Assignemnt 1

Qizheng Wang

1/26/2020

Question 1

Question 1a

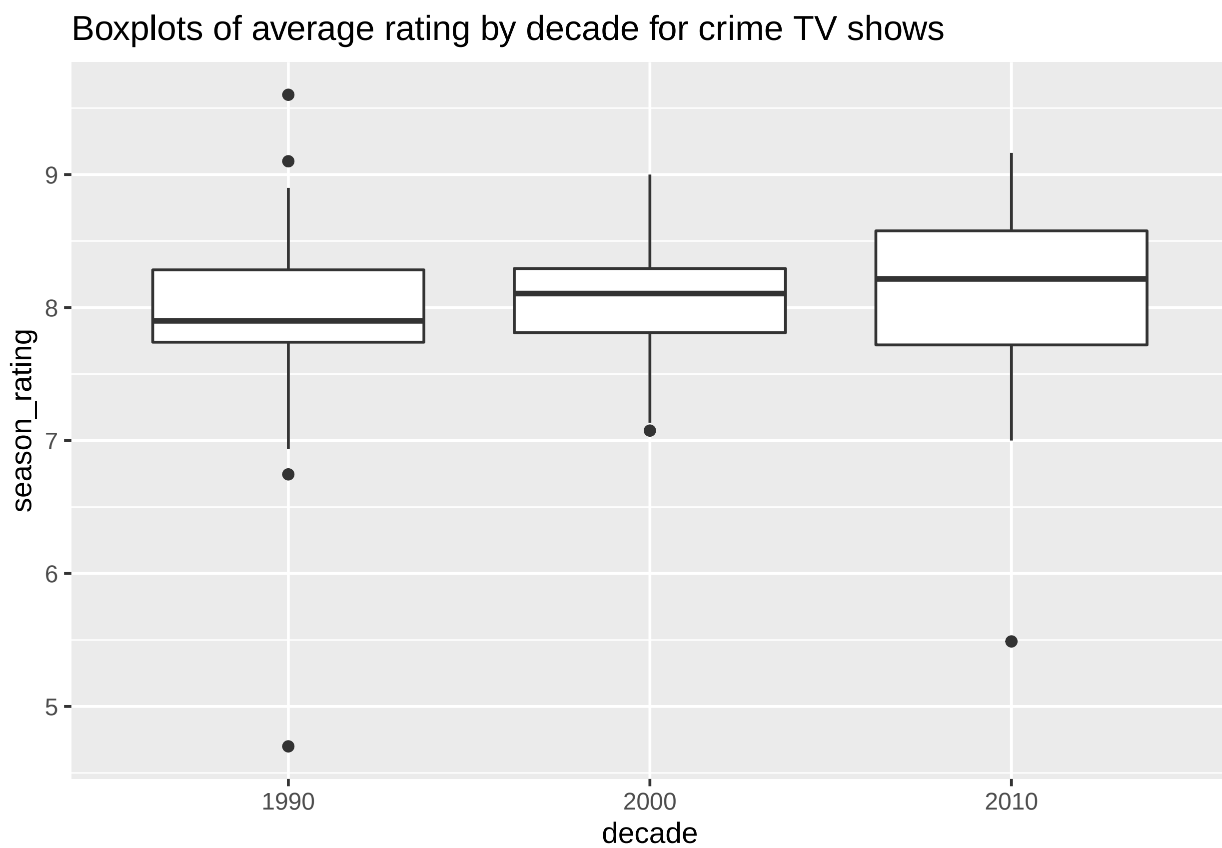
Equation for question of interest: $seasonNumber = \beta_1 * I_1 + \beta_2 * I_2 + \beta_3 * I_3 + \epsilon_i$ Anoava Assumptions: 1. Errors are independent 2. Erorrs are normoally distributed 3. Constant variance

Question 1b

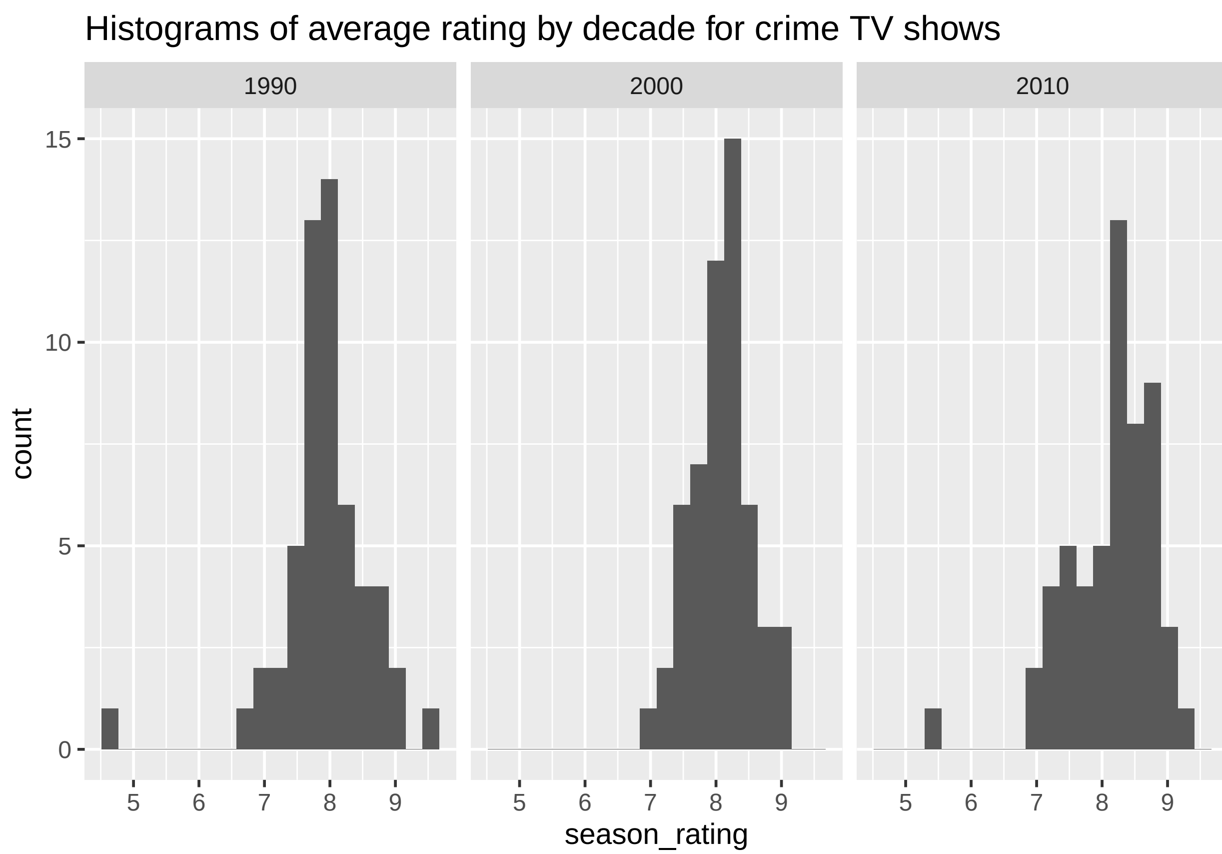
Null hypotheses: $\mu_1 = \mu_2 = \mu_3$ Alternative hypotheses: $\exists i \neq j, s.t\, \mu_i \neq \mu_j$ where μ_1 is the mean of season_rating when 1990s, μ_2 is the mean of season_rating when 2000s, and μ_3 is the mean of season_rating when 2010s

Question 1c

```r
crime_show_data <- readRDS("crime_show_ratings.RDS")

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) + gg
```



I prefer the Boxplot, becasue it's easier to compare the mean. How to improve:

Based on the plots, I think there is a significance difference between the means. The means are gradually increasing over the decades.

Question 1d

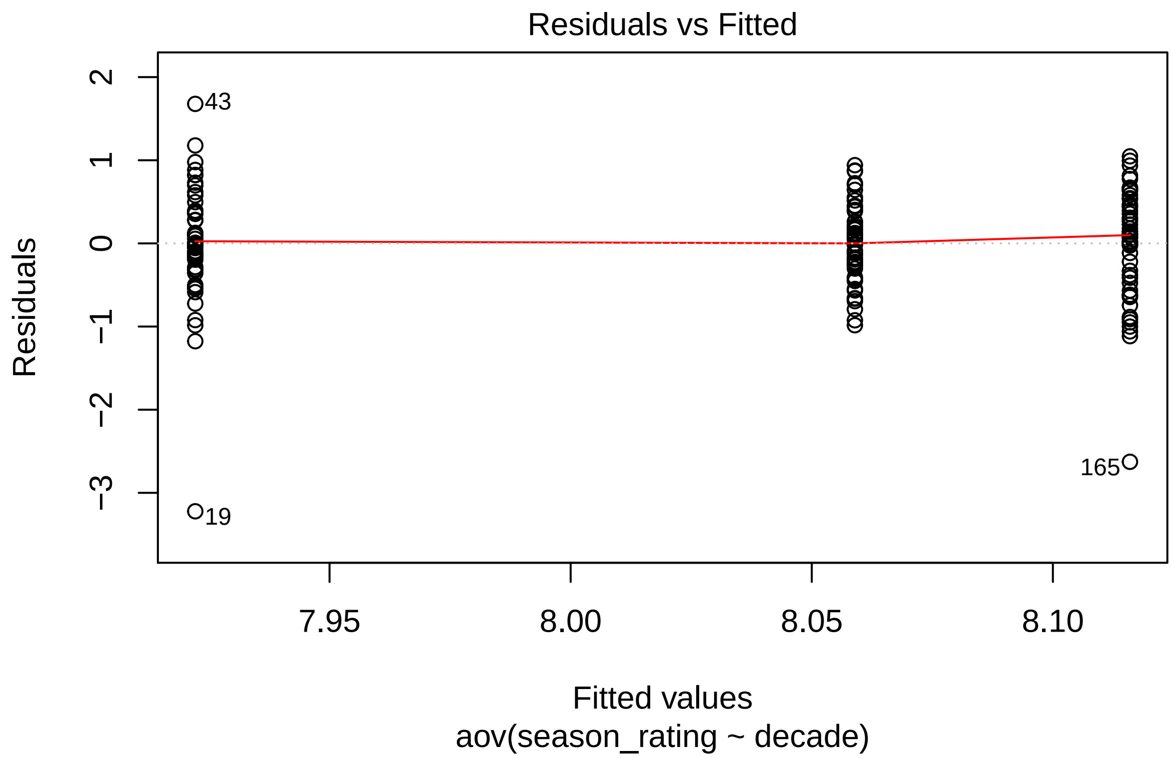
```
anova1d <- aov(season_rating~decade, data=crime_show_data)

summary(anova1d)
```

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

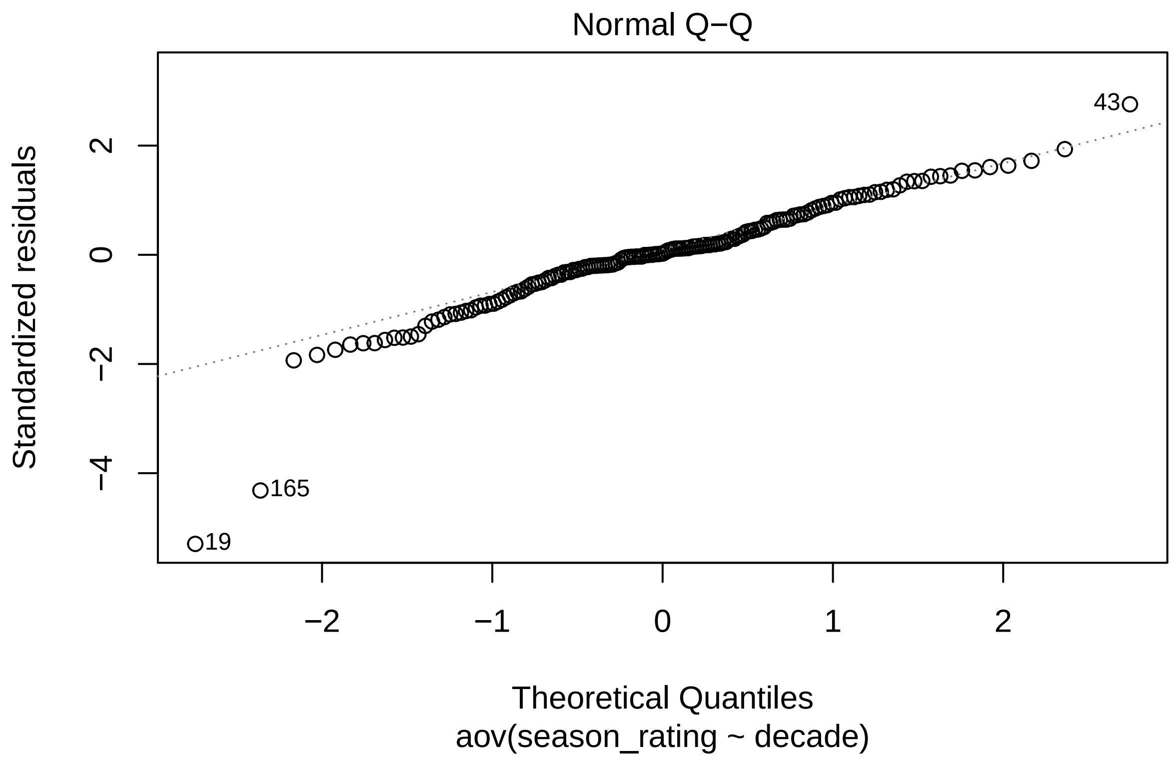
Based on the P-value = 0.238, $P - value \leq 0.05$, thus consider the result significant. We reject the null hypotheses, and there is evidence that the means of season rating differs by decades.

Question 1e

```
plot(anova1d, 1)
```




```r
plot(anova1d, 2)
```




```
crime_show_data %>% group_by(decade) %>% summarise(var_rating = sd(season_rating)^2)
```

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

```
0.4804055/0.2033781
```

```
## [1] 2.36213
```

Residuals vs Fitted plot does not have obvious pattern, thus the assumption for constant variance is satisfied.

Also, by rule of thumb: $s^2_{max}/s^2_{min} = 0.4804055/0.2033781 = 2.36213 \leq 3$ Thus the assumption of constant variance is probably satisfied.

The Normal Q-Q plot checks the normality. The plot forms nearly a straight line, thus the assumption of normality is satisfied.

Question 1f

```r
lm1f <- lm(season_rating~decade, data=crime_show_data)

summary(lm1f)
```

```
##
## Call:
## lm(formula = season_rating ~ 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 ***
## decade2000    0.1368     0.1171   1.168   0.2444
## decade2010    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
```

```r
7.9222+0.1368
```

```
## [1] 8.059
```

```r
7.9222+0.1938
```

```
## [1] 8.116
```

Interpret the coefficients from this linear model in terms of the mean season ratings for each de linear model:

Since there are no decade1990, thus decade1990 is the reference group, which means $\bar{\mu}_{1990} = 7.9222$.

The second coefficient $\beta_1 = 0.1368$ means $\bar{\mu}_{2000} - \bar{\mu}_{1990}$, thus $\bar{\mu}_{2000} = \bar{\mu}_{1990} + 0.1368 = 7.9222 + 0.1368 = 8.059$

The third coefficient $\beta_2 = 0.1938$ means $\bar{\mu}_{2010} - \bar{\mu}_{1990}$, thus $\bar{\mu}_{2010} = \bar{\mu}_{1990} + 0.1928 = 8.116$

Question 2

```r
smokeFile = 'smokeDownload.RData'
if(!file.exists(smokeFile)){
  download.file(  'http://pbrown.ca/teaching/303/data/smoke.RData', smokeFile) }
(load(smokeFile))
```

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

```
smokeFormats[
  smokeFormats[,'colName'] == 'chewing_tobacco_snuff_or',
  c('colName','label')]
```

```
##                      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
```

```
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=binomi

knitr::kable(summary(smokeModel)$coef, digits=3)
```

	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

```
logOddsMat = cbind(est=smokeModel$coef, confint(smokeModel, level=0.99))
```

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

```
oddsMat = exp(logOddsMat)
oddsMat[1,] = oddsMat[1,] / (1+oddsMat[1,])
rownames(oddsMat)[1] = 'Baseline prob'
knitr::kable(oddsMat, digits=3)
```

	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

$log\frac{\mu_i}{1-\mu_i} = X_i\beta$ Where X_i represents the age parameter, the rural or urban factor, and dummy variables for races.

Question 2b

Baselin pro in the table is exp(Intercept) when $X_1, X_2...X_n = 0$ which implies age=16, white race, M, lives in Urban area.

Question 2c

```
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
newData
```

```
##   Sex     Race ageC RuralUrban
## 1   M    white    0      Rural
## 2   M    white    0      Urban
## 3   M hispanic    0      Urban
## 4   F    black    0      Urban
## 5   F    asian    0      Urban
```

```
smokePred
```

```
##         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
```

```
expSmokePred = exp(smokePred[,c('fit','lower','upper')])

knitr::kable(cbind(newData[,-3],1000*expSmokePred/(1+expSmokePred)), digits=1)
```

Sex	Race	RuralUrban	fit	lower	
M	white	Rural	149.3	129.6	
M	white	Urban	63.0	49.9	
M	hispanic	Urban	31.5	23.0	
F	black	Urban	2.3	1.3	
F	asian	Urban	2.4	0.8	
Based	on the fit	, lower and u	pper, Wh	ite who	
Female	minorites	fit's value a	re 2.3+2	.4=4.7,	which divided by 1000 is smaller than the 0.5%. Thus it is reaso

Question 3

```r
fijiFile = 'fijiDownload.RData'
if(!file.exists(fijiFile)){
download.file('http://pbrown.ca/teaching/303/data/fiji.RData',fijiFile)}
(load(fijiFile))
```

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

```r
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(li
logRateMat = cbind(est=fijiRes$coef, confint(fijiRes, level=0.99))
```

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

```r
knitr::kable(cbind(summary(fijiRes)$coef,exp(logRateMat)),digits=3)
```

	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

```r
fijiSub$marriedEarly = fijiSub$ageMarried == '0to15'
fijiRes2 = glm(children ~ offset(logYears) + marriedEarly + ethnicity + urban,family=poisson(link=log),
logRateMat2 = cbind(est=fijiRes2$coef, confint(fijiRes2, level=0.99))
```

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

```r
knitr::kable(cbind(summary(fijiRes2)$coef,exp(logRateMat2)),digits=3)
```

	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

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

```
## 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
```

Question 3a

The model is $log(children) = log(Years) + X_i\beta$ Where $X_i\beta$ is indicator agedMarried, ethnicity, literacy, urban.

Question 3b

Yes, it is comparing nested models. Constraints: $\beta_{literacy} = 0$, ageMarried="0to15"

Question 3c

By comparing the est col of two models, we can see that the race est and urban est in the fijiRes2 model has slightly increase, while these columns represent the situation when other $\beta = 0$, meaning improving education and delaying marriage will result in having fewer children.