Predicting Insurance Costs

JR

August 14, 2018

Start by loading our libraries and reading in our data, then showing the first few rows with the head function.

```
library(rmarkdown)
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
--- tidyverse 1.2.1 --

## v ggplot2 2.2.1     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.5
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(caTools)
insurance <- read.csv('C:\\Users\\Jason\\Desktop\\R
Directory\\insurance.csv')

head(insurance)

##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Next, I'm checking the structure of the data and seeing if there are any missing values that we need to clean or deal with.

```
str(insurance)

## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
```

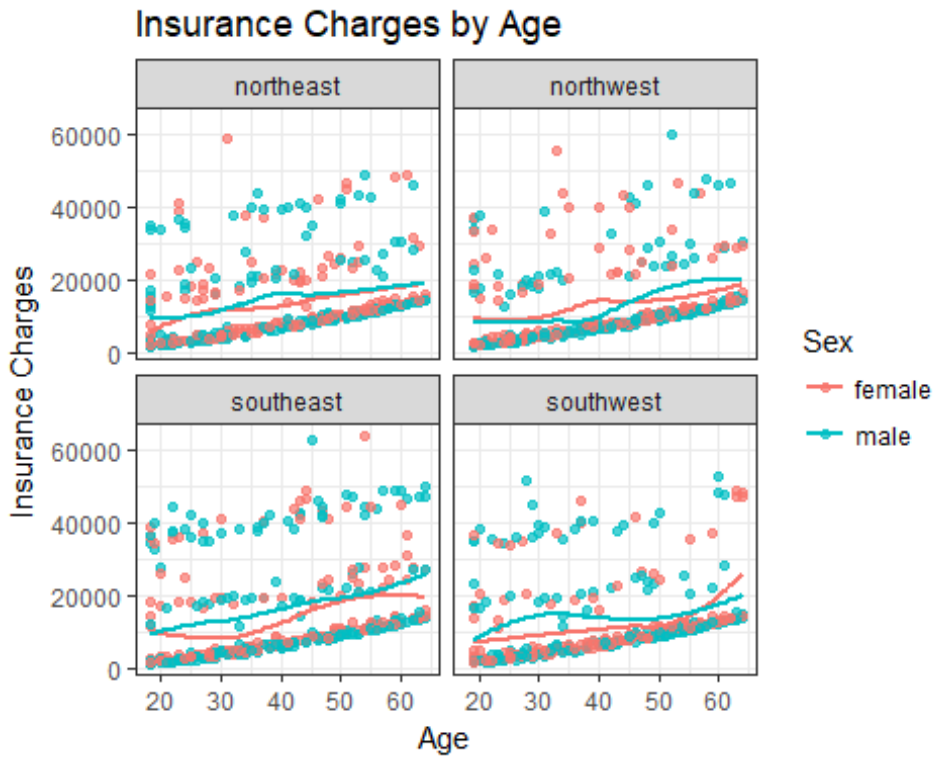
```
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3
2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...

any(is.na(insurance))

## [1] FALSE
```

Our variables are in correct formats and there are no missing values.

Let's do some exploratory data analysis!

First, I'm checking to see how insurance prices change with age and sex and also if there are any major differences in region

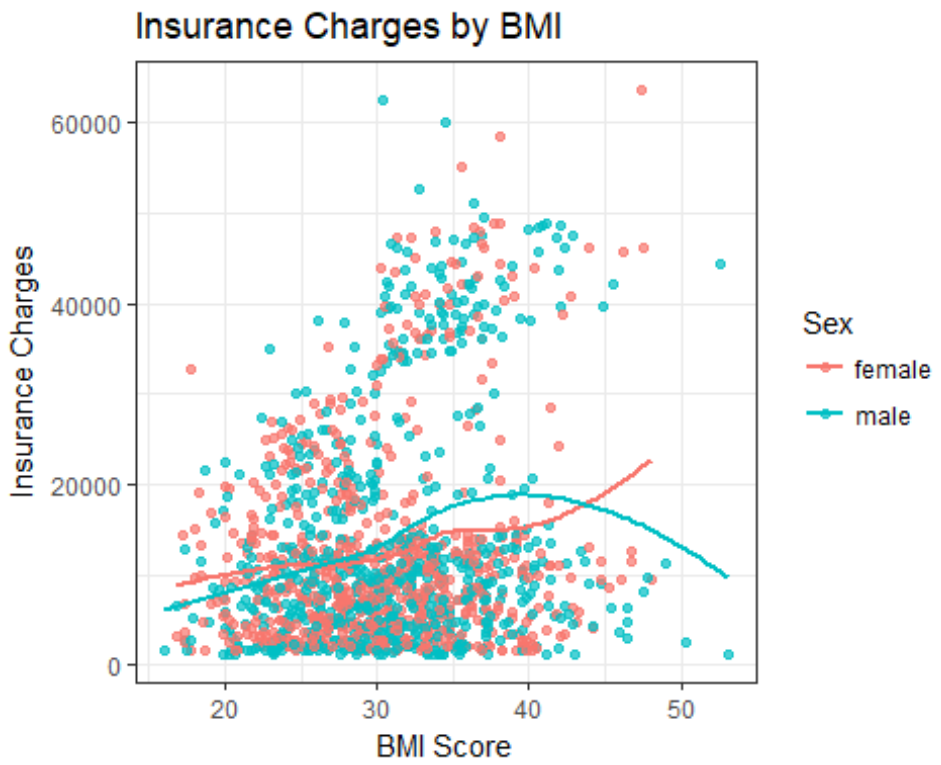
```
ggplot(insurance, aes(x = age, y = charges, color = sex)) +
  geom_point(alpha = 0.7) +
  theme_bw() +
  facet_wrap( ~ region) + labs(x = "Age",
                               y = 'Insurance Charges',
                               title = 'Insurance Charges by Age',
                               color = 'Sex')+
  geom_smooth(se = FALSE)
```



It seems like all regions follow approximately the same pattern and insurance costs increase as age increases, which makes sense. From first glance it doesn't look like there are any obvious differences in insurance costs for males or females of any age.

Next, I looked at whether BMI (Body Mass Index) was a factor in higher insurance prices. I expect that there should be some towards higher prices for higher BMI because BMIs over about 30 is a sign of being overweight or even a sign of obesity.

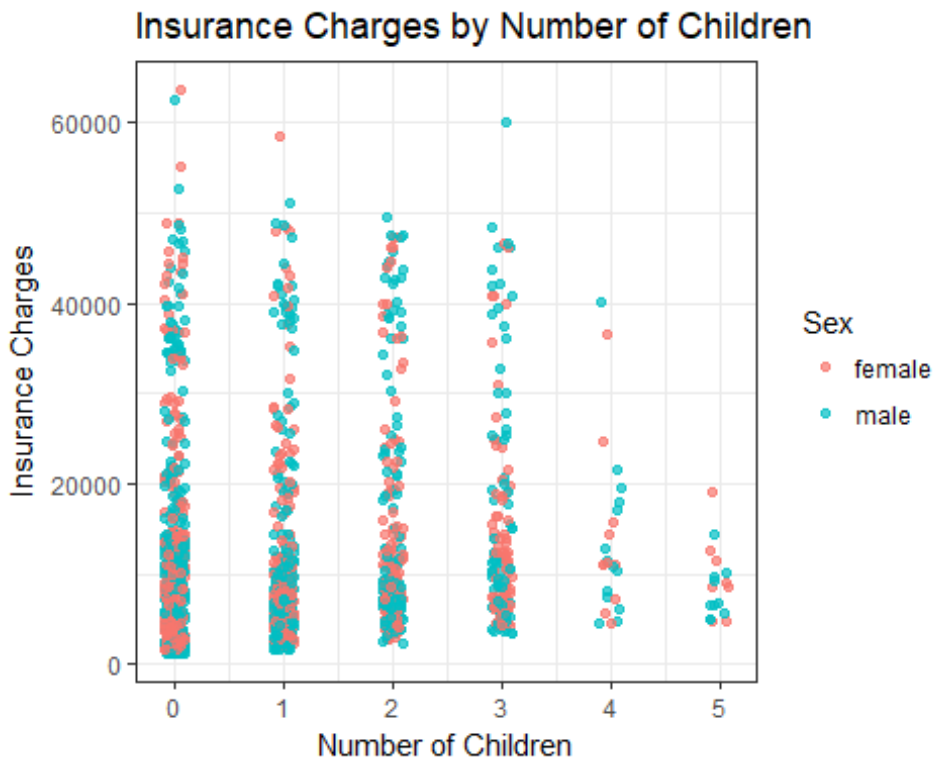
```
ggplot(insurance, aes(x = bmi, y = charges, color = sex)) +
  geom_point(alpha = 0.7) +
  theme_bw() + labs(x = "BMI Score",
                    y = 'Insurance Charges',
                    title = 'Insurance Charges by BMI',
                    color = 'Sex')+
                  geom_smooth(se = FALSE)
```



The results are somewhat surprising because I thought there would be a much clearer pattern showing higher insurance costs per higher BMI but there is only some positive trend. One possibility is that BMI isn't necessarily indicative of overall health or it may not matter to health insurer's actuarial tables.

Next, I decided to see if people with kids were charged more for insurance. I expect that there should be some increase to overall cost but adding to insurance is a marginal cost compared to initial one-person coverage

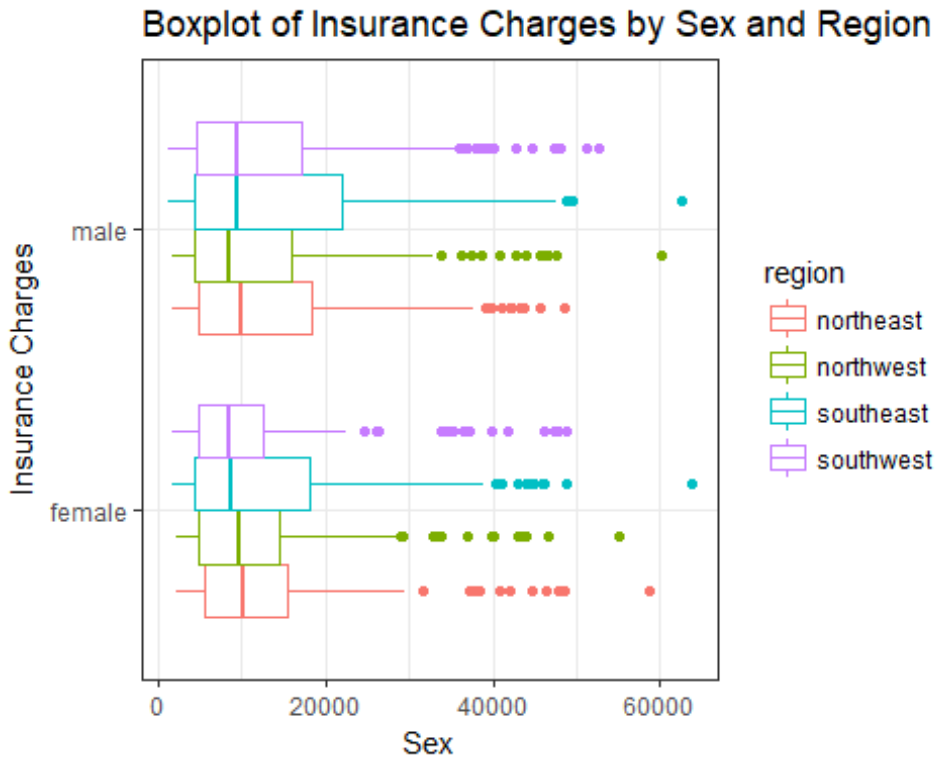
```
ggplot(insurance, aes(x = children, y = charges, color = sex))+
  geom_point(alpha = 0.7, position = position_jitter(width =0.1)) +
  theme_bw() + labs(x = "Number of Children",
                    y = 'Insurance Charges',
                    title = 'Insurance Charges by Number of Children',
                    color = 'Sex')
```



As expected, the more kids, the higher insurance 'base' but there aren't other patterns that show very much except that most people dont have more than 3 kids.

Finally, I looked at if men or women generally pay more for insurance.

```
ggplot(insurance, aes(x = sex, y = charges, color = region)) +
  geom_boxplot() +
  theme_bw() +
  coord_flip() +
  labs(x = 'Insurance Charges',
       y = 'Sex',
       title = 'Boxplot of Insurance Charges by Sex and Region')
```



The most noteable takeaway here is that men and women pay approximately the same median insurance costs. However, men have much larger ranges between the first and third quartiles than women in all regions, showing more variation of insurance costs.

Now that I know more about how the data looks I can start modelling. I am going to use a multivariable regression model to use all of our variables (age, sex, BMI, children, smoker, and region) to try to predict the insurance cost.

First I'll split our data into train/test sets using the caTools package

```
split <- sample.split(insurance$charges, SplitRatio = 0.75)
train <- subset(insurance, split == TRUE)
test <- subset(insurance, split == FALSE)
```

Now I can use the training set to built a model. I'll call summary on the model so we can see how the model works

```
model <- lm(charges ~ ., data = train)
summary(model)

##
## Call:
## lm(formula = charges ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11418  -2786  -1055   1314  30031
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11084.48    1149.21  -9.645  < 2e-16 ***
## age                251.00      13.63  18.421  < 2e-16 ***
## sexmale            -42.03     384.64  -0.109  0.91300
## bmi                319.42      33.29   9.594  < 2e-16 ***
## children           477.69     159.76   2.990  0.00286 **
## smokeryes        23733.80     480.15  49.429  < 2e-16 ***
## regionnorthwest   -549.66     551.60  -0.996  0.31926
## regionsoutheast  -1037.48     558.23  -1.859  0.06339 .
## regionsouthwest  -1071.11     549.29  -1.950  0.05146 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6052 on 994 degrees of freedom
## Multiple R-squared:  0.7451, Adjusted R-squared:  0.743
## F-statistic: 363.1 on 8 and 994 DF,  p-value: < 2.2e-16
```

From this model we can tell that age, BMI, number of children, and if the person smokes are statistically significant to our model. This is from the last where our P-value are less than 0.05. We can also see that our model explains about 75% of the insurance costs from the adjusted R-squared figure, also known as the 'goodness-of-fit'.

Now we can use our model to predict using the 'test' subset.

```
predicted_values <- predict(model, newdata = test)
```

Using correlation, let's see how the prediction did

```
actuals_preds <- data.frame(cbind(actuals = test$charges, predicteds =
predicted_values))

cor(actuals_preds)
```

```
##             actuals predicteds
## actuals    1.000000   0.875473
## predicteds 0.875473   1.000000
```

Another metric we can use to test our model (which was already calculated in the regression summary) is the sum of squares of residuals, which shows our R-squared.

```
SSE <- sum((test$charges - predicted_values) ^2)
SST <- sum((test$charges - mean(test$charges)) ^2)
1 - SSE/SST
```

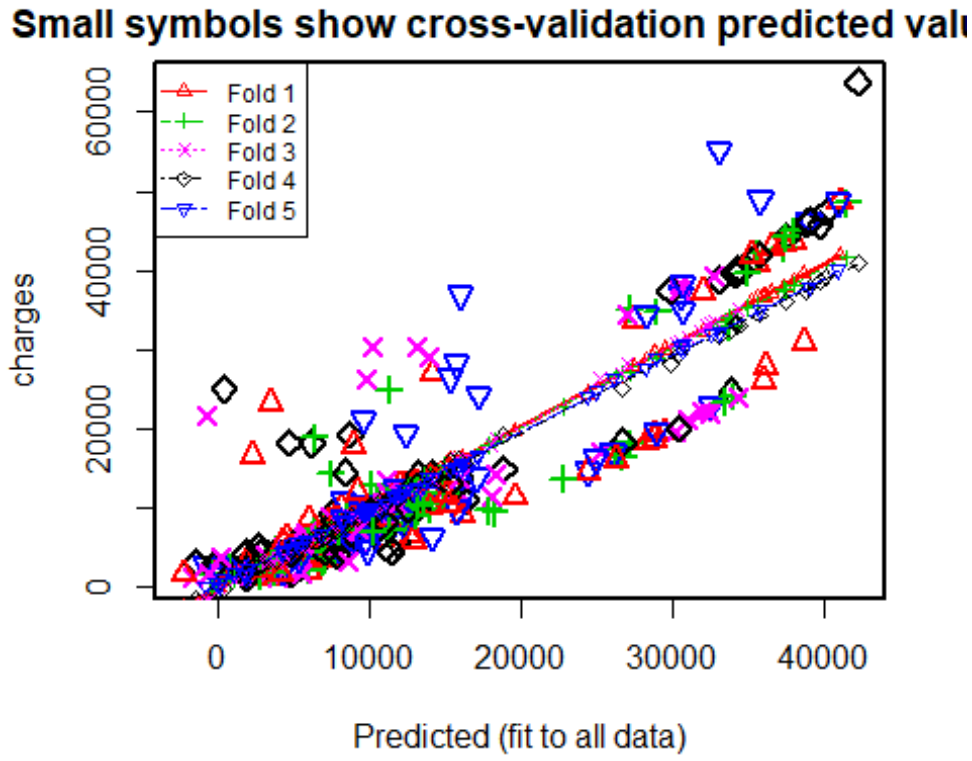
```
## [1] 0.7656397
```

The model predicted 76.4% of the insurance cost.

Finally, I'll use the K-Fold Cross Validation with 5 folds as one last way to measure the model. This function comes from the DAAG library.

```r
library(DAAG)
cv.lm(data = test, model, m = 5)
```

```
## Analysis of Variance Table
##
## Response: charges
##            Df   Sum Sq  Mean Sq F value Pr(>F)
## age         1 5.73e+09 5.73e+09  151.76 <2e-16 ***
## sex         1 1.24e+07 1.24e+07    0.33  0.567
## bmi         1 2.53e+09 2.53e+09   67.08  6e-15 ***
## children    1 1.21e+08 1.21e+08    3.22  0.074 .
## smoker      1 3.24e+10 3.24e+10  859.54 <2e-16 ***
## region      3 8.80e+07 2.93e+07    0.78  0.507
## Residuals 326 1.23e+10 3.77e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##
## fold 1
## Observations in test set: 67
##                12      38   43    45    68    76    87    90    91   104
## Predicted   36100   -1.59 4501 10788  7379 14657 38009 11748  5978 38622
## cvpred      36757   96.73 4469 10674  7416 14825 38700 11758  5731 39574
## charges     27809 2302.30 6272  6080  6389 11357 43579 11083  2027 30942
## CV residual -8948 2205.57 1804 -4595 -1027 -3469  4878  -675 -3704 -8632
##               117  119   131   137   147   163    225   260  283  340
## Predicted   19536 8812 12370  3338 35649 15612  29510 27415 4859 8657
```

```
## cvpred      19796 8727 12193  3479 36588 16218  30048 27838 4593 8639
## charges     11381 8601 12815  1261 40721 10451  19516 33750 4237 8233
## CV residual -8414 -125   622 -2218  4133 -5768 -10533  5912 -356 -406
##                343   375   380   398   404    418    445   469   498   531
## Predicted   13059  2837 14120  2197 13136  28853  35992  3417  9269 41023
## cvpred      12901  2611 14365  1962 13489  29790  36811  3023  9793 41937
## charges     13217  1392 27001 16586 10269  18608  26109 23289  8028 48676
## CV residual   317 -1220 12636 14624 -3220 -11182 -10702 20265 -1765  6739
##                555   564   581   597   616   631   641  648   700   709
## Predicted    8873 16086 11983  8705 36516 13950 12749 7253  7349  7596
## cvpred       8499 16327 11971  8756 37209 14518 13547 7147  7301  7464
## charges     17879  9059 12914  7640 42970 10086  6666 8252  3501  6113
## CV residual  9380 -7268   943 -1115  5761 -4433 -6880 1105 -3800 -1351
##                731   732   733   735   743   761    796   851   928   946
## Predicted    28854  8128  4917 15352 37355  6281  28264 32056  9141 15084
## cvpred       29599  8576  5265 16044 37895  6007  28735 32337  9534 15596
## charges      19362 10065  4235 14007 43254  3926  18311 37270 12032 11674
## CV residual -10237  1490 -1030 -2037  5359 -2081 -10424  4933  2498 -3922
##                968   970  978     985   990   1039  1068 1110  1112  1130
## Predicted   8105 11156 3287 4895.46 24411 2690.9 12830 5943 35214 -2376
## cvpred      8172 11492 3175 4921.77 24486 2268.6 12680 6120 36102 -2442
## charges     7518  8597 2903 4915.06 14572 2250.8  5757 8605 41949  1729
## CV residual -654 -2895 -272   -6.71 -9914  -17.8 -6923 2486  5848  4171
##               1138 1186   1200 1246   1307  1330  1336
## Predicted   1877 7827 4688.9 4202  26315 15118  4106
## cvpred      1576 7759 4976.2 4805  26448 15789  3809
## charges     3176 8604 4934.7 5615  16115 10325  1630
## CV residual 1600  844  -41.5  810 -10333 -5464 -2179
##
## Sum of squares = 2.51e+09    Mean square = 37469753    n = 67
##
## fold 2
## Observations in test set: 67
##                11   32    74  107   115  122  136   148  155  192   217
## Predicted   2464  999 13236 1959 13633 -443 1722 13999 7175 5318 11046
## cvpred      2592 1144 13411 1961 13807 -326 1873 14196 7349 5337 11512
## charges     2721 2198 11947 2332 11488 1706 2156  9878 7077 4884 10356
## CV residual  130 1054 -1464  370 -2318 2032  282 -4319 -272 -453 -1157
##               218    230   239   282   287  295   300   310   313   315
## Predicted    713 9130.4 26286 41549 18236 3459 11025 11206 35633 28953
## cvpred       840 9289.7 26584 41630 18440 3447 11490 11668 35695 28862
## charges     2484 9225.3 17353 48549  9433 3906  9249  7749 42125 34839
## CV residual 1644  -64.4 -9231  6919 -9007  460 -2240 -3919  6429  5977
##               316   328  370  385   392   477  497   547   553   591  609
## Predicted   12687 37290 3551 6987  5969 27146 3817  6940 10154 12488 4498
## cvpred      12853 37632 3681 7140  6403 27171 3833  7079 10174 12536 4635
## charges      9723 42761 3482 8303  2138 35148 4932  3269 12957 11842 4435
## CV residual -3131  5129 -200 1163 -4264  7976 1099 -3811  2783  -693 -200
##               628   638   642  676   692   711   742   776   811   818
## Predicted   10488 11255 33681 6347 12337  3480 27164 12737 11255  6988
```

```
## cvpred      10641 11436 34015 6773 12354  3608 27193 12910 11298  6981
## charges     11327 24915 32787 7223  8068  1728 18246 10560  9415  3598
## CV residual   686 13479 -1227  450 -4285 -1881 -8946 -2349 -1883 -3384
##                819   841   846   850   855   890   894   934     938   953
## Predicted    33421  2701 37988 13617 33975 14960 37458 11402 8514.97  5878
## cvpred       33501  2679 38087 14065 34062 15414 37525 11439 8975.39  6320
## charges      23401  1526 45009 10602 24107 11945 44203  7348 8965.80  4527
## CV residual -10100 -1152  6922 -3463 -9955 -3468  6677 -4091   -9.59 -1793
##                995  1007  1034  1038  1074  1087  1089  1124  1129  1141
## Predicted    26701  5401 22746 34892 12839 11820 17814  6340  7409 13036
## cvpred       27035  5540 22758 35252 13036 11987 17994  6504  7407 13206
## charges      16420  4438 13748 39726 12097 10796  9749 18903 14358  9048
## CV residual -10615 -1102 -9010  4473  -940 -1190 -8245 12399  6952 -4158
##               1228  1265  1269  1320  1327
## Predicted   11329 13230  3722  7983 10207
## cvpred      11491 13425  3727  8437 10388
## charges      7162 10371  1880  7202  7050
## CV residual -4329 -3054 -1847 -1235 -3338
##
## Sum of squares = 1.69e+09    Mean square = 25152541    n = 67
##
## fold 3
## Observations in test set: 67
##                 44   66    73   106   111   116  120   169   172   194  254
## Predicted    7851 1704 12114 26601 14134 13047 6265  4202 10098 12326 5397
## cvpred       8158 1950 12197 27023 13650 12301 6113  4325  9738 12233 5022
## charges      6314 1743 11742 17560 10825 30260 6686  2719  8117 12044 4261
## CV residual -1844 -207  -455 -9463 -2825 17959  574 -1606 -1622  -189 -762
##                262  271    281  294   303    352   371   388    412   491
## Predicted   25200 1171  32282 2027 15743 8937.6 11068 10211  30261  3289
## cvpred      26437  934  33091 2362 16176 9029.5 10846  9507  30861  3636
## charges     17085 1719  22332 2157 12266 8932.1 13415 30285  19595  1749
## CV residual -9352  785 -10759 -205 -3910  -97.4  2569 20778 -11267 -1887
##                496    503   525   538   552   566   610   619   644   697
## Predicted   1764  31876 30530 10564  4826  3222 32795 26954  7452 13919
## cvpred      1092  32379 31125 10738  5141  3333 33594 28374  7602 13811
## charges     1967  22218 38246  8825  3973  2128 39241 34440  4467 29186
## CV residual  875 -10161  7120 -1913 -1168 -1205  5647  6066 -3136 15375
##                701   708      716   724   769    794    801   821   835
## Predicted    4148 11706 12573.75  3853 18379  31059 6945.0 10796  9250
## cvpred       4638 11066 12153.69  3691 18795  31521 7145.8 10509  8812
## charges      2021 10264 12146.97  1263 14319  21196 7046.7  7446  5377
## CV residual -2617  -801    -6.72 -2427 -4476 -10325  -99.1 -3064 -3435
##                845   868  875   877   899  925   941   971  972   975  992
## Predicted   11871 18063 7327  9710  5457 5633 -1723 10840 4316  5318 8018
## cvpred      11189 18001 6383  9820  6089 5108 -2092 11029 4074  5238 7832
## charges     10072 11576 8891 26140  1635 6250  1122 10703 4992  2323 7145
## CV residual -1117 -6425 2508 16321 -4455 1143  3214  -326  918 -2915 -687
##               1028  1042   1046  1053  1106   1177  1187  1188 1195 1235
## Predicted    -768  -706  32499 10959 11802  34300 30517 15934 3066 8878
```

```
## cvpred      -1535 -1495  33336 10248 12125  35123 31089 16002 2930 9092
## charges     21595  1705  21881  9288 10339  23888 37465 13845 4134 8516
## CV residual 23130  3200 -11455  -960 -1786 -11236  6377 -2157 1204 -576
##              1239  1240  1264 1287 1288  1295
## Predicted   5737  8579  9166  179 5653 11141
## cvpred      4830  9221  9360 -206 5762 10314
## charges     6986  3238  7338 3733 5472 11931
## CV residual 2155 -5983 -2022 3939 -290  1617
##
## Sum of squares = 3.13e+09    Mean square = 46693070    n = 67
##
## fold 4
## Observations in test set: 67
##                33    47    51  123   144   180   185   190  198   200   220
## Predicted   3852  6797  4686 2895  6157 11405  9310  7523 9183 18827   356
## cvpred      3354  6803  4878 2745  6009 11480  8869  7245 8629 19299  -140
## charges     4688  3393  2211 2257 18158  8538  7731  4923 8517 14902 25082
## CV residual 1333 -3410 -2666 -488 12149 -2942 -1138 -2322 -112 -4398 25221
##                252   253   276   299   336   345    346  410   429   468
## Predicted   40094 37513 10008 33059 15894 16290 6867.5 5289 -1484 15639
## cvpred      39128 36124 10226 31821 16234 15552 6153.8 4863 -1318 15527
## charges     47305 44261  9716 38746 13823 10977 6184.3 4074  3167 12643
## CV residual  8177  8137  -510  6926 -2412 -4575   30.5 -789  4485 -2883
##                514   523   534   544   545   550   559   560   570  623
## Predicted   1872 13101  8653 42261 12326 38751 34474  4790 39735 7412
## cvpred      1939 13504  8333 40893 12512 37316 33137  4743 38678 7476
## charges     1256  9866 19215 63770 10231 45863 39983  1646 45702 9182
## CV residual -683 -3638 10882 22878 -2280  8547  6847 -3097  7024 1706
##                656   674   683  694   725  756   791  814  817  822   829
## Predicted   33859  8133 34284 1464 10569 5588 11852 2729 2116 -355 34092
## cvpred      32354  7726 33284 1477 10894 5828 11406 2963 2003 -315 33295
## charges     24667  6185 40104 2353 10106 5031  5662 4429 2843 2681 39597
## CV residual -7686 -1540  6819  876  -788 -796 -5744 1466  840 2996  6302
##                830   837   859   861   874   913   955   957  969  1037 1066
## Predicted   5335  6992  4658 39031  8819 14097 30382 35709 2080 29830 7068
## cvpred      5358  7163  4058 37864  8949 13932 29356 34417 2262 28232 7079
## charges     6117  4402 18218 46114  6849 14383 20010 42000 3280 37484 7045
## CV residual  759 -2761 14160  8249 -2100   451 -9347  7582 1018  9252  -33
##                1126  1162  1169  1184  1185  1204 1206  1218 1243  1273
## Predicted   13199 11086  8262 10131 26680 12488 2653  7791 1758  8346
## cvpred      13330 10555  8221 10443 25073 12932 2660  7240 1373  7646
## charges     14255  5124  4671  9447 18328  9964 5117  4058 4296 14478
## CV residual   925 -5431 -3551  -995 -6744 -2968 2456 -3182 2923  6833
##                1281  1313 1325  1326
## Predicted   10977 11409 4749 15279
## cvpred      10608 11452 4712 15871
## charges      8284  4536 4240 13143
## CV residual -2324 -6916 -472 -2728
##
## Sum of squares = 2.86e+09    Mean square = 42636854    n = 67
```

```
##
## fold 5
## Observations in test set: 67
##                  5   18     29    65  126   130   159   228    237   268
## Predicted    5738  972 -841.4 24353 4181  9717 30517 17130   97.5 16063
## cvpred       5971 1775   26.4 24113 4331  9437 30355 16412  616.2 15605
## charges      3867 2395 2775.2 14712 3385  6082 36950 24227 1615.8 14591
## CV residual -2104  620 2748.8 -9401 -945 -3355  6595  7816  999.6 -1014
##               277    312   344   347   348   362  364   406   444  483
## Predicted    -211   40.6 17110  8283 11525  6773 1718 16051 15724 1927
## cvpred        589   46.8 16772  8617 11595  6706 1634 15430 15232 2249
## charges      2804 1737.4 13982  4890  8334  4751 2598 11397 28288 1622
## CV residual 2215 1690.6 -2791 -3727 -3260 -1955  964 -4033 13056 -627
##                485   489   511  530   606   653   661   675    684   693
## Predicted    12766 35802 13791  234 12108 10105 14108 39018 9717.2  4277
## cvpred       12368 34927 13766 1042 11826  9994 13934 37916 9836.1  4549
## charges       9563 48885 11763 1708  9284  8281  6436 46201 9863.5  2362
## CV residual -2805 13958 -2003  666 -2543 -1714 -7499  8285   27.4 -2186
##                705   721  745   750   758   813   820   825   886   893
## Predicted    11013 15772 9719  5426 32577  8231 33127 11646 28907  9588
## cvpred       10609 15046 9792  5633 32094  8884 32136 11662 29008  9927
## charges       8931  9876 8827  3063 23065 11014 55135 12524 19720 10423
## CV residual -1678 -5171 -965 -2571 -9029  2129 23000   862 -9289   496
##                902   904   908   918   965  1002  1004  1014  1056  1060
## Predicted    40917 12220  9936 30711 15278 28215  9545 11954 11585  8148
## cvpred       40083 12250  9843 30703 14902 27677  9314 11817 11533  8181
## charges      48674  8126  7634 35069 26467 34473 21232  8765 10595  4463
## CV residual  8591 -4124 -2210  4366 11565  6795 11918 -3051  -938 -3719
##               1093 1096  1102  1142  1150 1176  1198  1203  1207  1226
## Predicted     5245 4806 11886 10591  9674 1818  8709  4262 15968  9868
## cvpred        5283 5008 11658  9969  9340 1677  8997  4505 14976  9618
## charges       3591 4561 11253  7955  5980 2154  5700  2055 36911  4796
## CV residual -1692 -447  -404 -2014 -3360  477 -3297 -2450 21935 -4822
##               1238  1253  1274  1275 1286  1289  1319
## Predicted    13063 24974  5157 26060 8198 30673 12375
## cvpred       12543 24726  5798 26319 8229 29931 12369
## charges      12224 16233  4747 17043 8535 38345 19497
## CV residual  -319 -8493 -1051 -9275  306  8414  7128
##
## Sum of squares = 2.82e+09    Mean square = 42073379    n = 67
##
## Overall (Sum over all 67 folds)
##       ms
## 38805119
```