SDM4 in R: Inferences for Regression (Chapter 25)

Nicholas Horton (nhorton@amherst.edu) and Sarah McDonald

June 16, 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 25: Inferences for Regression

Section 25.1: The population and the sample

```r
library(mosaic)
library(readr)
BodyFat <- read_csv("http://nhorton.people.amherst.edu/sdm4/data/Body_fat_complete.csv")
dim(BodyFat)
```

```
## [1] 250  16
```

```r
glimpse(BodyFat)
```

```
## Observations: 250
## Variables: 16
## $ `Body Density` <dbl> 1.07, 1.09, 1.04, 1.08, 1.03, 1.05, 1.05, 1.07,...
## $ PctBF          <dbl> 12.3, 6.1, 25.3, 10.4, 28.7, 20.9, 19.2, 12.4, ...
## $ Age            <int> 23, 22, 22, 26, 24, 24, 26, 25, 25, 23, 26, 27,...
## $ Weight         <dbl> 154, 173, 154, 185, 184, 210, 181, 176, 191, 19...
## $ Height         <dbl> 67.8, 72.2, 66.2, 72.2, 71.2, 74.8, 69.8, 72.5,...
## $ Neck           <dbl> 36.2, 38.5, 34.0, 37.4, 34.4, 39.0, 36.4, 37.8,...
## $ Chest          <dbl> 93.1, 93.6, 95.8, 101.8, 97.3, 104.5, 105.1, 99...
## $ Abdomen        <dbl> 85.2, 83.0, 87.9, 86.4, 100.0, 94.4, 90.7, 88.5...
## $ waist          <dbl> 33.5, 32.7, 34.6, 34.0, 39.4, 37.2, 35.7, 34.8,...
## $ Hip            <dbl> 94.5, 98.7, 99.2, 101.2, 101.9, 107.8, 100.3, 9...
## $ Thigh          <dbl> 59.0, 58.7, 59.6, 60.1, 63.2, 66.0, 58.4, 60.0,...
## $ Knee           <dbl> 37.3, 37.3, 38.9, 37.3, 42.2, 42.0, 38.3, 39.4,...
## $ Ankle          <dbl> 21.9, 23.4, 24.0, 22.8, 24.0, 25.6, 22.9, 23.2,...
## $ Bicep          <dbl> 32.0, 30.5, 28.8, 32.4, 32.2, 35.7, 31.9, 30.5,...
## $ Forearm        <dbl> 27.4, 28.9, 25.2, 29.4, 27.7, 30.6, 27.8, 29.0,...
## $ Wrist          <dbl> 17.1, 18.2, 16.6, 18.2, 17.7, 18.8, 17.7, 18.8,...
```

We can confirm the coefficients from the model on page 690.

```
BodyFatmod <- lm(PctBF ~ waist, data = BodyFat)
coef(BodyFatmod)
```

```
## (Intercept)       waist
##       -42.7         1.7
```

Section 25.2: Assumptions and conditions

We can regenerate the output and figures for the example on pages 692-696.

```
msummary(BodyFatmod)
```

```
##             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
```

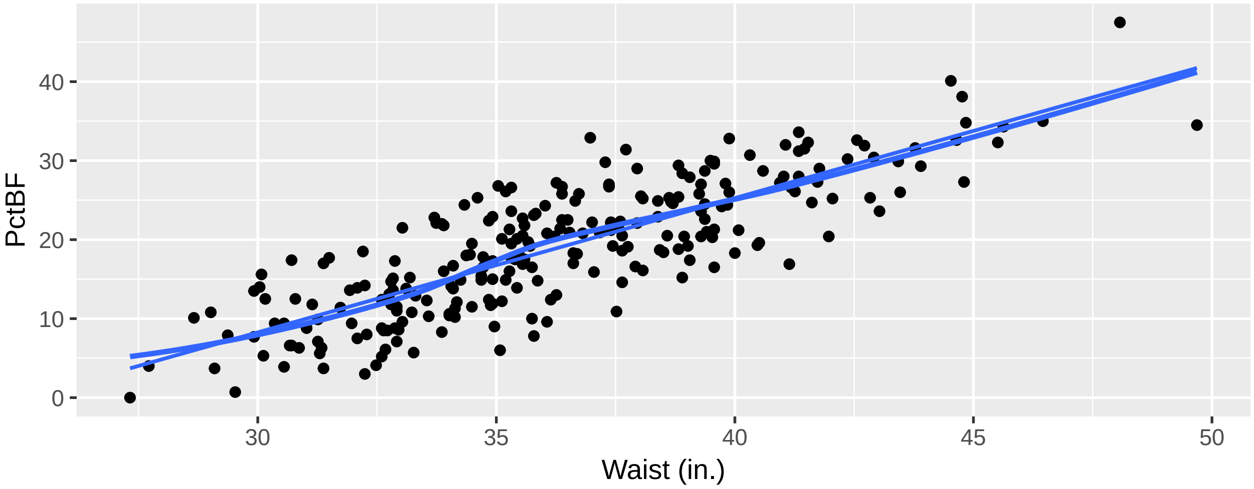
```
rsquared(BodyFatmod)
```

```
## [1] 0.678
```

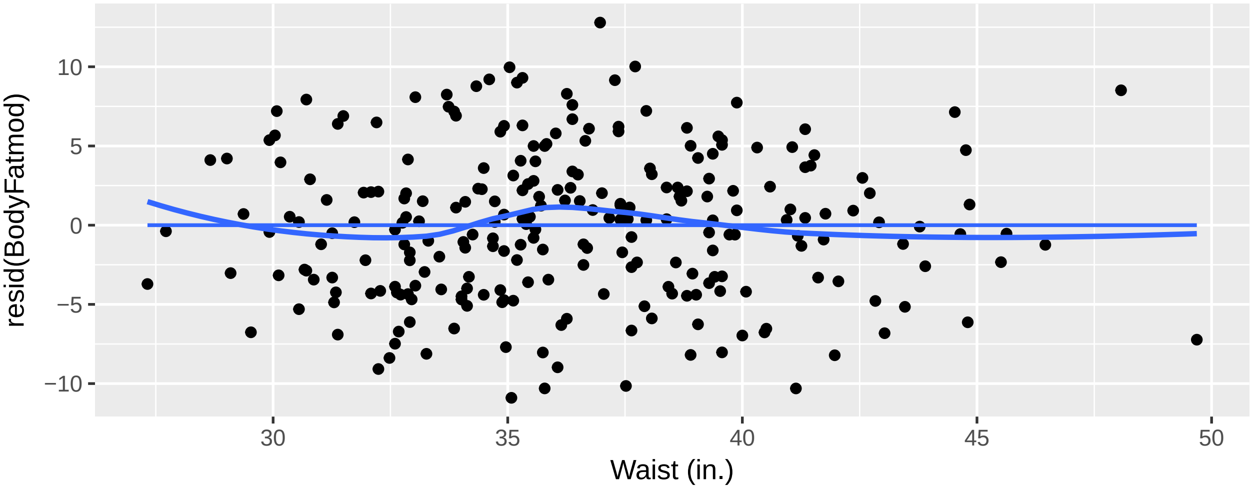
```
confint(BodyFatmod)     # see page 700
```

```
##              2.5 % 97.5 %
## (Intercept) -48.08 -37.38
## waist         1.55   1.85
```

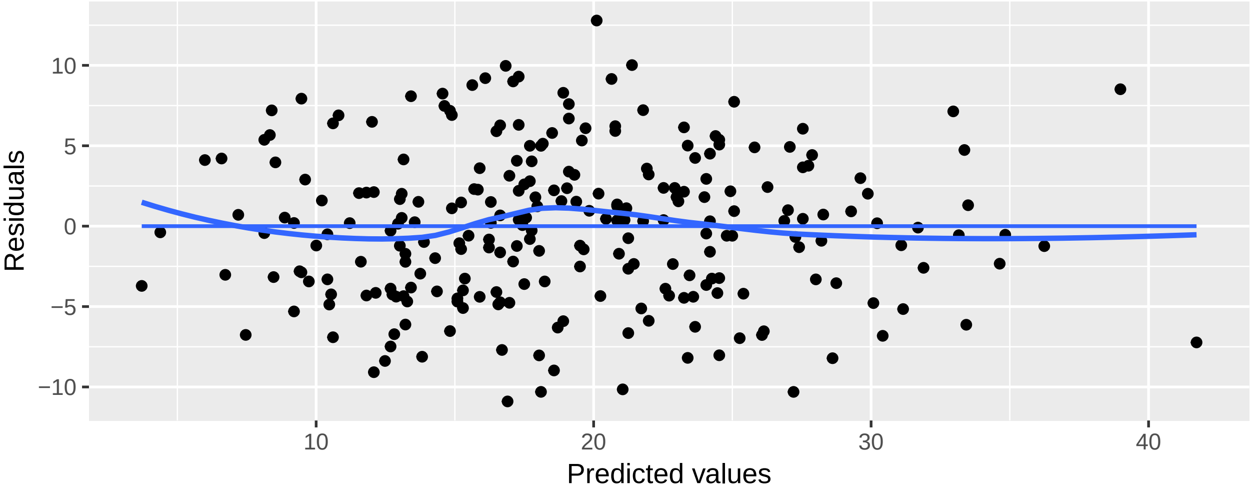
```
# Figure 25.4
gf_point(PctBF ~ waist, xlab = "Waist (in.)",
      data = BodyFat)  %>%  # see smoothers on p.92-93
  gf_lm() %>%
  gf_smooth(se = FALSE)
```



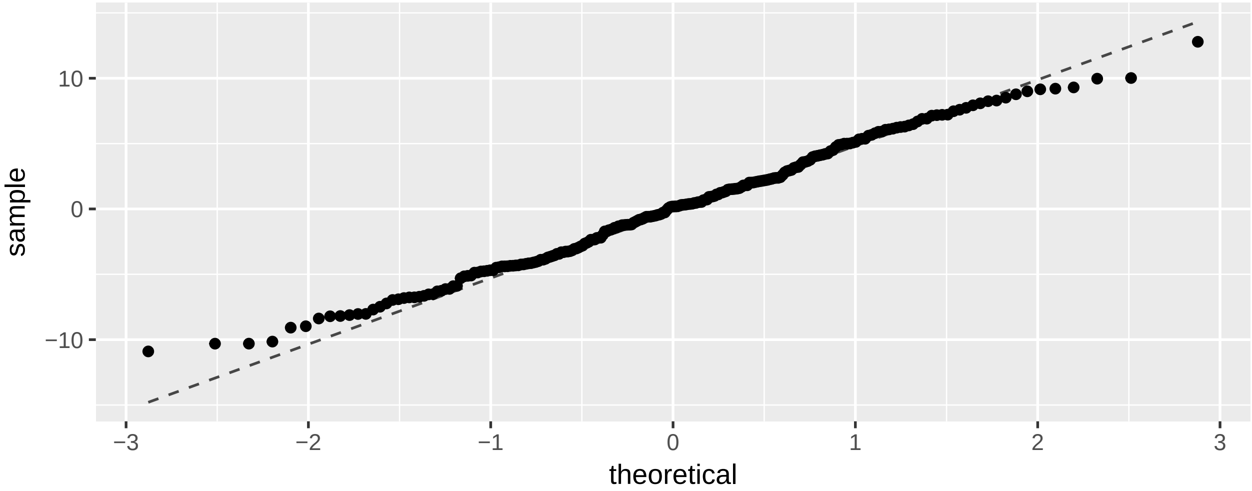
```
# Figure 25.5
gf_point(resid(BodyFatmod) ~ waist, xlab = "Waist (in.)",
       data = BodyFat) %>%
  gf_lm() %>%
  gf_smooth(se = FALSE)
```



```
# equiv of Figure 25.6   note that Figure 25.6 refers to the diamonds dataset
gf_point(resid(BodyFatmod) ~ fitted(BodyFatmod), xlab = "Predicted values",
       ylab = "Residuals",
       type = c("p", "r", "smooth"), data = BodyFat) %>%
  gf_lm() %>%
  gf_smooth(se = FALSE)
```



```
# Figure on bottom of page 695
gf_qq(~ resid(BodyFatmod)) %>%
  gf_qqline()
```

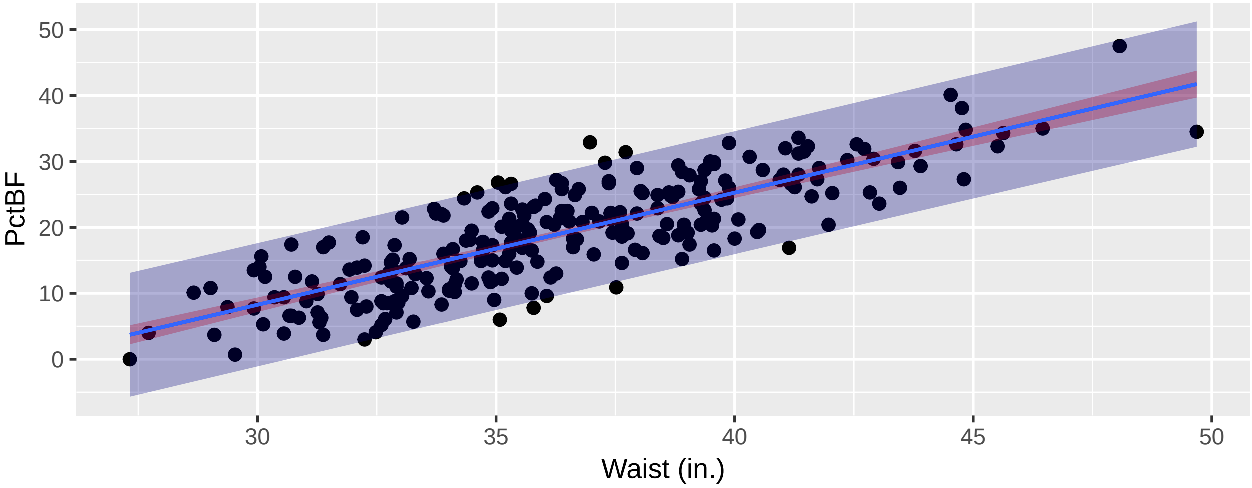


Section 25.6: Confidence intervals for predicted values

We can reproduce Figure 25.12 (page 707) using layers in ggformula.

```
library(broom)
gf_point(PctBF ~ waist, xlab = "Waist (in.)",
       panel = panel.lmbands, lwd = 2, cex = 0.2, data = BodyFat) %>%
  gf_lm(interval = "confidence", fill = "red") %>%
  gf_lm(interval = "prediction", fill = "navy")
```

```
## Warning: The plyr::rename operation has created duplicates for the
## following name(s): (`size`)
```



```
Craters <- read.csv("http://nhorton.people.amherst.edu/sdm4/data/Craters.csv")
dim(Craters)
```

```
## [1] 168   4
```

```
Craters <- mutate(Craters,
                  logDiam = log(Diam.km.),
                  logAge = log(age..Ma.))
Cratermod <- lm(logDiam ~ logAge, data = Craters)
favstats(~ logAge, data = Craters)   # note example in book has n=39
```

```
##    min   Q1 median   Q3  max mean   sd   n missing
##  -9.81 3.61   4.82 5.95 7.78 3.76 3.46 168       0
```

```
confpred <- predict(Cratermod, interval = "confidence")
intpred <- predict(Cratermod, interval = "prediction")
```

```
## Warning in predict.lm(Cratermod, interval = "prediction"): predictions on current data refer to _futu
```

```
select(Craters, -Name) %>%
  head(., 3)
```

```
##                              Location Diam.km. age..Ma. logDiam logAge
## 1                        Kansas, U.S.A.    0.015  1.0e-03   -4.20  -6.91
## 2 Western Australia,        Australia    0.024  2.7e-01   -3.73  -1.31
## 3                               Russia    0.027  5.5e-05   -3.61  -9.81
```

```
head(confpred, 3)
```

```
##       fit    lwr    upr
## 1 -2.1535 -2.766 -1.541
## 2 -0.0639 -0.399  0.271
## 3 -3.2362 -4.001 -2.471
```

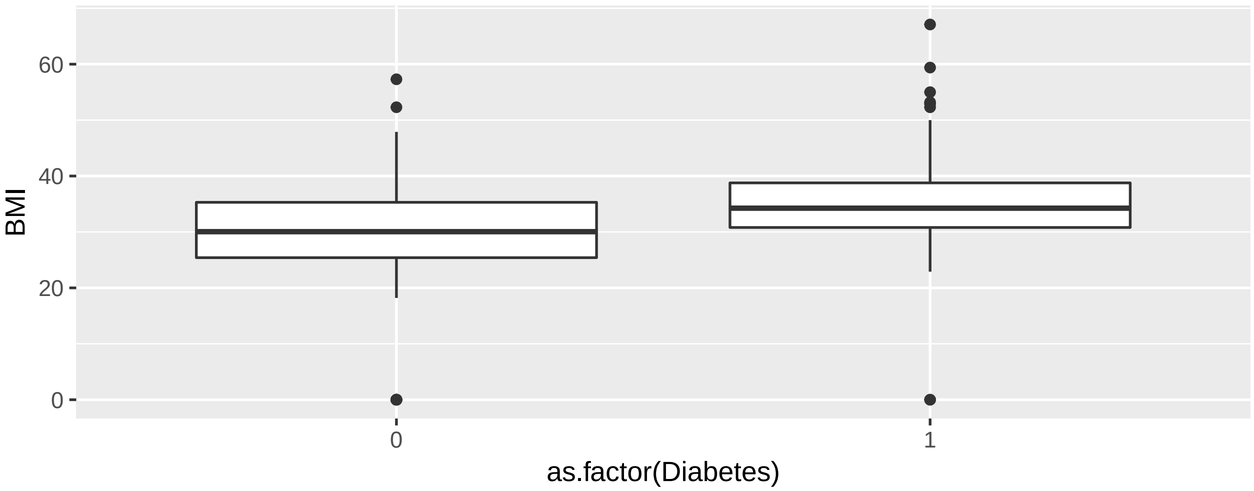
```
head(intpred, 3)
```

```
##       fit   lwr    upr
## 1 -2.1535 -4.68  0.368
## 2 -0.0639 -2.53  2.405
## 3 -3.2362 -5.80 -0.673
```

Section 25.7: Logistic regression

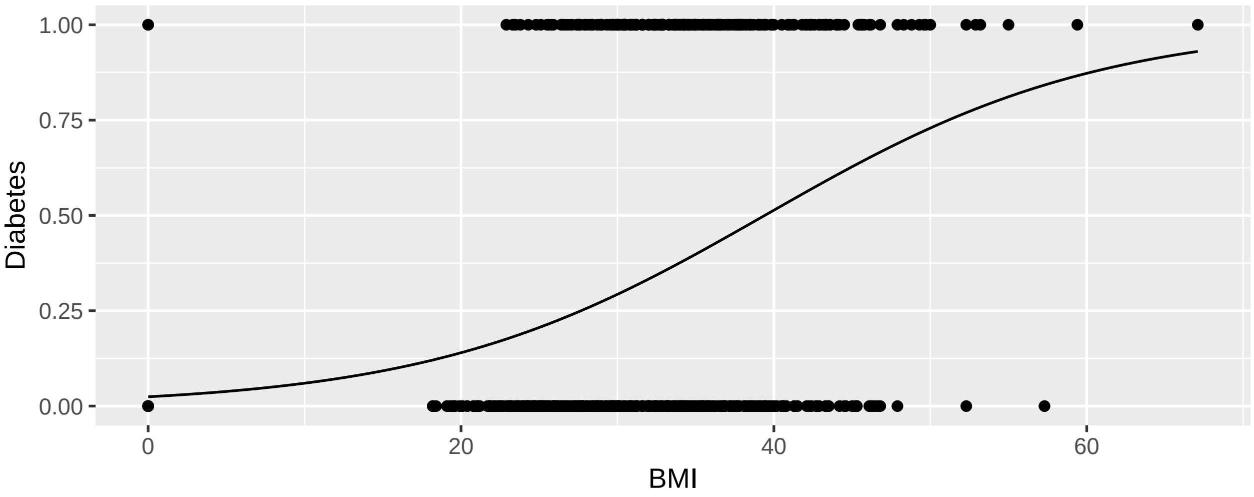
The Pima Indian dataset example is given on pages 708-712.

```
Pima <- read_csv("http://nhorton.people.amherst.edu/sdm4/data/Pima_Indians_Diabetes.csv")
Diabetes <- filter(Pima, BMI>0)  # get rid of missing values for BMI
gf_boxplot(BMI ~ as.factor(Diabetes), data = Pima)
```



```
pimamod <- glm(Diabetes ~ BMI, family = "binomial", data = Pima)
f2 <- makeFun(pimamod)
gf_point(Diabetes ~ BMI, data = Pima) %>%
gf_function(f2, add = TRUE)
```

```
## Warning: Ignoring unknown parameters: add
```



```
msummary(pimamod)
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.6864     0.4090   -9.01  < 2e-16 ***
## BMI            0.0935     0.0121    7.76  8.4e-15 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 920.71  on 766  degrees of freedom
## AIC: 924.7
##
## Number of Fisher Scoring iterations: 4
```