

Understanding R^2

Chapter 6, Lab 2

OpenIntro Biostatistics

Topics

- R^2 with simulated data
- R^2 with the PREVEND data

The correlation coefficient r measures the strength of the linear relationship between two variables. However, it is more common to measure the strength of a linear fit using r^2 , which is usually written as R^2 in the context of regression.

This lab first uses simulated data to explore the idea behind the quantity R^2 , then provides an example of using R^2 to assess the strength of the linear fit of a regression model.

The material in this lab corresponds to Section 6.3.2 of *OpenIntro Biostatistics*.

Introduction

The quantity R^2 describes the amount of variation in the response variable that is explained by the least squares line:

$$R^2 = \frac{\text{variance of predicted } y\text{-values}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}$$

R^2 can also be calculated using the following formula:

$$R^2 = \frac{\text{variance of observed } y\text{-values} - \text{variance of residuals}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(y_i) - \text{Var}(e_i)}{\text{Var}(y_i)}$$

R^2 with simulated data

A simulation can be conducted in which y -values are sampled according to a population regression model $y = \beta_0 + \beta_1 x + \epsilon$, where the parameters β_0 , β_1 , and the standard deviation of ϵ are known. Recall that ϵ is a normally distributed error term with mean 0 and standard deviation σ .

1. Run the following code chunk to simulate 100 (x, y) values, where the values for x are 100 numbers randomly sampled from a standard normal distribution and the values for y are determined by the population model $y_i = 100 + 25x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 5)$.

```
#set the seed
set.seed(2017)

#simulate values
x = rnorm(100)
error = rnorm(100, 0, 5)
y = 100 + 25*x + error
```

- a) Create a scatterplot of y versus x and add the line of best fit to the plot.
 - i. Does the line appear to be a good fit to the data?
 - ii. Why do the data points not fall exactly on a line, even though the data are simulated according to a known linear relationship between x and y ?
 - iii. How well does the regression line estimate the population parameters β_0 and β_1 ?
- b) From a visual inspection, does it seem that the R^2 for this linear fit is relatively high or relatively low?
- c) Run the code chunk shown in the template to create two histograms, one of the predicted y -values and one of the observed (i.e., simulated) y -values. Visually compare the variances of the two sets of values; do the predicted and observed y -values seem to have similar spread?
- d) Run the code chunk shown in the template to calculate R^2 from the following formula. What is the R^2 for this model?

$$R^2 = \frac{\text{variance of predicted } y\text{-values}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}$$

- e) Calculate the R^2 for the model using the following formula. Confirm that the value is the same as from using the formula in part d).

$$R^2 = \frac{\text{variance of observed } y\text{-values} - \text{variance of residuals}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(y_i) - \text{Var}(e_i)}{\text{Var}(y_i)}$$

- f) To have R print the R^2 of a linear model, use the `summary(lm())` function as shown in the template. Confirm that this value matches the ones from the previous calculations.

2. Simulate 100 new (x, y) values. Like before, the x values are 100 numbers randomly sampled from a standard normal distribution and the y values are determined by the population model $y_i = 100 + 25x_i + \epsilon_i$. For these data, however, the error term is distributed $N(0, 50)$.
 - a) Create a scatterplot of y versus x and add the line of best fit to the plot. Does the line appear to be a good fit to the data? How well does the regression line estimate the population parameters β_0 and β_1 ?
 - b) Run the code chunk shown in the template to create two histograms, one of the predicted y -values and one of the observed (i.e., simulated) y -values. Visually compare the variances of the two sets of values; do the predicted and observed y -values seem to have similar spread?
 - c) Based on the answers to parts a) and b), does it seem that the R^2 for this linear model is relatively high or relatively low?
 - d) Use any method to calculate R^2 for the linear model.
3. Run the code chunk shown in the template to simulate 100 new (x, y) values.
 - a) Fit a linear model predicting y from x to the data and calculate the R^2 for the model. Based on R^2 , does the model seem to fit the data well?
 - b) Plot the data and add the line of best fit. Evaluate whether the linear model is a good fit to the data; how does viewing the data change the conclusion from part a)?

R^2 with the PREVEND data

4. Run the code chunk shown in the template to load `prevend.samp`, the random sample of 500 individuals from the PREVEND data used in the previous labs in this chapter.
 - a) Plot RFFT scores versus age and confirm that these data seem reasonably linear.
 - b) What proportion of the variability in the observed RFFT scores is explained by the linear model predicting average RFFT score from age?
 - c) Evaluate the strength of the linear relationship between RFFT score and age. Does it seem like there might be other factors that explain the variability in RFFT score?