Quantile regression

Christopher F Baum

ECON 8823: Applied Econometrics

Boston College, Spring 2015

Motivation

Standard linear regression techniques summarize the average relationship between a set of regressors and the outcome variable based on the conditional mean function $E(y|x)$. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of y. Quantile regression provides that capability.

Analogous to the conditional mean function of linear regression, we may consider the relationship between the regressors and outcome using the conditional *median* function $Q_q(y|x)$, where the median is the 50^{th} percentile, or quantile q, of the empirical distribution. The quantile $q \in (0,1)$ is that y which splits the data into proportions q below and $1-q$ above: $F(y_q) = q$ and $y_q = F^{-1}(q)$: for the median, $q = 0.5$.

Motivation

Standard linear regression techniques summarize the average relationship between a set of regressors and the outcome variable based on the conditional mean function $E(y|x)$. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of y. Quantile regression provides that capability.

Analogous to the conditional mean function of linear regression, we may consider the relationship between the regressors and outcome using the conditional *median* function $Q_q(y|x)$, where the median is the 50^{th} percentile, or quantile q, of the empirical distribution. The quantile $q \in (0,1)$ is that y which splits the data into proportions q below and $1-q$ above: $F(y_q) = q$ and $y_q = F^{-1}(q)$: for the median, $q = 0.5$.

If ϵ_i is the model prediction error, OLS minimizes $\sum_i e_i^2$. Median regression, also known as least-absolute-deviations (LAD) regression, minimizes $\sum_i |e_i|$. Quantile regression minimizes a sum that gives asymmetric penalties $(1 - q)|e_i|$ for overprediction and $q|e_i|$ for underprediction. Although its computation requires linear programming methods, the quantile regression estimator is asymptotically normally distributed.

Median regression is more robust to outliers than least squares regression, and is semiparametric as it avoids assumptions about the parametric distribution of the error process.

If ϵ_i is the model prediction error, OLS minimizes $\sum_i e_i^2$. Median regression, also known as least-absolute-deviations (LAD) regression, minimizes $\sum_i |e_i|$. Quantile regression minimizes a sum that gives asymmetric penalties $(1 - q)|e_i|$ for overprediction and $q|e_i|$ for underprediction. Although its computation requires linear programming methods, the quantile regression estimator is asymptotically normally distributed.

Median regression is more robust to outliers than least squares regression, and is semiparametric as it avoids assumptions about the parametric distribution of the error process.

Just as regression models conditional moments, such as predictions of the conditional mean function, we may use quantile regression to model conditional quantiles of the joint distribution of y and x.

Let $\hat{y}(x)$ denote the predictor function and $e(x) = y - \hat{y}(x)$ denote the prediction error. Then

$$L(e(x)) = L(y - \hat{y}(x))$$

denotes the loss associated with the prediction errors. If $L(e) = e^2$, we have squared error loss, and least squares is the optimal predictor.

If $L(e) = |e|$, the optimal predictor is the conditional median, $\text{med}(y|x)$, and the optimal predictor is that $\hat{\beta}$ which minimizes $\sum_i |y_i - x_i'\beta|$.

Just as regression models conditional moments, such as predictions of the conditional mean function, we may use quantile regression to model conditional quantiles of the joint distribution of y and x.

Let $\hat{y}(x)$ denote the predictor function and $e(x) = y - \hat{y}(x)$ denote the prediction error. Then

$$L(e(x)) = L(y - \hat{y}(x))$$

denotes the loss associated with the prediction errors. If $L(e) = e^2$, we have squared error loss, and least squares is the optimal predictor.

If $L(e) = |e|$, the optimal predictor is the conditional median, $\text{med}(y|x)$, and the optimal predictor is that $\hat{\beta}$ which minimizes $\sum_i |y_i - x_i'\beta|$.

Just as regression models conditional moments, such as predictions of the conditional mean function, we may use quantile regression to model conditional quantiles of the joint distribution of y and x.

Let $\hat{y}(x)$ denote the predictor function and $e(x) = y - \hat{y}(x)$ denote the prediction error. Then

$$L(e(x)) = L(y - \hat{y}(x))$$

denotes the loss associated with the prediction errors. If $L(e) = e^2$, we have squared error loss, and least squares is the optimal predictor.

If $L(e) = |e|$, the optimal predictor is the conditional median, $\text{med}(y|x)$, and the optimal predictor is that $\hat{\beta}$ which minimizes $\sum_i |y_i - x_i'\beta|$.

Both the squared-error and absolute-error loss functions are symmetric; the sign of the prediction error is not relevant. If the quantile q differs from 0.5, there is an asymmetric penalty, with increasing asymmetry as q approaches 0 or 1.

Advantages of quantile regression (QR): while OLS can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

Furthermore, QR is invariant to monotonic transformations, such as $\log(\cdot)$, so the quantiles of $h(y)$, a monotone transform of y, are $h(Q_q(y))$, and the inverse transformation may be used to translate the results back to y. This is not possible for the mean as $E[h(y)] \neq h[E(y)]$.

Both the squared-error and absolute-error loss functions are symmetric; the sign of the prediction error is not relevant. If the quantile q differs from 0.5, there is an asymmetric penalty, with increasing asymmetry as q approaches 0 or 1.

Advantages of quantile regression (QR): while OLS can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

Furthermore, QR is invariant to monotonic transformations, such as $\log(\cdot)$, so the quantiles of $h(y)$, a monotone transform of y, are $h(Q_q(y))$, and the inverse transformation may be used to translate the results back to y. This is not possible for the mean as $E[h(y)] \neq h[E(y)]$.

Both the squared-error and absolute-error loss functions are symmetric; the sign of the prediction error is not relevant. If the quantile q differs from 0.5, there is an asymmetric penalty, with increasing asymmetry as q approaches 0 or 1.

Advantages of quantile regression (QR): while OLS can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

Furthermore, QR is invariant to monotonic transformations, such as $\log(\cdot)$, so the quantiles of $h(y)$, a monotone transform of y, are $h(Q_q(y))$, and the inverse transformation may be used to translate the results back to y. This is not possible for the mean as $E[h(y)] \neq h[E(y)]$.

Implementation

The quantile regression estimator for quantile q minimizes the objective function

$$Q(\beta_q) = \sum_{i:y_i \geq x_i'\beta}^{N} q|y_i - x_i'\beta_q| + \sum_{i:y_i < x_i'\beta}^{N} (1-q)|y_i - x_i'\beta_q|$$

This nondifferentiable function is minimized via the simplex method, which is guaranteed to yield a solution in a finite number of iterations. Although the estimator is proven to be asymptotically normal with an analytical VCE, the expression for the VCE is awkward to estimate. Bootstrap standard errors are often used in place of analytic standard errors.

Implementation

The quantile regression estimator for quantile q minimizes the objective function

$$Q(\beta_q) = \sum_{i:y_i \geq x_i'\beta}^{N} q|y_i - x_i'\beta_q| + \sum_{i:y_i < x_i'\beta}^{N} (1-q)|y_i - x_i'\beta_q|$$

This nondifferentiable function is minimized via the simplex method, which is guaranteed to yield a solution in a finite number of iterations. Although the estimator is proven to be asymptotically normal with an analytical VCE, the expression for the VCE is awkward to estimate. Bootstrap standard errors are often used in place of analytic standard errors.

The Stata command `qreg` estimates a multivariate quantile regression with analytic standard errors. By default the quantile is 0.5, the median. A different quantile may be specified with the `quantile()` option.

The `bsqreg` command estimates the model with bootstrap standard errors, retaining the assumption of independent errors but relaxing the assumption of identically distributed errors; thus they are analogous to robust standard errors in linear regression.

The Stata command `qreg` estimates a multivariate quantile regression with analytic standard errors. By default the quantile is 0.5, the median. A different quantile may be specified with the `quantile()` option.

The `bsqreg` command estimates the model with bootstrap standard errors, retaining the assumption of independent errors but relaxing the assumption of identically distributed errors; thus they are analogous to robust standard errors in linear regression.

The `iqreg` command performs interquantile range regression: regression of the difference in quantiles. By default, the quantiles (0.25, 0.75) produce interquartile range estimates. Bootstrap standard errors are produced.

The `sqreg` command produces QR estimates for several values of q simultaneously, allowing for differences between QR coefficients for different quantiles to be tested. Bootstrap standard errors are produced.

The `iqreg` command performs interquantile range regression: regression of the difference in quantiles. By default, the quantiles (0.25, 0.75) produce interquartile range estimates. Bootstrap standard errors are produced.

The `sqreg` command produces QR estimates for several values of q simultaneously, allowing for differences between QR coefficients for different quantiles to be tested. Bootstrap standard errors are produced.

Illustration

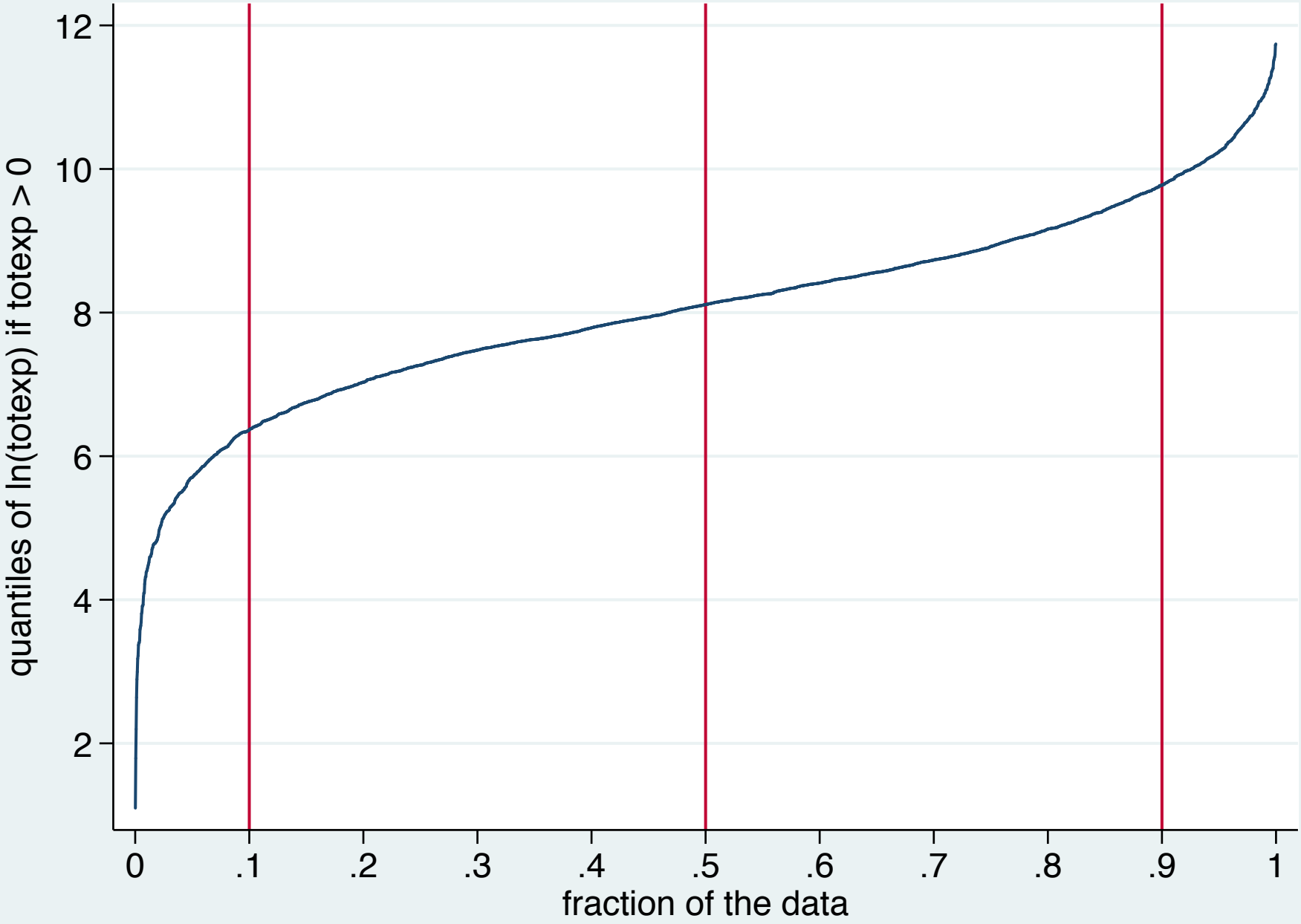
We illustrate using data from the Medical Expenditure Panel Survey (MEPS), modeling the log of total medical expenditure for Medicare (elderly) patients. Explanatory variables include an indicator for supplementary private insurance, a health status variable and three demographic measures: age, female, and white.

```
. use mus03data, clear
. drop if mi(ltotexp)
(109 observations deleted)
. su ltotexp suppins totchr age female white, sep(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ltotexp	2955	8.059866	1.367592	1.098612	11.74094
suppins	2955	.5915398	.4916322	0	1
totchr	2955	1.808799	1.294613	0	7
age	2955	74.24535	6.375975	65	90
female	2955	.5840948	.4929608	0	1
white	2955	.9736041	.1603368	0	1

Using Nick Cox's `qplot` command (*Stata J.*), we illustrate the empirical CDF of log total expenditures, which appears reasonably symmetric. Note that the 10th, 50th and 90th quantiles are roughly 6, 8, and 10 on the log scale.

```
. qplot ltotexp, recast(line) ylab(,angle(0)) ///
> xlab(0(0.1)1) xline(0.5) xline(0.1) xline(0.9)
```


quantiles of ln(totexp) if totexp > 0
12
10
8
6
4
2
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
fraction of the data

The median regression, with all covariates but `female` statistically significant:

```
. qreg ltotexp suppins totchr age female white, nolog
Median regression                               Number of obs =      2955
  Raw sum of deviations 3110.961 (about 8.111928)
  Min sum of deviations 2796.983                Pseudo R2     =    0.1009
```

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	.2769771	.0535936	5.17	0.000	.1718924	.3820617
totchr	.3942664	.0202472	19.47	0.000	.3545663	.4339664
age	.0148666	.0041479	3.58	0.000	.0067335	.0229996
female	-.0880967	.0532006	-1.66	0.098	-.1924109	.0162175
white	.4987457	.1630984	3.06	0.002	.1789474	.818544
_cons	5.648891	.341166	16.56	0.000	4.979943	6.317838

Given the equivariance property of QR (which depends on the correct specification of the conditional quantile function), we may calculate marginal effects in terms of the underlying level variable:

```
 . mat b = e(b)
. qui predict double xb
. qui gen double expxb = exp(xb)
. su expxb, mean
. mat b = r(mean) * b
. mat li b, ti("Marginal effects ($) on total medical expenditures")
b[1,6]:  Marginal effects ($) on total medical expenditures
       suppins      totchr         age      female       white       _cons
y1    1037.755   1477.2049   55.700813  -330.07346   1868.6593     21164.8
```

Implying, for instance, that expenditures increase by $55.70 per year, *cet. par.*, and that one more chronic condition (as captured by `totchr`) increases expenditures by $1,477.20, *cet. par.*

Given the equivariance property of QR (which depends on the correct specification of the conditional quantile function), we may calculate marginal effects in terms of the underlying level variable:

```
 . mat b = e(b)
. qui predict double xb
. qui gen double expxb = exp(xb)
. su expxb, mean
. mat b = r(mean) * b
. mat li b, ti("Marginal effects ($) on total medical expenditures")
b[1,6]:  Marginal effects ($) on total medical expenditures
       suppins      totchr         age      female       white       _cons
y1    1037.755   1477.2049   55.700813  -330.07346   1868.6593     21164.8
```

Implying, for instance, that expenditures increase by $55.70 per year, *cet. par.*, and that one more chronic condition (as captured by `totchr`) increases expenditures by $1,477.20, *cet. par.*

We may also compare OLS and QR estimates of this model at different quantiles:

```
 . eststo clear
. eststo, ti("OLS"): qui reg ltotexp suppins totchr age female white, robust
(est1 stored)
. foreach q in 0.10 0.25 0.50 0.75 0.90 {
  2.         eststo, ti("Q(`q´)"): qui qreg ltotexp suppins totchr age female w
> hite, q(`q´) nolog
  3. }
(est2 stored)
(est3 stored)
(est4 stored)
(est5 stored)
(est6 stored)
. esttab using 82303ht.tex, replace nonum nodep mti drop(_cons) ///
> ti("Models of log total medical expenditure via OLS and QR")
(output written to 82303ht.tex)
```

Table : Models of log total medical expenditure via OLS and QR

	OLS	Q(0.10)	Q(0.25)	Q(0.50)	Q(0.75)	Q(0.90)
suppins	0.257***	0.396***	0.386***	0.277***	0.149*	-0.0143
	(5.44)	(4.90)	(6.64)	(5.17)	(2.44)	(-0.16)
totchr	0.445***	0.539***	0.459***	0.394***	0.374***	0.358***
	(25.56)	(17.67)	(20.93)	(19.47)	(16.18)	(10.36)
age	0.0127***	0.0193**	0.0155***	0.0149***	0.0183***	0.00592
	(3.52)	(3.08)	(3.45)	(3.58)	(3.86)	(0.84)
female	-0.0765	-0.0127	-0.0161	-0.0881	-0.122*	-0.158
	(-1.65)	(-0.16)	(-0.28)	(-1.66)	(-2.01)	(-1.74)
white	0.318*	0.0734	0.338	0.499**	0.193	0.305
	(2.34)	(0.30)	(1.91)	(3.06)	(1.04)	(1.10)
N	2955	2955	2955	2955	2955	2955

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We see that the effect of supplementary insurance differs considerably, having a strong effect on expenditures at lower quantiles. The median estimate is similar to the OLS point estimate. For the health status variable, the effects are much stronger at lower quantiles, with the OLS effect quite far from the median estimate.

We can also formally test the equivalence of the quantile estimates across quantiles with `bsqreg`, which allows us to estimate the model for each of several quantiles in a single model, allowing for cross-equation hypothesis tests.

We see that the effect of supplementary insurance differs considerably, having a strong effect on expenditures at lower quantiles. The median estimate is similar to the OLS point estimate. For the health status variable, the effects are much stronger at lower quantiles, with the OLS effect quite far from the median estimate.

We can also formally test the equivalence of the quantile estimates across quantiles with `bsqreg`, which allows us to estimate the model for each of several quantiles in a single model, allowing for cross-equation hypothesis tests.

```
. qui sqreg ltotexp suppins totchr age female white, nolog q(0.1 0.25 0.5 0.75
> 0.9)
. test [q25=q50=q75]: suppins
 ( 1)  [q25]suppins - [q50]suppins = 0
 ( 2)  [q25]suppins - [q75]suppins = 0
       F(  2,  2949) =    7.41
            Prob > F =    0.0006
. test [q25=q50=q75]: totchr
 ( 1)  [q25]totchr - [q50]totchr = 0
 ( 2)  [q25]totchr - [q75]totchr = 0
       F(  2,  2949) =    5.10
            Prob > F =    0.0061
```

The estimates clearly reject equality of the estimated coefficients for the three quartiles in each case.

```
. qui sqreg ltotexp suppins totchr age female white, nolog q(0.1 0.25 0.5 0.75
> 0.9)
. test [q25=q50=q75]: suppins
 ( 1)  [q25]suppins - [q50]suppins = 0
 ( 2)  [q25]suppins - [q75]suppins = 0
       F(  2,  2949) =    7.41
            Prob > F =    0.0006
. test [q25=q50=q75]: totchr
 ( 1)  [q25]totchr - [q50]totchr = 0
 ( 2)  [q25]totchr - [q75]totchr = 0
       F(  2,  2949) =    5.10
            Prob > F =    0.0061
```

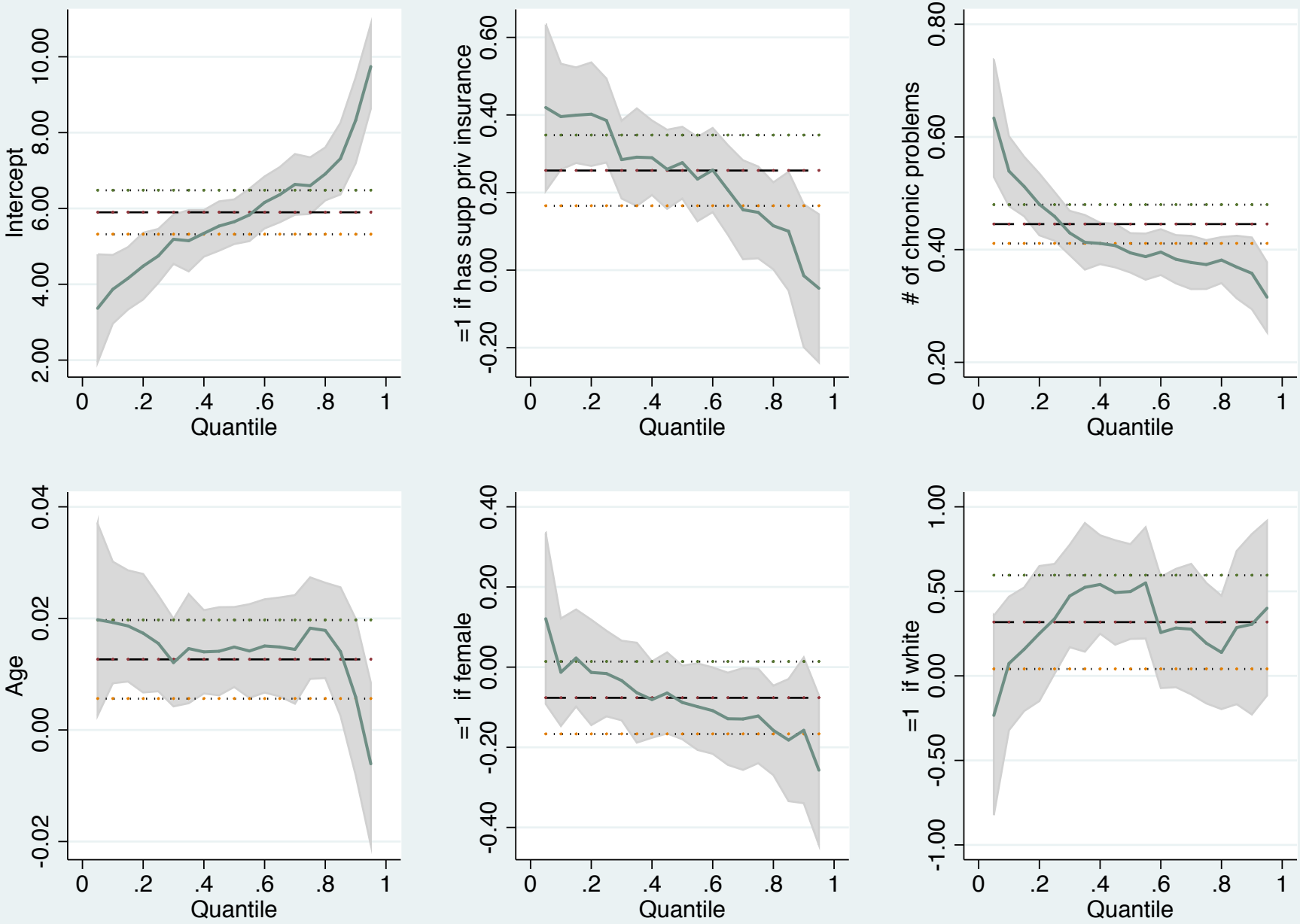
The estimates clearly reject equality of the estimated coefficients for the three quartiles in each case.

Using Azevedo's routine `grqreg`, available from SSC, we can view how each covariate's effects vary across quantiles, and contrast them with the (fixed) OLS estimates:

```
. qreg ltotexp suppins totchr age female white, q(.50) nolog
Median regression                               Number of obs =      2955
  Raw sum of deviations 3110.961 (about 8.111928)
  Min sum of deviations 2796.983                Pseudo R2     =    0.1009
```

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	.2769771	.0535936	5.17	0.000	.1718924	.3820617
totchr	.3942664	.0202472	19.47	0.000	.3545663	.4339664
age	.0148666	.0041479	3.58	0.000	.0067335	.0229996
female	-.0880967	.0532006	-1.66	0.098	-.1924109	.0162175
white	.4987457	.1630984	3.06	0.002	.1789474	.818544
_cons	5.648891	.341166	16.56	0.000	4.979943	6.317838

```
. grqreg, cons ci ols olsci reps(40)
```


Intercept
2.00
4.00
6.00
8.00
10.00
=1 if has supp priv insurance
-0.20
0.00
0.20
0.40
0.60
# of chronic problems
0.20
0.40
0.60
0.80
Age
-0.02
0.00
0.02
0.04
=1 if female
-0.40
-0.20
0.00
0.20
0.40
=1 if white
-1.00
-0.50
0.00
0.50
1.00
0
.2
.4
.6
.8
1
Quantile

The graph illustrates how the effects of private insurance and health status (number of chronic problems) vary over quantiles, and how the magnitude of the effects at various quantiles differ considerably from the OLS coefficient, even in terms of the confidence intervals around each coefficient.

Although quantile regression methods are usually applied to continuous-response data, it is possible to utilize them in the context of count data, such as would appear in a Poisson or negative binomial model. The `qcount` routine of Miranda, available from SSC, implements quantile count regression.

The graph illustrates how the effects of private insurance and health status (number of chronic problems) vary over quantiles, and how the magnitude of the effects at various quantiles differ considerably from the OLS coefficient, even in terms of the confidence intervals around each coefficient.

Although quantile regression methods are usually applied to continuous-response data, it is possible to utilize them in the context of count data, such as would appear in a Poisson or negative binomial model. The `qcount` routine of Miranda, available from SSC, implements quantile count regression.