STA303 - Assignment 1

Winter 2020

Due 2019-01-31

This assignment is worth 5% of your final grade. It is also intended as preparation for Test 1 (worth 20%) and your final exam, so making a good effort here can help you get up to 33% of your final grade. You will get your feedback on Assignment 1 before Test 1.

You should be able to do Question 1 by the end of week 1, Question 2 by the end of week 2, and Question 3 by the end of week 3.

- **Question 1** uses data about the TV ratings for crime shows, (crime_show_ratings.RDS). You will need to download this from the Assignment 1 Quercus page.
- **Question 2** uses smoking data, (smoking.RData) and instructions for obtaining the data are at the beginning of the question.
- **Question 3** uses Fiji birth data, (fiji.RData) and instructions for obtaining the data are at the beginning of the question.

Note: You can use whatever packages are useful to you, i.e., tidyverse is not required if you prefer base R or something else. Just make sure you show which packages you are loading in the libraries chunk. Some example code in this assignment is shown with tidyverse functions.

Libraries used:

```
library(tidyverse)
```

Question 1: ANOVA as a linear model

A random sample of 55 crime shows was taken from each decade (1990s, 2000s, 2010s). The following variables are provided in `crime_show_ratings.RDS`:

Variable	Description
season_number	Season of show
title	Name of show
season_rating	Average rating of episodes in the given season
decade	Decade this season is from (1990s, 2000s, 2010s)
genres	Genres this shows is part of

Question of interest: We want to know if the average season rating for crime shows is the same decade to decade.

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.

Question 1b

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

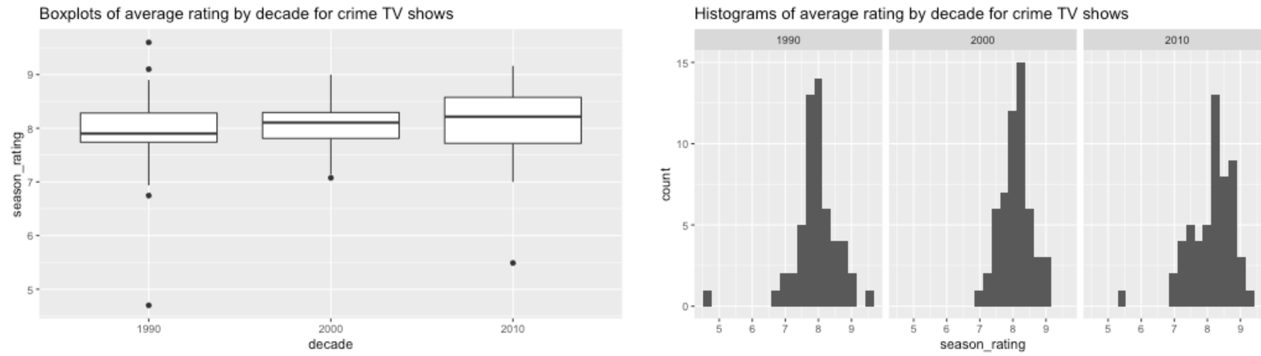
Question 1c

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
# load crimeshow data
# (have the .RDS downloaded to the same location your assignment .Rmd is saved)
crime_show_data <- readRDS("crime_show_ratings.RDS")

# 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")
```



(a) box

(b) hist

Figure 1: question 1c

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).

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(<name of anova object here>, 1)
plot(<name of anova object here>, 2)

# 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)
```

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 $\hat{\mu}_{2010s}$

Question 2: Generalised linear models - Binary

Data from the 2014 American National Youth Tobacco Survey is available on http://pbrown.ca/teaching/303/data, where there is an R version of the 2014 dataset `smoke.RData`, a pdf documentation file `2014-Codebook.pdf`, and the code used to create the R version of the data `smokingData.R`.

You can obtain the data with:

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

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

The `smoke` object is a `data.frame` containing the data, the `smokeFormats` gives some explanation of the variables. The `colName` and `label` columns of `smokeFormats` contain variable names in `smoke` and descriptions respectively.

```
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
```

Consider the following model and set of results

```
# 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

```r
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
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 $\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots$.

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.

Question 2c

If American TV is to believed, chewing tobacco is popular among cowboys, and cowboys are white, male and live in rural areas. In the early 1980s, when Dr. Brown was a child, the only Asian woman ever on North American TV was Yoko Ono, and Yoko Ono lived in a city and was never seen chewing tobacco. Consider the following code, and recall that a 99% confidence interval is roughly plus or minus three standard deviations.

```r
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
```

```r
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

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.

Question 3: Generalised linear models - Poisson

Data from the Fiji Fertility Survey of 1974 can be obtained as follows.

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

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

The `monthsSinceM` variable is the number of months since a woman was first married. We'll make the overly simplistic assumption that a woman's fertility rate is zero before marriage and constant thereafter until menopause. Only pre-menopausal women were included in the survey sample. The `residence` variable has three levels, with 'suva' being women living in the capital city of Suva. Consider the following code.

```
# 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
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), 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

```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

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 $\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots$.

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?

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.