آشنایی با مباحث تحلیل آماری

تعریف

آمار شاخه‌ای از ریاضیات است که به گردآوری، تحلیل، تفسیر، ارائه و سازمان‌دهی داده‌ها می‌پردازد. آمار را باید علم و عمل استخراج، بسط، و توسعهٔ دانش‌های تجربی انسانی با استفاده از روش‌های گردآوری، تنظیم، پرورش، و تحلیل داده‌های تجربی (حاصل از اندازه‌گیری و آزمایش) دانست.

زمینه‌های محاسباتی و رایانه‌ای جدیدتری همچون یادگیری ماشینی، و کاوش‌های ماشینی در داده‌ها، در واقع، امتداد و گسترش دانش گسترده و کهن از آمار به عهد محاسبات نو و دوران اعمال شیوه‌های ماشینی در همه‌جا می‌باشد. علم آمار، علم فن فراهم کردن داده‌های کمّی و تحلیل آن‌ها به منظور به‌دست‌آورن نتایجی که اگرچه احتمالی است، اما در خور اعتماد است.

تعریف علم آمار

علم آمار، مبتنی است بر دو شاخه آمار توصیفی و آمار استنباطی. در آمار توصیفی با داشتن تمام اعضا جامعه به بررسی خصوصیت‌های آماری آن پرداخته می‌شود در حالی که در آمار استنباطی با بدست آوردن نمونه‌ای از جامعه که خصوصیات اصلی جامعه را بیان می‌کند در مورد جامعه استباط آماری انجام می‌شود. در نظریهٔ آمار، اتفاقات تصادفی و عدم قطعیت توسط نظریهٔ احتمالات مدل‌سازی می‌شوند. در این علم، مطالعه و قضاوت معقول در بارهٔ موضوع‌های گوناگون، بر مبنای یک نمونه انجام می‌شود و قضاوت در مورد یک فرد خاص، اصلاً مطرح نیست.

کلمات کلیدی در علم آمار

- جامعه آماری
- نمونه آماری
- متغیر

ابزارهای مورد استفاده برای تهیه گزارش آماری

در آمار توصیفی از دو ابزار برای ترسیم یا گزارش ویژگی های موجود در داده ها استفاده می شود.

جداول آمار: متشکل از شاخص های آماری است.

شاخص های پراکندگی

شاخص های مرکزی

نمودار های آماری : انواع مختلفی دارد و بنا به نوع داده ها و اهداف محقق مورد استفاده قرار میگیرد.

شاخص های پراکندگی

شاخص های پراکندگی میزان پراکندگی مقادیر هر متغیر را در اطراف میانگین نشان می دهند.به وسیله این شاخص ها، می خواهیم بدانیم تا چه اندازه داده ها در اطراف نقطه تمرکز پراکنده اند.که مهمترین شاخص های پراکندگی عبارتند از:انحراف استاندارد، واریانس، ضریت تغییرات و دامنه تغییرات اشاره کرد.

دامنه تغییرات

فاصله میان بزرگترین و کوچکترین مقادیر در مجموعه دادها را اندازه گیری می کند. هر چه دامنه طولانی تر باشد، مجموعه داده ها گسترده تر است. دامنه نیز همانند میانگین تحت تأثیر داده های پرت قرار می گیرد و در چنین حالاتی یک معیار مناسب پراکندگی نیست. به علاوه، چون برای محاسبه دامنه فقط از دو اندازه بزرگترین مشاهده و کوچکترین مشاهده استفاده می شود معمولاً معیار رضایت بخشی برای پراکندگی به حساب نمی آید.

واریانس

میانگین مجذور تفاوت(انحراف)میان هر یک از مقادیر داده ها با میانگین آنها می باشد.چون تفسیر واریانس دشوار است در گزارش ها و پژوهش ها از انحراف استاندارد به جای واریانس استفاده می شود.انحراف استاندارد تفسیر میزان پراکندگی را آسان تر قابل فهم تر می سازد.

انحراف استاندارد(انحراف معیار):standard deviation

انحراف استاندارد مفیدترین ومتداول ترین شاخص پراکندگی است.مفیدبودن این شاخص به این دلیل است که با این شاخص می توان میزان پراکندگی هرتوزیع پیوسته رابرحسب واحد اندازه گیری نشان داد.این شاخص پایاترین ودقیق ترین شاخص پراکندگی است،که درمحاسبه ی آن ازکلیه ی اعداداستفاده می شود.واعمال ریاضی رامی توان درمورد آن انجام داد.این شاخص به منظورتعیین تغییرات یا پراکندگی توزیع نمره ها به کار برده می شود.ازاین شاخص می توان برای محاسبات آماری استفاده کرد.و به صورت گسترده ای درآماراستنباطی به کاربرده می شود.

محاسبه شاخص های مرکزی

کمیت هایی وجود دارند که می توانند به‌صورت کمی جامعه را معرفی نمایند بعضی از آنها محل تمرکز داده‌ها را معرفی می کنند که به آنها شاخص های مرکزی می گویند.همانطور که می دانید در محاسبات آماری لازم است که ویژگیها و موقعیت کلی داده‌ها تعیین شود. برای این منظور شاخصهای مرکزی محاسبه می‌شوند. مهم ترین شاخص های مرکزی عبارتند از : مد، میانه و میانگین که هر یک کاربرد خاص خود را دارا می‌باشند. یک شاخص مرکزی وقتی با ارزش است که دارای خواص زیر باشد:

-در محاسبه آن از تمام داده‌ها استفاده شود.

-دارای خصوصیات ساده قابل محاسبه باشد.

-به فرم ریاضی قابل محاسبه باشد.

نما یا مد

عبارت از عددی یا نمره ای که در توزیع فراوانی دارای بیشترین فراوانی است و از طریق مشاهده توزیع فراوانی و تعیین عددی که دارای بیشترین فراوانی است تعیین می شود.

موارد استفاده نما

۱- مقادیر ویژگی یا متغیر اسمی باشند.

۲- پژوهشگر علاقمند است عددی را که بیشتر تکرار شده است پیدا کند.

۳- پژوهشگر علاقمند است اطلاعاتی کلی و سریع درباره گرایشهای مرکزی بدست آورد.

میانگین

میانگین شناخته شده ترین و وسیع ترین مقدار متوسطی است که مورد استفاده قرار می گیرد و توصیف کننده مرکز توزیع فراوانی می باشد. در تحقیقاتی که مقیاس اندازه گیری داده‌ها حداقل فاصله‌ای است میانگین بهترین شاخص است.

متغیرها و داده‌ها:

متغیرهای کمّی: متغیرهایی هستند که قابل شمارش و اندازه گیری‌اند و حاصل سنجش آنها یک مقدار کمّی است

کمی پیوسته: کمیتی که بتواند بین دو مقدار خود تمامی اعداد حقیقی ممکن را بگیرد.

کمی گسسته: کمیتی که مقادیر آن شامل مجموعه شمارش پذیری از اعداد و یا زیر مجموعه ای از آن را اختیار کند

متغیرها و داده‌ها:

متغیرهای کیفی: متغیرهایی هستند که غیر قابل شمارش و اندازه گیری‌اند و حاصل سنجش آنها یک حالت و وضعیت است

کیفی اسمی

کیفی ترتیبی

همبستگی

محاسبه ارتباط میان مقادیر دو یا چند متغیر با یکدیگر

محاسبه همبستگی

ضریب همبستگی ابزاری آماری برای تعیین نوع و درجه رابطهٔ یک متغیر کمی با متغیر کمی دیگر است.

ضریب همبستگی، یکی از معیارهای مورد استفاده در تعیین همبستگی دو متغیر

مثال همبستگی

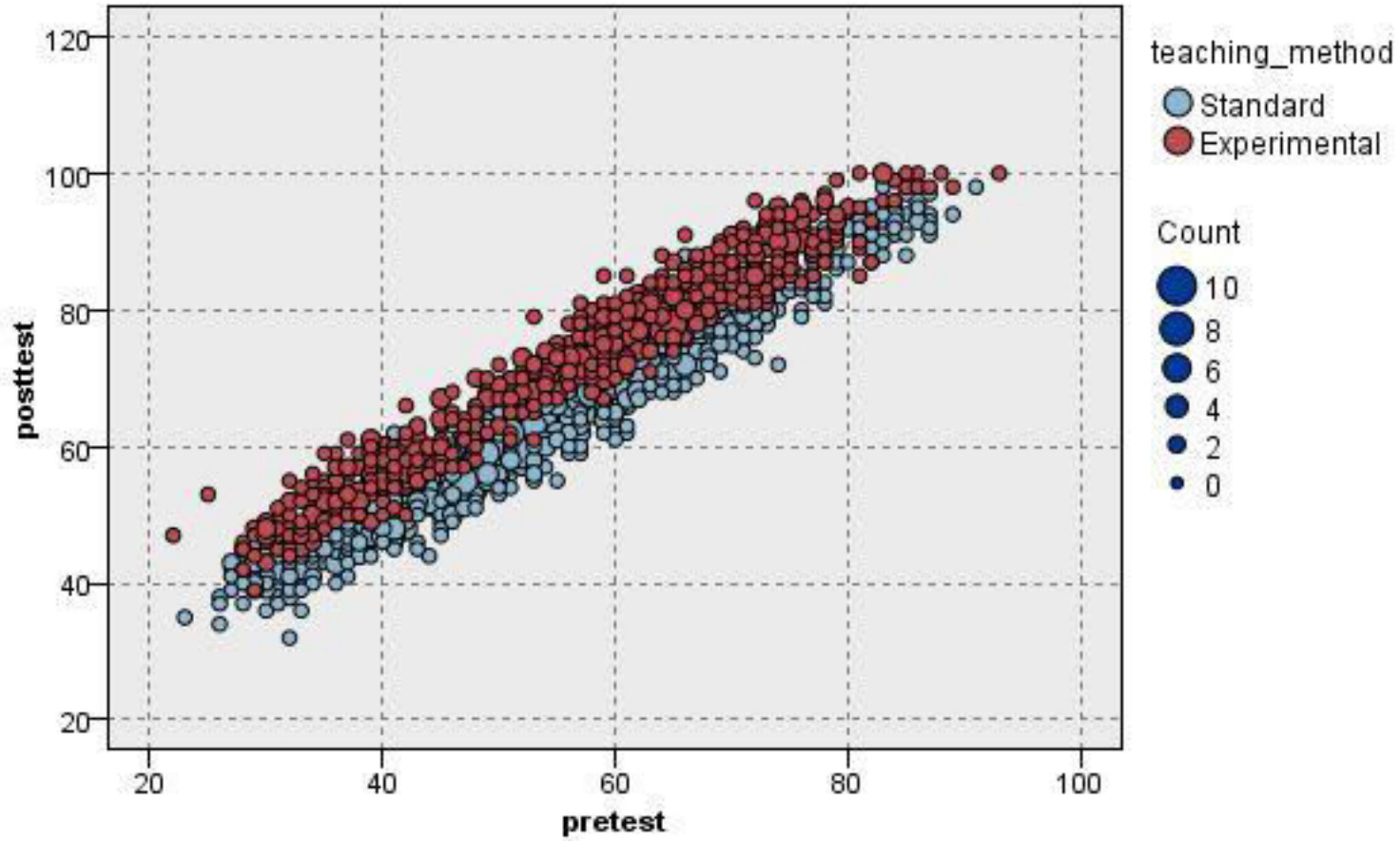
	IDNUM	Q1	Q2a	Q2b	Q2c	Q3a	Q3b	Q4a	Q4b	Q4c	Q4d	Q4e	Q4f	Q4g
IDNUM	1.000	0.005	-0.083	-0.070	-0.124	0.113	0.074	0.155	0.154	-0.009	-0.088	0.165	0.151	0.147
Q1	0.005	1.000	0.393	0.414	0.389	0.063	0.115	0.035	0.016	0.026	0.209	0.039	0.023	0.033
Q2a	-0.083	0.393	1.000	0.649	0.728	0.151	0.141	0.014	0.003	-0.005	0.193	0.007	0.001	0.001
Q2b	-0.070	0.414	0.649	1.000	0.836	0.172	0.251	0.035	0.030	0.017	0.293	0.016	0.031	0.019
Q2c	-0.124	0.389	0.728	0.836	1.000	0.280	0.327	0.025	0.018	0.010	0.273	0.003	0.011	0.008
Q3a	0.113	0.063	0.151	0.172	0.280	1.000	0.962	0.152	0.134	0.013	0.052	0.128	0.142	0.138
Q3b	0.074	0.115	0.141	0.251	0.327	0.962	1.000	0.152	0.136	0.018	0.072	0.126	0.147	0.142
Q4a	0.155	0.035	0.014	0.035	0.025	0.152	0.152	1.000	0.993	0.696	0.036	0.991	0.993	0.989
Q4b	0.154	0.016	0.003	0.030	0.018	0.134	0.136	0.993	1.000	0.705	0.046	0.992	0.995	0.992
Q4c	-0.009	0.026	-0.005	0.017	0.010	0.013	0.018	0.696	0.705	1.000	-0.043	0.703	0.695	0.700
Q4d	-0.088	0.209	0.193	0.293	0.273	0.052	0.072	0.036	0.046	-0.043	1.000	0.060	0.043	0.058
Q4e	0.165	0.039	0.007	0.016	0.003	0.128	0.126	0.991	0.992	0.703	0.060	1.000	0.993	0.992
Q4f	0.151	0.023	0.001	0.031	0.011	0.142	0.147	0.993	0.995	0.695	0.043	0.993	1.000	0.995
Q4g	0.147	0.033	0.001	0.019	0.008	0.138	0.142	0.989	0.992	0.700	0.058	0.992	0.995	1.000

تحلیل آماری چند متغیره

محتوای فایل

1. نمودار پراکندگی (Scatter plot)
2. Crosstab
3. توابع aggregation
4. آشنایی با مفهوم Group by
5. همبستگی
6. ماتریس همبستگی

نمودار پراکندگی



Crosstab برای تعداد سفارشات

Market					
Segment	Africa	Asia Pacific	Europe	LATAM	USCA
Consumer	2381	7362	6061	5321	5393
Corporate	1312	4321	3613	3053	3130
Home Office	894	2619	2055	1920	1855

Aggregation توابع

- Mean
- Sum
- Max
- Min
- Count
- Median

Crosstab با در نظر گرفتن یک تابع میانگین برای تعداد کالاها در هر سفارش

	Market				
Segment	Africa	Asia Pacific	Europe	LATAM	USCA
Consumer	2.311	3.424	3.567	3.731	3.704
Corporate	2.243	3.330	3.571	3.766	3.780
Home Office	2.369	3.438	3.599	3.736	3.720

Group By

تابع Group By برای بررسی خصوصیات مقادیر یک ویژگی کیفی (Categorical) جهت آشنایی دقیق‌تر مطابق با مقادیر ویژگی‌های کمی موجود در یک دیتاست است.

مثال:

	Market	Quantity_Mean	Discount_Mean	Shipping Cost_Mean
1	USCA	3.730	0.150	23.844
2	Asia Pacific	3.398	0.181	30.615
3	Europe	3.574	0.091	29.747
4	Africa	2.303	0.157	19.368
5	LATAM	3.743	0.136	22.831

تحلیل آماری تک متغیره

فرآیند داده کاوی

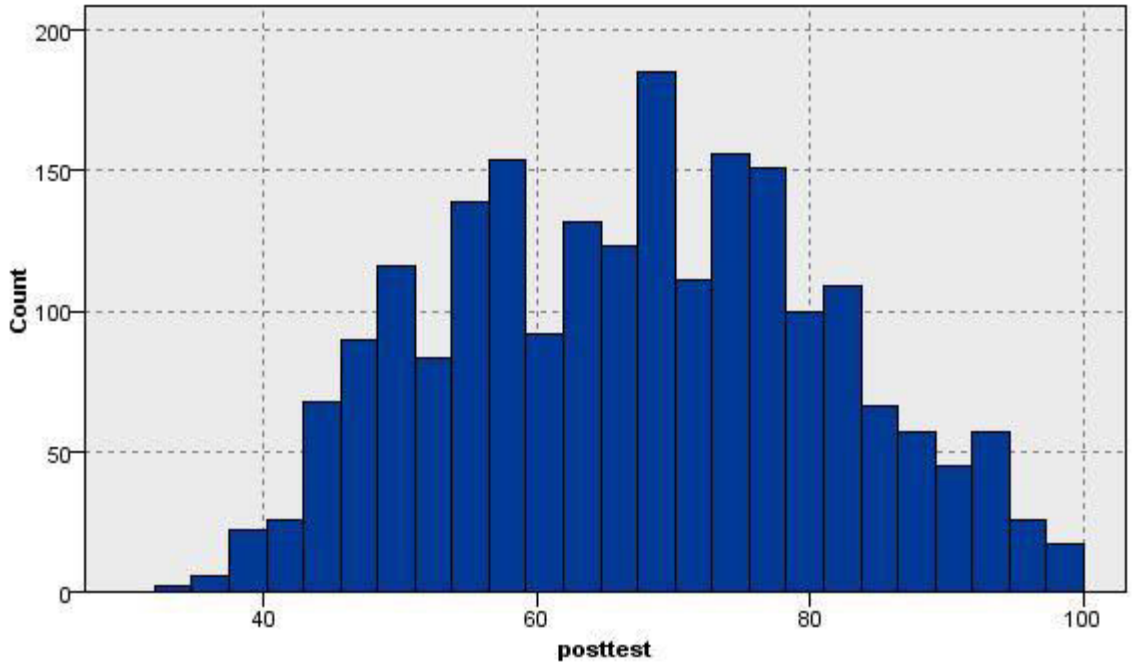
محتوای فایل

1. استفاده از تابع Group by برای محاسبه فراوانی یک متغیر کیفی
2. استفاده از نمودار هیستوگرام برای محاسبه فراوانی یک متغیر کمی

استفاده از تابع Group by برای محاسبه فراوانی یک متغیر کیفی

Value	Proportion	%	Count
1.000		74.17	1582
2.000		25.83	551

استفاده از نمودار هیستوگرام برای محاسبه فراوانی یک متغیر کمی



آزمون های آماری

محتوای فایل آموزشی

1. آزمون‌های میانگین
2. آزمون‌های ناپارامتری

آزمون‌های میانگین

1. مفهوم Significance
2. آزمون میانگین یک جامعه
3. آزمون مقایسه میانگین دو جامعه
4. آزمون زوجی
5. آزمون مقایسه میانگین چند جامعه (ANOVA)

مفهوم Significance

معنی داری که به اختصار آن را با sig نشان می‌دهیم، نشان دهنده خطا رخ داده در رد کردن فرضیه صفر (H0) است.

نام دیگر Signifcance، P-value است. هر چقدر مقدار P-value کمتر باشد (به صفر نزدیک‌تر باشد)، رد آن راحت‌تر است.

آلفا : سطح خطایی است که توسط ما در نظر گرفته می‌شود.

Sig < a ----> H0 : رد

Sig >= a ----> H0 : رد نمی‌شود

آزمون میانگین یک جامعه

One – sample T-test، آزمونی است که برای بررسی ادعای مربوط با میانگین یک متغیر به کار می‌رود.

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات تولید شده در باغ وی طی سال گذشته 10 تن بوده است.

H0 : میانگین = 10

H1 : میانگین <>10

آزمون مقایسه میانگین دو جامعه

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات تولید شده در باغ وی طی سال گذشته و امسال با هم برابر است.

H0 : میانگین وزن امسال = میانگین وزن سال قبل

H1 : میانگین وزن امسال <> میانگین وزن سال قبل

آزمون مقایسه میانگین دو جامعه مستقل

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات گوجه فرنگی و هویج تولید شده وی سال با هم برابر است.

H0 : میانگین وزن هویج =میانگین وزن گوجه

H1 : میانگین وزن هویج <> میانگین وزن سال گوجه

آزمون مقایسه میانگین چند جامعه (ANOVA)

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات گوجه فرنگی، پیاز و هویج تولید شده وی سال با هم برابر است.

H0 : میانگین وزن پیاز = میانگین وزن هویج =میانگین وزن گوجه

H1 : حداقل یکی از آن‌ها فرق کند

تعریف آزمون ناپارامتری

آزمون‌های ناپارامتری

- آزمون دو جمله‌ای
- آزمون علامت زوج – نمونه ای
- آزمون مک نمار
- آزمون کوکران
- آزمون من – ویتنی (آزمون U)
- کروسکال – والیس
- آزمون فریدمن

معادل آزمون های پارامتریک (میانگین) در آزمون‌های نا پارامتریک

آزمون پارامتریک	آزمون ناپارامتریک
میانگین یک جامعه (t-test)	دو جمله‌ای
مقایسه میانگین دو جامعه مستقل (t-test مستقل)	U
مقایسه میانگین چند جامعه	H (کروسکال والیس)

آزمون دو جمله‌ای

برای بررسی نسبت خاصی در جامعه به کار میرود:

مثال: بررسی درستی ادعای برابری نسبت کارشناسان و رئیس اداره های یک شرکت بیمه:

H0 : P= 50%

H1 : P <> 50%

آزمون علامت – زوج نمونه‌ای (ویل کاکسون)

مانند آزمون مقایسه میاگنین دو جامعه وابسته است که داری توزیع نرمال نمی‌باشند یا نمونه های موجود در دو جامعه کوچک است.

مثال: بررسی ادعای برابری حجم تولیدات سال گذشته با امسال:

H0 : mean= 0

H1 : mean <> 0

آزمون مک نمار

مثال: بررسی اینکه آیا تفاوت معناداری بین نظر دانشجویان قبل از ورود به دانشگاه و بعد از ورود به دانشگاه وجود دارد یا خیر؟

توضیحات	موافق	مخالف
نظر قبل از ورود به دانشگاه	181	19
نظر بعد از ورود به دانشگاه	50	150

H0 : mean= 0

H1 : mean <> 0

آزمون کوکران

مثال: بررسی اینکه آیا تفاوت معناداری بین نظر دانشجویان قبل از ورود به دانشگاه ، حین تحصیل و بعد از ورود به دانشگاه وجود دارد یا خیر؟

توضیحات	موافق	مخالف
نظر قبل از ورود به دانشگاه	181	19
حین تحصیل	120	80
نظر بعد از فارغ التحصیلی	50	150

H0 : mean= 0

H1 : mean <> 0

آزمون U

مانند آزمون مقایسه میاگنین دو جامعه مستقل است با این تفاوت که فرض نرمال بودن توزیع لازم نیست (توزیع نرمال ندارند.)

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات گوجه فرنگی و هویج تولید شده وی سال با هم برابر است.

H0 : میانگین وزن هویج =میانگین وزن گوجه

H1 : میانگین وزن هویج <> میانگین وزن سال گوجه

آزمون کروسکال والیس (آزمون H)

مثال: فرض کنید کشاورزی ادعا میکند میانگین وزن محصولات گوجه فرنگی، پیاز و هویج تولید شده وی سال با هم برابر است.

H0 : میانگین وزن پیاز = میانگین وزن هویج =میانگین وزن گوجه

H1 :حداقل یکی از آن‌ها فرق کند

نمونه گیری

محتوای فایل

- تعریف نمونه گیری به زبان ساده
- مزایا و معایب نمونه گیری در داده‌کاوی
- انواع روش نمونه گیری کاربردی در داده کاوی

تعریف نمونه گیری به زبان ساده

به جای پردازش و تحلیل کل دیتای موجود در دیتاست، زیر مجموعه از آن‌ها را برای ساخت مدل پیش بینی کننده استفاده کنیم.

مزایا نمونه گیری

- امکان افزایش دقت پیش بینی بر روی داده‌های جدید
- افزایش سرعت پردازش و تحلیل توسط ماشین

عیب نمونه گیری

- شناسایی اشتباه الگوهای پنهان و مؤثر در پیش بینی

آشنایی با روش‌های نمونه گیری پر کاربرد در داده کاوی

- نمونه گیری تصادفی ساده
- نمونه گیری طبقه‌ای

سری‌های زمانی

سری زمانی چیست؟

سری زمانی، مجموعه‌ای از داده‌ها است که بر حسب زمان یا متغیر دیگزی مرتب شده باشد.

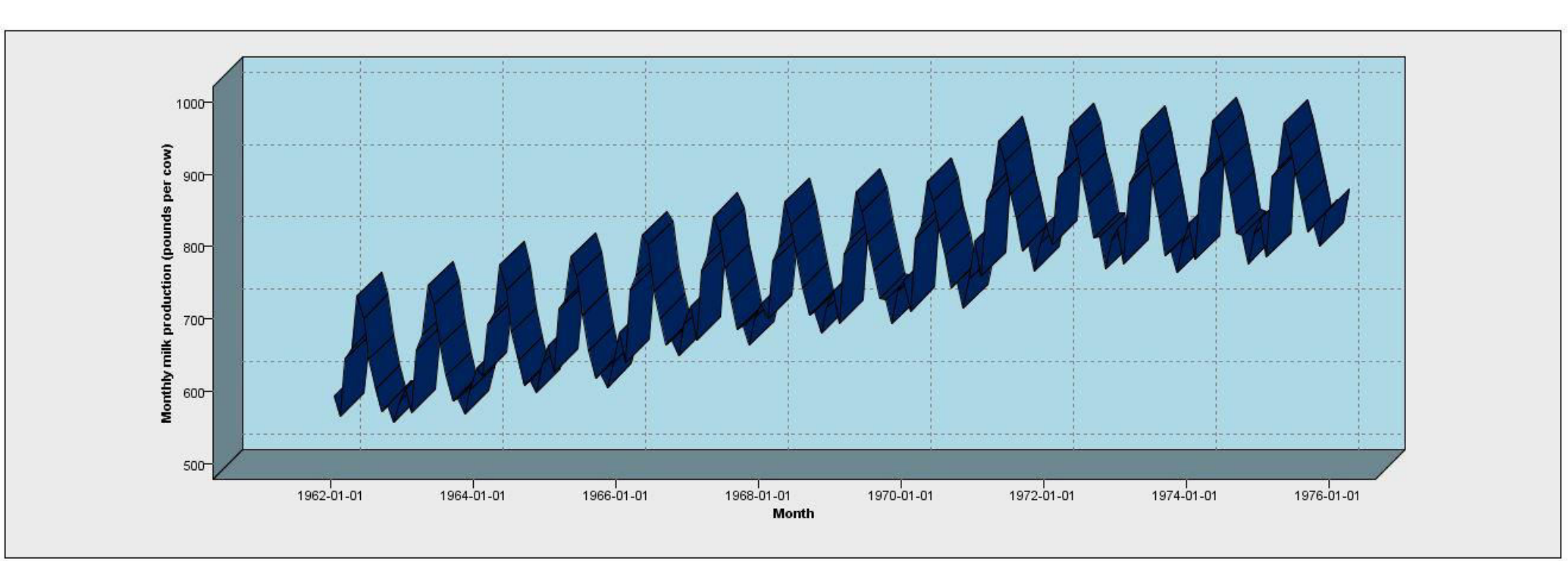
در سری‌های زمانی، دو فرآیند از اهمیت بسیار زیادی برخوردار هستند:

1) فرآیند Fitting (برازش)

2) فرآیند Forcasting (پیش بینی)

در فرآیند برازش از دیتای گذشته برای ساخت مدل استفاده می‌شود ولی فرآیند پیش بینی برای پیش بینی مقادیر نامعلوم با استفاده از مدل سری زمانی در بازه های زمانی آینده است.

وضعیت تولید شیر



مدل ARIMA

در آمار و اقتصادسنجی و به ویژه در آنالیز سری‌های زمانی، یک **میانگین متحرک خودهمبسته یکپارچه** (ARIMA) یک مدل گسترده‌تر از میانگین متحرک خودهمبسته (ARMA) است. این مدل‌ها در سریهای زمانی برای فهم بهتر مدل یا پیش‌بینی آینده به کار می‌روند. این مدل‌ها در جایی که داده‌ها غیر ایستا (non-stationary) باشند به کار می‌روند.

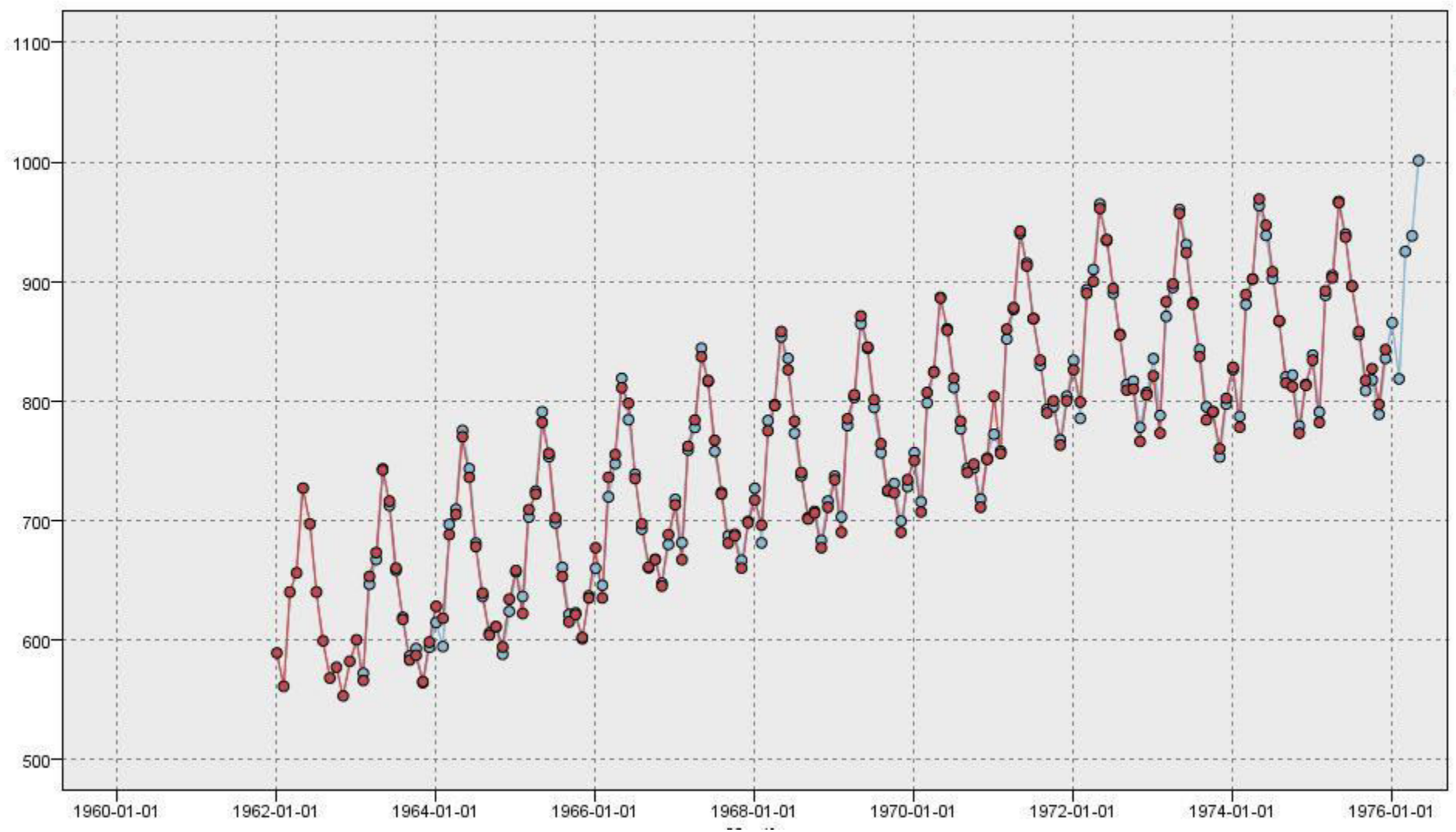
این مدل در اکثر موارد به صورت ARIMA(P,Q,D) نشان داده می‌شود که در آن p و d و q اعداد حقیقی غیرمنفی هستند که درجه خودهمبستگی، یکپارچگی و میانگین متحرک را معلوم می‌کنند.

شاخص های نیکویی برازش

- میانگین
- خطای معیار (SE)
- R2 ایستایی
- میانگین مجذور خطا (RMSE)
- میانگین قدر مطلق درصد خطا (MAPE)
- ماکزیمم قدر مطلق درصد خطا (MAX PE)
- میانگین قدر مطلق خطا (MAE)
- ماکزیمم قدر مطلق خطا (MAX AE)
- معیار اطلاعاتی نرمال شده بیز (Normalized BIC)

اهمیت کاهش ابعاد

- کاهش پچیدگی فرآیند داده کاوی
- کاهش حجم دیتاست
- افزایش سرعت پردازش
- افزایش دقت پیش بینی مدل


1100
1000
900
800
700
600
500
1960-01-01
1962-01-01
1964-01-01
1966-01-01
1968-01-01
1970-01-01
1972-01-01
1974-01-01
1976-01-01

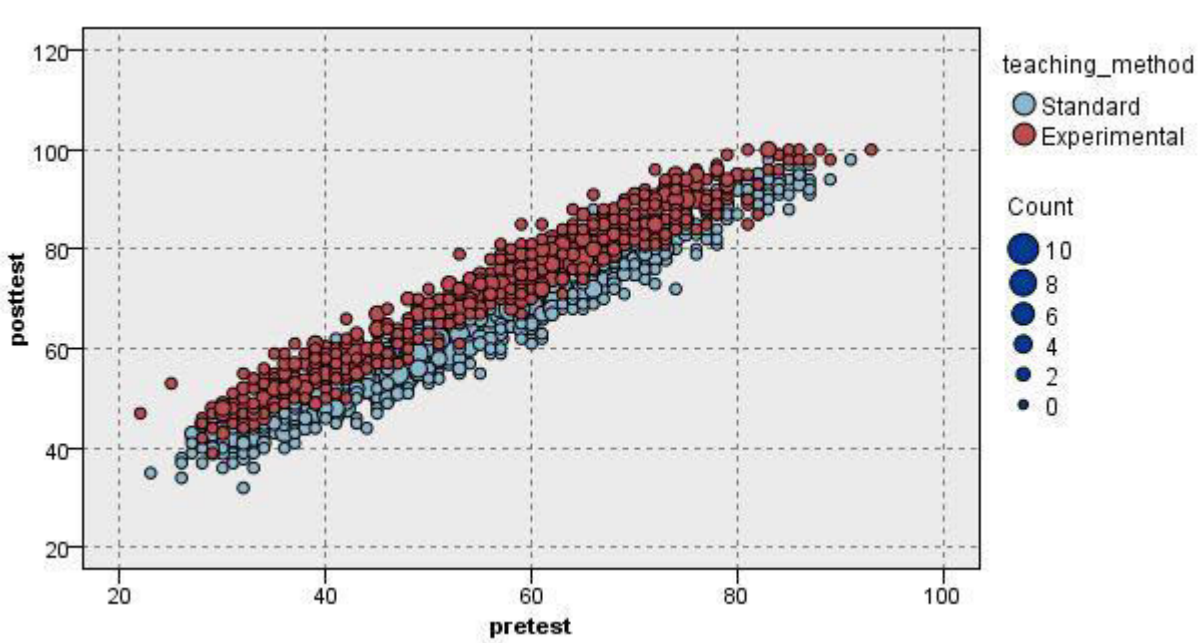
رگرسیون خطی

رگرسیو خطی چیست؟

رگرسیون نوعی روش یادگیری ماشین است که در بخش یادگیری نظارتی (Supervised learning) قرار می‌گیرد. به کمک رگرسیون، می‌تواند مقادیر یک متغیر وابسته (Dependent variable) به کمک یک یا چند متغیر مستقل (Independent variable) پیش بینی می‌شود.

$$Y^ = a + bx$$

رابطه خطی


posttest
120
100
80
60
40
20
20
40
60
80
100
pretest
teaching_method
Standard
Experimental
Count
10
8
6
4
2
0

مفاهیم مورد نیاز در رگرسیون خطی

R Square

خطای معیار تخمین

فرضیه آماری آزمون معنا داری کل مدل رگرسیون (به کمک جدول Anova)

مقدار ثابت (B0)

مقدار ضریب متغیر مستقل (B1)

فاصله اطمینان

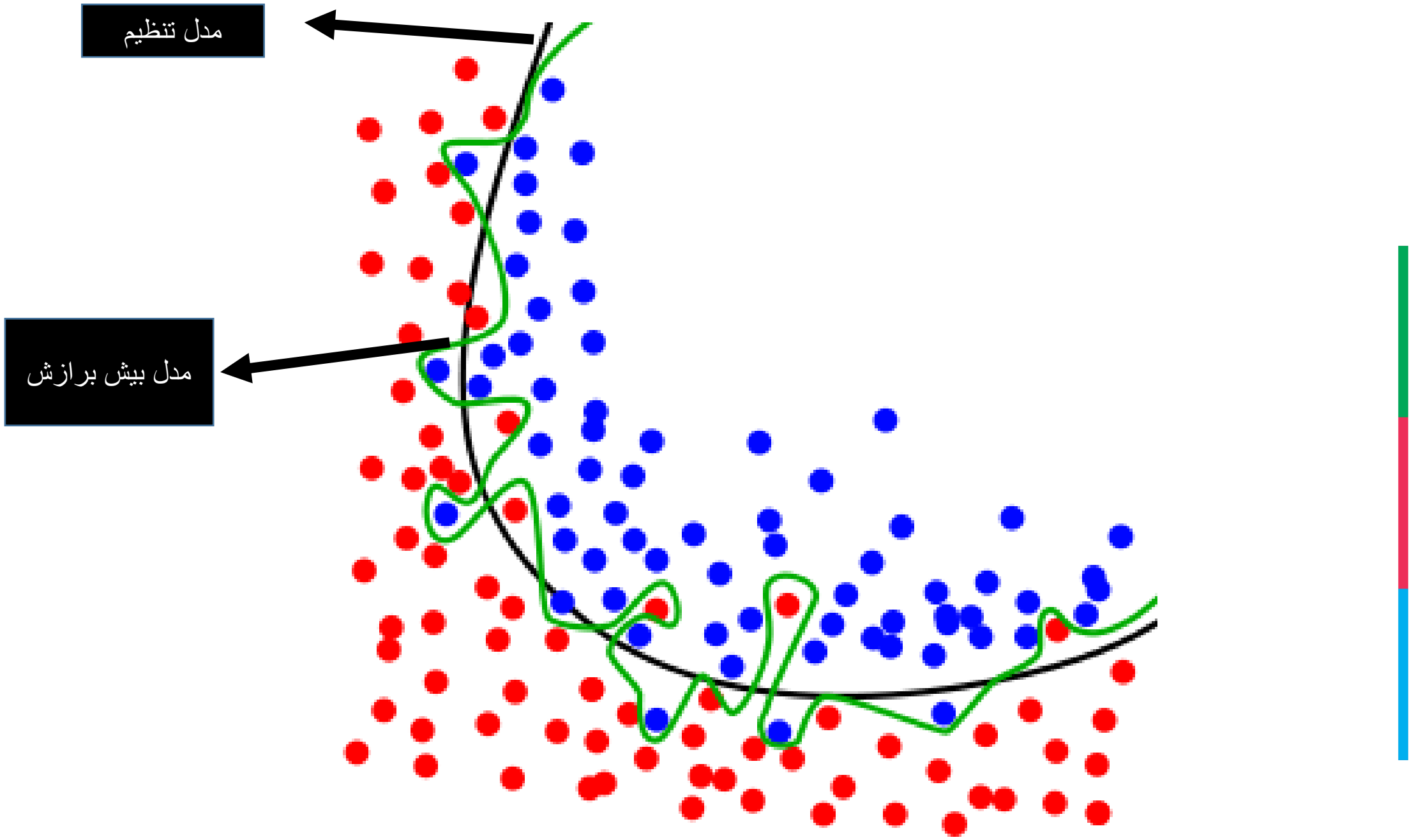
Regularization

تعریف Regularization: به فرآیند تولید با هذف جلوگیری از بیش برازش مدل (Over fitting) است.

بیش‌برازش Overfitting به پدیدهٔ نامطلوبی در آمار گفته می‌شود که در آن درجه آزادی مدل بسیار بیشتر از درجه آزادی واقعی انتخاب شده و در نتیجه اگرچه مدل روی داده استفاده شده برای یادگیری بسیار خوب نتیجه می‌دهد، اما بر روی داده جدید دارای خطای زیاد است.

برای مقابله با این پدیده از دور روش می‌توان استفاده کرد:

1) روش تنظیم کردن (Regularization)
2) روش Cross validation


مدل تنظیم
مدل بیش برازش

روش‌های مورد استفاده برای تنظیم کردن

نرم L1

L1- norm

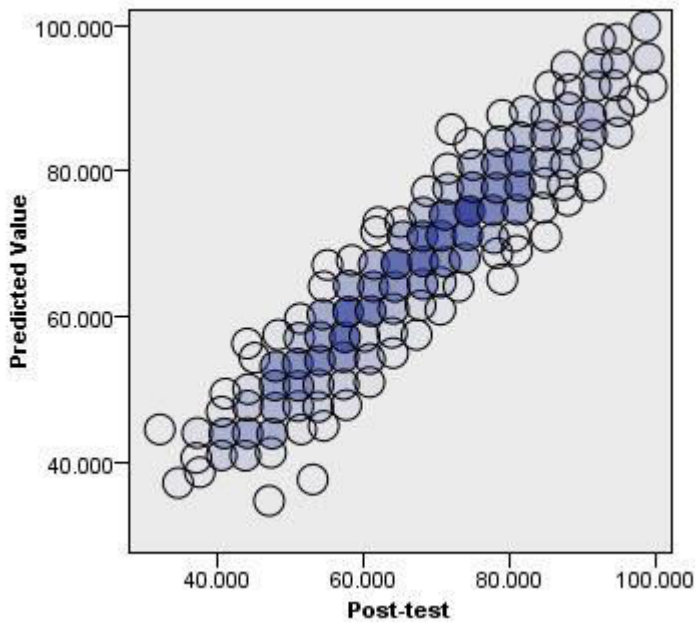
$$S = \sum_{i=1}^{n} |y_i - f(x_i)|.$$

نرم L2

L2 - norm

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

پیش بینی مقادیر متغیر هدف یا وابسته



تحلیل نهایی مدل

Model Term	Coefficient ▶	Sig.	Importance
Intercept	13.212	.000	
pretest_transformed	0.981	.000	1.000

نمره POST TEST

13.212 + 0.981* PRETEST