A1

Ishaan Nagi #1002452525

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------ tidyverse 1.3.0 --
```

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

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
crime_show_data <- readRDS("~/Google Drive/Uni/Winter 2020/STA303/A1/crime_show_ratings.RDS")
```

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.

$$\text{SeasonRating}_i = \beta_0 + \beta_1 \cdot \text{Decade}_{2000} + \beta_2 \cdot \text{Decade}_{2010}$$

- ANOVA assumptions:
  - Errors are independent (observations are independent).
  - Errors are normally distributed with $\text{E}[\epsilon_i] = 0$.
  - Constant Variance (homoscedasticity), $\text{var}[\epsilon_i = \sigma^2]$.

Question 1b

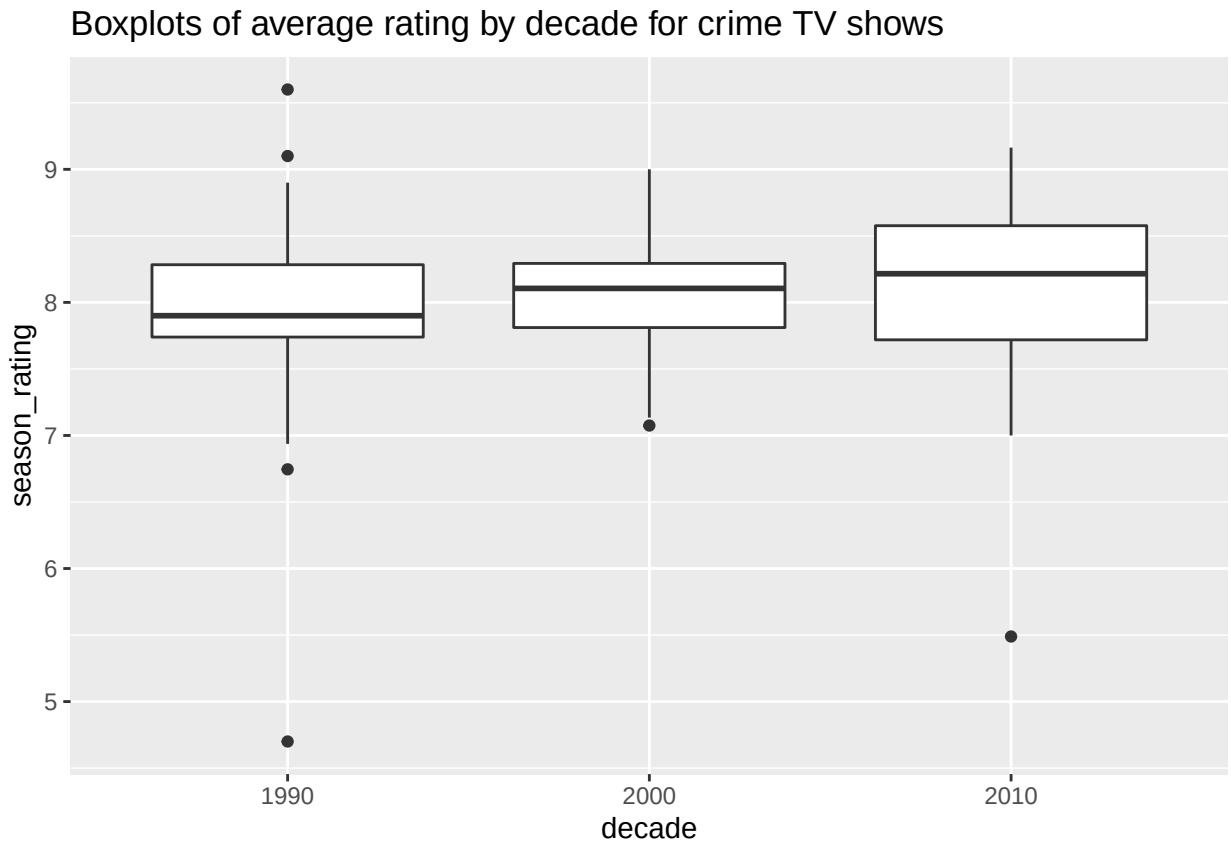
Write the hypotheses for an ANOVA for the question of interest in words. Make it specific to this context and question.

- H_0 : The average season rating for crime shows is the same between the decades 1990's, 2000's and 2010's. $[\mu_{1990} = \mu_{2000} = \mu_{2010}]$
- H_1 : The average season rating for crime shows vary from between the decades 1990's, 2000's and 2010's $[\mu_{1990}, \mu_{2000}$ and μ_{2010} not all equal].

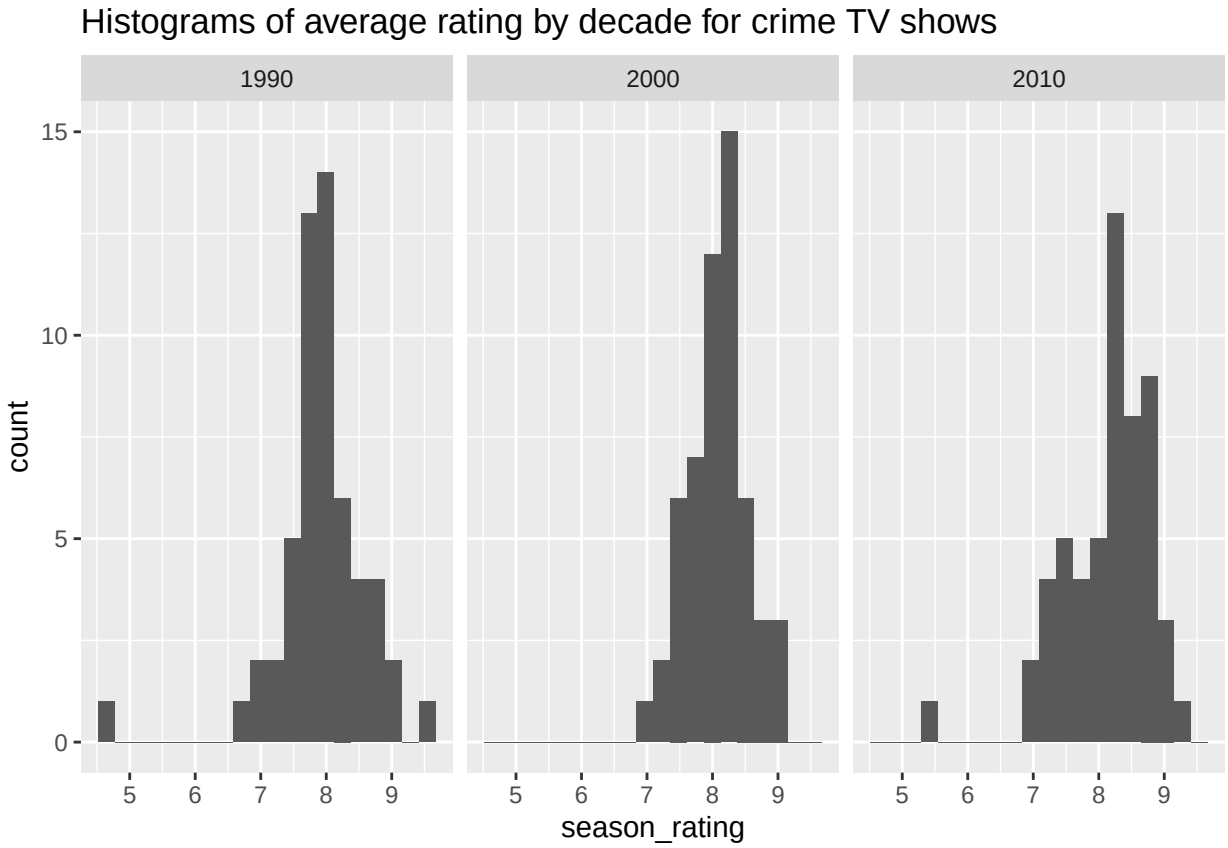
Question 1c

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?

```r
# 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")
```




```r
# 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")
```



- I prefer the boxplots since they indicate the summary of the data - the inter quartile range of the data, indicating the statistical dispersion between decades and where data points lie across groups. One way I would improve the boxplots, is by visually indicating skewness of the data.
- Purely based on the plots, the average ratings over the decade seem to be increasing.

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).

```
anova1 <- aov(season_rating ~ decade, data = crime_show_data)
summary(anova1)
```

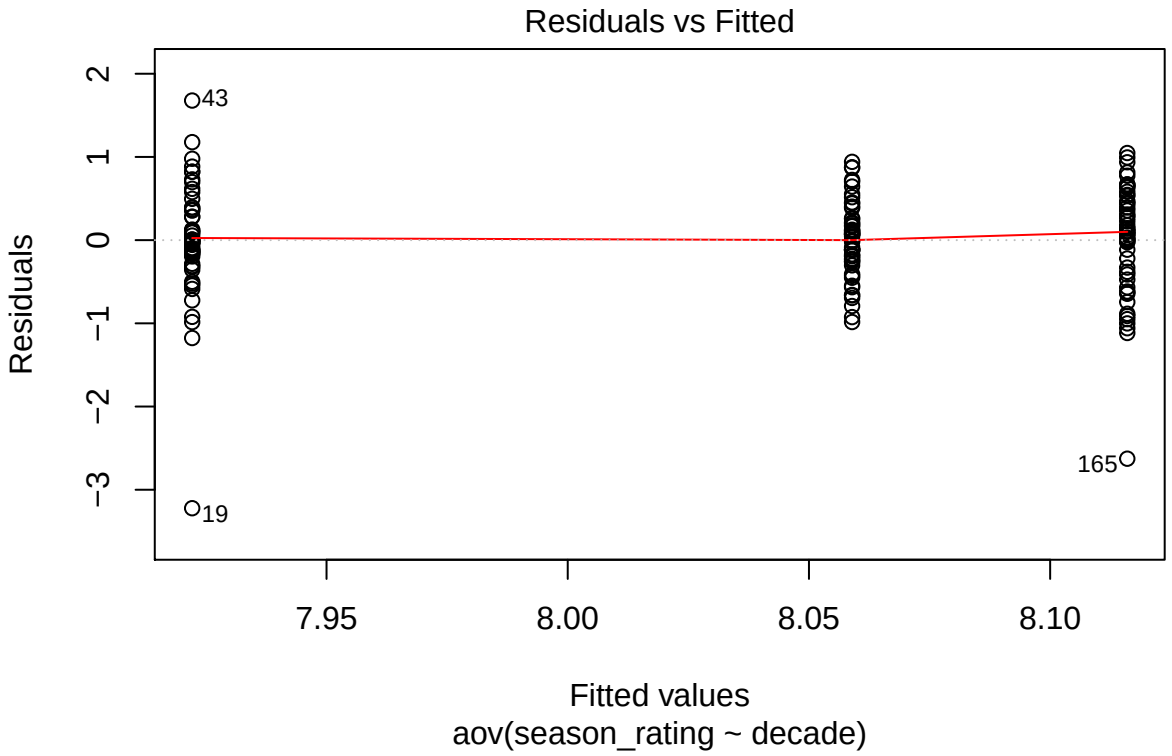
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## decade        2   1.09  0.5458   1.447  0.238
## Residuals   162  61.08  0.3771
```

- $\Pr(>F) > 0.05$, and thus we fail to reject the null hypothesis (H_0). [Based on 95% test]

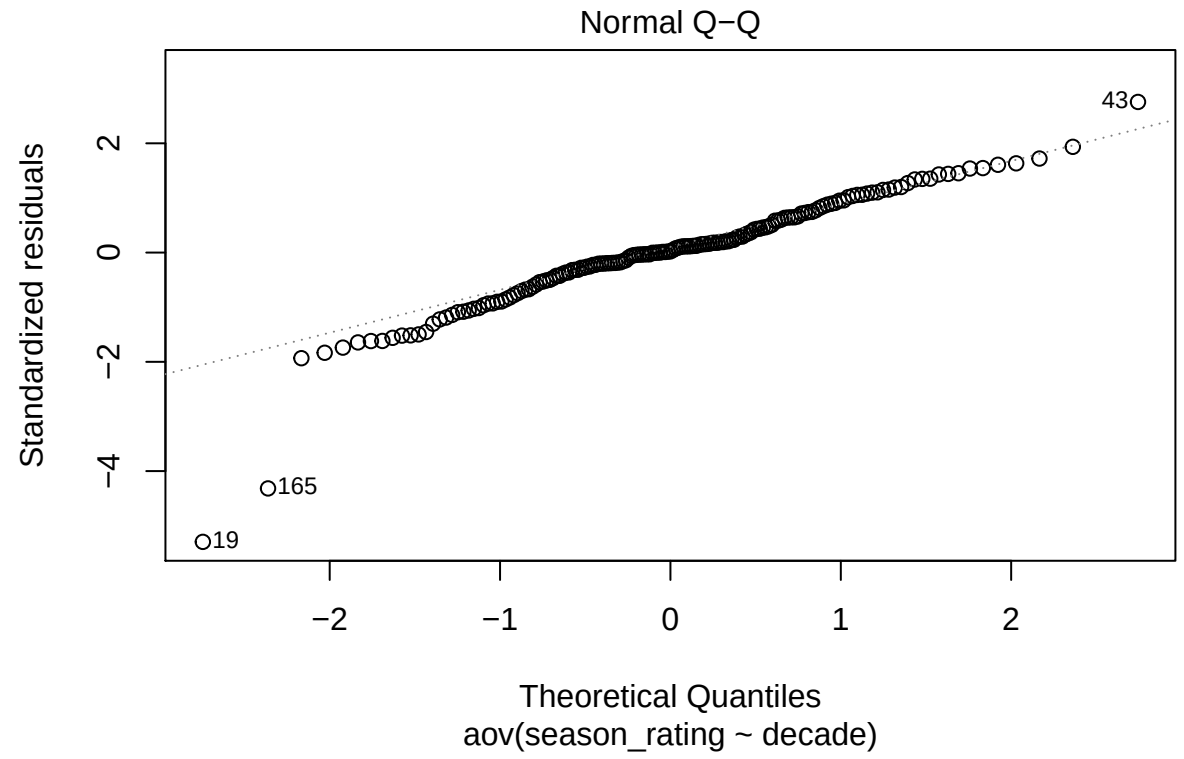
Question 1e

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. Note: there are specific tests for equality of variances, but for the purposes of this course we will just consider a rule of thumb from Dean and Voss (Design and Analysis of Experiments, 1999, page 112): if the ratio of the largest within-in group variance estimate to the smallest within-group variance estimate does not exceed 3, $s^2_{max}/s^2_{min} < 3$, the assumption is probably satisfied.

```r
# add your ANOVA object's name below (from Q1d)
plot(anova1, 1)
```




```r
plot(anova1, 2)
```



```r
# Note: this is the tidyverse way you can use a different method if you wish,
# but you're not required to write any code here
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
```

- Plot 1) Used to check the assumption of (homoscedasticity) constance variance of errors. The residuals are centered around the fitted line and do not seem to changing, thus agreeing with our assumption.
- Plot 2) We check the qq-plot to ascertain whether or not the normality of errors assumption holds. Majority of standaridized residuals seem to lie on the dotted line, agreeing with our assumption.

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, i.e., $\hat{\mu}_{1990s}$, $\hat{\mu}_{2000s}$, and $\hat{\mu}_{2010s}$.

```r
lm1 <- lm(season_rating ~ 0 + as.factor(decade), data=crime_show_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = season_rating ~ 0 + as.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|)
## as.factor(decade)1990   7.9222     0.0828   95.68   <2e-16 ***
## as.factor(decade)2000   8.0589     0.0828   97.33   <2e-16 ***
## as.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
```

- Based on the summary:
  - $\beta_0 = 7.992$, which is the season average rating for the decade 1990.
  - $\beta_1 = 8.0589$ which is the season average rating for the decade 2000.
  - $\beta_2 = 8.1160$ which is the season average rating for the decade 2010.

Question 2: Generalised linear models - Binary

Setup

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

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

- Get rid of 9, 10 year olds and missing age and race

```
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)), ]
#Set baseline age = 16
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))
```

```
## 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

Write down and explain the statistical model which smokeModel corresponds to, defining all your variables. It is sufficient to write $X_i\beta$ and explain in words what the variables in X_i are, you need not write $\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots$

- The smokeModel is trying to ascertain the probability that a subject has used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days, based on the attributes: ageC, RuralUrban, Race and Sex:
  - "ageC": is the difference in age as compared to the offset (16).
  - "RuralUrban": whether subject is located in a 'Urban' or 'Rural' area.
  - "Race": the ethnicity of the subject. ["white", "black", hispanic","asian", native" or "pacific"]
  - "Sex": Whether subject is Male or Female. ["M" or "F"]

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.

- The chosen baseline is a 16 years, urban, white and male.
- The "basline prob" refers to the basline odds that a subject from the basline population has used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days.

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.

- Based on the given table, it is expected that 149.3 out of a 1000 rural white males have used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days.
  - This outnumbers urban white males, urban hispanic males, urban black females and urban asian females with expected numbers (per 1000) 63.0, 31.5, 2.3 and 2.4 respectively.
  - Also note, the confidence intervals for the white, rural males does not overlap with any of the other categories.
  - Then we can say with reasonable certainty that rural white males are the group most likely to use chewing tobacco.
- Considering Females ethnic-minorities: it is expected that $(2.4+2.3 = 4.7)$ subjects used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days per thousand with Lower bound: $1.3 + 0.8 = 2.1$ and Upper Bound: $4.2 + 6.8 = 11$. [Following E$[X]$ property of Binomials]
  - Which is $= 0.47$ percent (per 100). Note: $0.47 < 0.5$.
  - However the 99% confidence interval ranges from 0.21% to 1.1%, so there isn't reasonable certainty that less than half of one percent of ethnic-minority urban women and girls chew tobacco.

Question 3: Generalised linear models - Poisson

Setup

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

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

```
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...
```

```
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

```
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...
```

```
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

Question 3a

Write down and explain the statistical model which fijiRes corresponds to, defining all your variables. It is sufficient to write $X_i\beta$ and explain in words what the variables in X_i are, you need not write $\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots$

FijiRes tries to ascertain the fertility rate of women as compared to the basline: married between the age 15 to 17, Fijian, literate and Rural).

- The model uses attributes 'ageMarried', 'ethnicity', 'literacy' and 'urban'
  - 'ageMarried' is the age of the subject at the time of marriage.
  - 'ethnicity' is the ethnicity of a subject. ["fijian", "indian", "european", "partEuropean", "pacificIslander", "routman", "chinese" and "other"]
  - 'literacy' whether or not the subject is literate.
  - 'urban' whether or not the subject is from "rural", "suva" or "otherUrban".

Question 3b

```
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
```

- Yes, the likelihood ratio test performed above is comparing nested models. 'fijiRes2' is a special case of 'fijiRes'.
- 'fijiRes2' wants to analyse the case where ageMarried' == '0to15' and the case where ageMarried =! '0to15', as compared to multiple levels of 'ageMarried' in 'fijiRes'. Also, 'fijiRes2' does not take into account 'literacy'.
- There is a constraint on the β_1, corresponding to 'marriedEarlyTRUE' which represents the change in the rate if a subject was married at the age of 14 or younger; The baseline case accounts for all other cases.

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 to influence their birth intervals. Supporters of each hypothesis are in agreement that fertility appears to be lower for women married before age 15, likely because these women would not have been fertile in the early years of their marriage.

Write a paragraph discussing the results above in the context of these two hypotheses.

Consider the fijiRes Model.

- Regarding Hypothesis 1:
  - The estimated fertility rate for women that are illeterate (literacy = no) is $0.984 \cdot 0.303$ (baseline fertility rate) $= 0.298152$, hence lower than literate women.
  - Disregarding women married from the age of 14 and younger, since both hyoptheses agree that fertility appears to be lower for women married before age 15, likely because these women would not have been fertile in the early years of their marriage.

  - All other fertility rates related to ageMarried, are greater than the baseline (1), suggesting that delaying marriage may not result in women choosing to have fewer children and increase the age gaps between their children.
  - However, note that all these estimates are insignificant. $Pr(>|z|)$ for most estimates is much greater than 0.05 (95% threshold). We cannot state any conclusions with high certainty. Also, the Likelihood ratio test in part b also tells us to consider the simpler model without literacy.
- Regarding hypothesis 2:
  - There is no way we can form a conclusion on fertility rates before and after 1974 or on the impact the lack of contraception without further data and analysis.