Employment Prestige After Biochemistry PhD: An Observational Study with Causal Inference Analysis

Richard Paul Yim

11 June 2023

Contents

1 Introduction

Academic job markets are known to be notoriously competitive with thousands of PhD students graduating each year in with science and engineering doctorates in the US [1]. In fact, this rate has only been increasing with no sign of decline. Yet, the number of academic positions, in particular postdoctoral positions, are not growing at a proportionate rate with most seeking industry position (i.e., academic job market is incredibly competitive). In this project we explore what activities if any have any causal effect on academic job market outcomes for those seeking academic careers in science and engineering.

1.1 Background

A 1979 paper by Long, Allison and McGinnis studied the academic placement of 239 male PhD biochemists [2]. The original data corresponding to this study is nearly twice as large including females as well. The paper was interested in investigating relationships with preemployment activities of these PhD biochemists. Controlling for covariates such as mentor productivity and PhD prestige the authors found no statistically significant association with the outcome of prestige of first academic position.

The reason prestige is important is because it there is high correlation with funding and prestige (i.e., correlation with more funding and money is more likely to lead to better job satisfaction) [3]. With this endpoint in mind the authors made three noteworthy conclusions:

1. PhD Department is much more important for first teaching position—inbreeding effect.
2. Bigger departments have poorer placement in terms of prestige—regression effect.
3. Preemployment activity is not as important as it would seem—measurement misspecification.

Of particular interest for this project is the third point where there may be a measurement misspecification in the characterization of "preemployment activities" since the authors only studied preemployment activities correponding to articles published, citations received and fellowship funding—all of which are highly correlated, which will be seen. Yet, this despite this potential misspecification we will attempt to reach similar conclusions through the lens of causal inference analysis in particular utilizing matching techniques and signed rank tests.

1.2 Problem Statement

With our background in mind, we aim to answer to following question:

> Is there a causal effect in first academic position prestige outcomes with respect to preemployment activities in fellowship received or number of articles published?

We study two "treatments" as fellowship received and whether the PhD student had published multiple articles or not. In the next section we will explore the relevant dataset along these two treatment specifications.

2 Data

In this section we perform exploratory data analysis on the original dataset, produce a working dataset—specifying the two different treatments specifications—, and explore missinginess in our Data. (Data source: https://socialsciences.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Long-PhDs.txt)

2.1 Original Dataset

In the original dataset there are $n = 408$ observations with six covariates. (Table 1 shows the five number summaries of the features in the dataset.)

Table 1: Five number summaries of all features in original dataset.

job	gender	phd	mentor
Min. :1.100	female:159	Min. :1.000	Min. : 0.00
1st Qu.:2.020	male :249	1st Qu.:2.482	1st Qu.: 4.00
Median :2.560		Median :3.360	Median : 19.00
Mean :2.629		Mean :3.201	Mean : 45.47
3rd Qu.:3.200		3rd Qu.:4.000	3rd Qu.: 56.25
Max. :4.800		Max. :4.800	Max. :532.00
NA's :99			

fellowship	articles	citations
no :156	Min. : 0.000	Min. : 0.00
yes:252	1st Qu.: 1.000	1st Qu.: 0.00
	Median : 2.000	Median : 8.00
	Mean : 2.277	Mean : 21.72
	3rd Qu.: 3.000	3rd Qu.: 29.25
	Max. :18.000	Max. :203.00

In this dataset we report no missingness in any of the covariates, but we do have missingness in the response for some covariates. We address this in the next subsection.

2.2 Working Dataset

We report no missing data at all in the covariates, but there is missing data in the prestige response (missing 99 response endpoints). Since the response is important to our analysis we only work with 309 data points throwing away the observations with missing data points—in theory we can perform multiple imputation, but exclude doing so since it's a bit of a distraction to the main focus of causal effects inference.

Instead, we observe differences in distributions for data points missing the response in Figure 1. Notice that in Figure 1, observations including the response generally have greater PhD prestige, mentor prestige/citations, greater number of published articles, and more citations. An explanation for this missingness is that there is a bias towards those who end up at teaching positions in academia who may be more willing to explain their job outcomes; a more simpler explanation is

the possibility that the dataset includes individuals that do not end up in academia, but may have went to industry or pursued other careers. In any case, we conclude that these data points are not relevant for our study.

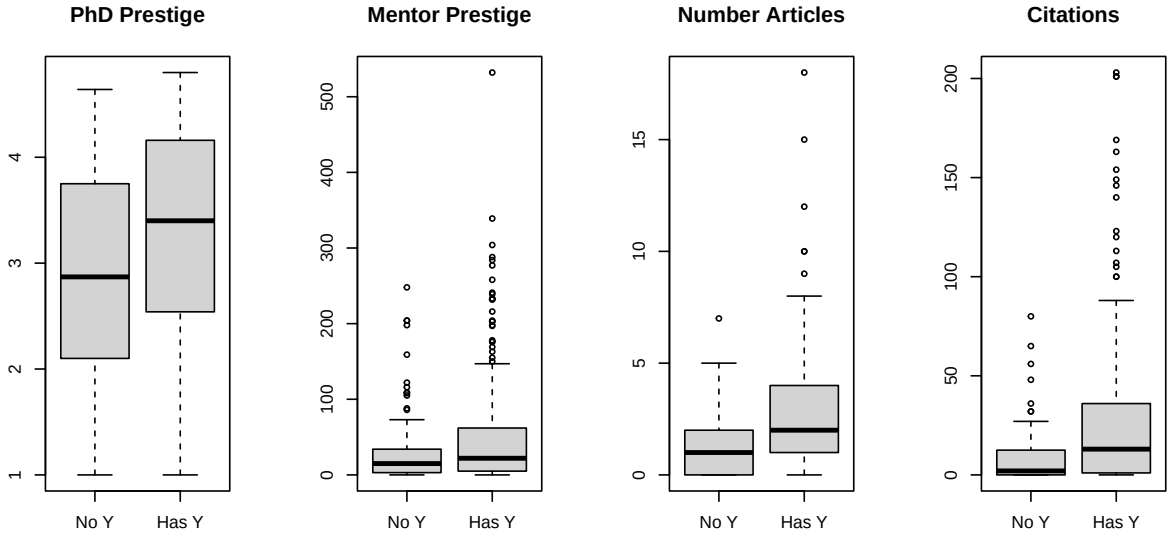

Figure 1: Boxplot differences between observations missing prestige response.

For additional data exploration we study any simple correlation structure that may be present between the covariates in the dataset. Figure 2 shows the corresponding correlation matrix and Pearson correlation values. The covariate that has the highest amount of correlation is between the number of citations and number of articles published. This makes sense since individuals with very no publications will obviously have no citations. Therefore, in all of our analysis we only include gender, PhD prestige and mentor citations in any sort of covariate analysis and propensity score values. (We include Table 2 in the appendix which includes the working data five number summaries.)

2.3 Treatment and Control

Addressing the measurement misspecification from the introduction, we consider a modified characterization of preemployment activity during a PhD and will serve as potential "treatments" in our observational study.

Namely, whether a PhD employee at an academic institution has multiple publications during their PhD program will be designated the "treatment" of a subject, and a student with only one or no publications as a control. Multiple is determined as two or more. For practical reasons we determine this to be a good treatment specification since it is an early indicator for PhD productivity which can cause PhDs to have more opportunities at prestigious institutions.

As an illustration, zero publications shows incompetence as a researcher; one publication shows competence, but otherwise indicates unproductive PhD career; two or more indicates some decent effort with research activities. Research is important as a PhD student since it is the primary reason

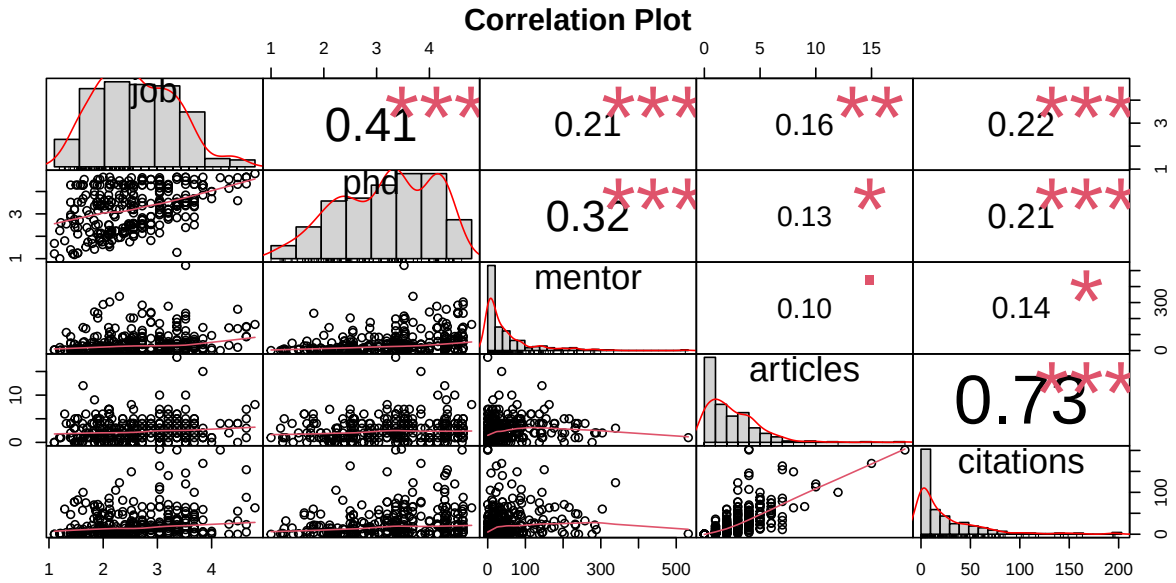

Figure 2: Correlation matrix with main diagonal as histogram; lower triangle corresponding to smoothing spline fit between features; upper triangle corresponding to pearson's correlation coefficient with the asterisks indicating level of significance for a naive simple linear regression fit.

one starts and finishes a PhD program, and with respect to prestige for first employment one would hope that a PhD student has agency in entering prestigious employment through preemployment publication activities.

3 Causal Inference

The key to causal inference is to control for any biases in the data and to isolate a signal of outcomes between a treatment and control group—we aim to put the data in a setting where we can assume no unmeasured confounders similar to a randzomized experiment. In particular, matching techniques are a primary tool for causal inference in observational studies. We apply both paired and full matching and follow with relevant sensitivity analyses.

3.1 Paired Matching

Paired optimal matching aims to generate comparable paired subjects of treatment and control, minimizing within pair differences to isolate any potential variation in treatment/control assignments—this is the basis for casual inference for observational studies data.

Optimal matching generates a distance matrix between subjects essentially mapping their similarities to a single scalar value where greater numerical score indicates less similarity, and small score indicates closer similarities. The resulting data structure is a bipartite graph of treatment and control subjects. Then a network flow optimization or auction algorithm is utilized to generate pairs traversing what is essentially a bipartite graph. We use `DiPs` package to create optimal matchings.

In particular, we use the robust Mahalanobis distance with a propensity score caliper, where the distance between a treatment and subject pair is

$$(\mathbf{x}_t - \mathbf{x}_s)^t \hat{\Sigma}^{-1} (\mathbf{x}_t - \mathbf{x}_s),$$

where $\hat{\Sigma}^{-1}$ is the inverse of the sample covariance/variance matrix.

In order to generate an optimal match we use a logistic propensity score caliper of the following form:

$$\text{logit}(\mathbb{E}(Y|X)) = X\beta = \beta_0 + \beta_1 \texttt{male}_i + \beta_2 \texttt{mentor}_i + \beta_3 \texttt{Phd}_i,$$

which is a binomial regression model with logit link (i.e., logistic regression). The idea behind adding this penalty is to have a measurement (caliper) for isolating any source of hidden biases. Since this is *not* a randomization study there is likely a lot of hidden bias. Creating a penalty via a propensity score caliper which should model likelihoods of being in a treatment/control group and hope to balance and minimize differences between paired treatment/control subjects.

After creating these matched pairs with calipers, the only covariate that seems to be relatively comparable between treatment and control is sex—PhD prestige and mentor are statistically significant and different between the treatment and control even after pairing. Following up, we find that a signed rank test produces a statistically significant difference in paired medians with a p-value of 0.0008673. However, since we have bad balancing from paired matching our result has low integrity.

As a follow-up on lack-of-integrity for this paired matching, if we perform sensitivity analysis, we find that the required factor needed for hidden biases to increase is very low where we see that the maximum probability in Table 2 is very high at just $\Gamma = 1.8$. This is consistent to our weak balancing. So it's difficult to draw any conclusions from optimal paired matching. In the next subsection, we attempt to remedy this with full matching.

Table 2: The gamma corresponds to the factor needed to increase the odds of exposure, and the minimum and maximum ranges correspond to the probabilities for discordance.

	gamma	minimum	maximum
1	1.00	0.00	0.00
2	1.20	0.00	0.01
3	1.40	0.00	0.04
4	1.80	0.00	0.26

3.2 Full Matching

There is a problem in our dataset where we have less controls than treatment units. We can characterize this notion with the `entire" number, which`represents the average number of controls available for matching to a treated subject with covariate `x`" [4]. Computing the five number summary of the entire number for our dataset we have that a super majority of the treated units have entire number less than 1, indicating that full matching may be better (see Table 3).

The key idea for full matching is that we divide the data into a collection of matched sets with variability in the number of treatments to control. However, even with full matching we retain the same sample size of 123 as seen in pair matching.

Furthermore, we will find that we have statistical significance for a signed rank test using a subsets over full match produced by `optmatch`, but again from our previous sensitivity analysis, there is a

lot variability that is attributed to hidden biases. There is not causal association to be made—our full match and paired match results are consistent with inconclusive results from the data.

Table 3: Entire number five number summary for full matching. Full matching is needed since at least 3/4 of observations have entire number less than 1.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	0.15	0.48	0.65	0.65	0.82	1.43

Table 4: Matching structure. We have an effective sample size of 122, which is essentially the same as the sample size for the paired matching

	7:1	6:1	5:1	4:1	3:1	2:1	1:1	1:2	1:3	1:4	1:5	0:1
1	1	3	1	8	12	16	44	6	1	3	2	1

4 Conclusions

The original authors performed an association study using very elegant and simple regression analysis. No associate was to be made of any significance for treatments related to preemployment PhD activity, in particular number of articles published.

We've reached the same conclusions here that there is no causal inference that can be made rigorously due to strong sensitivities in hidden biases that are not accounted for in this very limited dataset. An important point to be learned is that a study finding no association will prove to be very difficult to find any causal relationship (i.e., if there no relationship, there would be no directional relationship by definition). Yet, it is of sufficient interest to show, as we've done, that our results of inconclusive effects of preemployment activity are consistent with that of Long et al.'s results.

(Of course this isn't a message PhD students should quietly quit through all 4-6 years of their PhD because it has no effect on the prestige for their first academic job—it's just that there is no strong relationship between prestige and academic productivity, and that depending on location (e.g., midwest, east/west, south) prestige may not be at all important at a practical level for economic reasons.)

4.1 Other Considerations

The original study is fairly old, published in 1979 and restricted to only biochemist PhD graduates. Of further interest would be to extend this study to science and engineering PhD graduates working in industry. Information such as whether they had internships during their PhD program and other activities can be easily designed and manufactured from web crawlers/scrapers. This same analysis can readily be extended and we can attempt to intentionally sample data from sources such as LinkedIn to acquire information of first job out of PhD and say the revenue/market-rank of the firm that the person is working at. We can pull more features such as undergraduate institution and other degrees and more— and apply matching techniques at scale for hundreds of data points along many different treatment/control definitions.

5 Appendix

5.1 References

1. https://www.nsf.gov/nsb/sei/edTool/data/college-17.html
2. *Entrance Into the Academic Career*
3. *The Role of Early-Career University Prestige Stratification on the Future Academic Performance of Scholars*
4. Notes for Causal Inference in Observational Studies by Dylan Small and Ruoqi Yu—based off of Rosenbaum's *Observational Studies*

5.2 Additional Tables

Table 5: Working data five number summaries

job	gender	phd	mentor
Min. :1.100	female:105	Min. :1.000	Min. : 0.00
1st Qu.:2.020	male :204	1st Qu.:2.540	1st Qu.: 5.00
Median :2.560		Median :3.400	Median : 22.00
Mean :2.629		Mean :3.292	Mean : 49.59
3rd Qu.:3.200		3rd Qu.:4.160	3rd Qu.: 62.00
Max. :4.800		Max. :4.800	Max. :532.00

fellowship	articles	citations
no :102	Min. : 0.000	Min. : 0.00
yes:207	1st Qu.: 1.000	1st Qu.: 1.00
	Median : 2.000	Median : 13.00
	Mean : 2.563	Mean : 25.89
	3rd Qu.: 4.000	3rd Qu.: 36.00
	Max. :18.000	Max. :203.00

5.3 Miscellaneous (Randomization)

Since this project consists of observational data, we exclude randomization analysis, but include some code for reproducibility *if* this was a randomization experiment. (Again, we are just including code for reference not relevant/correct w.r.t to this data.)

5.4 Code Chunks

```
knitr::opts_chunk$set(echo = TRUE)
# ================================================================================
# read in data and quick data checks
```

```r
# ================================================================================
data <- read.table("Long-PhDs.txt")
data$gender <- factor(data$gender)
data$fellowship <- factor(data$fellowship)
head(data)
colnames(data)
summary(data)

# ================================================================================
# missingness
# ================================================================================
apply(data, 2, function(x) {
  mean(is.na(x)) # average missingness
} )
apply(data, 2, function(x) {
  sum(is.na(x)) # count missingness
} )
# response missingness structure
noResponseMult <- data[is.na(data$job),]
hasResponseMult <- data[!is.na(data$job),]
summary(hasResponseMult)
summary(noResponseMult)
par(mfrow=c(1,2))

# box plots on no response v response
par(mfrow=c(1,4))
boxplot(noResponseMult$phd, hasResponseMult$phd, names = c("No Y", "Has Y"))
title("PhD Prestige")
boxplot(noResponseMult$mentor, hasResponseMult$mentor, names = c("No Y", "Has Y"))
title("Mentor Prestige")
boxplot(noResponseMult$articles, hasResponseMult$articles, names = c("No Y", "Has Y"))
title("Number Articles")
boxplot(noResponseMult$citations, hasResponseMult$citations, names = c("No Y", "Has Y"))
title("Citations")

# ================================================================================
# working data
# ================================================================================
workingData <- hasResponseMult
dim(workingData)
summary(workingData)
colnames(workingData)

# ================================================================================
# correlation matrix
# ================================================================================
# correlations
```

```
library(xtable)
library("PerformanceAnalytics")
chart.Correlation(workingData[,-c(2,5)], histogram=TRUE, pch=19)
title("Correlation Plot",line=3, adj=0.45)
# create new response
workingData$multiple <- workingData$articles>=2
table(workingData$multiple)
library(xtable)
xtable(summary(workingData[,1:4]))
xtable(summary(workingData[,5:7]))

# ============================================================================
# Hidden bias and Overt bias
# ============================================================================
# t test on continuous covariates
t.test(workingData[!workingData$multiple,]$phd,
       workingData[workingData$multiple,]$phd)
t.test(workingData[!workingData$multiple,]$mentor,
       workingData[workingData$multiple,]$mentor)

# t test on discrete covariates
chisq.test(workingData$gender, workingData$multiple, correct=FALSE)
chisq.test(workingData$fellowship, workingData$multiple, correct=FALSE)
table(workingData$fellowship, workingData$multiple)

# ============================================================================
# Construct rank based Mahalanobis distance with propensity score caliper
# ============================================================================
# generate Mahalanobis distance matrix
X <- cbind(as.numeric(workingData$gender)-1,
         workingData$phd,
         workingData$mentor
         )
z <- workingData$multiple
p <- glm(z ~ X, family=binomial)$fitted.values
d <- cbind(workingData,p)
d$z <- z
library(DiPs)
library(exactRankTests)
dist <- maha_dense(z,X)
# add caliper
dist <- addcaliper(dist, d$z, d$p, c(-0.001,0.001), stdev=TRUE, penalty=1000)
# create matching
o <- match(!d$z, dist, workingData)
matcheddata <- o$data

# propensity score box plot
```

```
#boxplot(p[!d$z], p[d$z], names = c("Less than multiple", "Multiple"),
#        ylab="Propensity Score")
#title("Boxplots of Propensity Scores")
#wilcox.exact(p[!d$z], p[d$z], conf.int = TRUE)

# ============================================================================
# matching integrity
# ============================================================================
Xdata <- X
Xmatch <- cbind(as.numeric(matcheddata$gender)-1,
                matcheddata$phd,
                matcheddata$mentor
)
balance_tb=check(Xdata, Xmatch, workingData$multiple, matcheddata$multiple)

# t tests
covnames=rownames(balance_tb)
treatmat.after=Xmatch[matcheddata$multiple,]
controlmat.after=Xmatch[!matcheddata$multiple,]
t.test.pval.vec=rep(0,length(covnames))
for (i in 1:ncol(treatmat.after)){
  t.test.pval.vec[i]=
    t.test(treatmat.after[,i],controlmat.after[,i])$p.value
}
cbind(covnames,t.test.pval.vec)
#hist(matcheddata$phd)

# ============================================================================
# testing signed rank testing
# ============================================================================
md.control <- matcheddata[matcheddata$multiple,]$job
md.treated <- matcheddata[!matcheddata$multiple,]$job
wilcox.exact(md.control,
             md.treated,
             paired=TRUE,conf.int=TRUE)

library(exactRankTests)
md.control <- matcheddata[!matcheddata$multiple,]$job
md.treated <- matcheddata[matcheddata$multiple,]$job
wilcox.exact(md.treated,
             md.control,
             paired=TRUE,conf.int=TRUE)

# ============================================================================
# Sensitivity analysis for matching
# ============================================================================
sens.analysis.signedrank=function(diff,Gamma){
```

```
  rk=rank(abs(diff));
  s1=1*(diff>0);
  s2=1*(diff<0);
  W=sum(s1*rk);
  Eplus=sum((s1+s2)*rk*Gamma)/(1+Gamma);
  Eminus=sum((s1+s2)*rk)/(1+Gamma);
  V=sum((s1+s2)*rk*rk*Gamma)/((1+Gamma)^2);
  Dplus=(W-Eplus)/sqrt(V);
  Dminus=(W-Eminus)/sqrt(V);
  list(lowerbound=1-pnorm(Dminus),upperbound=1-pnorm(Dplus))
}
diff=md.treated-md.control
glist=c(1,1.2,1.4,1.8)
lead.sen.tb=matrix(nrow=length(glist),ncol=2)
for (g in 1:length(glist)){
  lead.sen.tb[g,]=unlist(sens.analysis.signedrank(diff,glist[g]))
}
gamma <- data.frame(gamma=glist,lead.sen.tb)
colnames(gamma) <- c("gamma","minimum","maximum")
xtable(print(gamma))

# ============================================================================
# Full matching
# ============================================================================
propscore.model <- glm(multiple~gender+phd+mentor, family=binomial(), data=workingData)
propscore.treated=predict(propscore.model,type="response")[propscore.model$y==1]
entire.number.treated=(1-propscore.treated)/propscore.treated
sumEntire <- matrix(summary(entire.number.treated),nrow=1)
colnames(sumEntire) <- c("Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", "Max.")
xtable(sumEntire)

library(optmatch)
# Full matching within a propensity score caliper.
mhd <- match_on(multiple ~ gender + phd + mentor, data = workingData) +
  caliper(match_on(propscore.model), width = 1)
matchvec <- fullmatch(mhd, data = workingData)
summary(matchvec)
workingData$matchvec=matchvec
xtable(t(stratumStructure(matchvec)))
effectiveSampleSize(matchvec)

# wilcoxon signed rank test
wilcox.exact(md.treated,
             md.control,
             conf.int=TRUE, subset=matchvec)

# ============================================================================
```

```
# Randomization
# ============================================================================
# Function for testing no treatment effect using the difference in
# sample means as the test statistic, rejecting for large values of
# the test statistic and using the Monte Carlo method with K draws
treat.effect.samplemean.montecarlo.test.func=function(treated.r,control.r,K){
  # Create vectors for r and Z, and find total number
  # in experiment and number of treated subjects
  r=c(treated.r,control.r);
  Z=c(rep(1,length(treated.r)),rep(0,length(control.r)));
  N=length(r);
  m=length(treated.r);
  # Observed test statistic
  obs.test.stat=mean(r[Z==1])-mean(r[Z==0]);
  # Monte Carlo simulation
  montecarlo.test.stat=rep(0,K);
  for(i in 1:K){
    treatedgroup=sample(1:N,m); # Draw random assignment
    controlgroup=(1:N)[-treatedgroup];
    # Compute test statistic for random assignment
    montecarlo.test.stat[i]=mean(r[treatedgroup])-mean(r[controlgroup]);
  }
  # Monte Carlo p-value is proportion of randomly drawn
  # test statistics that are >= observed test statistic
  pval=sum(montecarlo.test.stat>=obs.test.stat)/K;
  # 95% CI for true p-value based on Monte Carlo p-value
  lowerci=pval-1.96*sqrt(pval*(1-pval)/K);
  upperci=pval+1.96*sqrt(pval*(1-pval)/K);
  list(pval=pval,lowerci=lowerci,upperci=upperci,
       mc.test.stat=montecarlo.test.stat,
       obs.test.stat=obs.test.stat)
}

# For the job training experiment
treated.r.train <- workingData[workingData$multiple,]$job
control.r.train <- workingData[!workingData$multiple,]$job
res=treat.effect.samplemean.montecarlo.test.func(treated.r.train,
                                                 control.r.train,100000)
res$pval
res$lowerci
res$upperci
# Generate histogram of the test statistics
if(0){
  hist(res$mc.test.stat,xlab = "mc.test.stat",main="Histogram of mc.test.stat")
  arrows(res$obs.test.stat,10000,res$obs.test.stat,0,length=.1)
  text(res$obs.test.stat,12000,"Observed")
}
```