Homework 1 - Stat 495

Justin Papagelis

Due Weds, Sept. 14th by midnight

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

- https://www.rdocumentation.org/packages/broom/versions/0.4.1/topics/augment, CASI 1.1
- https://rdrr.io/cran/mosaic/man/resample.html, CASI 1.4
- https://datascienceplus.com/fitting-polynomial-regression-r/, CASI 1.1
- https://en.wikipedia.org/wiki/Permutation_test, CASI 1.4

PROBLEMS TO TURN IN: Vis 1, Vis 2, CASI 1.1, CASI 1.4, Portfolio Reflection

Visualization Problems (Adapted from an assignment by Prof. Horton)

Vis 1 - Compelling

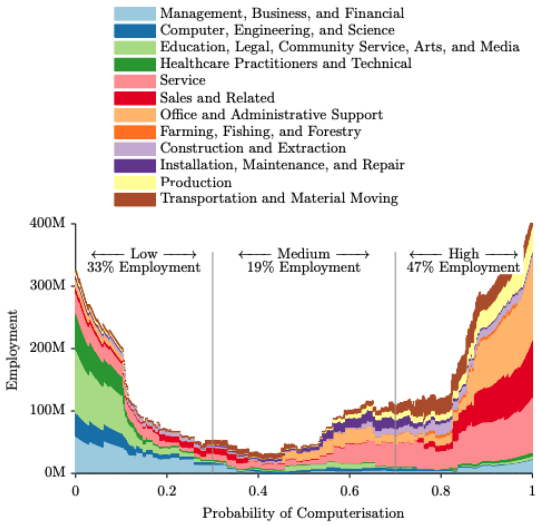


FIGURE 3. The distribution of BLS 2010 occupational employment over the probability of computerisation, along with the share in low, medium and high probability categories. Note that the total area under all curves is equal to total US employment.

Figure 1: Compelling Graphic

Frey, Carl Benedikt, and Michael Osborne. "The future of employment." (2013). Accessed 15 Sep. 2022. This graphic was found through Google Scholar

I found this graphic to be compelling because right from the start, it pops off of the page with lots of different colors. The graph contains many different components and is interesting to take in. In particular, I found it interesting how the different industries vary with the probability of computerization. Overall, this graphic combines a lot of different perspectives to look at while being visually appealing.

Vis 2 - Suboptimal

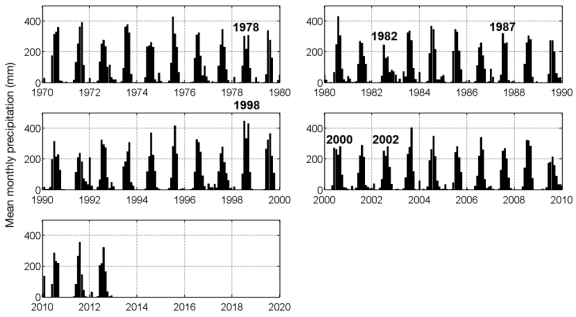


Fig. 3. Monthly mean total precipitation derived from weather stations of the coast region after applying the empirical orthogonal functional analysis for the period 1970–2012. The years with the occurrence of the uncommon midsummer drought are indicated in each plot

Vega-Camarena, José Pablo, et al. "Contrasting Rainfall Behavior between the Pacific Coast and the Mexican Altiplano." Climate Research, vol. 76, no. 3, 2018, pp. 225–40. JSTOR, https://www.jstor.org/stable/26626182. Accessed 15 Sep. 2022.

This graphic was found through JSTOR

I found this graphic to be suboptimal because it didn't really stand out to me as a reader and could definitely use some improvements. The type of graph that was used is not that helpful when compare the years because they do not all lie along a common axis. It is also difficult to see how distict the peaks are because there are not many axis labels.

CASI 1.1

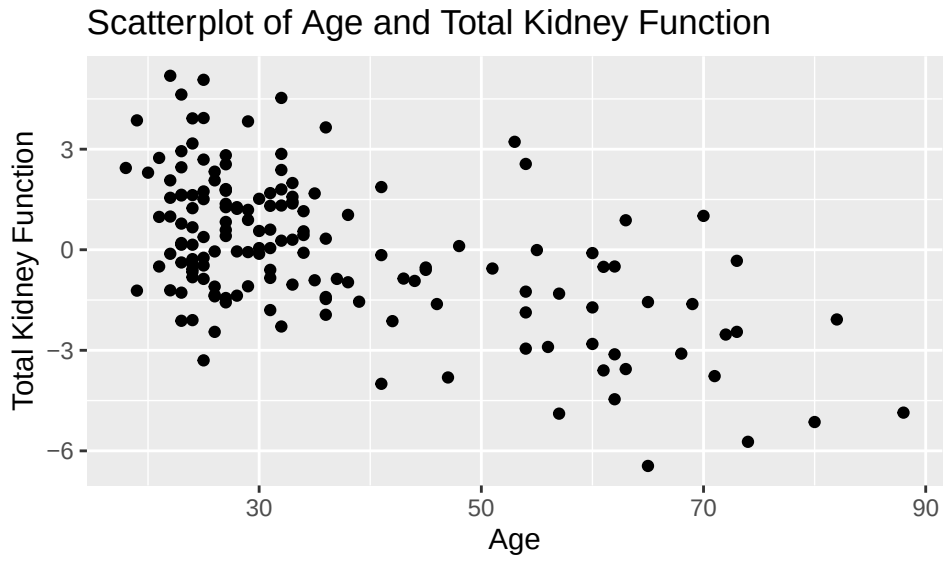
```
kidney <- read.table("http://web.stanford.edu/~hastie/CASI_files/DATA/kidney.txt", header = TRUE)
```

(a) Fit a cubic regression, as a function of age, to the kidney data of Figures 1.1 and 1.2, calculating estimates and standard errors at ages 20, 30, 40, 50, 60, 70, 80.

SOLUTION:

```
lm2 = lm(tot ~ poly(age, degree = 3), data = kidney)

ggplot(data = kidney, aes(x = age, y = tot)) +
  geom_point() +
  labs(title = "Scatterplot of Age and Total Kidney Function", x = "Age",
       y = "Total Kidney Function")
```




```
msummary(lm2)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.91e-04   1.45e-01    0.00     1.00
## poly(age, degree = 3)1 -1.56e+01   1.81e+00   -8.62  7.7e-15 ***
## poly(age, degree = 3)2 -4.76e-01   1.81e+00   -0.26     0.79
## poly(age, degree = 3)3 -1.08e-01   1.81e+00   -0.06     0.95
##
## Residual standard error: 1.81 on 153 degrees of freedom
## Multiple R-squared:  0.327,  Adjusted R-squared:  0.314
## F-statistic: 24.8 on 3 and 153 DF,  p-value: 3.86e-13
```

```
ages = data.frame(age = c(20, 30, 40, 50, 60, 70, 80))

pred = augment(x = lm2, data = kidney, newdata = ages, se_fit = TRUE) %>%
```

```
  rename("Age" = age, "Estimation" = .fitted, "Standard Errors"= .se.fit)

pred %>% knitr::kable(bookends = TRUE)
```

Age	Estimation	Standard Errors
20	1.249614	0.371693
30	0.508823	0.194624
40	-0.242607	0.259258
50	-1.014995	0.277694
60	-1.818660	0.334926
70	-2.663919	0.389150
80	-3.561092	0.668265

(b) How do the results compare with those in Table 1.1?

SOLUTION: Compared to the results in Table 1.1, the estimates using the cubic model are similar to the ones in Table 1.1. It does seem that the estimates tend to be lower than the other estimates for smaller ages using the cubic model. But as the ages increase, the estimated overall function of the kidneys is closer to the estimates of the first model. In general, the standard errors of the cubic model are higher than the standard errors of the estimates using the previous model.

CASI 1.4 - Slightly Modified

```
# Load and format data
leukemia_big <- read.csv("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
# says pictures from row 136
gene136 <- t(leukemia_big[136, ]) #says pictures from row 136
# Need to get ALL and AML tags in
type <- c(rep("ALL", 20), rep("AML", 14), rep("ALL", 27), rep("AML", 11))

# Set up dataset
leukemia <- data.frame(gene136, type)
leukemia <- rename(leukemia, gene136 = X136)
favstats(~ gene136 | type, data = leukemia)
```

```
##   type      min       Q1   median       Q3     max     mean       sd  n missing
## 1  ALL 0.210578 0.560344 0.732703 0.855127 1.63400 0.752479 0.275342 47       0
## 2  AML 0.324703 0.771426 0.967796 1.096250 1.42548 0.949973 0.243019 25       0
```

We want to see if there is a significant difference in mean gene expression for gene 136 for the ALL and AML groups.

(a) Record the means of the ALL and AML groups for the gene 136 data available for reference.

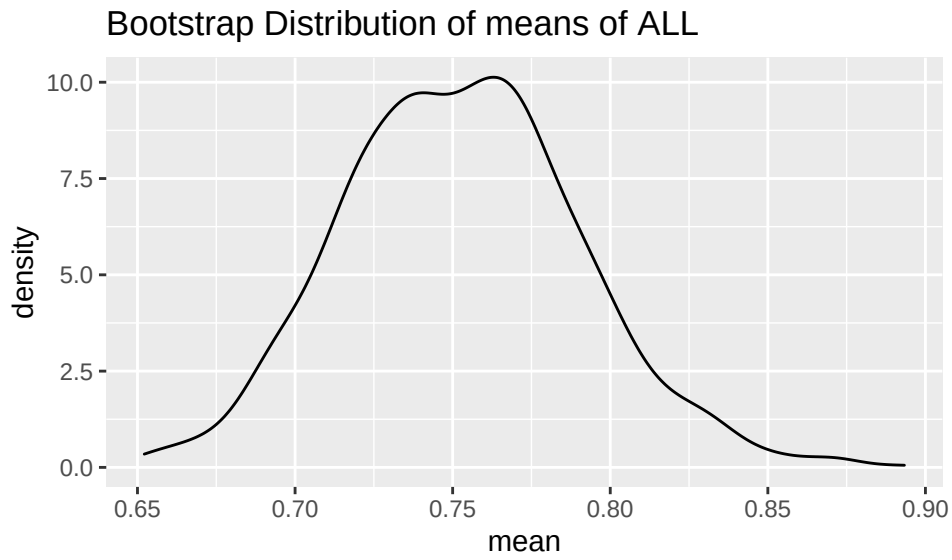
SOLUTION: The mean of the ALL group for the gene 136 is 0.752 and the mean of the AML group for the gene 136 is 0.950.

(b) Perform 1000 nonparametric bootstrap replications for the mean of ALL for gene 136. Describe the distribution of the resulting means. You can perform the bootstrap in any way you see fit (the functions do and resample might prove useful).

```
leukemia_ALL <- leukemia %>%
  filter(type == "ALL")

set.seed = 0
results_ALL <- do(1000)*mean(~gene136, data = mosaic::resample(leukemia_ALL, replace = T))

ggplot(data = results_ALL, aes(x = mean)) +
  geom_density() +
  labs(title = "Bootstrap Distribution of means of ALL")
```



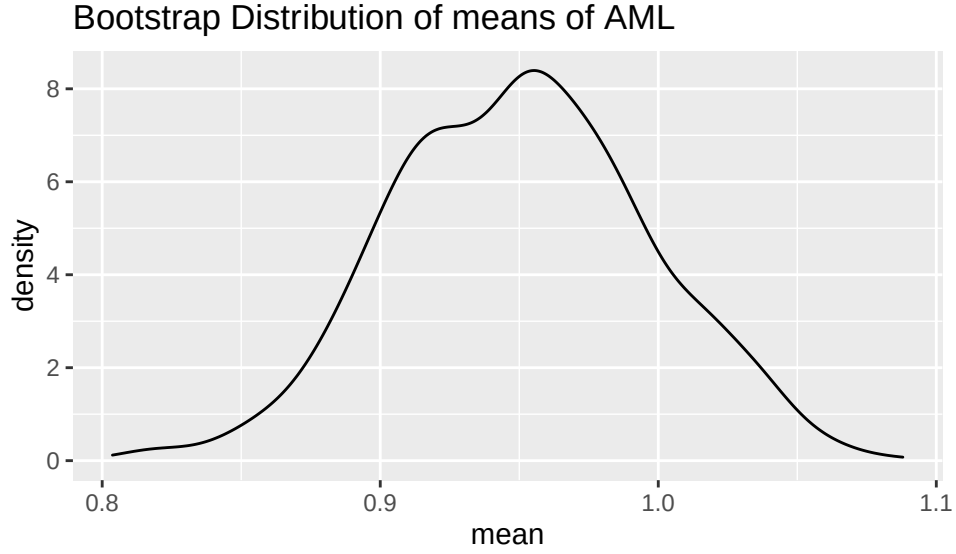
SOLUTION: The distribution of the means is unimodal and symmetric while centered around 0.76. The distribution appears to be relatively normal.

(c) Repeat (b) for AML.

```
leukemia_AML <- leukemia %>%
  filter(type == "AML")

set.seed = 0
results_AML <- do(1000)*mean(~gene136, data = mosaic::resample(leukemia_AML, replace = T))

ggplot(data = results_AML, aes(x = mean)) +
  geom_density() +
  labs(title = "Bootstrap Distribution of means of AML")
```



SOLUTION: The distribution of the means of AML is unimodal and relatively symmetric. It appears to be relatively normal. The distribution is centered around approximately 0.95.

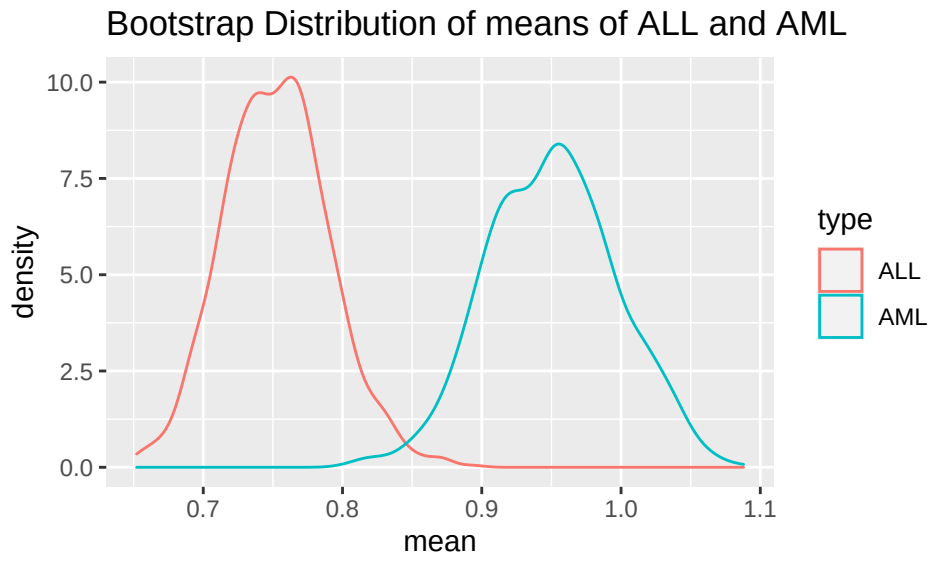
(d) Suggest an inference. In other words, what do your results in (b) and (c) suggest about whether there is a difference in means for the ALL and AML groups for gene 136?

```
results_ALL <- results_ALL %>%
  mutate(type = "ALL")

results_AML <- results_AML %>%
  mutate(type = "AML")

results_both <- rbind(results_ALL, results_AML)

ggplot(data = results_both, aes(x = mean, color = type)) +
  geom_density()  +
  labs(title = "Bootstrap Distribution of means of ALL and AML")
```



SOLUTION: From the density plots above, it appears that there is a difference in means for the ALL and the AML groups for gene 136. Specifically, a hypothesis could be that the mean of the AML group is greater than the mean of the ALL group for gene 136.

(e) Brainstorm an alternative way to approach the problem via a randomization/permutation test. Describe what you would do in a way that someone else could code it up. (You do not need to actually code this up, but you can if you want to see what the result is.)

SOLUTION: In order to compare the two groups directly, you could perform a 2-sample permutation test. First, we would calculate the difference in means between the two sample of gene 136 for AML and ALL. Next, we woudl combine the two samples into a single dataset. Then, we randomly assign the observations from the pooled dataset into either either AML or ALL. Then we calculate the difference in these means. We would repeat this until we have sufficient number of statistics (around 1000). This way, we can have lots of different permutations of ALL and AML for gene 136. The means we collected would be the distribution of of possible differences under a null hypothesis that the groups AML and ALL are the same. Then, a p-value can be calculated by finding the proportion of sampled permutations where the absolute value of the difference was greater than the difference in sample means from our original dataset. If we reject the null hypothesis, then we have significant evidence that the two groups, ALL and AML are not the same.

Portfolio Reflection

Look at our portfolio review and in-class activities. In a separate word or pdf document, in a few paragraphs, reflect on how the items in your portfolio demonstrate:

- how your statistical analytical skills have developed over time
- how your statistical writing skills have developed over time
- skills you have a solid grasp of (such as R code or visuals or regression)
- skills you would like to improve on

Then, set some goals for what you'd like to work on improving in future statistical reports. (Yes, you brainstormed some before, this is asking you to pick some to really focus on!)

Upload this portfolio reflection and goals document for future reports to your portfolio folder in your personal class repo.

Given what you are asked to include, I expect the document to have at least 3 paragraphs and contain at least 3 goals for future work.