## Data Inspection

We plot the x- and y- values of the data sets as a scatter plot for visual inspection. As seen in Figure 1 [Appendix], we observe that the y-values have variance increasing with the absolute value of x, in data sets data_1_1,2,3,5, and observe that the variance increases with x in data_1_4. This is symptomatic of conditional heteroskedaticity in errors, which leads to the (Ordinary Least Squares) OLS estimator to be inefficient. Due to the non-linear form of heteroscedasticity (such as the hour-glass shape of the errors), we use the White test over the Breusch-Pagan to inspect heteroscedasticity in errors. We observe that data_1_1,4,5 display heteroscedasticity to 5% significance level, data_1_3 to 10% significance level and data_1_2 to have homoscedastic errors (Table 1).

We then test for serially correlated errors using the Durbin-Watson test and observe that the data does not display serially correlated errors to 5% significance level (Table 1).

Table 1: Results of Durbin-Watson and White tests.

| Data set | Durbin-Watson Test* | White-Test (p-value)** |
|---|---|---|
| data_1_1 | 2.329 | 0.0099 |
| data_1_2 | 2.181 | 0.1888 |
| data_1_3 | 1.922 | 0.0806 |
| data_1_4 | 1.936 | 0.0054 |
| data_1_5 | 2.048 | 0.0014 |

* Test for serially correlated errors. At 5% significance level, the upper bound dU = 1.694 (k = 1, n = 100 (data_1_1/2)) and dU = 1.585 (k = 1, n = 50 (data_1_3/4/5))

** Test for Heteroskedastic errors

We also observe that the test for normal distribution of errors, using the Jarque-Bera test (values not cited here), showed that the errors were not normally distributed at 5% significance level.

## Estimators

Features such as heteroskedasticty reduce the efficiency of the OLS estimated parameters and we attempt to correct for that by using the Weighted Least Squares (WLS), the Feasible Generalized Least Squares (fGLS) and the Quantile Regression (QR) estimators. It should be noted that the OLS, WLS, fGLS and QR parameters are consistent for the same population parameters; they differ in their asymptotic efficiency.

We formulate the model of the data generating process as

$$y_i = a \cdot x_i + b.$$

**[1] The OLS estimator:** When the Gauss Markov conditions are satisfied the OLS estimator is BLUE. The loss function of the parameters is given as follows:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Solving the above yields,

$$\hat{a} = \frac{cov(x_i, y_i)}{var(x_i)} \quad , \qquad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$$

**[2] The WLS estimator**: The WLS estimator is efficient and unbiased but it depends on knowing the variance structure. We formulate the loss function as follows:

$$\min_{a,b} \sum_{i=1}^{n} \left( \frac{y_i}{w_i} - a\frac{x_i}{w_i} - \frac{b}{w_i} \right)^2$$

In our estimation, we propose two variance structures:

1. Due to the negative values in $\boldsymbol{x}$ in data sets data_1_1,2,3,5, we assume that the heteroscedasticity has the form $Var(u_i) = \sigma_i^2 x_i^2$ and model it using the weight structure $w_i = x_i$ . The intercept of the transformed model is the slope of the original model and that the slope of the transformed model is the intercept of the original one. The optimal coefficients turn out to be as follows:

$$\hat{a} = \frac{\sum_{i=1}^{N} \frac{y_i}{x_i} - \bar{y} \sum_{i=1}^{N} \frac{1}{x_i}}{N - \bar{x} \sum_{i=1}^{N} \frac{1}{x_i}}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x}$$

2. For data set data_1_4, we assume that the heteroscedasticity has the form $Var(u_i) = \sigma_i^2 x_i$ and model it using $w_i = x_i^{1/2}$. Since the weight is reciprocal of the variance, the WLS estimator is BLUE. The optimal coefficients turn out to be as follows:

$$\hat{a} = \frac{\sum_{i=1}^{N} x_i^{1/2} \sum_{i=1}^{N} y_i x_i^{-1/2} - \sum_{i=1}^{N} y_i x_i^{1/2} \sum_{i=1}^{N} x_i^{-1/2}}{\sum_{i=1}^{N} x_i^{-1/2} \left( \sum_{i=1}^{N} x_i^{1/2} + \sum_{i=1}^{N} x_i^{3/2} \sum_{i=1}^{N} x_i^{-1/2} \right)}$$

$$\hat{b} = \frac{\sum_{i=1}^{N} y_i x_i^{1/2} - \hat{a} \sum_{i=1}^{N} x_i^{3/2}}{\sum_{i=1}^{N} x_i^{1/2}}$$

**[3] The fGLS estimator:** When the variance structure is not known the fGLS estimator can be used. The estimator is unbiased though not efficient.

Procedure to correct for heteroscedasticity:

1. Run the regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ and obtain residuals on $\widehat{\boldsymbol{u}}$
2. Compute $\log(\widehat{\boldsymbol{u}}^2)$
3. Run the regression of $\log(\widehat{\boldsymbol{u}}^2)$ on $\boldsymbol{x}$: $\log(\widehat{\boldsymbol{u}}^2) = \delta_1 \boldsymbol{x} + \delta_{\boldsymbol{0}} = \widehat{\boldsymbol{g}}$. The $log$ ensures that the regression yields positive values since the dependent variable is the variance of errors.
4. Exponentiate the fitted values from 3: $\widehat{\boldsymbol{h}} = \exp(\widehat{\boldsymbol{g}})$
5. Estimate $\boldsymbol{y} = a \cdot \boldsymbol{x} + b$ using WLS with $\boldsymbol{W} = diag\{\widehat{\boldsymbol{h}}\}$.

This process yields the following loss functions:

$$\widehat{a_{fGLS1}} = \left( \boldsymbol{x}' \, \hat{\sigma}_{OLS}^{-\boldsymbol{2}} \, \boldsymbol{x} \right)^{-1} \boldsymbol{x}' \hat{\sigma}_{OLS}^{-\boldsymbol{2}} \boldsymbol{y}$$

$$\widehat{\boldsymbol{u}}_{fGLS1} = \boldsymbol{y} - \boldsymbol{x} \cdot \hat{a}_{fGLS1}$$

$$\hat{a}_{fGLS2} = \left( \boldsymbol{x}' \hat{\sigma}_{fGLS1}^{-\boldsymbol{2}} \, \boldsymbol{x} \right)^{-1} \boldsymbol{x}' \hat{\sigma}_{fGLS1}^{-\boldsymbol{2}} \, \boldsymbol{y}$$

**[4] The Quantile Estimator:** The quantile regression aims at estimating either the conditional median or other quantiles of the response variable. This contrasts with the method of least square which results in estimating that approximate the conditional mean of the response variable. The estimator is consistent and is more robust to outliers than the OLS estimator. The loss function can be expressed as follows:

$$\min_{a,b} \sum_{i:y_i \geq ax_i}^{n} q\,|y_i - ax_i - b| \quad + \sum_{i:y_i < ax_i}^{n} (1-q)\,|y_i - ax_i - b|$$

Since the loss function is piecewise linear, solving it is a linear programming problem.

The optimal quantiles of the QR estimator need to be established. We do this by examining the standard error of the slope yielded by the estimator for the following quantiles: [0.25, 0.45, 0.75, 0.95], and choose the quantile yielding the lowest standard error. At the 50% percentile the QR formulates into the Least Absolute Deviation (LAD) estimator. LAD gives equal emphasis to all observations, in contrast to OLS which, by squaring the residuals, gives more weight to large residuals. The LAD estimator however is frequently unstable, hence we refrain from including it in our estimator selection. Table 2 shows the optimal quantile for each data set.

Table 2: Quantile selection for Quantile Regression.

| Data set* | Standard error | Quantile |
|---|---|---|
| data_1_1 | 0.150 | 0.45 |
| data_1_2 | 0.098 | 0.45 |
| data_1_3 | 0.158 | 0.45 |
| data_1_4 | 0.529 | 0.75 |
| data_1_5 | 0.184 | 0.45 |

* The HC3 standard error is used in all data sets and estimators except in data_1_2.

## Assessment of Parameters

We examine the performance of the OLS, fGLS, WLS and QR estimators by examining the standard errors of their estimates of the slope when regressed on each of the five data sets. The criterion of standard errors is chosen because one would rather have that parameters were close to the truth with high probability as the sample size grows, *i.e.* their parameter estimates were consistent. The

standard errors for the data sets that exhibited heteroskedsticity are inconsistent. As a remedy, heteroscedasticity consistent (HC) standard errors are used (the HC3 standard error is particularly recommended). The most efficient estimator, i.e. one with the lowest standard error, is chosen as the best estimator for the data. The use of inferential statistics (such as the p-value) to determine the best estimator is not used because of the non-normality of errors as observed earlier.

As per the chosen criterion, the performance of the estimators is shown in Table 3 with the best estimator highlighted in grey.

Table 3: Standard error of Regression slope (\hat{a}) of OLS, fGLS, WLS and QR estimators

| Data set* | OLS | fGLS | WLS | QR |
|---|---|---|---|---|
| data_1_1 | 0.272 | 0.262 | 0.268 | 0.148 |
| data_1_2 | 0.160 | 0.160 | 0.413 | 0.097 |
| data_1_3 | 0.315 | 0.311 | 0.249 | 0.169 |
| data_1_4 | 0.724 | 0.405 | 0.444 | 0.529 |
| data_1_5 | 0.461 | 0.656 | 0.626 | 0.192 |

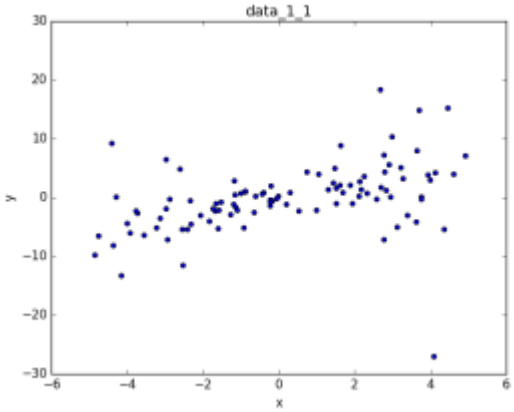
* The most efficient estimator for each data set is highlighted in grey.

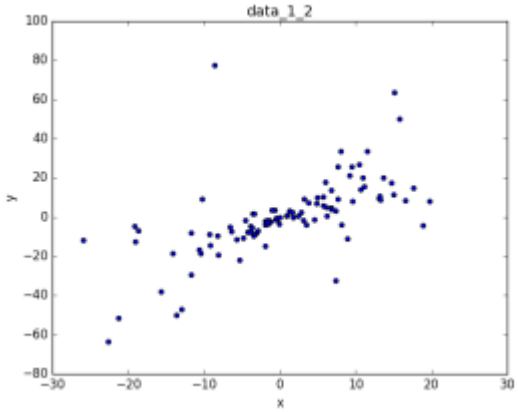
** The HC3 standard error is used in all data sets and estimators except in data_1_2.

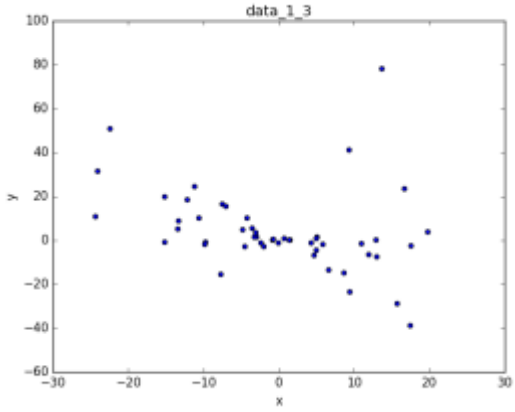
The results show that with exception to data set data_1_4 the QR estimator compares favorably over the other estimator -it yields estimates of the slope with the lowest standard error. The fits of the estimators are shown in Figure 2 [Appendix]. Tables A-D [Appendix] show other performance figures of the estimators over each of the data sets.

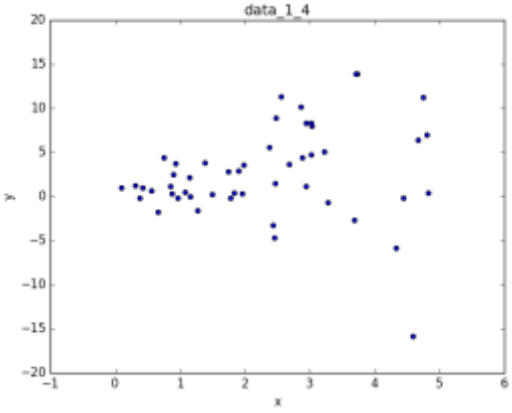
**APPENDIX**

Figure 1. Scatter plots of data points in data sets data_1_1,2,3,4,5



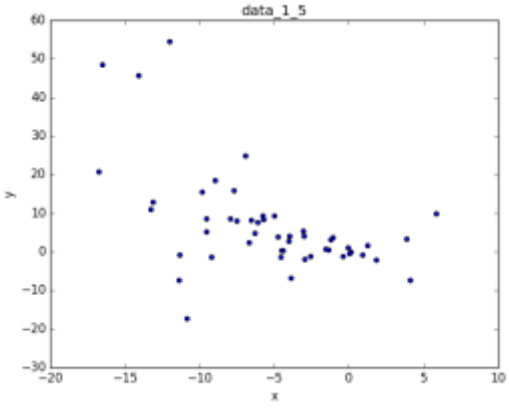



Figure 2. Regression fit of OLS, fGLS WLS and QR estimators.

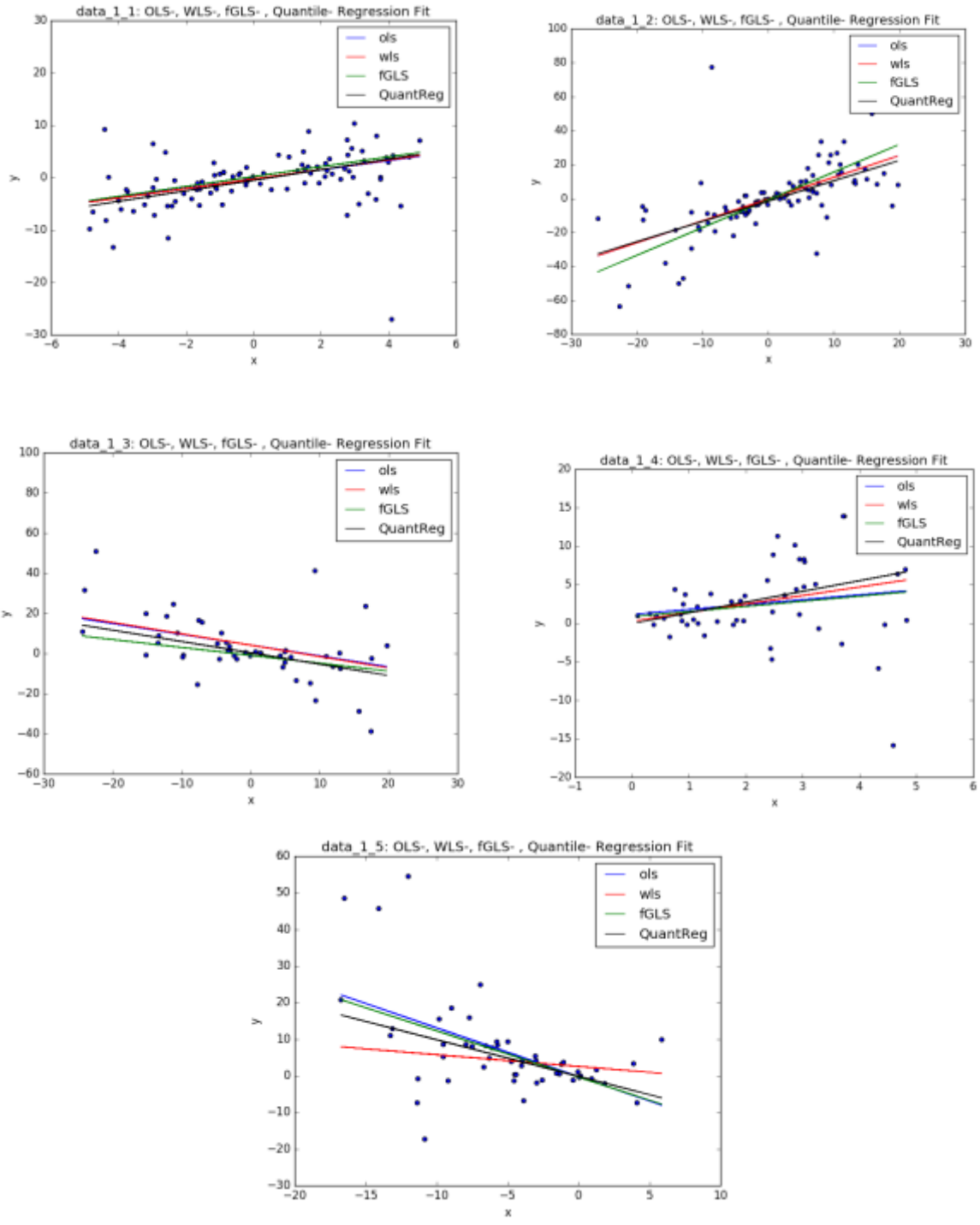


Table. A OLS Estimator

| | data_1_1 | data_1_2 | data_1_3 | data_1_4 | data_1_5 |
|---|---|---|---|---|---|
| std. err: x | 0.272 | 0.224 | 0.315 | 0.724 | 0.461 |
| std. err: const | 0.524 | 1.619 | 2.692 | 1.133 | 1.789 |
| x | 0.8808 | 1.2856 | -0.5378 | 0.6334 | -1.3331 |
| const | -0.2695 | -0.3959 | 4.1752 | 1.1481 | -0.2413 |
| z:x | 3.239 | 5.743 | -1.706 | 0.875 | -2.894 |
| z:const | -0.514 | -0.244 | 1.551 | 1.013 | -0.135 |
| P:x | 0.001 | 0 | 0.088 | 0.382 | 0.004 |
| P:z | 0.607 | 0.807 | 0.121 | 0.311 | 0.893 |
| R-Squared | 0.164 | 0.398 | 0.103 | 0.028 | 0.281 |
| Adj. R-Squared | 0.155 | 0.392 | 0.085 | 0.007 | 0.266 |
| Durbin-Watson Test | 2.329 | 2.181 | 1.922 | 1.936 | 2.048 |
| White-Test: F Stats. | 9.232 | 3.334 | 5.03697 | 10.4545 | 13.2078 |
| p-value | 0.0099 | 0.18878 | 0.080582 | 0.00537 | 0.001355 |

Table B. fGls Estimator

| | data_1_1 | data_1_2 | data_1_3 | data_1_4 | data_1_5 |
|---|---|---|---|---|---|
| std. err: x | 0.262 | 0.223 | 0.311 | 0.405 | 0.656 |
| std. err: const | 0.523 | 1.618 | 2.681 | 0.505 | 2.326 |
| x | 0.9069 | 1.2843 | -0.5652 | 1.1062 | -0.3203 |
| const | -0.2691 | -0.3952 | 4.1543 | 0.2676 | 2.5368 |
| z:x | 3.457 | 5.754 | -1.818 | 2.729 | -0.498 |
| z:const | -0.514 | -0.244 | 1.549 | 0.53 | 1.091 |
| P:x | 0.001 | 0 | 0.069 | 0.006 | 0.625 |
| P:z | 0.607 | 0.807 | 0.121 | 0.596 | 0.275 |
| R-Squared | 0.185 | 0.398 | 0.118 | 0.127 | 0.043 |
| Adj. R-Squared | 0.176 | 0.391 | 0.1 | 0.108 | 0.023 |
| Durbin-Watson Test | 2.33 | 2.182 | 1.927 | 1.964 | 1.59 |
| White-Test: F Stats. | 9.1722 | 3.335 | 4.98528 | 10.02998 | 16.6 |
| p-value | 0.01019 | 0.1887 | 0.082691 | 0.0066 | 0.0002 |

Table C. WLS Estimator

| | data_1_1 | data_1_2 | data_1_3 | data_1_4 | data_1_5 |
|---|---|---|---|---|---|
| std. err: x | 0.268 | 0.771 | 0.249 | 0.456 | 0.626 |
| std. err: const | 0.297 | 2.267 | 1.615 | 0.529 | 1.137 |
| x | 0.9426 | 1.6344 | -0.3931 | 0.6654 | -1.2645 |
| const | 0.1991 | -0.837 | -0.9092 | 0.849 | -0.375 |
| z:x | 3.52 | 2.119 | -1.578 | 1.458 | -2.02 |
| z:const | 0.671 | -0.369 | -0.563 | 1.606 | -0.33 |
| P:x | 0 | 0.034 | 0.114 | 0.145 | 0.043 |
| P:z | 0.502 | 0.712 | 0.573 | 0.108 | 0.742 |
| R-Squared | 0.12 | 0.138 | 0.052 | 0.083 | 0.146 |
| Adj. R-Squared | 0.111 | 0.129 | 0.032 | 0.064 | 0.129 |
| Durbin-Watson Test | 2.108 | 2.151 | 1.867 | 1.801 | 2.352 |
| White-Test: F Stats. | 9.0223 | 3.7099 | 4.84661 | 10.5956 | 13.456 |
| p-value | 0.01098 | 0.1564 | 0.08863 | 0.005 | 0.0012 |

Table D. Quantile Regression Estimator

| QuantReg | data_1_1 | data_1_2 | data_1_3 | data_1_4 | data_1_5 |
|---|---|---|---|---|---|
| quantile | 0.45 | 0.45 | 0.45 | 0.75 | 0.45 |
| std. err: x | 0.15 | 0.098 | 0.1584 | 0.529 | 0.1837 |
| std. err: const | 0.396 | 0.9443 | 1.7741 | 1.480 | 1.3907 |
| x | 1.0068 | 1.1974 | -0.5644 | 2.2508 | -1.0014 |
| const | -0.5055 | -1.5652 | 0.3427 | 0.5495 | -0.176 |
| t:x | 6.7313 | 12.463 | -3.514 | 4.255 | -5.098 |
| t:const | -1.2843 | -1.6575 | 0.1932 | 0.371 | -0.1266 |
| P:x | 0.000 | 0.000 | 0.008 | 0.000 | 0.000 |
| P:z | 0.2021 | 0.1006 | 0.8476 | 0.712 | 0.897 |
| Pseudo Rsquare | 0.185 | 0.306 | 0.104 | 0.2026 | 0.173 |