Gregory Kordas
Last update: October 12, 2020

Lecture 10
Quantile Regression

Abstract: We motivate quantile regression through some examples and present some asymptotic results useful for inference.

> Much of the early history of social statistics, strongly influenced by Quetelet, can be viewed as a search for the "average man" – that improbable man without qualities who could be comfortable with his feet in the ice chest and his hands in the oven. Some of this obsession can be attributed to the seductive appeal of the Gaussian law of errors. Everyone, as Poincare famously quipped, believes in the normal law of errors: the theorists because they believe it is an empirical fact, and the empiricists because they believe that it is a mathematical theorem. Once in the grip of this Gaussian faith, it suffices to learn about means. But sufficiency, despite all its mathematical elegance, should be tempered by a skeptical empiricism: a willingness to peer occasionally outside the cathedral of mathematics and see the world in all its diversity.
> – Roger Koenker, *Quantile Regression.* Cambridge University Press.

Consider the problem of evaluating the effect of surgery on patients. Let y be a health indicator and x be a binary (dummy) indicator of whether or not the patient has received surgery. Assuming for purposes of exposition, that x is the only relevant variable, we can write the simple regression model

$$y = h(x;\beta) + u.$$

If we assume that $E(u|x) = 0$, then $h(x;\beta)$ is a conditional mean function (i.e., $E(y|x) = h(x;\beta)$), while if $\mathrm{Med}(u|x) = 0$, then $h(x;\beta)$ is a conditional median function. Assuming that $h(x;\beta)$ is linear in x and β, we obtain a linear conditional expectation model $E(y|x) = \alpha + \beta X$, or a linear conditional median model $Med(y|x) = \alpha + \beta X$, depending on the kind of restriction we have put on the error term.

In the case of a linear expectation model, we can interpret β a *mean or average, treatment effect* (ATE), i.e., the effect of treatment on the mean of the conditional distribution of the health indicator y. If a median model is assumed, then β is a *median treatment effect* (MTE).

The information contained in the mean or median estimate is interesting, but *very limited.* If β is positive, for example, then we learn that the average person tends to benefit from this procedure, but we learn nothing about the rest of the distribution. To explore the effect of x on the *entire distribution of* y we need to look at other quantiles.

Figure 1 presents some stylized examples of an effect of a treatment x on a health indicator y. The first row of the graph presents the effect of a pure *location shift*, the second row of a pure *scale shift*, and the third row of a *location and scale shift.* Looking first at the location shift model, we see that the effect of treatment x on distribution of health is to shift it to the right (the solid line is the before and the dotted line is the after treatment distribution). In this case, the effect of treatment is the same for the entire distribution, so the *quantile treatment effect* (QTE), graphed to the right, is constant across all quantiles in $(0, 1)$.

The second raw presents the situation in the case where the treatment has a pure scale effect on the distribution of the health indicator. Here the mean (or median) of the distribution is unaffected by the treatment, but the rest of the distribution has changed. In particular, the QTE, graphed to the right, is negative at low quantiles and is strictly increasing as we move to higher quantiles of the conditional distribution. Clearly, a mean or median regression model would miss all this information as β would be insignificantly different from zero.

The last panel presents a combined location and scale effect. The QTE graphed to the right tells us that the treatment is beneficial for the mean person (dotted line) and for people belonging to high conditional quantiles, but is detrimental to those belonging to the lowest quantiles.

The simple location-scale models of Figure 1 are restrictive in that the QTE's so produced as monotone in the estimation quantile. This need not be so. Let $\tau \in (0, 1)$ be the estimation quantile and consider the *quantile regression model*

$$Q_\tau(y|x) = \alpha(\tau) + \beta(\tau)x,$$

where $\alpha(\tau)$ is the intercept and $\beta(\tau)$ is the slope of the τ-th quantile model. In our example, $\beta(\tau)$ is the QTE and can be any function of τ, monotone, constant, or U-shaped, as the situation may be.

In real applications, y depends on a host of variables, but this doesn't present any problems for us as we are free to add any number of regressors we wish, and write

$$Q_\tau(y|x) = x'\beta(\tau), \quad \tau \in (0, 1)$$

where $Q_\tau(y|x)$ is the τ-conditional quantile of y, x is a k-vector of regressors, and $\beta(\tau)$ is the k-vector of coefficients for the τ-th quantile model.

Quantile regression represents an extension of traditional estimation methods that allows for distinct quantile effects; see Koenker and Hallock (2001). The quantile model posits the τ-th quantile of y conditional on x to be, $Q_\tau(y|x) = \alpha(\tau) + x\beta(\tau), 0 < \tau < 1$.

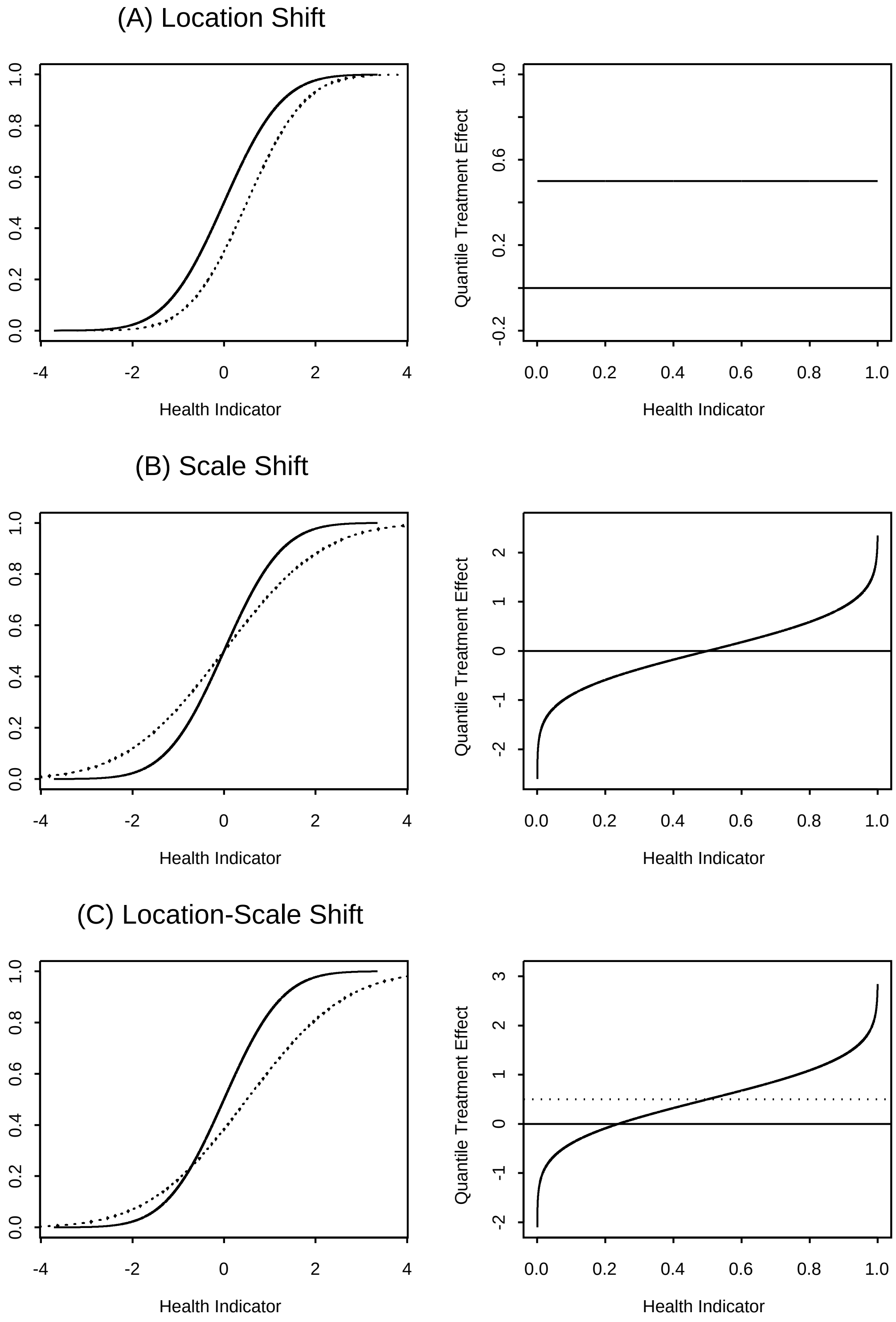

(A) Location Shift
(B) Scale Shift
(C) Location-Scale Shift
Health Indicator
Quantile Treatment Effect
-4
-2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
-0.2
-2
-1
1
2
3

Figure 1

If $\beta(\tau)$ is a constant β, the model reduces to the standard conditional expectation model, $E(y|x) = \alpha + x\beta$, with constant variance errors. When $\beta(\tau)$ depends on τ, the model allows the distribution of y to depend on x in different ways at different parts of the distribution. The traditional linear model can be viewed as a summary of all the quantile effects; that is,

$$\int_0^1 Q_\tau(y|x)\, d\tau = E(y|x).$$

Under this interpretation, traditional conditional mean analysis loses information due to its aggregation of possibly disparate quantile effects. Many different quantile paths, for example, can lead to $\beta_k = 0$. On the one hand, $\beta_k = 0$ can mean x_k does not matter – does not affect the distribution of y. But it can also mean there are *important but compensating quantile effects* relating y and x. In the latter case the single β statistic obscures information about quantile effects. This is especially important when scientific interest concerns differences in the way regressors affect different parts of the distribution. The details provided by the quantiles discriminate between what would be otherwise identical situations.

Example: As an example, consider the estimation of a salary equation. Log salary depends on education, experience and a host of other variables, among which is an indicator of union membership. We write the regression model

$$\log salary = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp^2 + \cdots + \beta_k union + u$$

where u are unobserved attributes like ability, ambition etc. We are interested in measuring the effect of all of theses variables on various parts of the conditional distribution of $\log salary$, and we are particularly interested in the "union effect". In particular, we would like to test the conjecture that unions may have a more beneficial effect on the salaries of people with low u (ability etc.) than those with high u. For this purpose, we write the quantile regression model

$$Q_\tau(\log salary|x) = \beta_0(\tau) + \beta_1(\tau) educ + \beta_2(\tau) exp + \beta_3(\tau) exp^2 + \cdots + \beta_k(\tau) union$$

end estimate it over a grid of quantiles $\tau \in (0, 1)$, say, $\tau = 0.10, 0.25, 0.50, 0.75$, and 0.90 (we could estimate a very fine grid but the paper mentioned below reported only these estimates). It is then customary to plot these estimates across the estimated quantiles, i.e., report the estimation results in graphs, one for each coefficient. For example, graphing $\beta_1(\tau)$ against τ we could see that it is increasing in τ (as the QTE's in Figure 1), so that education would have an increasing important effect as we move up the conditional quantiles of salaries.

Chamberlain (1988) estimated exactly such a model and reported estimates for various coefficients, but here we will only present his estimates for the *union* variable (the interested reader is referred to the paper). Figure 2 plots the $\beta(\tau)$ of the *union* variable

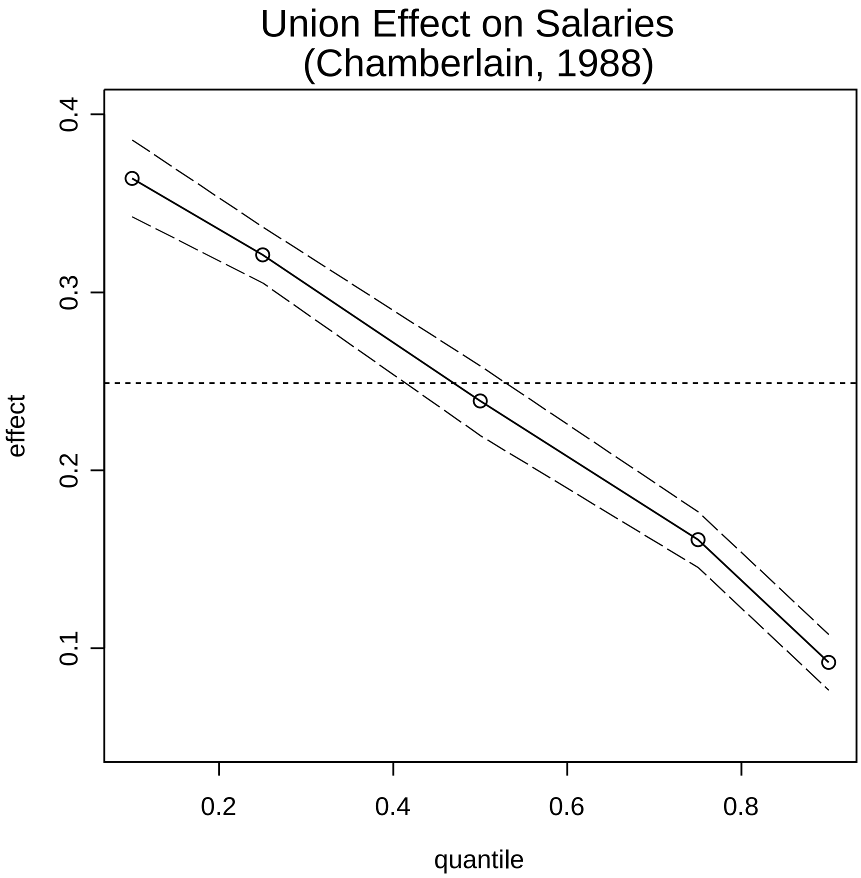


Figure 2. The union effect across quantiles.

(solid line) along with the OLS estimate (dotted line). The dashed lines around $\beta(\tau)$ are pointwise 95% confidence bands.

As we can see the union effect is very important for people low on the conditional distribution, but becomes smaller as we move up to higher conditional quantiles. Since the dependent variable is in logs, the union effect coefficient may be interpreted as the percent increase in wages due to union membership. We see that the mean effect of union membership is a 25% increase, which is similar to the $\tau = 0.50$ median effect. At the low $\tau = 0.10$ quantile this increase is 35%, while at the high $\tau = 0.90$ quantile union membership results in only 10% more salary. We see that union membership is beneficial across the board, but much more so for low ability workers. A single mean or median model would miss all this interesting structure! ■

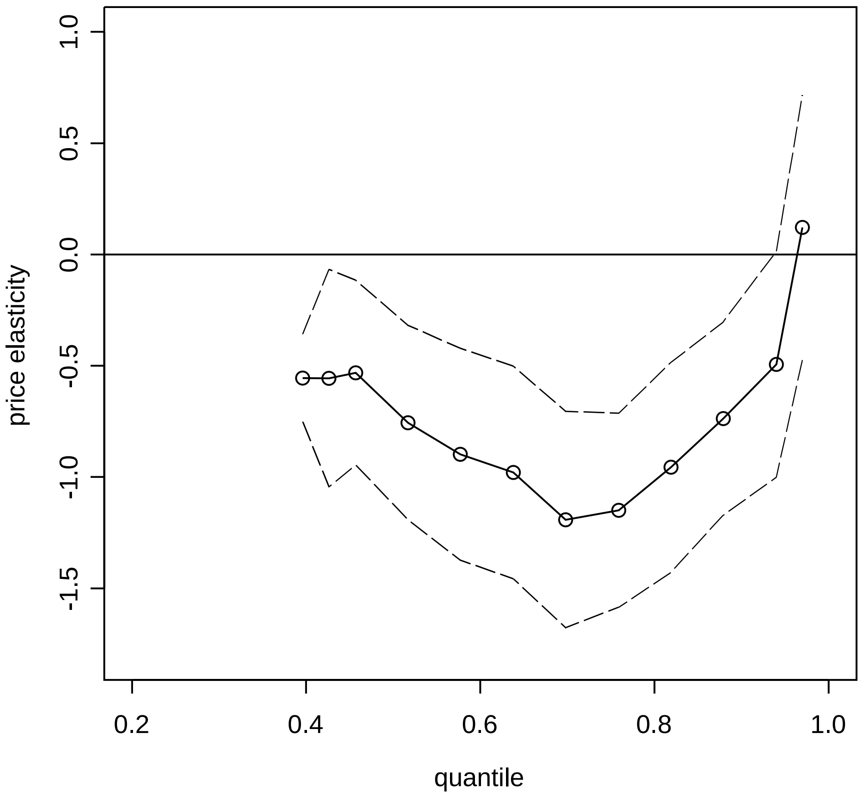


FIGURE 3. Alcohol price elasticity across qualtiles.

Example: As a final example of the usefulness of quantile regression in describing the effect of a set of explanatory variables on the conditional distribution of a response variable, consider estimating the demand for alcohol.

Manning *et. al.* (1995) report estimates of a simple model of the (Marshallian) demand for alcohol. Log-consumption is regressed against log price and and log-income for various quantiles, and the estimates are interpreted as quantile-specific elasticities of demand. Figure 3 presents the estimated price elasticity for various quantiles.

We see that price has a U-shaped effect. It is relatively unimportant for people belonging to low or high quantiles, and more important for people in the middle of the distribution (the estimation quantiles start at $\tau = 0.40$ instead of, say, $\tau = 0.10$, because 35% of the people reported zero consumption). These results make a lot of sense: People at the $\tau = 0.40$ to $\tau = 0.50$ quantiles are people that drink very little, and when they do they don't care too much about the price, so their demand is rather inelastic. People from $\tau = 0.60$ to $\tau = 0.80$ are the "social drinkers", people who drink more frequently.

These people have a very elastic demand. Finally, the people that are very high on the consumption of alcohol distribution have zero price elasticity. These people are addicted to drinking and are not responsive to price changes. ■

1. The Estimator and its Asymptotic Distribution

To formalize the discussion above, consider the regression model

$$y = x'\beta + u.$$

Coupled with a conditional quantile restriction $Q_\tau(u|x) = 0$ for a fixed quantile $\tau \in (0,1)$, we obtain the *linear quantile regression model*

$$Q_\tau(y|x) = x'\beta(\tau).$$

Given a random sample $\{y_i, x_i, i = 1, ..., n\}$ we can estimate $\beta(\tau)$ by minimizing a *weighted absolute deviations* objective function, i.e. by

$$\hat\beta(\tau) = \underset{\beta}{\operatorname{argmin}}\; n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

where

$$\rho_\tau(u) = u\,(\tau - I\{u < 0\})$$

is the *check function.* The check function generalizes the absolute value function. For $\tau = 0.5$,

$$\rho_{0.5}(u) = u\,(0.5 - I\{u < 0\}\} = \tfrac{1}{2}\,u\,(1 - 2I\{u < 0\}) = \tfrac{1}{2}\,u\ \operatorname{sgn}(u) = \tfrac{1}{2}\,|u|,$$

i.e., half the absolute value function, so $\beta(0.5)$ is conditional median (the $\frac{1}{2}$ factor simply rescales the objective function without affecting the minimizer). The rest of the quantiles are computed by over-weighting positive or negative deviations.

In order to see the purpose of the $\rho_\tau(.)$ function note that it takes the residuals $u_i = y_i - x_i'\beta$ as arguments. The sum in the minimization problem can therefore be rewritten as

$$n^{-1}\sum_{i=1}^{n} \rho_\tau(u_i) = n^{-1}\sum_{i=1}^{n} \tau\,|u_i|\,I\{u_i \geq 0\} + (1-\tau)\,|u_i|\,I\{u_i < 0\}.$$

We see that positive residuals associated with observation y_i above the suggested quantile regression hyperplane $x_i'\beta$ are given weight τ, while negative residuals, associated with observations y_i below the quantile regression hyperplane $x_i'\beta$, are weighted with $(1-\tau)$. For $\tau = 0.5$ positive and negative residuals are weighted equally, and an equal number of observations are above and below the optimal hyperplane. Then $x_i'\hat\beta(0.5)$ is the median regression hyperplane. When $\tau = 0.9$ each positive residual is weighted 9 times that of a negative residual with weight $1-\tau = 0.1$, and so for every observation above the hyperplane $x_i'\hat{\beta}(0.9)$ approximately 9 will be placed below it. Hence the $x_i'\hat{\beta}(0.9)$ hyperplane represents the 0.9-quantile. For an exact statement of this see Theorem 2.2 and Corollary 2.1 of Koenker (2005).

To do inference we need to know the asymptotic distribution $\hat{\beta}(\tau)$. But before presenting this result it is useful to first derive the asymptotic distribution of a *sample quantile.*

Let $\{y_i, i = 1, ..., n\}$ be a random sample of a variable y with distribution and density F and f, respectively. The τ-th sample quantile of y, is given by

$$\hat{q}_\tau = \operatorname*{argmin}_q \, n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - q).$$

For example, the median estimator is given by

$$\hat{q}_{0.5} = \operatorname*{argmin}_q \frac{1}{n}\sum_{i=1}^{n} |y_i - q| = \begin{cases} y_{(\frac{n-1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(y_{(\frac{n-1}{2})} + y_{(\frac{n}{2})}\right) & \text{if } n \text{ is even,} \end{cases}$$

where $y_{(i)}$ denotes the ith *order statistic* of y. The following theorem gives the asymptotic distribution of $\hat{q}_\tau$.

Theorem 1. *Let $\{y_i, i = 1, ..., n\}$ be an i.i.d. sample of a random variable y with distribution and density functions F and f, respectively, and let q_τ be its τth quantile. Provided that $f(F^{-1}(\tau)) =: f(q_\tau) \neq 0$,*

$$\sqrt{n}\,(\hat{q}_\tau - q_\tau) \xrightarrow{d} N\Big(0, \frac{\tau(1-\tau)}{f(q_\tau)^2}\Big).$$

Proof: We will only prove the median case $q := q_{0.5}$. Assume that n is odd so that $\hat{q} = y_{(\frac{n-1}{2})}$ (otherwise discard the last observation to make n odd – it will make no difference asymptotically). Without loss of generality, assume that $q = 0$, so that

$$\Pr\big(\sqrt{n}\,(\hat{q} - q) \leq a\big) = \Pr\Big(\hat{q} \leq \frac{a}{\sqrt{n}}\Big).$$

Let S_n denote the number of times $y_i - q, i = 1, ..., n$, exceed $a/\sqrt{n}$, and observe that

$$\hat{q} \leq \frac{a}{\sqrt{n}} \qquad \text{if and only if} \qquad S_n \leq \frac{n-1}{2}.$$

Also observe that $S_n \sim \text{Binomial}(p_n, n)$, with $p_n = 1 - F(a/\sqrt{n})$. Therefore, for $z = (S_n - E(S_n))/\sqrt{Var(S_n)}$,

$$\begin{aligned} \Pr\left(S_n \leq \frac{n-1}{2}\right) &= \Pr\left(z < \frac{\frac{1}{2}(n-1) - np_n}{\sqrt{np_n(1-p_n)}}\right) \\ &\to \Phi\left(\lim_{n\to\infty} \frac{\frac{1}{2}(n-1) - np_n}{\sqrt{np_n(1-p_n)}}\right), \end{aligned}$$

as $n \to \infty$, by the Lindeberg central limit theorem. Write

$$\begin{aligned}\frac{\frac{1}{2}(n-1) - np_n}{\sqrt{np_n(1-p_n)}} &= \frac{\sqrt{n}(\frac{1}{2} - p_n) - 1/(2\sqrt{n})}{\sqrt{p_n(1-p_n)}} \\ &= 2\sqrt{n}(\tfrac{1}{2} - p_n) + o_p(1) \\ &= 2a\frac{F(a/\sqrt{n}) - F(0)}{a/\sqrt{n}} + o_p(1) \\ &\to 2af(0),\end{aligned}$$

as $n \to \infty$, where the second line above uses $p_n(1-p_n) \to \frac{1}{4}$, and the third line uses the definition of p_n from above, i.e., $p_n = 1 - F(a/\sqrt{n})$, and $F(0) = \frac{1}{2}$. We have shown that for $q = q_{0.5}$,

$$\Pr(\sqrt{n}(\hat{q} - q) \leq a) \to \Phi(2af(q)),$$

or

$$\sqrt{n}(\hat{q} - q) \xrightarrow{d} N\left(0, \frac{1}{4f(q)^2}\right)$$

for $f(q) > 0$. The general result can be proven similarly, only this time the $\frac{1}{4}$ factor is replaced by $\tau(1-\tau)$ and $f(q_{0.5})$ by $f(q_\tau)$. ■

The factor $f(q_\tau)$ is the density of y evaluated at the τ-th quantile. If $f(q_\tau) = 0$ the τ-th quantile would have a zero density at that point, so no sample no matter how large could identify q_τ. The condition $f(q_\tau) \neq 0$ precludes this pathological case.

The following theorem gives the asymptotic distribution of regression quantiles.

Theorem 2. *Assume that*

(A1) The distribution functions $\{F_i\}$ are absolutely continuous, with continuous densities, $\{f_i(\xi)\}$, uniformly bounded away from 0 and ∞ at the points $\xi_i(\tau), i = 1, 2, \cdots$

(A2) There exist positive definite matrices D_0 and $D_1(\tau)$ such that

(i) $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n x_i x_i' = D_0 \equiv E\left[xx'\right]$;

(ii) $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n f_i(\xi_i(\tau)) x_i x_i' = D_1(\tau) \equiv E\left[f_i\left(F_i^{-1}(\tau)|x\right) xx'\right]$;

(iii) $\max_{i=1,\cdots,n} ||x_i||/\sqrt{n} \to 0$ *(this implies the Lindeberg condition).*

Then

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{d} N\left(0, \tau(1-\tau) D_1(\tau)^{-1} D_0 D_1(\tau)^{-1}\right).$$

Furthermore, if errors are i.i.d., $F_i = F, f_i = f$ for all i, and the above result simplifies to

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{d} N\left(0, \omega(\tau)^2 D_0\right),$$

where $\omega(\tau)^2$ is the variance of the τ-th unconditional quantile given by

$$\omega(\tau)^2 = \frac{\tau(1-\tau)}{f(F^{-1}(\tau))^2}.$$

Proof: Difficult – Take Econ 721 if you have to know. ■

The easiest way to understand this asymptotic result is to consider the homoskedastic case and compare it to the asymptotic distribution of the OLS coefficients. Recall that under homoskedasticity

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \sigma^2 E(xx')^{-1}\right)$$

so that the asymptotic variance of the OLS coefficient vector $\hat{\beta}$ is given by

$$\text{AVar}(\hat{\beta}) = \frac{\sigma^2}{n} E(xx')^{-1}.$$

From theorem 2, we see that under homoskedasticity

$$\sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau)) \xrightarrow{d} N\left(0, \omega(\tau)^2 E(xx')^{-1}\right)$$

so the asymptotic variance of $\hat{\beta}(\tau)$ is given by

$$\text{AVar}(\hat{\beta}(\tau)) = \frac{\omega(\tau)^2}{n} E(xx')^{-1}.$$

The difference between the two asymptotic variances is the constant by which the "regression factor" $E(xx')^{-1}$ is scaled. Since OLS is a conditional mean, the factor that appears in the asymptotic variance of $\hat{\beta}$ is the *variance of the sample mean* σ^2/n, whereas since QR is a conditional quantile, the factor that appears in the asymptotic variance of $\hat{\beta}(\tau)$ is the *variance of the sample quantile* $\omega(\tau)^2/n$.

2. Formulating Quantile Regression as a Linear Programming Problem

`Thie, Paul R. and Keough, Gerard E. (2008) -- An Introduction to Linear Programming and Game Theory, Wiley-Interscience.`

Linear programs (LPs) are predominantly analyzed and solved using the *standard form*

$$\min_z \; F(z) = c'z \quad \text{subject to} \quad Az \geq b, z \geq 0, \tag{1}$$

where c is an n-vector of constants (called the *objective constants*), z is a n-vector of variables (also called *instruments*), A is an $p \times n$ matrix of constants (called the *technological constants* of the problem), and b is is an p-vector of constants (called the *constraint constants*). Note that in the standard formulation of LP's, z is restricted to be positive, i.e. $z \in \mathbb{R}^n_+$, so in total we have $p + n$ constraints.

`Intriligator, Michael D.(1987) -- Mathematical Optimization and Economic Theory, Society for Industrial Mathematics, chapter 5.`

Each of the n nonnegativity constraints $z_i \geq 0, i = 1, ..., n$ defines a closed half space, and the intersection of all such half spaces is the nonnegative orthant of Euclidean n-space,

$\mathbb{R}^n_+$. Each of the p inequality constraints

$$\sum_{j=1}^{n} a_{ij} z_j \geq b_i, \quad i = 1, ..., p$$

also defines a closed half-space in $\mathbb{R}^n$, namely the set of points lying on, or on the appropriate side of, the hyperplane defined by

$$\left\{ z \in \mathbb{R}^n \left| \sum_{j=1}^{n} a_{ij} z_j = b_i, i = 1, ..., p \right. \right\}.$$

In general, the intersection of closed half-spaces in $\mathbb{R}^n$ is a *convex polyhedral set*, or, if bounded, a *convex polyhedron.*

The hyperplane boundaries are called *bounding faces*, and the points at which n or more bounding faces meet are called *vertices.* Each bounding face consists of all points at which one of the inequality or nonnegativity constraints is satisfied as an equality, and each vertex is a point at which n or more of the inequality constraints are satisfied as equalities.

The contours of the objective function are:

$$\left\{ z \in \mathbb{R}^n \left| c'z = \text{constant} \right. \right\},$$

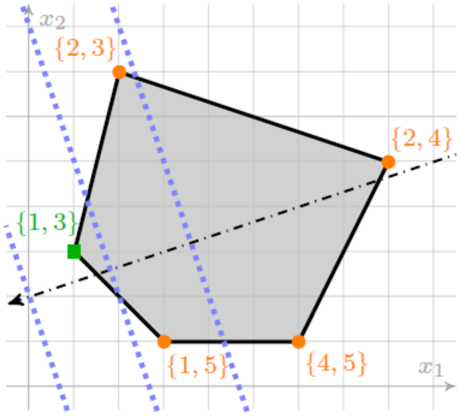
which is the equation of a hyperplane in $\mathbb{R}^n$. As the constant is varied the contour map is obtained as a series of parallel hyperplanes. The *preference direction* is the direction of steepest increase of the objective function and is given by the gradient vector:

$$\frac{\partial F}{\partial z} = c',$$

a row vector in $\mathbb{R}^n$ which is orthogonal to all contours through which it passes.

Geometrically, then, the linear programming problem is that of finding a point (or set of points) in $\mathbb{R}^n$ on that contour of the objective function lying furthest along the preference direction but within the convex polyhedral opportunity set. From the geometry it is apparent that if a solution exists, it cannot be an interior point but must rather lie on the boundary of the opportunity set—on one or more of the bounding faces or, equivalently, at one vertex, two vertices,. . . , n vertices and all points in between these vertices; i.e., all convex combinations of these vertices. The solution is obtained at the point(s) at which a contour hyperplane is a supporting hyperplane of the convex polyhedral opportunity set.

The (single) vertex solution, which is unique, and the two vertex (bounding face) solution, which is not unique, are illustrated in Fig. 2.4. In the latter case the common slope of the contours equals the slope of the highest possible bounding face hyperplane, a line in $\mathbb{R}^2$, so the solution occurs at two vertices and at all points on the line connecting these two vertices. In three space ($n = 3$), if a solution exists, it can be at a vertex point (the intersection of three or more bounding faces), along a line (the intersection of two bounding faces), or on a plane (a bounding face). While the solution need not be unique,



$$\begin{array}{rrl}
\text{minimize} & 3x_1 + x_2 & \\
\text{subject to} & x_1 + x_2 \geq 4 & \textcircled{1} \\
& -x_1 - 3x_2 \geq -23 & \textcircled{2} \\
& 4x_1 - x_2 \geq 1 & \textcircled{3} \\
& -2x_1 + x_2 \geq -11 & \textcircled{4} \\
& x_2 \geq 1 & \textcircled{5}
\end{array}$$

FIGURE 4

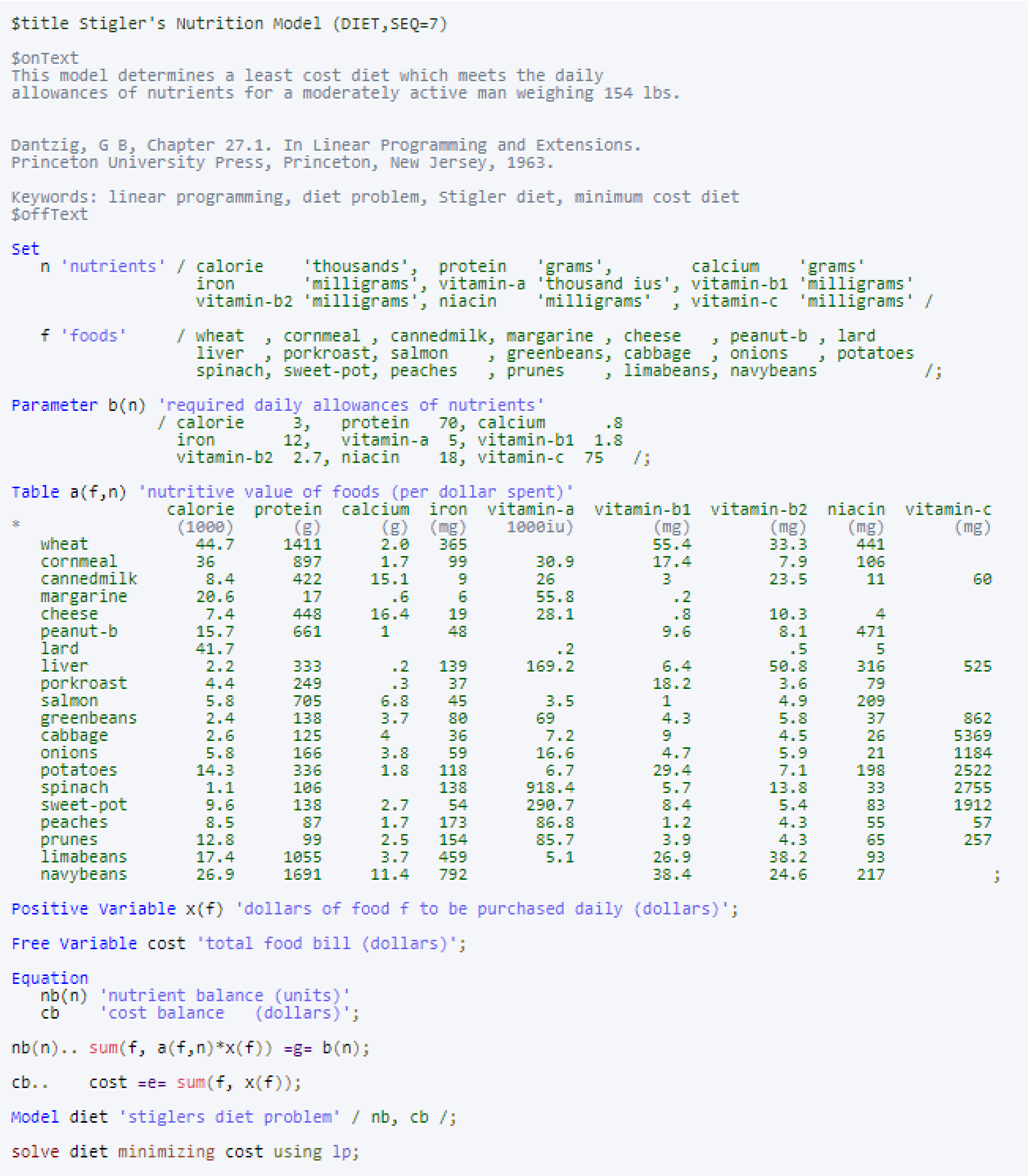
if a solution exists, the value of the objective function is unique. Also, from the convexity of the opportunity set and linearity of the objective function, by the local-global theorem of Sec. 2.3, a solution which is a local maximum is also a global maximum. Thus, if, in the opportunity set a vertex yields a higher value (or, more generally, no lower value) than all neighboring vertices, then it is a solution to the problem. This important property is the basis for the simplex algorithm, to be discussed below. Furthermore, if $n > p$ then solutions must occur at a vertex of the opportunity set at which $n|p$ or more of the instrument variables are equal to zero; i.e., there is at least one solution which has at most as many nonzero variables as there are inequality constraints.

Since the objective function is continuous and the opportunity set is closed, by the Weierstrass theorem a solution exists if the opportunity set is nonempty and bounded. Thus there are two circumstances in which there might not exist a solution to the linear programming problem. The first is that in which the constraints are inconsistent so the opportunity set is empty. If the opportunity set is nonempty and bounded then a solution exists and it must be a boundary solution. More generally, a solution exists if the opportunity set is nonempty and the objective function is bounded.

In general, then, there are three possible solutions for the linear programming problem: (i) a unique solution (at a vertex), (ii) infinitely many solutions (between two or more vertices), or (iii) no solution (if the opportunity set is empty or unbounded).

Example: The classical example of a standard LP problem is George Stigler's diet problem.

> **Stigler's diet problem:** For a moderately active man weighing 154 pounds, how much of each of 77 foods should be consumed on a daily basis so that the man's intake of 9 nutrients will be at least equal to the recommended dietary allowances (RDAs) suggested by the National Research Council in 1943, with the cost of the diet being minimal?



```
$title Stigler's Nutrition Model (DIET,SEQ=7)

$onText
This model determines a least cost diet which meets the daily
allowances of nutrients for a moderately active man weighing 154 lbs.


Dantzig, G B, Chapter 27.1. In Linear Programming and Extensions.
Princeton University Press, Princeton, New Jersey, 1963.

Keywords: linear programming, diet problem, Stigler diet, minimum cost diet
$offText

Set
   n 'nutrients' / calorie    'thousands',  protein   'grams',        calcium    'grams'
                   iron       'milligrams', vitamin-a 'thousand ius', vitamin-b1 'milligrams'
                   vitamin-b2 'milligrams', niacin    'milligrams'  , vitamin-c  'milligrams' /

   f 'foods'     / wheat  , cornmeal , cannedmilk, margarine , cheese   , peanut-b , lard
                   liver  , porkroast, salmon    , greenbeans, cabbage  , onions   , potatoes
                   spinach, sweet-pot, peaches   , prunes    , limabeans, navybeans           /;

Parameter b(n) 'required daily allowances of nutrients'
               / calorie     3,   protein   70, calcium      .8
                 iron       12,   vitamin-a  5, vitamin-b1  1.8
                 vitamin-b2  2.7, niacin    18, vitamin-c  75   /;

Table a(f,n) 'nutritive value of foods (per dollar spent)'
                calorie  protein  calcium  iron  vitamin-a  vitamin-b1  vitamin-b2  niacin  vitamin-c
*                (1000)      (g)      (g)  (mg)    1000iu)        (mg)        (mg)    (mg)       (mg)
   wheat           44.7     1411      2.0   365                   55.4        33.3     441
   cornmeal        36        897      1.7    99       30.9        17.4         7.9     106
   cannedmilk       8.4      422     15.1     9       26           3          23.5      11         60
   margarine       20.6       17       .6     6       55.8          .2
   cheese           7.4      448     16.4    19       28.1          .8        10.3       4
   peanut-b        15.7      661      1      48                    9.6         8.1     471
   lard            41.7                                 .2                      .5       5
   liver            2.2      333       .2   139      169.2         6.4        50.8     316        525
   porkroast        4.4      249       .3    37                   18.2         3.6      79
   salmon           5.8      705      6.8    45        3.5         1           4.9     209
   greenbeans       2.4      138      3.7    80       69           4.3         5.8      37        862
   cabbage          2.6      125      4      36        7.2         9           4.5      26       5369
   onions           5.8      166      3.8    59       16.6         4.7         5.9      21       1184
   potatoes        14.3      336      1.8   118        6.7        29.4         7.1     198       2522
   spinach          1.1      106            138      918.4         5.7        13.8      33       2755
   sweet-pot        9.6      138      2.7    54      290.7         8.4         5.4      83       1912
   peaches          8.5       87      1.7   173       86.8         1.2         4.3      55         57
   prunes          12.8       99      2.5   154       85.7         3.9         4.3      65        257
   limabeans       17.4     1055      3.7   459        5.1        26.9        38.2      93
   navybeans       26.9     1691     11.4   792                   38.4        24.6     217           ;

Positive Variable x(f) 'dollars of food f to be purchased daily (dollars)';

Free Variable cost 'total food bill (dollars)';

Equation
   nb(n) 'nutrient balance (units)'
   cb    'cost balance   (dollars)';

nb(n).. sum(f, a(f,n)*x(f)) =g= b(n);

cb..    cost =e= sum(f, x(f));

Model diet 'stiglers diet problem' / nb, cb /;

solve diet minimizing cost using lp;
```

Figure 5. GAMS implementation of Stigler's Diet Problem. GAMS Model Libraries, diet.gms : Stigler's Nutrition Model.

Figure 4 presents the GAMS implementation of the problem[1]. Of the total 77 food groups considered by Stigler, the GAMS program uses only 20. This is not a restriction

[1] Available at https://www.gams.com/latest/gamslib_ml/libhtml/gamslib_diet.html.

since the omitted foods are not consumed at all in the optimal schedule, so the solution to this restricted problem is the same as the solution to the complete model.

The nutrient RDAs required to be met in Stigler's experiment were calories, protein, calcium, iron, as well as vitamins A, B1, B2, B3, and C. The result was an annual budget allocated to foods such as evaporated milk, cabbage, dried navy beans, and beef liver at a cost of approximately $0.11 a day in 1939 U.S. dollars.

```
---- EQU nb  nutrient balance (units)

               LOWER     LEVEL    UPPER   MARGINAL

calorie        3.000     3.000     +INF      0.009
protein       70.000   147.414     +INF       .
calcium        0.800     0.800     +INF      0.032
iron          12.000    60.467     +INF       .
vitamin-a      5.000     5.000     +INF  4.0023E-4
vitamin-b1     1.800     4.120     +INF       .
vitamin-b2     2.700     2.700     +INF      0.016
niacin        18.000    27.316     +INF       .
vitamin-c     75.000    75.000     +INF  1.4412E-4

                       LOWER     LEVEL     UPPER    MARGINAL

---- EQU cb              .         .         .         1.000

cb  cost balance   (dollars)

---- VAR x  dollars of food f to be purchased daily (dollars)

              LOWER      LEVEL    UPPER   MARGINAL

wheat           .        0.030     +INF       .
cornmeal        .         .        +INF      0.489
cannedmilk      .         .        +INF      0.044
margarine       .         .        +INF      0.778
cheese          .         .        +INF      0.235
peanut-b        .         .        +INF      0.698
lard            .         .        +INF      0.626
liver           .        0.002     +INF       .
porkroast       .         .        +INF      0.893
salmon          .         .        +INF      0.652
greenbeans      .         .        +INF      0.615
```

```
cabbage          .       0.011     +INF        .
onions           .         .       +INF      0.555
potatoes         .         .       +INF      0.335
spinach          .       0.005     +INF        .
sweet-pot        .         .       +INF      0.350
peaches          .         .       +INF      0.758
prunes           .         .       +INF      0.667
limabeans        .         .       +INF      0.103
navybeans        .       0.061     +INF        .

                        LOWER     LEVEL     UPPER    MARGINAL

---- VAR cost           -INF      0.109     +INF        .

cost  total food bill (dollars)
```

wheat, liver, cabbage, spinach, and navy beans

In 2014, the Google chef Anthony Marco devised a recipe using a similar list of ingredients, that he called “Foie Linéaire à la Stigler” (Linear Liver à la Stigler). One Google employee described it as “delicious”.[2] ■

To arrive at a linear program on standard form the first problem is that in such a program (1) all variables z over which minimization is performed should be positive. To achieve this, residuals are decomposed into positive and negative parts using *slack variables*:

$$u_i = u_i^+ - u_i^-, \qquad i = 1, .., n$$

where $u_i^+ = \max\{0, u_i\} = |u_i| \cdot I\{u_i \geq 0\}$ is the positive part, and $u_i^- = \max\{0, -u_i\} = |u_i| \cdot I\{u_i < 0\}$ is the negative part of u. In the lingo of programming, the $u_i^+, i = 1, ..., n$, are called positive slack variables, and the $u_i^-, i = 1, .., n$, are called negative slack variables. The objective function can be written as

$$\begin{aligned} \sum_{i=1}^{n} \rho_\tau(u_i) &= \sum_{i=1}^{n} \tau\, u_i^+ + (1-\tau)\, u_i^- \\ &= \tau\, 1_n' u^+ + (1-\tau)\, 1_n' u^-, \end{aligned}$$

where $u^+ = (u_1^+, ..., u_n^+)'$, $u^- = (u_1^-, ..., u_n^-)'$, and 1_n is a n-vector of ones (the sumer vector). The residuals $u_i, i = 1, ..., n$, must satisfy the n constraints

$$y_i - x_i'\beta = u_i = u_i^+ - u_i^-, \qquad i = 1, ..., n.$$

[2] Orwant, Jon (2014),”Sudoku, Linear Optimization, and the Ten Cent Diet”, available at `https://ai.googleblog.com/2014/09/sudoku-linear-optimization-and-ten-cent.html`

This results in the following linear program

$$\begin{aligned} & \min_{\beta,u^+,u^-} \tau\, 1_n' u^+ + (1-\tau)\, 1_n' u^- && && (2) \\ \text{subject to} \quad & y_i - x_i'\beta - u_i^+ + u_i^- = 0, && i = 1, ..., n && (3) \\ & u_i^+ \geq 0, u_i^- \geq 0, && i = 1, ..., n && (4) \\ & \beta \in R^{k+1}. && && (5) \end{aligned}$$

The problem in (2)-(5) is almost of the form (1), the difference being that in proper LP programs all decision variables are positive, whereas here β is unrestricted, or, as we call it in LP jargon, it is *free*. To bring the formulation to the standard form we can again decompose β into positive and negative parts as follows

$$\beta = \beta^+ - \beta^-,$$

where $\beta^+ = \max\{0, \beta\}$ and $\beta^- = \max\{0, -\beta\}$, component-wise. This results in the following standard linear program

$$\begin{aligned} & \min_{\beta^+,\beta^-,u^+,u^-} \tau\, 1_n' u^+ + (1-\tau)\, 1_n' u^- && && (6) \\ \text{subject to} \quad & y_i - x_i'(\beta^+ - \beta^-) - u_i^+ + u_i^- = 0, && i = 1, ..., n && (7) \\ & u_i^+ \geq 0, u_i^- \geq 0, && i = 1, ..., n && (8) \\ & \beta_j^+ \geq 0, \beta_j^- \geq 0, && j = 1, ..., k. && (9) \end{aligned}$$

To see that the problem in (6)-(9) is a standard LP problem, we will write it in the form (1). To do that we need to specify the quantities c, z, A, and b. Let

$$b := y$$

and let

$$X(\beta^+ - \beta^-) + I_n u^+ - I_n u^- = [X, -X, I_n, -I_n] \begin{bmatrix} \beta^+ \\ \beta^- \\ u^+ \\ u^- \end{bmatrix} := A\, z,$$

where $X := [x_1', x_2', ..., x_n']'$ is the design matrix, I_n is the $n \times n$ identity matrix, and A and z are defined in the obvious way. The n constraints in (7) can now be written in matrix form as

$$Az = b.$$

Because β^+ and β^- enter the minimization problem only through the constraint in (9), an all zeros vector of dimension $(2k \times 1)$, denoted by 0_{2k}, must be introduced as part of the coefficient vector c, which can then be appropriately defined as

$$c := \begin{bmatrix} 0_{2k} \\ \tau \, 1_n \\ (1-\tau) \, 1_n \end{bmatrix}.$$

Then

$$c'z = 0_{2k}' \begin{bmatrix} \beta^+ \\ \beta^- \end{bmatrix} + \tau \, 1_n' u^+ + (1-\tau) \, 1_n' u^- = \sum_{i=1}^{n} \rho_\tau(u_i),$$

as desired.

Example:

```
base=read.table("http://freakonometrics.free.fr/rent98_00.txt",header=TRUE)
attach(base)
library(quantreg)
library(lpSolve)
tau <- 0.3

# Problem (1) only one covariate
X <- cbind(1,base$area)
K <- ncol(X)
N <- nrow(X)

A <- cbind(X,-X,diag(N),-diag(N))
c <- c(rep(0,2*ncol(X)),tau*rep(1,N),(1-tau)*rep(1,N))
b <- base$rent_euro
const_type <- rep("=",N)

linprog <- lp("min",c,A,const_type,b)
beta <- linprog$sol[1:K] - linprog$sol[(K+1):(2*K)]
beta
rq(rent_euro˜area, tau=tau, data=base)


# Problem (2) with 2 covariates
X <- cbind(1,base$area,base$yearc)
K <- ncol(X)
N <- nrow(X)

A <- cbind(X,-X,diag(N),-diag(N))
c <- c(rep(0,2*ncol(X)),tau*rep(1,N),(1-tau)*rep(1,N))
b <- base$rent_euro
const_type <- rep("=",N)

linprog <- lp("min",c,A,const_type,b)
```

```
beta <- linprog$sol[1:K] - linprog$sol[(K+1):(2*K)]
beta
rq(rent_euro˜ area + yearc, tau=tau, data=base)
```

Note how much slower the `lp` command of the `lpsolve` library is relative to the `rq` command of the `quantreg` library.

■

3. Modeling Melbourn's Maximum Daily Temperature

We consider the model

$$L(t) = \beta_0 + A \sin\left(\frac{2\pi}{365}(t - t^*)\right). \tag{10}$$

Since at $t = t^*$ the value of the expression inside the parenthesis above is 0 and $\sin 0 = 0$, we have $L(t^*) = \alpha_0$, so β_0 is the temperature at time t^*. Also, since for any two angle θ_1 and θ_2

$$\sin(\theta_1 - \theta_2) = \sin(\theta_1)\cos(\theta_2) - \cos(\theta_1)\cos(\theta_2) \tag{11}$$

we have that

$$\begin{aligned}
A \sin\left(\frac{2\pi}{365}(t - t^*)\right) &= A \sin\left(\frac{2\pi t}{365} - \frac{2\pi t^*}{365}\right) && (12)\\
&= A \sin\left(\frac{2\pi t}{365}\right) \cos\left(\frac{2\pi t^*}{365}\right) - A \cos\left(\frac{2\pi t}{365}\right) \cos\left(\frac{2\pi t^*}{365}\right) && (13)\\
&= \left[A \sin\left(\frac{2\pi t}{365}\right) - A \cos\left(\frac{2\pi t}{365}\right)\right] \cos\left(\frac{2\pi t^*}{365}\right) && (14)\\
&= \left[A \sin\left(\frac{2\pi t}{365}\right) - A \cos\left(\frac{2\pi t}{365}\right)\right] \left[1 - \left(\frac{2\pi t}{365}\right)^2 /2\right] && (15)\\
&= \beta_1 \sin\left(\frac{2\pi t}{365}\right) + \beta_2 \cos\left(\frac{2\pi t}{365}\right) && (16)
\end{aligned}$$

where we have defined

$$\beta_1 = A \cos\left(\frac{2\pi t^*}{365}\right) \quad \text{and} \quad \beta_2 = -A \cos\left(\frac{2\pi t^*}{365}\right) = -\beta_1. \tag{17}$$

If the starting time t^* is unknown, we can let β_1 and β vary freely and form their estimates retreive estimates of A and t^*.

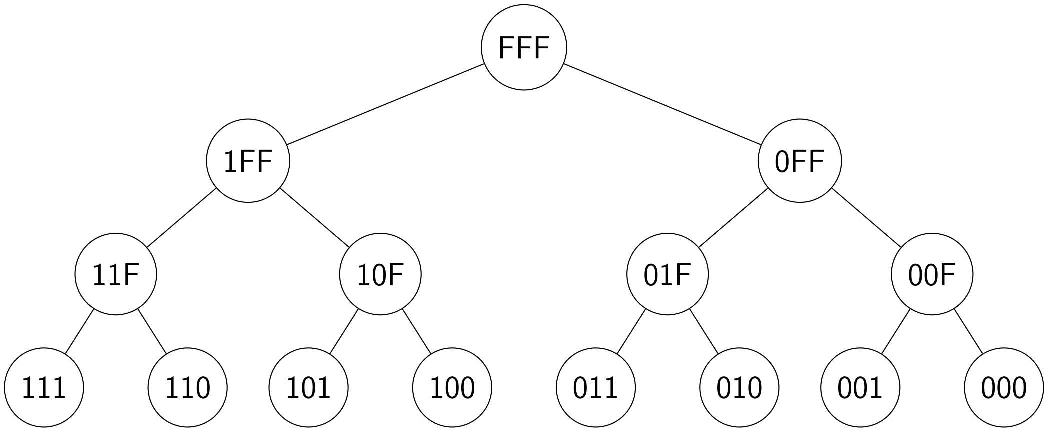
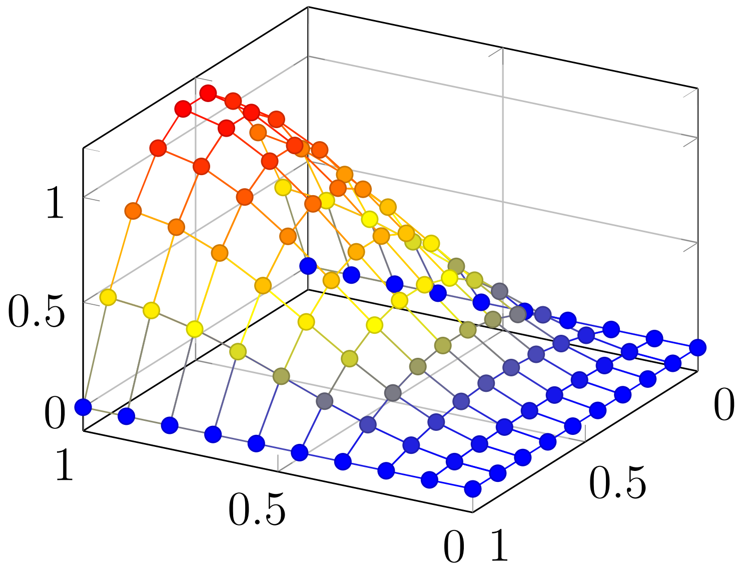
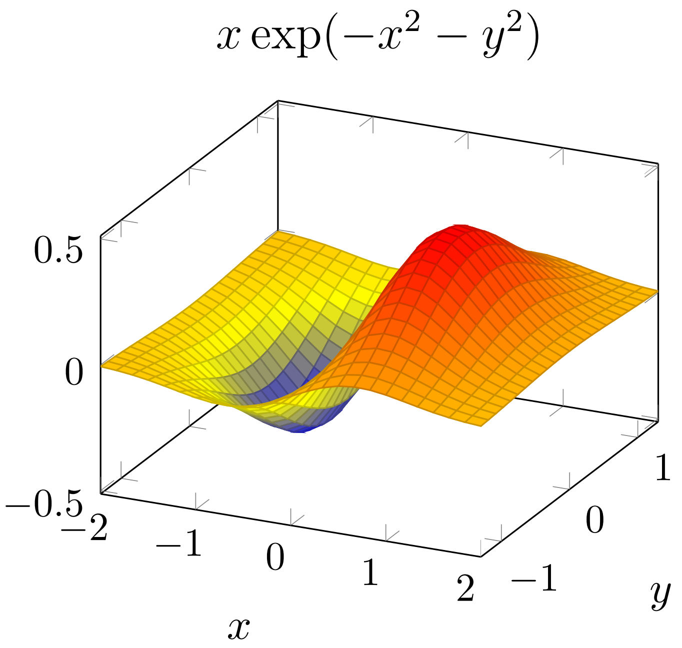
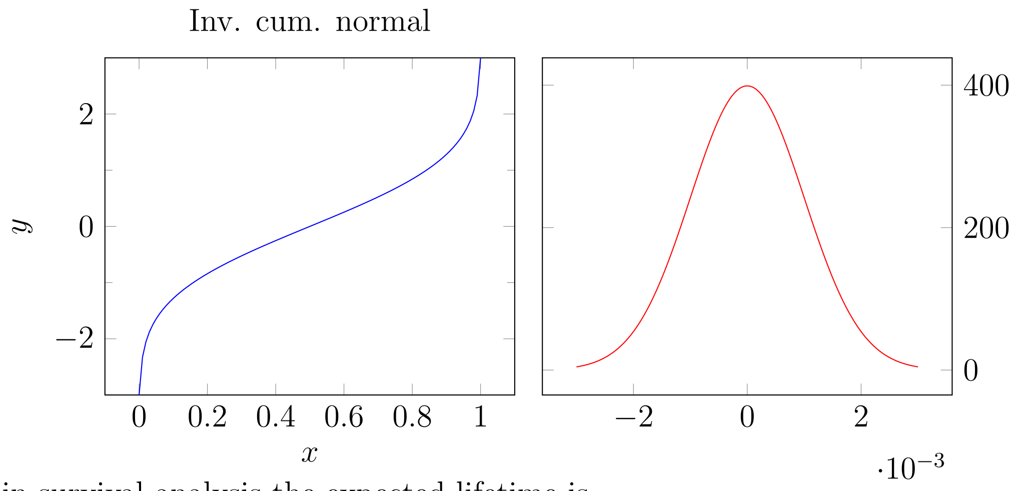


FIGURE 6. State-space tree for a sample of $n = 3$ observations. Each node is named according to the values of the 3 γ's, (γ_1, γ_2, and γ_3) in that state. F stands for "free", meaning that the corresponding γ takes any value in the $[0, 1]$ interval. Otherwise, the γ's are restricted to either 0 or 1. At the root node FFF all the γ's are free. The bottom line lists all $3! = 8$ feasible combinations of values for the γ's. For example, at the 10F node, $\gamma_1 = 1$, $\gamma_2 = 0$, and $\gamma_3 \in [0, 1]$.



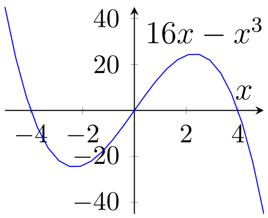


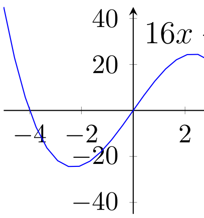


in survival analysis the expected lifetime is

$$\mu = \int_0^\infty S(t)dt$$

where the survival function is $S(t) = Pr(T > t) = 1 - F(t)$ measured from birth at $t = 0$. (It can easily be extended to cover negative values of t.)





References

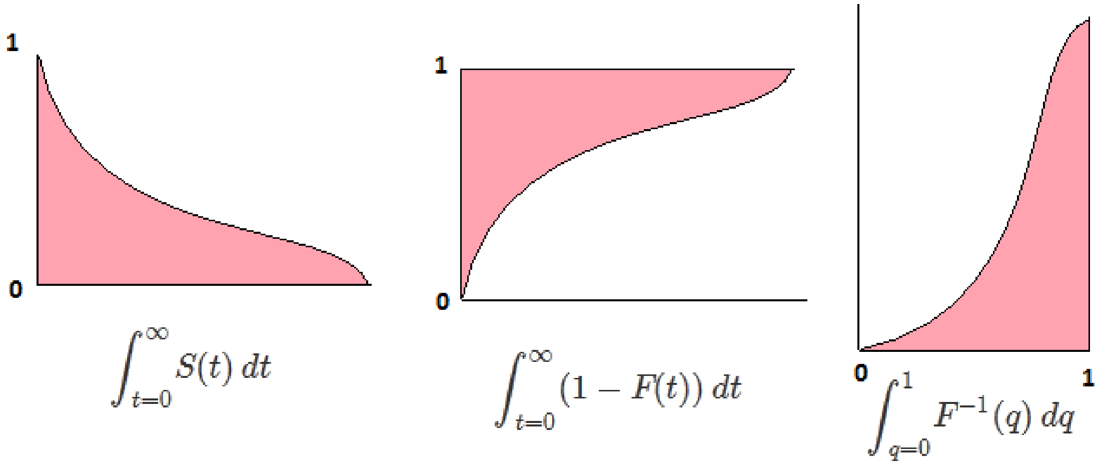


Figure 7. Various reflections of the same area expressed in alternative ways.

Furno Marilena, Cristina Davino, Domenico Vistocco (2013), *Quantile Regression: Theory and Applications*, Wiley.

Manning, W.G., L. Blumberg, and L.H. Moulton (1995), "The Demand for Alcohol: The Differential Response to Price", *Journal of Health Economics*, 14, 123–148.

Koenker, R. (2005), *Quantile Regression*, Econometric Society Monograph, Cambridge University Press, Cambridge.

Koenker R., and G. Bassett (1978), "Regression Quantiles", *Econometrica*, 46, 33-50.

Koenker, R. and K.F. Hallock (2001), "Quantile Regression", *Journal of Economic Perspectives*, vol. 15(4), 143-156.

Stigler George J. (1945) – The Cost of Subsistence, Journal of Farm Economics, 37, 1249-1258

Thie, Paul R. and Keough, Gerard E. (2008) – *An Introduction to Linear Programming and Game Theory*, Wiley-Interscience.

> I do not understand why statisticians commonly limit their inquiries to Averages and do not revel in more comprehensive views. Their souls seem so dull to the charm of variety as that of a native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once.
>
> — *Francis Galton*