POLS 503: Assigment 2

2017-04-21

This assignment works through an example in Yule (1899):

Yule (1899) is a published example multiple regression analysis in its modern form.[1]

Yule wrote this paper to analyze the effect of policy changes and implementation on pauperism (poor receiving benefits) in England under the English Poor Laws. In 1834, a new poor law was passed that established a national welfare system in England and Wales. The New Poor Law created new administrative districts (Poor Law Unions) to adminster the law. Most importantly, it attempted to standardize the provision of aid to the poor. There were two types of aid provided: in-relief or aid provided to paupers in workhouses where they resided, and out-relief or aid provided to paupers residing at home. The New Poor Law wanted to decrease out-relief and increase in-relief in the belief that in-relief, in particular the quality of life in workhouses, was a deterrence to poverty and an encouragement for the poor to work harder to avoid poverty.

Yule identifies that there are various potential causes of the change in rate of pauperism, including changes in the (1) law, (2) economic conditions, (3) general social character, (4) moral character, (5) age distribution of the population (pg. 250).

He astutely notes the following:

> If, for example, we should find an increase in the proportion of out-relief associated with (1) an increase in the proportion of the aged to the whole population, and also (2) an increase in the rate of pauperism, it might be legitimate to interpret the result in the sense that changes in out-relief and pauperism were merely simultaneous concomitants of changes in the proportion of aged-the change of pauperism not being a direct consequence of the change of administration, but both direct consequenices of the change in age distribution. It is evidently most important that we should be able to decide between two such differenit ilnterpretations of the same facts. This the method I have used is perfectly competernt to do — Yule (1899 pg. 250)

Setup

```
library("tidyverse")
library("modelr")
# devtools::install_github("jrnold/resamplr")
library("resamplr")
```

While only a subset of the original data of Yule (1899) was printed in the article itself, Plewis (2015) reconstructed the orginal data and Plewis (2017) replicated the original paper. This data is included in the package **datums**. This package is not on CRAN, but can be downloaded from github. **IMPORTANT** install the latest version of **datums** since a few fixes were recently made to the `pauperism` dataset.

```
# devtools::install_github("jrnold/datums")
library("datums")
```

The data for Yule (1899) is split into two data frames: `pauperism_plu` contains data on the Poor Law Unions (PLU), and `pauperism_year`, panel data with the PLU-year as the unit of observation.

```
pauperism <-
  left_join(datums::pauperism_plu, datums::pauperism_year,
```

[1] See Freedman (1997), Stigler (1990), Stigler (2016), and Plewis (2017) for discussions of Yule (1899).

```
            by = "ID") %>%
  mutate(year = as.character(year))
```

The data consist of 599 PLUs and the years: 1871, 1881, 1891 (years in which there was a UK census).

Yule (1899) is explcitly using regression for causal inference. The outcome variable of interest is:

- **Pauperism** the percentage of the population in receipt of relief of any kind, less lunatics and vagrants

The treatment (policy intervention) is the ration of numbers receiving outdoor relief to those receiving indoor relief.

- **Out-Relief Ratio:** the ratio of numbers relieved outdoors to those relieved indoors

He will control for two variables that may be associated with the treatment

- **Proportion of Old:** the proportion of the aged (65 years) to the whole population since the old are more likely to be poor.
- **Population:** in particular changes in population that may be proxying for changes in the economic, social, or moral factors of PLUs.

There is also **Grouping of Unions**, which is a locational classification based on population density that consists of Rural, Mixed, Urban, and Metropolitan.

Instead of taking differences or percentages, Yule worked with "percent ratio differences", $100 \times \frac{x_t}{x_{t-1}}$, because he did not want to work with negative signs, presumably a concern at the because he was doing arithmetic by hand and this would make calculations more tedious or error-prone.

Original Specification

Run regressions of `pauper` using the yearly level data with the following specifications. In Yule (1899), the regressions are

- *M1:* `paupratiodiff ~ outratiodiff + year + Type`
- *M2:* `paupratiodiff ~ outratiodiff + (popratiodiff + oldratiodiff) * (year + Type)`
- *M3:* `-1 + paupratiodiff ~ (outratiodiff + popratiodiff + oldratiodiff) * (year + Type)`
- *M4:* `paupratiodiff ~ (outratiodiff + popratiodiff + oldratiodiff) * (year + Type)`

1. Present the regressions results in a regression table
2. Interpret the coefficients for `outratiodiff` for each model.
3. Write the equations for each or all models, and describe the model with a sentence or two. Try to be as concise as possible. Look at recent journal articles for examples of the wording and format.
4. What is the difference between *M3* and *M4*. What are the pros and cons of each parameterization?
5. Conduct F-tests on the hypotheses:
6. All interactions in *M4* are 0
7. The coefficients on `outratiodiff` in *M4* are the same across years
8. The coefficients on `outratiodiff` in *M4* are the same across PLU Types
9. The coefficients on `outratiodiff` in *M4* are the same across PLU Types and years.

You can conduct F-tests with the function `anova(mod_unrestricted, mod_restricted)`.

5. Calculate the predicted value and confidence interval for the PLU with the median value of `outratiodiff`, `popratiodiff`, and `oldratiodiff` in each year and PLU Type for these models. Plot the predicted value and confidence interval of these as point-ranges.

6. As previously, calculate the predicted value of the median PLU in each year and PLU Type. But instead of confidence intervals include the prediction interval. How do the confidence and prediction intervals differ? What are their definitions?

Functional Forms

The regression line of the model estimated in Yule (1899) (ignoring the year and region terms and interactions) can be also written as

$$100 \times \frac{\texttt{pauper2}_t/\texttt{Popn2}_\texttt{t}}{\texttt{pauper2}_{t-1}/\texttt{Popn2}_{\texttt{t}-1}} = \beta_0 + \beta_1 \times 100 \times \frac{\texttt{outratio}_t}{\texttt{outratio}_{\texttt{t}-1}} + \beta_2 \times 100 \times \frac{\texttt{Popn65}_t/\texttt{Popn2}_t}{\texttt{Popn65}_{t-1}/\texttt{Popn2}_{t-1}} + \beta_3 \times 100 \times \frac{\texttt{Popn2}_t}{\texttt{Popn2}_{t-1}}$$

1. Write a model that includes only the log differences ($\log(x_t) - \log(x_{t-1})$) with only the `pauper2`, `outratio`, `Popn2`, and `Popn65` variables.
2. Estimate the model with logged difference predictors, Year, and month and interpret the coefficient on $\log(outratio_t)$.
3. What are the pros and cons of this parameterization of the model relative to the one in Yule (1899)? Focus on interpretation and the desired goal of the inference rather than the formal tests of the regression. Can you think of other, better functional forms?

Non-differenced Model

Suppose you estimate the model (*M5*) without differencing,

```
pauper2 ~ outratio + (Popn2 + Prop65) * (year + Type)
```

- Interpret the coefficient on `outratio`. How is this different than model *M2*?
- What accounts for the different in sample sizes in *M5* and *M2*?
- What model do you think will generally have less biased estimates of the effect of out-relief on pauperism: *M5* or *M2*? Explain your reasoning.

Substantive Effects

Read Gross (2014) and McCaskey and Rainey (2015). Use the methods described in those papers to assess the substantive effects of out-ratio on the rate of pauperism. Use the model(s) of your choosing.

Influential Observations and Outliers

Influential Observations for the Regression

For this use *M2*:

1. For each observation, calculate and explain the following:

- hat value (`hatvalues`)
- standardized error (`rstandard`)
- studentized error (`rstudent`)
- Cook's distance (`cooksd`)

2. Create an outlier plot and label any outliers. See the example here
3. Using the plot and rules of thumb identify outliers and influential observations

Influential Observations for a Coefficient

1. Run *M2*, deleting each observation and saving the coefficient for `outratiodiff`. This is a method called the jackknife. You can use a for loop to do this, or you can use the function `jackknife` in the package resamplr.
    1. For which observations is there the largest change in the coefficient on `outratiodiff`?
    2. Which observations have the largest effect on the estimate of `outratiodiff`? variable 本身的standard error 变化最大?
    3. How do these observations compare with those that had the largest effect on the overall regression as measured with Cook's distance?
    4. Compare the results of the jackknife to the `dfbeta` statistic for `outratiodiff`
2. Aronow and Samii (2015) note that the influence of observations in a regression coefficient is different than the the influence of regression observations in the entire regression. Calculate the observation weights for `outratiodiff`.
    1. Regress `outratiodiff` on the control variables
    2. The weights of the observations are those with the highest squared errors from this regression. Which observations have the highest coefficient values?
    3. How do the observations with the highest regression weights compare with those with the highest changes in the regression coefficient from the jackknife?

Omitted Variable Bias

An informal way to assess the potential impact of omitted variables on the coeficient of the variable of interest is to coefficient variation when covariates are added as a measure of the potential for omitted variable bias (Oster 2016). Nunn and Wantchekon (2011) (Table 4) calculate a simple statistic for omitted variable bias in OLS. This statistic "provide[s] a measure to gauge the strength of the likely bias arising from unobservables: how much stronger selection on unobservables, relative to selection on observables, must be to explain away the full estimated effect."

1. Run a regression without any controls. Denote the coefficient on the variable of interest as $\hat{\beta}_R$.
2. Run a regression with the full set of controls. Denote the coefficient on the variable of interest in this regression as $\hat{\beta}_F$.
3. The ratio is $\hat{\beta}_F/(\hat{\beta}_R - \hat{\beta}_F)$

Calculate this statistic for *M2* and interpret it.

Heteroskedasticity

Robust Standard Errors

1. Run *M2* and *M3* with a heteroskedasticity consistent (HAC) or robust standard error. How does this affect the standard errors on `outratio` coefficients? Use the **sandwich** package to add HAC standard errors (Zeileis 2004).

Multiple Regressions

1. Run the model with interactions for all years and types

```
lm(pauper2 ~ (outratio + Popn2 + Prop65) * year * Type - 1, data = pauperism)
```

2. For each subset of `year` and `type` run the regression

```
lm(pauper2 ~ outratio + Popn2 + Prop65)
```

3. Compare the coefficients, standard errors, and regression standard errors in these regresions.

To run the multiple regressions, save models as a list column `mod`, then save the results of `glance` and `tidy` as list columns:

```
all_interact <-
  crossing(Type = pauperism$Type, year = c(1881, 1891)) %>%
  mutate(mod = map2(year, Type,
                    function(yr, ty) {
                    lm(paupratiodiff ~ outratiodiff + popratiodiff + oldratiodiff,
                       data = filter(pauperism,
                                     year == yr,
                                     Type == ty))
                    })) %>%
  mutate(mod_glance = map(mod, broom::glance),
         mod_tidy = map(mod, broom::tidy))
```

Now extract parts of model. E.g. Standard errors of the regression:

```
all_interact %>%
  mutate(sigma = map_dbl(mod_glance, function(x) x$sigma)) %>%
  select(year, Type, sigma)
```

```
## # A tibble: 8 × 3
##    year         Type     sigma
##   <dbl>        <chr>     <dbl>
## 1  1881 Metropolitan  9.886436
## 2  1891 Metropolitan 24.790240
## 3  1881        Mixed 16.437527
## 4  1891        Mixed 17.403411
## 5  1881        Rural 13.801753
## 6  1891        Rural 17.078948
## 7  1881        Urban 19.523919
## 8  1891        Urban 25.557318
```

Weighted Regression

1. Run *M2* and *M3* as weighted regressions, weighted by the population (`Popn`) and interpret the coefficients on `outratiodiff` and interactions. Informally assess the extent to which the coefficients are different. Which one does it seem to affect more?
2. What are some rationales for weighting by population? See the discussion in Solon, Haider, and Wooldridge (2013) and Angrist and Pischke (2014).

Cross-Validation

When using regression for causal inference, model specification and choice should largely be based on avoiding omitted variables. Another criteria for selecting models is to use their fit to the data. But a model's fit to data should not be assessed using only the in-sample data. That leads to overfitting—and the best model would always be to include an indicator variable for every observation Instead, a model's fit to data can be assessed by using its out-of-sample fit. One way to estimate the *expected* fit of a model to *new* data is cross-validation.

We want to compare the predictive performance of the following models

```
mod_formulas <-
  list(
    m0 = paupratiodiff ~ 1,
    m1 = paupratiodiff ~ year + Type,
    m2 = paupratiodiff ~ outratiodiff + year + Type,
    m3 = paupratiodiff ~ outratiodiff + (popratiodiff + oldratiodiff) * (year + Type),
    m4 = -1  + paupratiodiff ~ (outratiodiff + popratiodiff + oldratiodiff) * (year + Type),
    m5 = paupratiodiff ~ (outratiodiff + popratiodiff + oldratiodiff) * year * Type
  )
```

Let's split the data into 10 (train/test) folds for cross-validation,

```
pauperism_nonmiss <-
  pauperism %>%
  filter(year %in% c(1881, 1891)) %>%
  select(paupratiodiff, outratiodiff, popratiodiff, oldratiodiff, year, Type, Region, ID) %>%
  tidyr::drop_na()
pauperism_10folds <-
  pauperism_nonmiss %>%
  resamplr::crossv_kfold(10)
```

For each model formula `f`, training data set `train`, and test data set, `test`, run the model specified by `f` on `train`, and predict new observations in `test`, and calculate the RMSE from the residuals

```
mod_rmse_fold <- function(f, train, test) {
  fit <- lm(f, data = as.data.frame(train))
  test_data <- as.data.frame(test)
  err <- test_data$paupratiodiff - predict(fit, newdata = test_data)
  sqrt(mean(err ^ 2))
}
```

E.g. for one fold and formula,

```
mod_rmse_fold(mod_formulas[[1]], pauperism_10folds$train[[1]],
              pauperism_10folds$test[[1]])
```

```
## [1] 19.5354
```

Now write a function that will calculate the average RMSE across folds for a formula and a cross-validation data frame with `train` and `test` list-columns:

```
mod_rmse <- function(f, data) {
  map2_dbl(data$train, data$test,
           function(train, test) {
             mod_rmse_fold(f, train, test)
           }) %>%
    mean()
}
```

```
mod_rmse(mod_formulas[[1]], pauperism_10folds)
```

```
## [1] 24.05803
```

Finall, we want to run `mod_rmse` for each formula in `mod_formulas`. It will be easiest to store this in a data frame:

```
cv_results <- tibble(
  model_formula = mod_formulas,
  .id = names(mod_formulas),
```

```
  # Formula as a string
  .name = map(model_formula,
              function(x) gsub(" +", " ", paste0(deparse(x), collapse = "")))
)
```

Use `map` to run `mod_rmse` for each model and save it as a list frame in the data frame,

```
cv_results <-
  mutate(cv_results,
         cv10_rmse = map(model_formula, mod_rmse, data = pauperism_10folds))
```

In the case of linear regression, the MSE of the Leave-one-out (n-fold) cross-validation can be analytically calculated without having to run n regressions.

```
loocv <- function(x) {
  mean((residuals(x) / (1 - hatvalues(x))) ^ 2)
}
```

We

```
cv_results <-
  mutate(cv_results,
         rmse_loo = map(mod_formulas, function(f) sqrt(loocv(lm(f, data = pauperism_nonmiss)))))
```

1. In the 10-fold cross validation, which model has the best out of sample prediction?
2. Using the LOO-CV cross-validation, which model has the best
3. Does the prediction metric (RMSE) and prediction task—predicting individual PLUs from other PLUs—make sense? Can you think of others that you would prefer?

Bootstrapping

Estimate the 95% confidence intervals of model with simple non-parametric bootstrapped standard errors. The non-parametric bootstrap works as follows:

Let $\hat{\theta}$ be the estimate of a statistic. To calculate bootstrapped standard errors and confidence intervals use the following procedure.

For samples $b = 1, ..., B$.

1. Draw a sample with replacement from the data
2. Estimate the statistic of interest and call it θ^*_b.

Let $\theta^* = \{\theta^*_1, \ldots, \theta^*_B\}$ be the set of bootstrapped statistics.

- standard error: $\hat{\theta}$ is $\text{sd}(\theta^*)$.
- confidence interval:
  - normal approximation. This calculates the confidence interval as usual but uses the bootstrapped standard error instead of the classical OLS standard error: $\hat{\theta} \pm t_{\alpha/2,df} \cdot \text{sd}(\theta^*)$
  - quantiles: A 95% confidence interval uses the 2.5% and 97.5% quantiles of θ^* for its upper and lower bounds.

Original model

```
mod_formula <- paupratiodiff ~ outratiodiff + (popratiodiff + oldratiodiff) * year * Type
mod_orig <- lm(mod_formula, data = pauperism_nonmiss)
```

```
bs_coef_se <-
  resamplr::bootstrap(pauperism_nonmiss, 1024) %>%
```

```
  # extract the strap column
  `[[`("sample") %>%
  # run
  map_df(function(dat) {
    lm(mod_formula, data = dat) %>%
    broom::tidy() %>%
    select(term, estimate)
  }) %>%
  # calculate 2.5%, 97.5% and sd of estimates
  group_by(term) %>%
  summarise(
    std.error_bs = sd(estimate),
    conf.low_bsq = quantile(estimate, 0.025),
    conf.low_bsq = quantile(estimate, 0.975)
  )
```

Now compare the std.error of the original and the bootstrap for `outratiodiff`

```
broom::tidy(mod_orig, conf.int = TRUE) %>%
  select(term, estimate, std.error) %>%
  filter(term == "outratiodiff") %>%
  left_join(bs_coef_se, by = "term")
```

```
##           term  estimate  std.error std.error_bs conf.low_bsq
## 1 outratiodiff 0.2274375 0.01433042    0.0188327    0.2685325
```

The bootstrap standard error is slightly higher. It is similar to the standard error generated using the heteroskedasticity consistent standard error.

```
sqrt(sandwich::vcovHC(mod_orig)["outratiodiff", "outratiodiff"])
```

```
## [1] 0.01985823
```

It is likely that there is correlation between the error terms of observations. At the very least, each PLU is included twice; these observations are likely correlated, so we are effectively overstating the sample size of our data. One way to account for that is to resample "PLUs", not PLU-years. This cluster-bootstrap will resample each PLU (and all its observations), rather than resampling the observations themselves.

```
pauperism_nonmiss %>%
  group_by(ID) %>%
  resamplr::bootstrap(1024) %>%
  # extract the strap column
  `[[`("sample") %>%
  # run
  map_df(function(dat) {
    lm(mod_formula, data = dat) %>%
    broom::tidy() %>%
    select(term, estimate)
  }) %>%
  # calculate 2.5%, 97.5% and sd of estimates
  group_by(term) %>%
  summarise(
    std.error_bs = sd(estimate),
    conf.low_bsq = quantile(estimate, 0.025),
    conf.low_bsq = quantile(estimate, 0.975)
  ) %>%
  filter(term == "outratiodiff")
```

```
## # A tibble: 1 × 3
##           term std.error_bs conf.low_bsq
##          <chr>        <dbl>        <dbl>
## 1 outratiodiff   0.01808499    0.2654708
```

However, this yields a standard error not much different than the Robust standard error.

1. Try bootstrapping "Region" and "BoothGroup". Do either of these make much difference in the standard errors.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics*. Princeton UP.

Aronow, Peter M., and Cyrus Samii. 2015. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60 (1). Wiley-Blackwell: 250–67. doi:10.1111/ajps.12185.

Freedman, David. 1997. "From Association to Causation via Regression." *Advances in Applied Mathematics* 18 (1). Elsevier BV: 59–110. doi:10.1006/aama.1996.0501.

Gross, Justin H. 2014. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science* 59 (3). Wiley-Blackwell: 775–88. doi:10.1111/ajps.12149.

McCaskey, Kelly, and Carlisle Rainey. 2015. "Substantive Importance and the Veil of Statistical Significance." *Statistics, Politics and Policy* 6 (1-2). Walter de Gruyter GmbH. doi:10.1515/spp-2015-0001.

Nunn, Nathan, and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101 (7): 3221–52. doi:10.1257/aer.101.7.3221.

Oster, Emily. 2016. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, September. Informa UK Limited, 0–0. doi:10.1080/07350015.2016.1227711.

Plewis, Ian. 2015. "Census and Poor Law Union Data, 1871-1891." SN 7822. UK Data Service; data collection. doi:10.5255/UKDA-SN-7822-1.

———. 2017. "Multiple Regression, Longitudinal Data and Welfare in the 19th Century: Reflections on Yule (1899)." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, February. Wiley-Blackwell. doi:10.1111/rssa.12272.

Solon, Gary, Steven Haider, and Jeffrey Wooldridge. 2013. "What Are We Weighting for?" National Bureau of Economic Research. doi:10.3386/w18859.

Stigler, Stephen M. 1990. *The History of Statistics: The Measurement of Uncertainty Before 1900*. HARVARD UNIV PR. http://www.ebook.de/de/product/3239165/stephen_m_stigler_the_history_of_statistics_the_measurement_of_uncertainty_before_1900.html.

———. 2016. *The Seven Pillars of Statistical Wisdom*. Harvard University Press. http://www.ebook.de/de/product/25237216/stephen_m_stigler_the_seven_pillars_of_statistical_wisdom.html.

Yule, G. Udny. 1899. "An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I.)." *Journal of the Royal Statistical Society* 62 (2). JSTOR: 249. doi:10.2307/2979889.

Zeileis, Achim. 2004. "Econometric Computing with Hc and Hac Covariance Matrix Estimators." *Journal of Statistical Software* 11 (1): 1–17. doi:10.18637/jss.v011.i10.