

IS4 in R: Linear Regression (Chapter 7)

Patrick Frenett, Vickie Ip, 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 <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 7: Linear Regression

Section 7.1: Least squares: the line of best fit

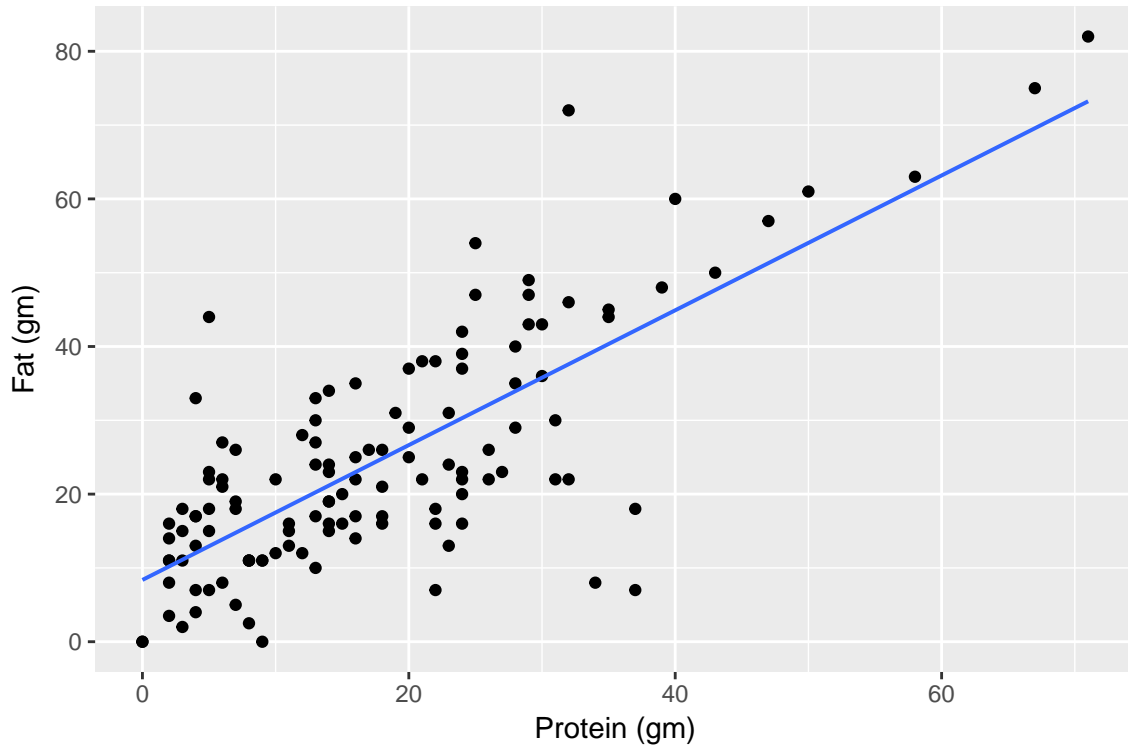
Figure 7.2 (page 179) displays a scatterplot of the Burger King data with a superimposed regression line.

```
library(mosaic); library(readr)
options(digits=3)
BK <- read_csv("https://nhorton.people.amherst.edu/sdm4/data/Burger_King_Items.csv")
```

```
names(BK)
```

```
## [1] "Item"      "Servingsize" "Calories"    "Fatcal"      "Fat"
## [6] "Sat"      "Transfat(g)" "Chol(mg)"    "Sodium(mg)"  "Carb(g)"
## [11] "Fiber(g)"  "Sugar(g)"    "Protein"
```

```
gf_point(Fat ~ Protein, data=BK) %>%
  gf_lm() %>%
  gf_labs(y="Fat (gm)", x="Protein (gm)")
```



We can calculate the residual for a particular value with 31 grams of protein.

```
BKmod <- lm(Fat ~ Protein, data=BK)
BKfun <- makeFun(BKmod)
BKfun(31) # predicted value for a item with 31 grams of protein
```

```
## 1
## 36.7
```

Section 7.2 The linear model

```
coef(BKmod)
```

```
## (Intercept) Protein
## 8.372 0.913
```

```
BKfun(0)
```

```
## 1
## 8.37
```

```
BKfun(32) - BKfun(31)
```

```
## 1
## 0.913
```

Section 7.3 Finding the least squares line

```
sx <- sd(~ Protein, data=BK); sx
```

```
## [1] 13.5
```

```
sy <- sd(~ Fat, data=BK); sy
```

```
## [1] 16.2
```

```
r <- cor(Protein ~ Fat, data=BK); r # same as cor(Fat ~ Protein)!
```

```
## [1] 0.761
```

```
r*sy/sx
```

```
## [1] 0.913
```

```
coef(BKmod)[2]
```

```
## Protein
```

```
## 0.913
```

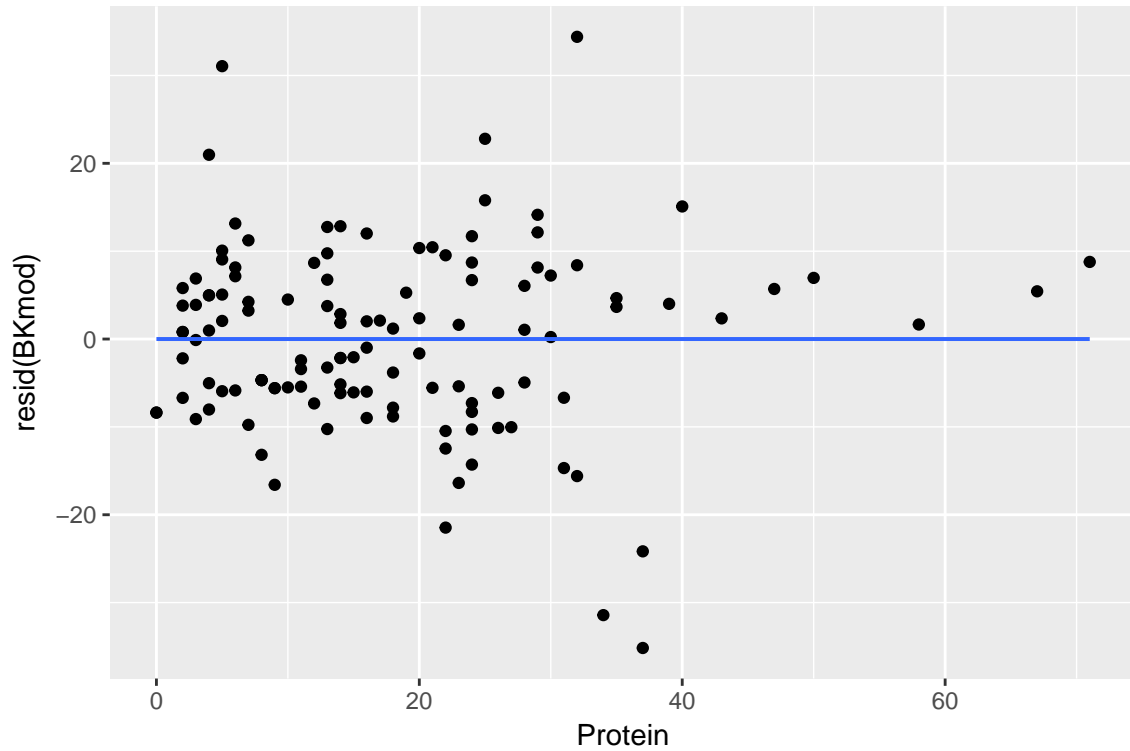
Section 7.4 Regression to the mean

Section 7.5 Examining the residuals

Figure 7.5 (page 188) displays the scatterplot of residuals as a function of the amount of protein.

The `summary` function generates a lot of output (much of which won't be familiar).

```
gf_point(resid(BKmod) ~ Protein, data=BK) %>%  
  gf_lm()
```



```
msummary(BKmod)
```

```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.3720     1.5991     5.24 7.1e-07 ***
## Protein      0.9134     0.0712    12.84 < 2e-16 ***
##
## Residual standard error: 10.6 on 120 degrees of freedom
## Multiple R-squared:  0.579, Adjusted R-squared:  0.575
## F-statistic: 165 on 1 and 120 DF, p-value: <2e-16
```

The residual standard error of 10.6 grams matches the value reported on page 194.

Section 7.6 R-squared: variation accounted for by the model

```
rsquared(BKmod)
```

```
## [1] 0.579
```

Section 7.7 Regression assumptions and conditions