Homework 3

Advanced Methods for Data Analysis (36-402)

Due Friday February 8, 2019, at 6:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as "*code*", and knitted or scanned/merged pdf or html as "*writeup*"). Make sure that everything you submit is readable.

1 From Homework 2

1.1 1f from "A Refresher in Linear Regression"

The data for this problem are in two files named `housetrain.csv` (containing training data) and `housetest.csv` (containing test data.) They are both comma-separated files with headers. You will need to read each of them into an *R* `data.frame`. The `read.csv` command will do this, if you get the syntax correct.

The data are from a census survey from several years ago. Each record (line in the file) corresponds to a small area called a *census tract*. The variables that appear in the data file have the following names:

- `Population`: The population of the census tract.
- `Latitude`: The number of degrees north of the equator where the census tract is located. South is negative, so latitude is between -90 and 90.
- `Longitude`: The number of degrees east of Greenwich where the census tract is located. West is negative, so longitude is between -180 and 180.
- `Median_house_value`: The median assessed value of houses (in thousands of dollars) in the census tract.
- `Median_household_income`: The median household income in the census tract.
- `Mean_household_income`: The average household income in the census tract.

The data are from census tracts in California and Pennsylvania.

The main goal of this problem is to model the relationship, if any, between `Median_house_value` (the response) and the other variables (potential predictors.) For all plots, use the option `,pch="."` because the default circles will overlap too much with such large data sets.

(f) (*Extra Credit Question*) Linear regression is about projecting the response vector $\vec{Y}$ onto the space spanned by linear combinations of the $\mathbf{X}$ variables. So the residual vector $\widehat{\vec{\epsilon}}$ is orthogonal to the fitted values $\widehat{\vec{Y}} = H\vec{Y}$, where $H = \mathbb{X}\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T$ (it is called "hat matrix" or "projection matrix"). Try the following:

1. Fit a regression of `Median_house_value` (Y) on `Population`($X1$) and `Median_household_income` ($X2$). Call the residuals $\widehat{\epsilon}_1$.
2. Fit a regression of `Mean_household_income`($X3$) on `Population`($X1$) and `Median_household_income`($X2$). Call the residuals $\widehat{\epsilon}_2$.
3. Fit a regression of $\widehat{\epsilon}_1$ on $\widehat{\epsilon}_2$.
4. Fit a regression of `Median_house_value`(Y) on `Population`($X1$), `Median_household_income`($X2$) and `Mean_household_income`($X3$).

What do you notice? Can you intuitively explain why this is the case?

Solution

```
set.seed(1000)
dat <- read.csv("housetrain.csv", header = TRUE)
fit5 <- lm(Median_house_value ~ Population + Median_household_income, data = dat)
eps1 <- resid(fit5)
fit6 <- lm(Mean_household_income ~ Population + Median_household_income, data = dat)
eps2 <- resid(fit6)
fit7 <- lm(eps1 ~ eps2)
fit8 <- lm(Median_house_value ~ Population + Median_household_income +
Mean_household_income, data = dat)
abs(coef(fit7)[["eps2"]] - coef(fit8)[["Mean_household_income"]])
1.5e-17
```

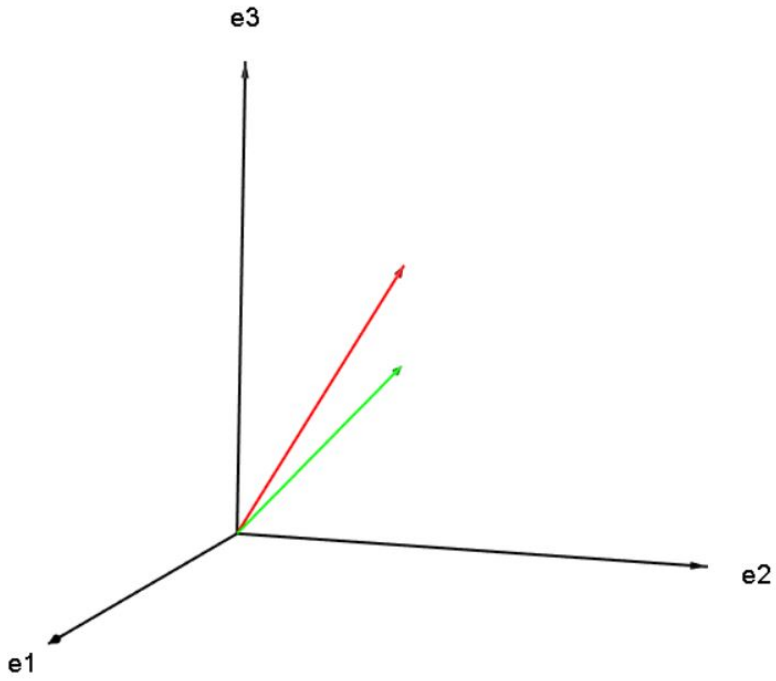
Notice that the slope coefficient for $\hat{\epsilon}_2$ is the same as that for Mean_household_income in the regression that includes all the variables. This is no coincidence, but let see why this is the case.

First, let us argue intuitively. The residual ϵ_1 of the regression Y on X_1 is what "is in" Y which cannot be explained by X_1. The residual ϵ_2 of X_1 on X_2 is what is in X_1 that cannot be explained by X_2. So regressing ϵ_1 on ϵ_2 yields a slope coefficient that indicates what is in Y that cannot be explained by X_1 but that can be explained by that part of X_2 that cannot be explained by X_1. Very convoluted! A better way to express this is to say: the slope of the regression of ϵ_1 on ϵ_2 is the additional contribution of X_2 on Y, after we have taken into account the contribution of X_1. But this is also exactly what the multiple regression of Y on X_1 and X_2 is doing.

Let's have simple example in mind. Suppose we observe 3 observations $(\vec{Y}, \mathbb{X})$: $(1, 1)$, $(2, 2)$ and $(3, 2)$. So, $\vec{Y} = (1, 2, 3)$ and $\mathbb{X} = (1, 2, 2)$. In general, we are used to see $\vec{Y}$ with n components and a matrix $\mathbb{X}$, which is $n \times p$.

The first two observations fit the model $Y = \beta X$ with $\beta = 1$ perfectly, but the third does

not. Check the picture below, the red vector is the $\vec{Y}$, which is not aligned with the green $\mathbb{X}$ vector $(1, 2, 2)$. So, we know the model will not be perfect.

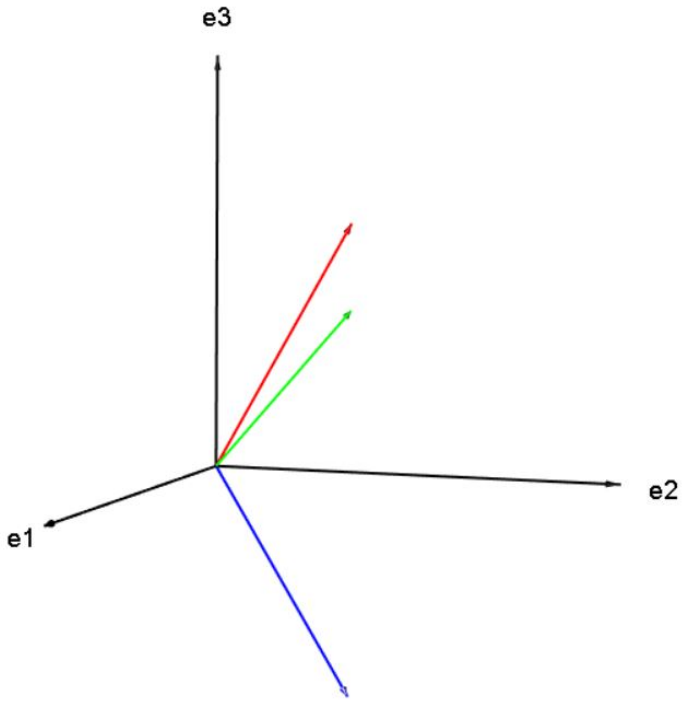


Intuitively, the closest vector to $\vec{Y}$ that is "aligned" with $\mathbb{X}$ can be found by projecting $\vec{Y}$ onto the vector $\mathbb{X}$. This is crucial, make sure you are convinced of this. To better visualize this projection, we can find a vector $\vec{\tau} = (\tau_1, \tau_2, \tau_3)$ orthogonal to $\mathbb{X}$ so that the space spanned by $(\vec{\tau}, \mathbb{X})$ (that is, the usual space with $\vec{\tau}$ as the ordinate and $\mathbb{X}$ as abscissa) contains $\vec{Y}$. This means we need to be able to write $\vec{Y} = \alpha_1\mathbb{X} + \alpha_2\vec{\tau}$. Because $\vec{\tau}$ must be orthogonal to $\mathbb{X}$, we require the dot product to be zero: $\tau_1 + 2\tau_2 + 2\tau_3 = 0$ [1] Then, to make sure that $\vec{Y}$ is in the space spanned by $(\mathbb{X}, \vec{\tau})$, we additionally require that

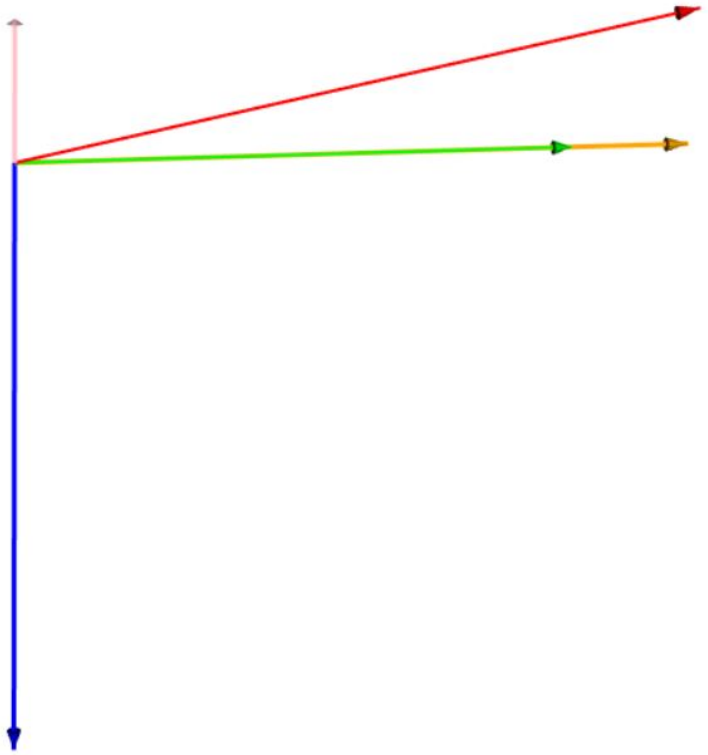
$$\begin{cases} \tau_1 + 2\tau_2 + 2\tau_3 = 0 \\ \alpha_1 + \alpha_2\tau_1 = 1 \\ 2\alpha_1 + \alpha_2\tau_2 = 2 \\ 2\alpha_1 + \alpha_2\tau_3 = 3 \end{cases}$$

Check that $\vec{\tau} = (1, 2, -5/2)$, $\alpha_1 = 11/9$ and $\alpha_2 = -2/9$ satisfies the system of equations above, as the picture below shows ($\vec{\tau}$ is in blue):

[1] Notice that covariance between random variables and dot product between vectors are kind of twins, which is why people often think of two vectors having zero covariance as being orthogonal. You can read more here.



If we reproduce the picture above just using the space $(\vec{\tau}, \mathbb{X})$, we get



And now? Well, we have our slope estimate from least squares, namely $\alpha_1 = 11/9$! The orange vector is $11/9X$ and the pink vector is $-2/9\tau$. What did we do? We just "split" $\vec{Y}$ into two vectors orthogonal to each other, one of which, for convenience, was just a scalar multiple of $\mathbb{X}$. This was convenient because it meant we could just read off the value of the slope coefficient.

We excluded the intercept from our discussion, but this is no problem: it just amounts to adding a vector $\mathbb{X}_2 = (1, 1, 1)$ into the picture, but this would get a little more messy. Excluding the intercept, we can check that R agrees with our calculation ($1.222 = 11/9$):

```
dat <- data.frame(y = c(1, 2, 3), x = c(1,2,2))
summary(lm(y ~ x - 1, data = dat))
Call:
lm(formula = y ~ x - 1, data = dat)

Residuals:
1      2      3
-0.222 -0.444  0.556

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
x    1.222      0.176    6.96     0.02 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.53 on 2 degrees of freedom
Multiple R-squared:  0.96,        Adjusted R-squared:  0.94
F-statistic: 48.4 on 1 and 2 DF,  p-value: 0.02
```

Equipped with this projecting business, we can then explain what we see in the example mentioned in the question. Consider the space spanned by orthogonal bases $(\vec{e}_1, \vec{e}_2, \vec{e}_3, \vec{e}_4)$ (great review of these concepts is Appendix B of Shalizi's book). We choose $\vec{e}_1$ to be aligned with $\mathbb{X}_1$, just as before: $\mathbb{X}_1 = \alpha\vec{e}_1$ (in the earlier picture, we took $\alpha = 1$ and $\vec{e}_1 = \mathbb{X}$). Then, we can express $\mathbb{X}_2$ as $\mathbb{X}_2 = \beta_1\vec{e}_1 + \beta_2\vec{e}_2$, $\mathbb{X}_3 = \tau_1\vec{e}_1 + \tau_2\vec{e}_2 + \tau_3\vec{e}_3$, and $\vec{Y} = \gamma_1\vec{e}_1 + \gamma_2\vec{e}_2 + \gamma_3\vec{e}_3 + \gamma_4\vec{e}_4$. You should be convinced that this is just a generalization of what we did earlier when we just had $(\vec{Y}, \mathbb{X})$. Regressing, Y on X_1 and X_2 means that the residual vector $\widehat{\epsilon}_1$ is just $\gamma_3\vec{e}_3 + \gamma_4\vec{e}_4$, because of orthogonality. Regressing X_3 on X_1 and X_2 means that the residual vector $\widehat{\epsilon}_2$ is just $\tau_3\vec{e}_3$. Regressing $\widehat{\epsilon}_1$ on $\widehat{\epsilon}_2$ means that the slope coefficient is $\frac{\gamma_3}{\tau_3}$ (again by orthogonality!). But this is exactly the slope coefficient for X_3 that we would get if we regress Y on (X_1, X_2, X_3).

Notice that none of our discussion involved randomness, we just found the best approximation of Y in terms of linear combinations of the Xs. Randomness is conceptually introduced, for example, to quantify how well we would do if we observe another X.

Why is minimizing the squared loss equivalent to projecting $\vec{Y} \in \mathbb{R}^n$ on the space spanned by the $\mathbb{X} \in \mathbb{R}^p$, namely vectors of the form $\beta\mathbb{X}$ for $\beta \in \mathbb{R}^p$? A famous theorem in geometry states that the orthogonal projection of $\vec{Y}$ on the space spanned by the $\mathbb{X}$s falls onto a point in the space of the $\mathbb{X}$s so that the distance between $\vec{Y}$ and that point is minimal. But what is the Euclidean distance between $\vec{Y}$ and a vector of the form $\beta\mathbb{X}$? Well, that is just $\sqrt{\sum_{i=1}^n (Y_i - \beta X_i)^2}$, hence the connection between "least squares" and projections.

Finally, you can read more about orthogonalizing each regressor and add them one at a time as a way to estimate coefficients at https://web.stanford.edu/~hastie/Papers/ESLII.pdf (pdf pages 73-74). The arguments justifying the procedure are the same as above. Indeed, when we compute the residual ϵ_2 of the regression X_2 and X_1, we effectively orthogonalize X_2 with respect to X_1.

1.2 2f from "Relaxing the Regression Assumptions"

Consider arbitrary random variables $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$ with *absolutely no assumptions relating the two*, and consider linearly regressing Y on X (in the population), with regression coefficients defined by

$$\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y), \quad \beta_0 = \mathbb{E}(Y) - \beta^T\mathbb{E}(X).$$

Conditional on X, our prediction for Y is hence $\beta_0 + \beta^T X$.

(f) (*Omitted Variables Problem*) In this part, we investigate what happens when we do not include a variable in the model, and this variable is related to both Y and other covariates. In particular, suppose that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, but we only observe (Y, X_1); so we run a linear regression of Y on X_1, and compute the ordinary least squares estimator $\widehat{\beta}_1$. Prove that $\widehat{\beta}_1$ is generally not a consistent estimator of β_1. Under what conditions would $\widehat{\beta}_1$ be consistent?

Recall that $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ *means that* $\forall \epsilon > 0$, $\mathbb{P}\left(|\widehat{\beta}_1 - \beta| > \epsilon\right) \to 0$ *as* $n \to \infty$.
Hint: You can proceed as follows:

1. *Identify what the estimator* $\widehat{\beta}_1$ *is in terms of* X_1, X_2 *and* ϵ.
2. *Can you think of a* ***very famous*** *law that would be useful to show convergence in probability? You may assume that if* $X_n \xrightarrow{p} X$, *then* $g(X_n) \to g(X)$ *for any continuous function (i.e. ratio). This is the so called continuous mapping theorem:* https://en.wikipedia.org/wiki/Continuous_mapping_theorem.

Solution

Notice that, by defintion of the OLS estimator, we have

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(X_1, Y)}{\widehat{\text{Var}}(X_1)}$$

where $\widehat{\text{Cov}}$ and $\widehat{\text{Var}}$ means that we are taking the sample analogues of covariance and variance.

Using the fact that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, we get

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\widehat{\text{Cov}}(X_1, X_2)}{\widehat{\text{Var}}(X_1)} + \frac{\widehat{\text{Cov}}(X_1, \epsilon)}{\widehat{\text{Var}}(X_1)}$$

Note that, by construction, we have $\widehat{\text{Cov}}(X_1, \epsilon) = 0$, so by the weak law of large numbers and an application of the continuous mapping theorem, we get:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

which shows how $\hat{\beta}_1$ is consistent only if $\text{Cov}(X_1, X_2) = 0$, which is rarely the case!

2 The Fisher's null hypothesis

In this exercise, you will learn about a famous idea by Sir RA Fisher: Fisher's null hypothesis.

Consider the table below; you can also find the observed X and Y values in the file `table2c.RData` on Canvas. Imagine that these values are the results of an experiment where 12 subjects are randomly assigned to treatment or control with 6 subjects in each group;

$C(X = 0)$ and $C(X = 1)$ denote the *potential outcomes* if a subject is not treated versus treated. Note that the subjects do not necessarily need to be equally split between treatment and control.

X	$C(X = 0)$	$C(X = 1)$	Y
0	0.2	*	0.2
0	0.7	*	0.7
0	0.3	*	0.3
0	0.9	*	0.9
0	0.1	*	0.1
0	0.1	*	0.1
1	*	1.1	1.1
1	*	0.4	0.4
1	*	1.2	1.2
1	*	0.5	0.5
1	*	2.6	2.6
1	*	1.7	1.7

Fisher's null hypothesis is that **the treatment effect is zero for all the subjects**:

$$H_0 : C_i(X_i = 1) = C_i(X_i = 0) \ \forall i \in \{1, \ldots, 12\}$$

Note that this is a strong null hypothesis: the treatment effect is exactly zero for all subjects, not just on average!

Under this null, we will compute a p-value through a permutation test. (For permutation tests, see e.g. All of Statistics, Sec 10.5, or the statistics textbook you used in 36-226.)

(2a) Choose a sensible test statistic, e.g. the absolute difference in average values of C between the treatment and control groups:

$$\begin{aligned} \widehat{\tau}_{\text{obs}} &= \left| \frac{1}{6} \sum_{i=1}^{12} X_i \ C_i(X_i = 1) - \frac{1}{6} \sum_{i=1}^{12} (1 - X_i) \ C_i(X_i = 0) \right| \\ &= \left| \frac{1}{6} \sum_{i=1}^{12} X_i Y_i - \frac{1}{6} \sum_{i=1}^{12} (1 - X_i) Y_i \right| . \end{aligned}$$

Compute the *observed* test statistic using the data in the table.

Solution

```
load("table2c.RData")
n <- nrow(dat); nt <- sum(dat$x); nc <- sum(1-dat$x)
get_tau <- function(treat_vec_indices, y0, y1, nt, nc){
        term1 <- mean(y1[treat_vec_indices])
        term2 <- mean(y0[-treat_vec_indices])
```

```
        tau <- abs(term1-term2)
        return(tau)
}
tauobs <- get_tau(which(dat$x==1), dat$y, dat$y, nt, nc)
print(tauobs)
[1] 0.87
## $
```

(2b) Fill out the table above under the Fisher's null hypothesis; that is, when H_0 is true.

Solution

X	$C(X=0)$	$C(X=1)$	Y
0	0.2	0.2	0.2
0	0.7	0.7	0.7
0	0.3	0.3	0.3
0	0.9	0.9	0.9
0	0.1	0.1	0.1
0	0.1	0.1	0.1
1	1.1	1.1	1.1
1	0.4	0.4	0.4
1	1.2	1.2	1.2
1	0.5	0.5	0.5
1	2.6	2.6	2.6
1	1.7	1.7	1.7

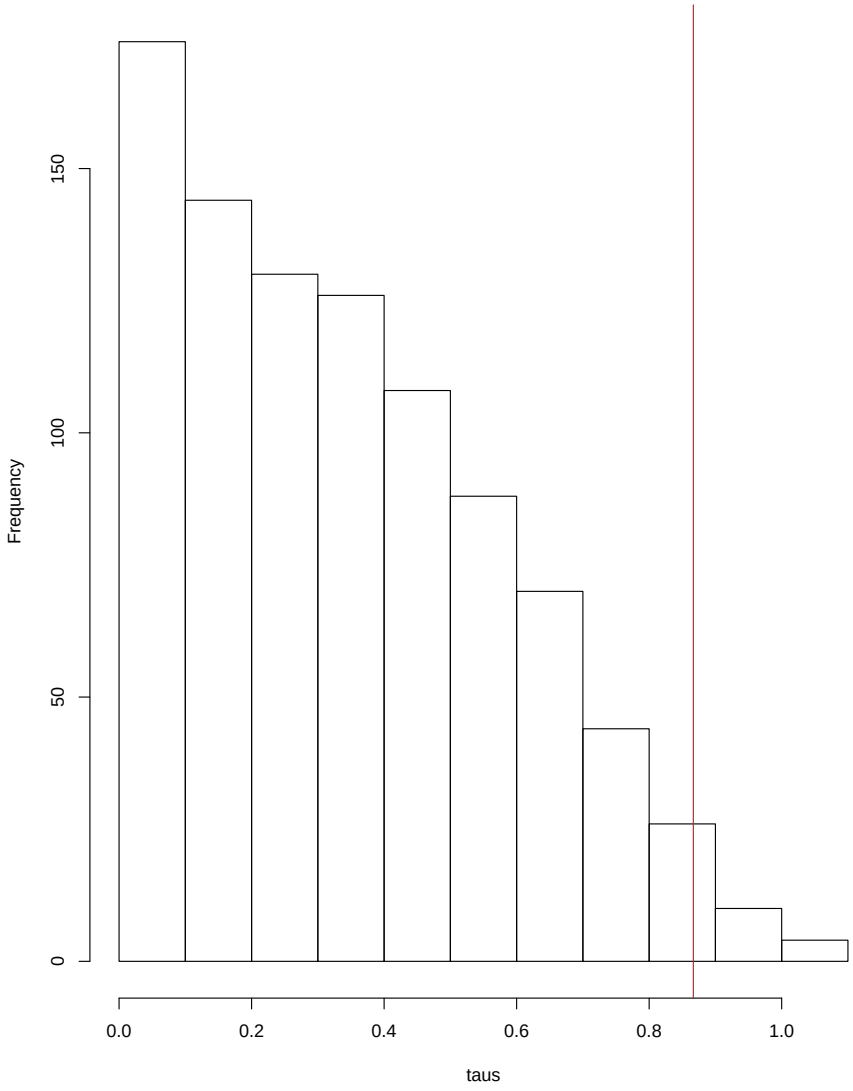
(2c) Load the table in `R` (file is `table2c.RData`) and generate all the possible permutations of the vector with X values. For each permutation, re-compute the test statistic of your choice and plot the resulting distribution with a histogram.

Solution

```
## Compute all permutations, n choose nt
all_perms <- combn(1:n, nt)

taus <- apply(all_perms, 2, get_tau, y0=dat$y, y1=dat$y, nt=nt, nc=nc)

hist(taus, main="")
abline(v=tauobs, col = "red")
```



(2d) Do you think there is a non-zero "treatment effect"? What is the observed p-value associated with the null hypothesis for this experiment?

Solution

```
pvalue <- mean(taus>=tauobs)
print(pvalue)
0.03246753
```

The pvalue is smaller than $\alpha = 0.05$, so we reject the Fisher's null hypothesis of zero treatment effect for all the units.

Alternatively, we could approximate the pvalue by sampling without replace from the treatment vector:

```
B <- 1e5
taus2 <- replicate(B, get_tau(sample(x=1:n, size=nt, replace=FALSE), dat$y,
                              dat$y, nt, nc))
pvalue <- mean(taus2>=tauobs)
print(pvalue)
0.03182
```

What we have essentially done is to randomly generate permutations of treatment vector, without actually listing all permutations. This is useful when the number of permutations is very large.

How much different will the approximated p-value from the true p-value? Here is a crude analysis (more refined analyses are possible). Let π_k denote a permutation. Notice that in AOS they think of $n!$ permutations, but here we could fix the fact that 6 units are in control group and 6 units are in treatment group. So the number of possible permutations is $\binom{12}{6} = 924$. Let p the true exact p-value, so $p = \mathbb{P}(t(\pi_k) \geq t(\pi_{obs})) = r/924$, for some integer r, where $t(\cdot)$ is a map from the space of permutations to $\mathbb{R}$ defining your test statistic. Notice how the smallest pvalue we could possibly observe is $1/924 \approx 0.001$. Let X denote the number of times the event $t(\pi_k) \geq t(\pi_{obs})$ occurs out of our B random samples. Let $\tilde{p}$ denote the estimated p-value, that is $\tilde{p} = X/B$. Then, $X \sim \text{Bin}(B, p)$. In particular, $\mathbb{E}(\tilde{p}) = p$ and $\text{Var}(\tilde{p}) = p(1-p)/B$. By an application of Chebyshev's inequality, we get $\mathbb{P}(|\tilde{p} - p| > \epsilon) \leq \frac{p(1-p)}{B\epsilon^2}$. So with probability at least $1 - \delta$, $|\tilde{p} - p| \leq \sqrt{\frac{p(1-p)}{B\delta}}$.

Remark 1: Notice that this procedure has an underlying assumption: the potential outcomes are treated as fixed quantities and the randomness comes solely from the treatment assignment vector. As a consequence, the inference is confined to the experiment. This is different from what we have covered in class, where we treated the potential outcomes (counterfactuals) as random variables, having their own distribution, albeit unobservable. Both approaches are widely used in practice; I hope you enjoyed learning about this framework!

3 No Paradox with Random Assignments

Let $X \in \{0, 1\}$, where $X = 1$ indicates that a subject is treated. Let Y and Z be continuous random variables. Y is an outcome of interest and Z denotes a collection of covariates. For example, Y could be a subject's cholesterol level, X could be the treatment variable for a new drug, and Z could denote various other lab values. Simpson's paradox is often stated for categorical covariates, but it could also happen in the continuous case. In particular, it reflects the fact that it is possible for the following inequalities to hold simultaneously:

$$\begin{aligned} \mathbb{E}(Y|Z, X = 1) &> \mathbb{E}(Y|Z, X = 0) \quad \text{almost everywhere,} \\ \mathbb{E}(Y|X = 1) &< \mathbb{E}(Y|X = 0) \end{aligned}$$

(a) Now show that it is impossible for the inequalities to hold simultaneously if $Z \perp\!\!\!\perp X$; that is, if Z and X are independent.

You can assume that $f_{Z|X=x}(z) > 0$ for $x \in \{0, 1\}$. You can also use without proof that the expectation is a monotone operator in the following sense:

Lemma 1. *If $X_1 \leq X_2$ almost everywhere (i.e. except on sets of probability 0), then $\mathbb{E}\{X_1\} \leq \mathbb{E}\{X_2\}$.*

Solution

Notice that $\mathbb{E}\{Y|Z, X = x\}$ is a random variable. In particular, it is a function of Z. So let us denote $g_x(Z) := \mathbb{E}\{Y|Z, X = x\}$. We know that $g_1(Z) > g_0(Z)$ almost everywhere, thus:

$$\begin{aligned}
\mathbb{E}\{Y|X = 1\} &= \int g_1(z) f_{Z|X=1}(z) dz \\
&= \int g_1(z) f_Z(z) dz \\
&= \mathbb{E}\{g_1(Z)\} \\
&> \mathbb{E}\{g_0(Z)\} \\
&= \int g_0(z) f_Z(z) dz \\
&= \int g_0(z) f_{Z|X=0}(z) dz \\
&= \mathbb{E}\{Y|X = 0\}
\end{aligned}$$

where we used the fact that $Z \perp\!\!\!\perp X$ for the second and fifth equalities and the Lemma for the inequality. Hence, if $Z \perp\!\!\!\perp X$, the two inequalities above cannot hold simultaneously.

(b) Construct a simple example where $Z \not\perp\!\!\!\perp X$ and the inequalities above hold simultaneously.

Solution

Let $Y \sim N(X + Z, 1)$, $Z \sim N(-2X + 1, 1)$ and $X \sim \text{Bern}(0.5)$. Then,

$$\mathbb{E}\{Y|Z, X = 1\} = Z + 1 > Z = \mathbb{E}\{Y|Z, X = 0\}$$

but

$$\mathbb{E}\{Y|X = 1\} = \int (z + 1) f(z|x = 1) dz = 0 < 1 = \int z f(z|x = 0) dz = \mathbb{E}\{Y|X = 0\}$$

as desired.