

# SDM4 in R: Linear Regression (Chapter 7)

Nicholas Horton ([nhorton@amherst.edu](mailto:nhorton@amherst.edu))

January 2, 2017

## 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](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>).

## Chapter 7: Linear Regression

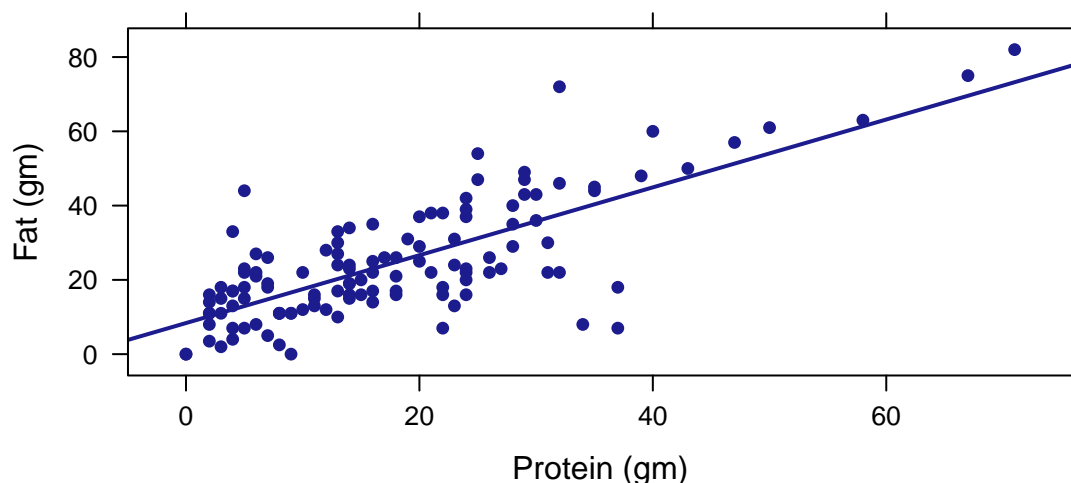
### Section 7.1: Least squares: the line of best fit

Figure 7.2 (page 183) displays a scatterplot of the Burger King data with a superimposed regression line.

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

```
## [1] "Item"          "Servingsize"  "Calories"     "FatCal"       "Fat"
## [6] "SatFat"       "Transfat"     "Chol"         "Sodium"       "Carb"
## [11] "Fiber"        "Sugar"        "Protein"
```

```
xyplot(Fat ~ Protein, ylab="Fat (gm)", xlab="Protein (gm)", type=c("p", "r"), data=BK)
```



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]
```

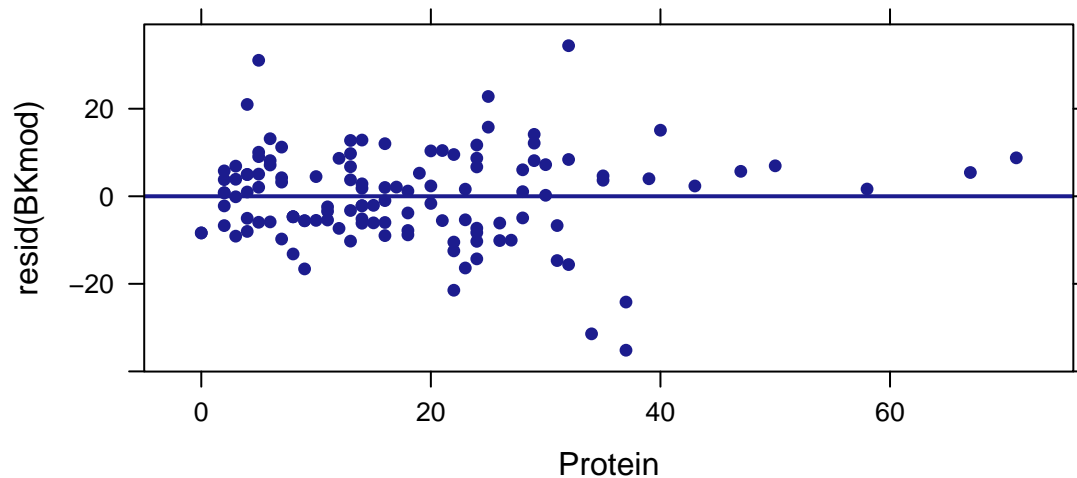
```
## Protein
##      0.913
```

## Section 7.4 Regression to the mean

## Section 7.5 Examining the residuals

Figure 7.5 (page 193) displays the scatterplot of residuals as a function of the amount of protein. The `msummary()` function generates a lot of output (much of which won't be familiar).

```
xyplot(resid(BKmod) ~ Protein, type=c("p", "r"), data=BK)
```



```
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