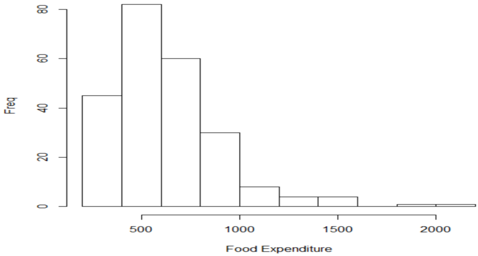
STATS762 Regression for Data Science

Quantile regression

May 1, 2019

Engel food expenditure

Ernst Engel surveyed food expenditures of 235 European working class households [1]



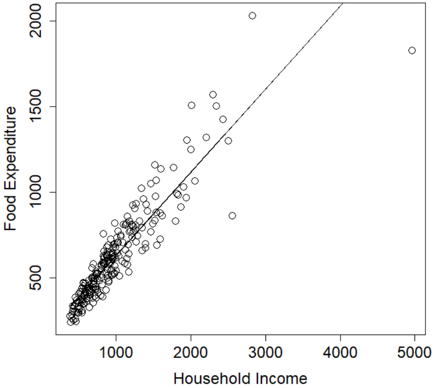
Food expenditure mean is 624.1501 and quantiles are

Quantile	0.10	0.25	0.50	0.75	0.90
	350.4664	429.6888	582.5413	743.8814	932.8867

[1] Engel, Ernst (1857). "Die Productions- und Consumtionsverhltnisse des Königreichs Sachsen". Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren. 89: 2829

Engel food expenditure

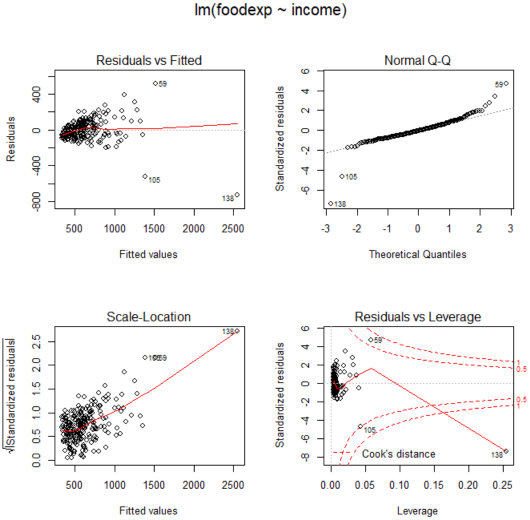
Relationship between household food expenditure and household income of 235 European working class households. [2]



[2] Engel, Ernst (1857). "Die Productions- und Consumtionsverhltnisse des Königreichs Sachsen". Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren. 89: 2829

Engel food expenditure

Relationship between household food expenditure and household income of 235 European working class households



Convenient methods

For the behaviour of y (outcome) conditional on x (predictor), consider regression $y_i = \mathbf{x}_i'\beta + \epsilon_i$, $i = 1, 2, ..., n$.

(a) Least squares (LS): Legendre (1805) [3]

- Minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2$ to obtain $\hat{\beta}$.
- $\mathbf{x}'\beta$ approximates the conditional mean of y given $\mathbf{x}$.

(b) Least absolute deviation (LDA): Boscovich (1755) [4]

- Minimizing $\sum_{i=1}^{n}|y_i - \mathbf{x}_i'\beta|$ to obtain $\hat{\beta}$.
- $\mathbf{x}'\beta$ approximates the conditional median of y given $\mathbf{x}$.

(c) Both the LS and LAD methods provide only partial description fo the conditional distribution of y.

[3] Legendre, Adrien-Marie (1805), Nouvelles mthodes pour la dtermination des orbites des cométes [New Methods for the Determination of the Orbits of Comets] (in French), Paris: F. Didot

[4] Boscovich, RJ. 1760. De recentissimis graduum dimensionibus, et figura, ac magnitudine terrae inde derivanda. Philosophiae Recentioris, a Benedicto Stay inRomano Archigynasis Publico Eloquentare Professore, vesibus traditae, Libri X, cum adnotianibus et Supplementis P. Rogerii Joseph Boscovich, S. J., 2: 406426

Quantile Regression

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.

We consider the relationship between the regressors and outcome using the conditional quantile function $Q_q(y|x)$, $0 < q < 1$.

Quantile Regression

The model prediction error for y_i is $e_i = y_i - \boldsymbol{x}_i'\beta$.

- LS minimizes $\sum_{i=1}^{n} e_i^2$.
- LAD minimizes $\sum_{i=1}^{n} |e_i|$ and this is equivalent to the median regression, $q = 0.5$.

Median regression is;

- More robust to outliers than LS regression.
- The optimal predictor is the conditional median, $med(yjx)$.

Let's use quantile regression to model conditional quantiles of the joint distribution of y and x.

Quantile Regression

Both the squared-error (LS) and absolute-error (LDA) 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.

Quantile regression minimizes a sum that gives asymmetric penalties $(1 - q)|e_i|$ for overprediction and $q|e_i|$ for underprediction.

Quantile Regression

The quantile regression estimator for quantile q minimizes the objective function

$$Q(\beta_q) = \sum_{i:y_i \geq \mathbf{x}_i'\beta}^{n} q|y_i - \mathbf{x}_i'\beta_q| + \sum_{i:y_i < \mathbf{x}_i'\beta}^{n} (1-q)|y_i - \mathbf{x}_i'\beta_q| \,.$$

Nondifferentiable function minimizing via the simplex method.

Quantile Regression

Advantages :

- More robust to non-normal errors and outliers.
- Impact of a covariate on the entire distribution of y, not merely its conditional mean.
- Invariant to monotonic transformation such the quantiles of $h(x)$, a monotone transform of y, are $h(Q_q(y))$, and the inverse transformation may be used to translate the results back to y.

Quantile Regression

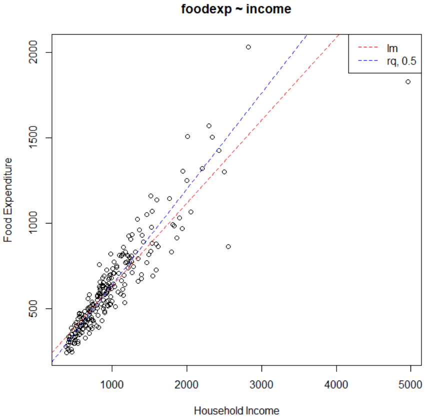
To fit the quantile regression using R:

```
library(quantreg)
out = rq( y ~ x , data, tau )
```

- rq-function fits the tau-quantile regression. By default tau=0.5.
- Usual commands are applicable; fitted.value, summary, predict and so on.

Engel food expenditure

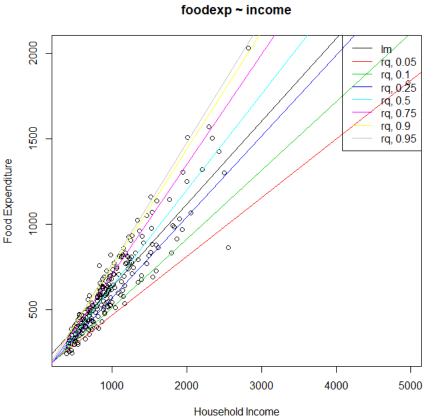
Comparison of LS fit and median regression (QR with $q = 0.5$).



Different median and mean regression fit due to asymmetry of the conditional density and partially by the strong effect of outliers.

Engel food expenditure

Comparison of LS fit and quantile regressions with various q values.



The conditional distribution is skewed to the left: the narrower spacing of the upper quantiles indicating high density and a short upper tail and the wider spacing of the lower quantiles indicating a lower density and longer lower tail.

Engel food expenditure

The output from summary is similar.

```
> data(engel)
> rqfit <- rq(foodexp ~ income, data = engel,tau=.5)
> summary(rqfit)

Call: rq(formula = foodexp ~ income, tau = 0.5, data = engel)

tau: [1] 0.5

Coefficients:
coefficients lower bd  upper bd
(Intercept)  81.48225     53.25915 114.01156
income        0.56018      0.48702   0.60199
```

Engel food expenditure

q	intercept	income
0.05	124.88004 (98.30212, 130.51695)	0.34336 (0.34333, 0.38975)
0.1	110.14157 (79.88753, 146.18875)	0.40177 (0.34210, 0.45079)
0.25	95.48354 (73.78608, 120.09847)	0.47410 (0.42033, 0.49433)
0.5	81.48225 (53.25915, 114.01156)	0.56018 (0.48702, 0.60199)
0.75	62.39659 (32.74488, 107.31362)	0.64401 (0.58016, 0.69041)
0.9	67.35087 (37.11802, 103.17399)	0.68630 (0.64937, 0.74223)
0.95	64.10396 (46.26495, 83.57896)	0.70907 (0.67390, 0.73444)

The $q = 0.9$ quantile regression curve displays the relationship of the food expenditure above 90% with Household Income for the population;

$$Q_{0.9}(\mathrm{FoodExpenditure}) = 67.35087 + 0.68630 * \mathrm{HouseholdIncome}\,.$$

Food expenditure below 10% with Household Income is

$$Q_{0.1}(\mathrm{FoodExpenditure}) = 110.14157 + 0.40177 * \mathrm{HouseholdIncome}\,.$$

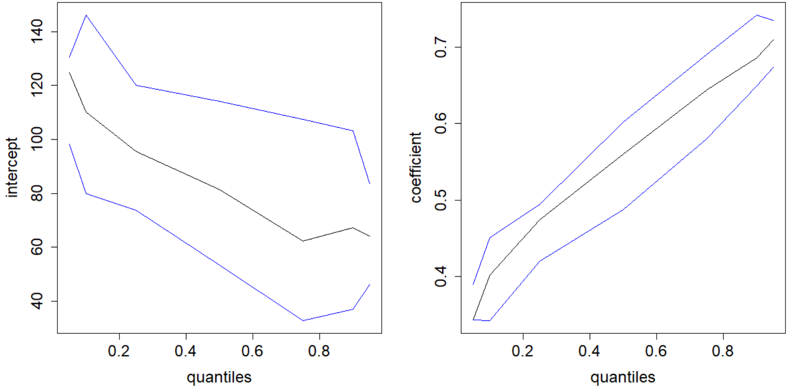
Engel food expenditure

Usual commands are used. For example, a 90% quantile of food expenditure given household income of 1520.

```
> rqfit <- rq(foodexp ~ income, data = engel,tau=.9)
> predict(rqfit,data.frame(income=1520))
1110.526
> rqfit$coefficients[1]+rqfit$coefficients[2]*1520
(Intercept)
1110.526
```

Engel food expenditure

Comparison of LS fit and quantile regressions with various q values.



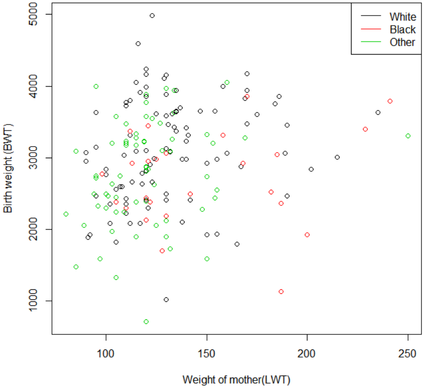
Birth Weight Study Data

Data contains information on 189 births in the obsetetrics clinic.[5]

Variable	Detail
ID	Identification code
LOW	Low birth weight indicator; 1 [BWT$\leq$2500g], 0 [BWT$>$2500g]
AGE	Age of mother
LWT	Weight of mother (lbs) at last menstrual period
RACE	Race group of mother; 1 [white], 2[black], 3[other]
SMOKE	Smoking status during pregnancy; 0 [no], 1 [yes]
PTL	Number of previous premature labors
HT	History of hypertension; 0 [no], 1 [yes]
UI	History of uterine irritability; 0 [no], 1 [yes]
FTV	Number of first trimester physician visits
BWT	Birth weight (grams)

[5] Hosmer D, Lemeshow S (2003). Applied logistic regression, 2nd edition. New York: John Wiley & Sons, Inc. Wiley.

Low Birth Weight Study Data



Relationship of the birth weight with weight of mother and ethic effect?

Birth Weight Study Data

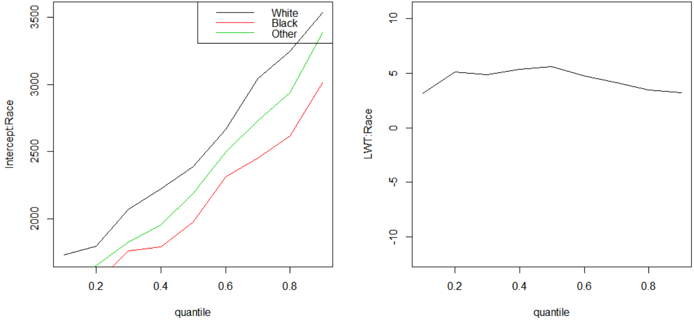
Model 1: BWT $\sim$ LWT + RACE

rq output

```
[1] "0.1 -quantile"
(Intercept)         LWT    RACEblack    RACEother
1730.888889    3.152778 -433.444444 -241.750000
[1] "0.2 -quantile"
(Intercept)         LWT    RACEblack    RACEother
1796.439024    5.146341 -278.463415 -146.975610
[1] "0.3 -quantile"
(Intercept)         LWT    RACEblack    RACEother
2065.857143    4.857143 -304.142857 -243.428571
[1] "0.4 -quantile"
(Intercept)         LWT    RACEblack    RACEother
2222.866667    5.386667 -431.266667 -266.533333
[1] "0.5 -quantile"
(Intercept)         LWT    RACEblack    RACEother
    2387.40        5.64      -414.92      -201.20
[1] "0.6 -quantile"
(Intercept)         LWT    RACEblack    RACEother
2663.229008    4.763359 -352.038168 -168.938931
[1] "0.7 -quantile"
(Intercept)         LWT    RACEblack    RACEother
3040.189655    4.155172 -589.724138 -309.655172
[1] "0.8 -quantile"
(Intercept)         LWT    RACEblack    RACEother
3246.595960    3.434343 -631.060606 -310.545455
[1] "0.9 -quantile"
(Intercept)         LWT    RACEblack    RACEother
3534.697674    3.224806 -521.875969 -151.674419
```

Birth Weight Study Data

Model 1: BWT $\sim$ LWT + RACE

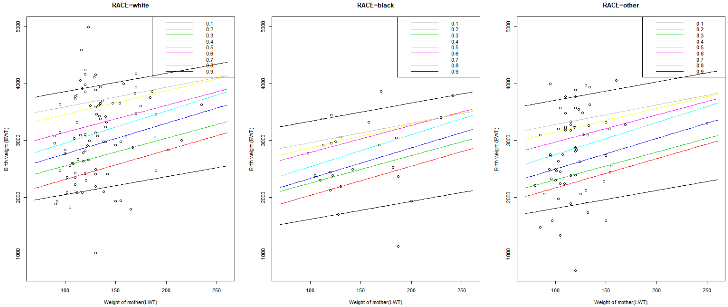


- Positive relation with weight of mother (LWT).
- Order of intercepts is White>Other>Black.

Birth Weight Study Data

Model 1: BWT $\sim$ LWT + RACE

Quantile regressions for the three groups.



Birth Weight Study Data

Model 1: BWT $\sim$ LWT + RACE

```
library(LogisticDx)
data(lbw)
> rqfit <- rq(BWT ~ LWT+RACE,data = lbw,tau=0.1)
> rqfit$coefficients
(Intercept)          LWT    RACEblack    RACEother
1730.888889     3.152778 -433.444444 -241.750000
> predict(rqfit,data.frame(LWT=130,RACE='white'))
1
2140.75
> 1730.88889+3.15278*130
[1] 2140.75
> predict(rqfit,data.frame(LWT=130,RACE='other'))
1
1899
> 1730.88889-241.75000+3.15278*130
[1] 1899
```

0.1 quantile of birth weight from a mother weighted 130 is 2140.75 when white and 1899 when other.

Low Birth Weight Study Data

Model 2: BWT $\sim$ LWT * RACE

```
> rqfit <- rq(BWT ~ LWT*RACE,data = lbw,tau=0.1)
> rqfit$coefficients
(Intercept)            LWT      RACEblack      RACEother
1471.753425       5.232877    2429.171948    -939.965546
LWT:RACEblack LWT:RACEother
-20.023921        5.851972
> predict(rqfit,data.frame(LWT=130,RACE='white'))
1
2152.027
> 1471.753425+5.232877*130
[1] 2152.027
> predict(rqfit,data.frame(LWT=130,RACE='other'))
1
1972.818
> 1471.753425-939.965546+(5.232877+5.851972)*130
[1] 1972.818
```

Birth Weight Study Data

Model 2: BWT $\sim$ LWT * RACE

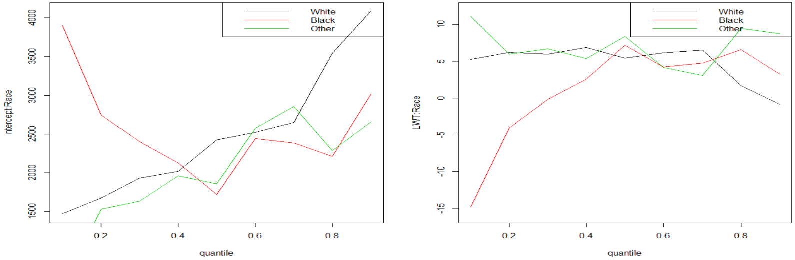
rq estimate

```
[1] "0.1 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 1471.753425    5.232877 2429.171948 -939.965546    -20.023921      5.851972
[1] "0.2 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
1669.9523810   6.2095238 1075.8253968 -144.8446886  -10.2984127    -0.2556777
[1] "0.3 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
1930.6619718   5.9859155  468.2648574 -297.6619718   -6.1566472     0.6940845
[1] "0.4 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 2016.157895    6.863158  110.933014  -59.824561     -4.272249     -1.476491
[1] "0.5 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 2423.500000    5.450000 -709.147059 -567.030612      1.726471      2.937755
[1] "0.6 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 2521.176471    6.164706  -81.824619   52.091822     -1.961002     -2.018364
[1] "0.7 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 2645.187500    6.531250 -261.937500  211.296371     -1.781250     -3.434476
[1] "0.8 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 3543.423729    1.677966 -1334.036341 -1256.876110     4.880592      7.774415
[1] "0.9 -quantile"
 (Intercept)         LWT   RACEblack   RACEother LWT:RACEblack LWT:RACEother
 4086.200000   -0.860000 -1073.378295 -1434.381818     4.084806      9.623636
```

Birth Weight Study Data

Model 2: BWT $\sim$ LWT * RACE

Intercept and coefficient by RACE.

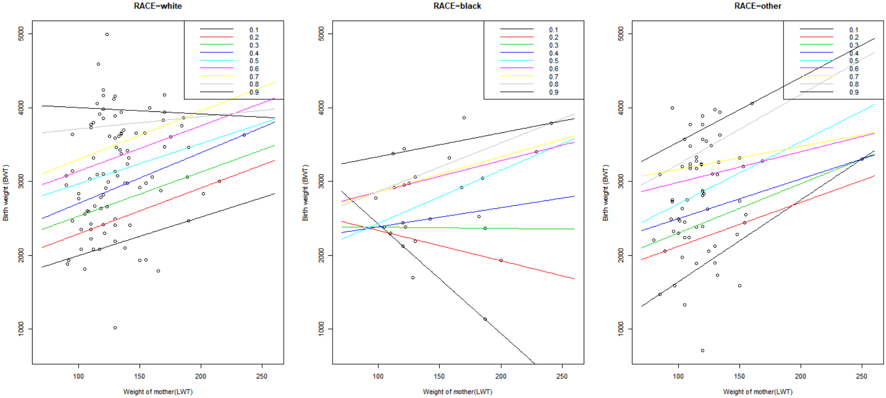


- Intercept order change about 0.4 quantile.
- In general positive relation with weight of mother (LWT).
- And what else do you observe?

Low Birth Weight Study Data

Model 2: BWT $\sim$ LWT * RACE

Quantile regressions for the three groups.



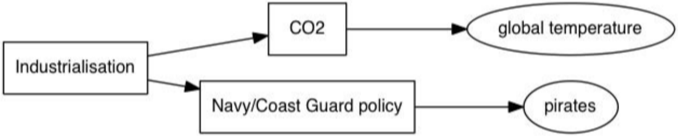
Casual graphs

- Graphical models used to encode assumptions about the data-generating process.
- Explain correlational and casual patterns.
- Helpful to explain relationship between factors.

Notations:

- $X \rightarrow Y$: X causes Y.
- $X \leftrightarrow Y$: X causes Y and Y causes X.
- $X - Y$: X covaries with Y.

Causal graph

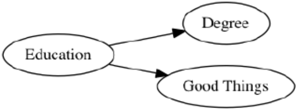


Causal graph

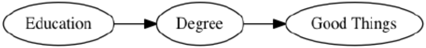
Education

We know that, compared to those whose formal education ends in high school, graduates have lower unemployment rates, higher salaries, better career prospects, and better health outcomes. (Chancellor's graduation speech)

You get a degree to learn stuff, which is useful

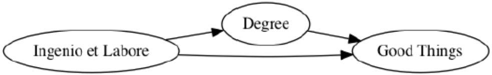


You get learn stuff to get a degree, which is useful

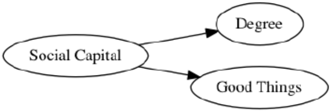


Causal graph

People who are smart and hard-working get degrees to prove this to potential employers, which is useful



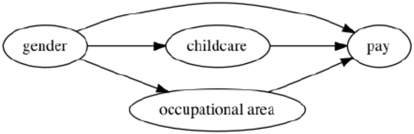
People who are the Right Sort of Person, which is useful, get degrees



Causal graph

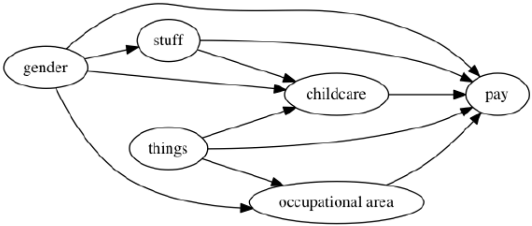
Gender gap in pay

But is it *really* an effect of gender?



Causal graph

Adding other relevant things and stuff



Still interesting questions about **mediation**: how much of the effect happens via specific paths