PHP2550: Practical Data Analysis

Assignment 3: Regression Analysis

Antonella Basso

November 4, 2022

1. Logistic Regression Paper

First, read the paper *Predicting lung cancer prior to surgical resection in patients with lung nodules* by Deppen et al. available on Canvas. This paper introduces a model called `TREAT` that is currently used in practice to predict lung cancer. Then, respond to the questions below.

a. Write a two-paragraph summary of the paper.

b. Compare the Mayo model to the TREAT model in terms of the initial goals of building the model, the population the training data represented, the variables included, and the resulting model.

c. How was missing data handled?

d. What measures or visuals were used to evaluate the models? How do we interpret these?

e. What were some limitations that the paper addressed?

Solution

a. This paper discusses a lung cancer research endeavor to develop a novel clinical predictive model aimed at identifying suspicious lung lesions in individuals with existing pulmonary nodules during preoperative evaluation. The development of the Thoracic Research Evaluation And Treatment (TREAT) model reflects an attempt to further reduce unnecessary operations (lung resections) for benign disease, improving upon current poorly calibrated models by estimating a lesion's "probability of malignancy at the point of surgical evaluation", rather than merely focusing on the improvement of screening and biopsy referrals in "general medical populations"—a valuable predictive task, which no existing model of this kind presently addresses. Moreover, the paper describes the corresponding study conducted to evaluate the performance of the TREAT model in comparison to the Mayo Clinic model with an emphasis on discrimination, calibration, overfitting, and overall diagnostic accuracy. Specifically, the model was implemented on a total of 492 eligible patients from the 606 reflected in VUMC's Thoracic Surgery Quality Improvement database and clinic records who received lung nodule (or mass) evaluations for known or suspected non-small cell lung cancer (NSCLC) from January 2005 to October 2010. For the purpose of examining the generalizability of the model, TREAT was subsequently validated on the 226 eligible patients identified within the 254 constituting the Tennessee Valley Veterans Affairs Cohort (VA)—a group of individuals receiving thoracic operations for confirmed or suspected lung cancer between January 2005 and December 2013.

   Having appealed to logistic regression for predicting malignancy, results showed lung cancer prevalences among the VUMC cohort and VA validation cohort of 72% and 93%, respectively. Moreover, in addition to finding a non-linear relationship, modeled as a restricted cubic spline, between lung cancer and smoking intensity (measured by pack-years), the TREAT model identified positive associations between risk of lung cancer and age; pre-operative lesion size and growth; previous cancer; and FDG-PET scan avidity. While internal validation for the TREAT model with optimism corrected bootstrapping (with

replacement, 500 iterations) resulted in an AUC of 0.87 (95%CI: $0.83 - 0.92$) and a Brier score of 0.12 (95%CI: $0.10 - 0.14$), external validation resulted in an AUC of 0.89 (95%CI: $0.79 - 0.98$) and a Brier score of 0.08 (95%CI: $0.06 - 0.10$)[1]—demonstrating the model's overall ability to discriminate between cancer and benign disease (AUC scores closer to 1), as well as its calibration, that is, the accuracy of its probabilistic predictions (Brier scores closer to 0).

b. As highlighted in part (a), while the Mayo Clinic model was designed to "evaluate nodules in patients selected from a general population who had a lesion found on imaging", the TREAT model was constructed with the more specific aim of estimating a lesion's "probability of malignancy at the point of surgical evaluation" for the purpose of aiding in the diagnosis of suspected lung cancer "without missing early stage disease". It is reasonable to infer that this difference in initial goals is related to the observed differences in model covariate choices and, subsequently, model performance with regards to calibration and overall diagnostic accuracy, as suggested in the paper. In particular, the Mayo model includes variables for an individual's age, smoking history (Y/N), and status of previous cancer, as well as their lesion's size, spiculation, and location. Conversely, the TREAT model includes variables for an individual's age, gender, BMI, COPD, smoking intensity (pack-years), hemoptysis, and status of previous cancer, as well as their lesion's size, growth, spiculation, location, and FDG-PET scan avidity—a set of predictors carefully chosen due to their relevance in either "previously published and validated models"; recent guidelines for surgical evaluation referral and diagnosis; or observations made by medical experts (thoracic surgeons). Of the six covariates found to be correlated with lung cancer by the TREAT model, and hence significant in predicting risk for malignancy—smoking intensity (in pack-years), age, previous cancer, FDG-PET scan avidity, and pre-operative lesion size and growth—only three of them were included in the Mayo model; suggesting the importance of the remaining variables for diagnostic accuracy given the TREAT model's superior performance. As mentioned in the paper, such additional variables "improved the discrimination between benign disease and lung cancer".

   Specifically, results showed that while the Mayo model, whose "published coefficients to estimate lung cancer risk" were used for comparative analysis, had AUC scores of 0.80 (95%CI: $0.75 - 0.85$) and 0.73 (95%CI: $0.60 - 0.85$) for the VUMC development and VA validation datasets, respectively, the bootstrapped TREAT model had corresponding AUC scores of 0.87 (95%CI: $0.83 - 0.92$) and 0.89 (95%CI: $0.79 - 0.98$)—reflecting the TREAT model's superior performance, as well as the Mayo model's decreasing accuracy in discriminating malignancy with increasing disease prevalence. Similarly, the Brier scores for the Mayo Clinic model were higher that those observed in the TREAT model; 0.17 (95%CI: $0.15 - 0.19$) and 0.18 (95%CI: $0.15 - 0.21$) for the VUMC and VA cohorts, respectively—also demonstrating the Mayo model's decreasing calibration with increasing lung cancer prevalence. Although the Mayo Clinic model, on average, overestimated the risk of lung cancer, predicting higher probabilities of disease among individuals with no lesion growth and non-avid FDG-PET scans compared to the TREAT model (i.e., factors not considered by the Mayo model but included in the TREAT model), researchers also point out that the Mayo model appeared to underestimate this risk in lower quintiles among the study populations, thereby limiting "its use in clinical practice for patients being evaluated for surgery". It is important to note however, that lung cancer prevalence in the VUMC cohort used to train the TREAT model was much higher (72%) than that reflected in the population used to train (23%) and validate (44%) the Mayo Clinic model. This, in tandem with the Mayo model's more general initial aims and limited number of significant predictors of malignancy, is the likely reason for its "poor calibration in patients referred for surgical evaluation" and inconsistent predictive discrimination compared to the TREAT model, making the latter better suited for "providing clinical guidance in estimating individual likelihood of lung cancer".

c. As stated by the authors and shown in Table 1, complete covariate data were available for $264/492 \approx 53.66\%$ of individuals in the VUMC development cohort and $136/226 \approx 60.18\%$ of individuals in the VA validation cohort, with those having confirmed cancer diagnoses being more likely to have complete observed data. Although missingness was most prevalent with regards to FDG-PET scan avidity and lesion growth, the remaining variables of interest had no more than 16% of their values missing for either dataset. In turn, researchers utilized only the observed data to conduct their analysis of

[1] Differs from the value of 0.13 reflected in the paper's abstract.

demographic variables and pre-specified lung cancer predictors prior to model fitting. Subsequently, researchers implemented multiple imputation techniques and predictive mean matching to correct for missingness in the datasets used to train and validate the TREAT model. As mentioned in the paper, the corresponding multiple imputation assumptions were examined following the methods of Potthoff et al., which were were determined to not be in conflict with the assumptions put in place by the missing data. Nonetheless, the presence of missing data, particularly as it pertains to the variables found to be significant predictors of lung cancer, poses potential problems regarding accuracy in the model's predictions. That is, despite the TREAT model's observed generalizability and superior performance, it is possible that having had access to more complete training and testing data would have further improved its diagnostic accuracy. In this light, researchers argue that the decision to include patients who did not have a surgical resection, instead having undergone radiographic surveillance, despite having increased the amount of missing data, actually "excluded a bias in the spectrum of risk encountered by clinicians" and minimized potential bias due to imbalance in the cohort (i.e., underrepresentation of this class of individuals). Notably, they claim that the observed pattern of missingness with regards to certain covariates (e.g., FEV1) may be attributed to the fact that certain tests are only typically performed on patients for whom resection is deemed likely, suggesting that data may not be in fact missing at ramdom (NMAR). Although this assumption contradicts that of the multiple imputation algorithm used in the study, subsequent sensitivity analysis confirmed that implementing the TREAT model only on complete cases yielded similar results to those observed when modeling on the imputed data.

d. Overall model performance on the VUMC and VA datasets were compared and assessed for both the TREAT and Mayo Clinic models on the basis of Brier scores and AUCs—a graph for which is given by Figure 2. While the former measure is used to evaluate the accuracy of a model's probabilistic predictions or "forecasts" regarding a set of events with observed binary outcomes, the latter is used to gauge a model's ability to distinguish between (positive and negative) outcome classes (e.g., lung cancer vs. benign disease). Formally, these can be interpreted as the mean squared error (MSE) between predicted probabilities and observed values, and the area under the receiver operating curve (ROC)[2], respectively. Although both measures range between 0.0 and 1.0, a well-calibrated model will see Brier scores close to 0.0, given that low values convey small discrepancies between probability forecasts and observed outcomes, and hence, little predictive error. Meanwhile, AUC values close to 1.0 suggest high class separability, with a value of or near 0.5 implying that the model has no discrimination capacity to distinguish between a positive and a negative class. Thus, AUC scores of 0.87 and 0.89 for the TREAT model on the VUMC and VA cohorts, demonstrate the model's strength in its ability to discriminate between lung cancer and benign disease—which is not only superior in comparison to that of the Mayo Clinic model, as previously mentioned, but more generalizable given that the Mayo model's diagnostic accuracy decreases with increased disease prevalence. In a similar vein, having displayed low and decreasing Brier scores of 0.12 and 0.07 with increased disease prevalence, both attests to the TREAT model's superior calibration and further supports its potential to generalize to a the population. Particularly, given that, in contrast to the Mayo model, the TREAT model's lower Brier scores indicate that it is able to model risk more effectively, with their decreasing behavior suggesting that it becomes less prone to predictive error in the face of growing lung cancer prevalence.

e. As mentioned in part (c), the presence of missingness in covariate data may have introduced bias into the TREAT model, reflecting one of the study's limitation in challenging the validity of its estimates. As discussed in the paper however, the observed patterns of missingness, which saw a shift in the corresponding (missing data) assumptions, rendered this potential for bias not immediately threatening to the model's reliability (although worthy of mention), especially when scrutinized under subsequent sensitivity analysis. An additional constraint identified by researchers was found to be particularly relevant with regards to the model's generalizability. Namely, the fact that the external validation cohort, in addition to having displayed a high prevalence of lung cancer, represents a predominantly male Veteran population with preoperative symptoms and disproportionately high smoking incidence, in tandem with the fact that the development data was obtained from a single academic medical center's database

[2]The curve obtained from plotting the model's true positive rate (i.e., "recall" or sensitivity") against its false positive rate (i.e., 1-"specificity"), as shown in Figure 2.

which embodied a carefully selected set of variables, could potentially restrict the pool of individuals to whom the model may apply. Specifically, despite not having observed a decline in evaluation measures (AUC and Brier scores) which is indicative of the model's capacity for generalizability, the potential for other populations of prospective lung resection recipients to comprise a diverse set of referral patterns and disease factors, all while reflecting varying disease prevalence, calls for further investigative work with such differences in mind to improve upon, validate, and further extend the generalizability of the TREAT model prior to its clinical use.

2. Regression Application

We will reproduce the results from the paper on abortion legislation from last assignment and extend the results. The data used for this problem comes from the American Community Survey (https://www.census.gov/programs-surveys/acs), scraped state health department websites (source http://www.johnstonsarchive.net/policy/abortion/), and scraped abortion provider locations (original source no longer online but the data was collected in 2017-2018). The file `reproductive health.csv` contains all this information by county while the file `state laws.csv` has whether a state is characterized as hostile under the conditions mentioned in the paper. I've also included the full preprocessing csv and original data on Canvas if you are interested (thanks to Rob Zielinski and Sarah Voter for providing their code from their final project last year).

First, fit a logistic regression model to predict whether a state is highly restrictive based on demographic information (you do not need to consider model building in this stage—use the same variables used in the paper). Then, use this model to build a propensity-score weighted, regression model to predict the rate of abortions per 1,000 women to replicate the result in the Brown et al paper. Compare your results to the original paper. You **do not** have to use cluster-robust standard errors.

Then, consider including other public health and economic factors in the model. The goal of this extension is to focus on the role that public services including public health insurance coverage, and specifically coverage of abortions under Medicaid, may play in influencing abortion rates.

Solution

I. **Model Replication**

With the goal of eliminating potential confounding due to factors observed in the data, we turn to propernsity score (PS) weighting (i.e., inverse probability (IP) weighting), following the paper, to obtain an unbiased causal effect estimate prior to recreating the model. Part of the appeal of this method comes from the fact that it offers a simple weight-based approach for creating a "pseudo-population" from our data in which legislative climate is independent of the observed demographic factors. In turn, this allows us to implement a propensity-score weighted regression model for predicting county-level abortion rates from legislative climate without the influence of demographic variables that may confound the causal effect.

Logistic Regression: PS/IP weighting.

1. Use logistic regression to estimate propensity of having a restrictive legislative climate given a set of demographic factors, X.
2. Weight outcomes according to units' inverse probability of exposure (i.e., propensity score $p(x) = \mathrm{P}(T = 1|X = x)$), such that a county's assigned weight is:

$$w(x) = \begin{cases} \frac{1}{p(x)}, & \text{if } T = 1; \\ \frac{1}{1-p(x)}, & \text{if } T = 0, \end{cases}$$

where T denotes the treatment variable `is_highly_restrictive__20X0`.

Following the specifications provided in the paper, which mention the use of county-specific demographic data—"including the percentage of the population in each race and ethnicity category, median income, and total female population"—we implement the following logistic regression model to derive county-level weights for a given year (2000, 2010, or 2020), with `white_pct` and `democrat_2008` as the racial and political reference groups, respectively.

```
glm(is_highly_restrictive__20X0 ~
        women + grad_pct +
        black_pct + native_american_pct +
        asian_pct + hispanic_pct +
```

```
            median_income + democrat_2008,
        family=binomial(link="logit"), data=abortion)
```

Seeing as insufficient data led researchers to exclude the year 2000 form their analysis, in tandem with the fact that the study period did not extend beyond the year 2014, we use the PS weights derived for the 2010 year.

Regression Model: Study model replication.

Subsetting the data to reflect the 18 states considered in the study's DiD model, we construct the following replica, leaving out the fixed effects. Model coefficients and their corresponding 95% confidence intervals are also provided below.

```
lm(abortion_rate_2010 ~
       dist_to_closest_facility_miles +
       is_highly_restrictive__2010,
   data=abortion_states,
   weights=abortion_states$ps_weight_2010)
```

	Est.	2.5%	97.5%
(Intercept)	4.2364279	4.0130590	4.4597969
dist_to_closest_facility_miles	-0.0175533	-0.0200441	-0.0150625
is_highly_restrictive___2010	-0.6420688	-0.8676963	-0.4164413

The model coefficient for the binary `is_highly_restrictive__2010` varibale suggests that adoption of a highly restrictive legislative climate is associated with an abortion rate decrease of approximately 0.64 abortions per 1,000 women, adjusting for distance. On the other hand, results from the study model discussed in the paper show that, compared to a less restrictive climate, a highly restrictive legislative climate corresponds to an abortion rate decrease of 0.48 (95% CI: -0.92, -0.04) abortions per 1,000 women. Although our results reflect a slight increase in legislative climate effect, our estimate lies within the confidence interval of the study model's estimate. Similarly, given that converse also holds, this difference is not immidiately concerning. Rather, it can most likely be attributed to our ommision of two-way fixed effects and cluster-robust standard errors. It should be noted however, that the CI for the estimate discussed in the paper is nearly twice as large as that shown here.

II. Model Extension

With the model extension objective to explore the role that public services; such as public health insurance coverage, including Medicaid; may have on abortion rates, we consider the following list of economic and public health-related factors that may be included in the replicated study model (above) to accomplish this.

```
interest_vars <- c("earnings_diff",
                   "pct_unemployed",
                   "pct_retirement_income",
                   "pct_public_assistance",
                   "mean_public_assistance_income",
                   "pct_food_stamps",
                   "pct_health_insurance_covered",
                   "pct_private_health_insurance",
                   "pct_public_health_insurance",
                   "pct_no_health_insurance",
```

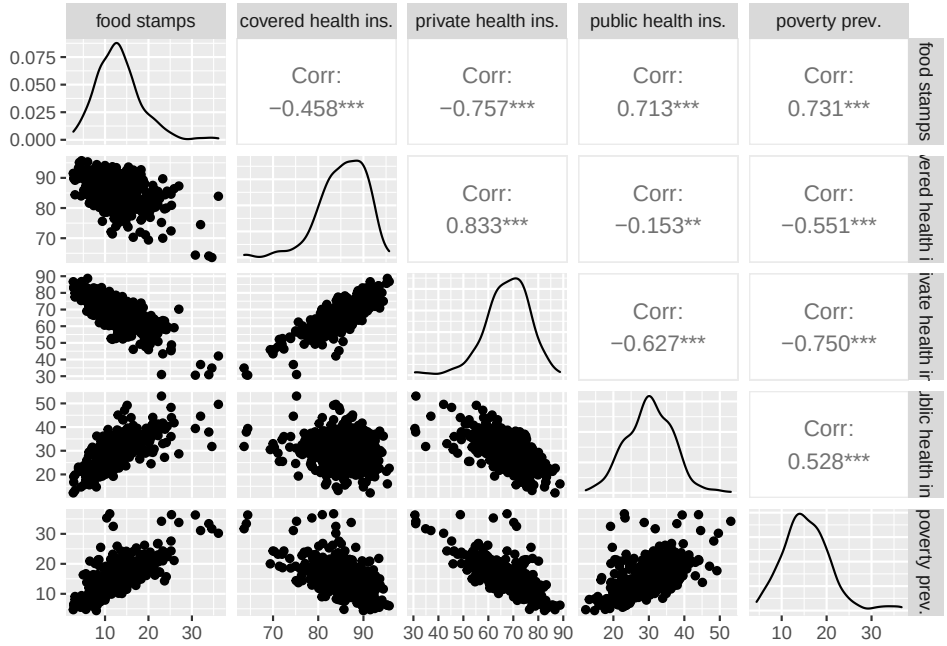
```
                    "pct_poverty_prev_12",
                    "medicaid_cover")
```

Prior to model building, we decide to look for possible correlations among our variables of interest to avoid potential multicollinearity. Specifically, we apply a function to each variable in the list to determine the corresponding set of factors for which strong correlation (set at a threshold of 0.7 in absolute terms) is present within this space.

```
var_corrs_f <- function(data, var){
  corrs <- cor(data, use="complete.obs")
  high_corr <- corrs[abs(corrs[,var])>0.7, ]
  return(rownames(high_corr))
}
```

Limiting the list of variables of interest to those that are strictly numeric for this purpose, we find that, of the remaining 11, a total of 6 factors were strongly correlated with at least one other numeric variable of interest—`"pct_food_stamps"`, `"pct_health_insurance_covered"`, `"pct_private_health_insurance"`, `"pct_public_health_insurance"`, `"pct_no_health_insurance"`, and `"pct_poverty_prev_12"`. Investigating their relationships more closely with the help of `ggpairs`, we find that `"pct_no_health_insurance"` is the direct inverse of `"pct_health_insurance_covered"`. Hence, we know to exclude the former in the extended model. Moreover, restricting our attention to positive correlations, we find that although `"pct_food_stamps"` is correlated with `"pct_public_health_insurance"`, the strength of its relationship with `"pct_poverty_prev_12"` is ever greater, suggesting that `"pct_food_stamps"` and `"pct_poverty_prev_12"` are likely strong predictors of one another. This being the case, it may be best to remove whichever factor appears least significant during model selection. However, given the fact that `"pct_food_stamps"` is still significantly correlated with `"pct_public_health_insurance"`, we may find neither `"pct_food_stamps"` nor `"pct_poverty_prev_12"` are actually worth keeping in the model; i.e., it is possible that, given the aforementioned relationships, we don't gain any significant information about abortion rates from their inclusion.

Nonetheless, these associations are all important to consider as we segway into the building the extended model. For reference, the `ggpairs` plot below, allows us to visualize the correlations found to be significant among our variables of interest. Surprisingly, we see that `"pct_private_health_insurance"` is strongly correlated with `"pct_health_insurance_covered"`, while not so much with `"pct_public_health_insurance"`. Although this could be due to what the variables actually represent—given the prominence of private health insurence in the U.S. for example, it would not be surprising to observe the majority of counties' health insurence coverage coming from private companies. It should be noted however, that the sum of `"pct_public_health_insurance"` and `"pct_private_health_insurance"`, does not equal, and is in fact larger than, `"pct_health_insurance_covered"` for most counties. For this reason, including all three variables in the model may not cause problems of multicollinearity, that is, if all happen to be significant, as we later show.



With these associations in mind, we now turn to building our extended model, employing a backward elimination model selection procedure. We begin with the full model (`study_model_ext_f`), including the predictors relevant to the study aims, and excluding `"pct_no_health_insurance"`, as well as `"pct_poverty_prev_12"` due to its immediate lack of significance in the model.

```
study_model_ext_f <- lm(abortion_rate_2010 ~
                          dist_to_closest_facility_miles +
                          is_highly_restrictive__2010 +
                          earnings_diff +
                          pct_retirement_income +
                          pct_unemployed +
                          pct_public_assistance +
                          mean_public_assistance_income +
                          pct_food_stamps +
                          pct_health_insurance_covered +
                          pct_private_health_insurance +
                          pct_public_health_insurance +
                          medicaid_cover,
                        data=abortion_states,
                        weights=abortion_states$ps_weight_2010)
```

Maintaining the same subset of data used for fitting the full model (`full_model_data <- study_model_ext_f$model`) for the purpose of more thorough model comparisons[3], we fit the first nested model (`study_model_ext1`), which differs from the full model in the exclusion of three terms; `"pct_food_stamps"`, `"pct_public_assistance"`, and `"mean_public_assistance_income"`; due to sequential insignificance under backward elimination. For the last model we consider (`study_model_ext2`), we remove two additional terms; `"pct_retirement_income"`, and `"pct_unemployed"` ; in a similar backward elimination process to reduce the size of the model by iteratively excluding the least significant covariates.

[3]Given that the subset of the data used for fitting each model becomes larger as we remove terms that displayed significant quantities of missing values, we proceed with the complete cases relevant to the covariates included in the full model. This allows for more more robust model comparisons between the full model and subsequent nested models via methods such as likelihood ratio tests (LRTs).

We now compare the three extended models by means of likelihood ratio tests (LRTs), adjusted R^2, and AIC values. Starting with the first approach, we perform a total of two LRTs; initially comparing the full model to the first nested model, and subsequently comparing the first and second nested models. The outcomes are as follows.

```
lrtest(study_model_ext_f, study_model_ext1):

  #Df  LogLik Df Chisq Pr(>Chisq)
1  14 -854.88
2  11 -843.65 -3 22.47  5.209e-05 ***


lrtest(study_model_ext1, study_model_ext2):

  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -843.65
2   9 -847.91 -2 8.5291    0.01406 *
```

At the 0.05 significance level, these results suggest that, in either case, the more complex model is preferable over the nested model; namely, `study_model_ext_f`. Specifically, in having rejected the null hypothesis that both models fit the data equally well in light of the omitted predictors having no significant effect over the response. By contrast, evaluating the three models on the basis of both their adjusted R^2 and AIC values, we find that the first nested model provides the better fit in each case. That is, looking at the table of values provided below, we see that `study_model_ext1` not only yields the highest adjusted R^2, but the lowest AIC value as well. Although there is little difference between derived adjusted R^2 values, the fact that `study_model_ext1` produced a value not only close to, but higher than that given by the full model, implies that the additional factors in the complex model do not help explain any of the variation present in the data. Moreover, the fact that `study_model_ext1` also displayed the lowest AIC value further suggests that, compared to the more complex and more simple model, it provides a better fit to the data.

Model	Adj. R^2	AIC
study_model_ext_f	0.3932461	1737.760
study_model_ext1	0.3948453	1709.291
study_model_ext2	0.3847810	1713.820

Based on all three measures used for comparison, we determine that `study_model_ext1` is the best-fit model, the coefficients for which are given below.

	Est.	2.5%	97.5%
(Intercept)	-0.0057121	-4.8532876	4.8418634
dist_to_closest_facility_miles	-0.0190868	-0.0251555	-0.0130182
is_highly_restrictive___2010	0.0596425	-0.4614115	0.5806965
earnings_diff	-0.0001235	-0.0001782	-0.0000688
pct_retirement_income	-0.1008964	-0.1833896	-0.0184032
pct_unemployed	0.0844815	-0.0005522	0.1695151
pct_health_insurance_covered	0.4515773	0.2859054	0.6172492
pct_private_health_insurance	-0.3638206	-0.4993886	-0.2282526
pct_public_health_insurance	-0.2437425	-0.3655213	-0.1219637
medicaid_coverTRUE	1.8930433	1.2575289	2.5285576

These results, suggest that although the variables `"pct_food_stamps"`, `"pct_public_assistance"`, and `"mean_public_assistance_income"`, may be associated with the outcome of interest, their inclusion in the model does not significantly improve the prediction of county-level abortion rates given the other model covariates. This may, in part, be due to the fact that several variables are strongly correlated, as shown previously. However, it is possible for our threshold of 0.7 for "significant correlation" may have been set too high to detect those possibly reflected here. Moreover, the fact that this model provides the best fit is also indicative of the fact that `"pct_retirement_income"` and `"pct_unemployed"` are significant, at least in the presence of the remaining factors, for predicting county-level abortion rates. More importantly however, this result validates the prominent role that public health services, especially Medicaid, and other economic factors play in accessibility to abortions. Specifically, the coefficient for Medicaid for example, indicates that its availability contributes to an increase of as much as 1.89 abortions per 1,000 women. Remarkably, this variable by itself has a more drastic impact on abortion rates than does the the state of a county's legislative climate.

Code Appendix

```
## Libraries
library(tidyverse)
library(lme4)
library(lmtest)
library(GGally)
library(kableExtra)
```

```
## Data
rep_health <- read.csv("/Users/antonellabasso/Desktop/PHP2550/Data/reproductive_health.csv")
state_laws <- read.csv("/Users/antonellabasso/Desktop/PHP2550/Data/state_laws.csv")
abortion <- left_join(rep_health, state_laws, by="state")

abortion <- abortion %>%
  mutate(grad_pct=college_grad/totalpop,
         women_pct=women/totalpop)

head(abortion)
```

```
## Logistic Regression (Propensity Scores)

# (1) using logistic regression to estimate propensity of exposure
# given a set of demographic factors
# (2) weighting units according to the inverse of corresponding likelihoods
# to remove potential confounding

# restrictive climate = outcome <- treatment/exposure

# with:
# white_pct = reference group
# republican_2008 = reference group

PS_LR_2000 <- glm(is_highly_restrictive__2000 ~
                    women + grad_pct +
                    black_pct + native_american_pct +
                    asian_pct + hispanic_pct +
                    median_income + democrat_2008,
                  family=binomial(link="logit"), data=abortion)

PS_LR_2010 <- glm(is_highly_restrictive__2010 ~
                    women + grad_pct +
                    black_pct + native_american_pct +
                    asian_pct + hispanic_pct +
                    median_income + democrat_2008,
                  family=binomial(link="logit"), data=abortion)

PS_LR_2020 <- glm(is_highly_restrictive__2020 ~
                    women + grad_pct +
                    black_pct + native_american_pct +
                    asian_pct + hispanic_pct +
                    median_income + democrat_2008,
                  family=binomial(link="logit"), data=abortion)

# summary(PS_LR_2000)
```

```r
# summary(PS_LR_2010)
# summary(PS_LR_2020)
```

```r
# propensity score (PS) weights
abortion$ps_weight_2000 <- ifelse(abortion$is_highly_restrictive__2000,
                                  1/PS_LR_2000$fitted.values,
                                  1/(1-PS_LR_2000$fitted.values))

abortion$ps_weight_2010 <- ifelse(abortion$is_highly_restrictive__2010,
                                  1/PS_LR_2010$fitted.values,
                                  1/(1-PS_LR_2010$fitted.values))

abortion$ps_weight_2020 <- ifelse(abortion$is_highly_restrictive__2020,
                                  1/PS_LR_2020$fitted.values,
                                  1/(1-PS_LR_2020$fitted.values))

# NOTE: Due to study specifications, we proceed with PS weights for 2010 only.

#checking balance
mean(abortion$ps_weight_2010) # ~ 2
```

```r
## Regression (Study) Model Replication

# states used in the study
states <- c("arizona", "delaware", "georgia", "illinois", "indiana", "michigan",
  "new york", "north carolina", "ohio", "oklahoma", "oregon", "pennsylvania",
  "south carolina", "texas", "utah", "vermont", "washington", "wisconsin")
abortion_states <- abortion %>% filter(state %in% states)

# replicated model without fixed effects
study_model <- lm(abortion_rate_2010 ~
                    dist_to_closest_facility_miles +
                    is_highly_restrictive__2010,
                  data=abortion_states,
                  weights=abortion_states$ps_weight_2010)
summary(study_model)

# model coefficients & 95% CI
study_model_coefs <- as.data.frame(cbind(Est.=coef(study_model),
                                         confint(study_model, level=0.95)))
```

```r
## Model Extension

# identifying variables of interest (non-demographic)
interest_vars <- c("earnings_diff",
                   "pct_unemployed",
                   "pct_retirement_income",
                   "pct_public_assistance",
                   "mean_public_assistance_income",
                   "pct_food_stamps",
                   "pct_health_insurance_covered",
                   "pct_private_health_insurance",
                   "pct_public_health_insurance",
                   "pct_no_health_insurance",
```

```
                  "pct_poverty_prev_12",
                  "medicaid_cover")

# numeric interest variables
interest_vars_num <- names(
  abortion_states[, which(names(abortion_states) %in% interest_vars)] %>%
    select_if(is.numeric))

# first looking for correlations in variables of interest to avoid multicollinearity

# function to obtain all variables sig. correlated with specified variable
var_corrs_f <- function(data, var){
  corrs <- cor(data, use="complete.obs")
  high_corr <- corrs[abs(corrs[,var])>0.7, ]
  return(rownames(high_corr))
}

# list of variables sig. correlated with each variable of interest
corr_list <- list()
for (i in 1:length(interest_vars_num)){
  corr_list[[i]] <- var_corrs_f(
    abortion_states[, which(names(abortion_states) %in% interest_vars_num)],
    interest_vars_num[i])
}

# only variables 5-10 had strong correlations
#interest_vars_num[5:10]
#corr_list[5:10]
```

```
# getting all variables correlated with those having strong correlations
corr_vars <- unique(c(unique(rapply(corr_list, unique)),
                      interest_vars_num[5:10]))

# visualizing with ggpairs

# all correlated vars
ggpairs_plot1 <- ggpairs(
  abortion_states[, which(names(abortion_states) %in% corr_vars)],
  columnLabels=c("food stamps", "covered health ins.", "private health ins.",
                 "public health ins.", "no health ins.", "poverty prev."))

# removing no health insurance (inverse of covered health insurance)
ggpairs_plot2 <- ggpairs(
  abortion_states[, which(names(abortion_states) %in%
                            corr_vars[corr_vars!="pct_no_health_insurance"])],
  columnLabels=c("food stamps", "covered health ins.", "private health ins.",
                 "public health ins.", "poverty prev."))

# surprisingly, covered h.i. is strongly correlated with private health insurance
# and public and private are not
# so we could just keep private and public (rather than covered and public)
ggpairs_plot3 <- ggpairs(
  abortion_states[, which(names(abortion_states) %in%
                            corr_vars[!corr_vars
```

```
                                    %in% c("pct_no_health_insurance",
                                           "pct_health_insurance_covered")])],
  columnLabels=c("food stamps", "private health ins.",
                 "public health ins.", "poverty prev."))

# poverty and food stamps seem to be strong predictors of one another
# we must eliminate one in the model (whichever is least significant, if any)
ggpairs_plot4 <- ggpairs(
  abortion_states[, which(names(abortion_states) %in%
                            corr_vars[!corr_vars
                                    %in% c("pct_no_health_insurance",
                                           "pct_health_insurance_covered",
                                           "pct_poverty_prev_12")])],
  columnLabels=c("food stamps",
                 "private health ins.", "public health ins."))

# food stamps is also strongly correlated with public insurance, so we may find that we don't need both
```

```
# building extended model

# backward selection
# remove variables based on correlations discovered
# use AIC and LRT to compare nested models

# full model
study_model_ext_f <- lm(abortion_rate_2010 ~
                        dist_to_closest_facility_miles +
                        is_highly_restrictive__2010 +
                        earnings_diff +
                        pct_retirement_income +
                        pct_unemployed +
                        pct_public_assistance +
                        mean_public_assistance_income +
                        pct_food_stamps +
                       # pct_poverty_prev_12 +           # MC (food stamps), less sig.
                        pct_health_insurance_covered +
                        pct_private_health_insurance +
                        pct_public_health_insurance +
                       # pct_no_health_insurance +       # MC (covered hi), NA
                        medicaid_cover,
                      data=abortion_states,
                      weights=abortion_states$ps_weight_2010)

# using data from the full model for comparison in LRT
full_model_data <- study_model_ext_f$model

# first nested model
study_model_ext1 <- lm(abortion_rate_2010 ~
                        dist_to_closest_facility_miles +
                        is_highly_restrictive__2010 +
                        earnings_diff +
                        pct_retirement_income +
                        pct_unemployed +
                        #pct_public_assistance +            # non sig.
```

```
                        #mean_public_assistance_income +    # non sig.
                        #pct_food_stamps +                  # non sig.
                        #pct_poverty_prev_12 +
                        pct_health_insurance_covered +
                        pct_private_health_insurance +
                        pct_public_health_insurance +
                        #pct_no_health_insurance +
                        medicaid_cover,
                      data=full_model_data,
                      weights=full_model_data$ps_weight_2010)

# second nested model
study_model_ext2 <- lm(abortion_rate_2010 ~
                        dist_to_closest_facility_miles +
                        is_highly_restrictive__2010 +
                        earnings_diff +
                        #pct_retirement_income +        # less sig.
                        #pct_unemployed +               # less sig.
                        #pct_public_assistance +
                        #mean_public_assistance_income +
                        #pct_food_stamps +
                        #pct_poverty_prev_12 +
                        pct_private_health_insurance +
                        pct_health_insurance_covered +
                        pct_public_health_insurance +
                        #pct_no_health_insurance +
                        medicaid_cover,
                      data=full_model_data,
                      weights=full_model_data$ps_weight_2010)

# summary(study_model_ext_f)
# summary(study_model_ext1)
# summary(study_model_ext2)
```

```
# model comparisons

# adjusted R squared
summary(study_model_ext_f)$adj.r.squared # 0.3932461
summary(study_model_ext1)$adj.r.squared # highest (0.3948453)
summary(study_model_ext2)$adj.r.squared # lowest (0.384781)
# suggests study_model_ext1 is preferable

# AIC
AIC(study_model_ext_f) # highest (1737.76)
AIC(study_model_ext1) # lowest (1709.291)
AIC(study_model_ext2) # 1713.82
# suggests study_model_ext1 is preferable

# LRT
lrtest(study_model_ext_f, study_model_ext1) # reject null
lrtest(study_model_ext1, study_model_ext2) # reject null
# suggests study_model_ext_f is preferable

# thus, we select study_model_ext1
```

```r
# predicted

# model coefficients & 95% CI
study_model_ext_coefs <- as.data.frame(cbind(Est.=coef(study_model_ext1),
                                             confint(study_model_ext1, level=0.95)))
```