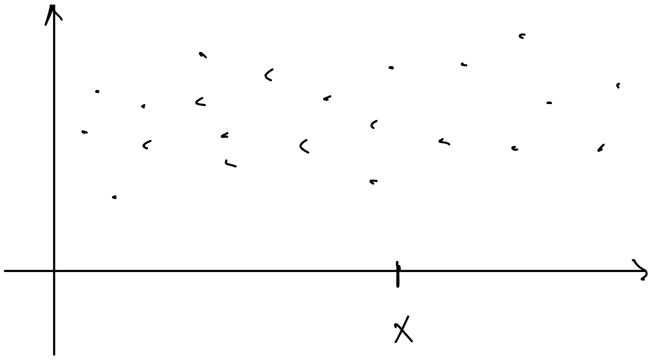
```
Stat 145, Wed 21-Apr-2021 -- Wed 21-Apr-2021
Biostatistics
Spring 2021


-------------------------------
Wednesday, April 21st 2021
-------------------------------
Due::   PS13 due at 11 pm

Other calendar items


-------------------------------
Wednesday, April 21st 2021
-------------------------------
Wk 12, We
Topic:: Regression reprised

least-squares regression:  \hat y = a + bx
 - identify slope as b, intercept as a
 - offers a "prediction" to value of y for given x
 - observed y vs. fitted/predicted \hat y-value
    residual = observed - predicted
      straight-line distance
      positive if data point is above line, negative if below
 - how data is used to choose a, b
    want to minimize sum of squared residuals
        sum r_i^2
    from calculus, obtain simple formulas
      b = r s_y / s_x   }
      a = ybar - b xbar }

Some data sets
 - spruces: http://scofield.site/teaching/data/csv/hesterberg/Spruce.csv
    lm(Di.change ~ Ht.change, data = spruces)
 - hdAndWine: .../teaching/data/csv/heartDiseaseDeathsAndWine.csv
 - Lock sets
    NFL_Malevolence
    Hurricanes  (double check it is the Lock5 one by this name)
```



True correlation (pop. param.): ρ

Sample " : r

population/sample slopes: β, b

$$b = r \cdot \frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

Big question: Does there exist an association between variables (both quantitative)

H_0: There isn't — in symbols $\rho = 0$ $(\beta = 0)$

The test of the hypotheses is called the "Model Utility Test".

H_a: Yes, there is — in symbols $\rho \neq 0$ $(\beta \neq 0)$

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

Thomas Scofield

April 20, 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

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

or equivalent stated as

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

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

Scatterplots; calculating b_1, r in R

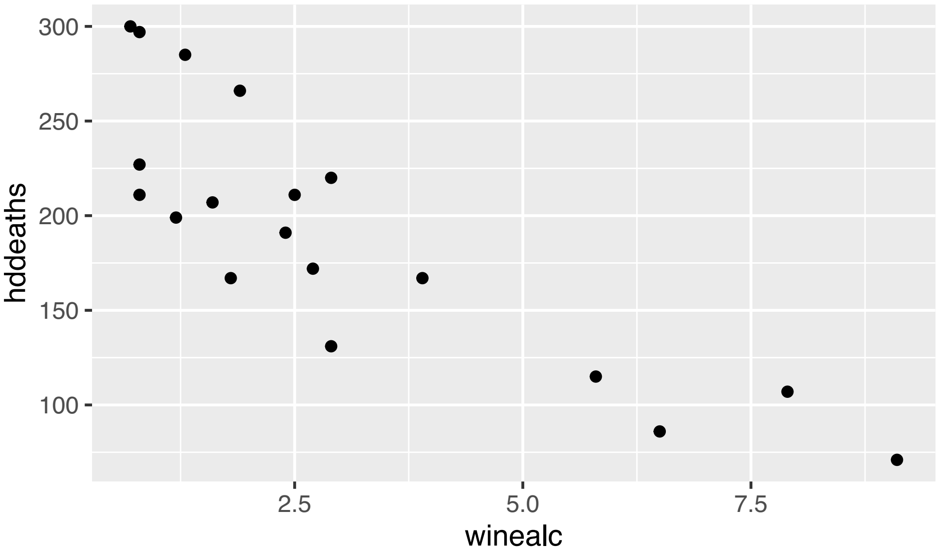
The data set found at http://scofield.site/teaching/data/csv/heartDiseaseDeathsAndWine.csv contains a variable `winealc` that measures wine consumption (measured in liters per person per year) in various countries and another variable `hddeaths` which measures heart disease mortality rates (deaths per 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
```

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]
```

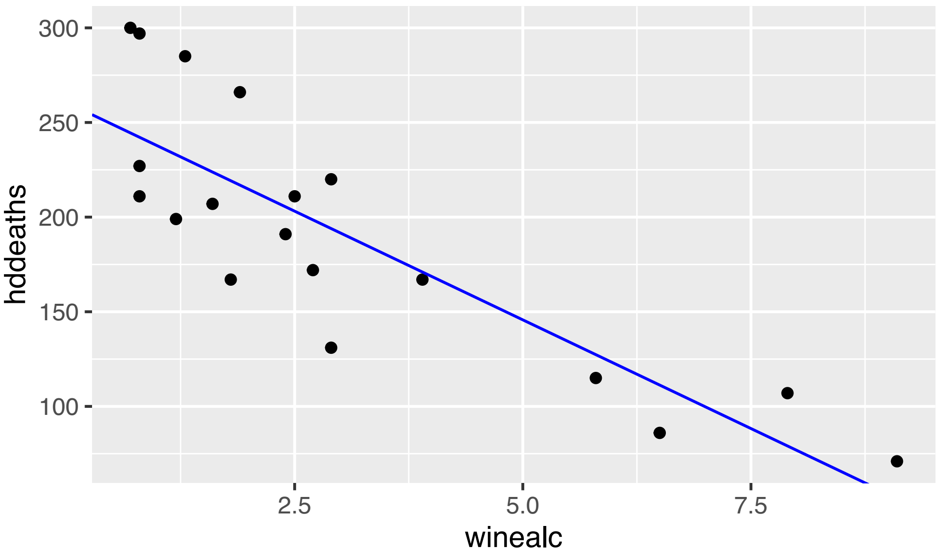
```
##   winealc
## -22.96877
```

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")
```

↑
In place of this command,
one could insert gf_lm( ).
No inputs are needed for this alternative line-producing command



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)
##        -9.047053
```

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)
```