Dehshini Ali

Problem set 1

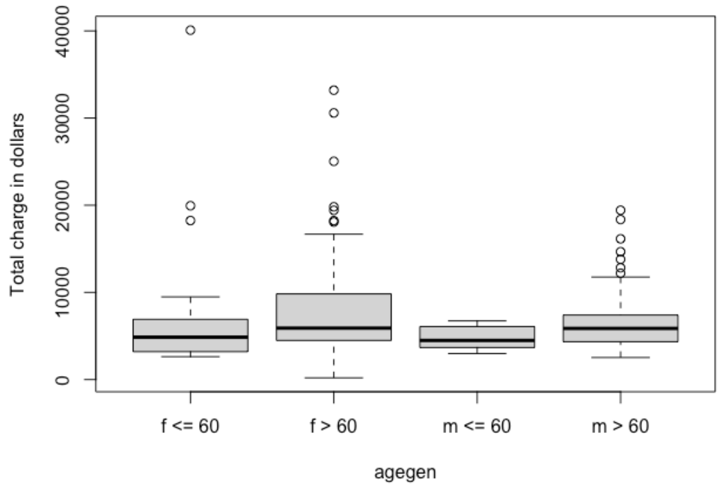
STEP A

1.

#create new variable "agegen", to indicate the four age-gender groups.

2. Inspect the data using side-by-side box plots:

Boxplot showing distributions of each age-gender category



A tibble: 4 × 7

agegen	obs	mean	median	min	max	sd
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
f <= 60	17	8321	4860	2629	40083	9662.
f > 60	83	8054.	5904	200	33197	5970.
m <= 60	10	4829.	4488.	2984	6731	1326.
m > 60	90	6668.	5866	2528	19427	3381.

Describe the differences across the four groups in median costs, spread and shape

The median cost for females>60 years is the highest at $5904 with the median cost for males>60, females<=60, and males<=60 at 5866, 4860, and 4488 respectively.

The costs for females have more variability and are more positively skewed than that for males.

3.

Analysis of Variance Table

Response: totchg

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(agegen)	3	1.619e+08	53968180	1.9412	0.1242
Residuals	196	5.449e+09	27800817		

```
>summary(model1)
Call:
lm(formula = totchg ~ as.factor(agegen), data = ce621)
Residuals:
   Min     1Q Median     3Q    Max
 -7854  -3061  -1438   1232  31762
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8321.0	1278.8	6.507	6.26e-10 ***
as.factor(agegen)f > 60	-267.1	1403.7	-0.190	0.8493
as.factor(agegen)m <= 60	-3492.3	2101.3	-1.662	0.0981 .
as.factor(agegen)m > 60	-1653.0	1394.4	-1.186	0.2372

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5273 on 196 degrees of freedom
Multiple R-squared:  0.02886,           Adjusted R-squared:  0.01399
F-statistic: 1.941 on 3 and 196 DF,  p-value: 0.1242
```

4.
Interpret each of the regression coefficients. Using regression, how do you test the overall hypothesis of no group differences? What is the difference between the results of the lm and glm commands?

The intercept (B_0) is the mean cost for the categorical group of F<=60.
The coefficient for f>60 (B_1) of -267.1 is the difference in means of the categorical groups f>60 and f<=60.
The coefficient for m<=60 (B_2) of -3492.3 is the difference in means of the categorical groups m<=60 and f<=60.
The coefficient for m>60 (B_3) of -1653.0 is the difference in means of the categorical groups m>60 and f<=60.

The global F test can be used to test the overall hypothesis of differences in the means of the groups.
H_0 : $B_1 = B_2 = B_3 = 0$
H_a : at least one of the groups mean is different. B_j != 0
F-statistic is 1.941 on 3 and 196 df,
p-value = 0.1242
fail to reject the null hypothesis and conclude that all categorical groups have similar mean costs.
There are 4 categorical groups. The number of pairwise tests that can be done is 4choose2 = 6.
Using Bonferroni's correction, corrected alpha = 0.05/6 = 0.0083.
Fail to reject the null since p of 0.1242 is greater than the corrected alpha.

The lm and the glm commands have the same results.

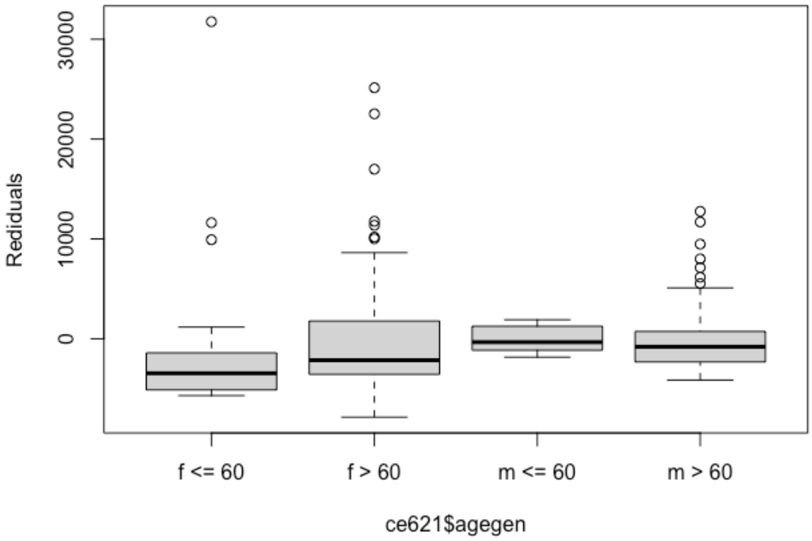
5.
Summarize your findings with respect to CE costs as a function of age and gender, as if for a health services journal.

The average cost of carotid endarterectomy was higher for females as compared to males with younger females spending more ($8321) than older females ($8054) and older males spending more ($6668) than younger males ($4829).
However, these means were not statistically significantly different from each other.

STEP B

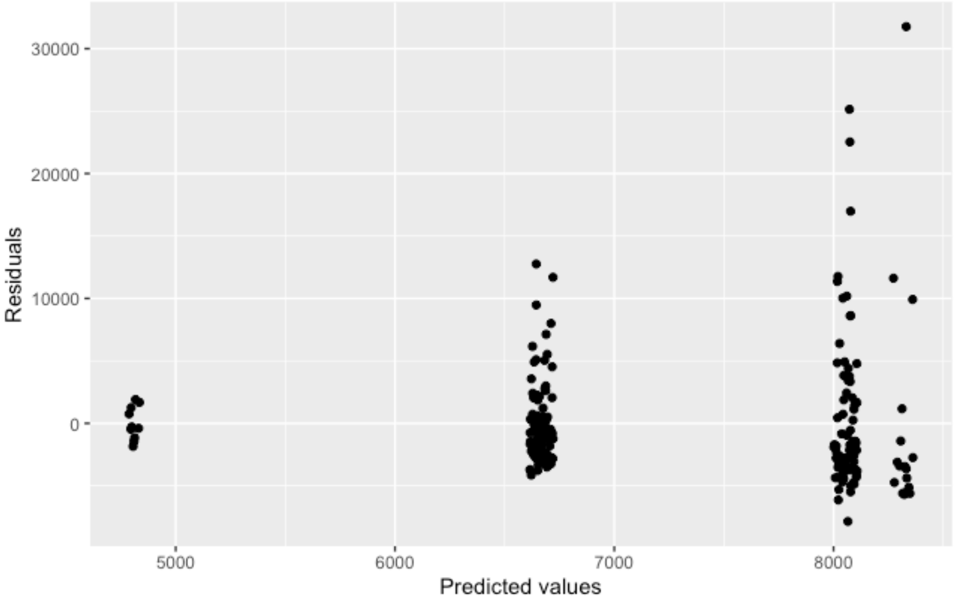
1.

Boxplot of residuals of the age-gen groups

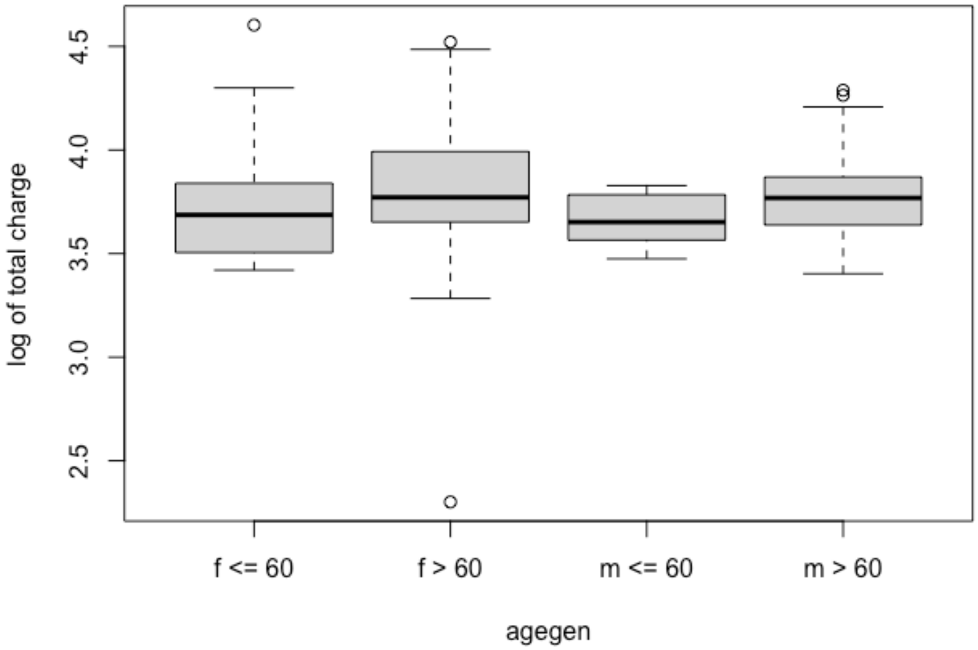


2.

Dot plot of the residuals of the age-gen categorical groups



3.



4. Do the within-group distributions appear more nearly normal and are their variances more nearly equal across groups using either the untransformed CE data in Step A or the transformed CE data in Step B?

The within group distributions are more nearly normal except for the group f<=60 which still looks positively skewed (skewed to the right).
The variability is now almost equal to each other, at least closer as compared to the box plot in step A. The variance in the group m<=60 remains smaller compared to the other groups.

STEP C

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 250 bootstrap replicates
CALL :
boot.ci(boot.out = results, type = "norm", index = 1)
Intervals :
Level      Normal
95%   ( 3593, 12591 )
Calculations and Intervals on Original Scale
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 250 bootstrap replicates
CALL :
boot.ci(boot.out = results, type = "norm", index = 2)
Intervals :
Level      Normal
95%   (-4828.3,  4709.4 )
Calculations and Intervals on Original Scale
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 250 bootstrap replicates
CALL :
boot.ci(boot.out = results, type = "norm", index = 3)
Intervals :
Level      Normal
95%   (-7780,  1236 )
Calculations and Intervals on Original Scale
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 250 bootstrap replicates
CALL :
boot.ci(boot.out = results, type = "norm", index = 4)
Intervals :
Level      Normal
95%   (-5989,  3070 )
Calculations and Intervals on Original Scale
```

C.I based on bootstrap std.errors.

Difference in means		original	95% Confidence interval	bias	std. error
(f<=60) – 0	t1*	8321.0000	(3593, 12591)	229.1168	2295.559
(f>60) – (f<=60)	t2*	-267.1205	(-4828.3, 4709.4)	-207.6785	2433.114
(m<=60) – (f<=60)	t3*	-3492.3000	(-7780, 1236)	-220.4808	2300.012
(m>60) – (f<=60)	t4*	-1653.0333	(-5989, 3070)	-193.6263	2311.066

C.I based on regression model

	2.5 %	97.5 %
(Intercept)	5799.016	10842.9840
as.factor(agegen)f > 60	-3035.358	2501.1166
as.factor(agegen)m <= 60	-7636.343	651.7426
as.factor(agegen)m > 60	-4402.908	1096.8418

3) Write a brief paragraph describing the differences in the distribution of expenditures among the four groups. Be numerate, focus on the mean expenditures, use confidence intervals and/or tests and discuss the differences in spread and shape as appropriate.

As compared to young females, the mean expenditure of CE is less than that for old females (-267 [-4828, 4709]), young males (-3492 [-7780, 1236]), and old males (-1653 [-5989, 3070]) The difference in means for the different groups compared to young females is however not statistically significant assuming the null hypothesis that (there is no difference between the group means) is true.

Codes Used

```
#read in the data as ce621
#create new variable "agegen", to indicate the four age-gender groups.
ce621 = ce621 %>% mutate(agegen=case_when(sex=="Male" & age<=60 ~ "m <= 60",
                                          sex=="Male" & age>60 ~ "m > 60",
                                          sex=="Female" & age<=60 ~ "f <= 60",
                                          sex=="Female" & age>60 ~ "f > 60"))

#inspect the data using boxplot
boxplot(totchg ~ agegen, data = ce621, ylab = "Total charge in dollars")

#summarize totchg according to groups
ce621 %>%
  group_by(agegen) %>%
  summarize(obs=n(),
                        mean=mean(totchg),
                        median=median(totchg),
                        min=min(totchg),
                        max=max(totchg),
                        sd=sd(totchg)
                        )

#linear regression of total charges on the age-gender groups
model1 = lm(totchg ~ as.factor(agegen), data = ce621 )
anova(model1)
summary(model1)

#boxplot the residuals for age-gender groups
boxplot(model1$residuals ~ ce621$agegen, ylab = "Rediduals")

#create new column for log of totchg
ce621 = ce621 %>% mutate(logtotchg=log10(totchg))
#boxplot the log of totchg for the different groups
boxplot(logtotchg ~ agegen, data = ce621, ylab = "log of total charge")

#create a new lm model for the logtotchg on the agegen groups
model3 = lm(logtotchg ~ as.factor(agegen), data = ce621)

boxplot(model3$residuals ~ ce621$agegen, ylab = "logResiduals")
```

```
#function to obtain regression coefficients. this will be used by the boot function.
regcoef = function(formula, data, i)
{
  fitline = lm(formula, data = data[i,])
  return(coefficients(fitline))
}
# bootstrapping with 250 replications
results2 = boot(data=ce621, statistic = regcoef, R=250, formula=totchg~agegen)

# get 95% confidence intervals from the bootstrap
boot.ci(results, type="norm", index=1) # intercept (f <=60)
boot.ci(results, type="norm", index=2) # f >60
boot.ci(results, type="norm", index=3) # m <=60
boot.ci(results, type="norm", index=4) # m >60

# get 95% confidence intervals from the regression model
confint(model1)
```