Quantile Regression

BIOS 6618

CU Anschutz

Quantile Regression

Motivation

We have primarily focused on linear regression where we model the mean of our outcome Y conditional on some predictors, $X_1, ..., X_p$.

However, there are certain cases where we may be concerned about the assumptions of linear regression:

- Outliers in our data
- Non-normal distribution of residuals
- Conditional distribution of Y is asymmetric
- Multiple modes
- Heteroscedastic variances

Some options to address these include data transformations, sensitivity analyses removing outliers, or using alternative modeling strategies. . .

Motivation

Quantile regression is one such alternative modeling strategy that is an extension of linear regression.[1]

Where linear regression models the conditional mean, quantile regression models the conditional quantile (e.g., median, 5th percentile, 75th percentile, etc.; often denoted by τ). This may be advantageous if our linear regression assumptions are not met, but it may also be useful if we are actually interested in modeling quantiles!

For example, one could construct growth curves by modeling different quantiles based on a dataset. Or, one may be interested in modeling across the distribution of the data to estimate associations in the tails.

[1] Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511754098

Extension from Linear Regression

In linear regression our model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i;\ \epsilon_i \sim N(0, \sigma^2_{Y|\mathbf{X}})$$

with a conditional mean given **X** of:

$$E(Y_i|\mathbf{X}) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$

Similarly, in quantile regression for any τth quantile we have:

$$Y_i = \beta_0^{(\tau)} + \beta_1^{(\tau)} X_{1i} + ... + \beta_k^{(\tau)} X_{ki} + \epsilon_i^{(\tau)};\ Q^{(\tau)}(\epsilon_i^{(\tau)}|\mathbf{X}) = 0$$

with a conditional quantile given **X** of:

$$Q^{(\tau)}(Y_i|\mathbf{X}) = \beta_0^{(\tau)} + \beta_1^{(\tau)} X_{1i} + ... + \beta_k^{(\tau)} X_{ki}$$

Extension from Linear Regression

In ordinary least squares (linear) regression, we estimated our regression coefficients based on minimizing the sum of squared errors:

$$SS_{Error} = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

For quantile regression of the median, we minimize the absolute sum of errors:

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

More generally, for any quantile τ we estimate $\hat{\beta}_k^{(\tau)}$ by minimizing the objective function:

$$Q(\beta^{(\tau)}) = \sum_{Y_i \geq \mathbf{X}\beta^{(\tau)}} \tau |Y_i - \mathbf{X}\beta^{(\tau)}| + \sum_{Y_i < \mathbf{X}\beta^{(\tau)}} (1-\tau)|Y_i - \mathbf{X}\beta^{(\tau)}|$$

However, this doesn't have a nice closed form like our OLS estimates & requires an algorithmic approach. We'll use the `quantreg` package in `R`.

Quantile versus Linear Regression

From a SAS proceedings, we have a nice summary comparing/contrasting the linear regression and quantile regression differences[2]:

Linear Regression	**Quantile Regression**
Models conditional mean $E(Y\|X)$	Models conditional quantiles $Q_\tau(Y\|X)$
Applies when n is small	Needs "sufficient" data
Often assumes normality	Is distribution agnostic
Does not preserve $E(Y\|X)$ under transformation	Preserves $Q_\tau(Y\|X)$ under transformation
Is sensitive to outliers	Is robust to response outliers
Is computationally inexpensive	Is computationally intensive

[2] Rodriguez, R. and Yao, Y. Five Things You Should Know about Quantile Regression (Paper SAS525-2017)

Need for "Sufficient" Sample Size

Whereas linear regression works well even when n is small (assuming assumptions are met), quantile regression needs a sufficient sample size for estimation.

What is sufficient?

The answer depends on what quantile(s) you are attempting to estimate! If you are estimating the median you may not need as large of a sample, but if you are estimating an extreme quantile more data is needed.

For example, if you only have $n = 20$, you may not have much confidence in estimating the 99th percentile. However, if $n = 2000$ we may be better able to estimate the 99th percentile.

Interpretation of Coefficients:

$$Q^{(\tau)}(Y_i|\mathbf{X}) = \beta_0^{(\tau)} + \beta_1^{(\tau)}X_{1i} + ... + \beta_k^{(\tau)}X_{ki}$$

Interpretation of coefficients:

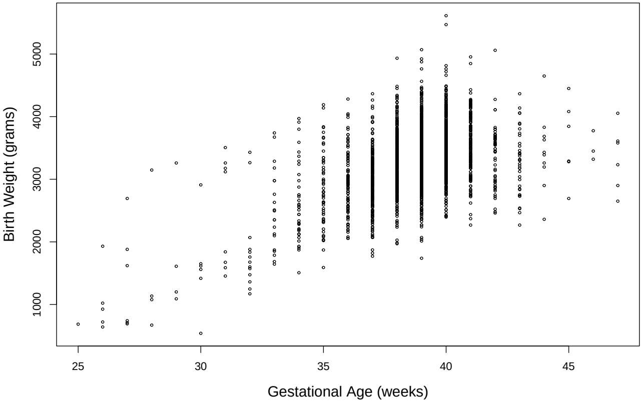
- Intercept: β_0 is the estimated value of Y for the τth quantile **when all other predictors**, $X_1, \ldots, X_k$, are equal to 0.
- Slope: β_j is the change in Y associated with a one-unit change in X_j for the τth quantile, assuming all other predictors are held constant.

Our interpretation is very similar to that of linear regression, except for it is with respect to estimated value of a given quantile or its potential difference between categorical predictors or for a one-unit change in a continuous predictor.

Quantile Regression Example

Quantile Regression in R

We will use the quantreg package in R for our example exploring the association between an outcome of infant birth weight (grams, dbwt) and a continuous predictor of gestational age (weeks, combgest) from a random sample of 2500 births from the 2018 US Natality Public Use File by estimating the 10th, 50th, and 99th percentiles:



quantreg Information

The `quantreg` package includes lots of options and can fit multiple quantiles/percentiles specified at once for fitting a quantile regression with the `rq()` function.

The `method` argument defines the algorithmic method used to estimate the parameters. Methods have different trade-offs for estimation, sample size, etc. and we will use the default `br` option.

Quantile Regression Estimates

```r
library(quantreg); library(ggplot2) # load packages
mod <- rq(dbwt ~ combgest, data=dat, tau=c(0.1,0.5,0.99)) # fit model
mod # print estimated coefficients
```

```
## Call:
## rq(formula = dbwt ~ combgest, tau = c(0.1, 0.5, 0.99), data = dat)
##
## Coefficients:
##              tau= 0.10  tau= 0.50 tau= 0.99
## (Intercept) -2855.3333 -1796.0000       774
## combgest      143.2222   131.3333        94
##
## Degrees of freedom: 2500 total; 2498 residual
```

The intercept in these models does not make scientific sense to interpret (i.e., predicted birth weight at a gestational age of 0 weeks).

For a 1 week increase in gestational age, the 10th/50th/90th percentile of birth weight increases by 143.2/131.3/94.0 grams.

Note the changing slope for different quantiles.

p-values for rq Object

```r
summary(mod, se='nid')[[2]] # summarize results for 0.5
```

```
## Warning in summary.rq(xi, U = U, ...): 47 non-positive fis

##
## Call: rq(formula = dbwt ~ combgest, tau = c(0.1, 0.5, 0.99), data = dat)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value       Std. Error  t value     Pr(>|t|)
## (Intercept) -1796.00000   215.01293    -8.35299     0.00000
## combgest      131.33333     5.48066    23.96304     0.00000
```

Using `summary.rq()` requires the specification of the `se` argument for how the standard errors will be calculated. For samples >1000, `nid` is the default.

We can see that p<0.001 (i.e., we reject the null hypothesis that $\beta_{combgest} = 0$), but no CIs are provided.

Note: warning *47 non-positive fis* is "generally harmless, leading to somewhat conservative (larger) estimate of the SEs" (http://www.econ.uiuc.edu/~roger/research/rq/FAQ).

Confidence Intervals for `rq` Object

If we really wanted confidence intervals, we would need to select an se option that facilitates it (e.g., `rank` if appropriate) or use a bootstrap approach:

```r
# bootstrap with 200 replicates for median beta coefficients
set.seed(6618) # set seed for reproducibility
boot_res <- boot.rq(x=mod$x, y=mod$y, tau=0.5, R=200)

# estimate 95% bootstrap percentile interval
sapply(1:ncol(boot_res$B),
       function(x) quantile(boot_res$B[,x], c(0.025, 0.975)))
```

```
##              [,1]     [,2]
## 2.5%  -2505.975 116.2271
## 97.5% -1201.888 150.0250
```

We are 95% confident that $\hat{\beta}_0$ is between -2506 and -1202 grams.

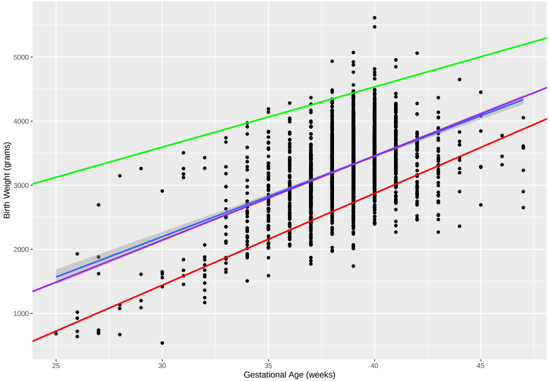
We are 95% confident that $\hat{\beta}_{combgest}$ is between 116 and 150 grams.

Note, we may wish to evaluate our bootstrap percentile interval performance (e.g., $|bias|/SE < 0.10$).

Plot Quantiles

We can also add our predicted quantile fits to our plot and add a linear regression fit for comparison:

```
ggplot(dat, aes(combgest,dbwt)) + geom_point() + geom_smooth(method='lm') +
  geom_abline(intercept=coef(mod)[1,1], slope=coef(mod)[2,1], size=1, color='red') +
  geom_abline(intercept=coef(mod)[1,2], slope=coef(mod)[2,2], size=1, color='purple') +
  geom_abline(intercept=coef(mod)[1,3], slope=coef(mod)[2,3], size=1, color='green') +
  xlab("Gestational Age (weeks)") + ylab("Birth Weight (grams)")
```



Closing Thoughts

This lecture serves as an introduction to quantile regression.

Your quantile regression can be more complex and include:

- Multiple predictors
- Non-linear terms (e.g., splines, polynomial terms)
- Nonparametric approaches to model estimation
- Censored data (e.g., survival or time-to-event data types)