Regression Analysis of Anthropometry Data: A Simulation Study of a Two-Stage Estimator

Stuart Sweeney*† Kathryn Grace‡

Abstract

Regression analysis of anthropometry data has a long history in public health research. Early work relied on conditional mean regression models, but given that most policy interest is in either the lower or upper tail of a distribution, recent studies have utilized either binary outcome regression (logistic or ordinal logistic) or quantile regression. If the errors of the index function underlying binary models have non-constant variance, it is well-known that parameter estimates are inconsistent. We present simulation results of a proposed two-stage estimator to adjust for heteroskedasticity of unknown form. The two-stage estimator appears to substantially reduce bias in both parameter estimates and predictive changes in prevalence.

Key Words: Quantile regression, ordinal regression, anthropometry, public health, heteroskedasticity, simulation study

1. Introduction

Studying variations in the measurements of people (height, weight, arm circumference, etc.), or anthropometry, has a long history of analysis in public health. The origins of many of the approaches in use today, including the ratio of weight in kilograms to squared height in meters – known now as the *body mass index* (BMI) and formerly as the Quetelet index – dates to work by Quetelet from the middle of the 19th century. Anthropometric data is widely used to classify and identify vulnerable populations over time and across space. An infant's weight at birth provides an indication of the mother's lifestyle during pregnancy, if a child's height-for-age is low relative to an international standard then researchers would classify the child as chronically undernourished, if a woman has a low mid-upper arm circumference the woman would be said to be suffering from significant caloric depletion.

In general, researchers are primarily interested in both the prevalence (during a specific time period or relative to a particular population) and in the determinants/correlates of sub-optimal anthropometric measures. Reflecting these interests, logistic regression using some external cut-point, is frequently employed despite the fact that the underlying distribution of most anthropometric measures is actually continuous. In a recently published article, Sweeney et al. (2012) show that in ignoring the underlying continuous distribution of anthroprometric measures the resulting statistical models are not able to capture important variation within a population. They further argue that with the use of two types of regression strategies, quantile regression and ordinal regression, that an improved understanding of the factors related to variation in anthropometric outcomes results. Basically, the use of quantile regression facilitates an analysis of the tails of the distribution and allows identification of complex distributional features. A two-stage estimator (based on a median regression) is then proposed which enables the use of the ordinal regression model. The focus of this paper is a simulation study of the two-stage estimator.

*Department of Geography, University of California, Santa Barbara, CA 93106-4060

†Institute for Social, Behavioral, and Economic Research, University of California, Santa Barbara, CA 93106

‡Department of Geography, University of Utah, Salt Lake City, UT

2. Regression Models

Throughout this section we assume the existence of some non-specific anthropometry measure, Y, and a covariate X measured for a large set of individuals. Early anthropometry research employed linear regression with the resulting parameter estimates measuring the shift in the conditional mean of Y given a one unit change X. Interpretation of the estimated linear coefficients and associated inferential tests requires that well known assumptions about the model specification and properties of the error distribution have been satisfied. If the assumptions are met then the relationship can between Y and covariates can be summarized with the few parameters estimated.

Within public health, anthropometry measures are typically employed as indirect indicators of current, past, or future poor health. For example, low height-for-age is an indirect indicator for chronic undernutrition, and a high BMI score is an indirect indicator of percent body fat, which in turn is a leading indicator of future chronic disease outcomes (Manson et al. 1995). The use of regression models in this context forces attention away from the mean of distributions and instead focuses on questions about the upper or lower tail, depending on the measure under study. Indeed, for many anthropometry measures there are national and international reference distributions with threshold values that define the range of tail values of concern. For example, we could define thresholds y^{t_1} and y^{t_2} such that individuals with $Y \leq y^{t_1}$ are extreme cases needing immediate interventions and those with $y^{t_1} < Y \leq y^{t_2}$ are of moderate concern.

Because of this focus on the tails of the distribution, conditional mean models subsequently gave way to the use of binary outcome regression models and quantile regression (for example, Wei et al. 2005). Binary outcome models are still conditional mean models but the outcome variable can now be one of the tail probabilities of interest. We can transform the continuous anthropometry measure Y into an ordinal outcome,

$$y_i^* = \begin{cases} 1 & \text{if } y_i \leq y^{t_1} \\ 2 & \text{if } y^{t_1} < y_i \leq y^{t_2} \\ 3 & \text{if } y_i > y^{t_2} \end{cases}$$

where 1 indicates extreme poor health, 2 indicate moderate poor health, and 3 indicates normal health. Regression models in this case are used to estimate the change in the probability of an individual being assigned to one of the categories conditional on their covariate value, $Pr(y_i^* \leq j|x_i) = F(x_i^T \gamma_j)$. One of the primary attractions of this approach is that the resulting estimates can be used to assess odds ratios or changes in prevalence conditional on covariate values. Since the anthropometry measures are usually collected to set policy goals in terms of prevalence outcomes, the interpretation of model results feed nicely into the policy context.

Quantile regression is a more recent approach in anthropometry studies but it has rapidly gained in popularity. In this case, parameter estimates measure the change in the τth quantile given a unit change in X,

$$Q_y[\tau|x_i] = \beta_0(\tau) + \beta_1(\tau)x_i.$$

Quantile regression loses the direct connection to prevalence because fixed thresholds such as y^{t_1} used in the binary model cannot be imposed. Instead, the share of population (τ) is fixed and variation in the quantile $Q_y[\tau|x_i]$ is assessed. Quantile regression has several desirable properties that yield important insights beyond prevalence. The parameter estimates are robust to outliers and the set of estimates $\hat{\beta}(\tau)$ with $\tau \in \{0, 1\}$ can be used to understand how the entire distribution of Y shifts and changes shape when conditioning on X (Koenker 2005).

One of the main findings in Sweeney et al. (2012), is that studies based on results using both quantile regression and ordinal regression can yield important insights that would be missed using either method in isolation. One problem in implementing an analysis using both models is that the conditional variance in Y over the domain of X is typically non-constant when dealing with individual level data such as that used in anthropometry. For quantile regression this does not cause a problem. Estimated standard errors will be slightly less efficient and the estimated coefficient $\hat{\beta}(\tau)$ are still consistent. But the estimated coefficients $\hat{\gamma}$ in binary models will be inconsistent (Greene 2003). Suppose we are interested in estimating a binary outcome model with continuous index function $y_i = x_i^T\gamma + \epsilon_i$. This can be formulated as

$$Pr(y_i < y^{t_1}|x_i) = Pr(x_i^T\gamma + y^{t_1} + \epsilon < 0) = Pr(\epsilon > x^T\gamma^*).$$

The non-central probability evaluation is absorbed into the intercept term. Under heteroskedasticity of an unknown form, $\sigma(x)$, we get instead,

$$y_i = x_i^T\gamma^* + \sigma(x_i)\epsilon_i$$

and the probability model is,

$$Pr(y_i < -y^{t_1}|x_i) = Pr\left(\epsilon_i > \frac{x_i^T\gamma^*}{\sigma(x_i)}\right)$$

The presence of $\sigma(x_i)$ in the probability evaluation is what leads to bias (finite samples) and inconsistency (large sample) of estimates. We suspect these problems are fairly widespread in applications of binary outcome models to anthropometry data.

The proposed solution in Sweeney et al. (2012) is to use the following estimation strategy:

1. Fit a median regression, recover the residuals, and use,

$$\hat{\sigma_i} = \sum_{j=1}^{n} W_{ij}|\hat{e}_{ij}| \quad i = 1, 2, \ldots, n$$

   to estimate the scaling effects, $\hat{\sigma_i}$.

2. Construct $\mathbf{y}^* = [I(\mathbf{y} \leq y^{t_1}), I(\mathbf{y} \leq y^{t_2})]$ and

$$\mathbf{X}^* = \mathbf{I}_2 \otimes (diag(\hat{\sigma}^{-1})\mathbf{X})$$

   and then fit an ordinal logit or probit model of $\mathbf{y}^*$ on $\mathbf{X}^*$.

3. Use average predictive comparisons to recover probabilities that are most relevant for the problem under study.

3. Simulation Study

The two-stage estimator proposed above is motivated by Zhao's (2001) research on the efficiency of the median regression estimator under heteroskedasticity of unknown form. In this case we are instead focused on using Zhao's strategy to estimate σ_i and then employ

that estimate in the binary outcome model to reduce bias. The simulation study that follows provides some sense of how the estimation strategy performs under different forms of heteroskedasticity.

We follow the same general Monte Carlo framework as Zhao (2001) but in addition to continuous covariates we also include a binary covariate. The three different heteroskedasticity functional forms used are:

Type I:

$$\sigma^{(1)} = a_1 e^{a_2|X'\beta|}$$

Type II:

$$\sigma^{(2)} = 1 + 3e^{-\frac{(X'\beta+5)^2}{100}}$$

Type III:

$$\sigma^{(3)} = \begin{cases} 0.75 & \text{if } min(X'\beta) \leq X'\beta < Q_{0.25}(X'\beta) \\ 1.5 & \text{if } Q_{0.25}(X'\beta) \leq X'\beta < Q_{0.5}(X'\beta) \\ 3 & \text{if } Q_{0.5}(X'\beta) \leq X'\beta < Q_{0.75}(X'\beta) \\ 2 & \text{if } Q_{0.75}(X'\beta) \leq X'\beta \leq max(X'\beta) \end{cases}$$

In the above equations, $a_1 = 1$, $a_2 = 0.2$, and $Q_p()$ is a quantile function. The design matrix includes elements $X = [1|x_1|x_2|x_3]$ where $x_1 = U + 0.2V_1$, $x_2 = 0.2U + V$, and $x_3 = I(V_2 < 0.4)$ are constructed from the random numbers, U a normal variate with mean 5 and standard deviation 9, V_1 a rectangular variate with bounds 0 and 4, and V_2 a rectangular variate with bounds 0 and 1. The parameter vector $\beta = \{0.1, -0.25, 0.25, 0.1\}$. Visual characterizations of each type of heteroskadasticity are displayed in Figure 1.

The results here are based on 400 simulation cycles with each simulated data set having 500 observations. For each simulation cycle, a pure disturbance term is defined $u = N(0, 1)$, and four continuous dependent variables are defined:

1. No heteroskedasticity: $y = X'\beta + u$
2. Heteroskedasticity I: $y^{(1)} = X'\beta + \sigma^{(1)}u$
3. Heteroskedasticity II: $y^{(2)} = X'\beta + \sigma^{(2)}u$
4. Heteroskedasticity IIII: $y^{(3)} = X'\beta + \sigma^{(3)}u$

For each of the continuous variables, discrete ordinal variables with three outcomes can be recovered using the thresholds -2 and 0.5 defined as y^*, $y^{*(1)}$, $y^{*(2)}$, and $y^{*(3)}$. The two-stage estimation is then pursued as described above. The first stage is a median regression using X and one of the three heteroskedastic continuous dependent variables. Residuals $\hat{u}$ are recovered from a median regression. The scaling vector, $\hat{\sigma}$ is estimated as $\sum_{j=1}^{n} W_{ij} f(\hat{e}_{ij})$ where the functional form $f(u)$ is either absolute value ($|u|$), squared residuals (u^2), or or a transformation between squaring and absolute value ($u^{4/3}$). Given, $\hat{\sigma}$, the design matrix is rescaled using $\mathbf{X}^* = \mathbf{I}_2 \otimes (diag(\hat{\sigma}^{-1})\mathbf{X})$. Estimates of the parameters of the ordinal models use a design matrix approach suggested by Winship and Mare (1984). Finally, estimates of β are stored from the ordinal regression. Below we present tables showing relative consistency ($\beta^{(two-stage)}/\beta^{(true)}$). An unbiased measure would have a relative consistency of 1, values less than one indicate downward bias, and greater than one upward bias.

For the ordinal regression models as specified, the expected relationships under no heteroskedasticity are: $\beta_{01} < \beta_{02}$, $\beta_{11} = \beta_{12} = -0.25$, $\beta_{21} = \beta_{22} = 0.25$, and $\beta_{31} = \beta_{32} = 0.1$. The results in Tables 1, 2, and 3 show in the first column the pure effects of heteroskedasticity of each type. The βs are consistently underestimated with bias of as much as 90%. The second column shows estimates after rescaling by the true $\sigma^{()}$. The last three columns in each table show the performance of the two-stage estimator using either $\hat{u}^2$, $\hat{u}^{4/3}$, or $|\hat{u}|$. The use of $\hat{u}^{4/3}$ yields the best results with much reduced bias compared to the unadjusted estimates (column 1). Notice that $\hat{u}^2$ tends to overestimate and on the most difficult type of heteroskedasticity (the step function, type III) the estimator occasionally explodes and means of the simulations are pushed to extremely high values (the medians are in the range of the other two-stage estimators).

Raw parameter estimates from an ordinal model are not directly interpretable (or are at least difficult to interpret). For each of the sets of simulated parameters, we used average predictive comparison (see Gelman and Pardoe 2007) to examine the marginal effects of a covariate on a particular type of probability. Those results are displayed in Figures 2, 3, and 4. The dot is the mean change in probability from a one unit change in the covariate. The horizontal lines indicate the 10th to 90th quantile of the simulated values. Notice that while the two-stage estimator is not perfect in terms of removing bias, the multiple comparisons involved in generating the average predictive comparisons tend to cancel out most of the remaining bias and appear quite robust to the presence of heteroskedasticity after the two-stage adjustment. In contrast the average predictive comparisons based on the unadjusted estimates clearly show bias. Policy interpretations based on the unadjusted estimates would thus yield the wrong expectations about changes in prevalence from an intervention on covariate X.

4. Conclusions

Anthropometric measures are some of the most important tools public health researchers have available to them as they facilitate quick and minimally invasive estimates of the health of a population. When it comes to statistically modeling the variation in specific anthropometric measures, however, some important considerations arise - namely issues related to biased and inconsistent estimates. Here we presented results from a two-stage estimator that accounts for some of the complex distributional features found in anthropometric data. If this approach is employed when anthropometric variation is of interest the resulting models are more likely to represent the true features of a population.

REFERENCES

Gelman, A. and Pardoe, I. (2007) "Average predictive comparisons for models with nonlinearity, interactions, and variance components" *Sociological Methodology*, 37(1), 23–51.

Greene W. (2003), *Econometric Analysis*, Upper Saddle River, NJ: Prentice Hall.

Koenker, R. (2005), *Quantile Regression*, Cambridge, UK: Cambridge University Press.

Manson, J.E., Willett, W., Stampfer, M.J., Colditz, G.A., Hunter, D.J., Hankinson, S.E., Hennekens, C.H., and Speizer, F.E. (1995), "Body Weight and Mortality among Women," *New England Journal of Medicine*, 333, 677–685.

Sweeney, S., Davenport, D., and Grace, K. (2012), "Combining insights from quantile and ordinal regression: Child malnutrition in Guatemala," *Economics and Human Biology* (In Press)
`http://dx.doi.org/10.1016/j.ehb.2012.06.001`

Wei, Y., Pere, A., Koenker, R. and He, X. (2005) "Quantile regression methods for reference growth charts," *Statistics in Medicine*, 25(8), 1369–1382.

Winship, C. and Mare, R. (1984) "Regression models with ordinal variables" *American Sociological Review*, 49(4), 512–525.

Zhao, Q. (2001) "Asymptotically efficient median regression in the presence of heteroskedasticity of unknown form" *Econometric Theory*, 17, 765–784.

Parm	no correction	true σ	$\hat{u}^2$ $\hat{\sigma}$	$\hat{u}^{4/3}$ $\hat{\sigma}$	$\lvert\hat{u}\rvert$ $\hat{\sigma}$
β_{01}	0.350	0.993	1.265	0.996	0.822
β_{02}	0.278	0.994	0.808	0.812	0.748
β_{11}	0.225	0.987	1.275	0.971	0.762
β_{12}	0.112	0.978	0.969	0.827	0.665
β_{21}	0.197	0.982	1.229	0.926	0.716
β_{22}	0.079	0.976	0.925	0.809	0.658
β_{31}	0.347	0.971	1.228	0.978	0.815
β_{32}	0.194	0.973	1.009	0.841	0.670

Table 1: Relative consistency for heteroskedasticity type I

Parm	no correction	true σ	$\hat{u}^2$ $\hat{\sigma}$	$\hat{u}^{4/3}$ $\hat{\sigma}$	$\lvert\hat{u}\rvert$ $\hat{\sigma}$
β_{01}	0.328	0.985	1.284	0.991	0.805
β_{02}	0.119	0.927	1.059	0.876	0.708
β_{11}	0.310	0.991	1.285	0.985	0.800
β_{12}	0.279	0.977	1.242	0.963	0.777
β_{21}	0.302	0.989	1.266	0.967	0.788
β_{22}	0.266	0.953	1.187	0.944	0.771
β_{31}	0.306	0.950	1.215	0.944	0.771
β_{32}	0.283	1.003	1.312	0.972	0.765

Table 2: Relative consistency for heteroskedasticity type II

Parm	no correction	true σ	$\hat{u}^2$ $\hat{\sigma}$	$\hat{u}^{4/3}$ $\hat{\sigma}$	$\lvert\hat{u}\rvert$ $\hat{\sigma}$
β_{01}	0.392	0.988	0.880	0.884	0.807
β_{02}	0.648	0.988	0.851	0.825	0.792
β_{11}	0.609	0.993	0.900	0.910	0.854
β_{12}	0.611	0.994	0.946	0.889	0.828
β_{21}	0.716	0.987	0.916	0.926	0.881
β_{22}	0.610	0.989	0.907	0.860	0.808
β_{31}	0.459	0.962	0.782	0.816	0.782
β_{32}	0.641	1.016	0.993	0.927	0.858

Table 3: Relative consistency for heteroskedasticity type III

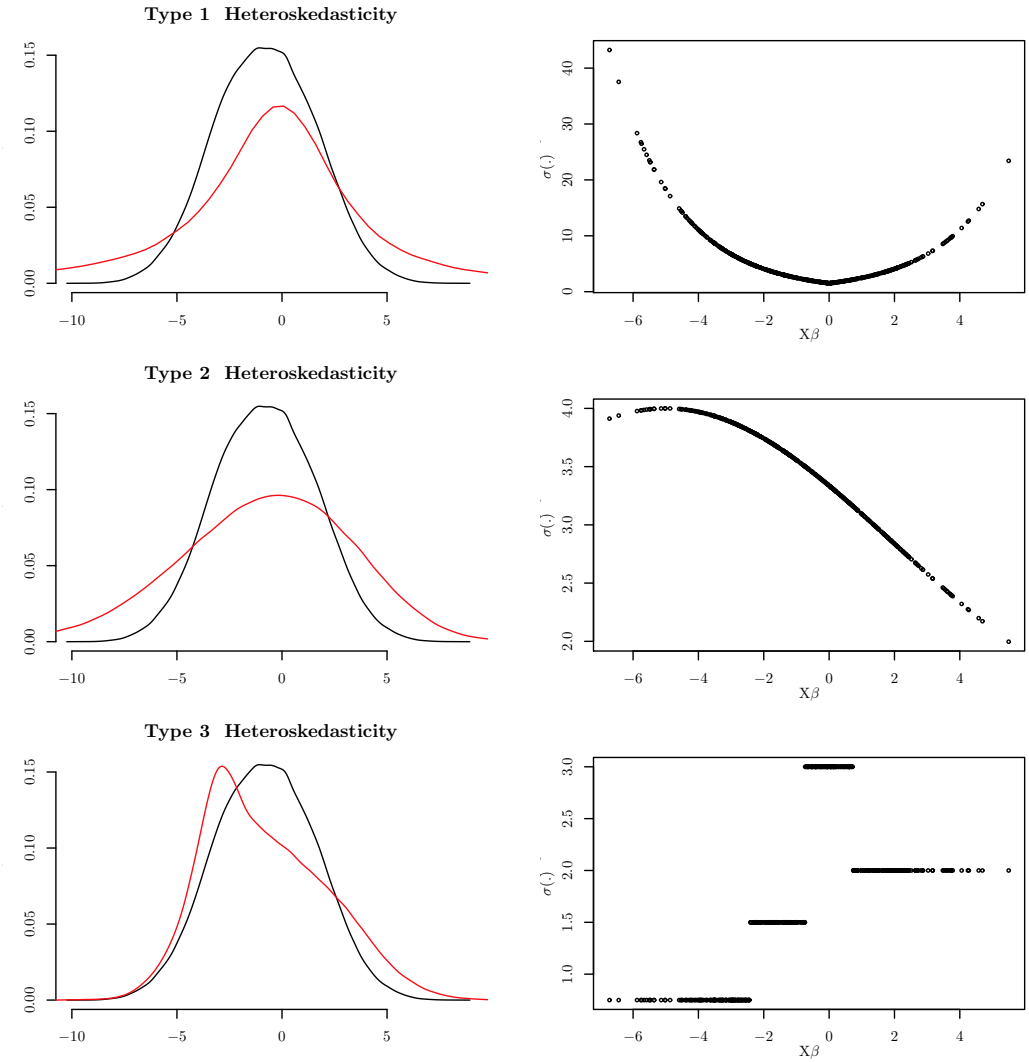

Figure 1: Three types of heteroskedasticity. In density curves the black line is pure disturbance and the red line is heteroskedastic.

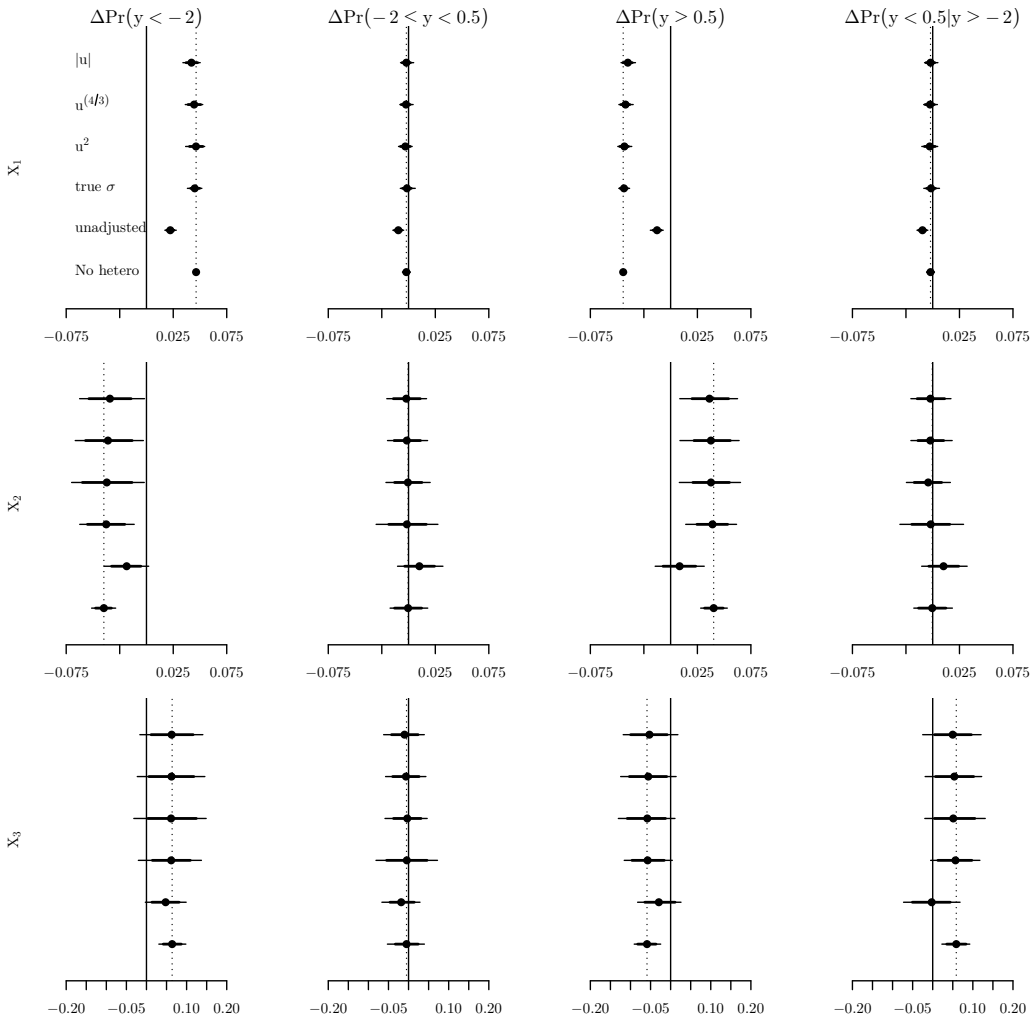

Figure 2: APCs for Type I heteroskedastcity.

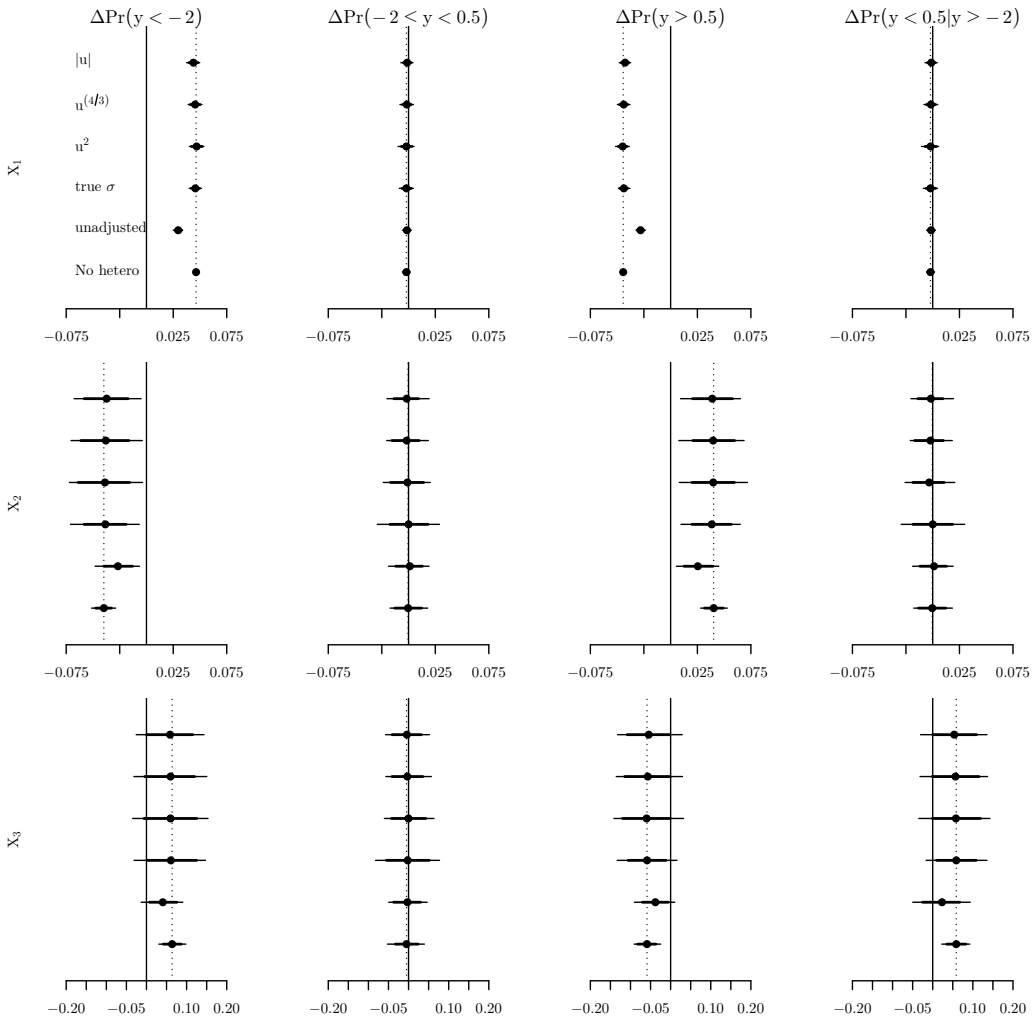

Figure 3: APCs for Type II heteroskedastcity.

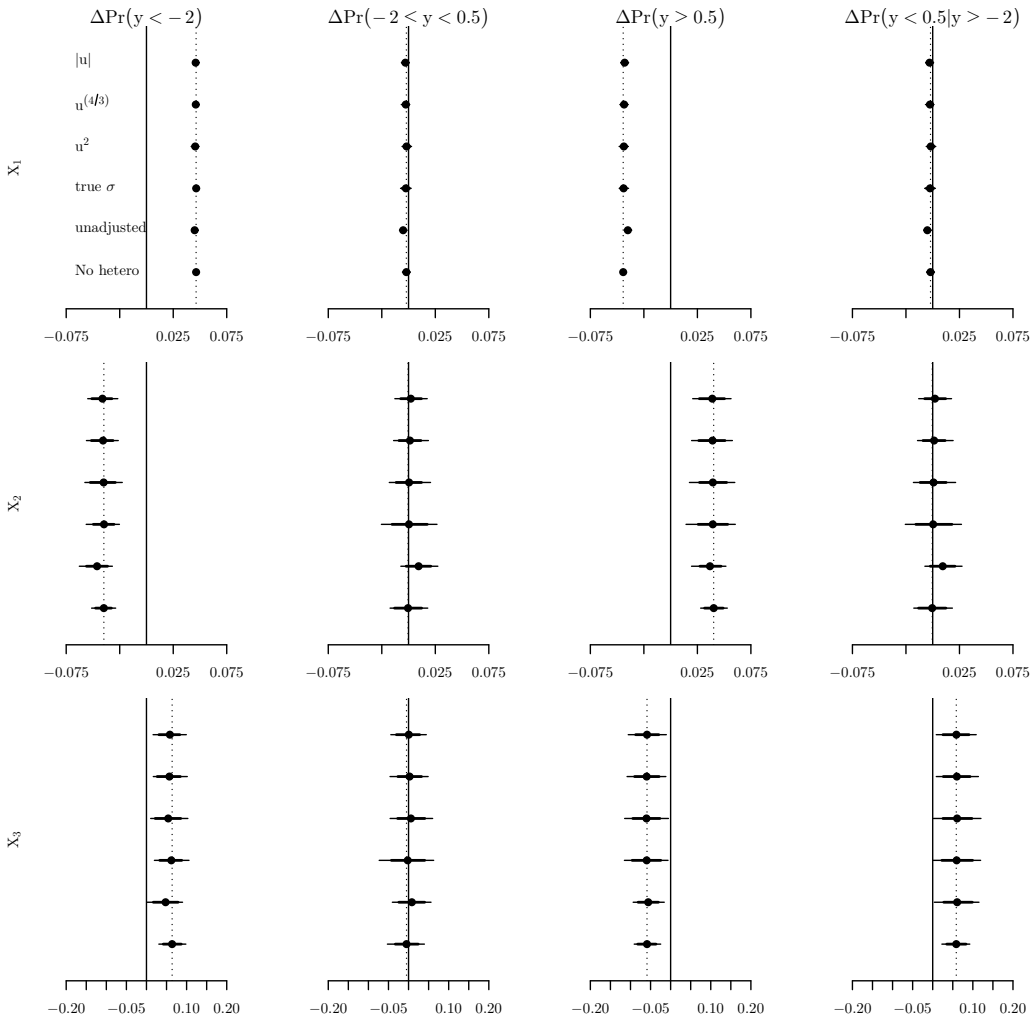

Figure 4: APCs for Type III heteroskedastcity.