7.2: Principal Components and Quantile Regression

Dr. Bean - Stat 5100

Suppose you are a dairy farmer trying to determine if a change in feed will lead to a significant increase in milk production, after accounting for the weight and heredity of the dairy cows. You notice that the model residuals are heteroskedastic across weight, though you determine that you can eliminate the heteroskedasticity with a log-transformation.

For this scenario, provide at least one argument in favor of OLS regression with the transformed data, vs quantile regression with the un-transformed data.

Arguments for OLS:

- Better suited for variable inference (exact solution with unbiased, minimum variance coefficient estimates).
- Much easier to fit computationally.
- Less model output (one model as opposed to many).

Arguments for Quantile Regression:

- More information rich: quantifies the differences in milk production across quantiles.
- Requires no variable transformation (easier to explain predictions).
- Outlier observations less influential.

Recall the form of the check loss functions. Why would a check loss function be more robust to outlier values than a squared loss function?

Outlier values are associated with large residuals. Squaring large residual values makes them especially large in the loss function. Check loss functions experience a linear increase, rather than a quadratic increase, in the penalty for outlier observations which gives outliers less influence in the optimization.

What would have to be true of the model residuals if the mean predictions (OLS) were significantly different from the median predictions of quantile regression?

Either the residuals are highly skewed, or there are outlier values having an undue influence on the estimated coefficients for the mean predictions.

What would be the issue with trying to estimate the quantile regression model associated with the 95th percentile with a small sample size? Would you have the same problem trying to estimate the model associated with the median?

A reliable estimate of the 95th percentile requires a lot of observations at or near each X-profile. Data "saturation" across all relevant X-profiles is difficult to achieve in small sample sizes. The median is a much more robust estimate that does not require nearly as much data saturation. As such, a quantile regression model fit to the median can be reasonably fit to small sample sizes. This is not true for some of the more extreme percentiles.