THE DANGER OF DATA SNOOPING IN SCIENTIFIC INFERENCE AND WHAT TO DO ABOUT IT

1 Introduction

2 Variable Selection

Variable selection as a form of model fitting is performed by finding which subset of the k input variables is required to model the dependent variable in the best and most succinct manner (Hayter, 2013). Since a simpler model is more explainable and requires fewer observations, researchers are motivated to perform variable selection the find the variables that most explain the underlying scientific phenomenon. Increasingly, the dataset might contain more variables than observations thus variable selection is necessary for inference. Textbooks would recommend variable selection procedures such as the following,

> *First, an experimenter starts by fitting all k input variables. If all variables are needed in the model, then no reduction is necessary. If one or more input variables has a p-value larger than 10%, the variable with the largest p-value (smallest absolute value of the t-statistic) is removed. The reduced model with k -1 input variables is then fitted, and the process is repeated.*(Hayter, 2013)

Typically, analysts proceed with their statistical analysis on the full dataset with the final selected model as if selection never happened. However, the selection procedure invalidates the inference because the variables with the most association to the dependent variable on this specific data are selected. By chance, variables with no true association to the dependent variable could have a high correlation on this specific dataset thus the subsequent inference (such as hypothesis testing) is likely to have type I error rate higher than the specified α.

*add examples of other types of variable selection and what the researchers did - more stepwise regression (forward and backward) - univariate p value based - Lasso + OLS - PCA (no paper yet)

Next, we will show empirically through a case study where the use forwards stepwise selection as detailed in the paper violates type I error controls.

2.1 Case Study: Forward Stepwise Selection In Psychology

Study: Frontiers in Psychology published the paper titled: *The Big Three Health Behaviors and Mental Health and Well-Being Among Young Adults: A Cross-Sectional Investigation of Sleep, Exercise, and Diet* claiming that getting good quality sleep, exercising, and eating more raw fruits and vegetables predicts better mental health and well- being in young adults (Wickham, Amarasekara, Bartonicek, & Conner, 2020). It was subsequently summarized and published in Science Daily on Dec. 16th, 2020 (ScienceDaily, 2020). In this section, we will demonstrate how the selection procedure used by the researchers could violate statistical inference. We first permute the data to break any association between the covariates and the outcomes, and then replicate the procedure to our best abilities and then show that more than α proportion of variables are significant under the global null.

Data: In a cross-sectional survey (Conner, 2021), 1,111 young adults (28.4% men) ages 18–25 from New Zealand and the United States answered an online survey measuring the health predictors of interests: typical sleep quantity and quality; physical activity; and consumption of raw and processed fruit and vegetables, fast food, sweets, and soda; Covariates: including demographics, socioeconomic status, body mass index, alcohol use, smoking, and health conditions; and the outcomes: measures of depressive symptoms [measured by the Center for Epidemiological Depression Scale (CES-D)] and well-being (measured by the Flourishing Scale) (Wickham et al., 2020).

Analysis procedure of the paper The procedure can be summarized into the following three steps. First all the covariates that correlated with either the predictors and/or the outcome measures are selected. Model 1 includes all the

Figure 1: The proportion of times with at least one false discovery (out of 200 trials) is much higher on the training set.

predictors of interest and the selected covariates. Second, quadratic factors for the sleep, activity, and diet variables were included to test for any non-linear associations with the outcomes and were retained only when significant. Model 2 adds the significant quadratic health predictors to model 1. Third, model 3 added the significant two-way interaction terms among the health behaviors to model 2.

Method: We first permuted the design matrix row-wise while keeping the outcome variable unchanged, thus breaking any correlation between the explanatory variables and the outcomes of interest (welling-being and depressive symptoms). Then we followed the data analysis process outlined in the paper. For model 1, we included all the continuous demographic covariates correlate with at least one predictor of interest or the outcome variable of interest and all the categorical covariates (since it is unclear how the authors defined correlation between categorical and continous variables). The selected covariates are used for Model 1 in the paper. For model 2, it is unclear if all quadratic terms were thrown in all at once or through a stepwise procedure. We will use a forward stepwise regression (that corresponds to $\alpha = 0.05$) that starts with all the demographic covariates and the health behaviors (sleep_quantity, sleep_quality, activity, raw_fruit_veggie, cooked_fruit_veggie, fastfood_daily, sweets_daily, soda_daily) and end with the starting model plus all the quadratic health behavior terms. Finally, we repeated the forward step-wise selection procedure to select the two-way interaction terms for model 3.

Results: Following the model selection process described in the research, we discovered many "significant" predictors in the shuffled dataset where no association should be found (Table 1). For depressive symptoms, *sleep quality squared* and *sleep quality* $\times$ *sweets daily* are were selected and are significant at the 0.05 level. For predicting well-being, *sleep quality* $\times$ *sweets daily* and *sleep quality* $\times$ *sweets daily* are significant at the 0.05 level. Additionally, the p-values for significance were not adjusted for multiple testing which also contributes to the non-reproducibility of the results.

Next, we present in Figure 1 the probability of having at least 1 significant quadratic term under the global null. Here we performed stepwise selection for the quadratic terms on the bootstrap samples of the shuffled dataset. As expected, the false discovery proportion is much higher than the $\alpha = 0.05$ threshold. Note that even as the sample size increases, the false discovery rate remained high.

Table 1: OLS results of the final model (model 3) on the shuffled dataset

	Dependent variable:	
	Depressive Symptoms	Flourishing
	(1)	(2)
Constant	−1.900 (2.116)	−0.113 (0.195)
age	0.276 (0.247)	−0.046** (0.023)
GenderDiverse	5.452* (3.109)	−0.264 (0.290)
Male	−0.283 (0.938)	0.023 (0.087)
Asian	2.768* (1.526)	−0.265* (0.142)
Black	−0.206 (1.812)	−0.281* (0.169)
Hispanic	1.642 (2.032)	0.130 (0.190)
Other	1.216 (1.073)	−0.164 (0.100)
sample = MTurk	−0.405 (1.379)	0.112 (0.129)
unemployed	0.185 (1.391)	−0.079 (0.130)
ses	0.580* (0.336)	−0.043 (0.031)
bmi	−0.060 (0.068)	0.001 (0.006)
Has no health conditions	0.494 (0.920)	−0.052 (0.086)
Anti-depressant	1.084 (1.034)	−0.063 (0.096)
Supplement	−0.907* (0.496)	0.041 (0.046)
NoFoodAllergy	−0.254 (1.097)	0.069 (0.102)
Eats some meats	1.570 (1.597)	0.019 (0.148)
alcohol_daily	0.694* (0.354)	−0.052 (0.033)
regsmoke1	−0.035 (1.449)	0.174 (0.135)
sleep_quantity	−0.206 (0.303)	0.062** (0.029)
sleep_quality	−0.736 (0.450)	0.025 (0.041)
activity	0.552** (0.224)	−0.022 (0.021)
raw_fv	−0.531** (0.263)	0.016 (0.025)
cooked_fv	0.572 (0.429)	−0.049 (0.042)
fastfood_daily	1.343 (1.696)	−0.149 (0.158)
sweets_daily	−0.211 (0.479)	0.011 (0.045)
soda_daily	0.036 (0.424)	0.002 (0.040)
sleep_quality squared	0.769** (0.328)	
fastfood_daily squared	−1.047 (0.672)	0.087 (0.063)
activity×cooked_fv	−0.339* (0.188)	
raw_fv×cooked_fv		0.026* (0.015)
fastfood_daily×sweets_daily	−0.976** (0.470)	0.097** (0.044)
sleep_quantity×sweets_daily	−0.509* (0.282)	0.058** (0.026)
sleep_quantity×raw_fv	0.213 (0.148)	
sleep_quantity×activity		−0.021* (0.013)
sleep_quantity×soda_daily		−0.035* (0.021)
sleep_quality×raw_fv		−0.029 (0.020)
Observations	1,111	1,111
R^2	0.052	0.042
Adjusted R^2	0.024	0.013
Residual Std. Error	12.841 (df = 1078)	1.195 (df = 1077)
F Statistic	1.849*** (df = 32; 1078)	1.445* (df = 33; 1077)
Note:		*p<0.1; **p<0.05; ***p<0.01

3 Variable Transformation

Variable transformation refers to the practice of applying a transformation function to the predictors or the outcome variable that changes their scale of measurement. The most frequent purpose of transformations is to achieve a mean function that is linear in the transformed scale. (Weisberg, 2013) This practice is especially common when the covariates or response variables are too spread out such as median income, commonly encountered in economics. Many also argued the importance of non-linear transformation in medical sciences to improve the fit and interpretability of results. Some nonlinear monotonic transformations target improving the symmetry of the variable, and the advantages of such transformation includes variance, stability, linearity (Shachar et al., 2018). Discretizing or dichotomizing continuous covariates is another common transformation in the medical field to since it makes simpler risk assessment or eligibility cut offs (Williams et al., 2006). For example the outcome risk measure can be categorized into high or low instead of on a continuous scale. Variable transformation is necessary in many applications for both statistical and practical reasons. However, p value based selection process of the "right" transformations can lead to biased inference due to the multiple comparison problem. We will describe two minimal-p based selection procedures from published studies and show how the selection procedure invalidates statistical inference.

3.1 Minimal P procedure for selecting power transformation: Boston dataset

Finding the right transformation of covariates or outcome variable that fit the data better is a seemingly innocuous and often necessary practice. We replicate the data analysis procedure to our best abilities as detailed in the paper *Hedonic housing prices and the demand for clean air* to show how such procedure invalidate inference. We operationalize a model with "better fit" as one with a higher R^2 score.

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}},$$

indicating the percentage of the variance in the dependent variable that the independent variables explain collectively. Selecting the λ (power of transformation) that maximizes R^2 is the same as selecting the λ that gives the lowest p-value in an F-test given the number of covariates is the same (which is the case in variable transformation). With simple algebra we can show that

$$\mathcal{F}\text{-statistics} = -\frac{n-p}{p-1}\frac{R^2}{R^2-1},$$

where n is the total number of observations and p is the number of parameters. The $\mathcal{F}$-statistics is a monotone function of R^2 on its defined domain $[0, 1]$. Thus a minimal-R^2 is equivalent to a minimal-p procedure. We demonstrate that the false discovery rate is inflated through the process of transformation selection when there are no association between the covariates and the outcome.

Study:

Data:

Analysis procedure A two stage transformation process took place. First, the transformation on the outcome variable *Medv* is selected. Since the median value of houses typically has high range, log scale might be more appropriate. However whether or not to log transform the dependent variable based is based on the fit of the two models (Harrison & Rubinfeld, 1978).

> "One of the major objectives in estimating the hedonic housing equation was to determine the best fitting functional form. Comparing models with either median value of owner-occupied homes (MV) or Log(MV) as the dependent variable, we found that the semi-log version provided a slightly better fit."

Once the version of the outcome variable was selected, the paper conducted a grid search over the "best" power transformation on the variable *NOX* while keeping the selected transformation of the outcome variable constant. NOX^p is then used as a covariate instead of NOX in the regression.

> "we concentrated on estimating a nonlinear term in *NOX*; i.e., we included NOX^p in the equation, where *p* is an unknown parameter. The statistical fit in the equation was best when *p* was set equal to 2. The exponent was estimated by performing a grid search over alternative parameter values for p in the term $\frac{NOX^{(p-1)}}{p-1}$"

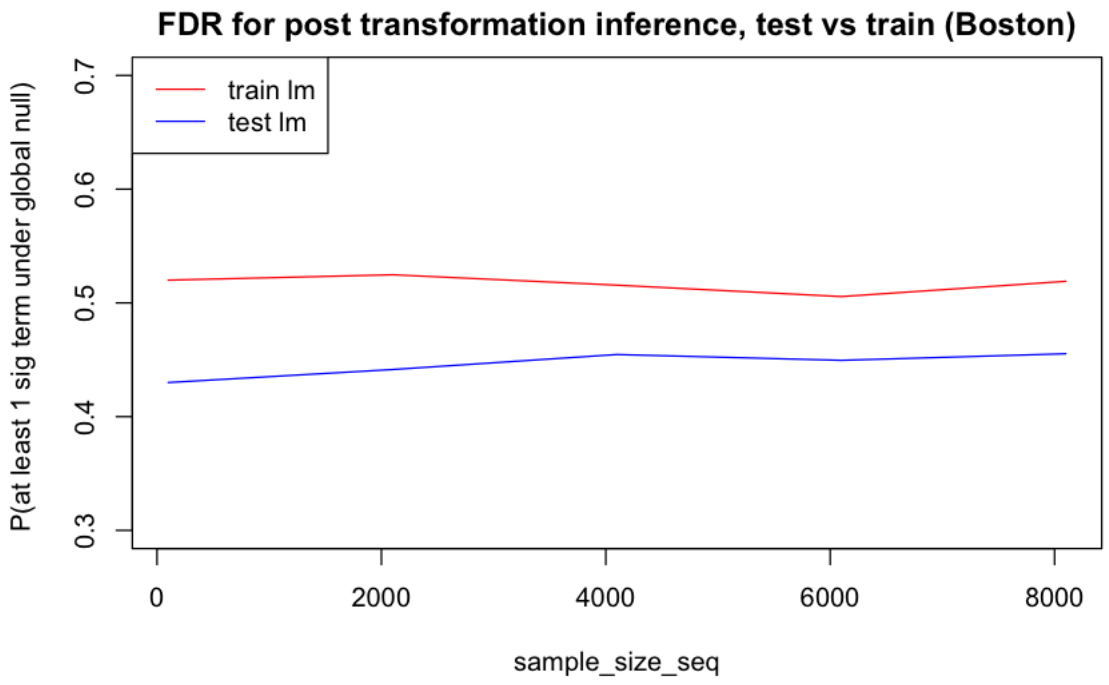

Figure 2: Selection leads to higher false discovery rate. The effect is constant over increasing sample sizes.

Method: We first permuted the Boston data set (n = 506) so that there is no association between the outcome variable Median Housing Value (MV) and the covariates: "age", "rmsq", "ldis", "lrad", "tax", "ptratio", "black", "lstat", "crim", "zn", "indus", "chas", "nox", "medv". We splitted the data into training and testing sets. No selection was done on the testing set. In the training set, selected the original outcome variable or the log transformed version based on which minimized the residual sum of squares regressing on the covariates. With the selected outcome variable, we do a grid search on the power p ($p \in \{1, 1/2, 0, 1/3, 1/2, 1\}$) such that $\Psi(NOX, p)$ minimizes the RSS of the selected outcome variable regressing on $\Psi(NOX, p)$ and other covariates. NOX^p is included in the final model. Since the paper did not detail the exact range of λ searched over, we will default to the recommendation from the textbook (Weisberg, 2013).

Result: We computed the proportion of times that at least 1 variable is significant (at $\alpha = 0.05$) under the global null for the training (data driven selection of the transformations on outcome variable $Medv$ and NOX) and testing (no transformation) set. As the figure shows that the selection procedure inflates the false discovery rate by 0.05 to 0.1 given the range of λ we searched over. The FDR inflation induced by the selection process will depend on the number of comparisons used in the grid search. Note that the problem with selection did not improve as the sample size increases.

4 Potential danger with consistent model selection procedures

Some argue that if a consistent variable/transformation selection procedure is used, there is no violation of inference validity asymptotically, the same variables will be selected and the false discovery rate will be controlled at the desired α level, *for any given fixed distribution of parameters*. However, it has been debunked that consistent variable selection has no effect on subsequent inference asymptotically(Leeb & Pötscher, 2005). In fact, the consistent selection procedure does not hold uniformly with respect to *all distributions of parameters*. A simple variable selection example demonstrating such result has been provided (Leeb & Pötscher, 2005). In this section, we will show that consistent variable *transformation* procedures can fail to give asymptotic guarantees with uniformity.

A consistent procedure: The textbook *Applied regression analysis* recommends transforming a single predictor or the outcome using the *scaled power transformation* defined as:

$$\Psi(X, \lambda) = \begin{cases} (X^\lambda - 1)/\lambda, \\ \log(X), \text{if } \lambda = 0 \end{cases}$$

, and selecting λ to minimize $RSS(\lambda)$ from a discrete set of real values (Weisberg, 2013). The selected power is denoted as $\hat{\lambda}$. It can be shown that minimizing the residual sum of squares is equivalent to minimizing negative log likelihood, implying that $\hat{\lambda}$ is the MLE estimator of λ^*. By the consistency of MLE procedures under certain regularity conditions, the transformation procedure is consistent.

A simple illustrating example: Let $\lambda^0 \in (0, 1)$;
$x_{1i}, x_{2i} \sim Unif(0, 1)$, and are independent of each other;
$\epsilon_i \sim Normal(0, 1)$

$$y_i = \frac{1}{2}x_{1i} + \frac{1}{2}\Psi(x_{2i}, \lambda^0) + \epsilon_i$$

We observed a total of n independent and identically distributed samples $y = \{y_1, ..., y_n\}$ and the covariates x_1, x_2. We want to find the appropriate transformation for x_2.

$$\hat{\lambda} = \underset{\lambda \in \{0,1\}}{\operatorname{argmin}} RSS(\hat{y}(\lambda)),$$

where $\hat{y(\lambda)} = \hat{\alpha}x_1 + \hat{\beta}\psi(x_2, \lambda)$, is the OLS result. Since the selection procedure is consistent, $\hat{\lambda}(n) \xrightarrow{p} \lambda^*, \lambda^* \in \{0, 1\}$. Note that $\lambda^* \neq \lambda^0$ regardless of the sample size, thus the asymptotic results refers to the convergence to λ^* only. For any fixed λ^0, we get that: $\hat{\beta}, \hat{\alpha} \xrightarrow{p} \beta^*, \alpha^*$ as $n \to \infty$. It seems as if the transformation procedure has made no impact on the parameters of interests. Indeed if we had applied the same methodology of breaking all associations in the data and then employ this transformation procedure, we would see no inflation of the false discovery rates on $\hat{\beta}, \hat{\alpha}$.

What is the problem: The devil lies in the fact the rate of convergence to λ^* is dependent on λ^0. If λ^0 is very close to 0 or 1, the problem is a lot easier (thus requiring less samples) than if it's somewhere in the middle. However we only observe a finite number of samples while not knowing what λ^0 is ahead of the time. We formalize this situation by introducing dependency on the sample size to the λ used in data generation: $\lambda^0(n) \to \lambda^0$ as $n \to \infty$. We show experimentally $P(\hat{\lambda}(n) = \lambda^*)$ can be quite low even when n is large.

Experiment set-up: Let $\lambda^0 \in [0.367, 0.467]$, $\lambda^0(n) = \lambda^0 + \frac{1}{3n}$, $\lambda^* = 1, \forall \lambda^0 \in [0.367, 0.467]$. We show that the proportion of times of the right power is selected out of 100 trials at different fixed sample sizes.

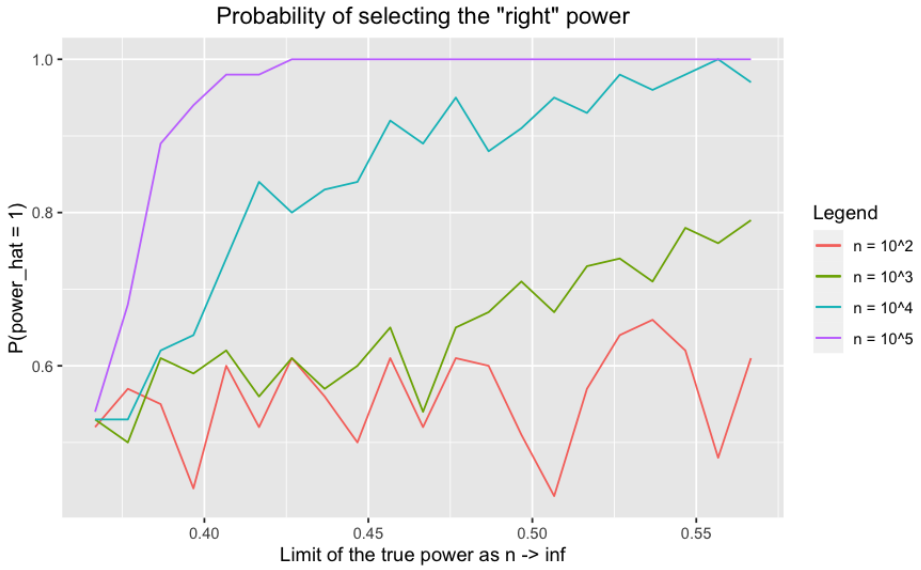

Figure 3: The probability of selecting the correct power transformation increases as the true power increases (the problem gets easier) and as the fixed sample size increases.

As shown in Figure 3, the probability of selecting the correct power transformation increases as the true power increases (the problem gets easier) or as the fixed sample size increases. When the problem is the most difficult ($\lambda^0 = 0.367$),

there is more than 40% chance that $\hat{\lambda} \neq \lambda^*$ even when $n = 100,000$.

Even if we do not care about estimating λ^0, the consistent selection procedure can invalidate assumptions on $\hat{\alpha}$.
Let $\hat{\alpha} := \hat{\alpha}(\hat{\lambda})$, $\alpha^* := \alpha(\lambda^*)$, $\alpha^0 := \alpha(\lambda^0) = \frac{1}{2}$.
We note that $\sqrt{n}(\hat{\alpha} - \alpha^*)$ should follow a normal distribution centered around 0 (α^* is close enough to $\alpha^0 = \frac{1}{2}$, thus $\frac{1}{2}$ is used in the plots below). However, if $\hat{\lambda} \neq \lambda^*$ then $\hat{\alpha} \to \alpha(\hat{\lambda}) \neq \alpha^*$. As we have shown in Figure 3, with a large sample size (n = 10,000) the wrong power can still be selected quite often and leading to a bimodal distribution. Figure 4 shows that with n = 10,000, $\sqrt{n}(\hat{\alpha} - \frac{1}{2})$ has a bimodal distribution when λ^0 is difficult to estimate and approaches a normal distribution centered around 0 when λ^0 get closer to 1.

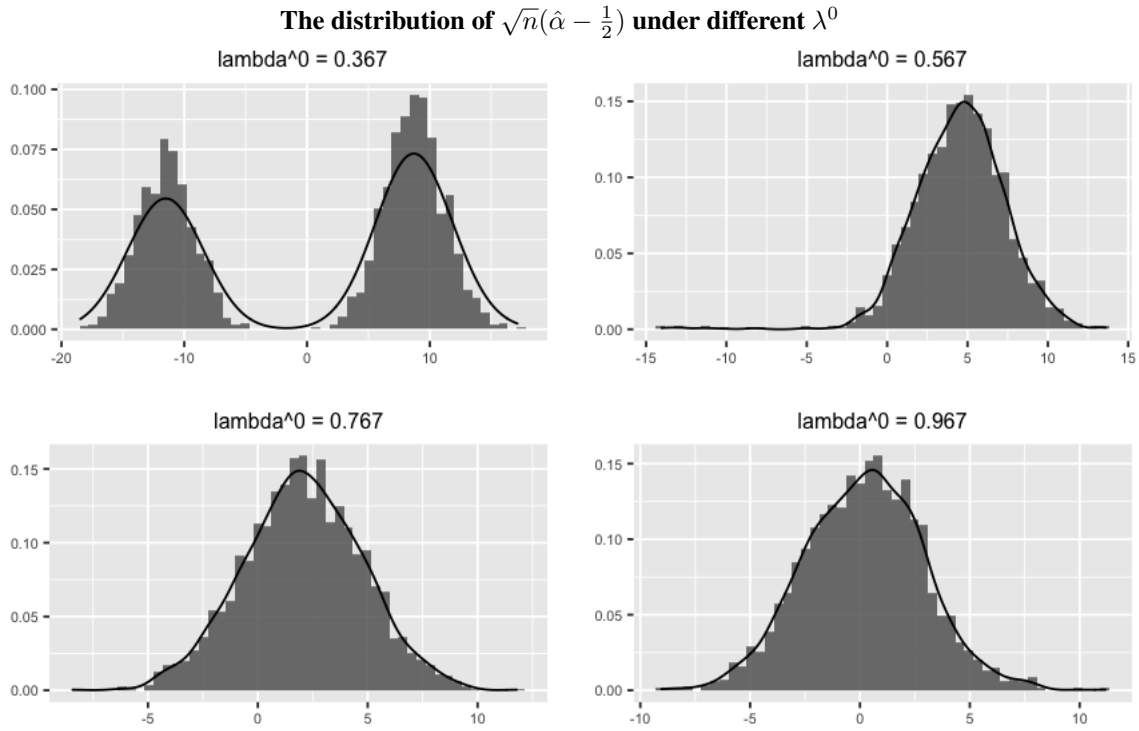

Figure 4: n = 10,000, number of trials = 2000

5 Remedies and Recommendations

There are in general three types of remedies for correcting for post-selection inference: sample splitting, simultaneous inference, and conditional selective inference. In the following section we explain their advantages and disadvantages and gives recommendations for valid inference after selection. where does data blurring/differential privacy approach to POSI fit in?

Sample Splitting: The most straight-forward and easy to implement method accounting for model selection. The idea is to randomly split the data into the train and the test set where the all the model selection procedures are done on the training set. Since the data is split randomly, independent of the data generating process, inference performed on the test set is valid for any selection procedure. Generality is the major *advantage* of this method since it makes no assumption on the type or sequence of selection procedures used. In applied data analysis, analysts often apply multiple selection procedures sequentially. For example, in the psychology case study (Wickham et al., 2020), predictors were first selected base on their correlations with the outcome variables, then the quadratic terms, lastly, the interactions of the selected predictors. Provided that the selection procedures were done only on the train split, sample splitting does not require any pre-commitment to a set of procedures. On the other hand, the *disadvantage* of sample splitting is that it does not work on dependent data such as networks or time series data. references on methods that does sample splitting for dependent data? Sample splitting also introduces randomness and reduces the effective sample size that can make

it less appealing to practitioners. Detailed procedure on sample splitting with bootstrapped confidence interval is in section 2.4.3 (Zhang, 2012).

simultaneous inference Q: the algo provided in the annual review paper relies on bootstrap, which has versions that works for dependent data but not for sample splitting? Q: - block bootstrap for time series and patch work bootstrap for networks exists but it doesn't guarantee crossvalidation works in those cases? - in what situations does that the uniform asymptotic linear representation assumption not hold Go through the stepwise example using the PoSI package (Buja & Zhang, 2020)does this work with dependent data?. Add references on replacing the bootstrap step. what does increasingly dense grid of Q mean in the Annual paper? a pakage that's still IP?: https://github.com/post-selection-inference/R More efficient algo (Kuchibhotla et al., 2020)(not sure if there is a package for it)

5.1 conditional selective inference

5.2 Other methods

5.3 Common pit-falls

To the researchers' credit, they are aware of the reproducibility crisis in psychology and write,

> "We also used 10-fold cross-validation to determine whether any interaction terms would be useful for predicting out-of-sample, above and beyond the no-interaction model. Cross-validation involves splitting data into several subsets or "folds" and then repeatedly fitting the model to all but one fold and testing the model on the leftover fold" (Wickham et al., 2020)

Through their cross-validation procedure, the researchers concluded that none of the interaction terms they included in the model helps with out-of-sample prediction. However, they did not subject the chosen quadratic terms to the same procedure. Moreover, the cross-validation was done on the same dataset that the 2-way interaction terms were chosen thus the estimated error from cross-validation is a biased estimate of the true test error. Lastly, statistical inference is a different problem than out-of-sample prediction.

5.4 Note on Robust vs Conventional variance on FDR control

Robust standard error (sandwich variance) is often recommended to practitioners as an assumption-lean alternative to the conventional standard error. However there are situations when statistical tests carried out using the robust standard error yields inflated false discovery rate and is less conservative then that of the conventional standard error estimates. For example, after we broke the association between the outcome variable and the covariates, the sandwich variance yields higher false discovery rate. This phenomenon is studied in the following simulation study, where we sample x covariates follow a log normal distribution, inference based on robust standard error lead to higher false discovery rate than that of the conventional standard error. When the linear model is correct, the robust variance can have higher variability thus leading to under coverage (Kauermann & Carroll, 2001). Experimentally, we showed that when $x_i \sim$ log normal distribution, the tests constructed based on robust error lead to higher FDR under the global null (the linear assumption is met). When x_i is distributed normally, the results from robust error agrees with that of the conventional error. model as approximation I section 13: when x is heavy tailed, robust variance under cover, show the graph when x normal and x log normal and comment on when should you use robust vs conventional and choose max of both is the conservative option

6 Conclusion

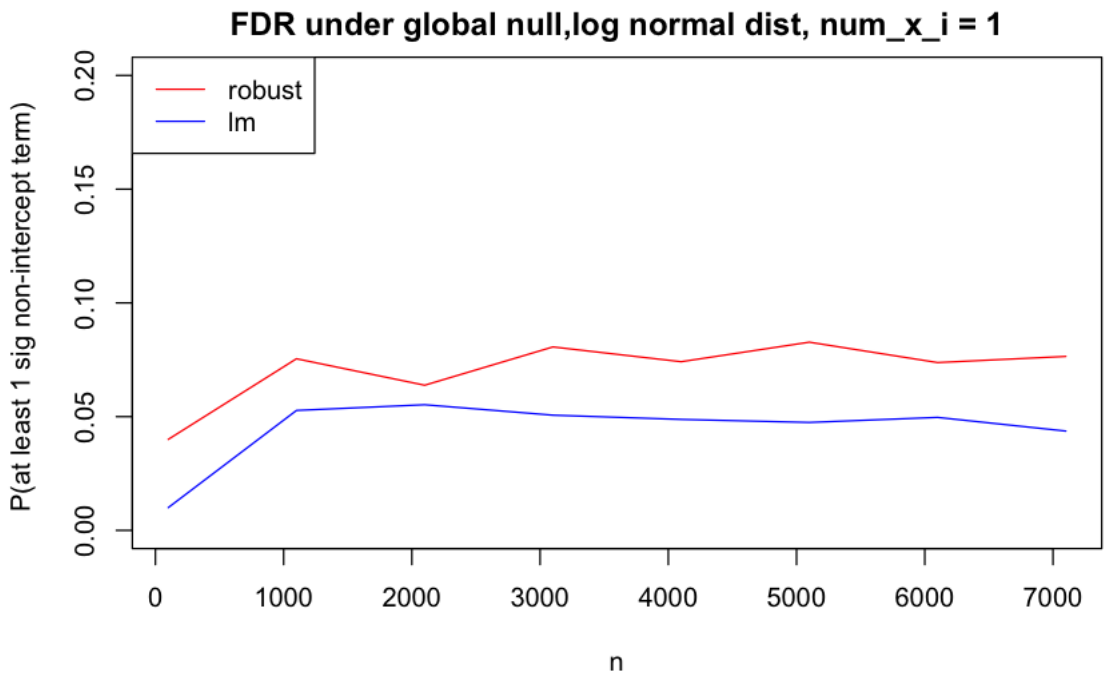

Figure 5: As sample size increases, the robust error still lead to higher FDR under the global null.

References

Buja, A., & Zhang, K. (2020). Posi: Valid post-selection inference for linear ls regression [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=PoSI` (R package version 1.1)

Conner, T. (2021, 1). *The lifestyle of young adults survey.*

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81-102. Retrieved from `https://www.sciencedirect.com/science/article/pii/0095069678900062` doi: https://doi.org/10.1016/0095-0696(78)90006-2

Hayter, A. J. (2013). *Probability and statistics for engineers and scientists*. Brooks/Cole, Cengage Learning.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396. Retrieved from `http://www.jstor.org/stable/3085907`

Kuchibhotla, A. K., Brown, L. D., Buja, A., Cai, J., George, E. I., & Zhao, L. H. (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics*, *48*(5), 2953–2981.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59.

ScienceDaily. (2020, 12). *Three pillars of mental health: Good sleep, exercise, raw fruits and veggies.*

Shachar, N., Mitelpunkt, A., Kozlovski, T., Galili, T., Frostig, T., Brill, B., … Benjamini, Y. (2018, 5). The importance of nonlinear transformations use in medical data analysis. *JMIR Medical Informatics*, *6*. Retrieved from `/pmc/articles/PMC5970282//pmc/articles/PMC5970282/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5970282/` doi: 10.2196/MEDINFORM.7992

Weisberg, S. (2013). Applied linear regression, 4th edition. , 368.

Wickham, S.-R., Amarasekara, N. A., Bartonicek, A., & Conner, T. S. (2020, 12). The big three health behaviors and mental health and well-being among young adults: A cross-sectional investigation of sleep, exercise, and diet. *Frontiers in Psychology*, *11*. doi: 10.3389/fpsyg.2020.579205

Williams, B. A., Mandrekar, J. N., Mandrekar, S. J., Cha, S. S., Furth, A. F., Williams, B., … Furth, M. F. A. (2006). Finding optimal cutpoints for continuous covariates with binary and time-to-event outcomes.

Zhang, K. (2012). *Valid post-selection inference*. University of Pennsylvania.

- what's the difference between the problem with consistency and the problem of finite sample estimation?
- is the dependency on n necessary?