Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding

Victor Veitch[1] and Anisha Zaveri[2]

[1]*Columbia University*
[2]*Weill Cornell Medicine*

Abstract

It is a truth universally acknowledged that an observed association without known mechanism must be in want of a causal estimate. However, causal estimation from observational data often relies on the (untestable) assumption of 'no unobserved confounding'. Violations of this assumption can induce bias in effect estimates. In principle, such bias could invalidate or reverse the conclusions of a study. However, in some cases, we might hope that the influence of unobserved confounders is weak relative to a 'large' estimated effect, so the qualitative conclusions are robust to bias from unobserved confounding. The purpose of this paper is to develop *Austen plots*, a sensitivity analysis tool to aid such judgments by making it easier to reason about potential bias induced by unobserved confounding. We formalize confounding strength in terms of how strongly the confounder influences treatment assignment and outcome. For a target level of bias, an Austen plot shows the minimum values of treatment and outcome influence required to induce that level of bias. Domain experts can then make subjective judgments about whether such strong confounders are plausible. To aid this judgment, the Austen plot additionally displays the estimated influence strength of (groups of) the observed covariates. Austen plots generalize the classic sensitivity analysis approach of Imbens [Imb03]. Critically, Austen plots allow *any* approach for modeling the observed data and producing the initial estimate. We illustrate the tool by assessing biases for several real causal inference problems, using a variety of machine learning approaches for the initial data analysis. Code is available at github.com/anishazaveri/austen_plots.

1 Introduction

The high costs of randomized controlled trials coupled with the relative availability of (large scale) observational data motivate attempts to infer causal relationships from observational data. For example, we might want to use electronic health record data to infer the expected outcome of intervening by assigning a particular treatment to a new patient. Causal inference from observational data must account for possible *confounders* that influence both treatment assignment and the outcome; e.g., wealth may be a common cause influencing whether a patient takes an expensive drug and whether they recover. Often, causal inference is based on the assumption of 'no unobserved confounding'; i.e., the assumption that the observed covariates include all common causes of the treatment assignment and outcome. This assumption is fundamentally untestable from observed data, but its violation can induce bias in the estimation of the treatment effect—the unobserved confounding may completely or in part explain the observed association. Our aim here is to develop a sensitivity analysis tool to aid in reasoning about potential bias induced by unobserved confounding.

Intuitively, if we estimate a large positive effect then we might expect the real effect is also positive, even in the presence of mild unobserved confounding. This is extremely useful

because it allows us to use observational data even in cases where we cannot strictly rule out all unobserved confounding. For example, consider the association between smoking and lung cancer. One could argue that this association arises from a genetic mutation that predisposes carriers to both an increased desire to smoke and to a greater risk of lung cancer. However, the association between smoking and lung cancer is large—is it plausible that some unknown genetic association could have a strong enough influence to explain the association? Answering such questions requires a domain expert to make a judgment about whether the confounding is "mild" relative to the "large" effect. In particular, the domain expert must translate judgments about the strength of the unobserved confounding into judgments about the bias induced in the estimate of the effect. The aim of sensitivity analysis is to formalize what is meant by the strength of unobserved confounding, and to show how to translate judgments about confounding strength into judgments about bias.

A prototypical example, due to Imbens [Imb03] (building on [RR83a]), illustrates the broad approach. The observed data consists of a treatment T (the patient took a blood pressure medication), an outcome Y (their blood pressure), and covariates X that may causally affect the treatment and outcome (socioeconomic status, age, baseline health measurements, etc.). We then posit an additional unobserved binary confounder U for each patient; e.g., a gene that affects both blood pressure and conscientiousness. Next, we suppose that the observed data and unobserved confounder were generated according to the following generative model:

$$\begin{aligned} U_i &\overset{\text{iid}}{\sim} \text{Bern}(1/2) \\ T_i \mid X_i, U_i &\overset{ind}{\sim} \text{Bern}(\text{sig}(\gamma X_i + \alpha U_i)) \\ Y_i \mid X_i, T_i, U_i &\overset{ind}{\sim} \text{Normal}(\tau T_i + \beta X_i + \delta U_i, \sigma^2). \end{aligned} \tag{1.1}$$

where $\text{sig} = \exp(\cdot)/(1+\exp(\cdot))$ is the sigmoid function. If U_i had been observed, we could estimate $(\hat{\tau}, \hat{\gamma}, \hat{\beta}, \hat{\alpha}, \hat{\delta}, \hat{\sigma}^2)$ from the data by maximum likelihood estimation, and report $\hat{\tau}$ as the estimate of the average treatment effect. Since U_i is not observed, it is not possible to identify the parameters from the data. Instead, we make (subjective) judgments about plausible values of α—how strongly U_i affects the treatment assignment—and δ—how strongly U_i affects the outcome. Contingent on plausible $\alpha = \alpha^*$ and $\delta = \delta^*$, the other parameters can be estimated by maximum likelihood. This yields an estimate of the treatment effect $\hat{\tau}(\alpha^*, \delta^*)$ under the presumed values of the sensitivity parameters.

This procedure was substantially improved by Imbens [Imb03] through two innovations. First, he re-parameterized the model so that the sensitivity parameters are the partial coefficients of determination for U_i instead of the raw regression coefficients. This introduces a standardized scale for the strength of influence, and facilitates comparison with observed covariates where the strength of influence can be measured from data. Second, he plotted the curve of all values of the sensitivity parameter that would yield given level of bias. This moves the analyst judgment from "what are plausible values of the sensitivity parameters?" to "are sensitivity parameters this extreme plausible?" We will adapt both innovations below.

The approach just outlined has a major drawback: it is completely reliant on the simple parametric specification (1.1). The assumed model is equivalent to assuming that, had U been observed, it would have been appropriate to use logistic regression to model treatment assignment, and linear regression to model the outcome. This assumption also implies a simple, parametric model for the relationships governing the observed data. This is out of step with modern practice, where we use flexible machine-learning methods to model these relationships. For example, the assumption forbids the use of neural networks or random forests, though such methods are often state-of-the-art for causal effect estimation [e.g., Dor+19; HB19; AI19]. In fact, the problem is worse still. Franks et al. [Fra+19] demonstrate that changing the assumed values of α^* and δ^* also changes how well the

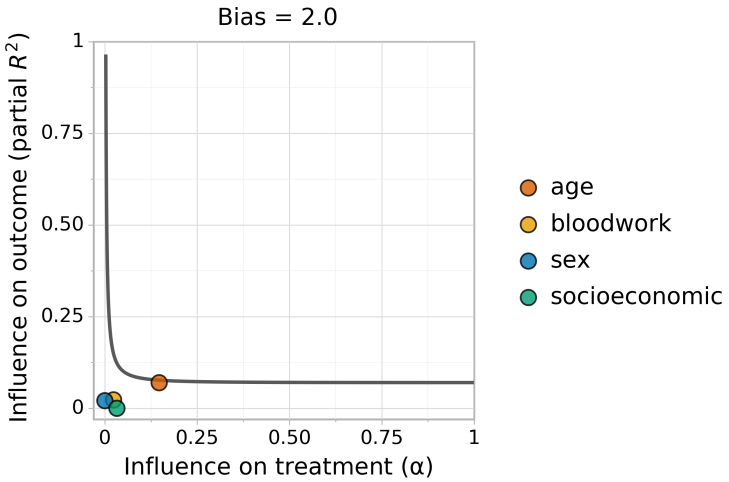

Figure 1: Austen plot showing how strong an unobserved confounder would need to be to induce a bias of 2 in an observational study of the effect of combination blood pressure medications on diastolic blood pressure [Dor+16]. We chose this bias to equal the nominal average treatment effect estimated from the data. We model the outcome with Bayesian Additive Regression Trees and the treatment assignment with logistic regression—Austen plots accommodate any choice of models. The curve shows all values treatment and outcome influence that would induce a bias of 2. The colored dots show the influence strength of (groups of) observed covariates, given all other covariates. For example, an unobserved confounder with as much influence as the patient's age might induce a bias of about 2.

assumed model fits the *observed* data. However, the point of sensitivity analysis is to aid us in domain-specific judgments about an assumption that is *not testable from data*. The approach lumps together poor model fit and unobserved confounding. In this sense, the fully-parametric cure may be worse than the unobserved-confounding disease.

There are a number of approaches that sidestep the parametric specification problem [e.g., Fra+19; She+11; VA11; DV15]. However, these methods require subjective judgments about sensitivity parameters that are more abstract than the ones used in Imbens' approach. See section 5 for a more detailed discussion. For example, Franks et al. [Fra+19] requires the analyst to make subjective judgments about posterior probability of treatment given the counterfactual outcomes. Such judgments may be possible and natural in some circumstances, but in many circumstances we'd prefer to make judgments on the more directly interpretable confounding strength parameters of the parametric approach (1.1).

Austen plots The purpose of this paper is to introduce *Austen plots*, an adaptation of Imbens' approach that fully decouples sensitivity analysis and modeling of the observed data. An example Austen plot is shown in Figure 1. The key elements of Imbens' approach are preserved. Plausibility judgments are made (i) post-hoc, (ii) on directly interpretable quantities—how much the confounder influences Y and T, and (iii) on a standardized scale that facilitates comparison to the influence strength of (groups of) observed covariates. Moreover, because the model underlying the Austen plot has no implications for the observed data, the data analysis may use absolutely any modeling approach, including any combination of machine-learning methods. Austen plots can then be constructed post-hoc to aid in reasoning about the potential impact of unobserved confounding.

Austen plots measure the confounder influence on the outcome as the partial R^2 of the confounder given the observed covariates. The confounder influence on treatment assignment is measured as

$$\alpha = \mathbb{E}[\mathrm{P}(T = 1 \mid U, x) \mid T = 1] - \mathbb{E}[\mathrm{P}(T = 1 \mid U, x) \mid T = 0]. \tag{1.2}$$

That is, α is the change in our belief about how likely a unit was to have gotten the treatment, given that they were actually observed to be treated (or not). If an unobserved confounder

strongly affects treatment assignment (with large u implying high probability of treatment), then actually seeing somebody treated suggests a high value of u, which in turn suggests they were likely to have been treated. Austen plots rely on a model of confounder influence to calculate bias from the confounder influence. This model, a simple generalization of Imbens' approach, is discussed below.

Figure 1, an Austen plot for an observational study of the effect of combination medications on diastolic blood pressure, illustrates the idea. A bias of 2 would suffice to undermine the qualitative conclusion that the blood-pressure treatment is effective. Examining the plot, an unobserved confounder as strong as age could induce this amount of confounding, but no other (group of) observed confounders has so much influence. Accordingly, if a domain expert thinks an unobserved confounder as strong as age is unlikely then they will think the treatment is likely effective. Or, if such a confounder is plausible, they will think the study fails to establish efficacy.

Notation For concreteness, we consider the estimation of the average effect of a binary treatment. The data are generated independently and identically $(Y_i, T_i, X_i, U_i) \overset{\text{iid}}{\sim} P$, where U_i is not observed and P is some unknown probability distribution. The average treatment affect (ATE) is

$$\texttt{ATE} = \mathbb{E}[Y \mid \text{do}(T = 1)] - \mathbb{E}[Y \mid \text{do}(T = 0)].$$

The use of Pearl's do notation indicates that the effect of interest is causal.

The traditional approach to causal estimation assumes that the observed covariates X contain all common causes of Y and T. If this 'no unobserved confounding' assumption holds, then the ATE is equal to parameter, τ, of the observed data distribution, where

$$\tau = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1] - \mathbb{E}[Y \mid X, T = 0]]. \tag{1.3}$$

The parameter τ can be estimated from a finite data sample. The general approach proceeds in two steps. First, we produce estimates $\hat{g}$ and $\hat{Q}$ for the propensity score g and the conditional expected outcome Q, where

Definition 1.1. The *propensity score* g is $g(x) = \text{P}(T = 1 \mid X = x)$ and the *conditional expected outcome* Q is $Q(t, x) = \mathbb{E}[Y \mid T = t, X = x]$.

In modern practice, Q and g are often estimated by fitting flexible machine learning models. The second step is to plug the estimated $\hat{Q}$ and $\hat{g}$ in to some downstream estimator $\hat{\tau}$. For example, following 1.3, the estimator

$$\hat{\tau}^{Q} = \frac{1}{n} \sum_i \hat{Q}(1, x_i) - \hat{Q}(0, x_i),$$

is a natural choice. Other estimators incorporate $\hat{g}$.

We are interested in the case of possible unobserved confounding. That is, where U causally affects Y and T. If there is unobserved confounding then the parameter τ is not equal to the ATE, so $\hat{\tau}$ is a biased estimate. Inference about the ATE then divides into two tasks. First, the statistical task: estimating τ as accurately as possible from the observed data. And, second, the causal (domain-specific) problem of assessing $\texttt{bias} = \texttt{ATE} - \tau$. Our aim is to reason about the bias induced by unobserved confounding—the second task—in a way that imposes no constraints on the modeling choices for $\hat{Q}$, $\hat{g}$ and $\hat{\tau}$ used in the initial analysis.

2 Sensitivity Model

Our sensitivity analysis should impose no constraints on how the *observed* data is modeled. However, sensitivity analysis demands some assumption on the relationship between the

observed data and the *unobserved* confounder. It is convenient to formalize such assumptions by specifying a probabilistic model for how the data is generated. The strength of confounding is then formalized in terms of the parameters of the model (the sensitivity parameters). Then, the bias induced by the confounding can be derived from the assumed model. Our task is to posit a generative model that both yields a useful and easily interpretable sensitivity analysis, and that avoids imposing any assumptions about the observed data.

To begin, consider the functional form of the sensitivity model used by Imbens [Imb03] (cf. (1.1)).

$$\operatorname{logit} \mathrm{P}(T = 1 \mid x, u) = h(x) + \alpha u \tag{2.1}$$

$$\mathbb{E}[Y \mid t, x, u] = l(t, x) + \delta u, \tag{2.2}$$

for some functions h and l. That is, the propensity score is logit-linear in the unobserved confounder, and the conditional expected outcome is linear. In principle, we could accomplish our task by finding a choice of h, l and distribution $\mathrm{P}(U \in \cdot)$ so that the implied marginal distribution on the observed data agrees with the actual distribution of the observed data.

The first key insight is that, by re-arranging (2.1) to solve for u and plugging in to (2.2), we see that the following functional form is an equivalent assumption:

$$\begin{aligned} \operatorname{logit} \mathrm{P}(T = 1 \mid x, u) &= h(x) + \alpha u \\ \mathbb{E}[Y \mid t, x, u] &= l(t, x) + \delta \operatorname{logit} \mathrm{P}(T = 1 \mid x, u). \end{aligned}$$

The important observation is that the unobserved confounder u only influences the data distribution through the propensity score. This means that we can circumvent the need to explicitly articulate the distribution of U and the function h by directly positing a distribution on $\mathrm{P}(T = 1 \mid x, u)$. This idea—treating the propensity score as a first-class citizen for sensitivity analysis—is the key to the analysis.

Definition 2.1. Let $\tilde{g}(x, u) = \mathrm{P}(T = 1 \mid x, u)$ denote the propensity score given observed covariates x and the unobserved confounder u.

The insight is that we can posit a sensitivity model by defining a distribution on $\tilde{g}$ directly. Our choice is:[1]

$$\tilde{g}(X, U) \mid X \sim \mathrm{Beta}(g(X)(1/\alpha - 1), (1 - g(X))(1/\alpha - 1)).$$

The sensitivity parameter α controls the influence of the unobserved confounder U on treatment assignment. When α is close to 0 then $\tilde{g}(X, U) \mid X$ is tightly concentrated around $g(X)$, and the unobserved confounder has little influence. That is, U minimally affects our belief about who is likely to receive treatment. Conversely, when α is close to 1 then $\tilde{g}$ concentrates near 0 and 1; i.e., knowing U would let us predict treatment assignment with near certainty.

With the $\tilde{g}$ model in hand, we define our sensitivity model:

Assumption 1 (Sensitivity Model)**.**

$$\begin{aligned} \tilde{g}(X, U) \mid X &\sim \mathrm{Beta}(g(X)(1/\alpha - 1), (1 - g(X))(1/\alpha - 1)) \\ T \mid X, U &\sim \mathrm{Bern}(\tilde{g}(X, U)) \\ \mathbb{E}[Y \mid T, X, U] &= Q(T, X) + \delta\big(\operatorname{logit} \tilde{g}(X, U) - \mathbb{E}[\operatorname{logit} \tilde{g}(X, U) \mid X, T]\big). \end{aligned}$$

[1] the peculiar Beta parameterization is so that α has the simple interpretation given in (1.2).

This model has been constructed to satisfy the requirement that the propensity score and conditional expected outcome are the g and Q actually present in the observed data:

$$\begin{aligned}\mathrm{P}(T=1 \mid X) &= \mathbb{E}[\mathbb{E}[T \mid X,U] \mid X] = \mathbb{E}[\tilde{g}(X,U) \mid X] = g(X)\\ \mathbb{E}[Y \mid T,X] &= \mathbb{E}[\mathbb{E}[Y \mid T,X,U] \mid T,X] = Q(T,X).\end{aligned}$$

The sensitivity parameters are α, controlling the dependence between the unobserved confounder the treatment assignment, and δ, controlling the relationship with the outcome. In effect, by making an assumption about the propensity score directly we have sidestepped the need to explicitly articulate the parts of the observed/unobserved relationship that are not actually relevant for the treatment effect estimation. This lets us respect the no-observable-implications desiderata while also maintaining a simple sensitivity model and easily interpretable sensitivity parameters.

Remark 2.2. We motivated the form of our sensitivity model by analogy to Imbens' model. A more general, but more complicated, derivation proceeds as follows. The propensity score carries all information in (X,U) that is relevant for causal identification [RR83b]. Accordingly, without loss of generality, the unobserved confounding bias may be assumed to be induced according to $\mathbb{E}[Y \mid X,T,U] = Q(T,X) + f(\mathrm{logit}(\tilde{g})) - \mathbb{E}[f(\mathrm{logit}(\tilde{g})) \mid X,T]$ for some function f. Our assumed model is equivalent to a first-order approximation of the general model in $\mathrm{logit}(\tilde{g}/g)$. If α is small (and f is smooth with bounded second derivatives) then it is possible to show the results that follow are approximately correct even for the fully general outcome model. However, for simplicity, we do not pursue this further.

Bias We now turn to calculating the bias in the estimate implied by a given choice of sensitivity parameters. By assumption, X and U together suffice to render the average treatment effect identifiable as:

$$\mathtt{ATE} = \mathbb{E}[\mathbb{E}[Y \mid 1,X,U] - \mathbb{E}[Y \mid 0,X,U]].$$

Plugging in our sensitivity model yields,

$$\mathtt{ATE} = \mathbb{E}[Q(1,X) - Q(0,X)] + \delta(\mathbb{E}[\mathrm{logit}\,\tilde{g}(X,U) \mid X, T=1] - \mathbb{E}[\mathrm{logit}\,\tilde{g}(X,U) \mid X, T=0]).$$

The first term is the observed-data estimate τ, so

$$\mathtt{bias} = \delta(\mathbb{E}[\mathrm{logit}\,\tilde{g}(X,U) \mid X, T=1] - \mathbb{E}[\mathrm{logit}\,\tilde{g}(X,U) \mid X, T=0]).$$

Then, by invoking Beta-Bernoulli conjugacy and standard Beta identities, we arrive at,

Theorem 2.3. *Under our sensitivity model, Assumption 1, an unobserved confounder with influence α and δ induces bias in the estimated treatment effect equal to*

$$\begin{aligned}\mathtt{bias} = \delta\mathbb{E}\big[&\psi\big(g(X)(1/\alpha - 1) + 1\big) - \psi\big((1-g(X))(1/\alpha - 1)\big)\\ &- \psi\big(g(X)(1/\alpha - 1)\big) + \psi\big((1-g(X))(1/\alpha - 1) + 1\big)\big],\end{aligned}$$

where ψ is the digamma function

In practice, can we estimate the bias by replacing the expectation (over X) with the mean over the data, and g by the estimated $\hat{g}$.

Reparameterization The model in the previous section provides a formalization of confounding strength and tells us how much bias is induced by a given strength of confounding. This lets us translate judgments about confounding strength to judgments about bias. For this to be useful, we must be able to make plausibility judgments about confounding strength, or, equivalently, the sensitivity parameters. To facilitate such judgments, we move to a more convenient reparameterization of the model.

Following Imbens [Imb03], we will re-express the outcome-confounder strength in terms of the partial coefficient of determination:

$$R^2_{Y,\text{par}}(\alpha,\delta) = \frac{\mathbb{E}(Y - Q(T,X))^2 - \mathbb{E}(Y - \mathbb{E}[Y \mid T,X,U])^2}{\mathbb{E}(Y - Q(T,X))^2}.$$

The advantages of this parameterization over δ are that $R^2_{Y,\text{par}}$ has a familiar interpretation—the proportion of previously unexplained variation in Y that is explained by the unobserved covariate U—and that it has a fixed, unitless scale—enabling easy comparisons with reference values.

The reparameterization depends on the unobserved quantity $\mathbb{E}[Y \mid T,X,U]$. To actually compute the reparameterization, we establish the following result:

Theorem 2.4. *Under our sensitivity model, Assumption 1, the outcome influence is*

$$R^2_{Y,\text{par}}(\alpha,\delta) = \delta^2 \sum_{t=0}^{1} \frac{\mathbb{E}\left[\psi_1\left(g(X)^t(1-g(X))^{1-t}(1/\alpha - 1) + 1[T=t]\right)\right]}{\mathbb{E}[(Y - Q(T,X))^2]},$$

where ψ_1 is the trigamma function.

We defer the proof to the supplementary material.

Notice that, since $R^2_{Y,\text{par}} \in [0,1)$, this expression restricts the possible values of δ. This is a particular manifestation of the requirement that the distribution for the observed data implied by the sensitivity model should match the actual distribution of the observed data. An advantage of the $R^2_{Y,\text{par}}$-parameterization is that it enforces the δ restriction automatically.

We do not re-parameterize the strength of confounding on treatment assignment because, by design, α is already interpretable and on a fixed, unitless scale.

Estimating bias In combination, Theorems 2.3 and 2.4 yield an expression for the bias in terms of α and $R^2_{Y,\text{par}}$. Namely, we rearrange Theorem 2.4 to solve for δ, and plug this expression into Theorem 2.3. Note that because the partial R^2 only depends on δ^2, the sign of the bias is ambiguous (i.e., we can only calculate the absolute value of the bias from α and $R^2_{Y,\text{par}}$).

In practice, we can estimate the bias induced by confounding by fitting models for $\hat{Q}$ and $\hat{g}$ and replacing the expectations by means over the data. To avoid problems associated with overfitting, we recommend a data splitting approach. Namely, split the data into k folds and, for each fold, estimate $Q(t_i,x_i)$ and $g(x_i)$ by fitting the $\hat{Q}$ and $\hat{g}$ models on the other $k-1$ folds.

Average treatment effect on the treated The same logic provides an estimate for the bias of the average treatment effect on the treated (ATT) under the sensitivity model. Namely, the expectation over X in Theorem 2.3 is conditioned on $T=1$. In practice, the bias can be estimated by taking the mean over only treated units. Note that the re-parameterization calculation does not change.

3 Calibration using observed data

The analyst must make judgments about the influence a hypothetical unobserved confounder might have on treatment assignment and outcome. To calibrate such judgments, we'd like to have a reference point for how much the observed covariates influence the treatment assignment and outcome. In the sensitivity model, the degree of influence is measured by

$R^2_{Y,\text{par}}$ and α. We want to measure the degree of influence of an observed covariate Z given the other observed covariates $X \backslash Z$.

For the outcome, this can be measured as:

$$R^2_{Y,X\backslash Z} := \frac{\mathbb{E}(Y - \mathbb{E}[Y \mid T, X\backslash Z])^2 - \mathbb{E}(Y - Q(T,X))^2}{\mathbb{E}(Y - \mathbb{E}[Y \mid T, X\backslash Z])^2}.$$

That is, the fraction of variation explained by Z not been previously explained by $X \backslash Z$. Computing this expression requires an estimate for $\mathbb{E}[Y \mid T, X\backslash Z]$. In practice, it is simplest to estimate the quantity by fitting a new regression model that predicts Y from T and $X \backslash Z$. Calling the fitted model $\hat{Q}_{X\backslash Z}$, we estimate

$$\hat{R}^2_{Y,X\backslash Z} = \frac{\frac{1}{n}\sum(y_i - \hat{Q}_{X\backslash Z}(t_i, x_i \backslash z_i))^2 - \frac{1}{n}\sum(y_i - \hat{Q}(t_i, x_i))^2}{\frac{1}{n}\sum(y_i - \hat{Q}_{X\backslash Z}(t_i, x_i \backslash z_i))^2}.$$

We now turn to the analogous estimate for the influence on treatment assignment. With α in the form of (1.2), it is not obvious how to compute this quantity. Accordingly, we re-express α in a more convenient form:

Theorem 3.1. *Under our sensitivity model, Assumption 1,*

$$\alpha = 1 - \frac{\mathbb{E}[\tilde{g}(X,U)(1 - \tilde{g}(X,U))]}{\mathbb{E}[g(X)(1 - g(X))]}.$$

We defer the proof to the appendix.

Then, to measure the influence of observed covariate Z on treatment assignment given $X \backslash Z$ we define $g_{X\backslash Z}(X \backslash Z) = \mathrm{P}(T = 1 \mid X \backslash Z)$ and

$$\alpha_{X\backslash Z} := 1 - \frac{\mathbb{E}[g(1-g)]}{\mathbb{E}[g_{X\backslash Z}(1 - g_{X\backslash Z})]}.$$

We then estimate this quantity by fitting a model for $g_{X\backslash Z}$ by predicting T from $X \backslash Z$ and then estimating

$$\hat{\alpha}_{X\backslash Z} = 1 - \frac{\frac{1}{n}\sum_i \hat{g}(x_i)(1 - \hat{g}(x_i))}{\frac{1}{n}\sum_i \hat{g}_{X\backslash Z}(x_i \backslash z_i)(1 - \hat{g}_{X\backslash Z}(x_i \backslash z_i))}.$$

Average treatment effect on the treated Reference values for the ATT can be computed in exactly the same way—i.e., it is not required to restrict the expectations to only the treated units. This is because the bias expression is given in terms of 'full data' α and $R^2_{Y,\text{par}}$.

Grouping covariates The estimated values $\hat{\alpha}_{X\backslash Z}$ and $\hat{R}^2_{Y,X\backslash Z}$ measure the influence of Z conditioned on all the other confounders. In some cases, this can be misleading. For example, if some piece of information is important but there are multiple covariates providing redundant measurements, then the estimated influence of each covariate will be small. To avoid this, we suggest grouping together related or strongly dependent covariates and computing the influence of the entire group in aggregate. For example, grouping income, location, and race together as 'socioeconomic variables'.

4 Examples

We now examine several examples of Austen plots for sensitivity analysis, showing: (1) We preserve the qualitative usefulness of Imbens' approach, without any modeling restrictions.

(2) Austen plots can aid judgments in the reliability of conclusions from real observational studies. (3) It matters what model is used for Q and g—so the ability to validly handle arbitrary models is important. (4) The bias estimates from the model tend to be conservative.[2]

Imbens' analysis To demonstrate the use of Austen plots, we follow Imbens [Imb03] and produce sensitivity plots for variations on the LaLonde job training data [LaL86]. We find that our conclusions about the effects of unobserved confounding are substantively the same as Imbens [Imb03]. That is, we arrive at sensible sensitivity conclusions while liberating ourselves from the need for parametric assumptions on the observed data.

The original purpose of the LaLonde job training data was to analyze the effect of a job training program on the annual earnings of a participant. The data consists of both an experimental (randomly assigned) part, and an observational sample from the Panel Study of Income Dynamics (PSID). Following Imbens [Imb03], we test on (1) the experimental sample, (2) the experimental treated with observational controls, (3) the same as 2, except with outcome defined as change in earnings since 1974, and (4) the same as 2, except individuals with high-earnings pre-treatment are dropped. We adjust for the same covariate set as Imbens: married, age, education, race, and earnings in 1974 and 1975. There are large differences in these background covariates between the experimental sample and the PSID controls—this is what makes causal inference from the combined experimental and PSID samples challenging.

Deviating from Imbens, we fit random forests for $\hat{Q}$ and $\hat{g}$. This demonstrates the sensitivity analysis in the case where the observed data model does not have a simple parametric specification. In particular, we split the data into 10 folds and predict $\hat{Q}(x_i, t_i)$ and $\hat{g}(x_i)$ for each unit in a given fold using random forests trained on the other folds.

Austen plots for these analyses are displayed in Figure 2. The experimental sample (panel A) is robust to unobserved confounding: inducing a bias of $1000 would require an unobserved confounder with a much stronger effect than any of the measured covariates or earning variables. By contrast, the non-experimental samples (panels B and C) are much more sensitive to unobserved confounding. Several of the covariates, if unobserved, would suffice to bias the estimate by $1000. Note that the sensitivity curves are the same for both B and C, since the outcome is just a linear transformation. Finally, the restricted sample (panel D) is both significantly more robust to bias than the full non-experimental samples, and the influence of the observed covariates is much reduced. Imposing the restriction makes the experimental and non-experimental populations much more similar. Accordingly, we might conclude that estimates in the restricted sample are less prone to bias in the sense that it seems less likely that there is some unobserved confounder with sufficient influence to meaningfully bias the result.

Practical relevance Figure 3 shows Austen plots for two effects estimated from observational data. The first study is based on data from the Infant Health and Development Program (IHDP) [BG+92], an experiment targeted at low-birth-weight, premature infants that provided child care and home visits. The study here measures the effect on an IQ test of the level of participation in IHDP child development centers in the two years preceding this test [Hil11, §6.1]. Level of participation is not randomly assigned, so Hill [Hil11] estimates the effect by using Bayesian Additive Regression Trees (BART) [Chi+10] to control for a range of pre-treatment covariates.

The second plot corresponds to the estimate of the effect of combination blood pressure medications on diastolic blood pressure described in [Dor+16]. The data is derived from an American survey that includes a variety of socioeconomic and health factors. We again use BART to adjust for these.

[2]Demonstration code is available at github.com/anishazaveri/austen_plots

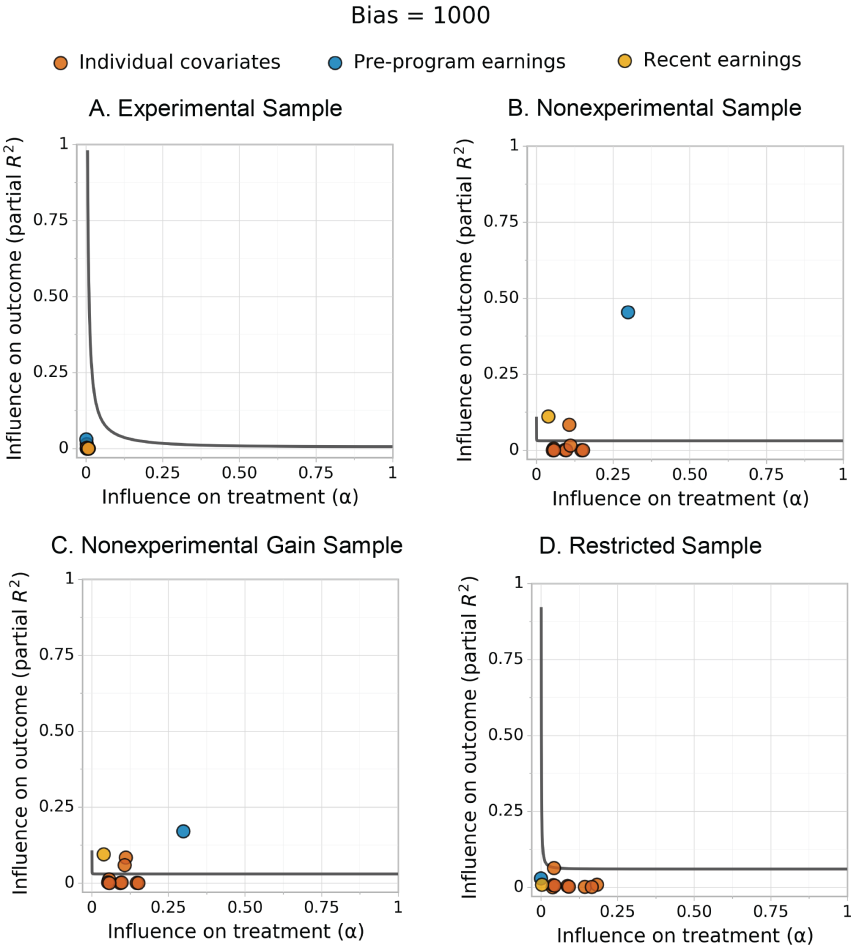

Figure 2: Austen plots preserve the qualitative conclusions of Imbens' analysis without imposing any restriction on the modeling of the observed data. In each plot, the black solid line indicates the partial R^2 and α values that would induce a bias of at least \$1000. Each plot also includes estimates for the strength of confounding for each of the nine covariates (red circles) as well as recent lag in earnings (RE75 and pos75, yellow circles), and the all pre-program earnings (RE74, pos74, RE75, pos75, green circles).

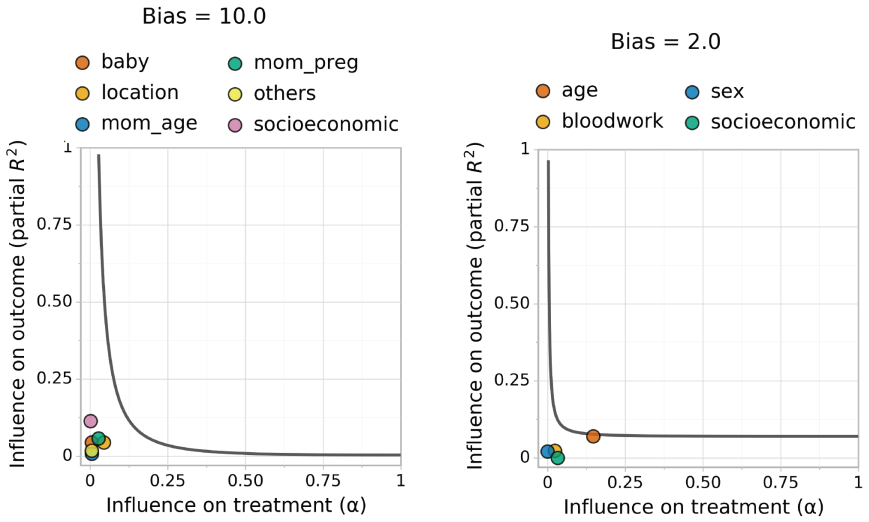

Figure 3: Austen plots are informative when applied to real data analysis. The left-hand plot is for the estimated effect of IHDP participation level on child IQ. The conclusions of this study seem robust to unobserved confounding—even the observed covariate groups do not have sufficient influence to undo the qualitative conclusion of the model. The right-hand plot is for the estimated effect of combination treatment on diastolic blood pressure. In this case, whether the study conclusions are reliable depends on whether an unobserved confounder as influential as age is credible—we should consult with an expert. In both cases, we model the outcome with Bayesian Additive Regression Trees, and the propensity score with logistic regression.

Table 1: The sensitivity model tends to be conservative in its bias estimates. Bias estimates for leaving out a confounding covariate are computed according to the sensitivity model (using the left-out covariate data) and by comparing non-parametric effect estimates from the full data (τ_x), and the left-out covariate data ($\tau_{x\setminus z}$). In all cases, the sensitivity model estimate is larger.

Study:	LaLonde Restricted	Blood Pressure	IHDP
Omitted covariate:	Education	Age	Socioeconomic
τ_x	2508.63	−2.33	12.72
$\tau_{x\setminus z}$	1982.54	−2.86	13.35
Nonparametric bias	526.09	0.53	−0.63
Sensitivity Model |bias|	986.90	1.91	0.75

The Austen plots are informative for these examples. In the first case, the Austen plot increases our confidence in the qualitative result. In the second case, it suggests we should be cautious about the conclusions unless unobserved confounders as strong as age are deemed unlikely.

Model specification We might hope that fully parametric approaches are 'good enough' in the sense that they yield sensitivity analyses with conclusions that are substantially the same irrespective of modeling details. Figure 4 shows this is not so. The conclusions of the sensitivity analysis depend on the observed data model, so we require a sensitivity model approach that can handle the model that best fits the observed data. Austen plots accomplish this by allowing arbitrary models for the observed data.

Sensitivity model conservatism The bias curves and influence estimates in Austen plots are contingent on the assumed sensitivity model, Assumption 1. Any sensitivity analysis must be predicated on some assumption about the influence of the unobserved confounder. We motivated our particular choice by simplicity and tractability. We also expect that our associated sensitivity model will often yield conservative values for bias; i.e., the bias

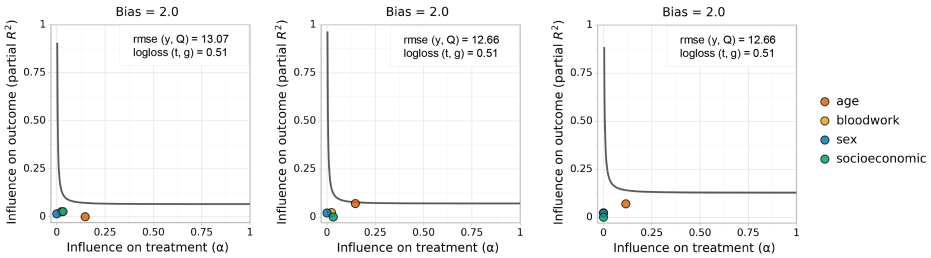

Figure 4: The qualitative conclusions of the sensitivity analysis depend on how the observed data is modeled. All three Austen plots are for estimating the effect of combination medications on blood pressure [Dor+16]. Both the sensitivity curves and the estimated influence of observed variables are affected by the modeling choice. The leftmost panel uses linear regression and logistic regression for the outcome and treatment models, respectively. The middle panel uses Bayesian Additive Regression Trees and logistic regression. The rightmost panel uses Bayesian Additive Regression Trees and a random forest predictor.

anticipated by the sensitivity model is higher than the true bias induced by the real, physical, mechanism of confounding. The reason is that we assume monotonically increasing bias in both treatment and outcome influence. In reality, hidden confounders can have more complicated relationships that 'cancel out'. For example, the effect of age in the blood pressure example might be: blood pressure increases with age, but young patients don't take their medication (preferring diet and exercise), middle age patients take it at a base rate, and old patients don't take the medication (fatalism). In this case, the young people cause the drug to look less effective (non-takers with low blood pressure), and old people cause the drug to look more effective (non-takers with high blood pressure). These effects cancel out somewhat, reducing the bias induced by failing to adjust for age. Assumption 1 does not allow for such cancellations.

To test conservatism, we create deliberately confounded datasets by removing an observed confounder from our baseline data. We compute the bias anticipated by our model, $\texttt{bias}(R^2_{Y,X\setminus Z}, \alpha_{X\setminus Z})$, using the measured influence strength of the covariate. We compute a non-parametric estimate of the bias by estimating the effect with the full data, estimating the effect with the deliberately confounded data, and taking the difference. The results are shown in table 1, and confirm the conservatism-in-practice. This increases our confidence that when an Austen plot suggests robustness to unobserved confounding we do indeed have such robustness.

5 Related Work

There are a wealth of approaches to sensitivity analysis. Informally, these divide into latent-confounder approaches in the style of Imbens', which are relatively easy to interpret, and alternative approaches that make weaker assumptions at the price of interpretability. Austen plots have the advantage that they preserve the key elements of the latent-confounder approach while also allowing arbitrary data modeling approaches. However, the assumptions and subjective judgments required for each analysis are quite different, and alternative approaches may be easier in some scenarios. In this sense, the zoo of sensitivity analysis methods are complimentary to each other and to the Austen plot approach.

The most closely related approaches to ours are sensitivity analysis based on parametric models in the style of Imbens [Imb03]. These typically assume some relatively simple parametric latent variable model, where the latent variable is the unobserved confounder, while also relaxing Imbens' fully-specified approach. Dorie et al. [Dor+16] extends an Imbens-like approach to accommodate BART as the outcome model by embedding model fitting in a (computationally intensive) simulation scheme. Zhang and Tchetgen Tchetgen

[ZT19] allow the outcome to be modeled by a generalized linear model, and also remove any assumption about the distribution of the unobserved confounder. Cinelli and Hazlett [CH20] allow for arbitrary kinds of confounders and propensity score models, but require that the outcome is modeled with linear regression. Cinelli et al. [Cin+19] make concrete assumptions about the edges of a causal DAG and use causal identification tools to translate those assumptions into effect (hence, bias) estimates. However, they assume that all relationships in the DAG are linear. Rosenbaum and Rubin [RR83a] assume a categorical covariate and a binary confounder. They don't impose any explicit additional constraints on the propensity score or outcome model, but their general approach requires 4 sensitivity parameters for each level of the observed covariate; making the sensitivity analysis practical requires further assumptions.

A different line of work relaxes parametric assumptions at the price of requiring the analyst to make judgments about more abstract sensitivity parameters. For example, Franks et al. [Fra+19] assume $\operatorname{logit} \mathrm{P}(T \mid X, Y^{(t)}) = \alpha_t(X) + \gamma_t Y^{(t)}$, where $Y^{(t)}$ are potential outcomes, and require the analyst to specify the sensitivity parameters γ_0, γ_1. That is, the analyst needs to assess the strength of the relationship between each of the potential outcomes and the treatment assignment. Shen et al. [She+11] present a method based on inverse propensity weighted estimators that is largely free of parametric assumptions. In their case, the sensitivity parameters are $\operatorname{var} \tilde{g}/g$ and the correlation between $\tilde{g}/g$ and the counterfactual outcomes. VanderWeele and Arah [VA11] derive very general bias formulas under the assumption of categorical covariates and categorical confounder. Their general analysis uses 4 sensitivity parameters per covariate level. They also offer a simplified version under stronger assumptions, where, in particular, a sensitivity parameter is the difference $\mathrm{P}(U = 1 \mid T = 1, x) - \mathrm{P}(U = 1 \mid T = 0, x)$—i.e., the analyst must make a judgment on the confounder-treatment strength in the anti-causal direction. Ding and VanderWeele [DV15] bound bias in estimate of the relative risk for a binary outcome allowing for a categorical confounder. The main strength of their approach is that the bound is essentially non-parametric. However, their treatment-confounder sensitivity parameter is the relative risk of the unobserved confounding given treatment, which is again in the anti-causal direction. In a different direction, Bonvini and Kennedy [BK19] propose a sensitivity model where most units are presumed to be unconfounded, but a portion may be arbitrarily confounded. The proportion of confounded units serves as the sensitivity parameter. Depending on the application, the subjective judgments required by these models may be harder or easier to assess than the confounding strength judgment required for Austen plots.

Another line of work assumes a model for the propensity score that's logit-linear in an unobserved confounder $U \in_{\mathrm{R}} [-1, 1]$, assumes an a priori bound on the coefficient of U, and derives the range of treatment effects consistent with this bound [e.g., Ros10; Zha+19; Yad+18]. See Yadlowsky et al. [Yad+18] for a particularly clear exposition. This approach has the advantage that it can avoid any parametric assumptions on the observed data, and avoid any modeling assumption relating U and Y. However, interpretation can be difficult relative to the parametric-model approaches; in particular, choosing the a priori bound can be challenging.

References

[AI19] S. Athey and G. W. Imbens. "Machine learning methods that economists should know about". In: *Annual Review of Economics* 1 (2019).

[BK19] M. Bonvini and E. H. Kennedy. "Sensitivity Analysis via the Proportion of Unmeasured Confounding". In: *arXiv e-prints* (2019).

[BG+92] J. Brooks-Gunn, F. ruey Liaw, and P. K. Klebanov. “Effects of early intervention on cognitive function of low birth weight preterm infants”. In: *The Journal of Pediatrics* 3 (1992).

[Chi+10] H. A. Chipman, E. I. George, and R. E. McCulloch. “Bart: bayesian additive regression trees”. In: *Ann. Appl. Stat.* 1 (2010).

[CH20] C. Cinelli and C. Hazlett. “Making sense of sensitivity: extending omitted variable bias”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1 (2020).

[Cin+19] C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. “Sensitivity analysis of linear structural causal models”. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019.

[DV15] P. Ding and T. J. VanderWeele. “Sensitivity analysis without assumptions.” In: *Epidemiology* (2015).

[Dor+16] V. Dorie, M. Harada, N. B. Carnegie, and J. Hill. “A flexible, interpretable framework for assessing sensitivity to unmeasured confounding”. In: *Statistics in Medicine* 20 (2016).

[Dor+19] V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. “Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition”. In: *Statist. Sci.* 1 (2019).

[Fra+19] A. M. Franks, A. D’Amour, and A. Feller. “Flexible sensitivity analysis for observational studies without observable implications”. In: *Journal of the American Statistical Association* 0 (2019).

[Hil11] J. L. Hill. “Bayesian nonparametric modeling for causal inference”. In: *Journal of Computational and Graphical Statistics* 1 (2011).

[HB19] P. Hünermund and E. Bareinboim. “Causal Inference and Data-Fusion in Econometrics”. In: *arXiv e-prints* (2019).

[Imb03] G. Imbens. “Sensitivity to exogeneity assumptions in program evaluation”. In: *The American Economic Review* (2003).

[LaL86] R. J. LaLonde. “Evaluating the econometric evaluations of training programs with experimental data”. In: *The American Economic Review* 4 (1986).

[RR83a] P. R. Rosenbaum and D. B. Rubin. “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 2 (1983).

[Ros10] P. Rosenbaum. *Design of Observational Studies*. 2010.

[RR83b] P. R. Rosenbaum and D. B. Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 1 (1983).

[She+11] C. Shen, X. Li, L. Li, and M. C. Were. “Sensitivity analysis for causal inference using inverse probability weighting”. In: *Biometrical Journal* 5 (2011).

[VA11] T. J. VanderWeele and O. A. Arah. “Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders”. In: *Epidemiology* 1 (2011).

[Yad+18] S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. “Bounds on the conditional and average treatment effect with unobserved confounding factors”. In: *arXiv e-prints* (2018).

[ZT19] B. Zhang and E. J. Tchetgen Tchetgen. “A Semiparametric Approach to Model-based Sensitivity Analysis in Observational Studies”. In: *arXiv e-prints* (2019).

[Zha+19] Q. Zhao, D. S. Small, and B. B. Bhattacharya. “Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 4 (2019).

A Proofs

Theorem 2.4. *Under our sensitivity model, Assumption 1, the outcome influence is*

$$R^2_{Y,\text{par}}(\alpha,\delta) = \delta^2 \sum_{t=0}^{1} \frac{\mathbb{E}\big[\psi_1\big(g(X)^t(1-g(X))^{1-t}(1/\alpha - 1) + 1[T=t]\big)\big]}{\mathbb{E}[(Y-Q(T,X))^2]},$$

where ψ_1 is the trigamma function.

Proof. First, we write:

$$\begin{aligned}
\mathbb{E}(Y - \mathbb{E}[Y \mid T,X,U])^2 &= \mathbb{E}(Y-Q(T,X))^2 \\
&\quad - 2\delta\mathbb{E}[(Y-Q(T,X))(\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])] \\
&\quad + \delta^2\mathbb{E}((\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T]))^2 \\
&= \mathbb{E}(Y-Q(T,X))^2 - \delta^2\mathbb{E}[\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T)]. \qquad \text{(A.1)}
\end{aligned}$$

Where we've used,

$$\begin{aligned}
&\mathbb{E}[(Y-Q(T,X))(\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])] \\
&= \mathbb{E}[\mathbb{E}[(Y-Q(T,X)) \mid T,X,U](\operatorname{logit}\tilde{g}(X,U) - \mathbb{E}[\operatorname{logit}\tilde{g}(X,U) \mid X,T])]]
\end{aligned}$$

and other standard conditional expectation manipulations.

The usefulness of (A.1) is that $\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T)$ has an analytic expression. Namely, by Beta-Bernoulli conjugacy, this is the variance of a logit-transformed Beta distribution. The analytic expression for this variance is,

$$\operatorname{var}(\operatorname{logit}\tilde{g}(X,U) \mid X,T) = \psi_1\big(g(X)(1/\alpha - 1) + T\big) + \psi_1\big((1-g(X))(1/\alpha - 1) + 1 - T\big),$$

where ψ_1 is the trigamma function. The claimed result follows by plugging in this expression into (A.1).

□

Theorem 3.1. *Under our sensitivity model, Assumption 1,*

$$\alpha = 1 - \frac{\mathbb{E}[\tilde{g}(X,U)(1-\tilde{g}(X,U))]}{\mathbb{E}[g(X)(1-g(X))]}.$$

Proof. The key insight is:

$$\begin{aligned}
\operatorname{var}(\tilde{g}) &= \mathbb{E}[\operatorname{var}(\tilde{g} \mid g)] + \operatorname{var}(\mathbb{E}[\tilde{g} \mid g]) \\
&= \mathbb{E}[\alpha g(1-g)] + \operatorname{var}(g),
\end{aligned}$$

where the first line is the law of total variance, and the second line uses the assumed Beta distribution of $\tilde{g} \mid g$. Accordingly,

$$\alpha = \frac{\operatorname{var}(\tilde{g}) - \operatorname{var}(g)}{\mathbb{E}[g(1-g)]}.$$

Now, observe that by the law of total variance,

$$\begin{aligned}
\operatorname{var}(T) &= \mathbb{E}[\operatorname{var}(T \mid g)] + \operatorname{var}(\mathbb{E}[T \mid g]) \\
&= \mathbb{E}[g(1-g)] + \operatorname{var}(g),
\end{aligned}$$

where we have used that $T \mid g$ is Bernoulli. By the same logic,

$$\operatorname{var}(T) = \mathbb{E}[\tilde{g}(1-\tilde{g})] + \operatorname{var}(\tilde{g}).$$

Whence,

$$\operatorname{var}(\tilde{g}) - \operatorname{var}(g) = \mathbb{E}[g(1-g)] - \mathbb{E}[\tilde{g}(1-\tilde{g})].$$

The result follows immediately. □