PS2

## Problem 3

#3a)

dtr = dtr[,-c(18,32)] #gets rid of "preterm" and "deadkids" covariates which aren't mentioned in the assignment   
fit1 = lm(dbirwt~., data = dtr)  
(summary(fit1))

##   
## Call:  
## lm(formula = dbirwt ~ ., data = dtr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1529.4 -310.9 -21.6 282.5 3344.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3159.03084 21.73122 145.368 < 2e-16 \*\*\*  
## alcohol -19.22708 9.79038 -1.964 0.049546 \*   
## anemia -6.87117 9.69125 -0.709 0.478320   
## cardiac -41.01329 13.16605 -3.115 0.001839 \*\*   
## chyper -61.49433 11.01403 -5.583 2.36e-08 \*\*\*  
## dfage 0.19812 0.22436 0.883 0.377225   
## dfeduc 2.67379 0.53585 4.990 6.05e-07 \*\*\*  
## diabete 75.40750 6.80698 11.078 < 2e-16 \*\*\*  
## disllb -0.14077 0.03939 -3.573 0.000352 \*\*\*  
## dlivord 24.35025 1.20809 20.156 < 2e-16 \*\*\*  
## dmage 4.66016 1.50225 3.102 0.001922 \*\*   
## dmar -27.87917 2.86945 -9.716 < 2e-16 \*\*\*  
## dmeduc 4.75905 0.59522 7.995 1.30e-15 \*\*\*  
## drink -4.84155 1.81560 -2.667 0.007662 \*\*   
## foreignb -16.69267 5.37183 -3.107 0.001887 \*\*   
## nprevist 16.25712 0.37801 43.007 < 2e-16 \*\*\*  
## pre4000 456.91942 8.70077 52.515 < 2e-16 \*\*\*  
## tobacco -190.99398 2.53133 -75.452 < 2e-16 \*\*\*  
## mblack -127.42371 7.48213 -17.030 < 2e-16 \*\*\*  
## motherr -94.53491 11.94874 -7.912 2.55e-15 \*\*\*  
## mhispan -71.39186 8.54875 -8.351 < 2e-16 \*\*\*  
## fblack -26.32857 7.28583 -3.614 0.000302 \*\*\*  
## fotherr -96.75534 11.80404 -8.197 2.48e-16 \*\*\*  
## fhispan -46.41558 7.88343 -5.888 3.92e-09 \*\*\*  
## adequac2 16.34015 4.01218 4.073 4.65e-05 \*\*\*  
## adequac3 32.47602 8.49687 3.822 0.000132 \*\*\*  
## tripre2 15.44724 4.23903 3.644 0.000268 \*\*\*  
## tripre3 40.55286 9.24514 4.386 1.15e-05 \*\*\*  
## tripre0 19.83494 12.79841 1.550 0.121191   
## first -49.92552 3.23650 -15.426 < 2e-16 \*\*\*  
## plural -581.90963 9.75522 -59.651 < 2e-16 \*\*\*  
## dmage2 -0.09537 0.02616 -3.646 0.000267 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 440.4 on 235231 degrees of freedom  
## Multiple R-squared: 0.09759, Adjusted R-squared: 0.09747   
## F-statistic: 820.6 on 31 and 235231 DF, p-value: < 2.2e-16

The ordinary OLS regression indicates that smoking during pregnancy is associated with a 191 gram decrease in infant birthweight, all else equal. This is significant at a very high significance level (extremely close to zero). This is likely not an unbiased estimate of the true effect, however, because we do not have any observations on births where the child was less than 2500 grams. Had these been included in the data, we would have expected a steeper regression line which would have led to a more extreme negative estimate of the effect of tobacco on birth weight.

#3b) \*See attached calculations sheets

#3c)

trfit = truncreg(dbirwt~., data = dtr, point = 2500, direction = "left")  
 #^ requires "truncreg package", truncated at 2500 from the left  
summary(trfit)

##   
## Call:  
## truncreg(formula = dbirwt ~ ., data = dtr, point = 2500, direction = "left")  
##   
##   
## Coefficients :  
## Estimate Std. Error t-value Pr(>|t|)   
## (Intercept) 3120.629820 25.261484 123.5331 < 2.2e-16 \*\*\*  
## alcohol -20.924655 12.037170 -1.7383 0.0821515 .   
## anemia -8.898618 11.372576 -0.7825 0.4339426   
## cardiac -46.755492 15.216624 -3.0727 0.0021216 \*\*   
## chyper -71.690401 12.765038 -5.6162 1.953e-08 \*\*\*  
## dfage 0.198595 0.259612 0.7650 0.4442905   
## dfeduc 2.780784 0.614206 4.5274 5.970e-06 \*\*\*  
## diabete 81.522759 7.646374 10.6616 < 2.2e-16 \*\*\*  
## disllb -0.162428 0.045221 -3.5918 0.0003284 \*\*\*  
## dlivord 27.357472 1.371509 19.9470 < 2.2e-16 \*\*\*  
## dmage 3.752420 1.741699 2.1545 0.0312041 \*   
## dmar -35.086095 3.350425 -10.4721 < 2.2e-16 \*\*\*  
## dmeduc 5.335018 0.681843 7.8244 5.107e-15 \*\*\*  
## drink -7.767759 2.547104 -3.0496 0.0022911 \*\*   
## foreignb -18.661023 6.210390 -3.0048 0.0026575 \*\*   
## nprevist 18.276308 0.429556 42.5470 < 2.2e-16 \*\*\*  
## pre4000 476.602694 9.395537 50.7265 < 2.2e-16 \*\*\*  
## tobacco -227.687078 3.030857 -75.1230 < 2.2e-16 \*\*\*  
## mblack -156.318524 8.852891 -17.6573 < 2.2e-16 \*\*\*  
## motherr -111.689517 14.086156 -7.9290 2.220e-15 \*\*\*  
## mhispan -85.299739 10.067926 -8.4724 < 2.2e-16 \*\*\*  
## fblack -28.551246 8.580668 -3.3274 0.0008766 \*\*\*  
## fotherr -114.858475 13.967816 -8.2231 2.220e-16 \*\*\*  
## fhispan -54.462927 9.279944 -5.8689 4.387e-09 \*\*\*  
## adequac2 15.514127 4.659358 3.3297 0.0008695 \*\*\*  
## adequac3 24.140939 10.121180 2.3852 0.0170703 \*   
## tripre2 19.132102 4.962085 3.8557 0.0001154 \*\*\*  
## tripre3 54.068833 11.074620 4.8822 1.049e-06 \*\*\*  
## tripre0 9.956931 15.921160 0.6254 0.5317152   
## first -58.635553 3.699245 -15.8507 < 2.2e-16 \*\*\*  
## plural -855.025328 15.758713 -54.2573 < 2.2e-16 \*\*\*  
## dmage2 -0.081507 0.030244 -2.6949 0.0070400 \*\*   
## sigma 471.771092 0.846293 557.4560 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Log-Likelihood: -1760100 on 33 Df

tobacest = trfit$coefficients[18]  
tobacse = coef(summary(trfit))[18,2] #gives std. error of tobacco estimate  
(tobacest)

## tobacco   
## -227.6871

(tobacse)

## [1] 3.030857

CItobtr = c(tobacest-tobacse\*1.96,tobacest+tobacse\*1.96) #calculates 95% CI  
(CItobtr)

## tobacco tobacco   
## -233.6276 -221.7466

## Problem 4

#4a)

dc = dfull[,-c(5,18,32)] #gets rid of uncensored birthweight, "preterm" and "deadkids" covariates which aren't mentioned in the assignment   
  
fit2 = lm(cdbirwt2500~., data = dc)  
(summary(fit2))

##   
## Call:  
## lm(formula = cdbirwt2500 ~ ., data = dc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1759.1 -321.1 -9.2 305.0 3391.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2994.06273 22.14499 135.203 < 2e-16 \*\*\*  
## alcohol -22.41149 9.75450 -2.298 0.021588 \*   
## anemia -32.73142 9.70757 -3.372 0.000747 \*\*\*  
## cardiac -49.42020 13.46942 -3.669 0.000243 \*\*\*  
## chyper -137.66248 10.82945 -12.712 < 2e-16 \*\*\*  
## dfage 0.04394 0.22964 0.191 0.848268   
## dfeduc 3.53092 0.55297 6.385 1.71e-10 \*\*\*  
## diabete 55.36499 6.99452 7.915 2.47e-15 \*\*\*  
## disllb -0.20753 0.04025 -5.156 2.53e-07 \*\*\*  
## dlivord 27.70157 1.24475 22.255 < 2e-16 \*\*\*  
## dmage 6.69487 1.53481 4.362 1.29e-05 \*\*\*  
## dmar -35.17267 2.93066 -12.002 < 2e-16 \*\*\*  
## dmeduc 5.83566 0.61410 9.503 < 2e-16 \*\*\*  
## drink -6.96337 1.66658 -4.178 2.94e-05 \*\*\*  
## foreignb -15.86726 5.53514 -2.867 0.004149 \*\*   
## nprevist 24.84049 0.37751 65.801 < 2e-16 \*\*\*  
## pre4000 473.51852 9.17416 51.614 < 2e-16 \*\*\*  
## tobacco -213.73563 2.57099 -83.134 < 2e-16 \*\*\*  
## mblack -138.76798 7.60035 -18.258 < 2e-16 \*\*\*  
## motherr -92.38990 12.29470 -7.515 5.73e-14 \*\*\*  
## mhispan -74.99944 8.73673 -8.584 < 2e-16 \*\*\*  
## fblack -40.42761 7.41291 -5.454 4.94e-08 \*\*\*  
## fotherr -104.63816 12.12976 -8.627 < 2e-16 \*\*\*  
## fhispan -53.76150 8.03972 -6.687 2.28e-11 \*\*\*  
## adequac2 30.12442 4.05670 7.426 1.12e-13 \*\*\*  
## adequac3 66.19177 8.48452 7.801 6.14e-15 \*\*\*  
## tripre2 26.63226 4.30160 6.191 5.98e-10 \*\*\*  
## tripre3 67.97058 9.30404 7.305 2.77e-13 \*\*\*  
## tripre0 -15.04631 12.06653 -1.247 0.212418   
## first -68.30725 3.33641 -20.473 < 2e-16 \*\*\*  
## plural -720.42450 7.47098 -96.430 < 2e-16 \*\*\*  
## dmage2 -0.15179 0.02674 -5.676 1.38e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 468 on 249968 degrees of freedom  
## Multiple R-squared: 0.1374, Adjusted R-squared: 0.1373   
## F-statistic: 1284 on 31 and 249968 DF, p-value: < 2.2e-16

This ordinary OLS regression indicates that smoking tobacco during pregnancy is associated with a 214 gram decrease in birthweight, all else equal. This is significant at a very high significance level. This is probably not an unbiased estimate of the true effect of smoking during pregnancy on birthweight for similar reasons as in the truncated model. In this case, we actually observe the birthweights that were truncated in the previous data, but they are limited to 2500 and can't go lower regardless of their true value. This gives more data points in the low birthweight area, which leads to a slightly steeper OLS line and a slightly more accurate estimate of the true value, though it is still incorrect.

#4b) \*See attached calculations sheet

#4c)

dc = dfull[,-c(5,18,32)] #gets rid of uncensored birthweight, "preterm" and "deadkids" covariates which aren't mentioned in the assignment   
  
fit2 = lm(cdbirwt2500~., data = dc)  
(summary(fit2))

##   
## Call:  
## lm(formula = cdbirwt2500 ~ ., data = dc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1759.1 -321.1 -9.2 305.0 3391.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2994.06273 22.14499 135.203 < 2e-16 \*\*\*  
## alcohol -22.41149 9.75450 -2.298 0.021588 \*   
## anemia -32.73142 9.70757 -3.372 0.000747 \*\*\*  
## cardiac -49.42020 13.46942 -3.669 0.000243 \*\*\*  
## chyper -137.66248 10.82945 -12.712 < 2e-16 \*\*\*  
## dfage 0.04394 0.22964 0.191 0.848268   
## dfeduc 3.53092 0.55297 6.385 1.71e-10 \*\*\*  
## diabete 55.36499 6.99452 7.915 2.47e-15 \*\*\*  
## disllb -0.20753 0.04025 -5.156 2.53e-07 \*\*\*  
## dlivord 27.70157 1.24475 22.255 < 2e-16 \*\*\*  
## dmage 6.69487 1.53481 4.362 1.29e-05 \*\*\*  
## dmar -35.17267 2.93066 -12.002 < 2e-16 \*\*\*  
## dmeduc 5.83566 0.61410 9.503 < 2e-16 \*\*\*  
## drink -6.96337 1.66658 -4.178 2.94e-05 \*\*\*  
## foreignb -15.86726 5.53514 -2.867 0.004149 \*\*   
## nprevist 24.84049 0.37751 65.801 < 2e-16 \*\*\*  
## pre4000 473.51852 9.17416 51.614 < 2e-16 \*\*\*  
## tobacco -213.73563 2.57099 -83.134 < 2e-16 \*\*\*  
## mblack -138.76798 7.60035 -18.258 < 2e-16 \*\*\*  
## motherr -92.38990 12.29470 -7.515 5.73e-14 \*\*\*  
## mhispan -74.99944 8.73673 -8.584 < 2e-16 \*\*\*  
## fblack -40.42761 7.41291 -5.454 4.94e-08 \*\*\*  
## fotherr -104.63816 12.12976 -8.627 < 2e-16 \*\*\*  
## fhispan -53.76150 8.03972 -6.687 2.28e-11 \*\*\*  
## adequac2 30.12442 4.05670 7.426 1.12e-13 \*\*\*  
## adequac3 66.19177 8.48452 7.801 6.14e-15 \*\*\*  
## tripre2 26.63226 4.30160 6.191 5.98e-10 \*\*\*  
## tripre3 67.97058 9.30404 7.305 2.77e-13 \*\*\*  
## tripre0 -15.04631 12.06653 -1.247 0.212418   
## first -68.30725 3.33641 -20.473 < 2e-16 \*\*\*  
## plural -720.42450 7.47098 -96.430 < 2e-16 \*\*\*  
## dmage2 -0.15179 0.02674 -5.676 1.38e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 468 on 249968 degrees of freedom  
## Multiple R-squared: 0.1374, Adjusted R-squared: 0.1373   
## F-statistic: 1284 on 31 and 249968 DF, p-value: < 2.2e-16

cnsfit = censReg(cdbirwt2500~., data = dc, left = 2500, right = Inf,method="BHHH")  
 #^requires "censReg" package, censored left at 2500, right at Inf  
(summary(cnsfit))

##   
## Call:  
## censReg(formula = cdbirwt2500 ~ ., left = 2500, right = Inf,   
## data = dc, method = "BHHH")  
##   
## Observations:  
## Total Left-censored Uncensored Right-censored   
## 250000 14737 235263 0   
##   
## Coefficients:  
## Estimate Std. error t value Pr(> t)   
## (Intercept) 2.942e+03 2.298e+01 128.056 < 2e-16 \*\*\*  
## alcohol -2.498e+01 1.013e+01 -2.467 0.013615 \*   
## anemia -4.011e+01 9.702e+00 -4.135 3.55e-05 \*\*\*  
## cardiac -5.376e+01 1.439e+01 -3.735 0.000188 \*\*\*  
## chyper -1.608e+02 9.290e+00 -17.308 < 2e-16 \*\*\*  
## dfage 7.916e-03 2.370e-01 0.033 0.973357   
## dfeduc 3.756e+00 5.837e-01 6.435 1.24e-10 \*\*\*  
## diabete 5.105e+01 6.428e+00 7.942 1.99e-15 \*\*\*  
## disllb -2.237e-01 4.106e-02 -5.448 5.09e-08 \*\*\*  
## dlivord 2.945e+01 1.267e+00 23.243 < 2e-16 \*\*\*  
## dmage 7.076e+00 1.591e+00 4.447 8.72e-06 \*\*\*  
## dmar -3.836e+01 3.046e+00 -12.592 < 2e-16 \*\*\*  
## dmeduc 6.230e+00 6.482e-01 9.612 < 2e-16 \*\*\*  
## drink -9.542e+00 1.533e+00 -6.225 4.83e-10 \*\*\*  
## foreignb -1.547e+01 5.779e+00 -2.678 0.007409 \*\*   
## nprevist 2.809e+01 3.322e-01 84.561 < 2e-16 \*\*\*  
## pre4000 4.787e+02 9.468e+00 50.559 < 2e-16 \*\*\*  
## tobacco -2.263e+02 2.718e+00 -83.265 < 2e-16 \*\*\*  
## mblack -1.480e+02 7.771e+00 -19.045 < 2e-16 \*\*\*  
## motherr -9.405e+01 1.273e+01 -7.390 1.46e-13 \*\*\*  
## mhispan -7.673e+01 9.108e+00 -8.424 < 2e-16 \*\*\*  
## fblack -4.512e+01 7.557e+00 -5.970 2.37e-09 \*\*\*  
## fotherr -1.088e+02 1.269e+01 -8.576 < 2e-16 \*\*\*  
## fhispan -5.742e+01 8.412e+00 -6.826 8.72e-12 \*\*\*  
## adequac2 3.705e+01 3.999e+00 9.265 < 2e-16 \*\*\*  
## adequac3 7.862e+01 8.452e+00 9.302 < 2e-16 \*\*\*  
## tripre2 2.863e+01 4.394e+00 6.516 7.23e-11 \*\*\*  
## tripre3 7.794e+01 9.671e+00 8.059 7.70e-16 \*\*\*  
## tripre0 -4.319e+01 1.180e+01 -3.660 0.000253 \*\*\*  
## first -7.405e+01 3.496e+00 -21.180 < 2e-16 \*\*\*  
## plural -9.111e+02 8.599e+00 -105.952 < 2e-16 \*\*\*  
## dmage2 -1.653e-01 2.761e-02 -5.986 2.15e-09 \*\*\*  
## logSigma 6.199e+00 1.361e-03 4553.633 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## BHHH maximisation, 14 iterations  
## Return code 2: successive function values within tolerance limit  
## Log-likelihood: -1809560 on 33 Df

tobacestcns = cnsfit$estimate[18]  
tobacsecns = coef(summary(cnsfit))[18,2]  
(tobacestcns)

## tobacco   
## -226.3155

(tobacsecns)

## [1] 2.718005

CItobcns = c(tobacestcns-tobacsecns\*1.96,tobacestcns+tobacsecns\*1.96)  
(CItobcns)

## tobacco tobacco   
## -231.6428 -220.9882

## Problem 5

#5a)

dfull = dfull[,-c(18,32,35)]#gets rid of "preterm", "deadkids", and "cdbirwt2500" covariates which aren't mentioned in the assignment   
  
fullfit = lm(dbirwt~., data = dfull)  
(summary(fullfit))

##   
## Call:  
## lm(formula = dbirwt ~ ., data = dfull)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3549.6 -303.0 17.8 334.5 3431.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2805.61513 25.35639 110.647 < 2e-16 \*\*\*  
## alcohol -25.19821 11.16907 -2.256 0.02407 \*   
## anemia -55.25044 11.11533 -4.971 6.68e-07 \*\*\*  
## cardiac -59.02826 15.42272 -3.827 0.00013 \*\*\*  
## chyper -209.61905 12.39990 -16.905 < 2e-16 \*\*\*  
## dfage -0.07689 0.26294 -0.292 0.76995   
## dfeduc 4.35438 0.63316 6.877 6.12e-12 \*\*\*  
## diabete 38.90363 8.00885 4.858 1.19e-06 \*\*\*  
## disllb -0.25344 0.04609 -5.499 3.82e-08 \*\*\*  
## dlivord 31.31721 1.42526 21.973 < 2e-16 \*\*\*  
## dmage 8.75244 1.75738 4.980 6.35e-07 \*\*\*  
## dmar -42.66961 3.35566 -12.716 < 2e-16 \*\*\*  
## dmeduc 6.00593 0.70315 8.541 < 2e-16 \*\*\*  
## drink -9.92581 1.90827 -5.201 1.98e-07 \*\*\*  
## foreignb -15.85210 6.33783 -2.501 0.01238 \*   
## nprevist 37.02335 0.43226 85.651 < 2e-16 \*\*\*  
## pre4000 480.98922 10.50457 45.789 < 2e-16 \*\*\*  
## tobacco -229.41852 2.94382 -77.932 < 2e-16 \*\*\*  
## mblack -154.08295 8.70253 -17.706 < 2e-16 \*\*\*  
## motherr -85.76344 14.07765 -6.092 1.12e-09 \*\*\*  
## mhispan -77.53224 10.00371 -7.750 9.20e-15 \*\*\*  
## fblack -53.10469 8.48791 -6.257 3.94e-10 \*\*\*  
## fotherr -107.06115 13.88879 -7.708 1.28e-14 \*\*\*  
## fhispan -59.67901 9.20561 -6.483 9.01e-11 \*\*\*  
## adequac2 67.95236 4.64499 14.629 < 2e-16 \*\*\*  
## adequac3 153.80668 9.71492 15.832 < 2e-16 \*\*\*  
## tripre2 27.16719 4.92541 5.516 3.48e-08 \*\*\*  
## tripre3 80.08478 10.65329 7.517 5.61e-14 \*\*\*  
## tripre0 -109.43095 13.81638 -7.920 2.38e-15 \*\*\*  
## first -82.01740 3.82024 -21.469 < 2e-16 \*\*\*  
## plural -982.86673 8.55440 -114.896 < 2e-16 \*\*\*  
## dmage2 -0.20353 0.03062 -6.647 3.00e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 535.9 on 249968 degrees of freedom  
## Multiple R-squared: 0.1549, Adjusted R-squared: 0.1548   
## F-statistic: 1478 on 31 and 249968 DF, p-value: < 2.2e-16

tobacestfull = fullfit$coefficients[18]  
tobacsefull = coef(summary(fullfit))[18,2]  
(tobacestfull)

## tobacco   
## -229.4185

(tobacsefull)

## [1] 2.943824

CIfull = c(tobacestfull-tobacsefull\*1.96,tobacestfull+tobacsefull\*1.96)  
(CIfull)

## tobacco tobacco   
## -235.1884 -223.6486

#5b)

efftobac = c(fit1$coefficients[18], tobacest,fit2$coefficients[18],cnsfit$estimate[18],tobacestfull)  
names(efftobac)= c("truncOLS", "truncModel", "censOLS", "censModel", "fullOLS")  
  
(efftobac)

## truncOLS truncModel censOLS censModel fullOLS   
## -190.9940 -227.6871 -213.7356 -226.3155 -229.4185

This table confirms some of the hypotheses stated above concerning the effects of truncated and censored data on estimated effects. Truncated data appears to lead to a stronger bias toward zero because there is no data pulling the OLS line toward its true value. Censored data has a similar effect, but it is less extreme because we still observe the data, just at a cutoff point instead of at its true value. The results provided here seem to indicate that the truncated model actually provides a better estimate of the true OLS estimate than the censored model. This may be because shifting the distribution of the observed data to account for the unobserved data (truncated model) better captures the true distribution than estimating a true distribution for observed data and only a mass point for unobserved data (censored model). In any case, both models are biased toward zero compared to the true OLS model, but it seems that they do a good job of predicting OLS.

## Problem 6

The paper by Chay and Powell discusses a few different kinds of semiparametric models that can be used to estimate effects when given censored data. The authors use earnings data on black and white US citizens in the 60s, which is censored right at the taxable maximum, to examine if there was any convergence in earnings inequality during the civil rights movement. Chay and Powell discuss censored regression first, stating that it is a useful tool when the distribution of the error term is known and has constant variance. This allows for the use of Maximum Likelihood Estimation. When the distribrution is unknown or the errors are thought to be heteroskedastic, other models, specifically censored least absolute deviation (CLAD), symmetrically censored least squares (SCLS), or identically censored least absolute deviations (ICLAD), can be used (given certain assumptions about the distribution of the error terms). CLAD essentially makes use of a median regression rather than a standard mean one. The idea is that if the median is within the uncensored range, it will be unaffected by censored data outside the range. SCLS re-censors the dependant variable so that it is symmetric around X'B, and estimates are obtained using least squares. ICLAD focuses on the differences between two censored dependent variables. These two dependent variables will become symmetric about their difference and estimates can be derived from minimizing the sum of squared residuals of these diferences across the observations. Chay and Powell apply these methods to the 60s earnings data to show how they deviate from ordinary OLS on censored data. The CLAD, SCLS, and ICLAD effects are significantly more extreme than OLS, OLS with the censored data deleted, and MLE estimates indicating that OLS and simple MLE are biased in predicting efects from censored data. In addition, these semiparametric models can provide some information about the sources of model misspecification using OLS and MLE. Overall, the "new" methods used illustrate a larger narrowing of the wage gap following the civil rights movement than could otherwise be estimated in a parametric approach.