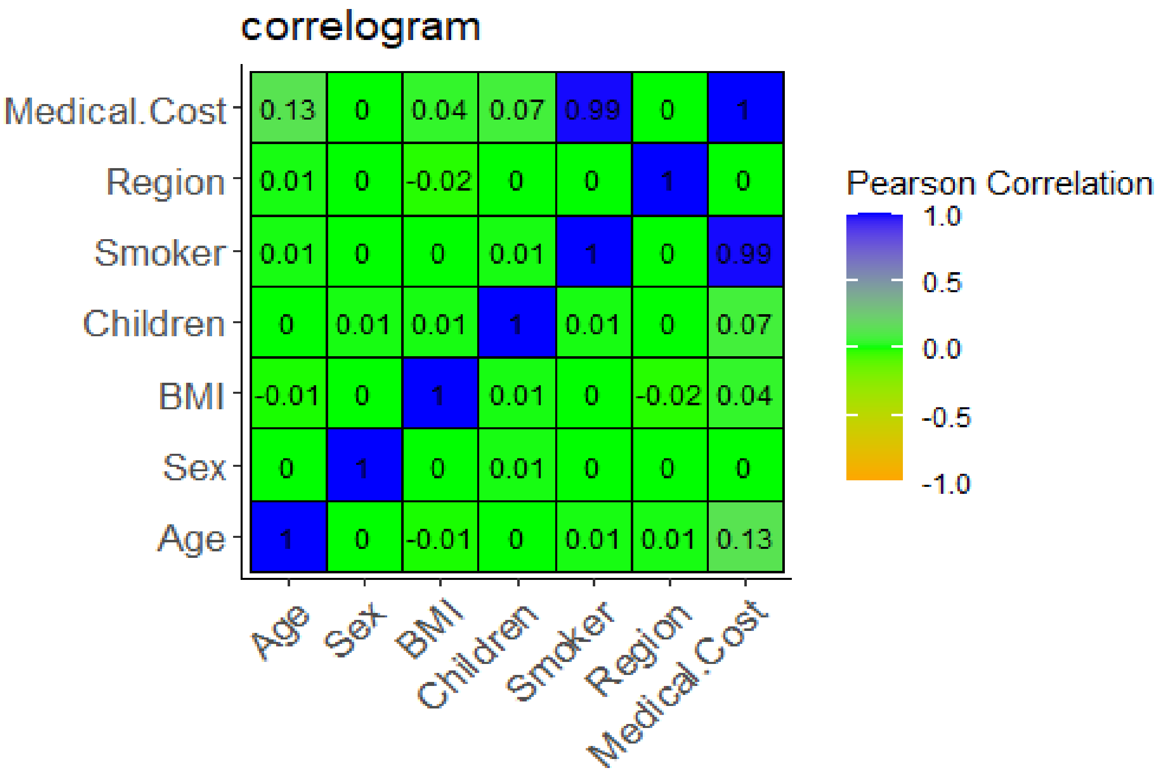
MEDICAL COST PREDICTION MODEL

Enock Bereka

2024-11-27

```
Load the necessary library and the dataset
library(tidyverse)

medical_costs <- read_csv("C:/Users/PC/OneDrive/Desktop/Data
Science/Datasets/medical costs.csv")

Data Manipulation


medical_costs <- rename(medical_costs,
                "Medical.Cost" = `Medical Cost`,)

Setting up Numerics
medical_costs$Sex <- as.factor(medical_costs$Sex)
medical_costs$Sex <- as.numeric(medical_costs$Sex)
medical_costs$Smoker <- as.factor(medical_costs$Smoker)
medical_costs$Smoker <- as.numeric(medical_costs$Smoker)
medical_costs$Region <- as.factor(medical_costs$Region)
medical_costs$Region <- as.numeric(medical_costs$Region)

Checking the ralationship between variables
library(ggcorrplot)
round(cor(medical_costs), 2)

##                 Age  Sex   BMI Children Smoker Region Medical.Cost
## Age            1.00 0.00 -0.01     0.00   0.01   0.01         0.13
## Sex            0.00 1.00  0.00     0.01   0.00   0.00         0.00
## BMI           -0.01 0.00  1.00     0.01   0.00  -0.02         0.04
## Children       0.00 0.01  0.01     1.00   0.01   0.00         0.07
## Smoker         0.01 0.00  0.00     0.01   1.00   0.00         0.99
## Region         0.01 0.00 -0.02     0.00   0.00   1.00         0.00
## Medical.Cost  0.13 0.00  0.04     0.07   0.99   0.00         1.00

cr <- round(cor(medical_costs), 2)

Visualizing our Correlatios
ggcorrplot(cr,title = "correlogram",lab_col = "black",
           lab = TRUE, legend.title = "Pearson Correlation",
           lab_size = 3, ggtheme = theme_classic(),
           outline.color = "black",
           colors = c("orange", "green", "blue"))
```



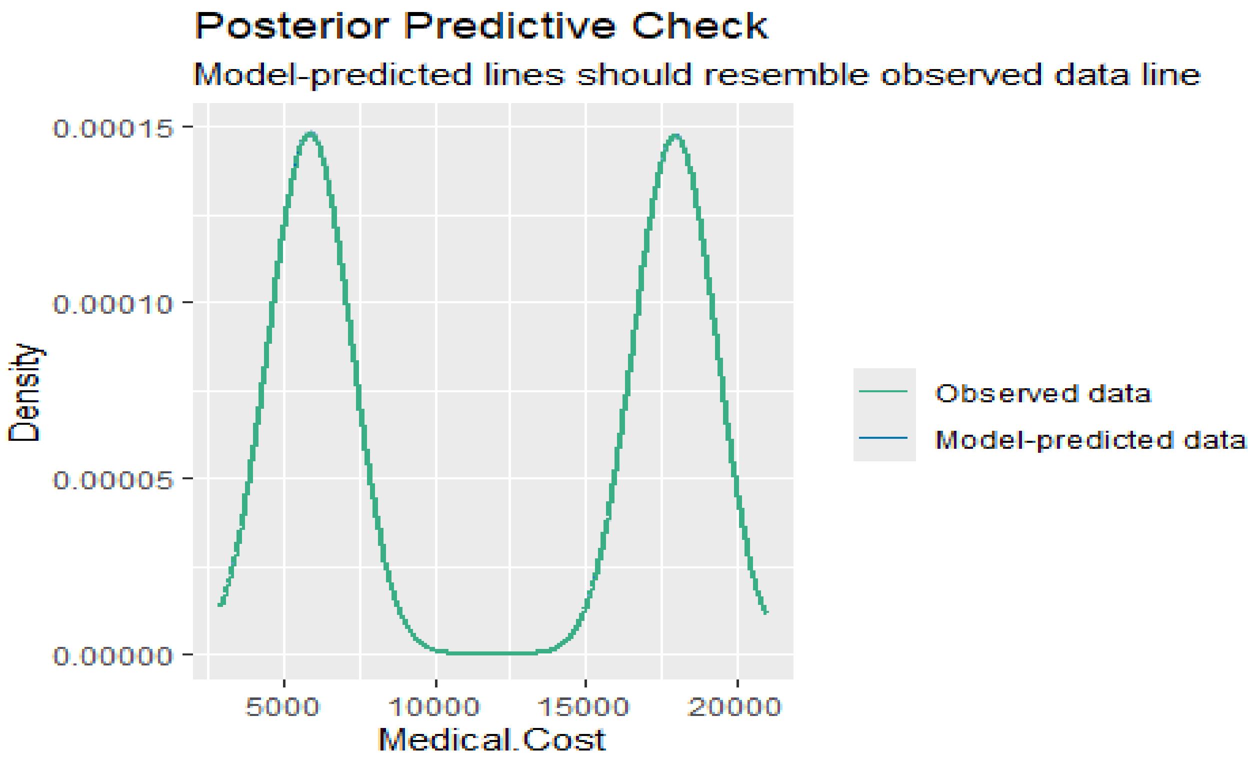
- There is a perfect positive correlation between Medical Cost and Smoker

- There is no correlation between Medical Cost and Region

- There is a weak positive correlation between "Children, BMI, Sex, Age" and Medical Costs.

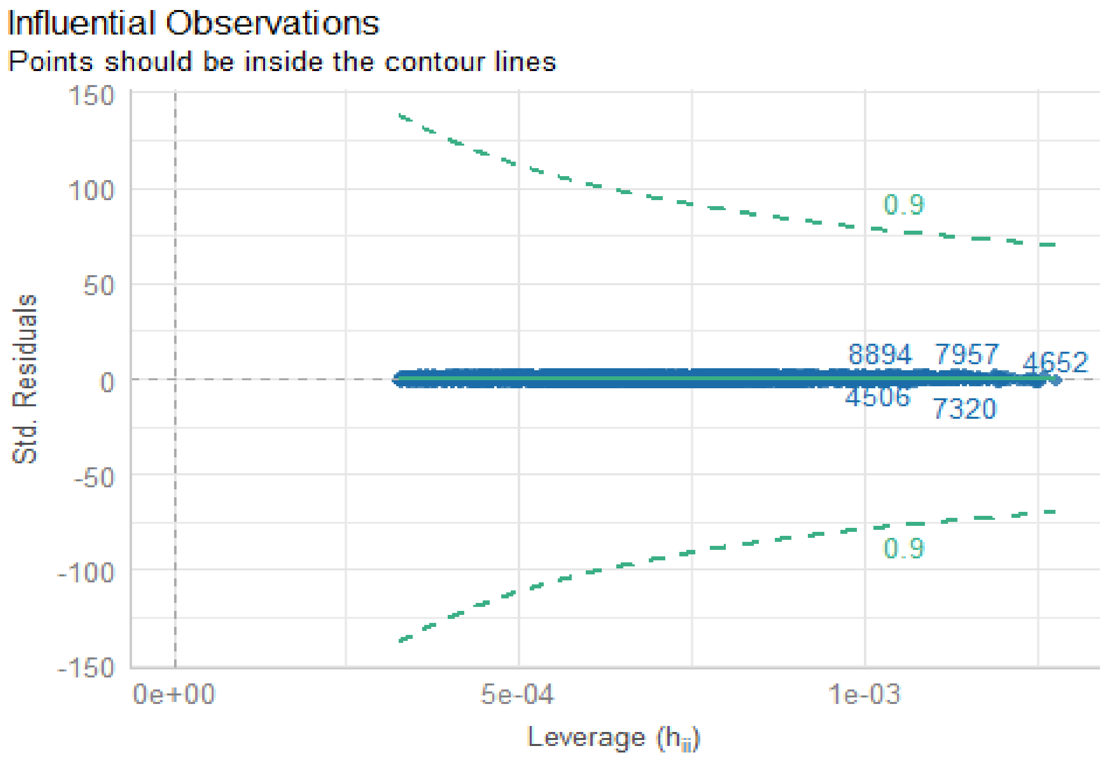
- There is no autocorrelation between predictors due to their low correlation among them.

```
Multivariable Analysis
Build the model
model <- lm(Medical.Cost~Age+Sex+BMI+Children+Smoker+
              Region, data = medical_costs)

Check model assumptions visually
library(performance)

check_predictions(model) %>% plot()
```



- The model predicted lines fits the observed data pretty well

```
check_outliers(model)

## OK: No outliers detected.
## - Based on the following method and threshold: cook (0.907).
## - For variable: (Whole model)

check_outliers(model) %>% plot()
```

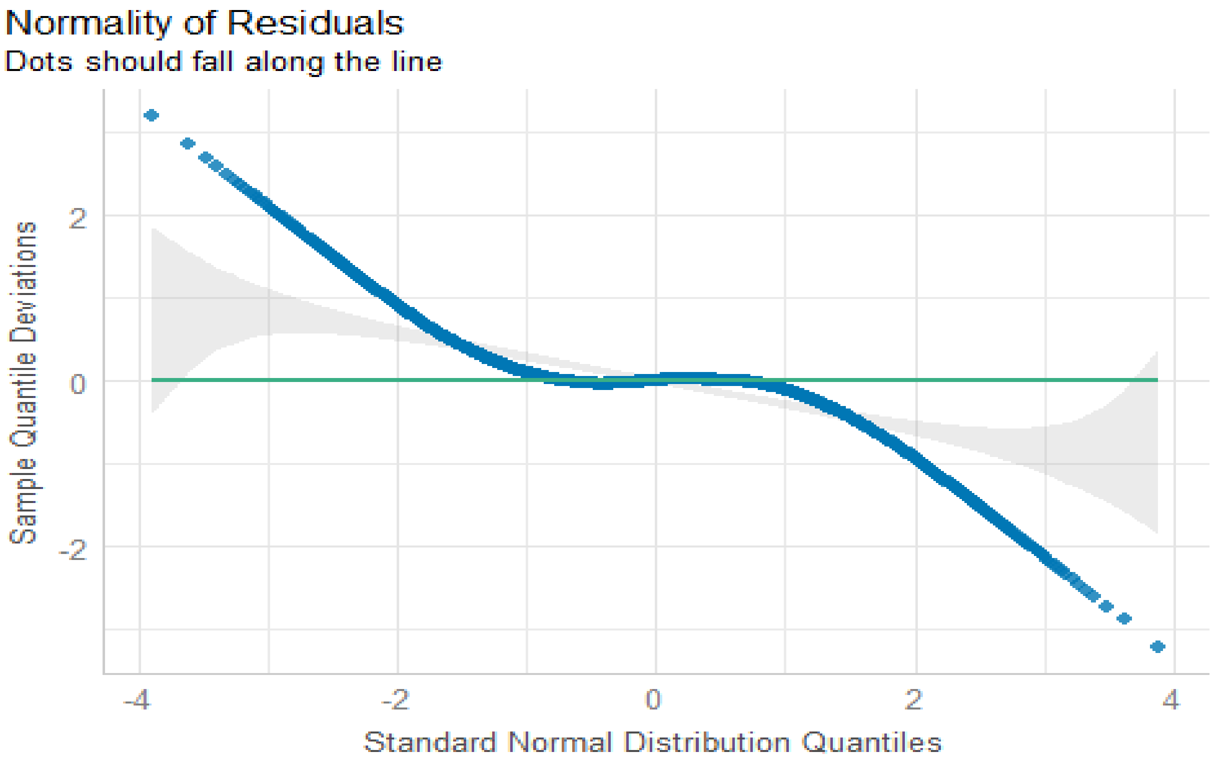


- `No influential outliers detected as all observations are inside the contour lines.`

```
check_autocorrelation(model)

## OK: Residuals appear to be independent and not autocorrelated (p = 0.652).

check_collinearity(model)

## # Check for Multicollinearity
##
## Low Correlation
##
##      Term  VIF       VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##       Age 1.00 [1.00,      Inf]         1.00      1.00     [0.00, 1.00]
##       Sex 1.00 [1.00,      Inf]         1.00      1.00     [0.00, 1.00]
##       BMI 1.00 [1.00, 7.59e+11]         1.00      1.00     [0.00, 1.00]
##  Children 1.00 [1.00, 5.61e+12]         1.00      1.00     [0.00, 1.00]
##    Smoker 1.00 [1.00,      Inf]         1.00      1.00     [0.00, 1.00]
##    Region 1.00 [1.00,      Inf]         1.00      1.00     [0.00, 1.00]
```

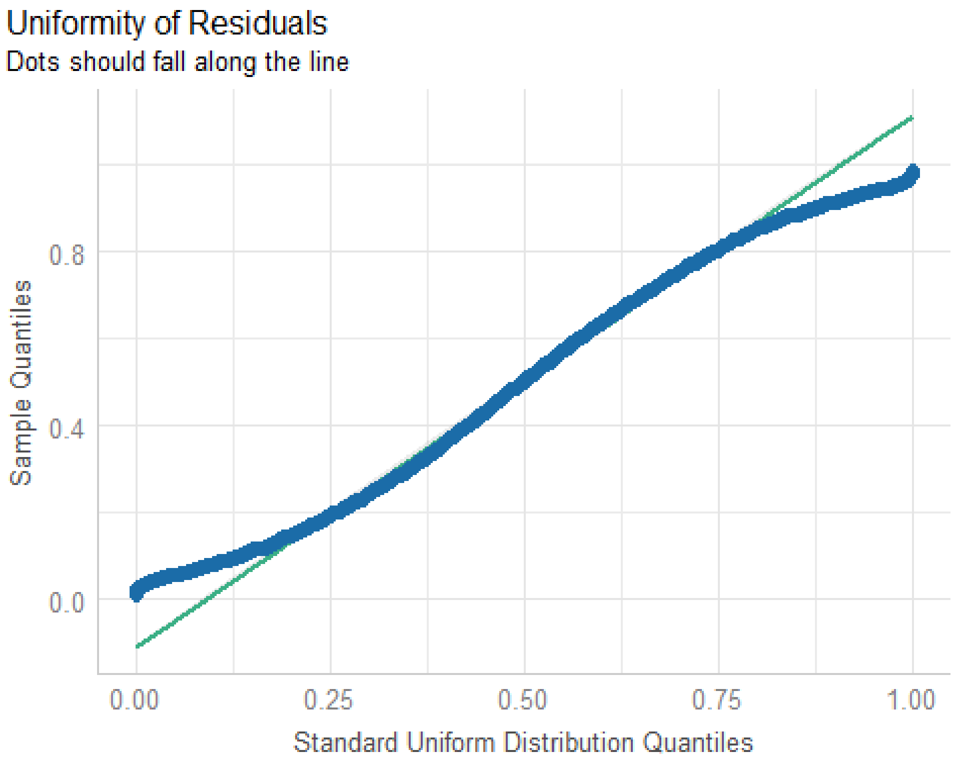
- `No multicollinearity detected as all variables has vif less than 5`

```
check_normality(model)

## Warning: Non-normality of residuals detected (p < .001).

check_normality(model) %>% plot()
```



- The residuals appear normally distributed as they stay inside the confidence intervals of the regression lines.

```
check_residuals(model)

## Ok: Uniformity of residuals is detected (p < .001).

check_residuals(model) %>% plot()
```



```
Get the summary of the model
summary(model)

##
## Call:
## lm(formula = Medical.Cost ~ Age + Sex + BMI + Children + Smoker +
##     Region, data = medical_costs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -522.62 -251.59    1.77  251.82  519.52
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -9472.2159    20.4092 -464.114   <2e-16 ***
## Age            50.1314     0.2100  238.725   <2e-16 ***
## Sex            -4.8282     5.7987   -0.833    0.405
## BMI            29.1122     0.4012   72.569   <2e-16 ***
## Children      202.3424     1.7042  118.733   <2e-16 ***
## Smoker      12001.1090     5.7991 2069.490   <2e-16 ***
## Region         -2.8173     2.5862   -1.089    0.276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 289.9 on 9993 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9977
## F-statistic: 7.299e+05 on 6 and 9993 DF,  p-value: < 2.2e-16
```

- The summary function did not give intuitive results hence I went on to apply emmeans function.

```
Solve most problems with emmeans
library(emmeans)

results <- emmeans(model, pairwise ~ Smoker, infer = T)
results

## $emmeans
##  Smoker emmean  SE   df lower.CL upper.CL  t.ratio p.value
##       1   5908 4.1 9993   5899.9     5916 1442.043  <.0001
##       2  17909 4.1 9993  17901.0    17917 4364.255  <.0001
##
## Results are averaged over the levels of: Sex
## Confidence level used: 0.95
##
## $contrasts
##  contrast          estimate  SE   df lower.CL upper.CL   t.ratio p.value
##  Smoker1 - Smoker2   -12001 5.8 9993   -12012   -11990 -2069.490  <.0001
##
## Results are averaged over the levels of: Sex
## Confidence level used: 0.95

#It provvides estimated averages with 95% CIs and p-values

Visualize model predictions
library(ggeffects)

ggeffect(model)

## $Age
## # Predicted values of Medical.Cost
##
## Age | Predicted |             95% CI
## -------------------------------------
##  15 |  10561.51 | 10549.14, 10573.87
##  20 |  10812.16 | 10801.59, 10822.74
##  30 |  11313.48 | 11306.03, 11320.92
##  35 |  11564.13 | 11557.82, 11570.45
##  40 |  11814.79 | 11809.07, 11820.52
##  45 |  12065.45 | 12059.60, 12071.29
##  50 |  12316.11 | 12309.47, 12322.74
##  65 |  13068.08 | 13056.92, 13079.23

##
## $Sex
## # Predicted values of Medical.Cost
##
## Sex | Predicted |             95% CI
## -------------------------------------
##   1 |  11901.33 | 11893.32, 11909.34
##   2 |  11896.50 | 11888.44, 11904.57
## 
## 
## $BMI
## # Predicted values of Medical.Cost
## 
## BMI | Predicted |            95% CI
## ------------------------------------
##  15 |  11537.85 | 11526.57, 11549.14
##  20 |  11683.41 | 11675.28, 11691.55
##  25 |  11828.98 | 11822.99, 11834.96
##  30 |  11974.54 | 11968.50, 11980.57
##  35 |  12120.10 | 12111.85, 12128.34
##  40 |  12265.66 | 12254.24, 12277.08
## 
## 
## $Children
## # Predicted values of Medical.Cost
## 
## Children | Predicted |            95% CI
## -----------------------------------------
##        0 |  11392.73 | 11382.63, 11402.84
##        1 |  11595.07 | 11587.49, 11602.65
##        2 |  11797.42 | 11791.49, 11803.34
##        3 |  11999.76 | 11993.84, 12005.68
##        4 |  12202.10 | 12194.53, 12209.67
##        5 |  12404.44 | 12394.35, 12414.54
## 
## 
## $Smoker
## # Predicted values of Medical.Cost
## 
## Smoker | Predicted |            95% CI
## ---------------------------------------
##    yes |   5907.98 |  5899.95,  5916.01
##    no  |  17909.09 | 17901.04, 17917.13
## 
## 
## $Region
## # Predicted values of Medical.Cost
## 
## Region    | Predicted |            95% CI
## ---------------------------------------
## northwest |  11903.16 | 11893.66, 11912.66
## northeast |  11900.34 | 11894.12, 11906.57
## southwest |  11897.53 | 11891.30, 11903.75
## southeast |  11894.71 | 11885.22, 11904.20
## 
```

```
ggpredict(model, terms = "Age")
```

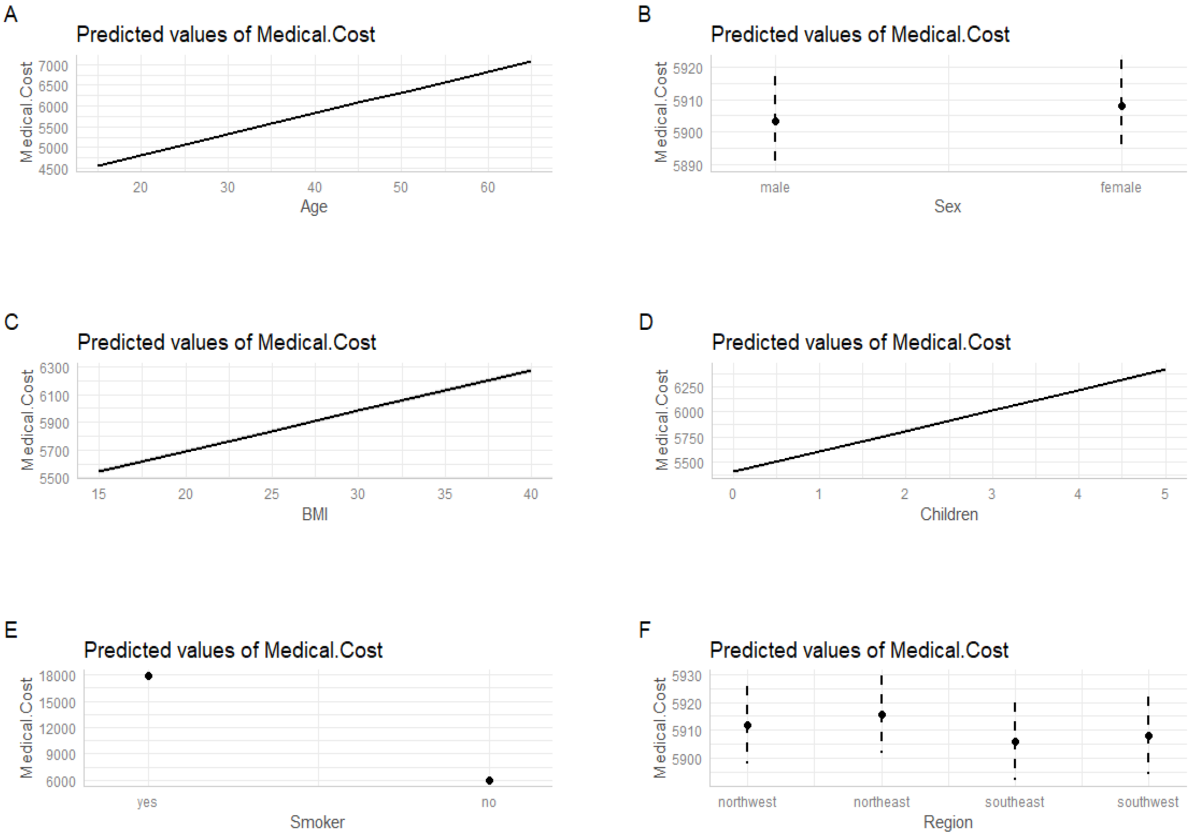
```
## # Predicted values of Medical.Cost
##
## Age | Predicted |             95% CI
## ------------------------------------
##  15 |  10561.51 | 10549.14, 10573.87
##  20 |  10812.16 | 10801.59, 10822.74
##  30 |  11313.48 | 11306.03, 11320.92
##  35 |  11564.13 | 11557.82, 11570.45
##  40 |  11814.79 | 11809.07, 11820.52
##  45 |  12065.45 | 12059.60, 12071.29
##  50 |  12316.11 | 12309.47, 12322.74
##  65 |  13068.08 | 13056.92, 13079.23
##
## Adjusted for:
## *      Sex =  female
## *      BMI = 27.40
## * Children =  2.50
## *   Smoker =  no
## *   Region =  southeast
```

```
library(sjPlot)
```

```
fancy_plot <- ggpredict(model) %>% plot() %>%
  sjPlot::plot_grid()
```

```
fancy_plot
```

```
## TableGrob (3 x 2) "arrange": 6 grobs
##          z     cells    name           grob
## Age      1 (1-1,1-1) arrange gtable[layout]
## Sex      2 (1-1,2-2) arrange gtable[layout]
## BMI      3 (2-2,1-1) arrange gtable[layout]
## Children 4 (2-2,2-2) arrange gtable[layout]
## Smoker   5 (3-3,1-1) arrange gtable[layout]
## Region   6 (3-3,2-2) arrange gtable[layout]
```



- Medical costs increases with increase in BMI.
- Medical costs increases with increase in the number of children.
- A person who smokes will incur higher medical costs compared to the one who do not smoke.
- Northeast region spend more money on medical costs as compared to other regions.
- Southeast region spend the least amount of money on medical costs compared to other regions.

Create a fancy table to check if your predictions are statistically significant

```
library(gtsummary)

fancy_table <- tbl_regression(
  model,
  add_pairwise_contrasts = T,
  pvalue_fun = ~style_pvalue(.x, digits = 3)) %>%
  bold_p()
fancy_table
```

Predictors	Beta	95% CI[1]	p-value
Age	50	50, 51	**<0.001**
Sex			
male - female	4.8	-16, 6.5	0.403
BMI	29	28, 30	**<0.001**
Children	202	199, 206	**<0.001**
Smoker			
no - yes	12,001	11,990, 12,013	**<0.001**
Region			
northwest - northeast	-3.9	-25, 17	0.965
southeast - northeast	-9.7	-31, 11	0.639
southeast - northwest	-5.8	-27, 15	0.895
southwest - northeast	-7.5	-28, 14	0.797

Predictors	Beta	95% CI[1]	p-value
southwest - northwest	-3.6	-25, 17	0.972
southwest - southeast	-2.2	-19, 23	0.993

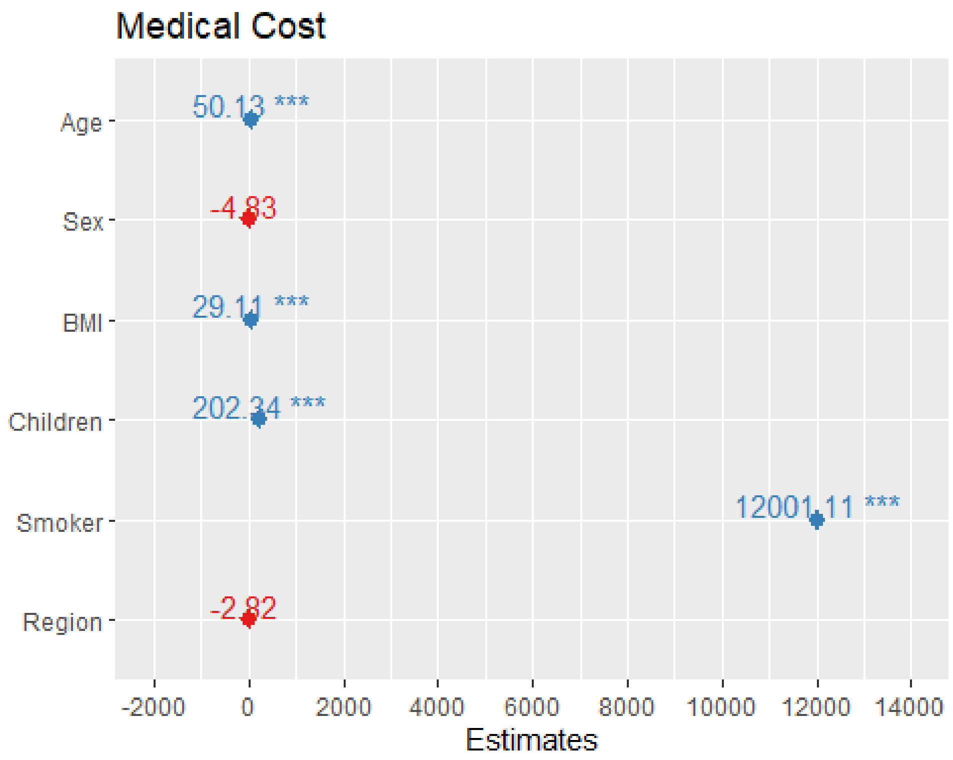
- A fancy table displays all pairwise comparisons for categorical predictors instead of only comparing everything to the reference category.
- We interpret the slopes Beta as the average change in Medical.Cost for a one unit change in any of the predictors while holding all other predictors fixed.
- Increasing the numeric predictor by one unit changes the estimated outcome by its better coefficient.
- For example in our model, an increase in age for one year, increases the medical cost by 50. This increase is statistically significant.
- An increase in BMI by one unit increases the medical cost by 29. This increase is statistically significant.
- An increase in the number of children by one increases the medical cost by 202. This increase is statistically significant.
- In a categorical predictors, each category is treated as a separate binary predictor a technique commonly known as “one-hot encoding”.
- For example if we changes our region from northwest to northeast, we shall have to use an additional of 3.9 on medical cost provided other predictors are held fixed.
- If we shifts our gender from male to female our medical costs increases by 4.8.
- If we shifts from normal to being smoker, our medical costs increases by 12001. The increase is statistically significant.

```
Get publication ready table
tab_model(model,
          show.reflvl = T,
          show.intercept = F,
          p.style = "numeric_stars")
```

	Medical.Cost		
Predictors	*Estimates*	*CI*	*p*
Age	50.13 ***	49.72 – 50.54	**<0.001**
BMI	29.11 ***	28.33 – 29.90	**<0.001**
Children	202.34 ***	199.00 – 205.68	**<0.001**
Region	-2.82	-7.89 – 2.25	0.276
Sex	-4.83	-16.19 – 6.54	0.405
Smoker	12001.11 ***	11989.74 – 12012.48	**<0.001**
Observations	10000		
R^2 / R^2 adjusted	0.998 / 0.998		

- The tab_model function generates a well formatted , publication ready table that not only presents easily interpretable p-values instead of cumbersome scientific notations but also reveals the 95% Cis instead of SE.

```
Visualize estimates
plot_model(model, show.values = TRUE, width = 0.2)
```



- This approach displays the model estimates along with their 95% Cis. And even include the significance stars that we need.

```
Produce the model equation
library(equatiomatic)

extract_eq(model)
```

$$\begin{aligned}\text{Medical. Cost} \\ = \alpha + \beta_1(\text{Age}) + \beta_2(\text{Sex}) + \beta_3(\text{BMI}) + \beta_4(\text{Children}) + \beta_5(\text{Smoker}) + \beta_6(\text{Region}) + \epsilon\end{aligned}$$

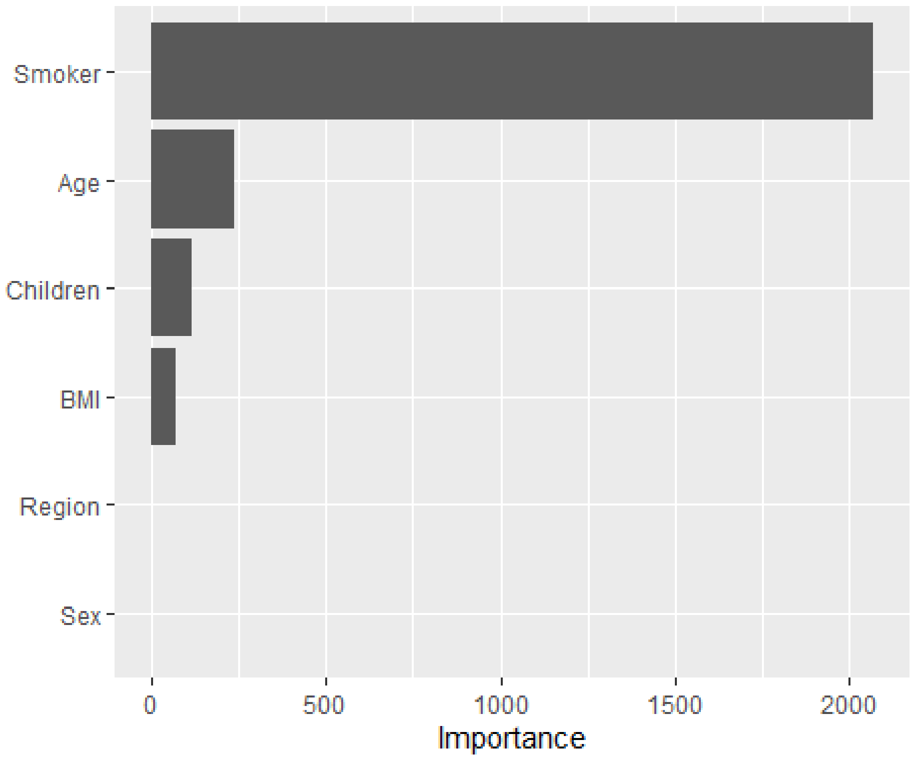
```
Checking variable importance
library(effectsize)

eff_size <- eta_squared(model) %>%
  mutate(Interpret = interpret_eta_squared(Eta2_partial))
eff_size

## # Effect Size for ANOVA
##
## Parameter | Eta2 (partial) |       95% CI |  Interpret
## -----------------------------------------------------
## Age       |           0.87 | [0.87, 1.00] |      large
## Sex       |       2.94e-03 | [0.00, 1.00] | very small
## BMI       |           0.36 | [0.35, 1.00] |      large
## Children  |           0.68 | [0.67, 1.00] |      large
## Smoker    |           1.00 | [1.00, 1.00] |      large
## Region    |       1.19e-04 | [0.00, 1.00] | very small
```

```
library(randomForest)

library(vip)
```

- Here we select the variable with the highest Eta2 as the top predictor in our model.
- Smoker is the top predictor of medical cost since it has the highest Eta2
- Sex is the least predictor of medical cost since it has the lowest Eta2

```
randomForest(Medical.Cost~Age+BMI+Smoker+Children+Sex+Region,
             medical_costs) %>%
  vip()
```

```
t-values as importance
vip(model)
```



- Smoker is the top predictor of medical costs while sex is the list predictor of medical costs as per our study

```


How well our model fits the data
performance(model)
```

```
## # Indices of model performance
##
## AIC       |      AICc |       BIC |    R2 | R2 (adj.) |    RMSE |   Sigma
## ---------------------------------------------------------------------------
## 1.418e+05 | 1.418e+05 | 1.418e+05 | 0.998 |     0.998 | 289.802 | 289.904
```

- R-square tells you how much of the total variance in your outcome that the model explains
- AIC is the amount of information our model lost, therefore the lower the AIC, the better the fit because less information is lost due model complexity
- Our model explains 99.8% of the variance in the medical cost making it an excellent model in making predictions.

```
How to interpret R-squared value
interpret_r2(0.998, rules = "cohen1988")

## [1] "substantial"
## (Rules: cohen1988)

?interpret_r2

Report model results
library(report)

report(model)

## We fitted a linear model (estimated using OLS) to predict Medical.Cost
with
## Age, Sex, BMI, Children, Smoker and Region (formula: Medical.Cost ~ Age +
Sex +
## BMI + Children + Smoker + Region). The model explains a statistically
## significant and substantial proportion of variance (R2 = 0.998, F(6, 9993)
=
## 7.30e+05, p < .001, adj. R2 = 0.998). The model's intercept, corresponding
to
## Age = 0, Sex = 0, BMI = 0, Children = 0, Smoker = 0 and Region = 0, is at
## -9472.22 (95% CI [-9512.22, -9432.21], t(9993) = -464.11, p < .001).
Within
## this model:
##
##   - The effect of Age is statistically significant and positive (beta =
50.13,
## 95% CI [49.72, 50.54], t(9993) = 238.72, p < .001; Std. beta = 0.11, 95%
CI
## [0.11, 0.11])
##   - The effect of Sex is statistically non-significant and negative (beta
=
## -4.83, 95% CI [-16.19, 6.54], t(9993) = -0.83, p = 0.405; Std. beta =
## -3.97e-04, 95% CI [-1.33e-03, 5.38e-04])
##   - The effect of BMI is statistically significant and positive (beta =
29.11,
## 95% CI [28.33, 29.90], t(9993) = 72.57, p < .001; Std. beta = 0.03, 95% CI
## [0.03, 0.04])
##   - The effect of Children is statistically significant and positive (beta
=
## 202.34, 95% CI [199.00, 205.68], t(9993) = 118.73, p < .001; Std. beta =
0.06,
## 95% CI [0.06, 0.06])
##   - The effect of Smoker is statistically significant and positive (beta =
## 12001.11, 95% CI [11989.74, 12012.48], t(9993) = 2069.49, p < .001; Std.
beta =
## 0.99, 95% CI [0.99, 0.99])
##   - The effect of Region is statistically non-significant and negative
(beta =
## -2.82, 95% CI [-7.89, 2.25], t(9993) = -1.09, p = 0.276; Std. beta = -
5.20e-04,
## 95% CI [-1.46e-03, 4.16e-04])
##
## Standardized parameters were obtained by fitting the model on a
standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```