

Evidence-based Health Care I (9) Correlation and Regression

Minato Nakazawa, Ph.D. (Department of International Health)

June 23, 2012

Contents

1	Relationship between the two quantitative variables	2
1.1	The difference between correlation and regression	2
1.2	Correlation analysis	2
1.3	Fitting a regression model	4
1.4	Testing the stability of estimated coefficients	7
2	Applied regression models	7
2.1	Multiple regression model	7
2.2	Evaluation of the goodness of fit	8
2.3	Points to be paid attention in fitting regression model	9
3	Furthur Readings	11

Corresponding to: Minato Nakazawa, Ph.D., Professor, Dept. International Health, Kobe Univ.
e-mail: minato-nakazawa@umin.net

Rev. 1. on 23rd June 2012: Modified for EZR.

1 Relationship between the two quantitative variables

Two well-known methods to examine the relationship between the two quantitative variables are calculating correlation and fitting the regression models.

First of all, drawing scattergram is necessary. Let's consider the relationship between height and the span of spread writing hand in survey data set.

In Rgui console, type `plot(Wr.Hnd ~ Height, data=survey)`. If you would like to see the relationship separately for males and females, use `pch=as.integer(Sex)` option.

In Rcmdr, select [Graphs], [Scatterplot...], then select Height as [x-variable] and Wr.Hnd as [y-variable], check off the box beside "Smooth Line", and click [OK]. Plotting by different markers for males and females, click the [Plot by groups...] button and select Sex before the clicking final [OK].

1.1 The difference between correlation and regression

The correlation means the strength of the relationship between 2 variables, and the regression means how much the variance of a variable can be explained by the variance of the other variable, by fitting the linear model.

1.2 Correlation analysis

The relationship shown as scatterplot may be apparent or spurious correlation. The researcher must always pay attention to it.

To show the strength of correlation, Pearson's product moment correlation coefficients are usually used. As nonparametric (using rank) version, the Spearman's rank correlation coefficients are also used.

The definition of the Pearson's correlation coefficient r between the 2 variables X and Y is,

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

The null-hypothesis that the r is not different from 0 can be tested using t_0 value defined as follows and t -distribution with $n - 2$ degree of freedom.

$$t_0 = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

In Rgui console, to calculate the Pearson's correlation coefficient between heights and spread writing hand spans with null-hypothesis testing, type as follows. For Spearman's rank correlation coefficients, `method="spearman"` option can be used.

```
cor.test(survey$Height, survey$Wr.Hnd)
```

In Rcmdr, select [Statistics], [Summaries], [Correlation test...], then select Height and Wr.Hnd (clicking with pressing **Ctrl**), and click [OK] (If you need Spearman's rank correlation, check the corresponding option). The following result will appear in the Output Window.

Pearson's product-moment correlation

```
data: survey$Height and survey$Wr.Hnd
t = 10.7923, df = 206, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5063486 0.6813271
sample estimates:
      cor
0.600991
```

The estimated r is 0.60 with 95% confidence interval of [0.50, 0.69]. The traditional criteria to judge the strength of correlation are, more than 0.7 'strong', 0.4-0.7 'moderate', 0.2-0.4 'weak'.

To calculate the correlations for males and females separately, we must make subset of the data. In Rgui console, it's easy to calculate those as follows.

```
males <- subset(survey, Sex=="Male")
cor.test(males$Height, males$Wr.Hnd)
females <- subset(survey, Sex=="Female")
cor.test(females$Height, females$Wr.Hnd)
```

In Rcmdr, select [Data], [Active data set], [Subset active data set], then type Sex=="Male" in the box below "Subset expression" and type males in the box below "Name for new data set" and click [OK]. The active data set will automatically change from survey to males. Then, select [Statistics], [Summaries], [Correlation test...], and select Height and Wr.Hnd (clicking with pressing **Ctrl**), and click [OK]. You will find the estimate of correlation coefficient between Height and Wr.Hnd with the result of null-hypothesis testing of correlation coefficient being zero for males.

To do similar calculation for females, at first you must change active dataset by clicking the box where males is shown as active dataset, and select survey and click [OK]. Again, select [Data], [Active data set], [Subset active data set], then type Sex=="Female" in the box below "Subset expression" and type females in the box below "Name for new data set" and click [OK]. The active data set will automatically change from survey to females. Then, select [Statistics], [Summaries], [Correlation test...], and select Height and Wr.Hnd (clicking with pressing **Ctrl**), and click [OK]. You will find the estimate of correlation coefficient between Height and Wr.Hnd with the result of null-hypothesis testing of correlation coefficient being zero for females.

1.3 Fitting a regression model

The principle of fitting regression models to observed data is that the variance of a dependent variable can be mostly explained by the variance of independent variables. If the explanatory power is enough, substituting the independent variables by actual values will serve a corresponding projection or estimation of the dependent variable. Reverse calculation is also possible, as in the case of so-called “working curve”. A working curve (but often line, sometimes with transformation) provides the equation as regression model for the series of observed absorptions for fixed concentrations. Usually a working line can be used when its explanatory power is more than 98%.

If the zero-adjustment is done by standard solution with zero concentration, the regression line must go through the origin (therefore, intercept must be zero), otherwise (zero-adjustment is done by pure water) the regression line may not go through the origin.

For example, let the series of absorption for the standard solutions with fixed concentrations (0, 1, 2, 5, 10 $\mu\text{g}/\ell$) as (0.24, 0.33, 0.54, 0.83, 1.32), when the zero-adjustment was done by pure water. If we denote the absorption variable as y and the concentration variable as x , the regression model can be written as $y = bx + a$. The coefficients a and b (a is called as “intercept” and b is called as “regression coefficient”) should be estimated by the least square method to find the set of a and b minimizing the sum of square errors,

$$f(a, b) = \sum_{i=1}^5 (y_i - bx_i - a)^2$$

Solving the equations that each partial differential of $f(a, b)$ by a and b equals 0, then we can obtain the following 2 equations.

$$b = \frac{\sum_{i=1}^5 x_i y_i / 5 - \sum_{i=1}^5 x_i / 5 \cdot \sum_{i=1}^5 y_i / 5}{\sum_{i=1}^5 x_i^2 / 5 - \left(\sum_{i=1}^5 x_i / 5 \right)^2}$$
$$a = \sum_{i=1}^5 y_i / 5 - b \cdot \sum_{i=1}^5 x_i / 5$$

Using these a and b values and measured absorption (for example 0.67), we can estimate the unknown concentration of sample solution. To note, the measured absorption of any sample must range within the values for standard solutions. The regression model has no guarantee to stand out range of standard solutions*¹.

In Rgui console, we can apply `lm()` (linear model) to estimate the fitted regression model as follows.

```
y <- c(0.24, 0.33, 0.54, 0.83, 1.32)
x <- c(0, 1, 2, 5, 10)
# apply linear model fitting
res <- lm(y ~ x)
```

*¹ Such an extrapolation is not recommended. Usually concentrating or diluting the solutions to remeasure the absorption is recommended.

```

# show the summary of result
summary(res)
# draw scattergram with regression line
plot(y ~ x)
abline(res)
# calculate the concentration corresponding to the absorption of 0.67
(0.67 - res$coef[1])/res$coef[2]

```

The summary of result is shown below.

```

Call:
lm(formula = y ~ x)

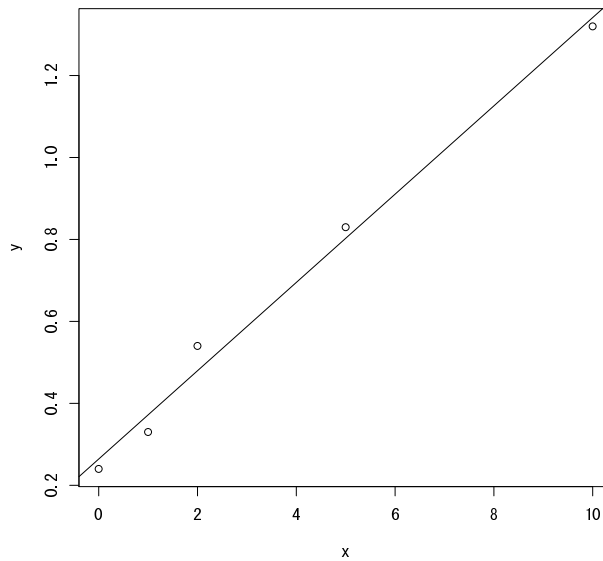
Residuals:
    1     2     3     4     5 
-0.02417 -0.04190  0.06037  0.02718 -0.02147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.26417    0.03090   8.549 0.003363 **
x            0.10773    0.00606  17.776 0.000388 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04894 on 3 degrees of freedom
Multiple R-squared:  0.9906,    Adjusted R-squared:  0.9875 
F-statistic: 316 on 1 and 3 DF,  p-value: 0.0003882

```

We can see that the estimated intercept was $a = 0.26417$ and regression coefficient was $b = 0.10773$, and the model explains 98.75% (0.9875) of the data, which is shown in Adjusted R-squared. The p-value means the result (significant probability) of testing the null-hypothesis (the extent how the variance of absorption can be explained by the model is similar to error variance).



The concentration for the absorption of 0.67 is given at last as 3.767084. Therefore, we can conclude that the concentration of the solution, of which absorption was 0.67, was $3.8 \mu\text{g}/\ell$.

In Rcmdr, the data must be entered as a Data Set. Select [Data], [New data set] and type workingcurve in the box of “Enter name for data set:”. After the “Data Editor” window appears, click [var1] and type y in the “variable name” box and check the radio button of “numeric” as type and press key. Next, change [var2] to [x] similarly. Then enter the data into each cells, and close the window (usually select [File], [Close]).

To draw scattergram with regression line, select [Graphs], [Scatterplot...], then select x as “x-variable” and y as “y-variable”. Check off the box beside “Smooth Line” and click [OK].

To apply a linear regression model fitting, select [Statistics], [Fit models], [Linear regression], then select y as “response variable” and x as “Explanatory variables”. Clicking [OK] leads to the summary of results shown in the Output Window.

For other situations than working curves, linear regression models can be applied in a similar manner. Let’s go back to the example of survey data set^{*2}. If we want to explain the variance of the span of writing hand by the height, we can apply the linear regression model to the survey data set by typing as follows in Rgui console.

```
res <- lm(Wr.Hnd ~ Height, data=survey)
summary(res)
```

In Rcmdr, after activating survey as already mentioned, select [Statistics], [Fit models], [Linear regression], then select Wr .Hnd as “response variable” and Height as “Explanatory variables”. Clicking [OK] gives the summary result.

^{*2} Of course, the MASS package must be loaded before using survey data set.

1.4 Testing the stability of estimated coefficients

When the response variable has virtually no relationship with the explanatory variable, the sums of squared residuals for many possible regression lines (any line on the centroid) may give almost same values. In other words, the estimated intercept and slope are very unstable in such situation. To evaluate the stability of parameters of regression line (regression coefficient b and intercept a), t_0 values are usually used. Let the relationship between y and x be expressed by the equation of $y = a_0 + b_0x + e$, and assume the error term e obeying the normal distribution with mean 0 and variance σ^2 , the estimated regression coefficient a would obey the normal distribution of mean a_0 , variance $(\sigma^2/n)(1 + M^2/V)$, where M and V are the mean and the variance of x . Then the sum of squared residuals Q divided by the variance of error σ^2 (say, Q/σ^2) obeys the chi-square distribution with degree of freedom $(n - 2)$. Therefore, the $t_0(a_0)$ defined as follows obeys the t -distribution with the degree of freedom $(n - 2)$.

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

However, to calculate this value, a_0 must be known. Under the null hypothesis of $a_0 = 0$, $t_0(0)$ calculated from the observed data is almost matching with $t_0(a_0)$ and obeys the t -distribution with degree of freedom $(n - 2)$. The absolute value of $t_0(0)$ is less than the 97.5% point of t -distribution at the 95% probability. We can also get the significance probability using the distribution function (cumulative probability density function).

Similarly, we can calculate $t_0(b)$ for regression coefficient as follows.

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

Using the relationship that the $t_0(b)$ obeys the t -distribution with degree of freedom $(n - 2)$, we can calculate the significance probability.

If the significant probability is very small (usually less than 5%, this criteria is called as the significance level of the test), we can say that a_0 or b_0 is significantly different from zero, which means the stability of estimated a_0 or b_0 .

In both Rgui console and Rcmdr, the significance probabilities are shown at the column of Pr(>|t|) .

2 Applied regression models

2.1 Multiple regression model

The explanatory variables can include two or more variables. In such case, the model is called as “multiple regression model”. There are some points to pay attention, but basically the explanatory variables can be given as the right terms of linear model, connected by +. For example, for the same data previously described, if you would like to explain the variance of the span of writing hand (Wr.Hnd) by the variance of height (Height) and the variance of the span of non-writing hand (NW.Hnd), you may type as follows in Rgui console.

```
res <- lm(Wr.Hnd ~ Height + NW.Hnd, data=survey)
summary(res)
```

In Rcmdr, select [Statistics], [Fit models], [Linear regression], then select Wr.Hnd as “response variable”, and Height and NW.Hnd with pressing **[Ctrl] key as “Explanatory variables”. Clicking [OK] leads to the summary of results shown in the Output Window.**

In the multiple regression model, the estimated regression coefficients are the “partial regression coefficients”, which adjust the effects of other explanatory variables on the response variable to obtain each explanatory variable’s own effect on the response variable. But the values of partial regression coefficients depend on the absolute scale of each variable, so that those cannot show the relative strength of effect on the response variable for each explanatory variable. For such comparison, the standardized partial regression coefficients can be used. In Rgui console, type as follows, then you obtain the estimates as stb for the standardized partial regression coefficients.

```
sdd <- c(0, sd(res$model$Height), sd(res$model$NW.Hnd))
stb <- coef(res)*sdd/sd(res$model$Wr.Hnd)
stb
```

The Rcmdr does not provide this as a menu item, but you can do so by editing the commands in script window, selecting lines and click [Submit].

2.2 Evaluation of the goodness of fit

It is always necessary to evaluate the goodness of fit of the regression model to the data.

After the least square estimation of a and b , we can define $z_i = a + bx_i$ for each x . Then $e_i = y_i - z_i$ can be considered as “residuals”. The residuals is the remaining part of the variance of y_i , that could not be explained by the regression model. Thus, the greater the residuals are, the worse the goodness of fit is. We would like to treat the both plus and minus residuals in its absolute distance, so that we can define the sum of squared residuals, Q , as follows.

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - z_i)^2$$

$$= \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2/n - \frac{\left(n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i\right)^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} / n$$

The sum of squared residuals Q is the scale to show badness of fit of the regression model. Q divided by n is called as residual variance (we will denote it as $\text{var}(e)$).

Among the values of $\text{var}(e)$ and the variance of y ($\text{var}(y)$) and Pearson’s product limit correlation coefficient r , the following equation always stands.

$$\text{var}(e) = \text{var}(y)(1 - r^2)$$

Therefore,

$$r^2 = 1 - \frac{\text{var}(e)}{\text{var}(y)}$$

Thus the closer to 1 r^2 is, the goodness of fit is higher. In that meaning, r^2 is called as “deterministic coefficient” or “the attributable proportion of x ”.

Since r^2 becomes larger depending on the number of explanatory variables, usually the r^2 will be adjusted for the degree of freedom. That is, Adjusted R-Squared in the summary of results.

As another indicator for the goodness of fit, the AIC (Akaike information criterion) is sometimes used (particularly in multiple regression models), which can be calculated in R, using the resulted object of linear model fitting (for example, `res` of the example above). In Rgui console, type `AIC(res)`, then you can get the AIC value. Here I don't explain any more, but many online materials and books can be found^{*3}.

2.3 Points to be paid attention in fitting regression model

The target variables may be measurements including error. In such situation, it is not valid to assume one as response variable and the other as explanatory variable. Generally speaking, if we can assume the direction of effect like the stature determining weight and not *vice versa*, the regression is possible where the stature is explanatory variable and the weight is response variable. However, when the explanatory variable includes measurement error, the explanatory power of the regression model become worse. In addition, the regression models with opposite combination of response variable and explanatory variable do not match. Thus, it is very important that the determination of which variable is response variable should be based on the direction of causal relationship, with enough reference to previous studies and clinical/biological knowledge.

Another point to be noticed is extrapolation of regression model for prediction. Especially the extrapolation should be avoided when you apply the working curve for prediction, because the linearity of the working curve is only confirmed within the range of standard material concentrations. The increase of absorbance tends to be smaller in higher concentration ranges due to saturation of molecules, the linearity is lost there. If you measure the samples with high concentration, you must dilute them into the ranges of standard materials.

Exercise

A built-in dataset `airquality` includes the air quality data in New York from May to September 1973. The variables are `Ozone` for ozone gas concentration in ppb, `Solar.R` for solar radiation in lang, `Wind` for wind speed in mph, `Temp` for atmospheric temperature in degree F, `Month` in number (5-9) and `Day` in number (1-31).

Let's fit the regression model for this data with ozone gas concentration as response variable and solar radiation as explanatory variable.

In Rgui console, enter the following 4 lines.

```
plot(Ozone ~ Solar.R, data=airquality)
res <- lm(Ozone ~ Solar.R, data=airquality)
abline(res)
summary(res)
```

^{*3} http://en.wikipedia.org/wiki/Akaike_information_criterion is the explanation in the Wikipedia.

In Rcmdr, at first, the airquality must be activated by select [Data], [Data in packages], [Read data set from an attached packages ...], then double-click datasets in the left panel and double-click airquality in the right panel, then click [OK].

To draw scattergram, [Graphs], [Scatterplot ...], then select Solar.R as x-variable and Ozone as y-variable. Check the box beside “Smooth Line” off, and click [OK]. Then you will get the scattergram with a regression line. To obtain the numerical result of regression model fitting, select [Statistics], [Fit models ...], [Linear regression], then select Ozone as Response variable and Solar.R as Explanatory variables, and click [OK]. Then you will see the result in the Output window.

Both give the same results as follows.

```
Call:
lm(formula = Ozone ~ Solar.R, data = airquality)

Residuals:
    Min       1Q   Median       3Q      Max
-48.292 -21.361  -8.864  16.373 119.136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.59873     6.74790   2.756 0.006856 **
Solar.R      0.12717     0.03278   3.880 0.000179 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 31.33 on 109 degrees of freedom
(42 observations deleted due to missingness)
Multiple R-Squared: 0.1213, Adjusted R-squared: 0.1133
F-statistic: 15.05 on 1 and 109 DF, p-value: 0.0001793
```

The fitted regression model is $Ozone = 18.599 + 0.127 \cdot Solar.R$, and the p-value shown in the bottom line is 0.0001793, as the result of F -test. Therefore, the fitting of the model can be judged as significant. However, the Adjusted R-squared shown above line is 0.11, which means only about 10% of the variance of the Ozone concentration can be explained by this model. We should judge the model is not so good.

To improve fitting, it may be possible to add more variables (for example, Wind and/or Temp) as explanatory variables as the multiple regression model. In Rgui console, you can easily do so by typing the next 3 lines. Then you see about 60% of Adjusted R-Squared value.

```
mres <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality)
summary(mres)
AIC(mres)
```

In Rcmdr, select [Statistics], [Fit models ...], [Linear model], then type Ozone at the box to the left of ~ and type Solar.R + Wind + Temp at the box to the right of ~, and click [OK].

3 Further Readings

- Armitage P, Berry G, Matthews JNS (2002) *Statistical Methods in Medical Research*, 4th ed., Blackwell Publishing.
- Bull K, Spiegelhalter DJ (1997) Tutorial in biostatistics: Survival analysis in observational studies. *Statistics in Medicine*, 16: 1041-1074.
- Faraway JJ (2006) *Extending the linear models with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall.
- <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf> (The introductory textbook for R Commander, provided by Prof. John Fox (McMaster Univ.), the developer of Rcmdr package.)
- Maindonald J, Braun J (2003) *Data analysis and graphics using R*, Cambridge Univ. Press.
- Nagelkerke N (1991) A note on a general definition of the coefficient of determination. *Biometrika*, 78: 691-692.
- R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Venables WN, Ripley BD (1999) *Modern Applied Statistics with S-PLUS. Third Edition*. Springer.
- EZR on Rcmdr: <http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmedEN.html>