W241 Final Project

Jason Papale
papale47@ischool.berkeley.edu

Collin Cunningham
collincun@ischool.berkeley.edu

Marcus A. Streips
mastreips@berkeley.edu

Abstract

An experimental design for testing the effectiveness of ”keyword stuffing” in resumes is presented along with a discussion of the potential impact that Applicant Tracking Systems (ATS) may have on the treatment and outcomes. The impact of the COVID-19 pandemic on results is discussed. Due to the lack of responses as a result of the pandemic, a pivot to a new experiment on hiking water awareness is presented as part of a replication study.

1 Introduction

Companies see far more applicants with the rise of online applications. A job board posting can see hundreds or thousands of applicants. Because having a human scan the surfeit of applications is costly and time consuming, companies have turned to applicant tracking systems (ATS) which offer efficient methods to scan candidates for relevant skills and experience.

With the growing ubiquity of ATSs, techniques rumored to beat these systems have similarly risen. One such method is to add hidden, job-relevant keywords in one's resume. Using white, small font in the footer listing SQL and TensorFlow, for example, would increase the candidate's chances for a data science position even though they have no experience with the tools. Then after the ATS passes on the resume to a human recruiter, they are able to speak to the standard text in the resume and can avoid directly lying about their experience. We sought to test this theory by designing an experiment to examine the following question:

Does adding invisible job-relevant keywords at the bottom of your resume increase chances of being interviewed?

2 Study Design

Our study began by creating a fictitious applicant named Eric Currie with an artificially generated profile picture and randomly generated name. Eric was a recent graduate of Georgia Institute of Technology with a high GPA and solid internship experience pursuing a career as a data analyst. We created two iterations of his resume: one resume lists his experience while avoiding an excessive number of potential keywords such as programming languages and technologies, the other was identical to the first except it listed hundreds of keywords in white, 6-point font at the bottom.

Because so many ATSs require the user to reenter experience into a form, we were forced to pivot to a job board that is based around “one-click applications” that automatically forwards resumes along to employers–however, we found this to be misleading as it often forwarded further screening questions to the email address to complete the application which had the potential to taint our study. Thus, we created two otherwise identical ZipRecruiter accounts corresponding to the two resumes. We decided to apply to jobs in Berkeley, Atlanta, Washington DC, and Austin because of the wealth of job listings and/or minimal impact from Corona at the time of the study's inception. For each of these cities, we identified a given number of postings and then randomly selected half to be in the treatment and the other half to be in control. We initially gave each application a two week window to respond, but adjusted this on the fly due to the paucity of responses.

In total, 252 applications were completed for this experiment which corresponds to the requisite power calculations performed in our preparation phase. We had only one true response. It was impossible to perform any meaningful analysis with the sole response.

2.1 Sample Size Calculation

The sample size was determined using the G*Power statistical analysis tool (Faul, et al. (2007), (2009)).[1] After conducting some research on the probabilities of an applicant getting an interview request, we conservatively estimated that probability to be 5 percent (Smith, 2013).[2] Using this proportion as the baseline conversion rate, we calculated the number of samples we would need to achieve a Power $(1-\beta)$ of 0.80, the chance we would be able to detect an Average Treatment Effect (ATE) of 10 percent, using Fisher's Exact-Test for comparing two independent binomial populations.

2.2 Hazard-Ratio Calculation

Given that we had to enforce an arbitrary cut-off date for this experiment (2 weeks) and that the time for a company to respond to an applicant can vary, we would have to perform our hypothesis testing on right censored data (Sestelo, 2017). Additionally, we would like our model to reflect the fact that the chance of an applicant receiving an offer is time-dependent on when the applicant applies relative to when the job was posted. As we are more concerned about comparing groups than about estimating the employer response time, we would center our analysis on the hazard ratio (the hazard, in this case, being the probability of receiving a request for an interview).

The Cox Proportional-Hazards (PH) Regression Model (Cox, 1972) is defined as:

$$h(t, \mathbf{X}) = h_0(t)e^{\Sigma_{j=1}^{p}\beta_j X_j} \tag{1}$$

Where the hazard h defined at time t is defined by a vector of explanatory variable coefficients $\mathbf{X}$. The Cox PH model is a semi-parametric in that the distribution of outcomes are unknown even if the regression parameters (betas) are. As such the outcomes are estimated using maximum likelihood estimates such as Expectation-Maximization (E-M) (Dempster et al., 1977). The resulting estimated survival curve can be evaluated using graphical approaches or a goodness-of-fit (GOF) test (see, e.g., Grambsch et al., 1994 and Schoenfeld, 1982).

We can write the hazard ratio of treatment i and control j as defined by their associated regression parameters β_i and β_j as:

$$\widehat{HR} = \frac{\hat{h}_i(t|\mathbf{X}_i)}{\hat{h}_j(t|\mathbf{X}_j)} = \frac{\hat{h}_0(t)\exp(\hat{\beta}\mathbf{X}_i)}{\hat{h}_0(t)\exp(\hat{\beta}\mathbf{X}_j)} = \exp(\hat{\beta}(\mathbf{X}_i - \mathbf{X}_j)) \tag{2}$$

The $(1-\alpha)\%$ confidence interveal is therefore

$$\exp(\hat{\beta}(\mathbf{X}_i - \mathbf{X}_j) \pm z_{1-\alpha/2}\widehat{se}(\hat{\beta}(\mathbf{X}_i - \mathbf{X}_j)) \tag{3}$$

The results of the Cox PH Model would be evaluated as follows:

- $HR = 1$: no effect
- $HR > 1$: increase in the hazard ratio
- $HR < 1$: reduction in the hazard ratio

In order to test the significance of a variable or an interaction term in the model, we can use the Wald Test. The null hypothesis of the Wald test is that coefficient $\beta_i = 0$, where the test statistic is

$$Z = \frac{\hat{\beta}_j - 0}{Std.Err(\hat{\beta}_j)} \sim N(0,1) \tag{4}$$

2.3 Pilot Study

To avoid egregious, irreparable mistakes in our final experiment, we ran an initial pilot study of 10 applications filled out directly on companies' websites. We received one response asking for an interview. During this phase, we were forced to manually re-enter the information from our resume which would likely negate the scan of information on the PDF resume submitted along with the application (which contained the hidden keywords). Therefore, we pivoted to the ZipRecruiter model described above. We ran a second check after the pivot and submitted an application and received an interview.

We felt optimistic that we would have enough responses given that the resume had generated interview requests in the pilot and the fact that we could complete applications in rapid succession

[1] G*power software available for download at http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

[2] "Why only 2 percent of applicants actually get interviews", Workopolis, November 10, 2016 https://careers.workopolis.com/advice/only-2-of-applicants-actually-get-interviews-heres-how-to-be-one-of-them/ (Accessed April 12, 2020).

using ZipRecruiter. There are two flaws in this generalization. Firstly, filling out applications on a company website is far more direct than the one-click applications; many of these were links third party recruiters use to ingest candidates into their systems for later use. Secondly, it appears that postings at the top of the pool of search results from which we did the single test application are promoted meaning they are either an urgent need or used by multi-level marketing schemes where recruiting is equivalent to sales. We fell victim to the latter, realizing our mistake too late.

2.4 Experiment

Designing an experiment in today's automated job market was a challenge given the layer of technology that exists between the applicant and the hiring manager. This technology has grown out of the need hiring managers have to be able to effectively and efficiently sort through the mass of applications they may receive for every job they post. The Wall Street Journal reports that resume screening software use is used by more than 90% of companies with more than 50,000 employees (98% for Fortune 500 companies)(Weber, 2012). [3] Experts estimate that, as of 2018, nearly 40 percent of all employers in the United States use an applicant tracking system (ATS) to screen candidates for their job openings.[4] Applicant tracking systems reject 75 percent of candidates, with 62 percent of companies using ATS reporting that some qualified candidates are being automatically filtered out of the vetting process by mistake according to a joint CareerArc/Future Workplace survey.[5]

This technological layer acts as an intermediary, filtering the data the manager receives, altering the experimental treatment in the process. For this reason, ensuring that each subject received the same treatment was a challenge. To address this concern, the team adopted a restricted admission strategy, limiting the job opportunity pool to only those available on one job application service (ZipRecruiter). This strategy did not control for the possibility that a particular employer had its own, separate system for accepting applications referred to it by the ZipRecruiter service. To help understand how the treatment may be altered by such systems, we conducted an analysis of the Applicant Tracking System (ATS) market as discussed below.

Knowing exactly how an application will be processed by a particular ATS system is challenging, given the proprietary nature of such systems. However, certain clues can be gleaned from service documentation, service portals and patent filings.[6] The job listing or application is one of the easiest ways to spot a specific ATS system.[7] For instance a link on the Starkbuck's career site, may display the following url:

```
>starbucks.com/careers/
corporate-careers/marketing
```

However, clicking the URL will result in being directed to an ATS url:

```
>starbucks.taleo.net/
careersection/1000222/jobsearch.
ftl?lang=en&LOCATION=200000194&
CATEGORY=200000027&src=CWS-11700
```

Careful scrutiny of the URL will help a researcher determine whether or not an application is being filtered by an ATS system. Knowing this information is useful for determining whether the application being submitted will be intercepted, processed and reviewed by an ATS system before reaching the employer. The top six ATS systems in the market (by market share) are:

1. Oracle/Taleo
2. Workday
3. SAP/SuccessFactors
4. IBM Kinexa/Brassring
5. ICIMS

[3]"Over 98 percent of Fortune 500 Companies Use Applicant Tracking Systems (ATS), Jon Shields, Jobscan Blog, June 20, 2018, https://www.jobscan.co/blog/fortune-500-use-applicant-tracking-systems/ (Accessed April 12, 2020).

[4]"The Secrets to Beating an Applicant Tracking System (ATS)", Trena Bell, CIO, April 17, 2018, https://www.cio.com/article/2398753/careers-staffing-5-insider-secrets-for-beating-applicant-tracking-systems.html (Accessed April 12, 2020).

[5]Bell, *Supra.*

[6]Example ATS patent filings: `https://docs.google.com/spreadsheets/d/1vQmhNv6CW9MHzl_CggXmzCF2iiRsM76yTmMtDWCmYQg/edit?usp=sharing`

[7]"How to Spot an ATS", Jon Shields, Jobscan Blog, September 14, 2017 https://www.jobscan.co/blog/spot-ats/, (Accessed April 12, 2020).

6. ADP

Understanding the ATS market enabled our team to conduct a patent literature review. Many interesting insights were gained on individual ATS systems. However, our patent search was unable to uncover any patent disclosures for the ZipRecruiter process. We were, therefore, forced to make assumptions based on what processes these systems share in common. ATS systems generally follow the following methodology when processing resumes for clients:

1. First, the software removes all formatting from the resume and scans for specific recognized keywords and key phrases.

2. Next, it sorts the content of the resume into individual categories like: Education, Contact Information, Skills, and Work Experience

3. Resumes with the highest scores relevant to the employer's specified keywords and phrases when combined with the necessary experience will be referred to the employer for further review.

During the processing phase, a parser extracts the raw text from the resume document, removes styling and formatting and labels the resulting tokenized text with one of the categories mentioned in phase 2. Once parsed and labeled, the ATS system uses algorithms to semantically match an applicant's resume to the job posting, assigning a relevancy score based on the employer's search terms and keywords, as well as the applicant's years of experience. These algorithms have become increasingly more sophisticated over the years, and are able to spot tactics applicants have used to try to "cheat the system", including: pasting keywords in white, pasting the entire job description in white, repeating the keywords as many times as possible, or adding a section labeled "keywords" where you stick various words from the job description.[8] It is likely, therefore, that our experimental treatment of listing hundreds of keywords in white, 6-point font at the bottom of the resume document would have been modified by the ZipRecuirter system. For these reasons, obtaining "compliance" from our subject (employers) in both treatment and control groups proved to be a challenge as did truly understanding the exact nature of the treatment the subject received. We had hoped to conduct a two-tail study to examine whether our treatment had either a positive or negative average treatment effect on the target outcome (number of job interviews offered). This would have allowed us to examine whether "gaming the system" had any appreciable impact (good or bad) on an applicant's chances of being contacted by a hiring manager. However, for the reasons discussed below, we did not have the opportunity.

2.5 Impact of COVID-19 on Experiment

The final and most impactful issue in our study was the onset of COVID-19 in the US. We ran our pilot study on March 1, 2020 when cases were 65 cases nationwide. We filled out applications for an 18 day period starting March 11. This period saw an increase in cases from 1,248 to 136,962. The Dow Jones Industrial Average saw record breaking drops losing 30% of its value. This economic trauma had devastating consequences on the job market and, in turn, our study. Almost 15% of selected job postings were actively removed from ZipRecruiter, and we were left with a single response – a response rate of 0.4%.

3 Replication Study

Due to the lack of responses in our experiment as a result of the COVID-19 pandemic, our team pivoted to assist another team run a replication of their experiment on desert hikers. The other team conducted a survey designed to test the impact of receiving probability data on the likelihood of water at various points on a desert trail and the amount of water carried on the hike. The experiment tested 12 different scenarios wherein the hike distances and the probability of water were varied. For instance, the baseline for the study was 5 miles to watering point, 5 miles to trail end and 90% probability of water being available (5/5/90). Examples of alternate scenarios tested by the team were : 5/5/10, 10/10/90 and 10/10/10. The team also collected data on covariates such as age, education and parental status. Using the Qualtrics survey platform, the team solicited over 1000 responses. Their analysis confirmed that

[8] "Beat the Robots: How to Get Your Resume Past the System & Into Human Hands", Regina Borsellino themuse, `https://www.themuse.com/advice/beat-the-robots-how-to-get-your-r-past-the-system-into-human-hands` (Accessed April 12, 2020).

there was a statistically significant impact that both distance and probability had on the amount of water carried. The amount of water carried increased proportionately with the amount of distance walked and the likelihood that water would not be available on the trail. Additionally, the study's authors concluded that there was a statistically significant impact of parental and martial status on the amount of water carried, which supported their hypothesis that married hikers, and hikers with children would be more risk averse and, therefore, more likely to care greater amounts of water on their hike.

In order to replicate this study, our team decided that we had to reduce the number of scenarios down to one given the amount of time we had left to collect data (about 2 weeks). In order to simplify the scenario and to generalize the results outside of the hiking context, we opted to focus our study on the following question:

To what degree does providing a probability (as opposed to a general statement) regarding the availability of a critical resource (e.g., water) affect its perceived necessity?

This design was setup to maximum the likelihood of compliance and reduce attrition, as the treatment was limited to answering one scenario. It was also designed to test the interesting psychological question about how the human brain interprets probability data when presented as a statistic (a percentage) versus a general statement of likelihood. The scenario, treatment and control used in the survey are shown in Figure 1.

In addition to the control and treatment survey questions, we included a default question block that asked each subject to answer a number of demographic questions including the following:

- Age
- Gender
- Height
- Weight
- Marital Status
- Parental Status
- Education
- Level of Physical Activity
- Hiking Experience (general)
- Hiking Experience (desert)
- Prior Existing Conditions

We did not collected any personally identifiable information (PII) such as name or contact information in order to avoid any issues with collected potentially protected and sensitive information relating to prior existing conditions. All survey participants were provided with a survey disclaimer that informed them of the purpose of the experiment, what data would be collected and how it would be used. Each survey subject confirmed they were at least 18 years old before they were admitted into the study. Any subjects that self-identified as having a prior existing condition that would preclude them from desert hiking were restricted from the study. Age, Height and Weight data were collected in order to calculate subject body mass index (BMI) as a general gauge of overall health.

The survey link was published as a post on the Facebook and LinkedIn social networks as well as distributed via email and discussion chats using an anonymous link (Figure 2). The experiment was conducted over 4 days and received 31 responses (14 control and 17 treatment) after being viewed over 545 on the LinkedIn Network (Figure 3).

3.1 Sample Size Calculations

Using the data from the prior study, we determined that the Cohen's D effect size was medium (0.48). Using this value, we determined an a priori sample size estimate using the given effect size, estimated α, required power $(1 - \beta)$ err prob., and group allocation 1:1 to estimate the required sample size for a t-test of difference of means between two independent groups to be **138** (Figure 4).

Given that we only had 31 responses, we know that any conclusions we draw will be underpowered. We calculated our post hoc power to be at most **0.25**, assuming a Cohen's D effect size of 0.48 estimated using data from the prior study (Figure 5). This meant that the probability of us detecting a statistically significant difference between the treatment and control groups was only 25 percent.

The power was even lower when a post hoc analysis was performed for a fixed model, linear

Scenario

You have been hiking for the past 4 weeks along the Pacific Crest trail through some beautiful, hot desert sections with lots of elevation change.

The past few days have been in the high 80s into the 90s, with the same forecast for the foreseeable future. And the water sources you have seen on your hike have been occasionally full. You are about to leave town again and have your backpack prepared for the trip. It weighs 35 pounds with all of your gear and food for the next 4 days.

You are sure there is water available at the spot you plan to camp tonight, but there is only 1 water source between here and there.

Control ("Usually")

It is 7.5 miles to the first water source, then another 7.5 miles to your camp site. According to the water report, there is usually water at the first water source.

Treatment ("75% Chance")

It is 7.5 miles to the first water source, then another 7.5 miles to your camp site. According to the water report, there is a 75% chance that there is water at the first water source.

Figure 1: Scenario with Treatment and Control

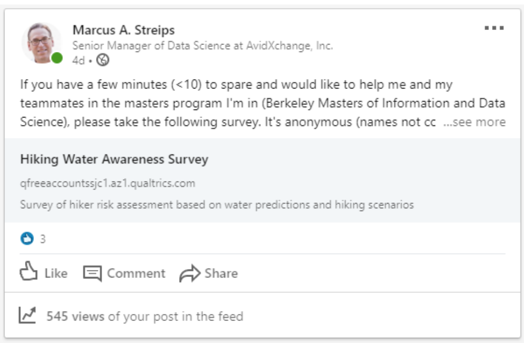

Figure 2: Survey Solicitation on LinkedIn

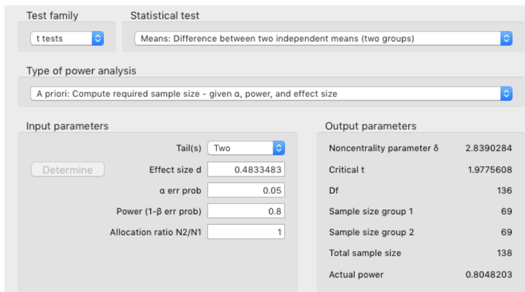

Figure 4: Required Sample Size Calculation

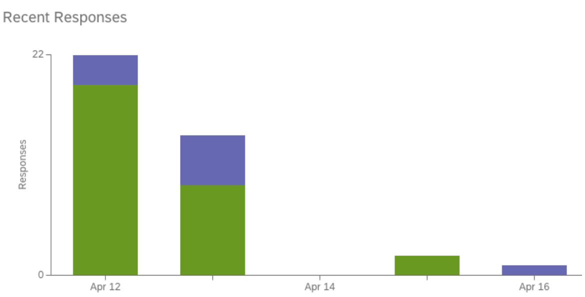

Figure 3: Distribution of Survey Responses

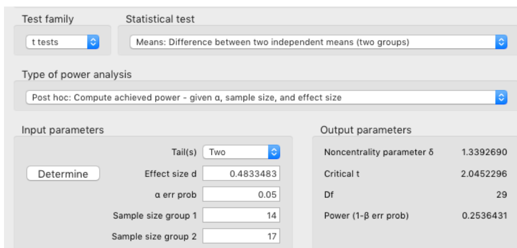

Figure 5: Post Hoc Power Calculation - T-Test

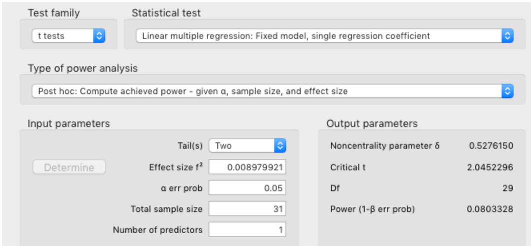

Figure 6: Post Hoc Power - Regression

```
	Welch Two Sample t-test

data:  Hiking_Data$Liters_of_Water_Control and Hiking_Data$Liters_of_Water_Treatment
t = 0.50561, df = 26.314, p-value = 0.6173
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9472276  1.5657150
sample estimates:
mean of x mean of y
 3.985714  3.676471
```

Figure 7: Welch Two Sample T-Test Results

regression with a single coefficient. Our regression had an R-squared value of 0.0089, from this we estimated an f^2 effect size of 0.0090. Using an α error estimate of 0.05, and a total sample size of 31 on a single predictor, the resulting power was calculated to be 0.08 (Figure 6). Such a lower power meant that we had very little chance (8 %) of detecting a statistically significant regression coefficient in our analysis.

3.2 Results of Replication Study

The experiment had an average treatment effect of -.31 liters when running Welch Two Sample T-Test, but the results were not statistically significant (p-value = .617) (Figure 7). The results were the same when conducting a simple linear regression with the amount of water carried as the target variable and the treatment flag as the observation (Figure 8). Based on these results, we fail to reject the null hypothesis that providing a probability regarding the availability of critical resources has an effect on the perceived necessity of those resources.

However, it is crucial to recognize that this finding is only applicable when the specific information provided is a 75% probability. Additional experimentation would be required to understand if this finding changes as the probability of having critical resources be available increases or decreases. Similarly, the findings of this experiment would support the hypothesis that, in this context, individuals may consider a 75% probability to be comparable to “usually.”

3.3 Covariate Analysis

A fixed effects, multiple regression analysis was performed on the covariates we collected as part of the default question block. We derived the BMI variable after calculating it from the height and weight data we collected from the survey using the

```
Call:
lm(formula = Hiking_Data$Liters_of_Water_Coalesced ~ Hiking_Data$Treatment_Flag)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1765 -1.3311  0.3143  0.9189  4.3235

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.9857     0.4475   8.906 8.53e-10 ***
Hiking_Data$Treatment_Flag1 -0.3092     0.6043  -0.512    0.613
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.674 on 29 degrees of freedom
Multiple R-squared:  0.008949,	Adjusted R-squared:  -0.02523
F-statistic: 0.2619 on 1 and 29 DF,  p-value: 0.6127
```

Figure 8: Simple Regression Results

formula [9]:

$$\text{BMI} = (\text{Weight} * 703/\text{Height}^2) \tag{5}$$

None of the coefficients from the multiple regression model were found to be statistically significant (p-value < 0.05) (Figure 9).

Given the small number of responses, we regrettably do not have enough power to support the findings of our analysis which are contrary to the findings of the original study. This conflict is most likely the result this replication study’s small samples size rather than any experimental error on the part of the original experiment we were trying to replicate.

4 Conclusions

It is our belief that even with our mistakes in performing this experiment, we would have had enough responses to perform a meaningful analysis and speak to our research question were it not for COVID-19. Personally, we know of many businesses that have frozen hiring during this period and many more that have begun layoffs. While it is unfortunate that our study concluded

[9] ”Calculating BMI Using the English System”, https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page5_2.html (Accessed April 26, 2020)

in this manner, this failure is overshadowed by the gravity of the pandemic. We hope that the businesses that failed to respond will weather the storm and the individuals that did not get the chance to review our resumes are healthy and safe.

As a result of the impact of COVID-19 on our experiment, we made a pivot to attempt a replication study of the an experiment being conducted by another team on the behavior of desert hikers. The result of that effort was hampered by the short time we had to collect data (2 weeks) and the number of responses we were able to elicit to our survey. None of our findings were able to replicate those of the original study, but this is more likely due to our study being severely under-powered than any true difference in experimental outcomes.

References

[Faul et al.2007] Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. 2007. *G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences*. Behavior Research Methods, 39, 175-191.

[Faul et al.2007] Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. 2009. *Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses*. Behavior Research Methods, 41, 1149-1160.

[Smith2013] Jacquelyn Smith 2013. *7 Things You Probably Didn't Know About Your Job Search*. Forbes, `https://www.forbes.com/sites/jacquelynsmith/2013/04/17/7-things-you-probably-didnt-know-about-your-job-search/#257729b38110` (Accessed April 12, 2020)

[Marta Sestelo 2017] Marta Sestelo 2017. *A Short Course on Survival Analysis Applied to the Financial Industry*. `https://bookdown.org/sestelo/sa_financial/` (Accessed April 12, 2020).

[Cox, D.R. 1972] D.R. Cox 1972. *Regression Models and Life-Tables (with Discussion)*. Journal of the Royal Statistical Society, Series B: Methodological 34: 187–220.

[Dempster et al. 1977] Dempster, A. P., N. M. Laird, and D. B. Rubin 1977. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society. Series B (Methodological) 39 (1)

[Grambsch et al. 1994] Grambsch, Patricia M., and Terry M. Therneau 1994. *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals*. Biometrika 81 (3). Oxford University Press: 515–26.

[Schoenfeld, David 1982] David Schoenfeld 1982. *Partial Residuals for the Proportional Hazards Regression Model*. Biometrika 69 (1): 239–41.

[Lauren Weber 2012] Lauren Weber 2012. *Your Resume vs. Oblivion*. `https://www.wsj.com/articles/SB10001424052970204624204577178941034941330` (Accessed April 12, 2020).

[American Psychological Association1983] American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

[Association for Computing Machinery1983] Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

[Chandra et al.1981] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

[Gusfield1997] Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

```
Call:
lm(formula = Hiking_Data$Liters_of_Water_Coalesced ~ Hiking_Data$Treatment_Flag +
    Hiking_Data$Gender + Hiking_Data$BMI + Hiking_Data$Age +
    Hiking_Data$Marital_Status + Hiking_Data$Have_Children +
    Hiking_Data$Education_Level + Hiking_Data$Physical_Activity_Level +
    Hiking_Data$Hiking_Backing_Experience_Level + Hiking_Data$Previous_Desert_Hiking_Experience)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8271 -0.9820  0.0000  0.9619  2.8584 

Coefficients:
                                                                                                             Estimate Std. Error t value Pr(>|t|)  
(Intercept)                                                                                                  10.13962    3.99539   2.538   0.0227 *
Hiking_Data$Treatment_Flag1                                                                                   0.22674    0.97408   0.233   0.8191  
Hiking_Data$GenderWoman                                                                                       0.11127    1.02510   0.109   0.9150  
Hiking_Data$BMI                                                                                              -0.13579    0.14055  -0.966   0.3493  
Hiking_Data$Age                                                                                              -0.04160    0.08823  -0.472   0.6441  
Hiking_Data$Marital_StatusSingle                                                                             -1.51865    1.20758  -1.258   0.2278  
Hiking_Data$Have_ChildrenYes                                                                                 -0.32300    1.38920  -0.233   0.8193  
Hiking_Data$Education_LevelGraduate level (some or graduate)                                                 -0.40146    1.08375  -0.370   0.7162  
Hiking_Data$Education_LevelSome college                                                                      -0.99260    2.44972  -0.405   0.6911  
Hiking_Data$Physical_Activity_LevelMinimal (1-2 Times per week)                                               1.53476    1.42017   1.081   0.2969  
Hiking_Data$Physical_Activity_LevelModerate (3-4 Times per week)                                              0.86359    1.25790   0.687   0.5028  
Hiking_Data$Hiking_Backing_Experience_LevelMinimal (I've gone hiking or backpacking a few times, for at least a mile at a time) -1.71957    2.08007  -0.827   0.4214  
Hiking_Data$Hiking_Backing_Experience_LevelModerate (I hike or backpack a few times a year, for at least a mile at a time)      -1.23215    0.90860  -1.356   0.1951  
Hiking_Data$Previous_Desert_Hiking_ExperienceYes                                                             -0.60297    1.11230  -0.542   0.5957  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.896 on 15 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.307,	Adjusted R-squared:  -0.2936 
F-statistic: 0.5112 on 13 and 15 DF,  p-value: 0.8843
```

Figure 9: Multiple Regression Model