STA303 Assignment 1

Manyi Luo - 1003799419

1/27/2020

```
## ── Attaching packages ─────────────────────────────────────── tidyver
se 1.3.0 ──
```

```
## ✓ ggplot2 3.2.1     ✓ purrr   0.3.3
## ✓ tibble  2.1.3     ✓ dplyr   0.8.3
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_con
flicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## # A tibble: 6 x 5
## # Groups:   decade [1]
##   season_number title                     season_rating decade genres
##           <dbl> <chr>                             <dbl> <chr>  <chr>
## 1             7 Matlock                            7.6  1990   Crime,Drama,Myst…
## 2             3 Homicide: Life on the St…          8.81 1990   Crime,Drama,Myst…
## 3             6 Law & Order                        8.01 1990   Crime,Drama,Myst…
## 4             5 Law & Order                        7.81 1990   Crime,Drama,Myst…
## 5             9 Columbo                            7.38 1990   Crime,Drama,Myst…
## 6            11 Columbo                            6.94 1990   Crime,Drama,Myst…
```

Question 1: ANOVA as a linear model

Question 1a:

Write the equation for a linear model that would help us answer our question of interest AND state the assumptions for the ANOVA.

The equation for a linear model that would help us answer our question of interest can be expressed below:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

whereas Y_{ij} is the jth observation from the ith group in general. Specifically in this case, i is grouped by decades (1990s, 2000s, 2010s), j represents each show within particular decade and Y_{ij} together represents each show's rating among decades. μ_i represents the average season rating of the ith decade. And ϵ_{ij} is the random error.

The assumptions for the ANOVA are the followings: (1). Errors are independent. (2). Errors are normal and $E(\epsilon_{ij}) = 0$. (3). Constant variance ($Var(\epsilon_{ij}) = \sigma^2$).

$$\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$$

Question 1b

Write the hypotheses for an ANOVA for the question of interest in words. Make it specific to this context and question.

The hypotheses for an ANOVA for the question of interest can be expressed as below:

$$H_0 : \mu_{1990} = \mu_{2000} = \mu_{2010},\ H_a : otherwise\ (\mu_i \neq \mu_j\, for\ some\ i\ and\ j,\ both\ i,\ j\ represent\ decades)$$

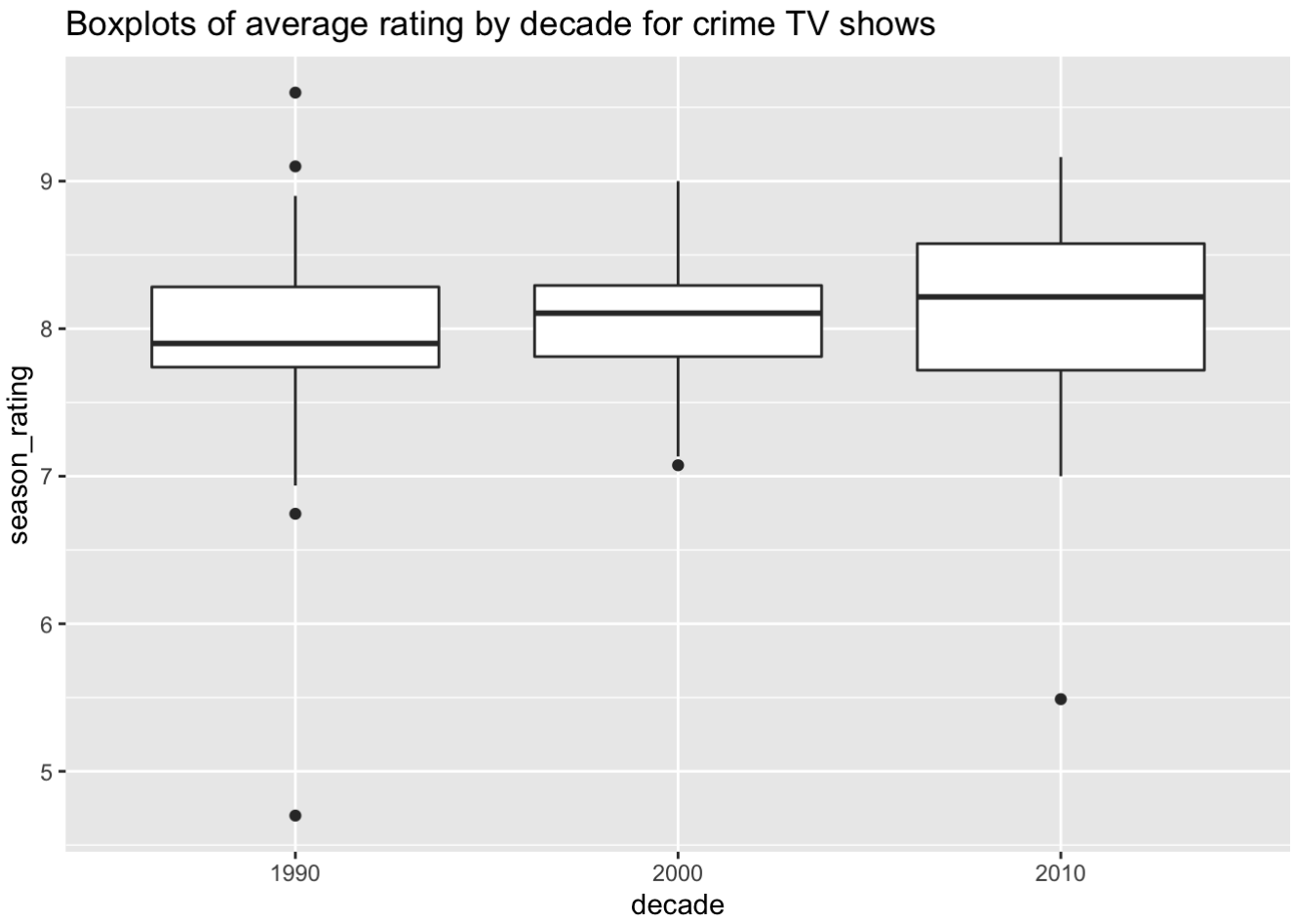
They can be explained in words as:

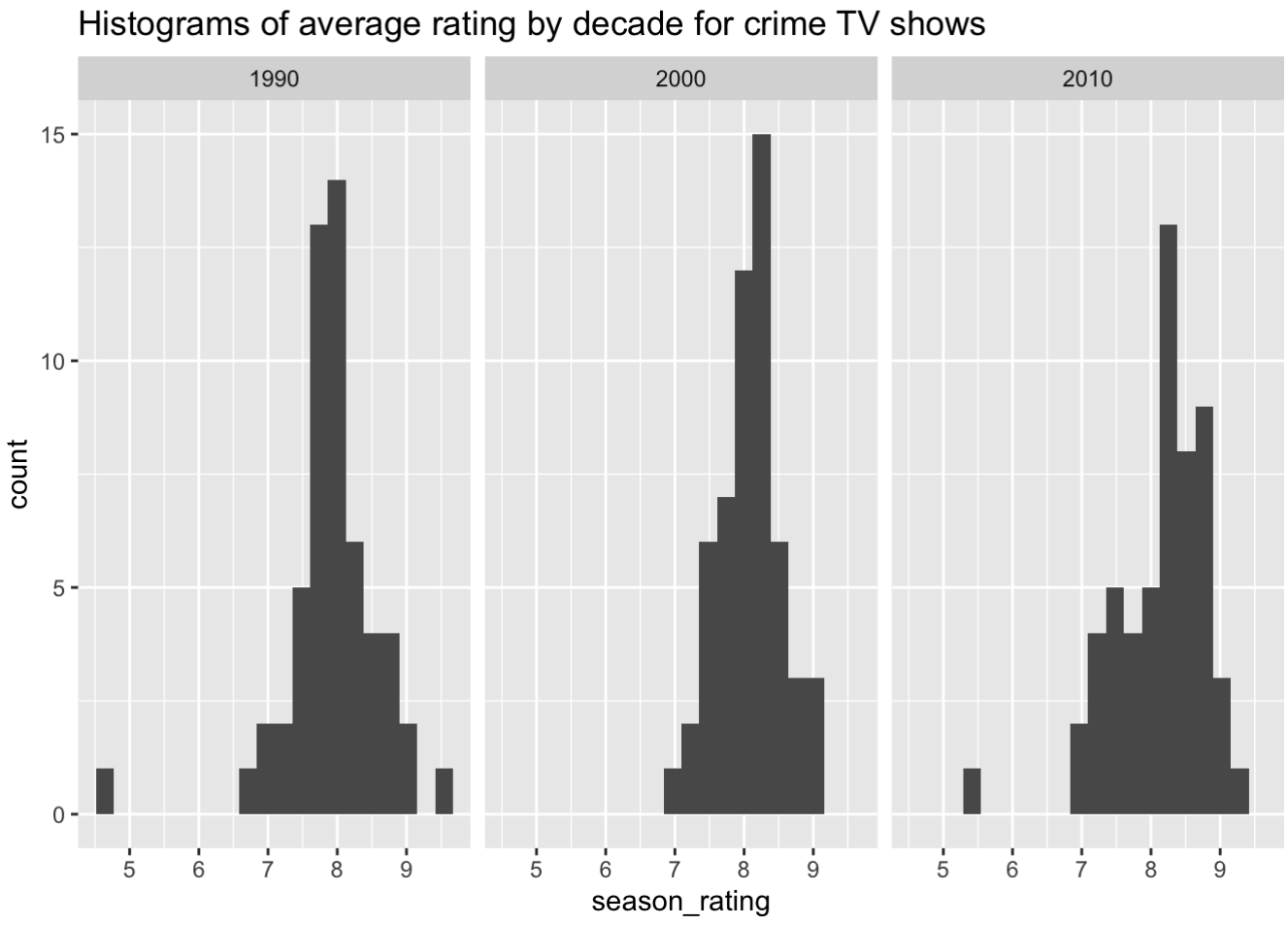
H_0: the average season rating for crime shows is the same from decade to decade.

H_a: at least one average season rating for crime shows is different from the other decade.

Question 1c

Make two plots, side-by-side boxplots and facetted historgrams, of the season ratings for each decade. Briefly comment on which you prefer in this case and one way you might improve this plot (you don't have to make that improvement, just briefly decribe it). Based on these plots, do you think there will be a significant difference between any of the means?


Boxplots of average rating by decade for crime TV shows
season_rating
9
8
7
6
5
1990
2000
2010
decade


Histograms of average rating by decade for crime TV shows
1990
2000
2010
count
15
10
5
0
5
6
7
8
9
season_rating

I personally perfer the first plot, which is the side by side boxplot. This is because we can compare the difference between average season ratings among decades more easily.

One improvement I would suggest is to add a horizontal line that represents the total average season rating for all decades.

And based on these plots, I don't think there will be a significant difference between any of the means.

Question 1d

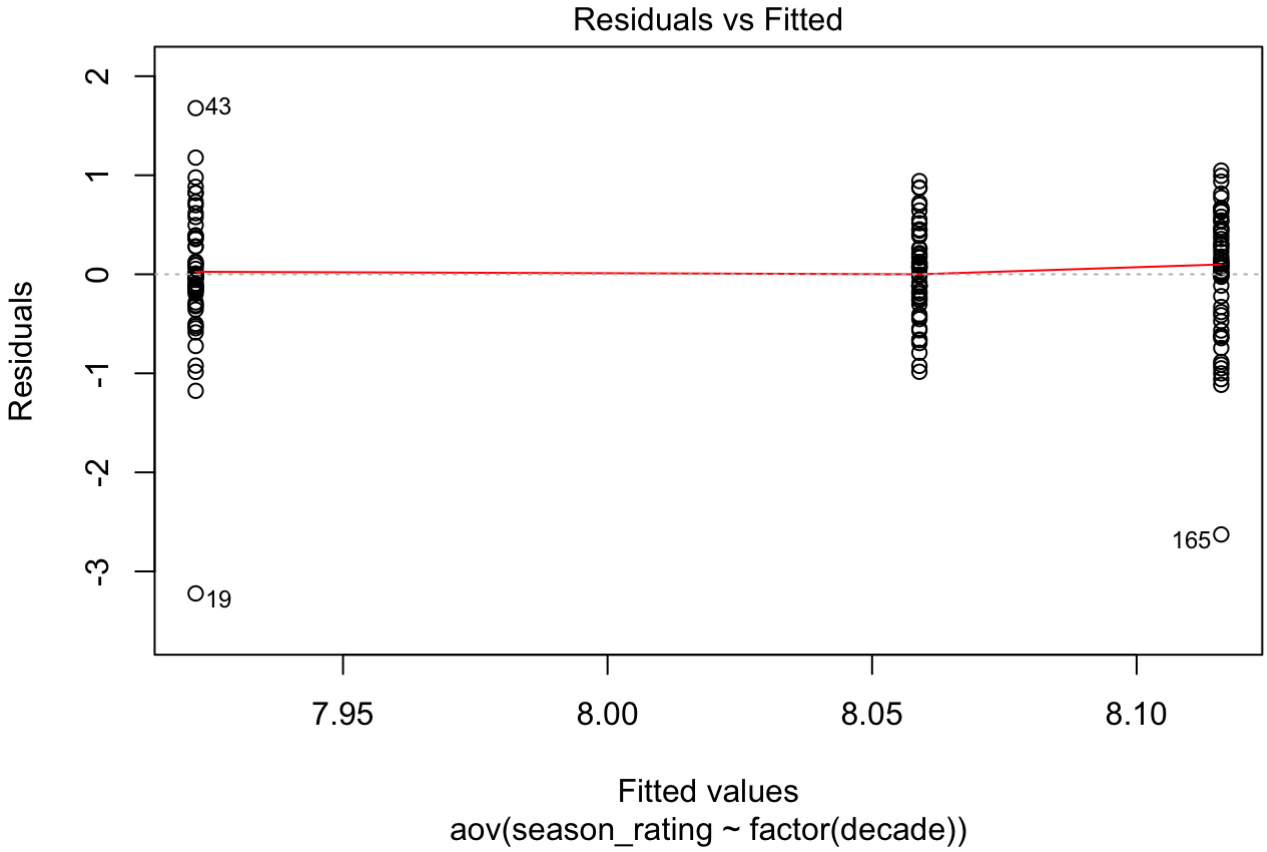
Conduct a one-way ANOVA to answer the question of interest above. Show the results of summary() on your ANOVA and briefly interpret the results in context (i.e., with respect to our question of interest).

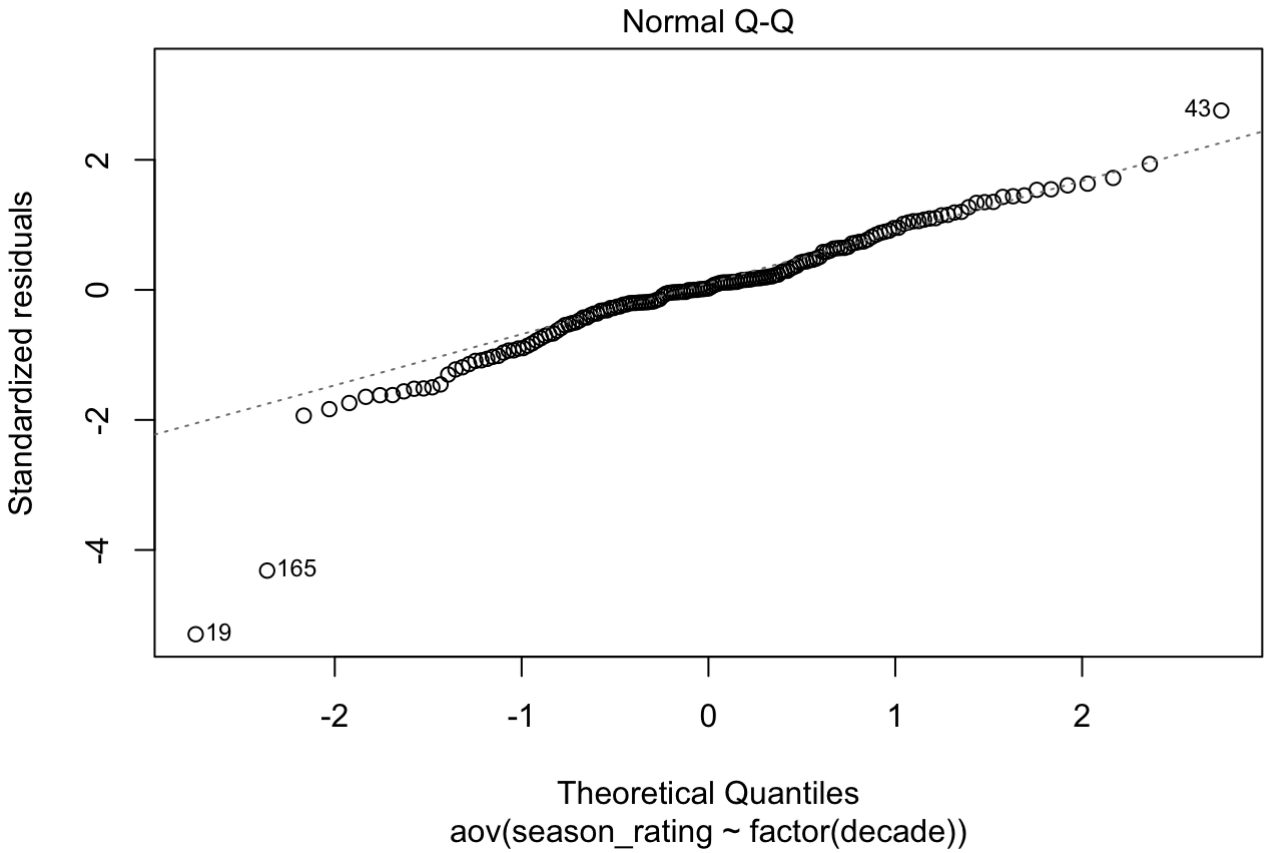
```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## factor(decade)   2   1.09  0.5458   1.447  0.238
## Residuals      162  61.08  0.3771
```

From the summary result above, we can see that the p_value under H_0 is true is 0.238, which is significantly larger than the critical value of 0.05. Therefore, our conclusion is insignificant and we failed to reject H_0. Under the context, this means we support the null hypothesis that the average season rating for crime shows is the same from decade to decade.

Question 1e

Update the code below to create two plots and the standard deviation of season rating by decade. Briefly comment on what each plot/output tells you about the assumptions for conducting an ANOVA with this data.


Residuals vs Fitted
Residuals
Fitted values
aov(season_rating ~ factor(decade))
43
19
165
2
1
0
-1
-2
-3
7.95
8.00
8.05
8.10


Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(season_rating ~ factor(decade))
43
165
19
2
0
-2
-4
-2
-1
0
1
2

```
## # A tibble: 3 x 2
##   decade var_rating
##   <chr>       <dbl>
## 1 1990        0.480
## 2 2000        0.203
## 3 2010        0.447
```

From the first plot (the residual plot), we can see that there is no pattern in residual points and the fitted line is nearly flat. Also, from the summary chart, we can see that the estimate of largest within group variance to the estimate of smallest within group variance does not exceed 3 (0.4804055/0.2033781 = 2.362 < 3). Both evidences support the third ANOVA assumption (constant variance).

From the second plot (the normal qq-plot), we can see that all observations form a linear trend and there are nearly no deviations in observations around the fitted line. This means that the residuals are normal distributed, which satisfies the second ANOVA assumption.

However, interpreting from the context, the first ANOVA assumption (Errors are independent) might be slightly violated. This is because decades are cummulative, as 2000 might be dependent on 1990 and 2010 might be dependent on 2000 as well as 1990.

Question 1f

Conduct a linear model based on the question of interest. Show the result of running summary() on your linear model. Interpret the coefficients from this linear model in terms of the mean season ratings for each decade. From these coefficients, calculate the observed group means for each decade.

The summary results are presented below:

```
##
## Call:
## lm(formula = season_rating ~ 0 + factor(decade), data = crime_show_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2222 -0.2589  0.0135  0.3862  1.6778
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## factor(decade)1990   7.9222     0.0828   95.68   <2e-16 ***
## factor(decade)2000   8.0589     0.0828   97.33   <2e-16 ***
## factor(decade)2010   8.1160     0.0828   98.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6141 on 162 degrees of freedom
## Multiple R-squared:  0.9943, Adjusted R-squared:  0.9942
## F-statistic:  9412 on 3 and 162 DF,  p-value: < 2.2e-16
```

From these coefficients, the estimated mean ratings for each decade are:
$\hat{\mu}_{1990} = 7.9222, \hat{\mu}_{2000} = 8.0589, \hat{\mu}_{2010} = 8.1160$

And from the last column, we can see that all the p-values for three decades are smaller than 2e^-16, which are significantly smaller than the critical value of 0.05. This shows that we have strong evidence to conclude that three estimated mean ratings above are nonzero.

A linear model can be conducted as below:

$$\hat{Y}_i = 7.9222I(D_i = 1990) + 8.0589I(D_i = 2000) + 8.1160(D_i = 2010)$$

whereas $\hat{Y}_i$ represents the estimated rating for a particular show and D_i represents the decade of the show. Indicator variables $I(D_i = 1990), I(D_i = 2000), I(D_i = 2010)$ equal to 1 if decades belongs to 1990, 2000 and 2010 respectively; otherwise, they equal to 0.

Eventhough we already knew the estimated mean ratings for each decade through coefficients above, we can still calculate them (as required by the prompt) using an alternative method:

```
##
## Call:
## lm(formula = season_rating ~ factor(decade), data = crime_show_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2222 -0.2589  0.0135  0.3862  1.6778
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.9222     0.0828  95.679   <2e-16 ***
## factor(decade)2000    0.1368     0.1171   1.168   0.2444
## factor(decade)2010    0.1938     0.1171   1.655   0.0998 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6141 on 162 degrees of freedom
## Multiple R-squared:  0.01756,    Adjusted R-squared:  0.005426
## F-statistic: 1.447 on 2 and 162 DF,  p-value: 0.2382
```

From the new regression results, we can see that the estimated mean ratings for decade 1990 is: $\hat{\mu}_{1990} = 7.9222$, as decade 1990 is used as the reference group. So the estimated mean ratings for decade 2000 and 2010 can be calculated using the new estimates in the first column: $\hat{\mu}_{2000} = 7.9222 + 0.1368 = 8.059, \hat{\mu}_{2010} = 7.9222 + 0.1938 = 8.116$, which is similar as what we got before.

Question 2: Generalised linear models - Binary

```
## [1] "smoke"        "smokeFormats"
```

```
##                     colName
## 151 chewing_tobacco_snuff_or
##                                                                          label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.700	0.082	-32.843	0.000
ageC	0.341	0.021	16.357	0.000
RuralUrbanRural	0.959	0.088	10.934	0.000
Raceblack	-1.557	0.172	-9.068	0.000
Racehispanic	-0.728	0.104	-6.981	0.000
Raceasian	-1.545	0.342	-4.515	0.000
Racenative	0.112	0.278	0.404	0.687
Racepacific	1.016	0.361	2.814	0.005
SexF	-1.797	0.109	-16.485	0.000

```
## Waiting for profiling to be done...
```

	est	0.5 %	99.5 %
Baseline prob	0.063	0.051	0.076
ageC	1.407	1.334	1.485
RuralUrbanRural	2.610	2.088	3.283
Raceblack	0.211	0.132	0.320
Racehispanic	0.483	0.367	0.628
Raceasian	0.213	0.077	0.466
Racenative	1.119	0.509	2.163
Racepacific	2.761	0.985	6.525
SexF	0.166	0.124	0.218

Question 2a

Write down and explain the statistical model which smokeModel corresponds to, defining all your variables.

The smokeModel corresponds to the logistic statistical model:

$$Y_i \sim Binomial(N_i, \mu_i)$$

whereas μ_i is the probability that a person consumes tobacco in each group i and N_i is the total number of people in each group i.

Also:

$$log(\frac{\mu_i}{1-\mu_i}) = x_i\beta$$

The reason why binomial logistic model is selected is because the response for each individual is either 0 or 1 in this case. And we cummulate the response for each group, so this situation satisfies binomial logistic model.

Variables are defined below based on their category (X_i):

1. X_1 (ageC) represents age, which is a numeric variable representing original age - 16.
2. X_2 (RuralUrban) is a indicator variable, which represents whether a person lives in rural (X_2 = 1) or urban area (X_2 = 0).
3. X_3 to X_7 are all indicator variables, which represent different races, including: black (X_3 = 1), hispanic(X_4 = 1), asian (X_5 = 1), native (X_6 = 1) and pacific (X_7 = 1). If an individual belongs to one of the X_3 to X_7, the other variables equal to 0.
4. X_8 (Sex) is also an indicator variable, which represents sex, includin female (X_4 = 1) and male (X_4 = 0).

Question 2b

Write a sentence or two interpreting the row “baseline prob” in the table above. Be specific about which subset of individuals this row is referring to.

“Baseline prob” is 0.063 for 16 years old white male, who lives in rural area that has ever tried smoking.

Question 2c

Write a short paragraph addressing the hypothesis that rural white males are the group most likely to use chewing tobacco, and there is reasonable certainty that less than half of one percent of ethnic-minority urban women and girls chew tobacco.

```
##         fit     se.fit     lower     upper
## 1 -1.740164 0.05471340 -1.904304 -1.576024
## 2 -2.699657 0.08219855 -2.946253 -2.453062
## 3 -3.427371 0.10692198 -3.748137 -3.106605
## 4 -6.053341 0.19800963 -6.647370 -5.459312
## 5 -6.041103 0.35209311 -7.097383 -4.984824
```

Sex	Race	RuralUrban	fit	lower	upper
M	white	Rural	149.3	129.6	171.4
M	white	Urban	63.0	49.9	79.2
M	hispanic	Urban	31.5	23.0	42.8
F	black	Urban	2.3	1.3	4.2
F	asian	Urban	2.4	0.8	6.8

From the first row of the second chart provided above, we can see that the estimated probability of white rural male who smokes is 149.3/1000 = 0.1493, which is the largest among other groups. This is also true when viewing white rural male’s confidence interval, as both lower bound and upper bound are the largest among other groups’ confidence interval. Therefore, we have no evidence to reject the first part of the hypothesis, and there is reasonable certainty that rural white males are the group most likely to use chewing tobacco.

Also we can see from the last two row of the second chart, the estimated probability of ethnic-minority urban women who smoke is 2.3+2.4/1000 = 0.0047, which is less than half of one percent. However, through viewing urban asian female's confidence interval, we can see it includes 5 (0.8, 6.8). So refering to the question, we are less certain to conclude that less than half of one percent of ethnic-minority urban women and girls chew tobacco.

Question 3: Generalised linear models - Poisson

Question 3a

Write down and explain the statistical model which fijiRes corresponds to, defining all your variables.

```
## [1] "fiji"     "fijiFull"
```

```
## Waiting for profiling to be done...
```

	Estimate	Std. Error	z value	Pr(>|z|)	est	0.5 %	99.5 %
(Intercept)	-1.181	0.017	-69.196	0.000	0.307	0.294	0.321
ageMarried0to15	-0.119	0.021	-5.740	0.000	0.888	0.841	0.936
ageMarried18to20	0.036	0.021	1.754	0.079	1.037	0.983	1.093
ageMarried20to22	0.018	0.024	0.747	0.455	1.018	0.956	1.084
ageMarried22to25	0.006	0.030	0.193	0.847	1.006	0.930	1.086
ageMarried25to30	0.056	0.048	1.159	0.246	1.057	0.932	1.195
ageMarried30toInf	0.138	0.098	1.405	0.160	1.147	0.882	1.462
ethnicityindian	0.012	0.019	0.624	0.533	1.012	0.964	1.061
ethnicityeuropean	-0.193	0.170	-1.133	0.257	0.824	0.514	1.242
ethnicitypartEuropean	-0.014	0.069	-0.206	0.837	0.986	0.822	1.171
ethnicitypacificIslander	0.104	0.055	1.884	0.060	1.110	0.959	1.276
ethnicityroutman	-0.033	0.132	-0.248	0.804	0.968	0.675	1.336
ethnicitychinese	-0.380	0.121	-3.138	0.002	0.684	0.492	0.920
ethnicityother	0.668	0.268	2.494	0.013	1.950	0.895	3.622
literacyno	-0.017	0.019	-0.857	0.391	0.984	0.936	1.034
urbansuva	-0.159	0.022	-7.234	0.000	0.853	0.806	0.902
urbanotherUrban	-0.068	0.019	-3.513	0.000	0.934	0.888	0.982

```
## Waiting for profiling to be done...
```

	Estimate	Std. Error	z value	Pr(>|z|)	est	0.5 %	99.5 %
(Intercept)	-1.163	0.012	-93.674	0.000	0.313	0.303	0.323
marriedEarlyTRUE	-0.136	0.019	-7.189	0.000	0.873	0.832	0.916
ethnicityindian	-0.002	0.016	-0.154	0.877	0.998	0.958	1.039
ethnicityeuropean	-0.175	0.170	-1.034	0.301	0.839	0.524	1.262
ethnicitypartEuropean	-0.014	0.068	-0.202	0.840	0.986	0.823	1.171
ethnicitypacificIslander	0.102	0.055	1.842	0.065	1.107	0.957	1.273
ethnicityroutman	-0.038	0.132	-0.285	0.775	0.963	0.672	1.330
ethnicitychinese	-0.379	0.121	-3.130	0.002	0.684	0.493	0.921
ethnicityother	0.681	0.268	2.545	0.011	1.976	0.907	3.667
urbansuva	-0.157	0.022	-7.162	0.000	0.855	0.808	0.904
urbanotherUrban	-0.066	0.019	-3.414	0.001	0.936	0.891	0.984

The statistical model that fijiRes corresponds to is the Poisson statistical model:

$$Y_i \sim Poisson(O_i, \mu_i)$$

This is because the response in this case envolves counting and its better for us to use Poisson model to explain the responses in terms of the predictors.

Also:

$$log(\mu_i/year) = X_i\beta$$

whereas Y_i represents the number of children which a woman i borns; μ_i represents the rate of children born in a year; O_i represents the offset term, which is the log number of years since the women first gets married; Last, X_i is a design matrix, which represents indicator variables (16 variables in total) include different age married (ageMarried0to15, ageMarried18to20……ageMarried30toInf), different ethnicities (ethnicityindian, ethnicityeuropean……ethnicityother), literate or not (literacy) and whether living in urban area (urbansuva, urbanotherUrban).

Question 3b

Is the likelihood ratio test performed above comparing nested models? If so what constraints are on the vector of regression coefficients β in the restricted model?

```
## Likelihood ratio test
##
## Model 1: children ~ offset(logYears) + marriedEarly + ethnicity + urban
## Model 2: children ~ offset(logYears) + ageMarried + ethnicity + literacy +
##     urban
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  11 -9604.3
## 2  17 -9601.1  6 6.3669     0.3834
```

Likelihood ratio test is performed when comparing nested models (model 1/fijiRES2 and model 2/fijiRes). Model1 is nested in Model 2, and we can see model 2 also include the indicator variable literacy. From the likelihood ratio test result, we can see that the p-value is 0.3834, which is significantly greater than 0.05. Therefore, our conclusion is to choose the simpler model, model 1; and the degree of literacy does not influence the number of children which a woman borns. The constraint on the vector of regression coefficients β in the restricted model is: $\beta_{literacy} = 0$.

Also, note that the variable marriedEarly in model 1 functions the same comparing to variable ageMarried in model 2. We can see that ageMarried further breaks down into ageMarried0to15, ageMarried18to20…… ageMarried30toInf; whereas marriedEarly equals to true when age married is 0 to 15 years old. The estimate for ageMarried0to15 is -0.119, which is close to the estimate for marriedEarlyTRUE (-1.163). This means married early (0 to 15 years old) will lead to a slight decrease in the number of children which a woman borns.

Question 3c

It is hypothesized that improving girls' education and delaying marriage will result in women choosing to have fewer children and increase the age gaps between their children. An alternate hypothesis is that contraception was not widely available in Fiji in 1974 and as a result there was no way for married women.

As already illustrated in Question 3b, literacy does not influence the number of children which a woman borns. Also, the p-value for literacy is 0.391 in model 1, which is significantly greater than the critical value of 0.05. This means that we have no evidence to conclude improving girls' education will result in women choosing to have fewer children.

Similarly, we can see that the p-value for marriedEarly is approximately zero, meaning that the estimate for marriedEarlyTRUE (-1.163) is significant. Therefore, we have strong evidence to say that married early (0 to 15 years old) will result in women choosing to have fewer children and increase the age gaps between their children. Therefore, both hypotheis is incorrect.

Information about contraception was not even provided by the data, so we can't make any conclusions about it.

```
#Question 1
library(tidyverse)
crime_show_data <- readRDS("/Users/mandy/Desktop/crime_show_ratings.RDS")
head(crime_show_data)

crime_show_data %>%
ggplot(aes(x = decade, y = season_rating)) +
geom_boxplot() +
ggtitle("Boxplots of average rating by decade for crime TV shows")
crime_show_data %>%
ggplot(aes(x = season_rating)) +
geom_histogram(bins=20) +
facet_wrap(~decade) +
ggtitle("Histograms of average rating by decade for crime TV shows")

one_way_anova <- aov(season_rating ~ factor(decade), data = crime_show_data)
summary(one_way_anova)

plot(one_way_anova, 1)
plot(one_way_anova, 2)
crime_show_data %>%
group_by(decade) %>%
summarise(var_rating = sd(season_rating)^2)

lm_1 = lm(season_rating ~ 0+factor(decade), data = crime_show_data)
summary(lm_1)
lm_1 = lm(season_rating ~ factor(decade), data = crime_show_data)
summary(lm_1)

#Question 2
smokeFile = 'smokeDownload.RData'
if(!file.exists(smokeFile)){
download.file(
'http://pbrown.ca/teaching/303/data/smoke.RData',
smokeFile)
}
(load(smokeFile))
smokeFormats[
smokeFormats[,'colName'] == 'chewing_tobacco_snuff_or',
c('colName','label')]
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)), ]
smokeSub$ageC = smokeSub$Age - 16
smokeModel = glm(chewing_tobacco_snuff_or ~ ageC + RuralUrban + Race + Sex,
data=smokeSub, family=binomial(link='logit'))
knitr::kable(summary(smokeModel)$coef, digits=3)
logOddsMat = cbind(est=smokeModel$coef, confint(smokeModel, level=0.99))
oddsMat = exp(logOddsMat)
oddsMat[1,] = oddsMat[1,] / (1+oddsMat[1,])
rownames(oddsMat)[1] = 'Baseline prob'
knitr::kable(oddsMat, digits=3)

newData = data.frame(Sex = rep(c('M','F'), c(3,2)),
Race = c('white','white','hispanic','black','asian'),
```

```
ageC = 0, RuralUrban = rep(c('Rural','Urban'), c(1,4)))
smokePred = as.data.frame(predict(smokeModel, newData, se.fit=TRUE, type='link'))[,1:2]
smokePred$lower = smokePred$fit - 3*smokePred$se.fit
smokePred$upper = smokePred$fit + 3*smokePred$se.fit
smokePred
expSmokePred = exp(smokePred[,c('fit','lower','upper')])
knitr::kable(cbind(newData[,-3],1000*expSmokePred/(1+expSmokePred)), digits=1)

#Question 3
fijiFile = 'fijiDownload.RData'
if(!file.exists(fijiFile)){
download.file(
'http://pbrown.ca/teaching/303/data/fiji.RData',
fijiFile)
}
(load(fijiFile))
fijiSub = fiji[fiji$monthsSinceM > 0 & !is.na(fiji$literacy),]
fijiSub$logYears = log(fijiSub$monthsSinceM/12)
fijiSub$ageMarried = relevel(fijiSub$ageMarried, '15to18')
fijiSub$urban = relevel(fijiSub$residence, 'rural')
fijiRes = glm(
children ~ offset(logYears) + ageMarried + ethnicity + literacy + urban,
family=poisson(link=log), data=fijiSub)
logRateMat = cbind(est=fijiRes$coef, confint(fijiRes, level=0.99))
knitr::kable(cbind(
summary(fijiRes)$coef,
exp(logRateMat)),
digits=3)
fijiSub$marriedEarly = fijiSub$ageMarried == '0to15'
fijiRes2 = glm(
children ~ offset(logYears) + marriedEarly + ethnicity + urban,
family=poisson(link=log), data=fijiSub)
logRateMat2 = cbind(est=fijiRes2$coef, confint(fijiRes2, level=0.99))
knitr::kable(cbind(
summary(fijiRes2)$coef,
exp(logRateMat2)),
digits=3)

lmtest::lrtest(fijiRes2, fijiRes)
```