزش و آزمون است. پیاده‌سازی این رویکرد، به همراه کد مربوطه، در مخزن گیت‌هاب پروژه ارائه شده است.

بحث در نتایج

۱- آیا رابطه‌ای بین x و y وجود دارد؟

در ابتدا فرض کردیم که x و y از طریق تابعی مانند $f(x)$ به یکدیگر مرتبط هستند. اما پرسش مهم این است که **آیا اساساً رابطه‌ای میان این دو متغیر وجود دارد که بتوان براساس آن چنین تابعی را فرض کرد؟**

این موضوع اهمیت دارد؛ زیرا می‌توانیم هر دو متغیر را روی یک نمودار رسم کنیم و با روش توضیح‌داده‌شده بهترین خط را بر داده‌ها برازش کنیم، **حتی اگر در واقعیت هیچ ارتباط علّی یا طبیعی میان آن‌ها وجود نداشته باشد.** برای مثال، می‌توان نرخ رشد جمعیت کشور را در برابر تعداد حفاری چاه‌های نفت و گاز رسم کرد، اما روشن است که چنین ارتباطی از نظر علمی نامعتبر و صرفاً حاصل تصادف داده‌هاست.

برای اینکه مشخص کنیم رابطه‌ی مشاهده‌شده واقعی است یا تنها نتیجه‌ی تصادف، از **آزمون فرض آماری** استفاده می‌کنیم. این فرآیند به‌طور کلی شامل چهار گام است:

گام اول) ساخت فرضیه‌ها

برای پاسخ به این پرسش، دو فرضیه زیر را مطرح می‌کنیم:

- H_0 بین این دو متغیر ارتباطی وجود ندارد؛ یعنی $\beta_1 = 0$.
- H_a بین دو متغیر ارتباط وجود دارد؛ یعنی $\beta_1 \neq 0$.

در آمار به فرضیه‌ی اول **فرضیه‌ی صفر**[۷] گفته می‌شود. نکته مهم این است که H_0 طوری تعریف می‌شود که در صورت صحیح‌بودن، **وضعیت پیش‌فرض** را تأیید کند. درواقع پیش‌فرض این است که ارتباطی بین دو متغیر وجود ندارد و ما به‌دنبال شواهدی برای کشف این ارتباط هستیم. بنابراین، رد شدن H_0 لحظه‌ای است که به **کشف وجود رابطه‌ی معنادار میان دو متغیر** می‌رسیم.

[7] Null Hypothesis

به H_a نیز **فرضیه‌ی جایگزین**[8] گفته می‌شود که دقیقاً عکس H_0 تعریف می‌شود. به‌طور خلاصه، هدف ما این است که در صورت وجود شواهد کافی، **H_0 را رد و H_a را بپذیریم**.

نکته‌ی بسیار مهم:

رد نشدن H_0 **به‌هیچ‌وجه** به معنای درست‌بودن آن نیست. برای مثال، اگر در این تحلیل نتوانیم H_0را رد کنیم، نمی‌توان نتیجه گرفت که قطعاً بین دو متغیر رابطه‌ای وجود ندارد؛ زیرا ممکن است علت آن **کم بودن تعداد نمونه‌ها یا کیفیت پایین داده‌ها** باشد. البته ممکن است واقعاً H_0 صحیح باشد. به‌طور خلاصه، **رد نشدن H_0 یعنی شواهد کافی برای رد آن در دست نیست.**

گام دوم) محاسبه‌ی آماره‌ی t[9]

برای بررسی رد یا عدم‌رد H_0 باید **آماره‌ی t** را محاسبه کنیم. در مسئله‌ی رگرسیون، فرضیه‌ی H_0 برابر است با $\beta_1 = 0$؛ زیرا اگر این ضریب صفر باشد، مقدار y عملاً **مستقل از مقدار** x خواهد بود (یعنی $y = \beta_0$).

بنابراین پس از برازش مدل رگرسیون، مقدار ضریب $\widehat{\beta_1}$ و خطای استاندارد آن $SE(\widehat{\beta_1})$ محاسبه می‌شود. آماره‌ی t از رابطه‌ی زیر به‌دست می‌آید:

$$T = \frac{(\widehat{\beta_1} - 0)}{SE(\widehat{\beta_1})} \qquad (6)$$

این آماره نشان می‌دهد که مقدار تخمینی ضریب β_1 چند برابر خطای استاندارد خود از مقدار صفر فاصله دارد. به‌صورت شهودی، هرچه مقدار قدرمطلق بزرگ‌تر باشد، شواهد قوی‌تری علیه فرضیه‌ی صفر وجود دارد.

اما سوال اصلی این است: چه مقدار بزرگی از T برای رد فرضیه‌ی صفر کافی است؟

گام سوم) محاسبه‌ی p-value

برای پاسخ به این پرسش، مقدار p محاسبه می‌شود. مقدار p احتمال مشاهده‌ی آماره‌ای به‌اندازه‌ی مقدار مشاهده‌شده (یا بزرگ‌تر از آن) است، مشروط بر این‌که فرضیه‌ی صفر صحیح باشد.

---

[8] Alternative Hypothesis

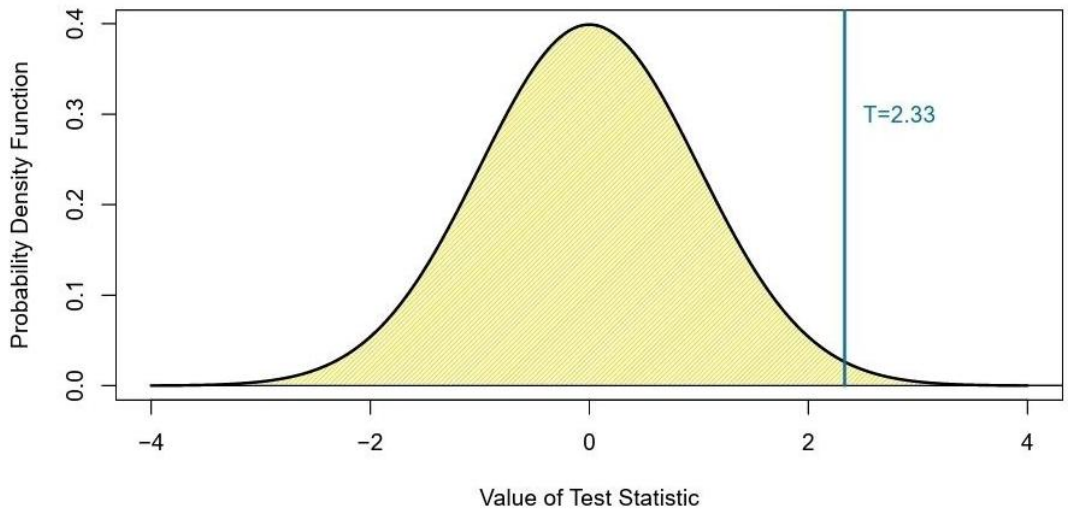
[9] t-statistic (or t-value)

به‌عبارت دیگر، اگر H_0 درست باشد، مقدار p نشان می‌دهد مشاهده‌ی چنین مقدار بزرگی از آماره‌ی T تا چه اندازه محتمل است.

اهمیت مقدار p در این است که یک کمیت بدون مقیاس و بین صفر و یک ارائه می‌دهد که امکان تصمیم‌گیری ساده‌تر را فراهم می‌کند، در حالی‌که مقدار T به‌تنهایی مقیاس مشخصی برای قضاوت ندارد.

به عنوان مثال اگر مقدار T محاسبه‌شده برابر ۲.۳۳ باشد، می‌توان آن را روی توزیع t (شکل ۶) قرار داد. مقدار p در واقع مساحت ناحیه‌ی سایه‌زده در دم توزیع است که نشان می‌دهد، *در صورت صحیح بودن فرضیه‌ی صفر*، احتمال آنکه آماره‌ی T مقداری برابر یا بزرگ‌تر از مقدار مشاهده‌شده بگیرد چقدر است.

هرچه این احتمال کوچک‌تر باشد، وقوع چنین مقداری صرفاً بر اثر تصادف داده‌ها بعیدتر است. بنابراین مقدار کوچک p نشان می‌دهد که شواهد قوی علیه فرضیه‌ی صفر وجود دارد.



شکل ۶ : توزیع t و نمایش موقعیت آماره‌ی t محاسبه‌شده؛ مساحت ناحیه‌ی دم توزیع متناظر با p-value است

گام چهارم) تصمیم‌گیری

در این مرحله مقدار p با یک سطح معناداری از پیش تعیین‌شده (مثلاً $\alpha = 0.05$) مقایسه می‌شود.

- اگر $p < \alpha$، فرضیه‌ی صفر رد می‌شود و وجود رابطه‌ی معنادار بین x و y نتیجه‌گیری می‌شود.
- اگر $p \geq \alpha$ ، شواهد کافی برای رد فرضیه‌ی صفر وجود ندارد.

انتخاب سطح معناداری به حوزه‌ی علمی و نوع مسئله بستگی دارد. در بسیاری از مسائل مهندسی مقدار ٪۵ رایج است، در حالی‌که در برخی حوزه‌ها مانند فیزیک یا علوم پایه، سطوح معناداری بسیار کوچک‌تری مورد استفاده قرار می‌گیرد.

نکته:

مقدار ۰٫۰۵ در واقع احتمال خطای نوع اول اســت؛ یعنی احتمال رد کردن اشــتباه یک فرضــیه‌ی درســت. توضیحات بیشتر درباره‌ی خطاهای نوع اول و دوم در گیتهاب آمده است.

نتایج مربوط به دیتاست مورد بررسی

برای دیتاست مورداستفاده، مقادیر محاسبه‌شده به‌صورت جدول ۱ هستند. مشاهده می‌کنید که:

- مقدار t برای ضریب تخلخل بزرگ است.
- مقدار p بسیار کوچک است.

بنابراین، H_0 **رد می‌شود** و نتیجه می‌گیریم که **بین تخلخل و** $A\sqrt{K}$ **رابطه‌ای معنادار وجود دارد.**

جدول ۱ خروجی آماری ضرایب خط رگرسیون

	Coef	Std err	t	P>|t|	[0.025 0.975]
Const	-0.0341	2.132	-0.016	0.987	[-4.239 4.171]
Porosity (%)	4.7680	0.199	23.918	0.000	[4.375 5.161]

۲- با فرض اینکه رابطه‌ی واقعی بین X و Y واقعاً خطی باشــد، ضــرایب تخمین زده شــده چقدر به مقادیر واقعی نزدیک هستند؟

در بخش روش دیدیم که اگر رابطه‌ی واقعی بین دو متغیر به‌صورت

$$f(x) = \beta_0 + \beta_1 x$$

باشد، مدل رگرسیون با استفاده از داده‌ها، تخمینی از آن تولید می‌کند:

$$f(x) \approx \hat{\beta}_0 + \hat{\beta}_1 x$$

حال پرسش مهم این است که **ضرایب تخمینی** $\hat{\beta}_0$ **و** $\hat{\beta}_1$ **تا چه حد به ضرایب واقعی** β_0 **و** β_1 **نزدیک‌اند؟**

برای کمی‌سازی عدم‌قطعیت در تخمین ضرایب، از **بازه‌ی اطمینان**[10] استفاده می‌شود. برای یک بازه‌ی اطمینان ۹۵٪ روابط زیر برقرار است:

[10] Confidence Interval

$$\widehat{\beta_0} \pm 2SE(\widehat{\beta_0})$$
$$\widehat{\beta_1} \pm 2SE(\widehat{\beta_1}) \quad (7)$$

نکته:

ضریب «۲» یک تقریب مناسب برای بازه‌ی ۹۵٪ در نمونه‌های بزرگ است. وقتی تعداد داده‌ها کمتر باشد، این عدد کمی تغییر می‌کند (مثلاً ۲.۰۲ یا ۲.۰۴) اما تفاوت معمولاً زیاد نیست.

تفسیر بازه‌ی اطمینان

بازه‌ی اطمینان ۹۵٪ به این معنی است که:

اگر فرآیند نمونه‌گیری را بارها تکرار کنیم، در ۹۵٪ مواقع بازه‌ای که با روابط بالا ساخته می‌شود، شامل مقدار واقعی βخواهد بود.

بنابراین **هرچه طول بازه کوچک‌تر باشد، تخمین دقیق‌تر است.** همچنین اگر بازه شامل عدد صفر *نباشد* (و کامل بالای صفر قرار داشته باشد)، نشان می‌دهد که ضریب با احتمال زیاد واقعاً غیرصفر است. این موضوع نیز به رد کردن فرضیه‌ی صفر کمک می‌کند.

نتیجه برای دیتاست موجود

مطابق ستون آخر جدول ۱، بازه‌ی اطمینان ۹۵٪ برای ضریب تخلخل به‌صورت زیر گزارش شده است:

$$[4.375, 5.161]$$

این بازه:

- **کوچک** است؛ یعنی تخمین $\hat{\beta}_1$ دقت بالایی دارد.
- **کاملاً مثبت** است و شامل صفر نمی‌شود؛ یعنی ضریب تقریباً با قطعیت متفاوت از صفر است.
- نشان می‌دهد که رابطه‌ی بین تخلخل و $A\sqrt{K}$ با داده‌های موجود بسیار **پایدار، معنادار و سازگار با مدل خطی** است.

۳- دقت کلی مدل چقدر است؟

حتی اگر در رابطه‌ی (۱)، تابع f دقیقاً معلوم باشد، به دلیل وجود جمله‌ی خطا ε، مقادیر تخمینی $\hat{y}$ همیشه دارای خطا خواهند بود. بنابراین لازم است با استفاده از چند معیار آماری، **دقت کلی مدل** را ارزیابی کنیم.

۱- معیار RSE

اولین معیار RSE[11] است که از رابطه‌ی زیر محاسبه می‌شود:

$$RSE = \sqrt{\frac{RSS}{n-2}} \tag{8}$$

در این رابطه:

- RSS مجموع مربعات باقی‌مانده‌ها است (در بخش مقدمه معرفی شد)،
- n تعداد داده‌ها است.

تفسیر RSE:

RSE تخمینی از **انحراف معیار خطاهای مدل** است و نشان می‌دهد به‌طور میانگین، مقادیر پیش‌بینی‌شده‌ی $\hat{y}$ چقدر از مقادیر واقعی فاصله دارند.

در مثال مورد بررسی، مقدار RSE برابر است با:

$$RSE = 5.849$$

یعنی مقدار تخمینی $A\sqrt{K}$ به‌طور متوسط حدود **۵.۸۵ واحد** با مقدار واقعی اختلاف دارد. اما آیا این مقدار قابل قبول است؟

تشخیص این موضوع به مقیاس مسئله بستگی دارد. برای قضاوت بهتر، می‌توان RSE را با مقادیر دیگر مقایسه کرد.

نسبت RSE به میانگین مقدار پاسخ

$$\frac{5.849}{50} \times 100 = 11.7\%$$

یعنی خطای متوسط مدل حدود **۱۱.۷٪** از مقدار متوسط $A\sqrt{K}$ است.

---

[11] Residual Standard Error

نسبت RSE به انحراف معیار پاسخ

$$\frac{5.849}{11.5} \times 100 = 50.86\%$$

این یعنی خط رگرسیون توانسته حدود **۵۱٪** از پراکندگی داده‌ها را نسبت به حالت «رسم یک خط افقی روی میانگین» توضیح دهد. این برداشت مکمل و هماهنگ با R^2 است.

توجه:

مقادیر میانگین و انحراف معیار پاسخ در کد نوشته شده محاسبه شده‌اند.

۲- معیار R^2

معیار مهم دیگر، ضریب تعیین یا R^2 است. این معیار عددی **بین ۰ و ۱** بوده و از رابطه‌ی زیر محاسبه می‌شود:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (9)$$

در این رابطه:

- TSS پراکندگی کل داده‌ها حول میانگین است و از نظر فرم مشابه RSS محاسبه می‌شود (رابطه‌ی (۴))، با این تفاوت که در آن به‌جای مقدار تخمینی $\hat{y}$، مقدار میانگین $\bar{y}$استفاده می‌شود،
- RSS بخش توضیح‌داده‌نشده‌ی پراکندگی پس از برازش مدل است.

تفسیر R^2:

این معیار سهمی از پراکندگی داده‌ها را نشان می‌دهد که مدل توانسته توضیح دهد. هرچه مقدار آن بزرگ‌تر باشد، مدل بهتر توانسته ساختار داده را بازسازی کند.

در دیتاست بررسی‌شده:

$$R^2 = 0.743$$

یعنی **۷۴.۳٪ از پراکندگی داده‌ها توسط مدل خطی توضیح داده شده است**. ۲۶٪ باقیمانده ممکن است ناشی از:

- متغیرهای حذف‌شده،
- خطاهای اندازه‌گیری،
- یا نویز ذاتی سیستم باشد.

نکته‌ی مهم:

برخلاف RSE که مقدار آن وابسته به مقیاس عددی Y است، R^2، عددی **بدون بعد و کاملاً نسبی** است. بنابراین برای **مقایسه‌ی مدل‌های مختلف** معیار مناسب‌تری به شمار می‌آید.

پیوست

- **استخراج روابط ضرایب خط رگرسیون**

فرض می‌کنیم داده‌های زوج (x_i, y_i) برای $i = 1, \cdots, n$ را داریم و مدل خطی زیر را فرض می‌کنیم:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

تابع خطای مجموع مربعات باقیمانده (RSS):

$$\mathrm{RSS}(\beta_0, \beta_1) = \Sigma_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

برای پیدا کردن β_0 و β_1 که RSS را کمینه می‌کنند، مشتق جزئی نسبت به هر یک می‌گیریم و برابر صفر قرار می‌دهیم (شروط لازم برای کمینه).

مشتق جزیی نسبت به $\boldsymbol{\beta_0}$

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Sigma_{i=1}^{n} y_i = n\beta_0 + \beta_1 \Sigma_{i=1}^{n} x_i$$

$$\frac{\Sigma_{i=1}^{n} y_i}{n} = \beta_0 + \beta_1 \frac{\Sigma_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \beta_0 + \beta_1 \bar{x} \Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

مشتق جزیی نسبت به $\boldsymbol{\beta_1}$

$$\frac{\partial RSS}{\partial \beta_1} = -2\left(\sum_{i=1}^{n}\frac{\partial \beta_0}{\partial \beta_1} + \Sigma_{i=1}^{n} x_i\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$= (-\Sigma_{i=1}^{n}\bar{x} + \Sigma_{i=1}^{n} x_i)\sum_{i=1}^{n}(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i) = 0$$

$$\beta_1 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$