Heteroskedastic Data Modeling

Jonathan Bryan

April 20, 2018

Problem

https://www3.nd.edu/~rwilliam/stats2/l25.pdf

Many statistical models assume constant variance of the error term when modeling a response variable against predictors. Heteroskedasticity breaks this assumption and can present unique challenges to inference and prediction. Heteroskedasticity occurs when the conditional variance of the response Y given the predictors X is also a function of t time or some vairable that has an ordered scale. In other words, the variance of the error term may change over time or with an ordered variable. Examples include increasing variation of measurement error of consumption habits with income (more money to buy various goods) and greater error variance of capital expenditures based on company size.

Approach

We simulate five datasets where $n = 300$, $p = 1, 2, 10, 100, 250$ and where the error term of the model is a linear function of one covariate in each dataset. We then simulate four more datasets where $n = 300$ and $p = 2, 10, 100, 250$ and half of the covariates are modestly correlated with each other $(0.2 \leq \rho \leq 0.5)$ but not including the covariate that is driving the heteroskedasticity, except for in the $p = 2$ case. These four datasets will help demonstrate the effect of multicollinearity on detection and modelling of heterskedastic data.

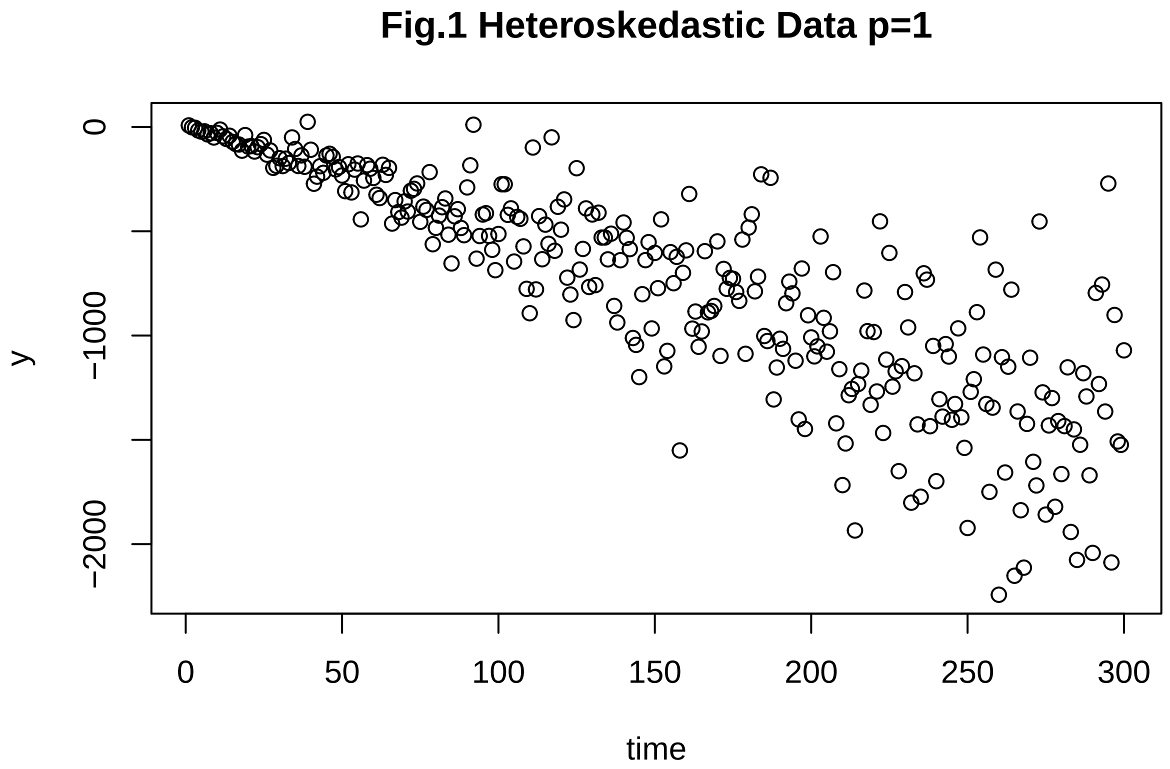
A survey of exploratory data analysis methods and test statistics is given to show how heteroskedasticity can be detected in low- and high-dimensional data and the effect of multicollinearity on detection. A comparison of ________ regression models is given to highlight advantages and disadvantages to different approaches.

```
## Warning: package 'plotly' was built under R version 3.4.4
```

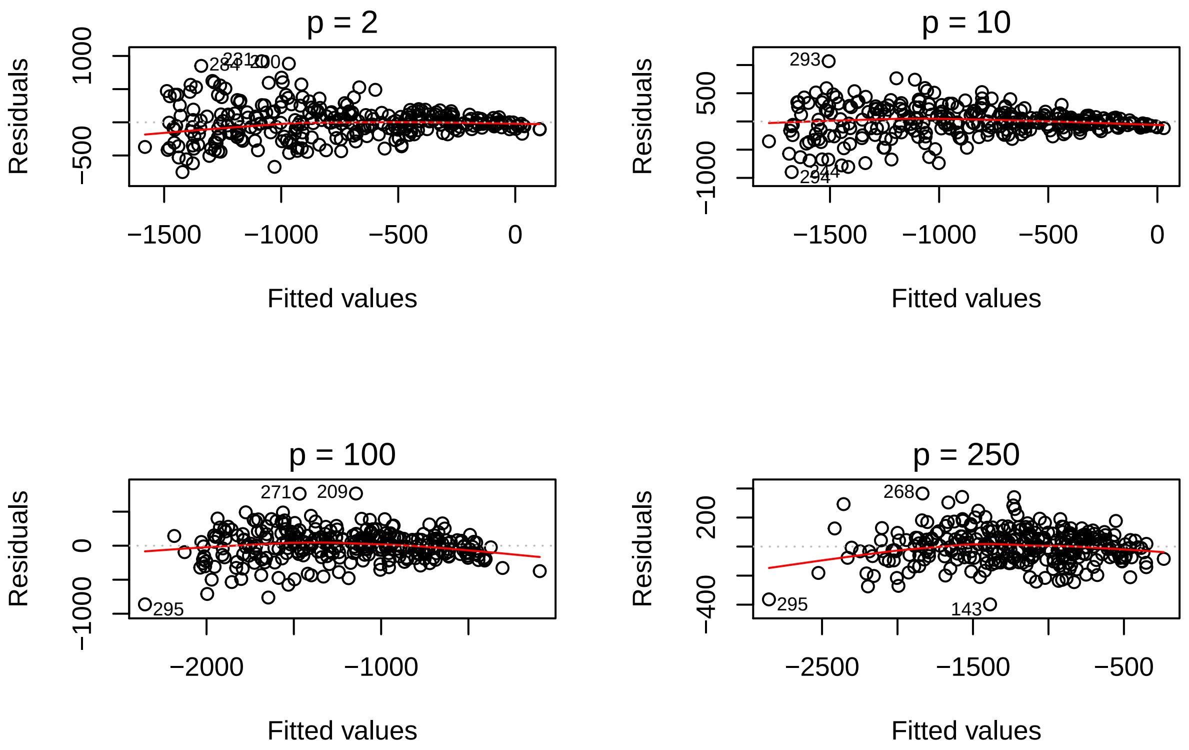
4) Heteroskedasticity tests: all tests with R support
5) Corrective Data Transformations
6) Regression modeling: low vs. high dimension and no vs. mullticol for each
7) Classification modeling: low vs. high dimension and no vs. mullticol for each

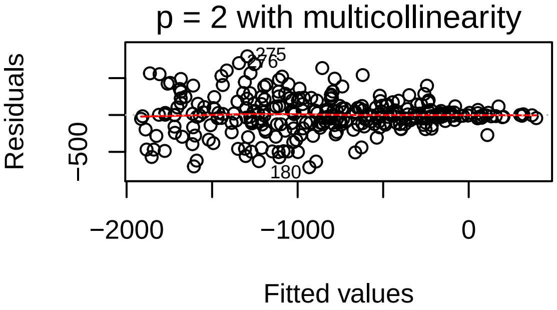
Visual Detection of Heteroskedasticity

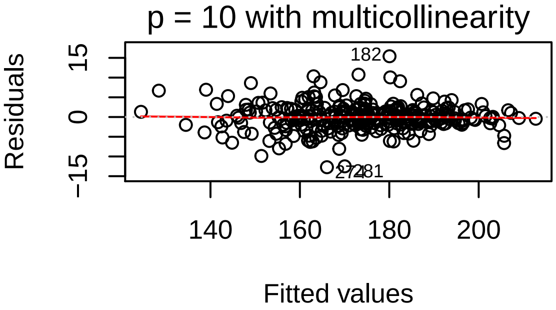
Response-Predictor Plots When using only one predictor to model the response, we can directly observe the behavior of the response variable along the scale of the predictor. We simulate a simple low dimensional dataset $(n = 300, p = 1)$, specified as $Y \sim N(\mu, \sigma^2)$ where $\mu = 10 + time * \beta$, $\beta = 2$, $\sigma^2 = 1.5 * time$ where the error term is indexed by a multiple of time. Notice that the dispersion of the Y response increase with time.

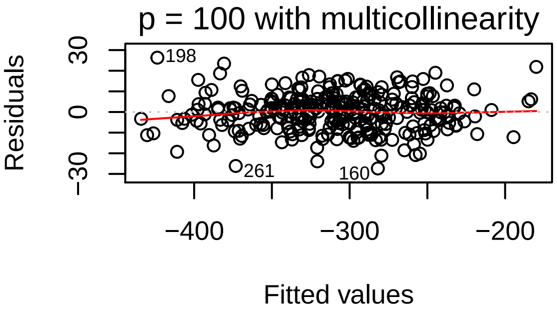


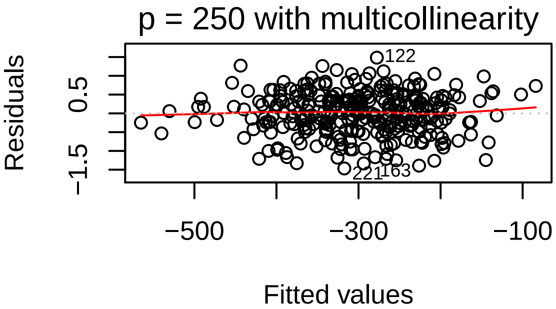
For multiple regression, rather than looking at several different plots of each variable against the response, residuals plots can be used to detect non-constant variance of the error term. If we oberve that the the variance of the residuals is increasing along the scale of the fitted values, this is a good indication that heterskedasticity is present in the data. We see that as $p \rightarrow n$ it is harder to visually detect heteroskedasticty. In addition, when there is even modest multicollinearity among the covariates ($p < 0.05 * n$), visual detection using residual plots becomes nearly impossible.


p = 2
Residuals
1000
500
0
−500
284
231
200
−1500
−1000
−500
0
Fitted values
p = 10
Residuals
500
0
−1000
293
294
244
−1500
−1000
−500
0
Fitted values
p = 100
Residuals
0
−1000
271
209
295
−2000
−1000
Fitted values
p = 250
Residuals
200
0
−400
268
295
143
−2500
−1500
−500
Fitted values









Statistical Tests of Heteroskedasticity

Goldfeld-Quandt test This test creates two subsets of the data to run separate ordinary least squares regressions. The separation point of the subsets can be arbitrary and do not have to contain all observations within each subset. The test assumes Gaussian errors and that both subsets of the data are full rank. The Goldfeld-Quandt test statistic is the ratio of the mean square errors of the regressions on the two subsets of the data. The test statistic is:

$$\frac{MSE_1}{MSE_2} \sim F_{n_1-k,\, n_2-k}$$

Where MSE_1 and MSE_2 are the mean sqaured error of the first and second half of the data respectively. The null hypothesis follows and F distribution with $n_1 - k,\ n_2 - k$ degrees of freedom. Increasing the number of observations dropped in the "middle" of the ordering will increase the power of the test but reduce the degrees of freedom for the test statistic(CITE). The Goldfeld-Qaundt test requires the be data ordered against a predictor. This may be a time-consuming step if there are many unordered predictors in the model. In addition, the test assumes the variance of the error term must be a monotonic function of the specified explanatory variable. For example, a if the variance of the error term is a nonlinar function of the explanatory variable the Goldfeld-Quandt test may erroneously accept the null hypothesis of homoskedastic errors (CITE).

Advantages

- Simple test procedure and statistic
- F-distribution is the asymptotic sampling distribution of the test statistic

Disadvantages

- Requires target explanatory variable to be ordered

- Does not detect heteroskedasticty contributes by an unknown variables
- Assumes variance of the error term is a monotonic function of the explanatory variable
- High-dimensional data

Table 1: Table 1. Statistical Tests for Heteroskedasticity

	1	2	10	100	250	2 MC	10 MC	100 MC	250 MC
Goldfeld-Quandt	0	0	0	0.000	NA	0	0	0.000	NA
Harrison-McCabe	0	0	0	0.000	NA	0	0	0.000	NA
Breusch-Pagan	0	0	0	0.224	0.282	0	0	0.246	0.629

White test

Why it breaksdown in high dimensions

http://www.jstor.org.proxy.lib.duke.edu/stable/pdf/2988471.pdf?refreqid=excelsior:40ccf64937d1b684ec45d40dc0c9d3fa

https://rpubs.com/cyobero/187387

Corrective Data Transformations

Statistical Modeling

In the linear regression setting, the maximum likelihood estimates of the coefficients remain unbiased in the presence of heteroskedasticity, however the standard error estimates are not which can effect whether a explanatory predictor is found to be significant. We see below that even with modest heterskedasticity our coefficient estimate and significance test for the time variable reflect the true model. However, when we increase the dimension of our data by one variable $p = 2$, we no longer have an accurate coefficient estimate for the additional variable and the only significant variable is time. When we further increase the dimensions of the data to $p = 10$ we lose two more signficant predictors.

```
#summary
#stargazer::stargazer(lm_p1.fit)
```

Regression models

https://lib.ugent.be/fulltxt/RUG01/002/376/288/RUG01-002376288_2017_0001_AC.pdf

Classification models

http://www.stat.columbia.edu/~gelman/research/published/aiepub.pdf