ORIGIN ≡ 1

Association and Correlation in Simple Regression

Once it is determined by ANOVA F or t-tests (testing H_0: $\beta = 0$) that an association between variables exists, summary statistics such as the coefficient of determination (r^2 or R^2) and coefficient of correlation (r) may prove helpful. More usefully, further inferences may be made concerning the degree of association. This may be done by testing of H_0: $\beta = \beta_0$ or providing confidence limits on β. Alternatively, one can test and provide confidence limits on the population correlation coefficient (ρ) via its sample estimate (r). Tests using correlation are particularly useful when the reseacher is unwilling to specify in advance which of two variables (X or Y) should be considered independent versus dependent.

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX19.1aR.txt")

X := ZAR⟨2⟩ **^Zar Example 19.1a**

Y := ZAR⟨3⟩

$X_{bar} := mean(X)$ $X_{bar} = 10.6833$

$Y_{bar} := mean(Y)$ $Y_{bar} = 7.5667$

$s_X := \sqrt{Var(X)}$ $s_X = 0.395$ $s_X^2 = 0.1561$

$s_Y := \sqrt{Var(Y)}$ $s_Y = 0.3499$ $s_Y^2 = 0.1224$

$n := length(Y)$ $n = 12$

$i := 1 .. n$

$$X = \begin{pmatrix} 10.4 \\ 10.8 \\ 11.1 \\ 10.2 \\ 10.3 \\ 10.2 \\ 10.7 \\ 10.5 \\ 10.8 \\ 11.2 \\ 10.6 \\ 11.4 \end{pmatrix} \quad Y = \begin{pmatrix} 7.4 \\ 7.6 \\ 7.9 \\ 7.2 \\ 7.4 \\ 7.1 \\ 7.4 \\ 7.2 \\ 7.8 \\ 7.7 \\ 7.8 \\ 8.3 \end{pmatrix}$$

ZAR =

	1	2	3
1	1	10.4	7.4
2	2	10.8	7.6
3	3	11.1	7.9
4	4	10.2	7.2
5	5	10.3	7.4
6	6	10.2	7.1
7	7	10.7	7.4
8	8	10.5	7.2
9	9	10.8	7.8
10	10	11.2	7.7
11	11	10.6	7.8
12	12	11.4	8.3

Corrected Sums of Squares:

$$L_{xx} := \sum_i \left(X_i - X_{bar}\right)^2 \qquad L_{yy} := \sum_i \left(Y_i - Y_{bar}\right)^2 \qquad L_{xy} := \sum_i \left(X_i - X_{bar}\right) \cdot \left(Y_i - Y_{bar}\right)$$

$L_{xx} = 1.7167$ $L_{yy} = 1.3467$ $L_{xy} = 1.3233$

Estimated Regression Coefficients for $Y = \alpha + \beta X$:

$b := \frac{L_{xy}}{L_{xx}}$ $b = 0.7709$ **< sample estimate of β**

$a := Y_{bar} - b \cdot X_{bar}$ $a = -0.6688$ **< sample estimate of α**

Estimated values of Y (Y_{hat}):

$Y_{hat_i} := a + b \cdot X_i$

Residuals:

$\varepsilon_i := Y_{hat_i} - Y_i$

Sum of Squares Decomposition of the Regression:

$$SS_T := \sum_i \left(Y_i - Y_{bar}\right)^2 \qquad SS_T = 1.3467$$

$$SS_R := \sum_i \left(Y_{hat_i} - Y_{bar}\right)^2 \qquad SS_R = 1.0201$$

$$SS_E := \sum_i \left(Y_i - Y_{hat_i}\right)^2 \qquad SS_E = 0.3265$$

Coefficient of Determination (R^2) and Coefficient of Correlation (r):

Values are defined by using corrected sums of squares or Sums of Squares in the Regression ANOVA table:

ANOVA Table for Linear Regression:

	SS	df	MS	
Regression:	$SS_R = 1.0201$	$df_R := 1$	$MS_R := \frac{SS_R}{df_R}$	$MS_R = 1.0201$
Residual:	$SS_E = 0.3265$	$df_E := n - 2$	$MS_E := \frac{SS_E}{df_E}$	$MS_E = 0.0327$
TOTAL:	$SS_T = 1.3467$	$df_T := n - 1$	$MS_T := \frac{SS_T}{df_T}$	$MS_T = 0.1224$

Coefficient of Correlation:

$r := \frac{L_{xy}}{\sqrt{L_{xx} \cdot L_{yy}}}$ $r = 0.8704$

$r := \sqrt{\frac{SS_R}{SS_T}}$ $r = 0.8704$ $r := \sqrt{1 - \frac{SS_E}{SS_T}}$ $r = 0.8704$ **< all equivalent**

Coefficient of Determination:

$R := r$ $R^2 = 0.7575$ **< Coefficient of Determination R^2 is r^2**

$R^2 := \frac{L_{xy}^2}{L_{xx} \cdot L_{yy}}$ $R^2 := \frac{SS_R}{SS_T}$ $R^2 := 1 - \frac{SS_E}{SS_T}$ **< equivalent**

Note: Values of r and R^2 range between -1 and 1. The closer R^2 is to -1 or 1, the stronger the linear relationship between the variables is *potentially* observed. R^2 or r near zero suggests no association. However, no single number can capture the situation exactly. It is possible, for data to show non-linear relationships, and for there to be high correlation/determination without necessarily a "good"regression fit or precision in prediction.

Correlation Coefficient Related to Sample Covariance & Standard Deviation:

$s_{xy} := \frac{L_{xy}}{(n-1)}$ $s_{xy} = 0.1203$ **< covariance of x with y**

$s_x := \sqrt{\frac{L_{xx}}{(n-1)}}$ $s_x = 0.395$ $s_X = 0.395$ **< calculated above for comparison**

$s_y := \sqrt{\frac{L_{yy}}{(n-1)}}$ $s_y = 0.3499$ $s_Y = 0.3499$ **< calculated above for comparison**

$r := \frac{s_{xy}}{s_x \cdot s_y}$ $r = 0.8704$ **< correlation coefficient in terms of covariance & standard deviations**

Correlation coefficient related to regression observed slope b (as estimate of β):

$b := r \cdot \sqrt{\frac{L_{yy}}{L_{xx}}}$ $b = 0.7709$ $b := r \cdot \frac{s_y}{s_x}$ $b = 0.7709$ **< equivalent**

$r := b \cdot \sqrt{\frac{L_{xx}}{L_{yy}}}$ $r = 0.8704$ $r := b \cdot \frac{s_x}{s_y}$ $r = 0.8704$ **< equivalent**

Adjusted Coefficient of Determination R_a^2:

$R_a := \sqrt{1 - \frac{MS_E}{MS_T}}$ $R_a^2 = 0.7333$ **< Adjusted Coefficient of Determination is the squared value.**

Note: R_a^2 is generally a better way of determining relative fit of a regression than using non-adjusted R^2.

One sample t-Test for ρ = 0:

This test, using ρ instead of β, is an equivalent alternative to the previous H_0: β=0 t-test.

Assumptions:

- **Variables X & Y are random samples from a Normal Distribution without Independent versus Dependent relationship.**
- **$Y_1, Y_2, Y_3, \ldots, Y_n$**
- **$X_1, X_2, X_3, \ldots, X_n$ with each value of X_i matched to Y_i**

Model:

$\rho = \beta(\sigma_x/\sigma_y)$ **< correlation coefficient ρ defined in terms of Regression slope β and standard deviations σ_x & σ_y.**

Hypotheses:

H_0: ρ = 0 **< No correlation**

H_1: $\rho \neq 0$ **< Two sided test (One Sided tests are also possible but rare)**

Standard Error of r (estimate of ρ):

$s_r := \sqrt{\frac{\left(1 - r^2\right)}{(n-2)}}$ $s_r = 0.1557$ **< standard error of r**

Test Statistic:

$t := \frac{r}{s_r}$ $t = 5.5893$

Sampling Distribution of Test Statistic t:

If Assumptions hold and H_0 is true, then $t \sim t_{(n-2)}$

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$C_1 := qt\left(\frac{\alpha}{2}, n-2\right)$ $C_1 = -2.2281$ $C_2 := qt\left(1-\frac{\alpha}{2}, n-2\right)$ $C_2 = 2.2281$

$C := |C_1|$

Decision Rule:

IF |t| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

$P := 2 \cdot pt(t, n-2)$ $P = 1.9998$ **< if** $t \leq 0$

$P := 2 \cdot (1 - pt(t, n-2))$ $P = 0.00023111$ **< if t > 0** **< Two sided case**

Prototype in R:

```
#SIMPLE LINEAR REGRESSION
#ZAR EXAMPLE 17.1
ZAR=read.table("c:/DATA/Biostatistics/ZarEX19.1aR.txt")
ZAR
attach(ZAR)

X=winglen
Y=taillen

#CREATING LINEAR MODEL LM:
LM=lm(Y~X)
Yhat=fitted(LM)
e=residuals(LM)
RESULTS=data.frame(X,Y,Yhat,e)
RESULTS

#ANOVA F-TEST FOR SLOPE:
anova(LM)

#t-TESTS FOR SLOPE & INTERCEPT:
summary(LM)
```

t statistic & P values are identical to the t-test for H_0: $\beta = 0$ **>**

$r = 0.8704$

$r^2 = 0.7575$

> summary(LM)

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26495 -0.11544  0.00903  0.13248  0.29757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6688     1.4744  -0.454 0.659773
X             0.7709     0.1379   5.589 0.000231 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1807 on 10 degrees of freedom
Multiple R-squared: 0.7575,     Adjusted R-squared: 0.7333
F-statistic: 31.24 on 1 and 10 DF,  p-value: 0.0002311
```

^ R^2 **^ R_a^2**

cor.test(X,Y, method="pearson",alternative="two.sided",conf.level=0.95)

> cor.test(X,Y, method="pearson",alternative="two.sided",conf.level=0.95)

$t = 5.5893$ **< t statistic above**
P value corresponds also >

CI corresponds to Fisher's
Z CI_ρ below >

$r = 0.8703546$ **< sample correlation >**

```
        Pearson's product-moment correlation

data:  X and Y
t = 5.5893, df = 10, p-value = 0.0002311
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5923111 0.9631599
sample estimates:
      cor
0.8703546
```

Fisher's One sample z-Test for $\rho = \rho_0$:

This test evaluates specific values of ρ_0 using Fisher's z-transformation approach.

Assumptions:

- Normal distribution for all variables used to compute correlation coefficient r.

Hypotheses:

$\rho_0 := 0.750$ **value must be explicitly stated.**

H_0: $\rho = \rho_0$ **< Correlation Coefficient ρ_0 not necessarily zero**

H_1: $\rho \neq \rho o$ **< Two sided test**

Fisher's Z-transformation:

$$z := \frac{1}{2} \cdot \ln\left(\frac{1+r}{1-r}\right) \qquad z = 1.3345$$

$$z_0 := \frac{1}{2} \cdot \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \qquad z_0 = 0.973$$

Test Statistic:

$$Z := (z - z_0) \cdot \sqrt{n-3} \qquad Z = 1.0848$$

Sampling Distribution of Test Statistic Z:

Z is distributed according to the *Standardized* Normal distribution of Z ~ N(0,1).

Critical Value of the Test:

$\alpha := 0.05$ **< Probability of Type I error must be explicitly set**

$$C_1 := \text{qnorm}\left(\frac{\alpha}{2}, 0, 1\right) \qquad C_1 = -1.96 \qquad C_2 := \text{qnorm}\left(1 - \frac{\alpha}{2}, 0, 1\right) \qquad C_2 = 1.96$$

$C := |C_1|$ $C = 1.96$ **^ Note use of N(0,1) here!**

Decision Rule:

IF |Z| > C, THEN REJECT H_0 OTHERWISE ACCEPT H_0

Probability Value:

$P := 2 \cdot \text{pnorm}(Z, 0, 1)$ $P = 1.722$ **< if** $t \leq 0$

$P := 2 \cdot (1 - \text{pnorm}(Z, 0, 1))$ $P = 0.27803$ **< if t > 0** **< Two sided case**

Confidence Interval for ρ:

$z_1 := z - C \cdot \frac{1}{\sqrt{n-3}}$ $z_2 := z + C \cdot \frac{1}{\sqrt{n-3}}$ $CI_z := (z_1 \quad z_2)$ $CI_z = (0.6812 \quad 1.9879)$

^ CI in transformed units of z

$CI_\rho := \left(\frac{e^{2 \cdot z_1} - 1}{e^{2 \cdot z_1} + 1} \quad \frac{e^{2 \cdot z_2} - 1}{e^{2 \cdot z_2} + 1} \right)$ $CI_\rho = (0.5923111 \quad 0.9631599)$ **< CI reverse transformed into units of ρ**

Prototype in R:

```
#FISHER'S Z TRANSFORMATION TEST:
rho0=0.750  #set test value
n=length(Y)
n
r=sqrt(summary(LM)$r.squared)
r
z=(1/2)*log((1+r)/(1-r))
z
z0=(1/2)*log((1+rho0)/(1-rho0))
z0
#TEST STATISTIC Z:
Z=(z-z0)*sqrt(n-3)
Z
#CRITICAL VALUE OF TEST:
alpha=0.05  #set type 1 error rate
C=abs(qnorm(alpha/2,0,1))
C
#PROBABILITY:
PA=2*pnorm(Z,0,1)
PA
PB=2*(1-pnorm(Z,0,1))
PB
P=min(PA,PB)
P
#CONFIDENCE INTERVAL FOR RHO
z1=z-C*(1/sqrt(n-3))
z1
z2=z+C*(1/sqrt(n-3))
z2
CILower=(exp(2*z1)-1)/(exp(2*z1)+1)
CILower
CIUpper=(exp(2*z2)-1)/(exp(2*z2)+1)
CIUpper
```

```
> r
[1] 0.8703546
> z=(1/2)*log((1+r)/(1-r))
> z
[1] 1.33454
> z0=(1/2)*log((1+rho0)/(1-rho0))
> z0
[1] 0.972955
> #TEST STATISTIC Z:
> Z=(z-z0)*sqrt(n-3)
> Z
[1] 1.084755

> #CRITICAL VALUE OF TEST:
> alpha=0.05  #set type 1 error rate
> C=abs(qnorm(alpha/2,0,1))
> C
[1] 1.959964
> #PROBABILITY:
> PA=2*pnorm(Z,0,1)
> PA
[1] 1.721970
> PB=2*(1-pnorm(Z,0,1))
> PB
[1] 0.2780303
> P=min(PA,PB)
> P
[1] 0.2780303

> #CONFIDENCE INTERVAL FOR RHO
>
> z1=z-C*(1/sqrt(n-3))
> z1
[1] 0.6812187
> z2=z+C*(1/sqrt(n-3))
> z2
[1] 1.987861
> CILower=(exp(2*z1)-1)/(exp(2*z1)+1)
> CILower
[1] 0.5923111
> CIUpper=(exp(2*z2)-1)/(exp(2*z2)+1)
> CIUpper
[1] 0.9631599
```