WEB APPENDIX 1.

Details on traffic density metric.

The roadway location and traffic data were obtained from Tele Atlas/Geographic Data Technology (GDT) Dynamap in 2005. The assignment of traffic count to links is straightforward for interstate freeways and other high-volume roads where count data are available for almost every link. On moderate and smaller roads, traffic count data are generally sparse, and imputation of link volumes is required. An extrapolation method based on roadway name, connectivity and distance was used to assign traffic count data to roadway links. Links were connected up to 5 km, 7 km, and 10 km from the traffic count locations for road Classes 4, 3, 2 and 1, respectively. Links with like names and within the specified distance were only assigned traffic count data when the links were connected. This extrapolation method produces consistent assignments of traffic volumes that have few gaps on the named roadways with count data; however, smaller local roads lacking count data are not included. Overall, with this method, volumes were assigned to 93% of Class 1 roads, 88% of Class 2 roads, 65% of Class 3 roads, and 7% of Class 4 roads. Local median volumes on class 1, 2, and 3 road were used for links not covered by the extrapolation model. Since the GDT traffic count data were mostly for 1995-2000 and the period of interest was 2000-2006, the counts were scaled up to represent 2003 traffic based on traffic based on county average vehicles-miles-traveled growth. Density plots were generated within a geographic information system using a linear decay function that approximates the fall-off of ambient concentrations with increasing distance away from roadways (*i.e.*, decays to background within approximately 300 meters). Traffic density maps were created using one parameterization for dispersion, in which density decreases by 90% at 300 m from the value at the edge of the roadway, which is consistent with data from numerous dispersion studies. GIS tools were used to extract the traffic densities from the map at the locations of the residences of the study population.

WEB APPENDIX 2.

Details on the calculation of the confidence intervals.

Based on assumptions related to the graph (25) (as well as sufficient experimentation in the exposure of interest in the target population), the parameter of the theoretical counterfactual distributions can be defined as a function of the observed data-generating distribution, say P_0. Thus, if we can consistently estimate the relevant components of P_0, then we can consistently estimate the so-called causal parameter of interest. Thus, the first goal is to estimate P_0 (such as the outcome regressed on the exposure and confounders) as nonparametrically (and thus with as little bias) as possible.

The goal of the T-MLE analysis is to augment the initial estimates of P_0 with a bias reduction step for the parameter of interest. The T in T-MLE is because the augmentation is specific to the parameter of interest, and so it “bends” the original estimate towards the goal, in our case estimating, at the population level, the predicted probability of term low birth weight had everyone been exposed to each quartile of traffic density.

The influence curve was used to derive standard errors for confidence intervals for the T-MLE and PIM estimates.

$SE(\psi_n) = \sqrt{\dfrac{\mathrm{var}_n[IC(Y_{i,}A_i,W_i;g_n,Q_n^1,\psi_n)]}{n}}$, where IC is the plug-in influence curve for this estimator, or:

$$IC(Y_{i,}A_i,W_i;g_n,Q_n^1,\psi_n) = Y_i - \left\{ \frac{I(A_i = a)}{g_n(a \mid W_i)} \left[Y_i - Q_n^1(a,W_i) \right] + Q_n^1(a,W_i) \right\} - \psi_n .$$

WEB APPENDIX 3.

R code for targeted maximum likelihood estimation and population intervention model estimates.

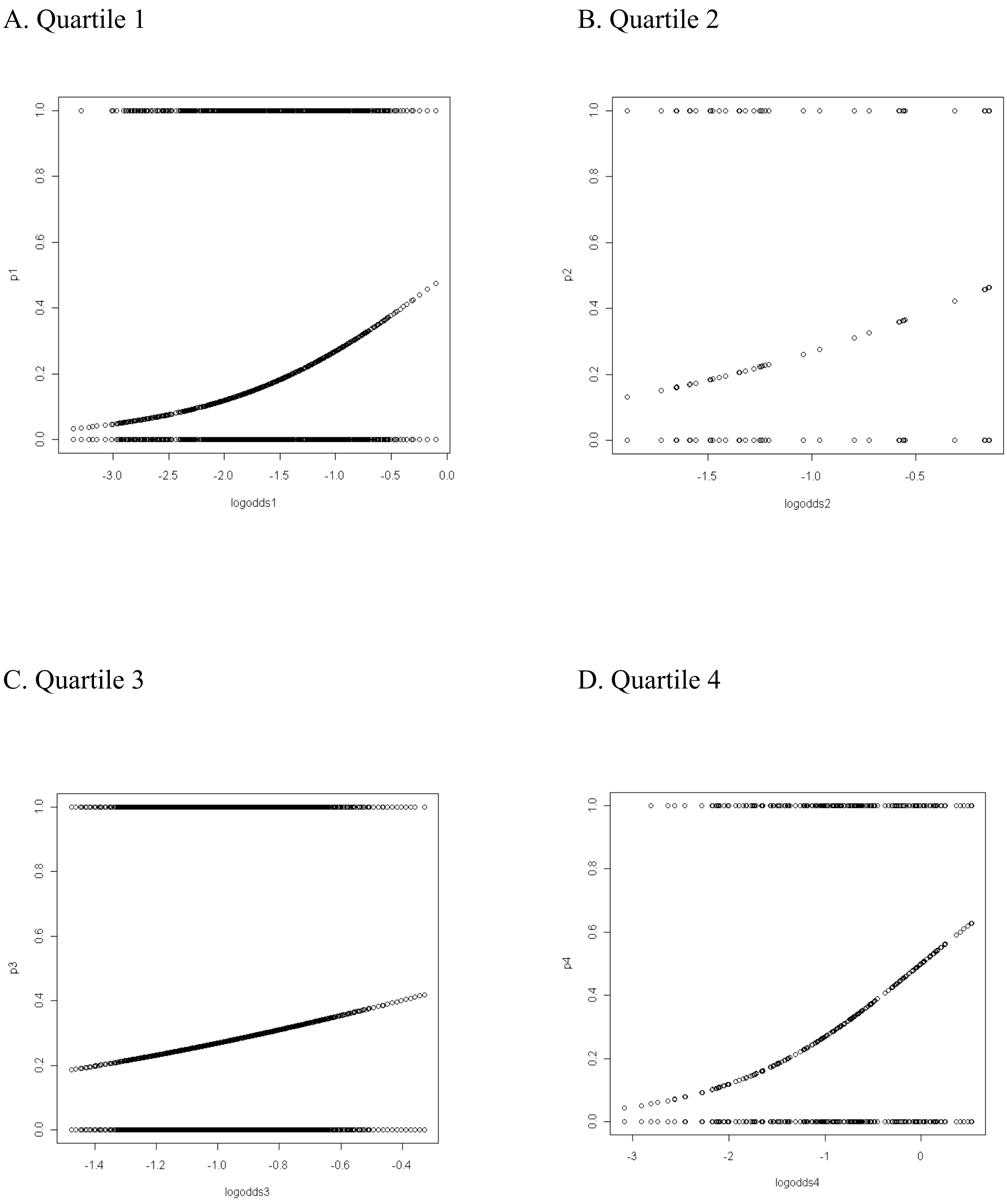
```
#------------------------------------------
# Program:       sage_tlbw_1.R
# Programmer:    Amy Padula
# Date Modified: 8-3-2011
# Description:
#
#        TMLE & PIM ESTIMATES
#        with D/S/A
#        traffic density --> tlbw
#------------------------------------------

# LOAD DSA PACKAGE
library(DSA)

# READ IN DATA ON CLUSTER COMPUTER
chaps1<- read.csv("chaps_full.csv",sep=",")
dim(chaps1)

## MASTER DATASET
keepdata0<-c("denq","denq1","denq2","denq3","denq4","ptb","prem","mom_ge35",
        "mom_le20","asian_mom","black_mom","hisp_mom","white_mom","other_mom
",
        "year","edum_cat","prem1","prem2","prem3","prem4","firstborn",
        "lowses","medi_cal","pren_care","fresno","kern","sanj","stan","tlbw")
data0<-as.data.frame(chaps1[,names(chaps1) %in% keepdata0])

names(data0)[names(data0)=="tlbw"]<-"y"
names(data0)[names(data0)=="denq1"]<-"a"

## CREATE DATASET FOR DSA  E[Y | A,W] - Q MODEL
cand.q <-c("y","a","mom_ge35","mom_le20","fresno","kern","sanj","stan",
        "asian_mom","black_mom","hisp_mom","white_mom","other_mom","year",
        "edum_cat","firstborn","lowses","medi_cal","pren_care")
data.q <- as.data.frame(data0[,names(data0) %in% cand.q])
#dim(data.q)
#head(data.q,20)

### E[Y | A,W] - Q MODEL
q.model <- DSA(y~a,family=binomial,data=data.q,maxsize=10,maxorderint=2,
        userseed=414,maxsumofpow=2,vfold=5,nsplits=10)
summary(q.model,family=binomial,data=data.q)
```

```
qaw<- predict(q.model,type="response",newdata=data0)

### G-COMP ESTIMATOR ON NEW DENSITY OF [Y | A,W]
### B0+B1(A=1)+B2W1+B3W2+...+EPSILON*H(A,W)
q1w<-predict(q.model,newdata=data.frame(a=1,y=data0[,"y"],
        mom_ge35=data0[,"mom_ge35"],mom_le20=data0[,"mom_le20"],
        fresno=data0[,"fresno"],kern=data0[,"kern"],sanj=data0[,"sanj"],
        stan=data0[,"stan"],asian_mom=data0[,"asian_mom"],
        black_mom=data0[,"black_mom"],hisp_mom=data0[,"hisp_mom"],
        white_mom=data0[,"white_mom"],other_mom=data0[,"other_mom"],
        year=data0[,"year"],edum_cat=data0[,"edum_cat"],
        firstborn=data0[,"firstborn"],lowses=data0[,"lowses"],
        medi_cal=data0[,"medi_cal"],pren_care=data0[,"pren_care"]))
mean.q1w<-mean(q1w)
print(mean.q1w)

### E(E(Y|A=1,W)-E(Y|A=0,W)) GCOMP ESTIMATE -> Y^1+COEF(Q*)
psi<-mean(1/(1+exp(-q1w)))
print(psi)

## CREATE DATASET FOR DSA  E[A | W] - G MODEL
cand.g <-c("a","mom_ge35","mom_le20","fresno","kern","sanj","stan",
        "asian_mom","black_mom","hisp_mom","white_mom","other_mom","year",
        "edum_cat","firstborn","lowses","medi_cal","pren_care")
data.g <- as.data.frame(data0[,names(data0) %in% cand.g])

### E[A | W] - G PART
g.model<- DSA(a~1,data=data.g,maxsize=10,maxorderint=2,userseed=414,
        maxsumofpow=2,family=binomial,vfold=5,nsplits=10)
summary(g.model,family=binomial,data=data.g)
gw<- predict(g.model,type="response",newdata=data0)
print(summary(gw))

#### H(A,W) - CLEVER COVARIATE
h<- ifelse(data0$a==1, (1/gw), (-1/(1-gw)))
print(summary(h))

### ONE STEP ESTIMATOR TO GET EPSILON [Y | A,W] - Q* MODEL
qs.model<-glm(data0$y~-1+offset(qaw)+h,family=binomial,data=data0)
summary(qs.model)
eps<-coefficients(qs.model)
print(summary(eps))

### E(E(Y|A=1,W)-E(Y|A=0,W)) TML ESTIMATE -> Y^1+COEF(Q*)*(H)
h.1<-as.vector(1/gw)
print(summary(h.1))
```

```
### ESTIMATES
#TMLE
tpsi<-mean(1/(1+exp(-(q1w+eps*h.1))))
print(tpsi)
#PIM
pim<-mean(data0$y-(1/(1+exp(-(q1w+eps*h.1)))))
print(pim)


### INFLUENCE CURVE
#TMLE
n<-length(data0$y)
tic<-((data0$y-qaw)*h+q1w-tpsi)
tvaric<-var(tic, na.rm=T)
tci.up<-tpsi+((1.96*tvaric)/sqrt(n))
tci.lo<-tpsi-((1.96*tvaric)/sqrt(n))
print(tci.up)
print(tci.lo)

#PIM
pic<-((data0$y-qaw)*h+q1w-pim)-data0$y-pim
pvaric<-var(pic, na.rm=T)
pci.up<-pim+((1.96*pvaric)/sqrt(n))
pci.lo<-pim-((1.96*pvaric)/sqrt(n))
print(pci.up)
print(pci.lo)

save.image(file="sage_tlbw_1.Rdata")
```

WEB FIGURE 1.

Plot of probability of treatment (exposure to traffic density) versus the log odds of treatment for each quartile of exposure.



A. Quartile 1

B. Quartile 2

C. Quartile 3

D. Quartile 4

WEB TABLE 1.

Characteristics of the SAGE population, San Joaquin Valley of California 2000-2006, by inclusion/exclusion in the final study population.

Covariates	Study population		Exclusions	
	N=237,031		N=31,434	
	N	%	N	%
Maternal age (years)				
<20	32,270	13.6	5,038	16.0
20-35	179,819	75.9	22,345	71.1
>35	24,942	10.5	4,051	12.9
Maternal race/ethnicity				
Asian	17,738	7.5	2,775	8.8
Black	11,560	4.9	2,239	7.1
Hispanic	132,605	55.9	18,132	57.7
White	71,522	30.2	7,808	24.8
Other	3,606	1.5	480	1.5
Maternal education				
No high school	28,027	11.8	3,894	12.4
Some high school	124,128	52.4	17,861	56.8
Some college	49,412	20.8	5,896	18.8
Bachelor’s or other degree	30,090	12.7	2,865	9.1
Missing	5,374	2.3	918	(2.9
Birth costs paid by Medi-Cal				
Yes	127,564	53.8	19,325	61.5
No	109,467	46.2	12,109	38.5
Low socioeconomic status[a]				
Yes	41,745	17.6	16,008	24.7
No	195,286	82.4	48,712	75.3
Parity				
0	83,819	35.4	10,452	33.2
>=1	153,212	64.6	20,982	66.8
Sex of infant				
Male	120,456	50.8	17,131	54.5
Female	116,575	49.2	14,303	45.5
Initiation of prenatal care				
1st trimester	192,905	81.4	23,883	76.0
2nd trimester	32,676	13.8	5,392	17.2
3rd trimester	7,317	3.2	1,041	3.3
Unknown	4,133	1.7	1,118	3.5
Year of birth				
2000	30,788	13.0	8,321	12.8
2001	31,707	13.4	8,191	12.6
2002	32,534	13.7	9,226	14.2
2003	33,082	14.0	10,281	15.8
2004	34,331	14.5	10,488	16.2

2005	35,567	15.0	8,758	13.5
2006	39,022	16.5	9,643	14.9
County of maternal residence				
Fresno	77,093	32.6	113,22	36.0
Kern	56,318	23.8	7,967	25.3
San Joaquin	59,680	25.2	6,941	22.1
Stanislaus	43,940	18.5	5,204	16.6

[a] Low socioeconomic status was defined as block group level unemployment >10%, income from public assistance >15% and families below poverty level >20% at the block group level from the 2000 census.