INF1344 Final

November 25, 2022

This is the analysis for Group 6, INF1344 Fall 2022.

[4]:
```
install.packages("car")
require("dplyr")
library(car)
#installing required packages
```

```
Installing package into ‘/opt/r’
(as ‘lib’ is unspecified)

also installing the dependencies ‘carData’, ‘nnet’, ‘pbkrtest’


Loading required package: dplyr


Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


Loading required package: carData


Attaching package: ‘car’


The following object is masked from ‘package:dplyr’:

    recode
```

[5]:
```
MCPD <- read.csv('insurance.csv', header =TRUE, sep = ',')
#reading the dataset
#dataset named MCPD, which stands for Medical Cost Personal Dataset
#noticing that there are 7 variables in this dataset, we will categorize them␣
 ↪first
#native categorical variables: sex, smoker, region
#native numerical variables: age, bmi, children, charge
#for the sake of this project, I am conducting a preliminary exploratory data␣
 ↪analysis
#the results here will be used to decide our resaerch question
#these comments ought to be remained internally within the group
```

[6]:
```
MCPD
#displaying our dataset
```

A data.frame: 1338 × 7

age <int>	sex <chr>	bmi <dbl>	children <int>	smoker <chr>	region <chr>	charges <dbl>
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622
46	female	33.440	1	no	southeast	8240.590
37	female	27.740	3	no	northwest	7281.506
37	male	29.830	2	no	northeast	6406.411
60	female	25.840	0	no	northwest	28923.137
25	male	26.220	0	no	northeast	2721.321
62	female	26.290	0	yes	southeast	27808.725
23	male	34.400	0	no	southwest	1826.843
56	female	39.820	0	no	southeast	11090.718
27	male	42.130	0	yes	southeast	39611.758
19	male	24.600	1	no	southwest	1837.237
52	female	30.780	1	no	northeast	10797.336
23	male	23.845	0	no	northeast	2395.172
56	male	40.300	0	no	southwest	10602.385
30	male	35.300	0	yes	southwest	36837.467
60	female	36.005	0	no	northeast	13228.847
30	female	32.400	1	no	southwest	4149.736
18	male	34.100	0	no	southeast	1137.011
34	female	31.920	1	yes	northeast	37701.877
37	male	28.025	2	no	northwest	6203.902
59	female	27.720	3	no	southeast	14001.134
63	female	23.085	0	no	northeast	14451.835
55	female	32.775	2	no	northwest	12268.632
23	male	17.385	1	no	northwest	2775.192
31	male	36.300	2	yes	southwest	38711.000
25	female	30.200	0	yes	southwest	33900.653
41	male	32.200	2	no	southwest	6875.961
42	male	26.315	1	no	northwest	6940.910
33	female	26.695	0	no	northwest	4571.413
34	male	42.900	1	no	southwest	4536.259
19	female	34.700	2	yes	southwest	36397.576
30	female	23.655	3	yes	northwest	18765.875
18	male	28.310	1	no	northeast	11272.331
19	female	20.600	0	no	southwest	1731.677
18	male	53.130	0	no	southeast	1163.463
35	male	39.710	4	no	northeast	19496.719
39	female	26.315	2	no	northwest	7201.701
31	male	31.065	3	no	northwest	5425.023
62	male	26.695	0	yes	northeast	28101.333
62	male	38.830	0	no	southeast	12981.346
42	female	40.370	2	yes	southeast	43896.376
31	male	25.935	1	no	northwest	4239.893
61	male	33.535	0	no	northeast	13143.337
42	female	32.870	0	no	northeast	7050.021
51	male	30.030	1	no	southeast	9377.905

1 Research Question

Is there any relationship between the primary beneficiary's age, and their individual medical costs billed by health insurance?

2 Research Hypothesis

Individual medical costs billed by health insurance is different for patients with younger age and older age, and sex has a moderating effect on the difference between the two groups.

3 Exploring the relationship between age and charge, with sex as a controlling varible

3.1 Variable Descriptive Statistics

3.1.1 Age

[7]:
```
range(MCPD$age)
#checking the maximum and minimum value for "age"
```

1. 18 2. 64

[8]:
```
length(MCPD$age)
#checking the count of "age"
```

1338

[9]:
```
median(MCPD$age)
#checking the median of "age"
```

39

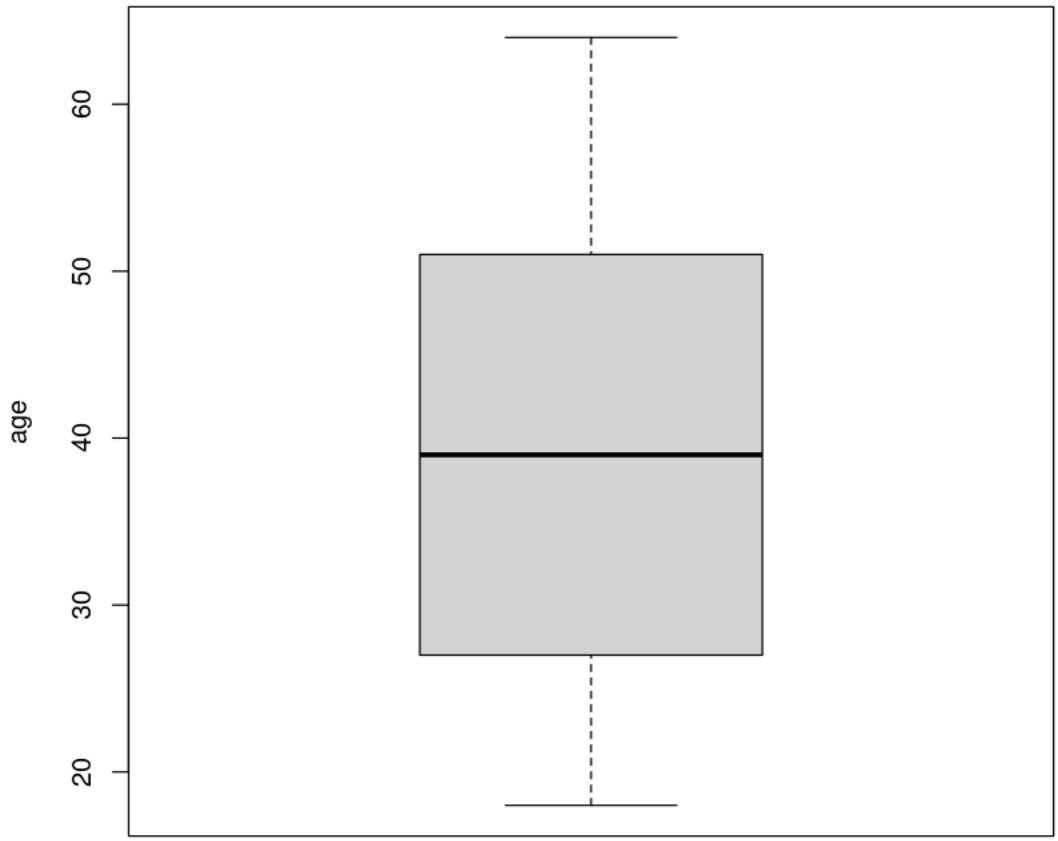
[10]:
```
mean(MCPD$age)
#checking the mean of "age"
```

39.2070254110613

[11]:
```
unique(MCPD$age)
#checking all the unique values of "age"
```

1. 19 2. 18 3. 28 4. 33 5. 32 6. 31 7. 46 8. 37 9. 60 10. 25 11. 62 12. 23 13. 56 14. 27 15. 52 16. 30 17. 34 18. 59 19. 63 20. 55 21. 22 22. 26 23. 35 24. 24 25. 41 26. 38 27. 36 28. 21 29. 48 30. 40 31. 58 32. 53 33. 43 34. 64 35. 20 36. 61 37. 44 38. 57 39. 29 40. 45 41. 54 42. 49 43. 47 44. 51 45. 42 46. 50 47. 39

[12]:
```
boxplot(MCPD$age,
        ylab = "age"
)
#checking if there are any outliers in "age"
```



Looks like there are no outliers.

3.1.2 Sex

[13]:
```
length(MCPD$sex)
#checking the count of "sex"
```

1338

[14]:
```
unique(MCPD$sex)
#checking all the unique values of "sex"
```

1. 'female' 2. 'male'

3.1.3 Charges

[15]:
```
range(MCPD$charges)
#checking the maximum and minimum value for "charges"
```

1. 1121.8739 2. 63770.42801

[16]:
```
length(MCPD$charges)
#checking the count of "charges"
```

1338

Noticing that the lenghths match for all three variables.

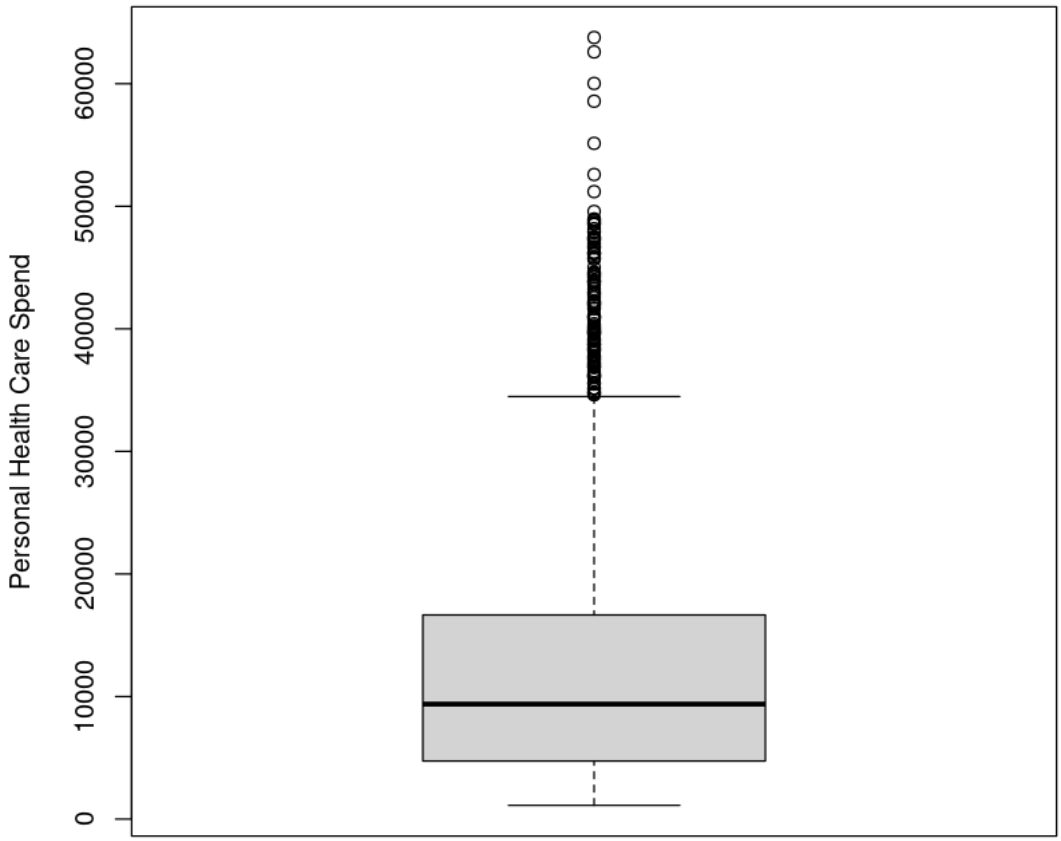
[17]:
```
median(MCPD$charges)
#checking the median of "charges"
```

9382.033

[18]:
```
mean(MCPD$charges)
#checking the mean of "charges"
```

13270.4222651413

[19]:
```
boxplot(MCPD$charges,
        ylab = "Personal Health Care Spend"
)
#checking if there are any outliers in "charges"
```



Data looks significantly skewed to the right. Thus cannot eliminate anyone as outliers (for now).

All three variables are recorded in standard manner, with no missing value or significant outlier. Therefore, we proceed with our analysis

3.2 Correlation Analysis

Correlation measures how strong two quantitative variables bivariate. Becasue we are investigating the relationship between age and personal health care cost, it is reasonable to for us to see if these two variable bivariate.

Ho: there is no lineaer correlation between patients' age and personal health care charges

Ha: there exists a lineaer correlation between patients' age and personal health care charges

[20]:
```
AgeChargescor <- cor.test(MCPD$age, MCPD$charges, method = "pearson")
AgeChargescor
```

```
        Pearson's product-moment correlation

data:  MCPD$age and MCPD$charges
t = 11.453, df = 1336, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2494139 0.3470381
sample estimates:
      cor
0.2990082
```

Because the p-value is smaller than 0.001, the correlation is significant. There exists a 0.2990082 correlation between age and personal health care charge.

3.3 Multiple Regression Analysis

Ho: age cannot be used to predict personal health care charge

Ha: age can be used to predict personal health care charge

First, we need to convert sex into a factor. (reason should be explained in the actual write-up)

[92]:
```
MCPD$sex.f <- as.factor(MCPD$sex)
```

"sex.f" is a variable with a factor class now, and we will verify it.

[93]:
```
print(class(MCPD$sex.f))
```

```
[1] "factor"
```

Now we run a multiple regression, with the dependent variable (DV) being "charge", independent variable (IV) being "age", and control variable (CV) being "sex"

[94]:
```
lmACS <- lm(charges ~ age + sex.f, data = MCPD)
print(summary(lmACS))
#lmACS stands for linear model for the relationship between Age, Charges, and␣
 ↪Sex.
#Great (or not great), age is significantly positively correlated with medical␣
 ↪charges; however, sex is a significant control variable.
```

```
Call:
lm(formula = charges ~ age + sex.f, data = MCPD)

Residuals:
   Min     1Q Median     3Q    Max
```

```
 -8821  -6947  -5511   5443  48203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2343.62     994.35   2.357   0.0186 *
age           258.87      22.47  11.523   <2e-16 ***
sex.fmale    1538.83     631.08   2.438   0.0149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11540 on 1335 degrees of freedom
Multiple R-squared:  0.09344,   Adjusted R-squared:  0.09209
F-statistic:  68.8 on 2 and 1335 DF,  p-value: < 2.2e-16
```

Both variables are statistically significant. For detailed interpretation, please refer to course slides.

3.4 T-Test

3.4.1 T-test for age and charges

Ho: age low = age high

Ha: age low age high

We have checked the mean and median of "age" for this sample, they are 39.2073 and 39, respectively; thus it makes sense if we divide "age" into 2 groups, the cutoff point being 39.

[95]:
```
ageL <- MCPD$charges[MCPD$age <= 39]
ageH <- MCPD$charges[MCPD$age > 39]
t.test(ageL, ageH)
```

```
        Welch Two Sample t-test

data:  ageL and ageH
t = -9.8047, df = 1335.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7528.464 -5018.126
sample estimates:
mean of x mean of y
 10157.22  16430.51
```

Awesome, the cost of health care spend is significantly different for younger and older samples!

3.4.2 T-test for sex and charges

Ho: male = female

Ha: male female

[96]:
```
chargesM <- MCPD$charges[MCPD$sex == "male"]
chargesF <- MCPD$charges[MCPD$sex == "female"]
t.test(chargesM, chargesF)
```

```
        Welch Two Sample t-test

data:  chargesM and chargesF
t = 2.1009, df = 1313.4, p-value = 0.03584
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   91.85535 2682.48932
sample estimates:
mean of x mean of y
 13956.75  12569.58
```

Significant, but to a less degree, suggesting it being a good controlling variable.

3.5 Two-Way ANOVA

3.5.1 Two-way ANOVA test

Ho: the mean personal health care charge is the same across all the groups

Ha: at least one group has a different mean personal health care charge than the other groups

First, we need to create a new categorical variable where age is categorized as "high" or "low" based on their values (i.e. high if the value is >39, low if value is <= 39)

[97]:
```
MCPD$age.c <- ifelse(MCPD$age>39, "high", "low")
View(MCPD)
```

age <int>	sex <chr>	bmi <dbl>	children <int>	smoker <chr>	region <chr>	charges <dbl>	sex.f <fct>	age.c <chr>
19	female	27.900	0	yes	southwest	16884.924	female	low
18	male	33.770	1	no	southeast	1725.552	male	low
28	male	33.000	3	no	southeast	4449.462	male	low
33	male	22.705	0	no	northwest	21984.471	male	low
32	male	28.880	0	no	northwest	3866.855	male	low
31	female	25.740	0	no	southeast	3756.622	female	low
46	female	33.440	1	no	southeast	8240.590	female	high
37	female	27.740	3	no	northwest	7281.506	female	low
37	male	29.830	2	no	northeast	6406.411	male	low
60	female	25.840	0	no	northwest	28923.137	female	high
25	male	26.220	0	no	northeast	2721.321	male	low
62	female	26.290	0	yes	southeast	27808.725	female	high
23	male	34.400	0	no	southwest	1826.843	male	low
56	female	39.820	0	no	southeast	11090.718	female	high
27	male	42.130	0	yes	southeast	39611.758	male	low
19	male	24.600	1	no	southwest	1837.237	male	low
52	female	30.780	1	no	northeast	10797.336	female	high
23	male	23.845	0	no	northeast	2395.172	male	low
56	male	40.300	0	no	southwest	10602.385	male	high
30	male	35.300	0	yes	southwest	36837.467	male	low
60	female	36.005	0	no	northeast	13228.847	female	high
30	female	32.400	1	no	southwest	4149.736	female	low
18	male	34.100	0	no	southeast	1137.011	male	low
34	female	31.920	1	yes	northeast	37701.877	female	low
37	male	28.025	2	no	northwest	6203.902	male	low
59	female	27.720	3	no	southeast	14001.134	female	high
63	female	23.085	0	no	northeast	14451.835	female	high
55	female	32.775	2	no	northwest	12268.632	female	high
23	male	17.385	1	no	northwest	2775.192	male	low
31	male	36.300	2	yes	southwest	38711.000	male	low
25	female	30.200	0	yes	southwest	33900.653	female	low
41	male	32.200	2	no	southwest	6875.961	male	high
42	male	26.315	1	no	northwest	6940.910	male	high
33	female	26.695	0	no	northwest	4571.413	female	low
34	male	42.900	1	no	southwest	4536.259	male	low
19	female	34.700	2	yes	southwest	36397.576	female	low
30	female	23.655	3	yes	northwest	18765.875	female	low
18	male	28.310	1	no	northeast	11272.331	male	low
19	female	20.600	0	no	southwest	1731.677	female	low
18	male	53.130	0	no	southeast	1163.463	male	low
35	male	39.710	4	no	northeast	19496.719	male	low
39	female	26.315	2	no	northwest	7201.701	female	low
31	male	31.065	3	no	northwest	5425.023	male	low
62	male	26.695	0	yes	northeast	28101.333	male	high
62	male	38.830	0	no	southeast	12981.346	male	high
42	female	40.370	2	yes	southeast	43896.376	female	high
31	male	25.935	1	no	northwest	4239.893	male	low
61	male	33.535	0	no	northeast	13143.337	male	high
42	female	32.870	0	no	northeast	7050.021	female	high
51	male	30.030	1	no	southeast	9377.905	male	high

A data.frame: 1338 × 9

Now we have both the IV and CV categorical variables, we can start our two-way ANOVA. Because we have 2 variables in interest, with 2 groups within each variable; we have 4 groups in total (a 2*2 design). These groups are "male * young", "male * old", "female * young", and "female * old". We will see how many samples there are for each group.

[98]:
```
sum(MCPD$sex.f == 'male' & MCPD$age.c == "low")
```

346

[99]:
```
sum(MCPD$sex.f == 'male' & MCPD$age.c == "high")
```

330

[100]:
```
sum(MCPD$sex.f == 'female' & MCPD$age.c == "low")
```

328

[101]:
```
sum(MCPD$sex.f == 'female' & MCPD$age.c == "high")
```

334

346 + 330 + 328 + 334 = 1338. We have all the samples included. This is an unbalanced design (different sample size per group), we will run ANOVA with a Type III sum of square (details should be included in the acutal write-up).

[102]:
```
anovaACS <- aov(charges ~ age.c * sex, data = MCPD)
Anova(anovaACS, type = "III")
print(summary(anovaACS))
#anovaACS stands for ANOVA of "age", "charges", and "sex".
```

A anova: 5 × 4

	Sum Sq <dbl>	Df <dbl>	F value <dbl>	Pr(>F) <dbl>
(Intercept)	83352874778	1	610.4361226	2.777815e-111
age.c	7023663006	1	51.4379093	1.221071e-12
sex	269333544	1	1.9724686	1.604189e-01
age.c:sex	15443197	1	0.1130985	7.366966e-01
Residuals	182152940873	1334	NA	NA

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
age.c         1 1.316e+10 1.316e+10  96.401 <2e-16 ***
sex           1 7.426e+08 7.426e+08   5.438 0.0198 *
age.c:sex     1 1.544e+07 1.544e+07   0.113 0.7367
Residuals  1334 1.822e+11 1.365e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

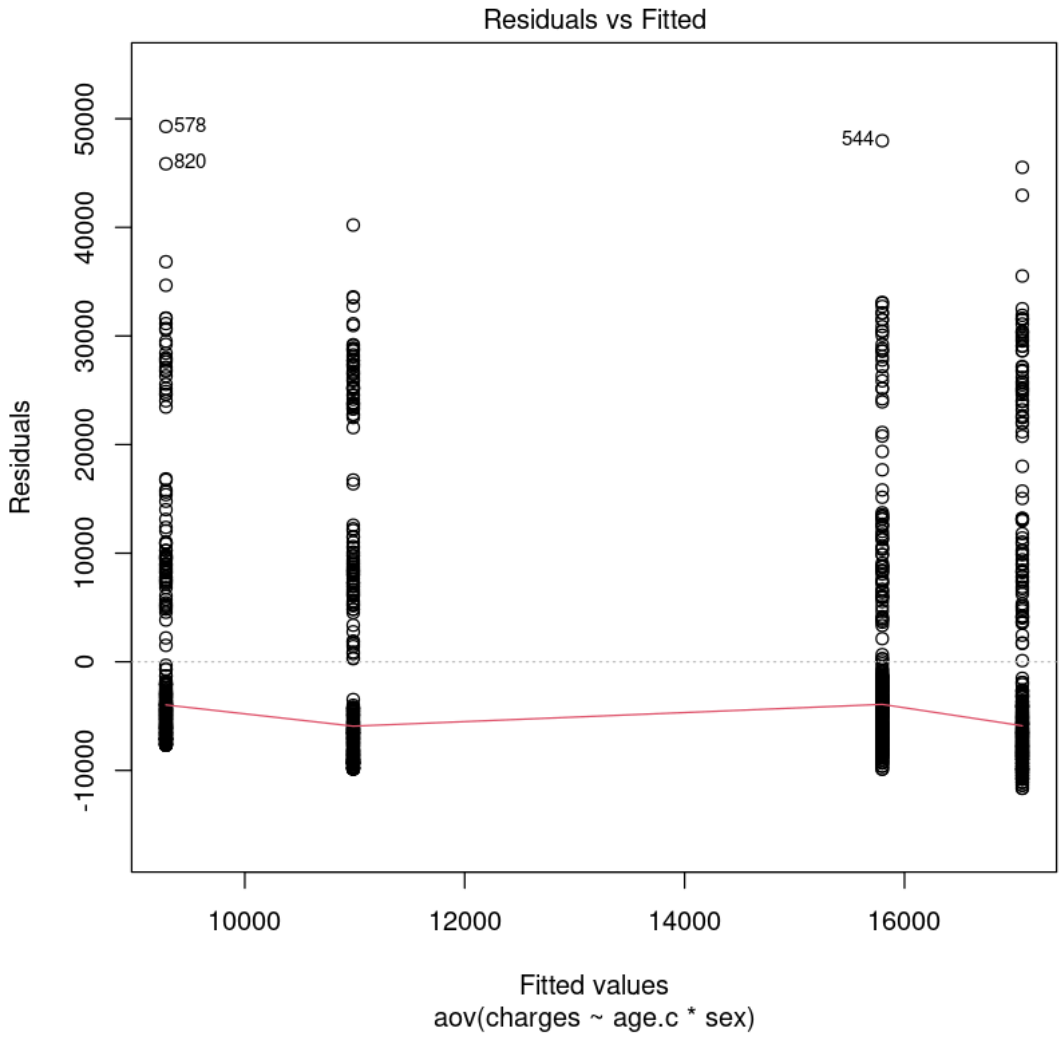
Both significant, and interpreted same as a regular ANOVA F value. Interaction is not, the interpretation is the following: "The p-value for the interaction between age.c and sex is 0.7367 (insignificant), which indicates that the relationships between age and health care charges does not depend on the patient's sex."

3.5.2 Assumptions of two-way ANOVA test

Two-way ANOVA test assumes that the observations within each cell are normally distributed and have equal variances. We will test the assumptions below. Assuming the independence of observation is true.

Checking the homogeneity of variance assumption

[103]:
```
plot(anovaACS, 1)
#Plotting the residuals versus fits plot to check the homogeneity of variance.
```



In the residuals versus fits plot, there is no evident relationships between residuals and fitted values (the mean of each groups), which is good. So, we can assume the homogeneity of variances. (because the range of y does not change significantly when x changes).

[104]:
```
leveneTest(charges ~ sex.f*age.c, data = MCPD)
#run a Levene's test to check the homogeneity of variances.
```

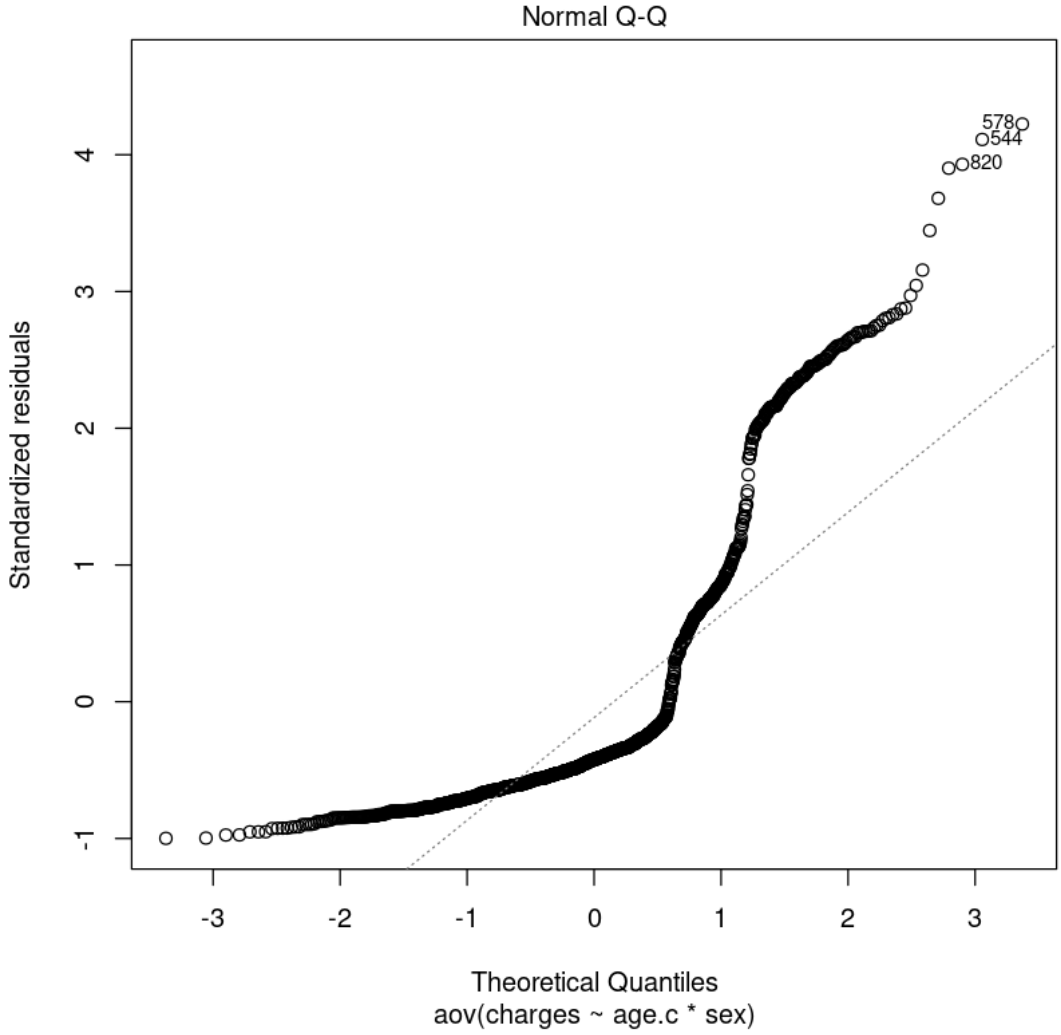
A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	3	3.976132	0.007806367
	1334	NA	NA

Ouch! Becuase the p-value of Levene's test is significant, suggesting that the assumption of equal variance is violated. However, we will keep using two-way ANOVA because firstly, Levene's test can be too sensitive when sample size gets larger; secondly, the groups sizes are almost equal, making the two-way ANOVA robust to the violation of this assumption; thirdly, equal variance is assumed from the residuals versus fits plot; lastly, we can use post-hoc test to verify if the group means are heterogeneious with Tukey's HSD tests. (I would like to have something like a Welch's ANOVA for our study, but it does not exist for a two-way ANOVA. We can find references supporting the above mentioned reasons.)

Checking the normality assumpttion

[105]:
```
plot(anovaACS, 2)
#plotting the normality plot of the residuals
```



Noticed that the residual quantile does not follow a straight line, the assumption is violated.

[106]:
```
anovaACS_residuals <- residuals(object = anovaACS)
shapiro.test(x = anovaACS_residuals)
#extract the residuals and run Shapiro's test
```

```
        Shapiro-Wilk normality test

data:  anovaACS_residuals
W = 0.74859, p-value < 2.2e-16
```

Because the p-value is significant, the assumption of normality is violated. Sad, very sad.

We will still use two-way ANOVA when this assumption is violated. The reason is quoted here "The assumption of normality is necessary for statistical significance testing using a two-way ANOVA. However, the two-way ANOVA is considered "robust" to violations of normality. This means that some violation of this assumption can be tolerated and the test will still provide valid results. Therefore, you will often hear of this test only requiring approximately normally distributed data. Furthermore, as sample size increases, the distribution can be quite non-normal and, thanks to the Central Limit Theorem, the two-way ANOVA can still provide valid results. " Retreived from https://www.amstatisticalconsulting.com/banking-fees-2-4/. We will need to find a more legit reference for the write-up.

3.5.3 Post-hoc analysis

Summary statistics for all four groups

[107]:
```
group_by(MCPD, sex.f, age.c) %>%
  summarise(
    count = n(),
    mean = mean(charges, na.rm = TRUE),
    sd = sd(charges, na.rm = TRUE)
)
#This will give us the means and standard diviations for all four gourps.
```

```
`summarise()` has grouped output by 'sex.f'. You can override using the
`.groups` argument.
```

A grouped_df: 4 × 5

sex.f <fct>	age.c <chr>	count <int>	mean <dbl>	sd <dbl>
female	high	334	15797.452	10640.81
female	low	328	9282.659	10656.64
male	high	330	17071.246	12767.76
male	low	346	10986.279	12473.85

Tukey multiple comparisons of means We have already proven both variables to be significantly different by ANOVA test. However, if we want to find out which group is not homogeneious with the other groups, we will still run a Tukey's test.

[108]:
```
tukey_test <- TukeyHSD(anovaACS)
print(tukey_test)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = charges ~ age.c * sex, data = MCPD)

$age.c
              diff       lwr       upr p adj
low-high -6273.295 -7526.715 -5019.875     0

$sex
```

```
                diff      lwr      upr     p adj
male-female 1489.841 236.3879 2743.295 0.0198643

$`age.c:sex`
                              diff        lwr       upr     p adj
low:female-high:female -6514.794 -8851.3487 -4178.239 0.0000000
high:male-high:female   1273.794 -1059.1862  3606.773 0.4967335
low:male-high:female   -4811.173 -7116.8602 -2505.486 0.0000006
high:male-low:female    7788.587  5445.0267 10132.148 0.0000000
low:male-low:female     1703.620  -612.7719  4020.013 0.2319951
low:male-high:male     -6084.967 -8397.7531 -3772.181 0.0000000
```

The interpretation here is straight forward. If the p-value is significant, the two groups' means are different. Because we have 4 groups, there are 6 pair-wise comparisons. Four of them do differ significantly from each other.

4 Short Comment

This concludes the analysis. Visualizations can be provided upon request. Please contact me (Justin) for additional support, or if you have any question, suggestions, or if you find any mistakes. You can also ask "Xinpei" for additional support. If you feel like helping this analysis, please try to find some literatures supporting the reasons I have mentioned for keeping applying two-way ANOVA when the assumptions were violated.

Thank you for reading :)