BIOS 662 Fall 2018

Linear Regression, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Multiple linear regression
- Measures of association
- Parametric/large N
  - Pearson correlation coefficient
- Nonparametric (i.e., rank based)
  - Spearman rank correlation coefficient
  - Kendall's τ

Multiple Linear Regression

Reasons for using multiple linear regression rather than just simple linear regression include:

- Determining the best set of variables with which to predict an outcome variable
- Allowing adjustment for potential confounders when investigating an exposure–disease association
- Investigating potential interactions between exposures associated with a disease
- Using a categorical predictor with more than two categories

Some of these reasons may apply simultaneously

Multiple Linear Regression Model

- Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i,$$

$$i = 1, 2, \ldots, N$$

- Data are $(Y_i, \boldsymbol{X}_i)$; $i = 1, 2, \ldots, N$, where $\boldsymbol{X}_i$ is a vector of length k

- Assumptions:

1. Linearity: each X variable is linearly associated with Y
2. The values of each X variable are fixed constants
3. ϵ_i iid $N(0, \sigma^2)$

Multiple Linear Regression Model

Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i,$$

$$i = 1, 2, \ldots, N$$

Interpretation of parameters:

- β_j is the change in the expected value of Y when the j^{th} X variable increases by one unit, with all the other X variables being held contact
- If the j^{th} X variable is dichotomous, that is, takes on only values in $\{0, 1\}$, this corresponds to the difference between $E(Y)$ when the value of the j^{th} X is 1 versus when it is 0

Matrix Formulation

- Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \ldots & X_{Nk} \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Matrix Formulation

- Linear model
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
- The least squares estimators are the solutions to the set of equations:
$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$
- Therefore, as in the simple linear regression case:
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$
- The coefficient of determination is now written as R^2 (rather than r^2); as before it is the proportion of the total variation attributable to regression (that is, explained by all the X variables together)

Analysis of Variance

- ANOVA table:

Source	df	SS	MS	F
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Residual	$N-k-1$	SSE	MSE = SSE/$(N-k-1)$	
Total	$N-1$	SST		

- The F test is for

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

versus

$$H_A : \text{at least one } \beta_j \neq 0$$

Multiple Linear Regression Example

- Consider the SBP and age example and suppose we want to investigate whether the association varies with gender
- Let:

  Y_i be the systolic blood pressure of person i

  X_{i1} be the age of person i

  X_{i2} be 1 if person i is male and 0 otherwise

  $X_{i3} = X_{i1} \cdot X_{i2}$

Multiple Linear Regression Example

```
proc reg;
   model sbp = age male;
```

```
                     Dependent Variable: sbp

              Number of Observations Read          40
              Number of Observations Used          40


                     Analysis of Variance

                              Sum of          Mean
Source              DF       Squares        Square    F Value    Pr > F

Model                2    2414.90795    1207.45397     288.31    <.0001
Error               37     154.95563       4.18799
Corrected Total     39    2569.86358
```

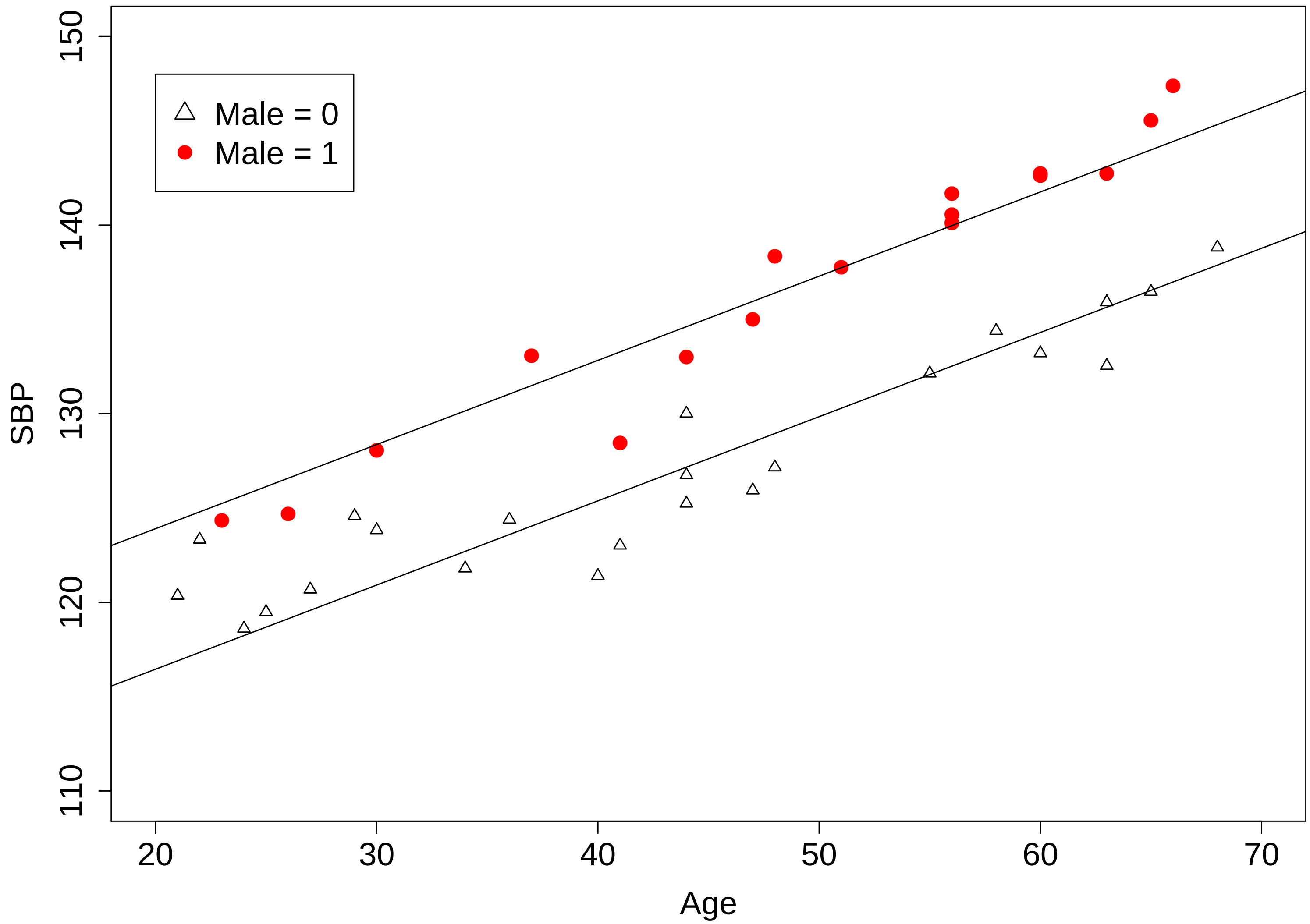
Multiple Linear Regression Example

Root MSE	2.04646	R-Square	0.9397
Dependent Mean	131.15651	Adj R-Sq	0.9364
Coeff Var	1.56032		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	107.52982	1.08737	98.89	<.0001
age	1	0.44634	0.02268	19.68	<.0001
male	1	7.44864	0.65488	11.37	<.0001

Multiple Linear Regression Example



Multiple Linear Regression Example

```
proc reg;
   model sbp = age male agemale;
```

Dependent Variable: sbp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2445.33277	815.11092	235.64	<.0001
Error	36	124.53081	3.45919		
Corrected Total	39	2569.86358			

Multiple Linear Regression Example

Root MSE	1.85989	R-Square	0.9515
Dependent Mean	131.15651	Adj R-Sq	0.9475
Coeff Var	1.41807		

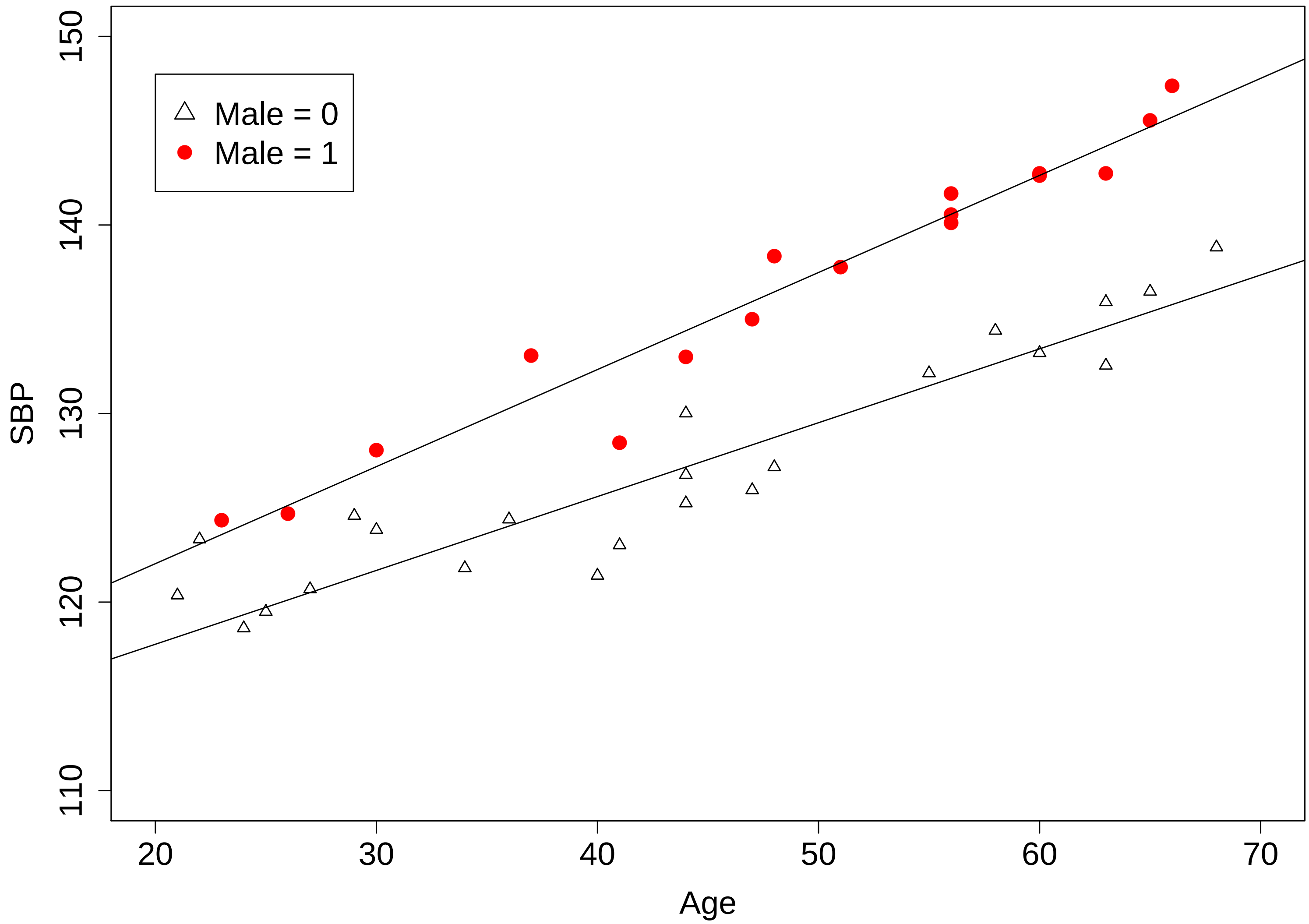
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	109.92861	1.27705	86.08	<.0001
age	1	0.39170	0.02765	14.17	<.0001
male	1	1.81501	1.99065	0.91	0.3680
agemale	1	0.12305	0.04149	2.97	0.0053

Multiple Linear Regression Example

```
> fit <- lm(sbp~age+male+agemale)
> summary(fit)

Call:
lm(formula = sbp ~ age + male + agemale)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.92860    1.27708  86.078  < 2e-16 ***
age           0.39170    0.02765  14.168  2.7e-16 ***
male          1.81503    1.99070   0.912  0.36797
agemale       0.12305    0.04149   2.966  0.00533 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.86 on 36 degrees of freedom
Multiple R-squared: 0.9515,     Adjusted R-squared: 0.9475
F-statistic: 235.6 on 3 and 36 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression Example



Multiple Linear Regression Example

Now suppose we use age in 10-year age groups

```
data sbp;
   set sbp;

agegroup=10*floor(age/10);

if 20 le age lt 30 then age2029=1;
   else age2029=0;
if 30 le age lt 40 then age3039=1;
   else age3039=0;
if 40 le age lt 50 then age4049=1;
   else age4049=0;
if 50 le age lt 60 then age5059=1;
   else age5059=0;
if 60 le age lt 70 then age6069=1;
   else age6069=0;
```

Multiple Linear Regression Example

```
proc reg;
   model sbp = agegroup;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1785.32369	1785.32369	86.47	<.0001
Error	38	784.53989	20.64579		
Corrected Total	39	2569.86358			

Root MSE	4.54376	R-Square	0.6947
Dependent Mean	131.15651	Adj R-Sq	0.6867
Coeff Var	3.46438		

Multiple Linear Regression Example

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	111.95282	2.18650	51.20	<.0001
agegroup	1	0.46554	0.05006	9.30	<.0001

Assumption here: SBP changes by the same amount from each age group to the next.

Multiple Linear Regression Example

```
proc reg;
   model sbp = age3039 age4049 age5059 age6069;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1873.06457	468.26614	23.52	<.0001
Error	35	696.79901	19.90854		
Corrected Total	39	2569.86358			

Root MSE	4.46190	R-Square	0.7289
Dependent Mean	131.15651	Adj R-Sq	0.6979
Coeff Var	3.40197		

Multiple Linear Regression Example

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	122.01318	1.57752	77.34	<.0001
age3039	1	4.21824	2.54367	1.66	0.1062
age4049	1	6.56457	2.07327	3.17	0.0032
age5059	1	15.76066	2.40970	6.54	<.0001
age6069	1	17.78679	2.11646	8.40	<.0001

Multiple Linear Regression Example

```
proc reg;
   model sbp = age2029 age3039 age4049 age5059;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	139.79997	1.41098	99.08	<.0001
age2029	1	-17.78679	2.11646	-8.40	<.0001
age3039	1	-13.56855	2.44388	-5.55	<.0001
age4049	1	-11.22222	1.94954	-5.76	<.0001
age5059	1	-2.02613	2.30411	-0.88	0.3852

Multiple Linear Regression Example

```
proc glm;
   class agegroup;
   model sbp = agegroup / solution;
   lsmeans agegroup;
```

```
           The GLM Procedure

        Class Level Information

Class          Levels    Values

agegroup            5    20 30 40 50 60


Number of Observations Read          40
Number of Observations Used          40
```

Multiple Linear Regression Example

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		139.7999661	B	1.41097637	99.08	<.0001
agegroup	20	-17.7867909	B	2.11646455	-8.40	<.0001
agegroup	30	-13.5685533	B	2.44388276	-5.55	<.0001
agegroup	40	-11.2222160	B	1.94954402	-5.76	<.0001
agegroup	50	-2.0261338	B	2.30411476	-0.88	0.3852
agegroup	60	0.0000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Multiple Linear Regression Example

```
   The GLM Procedure
  Least Squares Means

agegroup       sbp LSMEAN

20             122.013175
30             126.231413
40             128.577750
50             137.773832
60             139.799966
```

Correlation

- The *correlation* between random variables X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Note:

$$\rho = \frac{\beta_{y \cdot x} \sigma_X}{\sigma_Y} = \frac{\beta_{x \cdot y} \sigma_Y}{\sigma_X}$$

Correlation

- Estimate ρ by
$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (X_i - \bar{X})^2}} = \frac{[XY]}{\sqrt{[X^2][Y^2]}}$$
the *sample Pearson product moment correlation coefficient*

- One can show that
$$r = \hat{\beta}_{y \cdot x} \frac{s_X}{s_Y} = \mathrm{sign}(\hat{\beta}_{y \cdot x}) \sqrt{r^2}$$
where r^2 is as in the first set of notes on regression, i.e., the proportion of total variation attributable to regression

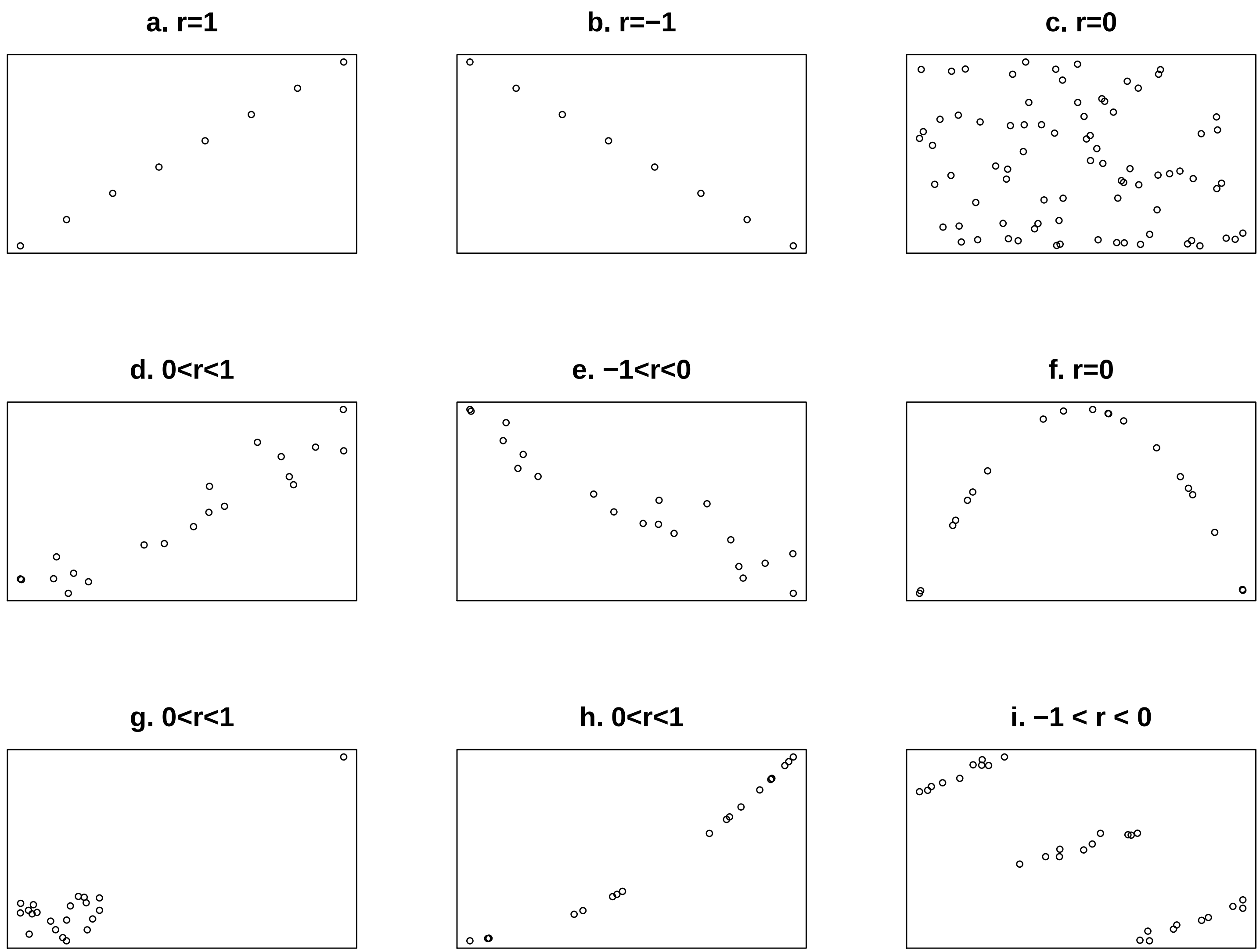
Correlation

- The correlation coefficient r has the following properties:
  - $r \in [-1, 1]$
  - $r = 1$ iff all observations lie on a straight line with positive slope
  - $r = -1$ iff all observations lie on a straight line with negative slope
  - it is invariant under multiplication and addition of constants to X or Y
  - it measures *linear association* between two variables
  - it tends to be close to zero if there is no linear association, even if there is a strong non-linear association

Demonstrating Correlation Properties Using R

```
> x <- 1:11
> y <- x

> cor(y,x)
[1] 1

> cor(y,3*x)
[1] 1

> cor(y/100,3*x+10)
[1] 1

> cor(y,x^2)
[1] 0.9739695

> x <- c(-5:5)
> cor(y,x^2)
[1] 0
```

Correlation: Figure 9.11



Correlation

- The test statistic

$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}} \sim t_{N-2}$$

  can be used to test $H_0 : \rho = 0$

- Claim: this test is equivalent to testing $H_0 : \beta_{y \cdot x} = 0$
- Proof of claim on next couple of pages

Correlation

- First note that

$$\begin{aligned}(N-2)s_{y\cdot x}^2 &= \text{SSE} = \text{SST} - \text{SSR} \\ &= \text{SST}\left(1 - \frac{\text{SSR}}{\text{SST}}\right) \\ &= [Y^2]\left(1 - \frac{[XY]^2}{[Y^2][X^2]}\right) \\ &= (N-1)s_Y^2(1-r^2)\end{aligned}$$

- Next recall that

$$\hat{\beta}_{y\cdot x} = \frac{[XY]}{[X^2]}$$

Correlation

- Then

$$
\begin{aligned}
t &= \frac{\hat{\beta}_{y\cdot x}}{s_{y\cdot x}/\sqrt{[X^2]}} = \frac{[XY]/[X^2]}{s_{y\cdot x}/\sqrt{[X^2]}} \\
&= \frac{[XY]/\sqrt{[X^2]}}{s_{y\cdot x}} = \frac{r\sqrt{[Y^2]}}{s_{y\cdot x}} \\
&= \frac{r s_Y \sqrt{N-1}}{\sqrt{(1-r^2)s_Y^2(N-1)/(N-2)}} \\
&= \frac{r}{\sqrt{(1-r^2)/(N-2)}}
\end{aligned}
$$

Correlation

- In general,

$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}} \sim t_{N-2} \tag{1}$$

if

1. (X, Y) bivariate normal (Section 9.3.3 of the text), or
2. $Y|X$ is normally distributed with constant variance (that is, the usual regression model holds)

- (1) holds approximately for large N (cf. Graybill, 1976, Section 6.10)

Correlation Example

- Cholesterol was measured in 100 spouse pairs
- If there is no environmental effect (e.g., shared diet) on cholesterol we would expect $\rho = 0$
- $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$
- $t_{98,0.975} = 1.98$, so $C_{0.05} = \{t : |t| > 1.98\}$
- Observed $r = 0.25$, so that

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} = \frac{0.25}{\sqrt{(1 - 0.25^2)/98}} = 2.556$$

- $p = 2 \times \{1 - F_{t_{98}}(2.556)\} = 0.0121$

Correlation Example: SAS

```
proc corr;
   var x y;
```

```
              Pearson Correlation Coefficients, N = 100
                     Prob > |r| under H0: Rho=0

                                 x              y

              x            1.00000        0.25000
                                           0.0121

              y            0.25000        1.00000
                            0.0121
```

Correlation Using Fisher's Transformation

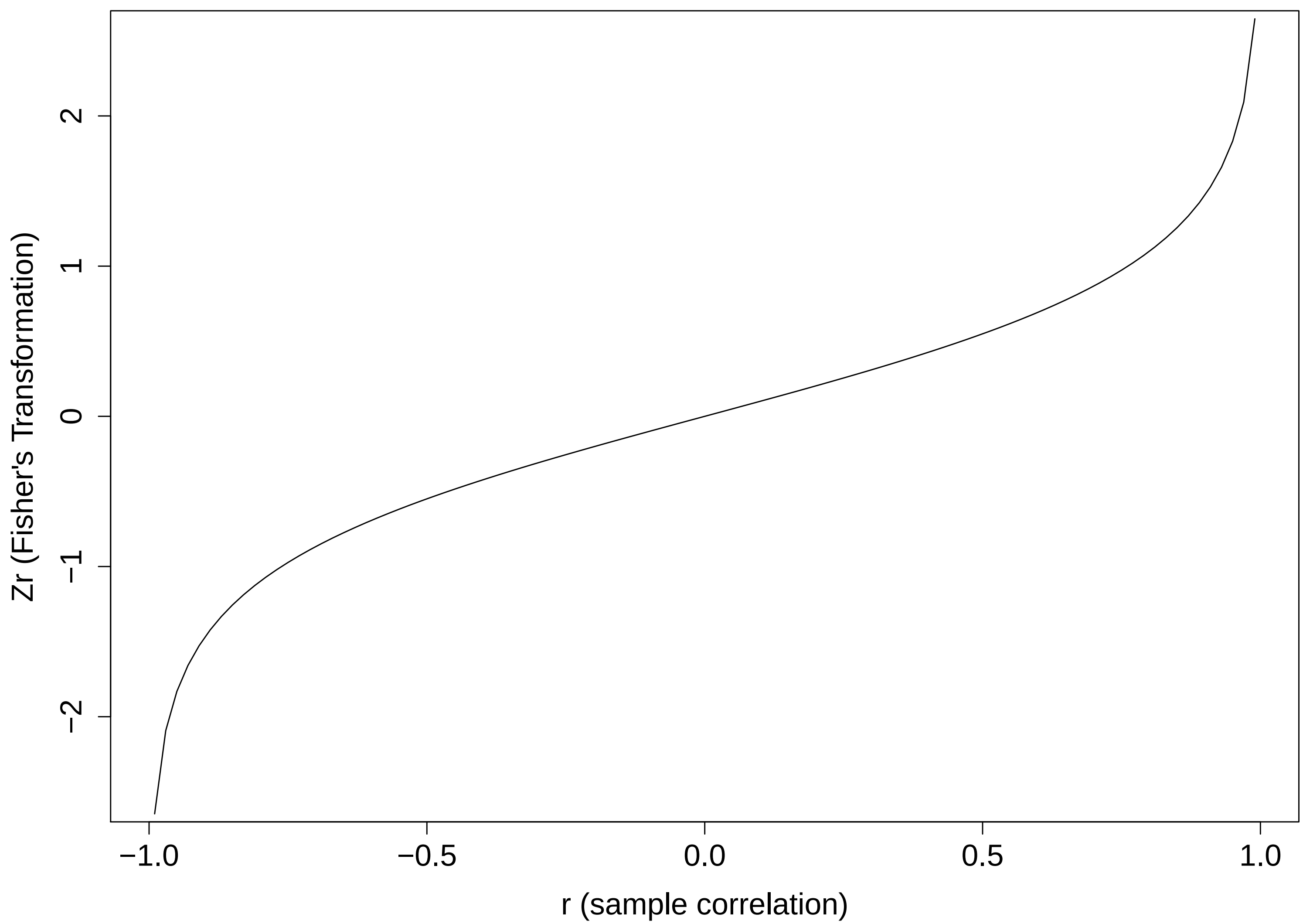
- R. A. Fisher developed a test of $H_0 : \rho = \rho_0$
- He showed that

$$z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{N-3}\right)$$

- Under $H_0 : \rho = \rho_0$

$$z = \frac{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{1/(N-3)}} \sim N(0,1)$$

Correlation: Fisher's Transformation



Using Fisher's Transformation: Example

- Cholesterol example
- $N = 100, \;\; r = 0.25$
- $H_0 : \rho = 0$

$$z_r = \frac{1}{2} \log \left(\frac{1.25}{0.75} \right) = 0.2554$$

$$z = \frac{0.2554 - 0}{\sqrt{1/97}} = 2.5155$$

$$p = 2 \times \{1 - \Phi(2.515)\} = 0.0119$$

Correlation Using Fisher's Transformation

- The Fisher transformation can be used for a CI for ρ

$$z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \;\Rightarrow\; e^{2z_r} = \frac{1+r}{1-r} \;\Rightarrow\; r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$$

$$z_L = z_r - z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$z_U = z_r + z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$r_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}; \quad r_U = \frac{e^{2z_U}-1}{e^{2z_U}+1}$$

Using Fisher's Transformation: Example

- 95% CI when $r = 0.25$ and $n = 100$

$$(z_L, z_U) = 0.2554 \pm 1.96/\sqrt{97} = (0.0564, 0.4544)$$

$$r_L = \frac{e^{2\times 0.0564} - 1}{e^{2\times 0.0564} + 1} = 0.0563$$

$$r_U = \frac{e^{2\times 0.4544} - 1}{e^{2\times 0.4544} + 1} = 0.4255$$

Correlation Using Fisher's Transformation: SAS

```
proc corr fisher(biasadj=no);
  var x y;
```

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z
x	y	100	0.25000	0.25541

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	95% Confidence Limits		p Value for H0:Rho=0
x	y	0.056350	0.425524	0.0119

Correlation Using Fisher's Transformation: R

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 2.556, df = 98, p-value = 0.01212
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05634962 0.42552363
sample estimates:
      cor
0.2500007
```

Correlation Using Fisher's Transformation

- Comparing two correlations: Two independent samples

$$H_0 : \rho_1 = \rho_2 \text{ vs. } H_A : \rho_1 \neq \rho_2$$

- Compute z_{r_1} and z_{r_2}

$$\text{Var}(z_{r_1} - z_{r_2}) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

- Thus under H_0

$$z = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \sim N(0, 1)$$

Using Fisher's Transformation: Example

- If blood pressure level is inherited, one would expect the correlation between blood pressure of mothers and their natural children to be greater than between mothers and their adopted children
- In a study, 1000 mothers and one of their randomly chosen natural children had their blood pressure measured
- In a separate sample, 100 mothers and their adopted children also had their BP measured

Using Fisher's Transformation: Example cont.

- Let

  ρ_1 = population correlation for natural pairs

  ρ_2 = population correlation for adopted pairs

- Hypotheses

$$H_0 : \rho_1 = \rho_2 \text{ vs. } H_A : \rho_1 > \rho_2$$

- Critical region

$$C_{0.05} = \{z : z > 1.645\}$$

Using Fisher's Transformation: Example cont.

- $r_1 = 0.32; \;\; r_2 = 0.06$
- $z_{r_1} = 0.3316; \;\; z_{r_2} = 0.0601$
- Thus

$$z = \frac{0.3316 - 0.0601}{\sqrt{\frac{1}{997} + \frac{1}{97}}} = 2.55$$

- So we reject the null hypothesis and conclude that blood pressure levels appear to have an inherited component

Correlation Homogeneity

- Testing the homogeneity of k correlations
- Fisher's transformation can be used to test the hypothesis that several correlations are equal

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$$

vs.

$$H_A : \text{ at least one inequality}$$

Correlation Homogeneity

- Let

$$T_1 = \sum_{i=1}^{k} (n_i - 3) z_{r_i}$$

and

$$T_2 = \sum_{i=1}^{k} (n_i - 3) z_{r_i}^2$$

- Under H_0

$$H = T_2 - \frac{T_1^2}{\sum (n_i - 3)} \sim \chi^2_{k-1}$$

- Cf. Graybill (1976, p. 405)

Correlation Homogeneity: Example

- Does the correlation between LDL-cholesterol and HDL-cholesterol change with age in women not taking hormones?

Age	n	r	z_r
20-29	277	-0.08	-0.0802
30-39	479	-0.25	-0.2554
40-49	508	-0.19	-0.1923
50-59	373	-0.18	-0.1820
60-69	216	-0.15	-0.1511

Correlation Homogeneity: Example cont.

- Null hypothesis $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$
- Critical region $C_{0.05} = \{H : H > 9.49\}$
- Compute test statistic

$$T_1 = 274(-0.0802) + \cdots + 213(-0.1511) = -340.200$$

$$T_2 = 274(-0.0802)^2 + \cdots + 213(-0.1511)^2 = 68.614$$

$$H = 68.614 - \frac{(-340.200)^2}{1838} = 5.65$$

- So we do not reject the null hypothesis; the correlation between LDL-cholesterol and HDL-cholesterol does not appear to change with age

Rank Correlation Coefficients

- Using ranks makes statistics robust to outliers
- Spearman rank correlation, Kendall's τ
- Nonparametric measures of association

Spearman Rank Correlation

1. Ys and Xs are ranked from 1 to N separately
2. The correlation of the ranks is then computed

Spearman Correlation: Example

- Ten children are ranked according to their mathematical and musical abilities

Child	Math	Music
A	7	5
B	4	7
C	3	3
D	10	10
E	6	1
F	2	9
G	9	6
H	8	2
I	1	8
J	5	4

Spearman Correlation

- Let R_{1i} and R_{2i} be the ranks of the Y_i and X_i, respectively
- Spearman correlation coefficient

$$r_s = \frac{\sum(R_{1i}-\bar{R}_1)(R_{2i}-\bar{R}_2)}{\sqrt{\sum_i(R_{1i}-\bar{R}_1)^2 \sum_i(R_{2i}-\bar{R}_2)^2}}$$

$$= 1 - \frac{6\sum d_i^2}{N^3-N}$$

where $d_i = R_{1i} - R_{2i}$

- The form of r_s containing $\sum d_i^2$ is not correct if ties are present
- Note:

$R_{1i} = R_{2i}$ for all $i \Rightarrow d_i = 0$ for all $i \Rightarrow r_s = 1$

Spearman Correlation

- Suppose N is odd and $N = 2m + 1$
- Then the most extreme discordant rankings are

i	1	2	$\cdots$						N
R_{1i}	1	2	$\cdots$	m	$m+1$	$m+2$	$\cdots$	$2m$	$2m+1$
R_{2i}	$2m+1$	$2m$	$\cdots$	$m+2$	$m+1$	m	$\cdots$	2	1
d_i	$-2m$	$2-2m$	$\cdots$	-2	0	2	$\cdots$	$2m-2$	$2m$

Spearman Correlation

- Under this configuration

$$
\begin{aligned}
\sum_{i=1}^{N} d_i^2 &= 4m^2 + 4(m-1)^2 + \cdots + 4(1)^2 + 0 \\
&\qquad + 4(1)^2 + \cdots + 4(m-1)^2 + 4m^2 \\
&= 8\sum_{j=1}^{m} j^2 \\
&= 8m(m+1)(2m+1)/6 \\
&= \left(4 \times \tfrac{N-1}{2} \times \tfrac{N+1}{2} \times N\right)/3 \\
&= (N^3 - N)/3
\end{aligned}
$$

Spearman Correlation

- Thus
$$r_s = 1 - \frac{6(N^3 - N)}{3(N^3 - N)} = 1 - 2 = -1$$
- In a similar way, it can be shown that if N is even, the most extreme rankings give $r_s = -1$
- So:

$r_s = 1$ if perfect agreement in the ranks

$r_s = -1$ if perfect disagreement in the ranks

Spearman Correlation

Child	Math	Music	d
A	7	5	2
B	4	7	−3
C	3	3	0
D	10	10	0
E	6	1	5
F	2	9	−7
G	9	6	3
H	8	2	6
I	1	8	−7
J	5	4	1

- Spearman correlation

$$r_s = 1 - \frac{6(2^2 + (-3)^2 + \cdots + 1^2)}{10^3 - 10} = 1 - \frac{6(182)}{990} = -0.103$$

Spearman Correlation: SAS and R

```
proc corr spearman;
   var math music;
```

```
              Spearman Correlation Coefficients, N = 10
                      Prob > |r| under H0: Rho=0

                               math          music

                math         1.00000       -0.10303
                                             0.7770

                music       -0.10303        1.00000
                              0.7770
```

```
> cor(math,music,method="spearman")
[1] -0.1030303
```

Spearman Correlation

- The Spearman correlation coefficient can be used to test the null hypothesis of independence

$$H_0 : X \perp Y \;\text{ vs. }\; H_A : X \not\perp Y$$

that is, $H_A : X$ and Y not independent

- Distribution of r_s under H_0 is derived using a permutation-based argument
- We can list the R_{1i} in ascending order
- There are $N!$ possible orderings of the R_{2i}
- Under H_0, each of these orderings is equally likely

Spearman Correlation

- Example: $N = 3$

R_{1i}	1	2	3	$\sum d_i^2$	r_s
R_{2i}	1	2	3	0	1.0
R_{2i}	1	3	2	2	0.5
R_{2i}	2	1	3	2	0.5
R_{2i}	2	3	1	6	-0.5
R_{2i}	3	1	2	6	-0.5
R_{2i}	3	2	1	8	-1.0

Spearman Correlation

- CDF of r_s

k	$\Pr[r_s \leq k]$
-1.0	$1/6$
-0.5	$1/2$
0.5	$5/6$
1.0	1

- Text Table A.12, p. 838, gives the two sided critical values for testing $H_0 : X \perp Y$
- If N is large (> 10; Neter et al. 1996, page 652),

$$t_s = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t_{N-2}$$

Spearman Correlation: Example

- Example: math (X) and music (Y)
- $N = 10; \;\; r_s = -0.1030$
- From Table A.12, $C_{0.05} = \{r_s : |r_s| > 0.648\}$
- Assume $N = 10$ is large enough to use the t approximation
- $C_{0.05} = \{t_s : |t_s| > t_{8,0.975} = 2.306\}$
- $t_s = \frac{-0.1030\sqrt{8}}{\sqrt{1-(-0.1030)^2}} = -0.2930$
- $p = 2 \times \Pr[t_8 < -0.2929] = 0.7771$

Spearman Correlation: Ties

- In the presence of ties, ranks are replaced by midranks
- However, critical values in Table A.12 are only approximate
- If N is large, use t_s as before; i.e.,

$$t_s = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t_{N-2}$$

Kendall's τ

- Kendall's τ: Another rank correlation statistic
- Data: (X_i, Y_i) for $i = 1, 2, \ldots, N$
- Definitions: Two pairs of observations are
  concordant if $(X_i - X_j)(Y_i - Y_j) > 0$
  discordant if $(X_i - X_j)(Y_i - Y_j) < 0$

Kendall's τ

- Let p_c be the probability that a randomly chosen pair of observations is concordant; and p_d the probability that they are discordant; then

$$\tau = p_c - p_d$$

- Note:

  $-1 \leq \tau \leq 1$

  if X and Y are independent, $\tau = 0$

Kendall's τ

- There are $\binom{N}{2}$ pairs of observations
- Let P be the number of concordant pairs
- Let Q be the number of discordant pairs
- The estimate of τ is
$$r_k = \frac{P - Q}{\binom{N}{2}} = 1 - \frac{2Q}{\binom{N}{2}} = \frac{2P}{\binom{N}{2}} - 1$$
- The last two terms assume no ties, so that $P + Q = \binom{N}{2}$
- Replacing Xs and Ys with their ranks does not change τ

Kendall's $\boldsymbol{\tau}$

- $H_0 : \tau = 0$ vs. $H_A : \tau \neq 0$
- The distribution of r_k under H_0 is computed using permutation principles
- As with r_s, there are $N!$ equally likely outcomes
- Kendall, *Rank Correlation Methods*, Hafner Publishing, 1962, gives a table of the distribution of $P - Q$ for $4 \leq N \leq 10$

Kendall's $\boldsymbol{\tau}$

- Upper one-sided critical values of r_k
- Note that the distribution of r_k is symmetric about 0

N	0.05	0.025
5	0.80	1.00
6	0.73	0.87
7	0.62	0.71
8	0.57	0.64
9	0.50	0.56
10	0.42	0.51

Kendall's τ: Example

- Cigarette consumption and lung cancer mortality in England and Wales, 1930-1969

Period	$\log_{10}$ mortality	$\log_{10}$ tobacco (lb/person)
1930-34	-2.35	-0.26
1935-39	-2.20	-0.03
1940-44	-2.12	0.30
1945-49	-1.95	0.37
1950-54	-1.85	0.40
1955-59	-1.80	0.50
1960-64	-1.70	0.55
1965-69	-1.58	0.55

Kendall's $\boldsymbol{\tau}$

- $C_{0.05} = \{r_k : |r_k| \geq 0.64\}$
- Observation 1: $(-2.35, -0.26)$

  Observation 2: $(-2.20, -0.03)$

  $\{-2.35 - (-2.2)\}\{-0.26 - (-0.03)\} > 0 \;\Rightarrow\;$ concordant
- Observation 1 and observation 3:

  $\{-2.35 - (-2.12)\}(-0.26 - 0.3) > 0 \;\Rightarrow\;$ concordant
- $P - Q = 27 \;\Rightarrow$

$$r_k = \frac{27}{\binom{8}{2}} = \frac{27}{28} = 0.96$$

Kendall's $\boldsymbol{\tau}$

- If N is sufficiently large (≥ 10), under $H_0 : \tau = 0$

$$r_k \sim N\left(0,\ \frac{2(2N+5)}{9N(N-1)}\right)$$

$$P - Q \sim N\left(0,\ \frac{N(N-1)(2N+5)}{18}\right)$$

or

$$Z = \frac{P - Q}{\sqrt{\frac{N(N-1)(2N+5)}{18}}} \sim N(0,1)$$

Kendall's $\boldsymbol{\tau}$

- If there are tied observations, r_k cannot be 1 or -1.
- Let

$$t_x = \frac{1}{2}\sum_i t_{xi}(t_{xi} - 1) \quad \text{and} \quad t_y = \frac{1}{2}\sum_i t_{yi}(t_{yi} - 1)$$

where t_{zi} denotes the number of observations in the i^{th} set of ties for $z = x, y$

Kendall's $\boldsymbol{\tau}$

- Let

$$W = \sqrt{\left(\frac{1}{2}N(N-1) - t_x\right)\left(\frac{1}{2}N(N-1) - t_y\right)}$$

- Define

$$r_{k_b} = \frac{P - Q}{W}$$

This statistic is known as *Kendall's* τ_b

Kendall's $\boldsymbol{\tau}$: Tobacco Example Revisited

- Recall that $N = 8$ and there was one set of ties (of size 2) for the tobacco variable
- Thus

$$W = \sqrt{\left(\frac{1}{2}8(8-1)\right)\left(\frac{1}{2}8(8-1)-1\right)}$$

- Yielding

$$r_{k_b} = \frac{27}{\sqrt{28 \times 27}} = 0.98198$$

Kendall's τ: Tobacco Example cont.

- SAS

```
proc corr kendall;
   var mortality tobacco;
```

```
Kendall Tau b Correlation Coefficients, N = 8
         Prob > |tau| under H0: Tau=0

               mortality       tobacco

mortality        1.00000       0.98198
                                0.0008

tobacco          0.98198       1.00000
                  0.0008
```

Kendall's τ: Tobacco Example cont.

- R

```
> cor(mortality, tobacco, method="kendall")
[1] 0.9819805

> cor.test(mortality, tobacco, method="kendall")

        Kendall's rank correlation tau

data:  mortality and tobacco
z = 3.3662, p-value = 0.000762
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.9819805

Warning message:
In cor.test.default(mortality, tobacco, method = "kendall") :
  Cannot compute exact p-value with ties
```

Kendall's $\boldsymbol{\tau}$

- Kendall's score $P - Q$

$$r_{k_a} = \frac{P - Q}{\binom{N}{2}}$$

and

$$r_{k_b} = \frac{P - Q}{W}$$

- Tests based on r_{k_a} and r_{k_b} are equivalent
- Asymptotic variance of $P - Q$ under H_0 is given on page 336 of the text

$$Z = \frac{P - Q}{\sqrt{\mathrm{Var}(P - Q)}} \sim N(0, 1)$$

Kendall's τ: Example

- In general, $\text{Var}(P-Q)$ equals

$$\frac{N(N-1)(2N+5)}{18} - \sum_i \frac{t_{xi}(t_{xi}-1)(2t_{xi}+5)}{18} - \cdots$$

- For tobacco example, $\text{Var}(P-Q)$ is

$$\frac{8(8-1)(2\cdot 8+5)}{18} - 0 - \frac{2(2-1)(2\cdot 2+5)}{18} + 0 + 0 = 64.333$$

- Thus

$$z = \frac{27}{\sqrt{64.333}} = 3.366$$

yielding $p = 2 \cdot \{1 - \Phi(3.366)\} = 0.0008$

Correlation: Summary/Remarks

- r is appropriate if (X, Y) bivariate normal; sensitive to outliers, major(?) departures from normality
- Nonparametric alternatives: r_s and r_k
- If (X, Y) bivariate normal with correlation ρ,

$$r \xrightarrow{p} \rho \qquad r_s \xrightarrow{p} \frac{6}{\pi}\arcsin(\rho/2) \qquad r_k \xrightarrow{p} \frac{2}{\pi}\arcsin(\rho)$$

  (Kraemer, 1998 “Rank Correlation” *Encyclopedia of Biostatistics*)

- ARE of r_s and r_k compared to r: $9/\pi^2 = 0.912$ (Conover, 1980 *Practical Nonparametric Statistics*)