Regression Analysis of Medical Charges

Presented by:
David Oniani, Madeline Pope, Zach Sturgeon

Background

Data Attributes

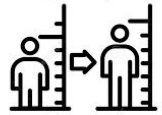
- Personal medical costs of 1,339 random individuals
- Dataset was collected from Kaggle (Thanks Miri Choi)

Question we would like to answer:

- Can we accurately predict what an individual's medical charges would be if we attain these 6 variables from the individual?
- What factors best predict an individual's medical charges

Variables

BMI

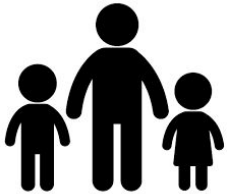


numerical

Sex

categorical

Children



categorical

Smoker

categorical

Age

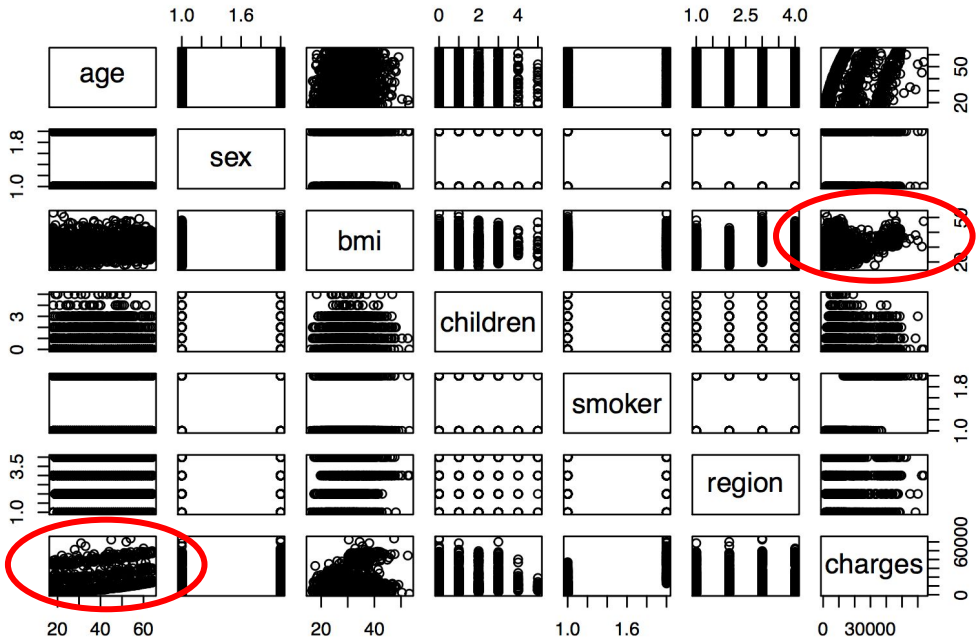
numerical

Region

categorical

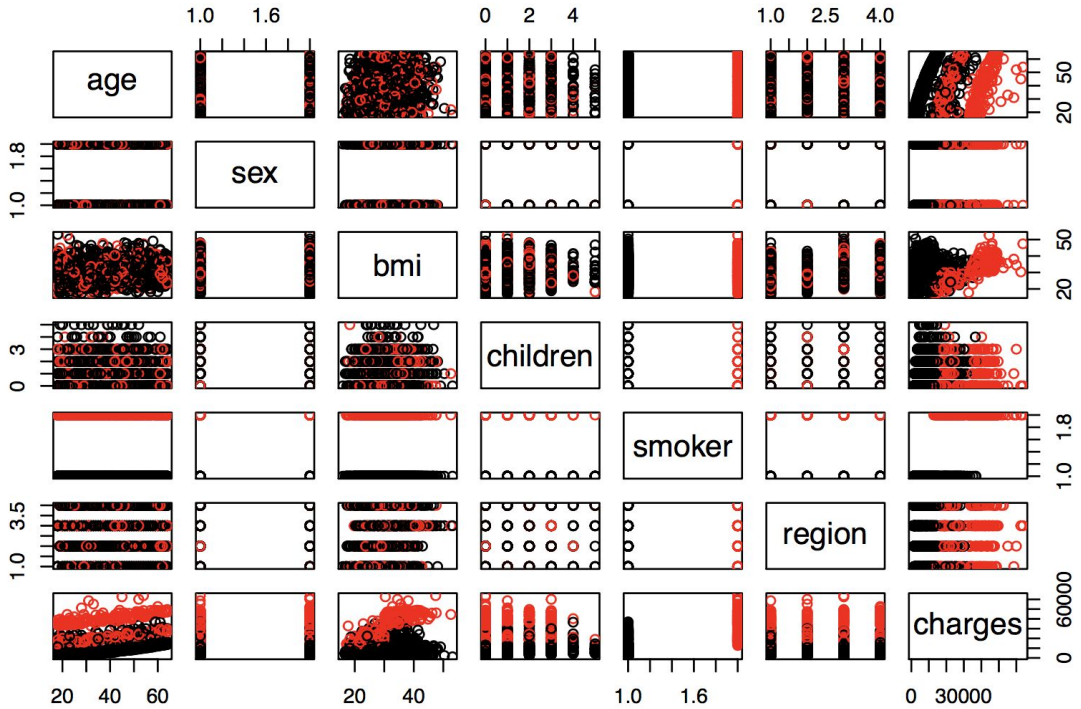
Medical Charges

Correlation Plots



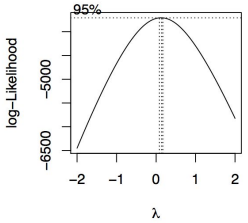
- Plots that caught our eyes are circled in red
- Expected correlation between age and charges
- Children behaves as a categorical variable
- Correlation between BMI and charges is unclear

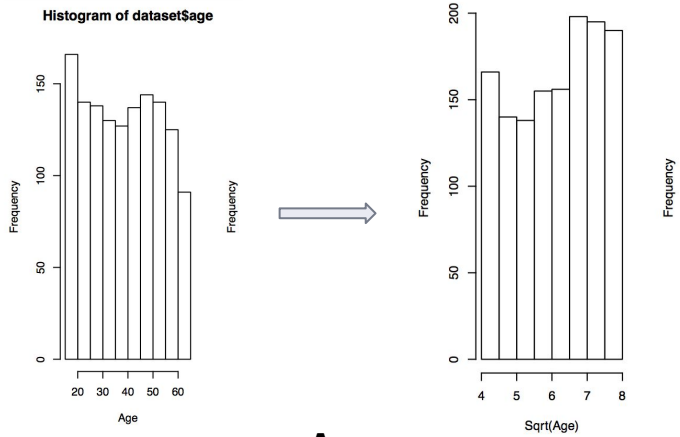
Correlation Plots: Colored by Smoking Status



- Black = Non - Smoker
- Red = Smoker
- Smoking status appears highly correlated with charges

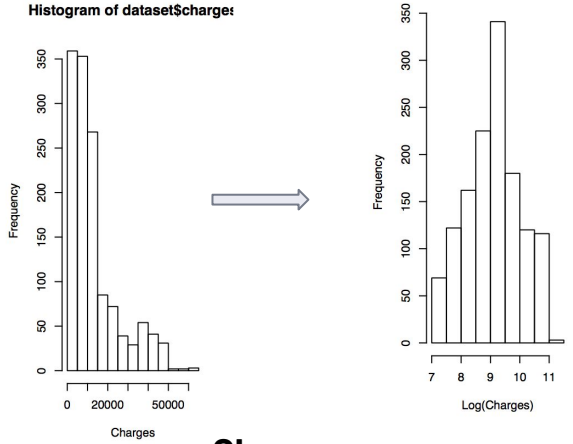
Transformations





Age

- Original histogram of age showed a right skewed distribution
- Ultimately chose to square root transform age to achieve a less skewed distribution



Charges

- Original histogram of charges showed a right skewed distribution
- Log transformation of charges normalized the distribution

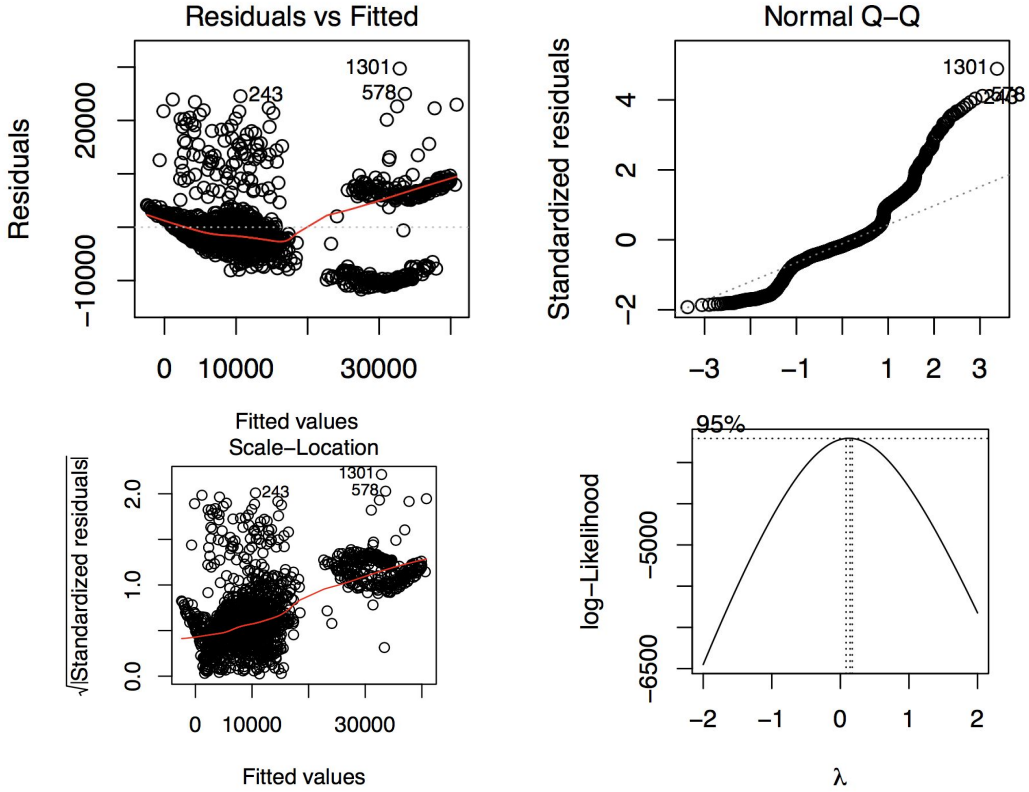
First Order Model

```
allFit0 <- lm(charges ~ sqrt(age) + sex + bmi + children + smoker + region, data=dataset)
summary(allFit0)
```

```
##
## Call:
## lm(formula = charges ~ sqrt(age) + sex + bmi + children + smoker +
##     region, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11674  -2891  -1017   1556  29716
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -20813.44    1219.58 -17.066   <2e-16 ***
## sqrt(age)         3075.79     145.78  21.099   <2e-16 ***
## sexmale           -131.83     334.90  -0.394   0.6939
## smokeryes        23837.59     415.56  57.362   <2e-16 ***
## regionnorthwest   -351.16     479.06  -0.733   0.4637
## regionsoutheast  -1044.47     481.49  -2.169   0.0302 *
## regionsouthwest   -966.36     480.73  -2.010   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6098 on 1329 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.7465
## F-statistic: 493.1 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- 74.56% of the variation in charges is explained by the model
- On average, predictions of the model are $6,098 away from the real value

Residual and Box-Cox



- Weak Evidence that both sets of residuals are coming from normal distributions (Q-Q plot)
- Square Root of Standardized Residuals VS Fitted values shows a set of lines which look like a curved line. This suggests that the residuals are not spread equally along the ranges of predictors and that the variance is not constant.
- Box-Cox reinforced observation that log transforming charges is acceptable

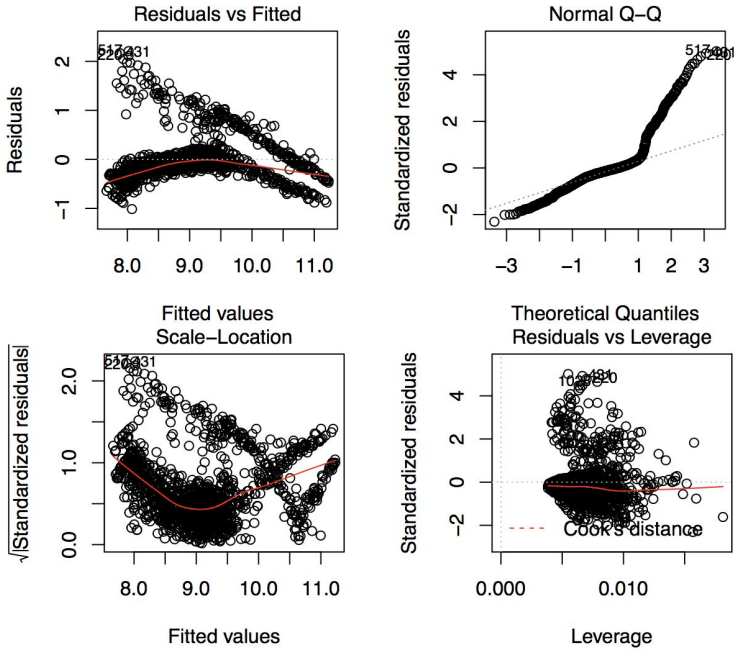
Intermediate Model

```
allFit1 <- lm(log(charges) ~ sqrt(age) + sex + bmi + children + smoker + region, data=dataset)
summary(allFit1)
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + sex + bmi + children +
##     smoker + region, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01382 -0.20734 -0.06487  0.05970  2.21077
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.788769   0.088497  65.412  < 2e-16 ***
## sqrt(age)        0.422666   0.010578  39.956  < 2e-16 ***
## sexmale         -0.074988   0.024301  -3.086 0.002072 **
## bmi              0.013616   0.002087   6.525 9.62e-11 ***
## children         0.091960   0.010071   9.131  < 2e-16 ***
## smokeryes        1.553346   0.030155  51.512  < 2e-16 ***
## regionnorthwest -0.063431   0.034763  -1.825 0.068275 .
## regionsoutheast -0.157451   0.034939  -4.506 7.17e-06 ***
## regionsouthwest -0.129619   0.034884  -3.716 0.000211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4425 on 1329 degrees of freedom
## Multiple R-squared:  0.7698, Adjusted R-squared:  0.7685
## F-statistic: 555.7 on 8 and 1329 DF,  p-value: < 2.2e-16
```

- When least significant predictors were removed (sex and region) model performance decreased
- Predictions of the model are .4425 log(dollars) away from the real charge value
- 76.85% of the variation in charges can be explained by this model

Residual Analysis



- These plots still show that the spread of residuals are not consistent along the range of predictors, but they are improved from our initial model

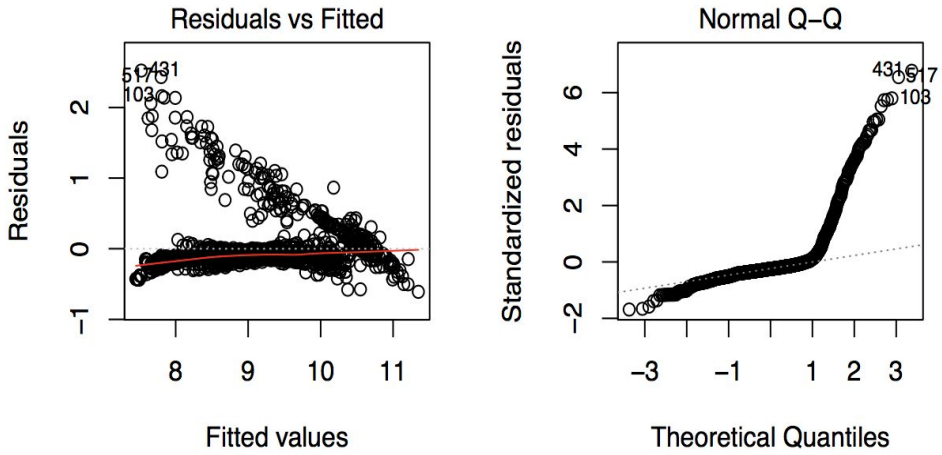
Stepwise Regression

- Yielded the same model
- Let's try some interaction effects

```
## Start:  AIC=-2175.37
## log(charges) ~ sqrt(age) + as.factor(sex) + bmi + as.factor(children) +
##     as.factor(smoker) + as.factor(region)
##
##                       Df Sum of Sq    RSS      AIC
## <none>                               258.18 -2175.37
## - as.factor(sex)       1      1.88 260.06 -2167.67
## - as.factor(region)    3      4.75 262.94 -2156.96
## - bmi                  1      8.14 266.32 -2135.82
## - as.factor(children)  5     18.33 276.51 -2093.62
## - sqrt(age)            1    311.58 569.76 -1118.26
## - as.factor(smoker)    1    517.24 775.42  -705.91
```

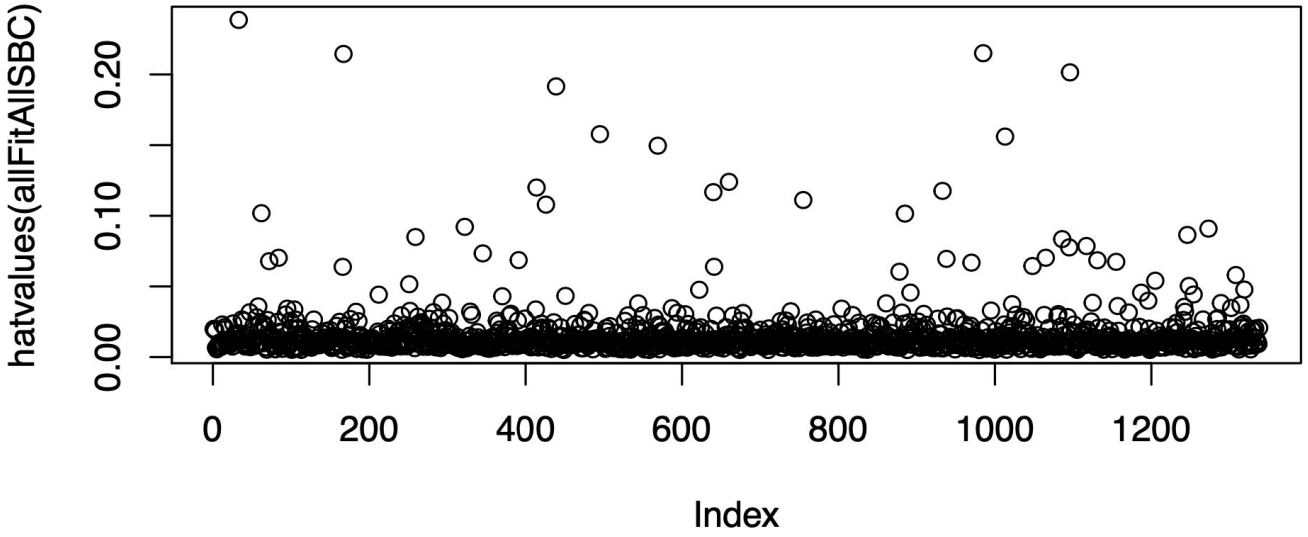
Interaction Effects

```
##
## Residual standard error: 0.3745 on 1317 degrees of freedom
## Multiple R-squared:  0.8366, Adjusted R-squared:  0.8341
## F-statistic: 337.2 on 20 and 1317 DF,  p-value: < 2.2e-16
```



- Interaction effects are incorporated in the stepwise regression
- 83.41% of the variation in charges is explained by the model
- Residuals vs fitted values shows much less curvature than before

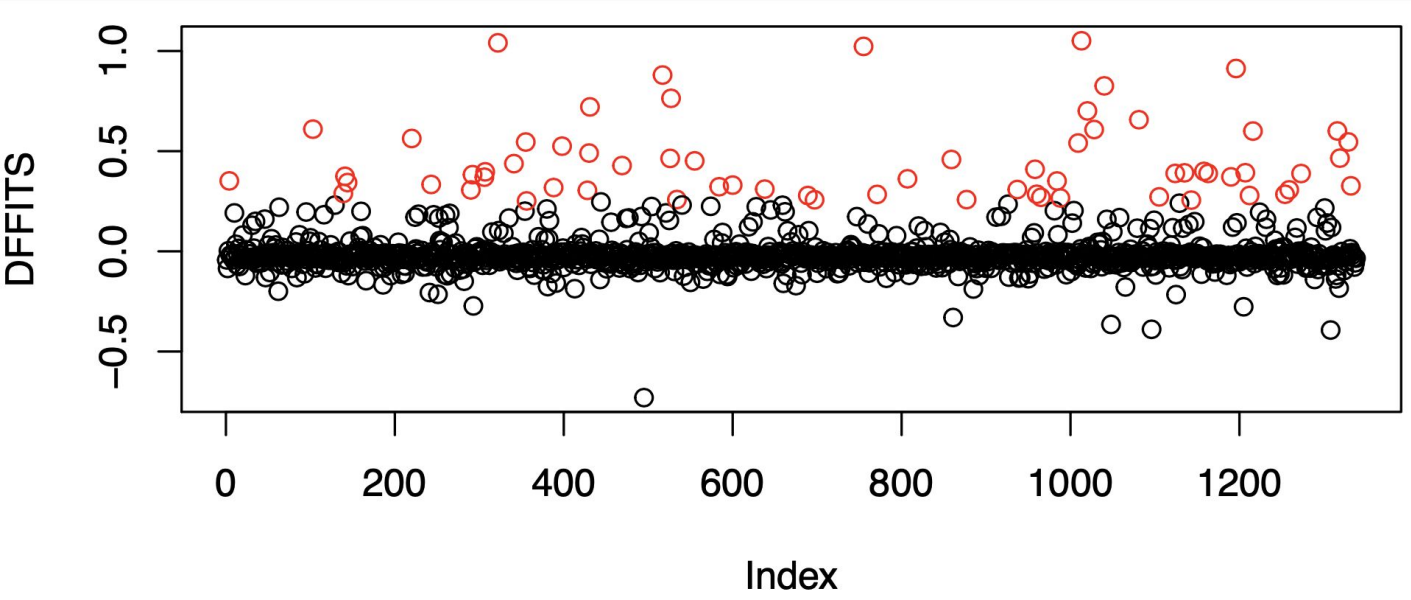
Hat Matrix Diagonals



- There are a few points over the threshold
- As the threshold for Hat Matrix Diagonals is $\frac{2p}{n}$, a few of these points are not far enough from the threshold to be of concern

DFFITS

- Possible leverage points are colored red
- Threshold for coloring is $2\sqrt{\frac{p}{n}}$



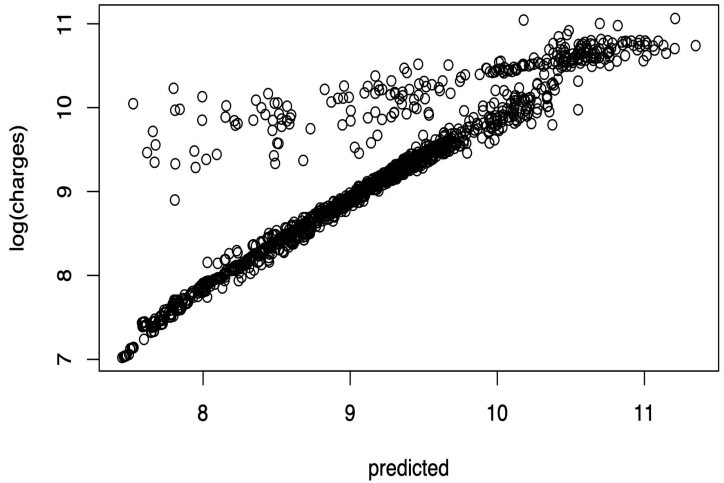
VIF (Variance Inflation Factor)

- All values less than 5, no indication of problematic collinearity

```
##                                      GVIF Df GVIF^(1/(2*Df))
## centeredSqrtAge                  2.905291  1        1.704492
## as.factor(sex)                   1.014229  1        1.007089
## centeredBMI                      1.398220  1        1.182463
## as.factor(children)              1.166555  5        1.015525
## as.factor(smoker)                1.021240  1        1.010564
## as.factor(region)                1.126837  3        1.020102
## centeredSqrtAge:as.factor(sex)   2.050878  1        1.432089
## centeredSqrtAge:as.factor(children) 1.972008  5        1.070264
## centeredSqrtAge:as.factor(smoker)   1.281435  1        1.132005
## centeredBMI:as.factor(smoker)    1.296649  1        1.138705
```

10-Fold Cross Validation

```
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 1337, 1337, 1337, 1337, 1337, 1337, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.3776655  0.831198  0.2112711
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```



- 83.11% of the variation in charges is explained by the model
- Actual VS Predicted plot is approximately

$$y = x$$

Thank You!