STAT370 (Theoretical Statistics) Takehome exam 2

Keane Ng

May 13, 2020

This takehome exam is like a homework, albeit somewhat longer than usual. The only difference is that you cannot interact with anyone except me as instructor.

The assignment is due by midnight on Wednesday, May 13th via Gradescope. The deadline for late submission (with penalty) is the end of the day on Thursday, May 14th.

Honor code

Amherst College's Honor Code states that:

Every person's education is the product of his or her intellectual effort and participation in a process of critical exchange. Amherst cannot educate those who are unwilling to submit their own work and ideas to critical assessment. Nor can it tolerate those who interfere with the participation of others in the critical process. Therefore, the College considers it a violation of the requirements of intellectual responsibility to submit work that is not one's own or otherwise to subvert the conditions under which academic work is performed by oneself or by others.

For this assignment, you may not interact with anyone except the instructor. You may refer to your notes and other materials that you have used to date, including the documentation for functions and commands in R and sample solutions on the course GitHub repo or one or more of your group GitHub repos. All other resources that you consult must be listed below (or write "no additional resources used").

To signify your agreement with these stipulations please include your full legal name below.

Additional resources used: no additional resources used.

Electronic signature: Keane Ng

Assumptions for the one sample t-test

We are often warned of the perils of small samples and non-normality when using procedures like the one sample t-test. In this question you will explore the robustness of the two-sided one-sample t-test when the null hypothesis is that the population mean is 1 but the underlying data are exponential with rate parameter 1. Please assume an alpha level of 0.05 with n = 10 observations per group and carry out at least 10,000 simulations.

Your goal is to answer the following question: What is the impact of misspecification of the underlying data distribution (as described above) on the alpha level of the test. Is the test equally likely to reject on both sides? Does it maintain appropriate overall coverage?

Please preface your solution with your all needed code (to avoid distractions in the report to follow in the next section).

SOLUTION:

```r
set.seed(1000)
numSims = 10000
rate_param = 1
samp_Size = 10
sig_level = 0.05
perform_t_test_exp <- function(param = rate_param, size = samp_Size) {
  my_exp_Sample = rexp(size, rate = param)
  outcome <- t.test(my_exp_Sample, mu = 1, alternative = "two.sided")
  return (outcome)
}
perform_t_test_norm <- function(size = samp_Size) {
  my_norm_Sample = rnorm(size, mean = 1, sd = 1)
  outcome <- t.test(my_norm_Sample, mu = 1, alternative = "two.sided")
  return (outcome)
}
results_exp <- do(numSims)*perform_t_test_exp()
results_norm <- do(numSims)*perform_t_test_norm()
summary_exp <- results_exp %>% select(t, df, p.value, lower, upper)
summary_exp <- mutate(summary_exp, reject = (p.value < sig_level))
summary_exp <- mutate(summary_exp, reject.lower = (t < qt(0.025, 9)))
summary_exp <- mutate(summary_exp, reject.upper = (t > qt(0.975, 9)))
summary_norm <- results_norm %>% select(t, df, p.value, lower, upper)
summary_norm <- mutate(summary_norm, reject = (p.value < sig_level))
summary_norm <- mutate(summary_norm, reject.lower = (t < qt(0.025, 9)))
summary_norm <- mutate(summary_norm, reject.upper = (t > qt(0.975, 9)))
tab_norm <- table(summary_norm$reject)
tab_norm <- as.data.frame(tab_norm)
tab_exp <- table(summary_exp$reject)
tab_exp <- as.data.frame(tab_exp)
tab_norm$Freq2 <- tab_exp$Freq
tab_norm = tab_norm[, 2:3]
colnames(tab_norm) = c("Normal distribution", "Exponential distribution")
tab_norm <- rbind(tab_norm, c(sum(summary_norm$reject.lower),
                              sum(summary_exp$reject.lower)))
tab_norm <- rbind(tab_norm, c(sum(summary_norm$reject.upper),
                              sum(summary_exp$reject.upper)))
rownames(tab_norm) = c("Not rejected", "Rejected",
                       "Small t-statistic", "Large t-statistic")
test_exp_alpha <- binom.test(summary_exp$reject,
  p = 0.05, alternative = c("two.sided"), conf.level = 0.99)
rej_vector <- c(rep(TRUE, sum(summary_exp$reject.lower)),
                rep(FALSE, sum(summary_exp$reject.upper)))
test_exp_equal_sides <- binom.test(rej_vector,
  p = 0.5, alternative = c("two.sided"), conf.level = 0.99)
```

Our overall result is that we have evidence to suggest that the α level increases and the coverage rate hence decreases, since $1-\alpha$ is the coverage rate. Our 99% confidence interval for the α level when the data are distributed exponentially is (0.0984, 0.1144) and hence our 99% confidence interval for the new coverage rate is (0.8856, 0.9016). Furthermore, we have strong evidence to believe that the test no longer equally rejects on both sides; instead, it reject samples with low means much more often than sample with high means. A 99% confidence interval generated for the probability that a sample has a low mean given that it is rejected is (0.9314, 0.9665), so we have good reason to believe that low sample means are more often rejected than high sample means. The detailed explanation of how we arrived at these results is contained in the following report.

Present your findings in the form of a short report that would be comprehensible to a student who has completed an introductory statistics course. You may include at most one figure and one table to accompany the report. Be sure to pay attention to code quality, formatting, and typos. Include a 99% confidence interval for your estimation of the true rejection rate as part of your solution.

SOLUTION:

The one-sample t-test assumes that the underlying distribution of the sample data is approximately normal. Even in the case when the sample data are not normally distributed but the sample size is large enough, the central limit theorem implies that the results of the one-sample t-test are still approximately valid. However, when neither condition is fulfilled, that is, when the sample size is small and the underlying distribution is not normal, the significance level, i.e. the probability that we reject the null hypothesis when it is true, may no longer be maintained at $\alpha = 0.05$, and hence the coverage rate may no longer be equal to $100(1-\alpha)\% = 95\%$ as desired. By the duality of confidence intervals and hypothesis tests, this is equivalent to saying that when the null hypothesis is true, our test no longer rejects the $0.05 = 5\%$ most extreme values generated under the null hypothesis; we may reject too many or too few of them, and the confidence intervals we generate contain the true mean either too often or not often enough. To investigate, we simulate 10000 samples of size 10 where the data is generated by an exponential distribution but the null hypothesis is still true: the rate parameter is 1 and hence the population mean is indeed equal to 1. We conduct the one-sample t-test for each sample and note how many times it rejects the null hypothesis in favour of the alternative hypothesis that the true mean is not equal to 1. Since our data is generated under the null hypothesis, we should expect a rejection rate equal to the significance level of $\alpha = 0.05$; this would also mean that 95% of confidence intervals for the true difference between means generated by our test should contain the true value, that is, $\mu = 1$. For contrast, we also repeated the simulation with normally distributed samples. Below is a summary table of the results of our simulations:

```
kable(tab_norm)
```

	Normal distribution	Exponential distribution
Not rejected	9526	8938
Rejected	474	1062
Small t-statistic	231	1010
Large t-statistic	243	52

After simulating, we check each of the 10000 t-tests that were performed on the exponentially distributed data to see how many of them rejected the null hypothesis that $\mu = 1$. If the α level of the test is to be maintained at 0.05, 0.05 or about 500 of the t-tests should reject the null hypothesis, since the significance level of the test is the probability that the null hypothesis is rejected when it is true. Furthermore, we also check the number of t-statistics that were rejected because they were too small or too large. If the test is equally likely to reject on both sides, we should see that the number of t-statistics that were small enough to reject should be approximately the same as the number of t-statistics that were large enough to reject.

From the table above, we see that the rejection rate for the exponentially distributed data is $\frac{1062}{10000} = 10.62\%$, nowhere close to our desired 5%. This also means that for the exponentially distributed data, our empirical coverage rate is $100\% - 10.62\% = 89.38\%$, much lower than the 95% we anticipated. In contrast, the normally distributed data have a rejection rate of $\frac{474}{10000} = 4.74\%$, which is close to the expected 5%. How strong is this is as evidence that if the data are distributed exponentially, the α level, i.e. the probability of rejecting the null hypothesis given that it is true, is no longer 0.05, and hence the coverage is also no longer 95%? We can generate a 99% confidence interval for the true rejection rate based on our data by conducting a secondary hypothesis test. Since each t-test has either a reject or not-reject outcome, the number of rejections is a binomial random variable with 10000 trials and probability of rejection r which we wish to measure, and we will test the null hypothesis that $r = 0.05$ against the alternative hypothesis that $r \neq 0.05$. For this secondary test, we will adopt a significance level of 0.01.

```
signif(test_exp_alpha$p.value, 4)
```

```
## [1] 5.095e-113
```

We get an exceedingly small p-value (< 0.0001), which indicates that we have sufficient evidence that the true rejection rate is no longer 5%, so the coverage is also no longer 95%. We can generate a 99% confidence interval for the true rejection rate:

```r
round(test_exp_alpha$conf.int, 4)
```

```
## [1] 0.0984 0.1144
## attr(,"conf.level")
## [1] 0.99
## attr(,"method")
## [1] "Score"
```

Our 99% confidence interval for the true value of α is (0.0984, 0.1144), which indeed does not contain 0.05; we have significant evidence to suggest that the α level is higher than 0.05. This also yields us a 99% confidence interval for the new coverage rate: (0.8856, 0.9016); we have strong evidence to suggest the actual coverage rate is lower than 0.95. This is a problem because it suggests that our coverage interval is too small for our desired significance level of 0.05; we are no longer rejecting the most extreme 5% of samples generated under the null hypothesis. This is a problem because if we were to use the test to make conclusions about other samples that indeed have mean equal to 1 but the data are not normally distributed, we may end up rejecting the null hypothesis while it is true more often than we would like.

Finally, we check to see if the test rejects an approximately equal number of samples with means that are too high or too low. Since the t-statistic is calculated by $t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$, high values of the sample mean correspond to high values of t, while low values of the test statistic correspond to low values of t. Hence, we have to check if the t-test is rejecting an approximately equal number of high and low t-statistics. From our data, the test rejected 1010 t-statistics for being too low, and only 52 t-statistics for being too high. In contrast, the t-test on normally distributed data rejected 231 small t-statistics and 243 large t-statistics, which is a much more even distribution. Once again, to check how strong this is as evidence that rejection on either side is not equally likely, we can conduct a binomial test with the null hypothesis being that if the t-test rejects, the probability of a t-statistic being rejected for being too low is equal to the probability that a t-statistic is rejected for being too high, i.e. both have probability 0.5, versus the alternative that the probability is not 0.5.

```r
signif(test_exp_equal_sides$p.value, 4)
```

```
## [1] 3.394e-231
```

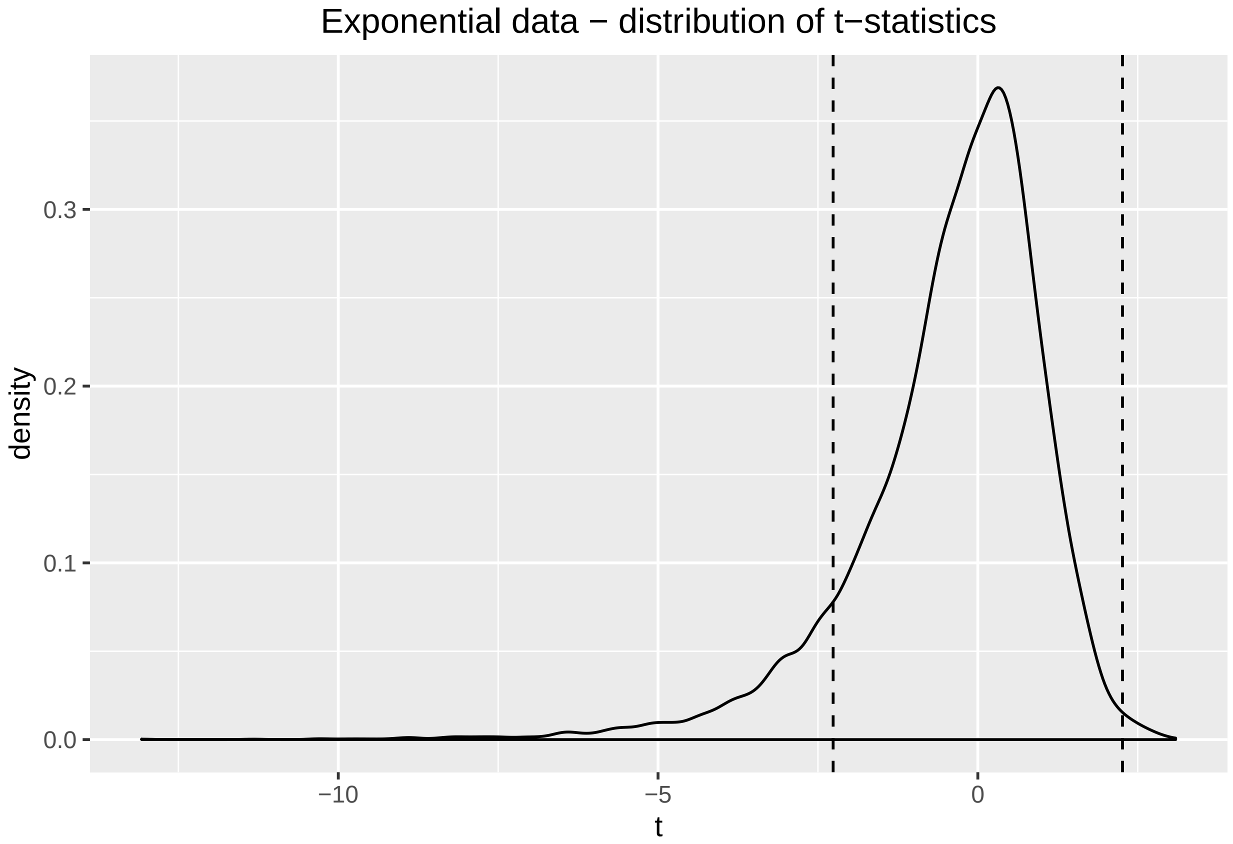
Again, we obtain an exceedingly low p-value (< 0.0001), which indicates that we have sufficient evidence that the probability of a rejection because of a low t-statistic is higher than the probability of rejection because of a high t-statistic. Hence, we have statistical evidence to conclude that there is no longer an equal probability of the t-test rejecting on both sides.

```r
round(test_exp_equal_sides$conf.int, 4)
```

```
## [1] 0.9314 0.9665
## attr(,"conf.level")
## [1] 0.99
## attr(,"method")
## [1] "Score"
```

A 99% confidence interval for the probability of the t-statistic being below the lower cutoff given that it is rejected is (0.9314, 0.9665), so it is exceedingly likely that the true probability of a t-statistic being rejected for being too low is much higher than the true probability that a t-statistic is rejected for being too high. A good way of visualizing this would be to create a density plot of the 10000 t-statistics generated by our simulation. In the figure below, we display the distribution of simulated t-statistics together with dashed lines that represent the cutoff values of t for each sample. If a sample has a t-statistic that is to the left of the left dashed line, it is low enough to reject; if the t-statistic is to the right of the right dashed line, it is high enough to reject. If rejection were equally likely on both sides, we should see that the area below the lower cutoff t-statistic is approximately the same as the area above the upper cutoff t-statistic.

```r
ggplot(results_exp, aes(x = t)) + geom_density() +
  geom_vline(xintercept = qt(0.975, 9), linetype = "dashed") +
  geom_vline(xintercept = qt(0.025, 9), linetype = "dashed") +
  ggtitle("Exponential data - distribution of t-statistics") +
  theme(plot.title = element_text(hjust = 0.5))
```



Indeed, significantly more of the t-statistics are distributed below the lower cutoff value than the upper cutoff value, meaning that we have evidence to suggest that the t-test is more likely to reject on the side of low t-statistics than high t-statistics. In other words, the t-test detects low sample means as being statistically significant too often, and does not detect high sample means as being statistically significant as often as we would like.

In conclusion, our simulation provides us strong evidence to believe that a non-normal underlying distribution can change the α level and hence the coverage as well. In this case, our empirical α level, i.e. the probability that we reject the null hypothesis when it is true, turned out to be 0.1062, much higher than the 0.05 we anticipated, which caused our coverage to be 89.38%, lower than desired. We found significant evidence to suggest that the t-test has an α level higher than desired therefore has a coverage probability is lower than expected; furthermore, there is also significant evidence to suggest that it does not have equal probability of rejecting on both sides but instead rejects more low t-statistics or sample means than high t-statistics or sample means.

ANOVA table

Use the R output below to create an ANOVA table:

```
msummary(lm(y ~ x, data = z))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.36327    2.78769  11.609   <2e-16 ***
age          0.01359    0.07643   0.178    0.859

Residual standard error: 12.53 on 451 degrees of freedom
Multiple R-squared:  7.007e-05, Adjusted R-squared:  -0.002147
F-statistic: 0.03161 on 1 and 451 DF,  p-value: 0.859
```

SOLUTION:

Using the output from the lm() function, we can see that the residual standard error, $RSE = 12.53 = \sqrt{\frac{RSS}{n-2}}$ where $n-2 = 451$ is the number of degrees of freedom; it must be the case that $n = 453$. Calculating the sum of squared errors yields $RSS = (n-2)RSE^2 = 70807.4$. We also know that $MSE = \frac{SSE}{n-2} = 157.00$. Now the F statistic is given by the output as 0.03161 so that $0.03161 = F^* = \frac{MSR}{MSE} \implies MSR = 4.9628$. Now MSR is the regression sum of squares, SSR, divided by the degrees of freedom of the intercept, which is just 1, so we have $SSR = MSR = 4.9628$. We now compute the total sum of squares $SST = SSR + RSS = 4.9628 + 70807.4 = 7.0812.4$. We verify that the R^2 value is indeed the multiple R-squared in the R output: $\frac{4.9628}{70812.4} = 7.008 \times 10^{-5} \approx 7.007 \times 10^{-5}$. Note that the p-value of the hypothesis test that $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ is 0.859, already given in the output. In particular, there is insufficient evidence to conclude that the dependent variable is truly dependent on the independent variable.

```
#ANOVA TABLE GOES HERE:
kable(table)
```

Source of variation	Deg. freedom	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	4.9628	4.9628	0.03161	0.859
Residuals	451	70807.4	157.0009		
Total	452	70812.4			

Power

a) Use at least 10,000 simulations to estimate the power of a test of $H_0 : \beta_1 = 0$ in a simple linear regression model if the alternative is that $\beta_1 = 4$ according to the following parameters:

- $\beta_0 = 2$
- $\alpha = 0.01$
- n = 12
- $\sigma = 1$
- the x values are evenly spaced from 0 to 10, and
- the linear model is correctly specified

SOLUTION:

```r
set.seed(1000)
numSim <- 10000
test_power <- function(beta_1 = 4, beta_0 = 2, alpha = 0.01, n = 12, sigma = 1,
                       x_lower = 0, x_upper = 10, num_x = 12) {
  x_vals <- seq(from = x_lower, to = x_upper, length.out = num_x)
  errs <- rnorm(num_x, mean = 0, sd = sigma)
  y_vals <- beta_0 + beta_1*x_vals + errs
  sample_data <- data.frame(cbind(x_vals, y_vals))
  my_model <- lm(y_vals ~ x_vals, data = sample_data)
  return (confint(my_model, c("x_vals"), level = 0.99)[1,1] > 0)
}
test_power_results <- do(numSim) * test_power()
x <- binom.test(test_power_results$test_power,
                alternative = c("two.sided"), conf.level = 0.99)
table(test_power_results)
```

```
## test_power_results
##  TRUE
## 10000
```

```r
signif(x$conf.int, 4)
```

```
## [1] 0.9995 1.0000
## attr(,"conf.level")
## [1] 0.99
## attr(,"method")
## [1] "Score"
```

Above, we carry out 10000 simulations of data that follow the statistical model $y_i = \beta_0 + \beta_1 x_i + e_i$ with $\beta_0 = 2, \beta_1 = 4, E(e_i) = 0, Var(e_i) = 1^2$. We see that the test of $H_0 : \beta_1 = 0$ versus the alternative $H_1 : \beta_1 \neq 0$ rejects the null hypothesis when it is false 10000 times out of 10000; our 99% confidence interval for the true power is (0.9995, 1). We have strong evidence to suggest that the true power is very close to 1.

Does the true value of β_0 matter?

SOLUTION:

The confidence intervals and hypothesis tests for β_1 are generated based upon the assumption that $\frac{\hat{\beta_1}-\beta_1}{s_{\hat{\beta_1}}}$ follows a t-distribution with $n-2$ degrees of freedom. We have that the estimated slope, $\hat{\beta_1}$ is unbiased,

$$E(\hat{\beta_1}) = \beta_1$$

$$Var(\hat{\beta_1}) = \frac{n\sigma^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

The true value of β_0 does not matter because the distribution of $\hat{\beta_1}$ does not depend on β_0.

Simpson's paradox

Create a synthetic (artificial) data set with three variables y, x1, and x2 such that:

- the coefficient on x1 in `lm(y ~ x1)` is positive, but
- the coefficient on x1 in `lm(y ~ x1 + x2)` is negative.

SOLUTION:

We create a hypothetical data set where there are two classes, class A with 15 males and 5 females, and class B with 5 males and 15 females. Both classes take a standardized test and the results are analyzed for correlation between gender and score, then correlation between gender and score. Let gender = 1 denote a male student and gender = 0 denote a female student

```r
set.seed(1000)
Gender <- c(rep(1, 20), rep(0, 20))
Class <- c(rep("A", 15), rep("B", 5), rep("A", 5), rep("B", 15))
Score <- c(rep(65, 15), rep(35, 5), rep(70, 5), rep(40, 15))
Score = Score + rbinom(40, size = 20, prob = 0.5)
scores <- data.frame(cbind(Gender, Score, Class), stringsAsFactors = FALSE)
lmGender <- lm(Score ~ Gender, data = scores)
lmGender$coef
```

```
## (Intercept)     Gender1
##       57.00       10.15
```

```r
lmOverall <- lm(Score ~ Gender + Class, data = scores)
lmOverall$coef
```

```
## (Intercept)     Gender1      ClassB
##   79.675000   -4.966667  -30.233333
```

In the first model, where we use gender as the only predictor of score, it seems as though having gender = 1 (male) correlates to higher test scores, since the coefficient of gender is positive. In particular, our model predicts that males score 10.15 points higher than females on average. However, this is actually because of the gender imbalances of the two classes; many more males are assigned to the high performing class than the low performing class, while more females are assigned to the low performing class. Using a model of Score ~ Gender + Class yields a coefficient of -4.967 for gender, meaning that after accounting for the gender distribution of classes, female students perform 4.967 points higher than males on average. In other words, females score worse than males when the sample is considered as a whole, but in each individual class, female students outperformed their male classmates.

Persistence

Take one homework problem you have worked on this semester that you struggled to understand and solve, and explain how the struggle itself was valuable. In the context of this question, describe the struggle and how you overcame the struggle. You might also discuss whether struggling built aspects of character in you (e.g. endurance, self-confidence, competence to solve new problems) and how these virtues might benefit you in later ventures.

SOLUTION:

In individual HW03, I was tasked with plotting an ecdf, histogram and normal probability plot for a given dataset. While this seems like a simple task, I confess that I spent an outsize amount of time and effort in getting the plots to look the way I wanted them to. Despite having experience in coding with several languages, I often struggle to understand the R documentation; compared to documentation for some other languages, I find it difficult to understand and often lacking in necessary information or detail. I was considerably frustrated with ggplot and other packages, but I did not want to give up and submit a half-baked solution, not only because it would not meet my personal standards, but also because I knew that even if I put off learning how to make the plots now, it would still be a necessary skill to master in the future. Eventually, I managed to get the plots to look just the way I wanted, with clear overlays on each plot for the normal probability distribution, allowing one to easily visualize whether the dataset was indeed approximately normally distributed or not. Persisting not only allowed me to produce work of a quality I was satisfied with, but more importantly allowed me to gain more confidence and competence in working with R packages. As I move on to further courses or projects that require data analysis skills, having R as an alternative to Python offers me the flexibility to choose whichever language best suits the task at hand.

Creativity

Give one example of a statistical idea from this class that you found creative, and explain what you find creative about it. For example, you can choose an instance of creativity you experienced in your own problem-solving, or something you witnessed in another person's definition or reasoning.

SOLUTION:

In individual HW08, I was given a dataset of heart rate and body temperature. After using linear regression to model the relationship between heart rate and body temperature for both males and females separately, I was tasked with testing whether the two different models had the same slope. After being puzzled for quite some time, I sought help from my instructor, who told me to use a combined model including an interaction term, rate $\sim$ temperature + gender + temperature $*$ gender. Initially, I did not understand the rationale behind this. However, after some Googling and much thought, it dawned upon me that this was a slick way to verify if the difference in slopes between the two groups is indeed statistically significant or not. Since the t-test already provides us a way to check if the coefficient of a given regression term is significantly different from 0, we can use it to check if the interaction term is significant, and hence determine whether the relationship between heart rate and temperature is conditional on gender. To me, this was a very creative way to arrive at the desired result with as little effort as possible; an alternative would be to conduct an F-test from scratch, which would be much more tedious for little or no benefit.