1) Quantile Regression

Vitor Kamada

August 2019

Reference

Angrist & Pischke (2009). **Mostly Harmless Econometrics: An Empiricist's Companion.** Ch 7.

https://ebookcentral.proquest.com/lib/wayne/detail.action?docID=475846

Wooldridge (2010). **Econometric Analysis of Cross Section and Panel Data.** Ch 12.10

https://ebookcentral.proquest.com/lib/wayne/detail.action?docID=3339196&

Population Quantiles

$$Y \sim N(0,1)$$

$$q = Pr[Y \leq \mu_q] = F_y(\mu_q)$$

$$\mu_q = F_y^{-1}(q)$$

$$\mu_{0.5} = 0, \text{ then } Pr[Y \leq 0] = 0.5$$

$$\mu_{0.975} = 1.96, \text{ then } Pr[Y \leq 1.96] = 0.975$$

OLS vs Median and Quantile Regression

$$\sum_i u_i^2$$

$$\sum_i \mid u_i \mid$$

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

Standard Errors

$$\hat{\beta}_q \overset{a}{\sim} N(\beta_q, A^{-1}BA^{-1})$$

$$A = \sum_i q(1-q)x_i x_i'$$

$$B = \sum_i f_{u_q}(0|x_i)x_i x_i'$$

$f_{u_q}(0|x_i)$: conditional density of the error term $u_q = y - x'\beta_q$ at $u_q = 0$

Interpretation of Conditional Quantile Coefficients

$$Q_q(y_i|x_i) = \beta_1 + \beta_2 x_i + F_{u_i}^{-1}(q)$$

If errors are iid, then

$$F_{u_i}^{-1}(q) = F_{u}^{-1}(q)$$

$$Q_q(y_i|x_i) = \{\beta_1 + F_{u}^{-1}(q)\} + \beta_2 x_i$$

$$\frac{\partial Q_q(y|x)}{\partial x_j} = \beta_{qj}$$

Angrist et al. (2006): Returns to Schooling

$$Ln(wage) = \beta_q educ + Xs + u$$

		Desc. Stats.		Quantile Regression Estimates					OLS Estimates	
Census	Obs.	Mean	SD	0.1	0.25	0.5	0.75	0.9	Coeff.	Root MSE
1980	65,023	6.4	.67	.074 (.002)	.074 (.001)	.068 (.001)	.070 (.001)	.079 (.001)	.072 (.001)	.63
1990	86,785	6.5	.69	.112 (.003)	.110 (.001)	.106 (.001)	.111 (.001)	.137 (.003)	.114 (.001)	.64
2000	97,397	6.5	.75	.092 (.002)	.105 (.001)	.111 (.001)	.120 (.001)	.157 (.004)	.114 (.001)	.69

The sample includes US born white and black men aged 40-49

All models control for race and potential experience

Working Class Belgian Households in 1857

Roger & Hallock (2001). "Quantile Regression". Journal of Economic Perspectives, Vol 15(4),143–156

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
```

Engel dataset

```
data = sm.datasets.engel.load_pandas().data
data.head()
```

```
data.describe()
```

	income	foodexp
0	420.157651	255.839425
1	541.411707	310.958667
2	901.157457	485.680014
3	639.080229	402.997356
4	750.875606	495.560775

	income	foodexp
count	235.000000	235.000000
mean	982.473044	624.150111
std	519.230879	276.456997
min	377.058369	242.320202
25%	638.875788	429.688763
50%	883.984917	582.541251
75%	1163.986672	743.881432
max	4957.813024	2032.679190

OLS

```
ols = smf.ols('foodexp ~ income', data).fit()
print(ols.summary())
```

```
                coef    std err          t      P>|t|
-------------------------------------------------------
Intercept   147.4754     15.957      9.242      0.000
income        0.4852      0.014     33.772      0.000
```

```
ols = smf.ols('foodexp ~ income', data).fit()
ols_ci = ols.conf_int().loc['income'].tolist()
ols = dict(a = ols.params['Intercept'],
           b = ols.params['income'],
           lb = ols_ci[0],
           ub = ols_ci[1])
print(ols)
```

```
{'a': 147.47538852370573, 'b': 0.48517842367692354, 'lb': 0.4568738130184233, 'ub': 0.51348303433
```

Least Absolute Deviation

```python
mod = smf.quantreg('foodexp ~ income', data)
medianReg = mod.fit(q=.5)
print(medianReg.summary())
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                foodexp   Pseudo R-squared:               0.6206
Model:                       QuantReg   Bandwidth:                       64.51
Method:                 Least Squares   Sparsity:                        209.3
Date:                Mon, 29 Jul 2019   No. Observations:                  235
Time:                        16:49:51   Df Residuals:                      233
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     81.4823     14.634      5.568      0.000      52.649     110.315
income         0.5602      0.013     42.516      0.000       0.534       0.586
```

Many Quantiles between .05 and .95

```
quantiles = np.arange(.05, .96, .1)
def fit_model(q):
    res = mod.fit(q=q)
    return [q, res.params['Intercept'],
            res.params['income']] + \
            res.conf_int().loc['income'].tolist()

models = [fit_model(x) for x in quantiles]
models = pd.DataFrame(models,
             columns=['q', 'a', 'b', 'lb', 'ub'])

print(models)
```

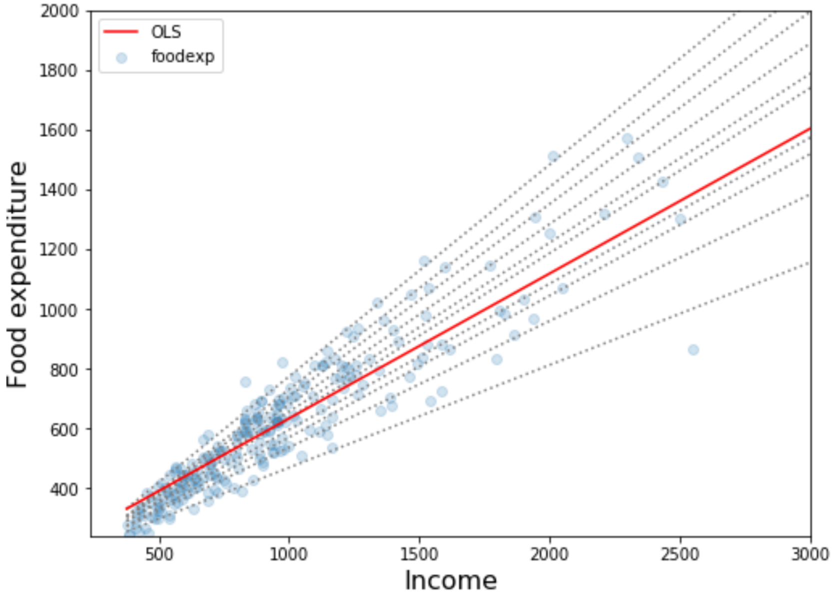
Result

	q	a	b	lb	ub
0	0.05	124.880099	0.343361	0.268632	0.418090
1	0.15	111.693660	0.423708	0.382780	0.464636
2	0.25	95.483539	0.474103	0.439900	0.508306
3	0.35	105.841294	0.488901	0.457759	0.520043
4	0.45	81.083647	0.552428	0.525021	0.579835
5	0.55	89.661370	0.565601	0.540955	0.590247
6	0.65	74.033433	0.604576	0.582169	0.626982
7	0.75	62.396584	0.644014	0.622411	0.665617
8	0.85	52.272216	0.677603	0.657383	0.697823
9	0.95	64.103964	0.709069	0.687831	0.730306

Plotting 10 Quantile Regression Models

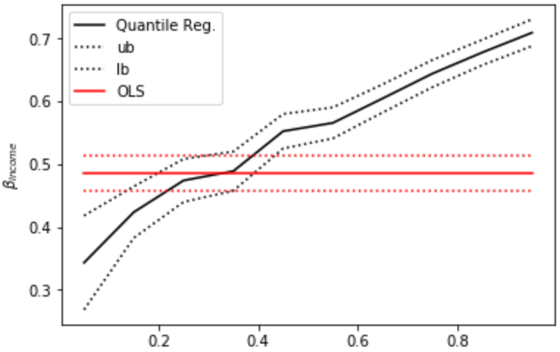
```
x = np.arange(data.income.min(), data.income.max(), 50)
get_y = lambda a, b: a + b * x

fig, ax = plt.subplots(figsize=(8, 6))

for i in range(models.shape[0]):
    y = get_y(models.a[i], models.b[i])
    ax.plot(x, y, linestyle='dotted', color='grey')

y = get_y(ols['a'], ols['b'])

ax.plot(x, y, color='red', label='OLS')
ax.scatter(data.income, data.foodexp, alpha=.2)
ax.set_xlim((240, 3000))
ax.set_ylim((240, 2000))
legend = ax.legend()
ax.set_xlabel('Income', fontsize=16)
ax.set_ylabel('Food expenditure', fontsize=16);
```

Quantile Regressions and OLS



Quantiles of the Conditional Food Expenditure

```
n = models.shape[0]
p1 = plt.plot(models.q, models.b, color='black', label='Quantile Reg.')
p2 = plt.plot(models.q, models.ub, linestyle='dotted', color='black')
p3 = plt.plot(models.q, models.lb, linestyle='dotted', color='black')
p4 = plt.plot(models.q, [ols['b']] * n, color='red', label='OLS')
p5 = plt.plot(models.q, [ols['lb']] * n, linestyle='dotted', color='red')
p6 = plt.plot(models.q, [ols['ub']] * n, linestyle='dotted', color='red')
plt.ylabel(r'$\beta_{income}$')
plt.xlabel('Quantiles of the conditional food expenditure distribution')
plt.legend()
plt.show()
```



Abadie, Angrist, and Imbens (2002)

$$Earnings_i = \rho JTPA_i + Xs + u_i$$

The Job Training Partnership Act (JTPA): subsidized training to disadvantaged American workers in the 1980s.

6,102 women and 5,102 men

60% of those offered training actually received JTPA services.

Z: randomly assigned offer of JTPA services

OLS and Quantile Regression

		Quantile				
Variable	OLS	.15	.25	.50	.75	.85
Training effect	3,754	1,187	2,510	4,420	4,678	4,806
	(536)	(205)	(356)	(651)	(937)	(1,055)
% Impact of training	21.2	135.6	75.2	34.5	17.2	13.4
High school or GED	4,015	339	1,280	3,665	6,045	6,224
	(571)	(186)	(305)	(618)	(1,029)	(1,170)
Black	−2,354	−134	−500	−2,084	−3,576	−3,609
	(626)	(194)	(324)	(684)	(1087)	(1,331)
Hispanic	251	91	278	925	−877	−85
	(883)	(315)	(512)	(1,066)	(1,769)	(2,047)
Married	6,546	587	1,964	7,113	10,073	11,062
	(629)	(222)	(427)	(839)	(1,046)	(1,093)
Worked < 13	−6,582	−1,090	−3,097	−7,610	−9,834	−9,951
weeks in past year	(566)	(190)	(339)	(665)	(1,000)	(1,099)
Constant	9,811	−216	365	6,110	14,874	21,527
	(1,541)	(468)	(765)	(1,403)	(2,134)	(3,896)

2SLS and Quantile Treatment Effect (QTE)

		Quantile				
Variable	2SLS	.15	.25	.50	.75	.85
Training effect	1,593	121	702	1,544	3,131	3,378
	(895)	(475)	(670)	(1,073)	(1,376)	(1,811)
% Impact of training	8.55	5.19	12.0	9.64	10.7	9.02
High school or GED	4,075	714	1,752	4,024	5,392	5,954
	(573)	(429)	(644)	(940)	(1,441)	(1,783)
Black	−2,349	−171	−377	−2,656	−4,182	−3,523
	(625)	(439)	(626)	(1,136)	(1,587)	(1,867)
Hispanic	335	328	1,476	1,499	379	1,023
	(888)	(757)	(1,128)	(1,390)	(2,294)	(2,427)
Married	6,647	1,564	3,190	7,683	9,509	10,185
	(627)	(596)	(865)	(1,202)	(1,430)	(1,525)
Worked <13 weeks in past year	−6,575	−1,932	−4,195	−7,009	−9,289	−9,078
	(567)	(442)	(664)	(1,040)	(1,420)	(1,596)
Constant	10,641	−134	1,049	7,689	14,901	22,412
	(1,569)	(1,116)	(1,655)	(2,361)	(3,292)	(7,655)