Regression Lines differ by sample (Chapter 20)

P.B. Matheson adopted from Nicholas Horton

December 13, 2020

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fifth Edition of *Intro Stats* (2018) by De Veaux, Velleman, and Bock. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at http://nhorton.people.amherst.edu/is5.

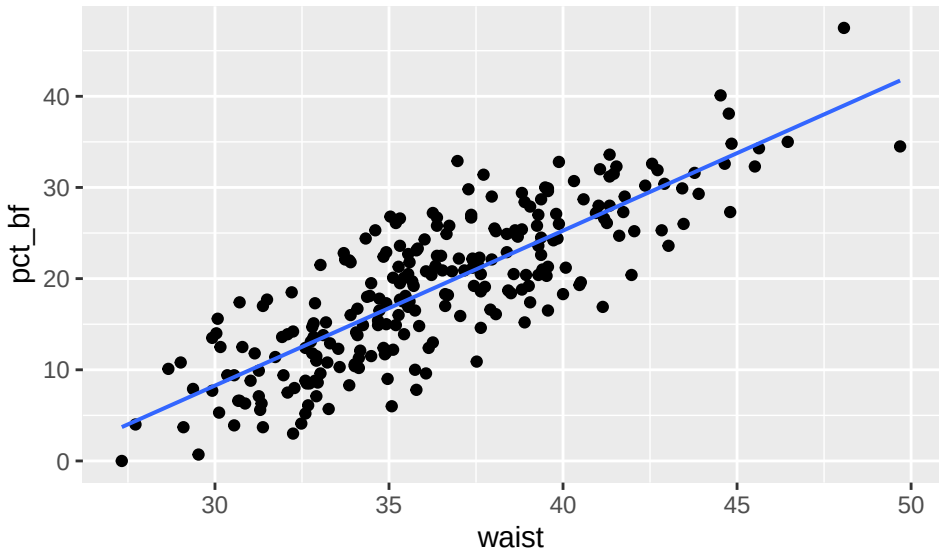
This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (https://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 20: Inferences for Regression

```
library(mosaic)
library(readr)
library(janitor)
BodyFat <- read_csv("http://nhorton.people.amherst.edu/is5/data/Bodyfat.csv") %>%
  janitor::clean_names()
```

By default, `read_csv()` prints the variable names. These messages have been suppressed using the `message=FALSE` code chunk option to save space and improve readability. Here we use the `clean_names()` function from the `janitor` package to sanitize the names of the columns (which would otherwise contain special characters or whitespace).

```
gf_point(pct_bf ~ waist, data=BodyFat) %>%
  gf_lm()
```



Random Matters: Slopes Vary

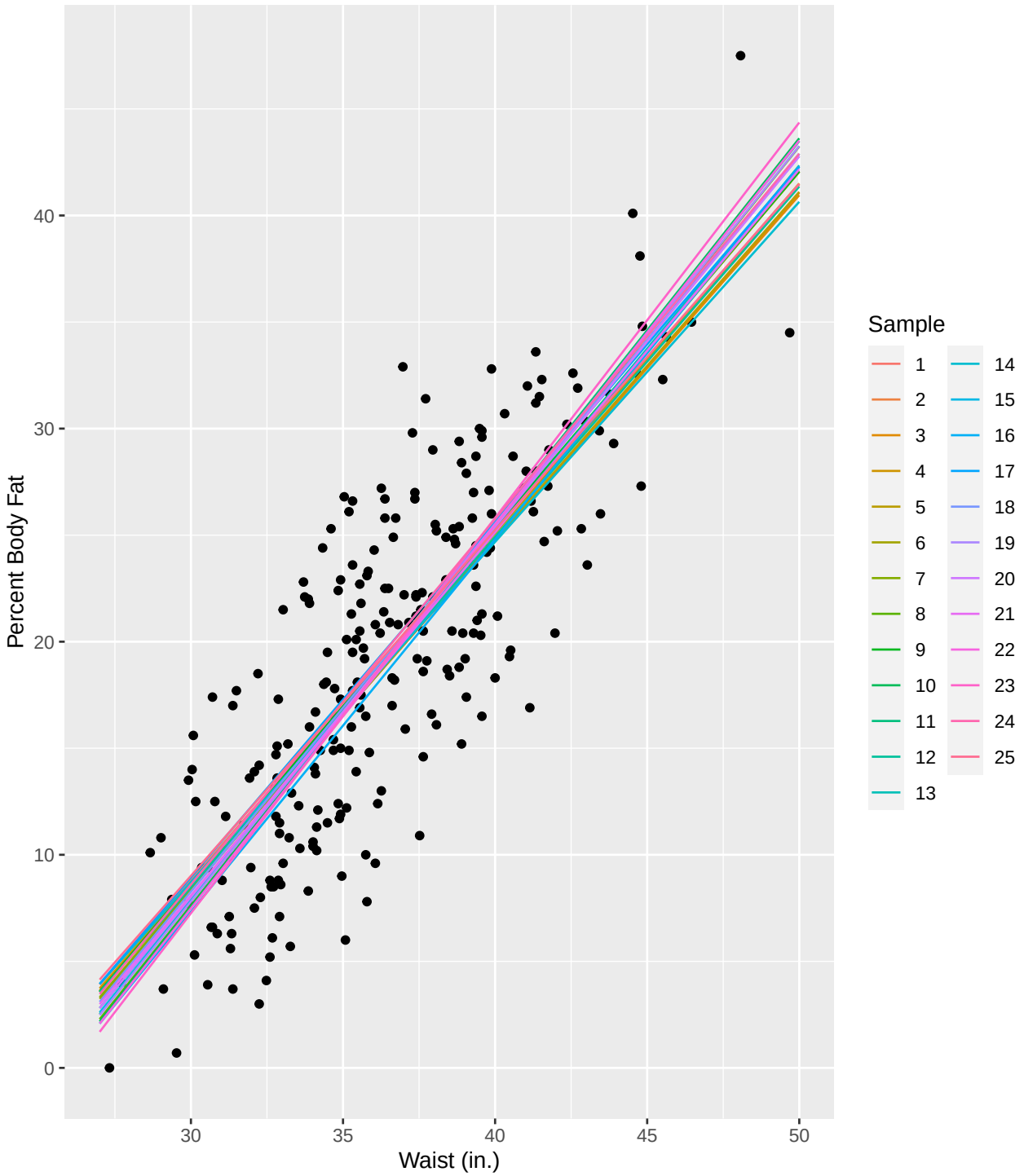
```
origsample <- lm(pct_bf ~ waist, data = BodyFat)

msummary(origsample)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.73413    2.71651  -15.73   <2e-16 ***
## waist         1.69997    0.07431   22.88   <2e-16 ***
##
## Residual standard error: 4.713 on 248 degrees of freedom
## Multiple R-squared:  0.6785, Adjusted R-squared:  0.6772
## F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

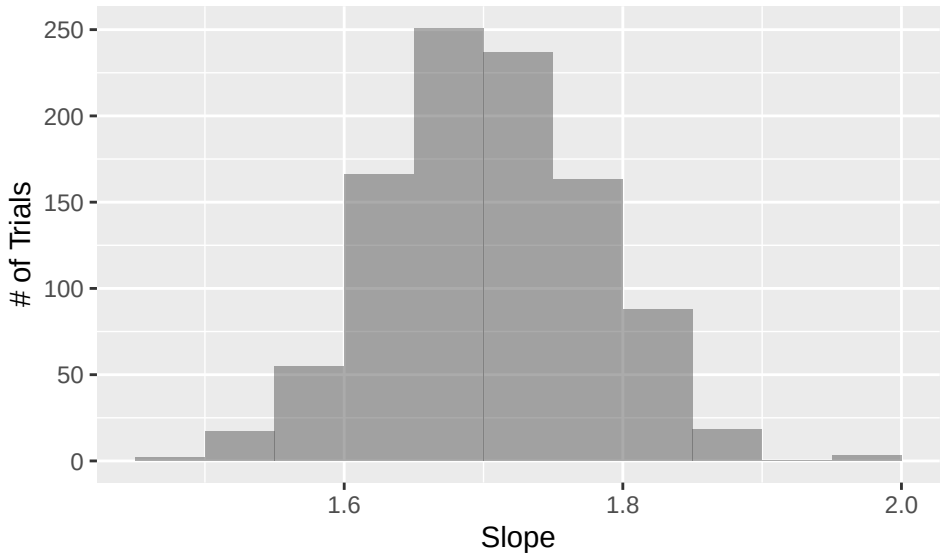
```
numsamp <- 25 # It's too messy to do any more than 25
slopesdata <- do(numsamp) * lm(pct_bf ~ waist, data = resample(BodyFat))
```

For more information about `resample()`, refer to the resample vignette in mosaic.

```
slopesdata <- slopesdata %>%
  mutate(at27 = Intercept + waist * 27, at50 = Intercept + waist * 50, color = as.factor(1:25))
# Figure 20.4, page 644
gf_point(pct_bf ~ waist, data = BodyFat) %>%
  gf_segment(at27 + at50 ~ 27 + 50, data = slopesdata, color = ~color) %>%
  gf_labs(color = "Sample", x = "Waist (in.)", y = "Percent Body Fat")
```




```
numsamp <- 1000 # To see the shape of the histogram
slopesdata <- do(numsamp) * lm(pct_bf ~ waist, data = resample(BodyFat))
# Figure 20.5
gf_histogram(~waist, data = slopesdata, binwidth = .05, center = .025) %>%
  gf_labs(x = "Slope", y = "# of Trials")
```



For the histogram, we use 1,000 trials.