Regression and Prediction

Fabian Peri

October 8, 2018

```
# packages needed for chapter 4
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(ascii)
```

```
##
## Attaching package: 'ascii'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(splines)
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.8-23. For overview type 'help("mgcv-package")'.
```

```
# Import the datasets needed for chapter 3
PSDS_PATH <- file.path('C:/Users/fabia/Desktop', 'psds_data')

## Import datasets needed for chapter 4
lung <- read.csv(file.path(PSDS_PATH, 'data', 'LungDisease.csv'))

zhvi <- read.csv(file.path(PSDS_PATH, 'data', 'County_Zhvi_AllHomes.csv'))
zhvi <- unlist(zhvi[13,-(1:5)])
dates <- parse_date_time(paste(substr(names(zhvi), start=2, stop=8), "01", sep="."), "Ymd")
zhvi <- data.frame(ym=dates, zhvi_px=zhvi, row.names = NULL) %>%
  mutate(zhvi_idx=zhvi/last(zhvi))

house <- read.csv(file.path(PSDS_PATH, 'data', 'house_sales.csv'), sep='\t')
# house <- house[house$ZipCode > 0, ]
# write.table(house, file.path(PSDS_PATH, 'data', 'house_sales.csv'), sep='\t')
```

```
## Code for Figure 1
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0401.png'),  width = 4, height=4, units='in', res=300
par(mar=c(4,4,0,0)+.1)
plot(lung$Exposure, lung$PEFR, xlab="Exposure", ylab="PEFR")
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.1
model <- lm(PEFR ~ Exposure, data=lung)
model
```

```
##
## Call:
## lm(formula = PEFR ~ Exposure, data = lung)
##
## Coefficients:
## (Intercept)     Exposure
##     424.583       -4.185
```

```
## Code for figure 2
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0402.png'), width = 350, height = 350)
par(mar=c(4,4,0,0)+.1)

plot(lung$Exposure, lung$PEFR, xlab="Exposure", ylab="PEFR", ylim=c(300,450), type="n", xaxs="i")
abline(a=model$coefficients[1], b=model$coefficients[2], col="blue", lwd=2)
text(x=.3, y=model$coefficients[1], labels=expression("b"[0]),  adj=0, cex=1.5)
x <- c(7.5, 17.5)
y <- predict(model, newdata=data.frame(Exposure=x))
segments(x[1], y[2], x[2], y[2] , col="red", lwd=2, lty=2)
segments(x[1], y[1], x[1], y[2] , col="red", lwd=2, lty=2)
text(x[1], mean(y), labels=expression(Delta~Y), pos=2, cex=1.5)
text(mean(x), y[2], labels=expression(Delta~X), pos=1, cex=1.5)
text(mean(x), 400, labels=expression(b[1] == frac(Delta ~ Y, Delta ~ X)), cex=1.5)
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.2
fitted <- predict(model)
resid <- residuals(model)
```

```
## Code for figure 3
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0403.png'),  width = 4, height=4, units='in', res=300
par(mar=c(4,4,0,0)+.1)

lung1 <- lung %>%
  mutate(Fitted=fitted,
         positive = PEFR>Fitted) %>%
  group_by(Exposure, positive) %>%
  summarize(PEFR_max = max(PEFR),
            PEFR_min = min(PEFR),
            Fitted = first(Fitted)) %>%
  ungroup() %>%
  mutate(PEFR = ifelse(positive, PEFR_max, PEFR_min)) %>%
  arrange(Exposure)

plot(lung$Exposure, lung$PEFR, xlab="Exposure", ylab="PEFR")
abline(a=model$coefficients[1], b=model$coefficients[2], col="blue", lwd=2)
segments(lung1$Exposure, lung1$PEFR, lung1$Exposure, lung1$Fitted, col="red", lty=3)
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.3
head(house[, c("AdjSalePrice", "SqFtTotLiving", "SqFtLot", "Bathrooms",
               "Bedrooms", "BldgGrade")])
```

```
##   AdjSalePrice SqFtTotLiving SqFtLot Bathrooms Bedrooms BldgGrade
## 1       300805          2400    9373      3.00        6         7
## 2      1076162          3764   20156      3.75        4        10
## 3       761805          2060   26036      1.75        4         8
## 4       442065          3200    8618      3.75        5         7
## 5       297065          1720    8620      1.75        4         7
## 6       411781           930    1012      1.50        2         8
```

```
## Code snippet 4.4
house_lm <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
                 Bedrooms + BldgGrade,
               data=house, na.action=na.omit)
```

```
## Code snippet 4.5
house_lm
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
##     Bedrooms + BldgGrade, data = house, na.action = na.omit)
##
## Coefficients:
##   (Intercept)  SqFtTotLiving        SqFtLot      Bathrooms       Bedrooms
##    -5.287e+05      2.127e+02     -1.430e-02     -1.823e+04     -4.657e+04
##     BldgGrade
```

```
##     1.088e+05
```

```
## Code snippet 4.6
summary(house_lm)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
##     Bedrooms + BldgGrade, data = house, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1950841  -114032   -21451    83578  9549956
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.287e+05  1.443e+04 -36.629  < 2e-16 ***
## SqFtTotLiving  2.127e+02  3.401e+00  62.552  < 2e-16 ***
## SqFtLot       -1.430e-02  5.760e-02  -0.248    0.804
## Bathrooms     -1.823e+04  3.225e+03  -5.654 1.58e-08 ***
## Bedrooms      -4.657e+04  2.329e+03 -19.999  < 2e-16 ***
## BldgGrade      1.088e+05  2.164e+03  50.266  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 259400 on 27057 degrees of freedom
## Multiple R-squared:  0.5348, Adjusted R-squared:  0.5348
## F-statistic:  6222 on 5 and 27057 DF,  p-value: < 2.2e-16
```

```
## Code snippet 4.7
house_full <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
                   Bedrooms + BldgGrade + PropertyType + NbrLivingUnits +
                   SqFtFinBasement + YrBuilt + YrRenovated + NewConstruction,
                 data=house, na.action=na.omit)
```

```
## Code snippet 4.8
step_lm <- stepAIC(house_full, direction="both")
```

```
## Start:  AIC=671316
## AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms + Bedrooms +
##     BldgGrade + PropertyType + NbrLivingUnits + SqFtFinBasement +
##     YrBuilt + YrRenovated + NewConstruction
##
##                    Df  Sum of Sq        RSS    AIC
## - NbrLivingUnits    1 3.6803e+09 1.6030e+15 671314
## - YrRenovated       1 1.2789e+10 1.6030e+15 671314
## - SqFtLot           1 2.5471e+10 1.6030e+15 671314
## - NewConstruction   1 7.1632e+10 1.6030e+15 671315
## <none>                           1.6030e+15 671316
## - SqFtFinBasement   1 2.8579e+11 1.6033e+15 671319
## - PropertyType      2 7.8637e+12 1.6108e+15 671444
## - Bathrooms         1 1.0095e+13 1.6131e+15 671484
## - Bedrooms          1 2.9035e+13 1.6320e+15 671800
## - SqFtTotLiving     1 1.4207e+14 1.7450e+15 673612
## - YrBuilt           1 1.4711e+14 1.7501e+15 673690
## - BldgGrade         1 2.3338e+14 1.8364e+15 674993
```

```
##
## Step:  AIC=671314.1
## AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms + Bedrooms +
##     BldgGrade + PropertyType + SqFtFinBasement + YrBuilt + YrRenovated +
##     NewConstruction
##
##                   Df  Sum of Sq        RSS    AIC
## - YrRenovated      1 1.2524e+10 1.6030e+15 671312
## - SqFtLot          1 2.5211e+10 1.6030e+15 671313
## - NewConstruction  1 7.2192e+10 1.6031e+15 671313
## <none>                          1.6030e+15 671314
## + NbrLivingUnits   1 3.6803e+09 1.6030e+15 671316
## - SqFtFinBasement  1 2.8911e+11 1.6033e+15 671317
## - PropertyType     2 7.8769e+12 1.6109e+15 671443
## - Bathrooms        1 1.0152e+13 1.6131e+15 671483
## - Bedrooms         1 2.9229e+13 1.6322e+15 671801
## - SqFtTotLiving    1 1.4222e+14 1.7452e+15 673613
## - YrBuilt          1 1.4802e+14 1.7510e+15 673702
## - BldgGrade        1 2.3544e+14 1.8384e+15 675021
##
## Step:  AIC=671312.3
## AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms + Bedrooms +
##     BldgGrade + PropertyType + SqFtFinBasement + YrBuilt + NewConstruction
##
##                   Df  Sum of Sq        RSS    AIC
## - SqFtLot          1 2.5083e+10 1.6030e+15 671311
## - NewConstruction  1 7.1293e+10 1.6031e+15 671311
## <none>                          1.6030e+15 671312
## + YrRenovated      1 1.2524e+10 1.6030e+15 671314
## + NbrLivingUnits   1 3.4152e+09 1.6030e+15 671314
## - SqFtFinBasement  1 2.9330e+11 1.6033e+15 671315
## - PropertyType     2 7.8650e+12 1.6109e+15 671441
## - Bathrooms        1 1.0238e+13 1.6132e+15 671483
## - Bedrooms         1 2.9219e+13 1.6322e+15 671799
## - SqFtTotLiving    1 1.4221e+14 1.7452e+15 673611
## - YrBuilt          1 1.6196e+14 1.7650e+15 673915
## - BldgGrade        1 2.3548e+14 1.8385e+15 675020
##
## Step:  AIC=671310.7
## AdjSalePrice ~ SqFtTotLiving + Bathrooms + Bedrooms + BldgGrade +
##     PropertyType + SqFtFinBasement + YrBuilt + NewConstruction
##
##                   Df  Sum of Sq        RSS    AIC
## - NewConstruction  1 6.3500e+10 1.6031e+15 671310
## <none>                          1.6030e+15 671311
## + SqFtLot          1 2.5083e+10 1.6030e+15 671312
## + YrRenovated      1 1.2396e+10 1.6030e+15 671313
## + NbrLivingUnits   1 3.1669e+09 1.6030e+15 671313
## - SqFtFinBasement  1 2.8652e+11 1.6033e+15 671314
## - PropertyType     2 7.8468e+12 1.6109e+15 671439
## - Bathrooms        1 1.0215e+13 1.6132e+15 671481
## - Bedrooms         1 2.9451e+13 1.6325e+15 671801
## - SqFtTotLiving    1 1.4593e+14 1.7490e+15 673667
## - YrBuilt          1 1.6199e+14 1.7650e+15 673914
```

```
## - BldgGrade        1 2.3547e+14 1.8385e+15 675018
##
## Step:  AIC=671309.8
## AdjSalePrice ~ SqFtTotLiving + Bathrooms + Bedrooms + BldgGrade +
##     PropertyType + SqFtFinBasement + YrBuilt
##
##                   Df  Sum of Sq        RSS    AIC
## <none>                          1.6031e+15 671310
## + NewConstruction  1 6.3500e+10 1.6030e+15 671311
## + SqFtLot          1 1.7290e+10 1.6031e+15 671311
## + YrRenovated      1 1.1567e+10 1.6031e+15 671312
## + NbrLivingUnits   1 3.7093e+09 1.6031e+15 671312
## - SqFtFinBasement  1 2.6805e+11 1.6033e+15 671312
## - PropertyType     2 8.5458e+12 1.6116e+15 671450
## - Bathrooms        1 1.0235e+13 1.6133e+15 671480
## - Bedrooms         1 2.9483e+13 1.6326e+15 671801
## - SqFtTotLiving    1 1.4722e+14 1.7503e+15 673686
## - YrBuilt          1 1.7535e+14 1.7784e+15 674117
## - BldgGrade        1 2.3572e+14 1.8388e+15 675020
```

```
step_lm
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving + Bathrooms + Bedrooms +
##     BldgGrade + PropertyType + SqFtFinBasement + YrBuilt, data = house,
##     na.action = na.omit)
##
## Coefficients:
##               (Intercept)              SqFtTotLiving
##                 6227632.22                     186.50
##                  Bathrooms                  Bedrooms
##                    44721.72                  -49807.18
##                  BldgGrade  PropertyTypeSingle Family
##                  139179.23                   23328.69
##     PropertyTypeTownhouse            SqFtFinBasement
##                   92216.25                       9.04
##                   YrBuilt
##                   -3592.47
```

```
lm(AdjSalePrice ~  Bedrooms, data=house)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ Bedrooms, data = house)
##
## Coefficients:
## (Intercept)     Bedrooms
##      109860       136402
```

```
# WeightedRegression
## Code snippet 4.9
house$Year = year(house$DocumentDate)
house$Weight = house$Year - 2005
```

```
## Code snippet 4.10
house_wt <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
                Bedrooms + BldgGrade,
               data=house, weight=Weight, na.action=na.omit)
round(cbind(house_lm=house_lm$coefficients,
            house_wt=house_wt$coefficients), digits=3)
```

```
##                  house_lm    house_wt
## (Intercept)   -528724.348 -580378.015
## SqFtTotLiving     212.708     229.945
## SqFtLot            -0.014      -0.181
## Bathrooms      -18233.212  -23335.486
## Bedrooms       -46574.193  -54234.376
## BldgGrade      108780.111  116037.063
```

```
# Factor Variables
## Code snippet 4.11
head(house[, 'PropertyType'])
```

```
## [1] Multiplex     Single Family Single Family Single Family Single Family
## [6] Townhouse
## Levels: Multiplex Single Family Townhouse
```

```
## Code snippet 4.12
prop_type_dummies <- model.matrix(~PropertyType -1, data=house)
head(prop_type_dummies)
```

```
##   PropertyTypeMultiplex PropertyTypeSingle Family PropertyTypeTownhouse
## 1                     1                         0                     0
## 2                     0                         1                     0
## 3                     0                         1                     0
## 4                     0                         1                     0
## 5                     0                         1                     0
## 6                     0                         0                     1
```

```
## Code snippet 4.13
lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
     Bedrooms +  BldgGrade + PropertyType, data=house)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
##     Bedrooms + BldgGrade + PropertyType, data = house)
##
## Coefficients:
##               (Intercept)             SqFtTotLiving
##                 -4.409e+05                 2.072e+02
##                    SqFtLot                 Bathrooms
##                 -2.314e-02                -1.500e+04
##                   Bedrooms                 BldgGrade
##                 -4.957e+04                 1.122e+05
## PropertyTypeSingle Family      PropertyTypeTownhouse
##                 -9.819e+04                -1.189e+05
```

```
## Code snippet 4.14
table(house$ZipCode)
```

```
##
##    -1  9800 89118 98001 98002 98003 98004 98005 98006 98007 98008 98010
##  4374     1     1   358   180   241   293   133   460   112   291    56
## 98011 98014 98019 98022 98023 98024 98027 98028 98029 98030 98031 98032
##   163    85   242   188   455    31   366   252   475   263   308   121
## 98033 98034 98038 98039 98040 98042 98043 98045 98047 98050 98051 98052
##   517   575   788    47   244   641     1   222    48     7    32   614
## 98053 98055 98056 98057 98058 98059 98065 98068 98070 98072 98074 98075
##   499   332   402     4   420   513   430     1    89   245   502   388
## 98077 98092 98102 98103 98105 98106 98107 98108 98109 98112 98113 98115
##   204   289   106   671   313   361   296   155   149   357     1   620
## 98116 98117 98118 98119 98122 98125 98126 98133 98136 98144 98146 98148
##   364   619   492   260   380   409   473   465   310   332   287    40
## 98155 98166 98168 98177 98178 98188 98198 98199 98224 98288 98354
##   358   193   332   216   266   101   225   393     3     4     9
```

```
## Code snippet 4.15
zip_groups <- house %>%
  mutate(resid = residuals(house_lm)) %>%
  group_by(ZipCode) %>%
  summarize(med_resid = median(resid),
            cnt = n()) %>%
  # sort the zip codes by the median residual
  arrange(med_resid) %>%
  mutate(cum_cnt = cumsum(cnt),
         ZipGroup = factor(ntile(cum_cnt, 5)))
house <- house %>%
  left_join(select(zip_groups, ZipCode, ZipGroup), by='ZipCode')


# correlated variables
# Code snippet 4.15
step_lm$coefficients
```

```
##               (Intercept)             SqFtTotLiving
##              6.227632e+06              1.865012e+02
##                 Bathrooms                  Bedrooms
##              4.472172e+04             -4.980718e+04
##                 BldgGrade PropertyTypeSingle Family
##              1.391792e+05              2.332869e+04
##     PropertyTypeTownhouse           SqFtFinBasement
##              9.221625e+04              9.039911e+00
##                   YrBuilt
##             -3.592468e+03
```

```
# Code snippet 4.16
update(step_lm, . ~ . -SqFtTotLiving - SqFtFinBasement - Bathrooms)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ Bedrooms + BldgGrade + PropertyType +
##     YrBuilt, data = house, na.action = na.omit)
##
## Coefficients:
##               (Intercept)                   Bedrooms
##                   4834680                      27657
```

```
##                 BldgGrade  PropertyTypeSingle Family
##                    245709                     -17604
##     PropertyTypeTownhouse                    YrBuilt
##                    -47477                      -3161
```

```
#  ConfoundingVariables
## Code snippet 4.17
lm(AdjSalePrice ~  SqFtTotLiving + SqFtLot +
     Bathrooms + Bedrooms +
     BldgGrade + PropertyType + ZipGroup,
   data=house, na.action=na.omit)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
##     Bedrooms + BldgGrade + PropertyType + ZipGroup, data = house,
##     na.action = na.omit)
##
## Coefficients:
##               (Intercept)              SqFtTotLiving
##                -6.737e+05                  1.938e+02
##                   SqFtLot                  Bathrooms
##                 3.526e-01                  8.121e+03
##                  Bedrooms                  BldgGrade
##                -3.953e+04                  1.048e+05
## PropertyTypeSingle Family      PropertyTypeTownhouse
##                 2.827e+03                 -2.126e+04
##                 ZipGroup2                  ZipGroup3
##                 5.913e+04                  1.029e+05
##                 ZipGroup4                  ZipGroup5
##                 1.727e+05                  3.340e+05
```

```
#  Interactions
## Code snippet 4.18
lm(AdjSalePrice ~  SqFtTotLiving*ZipGroup + SqFtLot +
     Bathrooms + Bedrooms +
     BldgGrade + PropertyType,
   data=house, na.action=na.omit)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ SqFtTotLiving * ZipGroup + SqFtLot +
##     Bathrooms + Bedrooms + BldgGrade + PropertyType, data = house,
##     na.action = na.omit)
##
## Coefficients:
##               (Intercept)              SqFtTotLiving
##                -4.990e+05                  1.030e+02
##                 ZipGroup2                  ZipGroup3
##                -3.794e+04                  5.436e+04
##                 ZipGroup4                  ZipGroup5
##                -2.666e+03                 -1.722e+05
##                   SqFtLot                  Bathrooms
##                 5.535e-01                  4.197e+02
##                  Bedrooms                  BldgGrade
##                -3.919e+04                  1.086e+05
```

```
## PropertyTypeSingle Family      PropertyTypeTownhouse
##                  4.530e+03                 -2.333e+04
##   SqFtTotLiving:ZipGroup2    SqFtTotLiving:ZipGroup3
##                  4.999e+01                  1.599e+01
##   SqFtTotLiving:ZipGroup4    SqFtTotLiving:ZipGroup5
##                  8.230e+01                  2.356e+02
```

```
head(model.matrix(~C(PropertyType, sum) , data=house))
```

```
##   (Intercept) C(PropertyType, sum)1 C(PropertyType, sum)2
## 1           1                     1                     0
## 2           1                     0                     1
## 3           1                     0                     1
## 4           1                     0                     1
## 5           1                     0                     1
## 6           1                    -1                    -1
```

```
# outlier anaysis
## Code snippet 4.19
house_98105 <- house[house$ZipCode == 98105,]
lm_98105 <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
                Bedrooms + BldgGrade, data=house_98105)
```

```
## Code snippet 4.20
sresid <- rstandard(lm_98105)
idx <- order(sresid, decreasing=FALSE)
sresid[idx[1]]
```

```
##     24333
## -4.326732
```

```
resid(lm_98105)[idx[1]]
```

```
##     24333
## -757753.6
```

```
## Code snippet 4.21
house_98105[idx[1], c('AdjSalePrice', 'SqFtTotLiving', 'SqFtLot',
                      'Bathrooms', 'Bedrooms', 'BldgGrade')]
```

```
##       AdjSalePrice SqFtTotLiving SqFtLot Bathrooms Bedrooms BldgGrade
## 24333       119748          2900    7276         3        6         7
```

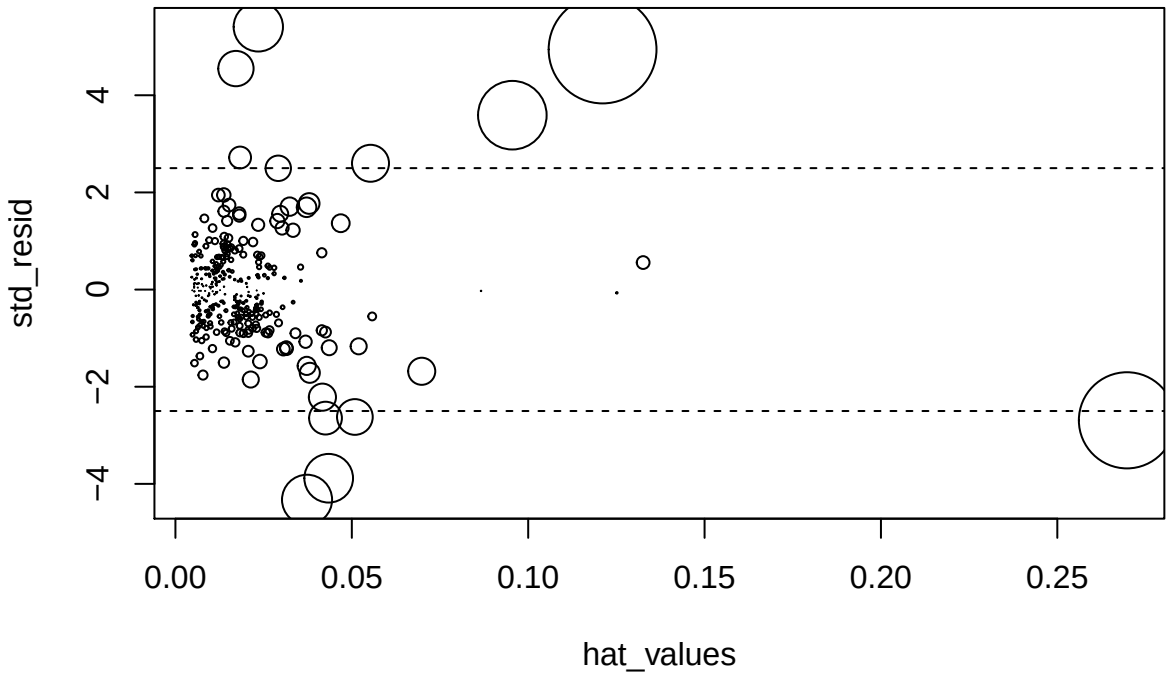
```
# Figure 4-5: Influential data point in regression
seed <- 11
set.seed(seed)
x <- rnorm(25)
y <- -x/5 + rnorm(25)
x[1] <- 8
y[1] <- 8

png(filename=file.path(PSDS_PATH, 'figures', 'psds_0405.png'),  width = 4, height=4, units='in', res=300
par(mar=c(3,3,0,0)+.1)
plot(x, y, xlab='', ylab='', pch=16)
model <- lm(y~x)
abline(a=model$coefficients[1], b=model$coefficients[2], col="blue", lwd=3)
model <- lm(y[-1]~x[-1])
abline(a=model$coefficients[1], b=model$coefficients[2], col="red", lwd=3, lty=2)
```

```
dev.off()
```

```
## pdf
##   2
```

```
# influential observations
## Code snippet 4.22
std_resid <- rstandard(lm_98105)
cooks_D <- cooks.distance(lm_98105)
hat_values <- hatvalues(lm_98105)
plot(hat_values, std_resid, cex=10*sqrt(cooks_D))
abline(h=c(-2.5, 2.5), lty=2)
```




```
## Code for Figure 4-6
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0406.png'), width = 4, height=4, units='in', res=300)
par(mar=c(4,4,0,0)+.1)
plot(hat_values, std_resid, cex=10*sqrt(cooks_D))
abline(h=c(-2.5, 2.5), lty=2)
dev.off()
```

```
## pdf
##   2
```

```
## Table 4-2

lm_98105_inf <- lm(AdjSalePrice ~ SqFtTotLiving + SqFtLot +
                     Bathrooms +  Bedrooms + BldgGrade,
                   subset=cooks_D<.08, data=house_98105)
```

```
df <- data.frame(lm_98105$coefficients,
                 lm_98105_inf$coefficients)
names(df) <- c('Original', 'Influential Removed')
ascii((df),
      include.rownames=TRUE, include.colnames=TRUE, header=TRUE,
      digits=rep(0, 3), align=c("l", "r", "r") ,
      caption="Comparison of regression coefficients with the full data and with influential data remove
```

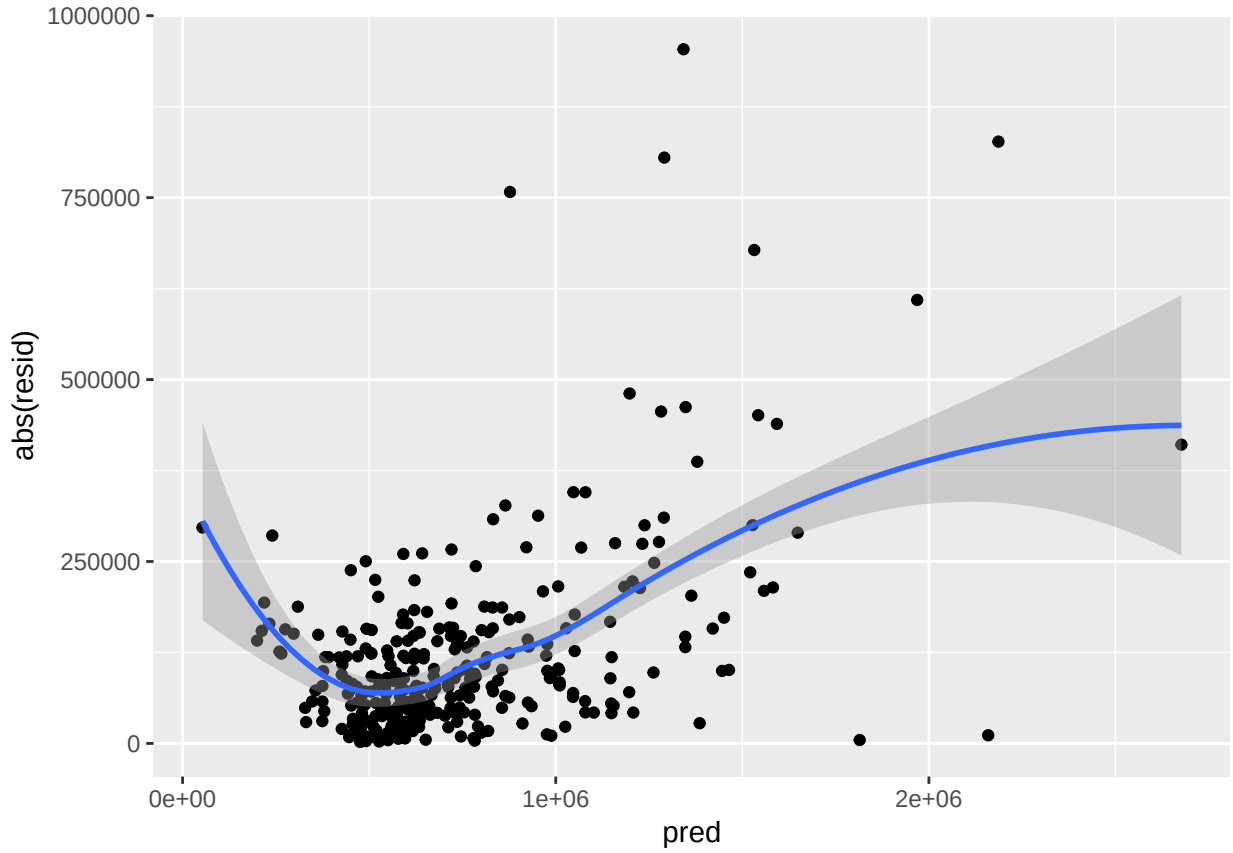
```
## Warning in rep(rownames, length = nrow(x)): 'x' is NULL so the result will
## be NULL
```

```
## Warning in rep(colnames, length = ncol(x)): 'x' is NULL so the result will
## be NULL
```

```
## .Comparison of regression coefficients with the full data and with influential data removed
## |=====================================================
## 1.1+| >h| Original >h| Influential Removed
## <| (Intercept)   >| -772550  >| -647137
## <| SqFtTotLiving >| 210      >| 230
## <| SqFtLot       >| 39       >| 33
## <| Bathrooms     >| 2282     >| -16132
## <| Bedrooms      >| -26320   >| -22888
## <| BldgGrade     >| 130000   >| 114871
## |=====================================================
```

```
## heteroskedasticity
## Code snippet 4.23
df <- data.frame(
  resid = residuals(lm_98105),
  pred = predict(lm_98105))
ggplot(df, aes(pred, abs(resid))) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## Code for figure 4-7
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0407.png'), width = 4, height=4, units='in', res=300)

ggplot(df, aes(pred, abs(resid))) +
  geom_point() +
  geom_smooth() +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
dev.off()
```

```
## pdf
##   2
```

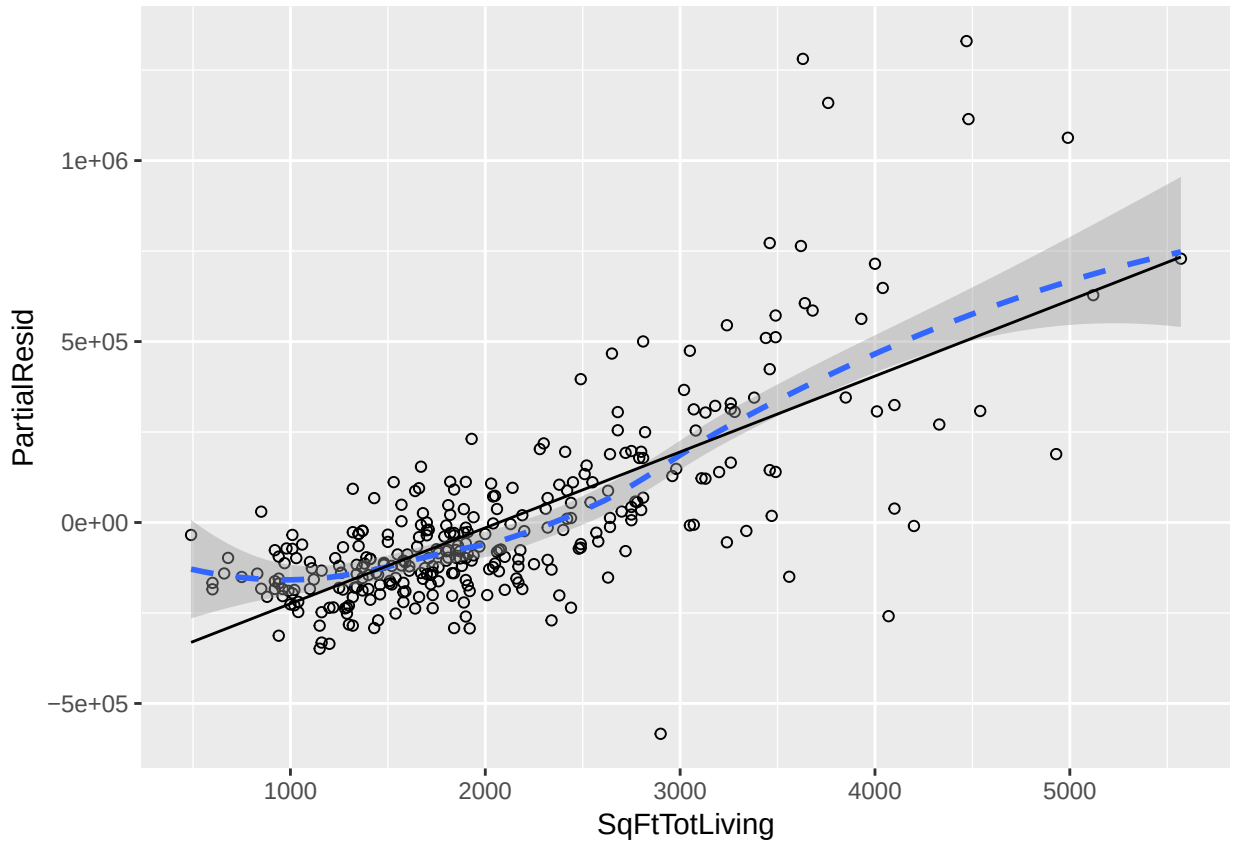
```
## Code for figure 4-8
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0408.png'), width = 4, height=4, units='in', res=300)
par(mar=c(4,4,0,0)+.1)
hist(std_resid, main='')
dev.off()
```

```
## pdf
##   2
```

```
## partial residuals plot
## Code snippet 4.24
terms <- predict(lm_98105, type='terms')
partial_resid <- resid(lm_98105) + terms
```

```
## Code snippet 4.25
df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                 Terms = terms[, 'SqFtTotLiving'],
                 PartialResid = partial_resid[, 'SqFtTotLiving'])
ggplot(df, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  geom_line(aes(SqFtTotLiving, Terms))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## Code for figure 4-9
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0409.png'),  width = 4, height=4, units='in', res=300

df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                 Terms = terms[, 'SqFtTotLiving'],
                 PartialResid = partial_resid[, 'SqFtTotLiving'])
ggplot(df, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  theme_bw() +
  geom_line(aes(SqFtTotLiving, Terms))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.26

lm(AdjSalePrice ~ poly(SqFtTotLiving, 2) + SqFtLot +
   BldgGrade + Bathrooms +  Bedrooms,
   data=house_98105)
```

```
##
## Call:
## lm(formula = AdjSalePrice ~ poly(SqFtTotLiving, 2) + SqFtLot +
##     BldgGrade + Bathrooms + Bedrooms, data = house_98105)
##
## Coefficients:
##              (Intercept)  poly(SqFtTotLiving, 2)1  poly(SqFtTotLiving, 2)2
##                -402530.47               3271519.49                776934.02
##                   SqFtLot                BldgGrade                Bathrooms
##                     32.56                135717.06                 -1435.12
##                  Bedrooms
##                  -9191.94
```

```
lm_poly <- lm(AdjSalePrice ~  poly(SqFtTotLiving, 2) + SqFtLot +
                BldgGrade +  Bathrooms +  Bedrooms,
              data=house_98105)
terms <- predict(lm_poly, type='terms')
partial_resid <- resid(lm_poly) + terms

## Code for Figure 4-10
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0410.png'), width = 4, height=4, units='in', res=300)

df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                 Terms = terms[, 1],
                 PartialResid = partial_resid[, 1])
ggplot(df, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  geom_line(aes(SqFtTotLiving, Terms))+
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.27
knots <- quantile(house_98105$SqFtTotLiving, p=c(.25, .5, .75))
lm_spline <- lm(AdjSalePrice ~ bs(SqFtTotLiving, knots=knots, degree=3) +  SqFtLot +
                  Bathrooms + Bedrooms + BldgGrade,  data=house_98105)


terms1 <- predict(lm_spline, type='terms')
partial_resid1 <- resid(lm_spline) + terms

## Code for Figure 4-12
```

```
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0412.png'), width = 4, height=4, units='in', res=300)

df1 <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                  Terms = terms1[, 1],
                  PartialResid = partial_resid1[, 1])
ggplot(df1, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  geom_line(aes(SqFtTotLiving, Terms))+
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
dev.off()
```

```
## pdf
##   2
```

```
## Code snippet 4.27
lm_gam <- gam(AdjSalePrice ~ s(SqFtTotLiving) + SqFtLot +
                Bathrooms +  Bedrooms + BldgGrade,
              data=house_98105)
terms <- predict.gam(lm_gam, type='terms')
partial_resid <- resid(lm_gam) + terms

## Code for Figure 4-13
png(filename=file.path(PSDS_PATH, 'figures', 'psds_0413.png'), width = 4, height=4, units='in', res=300)
df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                  Terms = terms[, 5],
                  PartialResid = partial_resid[, 5])
ggplot(df, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  geom_line(aes(SqFtTotLiving, Terms))  +
  theme_bw()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
dev.off()
```

```
## pdf
##   2
```