# Unified Framework for Post-selection Inference

Arun Kumar Kuchibhotla

# Abstract

The development of the classical theory of mathematical statistics is based on the philosophy that all the variables used for modeling is fixed prior to the collection of data. As soon as the number of variables available becomes larger than 15, it is desirable in practice to select (based on the data) a parsimonious set of variables for practical interpretation or scientific explanation. Once such a variable selection is performed based on the data, the classical theory is (bluntly speaking) useless for inference (that is, confidence regions or testing hypothesis) and in fact very misleading. Another example along the same lines involves selecting an optimal tuning parameter using a cross-validation-type criterion in lasso or smoothing splines. In this example, the final estimator is based on the data-driven tuning parameter but somehow most of the theory is based on fixed (non-random) tuning parameter.

In both the examples mentioned above, the final estimator is obtained after a data-driven selection (either variables or some tuning parameter). In this thesis, these problems are called post-selection problems and the main focus of this thesis is to provide a unified framework for a valid post-selection inference, that is, valid inference based on the estimators obtained post-selection. The primary focus of the thesis is on the first type of selection: variable/model selection in regression problems. The unified framework does allow for the second type of selection too.

Valid post-selection inference has been a topic of research interest at least since 1960's but has received a increasing attention in the recent times. Invalidity of classical inference post-selection problems may not only be due to the selection but also due to misspecification of model. Misspecification is a very natural outcome of model selection since the selected model cannot always be guaranteed to match the truth. If such a guarantee exists, then the post-selection problem does not require further study. Most of the literature on valid post-selection inference has concentrated on the assumption of a true parametric model.

In this thesis, valid post-selection inference is provided under no parametric assumptions. The simplest setting in this thesis is when the observations are independent satisfying certain moment restrictions (and no further model/distributional assumptions). Extensions to various dependent settings are also given. Throughout, the total number of variables available is allowed to grow with the sample size and can be almost exponential in the sample size.

# Contents

**1 Motivating Examples** **3**
1.1 Practice of Statistics in Textbooks/Education . . . . . 4
1.1.1 Case Study TE1 . . . . . 5
1.1.2 Case Study TE2 . . . . . 5
1.1.3 Case Study TE3 . . . . . 6
1.2 Practice of Statistics in Literature . . . . . 7
1.2.1 Case Study L1 . . . . . 7
1.2.2 Case Study L2 . . . . . 7
1.2.3 Case Study L3 . . . . . 8
1.2.4 Case Study L4 . . . . . 9
1.2.5 Case Study L5 . . . . . 10
1.3 Some Observations and Formulation of the Problem . . . . . 10

**Bibliography** **13**

# Chapter 1

# Motivating Examples

In 2005, the Stanford epidemiologist Ioannidis made a dramatic claim: "most published research findings are false." The claim is largely believed to be true. It gave rise to the term "reproducibility/replicability crisis." Several factors have been identified as possible causes, first among them publication bias, that is, the fact that null findings tend not to get published. This is an institutional problem whose solution is the reform of publication policies. Another contributing factor, closer to home for us statisticians, is the breakdown of the classical statistical inference framework under the current practice of statistics. In the modern practice of statistics, data analysts tend to use many forms of data exploration before applying statistical inference, and this is a problem that requires our serious attention. This aspect of the replicability crisis will be the backdrop of the thesis. The main goal of the thesis is to provide a unified framework for resolving the problem of **V**alid **I**nference after **D**ata **E**xploration (VIDE).

Classical statistical inference framework is built to provide valid statistical conclusions when the hypotheses to test and the model to fit are decided without the involvement of the data at hand. The practice of statistics does not follow this sequence, as will be shown in this chapter. This deviation from the classical inference framework can drastically invalidate the conclusions. For an illustration of the drastic invalidity when the hypothesis to test is chosen based on the data, consider the following example:

1. Generate 500 observations from $(Y_1, X_1), \ldots, (Y_n, X_n) \overset{iid}{\sim} N(0, I_{p+1})$, for some $p \geqslant 1$. In this distribution, $Y_i$'s are independent of $X_i$'s.

2. Select one covariate which is the most correlated with the response, that is,

$$\widehat{j} \;:=\; \underset{1 \leqslant j \leqslant p}{\arg\max} \left| \widehat{\mathrm{corr}}(Y, X_j) \right|.$$

   Here $\widehat{\mathrm{corr}}$ is computed based on the 500 observations.

3. Compute the least squares estimator

$$(\widehat{\alpha}_{\widehat{j}},\, \widehat{\beta}_{\widehat{j}}) \;:=\; \underset{(\theta_1, \theta_2)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta_1 - \theta_2 X_{i,j})^2,$$

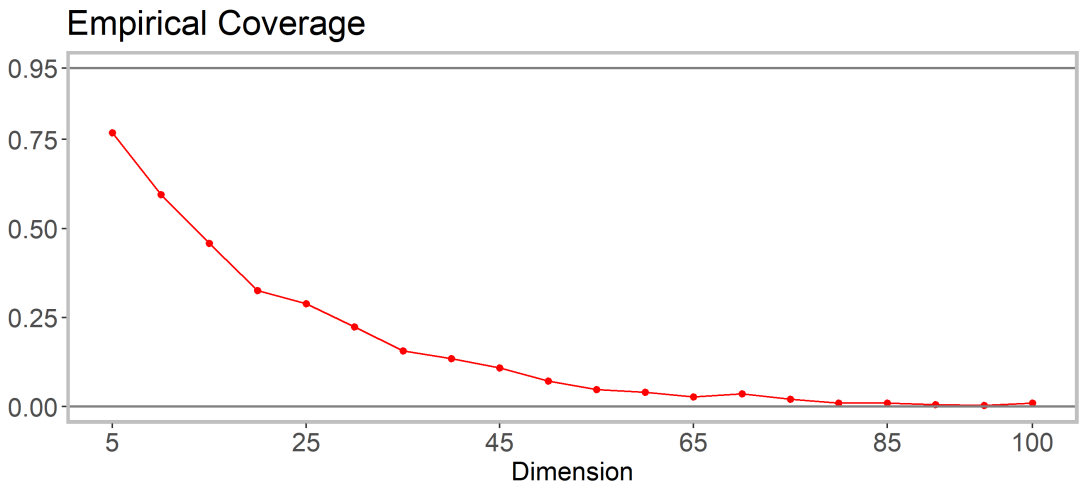


Figure 1.1: Attained level of a classical test when the hypothesis is selected after data exploration.

where $X_{i,j}$ represents the $j$-th coordinate of $X_i$.

4. Test the hypothesis $H_{0,\widehat{j}}$ of insignificant coefficient based on the estimator $\widehat{\beta}_{\widehat{j}}$. The classical test of level 0.05 in this case is

$$\text{Reject } H_{0,\widehat{j}} \text{ if } \left| \frac{n^{1/2}\widehat{\beta}_{\widehat{j}}}{\widehat{\sigma}_{\widehat{j}}} \right| \geqslant 1.96. \tag{1.1}$$

Here $\widehat{\sigma}_{\widehat{j}}/n^{1/2}$ is the classical estimator of the standard error of $\widehat{\beta}_{\widehat{j}}$ (disregarding the randomness of $\widehat{j}$). This test is same as looking at the summary of `lm(Y ~ X`$_{\widehat{j}}$`)` and taking the decision of reject if the $p$-value is less than 0.05.

Because the response is uncorrelated with all the covariates, one might naively expect that the test (1.1) controls Type I error at 0.05. Figure 1.1 shows the attained true level of the test (1.1). Need to plot 1−coverage. This example shows that the classical statistical procedures do not solve the VIDE problem and requires a non-trivial adjustment.

There are many ways of mathematically formalizing the VIDE problem. In the following sections, I will provide a few examples from textbooks and published research. Then the VIDE problem will be formalized mathematically at the end of this chapter and will be solved in the forthcoming chapters.

## 1.1 Practice of Statistics in Textbooks/Education

In this section, I present several examples from textbooks and educational journals where the full procedure described involves testing hypotheses obtained after data exploration.

### 1.1.1 Case Study TE1: Moore and McCabe (1998)

Example 11.1 of Moore and McCabe (1998) introduces the GPA dataset in the anticipation of predicting the college GPA of students based on the high school scores in math (HSM), science (HSS), and english (HSE). In the process of refining the basic linear model of GPA on HSM, HSS, HSE, the authors (page 724) write

> Because the variable HSS has the largest $P$-value of the three explanatory variables and therefore appears to contribute the least to our explanation of GPA, we rerun the regression using only HSM and HSE as explanatory variables. The $F$ statistic indicates that we reject the null hypothesis that the regression coefficients for the two explanatory variables are both zero. The $P$ value is still 0.0001.

This is similar to what was done in the illustrative example. The results of the first linear model suggested the next hypothesis to test and hence is obtained as a result of data exploration. This invalidates the classical F-test. A side point to note here is that the second linear model (with HSM and HSE) might be misspecified, that is, may not satisfy all the assumptions of the classical linear model.

### 1.1.2 Case Study TE2: Stine and Foster (2013)

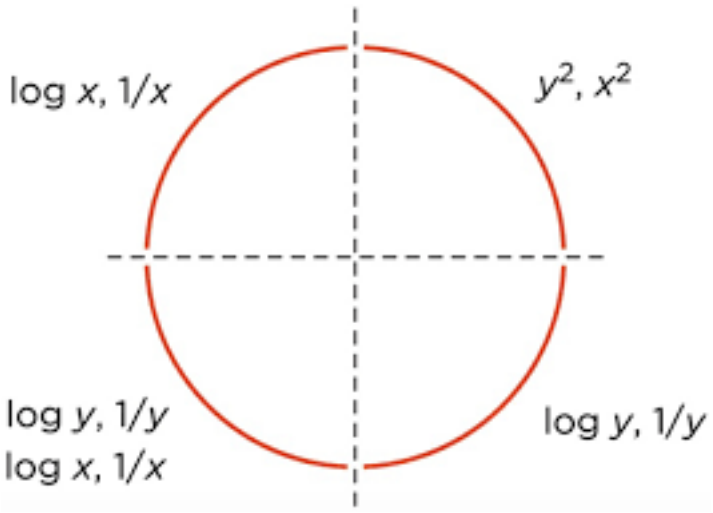


Figure 1.2: Tukey's Ladder of Transformation. The Four Quadrant Approach.

In the context of fitting a curve to bivariate data, Stine and Foster (2013, page 515) write

> Deciding on a transformation requires several skills. First, think about the context of the problem: why should the association by linear? Then, once you see curvature in the scatterplot, compare the curvature to the bending

> patterns shown in Figure 1.2. Among the choices offered, find the one that captures the curvature of the data and produces an interpretable equation. **Above all, don't be afraid to several.**[1] Picking a transformation requires practice, and you may need to try several to find one that is interpretable and captures the pattern in the data.

Unlike the suggestion of Moore and McCabe in Section 1.1.1, the suggestion here more dangerous in the sense that it (actively) advises the data analyst to make subjective decisions on what transformations to try and use in the final model. Because this advise is based on visualization, it is not possible to mathematically analyze the selection method. (The suggestion in 1.1.1 is backward elimination which is analytically precise and is possibly amenable to mathematical analysis.)

### 1.1.3 Case Study TE3: Pardoe (2008)

This paper from the Journal of Statistics Education is written to address the challanges of teaching complicated aspects of linear regression modeling using Oregon realtor data. The paper, however, spells out the details of how linear regression modeling is usually taught in basic courses and this way invalidates the classical inference so badly that it may not be possible to adjust for it.

The paper models the price of a home in terms of 12 features of the home including Bed and Bath. Section 3 of the paper fits, model 1, a linear regression for price on all 12 features. Then the author writes

> However, the residuals of model 1 fail to satisfy the zero mean (linearity) assumption in a plot residuals versus Age, displaying a relatively pronounced curved pattern. $\cdots$ To attempt to correct this failing, we will add an $\text{Age}^2$ transformation to the model, which as discussed above was also suggested from the realtor's experience. The finding that the residual plot with Age has a curved pattern does not necessarily mean that an $\text{Age}^2$ transformation will correct this problem, but it is certainly worth trying.

This might look similar to the example of transformations in Section 1.1.2 but note that the decision to try transformations is done based on the data. Further, the decision of trying the square transformation is adhoc and is not chosen from a starting family of transformations (as in Section 1.1.2). The modeling, in the paper, does not stop there and the paper proceeds

> In addition, both Bath and Bed have relatively large individual $t$-test $p$-values in model 1, which appears to contradict the notion that home proces should increase with the number of bedrooms and bathrooms. ... The instructor can guide the students in seeing that to model such a relationship we need to add a Bath$\times$Bed interaction term to the model.

[1] Emphasis added here.

This along with the square transformation of Age is called model 2. The author looks at the classical summary table as if no exploration has been done and writes

> However, the model includes some terms with large individual $t$-test $p$-values, suggesting that perhaps it is more complicated that it needs and ...,

and constrcuts a more refined model along with an interpretation.

This case study shows how fluid the modeling process is and how much it differs the classical mathematical framework of inference for linear regression. The classical framework requires fitting one model (decided apriori) and then infer.

## 1.2 Practice of Statistics in Literature

The previous section has shown various examples of how the practice of statistics differs from the mathematical inference framework in textbooks and education. To further illustrate the practice of data exploration, I will now present a few case studies from the literature.

### 1.2.1 Case Study L1: Harrison Jr and Rubinfeld (1978)

This is the paper that introduced the well-known Boston housing data[2]. Although forgotten in the subsequent use of this data, the data was collected to measure the willingness to pay for clean air. Boston is divided into census tracts and in each tract, the median (MV) of the property value of homes among those in the tract. The concentration of nitrogen oxide (NOX) is used as a proxy for clean air and the response/dependent variable is MV. The dataset includes 12 more confounders that are adjusted for in the linear models. In fitting a model for MV, the authors write (on page 86)

> 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 semilog version provided a slightly better fit. Using Log(MV) as the dependent variable, we concentrated on estimating a nonlinear term in NOX, i.e., we included $\text{NOX}^p$ in the equation, where $p$ is an unknown parameter. $\cdots$
>
> The statistic fit in the equation was best when $p$ was set equal to 2.0, i.e., when $\text{NOX}^2$ was in the equation. $\cdots$ The NOX variable has a negative sign and is highly significant.

The authors essentially explored the dataset to obtain the "right" transformations for the response and the covariate of interest. Further they ignore the exploration and report the statistical significance of coefficient without any adjustment.

[2]The data in this paper seems to be wrongly coded in a few places. See Gilley et al. (1996) for details.

### 1.2.2 Case Study L2: Whittingham et al. (2006)

The authors of this paper do not use classical inference after data exploration but show that this practice with stepwise form of data exploration is more prevelant in ecology and animal behavior. The authors describe the dangers of using stepwise selection methods for inference and surveys several papers from the literature to illustrate that this is common practice. On page 1184, the authors write

> A second problem with stepwise multiple regression is more widely recognized and yet appears not to have deterred many ecologists from using the technique. ⋯ In particular, it is easy to overlook the fact that a single stepwise regression does not represent one hypothesis test but, rather, involves a large number of tests. This inevitably inflates the probability of Type I errors (false positive results). ⋯ Finally, owing to the selection of variables to include on the basis of the observed data, the distribution of the F-statistic is also affected, invalidating tests of the overall statistical significance of the final model.

For a similar paper with practical recommendations, see Lydersen (2014). For a paper recommending variable selection without any indication of its consequences, see Chowdhury and Turin (2020).

### 1.2.3 Case Study L3: Wiens et al. (2015)

The authors of the paper develops predictive models to predict post-discharge mortality which is defined as a binary random variable taking value 1 if the child dies within six months of discharge. The statistical analysis is based on 1307 enrolled participants. In the statistical analysis section of the paper, the authors write[3]

> All variables were assessed using **univariate logistic regression to determine their level of association with the primary outcome**. Continuous variables were assessed for model fit using the Hosmer-Lemeshow test. Missing data was imputed by the method of multivariate imputation using chained equations. Following univariate analysis, candidate models were generated using a **stepwise selection procedure minimising Akaike's Information Criterion (AIC)**. This method is considered asymptotically equivalent to cross-validation and bootstrapping. All models generated in this sequence having **AIC values within 10% of the lowest value** were considered as reasonable candidates. The final selection of a model was judged on **model parsimony** (the simpler the better), **availability of the predictors** (with respect to minimal resources and cost), and the **attained sensitivity** (with at least 50% specificity).

[3] Emphasis added here.

It is easy to recognize that the the process described above is not reproducible (because of subjectivity and incomplete details) and hence all the inference for the models reported in Table 3 of the paper is invalid.

### 1.2.4 Case Study L4: Nersisyan et al. (2019)

The authors of this paper study the inheritance patterns of Telomere length. Telomore is the end cap of a chromosome (as shown in Figure 1.3) and reduces in length as the cell divides. Hence, the length of the telomere acts a biological age of a person. The first

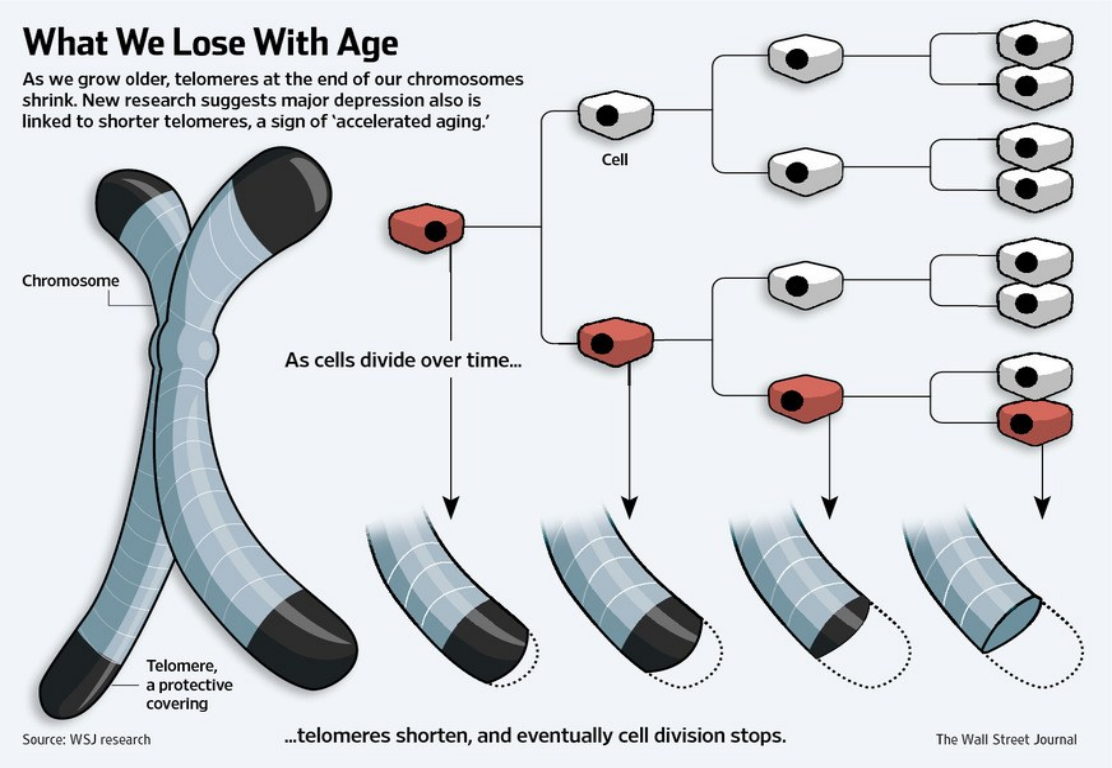


Figure 1.3: Telomere Length

problem the authors study is associating an offspring's mean telomere length (MTL) to the age and sex of the offspring along with the telomere length of the parents (mother's mMTL and father's fMTL) and the age of the parents at conception (mother's MAC and father's PAC). In developing the models, the regression analysis section of the methods in the paper mentions[4]

> Multivariate linear regression (MLR) analysis was performed to evaluate the correlation between MTL and age and sex in the studied population. ⋯ Two of the families with missing data for the mother were removed, and two

[4]Emphasis added here.

families with discordant age diferences at the time of data collection and at conception were also discarded. Overall, MLR were done on 246 families. A set of pairwise regressions on the predictors were performed to estimate dependence between variables, and interaction terms were introduced for correlated predictors. The MLR models were tested by **sequential introduction of predictors and interaction terms**. The best model was chosen based on maximization of the adjusted R square term: ultimately, from the **three best models with similar adjusted R squared values the simplest one was chosen**.

### 1.2.5 Case Study L5: Bolt et al. (2016)

The authors of this paper examine the variables that are significantly associated with communication in every day activities, or comminicative participation, in adult survivors of head and neck cancer (HNC). In the statistical analysis section (page 1148) of the paper, the authors write

> The associations of the 17 variables with communicative participation were examined with multiple linear regression analysis in SPSS, version 18.0 (IBM). Communicative participation, age, time since diagnosis, and self-reported cognitive function were continuous variables; all others were categorical variables. Throughout the process of backward stepwise regression, model fit was analyzed with an overall regression $F$ statistic. Individual variables with regression coefficients significant at $P < .05$ were retained in the model.

Because the final selected set of variables are obtained through data exploration, they are cannot be confirmed as significantly associated variables using the classical tests.

## 1.3 Some Observations and Formulation of the Problem

In all of the works reported above, the method of data analysis constitutes the following: Have a question of interest, get the dataset, explore the data to find a good model to fit or find the subset of covariates to be used in the model or find the transformations for variables to be used in the model, and then fit the model to draw inference or statistical conclusions. For example, in the context of fitting a linear regression with a treatment variable. The question of interest could be "is there a non-zero treatment effect?" In presence of confounders, one might select a subset of confounders to be used in the final model or one might select the transformation for the response/confounders. Then fit the model with selected set of confounders and transformations.

The illustrative example discussed in the beginning of the chapter (Figure 1.1) shows that in this practice classical test or confidence regions cannot be used for reliable conclusions. A reasonable mathematical formulation of the problem (in case of linear regression) could be as follows: Suppose we have observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$.

1. For each $\mathrm{M} \subseteq \{1, 2, \ldots, p\}$, define the "target"

$$\beta_{\mathrm{M}} \;:=\; \underset{\theta \in \mathbb{R}^{|\mathrm{M}|}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(Y_i - X_{i,\mathrm{M}}^{\top}\theta)^2\right].$$

2. Based on the data, select a subset $\widehat{\mathrm{M}} \subseteq \{1, 2, \ldots, p\}$ of covariates using whatever method of practitioner's choice. (This freedom in the selection method should be allowed to solve the problems in the practical scenarios elluded to above.)

3. Calculate the estimator

$$\widehat{\beta}_{\widehat{\mathrm{M}}} \;:=\; \underset{\theta \in \mathbb{R}^{|\widehat{\mathrm{M}}|}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_{i,\widehat{\mathrm{M}}}^{\top}\theta)^2.$$

   This estimator "targets" $\beta_{\widehat{\mathrm{M}}}$ (the evaluation of the map $\mathrm{M} \mapsto \beta_{\mathrm{M}}$ at $M = \widehat{\mathrm{M}}$). This fact will be shown in later chapters. In all of the case studies presented before, the practitioners use the estimator $\widehat{\beta}_{\widehat{\mathrm{M}}}$ for inference or statistical conclusions.

4. Because $\widehat{\beta}_{\widehat{\mathrm{M}}}$ targets $\beta_{\widehat{\mathrm{M}}}$, inference based on $\widehat{\beta}_{\widehat{\mathrm{M}}}$ is inference for $\beta_{\widehat{\mathrm{M}}}$ and hence the VIDE problem in this case is to construct a valid confidence region $\widehat{\mathcal{R}}_{\widehat{\mathrm{M}}}$ in that it satisfies

$$\liminf_{n \to \infty} \mathbb{P}\left(\beta_{\widehat{\mathrm{M}}} \;\in\; \widehat{\mathcal{R}}_{\widehat{\mathrm{M}}}\right) \;\geqslant\; 1 - \alpha, \tag{1.2}$$

   *irrespective* of how $\widehat{\mathrm{M}} \subseteq \{1, 2, \ldots, p\}$ is obtained based on the data.

Selection of variables is only one of many outcomes of data exploration. As described above, variable transformation can also be seen as an outcome of data exploration. For each transformation $g : \mathbb{R} \to \mathbb{R}$, define the "target"

$$\beta_{g} \;:=\; \underset{\theta \in \mathbb{R}^{p}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(g(Y_i) - X_i^{\top}\theta)^2\right].$$

Similarly, the estimator $\widehat{\beta}_{g}$ is obtained as

$$\widehat{\beta}_{g} \;:=\; \underset{\theta \in \mathbb{R}^{p}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} (g(Y_i) - X_i^{\top}\theta)^2.$$

Based on the data, the practitioner chooses a transformation $\widehat{g} \in \mathcal{G}$ from a class of transformations. The class of Box-Cox transformations is one such example: $\{y \mapsto (y^{\lambda} - 1)/\lambda : \lambda > 0\}$. The VIDE problem in this case is to construct a valid confidence region $\widehat{\mathcal{R}}_{\widehat{g}}$ for $\beta_{\widehat{g}}$ in that it satisfies

$$\liminf_{n \to \infty} \mathbb{P}\left(\beta_{\widehat{g}} \;\in\; \widehat{\mathcal{R}}_{\widehat{g}}\right) \;\geqslant\; 1 - \alpha, \tag{1.3}$$

irrespective of how $\widehat{g} \in \mathcal{G}$ is obtained based on the data.

The VIDE problems (1.2) and (1.3) represent the prototypical problems solved in this thesis. The extension to the case of logistic, Poisson, and Cox regression models. The general VIDE problem can be described as follows. Suppose $Z_1, \ldots, Z_n$ are observations taking values in a set $\mathcal{Z}$. Consider a universe $\mathcal{Q}$ of all possible selections and for every $q \in \mathcal{Q}$ define the estimator

$$\widehat{\theta}_q \;:=\; \underset{\theta \in \Theta_q}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \ell_q(\theta, Z_i),$$

for a loss function $\ell_q(\cdot, \cdot)$ and a "parameter" set $\Theta_q$ (that could possibly depend on $q$). The data analyst can now choose an element $\widehat{q} \in \mathcal{Q}$ and the inference is to be based on the estimator $\widehat{\theta}_{\widehat{q}}$. The VIDE problem is to construct a confidence region $\widehat{\mathcal{R}}_{\widehat{q}}$ such that

$$\liminf_{n \to \infty} \mathbb{P}\left(\theta_{\widehat{q}} \;\in\; \widehat{\mathcal{R}}_{\widehat{q}}\right) \;\geqslant\; 1 - \alpha, \tag{1.4}$$

*irrespective* of how $\widehat{q} \in \mathcal{Q}$ is chosen based on the data. Here the "target" $\theta_{\widehat{q}}$ is defined as the evaluation of the map $q \mapsto \theta_q$, at $q = \widehat{q}$, given by

$$\theta_q \;:=\; \underset{\theta \in \Theta_q}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell_q(\theta, Z_i)].$$

As an example, one can take $Z_i = (X_i, Y_i)$, $\mathcal{Q} = \{M : M \subseteq \{1, 2, \ldots, p\}\}$, for $q = \mathrm{M} \in \mathcal{Q}$, $\Theta_q = \mathbb{R}^{|\mathrm{M}|}$, and $\widehat{\theta}_q = \widehat{\beta}_{\mathrm{M}}$. A second example is where $\mathcal{Q} = \{g : \mathbb{R} \to \mathbb{R} : g \in \mathcal{G}\}$, for $q = g \in \mathcal{G}$, $\Theta_q = \mathbb{R}^p$, and $\widehat{\theta}_q = \widehat{\beta}_g$.

The most important assumption of the VIDE framework is that the universe of estimators $\{\widehat{\theta}_q : q \in \mathcal{Q}\}$ is prefixed and is not allowed to depend on the data. For instance, one cannot choose $\ell_q(\cdot, \cdot)$, or $\Theta_q$, or $\mathcal{Q}$ based on the data.

**Some Limitations of the Framework** The formulation of the VIDE problem in (1.4) is very general but still has some limitations and does not cover certain types of exploration that would be considered reasonable/intuitive.

For instance, consider the following data exploration. Start with the observations $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$. Explore the data to find that modeling the response in terms of the covariates is not enough and pairwise interactions are needed to get a better model. Then choose a model $\widehat{\mathrm{M}} \subseteq \{1, 2, \ldots, p, (1, 2), (1, 3), \ldots, (p-1, p)\}$. Perform linear regression and draw some statistical conclusions. This does not fit into the problem formulation as in (1.2) and (1.3). The reason being that the decision of adding more covariates (such as pairwise interactions) is based on the data and the analyst could have taken a decision of adding more transformed variables instead of interactions. It will be shown in Chapter **??** that the VIDE problem is impossible to solve for these more general data exploration procedures.

# Bibliography


Bolt, S., Eadie, T., Yorkston, K., Baylor, C., and Amtmann, D. (2016). Variables associated with communicative participation after head and neck cancer. *JAMA Otolaryngology–Head & Neck Surgery*, 142(12):1145–1151.

Chowdhury, M. Z. I. and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1).

Gilley, O. W., Pace, R. K., et al. (1996). On the harrison and rubinfeld data. *Journal of Environmental Economics and Management*, 31(3):403–405.

Harrison Jr, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.;(United States)*, 5(1).

Lydersen, S. (2014). Statistical review: frequently given comments. *Annals of the rheumatic diseases*, pages annrheumdis–2014.

Moore, D. and McCabe, G. (1998). *Introduction to the Practice of Statistics*. W. H. Freeman.

Nersisyan, L., Nikoghosyan, M., and Arakelyan, A. (2019). Wgs-based telomere length analysis in dutch family trios implicates stronger maternal inheritance and a role for rrm1 gene. *Scientific Reports*, 9(1):1–9.

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, 16(2).

Stine, R. and Foster, D. (2013). *Statistics for Business: Decision Making and Analysis*. Always learning. Pearson Education.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of animal ecology*, 75(5):1182–1189.

Wiens, M., Kumbakumba, E., Larson, C., Ansermino, J., Singer, J., Kissoon, N., Wong, H., Ndamira, A., Kabakyenga, J., Kiwanuka, J., et al. (2015). Postdischarge mortality in children with acute infectious diseases: derivation of postdischarge mortality prediction models. *BMJ open*, 5(11):e009449.