Inference for Regression: Test for Linear Association Between Two Quantitative Variables

Thomas Scofield

November 21, 2021

The Model Utility Test

There are things you can do whenever you have bivariate quantitative data, such as

- produce a scatterplot of the data.
- calculate the (sample) correlation r.
- find the slope b_1 and intercept b_0 (both *sample statistics*) of the least squares regression line.

As we discussed in Chapter 2, the correlation is not always *meaningful*. But, in this chapter, we will assume that it is—that we have variables X and Y where the average response value $\mu_Y(x)$ at any particular value of X is given linearly as

$$\mu_Y(X) = \beta_0 + \beta_1 X,$$

making it meaningful to discuss the true correlation ρ.

An association between X and Y exists if the true slope $\beta_1 \neq 0$ or, equivalently, if the true correlation $\rho \neq 0$. Otherwise the variables are independent, meaning X has no value in predicting Y. To conduct a test, called the **model utility test**, of

$$\mathbf{H}_0 \colon \ \beta_1 = 0 \qquad \text{vs.} \qquad \mathbf{H}_a \colon \ \beta_1 \neq 0,$$

or equivalent stated as

$$\mathbf{H}_0 \colon \ \rho = 0 \qquad \text{vs.} \qquad \mathbf{H}_a \colon \ \rho \neq 0,$$

we will need sample data, producing a sample slope b_1 or sample correlation r.

Scatterplots; calculating b_1, r in R

The data set found at http://scofield.site/teaching/data/csv/heartDiseaseDeathsAndWine.csv contains a variable `winealc` that measures wine consumption (measured in liters per person per year) in various countries and another variable `hddeaths` which measures heart disease mortality rates (deaths per thousand). We import this data, view a scatterplot, and calculate the sample values.

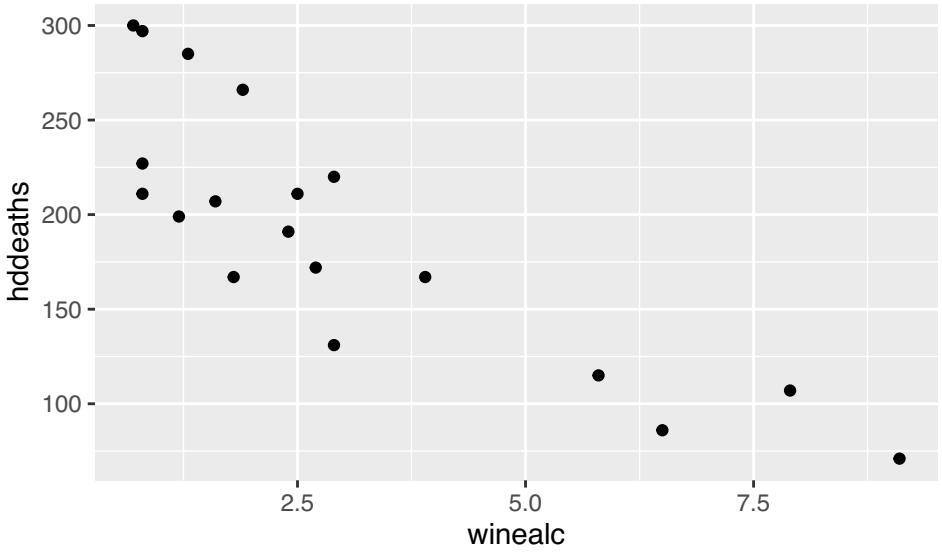
```
hdAndWine <- read.csv("http://scofield.site/teaching/data/csv/heartDiseaseDeathsAndWine.csv")
head(hdAndWine)
```

```
##       country winealc hddeaths
## 1   Australia     2.5      211
## 2 Netherlands     1.8      167
## 3     Austria     3.9      167
## 4 New Zealand     1.9      266
## 5     Belgium     2.9      131
## 6      Norway     0.8      227
```

From sampled data we can glean statistics like
r (sample correlation)
b_0 (sample y-intercept)
b_1 (sample slope)

In making a scatterplot, we must decide which variable (between `winealc` and `hddeaths`) to consider explanatory, placing it on the horizontal axis. Either could serve in that role, but in most discussions involving these variables, it is the alcohol consumption that people generally adopt as explanatory. So, it appears on the right side of the tilde in the command

```
gf_point(hddeaths ~ winealc, data=hdAndWine)
```



We can get the coefficients (intercept b_0 and slope b_1) of the best-fit line for the data via the command

```
lm(hddeaths ~ winealc, data=hdAndWine)
```

```
##
## Call:
## lm(formula = hddeaths ~ winealc, data = hdAndWine)
##
## Coefficients:
## (Intercept)      winealc
##      260.56       -22.97
```

Note that, by adding `$coefficients`, the output is less "wordy",

```
lm(hddeaths ~ winealc, data=hdAndWine)$coefficients
```

```
## (Intercept)     winealc
##   260.56338   -22.96877
```

and by altering this to '$coefficients[2]$ we obtain just the slope.

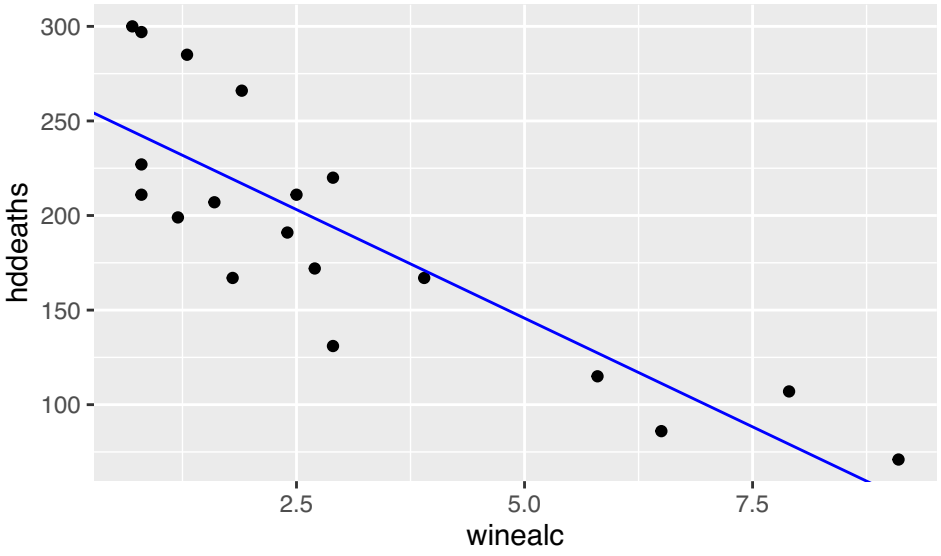
```
lm(hddeaths ~ winealc, data=hdAndWine)$coefficients[2]
```

```
##   winealc
## -22.96877
```

This last version, isolating the response to *slope* only, will be helpful in generating randomization distributions for b_1.

We can overlay the best-fit line by "piping" the scatterplot to the `gf_abline()` command with specified slope and intercept:

```
gf_point(hddeaths ~ winealc, data=hdAndWine) %>%
  gf_abline(slope = ~-22.97, intercept = ~260.56, color="blue")
```



It is simpler, and achieves the same thing as above, to pipe the scatterplot to `gf_lm()`.

```
gf_point(hddeaths ~ winealc, data=hdAndWine) %>% gf_lm(color="blue")
```

To calculate the sample correlation, instead, we change `lm()` to `cor()`:

```
cor(hddeaths ~ winealc, data=hdAndWine)
```

```
## [1] -0.8428127
```

Question

Would we get the same slope and intercept if we exchanged the roles of the variables, in this case making `hddeaths` the explanatory variable? Would we get the same correlation?

Randomization

Randomization distributions are meant to simulate the null distribution—what sort of values we expect, and how frequently, out of our sample statistic when the null hypothesis (no association between the quantitative variables) holds. We simulate it by shuffling one of the variables.

Randomization distribution for b_1: In the context of our *wine-and-heart-disease-deaths* data, one randomization statistic b_1 arises from

```
lm(hddeaths ~ shuffle(winealc), data=hdAndWine)$coefficients[2]
```

```
## shuffle(winealc)
##         4.161994
```

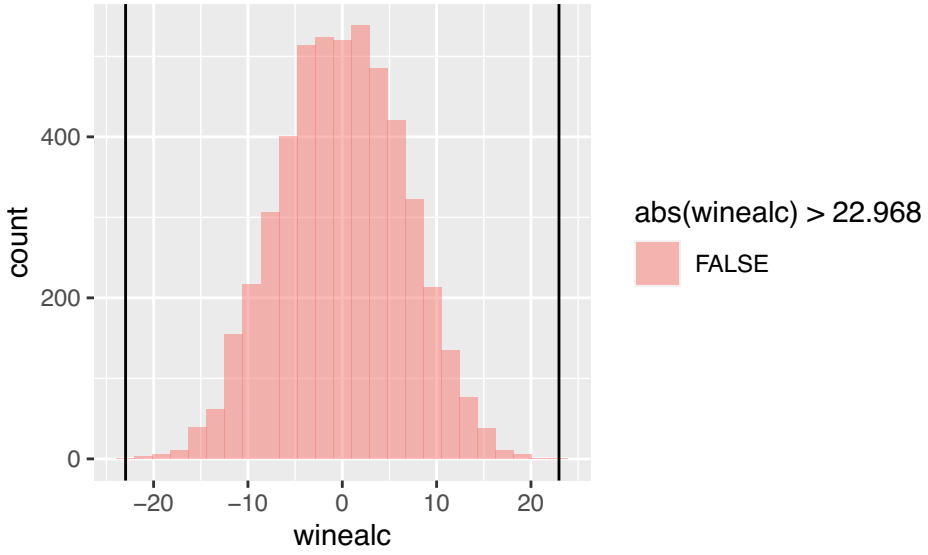
We get an approximate P-value when we generate lots of these randomization statistics, locate our test statistic (the slope for the original data), and determining how often something that extreme (or more so) occurs:

```
manyb1s <- do(5000) * lm(hddeaths ~ shuffle(winealc), data=hdAndWine)$coefficients[2]
head(manyb1s)
```

```
##      winealc
## 1  12.158335
## 2  -4.762228
## 3 -10.205138
## 4   7.933487
## 5 -10.321564
## 6   3.098284
```

This randomization distribution with the test statistic can be visualized in a manner like we have used before:

```
gf_histogram(~winealc, data=manyb1s, fill = ~abs(winealc)>22.968) %>%
  gf_vline(xintercept = ~22.968) %>%
  gf_vline(xintercept = ~-22.968)
```



This graph makes it appear that occurrences of $b_1 = 22.96877$, or even further distant from 0, are extremely rare. Indeed, if we count how often it happened in our 5000 tries, we have

```
2 * nrow( filter(manyb1s, winealc < -22.968) ) / 5000
```

```
## [1] 0
```

Randomization distribution for r: Consider another dataset, the **RestaurantTips** data from the `Lock5withR` package. The question here is whether there is an association between a bill for the meal at a restaurant, and the tip as a percentage-of-the-bill-as-tip (variable name `PctTip`). There are other ways to state this question of association. In the text, the Locks state it (roughly) as, "Is `Bill` an effective predictor of the size of the tip as a percentage of the bill?" The null hypothesis says, "no, it isn't", or $\rho = 0$.

To generate a randomization distribution for r under this null hypothesis, we start with the command that generates the r from the original data:

```
cor(PctTip ~ Bill, data=RestaurantTips)
```

```
## [1] 0.1352976
```

That is our test statistic.

The generation of a single randomization statistic r comes from shuffling one of the variables:

```
cor(PctTip ~ shuffle(Bill), data=RestaurantTips)
```

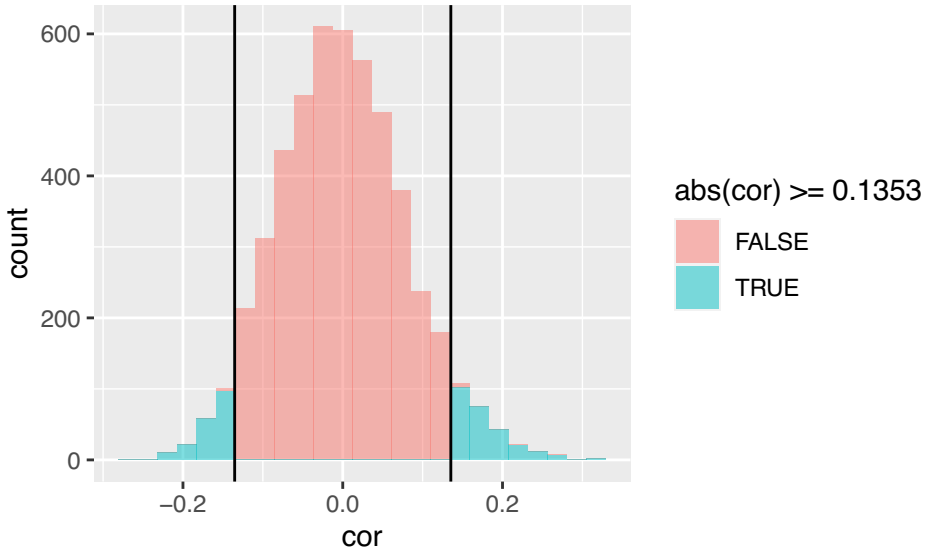
```
## [1] 0.001997988
```

If we do this often, we get a randomization distribution for r. We locate our test statistic on this distribution and use it as a boundary to determine the P-value.

```
manyCors <- do(5000) * cor(PctTip ~ shuffle(Bill), data=RestaurantTips)
head(manyCors)
```

```
##           cor
## 1 -0.04941720
## 2  0.01103168
## 3 -0.09646720
## 4 -0.07947382
## 5 -0.09839163
## 6  0.04777958
```

```
gf_histogram(~cor, data=manyCors, fill= ~abs(cor) >= 0.1353) %>%
  gf_vline(xintercept = ~0.1353) %>%
  gf_vline(xintercept = ~-0.1353)
```




```
nrow( filter(manyCors, abs(cor) >= 0.13529) ) / 5000
```

```
## [1] 0.0902
```

This P-value is not smaller than $\alpha = 0.05$, so at that level we fail to reject the null hypothesis, that the true correlation ρ (and the true slope β_1) is zero. That is, if someone tends to think that patrons of restaurants tip the same percentage regardless of the overall tab, we have not found here evidence sufficient to conclude otherwise.

Exercise

Throughout the discussion above, it has been implied that it didn't matter which test statistic you randomized, r or b_1. Convince yourself that this is the case by playing with randomization distributions and P-values for bivariate quantitative data in StatKey. The confirmation that "it does not matter" would be that, when working with a fixed dataset, when you generate a P-value corresponding to your test statistic r, it is roughly the same as the P-value corresponding to your test b_1.

This reflects where class ended on Tues., Nov. 23