رگرسیون خطی ساده

متغیرهای بکار رفته در رگرسیون متغیرهای مستقل (Independent) و متغیر وابسته (Dependent) می‌باشد.

Y	X
متغیر وابسته	متغیر مستقل
$\hat{Y} = b_0 + b_1\hat{X}$	

متغیرهای وابسته حتما باید کمّی باشند ولی متغیرهای مستقل هم می‌توانند کمّی و هم کیفی باشند. متغیرهای کیفی که در رگرسیون بکار برده می‌شوند به عنوان متغیر مجازی (Dummy) مدنظر قرار می‌گیرند که مقادیر صفر و یک را به خود اختصاص می‌دهند. مثلا داشتن پارکینگ ۱ و نداشتن پارکینگ صفر.

زمانی که متغیرها بیش از دو حالت باشند نیز باید از چند متغیر مجازی استفاده شود. به عنوان مثال در متغیر مکانی:

مکان: شمال، جنوب، شرق و غرب که به این صورت تعریف می‌گردد:

$$X_3\begin{cases}1 & \text{شمال}\\ 0 & \text{بقیه}\end{cases}, X_4\begin{cases}1 & \text{جنوب}\\ 0 & \text{بقیه}\end{cases}, X_5\begin{cases}1 & \text{شرق}\\ 0 & \text{بقیه}\end{cases}, X_6\begin{cases}1 & \text{غرب}\\ 0 & \text{بقیه}\end{cases}$$

اگر متغیر وابسته کیفی باشد تنها راه، بکارگیری رگرسیون لجستیک است.

متغیرهای مستقل نباید با یکدیگر همبستگی داشته باشند.

مثال: یک شرکت حمل و نقل رابطه بین زمان طی شده و مسافت طی شده را به صورت زیر ثبت نموده است.

y	۹,۳	۴,۸	۸,۹	۶,۵	۴,۲	۶,۲	۷,۴	۶	۷,۶	۶,۱
x_1	۱۰۰	۵۰	۱۰۰	۱۰۰	۵۰	۸۰	۷۵	۶۵	۹۰	۹۰

جواب:

متغیر مسافت مستقل و متغیر زمان وابسته می‌باشد.

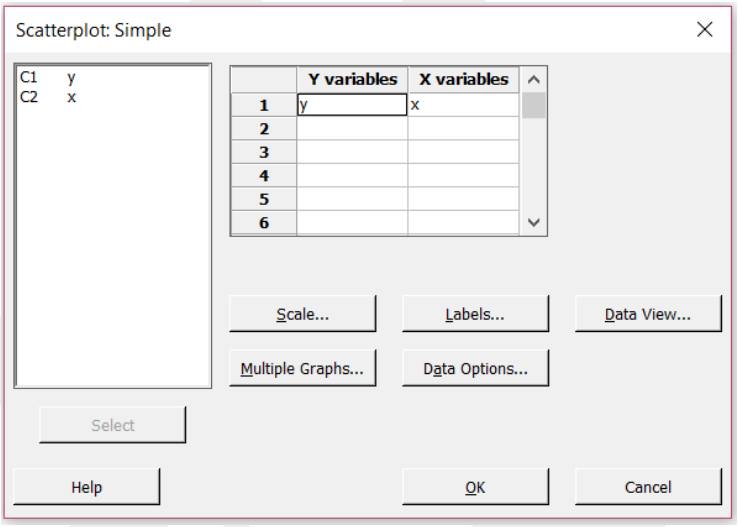
ابتدا نمودار (x,y) متغیرها را ترسیم نموده تا وضعیت خطی بودن یا غیر خطی بودن را مورد بررسی قرار دهیم.

سپس خط رگرسیون داده های فوق را محاسبه می‌کنیم.

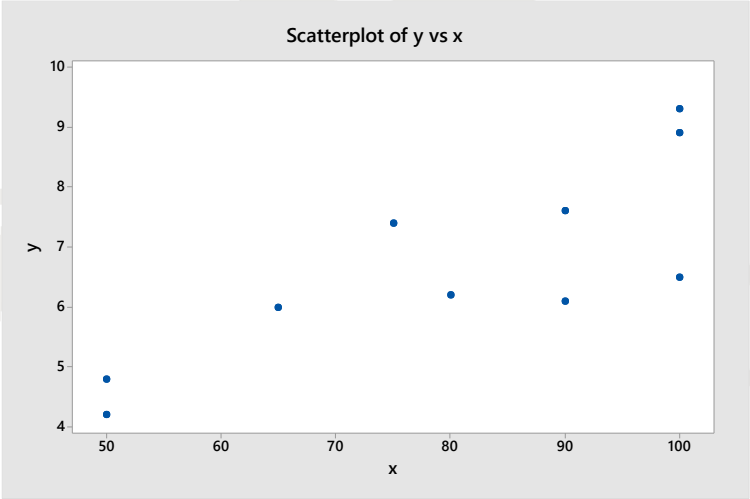
ترسیم نمودار x و y با نرم افزار Minitab

Graph > Scatterplots

نمودار Simple را انتخاب می‌نماییم.



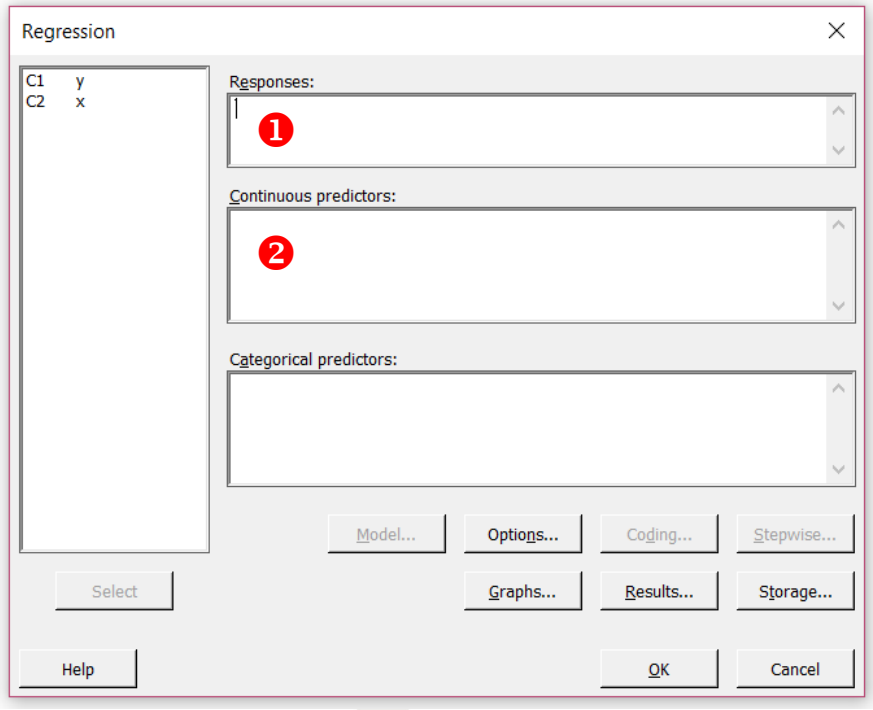
متغیرهای x و y را معرفی نموده و نمودار را ترسیم می‌کنیم.



در مثال فوق، پراکندی داده‌ها نشان می‌دهد که امکان رسم رگرسیون خطی مقدور است.

رگرسیون خطی نرم افزار Minitab

Stat > Regression> Regression> Fit Regression Model...



Responses ❶

در این قسمت متغیر وابسته (در اینجا y) وارد می‌شود.

Continuous predictors ❶

در این قسمت متغیر مستقل (در اینجا x) وارد می‌شود.

خروجی نرم افزار در مثال فوق:

Regression Analysis: y versus x

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	15.871	15.8713	15.81	0.004
x	1	15.871	15.8713	15.81	0.004
Error	8	8.029	1.0036		
Lack-of-Fit	4	2.137	0.5343	0.36	0.825
Pure Error	4	5.892	1.4729		
Total	9	23.900			

```
Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
1.00179  66.41%     62.21%      48.89%
```

$R\text{-}sq$ یا (R^2) «ضریب تعیین» می‌باشد که برای قضاوت در خصوص مناسب بودن مدل رگرسیون از نظر تشریح پراکندگی داده‌ها توسط مدل مورد استفاده قرار می‌گیرد.

$R\text{-}sq\ (adj)$ «ضریب تعیین تعدیل شده» بوده که معیار بهتری نسبت به $R\text{-}sq$ است

```
Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    1.27     1.40     0.91    0.390
x         0.0678   0.0171     3.98    0.004  1.00
```

می‌توانیم آزمونی برای معنی داری یا عدم معنی داری متغیر x در این معادله رگرسیون نوشت.

با در نظر گیری معادله $y = b_0 + b_1 x_1$ داریم.

$$x_1 \begin{cases} H_0: b_1 = 0 \\ H_1: b_1 \neq 0 \end{cases}$$

با توجه به اینکه $P - value = 0.004 < \alpha = 0.05$ در نتیجه فرضیه H_0 رد شده و متغیر x_1 در این معادله ضریبی غیر صفر داشته و معنی دار می‌باشد.

```
Regression Equation

y = 1.27 + 0.0678 x
```

به ازای هر یک کیلومتر مسیر طی شده انتظار می‌رود به میزان ۰,۰۶۷۸ به زمان اضافه شود.

در مثال فوق متغیر جدید با عنوان تعداد توقفات خودروها اضافه می‌گردد که نتیجه بدین شرح تغییر می‌کند.

y	۹,۳	۴,۸	۸,۹	۶,۵	۴,۲	۶,۲	۷,۴	۶	۷,۶	۶,۱
x_1	۱۰۰	۵۰	۱۰۰	۱۰۰	۵۰	۸۰	۷۵	۶۵	۹۰	۹۰
x_2	۴	۳	۴	۲	۲	۲	۳	۴	۳	۲

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.573142  90.38%     87.63%      80.76%
```

نسبت به حالت قبل، هر دو مولفه «ضریب تعیین» و «ضریب تعیین تعدیل شده» بهبود یافتند.

```
Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   -0.869    0.952    -0.91    0.392
x1        0.06113  0.00989     6.18    0.000  1.03
x2          0.923    0.221     4.18    0.004  1.03
```

ملاحظه می‌گردد که $P - value$ هر دو متغیر x_1 و x_2 از α کوچکتر بوده و هر دو متغیر در این معادله معنی دار می‌باشند.

```
Regression Equation

y = -0.869 + 0.06113 x1 + 0.923 x2
```

به ازای افزایش یک توقف(x_2)، زمان طی شده (y) به میزان ۰,۹۲۳ ساعت افزایش می‌یابد با توجه به ثابت ماندن طول مسیر(x_1)

به مثال فوق یک متغیر مجازی جدید اضافه می‌گردد.

y	۹,۳	۴,۸	۸,۹	۶,۵	۴,۲	۶,۲	۷,۴	۶	۷,۶	۶,۱
x_1	۱۰۰	۵۰	۱۰۰	۱۰۰	۵۰	۸۰	۷۵	۶۵	۹۰	۹۰
x_2	۴	۳	۴	۲	۲	۲	۳	۴	۳	۲
x_3	۱	۰	۰	۰	۱	۰	۱	۱	۰	۱

```
Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.615103  90.50%     85.75%      76.46%
```

نسبت به حالت قبل (شرایط دوم مثال)، مولفه «ضریب تعیین» بهبود یافت اما «ضریب تعیین تعدیل شده» کمتر شده است.

```
Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   -0.95     1.06    -0.89    0.405
x1        0.0619   0.0109     5.66    0.001  1.09
x2         0.913    0.240     3.80    0.009  1.05
x3         0.112    0.404     0.28    0.790  1.08
```

ملاحظه می‌گردد که $P-value$ دو متغیر x_1 و x_2 از α کوچکتر بوده اما متغیرx_3 بزرگتر می‌باشد. در نتیجه متغیر x_3 در این معادله معنی دار نبوده و باید این متغیر را از رگرسیون خارج نمود.

```
Regression Equation

y = -0.95 + 0.0619 x1 + 0.913 x2 + 0.112 x3
```

مقایسه تاثیر متغیرهای مستقل در رگرسیون

در مثال فوق با توجه به وجود ۳ متغیر مستقل و ۱ متغیر وابسته حالات متفاوتی از بکارگیری متغیرهای مستقل وجود دارد که می‌توان آنها را از روش زیر مقایسه نمود:

Stat > Regression> Regression> Best Subsets ...

خروجی نرم افزار در مثال فوق:

Best Subsets Regression: y versus x1, x2, x3

Response is y

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	x 1	x 2	x 3
1	66.4	62.2	48.9	15.2	1.0018	X		
1	37.8	30.1	0.0	33.3	1.3627		X	
2	90.4	87.6	80.8	2.1	0.57314	X	X	
2	67.7	58.5	34.5	16.4	1.0502	X		X
3	90.5	85.8	76.5	4.0	0.61510	X	X	X

با بررسی *R-sq (adj)* ملاحظه می‌گردد که بهترین نتیجه زمانی حاصل می‌شود که متغیرهای مستقل بکار گرفت شده دو متغیر x_1 و x_2 باشند.