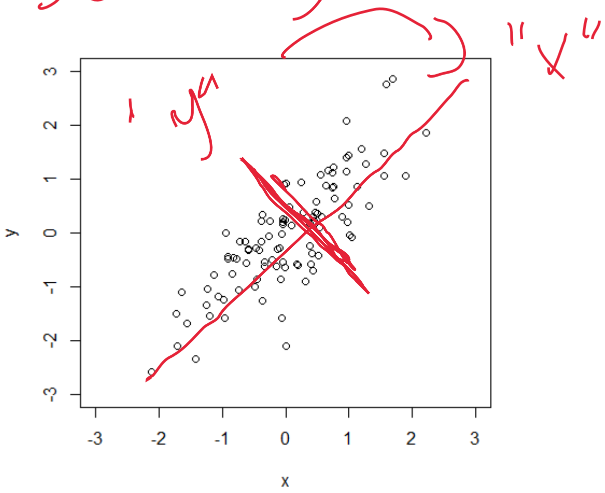
7.1: Principal Components and Quantile Regression

Dr. Bean - Stat 5100

1 Principal Components (PC) Regression

↳ dimensionality reduction



Principal Components is essentially a **re-projection** of the data into a new space where each axis follows the direction of the **highest variance** of the data, in descending order.

- Each component is a linear combination of the original X variables.

$$
\begin{aligned}
PC_{i,1} &= a_1^{(1)} X_{i,1} + \cdots + a_{p-1}^{(1)} X_{i,p-1} \\
PC_{i,2} &= a_1^{(2)} X_{i,1} + \cdots + a_{p-1}^{(2)} X_{i,p-1} \\
&\vdots \\
PC_{i,p-1} &= a_1^{(p-1)} X_{i,1} + \cdots + a_{p-1}^{(p-1)} X_{i,p-1}
\end{aligned}
$$

- Components derived from eigenvalues/eigenvectors of the matrix $X^T X$
- Often used as a form of dimensionality reduction.

Nice Mathematical Properties:

- $\rho(PC_j, PC_k) = 0 \quad \forall j \neq k$ → most desirable for OLS
- $\sum_j \left(a_j^{(k)}\right)^2 = 1$
- $Var(PC_1) \geq Var(PC_2) \geq \cdots \geq Var(PC_{p-1})$

1.1 Model

$$Y_i = \beta_0 + \beta_1 PC_{i,1} + \cdots + \beta_{p-1} \overset{PC}{\cancel{X}}_{i,p-1} + \epsilon_i$$

estimated with

$$\hat{Y} = \beta_0 + b_1 PC_{i,1} + \cdots + b_{p-1} \overset{PC}{\cancel{X}}_{i,p-1}.$$

Note: The PC's only consider relationships among the X variables and do not depend on Y.

Consequently, *all* of the principal components should be considered, not just the first few.

Why might dropping "low variance" principal components hurt our regression model?

Its possible that a "low variance" PC has high explanatory power for Y,

1.2 Pros and Cons

- **Pro:** Guaranteed uncorrelated predictors → no multicollinearity → meaningful model *coefficients.*
- **Con:** No guarantee of meaningful *variables.*

PC is not popular among statisticians "dark art"

```
proc princomp data=<dataset> standard out=<output dataset>;
var <all x variables>;
run;

proc reg data=PCout;
  model <y-variable> = Prin1-Prin<lastnumber> / vif;
run;
```

2 Quantile Regression

(known)

Quantiles: a set of q ranges for which there is an equal probability of an observation falling into each range.

value q for which there are X% of obs less than.

- Example: the median splits the observations into two groups, where 50% of the observations fall in each group.

In ordinary least squares (OLS) our goal was to model the **mean** of the response variable Y.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \qquad \cancel{X\epsilon} \quad \epsilon \sim N(0, \sigma^2)$$

with

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1}$$

based on the assumption that the model residuals were unbiased, normally distributed, with constant variance.

If the variance was not constant across Y, we were forced to consider variable transformations (Handout 2.2) or weighted least squares regression (Handout 4.2).

In quantile regression, heterskedasticity is seen as an **opportunity** to be pursued, rather than a **problem** to be fixed.

2.1 Motivating Example

(Groups) Looking at Figure ~~??~~ 1, how does the relationship between a Husband's education level and income change *across the quantiles of income*?

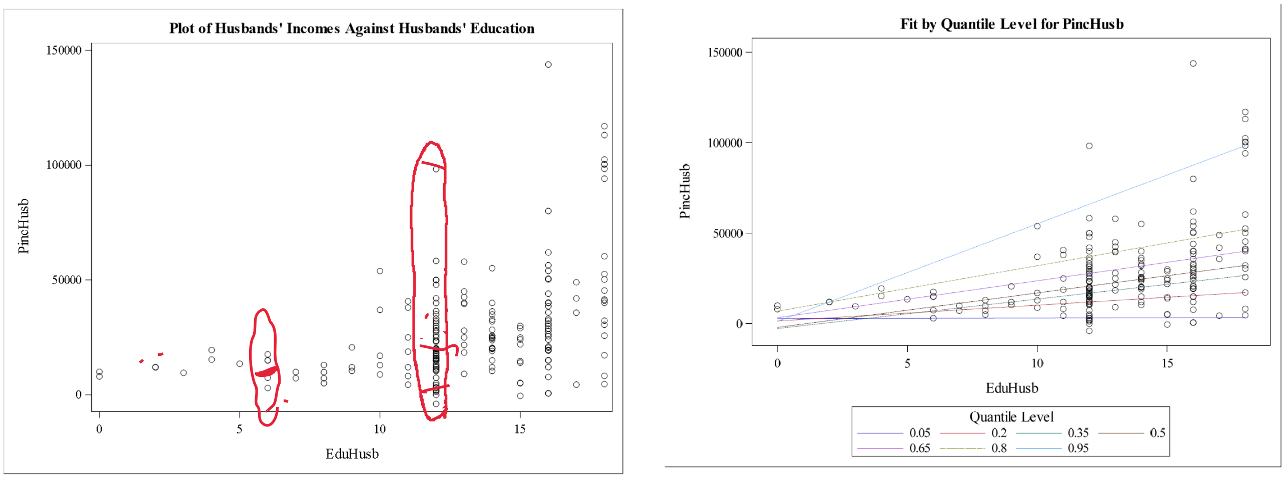


Figure 1: (Left) Plot of education vs income level. (Right) Series of quantile regression lines overlaid on the scatterplot of education and income.

Education matters more to high earners than low earners for any given level of education.

Simply trying to model effect of education on *average* income doesn't tell the full story.

2.2 Model

In ordinary least squares (OLS) regression, we assume the model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where estimated coefficients b_j were selected to minimize $\sum_i \left(Y_i - \hat{Y}_i\right)^2$.

In contrast, Quantile Regression selects $b_k(\tau)$ to minimize $\sum_i \rho_\tau \left(Y_i - Q_\tau(Y_i)\right)$, where

$$Q_\tau(Y_i) = b_0(\tau) + b_1(\tau) X_{i,1} + \cdots + b_{p-1}(\tau) X_{i,p-1}$$

τ	a quantile from 0 to 1
$b_j(\tau)$	estimated coefficient (a function of the quantile)
$Q_\tau(Y_i)$	the estimated τ quantile for the X-profile $X_{i,1}, \ldots, X_{i,p-1}$
$\rho_\tau(r)$	a "check loss" function $\rho_\tau(r) = \max\{\tau r, (\tau - 1)r\}$

(Groups) For the check loss function in Figure 2, X is the value of the residual and Y is the penalty associated with the residual. Knowing this, how do the "penalties" for $\tau = 0.3$ and $\tau = 0.9$ differ? Why does this difference seem reasonable?

Check loss function: encourages us to not fit the middle of the points.

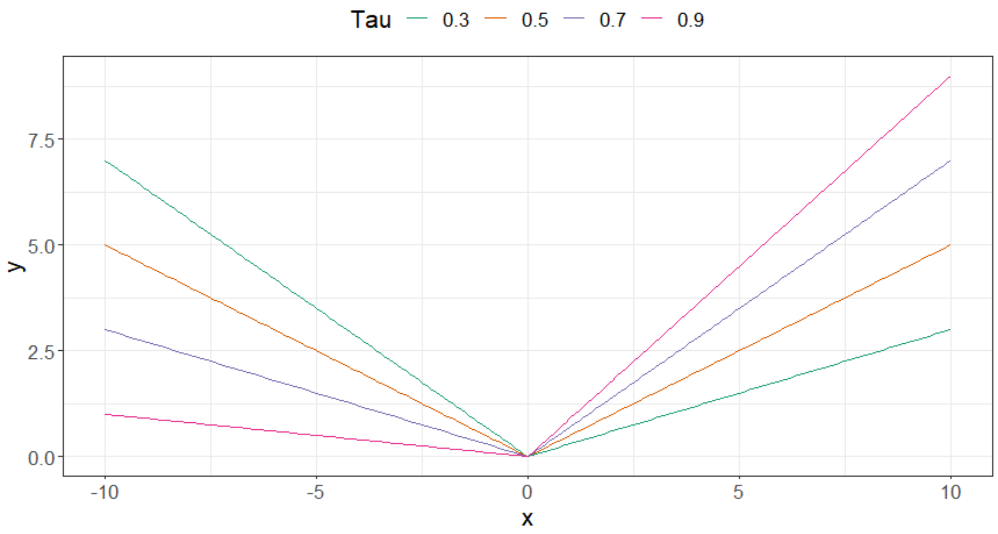


Figure 2: Check loss functions for various quantiles.

2.3 Comparison to OLS

OLS	**Quantile Regression**
Predicts conditional mean $E(Y\|X_1, X_2, \ldots)$	✓ Predicts conditional *distribution* (via quantiles)
Error terms must meet distributional assumptions	✓ No assumptions for error terms
Sensitive to outliers	✓ Robust to outliers (except extreme quantiles)
Effectively applied to small samples ✓	Data-hungry
Computationally cheap ✓	Computationally expensive (no closed-form solution/requires lots of quantile models).

2.4 SAS Code

```
proc quantreg data=<dataset>;
model <model statement> /
quantile= <low> to <high> by <increment>
plot=quantplot;
run;

proc quantselect data=<dataset> plots=coefficients;
```

```
model <model statement>
/ quantile = 0.1 0.5 0.9 selection=lasso(sh=<step horizon (integer)>);
partition fraction(validate=0.3);
run;
```

2.5 Good Resources

- Rodriguez, Bob and Yao, Yonggang (2017) “Give Things You Should Know About Quantile Regression” https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf
- Rodriguez, Bob (2018) “Three Things you Should Know About Quantile Regression” https://www.youtube.com/watch?v=CUOofd3hSOA