SDM4 in R: Regression Wisdom (Chapter 8)

Nicholas Horton (nhorton@amherst.edu)

June 4, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fourth Edition of *Stats: Data and Models* (2014) 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/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.

Chapter 8: Regression Wisdom

Section 8.1: Examining residuals

Figure 8.1 (page 220) displays the scatterplot of heart rate vs duration for the Penguins dataset (along with a superimposed regression line and a smoother).

```
library(mosaic)
library(readr)
options(digits = 3)
Penguins <- read_csv("http://nhorton.people.amherst.edu/sdm4/data/Penguins.csv")
gf_point(DiveHeartRate ~ Duration, ylab = "Dive Heart Rate (bpm)",
         xlab="Duration (mins)", data = Penguins) %>%
  gf_lm() %>%
  gf_smooth(se = FALSE)
```

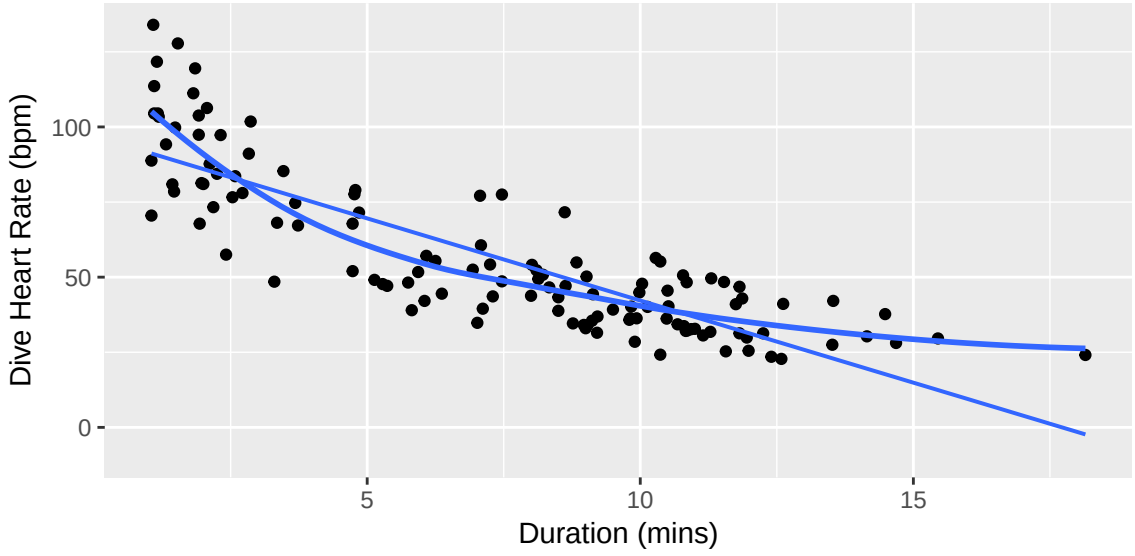


Figure 8.2 (page 220) displays the residuals from a linear regression model as a function of duration.

```
Penguinmod <- lm(DiveHeartRate ~ Duration, data = Penguins)
gf_point(resid(Penguinmod) ~ Duration, data = Penguins) %>%
  gf_lm() %>%
  gf_smooth(se = FALSE)
```

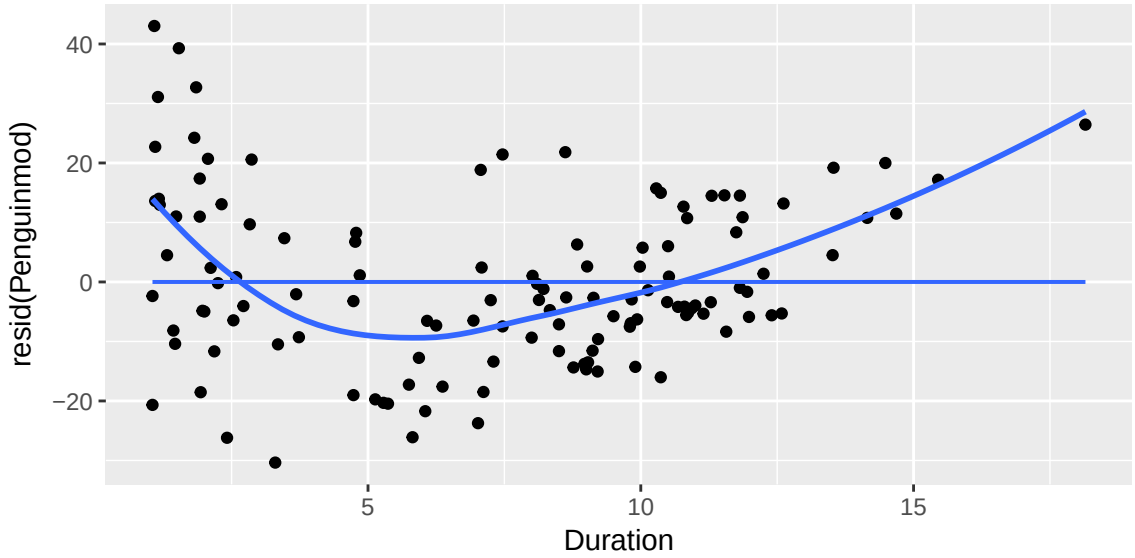


Figure 8.3 (page 221) displays the histogram of residuals for the cereal data from Chapter 7.

```
Cereals <- read_csv("http://nhorton.people.amherst.edu/sdm4/data/Cereals.csv")
Cerealmod <- lm(calories ~ sugars, data = Cereals)
gf_histogram(~ resid(Cerealmod), binwidth = 7.5)
```

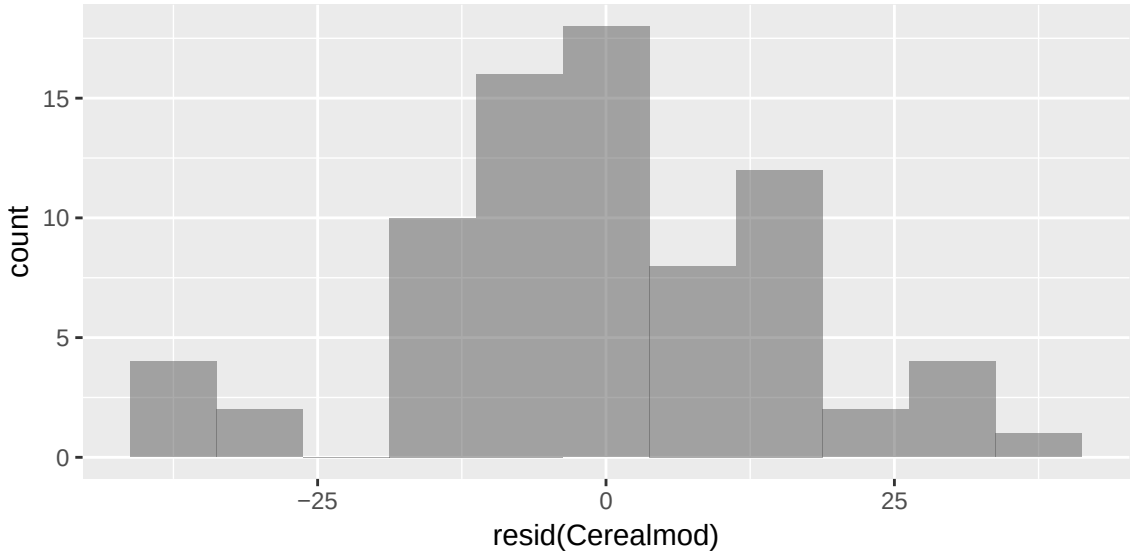
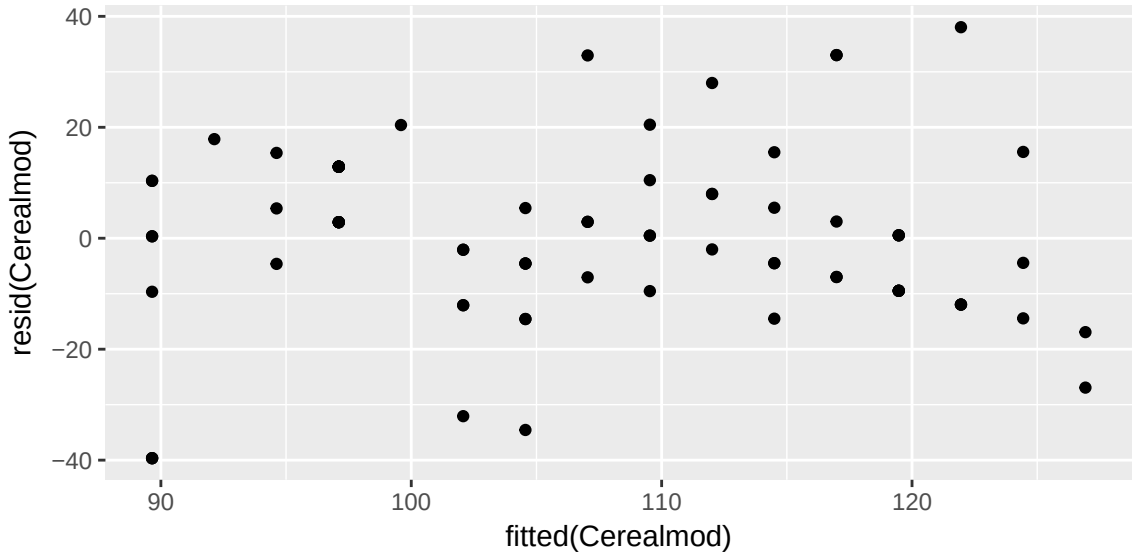


Figure 8.4 (page 221) displays a scatterplot of the residuals vs predicted values. Without jittering, the display has an odd pattern.

```
gf_point(resid(Cerealmod) ~ fitted(Cerealmod))
```



By adding some random noise we can more easily observe values that are shared by more than one cereal.

```
gf_point(jitter(resid(Cerealmod)) ~ jitter(fitted(Cerealmod)))
```

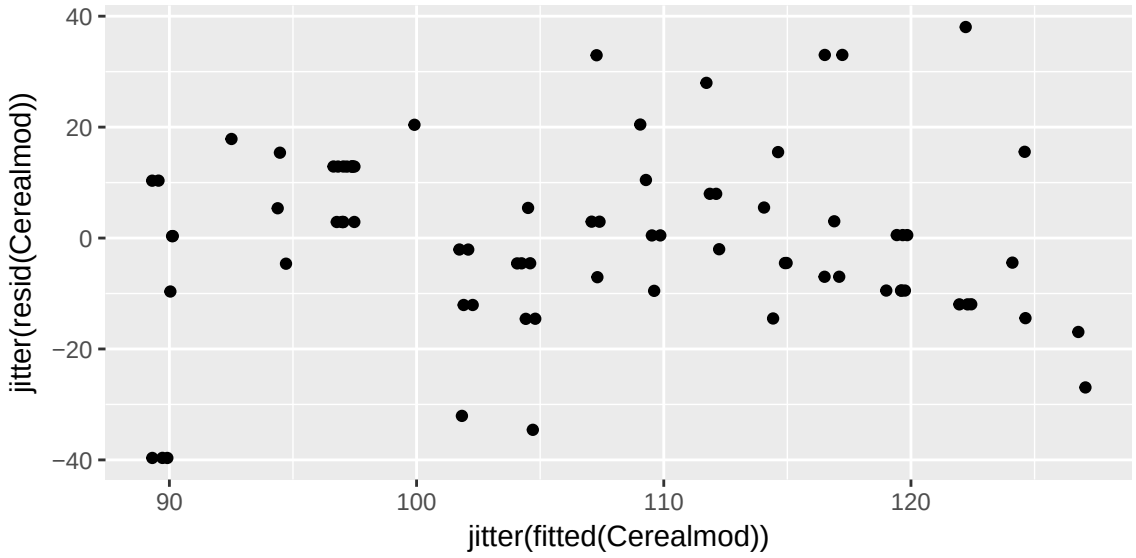


Figure 8.5 (page 222) displays the scatterplot stratified by what shelf it is displayed on at the store.

```
tally(~ shelf, data=Cereals)
```

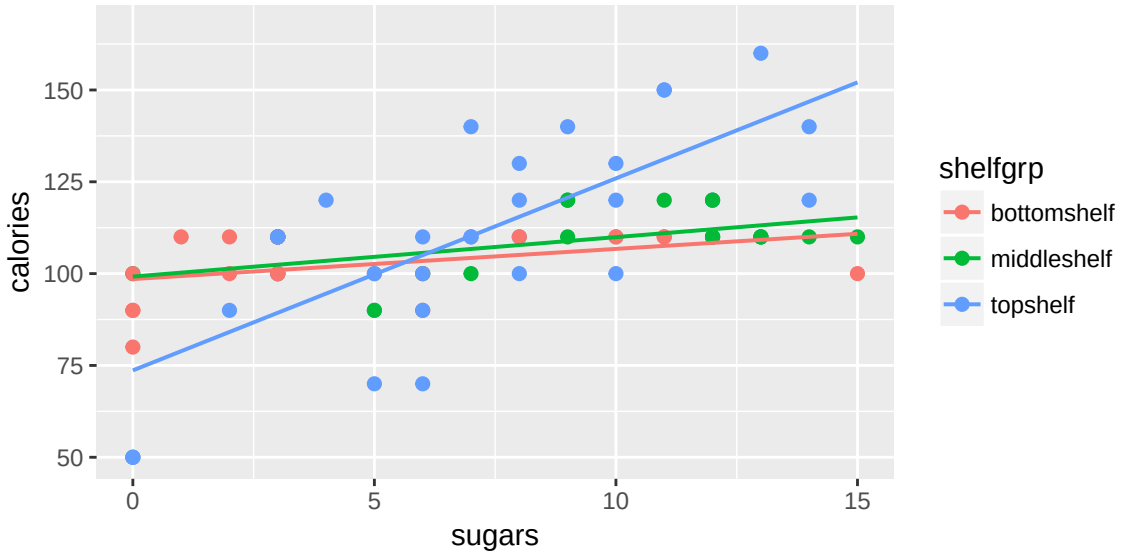
```
## shelf
##  1  2  3
## 20 21 36
```

```
Cereals <- mutate(Cereals, shelfgrp =
  derivedFactor(
    bottomshelf = shelf == 1,
    middleshelf = shelf == 2,
    topshelf = shelf == 3
  )
```

```
)
tally(~ shelfgrp, data = Cereals)
```

```
## shelfgrp
## bottomshelf middleshelf    topshelf
##          20          21          36
```

```
gf_point(calories ~ sugars, color = ~ shelfgrp,
  lwd=2, data = Cereals) %>%
  gf_lm()
```



Section 8.2: Extrapolation and reaching beyond the data

Section 8.3: Outliers, leverage, and influence

Section 8.4: Lurking variables and causation

Section 8.5: Working with summary values