Project in Statistics
Semiparametric Models and Efficient Estimation
Part 3

Exam number # 11

June 17, 2022

Contents

1 Introduction

In this project, we will consider estimating the causal effect of a binary treatment R on a binary outcome Y, where the effect is quantified by the marginal log odds ratio. The treatment R is randomized and independent of additional baseline covariates X. To give an introduction to what we intended to solve in this report, we first establish the problems arising in the first two reports written.
In the first report we investigated two models for this setup - a linear model,

$$\mathbb{E}[Y \mid R, X] = \alpha + \beta R + \gamma X$$

and a binary logistic model

$$P(Y \mid R, X) = \text{expit}(\alpha + \beta R + \gamma X)$$

where the parameter of interest is β. In the linear model we found the marginal and the conditional treatment effect to coincide with β, while in the logistic model, the marginal treatment effect is different from the conditional treatment effect. This highlights, that the binary logistic model is non-collapsible, whereas the linear model is. The consequence is, that we cannot use other baseline covariates X in a logistic regression to gain efficiency of our estimates, if we want to estimate the marginal treatment effect in the full model containing X.
In the second report, we derived an efficient estimator of the treatment effect, without including baseline covariates and implemented the one-step estimator to solve the efficient score function. In this report we will derive the efficient semi-parametric estimator of the marginal log odds ratio (marginal treatment effect) in the full model with baseline covariates. That is, we will solve the problem of non-collapsibility of the logistic model. More specifically, we derive the efficient semi-parametric estimator and run a comparative simulation study and apply our efficient estimator on real data from an RCT, which studies the effect on six-year survival of liver disease patients of a treatment drug.

2 Deriving the efficient influence functions

2.1 The naive glm estimator

Consider the setup from the introduction, with $p_r = P(Y = 1 \mid R = r)$ for $r = 0, 1$. The target parameter of this problem is

$$\beta = g\left(p_0, p_1\right) \equiv \log \frac{p_1 / \left(1 - p_1\right)}{p_0 / \left(1 - p_0\right)}$$

This is equivalent to the restricted moment model $\mathbb{E}[Y \mid R] = \text{expit}(\alpha + R\beta)$. Simple estimators of p_0 and p_1 are the treatment specific sample averages,

$$\hat{p}_0 = \frac{\sum_{i=1}^n Y_i(1 - R_i)}{\sum_{i=1}^n (1 - R_i)} \ , \quad \hat{p}_1 = \frac{\sum_{i=1}^n Y_i R_i}{\sum_{i=1}^n R_i}.$$

A first estimator of β is

$$\hat{\beta}_n = g\left(\hat{p}_0, \hat{p}_1\right)$$

Using a first order Taylor expansion of $\hat{\beta}_n$ we get

$$\hat{\beta}_n = g\left(\hat{p}_0, \hat{p}_1\right) = g\left(p_0, p_1\right) + \dot{g}_0\left(p_0, p_1\right)\left(\hat{p}_0 - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\left(\hat{p}_1 - p_1\right) + \hat{R}$$

where

$$\begin{aligned}\hat{R} &= \ddot{g}_0(p_{*0}, p_{*1})(\hat{p}_0 - p_0)^2 + \ddot{g}_1(p_{*0}, p_{*1})(\hat{p}_1 - p_1)^2 \\ &= \frac{1 - 2p_{*0}}{p_{*0}^2(1 - p_{*0}^2)^2}(\hat{p}_0 - p_0)^2 + \frac{2p_{*1} - 1}{p_{*1}^2(1 - p_{*1}^2)^2}(\hat{p}_1 - p_1)^2\end{aligned}$$

where (p_{*0}, p_{*1}) denotes an intermediate value between (p_0, p_1) and $(\hat{p}_0, \hat{p}_1)$.
We want to control the error term and prove, that $\sqrt{n}\hat{R} \to 0$ in probability to establish the influence function for this estimator. We prove the convergence statement for each term in $\hat{R}$. To do this, we need consistency of $\hat{p_0}$, i.e.,

$$\hat{p_0} - p_0 \xrightarrow{P} 0$$

to conclude

$$\frac{1-2p_{*0}}{p_{*0}^2(1-p_{*0}^2)^2} \xrightarrow{P} \frac{1-2p_0}{p_0^2(1-p_0^2)^2} =: C.$$

And as $\hat{p_0}$ is the MLE, we get asymptotic normality, i.e.,

$$\sqrt{n}(\hat{p_0} - p_0) \xrightarrow{D} \mathcal{N}(0, \mathbb{E}[\phi_0^2])$$

Where ϕ_0 is the influence function for $\hat{p}_0$, which is explained below. Collecting these statements finally gives us, that

$$\underbrace{\frac{1-2p_{*0}}{p_{*0}^2(1-p_{*0}^2)^2}}_{\xrightarrow{P} C}(\hat{p}_0 - p_0)\underbrace{\sqrt{n}(\hat{p}_0 - p_0)}_{\xrightarrow{D}\mathcal{N}(0,\mathbb{E}[\phi_0^2])} \xrightarrow{P} 0$$

Where the overall convergence to 0 in probability follows by Slutsky. The same method is used for the other second term. We are now ready to find our influence function, as we expand $\sqrt{n}\left(\hat{\beta}_n - \beta\right)$ to get

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n - \beta\right) &= \sqrt{n}\left(\dot{g}_0\left(p_0, p_1\right)\left(\hat{p}_0 - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\left(\hat{p}_1 - p_1\right)\right) \\
&= \sqrt{n}\left(\dot{g}_0\left(p_0, p_1\right)\left(\frac{\sum_{i=1}^n Y_i(1-R_i)}{\sum_{i=1}^n (1-R_i)} - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\left(\frac{\sum_{i=1}^n Y_i R_i}{\sum_{i=1}^n R_i} - p_1\right)\right) + \sqrt{n}\hat{R} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{g}_0\left(p_0, p_1\right)\frac{(1-R_i)}{(1-\delta)}\left(Y_i - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\frac{R_i}{\delta}\left(Y_i - p_1\right) + o_p(1)
\end{aligned}$$

We can directly see our influnce function ϕ to be

$$\phi(Z) = \dot{g}_0\left(p_0, p_1\right)\frac{(1-R)}{(1-\delta)}\left(Y - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\frac{R}{\delta}\left(Y - p_1\right).$$

We also find the influence functions for p_0 and p_1. These are derived, as

$$\sqrt{n}\left(\hat{p}_0 - p_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{(1-R_i)}{(1-\delta)}\left(Y_i - p_0\right) + o_p(1)$$

and

$$\sqrt{n}\left(\hat{p}_1 - p_1\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{R_i}{\delta}\left(Y_i - p_1\right) + o_p(1)$$

such that $\phi_0(Z) = \frac{(1-R)}{(1-\delta)}\left(Y - p_0\right)$ and $\phi_1(Z) = \frac{R}{\delta}\left(Y - p_1\right)$. From this, the influence function for β is a weighted influence function of p_0 and p_1,

$$\phi(Z) = \dot{g}_0\left(p_0, p_1\right)\phi_0(Z) + \dot{g}_1\left(p_0, p_1\right)\phi_1(Z).$$

We could just as well chose to focus on estimating p_0 and p_1, but our parameter of interest is β. Now we have an initial influence function to use, when we want to find the efficient influence function. However, we still need to find the tangent space of our model.

2.2 The efficient influence function

If we put no restrictions of the distribution on Z we are in the non-parametric setting and by theorem 4.5 in Tsiatis, we can write the Hilbert space of zero-mean functions as

$$\mathcal{H} = \mathcal{T}_1 \oplus \mathcal{T}_2 \oplus \mathcal{T}_3$$

where

$$\begin{aligned}
\mathcal{T}_1 &= \{\alpha_1(X) : \mathbb{E}\left[\alpha_1(X)\right] = 0\} \\
\mathcal{T}_2 &= \{\alpha_2(R, X) : \mathbb{E}\left[\alpha_2(R, X) \mid X\right] = 0\} \\
\mathcal{T}_3 &= \{\alpha_3(Y, R, X) : \mathbb{E}\left[\alpha_3(Y, R, X) \mid R, X\right] = 0\}
\end{aligned}$$

However, we assume independence between X and the treatment assignment R, from which we get a semi-parametric model, as the distribution

$$p_{R \mid X}(r \mid x) = p_R(r) = \delta^r(1-\delta)^r$$

is completely specified by the treatment assignment being binary and δ known. We are now ready to find the space of all influence functions, by finding the tangent space $\mathcal{T}$. As δ is assumed known, the score of this part of the likelihood is 0, and so $\mathcal{T}_2$ gives no contribution to $\mathcal{H}$. Thus

$$\mathcal{T}^\perp = \mathcal{T}_2$$

Using Theorem 4.5 we can represent $\mathcal{T}_2$ as

$$\{h(R, X) - \mathbb{E}[h(R, X) \mid X]\}$$

for all square integrable functions $h(\cdot)$ of R, X. From R being binary we can write

$$h(R, X) = Rh_1(X) + h_2(X)$$

and as $R \perp\!\!\!\perp X$,

$$h(R, X) - \mathbb{E}[h(R, X) \mid X] = Rh_1(X) + h_2(X) - \mathbb{E}[R \mid X]h_1(X) - h_2(X) = (R - \delta)h_1(X)$$

where h_1 and h_2 are functions of X. Therefore

$$\mathcal{T}^\perp = \{(R - \delta)h_*(X) \ : \ h_*(X) \text{ arbitrary}\}$$

Thus the space of all influence functions are given by

$$\left\{\dot{g}_0\left(p_0, p_1\right)\frac{(1-R)}{(1-\delta)}\left(Y - p_0\right) + \dot{g}_1\left(p_0, p_1\right)\frac{R}{\delta}\left(Y - p_1\right) + (R - \delta)h_*(X) : h_*(X) \text{ arbitrary}\right\}$$

While the naive estimator with ϕ as its influence function does not depend on the baseline covariates X, the space of all influence function does contain functions, that do. We are now in a position to derive the efficient influence function by using theorem 4.3 from Tsiatis, as

$$\phi_{\text{eff}}(Z) = \phi(Z) + \Pi\left(\phi(Z) \mid \mathcal{T}^\perp\right).$$

Utilizing theorem 4.5 we have

$$\Pi\left(h(\cdot) \mid \mathcal{T}^\perp\right) = \Pi\left(h(\cdot) \mid \mathcal{T}_2\right) = \mathbb{E}(h(\cdot) \mid R, X) - \mathbb{E}(h(\cdot) \mid X)$$

Now we only need to compute $\Pi\left(\phi(Z) \mid \mathcal{T}^{\perp}\right)$,

$$\begin{aligned}\Pi\left(\phi(Z) \mid \mathcal{T}^{\perp}\right) &= \Pi\left(\dot{g}_0\left(p_0, p_1\right) \frac{(1-R)}{(1-\delta)}\left(Y-p_0\right)+\dot{g}_1\left(p_0, p_1\right) \frac{R}{\delta}\left(Y-p_1\right) \mid \mathcal{T}^{\perp}\right) \\ &= \dot{g}_0\left(p_0, p_1\right)\left[\mathbb{E}\left(\frac{(1-R)}{(1-\delta)}\left(Y-p_0\right) \mid R, X\right)-\mathbb{E}\left(\frac{(1-R)}{(1-\delta)}\left(Y-p_0\right) \mid X\right)\right] \\ &+ \dot{g}_1\left(p_0, p_1\right)\left[\mathbb{E}\left(\frac{R}{\delta}\left(Y-p_1\right) \mid R, X\right)-\mathbb{E}\left(\frac{R}{\delta}\left(Y-p_1\right) \mid X\right)\right]\end{aligned}$$

Considering each term, we can use the tower property and get,

$$\mathbb{E}\left(\frac{(1-R)}{(1-\delta)}\left(Y-p_0\right) \mid R, X\right)-\mathbb{E}\left(\frac{(1-R)}{(1-\delta)}\left(Y-p_0\right) \mid X\right)$$

Using that the term $(1-R)$ is 0 on $(R=1)$ we only need to condition on $R=0$ and move $(1-R)$ outside the expectation by independece.

$$\frac{(1-R)}{(1-\delta)} \mathbb{E}\left(\left(Y-p_0\right) \mid R=0, X\right)-\mathbb{E}\left(\frac{(1-R)}{(1-\delta)} \mathbb{E}\left(\left(Y-p_0\right) \mid R=0, X\right) \mid X\right)$$

By independence of R and X we get

$$\frac{(1-R)}{(1-\delta)} \mathbb{E}\left(\left(Y-p_0\right) \mid R=0, X\right)-\mathbb{E}\left(\left(Y-p_0\right) \mid R=0, X\right)$$

i.e.,

$$\frac{(\delta-R)}{(1-\delta)} \mathbb{E}\left(\left(Y-p_0\right) \mid R=0, X\right)$$

Using the same argumentation on the second term, we get

$$\begin{aligned}&\mathbb{E}\left(\frac{R}{\delta}\left(Y-p_1\right) \mid R, X\right)-\mathbb{E}\left(\frac{R}{\delta}\left(Y-p_1\right) \mid X\right) \\ &=\frac{R-\delta}{\delta} \mathbb{E}\left(\left(Y-p_1\right) \mid R=1, X\right)\end{aligned}$$

Thus

$$\begin{aligned}\phi_{\text{eff}}(Z) &= \dot{g}_1\left(p_0, p_1\right) \frac{R}{\delta}\left(Y-p_1\right)+\dot{g}_0\left(p_0, p_1\right) \frac{(1-R)}{(1-\delta)}\left(Y-p_0\right) \\ &- \dot{g}_1\left(p_0, p_1\right) \frac{R-\delta}{\delta}\left(\mathbb{E}(Y \mid R=1, X)-p_1\right)+\dot{g}_0\left(p_0, p_1\right) \frac{R-\delta}{1-\delta}\left(\mathbb{E}(Y \mid R=0, X)-p_0\right).\end{aligned}$$

2.3 A flexible estimator using a polynomial basis expansion.

Instead of finding the most efficient, we can find an estimator, that is more efficient than the naive estimator, by projecting down on a flexible subspace of $\mathcal{T}^{\perp}$, namely,

$$\tilde{\mathcal{T}}^{\perp}=\left\{(R-\delta) \theta^{\top} q(X): \theta \in \mathbb{R}^d\right\},$$

where $q(X) \in \mathbb{R}^d$ and $q(X)=\left(1, X, X^2, X^3\right)^{\top}$. The new influence function is

$$\tilde{\phi}(R, Y)=\phi(R, Y)-\Pi\left\{\phi(R, Y) \mid \tilde{\mathcal{T}}^{\perp}\right\}$$

The idea behind this estimator, denoted $\tilde{\beta}$, is, that we project down on a flexible space, which contains all taylor approximations up to the fourth order, and by the multivariate Pythagorean theorem 3.3 in Tsiatis this yields a more asymptotic efficient estimator than the naive glm $\hat{\beta}$, as

$$\begin{aligned}\text{Var}(\tilde{\phi}(R,Y)) &= \text{Var}\left(\phi(R,Y) - \Pi\left\{\phi(R,Y) \mid \tilde{\mathcal{T}}^{\perp}\right\}\right) \\ &= \text{Var}(\phi(R,Y)) - \text{Var}\left(\Pi\left\{\phi(R,Y) \mid \tilde{\mathcal{T}}^{\perp}\right\}\right) \\ &\leq \text{Var}(\phi(R,Y))\end{aligned}$$

To find an expression of for the projection term, we use example 1 in section 2.4 in Tsiatis and get the unique projection to be

$$\begin{aligned}\Pi\left\{\phi(R,Y) \mid \tilde{\mathcal{T}}^{\perp}\right\} &= \mathbb{E}[\phi(R,Y)(R-\delta)q(X)]\left(\mathbb{E}\left[(R-\delta)^2 q(X)q(X)^T\right]\right)^{-1}(R-\delta)q(X) \\ &= \mathbb{E}[\phi(R,Y)(R-\delta)q(X)]\frac{1}{\delta(1-\delta)}\left(\mathbb{E}\left[q(X)q(X)^T\right]\right)^{-1}(R-\delta)q(X) \\ &= (R-\delta)\theta_0^T q(X)\end{aligned}$$

where $\theta_0 = \frac{1}{\delta(1-\delta)}\mathbb{E}[\phi(R,Y)(R-\delta)q(X)]\left(\mathbb{E}\left[q(X)q(X)^T\right]\right)^{-1}$.

3 Estimation

3.1 Naive estimator

The naive estimator is, as written in the introduction, is found by plugging in estimates of p_0 and p_1 into g, i,e,

$$\hat{p}_0 = \frac{\sum_{i=1}^n Y_i(1-R_i)}{\sum_{i=1}^n (1-R_i)}\ , \quad \hat{p}_1 = \frac{\sum_{i=1}^n Y_i R_i}{\sum_{i=1}^n R_i}.$$

such that

$$\hat{\beta}_n = g\left(\hat{p}_0, \hat{p}_1\right)$$

This is the same estimate as we would get by fitting a logistic glm of the response on the treatment.
To estimate the asymptotic variance, we use the influence function and the LLN, to get the estimator,

$$\hat{\phi} = n^{-1}\sum_{i=1}^n \hat{\phi}\left(Z_i; \tilde{P}\right)^2.$$

of the asymptotic variance, where $\tilde{P}$ signifies, that we plug in estimates of p_0 and p_1 and δ.

3.2 Efficient estimator

To construct a one-step estimator of the efficient estimate, we use $\hat{\beta}$, by

$$\hat{\beta}_{eff} = \hat{\beta} + \frac{1}{n}\sum_{i=1}^n \phi_{eff}\left(Z_i; \tilde{P}\right)$$

However, the efficient estimator needs estimates of $\mathbb{E}[Y \mid R=1, X]$ and $\mathbb{E}[Y \mid R=0, X]$. These are estimated by fitting a logistic regression model Y on R and X and then predicting each term for each X. We also fit the efficient model using a mis-specified model for $\mathbb{E}[Y \mid R=1, X]$ and $\mathbb{E}[Y \mid R=0, X]$ using a normal regression model regressing Y on X and R, instead of the true DGP logistic regression.
We again use the influence function and the LLN, to get the estimator for the asymptotic variance,

$$\hat{\phi}_{eff} = n^{-1} \sum_{i=1}^{n} \phi_{eff} \left(Z_i; \tilde{P} \right)^2 .$$

3.3 Estimator using a polynomial basis expansion

We can use the naive estimate $\hat{\beta}$ to construct a one-step estimator

$$\tilde{\beta} = \hat{\beta} + \frac{1}{n} \sum_{i=1}^{n} \phi \left(Z_i; \hat{\theta}_0, \tilde{P} \right)$$

where we estimate θ_0 by

$$\hat{\theta}_0 = \frac{1}{\delta(1-\delta)} \left\{ \sum_{i=1}^{n} q\left(X_i\right) q\left(X_i\right)^\top \right\}^{-1} \sum_{i=1}^{n} \left(R_i - \delta\right) \phi\left(R_i, Y_i\right) q\left(X_i\right)$$

To estimate the variance we again use the influence function to get the estimator

$$\tilde{\phi} = n^{-1} \sum_{i=1}^{n} \tilde{\phi} \left(Z_i; \hat{\theta}_0, \widetilde{P} \right)^2$$

4 Simulation study results

We now want to asses the performance of the estimators presented above. We do this by simulating data from the following DGP.

$$P(Y=1 \mid R, X) = \text{expit}(-0.5 + 0.3R + \gamma X)$$

where R is binary with $P(R=1) = 0.5$ and X is $\mathcal{N}(0,1)$ distributed and independent of R. For each effect size $\gamma \in \{0, \text{-log(4)}, \text{-log(6)}\}$ and $n \in \{200, 400\}$ we simulate 20.000 datasets of size n and γ and estimate the marginal log odds. We also estimate the standard error of our effect estimates in each run and take the mean. Furthermore we compute the bootstrapped standard deviation. This is done for both the naive model, the efficient estimate, the polynomial estimate and the mis-specified efficient model. We also compute the true marginal means by using numerical integration. To find an expression of the marginal log odds, we use the tower property, as

$$p_R = P(Y=1 \mid R) = \mathbb{E}[P(Y=1 \mid R, X) \mid R] = \mathbb{E}[\text{expit}(\alpha + \beta R + \gamma X) \mid R]$$

The marginal log odds ratio becomes

$$\frac{\mathbb{E}[\text{expit}(\alpha + \beta + \gamma X)] / \mathbb{E}[h(\alpha + \beta + \gamma X)]}{\mathbb{E}[\text{expit}(\alpha + \gamma X)] / \mathbb{E}[h(\alpha + \gamma X)]}$$

where $h(x) = \frac{1}{1+\exp(x)} = 1 - \text{expit}(x)$. These expectations can be computed using numerical integration for known β and γ. The numerical integration is done using Julia and the package QuadGK.jl
The results are presented in the table below.

	n = 200			n = 400		
γ	0	-log(4)	-log(6)	0	-log(4)	-log(6)
True marginal effect	**0.300**	**0.220**	**0.194**	**0.300**	**0.220**	**0.194**
Naive						
mean $\hat{\beta}$	0.304	0.222	0.198	0.302	0.224	0.194
sd $\hat{\beta}$	0.295	0.288	0.289	0.204	0.203	0.204
mean(s.e($\hat{\beta}$))	0.290	0.288	0.287	0.204	0.203	0.202
Efficient						
mean $\hat{\beta}_{eff}$	0.305	0.219	0.194	0.302	0.221	0.196
sd $\hat{\beta}_{eff}$	0.294	0.248	0.233	0.207	0.174	0.163
mean(s.e($\hat{\beta}_{eff}$))	0.290	0.246	0.231	0.204	0.174	0.163
Polynomial expansion						
mean $\tilde{\beta}$	0.301	0.224	0.195	0.301	0.223	0.196
sd $\tilde{\beta}$	0.293	0.250	0.236	0.206	0.175	0.163
mean(s.e($\tilde{\beta}$))	0.288	0.246	0.231	0.204	0.174	0.163
Mis-specified efficient						
mean $\hat{\beta}_{mis}$	0.303	0.220	0.199	0.303	0.219	0.197
sd $\hat{\beta}_{mis}$	0.291	0.252	0.238	0.207	0.175	0.167
mean(s.e($\hat{\beta}_{mis}$))	0.290	0.248	0.235	0.204	0.175	0.165

The table shows, that we gain efficiency over the naive glm estimator, by using the efficient estimator, which includes baseline covariates. We also gain efficiency with the mis-specified and the polynomial basis model. The difference however, is that the misspecified model and the polynomial model have higher or similar estimated standard error and standard deviation, than the efficient. It is no surprise, that we almost get the same estimates from the polynomial model, as we project our influence down to a very flexible and large space and for the polynomial model we do not need to model $\mathbb{E}[Y \mid R = 1, X]$ and $\mathbb{E}[Y \mid R = 0, X]$, which is preferred, since we might mis-specify these. But as the study shows, it is difficult to mis-specify the conditional expectations enough, to get very different estimates. This might just be caused by the fact that the DGP is very simple and it is not unimaginable that this would change for more complex DGP. We also see, that for $n = 400$ we get much more stable estimates and lower standard error and sample standard deviations for all, but the naive estimator. Another simple observation is, that when there is no effect of X , i.e, $\gamma = 0$ we get the same estimates across all four methods.

5 Application to RCT data

We also apply our derived efficient estimator to a real world dataset. The data comes from a study on people with liver disease, which were randomly assigned treatment. The following covariates were recorded for each patient

- Y : response status at exit (0 : alive after six years, 1: death or liver transplantation before six years)
- R : randomized treatment allocation (0: placebo, 1: UDC)

- X : centered log bilirubin measured at baseline (log umol/L).

The parameter of interest is the marginal log odds ratio, which by the randomized setup is the causal effect of the treatment R on the status Y. We fit three models to estimate the marginal log odds. First we fit a fully parametrized logistic regression model using only the treatment indicator. Secondly we fit the efficient estimator using the one-step estimator described above. Finally we estimate the treatment effect using the polynomial basis estimator. The table below shows results for the estimation of the marginal log odds ratio.

		95 % CI	
	Marginal log OR estimate	lower	upper
Naive	-0.397	-0.869	**0.075**
Efficient	-0.653	-1.050	-0.256
Polynomial	-0.697	-1.089	-0.304

From the efficient estimator we estimate odds for getting a death or liver transplantation after six years is 0.520 (CI: $0.350, 0.774$) times lower for those, who received the treatment, compared to those who got the placebo using the efficient estimator. This estimate yields a significant effect on a 5% level of the treatment. From the polynomial basis estimator we get 0.498 (CI: $0.337, 0.738$) as our odds estimate. The important thing to note is, that if we include the baseline measurement of the log-bilirubin level, we get a larger effect size of the treatment as well as a significant estimate, compared to the fully parametrized logistic regression model, which yields a un-significant estimate of the odds ratio to be 0.672 (CI: $0.420, 1.08$). That including the covariate has an effect also gives preliminary evidence to the fact, that bilirubin level has a direct effect on the response.

6 Conclusion

In this report we investigated how to estimate the causal effect of a binary treatment R on a binary outcome Y, when we want to include information of relevant covariates X to gain efficiency. We did this by deriving the efficient influence function for a semi-parametric model, which allowed us to use relevant baseline covariates X in our estimation to achieve greater efficiency, compared to a standard fully parametrised logistic regression model. This solves the problem of non-collapsibility of the logistic model using the theory of semi-parametric estimators.

More specifically, we have derived the efficient estimator for the parameter of interest, as well as another flexible estimator using polynomial basis functions and investigated their performance on simulated data. The simulation study shows, that the efficient estimator correctly estimates the true marginal effect with a lower standard error, than the logistic regression model, when including baseline covariates. We also saw, that the polynomial estimator, yielded equally as efficient estimates as the most efficient, without the drawback of having to model conditional expectations, as in the efficient estimator. However, even though the conditional expectations in the efficient estimator is mis-specified, the estimator is still almost efficient as the efficient.

Finally we applied our estimator to a real world data set from an treatment-placebo RCT, where we estimate the effect of a drug on six-year survival. We found, that our efficient estimator estimates a significant effect of the drug on survival, while the standard logistic regression model does not.

Although the estimator developed in this report is able to solve the problem of non-collapsibility in the logistic setting, we still need the covariate X to have an effect on the treatment to be

able to get more efficient estimates. How to select the influential covariates is not touched upon in this report, but guidelines can be found in Hauck et. al.

7 Appendix

Reproducible code is attached as well as a README containing required packages to run the scripts.

References

[1] Van der Vaart, A. W. (2000). *Asymptotic statistics (Vol. 3)*. Cambridge university press.

[2] Tsiatis AA. (2006) *Semiparametric Theory and Missing Data.* Springer Verlag, New York

[3] Walter W. Hauck and Susan Anderson and Sue M. Marcus, (1998), *Should we adjust for covariates in nonlinear regression analyses of randomized trials?*, Controlled clinical trials, volume=19,3 pages 249 - 56