IS5 in R: Inferences for Regression (Chapter 20)

Margaret Chien and Nicholas Horton (nhorton@amherst.edu)

July 11, 2018

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. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. 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 (http://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

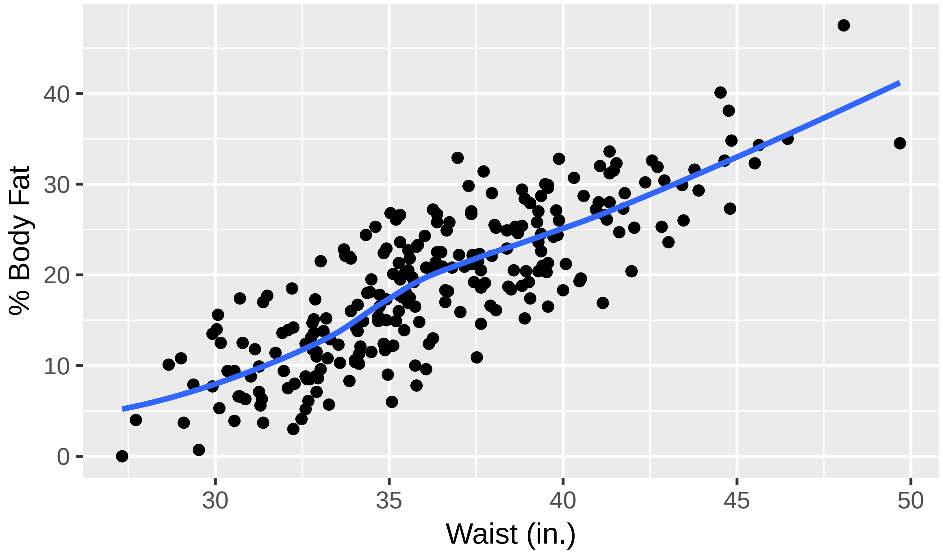
```r
library(mosaic)
library(readr)
library(janitor)
BodyFat <- read_csv("http://nhorton.people.amherst.edu/is5/data/Bodyfat.csv") %>%
  clean_names()
```

```
## Parsed with column specification:
## cols(
##   Density = col_double(),
##   Pct.BF = col_double(),
##   Age = col_integer(),
##   Weight = col_double(),
##   Height = col_double(),
##   Neck = col_double(),
##   Chest = col_double(),
##   Abdomen = col_double(),
##   Waist = col_double(),
##   Hip = col_double(),
##   Thigh = col_double(),
##   Knee = col_double(),
##   Ankle = col_double(),
##   Bicep = col_double(),
##   Forearm = col_double(),
##   Wrist = col_double()
## )
```

By default, `read_csv()` prints the variable names. These messages can be 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).

```
# Figure 20.1, page 642
gf_point(pct_bf ~ waist, data = BodyFat) %>%
  gf_smooth() %>% # to show linear relationship
  gf_labs(x = "Waist (in.)", y = "% Body Fat")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

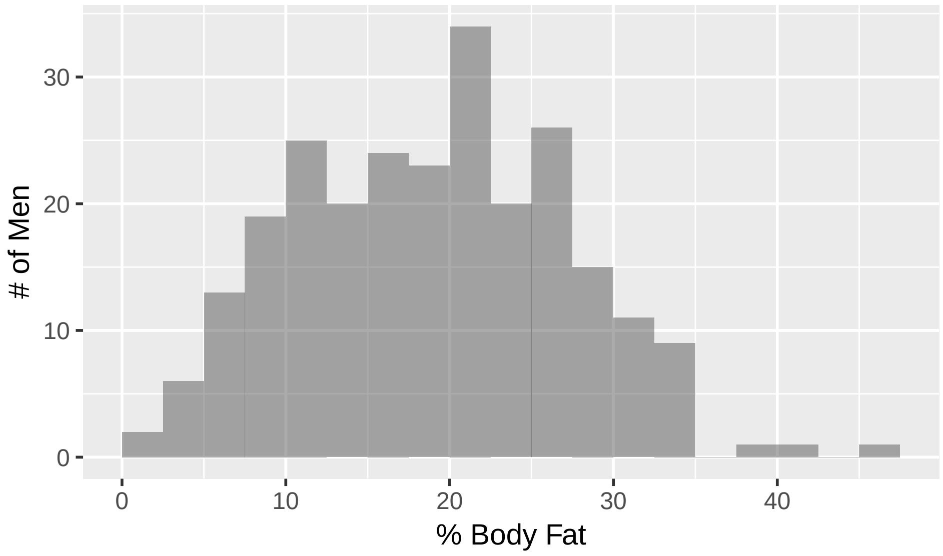


Section 20.1: The Regression Model

```
lm(pct_bf ~ waist, data = BodyFat)
```

```
##
## Call:
## lm(formula = pct_bf ~ waist, data = BodyFat)
##
## Coefficients:
## (Intercept)        waist
##      -42.73         1.70
```

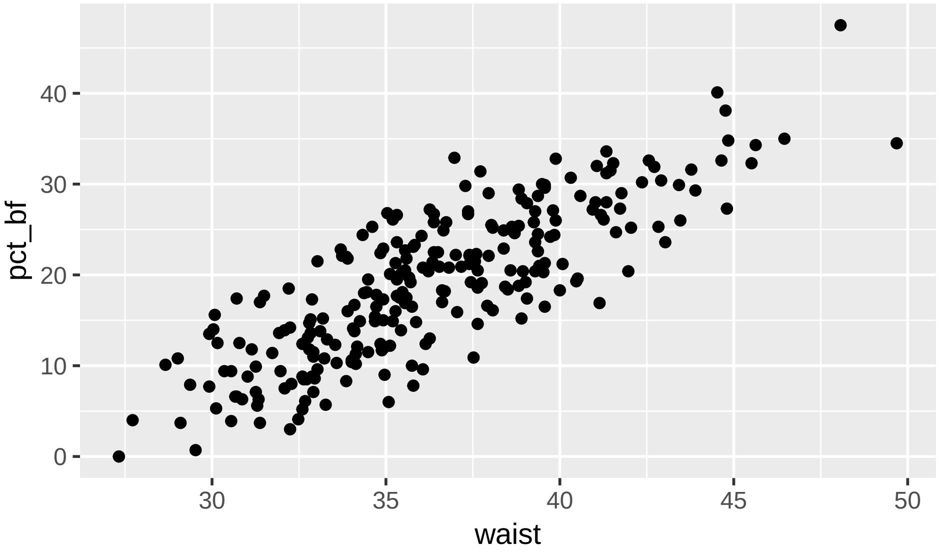
```
# Figure 20.2
gf_histogram(~ pct_bf, data = BodyFat, binwidth =  2.5, center = 1.25) %>%
  gf_labs(x = "% Body Fat", y = "# of Men")
```



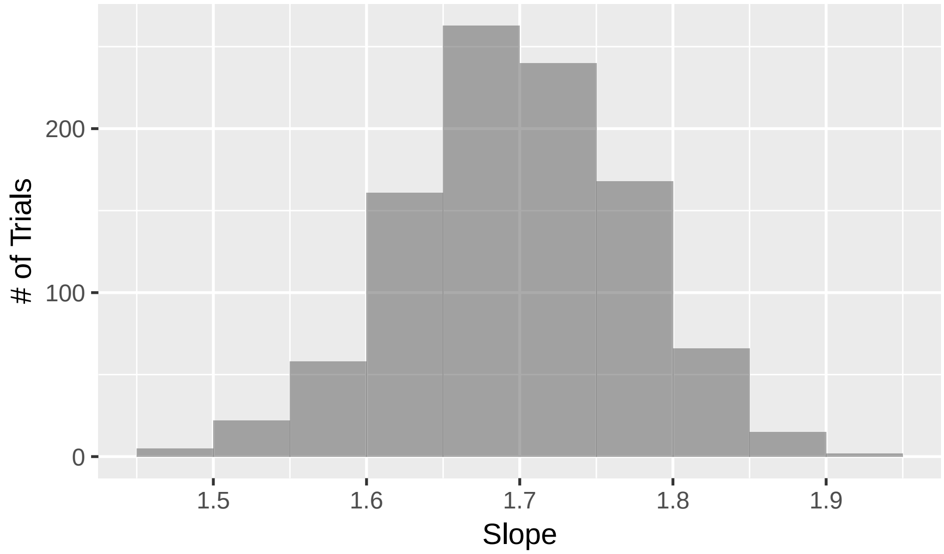
XX NH not sure if I can make Figure 20.3 (page 643)

Random Matters: Slopes Vary

```
numsamp <- 1000
slopesdata <- do(numsamp) * lm(pct_bf ~ waist, data = resample(BodyFat))
gf_point(pct_bf ~ waist, data = BodyFat) #%>%
```

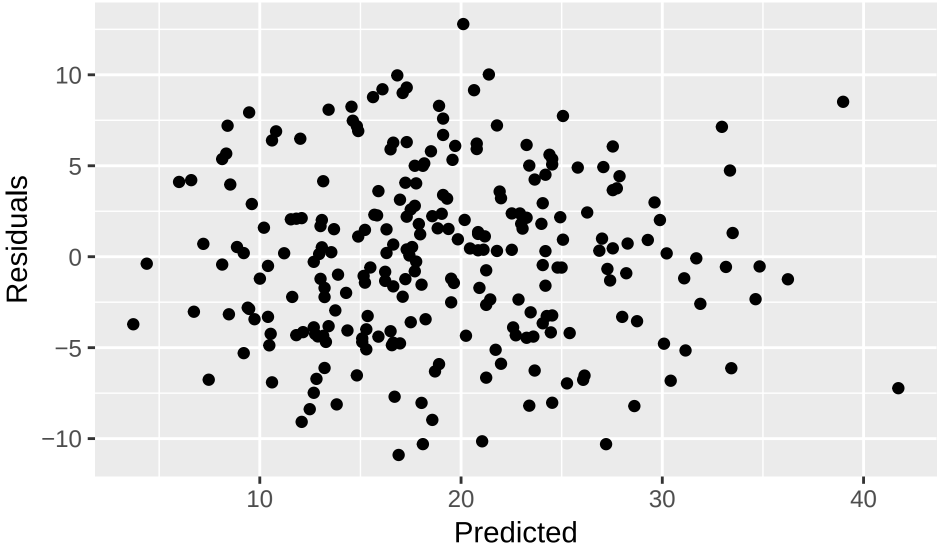


```
#  gf_abline(slope = ~ waist, intercept = ~ Intercept, slopesdata) # gf_coefline()?
gf_histogram(~ waist, data = slopesdata, binwidth = .05, center = .025) %>%
  gf_labs(x = "Slope", y = "# of Trials")
```



Section 20.2: Assumptions and Conditions

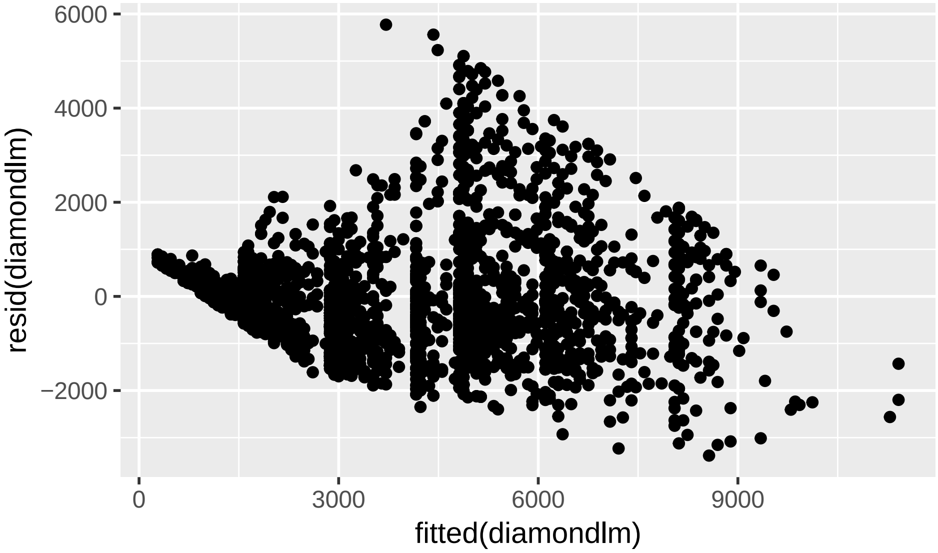
```
# Figure 20.6 is the same as Figure 20.1
# Figure 20.7 (page 645)
bodyfatlm <- lm(pct_bf ~ waist, data = BodyFat)
gf_point(resid(bodyfatlm) ~ fitted(bodyfatlm)) %>%
  gf_labs(x = "Predicted", y = "Residuals")
```



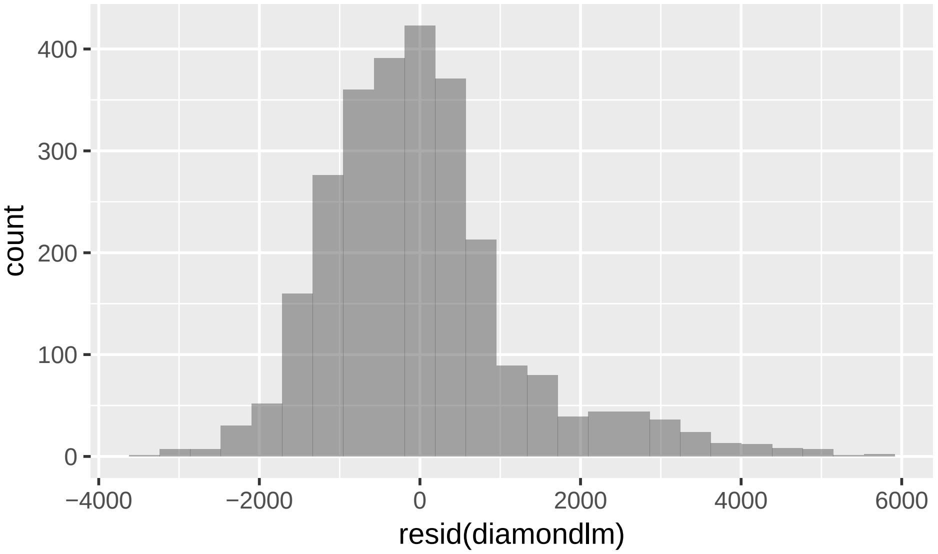
```
Diamonds <- read_csv("http://nhorton.people.amherst.edu/is5/data/Diamonds.csv") %>%
  clean_names()
```

```
## Parsed with column specification:
## cols(
##   Price = col_integer(),
##   `Carat Size` = col_double(),
##   Color = col_character(),
##   Clarity = col_character(),
##   Cut = col_character()
## )
```

```
diamondlm <- lm(price ~ carat_size, data = Diamonds) # there's no carat weight varible?
# Figure 20.8, page 646
gf_point(resid(diamondlm) ~ fitted(diamondlm))
```



```
# Figure 20.9
gf_histogram(~ resid(diamondlm))
```



Section 20.3: Regression Inference and Intuition

Section 20.4: The Regression Table

Section 20.5: Multiple Regression Inference

Section 20.6: Confidence and Prediction Intervals

Section 20.7: Logistic Regression

Section 20.8: More About Regression