رگرسیون خطی ساده :

اگر دو یا چند متغیر مستقل از هم نباشند و بین آنها رابطه‌ای وجود داشته باشد ارتباط بین متغیرها را با آزمون‌های مختلفی می‌توان بررسی کرد، اما شدت و میزان این ارتباط و جهت آنها را بایستی با آزمون رگرسیون تعیین گردد. وجود رابطه آماری بین دو متغیر بیشتر از همه در نموداری به نام نمودار پراکنش نمایان می‌شود.

رگرسیون به معنی تعیین روابط نادقیق بین متغیرهای آماری و تحلیل این روابط می باشد. تحلیل رگرسیون به دلایل زیر اهمیت دارند:

۱. بررسی چگونگی تاثیر متغیرها بر یکدیگر

۲. تعیین رابطه موجود بین متغیرها جهت کاربرد در تخمین های مشابه در آینده

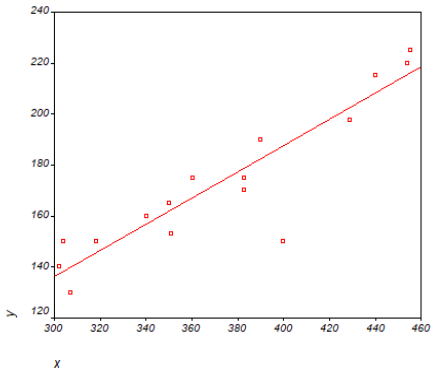
جهت تحلیل رگرسیون مراحل زیر به ترتیب انجام می گیرد:

در این بخش فرض می شود متغیر Y به عنوان متغیر پاسخ یا وابسته و X به عنوان متغیر کنترل و یا مستقل مورد مطالعه پژوهشگر است که هدف یافتن یک مدل ریاضی بین این دو متغیر براساس مشاهدات حاصل از یک نمونه تصادفی از این دو متغیر می باشد. برای رسیدن به این هدف باید مراحل زیر انجام شود.

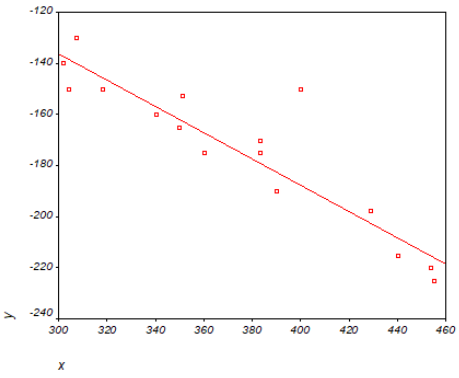
- رسم نمودار پراکنش داده ها
- محاسبه ضریب همبستگی خطی پیرسون
- برازش مدل مناسب
- آزمون فرض معنی داری مدل
- تحلیل باقیمانده ها

- **رسم پراکنش داده ها:** اگر هیچگونه اطلاعی از رابطه بین دو متغیر وجود نداشته باشد، بهترین کار رسم نمودار پراکنش داده ها می باشد که تا حدودی رابطه موجود را نشان می دهد. برای بررسی همبستگی یا رابطه بین دو متغیر ابتدا لاًزم است که مقادیر آن

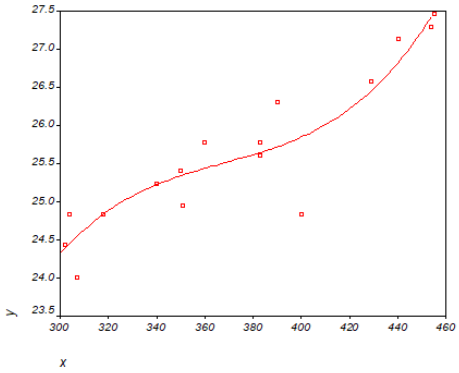
دو صفت کمی‌را در یک نمودار رسم کنیم. معمولاً در یک نمودار اگر مقادیر متغیر مستقل روی محور طول‌ها و مقادیر متغیر وابسته روی محور عرض‌ها رسم گردد، به آن نمودار پراکنش گویند. وقتی با استفاده از نمودار پراکنش به بررسی رابطه بین دو متغیر می‌پردازیم، نمودار های زیر ممکن است به وجود آید.



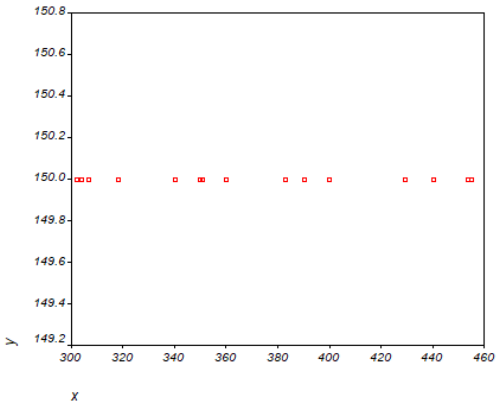
الف-ضریب همبستگی مثبت



ب- ضریب همبستگی منفی



ج- ضریب همبستگی مثبت



د- ضریب همبستگی صفر

محاسبه ضریب همبستگی خطی:

در محاسبه ضریب همبستگی خطی نقش اصلی را کوواریانس نمونه دارد بنابراین ابتدا این شاخص تعریف و تفسیر می شود و سپس ضریب همبستگی خطی تعریف و تفسیر خواهد شد. فرض کنید نمونه‌ای تصادفی به حجم n از دو متغیر X و Y را با نماد

$(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ نشان دهیم در این صورت کواریانس نمونه‌ای بین دو متغیر به صورت زیر محاسبه می‌شود.

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- ✓ کواریانس جهت همبستگی بین دو متغیر را نشان می‌دهد ولی با دانش از مقدار آن نمی‌توان در مورد شدت همبستگی تصمیم‌گیری نمود. اگر کواریانس دو متغیر مثبت باشد نشان دهنده همبستگی مستقیم (مثبت) بین دو متغیر و اگر مقدار آن منفی باشد، نشان دهنده همبستگی معکوس بین دو متغیر می‌باشد. واضح است که اگر کواریانس بین دومتغیر صفر شود، آن دو متغیر ناهمبسته خواهند بود.
- ✓ ضریب همبستگی بیان کننده ارتباط خطی بین دو متغیر است.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - nx^{-2})(\sum y_i^2 - ny^{-2})}} = \frac{S_{xy}}{S_x S_y}$$

که در آن S_x انحراف معیار متغیر کنترل و S_y انحراف معیار متغیر پاسخ هستند.

- مقدار r تقریباً برابر ۱ بیان می‌کند که یک رابطه خطی کامل بین متغیرهای کنترل و پاسخ با شیب مثبت وجود دارد.
- مقدار r تقریباً برابر ۱- بیان می‌کند که یک رابطه خطی کامل بین متغیرهای کنترل و پاسخ با شیب منفی وجود دارد.
- مقدار r تقریباً برابر صفر بیان می‌کند که هیچگونه رابطه خطی بین متغیرهای پاسخ و کنترل برقرار نیست.

برای معنی داری ضریب همبستگی بایستی فرضیات زیر مورد آزمون قرار می گیرند:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

که در آن ρ ضریب هبستگی جامعه است. $RH_0 \Leftrightarrow Sig < \alpha$

رد فرضیه صفر بیان می کند که یک رابطه خطی معنی دار بین متغیر وابسته و مستقل برقرار است.

- **برازش مدل مناسب:**

وقتی که وجود رابطه بین متغیرها تأیید شد، بایستی مدلی را برازش داد. منظور از برازش مدل مناسب تعیین رابطه موجود(رگرسیون بین متغیرهای کنترل (X) و پاسخ (Y)) می باشد. در صورتی که رابطه فقط شامل یک متغیر کنترل باشد، تحت عنوان معادله رگرسیون ساده و اگر دو یا بیش از دو متغیر کنترل را شامل شود به عنوان معادله رگرسیون چند متغیره شناخته می شود. رگرسیون اعم از ساده یا چند متغیره به شکل خطی و یا غیر خطی تعیین می شود.

جهت نوشتن معادله رگرسیون نیاز به برآورد پارامترهای معادله می باشد (α , β):

مدل خطی ساده:

$$y_i = \alpha + \beta x_i + e_i$$

برازش معادله فوق به n مشاهده زوج $(x_1, y_1),(x_2, y_2),.....(x_n, y_n)$ **به روش کمترین مربعات خطا به صورت زیر انجام میشود. در این روش مسئله مینمم کردن کمیت زیر به عنوان یک معیار دقت (یعنی کمترین خطا) می باشد. با قیمانده ها که تفاوت مقدار مشاهده شده با مقدار برازش شده را نشان میدهد عبارتند از:**

$$e_i = y_i - \hat{y}_i, \quad i = 1,2,.....n$$

مجموع مربعات خطا $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ **که درآن**

$$\hat{y}_i = a + bx_i \ , i = 1,2,.....,n$$

معادله خط کمترین مربعات خطا میباشد. و ضرایب از روابط زیر تعیین میشوند.

$$\hat{\beta} = b = \frac{SS_{xy}}{SS_x}, \ \hat{\alpha} = a = \bar{y} - b\bar{x}$$

که در آن

$$\hat{\beta} = \frac{SS_{xy}}{SS_x} \quad \text{و} \quad SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{و} \quad SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y}) \ (x_i - \bar{x})$$

ضریب تعیین:

✓ کمیتی است که میزان دقت مدل برازش داده شده را اندازه گیری و بیان می نماید. مقادیر نزدیک به ۱ دقت بالا و مقادیر نزدیک صفر دقت پایین مدل را نشان میدهد.

در تحلیل رگرسیون تغییرات کل سهم خطا از تغییرات کل یا تغییرات بیان نشده و سهم مدل از تغییرات کل یا تغییرات بیان شده به ترتیب با نمادهای $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$ و $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ و $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ نشان داده میشود. قابل توجه است که رابطه زیر بین این سه کمیت همواره برقرار است.

$$S_y = SS_R + SS_E$$

و در آن $i = 1,2,.....,n$, $e_i = y_i - \hat{y}_i$ باقیمانده ها هستند که نماینده خطاها در تحلیل رگرسیون هستند.

$$R^2 = \frac{SS_R}{SS_y} = 1 - \frac{SS_E}{SS_y}$$

بدیهی است که $0 \le R^2 \le 1$. علاوه بر این فرمول محاسباتی ضریب تعیین عبارت است از. $R^2 = r^2$.

آزمون معنی دار مدل:

برای آزمون معنی داری مدل فرضیه های آماری زیر که فرضیه های معنی داری مدل معروف هستند باید مورد بررسی قرار گیرد.

$$\boldsymbol{H_0: \beta = 0}$$
$$\boldsymbol{\underline{H_1: \beta} \neq 0}$$

با توجه به رابطه $r = \hat{\beta}\sqrt{\frac{SS_y}{SS_x}}$ بین شیب معادله کمترین مربعات خط و ضریب همبستگی خطی نمونه مشاهده می شود که: $r = 0 \Leftrightarrow \hat{\beta} = 0$

این رابطه ارتباط مستقیم شیب خط کمترین مربعات خطا و ضریب همبستگی پیرسن را بیان می کند که به دلیل همین رابطه ضریب همبستگی خطی پیرسن نامیده می شود.

T-Test : مـدل معنی دار است اگر و تنها اگر

$$RH_0 \Leftrightarrow |t| > t_{\frac{\alpha}{2}}(n-2)$$
$$RH_0 \Leftrightarrow Sig < \alpha$$

F-Test

برای انجام از این آزمون از جدول ANOVA تحلیل واریانس که به صورت زیر رسم می شود استفاده می کنیم.

$$RH_0 \Leftrightarrow F > F_{\alpha}(n-2)$$
$$RH_0 \Leftrightarrow Sig < \alpha$$

	SS	df	MS	F	Sig
Model	**SSR**	۱	**MSR=** $\frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	**p**
Error	**SSE**	**n-2**	**MSE=** $\frac{SSE}{n-2}$		
Total	**SSY**	**n-1**			

برای اینکه بدانیم مدل نیاز به عرض از مبدأ دارد، یا خیر، بایستی فرض زیر مورد آزمون قرار گیرد.

$$\boldsymbol{H_0: \alpha = 0}$$
$$\boldsymbol{H_1: \alpha \neq 0}$$

T-Test : وجود عرض از مبدا در مدل الزامی است اگر و تنها اگر

$$RH_0 \Leftrightarrow |t| > t_{\frac{\alpha}{2}}(n-2)$$
$$RH_0 \Leftrightarrow Sig < \alpha$$

ولی هنوز تحلیل پایان نیافته و باید درستی شرایط و یا پذیره های اصلی به شرح ذیل باید تایید و یا رد شود.در صورت تایید تمام این پذیره ها مدل برازش شده معتبر و قابل استفاده خواهد بود.

- **آزمون‌های درستی تشخیص(آنالیز باقیمانده ها):**

باقیمانده‌های $e_1, e_2, \ldots, e_n$ که از رابطه $e_i = y_i - \hat{y}_i,\ i = ۱,۲,...,n$ محاسبه می شوند

نقش یک نمونه تصادفی از خطا را ایفا می‌کنند، بنابراین برای آزمون فرضیه‌های اساسی از مقادیر e_i استفاده کرده و آزمون‌های مربوط را انجام می‌دهیم.

۱- نرمال بودن خطاها: برای تشخیص نرمال بودن خطاها نمودارهای زیر را رسم می‌کنیم.

هیستوگرام باقیمانده‌ها: در صورت نرمال بودن توزیع مانده ها، هیستوگرام توزیع نرمال تبعیت می‌کند.

نمودار احتمال نرمال (NPP): در صورت صحیح بودن فرض توزیع نرمال، داده ها روی یک خط راست قرار می‌گیرند.

۲- آزمون میانگین صفر و واریانس ثابت

$$E(\varepsilon) = 0$$

$$Var(\varepsilon) = \sigma^2$$

- رسم نمودار پراکنش مقادیر پیش بینی استاندارد شده در مقابل مانده های استاندارد شده: چنانچه نمودار پراکنش شکل و روند خاصی نشان ندهد، فرض میانگین صفر و واریانس ثابت تایید می شود.

- رسم نمودار پراکنش مقادیر پیش بینی استاندارد شده در مقابل مانده های حذفی: برای باقیمانده های با میانگین صفر و واریانس ثابت، نمودار شکل و روند خاصی نشان نمی دهد.

- رسم نمودار پراکنش مقادیر پیش بینی استاندارد شده در مقابل مانده های استیودنت شده حذفی: پراکنش بدون شکل و روند خاص موید فرض میانگین صفر و واریانس ثابت می باشد.

داده‌های پرت: داده‌هایی که باقیمانده استاندارد شده آنها خارج از بازه ۲- و ۲ باشد، داده پرت و اگر خارج از بازه ۳- و ۳ قرار گیرند، داده‌های بسیار پرت نامیده می‌شوند. باید دلیل وجود داده های پرت بررسی شود. وقتی نوع عامل مشخص شد، آنگاه می‌توان تصمیم گرفت که داده‌های پرت بایستی حذف شوند، یا نه.

مدلهای غیر خطی:

فرض وجود رابطه خطی بین متغیر پاسخ و کنترل نقطه شروع در تحلیل رگرسیون میباشد در صورتیکه در بسیاری از مسایل مورد تحقیق یک رابطه غیر خطی برقرار می باشد که این موضوع را با رسم نودار پراکنش می توان تشخیص داد. برای برازش مدلهای غیر خطی علاوه بر مدلهای موجود در نرم افزار ها با استفاده از تبدیلهای مناسب ابتدا مدل به یک مدل غیر خطی تبدیل شده سپس تحلیل انجام می شود.

در جدول زیر برخی از مدلهای غیر خطی پر کاربرد که با تبدیل به یک مدل خطی قابل برازش به داده ها می باشند ارائه شده است.

	مدل خطی جدید	تبدیل مناسب	مدل غیر خطی
۱	$y' = \alpha' + \beta x$	$y' = \ln(y),\ \alpha' = \ln(\alpha),$	$y = \alpha.\exp(\beta x)$
۲	$y' = \alpha' + \beta' x$	$y' = \ln(y),\ \alpha' = \ln(\alpha), \beta' = \ln(\beta)$	$y = \alpha\beta^{x}$
۳	$y' = \alpha' + \beta x'$	$y' = \ln(y),\ \alpha' = \ln(\alpha), x' = \ln(x)$	$y = \alpha x^{\beta}$
۴	$y' = \alpha + \beta x$	$y' = \frac{1}{y}$	$y = \frac{1}{\alpha + \beta x}$
۵	$y = \alpha + \beta x'$	$x' = \ln(x)$	$y = \alpha + \beta \ln(x)$
۶	$y = \alpha + \beta x'$	$x' = f(x)$	$y = \alpha + \beta f(x)$

Simple Linear Regression:

-Scatter Plot:

Step 1. Enter your data in two variables: x **and y.**

Step 2. Choose **Graph**s in menu------**Interactive**------**transfer x into X Axis and y into Y Axis------OK .**

- Correlation Coefficient and Estimated Regression Line:

Step 1 Choose Analyze -----Correlate-----Bivariate.

Step 2. In dialog box of bivariate, transfer variables **x** and **y** into **variables** then check Pearson, ----two tailed---OK.

Step3. Choose **Analyze----Regression----Linear**. In dialog box of **Linear transfer x and y into independents and dependent respectively**. **OK.**

Nonlinear Regression: -

Step 1. **Choose analyze----Regression----Curve estimations.**

Step 2. In this dialog box transfer x and y into independent and dependent respectively -----choose your models in this box ----check Display ANOVA tables---OK.

Residuals analyze:

Select Plots option in dialog box of Linear ---Select your variables and check Histogram and NPP.

مثال ۱-: یکی از موارد مورد مطالعه در بحث لرزه‌خیزی هر ناحیه ای بررسی رابطه بین بزرگای زمین لرزه و پارگی گسل می‌باشد که تا کنون روابط مختلفی بدین منظور ارائه شده است. بدین منظور داده‌های ۲۲ زمین لرزه که طول گسیختگی گسل عامل آنها در دست می‌باشد تا رابطه‌ای بین این پارامترها بدست آید.

M	Fr
۷.۶	۲۱
۷	۳۰
۷.۷	۷۰
۷.۷	۷۵
۷	۷۵
۶.۶	۱۸
۷.۴	۶۷
۶.۷	۱۷
۷.۳	۶۷
۷.۲	۳۳
۶.۴	۳۰
۶.۸	۲۳
۶.۹	۳۰
۶.۶	۴۸
۷.۲	۸۲
۷.۴	۸۰
۵.۸	۹
۷.۴	۹۲
۷.۱	۶۰

دکتر محمد امینی

۶۵	۷.۱
۷۵	۷.۳
۴	۶

مثال ۲- در نمونه برداری از چاههای آب دشت مشهد در سال ۱۳۸۹ غلظت سدیم و کلر برحسب میلی اکی والان گرم(جدول ۵-۱) اندازه گیری شده است. آیا رابطه ای بین غلظت این دو عنصر وجود دارد ؟

جدول ۵-۱- غلظت سدیم و کلر برحسب میلی اکی والان گرم چاههای آب مشهد در سال ۱۳۸۹

ردیف	cl	na		ردیف	cl	na
۱	۴.۳	۹.۲		۲۹	۱	۲.۶
۲	۲.۲	۸.۱		۳۰	۳.۵	۶.۸
۳	۸.۳	۱۲.۶		۳۱	۰.۹	۲.۷
۴	۷	۱۰.۱		۳۲	۰.۵	۰.۳
۵	۱۹	۲۵.۶		۳۳	۲	۴.۲
۶	۷	۱۰.۲		۳۴	۴.۵	۱۳.۶
۷	۲۰	۲۷.۲		۳۵	۱.۶	۴.۳
۸	۱۲	۱۸.۵		۳۶	۱	۱.۵
۹	۱۲.۵	۱۳.۷		۳۷	۰.۴	۱.۳
۱۰	۱۲	۱۴		۳۸	۴	۶.۴
۱۱	۱۵	۲۰.۶		۳۹	۲.۸	۲.۹
۱۲	۲۰	۲۱.۱		۴۰	۱.۴	۳.۷
۱۳	۳	۳.۲		۴۱	۱	۲.۱
۱۴	۱.۵	۳.۴		۴۲	۴.۸	۶
۱۵	۱	۱.۲		۴۳	۳.۵	۴.۳
۱۶	۱.۵	۲.۵		۴۴	۲.۸	۸.۴
۱۷	۰.۸	۰.۷		۴۵	۷.۳	۱۲.۱

۱۸		۰.۳	۰.۲	۴۶		۲.۸	۳.۷
۱۹		۱.۳	۱.۵	۴۷		۲۸	۱۷.۳
۲۰		۱۷	۲۰.۳	۴۸		۲.۵	۷
۲۱		۰.۶	۰.۶	۴۹		۵	۹.۹
۲۲		۰.۷	۰.۸	۵۰		۲۰	۳۱
۲۳		۲	۴.۲	۵۱		۹.۳	۱۳.۹
۲۴		۱.۳	۲.۱	۵۲		۲۷	۳۷.۲
۲۵		۰.۸	۱.۴	۵۳		۲۵	۳۲.۵
۲۶		۰.۹	۱	۵۴		۳۸	۴۵.۷
۲۷		۰.۸	۲.۶	۵۵		۸.۸	۱۵.۲
۲۸		۰.۵	۱.۵	۵۶		۱۸	۲۸

مثال ۳-شرکت آب منطقه ای مازندران در یک آزمایش رفت پمپاژ چاه عمیق در چاه شماره EX6 نکا واقع در استان مازندران تراز سطح آب زیرزمینی بر اساس زمان را به صورت جدول زیر تهیه کرده است.بستگی داده ها را از طریق رگرسیون غیر خطی نمایش دهید.

زمان (دقیقه)	سطح آب (متر)
۰	۳۰,٦
۱	۳۰,٥٦
۲	۳۰,٥٦
۳	۳۰,٥٥
٤	۳۰,٥٥
٥	۳۰,٥٤
٦	۳۰,٥۲
۷	۳۰,٥۲
۸	۳۰,٥
۹	۳۰,٥
۱۰	۳۰,٤۹
۱٥	۳۰,٤۸
۲۰	۳۰,٤۸
۲٥	۳۰,٤۷
۳۰	۳۰,٤۷
٤۰	۳۰,٤۷
٥۰	۳۰,٤٦
٦۰	۳۰,٤٦