Multiple Regression (Chapter 25)

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

June 19, 2018

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 https://nhorton.people.amherst.edu/is4.

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 25: Multiple Regression

Section 25.1: Two Predictors

The table on page 743 displays the results from the multiple regression model.

```
library(mosaic)
library(readr)
options(digits = 3)
BodyFat <- read_csv("https://nhorton.people.amherst.edu/sdm4/data/Body_fat_complete.csv")
BodyFatmod <- lm(PctBF ~ waist + Height, data = BodyFat)
msummary(BodyFatmod)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1009     7.6861   -0.40     0.69
## waist         1.7731     0.0716   24.77  < 2e-16 ***
## Height       -0.6015     0.1099   -5.47  1.1e-07 ***
##
## Residual standard error: 4.46 on 247 degrees of freedom
## Multiple R-squared:  0.713,  Adjusted R-squared:  0.711
## F-statistic:  307 on 2 and 247 DF,  p-value: <2e-16
```

We can use this model to generate predicted values.

```
BodyFatfun <- makeFun(BodyFatmod)
BodyFatfun(waist = 0, Height = 0)  # returns intercept
```

```
##    1
## -3.1
```

```
BodyFatfun(waist = 30, Height = 70)
```

```
##    1
## 7.98
```

```
-3.101 + 1.773*30 - 0.602*70
```

```
## [1] 7.95
```

Section 25.2: What Multiple Regression Coefficients Mean

Figure 25.1 on page 744 displays the scatterplot of percent body fat against height.

```
gf_point(PctBF ~ Height, data = BodyFat)
```

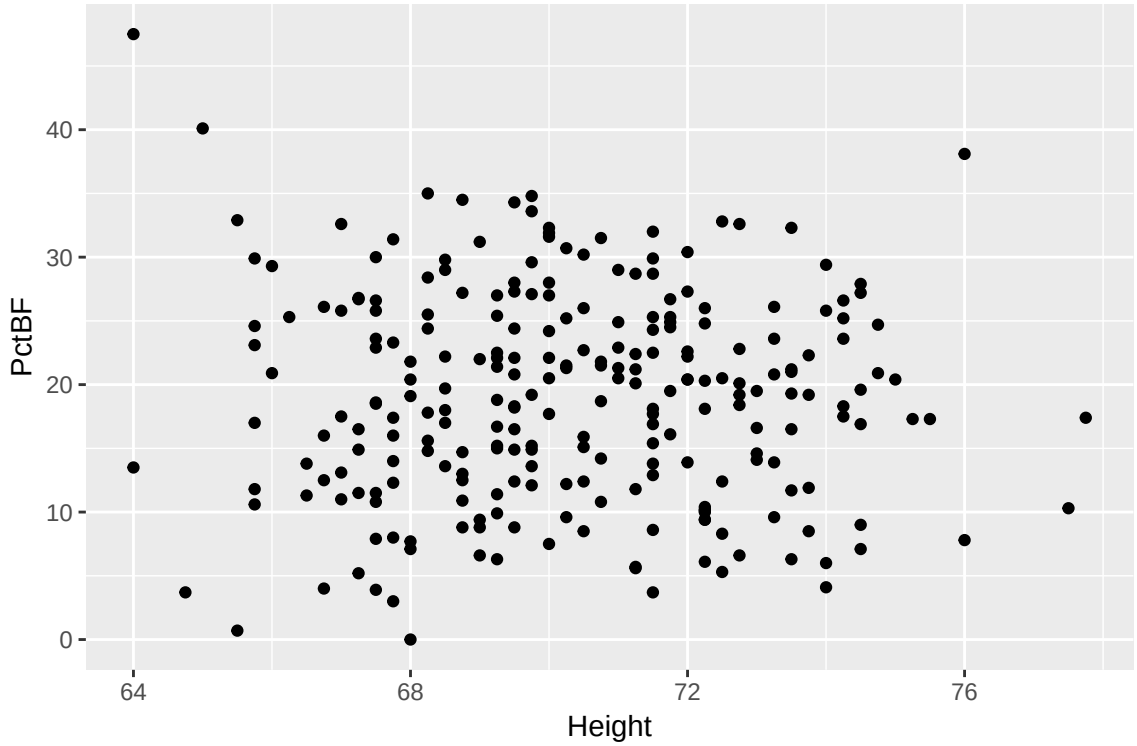


Figure 25.2 (page 745) displays the scatterplot for a subset of the data (men with waist sizes between 36 and 38 inches).

```
gf_point(PctBF ~ Height, data = filter(BodyFat, waist > 36, waist < 38)) %>%
  gf_lm()
```

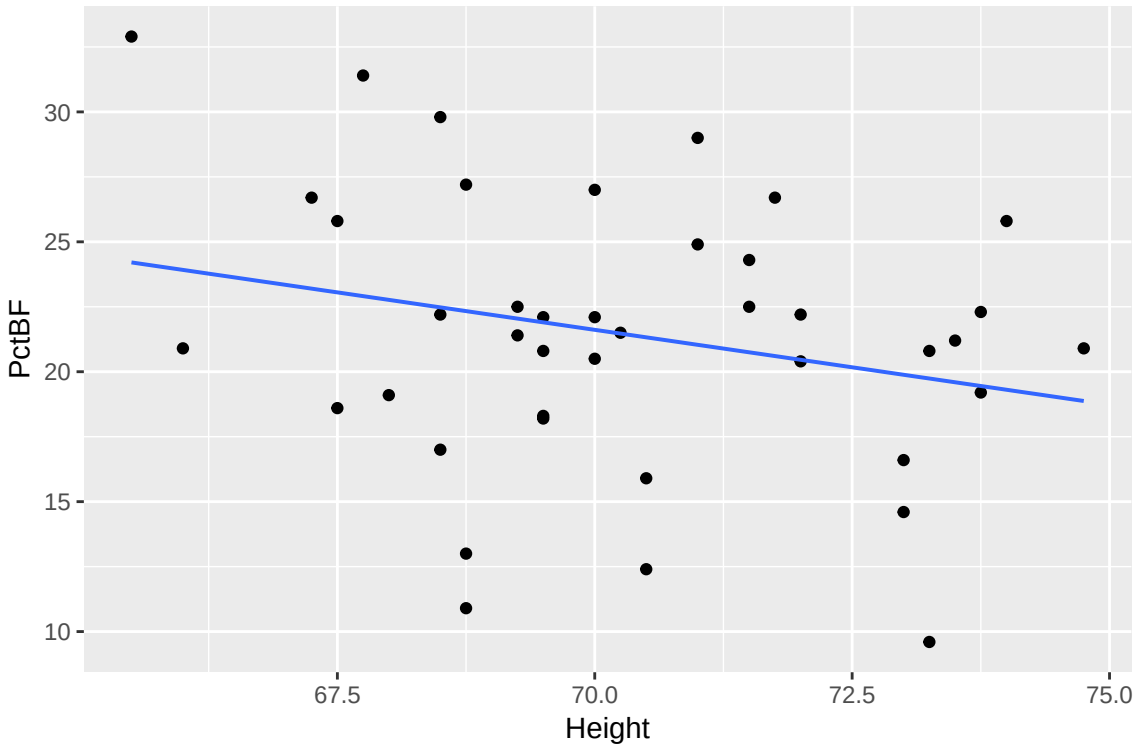
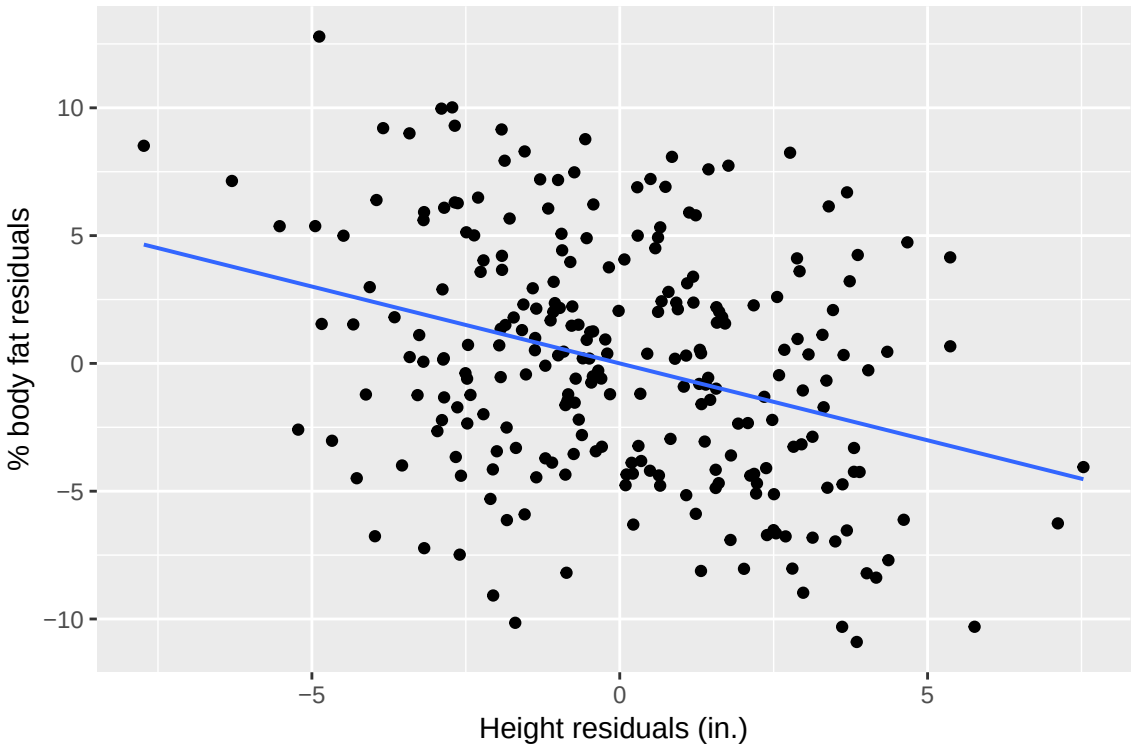


Figure 25.3 (page 745) displays the partial regression plot for weight.

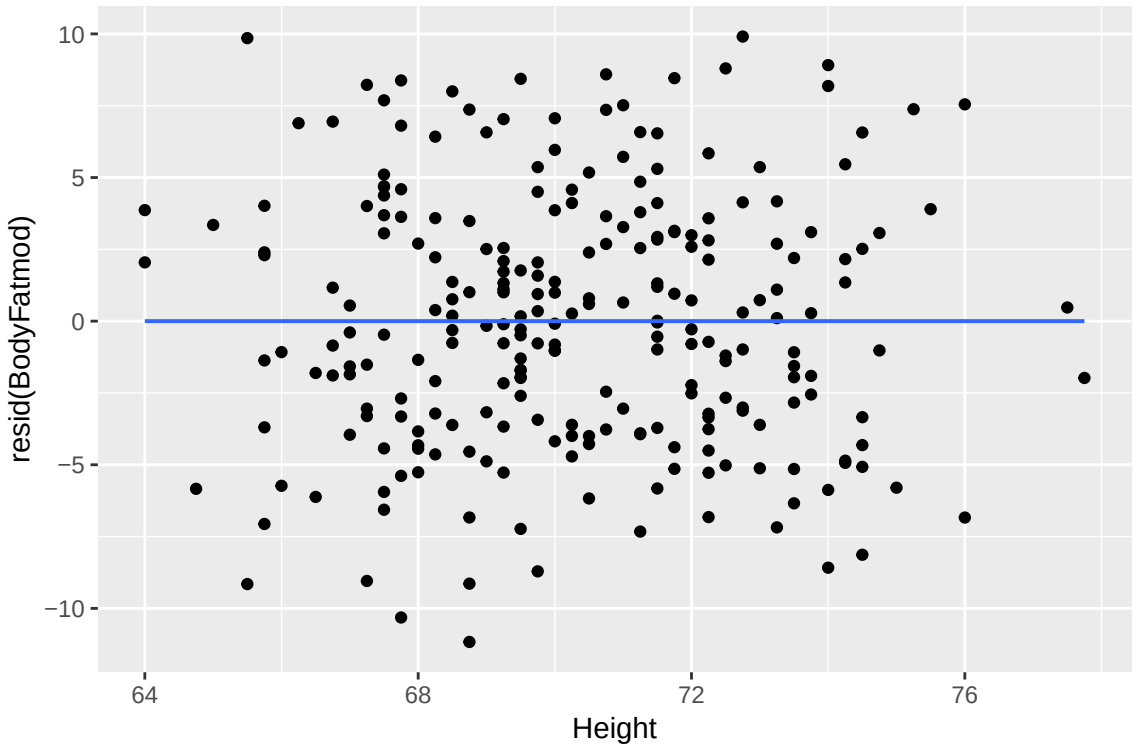
```
BodyFatwaist <- lm(PctBF ~ waist, data = BodyFat)
BodyFatheight <- lm(Height ~ waist, data = BodyFat)

gf_point(resid(BodyFatwaist) ~ resid(BodyFatheight)) %>%
  gf_lm() %>%
  gf_labs(y = "% body fat residuals", x = "Height residuals (in.)")
```



Section 25.3: The Multiple Regression Model

Figure 25.4 (page 747) displays scatterplots of residuals vs. height and waist, respectively.

```
gf_point(resid(BodyFatmod) ~ Height, data = BodyFat) %>%
  gf_lm()
```




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

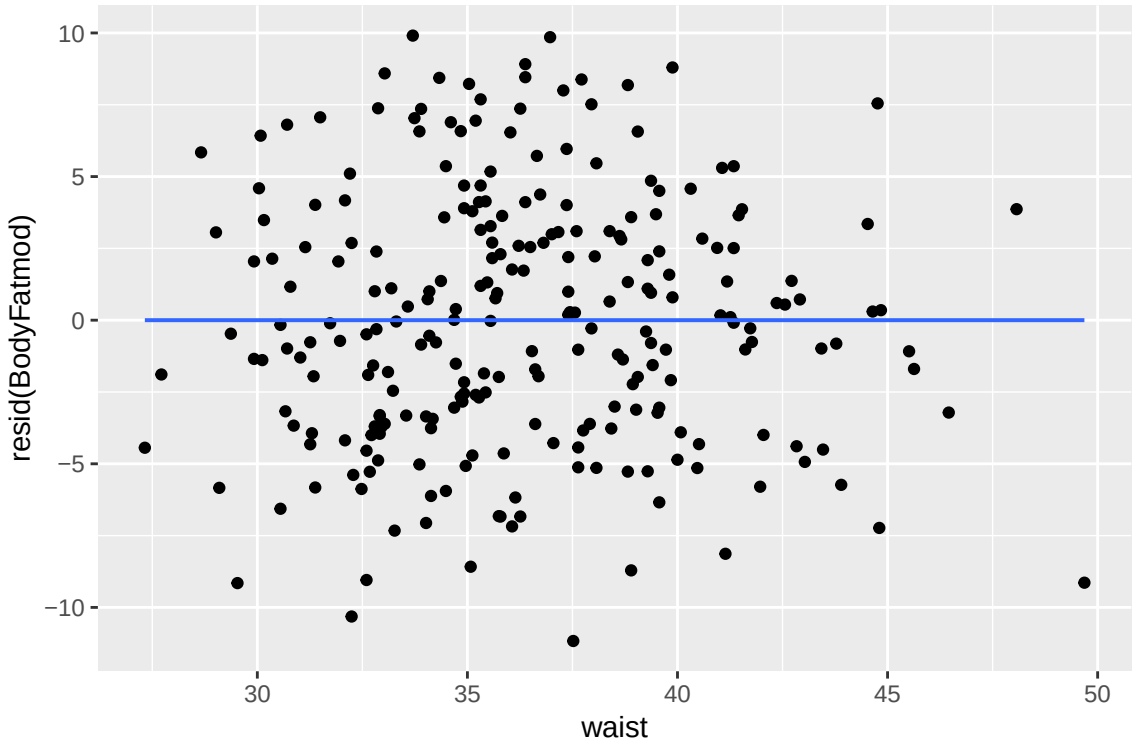
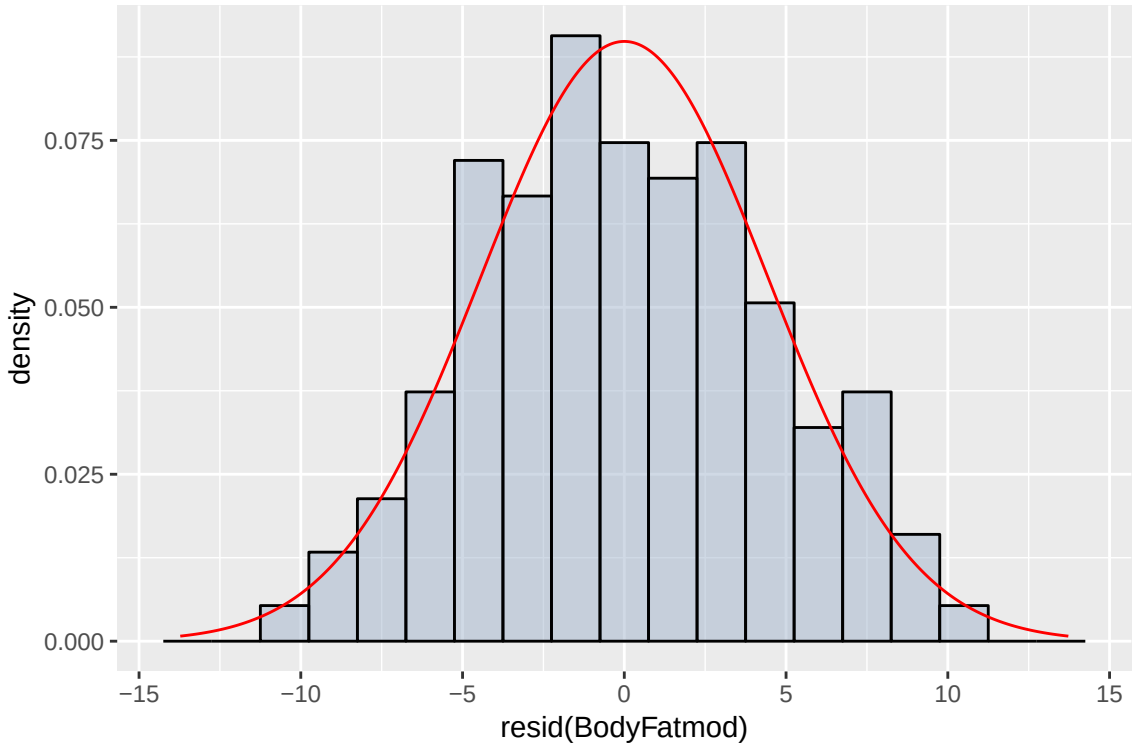
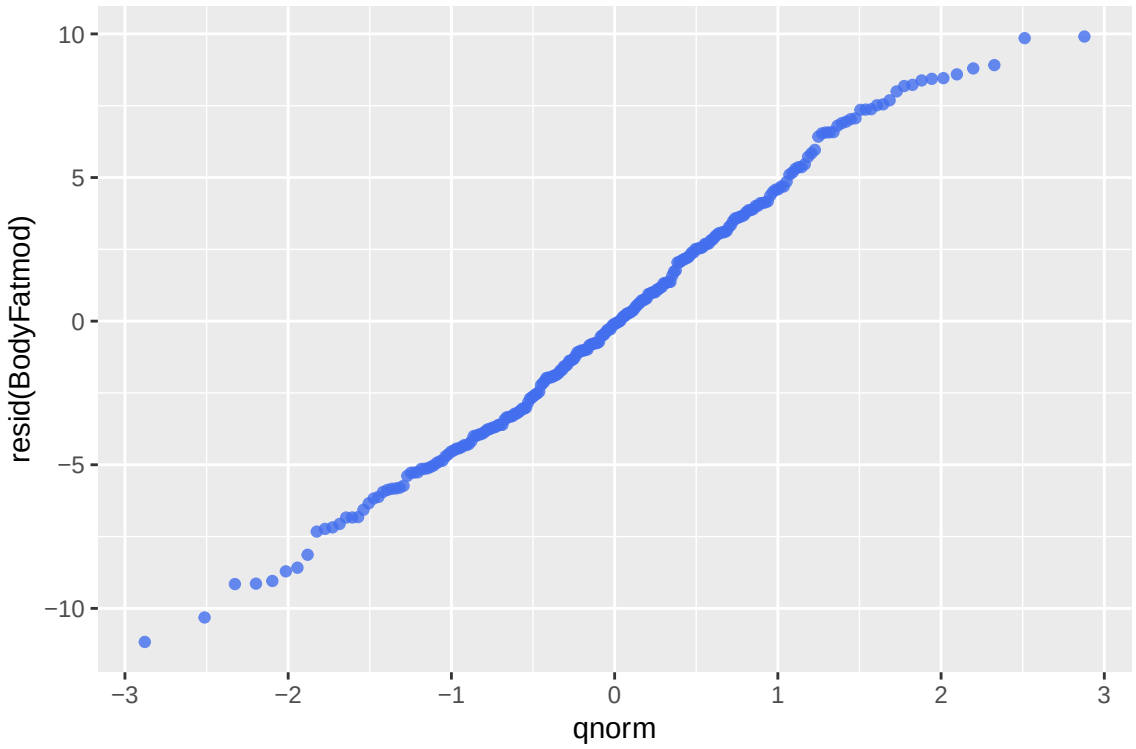


Figure 25.5 (page 747) displays histogram and qq plot of the residuals.

```
res <- resid(BodyFatmod)
me <- mean(res)
std <- sd(res)

gf_histogram(..density.. ~ resid(BodyFatmod),
             binwidth = 7.5/5, fill = "lightsteelblue3", apha = 0.8, col = TRUE) %>%
  gf_dist("norm", mean = me, sd = std, col="red")
```



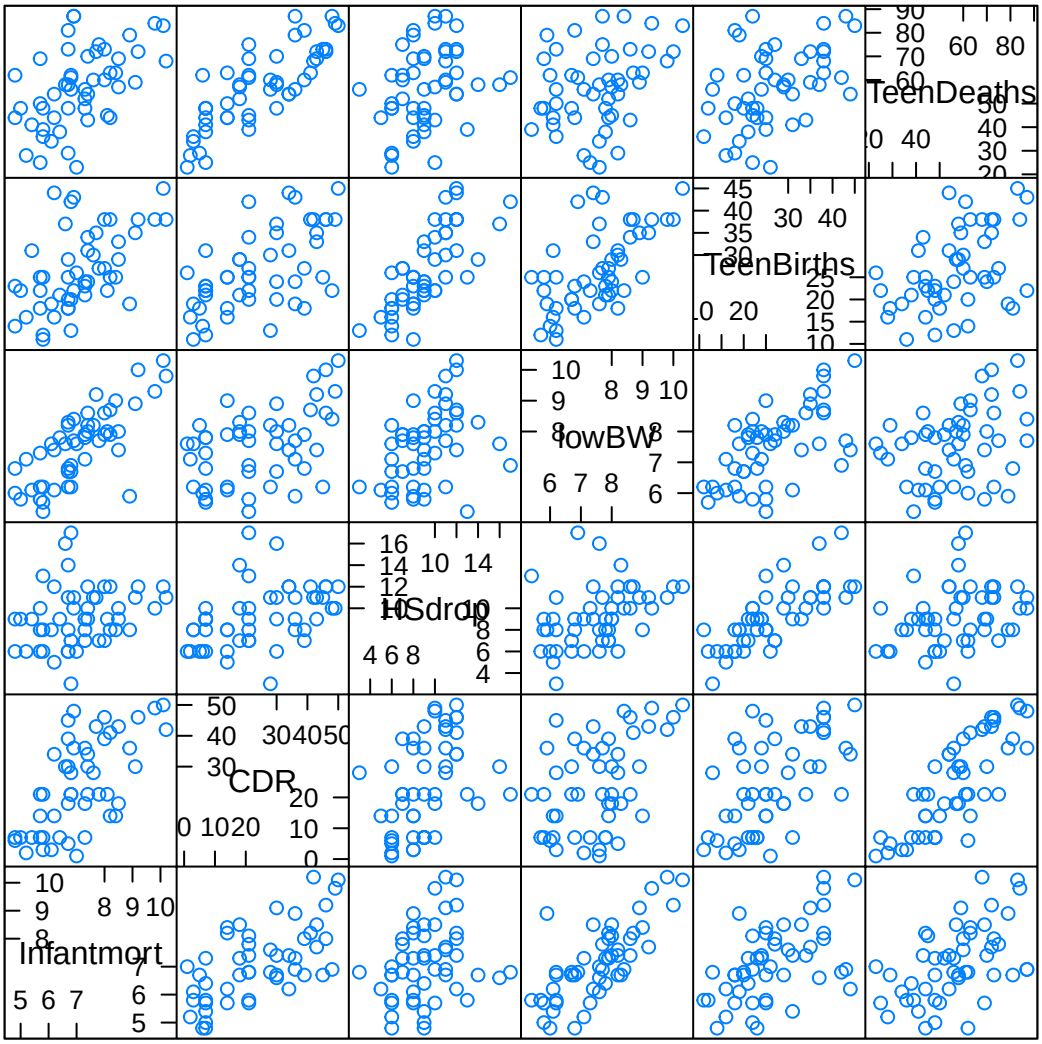

```
gf_qq(~ resid(BodyFatmod), alpha = 0.8, col = "royalblue2") %>%
  gf_labs(x = "qnorm", y = "resid(BodyFatmod)")
```



Section 25.4: Multiple Regression Inference

Figure 25.6 (page 754) displays the scatterplot matrix infant mortality data.

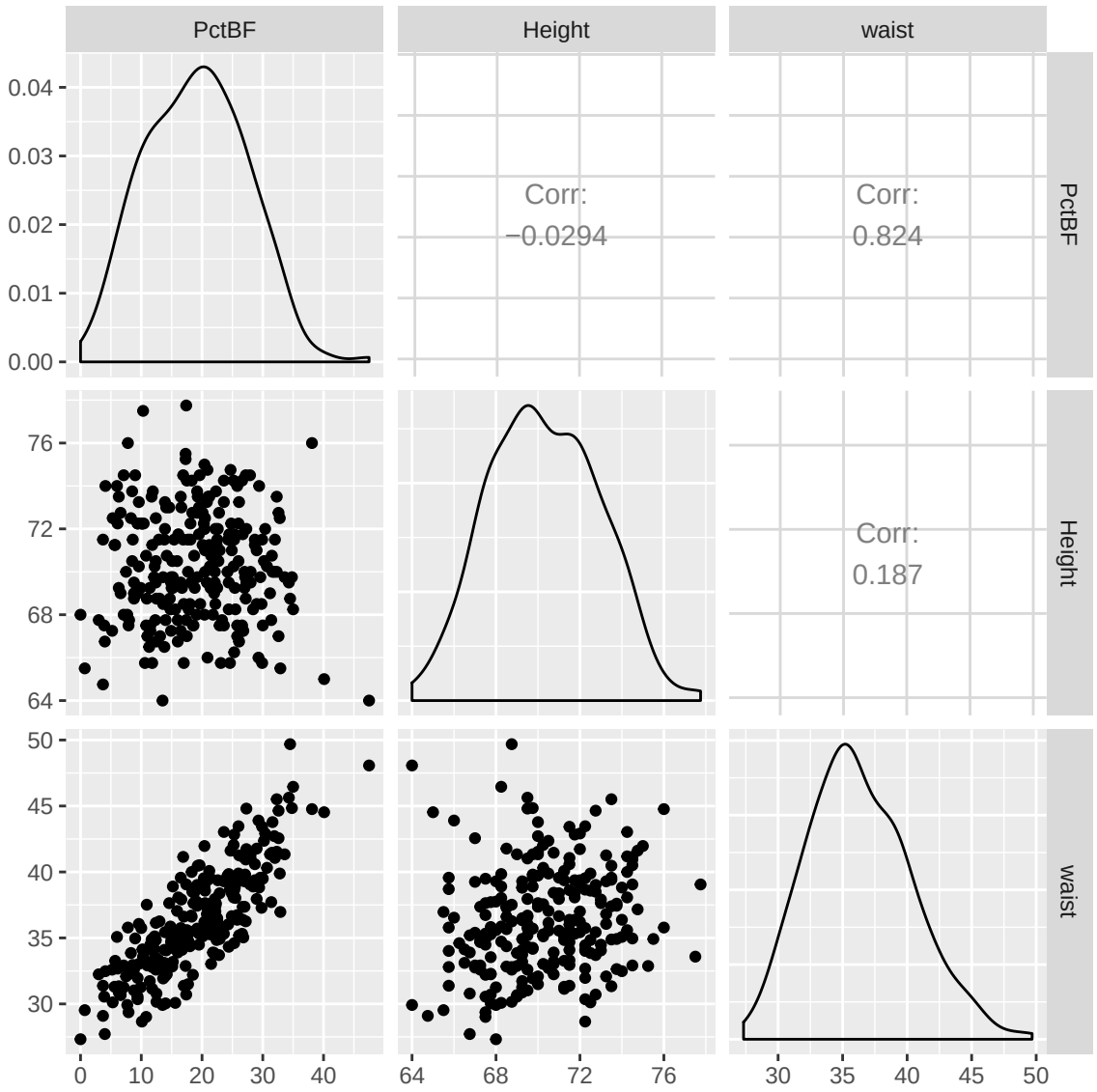
```
InfantMortality <- read_csv("https://nhorton.people.amherst.edu/sdm4/data/Infant_Mortality.csv")
splom(select(InfantMortality, -State))
```



Scatter Plot Matrix

In addition, we display a scatterplot matrix for the motivating example from the chapter (BodyFat) using the `GGally` package.

```
subsetBodyFat <- select(BodyFat, PctBF, Height, waist)
library(GGally)
ggpairs(subsetBodyFat)
```



Section 25.5: Comparing Multiple Regression Models

We may want to compare which of our models provides the most parsimonious fit to these data.

```
msummary(BodyFatheight)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.8864     1.4848   44.37   <2e-16 ***
## waist         0.1216     0.0406    2.99    0.003 **
##
## Residual standard error: 2.58 on 248 degrees of freedom
## Multiple R-squared:  0.0349, Adjusted R-squared:  0.031
## F-statistic: 8.96 on 1 and 248 DF,  p-value: 0.00305
```

```
msummary(BodyFatwaist)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.7341     2.7165   -15.7   <2e-16 ***
## waist         1.7000     0.0743    22.9   <2e-16 ***
##
## Residual standard error: 4.71 on 248 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.677
## F-statistic:  523 on 1 and 248 DF,  p-value: <2e-16
```

```
msummary(BodyFatmod)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.1009     7.6861   -0.40     0.69
## waist         1.7731     0.0716   24.77  < 2e-16 ***
## Height       -0.6015     0.1099   -5.47  1.1e-07 ***
##
## Residual standard error: 4.46 on 247 degrees of freedom
## Multiple R-squared:  0.713,  Adjusted R-squared:  0.711
## F-statistic:  307 on 2 and 247 DF,  p-value: <2e-16
```

The adjusted R-squared value of 0.711 is considerably higher for the model with both predictors (though the model with just waist has an adjusted R-squared value of 0.677).