**Coefficient of determination**

The purpose of the coefficient of determination is to assess how much the errors in the prediction of Y can be reduced by using x.

Suppose we did not have knowledge of the independent (explanatory) variable x, but still wanted to predict the dependent (response) variable Y.

A measure of "how good"  does for the prediction is

Σ(observed y - predicted y)2



= TSS

This is a measure of the “total” variability of the yi’s. TSS is the “total sum of squares” and corresponds to what we have seen earlier in this course.

Note that some books refer to this quantity as SS(Total) or Syy.

Suppose we again use the knowledge of x to predict y. We would then use  to predict Y.

A measure of "how good"  does is

Σ(observed y - predicted y)2



= SSE

How much do we improve by using this additional knowledge of x?

* TSS – SSE = 
* The percent reduction in total sums of squares is the coefficient of determination:



Notes:

1. Some books will denote this quantity as .
2. 0 ≤  ≤ 1
3. 100×R2% of the variation in Y can be “explained” by using x to predict Y; i.e., the error in predicting Y can be reduced by 100×R2% when the regression model is used instead of just .
4. The capital letter for “R2” is not meant to represent it as a random variable. Rather, this is by far the most often way it is represented.
5. R2 is a measure of "fit" for the sample regression model

0 0.25 0.5 0.75 1.0

Bad Fit Good Fit

Example: Sales and Advertising

One can compute TSS =  = 6 and SSE =  = 1.1. Then



* 81.67% of the variation in sales can be explained by using advertising to predict sales.
* 81.67% of variation in sales is due to advertising



Example: College and HS GPA(gpa\_regression.R, gpa.csv)

From earlier,

> summary(object = mod.fit)

Call:

lm(formula = College.GPA ~ HS.GPA, data = gpa)

Residuals:

Min 1Q Median 3Q Max

-0.55074 -0.25086 0.01633 0.24242 0.77976

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.0869 0.3666 2.965 0.008299 \*\*

HS.GPA 0.6125 0.1237 4.953 0.000103 \*\*\*

---

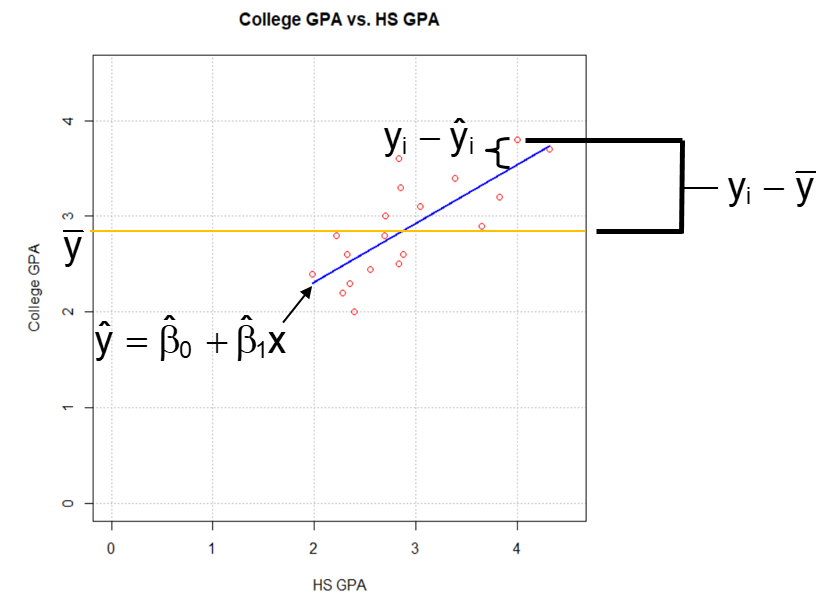
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3437 on 18 degrees of freedom

Multiple R-squared: 0.5768, Adjusted R-squared: 0.5533

F-statistic: 24.54 on 1 and 18 DF, p-value: 0.0001027

* 57.68% of the variation in College GPA can be explained by using HS GPA to predict college GPA.
* The model fits the data o.k.
* See the plot below showing what R2 measures:



Example: College GPA and Pizza (gpa\_regression.R, College\_GPA\_pizza.csv)

From earlier,

> summary(mod.fit2)

Call:

lm(formula = College.GPA ~ pizza, data = gpa2)

Residuals:

Min 1Q Median 3Q Max

-0.92519 -0.39645 -0.06745 0.41992 0.94080

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.9417 0.2082 14.132 3.48e-11 \*\*\*

pizza -0.0165 0.0358 -0.461 0.651

---

Signif. codes:

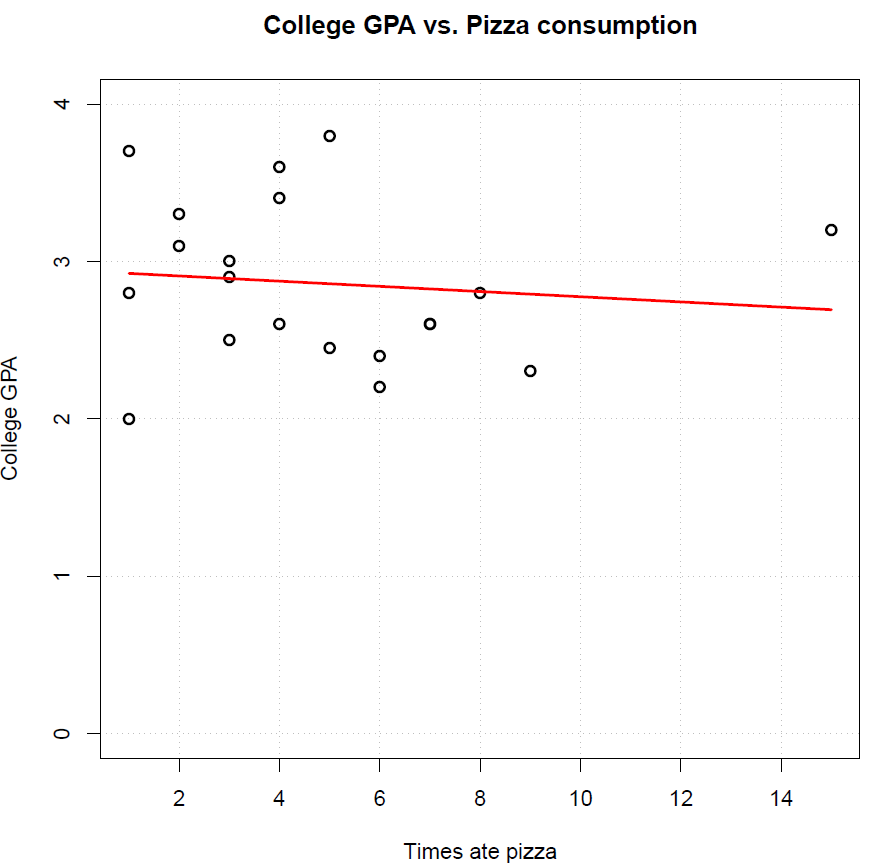
0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5252 on 18 degrees of freedom

Multiple R-squared: 0.01166, Adjusted R-squared: -0.04325

F-statistic: 0.2123 on 1 and 18 DF, p-value: 0.6505

* 1.2% of the variation in college GPA can be explained by using the number of times a student eats pizza to predict college GPA.
* The model fits the data poorly.



## Correlation coefficient

There is a closely related measure to the coefficient of determination called the coefficient of correlation. Some books may refer to it as the correlation coefficient, the Pearson correlation coefficient, or just simply the correlation. The sample estimate is

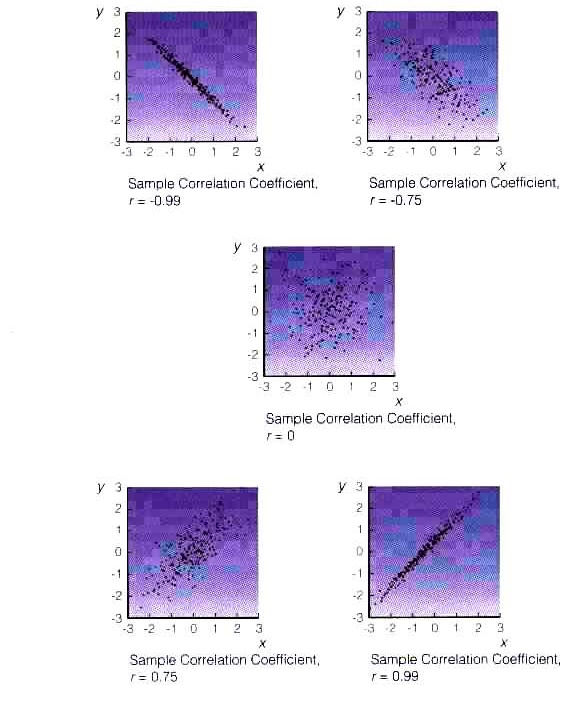


Comments:

* The square of r is the coefficient of determination R2! Note that this equivalence only occurs when there is only one independent variable in the model (and this variable is represented by one term).
* I use a lowercase letter for “r” simply because this is the way it is most often represented.
* What are the advantages to using r instead of R2?
* No regression model is needed
* With -1 ≤ r ≤ 1, a negative value says there is “negative” dependence and a positive value indicates there is “positive” dependence.



Below are some examples from a book to help demonstrate this measure:



Similar to how we had  as an estimate of the population parameter μ, we also have r as an estimate of a population parameter ρ (the greek letter “rho”). This parameter measures the dependence between x and y in the population:



One can show that



has a t distribution with ν = n – 2 degrees of freedom, where I use a capitol R to represent a random variable for the correlation. Thus,



Given this probability result, we now have a way to perform hypothesis tests:

Test statistic method:

1. Ho: ρ = 0 (no linear relationship)

Ha: ρ ≠ 0 (linear relationship)

1. Calculate the test statistic:



1. State the critical value: ±tα/2, n-2
2. Decide whether or not to reject Ho
3. State a conclusion in terms of the problem

Reject Ho – x is linearly related to Y.

Don’t Reject Ho – There is not sufficient evidence to show that x is linearly related to Y.

\_\_\_\_ means to put in what x and Y are in the problem

P-value method:

1. Ho: ρ = 0 (no linear relationship)

Ha: ρ ≠ 0 (linear relationship)

1. Calculate the p-value: p-value = 2×P(T>|t|) where T is a random variable with a t-distribution with ν = n – 2 and t is the compute value of the test statistic.
2. State α
3. Decide whether or not to reject Ho
4. State a conclusion in terms of the problem

NOTE: One can show that hypothesis tests for ρ = 0 and β1 = 0 are equivalent!

A confidence interval for ρ can be found as well. The typical method is through what is often called a “Fisher transformation”. Simply, Sir Ronald Fisher showed that



can be approximated well by a standard normal distribution for a large sample. Through using some algebra, one can show that a (1 – α)100% confidence interval for  is



If we let “L” and “U” be the lower and upper bounds of the interval above, the (1 – α)100% confidence interval for ρ is



Example: College and HS GPA (gpa\_regression.R, gpa.csv)

> cor(x = gpa$HS.GPA, y = gpa$College.GPA)

[1] 0.7594879

> cor(x = gpa)

HS.GPA College.GPA

HS.GPA 1.0000000 0.7594879

College.GPA 0.7594879 1.0000000

The output gives us r = 0.7595. There is a strong positive linear correlation in the sample, which makes sense given what we have seen with the scatter plot before. Note that the second example above gives what is often called a “correlation matrix”.

Is there a linear relationship between HS GPA and college GPA in the population? Use α = 0.05.

> cor.test(x = gpa$HS.GPA, y = gpa$College.GPA, alternative

= "two.sided", method = "pearson", conf.level = 0.95)

Pearson's product-moment correlation

data: gpa$HS.GPA and gpa$College.GPA

t = 4.9533, df = 18, p-value = 0.0001027

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.4774242 0.8996470

sample estimates:

cor

0.7594879

1. Ho: ρ=0

Ha: ρ≠0

2) 

NOTE: This is the same test statistic value as for the β1 = 0 hypothesis test

1. ±t0.025, 18 = ±2.101

> qt(p = 1 - 0.05/2, df = 18)

[1] 2.100922

4) Reject Ho because 4.95 > 2.101

5) There is a linear relationship between HS GPA and college GPA.

Notes:

* The method = “pearson” is the default in cor.test(). This name comes about through the correlation coefficient often being referred to as the “Pearson” correlation coefficient.
* 0.75952 = 0.5768 = R2
* The confidence interval given in the output is 0.4774 < ρ < 0.8996. This also can be found using the code below:

> r <- cor(x = gpa$HS.GPA, y = gpa$College.GPA)

> fisher <- 0.5\*log((1+r)/(1-r))

> n <- nrow(gpa)

> L <- fisher - qnorm(p = 1 - 0.05/2)/sqrt(n - 3)

> U <- fisher + qnorm(p = 1 - 0.05/2)/sqrt(n - 3)

> data.frame(lower = (exp(2\*L) - 1)/(exp(2\*L) + 1), upper

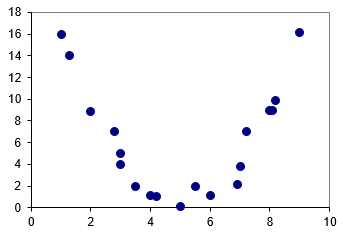
= (exp(2\*U) - 1)/(exp(2\*U) + 1))

lower upper

1 0.4774242 0.899647

Additional notes about correlation:

1. Suppose r = -0.02 from the data in the scatter plot below. Does this mean that x is not related strongly to Y in the sample?



NO!

Because r is close to zero, there is not a strong linear relationship in the sample; however, there appears to be a strong quadratic relationship in the sample.

Remember: r only measures the degree of a linear relationship

1. Strong correlation does not necessarily imply that x causes Y or vice versa. Strong correlation only means there is a linear relationship, not a causal relationship. Below is an example from another book to help illustrate this:

