ST552 Homework 5

Nick Sun

February 20, 2019

Faraway 3.6

39 MBA students were asked about happiness and how this is related to income and social health.

a) Which predictors were statistically significant at the 1% level?

```
##
## Call:
## lm(formula = happy ~ ., data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7186 -0.5779 -0.1172  0.6340  2.0651
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.072081   0.852543  -0.085   0.9331
## money        0.009578   0.005213   1.837   0.0749 .
## sex         -0.149008   0.418525  -0.356   0.7240
## love         1.919279   0.295451   6.496 1.97e-07 ***
## work         0.476079   0.199389   2.388   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 34 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.6761
## F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

Only one of the predictor variables is significant at the $\alpha = .01$ level: `love`.

b) Are there potential violations in the assumptions used to perform the t-tests

Using the `table` function and also making a histogram:

```
##
##  2  3  4  5  6  7  8  9 10
##  1  1  4  5  2  8 14  3  1
```

One of the assumptions of using a t-test is normality of the response and normality of $\hat{\beta}$. The problem here is that the response only has discrete values and furthermore, the distribution of these values is not symmetric. Therefore, the response is not normal. A lack of normality in the distribution of the response will manifest as a lack of normality in the distribution of errors which prevents the t-test from being *exact*. We're probably still ok with using the t-test though since the t-test is pretty robust to violations of normality.

(c) Use permutation procedure to test the significance of `money`

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
## [1] 0.0762
```

We get a p-value of .0776 for the permuation approach which does not allow us enough evidence to reject at the $\alpha = .05$ level.

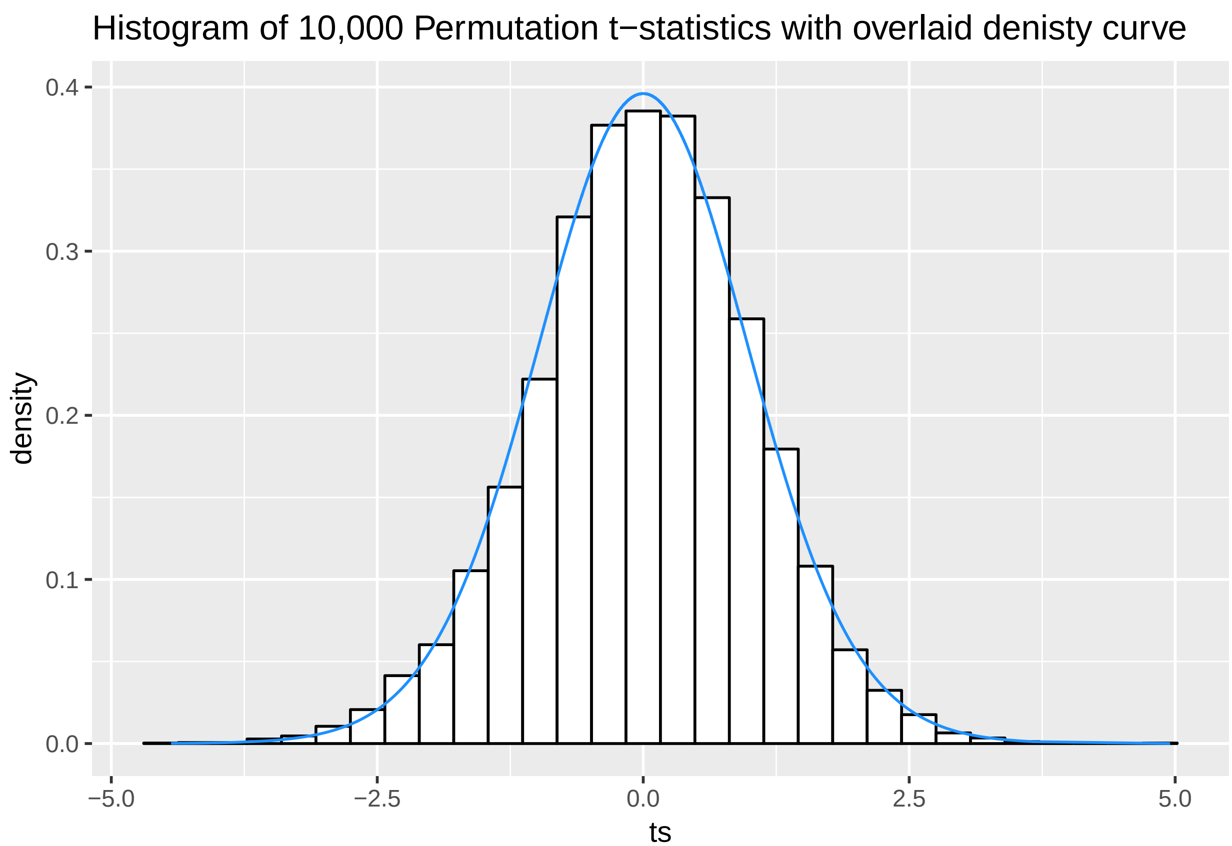
Compare this to the p-value we got for the t-test: .0749. They are actually pretty darn close I'd say.

(d) and (e) histogram and density curve of permutation t-statistics

```
grid <- seq(-3, 3, length=300)
tdensity <- dt(tstats, df = 34)
```

```r
tdata <- data.frame(ts = tstats, td = tdensity)
ggplot(tdata, aes(x = ts)) + geom_histogram(aes(y=..density..), color = "black", fill = "white") + geom_
  ggtitle("Histogram of 10,000 Permutation t-statistics with overlaid denisty curve")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(e) Construct a 95% Bootstrap Confidence Interval, permutation-style

```
##     Intercept       money        sex     love      work
## 5%  -1.377437 0.001453558 -0.7864842 1.464601 0.1760038
## 95%  1.265641 0.017652713  0.4986829 2.364874 0.7848325
```

```
##       Intercept        money        sex     love      work
## 2.5%  -1.626392 2.716464e-05 -0.8997172 1.372887 0.1231986
## 97.5%  1.513561 1.887393e-02  0.6112587 2.447659 0.8435497
```

The 95% confidence interval contains 0 which means that we will fail to reject $H_0 : \beta_{\text{money}} = 0$ at the $\alpha = .05$ level. This corresponds with our initial output from the `summary(happy_lm)` function where we also failed to reject.

Faraway 5.1

Fitting all 8 models using the **teengamb** data, oh boy~

```
##                                models sex coefficient p-value
## 5                  gamble, sex, status         -35.709  0.0005
## 3          gamble, sex, status, verbal         -33.752  0.0011
## 7                  gamble, sex, verbal         -27.722  0.0020
## 8                          gamble, sex         -25.909  0.0044
## 2          gamble, sex, status, income         -24.339  0.0045
## 4          gamble, sex, income, verbal         -22.960  0.0015
## 1 gamble, sex, status, income, verbal         -22.118  0.0101
## 6                  gamble, sex, income         -21.634  0.0027
```

The above is a table showing all 8 models, the coefficient for sex, and the respective p-value. For the most part, the effect of *sex* is fairly stable. It is consistently negative and in the range of $(-21.63, -35.709)$. This tells us at the very least that regardless of the model, women on average gamble less money than men, after accounting for other variables. Additionally, in each model the t-statistic p-value is significant at $\alpha = .05$ and for most of the cases also significant at $\alpha = .01$. Regardless of what model we pick, *sex* is considered an important variable and worth having in the model.

Interestingly, the effect of *sex* is noticeably higher in model 5 which has *sex* and *status* and in model 3 which has *sex*, *status*, and *verbal*. The interpretation for this might be that for these models, after accounting for *status* and *verbal* score, the difference between the sexes is greater than it would be when accounting for other potential covariates. Noticeably, there is a decrease in the magnitude of the effect of sex from model 5 to model 2 where we add *income*. Adding *income* and accounting for in our model somehow diminishes the effect of gender in the amount gambled.

Overall though, the main point stands - we have evidence to say that on average women gamble less than men.

Question 3: Nike Vaporflys! The shoe to beat

If I cared about running, I might be inclined to buy a pair.

(a) Is the difference practically significant?

For serious runners, even a 1% increase in speed is very good! Since the calibre of runners explored in the article are for the most part seem experienced, I think any advantage that the Vaporflys could provide would be practically significant.

(b) Brief description of where the data came from

The authors cite public race reports and Strava. Strava is self-branded as a social network for athletes. Strava itself collects data on the runners from smartphones or satellite watches, so presumably there is an app which records the runner's time and progression.

Notably, only a third of runners on Strava reported the shoe that they wore.

(c) Measuring shoe effects using statistical methods

What variables do the authors account for?

The authors account for age, gender, race history, and training done prior to the race.

One variable that might be confounding would be race history. Experienced runners are probably more likely to invest more in running equipment i.e. expensive Vaporfly shoes. Experienced runners are also probably likely to have coaches and trusted training techniques for improving their race times *so* when race time actually comes they have a greater chance than less experienced runners of getting a PR.

Specify a regression model you might start with to repeat their analysis

$$\text{change in race time}_i = \text{Vaporfly}_i + \text{previous race time}_i + \text{training miles}_i + \text{age}_i + \text{gender}_i + \text{weather conditions}_i$$

where Vaporfly_i is an indicator variable that takes on 1 when the runner is wearing Vaporflys and 0 if they are not.

(d) How did they calculate change in performance?

The section begins by comparing two specific runners who had similar race times to each other *and* ran in the same races (2017 and 2018 Boston Marathon). The actual visualization though has a different methodology; they do not compare pairs of similar runners and instead group together runners who switched to Vaporflys between 2017 and 2018 and then compared their change in performance with those who did not. The assumption here is that since we are comparing the same runner at different time points, we automatically get a easy-to-interpret comparison between similar race times without having to worry about matching runners who are similar in experience and training.

(e) How do the approaches in (c) and (d) deal with possible confounders?

The statistical model from *(c)* accounts for confounder by trying to include a bunch of covariates in the model that are probably related to both race time and shoe selection, ex: *age*, *gender*, *race experience*, etc. The con of this is that it is possible to still leave out confounders.

The direct comparison between runners in the same two races accounts for these unknown confounders by actually observing the runners on race day - conditions like weather and route are automatically factored into this comparison. However, this approach still is not perfect since it is possible for the runners to have completely different training regimens. The authors also note that runners could be saving their special shoes (like Vaporflys) for a day where they are feeling fast and definitely trying to PR. This would be a confounder that we cannot account for using this direct comparison.

The other direct comparison method of looking at the same runner between different years of the same race should hopefully eliminate training regimen as a confounder since a runner's training schedule probably doesn't change too much between races. However, now we introduce the obvious confounder of runners naturally improving over time - it is impossible to parse out if these runners are faster because they wore Vaporflys or if they are faster because they are *faster* than they were a year ago.

(f) What justification do the authors have?

The authors basic line of reasoning is that they tried a bunch of different methods and different comparisons and Vaporflys and the runners who wear them always come out on top. Different methods account for different confounders so potential problems with one analysis, for example, saving special shoes for especially fast races was accounted for by certain techniques even though it was missed by others. This kind of multiple analysis would be bad if the authors were hunting for some kind of association, but this scenario is different since the reason they are doing all these analyses is to corroborate a result.