Statistical Properties of a Univariate Feature Relevance Estimator in the Presence of Missing Data

Kristin Bennett[1], Isabelle Guyon[2], and Borja Seijo-Pardo[3]

***Abstract*—The t-statistic is frequently used to rank continuous features/variables in order of "relevance" to a binary outcome/target, and p-values are often cited as a metric of variable significance, in the context of univariate testing. We investigate statistical bias that can be introduced when the variable being tested has missing values that are imputed by regression using an auxiliary variable. We derive the variance of the difference of the means of the two outcomes estimated using the imputed variables and show that the pooled variance used in the t-statistic underestimates this variance. Thus applying the original two-sample t-test statistic for feature selection for problems with imputed will yield false positives due to the underestimation of variance. We propose a modified statistic reducing such bias. We make no assumptions about the joint distributions of variables being tested and the auxiliary variable. We suggest improved statistics. The appropriate statistics for testing the difference of means for feature selection when data is imputed remains an open question.**

I. INTRODUCTION

Missing values occur very frequently in data and have been the subject of extensive studies [1], [2], [3], [4], [5], [6], [7]. Practitioners often favor imputing missing values before applying a generic algorithm for classification, regression, and/or feature selection. Possible biases introduced by such procedures can be alleviated with multiple imputation [7], [8], [9], [10]. We are more particularly interested in the problem of feature selection. As a case study, we investigate the bias that can be introduced by imputing missing value by regression of continuous helper variable H (auxiliary variable with complete data) onto a continuous "source variable" S (variable of interest having missing values), for which we want to test the relevance to a binary "target variable" $T \in \{1, 2\}$ (a 2-class problem). We assume data missing completely at random (MCAR) to separate the problems of bias introduced by the "missingness mechanism" and bias introduced by imputation. As a criterion of relevance, we use the simple T-statistic, which computed the ratio of: (1) The difference in the means of S for the 2 classes $\bar{S_1} - \bar{S_2}$, and (2) the standard deviation of the difference of the means, which assuming equal class variance σ^2, is approximately $\sqrt{2/n}\sigma$, where n is the number of samples.

One way of dealing with missing values is to ignore all the n_{mis} missing samples and compute the t-statistic on the basis only of the n_{obs} observed samples. It is known [5] that asymptotically this will lead to unbiased results for when values are missing completely at random. But discarding samples with missing values results in a loss of statistical power (risk of false negatives). Thus imputation (*e.g.* by regression) is tempting. However, in the context of feature selection, there are at least three types of biases that can be introduced with imputation by regression:

- **Optimistic sample count**: a simple imputation with the mean of S (which could be thought of as a regression with a constant model) allows us to use all n samples instead of n_{obs}. However, this leads to a fake increase in the number of samples artificially inflating the t-statistic and learning to too small estimates of the p-value.
- **Optimistic variance estimate by ignoring the regression residual**: imputing with the mean of S or with an estimate of the expected value of S give H (regression) leads to an under-estimation of the variance of S because imputed values have less variance than real values. To avoid that, imputed values should be drawn from $P(S|H)$ to take into account the intrinsic noise (captured e.g. by the residual in a least-square regression), as performed in multiple imputations.
- **Optimistic variance estimate by ignoring the uncertainty on the regression parameter(s)**: when imputing by regression, we must estimate the parameters of the regression model(*e.g.* the slope a of a linear regression) using the n_{obs} available values. Our estimator of the variance of S using values imputed by regression must also take into account the uncertainty on our estimate of coefficient a.

In what follows, we derive adjusted t-statistics accounting for such biases. Our purpose is mostly didactic: we want to arrive at a simple formula, which includes terms accounting for systematic errors introduced by "intrinsic noise" (regression residual) and "finite training sample" (regression parameter error bar). We demonstrate that for even simple case is known with certainty, the usual t-test is optimistic due to underestimation of the increased variance arising from imputation.

Thus, we place ourselves in a simple case in which the t-statistic models well the data generating process, assumed to be:

- T is a Bernouilli process, $T \in \{1, 2\}$ with given a priori probabilities $P(T = 1)$ and $P(T = 2)$.
- S is drawn given T with: $P(S|T) \sim \setminus(\mu_T, \sigma_T)$.

H is a variable correlated with S. We use linear regression to

*This work was supported by the Paris-Saclay Center for Data Science.

[1]K. Bennett is Associate Director of Institute of Data Exploration and Application, and Professor of Mathematical Sciences at Rensselaer Polytechnic, Troy, NY 45435, 12180 bennek@rpi.edu

[2] Isabelle Guyon is Professor of Informatics at Université Paris-Saclay, France.

[3]Borja Seijo-Pardo is a Ph.D. student in Department of Computer Science - University of A Coruña, Spain

perform imputation:

$$\hat{S} = aH + b$$

Variables T, S, and H are jointly observed.

Other authors have addressed the related problem of variance estimation in the presence of imputation of missing data in a more detailed and general way (e.g. [11], [12], [13]), not directly relating it to the problem of feature selection. These could be useful follow up readings to generalize our findings to other univariate feature selection statistics. Our purpose in this paper is to arrive at a formula, which exhibits terms showing schematically the influence of the various types of biases, to alert the machine learning audience to such problems, without relatively simple calculations.

In Section II, we first compute a bias correction to the t-statistic assuming that coefficient a is known *a priori*, then derive a corresponding test (Section III). In Section IV, we then relax that assumption and take into account the variability of a due to finite training data.

II. Derivation of statistic for fixed regression coefficient

Let us first describe notations and assumptions:

$T \in \{1, 2\}$ are the two classes. There are n_T points in each class.

The samples of interest S_{Ti}, $i = 1 \dots n_T$ are the i.i.d samples in class T with mean μ_T and variance σ_T. The set of n_{obsT} observed samples in Class T are obs_T. The set of n_{misT} missing samples in Class T are mis_T.

The helper samples are H_{Ti}, $i = 1 \dots n_T$ are the i.i.d samples in class T. These are all observed, however we can still index them using the above index scheme as necessary.

We create estimates of $S_{Ti}, i \in mis_T$ using linear regression on S. This creates the variable $\hat{S}_{Ti} = aH_{Ti} + b + \epsilon_i$. For the moment we ignore the mechanism of how the coefficients of $\hat{S}$ are estimated. We assume ϵ are i.i.d. with mean 0 and standard deviation σ_ϵ. We assume a and σ_e are known.

We estimate μ_T using the imputed data by

$$\bar{S}_T = \frac{\sum_{i \in obs_T} S_{Ti} + \sum_{j \in mis_T} \hat{S}_{Ti}}{n_T} \tag{1}$$

We seek to test the alternative hypothesis that $H_1 : \mu_1 \neq \mu_2$ against the null hypotheses $H_0 : \mu_1 = \mu_2$.

Using only the observed data and the assumption that the standard deviations are equal for the two classes this can be accomplished using the two sample T-Test with common standard deviation:

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sigma_p \sqrt{\frac{1}{n_{obs1}} + \frac{1}{n_{obs2}}}} \tag{2}$$

where the pooled standard deviation is $\sigma_p = \sqrt{\frac{(n_{obs1}-1)\sigma^2_{obs1} + (n_{obs2}-1)\sigma^2_{obs2}}{n_{obs1}+n_{obs2}-2}}$.

We now want to see how this test changes when the incorporate the imputed data. To do this we begin by deriving the mean and variance of our estimate of $\bar{S}_1 - \bar{S}_2$ (after imputation).

A. Derivation of mean and variance of the estimate

We want to compute the expected value of the difference of the adjusted means, $E(\bar{S}_1 - \bar{S}_2)$.

To do this we first derive a result for any $j \in m_T$.

$$\begin{aligned}
E(\hat{S}_{Tj}) &= E(\hat{a}H_{Tj} + \hat{b} + \epsilon_j) && (3)\\
&= E(\hat{a}H_{Tj}) + E(\hat{b}) + E(\epsilon_j) && (4)\\
&= E(\hat{a}H_{Tj}) + b + 0 && (5)\\
&= cov(\hat{a}, h_{Tj}) + E(\hat{a})E(h_{Tj}) + b && (6)\\
&= a\mu_{HT} + b \text{ if a is fixed} && (7)
\end{aligned}$$

The later comes from the fact that OLS creates unbiased estimates of a and b and for fixed a, $cov(\hat{a}, h_{Tj}) = 0$. If a is calculated then $cov(\hat{a}, h_{Tj}) = cov(g(S_{obs1}, H_{obs1}), H_{Tj})$ may not be zero, where g is a fixed function of the observed data.

Using this result we can find the expected difference of the estimates (abbreviating obs with o and mis with m):

$$\begin{aligned}
&E(\bar{S}_1 - \bar{S}_2) = \\
&E\left(\frac{\sum_{i \in o_1} S_{1i} + \sum_{j \in m_1} \hat{S}_{1j}}{n_1} - \frac{\sum_{i \in o_2} S_{2i} + \sum_{j \in m_2} \hat{S}_{2j}}{n_2}\right)\\
&= \frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}E(\hat{S}_1) - \frac{n_{o2}}{n_2}\mu_2 - \frac{n_{m2}}{n_2}E(\hat{S}_2)\\
&= \frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}(cov(\hat{a}, h_1) + E(\hat{a})E(h_1) + b))\\
&- \frac{n_{o2}}{n_2}\mu_2 - \frac{n_{m2}}{n_2}(cov(\hat{a}, h_2) + E(\hat{a})E(h_2) + b))
\end{aligned} \tag{8}$$

If we assume that a is fixed or that $\frac{n_{m1}}{n_1}(cov(\hat{a}, h_1) = \frac{n_{m2}}{n_2}(cov(\hat{a}, h_2)$, then the expression simplifies to

$$\begin{aligned}
&E(\bar{S}_1 - \bar{S}_2) = \\
&\left(\frac{n_{o1}}{n_1}\mu_1 + \frac{n_{m1}}{n_1}a\mu_{H1}\right) - \left(\frac{n_{o2}}{n_2}\mu_2 + \frac{n_{m1}}{n_2}a\mu_{H2}\right)
\end{aligned} \tag{9}$$

If we assume that the sample sizes for missing and observed are the same for both classes, this simplifies to

$$E(\bar{S}_1 - \bar{S}_2) = (f_o\mu_1 + f_m a\mu_{H1}) - (f_o\mu_2 + f_m a\mu_{H2}) \tag{10}$$

where f_o and f_m are the fractions of observed and missing data respectively.

Now let's derive the variance. We define

$$Q_o = \frac{\sum_{i \in o_1} S_{1i}}{n_1} - \frac{\sum_{i \in o_2} S_{2i}}{n_2} \tag{11}$$

and

$$Q_m = \frac{\sum_{j \in m_1} \hat{S}_{1j}}{n_1} - \frac{\sum_{j \in m_2} \hat{S}_{2j}}{n_2} \tag{12}$$

The result also exploits the facts that $S_i \perp S_j$ and $S_i \perp \hat{S}_j$ for $i \neq j$, and $S_i \perp H_j$ and $\hat{a} \perp H_j$for $i \in obsT$, $H_j \in misT$.

$$
\begin{aligned}
var(\bar{S}_1 - \bar{S}_2) &= var(Q_o + Q_m) \\
&= var(Q_o) + var(Q_m) + 2cov(Q_o, Q_m) \\
&= \frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) \\
&+ \frac{n_{m1}}{n_1^2} var(\hat{S}_1) + \frac{n_{m2}}{n_2^2} var(\hat{S}_2) + 2cov(Q_o, Q_m) \\
&= \frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) \\
&+ \frac{n_{m1}}{n_1^2} var(aH_1 + \epsilon) + \frac{n_{m2}}{n_2^2} var(aH_2 + \epsilon) \\
&+ 2cov(Q_o, Q_m)
\end{aligned} \tag{13}
$$

For the case of a fixed. We can exploit $cov(Q_o, Q_m) = 0$. This covariance may not be 0 for estimated $\hat{a}$.

We can further simplify the results

$$
\begin{aligned}
&var(\bar{S}_1 - \bar{S}_2) = \\
&\frac{n_{o1}}{n_1^2}\sigma_1^2 + \frac{n_{o2}}{n_2^2}\sigma_2^2 + \frac{n_{m1}}{n_1^2}(a^2\sigma_{H1}^2 + \sigma_\epsilon^2) + \frac{n_{m2}}{n_2^2}(a^2\sigma_{H2}^2 + \sigma_\epsilon^2)
\end{aligned} \tag{14}
$$

If we assume a fixed, classes have same variances, and equal sample sizes for both classes, this reduces to

$$
var(\bar{S}_1 - \bar{S}_2) = \frac{2}{n}(f_o\sigma^2 + f_m(a^2\sigma_H^2 + \sigma_\epsilon^2)) \tag{15}
$$

The distribution of this statistic would depend on the assumptions of the problem.

III. Creating the test for fixed a

We use as a our statistic (for the case of a fixed so far):

$$
t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(f_o\sigma^2 + f_m(a^2\sigma_H^2 + \sigma_\epsilon^2)}} \tag{16}
$$

with the variables replaced by their corresponding sample estimates.

If we further assume S and H have same variance and $a = 1$, then

$$
t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(\sigma_{all}^2 + f_m\sigma_\epsilon^2)}} \tag{17}
$$

where σ_{all}^2 is the sample covariance of observed S_T and imputed $\hat{S}_T$ combined.

IV. Creating the test for estimated a

The coefficients of the regression function have a closed form so we can figure out their distributions.

Define $\bar{S}_o$ as the sample mean of $S_i, i \in obs$, $\bar{H}_o$ and $var(H_o)$ as the sample mean and variance of $H_i, i \in obs$, and $cov(H_o, S_o)$ as the corresponding sample covariance. Then

$$
\hat{a} = \frac{cov(H_o, S_o)}{var(H_o)} \tag{18}
$$

$$
\hat{b} = \bar{S}_o - \hat{a}\bar{H}_o \tag{19}
$$

$$
E(\hat{a}) = a \tag{20}
$$

$$
var(\hat{a}) = \frac{\sigma_\epsilon^2}{n_o\sigma_H^2} \tag{21}
$$

where $\sigma_R^2 = \frac{\sum_{i\in o}(S_i - \hat{s_i})}{n_o - 2}$.

Note that we can remove b by normalizing observed data so let's assume we do that and neglect b from further calculations.

We need a key result here on the variance of the estimated S for both classes:

$$
\begin{aligned}
var(\hat{S}_{Ti}) &= var(\hat{a}H_{Ti} + \epsilon) \\
&= var(\hat{a}H_{Ti}) + var(\epsilon) \\
&= var(\hat{a})var(H_{Ti}) + var(\hat{a})E(H_{Ti})^2 \\
&+ var(H_{Ti})E(\hat{a})^2 + var(\epsilon) \\
&= \frac{\sigma_\epsilon^2}{n_o\sigma_{H_o}^2}\sigma_{H_T}^2 + \frac{\sigma_\epsilon^2}{n_o\sigma_{H_o}^2}\mu_{H_T}^2 + \sigma_{H_T}^2 a^2 + \sigma_\epsilon^2 \\
&= \sigma_{H_o}^2 a^2 + \left[1 + \frac{1}{n_o} + \frac{\mu_{H_T}^2}{n_o\sigma_{H_o}^2}\right]\sigma_\epsilon^2
\end{aligned} \tag{22}
$$

Note this depends on $var(XY) = var(X)(var(Y) - var(X)E(Y)^2 - Var(Y)E(X)^2$

The last step assumes $\sigma_{H_o}^2 = \sigma_{H_T}^2$, i.e. that the variance of H is the same for all T. Estimation of the a increases the impact of the residual on the variance, but this additional impact goes to 0 as $n_0 \to \infty$.

Note

$$
\begin{aligned}
&var(Q_o) = \\
&\frac{n_{o1}}{n_1^2} var(S_1) + \frac{n_{o2}}{n_2^2} var(S_2) = \frac{n_{o1}}{n_1^2}\sigma_1^2 + \frac{n_{o2}}{n_2^2}\sigma_2^2
\end{aligned} \tag{23}
$$

since the $cov(S_1, S_2) = 0$. and

$$
\begin{aligned}
&var(Q_m) = \\
&\frac{n_{m1}}{n_1^2} var(\hat{S}_1) + \frac{n_{m2}}{n_2^2} var(\hat{S}_2) - \frac{2}{n_1 n_2} cov(\sum_{i\in m_1} \hat{S}_{1i}, \sum_{j\in m_2} \hat{S}_{2j})
\end{aligned} \tag{24}
$$

Here we assume that the covariance term is negligible since its limit is 0 as the sample size grows and the only dependence between $\hat{S}_1$ and $\hat{S}_2$ is through $\hat{a}$.

Under are simplifying assumptions that that S and H have the same variance and a=1, this becomes

$$\begin{aligned} var(Q_o) &= \frac{2f_o}{n}\sigma^2 \\ var(Q_m) &= \frac{2f_m}{n}\left[\sigma^2 + \left[1 + \frac{(2\sigma^2 + \mu_{H_1}^2 + \mu_{H_2}^2)}{2n_o\sigma^2}\right]\sigma_\epsilon^2\right] \end{aligned} \tag{25}$$

Thus under the assumptions that the variance of S and H are equal and that the sample size is sufficiently large,

$$\begin{aligned} var(\bar{S}_1 - \bar{S}_2) &= var(Q_o + Q_m) \\ &= var(Q_o) + var(Q_m) + 2cov(Q_o, Q_m) \\ &= \frac{2}{n}\left[\sigma^2 + f_m\left[1 + \frac{(2\sigma^2 + \mu_{H_1}^2 + \mu_{H_2}^2)}{2n_o\sigma^2}\right]\sigma_\epsilon^2\right] \\ &\qquad + 2cov(Q_o, Q_m) \\ &\approx \frac{2}{n}\left[\sigma^2 + f_m\left[1 + \frac{(2\sigma^2 + \mu_{H_1}^2 + \mu_{H_2}^2)}{2n_o\sigma^2}\right]\sigma_\epsilon^2\right] \end{aligned} \tag{26}$$

Neglecting the covariance will result in an underestimate of the variance of $\bar{S}_1 - \bar{S}_2$. Recall that Q_0 is the difference in the estimates of the means of the classes on the observed data and that Q_m is the differences in the means of the classes on the imputed data. Thus Q_0 and Q_m will be positively correlated. As the number of observed samples goes to infinity for a fixed f_m, then Q_o and a rapidly converge to constants. Thus the problem converges to the case of a known which has the $cov(Q_0, Q_m) = 0$ so the covariance term would have little impact for problems with large numbers of observed data.

This suggests a statistic of the form

$$t = \frac{\bar{S}_1 - \bar{S}_2}{\sqrt{\frac{2}{n}(\sigma_{all}^2 + f_m\left[1 + \frac{\alpha}{n_o}\right]\sigma_\epsilon^2)}} \tag{27}$$

with $\alpha > 0$ chosen appropriately. The best statistic and the distribution of that statistic depends on assumptions one the joint distribution of S and H, and thus remains an open question. But clearly when used with imputed data, the typical t-test may induce false positive errors due to underestimation of the variance.

ACKNOWLEDGEMENTS

We thank our colleagues of the Center for Data Science and in particular Balázs Kégl for supporting the launching event of the Paris-Saclay workgroup on missing data and for supporting the visit of Kristin Bennett. We are particularly grateful to Julie Josse for her guidance and for many stimuating discussions with Alain-Jacques Valleron, Amparo Alonso-Betanzos, Verónica Bolón, and Mehreen Saeed.

REFERENCES

[1] J. Pearl and K. Mohan, "Recoverability and testability of missing data: Introduction and summary of results," Tech. Rep., 2013. [Online]. Available: http://ftp.cs.ucla.edu/pub/stat_ser/r417.pdf

[2] I. Shpitser, K. Mohan, and J. Pearl, "Missing data as a causal and probabilistic problem," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, Amsterdam, Netherlands, 2015, pp. 802–811.

[3] K. Mohan, J. Pearl, and T. Jin, "Missing data as a causal inference problem," in *Proceedings of the Neural Information Processing Systems Conference (NIPS)*, 2013. [Online]. Available: https://ssrn.com/abstract=2343794

[4] C. K. Enders, *Applied missing data analysis*. Guilford Press, 2010.

[5] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[6] J. W. Graham, "Missing data analysis: Making it work in the real world," *Annual review of psychology*, vol. 60, pp. 549–576, 2009.

[7] N. T. Longford, *Missing data and small-area estimation: Modern analytical equipment for the survey statistician*. Springer Science & Business Media, 2006.

[8] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[9] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2012.

[10] P. Royston *et al.*, "Multiple imputation of missing values: update," *Stata Journal*, vol. 5, no. 2, p. 188, 2005.

[11] J. C. Deville and C. E. Särndal, "Variance estimation for the regression imputed Horvitz-Thompson estimator," vol. 10, pp. 381–394, 1994.

[12] J. N. K. Rao, "Variance estimation in the presence of imputation for missing data," in *AMSTAT*, 2000.

[13] J.-K. Kim, "Variance estimation after imputation," vol. 27, pp. 75–83, 2001.