IPS9 in R: Multiple regression (Chapter 11)

Nicholas Horton (nhorton@amherst.edu)

January 19, 2019

Introduction and background

These documents are intended to help describe how to undertake analyses introduced as examples in the Ninth Edition of *Introduction to the Practice of Statistics* (2017) by Moore, McCabe, and Craig.

More information about the book can be found here. The data used in these documents can be found under Data Sets in the Student Site. 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/ips9/.

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 11: Multiple Regression

This file replicates the analyses from Chapter 11: Multiple Regression.

First, load the packages that will be needed for this document:

```
library(mosaic)
library(readr)
```

Section 11.1: Inference for multiple regression

Example 11.1: Predicting early success in college

```
GPA <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter11/EG11-01GPA.csv")
# Figure 11.1, page 609
head(GPA)
```

```
## # A tibble: 6 x 9
##     OBS   GPA   HSM   HSS   HSE  SATM SATCR  SATW   SEX
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1  3.84    10    10    10   630   570   590     1
## 2     2  3.97    10    10    10   750   700   630     0
## 3     3  3.49     8    10     9   570   510   490     1
## 4     4  1.95     6     4     8   640   600   610     0
## 5     5  2.59     8    10     9   510   490   490     1
## 6     6  3        7    10    10   660   680   630     0
```

Section 11.2: A Case Study

```
# Figure 11.2, page 619
favstats(~ GPA, data = GPA)
```

```
##   min     Q1 median   Q3 max     mean        sd   n missing
##  0.03 2.3025  2.975 3.45   4 2.842133 0.8178992 150       0
```

```
favstats(~ HSM, data = GPA)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##    2  8      9 10  10 8.586667 1.461757 150       0
```

```
favstats(~ HSS, data = GPA)
```

```
##  min Q1 median Q3 max mean       sd   n missing
##    4  8      9 10  10  8.8 1.395102 150       0
```

```
favstats(~ HSE, data = GPA)
```

```
##  min Q1 median Q3 max     mean      sd   n missing
##    4  8      9 10  10 8.833333 1.26606 150       0
```

```
favstats(~ SATM, data = GPA)
```

```
##  min  Q1 median    Q3 max  mean       sd   n missing
##  460 570    630 677.5 800 623.6 74.83566 150       0
```

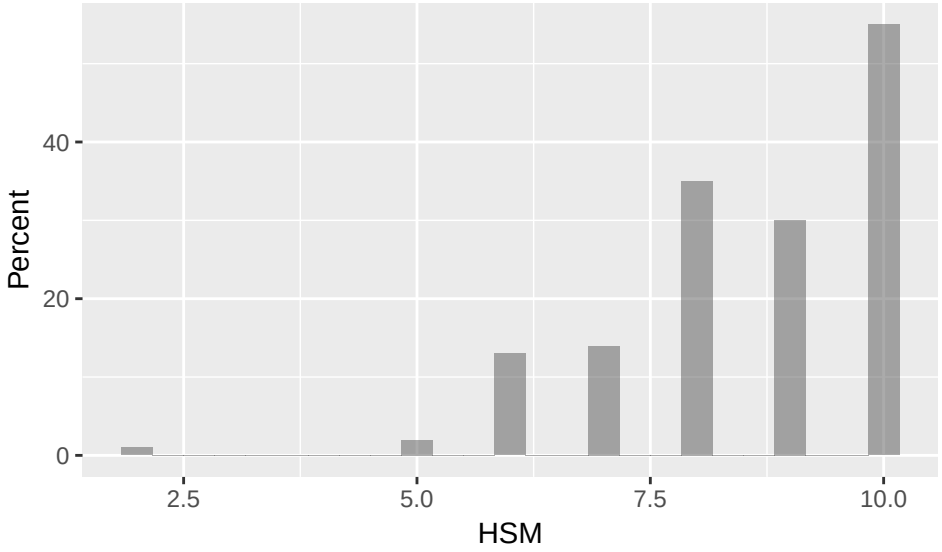
```
favstats(~ SATCR, data = GPA)
```

```
##  min    Q1 median  Q3 max  mean       sd   n missing
##  330 512.5    560 630 800 573.8 87.62083 150       0
```

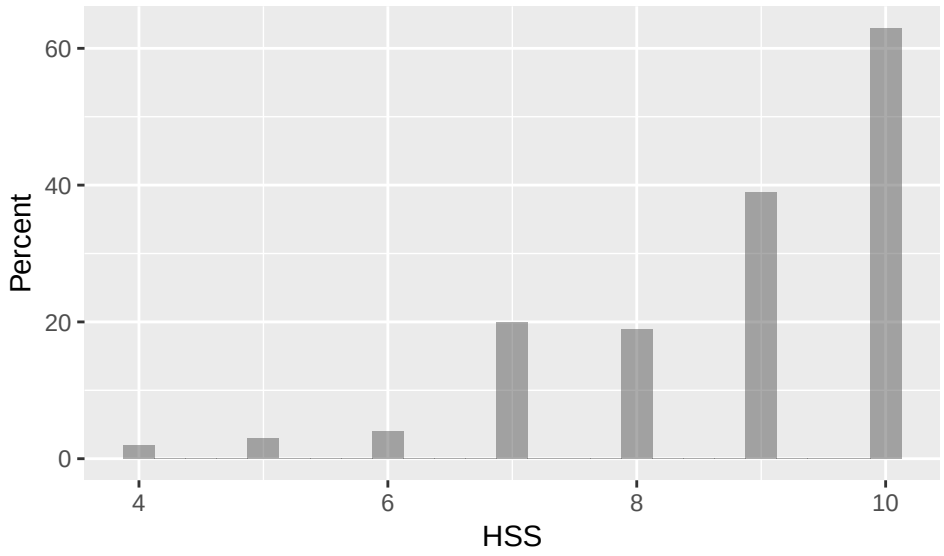
```
favstats(~ SATW, data = GPA)
```

```
##  min  Q1 median  Q3 max  mean       sd   n missing
##  350 490    560 620 770 562.6 80.08745 150       0
```

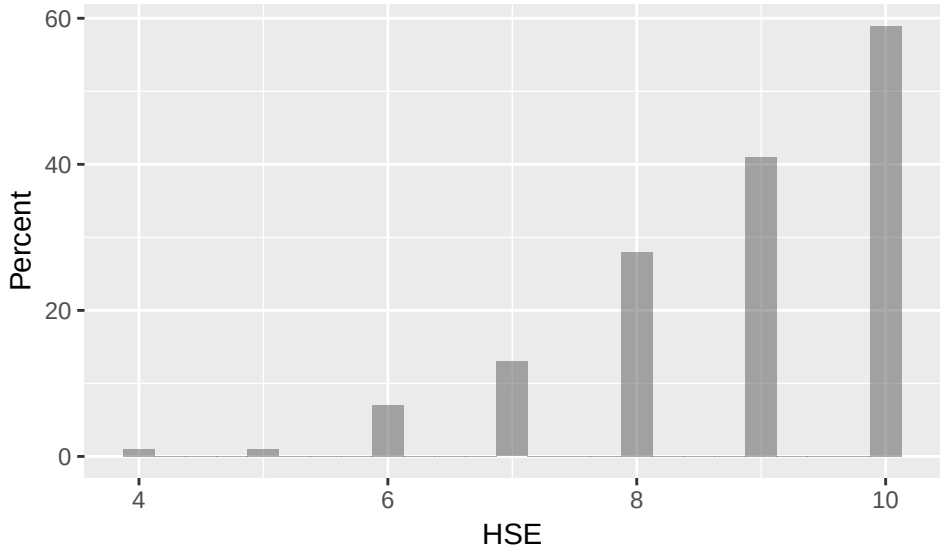
```
# Figure 11.3, page 620
gf_histogram(~ HSM, data = GPA) %>%
  gf_labs(y = "Percent")# Doesn't look great
```




```
gf_histogram(~ HSS, data = GPA) %>%
  gf_labs(y = "Percent")
```




```
gf_histogram(~ HSE, data = GPA) %>%
  gf_labs(y = "Percent")
```



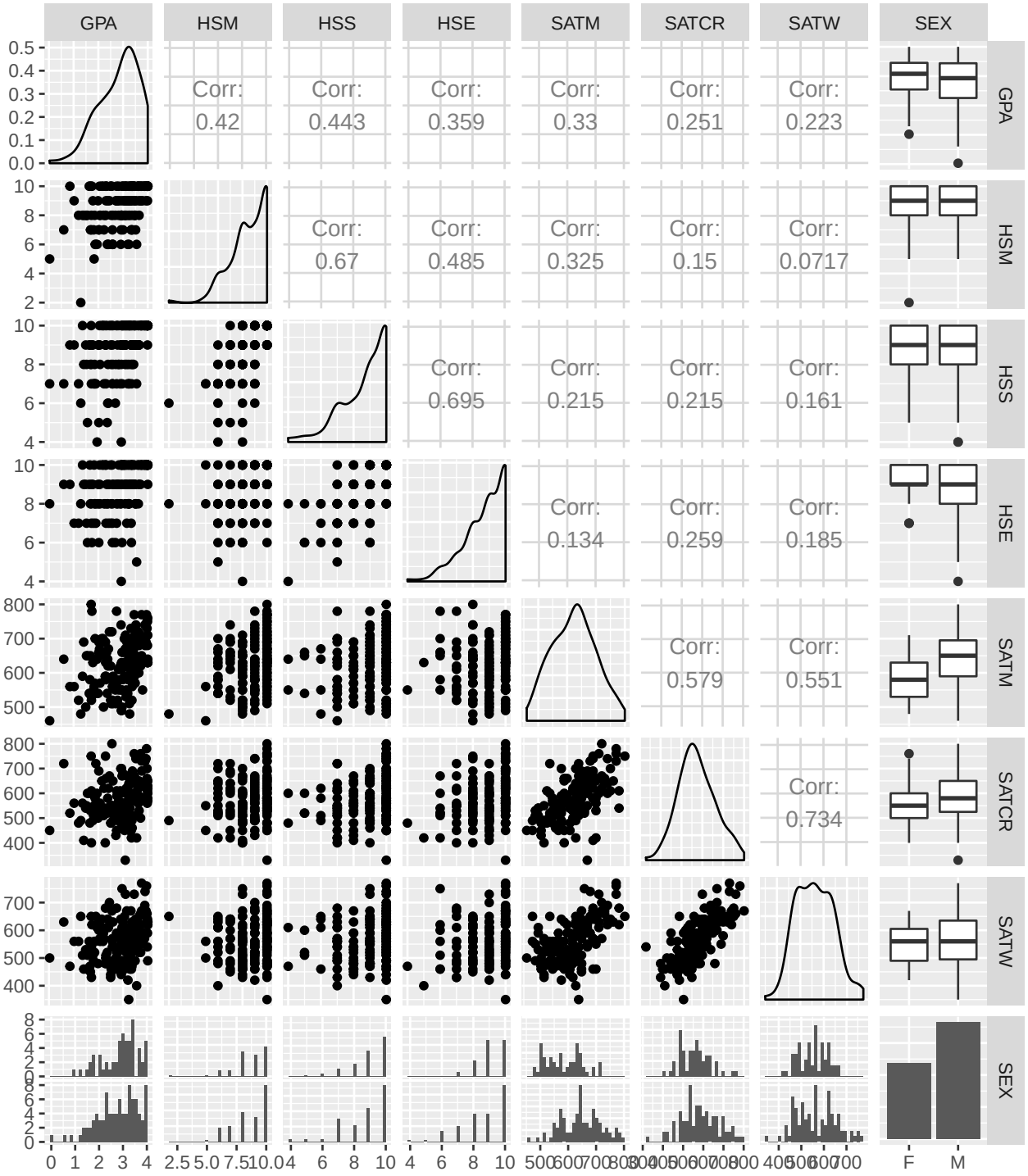
Relationships between pairs of variables

```
# Figure 11.4, page 621
options(digits = 2)
cor(GPA)
```

```
##          OBS    GPA    HSM   HSS  HSE   SATM SATCR   SATW    SEX
## OBS    1.000 -0.018  0.059 0.026 0.12 -0.083 -0.04 -0.057  0.093
## GPA   -0.018  1.000  0.420 0.443 0.36  0.330  0.25  0.223  0.089
## HSM    0.059  0.420  1.000 0.670 0.48  0.325  0.15  0.072 -0.034
## HSS    0.026  0.443  0.670 1.000 0.70  0.215  0.22  0.161  0.096
## HSE    0.117  0.359  0.485 0.695 1.00  0.134  0.26  0.185  0.182
## SATM  -0.083  0.330  0.325 0.215 0.13  1.000  0.58  0.551 -0.408
## SATCR -0.040  0.251  0.150 0.215 0.26  0.579  1.00  0.734 -0.151
## SATW  -0.057  0.223  0.072 0.161 0.19  0.551  0.73  1.000 -0.098
## SEX    0.093  0.089 -0.034 0.096 0.18 -0.408 -0.15 -0.098  1.000
```

Example 11.13: Pairwise relationships among variables in the GPA data set

```
GPA <- read_csv("https://nhorton.people.amherst.edu/ips9/data/chapter11/EX11-13GPA.csv")
GPA <- GPA %>%
  mutate(SEX = ifelse(SEX == 1, "F", "M"))
library(GGally)
# Figure 11.5
GPA %>% select(-OBS) %>%
  GGally::ggpairs()
```



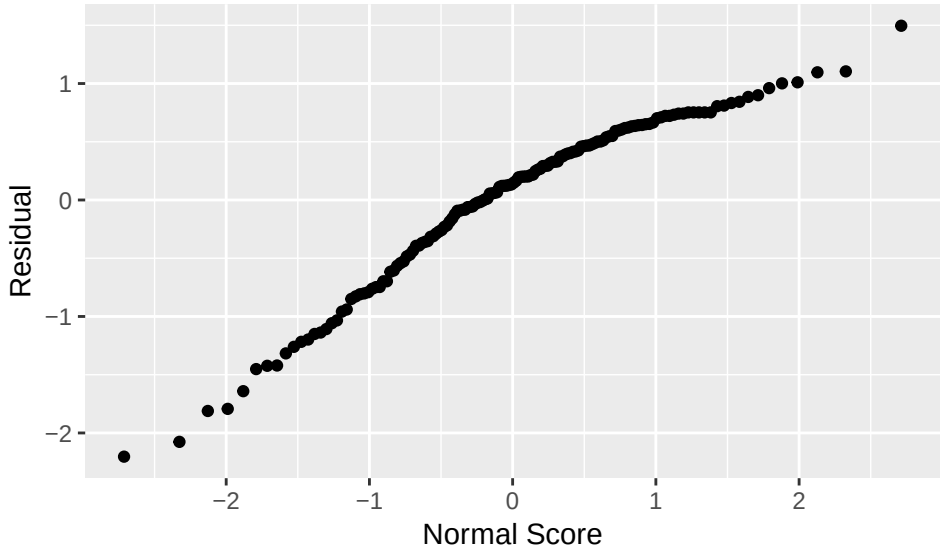
Regression on high school grades

```
# Figure 11.6, page 623
gpamultlm <- lm(GPA ~ HSM + HSE + HSS, data = GPA)
msummary(gpamultlm)
```

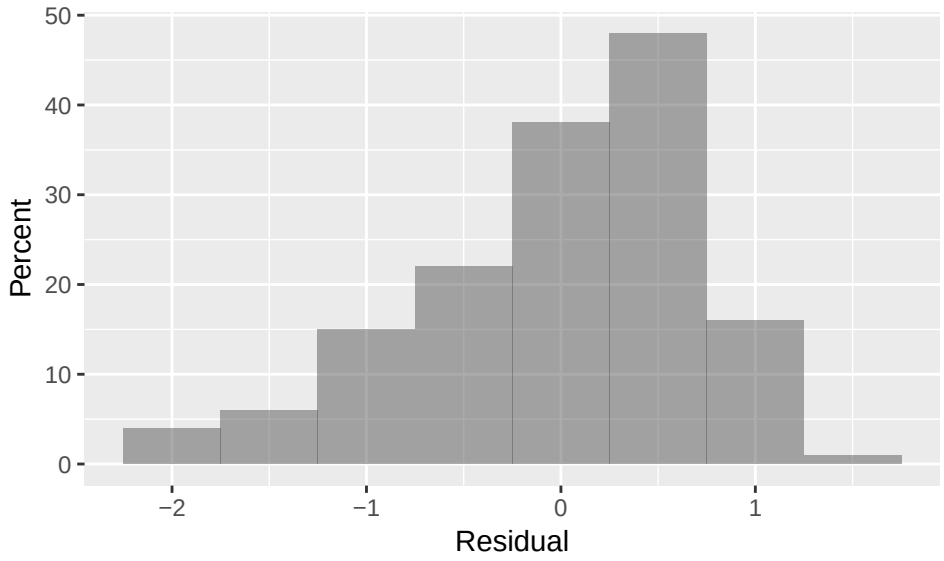
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0693     0.4537    0.15    0.879
## HSM           0.1232     0.0549    2.25    0.026 *
## HSE           0.0585     0.0654    0.89    0.373
## HSS           0.1361     0.0700    1.95    0.054 .
##
## Residual standard error: 0.73 on 146 degrees of freedom
## Multiple R-squared:  0.228,  Adjusted R-squared:  0.212
## F-statistic: 14.4 on 3 and 146 DF,  p-value: 3.03e-08
```

Examining the residuals

```
# Figure 11.7, page 625
gf_qq(~ resid(gpamultlm)) %>%
  gf_labs( x = "Normal Score", y = "Residual")
```




```
gf_histogram(~ resid(gpamultlm), binwidth = .5) %>%
  gf_labs(x = "Residual", y = "Percent")
```



Example 11.14: Residual plots for the GPA analysis

Refining the model

```
# Figure 11.8, page 626
gpamultlm2 <- lm(GPA ~ HSM + HSS, data = GPA)
msummary(gpamultlm2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2570     0.4019    0.64   0.5236
## HSM           0.1250     0.0548    2.28   0.0240 *
## HSS           0.1718     0.0574    2.99   0.0032 **
##
## Residual standard error: 0.73 on 147 degrees of freedom
## Multiple R-squared:  0.224,  Adjusted R-squared:  0.213
## F-statistic: 21.2 on 2 and 147 DF,  p-value: 8.41e-09
```

Regression using all variables

```
# Figure 11.9
gpasatlm <- lm(GPA ~ SATM + SATCR + SATW, data = GPA)
msummary(gpasatlm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.58e-01   5.67e-01    0.81   0.4202
## SATM        3.01e-03   1.07e-03    2.81   0.0056 **
## SATCR       8.03e-04   1.13e-03    0.71   0.4767
## SATW        7.88e-05   1.20e-03    0.07   0.9479
##
## Residual standard error: 0.78 on 146 degrees of freedom
## Multiple R-squared:  0.114,  Adjusted R-squared:  0.0961
## F-statistic: 6.28 on 3 and 146 DF,  p-value: 0.000489
```

```
# Figure 11.10, page 628
gpaalllm <- lm(GPA ~ SATM + SATCR + SATW + HSS + HSE + HSM, data = GPA)
msummary(gpaalllm)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.186783   0.616408   -1.93    0.056 .
## SATM         0.001989   0.001057    1.88    0.062 .
## SATCR        0.000157   0.001049    0.15    0.881
## SATW         0.000474   0.001117    0.42    0.672
## HSS          0.130097   0.068768    1.89    0.061 .
## HSE          0.056791   0.065681    0.86    0.389
## HSM          0.091477   0.057181    1.60    0.112
##
## Residual standard error: 0.71 on 143 degrees of freedom
## Multiple R-squared:  0.273,  Adjusted R-squared:  0.242
## F-statistic: 8.95 on 6 and 143 DF,  p-value: 2.69e-08
```

```
# Figure 11.11, page 631
MASS::stepAIC(gpaalllm)
```

```
## Start:  AIC=-95
## GPA ~ SATM + SATCR + SATW + HSS + HSE + HSM
##
##         Df Sum of Sq  RSS   AIC
## - SATCR  1     0.011 72.5 -97.1
## - SATW   1     0.091 72.6 -96.9
## - HSE    1     0.379 72.8 -96.3
## <none>               72.5 -95.1
## - HSM    1     1.297 73.8 -94.5
## - SATM   1     1.795 74.3 -93.5
## - HSS    1     1.814 74.3 -93.4
##
## Step:  AIC=-97
## GPA ~ SATM + SATW + HSS + HSE + HSM
##
##        Df Sum of Sq  RSS   AIC
## - SATW  1     0.201 72.7 -98.7
## - HSE   1     0.409 72.9 -98.3
## <none>              72.5 -97.1
## - HSM   1     1.288 73.8 -96.5
## - HSS   1     1.813 74.3 -95.4
## - SATM  1     2.075 74.6 -94.9
##
## Step:  AIC=-99
## GPA ~ SATM + HSS + HSE + HSM
##
##        Df Sum of Sq  RSS   AIC
## - HSE   1      0.51 73.2 -99.6
## <none>              72.7 -98.7
## - HSM   1      1.12 73.8 -98.4
## - HSS   1      1.91 74.6 -96.8
## - SATM  1      4.30 77.0 -92.1
##
## Step:  AIC=-100
## GPA ~ SATM + HSS + HSM
##
##        Df Sum of Sq  RSS   AIC
## <none>              73.2 -99.6
## - HSM   1      1.19 74.4 -99.2
```

```
## - SATM  1      4.21 77.4 -93.3
## - HSS   1      4.76 77.9 -92.2
```

```
##
## Call:
## lm(formula = GPA ~ SATM + HSS + HSM, data = GPA)
##
## Coefficients:
## (Intercept)         SATM          HSS          HSM
##    -0.88715      0.00237      0.17252      0.08505
```