Predicting High Healthcare Costs

Myron Keith Gibert Jr

2022-11-07

Contents

Introduction

Data for this project can be found at:

Setup

Sets chunk options. Installs packages (if needed) and then loads them into R.

Read data

The data is in .csv format, so we'll load in using read.csv

```
# Read in the csv file.  Set the first line of the file as the header for the table

data <- read.csv("insurance.csv",header=TRUE)

head(data,5)
```

```
##   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
```

I want age, bmi, children, and charges to be numeric, so let's convert them.

Modify factor classes

```
data_fixed <- data %>%
        mutate(age = as.numeric(age),
               bmi = as.numeric(bmi),
               children = as.numeric(children),
               charges = as.numeric(charges))

sapply(data_fixed,class)
```

```
##         age         sex         bmi    children      smoker      region
##   "numeric" "character"   "numeric"   "numeric" "character" "character"
##     charges
##   "numeric"
```

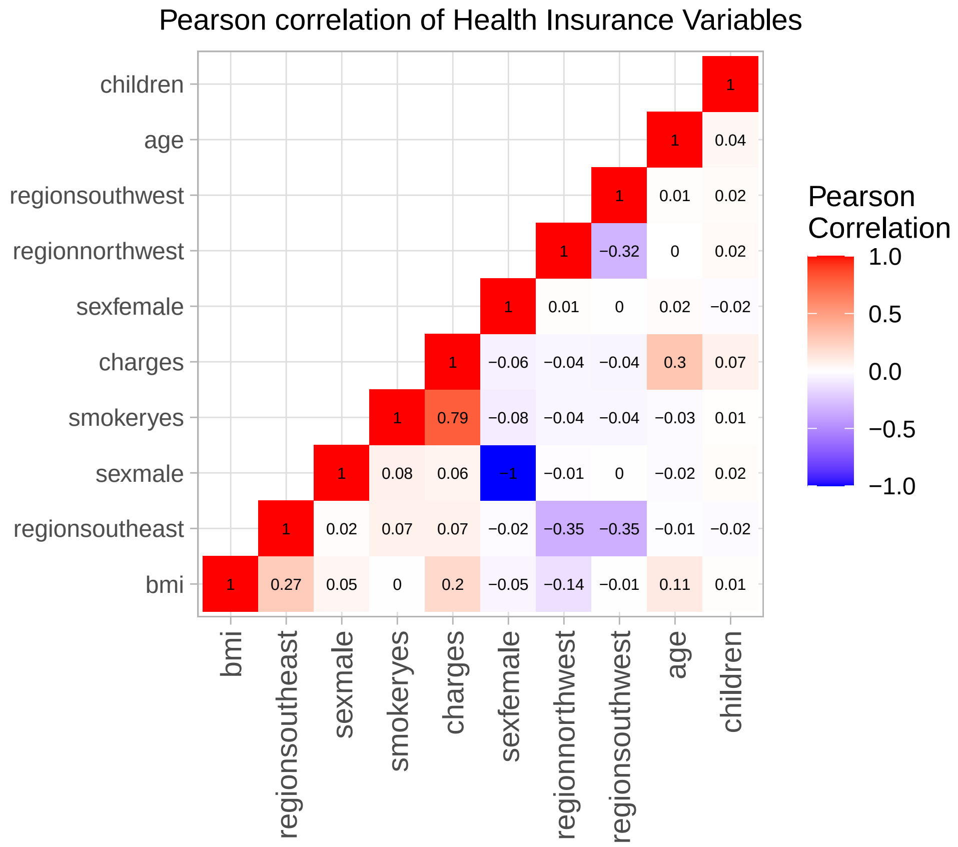
Exploring whether any variable strongly correlates with high patient expenses.

Write a function to generation correlation matrix for plotting

```r
process_cormat <- function(x){

dd <- as.dist((1-x)/2)
hc <- hclust(dd)
x <-x[hc$order, hc$order]

x[upper.tri(x)] <- NA
  return(x)

}
```

Simple correlation plot

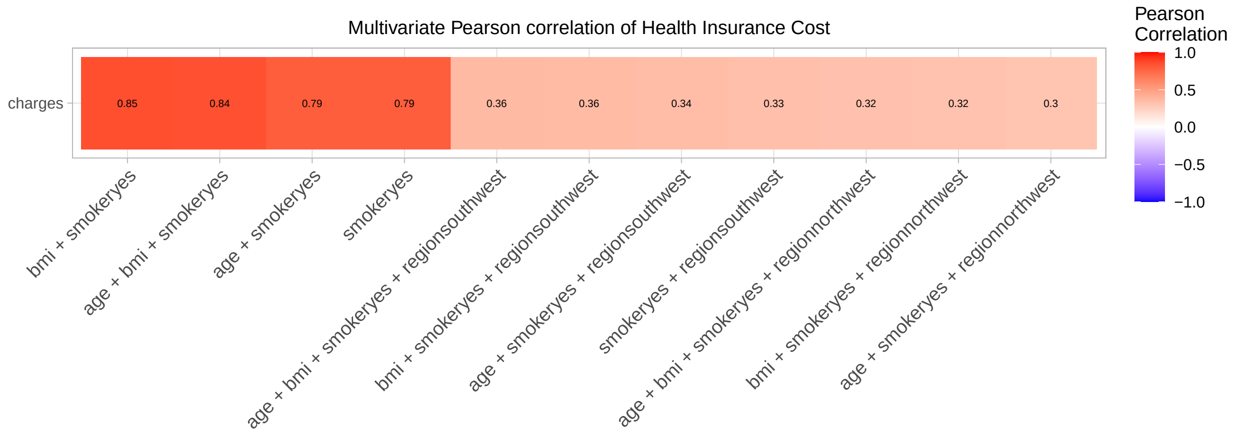
```r
corrdata <- model.matrix(~0+., data=data_fixed) %>%
  cor(use="pairwise.complete.obs")

corrdata_processed <- process_cormat(corrdata)

### Melt
corrdata_tall <- melt(corrdata_processed, na.rm = TRUE)

p <- ggplot(corrdata_tall,aes(Var1,Var2,fill=as.numeric(value))) +
  geom_tile() +
  geom_text(size=rel(2.0),aes(label=round(as.numeric(value),2))) +
  ggtitle("Pearson correlation of Health Insurance Variables") +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme(axis.text.x=element_text(size=rel(1.2),
                                 angle = 90,
                                 vjust = 0.5,
                                 hjust = 1),
        plot.title = element_text(size=rel(1.0),
                                  hjust = 0.5),
        axis.title=element_blank()) +
  coord_fixed()

p
```



Being a smoker is very likely to result in higher health costs. Age and BMI have weaker correlations. Let's take a look at some combinations!

Multiple correlation plot (Charges vs Smoker+Other Variables)

```
## Using the model matrix function. ~ to set comparison,
## 0+ to include intercept (omit if not needed), and "." to indicate that we
## wish to compare all variables.

corrdata <- model.matrix(~0+charges*age*bmi*smoker*., data=data_fixed) %>%
  cor(use="pairwise.complete.obs")

corrdata_processed <- process_cormat(corrdata)

### Melt
corrdata_tall <- melt(corrdata_processed, na.rm = TRUE) %>%
  mutate(detector = str_detect(Var1,"charges"),Var3 = str_replace_all(Var1,":"," + ")) %>%
  filter(Var2 == "charges" & detector == FALSE & abs(value) >= 0.3)

p <- ggplot(corrdata_tall,aes(reorder(Var3,desc(value)),Var2,fill=as.numeric(value))) +
  geom_tile() +
  geom_text(size=rel(2.0),aes(label=round(as.numeric(value),2))) +
  ggtitle("Multivariate Pearson correlation of Health Insurance Cost") +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme(axis.text.x=element_text(size=rel(1.2),
                                 angle = 45,
                                 hjust = 1),
        plot.title = element_text(size=rel(1.0),
                                  hjust = 0.5),
        axis.title=element_blank()) +
  coord_fixed()

p
```



Having a high BMI while being a older smoker seems to correlate very highly with having high health costs. However, these variables might not be worth including over just looking at smoking alone, since it only leads to a slight increase in the correlation.

Let's use regression modeling using the three variables (age,bmi,smoker status) and compare the model to a simple regression model for smoker status alone to see if using all three is actually better.

Modeling Healthcare costs using Age, BMI, and Smoker status and comparing to Simple Regression using only Smoker status.

Split data into testing and training sets

```
set.seed(1234)

data_subset_split <- initial_split(data_fixed,prop=0.80)

data_subset_training <- training(data_subset_split)
```

Construct a multiple regression model using Age, BMI, and Smoker status to predict cost

```
lm_model <- linear_reg() %>%
            set_engine('lm') %>% # adds lm implementation of linear regression
            set_mode('regression')

lm_fit <- lm_model %>%
          fit(charges ~ age+bmi+smoker, data = data_subset_training)

fitdata <- tidy(lm_fit)

fitdata
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  -10994.    1046.      -10.5  1.14e- 24
## 2 age             260.      13.5      19.3  3.29e- 71
## 3 bmi             302.      30.5       9.87 4.67e- 22
## 4 smokeryes     24088.     469.       51.4  9.19e-291
```

Test model

```
data_subset_testing <- testing(data_subset_split)

charge_predictions_train <- predict(lm_fit,new_data = data_subset_training)

charge_test_results_train <- data_subset_training %>%
  dplyr::select(age,bmi,smoker,charges) %>%
  bind_cols(charge_predictions_train)

charge_predictions_test <- predict(lm_fit,new_data = data_subset_testing)

charge_test_results_test <- data_subset_testing %>%
  dplyr::select(age,bmi,smoker,charges) %>%
  bind_cols(charge_predictions_test)
```

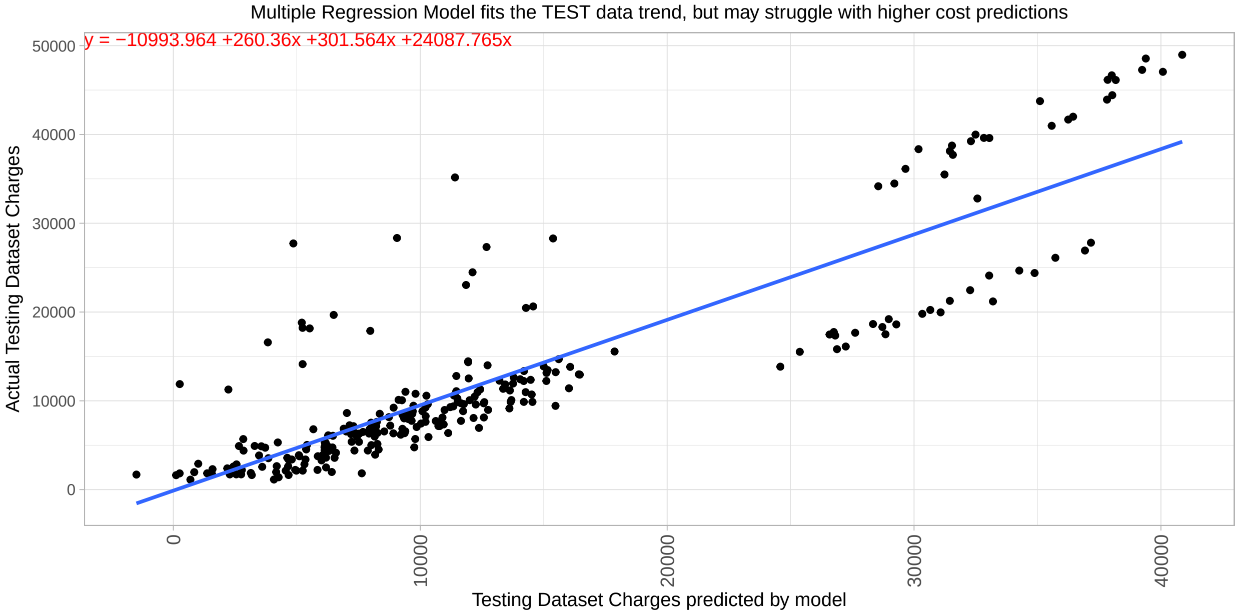
Write Equation to Extract Regression Equation

```
extract_eq <- function(lm_fit,digits){
        fitdata <- tidy(lm_fit) %>%
                mutate(operator = ifelse(estimate >= 0,"+","-"),
                       variable = ifelse(term == "(Intercept)","","x"))

        equation_result <- "y = "
        i <- 1

        for(i in 1:length(fitdata$term)){

        equation_result <- paste(
          equation_result,
          fitdata$operator[i],
          round(abs(fitdata$estimate[i]),digits),
          fitdata$variable[i],
          " ",sep="")

        }

        equation_result
}
```

Evaluate Model

Visualize Predicted vs Actual Results on the Testing Set

```
reg_equation <- extract_eq(lm_fit,3)

p <- ggplot(charge_test_results_test,(aes(.pred,charges))) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  xlab("Testing Dataset Charges predicted by model") +
  ylab(expression("Actual Testing Dataset Charges")) +
  ggtitle("Multiple Regression Model fits the TEST data trend,
          but may struggle with higher cost predictions") +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.2),
                                 angle = 90,
                                 vjust = 0.5,
                                 hjust = 1),
        plot.title = element_text(size=rel(1.0),
                                  hjust = 0.5),
        legend.position = "none") +
annotate('text',
         label=reg_equation,
         x=-Inf,
         y=Inf,
         hjust=0,
         vjust=1,
         color="red")

p
```



Generate Alternative Models

Alternative Model 1: Simple regression model using Smoker status to predict cost

```
lm_model <- linear_reg() %>%
            set_engine('lm') %>% # adds lm implementation of linear regression
            set_mode('regression')

lm_fit <- lm_model %>%
          fit(charges ~ smoker, data = data_subset_training)

fitdata <- tidy(lm_fit)

fitdata
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    8514.      260.      32.8 1.12e-163
## 2 smokeryes     23871.      570.      41.9 1.78e-227
```

```
charge_predictions_smoker <- predict(lm_fit,new_data = data_subset_testing)

rsq_smoker <- data_subset_testing %>%
  dplyr::select(smoker,charges) %>%
  bind_cols(charge_predictions_smoker) %>%
  rsq(truth = charges, estimate = .pred) %>%
  mutate(dataset = "Test (Smoker only)") %>%
  dplyr::select(dataset,.estimate)
```

Alternative Model 2: Multiple regression model using BMI + Smoker status to predict cost

```
#BMI + Smoker

lm_model <- linear_reg() %>%
            set_engine('lm') %>% # adds lm implementation of linear regression
            set_mode('regression')

lm_fit <- lm_model %>%
          fit(charges ~ bmi+smoker, data = data_subset_training)

fitdata <- tidy(lm_fit)

fitdata
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   -2657.    1105.      -2.41 1.63e-  2
## 2 bmi             366.      35.2     10.4  4.24e- 24
## 3 smokeryes     23802.     544.      43.8  1.80e-240
```

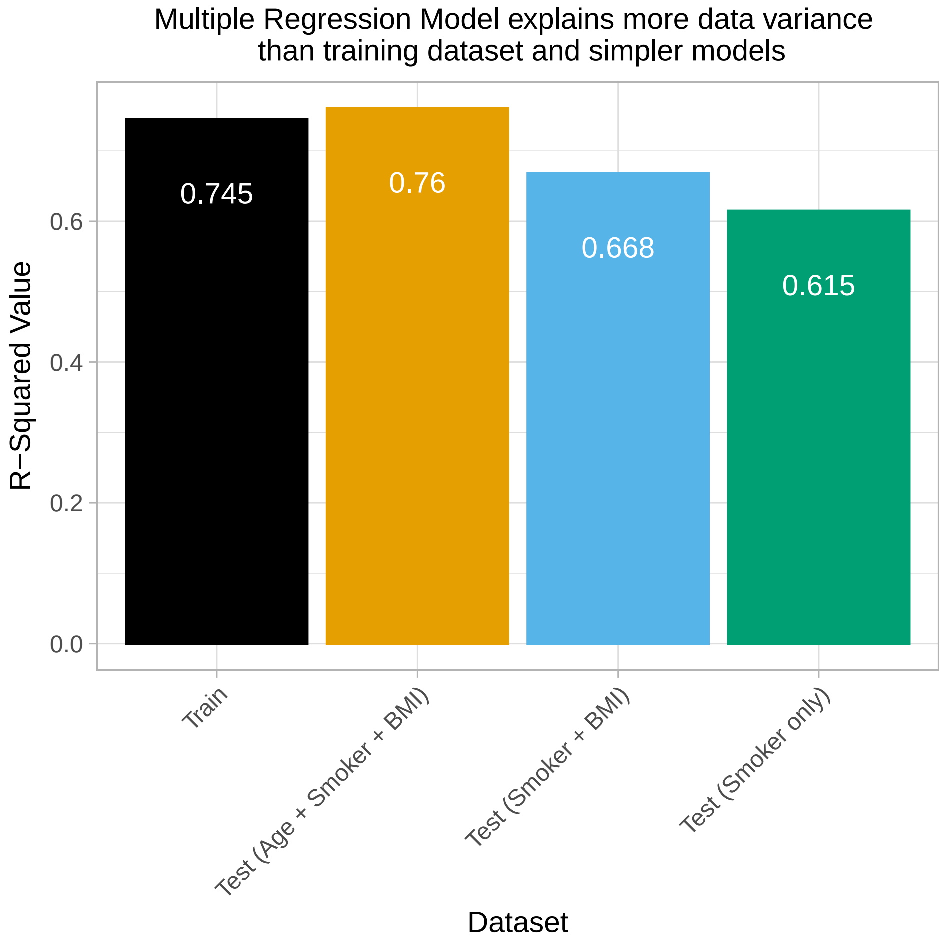
```
charge_predictions_bmismoker <- predict(lm_fit,new_data = data_subset_testing)

rsq_bmismoker <- data_subset_testing %>%
  dplyr::select(smoker,charges) %>%
  bind_cols(charge_predictions_bmismoker) %>%
  rsq(truth = charges, estimate = .pred) %>%
  mutate(dataset = "Test (Smoker + BMI)") %>%
  dplyr::select(dataset,.estimate)
```

Compare Testing Model for Healthcare Costs to alternatives using R-Squared Method

Combine model results

```
rsq_train <- charge_test_results_train %>%
  rsq(truth = charges, estimate = .pred) %>%
  mutate(dataset = "Train") %>%
  dplyr::select(dataset,.estimate)

rsq_test <- charge_test_results_test %>%
  rsq(truth = charges, estimate = .pred) %>%
  mutate(dataset = "Test (Age + Smoker + BMI)") %>%
  dplyr::select(dataset,.estimate)

rsq_combined <- rbind(rsq_train,rsq_test,rsq_smoker,rsq_bmismoker) %>%
  mutate(dataset = fct_relevel(dataset,
                               "Train",
                               "Test (Age + Smoker + BMI)",
                               "Test (Smoker + BMI)",
                               "Test (Smoker only)",))

p <- ggplot(rsq_combined,(aes(dataset,.estimate,fill=dataset,color=dataset))) +
  geom_bar(stat="identity",position="dodge") +
  geom_text(size=rel(4.0),
            vjust=4,color="white",
            aes(label=round(as.numeric(.estimate),3))) +
  xlab("Dataset") +
  ylab(expression("R-Squared Value")) +
  ggtitle("Multiple Regression Model explains more data variance
          \n than training dataset and simpler models") +
  theme_light() +
  scale_color_colorblind() +
  scale_fill_colorblind() +
  theme(axis.text.x=element_text(size=rel(1.0),
                                 angle = 45,
                                 hjust = 1),
        plot.title = element_text(size=rel(1.0),
                                  hjust = 0.5),
        legend.position = "none")

p
```

Plot results



Discussion

Conclusions

Additional Caveats to Consider