IS4 in R: Scatterplots, Association, and Correlation (Chapter 6)

Patrick Frenett and Nicholas Horton (nhorton@amherst.edu)

July 17, 2017

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Intro Stats* (2013) 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://www.amherst.edu/~nhorton/sdm4.

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.

Note that some of the figures in this document may differ slightly from those in the IS4 book due to small differences in datasets. However in all cases the analysis and techniques in R are accurate.

Chapter 6: Scatterplots, Association, and Correlation

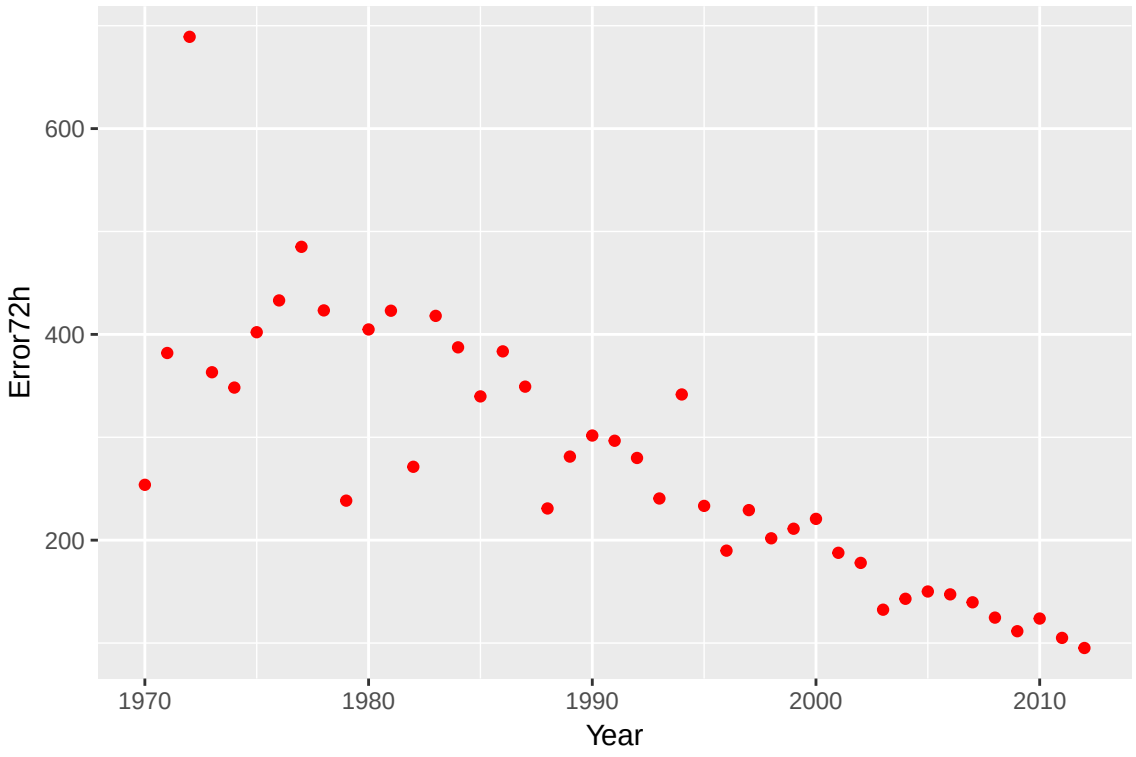
Section 6.1: Scatterplots

Figure 6.1 (page 150) displays the scatterplot of the average tracking error over time.

```
library(mosaic); library(readr); library(ggformula)
options(digits=3)
Hurricanes <-
  read_csv("http://www.amherst.edu/~nhorton/sdm4/data/Tracking_hurricanes_2012.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Error24h = col_double(),
##   Error48h = col_double(),
##   Error72h = col_double()
## )
```

```
gf_point(Error72h ~ Year, ylab="Prediction error (nautical miles)", data=Hurricanes, col="red")
```



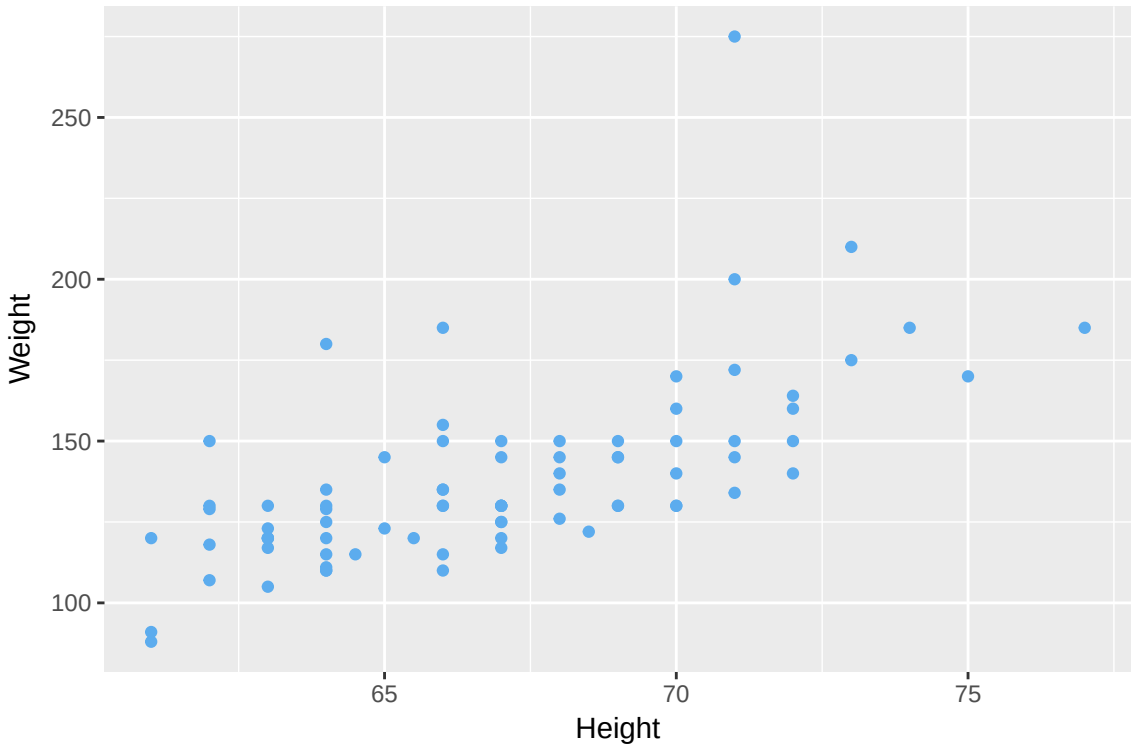
Section 6.2: Correlation

Figure 6.2 (page 153) displays the scatterplot of weight vs. height for a sample of students from statistics classes.

```
HtWt <- read_csv("http://www.amherst.edu/~nhorton/sdm4/data/Heights_and_Weights.csv")
```

```
## Parsed with column specification:
## cols(
##   Weight = col_integer(),
##   Height = col_double()
## )
```

```
gf_point(Weight ~ Height, ylab= "Weight (lbs)", xlab="Height (in)", data=HtWt, col="steelblue2")
```




```
cor(Weight ~ Height, data=HtWt)
```

```
## [1] 0.644
```

Kendall's Tau and Spearman's Rho

```
cor(Weight ~ Height, method="kendall", data=HtWt)
```

```
## [1] 0.545
```

```
cor(Weight ~ Height, method="spearman", data=HtWt)
```

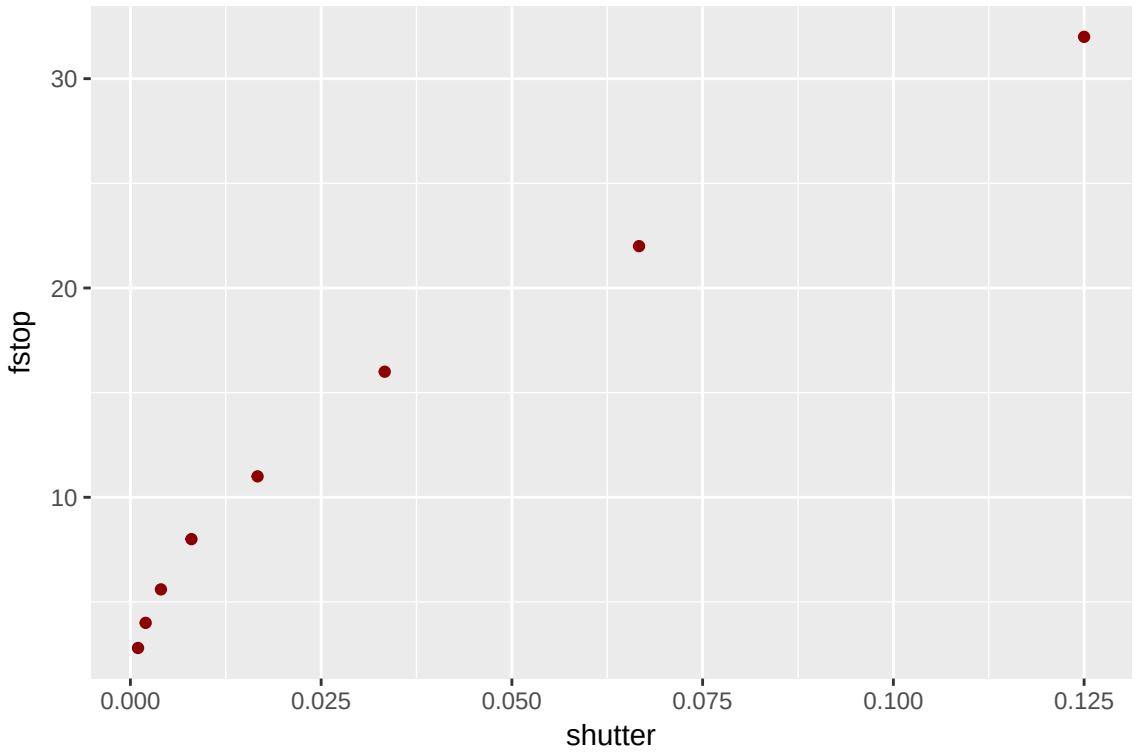
```
## [1] 0.697
```

Section 6.3: Warning: Correlation does not always equal Causation

Section 6.4: Straightening scatterplots

Since the dataset is so small for Figure 6.10 (page 161) we can enter it by hand.

```
fstop <- c(2.8, 4, 5.6, 8, 11, 16, 22, 32)
shutter <- c(1/1000, 1/500, 1/250, 1/125, 1/60, 1/30, 1/15, 1/8)
lenses <- data.frame(fstop, shutter)
gf_point(fstop ~ shutter, ylab = "f/stop", xlab = "Shutter Speed (sec)", data=lenses, col="darkred")
```



A new transformed variable can be added using the `mutate` function.

```
lenses <- mutate(lenses, fstopsq = fstop*fstop)
gf_point(fstopsq ~ shutter, ylab = "f/stop (squared)", xlab="Shutter Speed (sec)", data=lenses, col="fi
```