Week 14 Quantile Regression[1]

STAT 111

Zhengfei Li (Alex)

Swarthmore College Instructed by Prof. Phil Everson

2025-04-30

[1] Acknowledge Prof. Daifeng He's ECON135 Syllabus for suggested readings

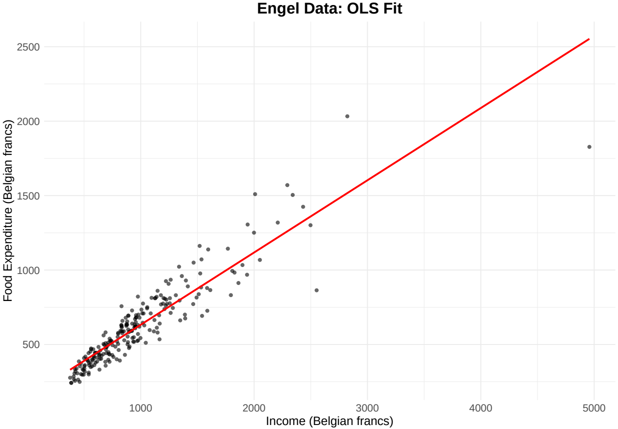
The Engel Dataset

- **Description:** 235 sampled households on annual household income (`income`) and food expenditure (`foodexp`) for 1857 Belgian working-class households in Belgian francs.
- **Source:** Available in R's `quantreg` package `data(engel)`
- **Goal:** We wish to capture differential spending behavior at the low versus high end of the income distribution

income	foodexp
420.1577	255.8394
541.4117	310.9587
901.1575	485.6800
639.0802	402.9974
750.8756	495.5608

OLS Fit & Its Limitations

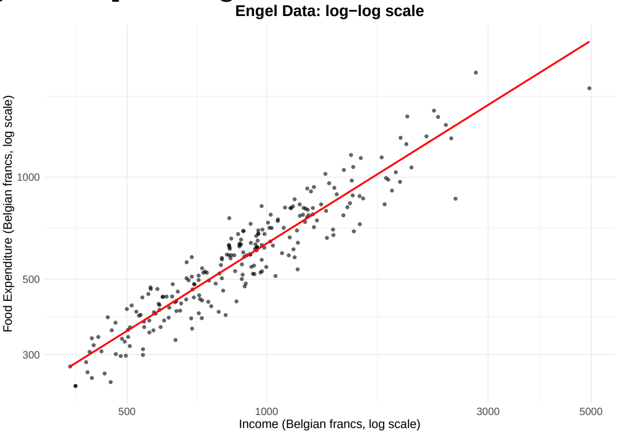
- `FoodExp ~ Income`



- **OLS:** Conditional mean $E[\text{foodexp} \mid \text{income}]$ via minimizing squared error.
- **Limitation:** Non-constant variance.

OLS on log-log scale

- `log(FoodExp) ~ log(Income)`



- stabilizes variance and linearizes correlation
- Restricted functional-form
- Assume multiplicative error

Quantile Regression

- **Target:** The τ-th conditional quantile $Q_\tau(Y \mid X)$, defined by $P(Y \leq Q_\tau \mid X) = \tau$.

$$Q_\tau(Y \mid X) \; = \; X^\top \beta_\tau$$

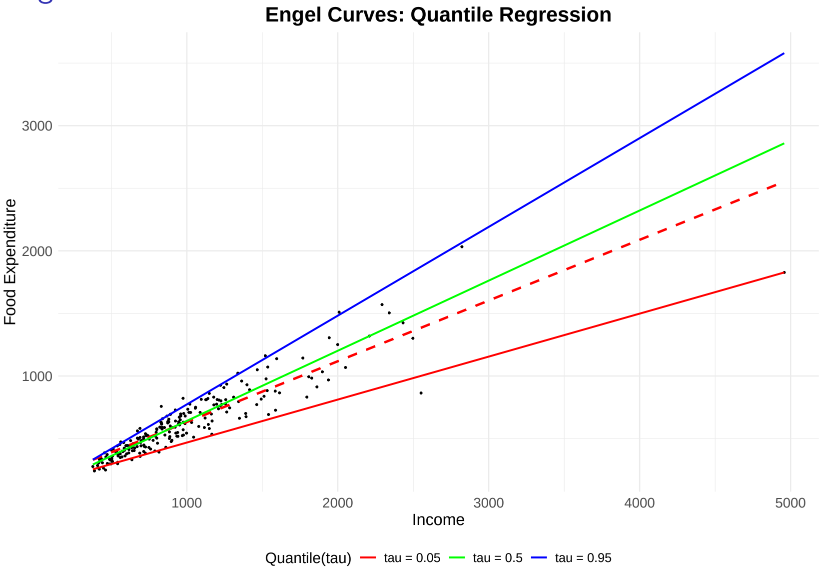
Quantile Regression with Engel Dataset

```
ols_mod <- lm(foodexp ~ income, data = engel)
taus   <- c(0.05, 0.5, 0.95)
qr_mod <- rq(foodexp ~ income, tau = taus, data = engel)
ols_sum    <- summary(ols_mod)
qr_sum  <- summary(qr_mod, se = "boot", B = 999)
```

Table 2: Coefficient estimates with two-decimal precision

Term	Model	Estimate	SE	p_value
income	OLS	0.49	0.01	0
income	tau=0.05	0.34	0.04	0
income	tau=0.5	0.56	0.03	0
income	tau=0.95	0.71	0.03	0

Engel Curves



Quantile Regression Estimators

- **Assumptions:** Linearity, additivity (in our setup here), independence of observation, large sample size
- **Estimator:**

$$\hat{\beta}_\tau = \arg\min_b \sum_{i=1}^{n} \rho_\tau(y_i - X_i' b),$$

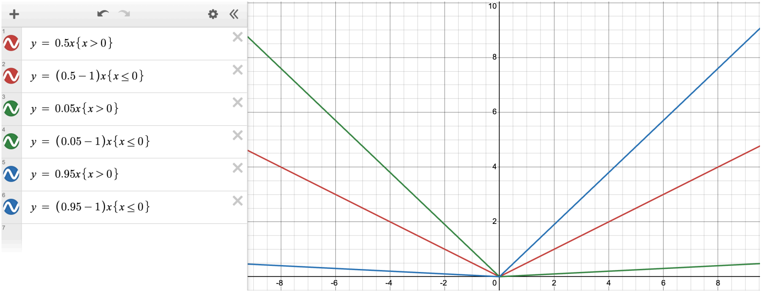
  where $\rho_\tau(u) = u\,(\tau - I\{u < 0\})$ is the check loss.
- **Interpretation:** Coefficients shift the τ-th conditional quantile, revealing heterogeneous covariate effects across the distribution.
- Coefficient usually calculated by Linear Programming

Check Loss Function

- **Estimator:**

$$\hat{\beta}_\tau = \arg\min_b \sum_{i=1}^{n} \rho_\tau(y_i - X_i'b),$$

where $\rho_\tau(u) = u\,(\tau - I\{u < 0\})$ is the check loss.



Why the Check Loss Recovers Quantiles

- **Estimator:**

$$\hat{\beta}_\tau = \arg\min_b \sum_{i=1}^{n} \rho_\tau(y_i - X_i' b),$$

where $\rho_\tau(u) = u\,(\tau - I\{u < 0\})$ is the check loss.

1. **Population objective:**

$$L(c) = E\big[\rho_\tau(Y - c)\big].$$

2. **First-order condition:**

$$\frac{d}{dc} L(c) = P(Y \le c) - \tau.$$

3. **Solution:** Setting derivative to zero gives $P(Y \le c^*) = \tau$, so we have unique minimizer $c^* = Q_\tau(Y)$.

Asymptotic Normality & SEs

- Under regularity (i.i.d, moment existence, continuity, etc),

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N\Big(0, \frac{\tau(1-\tau)}{f_{Y|X}(Q_\tau)^2}\,\Omega^{-1}\Big),$$

  where $\Omega = E[XX']$ and $f_{Y|X}$ is the conditional density at Q_τ
  .
- **Standard errors:** Estimate the sparsity $s(\tau) = 1/f_{Y|X}(Q_\tau)$ via:
  - kernel: $\hat{f}_{Y|X}(Q_\tau \mid x) = \frac{1}{n\,h}\sum_{i=1}^{n} K\Big(\frac{y_i - X_i{}'\hat{\beta}_\tau}{h}\Big)$
  - or bootstrap: take empirical SD from coefficient estimates of repeated sample draw with replacement
  - `summary.rq()` in **quantreg** provides multiple options.

Summary: Comparing CEF vs CQF

Feature	CEF (OLS)	CQF (Quantile)
Target	$E[Y \mid X]$	$Q_\tau(Y \mid X)$
Loss	$(Y - m)^2$	$\rho_\tau(Y - q)$
Robustness	Sensitive to outliers, variance	Robust to outliers, tail-specific effects
Assumptions	Homoskedasticity for efficiency	No need for constant variance, Error distribution shape unspecified
Inference	Closed-form SEs via OLS theory	Requires sparsity estimation or bootstrap

A taste of Bayesian Quantile Regression

- **Prior**: Common prior for β could be uniform or gaussian. Inverse-gamma prior is typically used on σ.
- **Asymmetric Laplace pseudo-likelihood**
  - $f(u \mid \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp[-\rho_\tau(u)/\sigma]$,where $\rho_\tau(u) = u(\tau - I\{u < 0\})$, parameterized by μ, scale σ, and τ
- **Posterior**
  - $p(\beta, \sigma, ... \mid y) \propto p(\beta)\, p(\sigma)\, \prod_{i=1}^{n} f(y_i - x_i^\top \beta \mid \sigma, \tau)$,
- There is no closed form for the posterior distribution, but usually explored via MCMC.

References

1. Koenker, R., & Hallock, K. F. (2001). Quantile Regression. The Journal of Economic Perspectives, 15(4), 143–156. http://www.jstor.org/stable/2696522
2. Angrist, J. D., & Jörn-Steffen Pischke. (2008). Mostly harmless econometrics: An empiricist's companion. (1st ed., pp. xiii–xiii). Princeton University Press. https://doi.org/10.1515/9781400829828
3. Wikipedia (2025). "Quantile regression."