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 VAIRABLES

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".

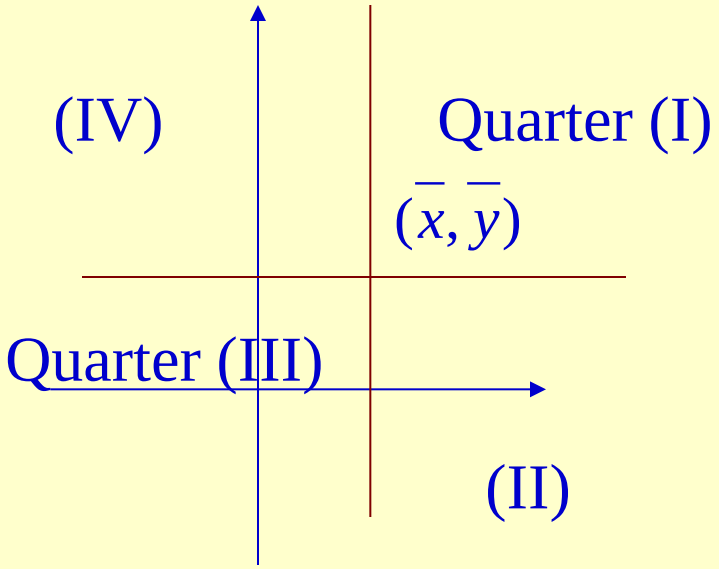
SCATTER DIAGRAM

- 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.

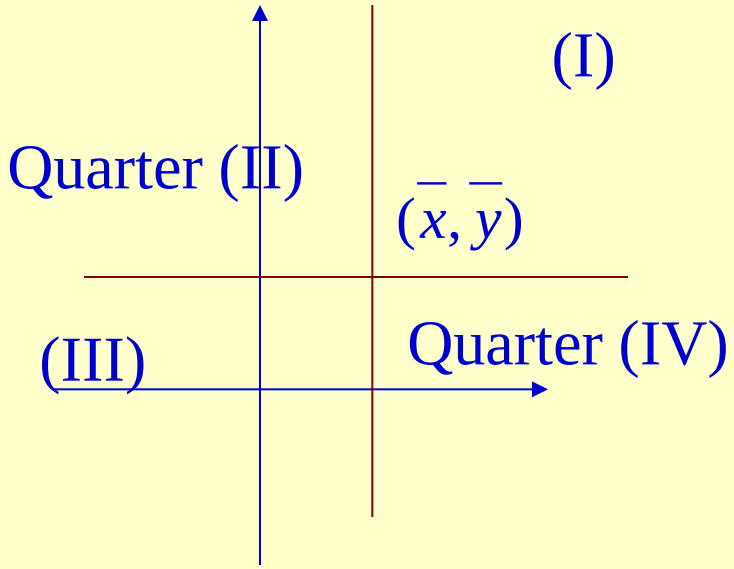
SCATTER DIAGRAM

- 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 n; one of which is often mentioned as **the "short-cut" formula**:

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

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

$$\mathbf{r = \frac{s_{xy}}{s_x s_y}}$$

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:

$$r(cX + d, aY + b) = r(X, Y)$$

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.

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

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

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:

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

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

$$\sigma_{xy}=Cov(X,Y)$$

$$=E[(X-\mu_x)(Y-\mu_y)]$$

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 α=.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-value < .001.**

In recent years, it has become increasingly clear that the magitude 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 is by first making the "Fisher's z transformation"; the distribution of z is normal if the sample size is not too small

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

$$z \in \text{Normal}$$

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

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

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

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% Confedence 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% Confedence 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% Confedence 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 $\sigma_{y|x}$:

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

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

$$\boldsymbol{\sigma^2_{y|x}=(1-\rho^2)\sigma^2_y}$$

Again, since $Var(Y|X)=(1-\rho^2)Var(Y)$, 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":

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

In the bivariate normal model, when ρ = 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 ρ≠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. 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"**

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.

$$\rho = \frac{\sum (R - \bar{R})(S - \bar{S})}{\sqrt{[\sum (R - \bar{R})^2][\sum (S - \bar{S})^2]}}$$

$$= 1 - \frac{6\sum (R - S)^2}{n(n^2 - 1)}$$

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

$$\rho = 1 - \frac{6\sum (R-S)^2}{n(n^2-1)}$$

$$= 1 - \frac{(6)(560.5))}{(12)(144-1)}$$

$$= -.96$$

$$r = -.946$$

The relationship between Pearson's r and Spearman's ρ is similar to that between the two-sample t-test and the Wilconxon 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;
```

CAUTION ON EXTRAPOLATION

SCOPE OF THE MODEL

In formulating a regression model, we need to restrict the "coverage" of the model to some interval of values of the independent variable X; this is determined either by the design or the availability of data at hand. The shape of the regression function outside this range would be in doubt because the investigation provided no evidence as to the nature of the statistical relation outside this range. In short, one should not do any extrapolation.

The National Health and Nutrition Examination Study (or NHANES) tracks the health data of a large , representative sample of Americans , covering everything from hearing loss to sexually transmitted infections. For example, it gives very good data for the proportion of Americans who are overweight, which is defined as having body-mass index of 25 or higher.

There's no question that the prevalence of overweight has increased in recent decades. In the early 1970s, it was under 50%. By the early 1990s, that figure had risen to almost 60%, and by 2008 almost three-quarters of the U.S. population was overweight.

"Will all Americans become overweight"?

Youfa Wang and colleagues published an article in *Obesity* claiming that, yes, by the year 2048.

You can plot the prevalence of obesity against time and generate a linear regression line. In 2048, that line crosses 100%. And that's why Wang wrote that all Americans will be over weight in 2048, if the current trends continue.

There are two problems here.
The obvious one is the fatal extrapolation. You can easily make up a few counter example: (a) The line crosses 100% in 2048; so can we say that, by 2060, 109% of Americans would be overweight? (b) If we apply the same method to Black men, whose overweight prevalence is a bit smaller than that of the average American, its line will crosses 100% in 2095. If not all Black men are overweight in 2048, how can we say that ALL American are overweight in 2048?

In addition, sometimes lines are straight locally but curved globally. That's the case of proportion, as a dependent variable. It fits a logistic curve; that is, log([p/(1-p)] is a linear function of time (or whatever used as independent variable). If you can plot the prevalence of obesity against time, it might look like a straight line in the middle range of proportion but not so at both ends. Using wrong model contributes making extrapolation even worse.

In addition, sometimes lines are straight locally but curved globally. That's the case of proportion, as a dependent variable. It fits a logistic curve; that is, log([p/(1-p)] is a linear function of time (or whatever used as independent variable). If you can plot the prevalence of obesity against time, it might look like a straight line in the middle range of proportion but not so at both ends. Using wrong model contributes making extrapolation even worse.

The problem of linear extrapolation was warned by Mark Twain in “ Life on the Mississippi”:

“ The Mississippi river was twelve hundred and fifteen miles. In the space of one hundred and seventy-six years, the Lower Mississippi has shortened itself two hundred and forty-two miles; and average of one mile and one third per year. Therefore, any person can see that seven hundred and forty-two years from now the Mississippi will be only a mile and three-quarters long!”

CORRELATION AND CAUSATION

As we mentioned, all Null Hypotheses are likely false; everything is perhaps correlated to everything else. So people do not report all of these correlations. When you read a report that one thing is correlated with another, perhaps that correlation is “strong enough” to be worth reporting. But what? Does it pass a statistical test of significance?

There's something slippery about understanding of correlation and causation. When we say that good cholesterol HDL is correlated with a lower risk of heart attack, we're making a factual statement that "if you've got a high level of HDL cholesterol, you're less likely to have a heart attack". That does not necessarily men that the HDL is "doing something" – like scrubbing your arterial walls causing your cardiovascular health to improve.

It might be that HDL and heart attack are correlated for a different reason; say, some unknown factor tends both to increase HDL and decrease the risk of cardiovascular events. If that's the case, an HDL-increasing drug might or might not prevent heart attack. If the drug affects HDL by way of that mysterious factor, it would probably help your heart; but it boosts DDL by some other way, it would not help at all.

Back to the relationship between smoking and lunch cancer. At the turn of the twentieth century, lung cancer was an extremely rare disease. But by the late 1940's, the disease accounted for nearly 20% of cancer deaths among British men. Lung cancer was on the rise but no one was sure what to blame. Maybe it was smoke from factories, maybe increased levels of car exhaust? Or maybe it was cigarette smoking whose popularity had exploded during the very same period?

Cigarette smoking was emerged from the famous Case-control study by Doll and Hill in 1950. They show that smoking and lung cancer are correlated and the association got stronger as smoking got heavier. Does smoking cause lung cancer? Doll and Hill's data showed that some heavy smokers do not get lung cancer and some nonsmokers do. So the association was not of strict determination.

The famous Berkson, an MD, put it this way " Cancer is a biologic, not a statistical problem. Statistics can soundly play an ancillary role in its elucidation. But if biologists permit statisticians to become arbiters biologic questions, scientific disaster is inevitable".

Doll and Hill's study does not show that smoking causes lung cancer; as they write "the association would occur if carcinoma of the lung caused people to smoke or if both attributes were end-effects of a common cause". Of course, it is not reasonable to say that a tumor would go back in time and causes someone to smoke. But the problem of a common cause was more troubling at the time.

Even R. A. Fisher, the founding father of modern statistics was a skeptic of the tobacco-cancer link. He questioned "Is it possible that pre-cancerous condition – which must exist and is known to exist for years before become cancerous – is one of the causes of cigarette smoking? I don't think it can be excluded. I don't think we know enough to say it's a cause but pre-cancerous condition is involving certain amount of chronic inflammation …"

Over the course of the 1950s and 1960s, scientific knowledge and opinion on the relationship between smoking and lung cancer steadily converged toward consensus; the association between smoking and lung cancer had appeared consistently across study after study leading to several Surgeon General's reports implicating cigarette smoking as the cause lung and several other cancers. Scientists even identified several tobacco-specific carcinogens, e.g. NNN, and proved that they cause cancers in animals.

Due As Homework

Given n pairs of numbers (x,y) on two variables X and Y (n is the sample size); consider the following four (4) questions:

a) Draw a Scatter Diagram to show the association, if any, between these two variables and calculate the Pearson's coefficient of correlation r. Does it look "linear"?

b) Testing for possible independence between X and Y, at the %5 level of significance

c) Calculate the 95% Confidence Interval for the Population Coefficient of Correlation

d) Calculate the Spearman's rho; is it close to the value of Pearson's in (a)?

(Show the formulas you use and SAS programs)

#4.1 We have a data set on 100 infants (File: "Infants"); the response or dependent variable Y is the infant's Birth Weight. Data on 5 other factors are included: Head Circumference, Length, Mother's Age, Gestational Weeks (the length of pregnancy), and Toxemia (0/1; toxemia is a pregnancy condition resulting from metabolic disorder). Answer the above 4 questions with X = Gestational Weeks.

#4.2 It has been generally known that respiratory function may decline with age. To study this possibility, We consider a data set consisting of age (years) and vital capacity (VC, liters) for each of 44 men working in the cadmium industry but have not been exposed to cadmium fumes. Answer the 4 questions (a)-(d) of Exercise 1.1 using dataset "Vital Capacity" with variables X = Age and Y = (100)(Vital Capacity) plus the following question:

e) Can we say that the correlation coefficient between Age and Y is the same as the correlation coefficient between Age and vital Capacity, why or why not?