PS3

## Problem 1

### a)

fit1 = lm(dbirwt~tobacco,data = d)  
(summary(fit1))

##   
## Call:  
## lm(formula = dbirwt ~ tobacco, data = d)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3199.2 -307.2 32.8 349.4 3409.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3426.194 1.271 2694.90 <2e-16 \*\*\*  
## tobacco -260.580 2.960 -88.04 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 574.1 on 249998 degrees of freedom  
## Multiple R-squared: 0.03007, Adjusted R-squared: 0.03007   
## F-statistic: 7751 on 1 and 249998 DF, p-value: < 2.2e-16

The coefficient on the smoking indicator would only represent a causal efect on infant birthweight if whether or not a mother smoked during pregnancy was determined completely at random. This would make the treatment (smoking) independent and therefore uncorrelated with any other variable or possible confounder, and allow us to compare the treatment and control groups under the assumption that they have the same characteristics on average. Clearly, this is not a plausible assumption in this case because whether or not a mother smokes during pregnancy is not determined at random, and is probably related to other attributes and health risks (nor could it be assigned at random, ethically speaking).

### b)

db = d[,-c(18,32,35)]#gets rid of "preterm", "deadkids", and "cdbirwt2500" covariates   
  
fit2 = lm(dbirwt~., data = db)  
(summary(fit2))

##   
## Call:  
## lm(formula = dbirwt ~ ., data = db)  
##   
## 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

Again, the coefficient on the smoking indicator would only represent a causal effect if whether or not a mother smokes during pregnancy was determined completely randomly. In this case, this is an unreasonable assumption because this is not something that can be randomly assigned (nor is it in real life), and smoking during pregnancy is probably correlated with other covariates in this data set (such as alcohol, education, etc.). The tobacco coefficient is roughly 30 units lower in this regression than in the previous one. This is probably because it is correlated with some of the other covariates. In essence, the overall effect is distributed between the highly correlated covariates, whereas it is only captured by the smoking indicator in the regression above.

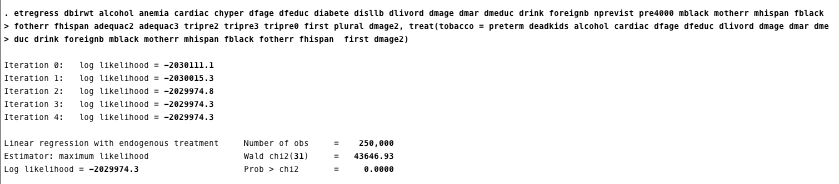
### c)

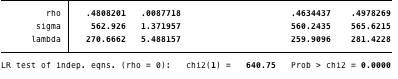
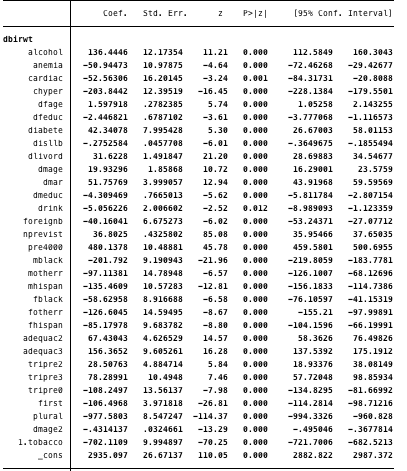
\*See attached sheet

### d)

These two variables would be valid exclusion restrictions if they only affect a child's birth weight through whether or not a mother smoked during pregnancy. In other words, if a premature birth or dead child affected a mother's chances of smoking during her next pregnancy, but didn't affect the next child's birthweight directly.

### e)





These results were obtained using MLE (selection equation coefficients left out to save space). One advantage of using Heckman's two-step method over MLE is that it is much easier to implement. On the other hand, MLE provides asymptotically efficient estimators (assuming normality is satisfied), and confines rho to its normal [-1,1] boundaries. According to this model, smoking during pregnancy causes a 702 gram decrease in birthweight, with a 95%CI of [-722,-683]. The estimated rho tells us that there is a correlation of .481 between the errors in the outcome equation and the errors in the selection equation. Since we have a high Chi-Square value for this rho in the test of independence, we can reject the hypothesis that the outcome and selection equations are independent. Therefore, our exclusion restrictions may actually have some effect on infant birthweight on their own, rather than only through smoking during pregnancy.

### f)

I do not trust the results from this partiular selection model. The magnitude of the effect of smoking during pregnancy on infant birthweight seems way too high to be realistic, especially considering some of the past analysis we have performed on this subject. It is not safe to trust this estimated effect because there are likely flaws in the exclusion restrictions that we included in the selection equation. Clearly these are not valid exclusions because previous premature births or previously birthing a baby that dies will have an effect on the "healthiness" (birthweight) of the next child in some way other than through smoking alone. For example, preterm and deadkids are likely variables related to the overall child-bearing ability of the mother, and so they will affect infant birthweight regardless of whether or not the mother smoked during pregnancy.

### g)

One valid instrument for this problem may be the degree of tobacco advertising the mother is exposed to. We could imagine that this would have an effect on whether or not she smokes during pregnancy (maybe due to habit relapse from more extreme hormones, for example), but that it wouldn't affect the child's birthweight directly. This wouldn't be a perfect IV because being a past smoker would probably have something to do with the effect of the tobacco advertising. If this variable were used in an IV model, we would no longer be estimating the ATE of smoking during pregnancy on birthweight for all mothers, but instead we would get distinct ATEs of smoking during pregnancy on birthweight for mothers that were exposed to varying degrees of tobacco advertisements. In this case, the exclusion restriction assumption would be that being exposed to tobacco ads affects whether or not a mother smokes during pregnancy, but not the infant's birthweight. Monotonicity would state that tobacco advertisements cannot have a negative effect on whether or not a mother smokes during pregnancy, which may be a poor assumption for this instrument since tobacco ads could also be perceived as a deterrent to smoking during pregnancy for some mothers.

## Problem 2

### a)

fit3 = lm(dbirwt~tobacco,data = d1)  
(summary(fit3))

##   
## Call:  
## lm(formula = dbirwt ~ tobacco, data = d1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3196.4 -304.4 35.6 347.6 2785.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3423.360 1.987 1722.48 <2e-16 \*\*\*  
## tobacco -266.029 4.762 -55.87 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 571.1 on 99998 degrees of freedom  
## Multiple R-squared: 0.03027, Adjusted R-squared: 0.03026   
## F-statistic: 3121 on 1 and 99998 DF, p-value: < 2.2e-16

CI1 = c(fit3$coefficients[2]-1.96\*coef(summary(fit3))[2,2],fit3$coefficients[2]+1.96\*coef(summary(fit3))[2,2])  
(CI1)

## tobacco tobacco   
## -275.3615 -256.6956

fit4 = lm(dbirwt~.,data = d1)  
(coef(summary(fit4))[18,])

## Estimate Std. Error t value Pr(>|t|)   
## -231.976451 4.728613 -49.058031 0.000000

CI2 = c(fit4$coefficients[18]-1.96\*coef(summary(fit4))[18,2],fit4$coefficients[18]+1.96\*coef(summary(fit4))[18,2])  
(CI2)

## tobacco tobacco   
## -241.2445 -222.7084

The estimated effect of smoking on birthweight is -267 and -232 for regressions one and two, respectively.

### b)

fit5 = rq(dbirwt~tobacco,tau=0.5, data = d1)  
 #requires "quantreg" package, tau=0.5 for median regression  
(summary(fit5))

##   
## Call: rq(formula = dbirwt ~ tobacco, tau = 0.5, data = d1)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## Value Std. Error t value Pr(>|t|)   
## (Intercept) 3459.00000 1.66250 2080.59666 0.00000  
## tobacco -259.00000 4.48226 -57.78330 0.00000

CI3 = c(fit5$coefficients[2]-1.96\*coef(summary(fit5))[2,2],fit5$coefficients[2]+1.96\*coef(summary(fit5))[2,2])  
(CI3)

## tobacco tobacco   
## -267.7852 -250.2148

The estimated coefficient for the smoking indicator using median regression is -259. This estimated coefficient does not represent a causal effect of smoking on birth weight. It is similar to an OLS regression but uses a different minimization condition (LAD), so we do not get causality. We can not infer the effect of smoking during pregnancy on the unconditional median of the birth weight distribution using this estimated coefficient. Instead, this coefficient gives the the effect of smoking during pregnancy on the median of the birthweight distribution conditional on the smoking indicator.

### c)

fit6 = rq(dbirwt~.,tau=0.5,data=d1)

(summary(fit6))

##   
## Call: rq(formula = dbirwt ~ ., tau = 0.5, data = d1)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## Value Std. Error t value Pr(>|t|)   
## (Intercept) 3043.57365 41.06448 74.11693 0.00000  
## alcohol -36.47542 25.81805 -1.41279 0.15772  
## anemia -18.85640 14.64652 -1.28743 0.19795  
## cardiac -25.33957 21.10688 -1.20054 0.22993  
## chyper -109.60915 25.30401 -4.33169 0.00001  
## dfage 0.03799 0.43734 0.08686 0.93078  
## dfeduc 5.34229 1.04401 5.11708 0.00000  
## diabete 69.05360 17.57617 3.92882 0.00009  
## disllb -0.21286 0.07452 -2.85616 0.00429  
## dlivord 25.14052 2.44618 10.27748 0.00000  
## dmage 2.66732 2.82992 0.94255 0.34592  
## dmar -45.63106 5.69771 -8.00867 0.00000  
## dmeduc 5.20409 1.15768 4.49527 0.00001  
## drink -11.90594 5.21043 -2.28502 0.02231  
## foreignb -27.58578 9.98509 -2.76270 0.00573  
## nprevist 25.08754 0.73442 34.15973 0.00000  
## pre4000 474.51251 14.02622 33.83040 0.00000  
## tobacco -227.74454 5.18824 -43.89627 0.00000  
## mblack -134.60064 14.13978 -9.51929 0.00000  
## motherr -90.65273 20.34897 -4.45490 0.00001  
## mhispan -103.35123 16.87736 -6.12366 0.00000  
## fblack -64.23100 13.67897 -4.69560 0.00000  
## fotherr -105.64625 20.18445 -5.23404 0.00000  
## fhispan -36.80249 14.93608 -2.46400 0.01374  
## adequac2 26.65863 7.89146 3.37816 0.00073  
## adequac3 61.46341 17.06990 3.60069 0.00032  
## tripre2 35.52259 8.24012 4.31093 0.00002  
## tripre3 63.73929 17.85951 3.56893 0.00036  
## tripre0 -50.77035 27.08784 -1.87429 0.06089  
## first -77.14674 6.30641 -12.23307 0.00000  
## plural -911.44478 16.79302 -54.27521 0.00000  
## dmage2 -0.09199 0.04907 -1.87480 0.06082

CI4 = c(fit6$coefficients[18]-1.96\*coef(summary(fit6))[18,2],fit6$coefficients[18]+1.96\*coef(summary(fit6))[18,2])  
(CI4)

## tobacco tobacco   
## -237.9135 -217.5756

I would interpret the coefficient on tobacco (-228) as the effect of smoking during pregnancy on the median of the conditional brthweight distribution (conditional on all covariates). This is not a causal effect for reasons similar to those stated above. The effect estimated by median regression is higher by about four grams than the effect estimated by OLS above (-228 vs. -232). This tells me that the mean effect of smoking on birthweight is slightly more extreme than the median effect. This may be because there are a few outliers that pull the mean downwards.

### d)

R documentation states that rq uses the sandwich method to compute the asymptotic vcv. While STATA has this option, it defaults to the Koenker and Bassett method, which assumes that errors are distributed i.i.d. This is not a requirement for the sandwich estimator. I like the sandwich estimator better because it does not impose this restriction, and can calculate the asymptotic variance regardless of whether or not the assumption is true. The main drawback is that negative estimates can occur using the sandwich estimator (and they did in a few cases). While this isn't a big deal in my case, it can cause issues if there are a lot of negative estimates. The sandwich estimator used by R allows for heteroskedasticity, since it does not impose the i.i.d. restriction (fur(0|X)), while the default stata estimate does not allow for heteroskedasticity (fur(0)).

### e)

Bootstrapping estimates the asymptotic variance without imposing the restriction that the errors are i.i.d. (in other words, it uses fur(0|X)).

set.seed(100)  
n=nrow(d1)  
est = matrix(nrow = 32,ncol = 10)  
for(i in 1:10) {  
 sam = sample(1:n, replace = TRUE)  
 ds = d1[sam,]  
 fit = rq(dbirwt~.,tau=0.5,data = ds)  
 est[,i] = fit$coefficients  
}

(setob = sd(est[18,]))

## tobacco

## 4.254971  
vcv = var(t(est))

This method gives us an estimate of the standard error on tobacco of 4.254971. This is pretty close to the quantile regression output of 5.188, especially considering that we only used 10 replications.

### f)

tobac = fit6$coefficients[18]  
alc = fit6$coefficients[2]  
(test = (-300-tobac-alc)/(sqrt(setob^2+sd(est[2,]^2))))

## tobacco   
## -1.377377

The test statistic for this null hypothesis is -1.38, so we can not reject the null that the effects of tobacco and alcohol together are less than or equal to -300 (critical value of 1.645 for one tail).

## 

## Problem 3

### a)

(fit7 = rq(dbirwt~tobacco,tau=0.1,data = d1))

## Call:  
## rq(formula = dbirwt ~ tobacco, tau = 0.1, data = d1)  
##   
## Coefficients:  
## (Intercept) tobacco   
## 2778 -283   
##   
## Degrees of freedom: 100000 total; 99998 residual

(fit8 = rq(dbirwt~tobacco,tau=0.25,data = d1))

## Call:  
## rq(formula = dbirwt ~ tobacco, tau = 0.25, data = d1)  
##   
## Coefficients:  
## (Intercept) tobacco   
## 3119 -263   
##   
## Degrees of freedom: 100000 total; 99998 residual

(fit9 = rq(dbirwt~tobacco,tau=0.5,data = d1))

## Call:  
## rq(formula = dbirwt ~ tobacco, tau = 0.5, data = d1)  
##   
## Coefficients:  
## (Intercept) tobacco   
## 3459 -259   
##   
## Degrees of freedom: 100000 total; 99998 residual

(fit10 = rq(dbirwt~tobacco,tau=0.75,data = d1))

## Call:  
## rq(formula = dbirwt ~ tobacco, tau = 0.75, data = d1)  
##   
## Coefficients:  
## (Intercept) tobacco   
## 3771 -256   
##   
## Degrees of freedom: 100000 total; 99998 residual

(fit11 = rq(dbirwt~tobacco,tau=0.9,data = d1))

## Call:  
## rq(formula = dbirwt ~ tobacco, tau = 0.9, data = d1)  
##   
## Coefficients:  
## (Intercept) tobacco   
## 4082 -255   
##   
## Degrees of freedom: 100000 total; 99998 residual

The estimated coefficients on tobacco get closer to zero (at a slowing pace) as we increase the quantile. These coefficients can be interpreted as the effect of smoking during pregnancy on their respective conditional quantiles of infant birthweight, conditioning on the smoking indicator. We can not infer the effect of smoking during pregnancy on the unconditional quantiles of the birth weight distribution from these estimates.

### b)

fit12 = rq(dbirwt~.,tau=0.1,data = d1)

fit13 = rq(dbirwt~.,tau=0.25,data = d1)

fit14 = rq(dbirwt~.,tau=0.5,data = d1)

fit15 = rq(dbirwt~.,tau=0.75,data = d1)

fit16 = rq(dbirwt~.,tau=0.9,data = d1)

(b.1 = fit12$coefficients[18]) #.1

## tobacco   
## -253.7774

(b.25 = fit13$coefficients[18]) #.25

## tobacco   
## -231.8153

(b.5 = fit14$coefficients[18]) #.5

## tobacco   
## -227.7445

(b.75 = fit15$coefficients[18]) #.75

## tobacco   
## -220.5429

(b.9 = fit16$coefficients[18]) #.9

## tobacco   
## -228.9399

After controlling for the other covariates, the coefficient on the smoking indicator is lower in magnitude than it otherwise is without controlling for the other covariates. Again, we see that the effect gets closer to zero at a slowing rate as we move up in the quantiles. Interestingly, the effect at the .90 quantile is more extreme than those at both the .75 and the .5 quantiles. These coefficients can be interpreted as the effect of smoking during pregnancy on infant birthweight, conditional on the other covariates, at respective birthweight quantiles. These effects are generally less extreme than the one we calculated using OLS in 2a), except for at the .10 quantile (though the .25 quantile is very close to the OLS estimate). The results from the analysis above imply that smoking during pregnancy has a greater negative effect on the health of a baby if the baby is prone to a low birth weight to begin with, conditional on the other covariates.

### c)

se.1 = coef(summary(fit12))[18,2]

se.25 = coef(summary(fit13))[18,2]  
se.5 = coef(summary(fit14))[18,2]  
se.75 = coef(summary(fit15))[18,2]  
se.9 = coef(summary(fit16))[18,2]

(anova(fit12,fit13,fit14,fit15,fit16,joint=FALSE))

## Quantile Regression Analysis of Deviance Table  
##   
## Model: dbirwt ~ alcohol + anemia + cardiac + chyper + dfage + dfeduc + diabete + disllb + dlivord + dmage + dmar + dmeduc + drink + foreignb + nprevist + pre4000 + tobacco + mblack + motherr + mhispan + fblack + fotherr + fhispan + adequac2 + adequac3 + tripre2 + tripre3 + tripre0 + first + plural + dmage2  
## Tests of Equality of Distinct Slopes: tau in { 0.1 0.25 0.5 0.75 0.9 }  
##   
## Df Resid Df F value Pr(>F)   
## alcohol 4 5e+05 1.8034 0.1250173   
## anemia 4 5e+05 1.7007 0.1466902   
## cardiac 4 5e+05 0.5434 0.7038644   
## chyper 4 5e+05 10.8759 8.141e-09 \*\*\*  
## dfage 4 5e+05 1.0754 0.3667192   
## dfeduc 4 5e+05 2.6762 0.0300902 \*   
## diabete 4 5e+05 14.9559 3.165e-12 \*\*\*  
## disllb 4 5e+05 1.1699 0.3217630   
## dlivord 4 5e+05 1.2453 0.2892511   
## dmage 4 5e+05 1.8445 0.1172067   
## dmar 4 5e+05 4.0265 0.0028805 \*\*   
## dmeduc 4 5e+05 0.7786 0.5388389   
## drink 4 5e+05 1.1963 0.3100729   
## foreignb 4 5e+05 1.7117 0.1442199   
## nprevist 4 5e+05 75.4625 < 2.2e-16 \*\*\*  
## pre4000 4 5e+05 0.6070 0.6575654   
## tobacco 4 5e+05 3.3762 0.0090567 \*\*   
## mblack 4 5e+05 0.6472 0.6287790   
## motherr 4 5e+05 1.1221 0.3439406   
## mhispan 4 5e+05 1.1777 0.3182860   
## fblack 4 5e+05 2.1686 0.0697770 .   
## fotherr 4 5e+05 4.0014 0.0030117 \*\*   
## fhispan 4 5e+05 2.1655 0.0701281 .   
## adequac2 4 5e+05 2.3037 0.0559515 .   
## adequac3 4 5e+05 1.4597 0.2115103   
## tripre2 4 5e+05 6.3642 4.073e-05 \*\*\*  
## tripre3 4 5e+05 5.3166 0.0002804 \*\*\*  
## tripre0 4 5e+05 4.1500 0.0023114 \*\*   
## first 4 5e+05 10.2146 2.878e-08 \*\*\*  
## plural 4 5e+05 11.0657 5.663e-09 \*\*\*  
## dmage2 4 5e+05 1.8072 0.1242740   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Given the small p-value on the tobacco indicator, we can safely reject the null hypothesis that the coefficients on the tobacco indicator are the same at each quantile tested. Similar results can be seen for the variables chyper, diabete, dmar, fotherr, tripre, first, and plural.

### d)

(fit12$coefficients[29]) #.1

## tripre0   
## -503.1754

(fit13$coefficients[29]) #.25

## tripre0   
## -127.6602

(fit14$coefficients[29]) #.5

## tripre0   
## -50.77035

(fit15$coefficients[29]) #.75

## tripre0   
## -38.12517

(fit16$coefficients[29]) #.9

## tripre0   
## -24.95823

The coefficients on tripre0 at different quantiles follow a similar pattern to the smoking indicator, however the overall change across quantiles is far more extreme. These coefficients can be interpreted as the effect of not receiving parental care during pregnancy on infant birthweight, controlling for the covariates, at respective infant birthweight quantiles. Compared to the OLS estimate in 2a) of -120, only quantiles .10 and .25 have a value more extreme in magnitude (quantile .10 is almost 4.5x bigger). This suggests that the OLS estimate for tripre0 was pulled significantly by birthweights near and below the .10 quantile, especially considering that the conditional effect at the median was about -51. The results from this analysis imply that the effect of not receiving parental care on a child's birthweight is far, far greater for children with low birthweights than it is for children with high birthweights, conditional on the other covariates.

## Problem 4

The article by Angrist and Krueger (2001) outlines various uses for instrumental variables (IVs). First, the authors discuss P.G. Wright's use of IVs to obtain estimates of the elasticity of supply and demand for a wide range of products. The idea was that the IVs affected cost conditions (supply) without affecting demand, or vice versa, and could therefore be seen as a "shifter" of its respective curve. While Wright averaged IVs, using 2SLS provides a more efficient estimate of elasticity (Theil 1953).

IVs are also helpful for overcoming measurement error, which results in understated regression estiamtes. In this context, an IV that is uncorrelated with the measurement error but is correlated with a correctly measured covariate provides a consistent regression estiamte. This can be applied in a similar manner to models that are thought to have omitted variables. An appropriate IV in this case would be a variable that is not correlated with the omitted variable (like the measurement error), but is correlated with the covariate in question. Incorporating this IV into an IV regression would provide a consistent estimate of the covariate in question (one correlated with the IV). The authors go on to provide an example of this using compulsory schooling laws (which we worked with in Prof. James' class!). The authors then discuss some of the issues associated with using IVs to obtain estimates. The first is that not every individual's dependent variable is always affected by the instrument. The example they provide relates to the Vietnam War data we looked at in class. The IV in this case is whether or not an individual was drafted, which was thought to affect their veteran status but not future earnings. The issue is that there were individuals that would have served in the military regardless of if they were drafted, so the estimates obtained really only apply to individuals who were affected by the draft. Finally, the authors discuss the difficulties behind creating a good instrument. These include making an instrument that is truly uncorrelated with the omitted variable and the error term, and strongly correlated with the desired endogenous variables.

## 

## Problem 5

The article by Koenker and Hallock (2001) discusses some of the methods behind, and uses of, quantile regression. First The authors perform a quantile regression of the market value of a firm's equity on executive compensation. They find not only that compensation tends to rise with firm size, but that variability in compensation does as well. Koenker and Hallock go on to introduce the theory behind quantile regression. It is very similar to OLS, except that, when looking at the median, the problem becomes minimizing the sum of the absolute value of the residuals. When looking at different quantiles besides the median, residuals are weighted depending on the quantile.

These methods were used by Engel (1857) to show the dispersion of food expenditure as household income increases. Next, the authors discuss quantile regression in the context of our problem: the effects of various mother attributes and risk factors on infant birthweight. This analysis confirms some of the results that we reached above. The main takeaway here is that there is rarely a constant effect of these covariates on birthweight at different birthweights. Instead, the effects change depending on the infant's birthweight. Koenker and Hallack explain in the next section that because of this phenomenon, there is a strong case for using the results of quantile regression more in emperical work, rather than simply looking at the conditional mean provided by OLS. This shift is also becoming backed by advances in statistical software. All in all, quantile regression can be thought of as a disciple of OLS that helps us examine data even more closely.