STA035B Midterm 2, Winter 2024

Name _____

Student ID _____

Section _____

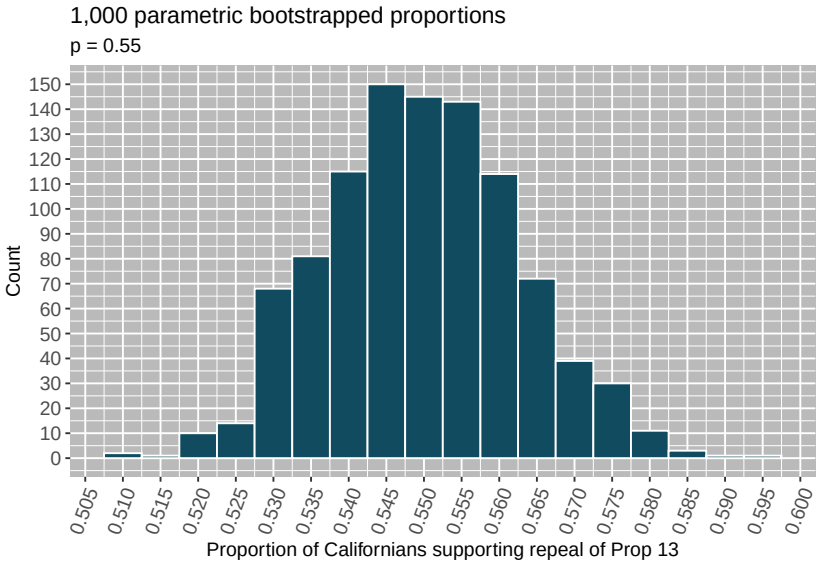
problem	points
1a	
1b	
2a	
2b	
3a	
3b	
3c	
4.1	
4.2	
5a	
5b	
total	

Problem 1

Suppose we are running a poll on whether or not a majority of Californians would like to repeal Proposition 13. The pollster takes a random sample of 1,563 people and reports that 58% of the people supported repeal. Suppose that in order for the repeal to go forward, 55% or more of voters must approve of the repeal.

(a) What are the null and alternative hypotheses for evaluating whether these data provide convincing evidence that, if voted on, Proposition 13 would be repealed in the US? (5 points)

(b) A parametric bootstrap simulation with 1,000 bootstrap samples was run and the resulting null distribution is displayed in the histogram below. Estimate the p-value using this distribution and conclude the hypothesis test in the context of the problem. (10 points)

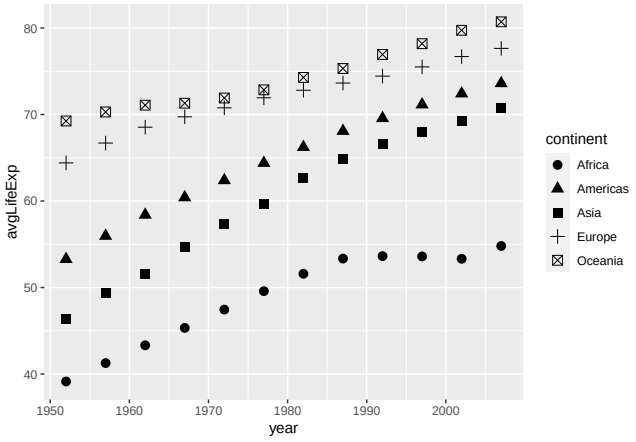


Problem 2

Consider the dataset `gapminder`, which has observations of the life expectancy, population, and gdp per capita (in dollars, $) for years between 1952 and 2007 for different countries. A sample of 5 rows from this table is given as follows.

```
# A tibble: 5 x 6
  country      continent  year lifeExp       pop gdpPercap
  <fct>        <fct>     <int>   <dbl>     <int>     <dbl>
1 Denmark      Europe     1982    74.6   5117810    21688.
2 Egypt        Africa     1982    56.0  45681811     3504.
3 Brazil       Americas   2002    71.0 179914212     8131.
4 Finland      Europe     1997    77.1   5134406    23724.
5 Burkina Faso Africa     1962    37.8   4919632      723.
```

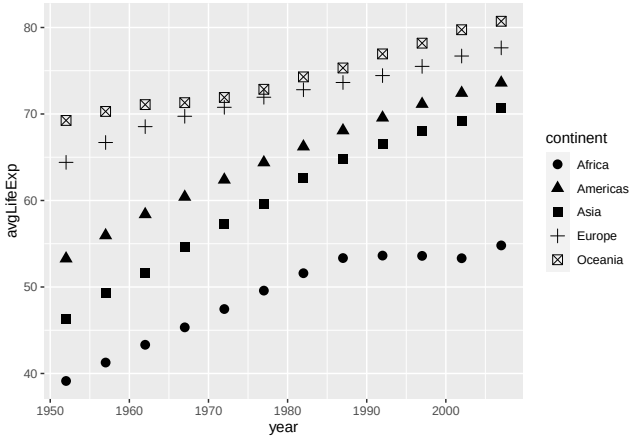
For each of the following, determine whether or not the code correctly computes the average life expectancy per year for each continent from 1952-2007 and makes the plot appearing below. If the code does not correctly do the computation and plot, explain what is wrong with the code.



(a) 3 points

```
gapminder %>%
  group_by(year) %>%
  summarize(avgLifeExp = mean(lifeExp, na.rm=TRUE)) %>%
  ggplot(aes(x = year, y = avgLifeExp)) +
  geom_point(aes(shape = continent))
```

(The text and plot from the previous page are copied below for your convenience.) Determine whether or not the code correctly computes the average life expectancy per year for each continent from 1952-2007 and makes the plot appearing below. If the code does not correctly do the computation and plot, explain what is wrong with the code.



(b) 3 points

```
gapminder %>%
  group_by(continent, year) %>%
  summarize(avgLifeExp = mean(lifeExp, na.rm=TRUE)) %>%
  ggplot(aes(x = year, y = avgLifeExp)) +
  geom_point()
```

Problem 3

Consider again the gapminder dataset from the previous problem.

Part (a), 10 points Suppose we wanted to produce a linear model which uses the gapminder data to try to predict the life expectancy of a country by its GDP per capita in 2007,

```
lm(lifeExp ~ gdpPercap,
   data = gapminder %>% filter(year == 2007))
```

```
Call:
lm(formula = lifeExp ~ gdpPercap, data = gapminder %>% filter(year ==
    2007))

Coefficients:
(Intercept)    gdpPercap
  5.957e+01    6.371e-04
```

Describe the linear model provided by the above code. Write the resulting formula for predicting life expectancy using the GDP per capita, and interpret each of the quantities in the formula.

Part (b), 3 points Suppose we consider a country whose GDP per capita is \$10,000. According to the linear model above, what is the predicted life expectancy?

Part (c), 5 points The variable "continent" in `gapminder` is a factor with 5 levels: Africa, Americas, Asia, Europe, and Oceania. Suppose you fit the following linear model to predict life expectancy by continent in the year 2007,

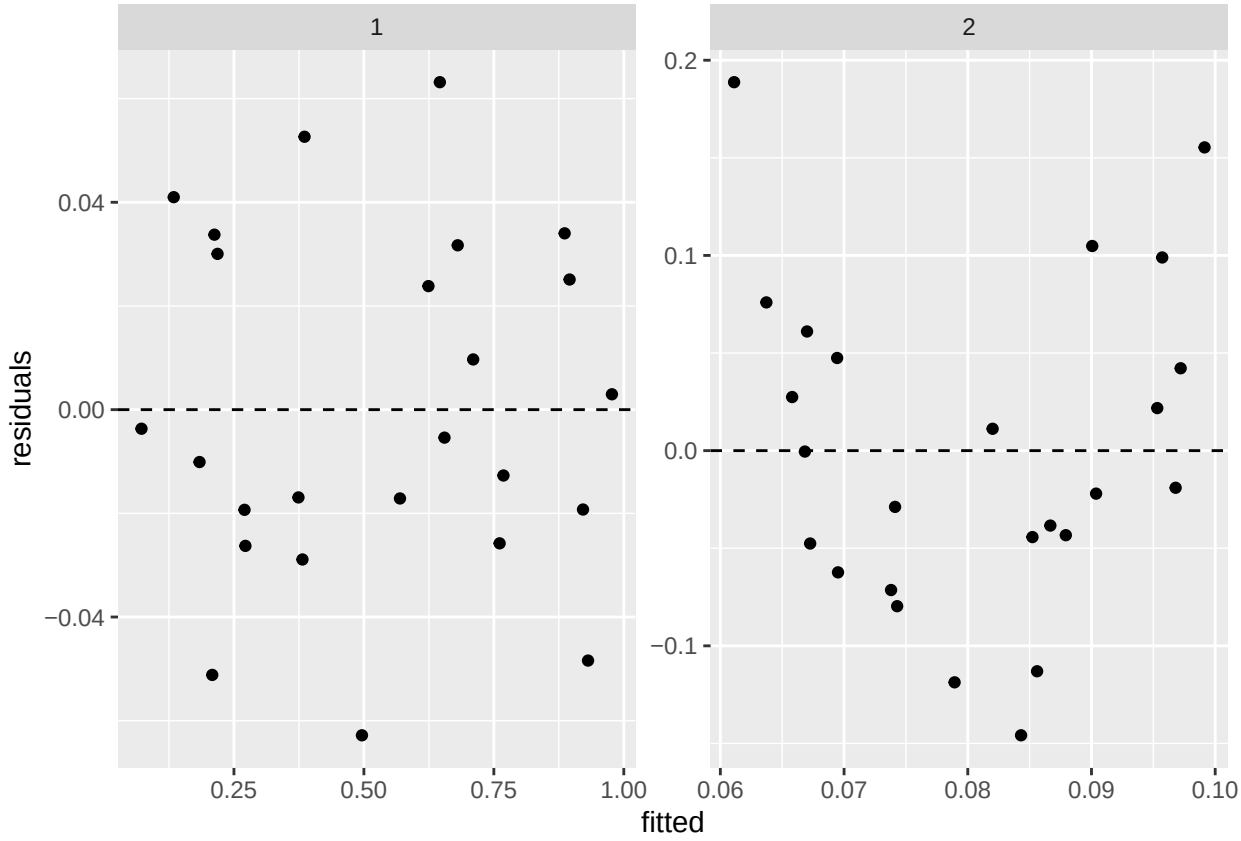
```
lm(lifeExp ~ continent,
   data = gapminder %>% filter(year == 2007))
```

How many independent variables are in the resulting linear model? Circle your answer and provide an explanation.

(A) 1
(B) 2
(C) 3
(D) 4
(E) 5

Problem 4

Consider the following residual plots.



For each of the residual plots above, describe which aspects of the plot (if any) indicates that a linear model is appropriate for modeling the data, and which aspects (if any) seem concerning for using a linear model.

Plot 1, 3 points

Plot 2, 3 points

Problem 5

Suppose we know that the scores on the midterm exam approximately followed a normal distribution, with an average score of 83 and a standard deviation of 5.

Part (a), 3 points Which of the following correctly computes the score corresponding to the 99th percentile?

(A) `pnorm(0.99, mean = 83, sd = 5)`
(B) `qnorm(0.99, mean = 83, sd = 5)`
(C) `qnorm(0.01, mean = 83, sd = 5)`
(D) `pnorm(0.01, mean = 83, sd = 5)`

Part (b), 3 points Suppose Sal gets a 70 on the exam. Approximately what percent of students scored above Sal on the exam?

(A) Between 50% and 84%
(B) Between 84% and 95%
(C) Between 97% and 99.8%
(D) Between 99.8% and 100%

Page for scratch work