

PubH 7405: REGRESSION ANALYSIS



Correlation Analysis

CORRELATION & REGRESSION

- We have 2 continuous measurements made on each subject, one is the response variable Y, the other predictor X. There are two types of analyses:
- **Correlation**: is concerned with the association between them, **measuring the strength** of the relationship; **the aim is to determine if they are correlated – the roles are exchangeable.**
- **Regression**: To **predict** response from predictor.

ROLES OF VARIABLES

In Regression Analysis, each has a well-defined role; we'll predict "response Y" from a given value of "predictor X"

In Correlation Analysis, the roles of "X" and "Y" are exchangeable; in the coefficient of correlation "r" is symmetric with respect to X and Y: we get the same result regardless of which one is X – no special "label".

SCATTER DIAGRAM

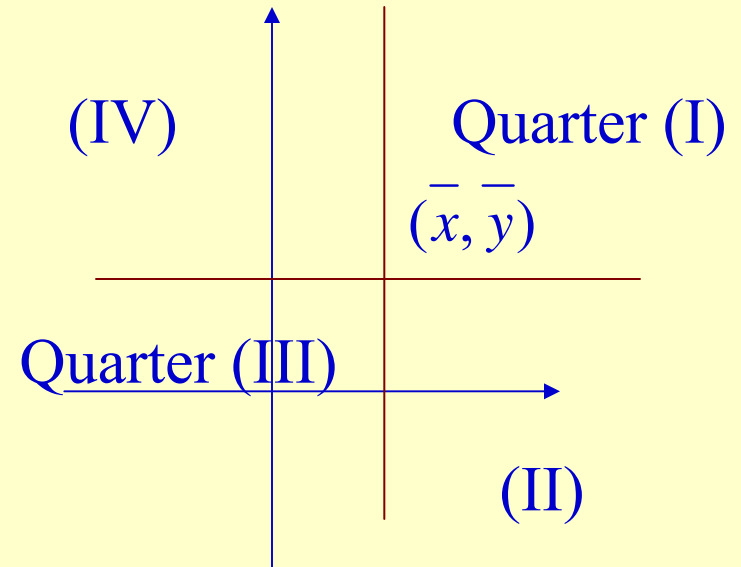
- In quarters I and III,

$$(x - \bar{x})(y - \bar{y}) > 0$$

- For positive association,

$$\sum (x - \bar{x})(y - \bar{y}) > 0$$

- For stronger relationship most of the dots, being closely clustered around the line, are in these two quarters; the above sum is large.

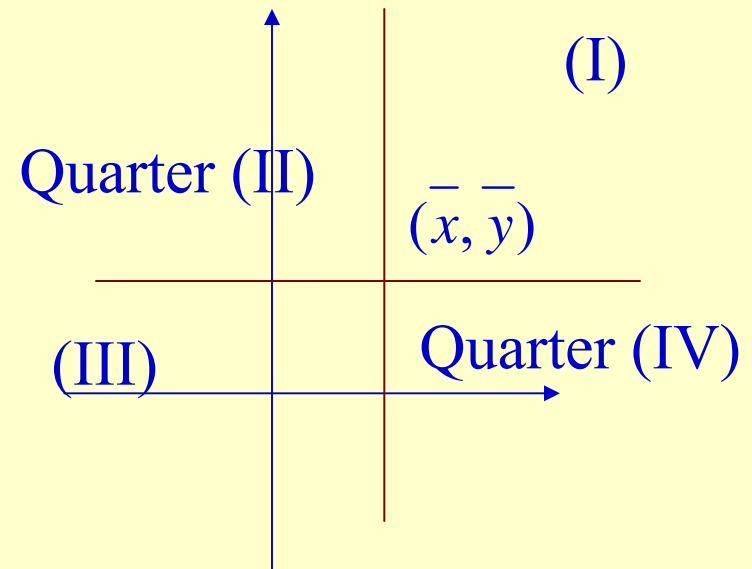


SCATTER DIAGRAM

- In quarters II and IV,
 $(x - \bar{x})(y - \bar{y}) < 0$
- For negative association,

$$\sum (x - \bar{x})(y - \bar{y}) < 0$$

- For stronger relationship most of the dots, being closely clustered around the line, are in these two quarters; the sum is a large negative number.



SUMMARY

- The “**sum of products**” $\sum (x - \bar{x})(y - \bar{y})$ summarizes the “**evidence**” of the relationship under investigation; It is zero or near zero for weak associations and is large, negative or positive, for stronger associations. **The sum of products can be used as a measure of the strength of the association itself.**
- However, it is “**unbounded**” making it hard to use because we cannot tell if we have a strong association (**how large is “large”?**).
- We need to “**standardize**” it.

COEFFICIENT OF CORRELATION

With a standardization, we obtain a statistic r is called the **Correlation Coefficient** measuring the strength of the relationship:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

WE can “explain” the denominator as necessary for standardization: to obtain a statistic in $[-1, 1]$.

There are many different ways to express the coefficient of correlation r ; one of which is often mentioned as **the “short-cut” formula**:

$$\begin{aligned} r &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}} \\ &= \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}][\sum y^2 - \frac{(\sum y)^2}{n}]}} \\ \mathbf{r} &= \frac{\mathbf{s_{xy}}}{\mathbf{s_x s_y}} \end{aligned}$$

s_{xy} is the “sample covariance” of X and Y

Another very useful formula is to express the coefficient of correlation r as the “Average Product” in “standard units” – where s_x and s_y are the (sample) standard deviations of X and Y , respectively:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

The following is an important and very interesting characteristic:

$$\mathbf{r(cX + d, aY + b) = r(X, Y)}$$

we can prove by showing that $(u = c*x + d)$ is the same as x in standard units & $(v = a*y + b)$ is the same as y in standard units; e.g.

$$\frac{\mathbf{u - \bar{u}}}{\mathbf{s_u}} = \frac{\mathbf{x - \bar{x}}}{\mathbf{s_x}}$$

$$\frac{\mathbf{v - \bar{v}}}{\mathbf{s_v}} = \frac{\mathbf{y - \bar{y}}}{\mathbf{s_y}}$$

CORRELATION MODEL

“Correlation Data” are often cross-sectional or observational. Instead of a regression model, one should consider a “correlation model”; the most widely used is the “Bivariate Normal Distribution” with density:

$$f(X, Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) + \left(\frac{Y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$$\begin{aligned}\sigma_{xy} &= \text{Cov}(X, Y) \\ &= E[(X - \mu_x)(Y - \mu_y)]\end{aligned}$$

σ_{xy} is the Covariance and ρ is the Coefficient of Correlation between the two random variables X and Y; ρ is estimated by the (sample) Coefficient of Correlation r .

The classic work of **Pearson** (Biometrika, 1909) and Fisher (Biometrika, 1915; Metron, 1921) led to the **(Pearson's, product moment) coefficient of correlation r** which generated a steady stream of development of measures of association and correlation coefficients appropriate in different contexts (latest: Kraemer, SMMR, **2006**).

At the beginning of the 20th century, correlation analysis was one of the most common statistical analyses, especially in **health psychology and epidemiology**. **Conversely, the use of coefficient of correlation r has had major influence on medical policy decision making.**

HOW STRONG IS A CORRELATION?

- The Coefficient of Determination r^2 , when expressed as percentage, represents the proportion of the degree of variation among the values of one variable which is accounted by its relationship with the other variable.
- When $r^2 > 50\%$, one variable is responsible for more than half of the variation in the other; the relationship is obviously strong.
- A correlation with $r > .7$ is therefore conventionally considered as a strong.

TESTING FOR INDEPENDENCE

- The Coefficient of Correlation r measures the strength of the relationship between two variables, say the Mother's Weight and her Newborn's Birth Weight. But r is only a **Statistic**; it is an Estimate of an unknown Population Coefficient of Correlation ρ (rho), the same way the sample \bar{x} is used as an estimate of the Population mean μ .
- The basic question is concerned: $H_0: \rho = 0$; only when H_0 is true, the two variables are not correlated.

Try to separate a “statistic” from a “parameter”. When $r = 0$, it only imply that values of the two factors, as measured from **that sample**, are not related. But you can’t generalize that yet (what you found might happen by chance, if you do it again you might not see it again); **Only when $\rho = 0$** , we can conclude that the factors are not related - **population-wise** .

TESTING FOR INDEPENDENCE

- The Coefficient of Correlation r measures the strength of the relationship between two variables; but as **statistic** it involves “random variation” in its sampling distribution. We are interested in knowing if we can conclude that: $\rho \neq 0$, that the two variables under investigation are really correlated - not just by chance.
- It is a two-sided Test of the Null Hypothesis of No Association, $H_0: \rho = 0$, against $H_A: \rho \neq 0$ of Real Association (you can do it as one-sided too).

TESTING FOR INDEPENDENCE

- The Test Statistic is:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

- It is the same t-test as used in the comparison of two Population Means; the Degree of Freedom is: $df = n-2$ (same way to form your rejection decision and to calculate p-value) .

EXAMPLE

- For the Birth-Weight problem, we have: $n=12$ and $r = -.946$ leading to:

$$t = (-.946) \sqrt{\frac{12-2}{1-(-.946)^2}}$$

$$t = -9.23$$

- At $\alpha=.05$ & $df = 10$, the tabulated coefficient is 2.228 indicating that the Null Hypothesis should be rejected ($t=-9.23 < -2.228$); (two-sided) $p\text{-value} < .001$.

In recent years, it has become increasingly clear that the magnitude and direction of correlation coefficient is more important, not merely whether or not the association is “statistically significant” (Hunter, 1997); Schmidt, 1996). It has been suggested that the **Null Hypothesis of independence in never true** (Meehl, 1967; Jones and Tukey, 2000). Consequently, a “statistically significant association” **only means** the sample size and the design were good enough to detect a non-random association between X and Y, not the strength of the association in necessarily of any clinical significance. **You can say that p-value is a statement about the data, not a statement about the strength of an association.**

The Coefficient of Correlation ρ between the two random variables X and Y is estimated by the (sample) Coefficient of Correlation r but the sampling distribution of r is far from being normal. Confidence intervals of r is by first making the “**Fisher’s z transformation**”; the distribution of z is normal if the sample size is not too small

$$\mathbf{z} = \frac{1}{2} \ln \left(\frac{1 + \mathbf{r}}{1 - \mathbf{r}} \right)$$

$\mathbf{z} \in \mathbf{Normal}$

$$\mathbf{E}(\mathbf{z}) = \frac{1}{2} \ln \left(\frac{1 + \boldsymbol{\rho}}{1 - \boldsymbol{\rho}} \right)$$

$$\sigma^2(\mathbf{z}) = \frac{1}{\mathbf{n} - 3}$$

$$\mathbf{r} = \frac{\exp(2\mathbf{z}) - 1}{\exp(2\mathbf{z}) + 1}$$

EXAMPLE #1: Birth Weight Data

x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

$$r = -.946$$

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

$$= -1.792$$

$$\sigma(z) = \sqrt{\frac{1}{12-3}}$$

$$= .333$$

$$-1.792 \pm (1.96)(.333) = (-2.445, -1.139)$$

95% Confidence Interval of ρ is :

$$\frac{\exp(2z) - 1}{\exp(2z) + 1} = (-.985, -.814)$$

$$r = -.946$$

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

$$= \frac{1}{2} \ln\left(\frac{1-.946}{1+.946}\right) = -1.792$$

$$\sigma(z) = \sqrt{\frac{1}{12-3}} = .333$$

$$-1.792 \pm (1.96)(.333) = (-2.445, -1.139)$$

95% Confidence Interval of ρ is (-.985,-.814):

$$\frac{\exp[(2)(-2.445)]-1}{\exp[(2)(-2.445)]+1} = -.985$$

$$\frac{\exp[(2)(-1.139)]-1}{\exp[(2)(-1.139)]+1} = -.814$$

EXAMPLE #2: AGE & SBP

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

$$r = .564$$

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

$$= .639$$

$$\sigma(z) = \sqrt{\frac{1}{15-3}}$$

$$= .289$$

$$.639 \pm (1.96)(.289) = (.072, 1.205)$$

95% Confidence Interval of ρ is :

$$\frac{\exp(2z) - 1}{\exp(2z) + 1} = (.073, .835)$$

CONDITIONAL DISTRIBUTION

$$f(X,Y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right) + \left(\frac{Y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

Theorem :

The conditional distribution of Y for any given $X=x$ is normal with mean $\beta_0 + \beta_1 x$ and standard deviation $\sigma_{y|x}$:

$$\beta_0 = \mu_y - \mu_x \rho \frac{\sigma_y}{\sigma_x}$$

$$\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$$

$$\sigma_{y|x}^2 = (1-\rho^2)\sigma_y^2$$

Again, since $\text{Var}(Y|X) = (1 - \rho^2)\text{Var}(Y)$, ρ is both a measure of linear association and a measure of “variance reduction” (in Y associated with knowledge of X) – that’s why we called r^2 , an estimate of ρ^2 , the “coefficient of determination”.

CORRELATION & REGRESSION

- Suppose that we select a random sample of observations $\{(x,y)\}$ from the bivariate normal distribution and wish to make conditional inferences about Y , given $X = x$.
- The previous results of the “normal regression model” is entirely applicable because:
 - (1) The Y observations are independent
 - (2) The Y observations when X is considered given or fixed are distributed as normal with constant variance $\sigma_{y|x}^2$ and mean: $E(Y_2|Y_1) = \beta_0 + \beta_{1x}$

The conditional distribution also explains a (future) result: the relationship between the (estimated) “slope” and the Pearson’s “coefficient of correlation”:

$$\mathbf{b}_1 = \mathbf{r} \frac{\mathbf{s}_y}{\mathbf{s}_x}$$

In the bivariate normal model, when $\rho = 0$, the exact distribution of r is such that the following t-statistic has a t-distribution with $(n-2)$ degrees of freedom. And this distribution is remarkably robust to deviations from the assumption of bivariate normality:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

When $\rho \neq 0$, the exact distribution of r is unknown. The Fisher's transformation has been widely used, the transformed statistic Z is approximately distributed as normal with variance equal to $1/(n-3)$:

$$Z = \frac{1}{2} \ln \left[\frac{1+r}{1-r} \right]$$

However, the **non-null distribution** **may not be that robust** to deviations from marginal normality, to unequal conditional variances, to non-linear relationship, & to presence of outliers.

IMPLICATION

- **If we only need to test for independence (and/or draw regression inferences), diagnostics may not be very crucial because the methods are robust.**
- **But if we focus on confidence interval estimation of the coefficient of correlation we should pay attention to diagnostics because the method (Fisher's transformation) may not be robust. We will do more of these in the following weeks.**

We end up having “a correlation analysis” and “a regression analysis” that go hand-in-hand:

$$b_1 = r \frac{s_y}{s_x}$$

From this simple result, we can see that “ b_1 ” and “ r ” are of the same sign – and both are equal to zero at the same time; they measure the same thing but on different “scale”.

We end up having “a correlation analysis” and “a regression analysis” that go hand-in-hand. But, in general, it does not have to be that way

A measure of the strength of an association, say θ , needs only satisfying the following conditions: (1) It is **an unit-free measure** that range from -1 to +1, (2) If X and Y are independent, $\theta = 0$, and (3) If one can perfectly predict Y from X, or X from Y, $\theta = 1$

Besides the (**Pearson's**) coefficient of correlation r , we have (i) **Spearman's rho** and (ii) **Kendall's tau**.

Spearman's rho and Kendall's tau are nonparametric statistics; statistics/methods based on "ranks"

INTRACLASS CORRELATION

In statistics, the intraclass correlation (or the *intraclass correlation coefficient*, abbreviated ICC) is a descriptive statistic that can be used when quantitative measurements are made on units that are organized into groups. It describes how strongly units in the same group resemble each other. While it is viewed as a type of correlation, unlike most other correlation measures it operates on data structured as groups, rather than data structured as paired observations.

ICC: EXAMPLES

The *intraclass correlation* is commonly used to quantify the degree to which individuals resemble each other in terms of a quantitative trait (see **heritability**). Another prominent application is the assessment of **consistency or reproducibility of quantitative measurements made by different observers measuring the same quantity.**

INTRACLASS VS. INTERCLASS

The earliest work on intraclass correlation focused on the case of paired measurements (e.g. Quantitative trait of identical twins), and the first intraclass correlation (ICC) statistics to be proposed were modifications of the interclass correlation (Pearson correlation). The key difference between this ICC and the interclass (Pearson) correlation is that the data are pooled to estimate the mean and variance; therefore, its degree of freedom is $(2n-1)$ for variance and $(2n-2)$ for test of independence.

Suppose the data set consists of n pairs of observations expressing a possible relationship between two continuous variables. We characterize the strength of such a relationship by calculating the coefficient of correlation r called the Pearson's correlation coefficient. Like other common statistics, such as the mean and the standard deviation s , the correlation coefficient r is very sensitive to “extreme observations”. We may be interested in calculating a measure of association that is more robust with respect to outlying values. There are not one but two nonparametric procedures: the “Spearman's rho” and the “Kendall's tau” rank correlation coefficients. Spearman's rho is more similar to Pearson's r .

Spearman's rho

The Spearman's rank correlation is a direct nonparametric counterpart of the Pearson's correlation coefficient. To perform this procedure, we first arrange the “x” values from smallest to largest and assign a “rank” from 1 to n for each value; let R_i be the rank of value x_i . Similarly, we arrange the “y” values also from smallest to largest and assign a rank from 1 to n for each value; let S_i be the rank of value y_i . If there are tied observations, we assign an “average rank” averaging the ranks that the tied observations jointly take. For example, if the second and third measurements are equal, they both are assigned 2.5 as their common rank. The next step is to replace, in the formula of the Pearson's correlation coefficient r , x_i by its rank R_i and y_i by its rank s_i .

$$\begin{aligned}\rho &= \frac{\sum (R - \bar{R})(S - \bar{S})}{\sqrt{[\sum (R - \bar{R})^2][\sum (S - \bar{S})^2]}} \\ &= 1 - \frac{6 \sum (R - S)^2}{n(n^2 - 1)}\end{aligned}$$

It's still symmetric; doesn't matter which rank is R; the second formula is easier to use in hand calculations.

NUMERICAL EXAMPLE

x (oz)	R=Rank(x)	y (%)	S=Rank(y)	R-S	(R-S) ²
112	10	63	3	7	49
111	9	66	4	5	25
107	8	72	5.5	2.5	6.25
119	12	52	2	10	100
92	4	75	7	-3	9
80	1	118	11	-10	100
81	2	120	12	-10	100
84	3	114	10	-7	49
118	11	42	1	10	100
106	7	72	5.5	1.5	2.25
103	6	90	8	-2	4
94	5	91	9	-4	16
Totals					560.5

$$\begin{aligned}
 \rho &= 1 - \frac{6 \sum (R - S)^2}{n(n^2 - 1)} \\
 &= 1 - \frac{(6)(560.5)}{(12)(144 - 1)} \\
 &= -.96 \\
 \mathbf{r} &= \mathbf{-.946}
 \end{aligned}$$

The relationship between Pearson's r and Spearman's ρ is similar to that between the two-sample t-test and the Wilcoxon test: replacing observed values by their ranks.

Kendall's tau

Unlike the Spearman's rho, the other rank correlation - the Kendall's tau rank correlation is defined and calculated very differently, even though they often yield similar numerical results. In practical applications, you could rely on SAS; it could be as follows:

```
PROC CORR PEARSON SPEARMAN KENDALL;  
VAR XNAME YNAME;
```

CORRELATION (& Scatter Diagram)

```
options ls=79;
title "Descriptive Statistics for SBP versus Age";
data SBP;
input X Y;
  label X = 'Age'
       Y = 'Blood Pressure';
cards;
42 130
46 115
...
75 165
;
```

```
proc CORR data=SBP;
```

```
run;
```

```
proc plot data=SBP;
```

```
  plot y*x='*';
```

```
run;
```

Proc CORR gives the coefficient of correlation r (& the p -value)

Proc PLOT provides the Scatter Diagram; could choose symbol to plot.

Specify Notation for the graph

Simple Statistics

Variable	N	Mean	Std Dev	Sum
X	15	65.600000	15.592123	984.000000
Y	15	146.200000	19.479660	2193.000000

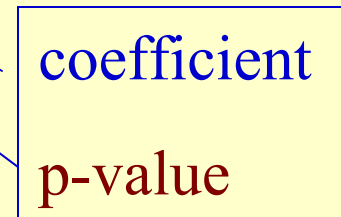
Simple Statistics

Variable	Minimum	Maximum	Label
X	41.000000	85.000000	Age
Y	100.000000	165.000000	Blood Pressure

OUTPUT

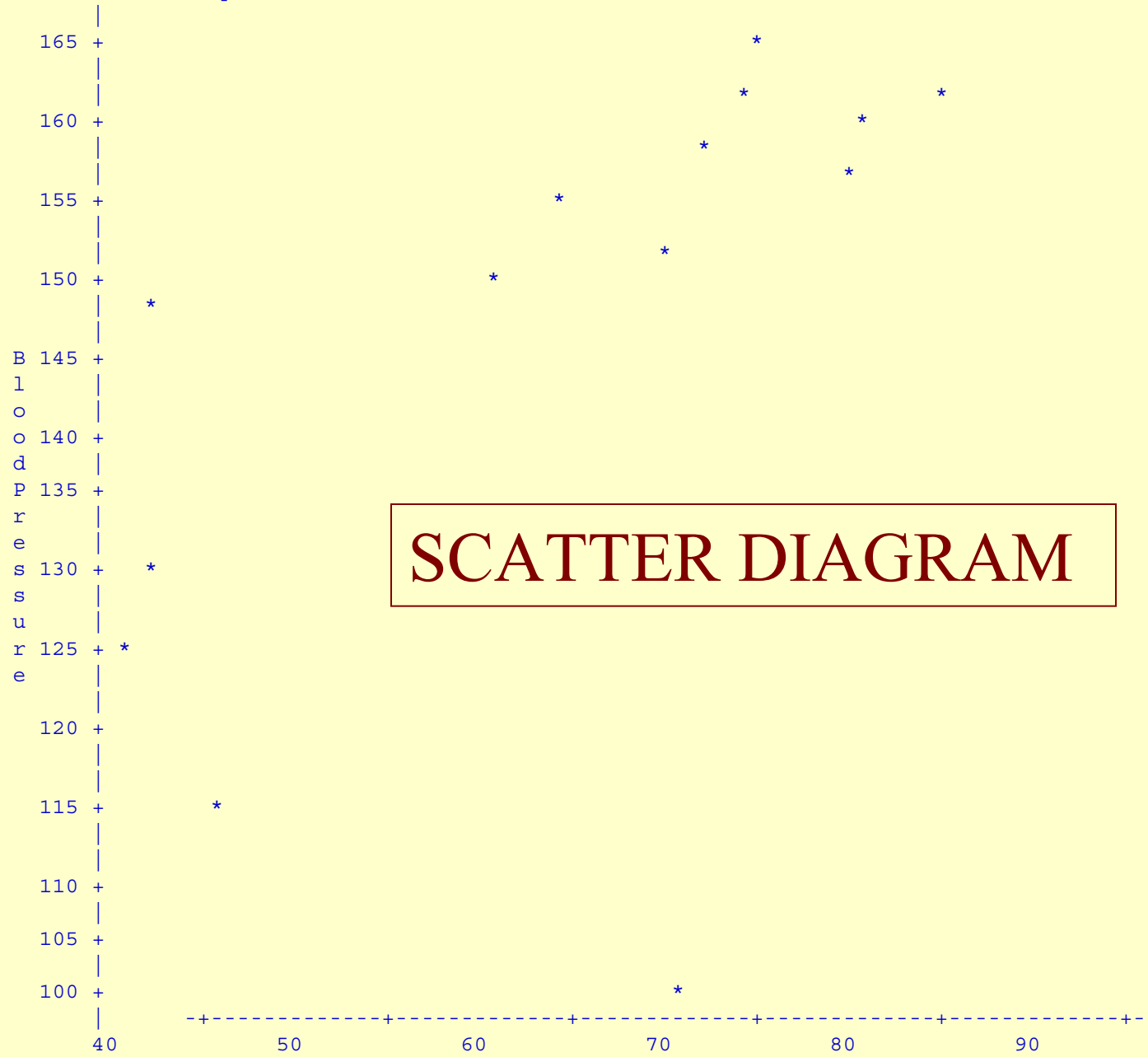
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 15

	X	Y
X	1.00000	0.56422
Age	0.0	0.0285
Y	0.56422	1.00000
Blood Pressure	0.0285	0.0



Note: results are symmetric

Plot of Y*X. Symbol used is '*'.

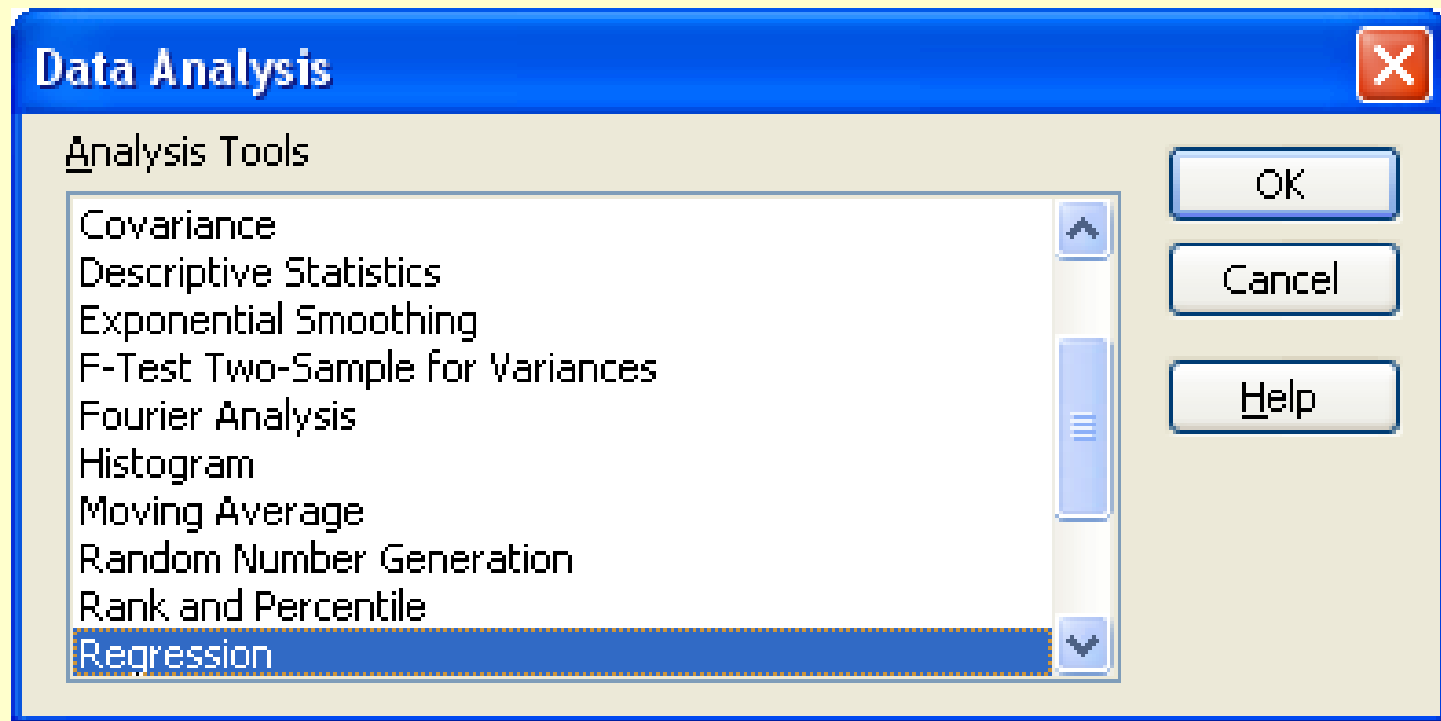


SCATTER DIAGRAM

The “Scatter Diagram” would help to see the suitability of using Pearson’s r ; for example, in the presence of some suspected outliers (suspected but not so sure to delete), Spearman’s ρ would be a necessary backup.

Excel: ANALYSIS

(1) click the *Tools* then (2) *Data Analysis*; among functions available, choose *Regression*.



SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.564224332
R Square	0.318349097
Adjusted R Square	0.265914412
Standard Error	16.68993768
Observations	15

You got “r” but labeled as “Multiple R”; be aware, the “sign” may be wrong i.e. “negative” – look at regression results, **r has the same sign as the “slope” (Excel is aimed at Regression, not Correlation alone).**

Due As Homework

4.1 The following data were collected during an experiment in which 10 laboratory animals were inoculated with a pathogen. The variables are Time after inoculation (X, in minutes) and Temperature (Y, in Celsius degrees).

X	y
24	38.8
28	39.5
32	40.3
36	40.7
40	41.0
44	41.1
48	41.4
52	41.6
56	41.8
60	41.9

- Draw a Scatter Diagram to show the association, if any, between these two variables and calculate the Pearson's coefficient of correlation r
- Testing for possible independence between X and Y, at the %5 level of significance
- Calculate the 95% Confidence Interval for the Population Coefficient of Correlation
- Calculate the Spearman's rho

(Show the formulas you use and SAS programs)

4.2 Answer the 4 questions of Exercise 4.1 using dataset "Infants" with variables X = Gestational Weeks and Y = Birth Weight.

Only #4.2 is required