

# A1 Soln

Xiangyu Kong

15/01/2020

## Question 1

### Read Data

```
# Read the crime show ratings data
crime_show_file = "crime_show_ratings.RDS"
crime_show_data = readRDS(crime_show_file)
```

### Question 1.a

Let  $y_i$  denote season rating for sample  $i$ .

Let  $x_{i,2000}$  be indicator variable that is set to 1 if the decade for the sample  $i$  is 2000, 0 otherwise.

Let  $x_{i,2010}$  be indicator variable that is set to 1 if the decade for the sample  $i$  is 2010, 0 otherwise.

Equation for linear model:

$$y_i = \beta_0 + \beta_1 x_{i,2000} + \beta_2 x_{i,2010} + \epsilon_i$$

Anova Assumptions:

1. Errors ( $\epsilon_i$ ) are independent
2. Errors are normally distributed with  $E[\epsilon_i] = 0$
3. Errors have constant variance  $var[\epsilon] = \sigma^2$

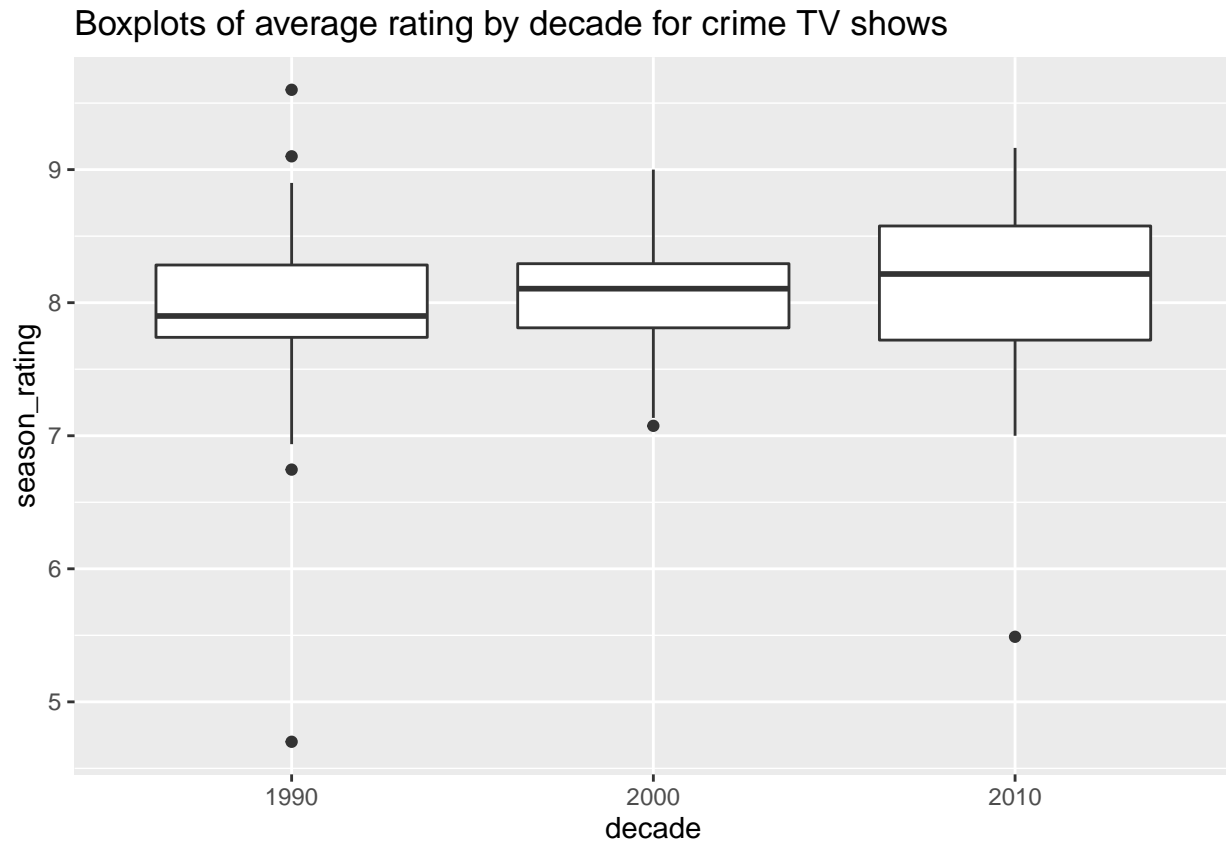
### Question 1.b

The hypotheses for ANOVA are listed and can be described as follows:

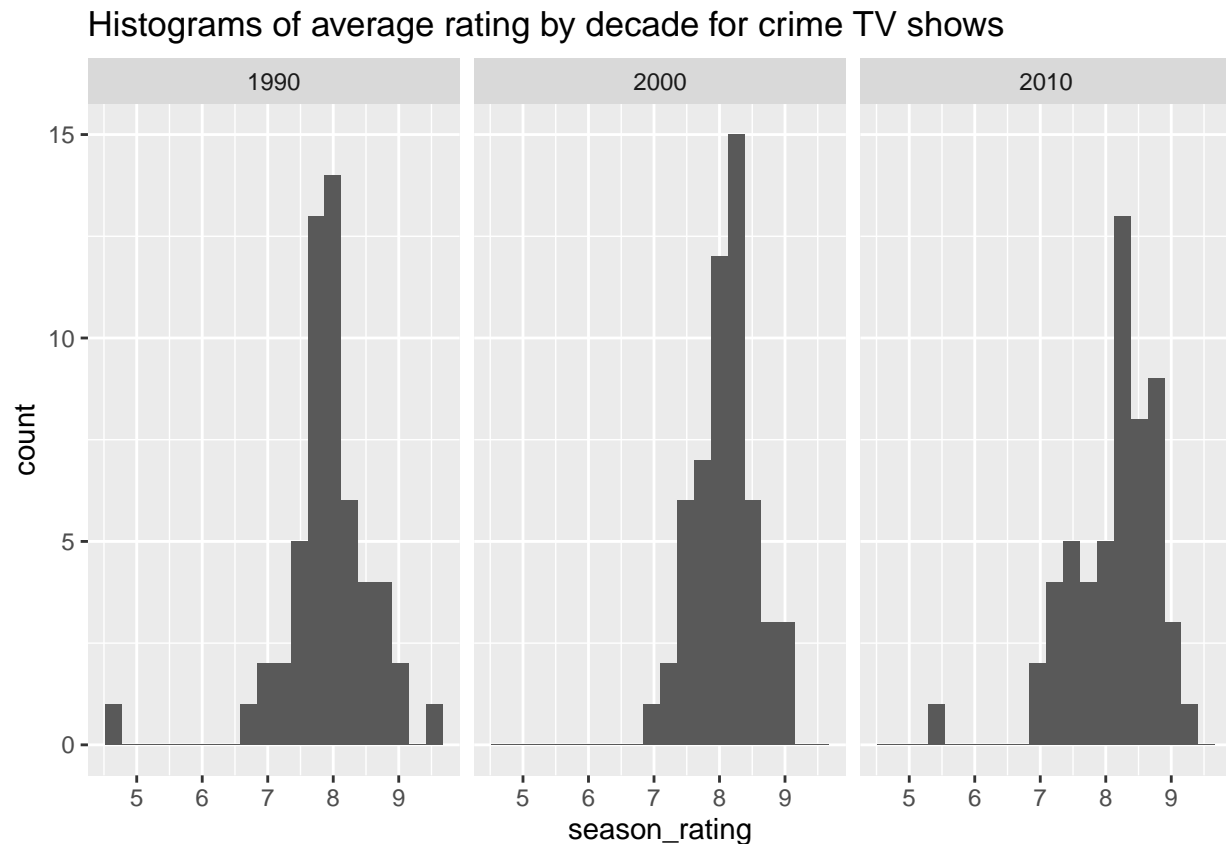
- $H_0$ :  $\mu_{1990} = \mu_{2000} = \mu_{2010}$ : The mean season rating for crime shows are the same across different decades. I.e. Different decades does not have effect over season ratings
- $H_1$ : at least one mean is different from the others: The mean season rating for crime shows are different across different decades. I.e. Different decades has at least some effect over season ratings

### Question 1.c

```
# Side by side box plots
crime_show_data %>% ggplot(aes(x = decade, y = season_rating)) +
  geom_boxplot() + ggtitle("Boxplots of average rating by decade for crime TV shows")
```



```
# Facetted histograms
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")
```



The box plot provides a better visualization of the data because it shows comparison across three decades' basic statistics (maximum, minimum, quartiles, median) side by side.

On the other side, with the histograms, it is harder to tell which decade has a higher rating because it only provides visualization over frequencies within each decade, and provides a relatively poor visualization for comparing between different decades.

One improvement for the box plot could be to cleaning the data before plotting. In the plot, we observe that there are some outliers, especially with 1990 and 2010. Removing those outliers may provide a even better visualization.

According to the box plot, we can see that the boxes are roughly on the same level. There is no sign of extremely skewed data except for some outliers, so their means are similar to the median (all around 8). Thus it does not suggest a significant difference between the means.

### Question 1.d

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

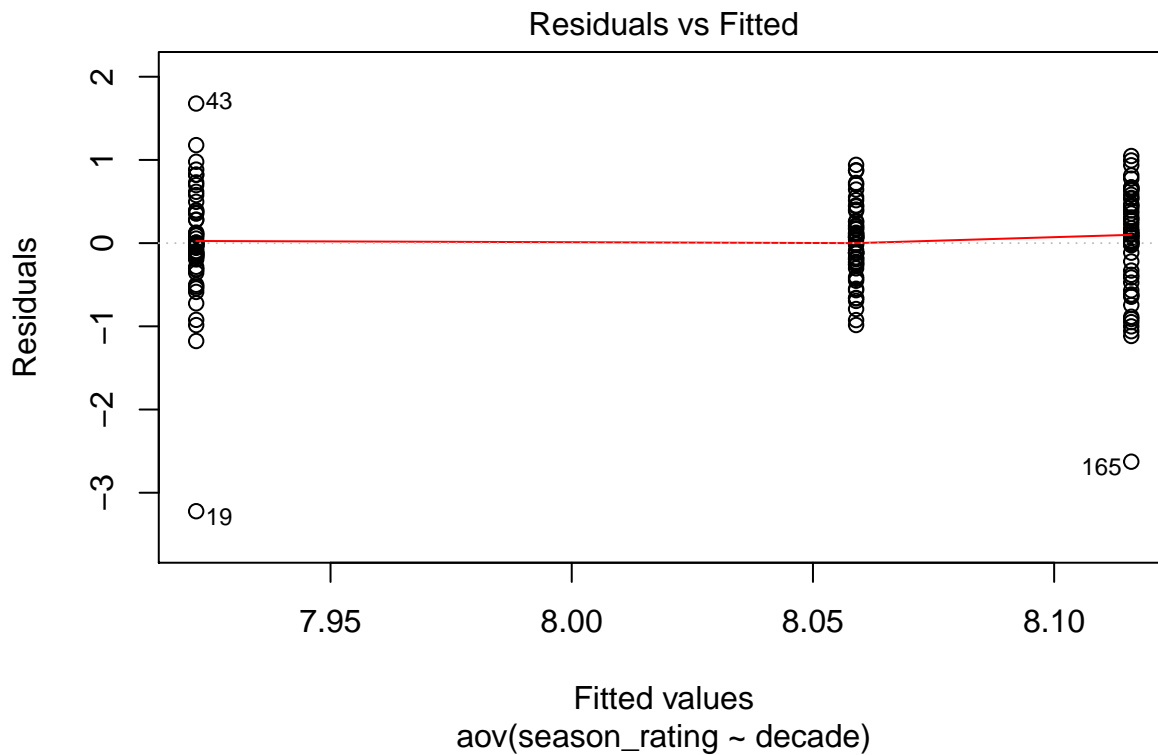
summary(one_way_anova)
```

```
##              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
```

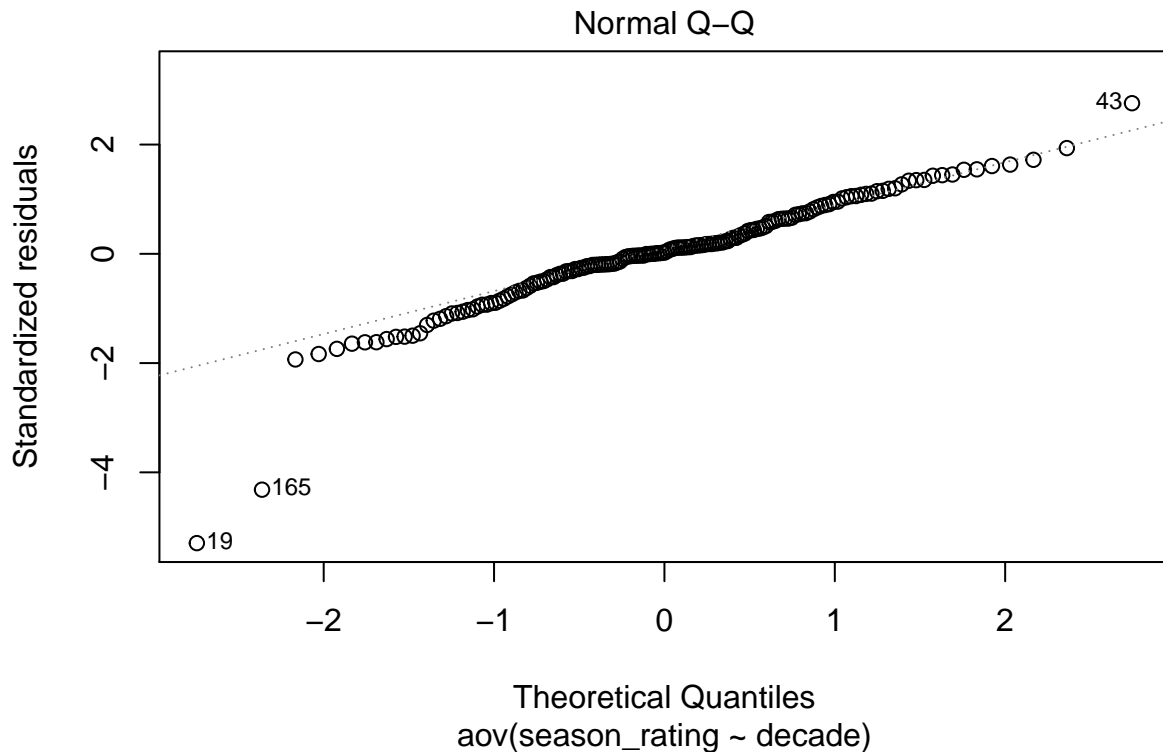
From the one way anova we can see that the p-value for the F-test is 0.238. We can interpret this as: The probability of observing the current sample given the assumption that the three decades having the same mean season ratings is 0.238.

### Question 1.e

```
plot(one_way_anova, 1)
```



```
plot(one_way_anova, 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
```

The first plot is the Residual vs Fitted plot. The plot shows that except for some outliers, the residuals are roughly randomly scattered around the 0-line, and does not indicate any pattern. This shows that the data follows a linear relationship, have equal error variances, and have a few outliers.

The second plot is the normal q-q plot. From the plot, we can see that except for points 19, 165, and 43, the points form a relatively straight line, indicating that the data follows a normal distribution with a few outliers.

From the standard deviations, we calculate that the ratio of the largest within-group and biggest within-group variance estimate is  $\frac{0.480}{0.203} = 2.365 < 3$ . According to the rule of thumb from Dean and Voss, the assumption for equality of variances is satisfied.

## Question 1.f

```
lm1 = lm(season_rating ~ decade, data = crime_show_data)

summary(lm1)

##
## 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
```

The linear model can be expressed as

$$y = \beta_0 + \beta_1 x_{2000} + \beta_2 x_{2010}$$

where  $y$  is the season rating,  $\beta_i$ s are the coefficients,  $x_{2000}$  is the indicator variable for decade 2000, and  $x_{2010}$  is the indicator variable for decade 2010.

$\beta_0$  is the intercept of the regression line, which is equal to the sample mean for decade 1990.

$\beta_1$  is the amount of score increase when the indicator variable  $x_{2000}$  is set to 1.

$\beta_2$  is the amount of score increase when the indicator variable  $x_{2010}$  is set to 1.

Then the sample mean for decade 1990  $\hat{\mu}_{1990} = \beta_0 = 7.9222$ .

The sample mean for decade 2000  $\hat{\mu}_{2000} = \beta_0 + \beta_1 = 7.9222 + 0.1368 = 8.059$ .

The sample mean for decade 2010  $\hat{\mu}_{2010} = \beta_0 + \beta_2 = 7.9222 + 0.1938 = 8.116$ .

## Question 2

### Read Data

```
# Read the crime show ratings data
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
# get rid of 9, 10 year olds and missing age and race
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)
```

	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))
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

	est	0.5 %	99.5 %
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 2.a

The statistical model that corresponds to `smokeModel` is

$$Y_i \sim \text{Binomial}(N_i, \mu_i)$$

$$h(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = X_i^T \beta$$

where

- $\mu_i$  is the sample mean  $E(Y_i)$
- $N_i$  is the population size
- $h(\mu_i)$  is the logit link function

Within  $X_i$ , there is an intercept of all 1's, a numeric variable of age centered around 16, and indicator variables for Region (Rural / Urban), Race (Black, Hispanic, Asian, Native, Pacific), and Sex (Male, Female).

## Question 2.b

For the baseline prob row,

- The value under `est` refers to the probability of observing the subset of individuals who are 16-year-old urban white males that have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days.
- The value under 0.5% and 99.5% represents the 99% confidence interval for the estimated probability

## Question 2.c

```
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
```

```
##      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	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

The claim that rural white males are the group most likely to use chewing tobacco is likely to be true. This is because the confidence interval of rural white males do not overlap with any other group, and have the highest fit value.

The claim that less than half of one percent of ethnic-minority urban women and girls chew tobacco can neither be confirmed or denied base on the data. Considering black and asian urban female, the fitted probability of them chewing tobacco is in 99% confidence interval of (0.13%, 0.42%) and (0.08%, 0.68%). We can see that the 99% confidence interval for black urban women is less than half of one percent. However, the 99% confidence interval for asian urban women includes 0.5% of its population. Thus we cannot conclude the claim is true.

## Question 3

### Read Data

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

## [1] "fiji"      "fijiFull"

# get rid of newly married women and those with missing
# literacy status
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)
```

	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

	Estimate	Std. Error	z value	Pr(> z )	est	0.5 %	99.5 %
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

```
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)
```

	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

```
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 3.a

The statistical model that corresponds to `fijiRes` is

$$Y_i \sim \text{Poisson}(\lambda)$$

$$h(\mu_i) = h(\lambda_i) = \log\left(\frac{\lambda_i}{O_i}\right) = X_i^T \beta$$

where

- $\mu_i$  is the sample mean  $E(Y_i)$
- $O_i$  is the offset term, which is the number of months since married.
- $h(\mu_i)$  is the log link function

Within  $X_i$ , there is an intercept of all 1's, and indicator variables for Age range of marriage (15-18, 0-15, 18-20, 0-22, 22-25, 25-30, 30+), ethnicity (Fijian, Indian, European, Part Eruopean, Pacific Islander, Routman, Chinese, Others) + literacy (Yes, No) + urban (Rural, Suva, Other Urban).

The intercept represents predicted rates of children per month for the subset of females who are between 15 to 18 years old, Fijian, rural and literate.

### Question 3.b

The model `fijiRes2` is nested within `fijiRes`.

`fijiRes2` can be viewed as `fijiRes` stripping ethnicity, and only retaining one age group indicator variable. The `marriedEarly` variable can simply be seen as the indicator variable for Age range of marriage 0-15. Thus the coefficients  $\beta$  for `fijiRes2` will have 6 less rows and is a subset of the coefficients for `fijiRes`.

### Question 3.c

From the `lmtest` result above, we can see that the model being compared is `fijiRes2` and `fijiRes`.

`fijiRes` takes into account for women of different ages and their literacy.

`fijiRes2` is nested within `fijiRes`, and does not account for women of different ages above 15 years old, and their literacy.

The p-value for the test is 0.3834. This is not significant enough to reject the null hypothesis that adding literacy and age range improves how well the model explains the data.

Thus the second claim is likely to be true.