Statistical Method Selection Matters: Vanilla Methods in Regression May Yield Misleading Results

Jan Kalina[1]

Abstract: The primary aim of this work is to illustrate the importance of the choice of the appropriate methods for the statistical analysis of economic data. Typically, there exist several alternative versions of common statistical methods for every statistical modeling task and the most habitually used ("vanilla") versions may yield rather misleading results in non-standard situations. Linear regression is considered here as the most fundamental econometric model. First, the analysis of a world tourism dataset is presented, where the number of international arrivals is modeled for 140 countries of the world as a response of 14 pillars (indicators) of the Travel and Tourism Competitiveness Index. Heteroscedasticity is clearly recognized in the dataset. However, the Aitken estimator, which would be the standard remedy in such a situation, is revealed here to be very inappropriate; regression quantiles represent a much more suitable solution here. The second illustration with artificial data reveals standard regression quantiles to be unsuitable for data contaminated by outlying values; their recently proposed robust version turns out to be much more appropriate. Both illustrations reveal that choosing suitable methods represent an important (and often difficult) part of the analysis of economic data.

Keywords: linear regression, assumptions, non-standard situations, robustness, diagnostics.
JEL Classification: C14, C12, C21

1 Introduction

The main purpose of this paper is to illustrate that it is very important to pay attention to choosing the appropriate methods for the analysis of economic data. We believe that users of statistical methods should not rely on cookbooks (strict unambiguous instructions how to proceed in different situations) but rather on common sense and logical thinking (Long and Teetor, 2019). This idea will be illustrated on the linear regression model, which is the most fundamental statistical model in economics. The linear regression is considered throughout the paper in the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

which may be expressed in the matrix notation as $Y = X\beta + e$. Here, $\beta = (\beta_0, \ldots, \beta_p)^T$ is the vector of parameters and $e_1, \ldots, e_n$ are independent identically distributed random errors. One of the assumptions for using the least squares is homoscedasticity, i.e. the situation with the same variance for all the random errors. Formally, we can express homoscedasticity as $var\ e_i = \sigma^2$ for $i = 1, \ldots, n$. The most common estimator of the parameters in (1) is the least squares estimator.

Fitting linear regression is typically accompanied by significance tests, which are based on the assumption of normally distributed errors. Even without this assumption, the least squares estimator of $\boldsymbol{\beta}$ is the best linear unbiased estimator; however, hypothesis tests about $\boldsymbol{\beta}$ do require normal errors. It is known that the least squares estimator may be very unsuitable under the presence of outliers in the data (Jurečková et al., 2019). Outliers represent a commonly appearing problem in the practical analysis of economic data (Kalina et al., 2019). Alternative methods include robust (or nonparametric) estimation and corresponding tests (Saleh et al., 2012).

Modeling trend under heteroscedasticity (i.e. if homoscedasticity is violated) is usually performed using the Aitken model, i.e. considering a specific model for explaining the particular form of heteroscedasticity (Fox, 2019). It is less frequent to use regression quantiles instead (Gneiting et al., 2023). Our aim is to illustrate the possibly harmful effect of using inappropriate statistical methods. Instead of the most habitually used ("vanilla") methods, practitioners should

[1] The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 00 Prague 8, Czech Republic & The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic, kalina@cs.cas.cz

perform method selection, i.e. search for more appropriate methods. Model selection is shown in the paper to represent an important (and often difficult) part of the statistical data analysis.

Section 2 is devoted to heteroscedasticity modeling applied to a world tourism dataset. Although the data are clearly revealed as heteroscedastic, the Aitken estimator (i.e. the standard remedy in such a situation) is revealed here to be very inappropriate. As an alternative, regression quantiles represent a much more suitable solution here. In Section 3, standard regression quantiles are shown to be unsuitable for an artificially created dataset contaminated by outlying values. A robust version of the regression quantiles is much more appropriate here. Section 4 concludes the paper.

2 Heteroscedasticity in the linear regression model

From a general perspective, linear regression modeling should always be accompanied by diagnostic tools (Sasaki and Wang, 2023). These also include a test of a possible heteroscedasticity. Modeling heteroscedasticity will be illustrated in this section on a world tourism dataset. The Aitken estimator is revealed here to be inappropriate for the considered data, while regression quantiles represent a much more suitable solution.

The TTCI (Travel and Tourism Competitiveness Report) dataset previously analyzed in Kalina and Vidnerová (2022) will be analyzed in this section. The study is focused on the relationship between tourism performance and tourism competitiveness. Particularly, we model the number of international arrivals (in millions, denoted as *Y*) as a response of $p = 14$ pillars of TTCI. We work with $n = 140$ countries, for which both the response and the TTCI pillars are available. The regressors are used from Calderwood and Soshkin (2019) and the values of the response originally published by World Bank are taken from Roser (2020). The response *Y* comes from the year 2016; because the response is not available for a few of developing countries for 2016, we had to take their response from the last available year. Our analysis revealed the dataset to contain 5 outliers. However, because the heteroscedasticity tests performed here do not reveal to be much influenced by the presence of outliers, we omit a more detailed discussion of robustness issues. Standard *t*-tests reveal pillar 14 (Cultural Resources and Business Travel), pillar 11 (Ground and Port Infrastructure), and pillar 12 (Tourist Service Infrastructure) to be the most relevant predictors for explaining *Y*.

Visual criteria are the simplest tools for detecting heteroscedasticity, which is clearly detected in Figure 1. More formally, we perform two standard heteroscedasticity tests, namely the Breusch-Pagan test and the White test (White, 1980). These tests attempt to explain the heteroscedasticity by fitting residuals of (1) against selected (or all) regressors (Li and Yao, 2019). The results are given in Table 1 and the heteroscedasticity is detected here to be very strong. The *p*-values are given only with 4 decimal points for the White test in the package het.test of R software. Asymptotic tests for the robust least weighted squares (LWS) estimator of Víšek (2011) are also presented in Table 1.

In general, if heteroscedasticity is confirmed in a given regression model, it may be recommended to estimate the regression parameters by means of Aitken estimator, which is commonly denoted as the generalized least squares estimator. Let us consider the Aitken model in the form

$$\frac{Y_i}{\sqrt{k_i}} = \frac{\gamma_0}{\sqrt{k_i}} + \gamma_1 \frac{X_{i1}}{\sqrt{k_i}} + \cdots + \gamma_p \frac{X_{ip}}{\sqrt{k_i}} + e_i^*, \qquad i = 1, \ldots, n, \tag{2}$$

with parameters $\gamma_0, \ldots, \gamma_p$ as in formula (9.11) of Greene (2012). The random errors in (2) are denoted as $e_1^*, \ldots, e_n^*$ to stress that they are different from the errors $e_1, \ldots, e_n$ in the original model (1). Here, we consider the general form (2) with a specific choice of the constants

$$\sqrt{k_i} = \hat{Y}_i, \qquad i = 1, \ldots, n, \tag{3}$$

where $\hat{Y}_i$ is the estimated value of Y_i obtained by the least squares in the original model (1). The model (2) does not contain an intercept and the parameters are estimated by ordinary least squares. The choice (3) corresponds to the assumption that $var\ e_i = \left(\sigma \hat{Y}_i\right)^2$, which seems reasonable for the given data. The construction (3) is then motivated by the fact that it ensures a homoscedastic model, particularly

$$var\ e_i^* = var\ \frac{e_i}{\hat{Y}_i} = \left(\frac{1}{\hat{Y}_i}\right)^2 \cdot \left(\sigma \hat{Y}_i\right)^2 = \sigma^2. \tag{4}$$

Table 1 TTCI dataset: results of heteroscedasticity tests.

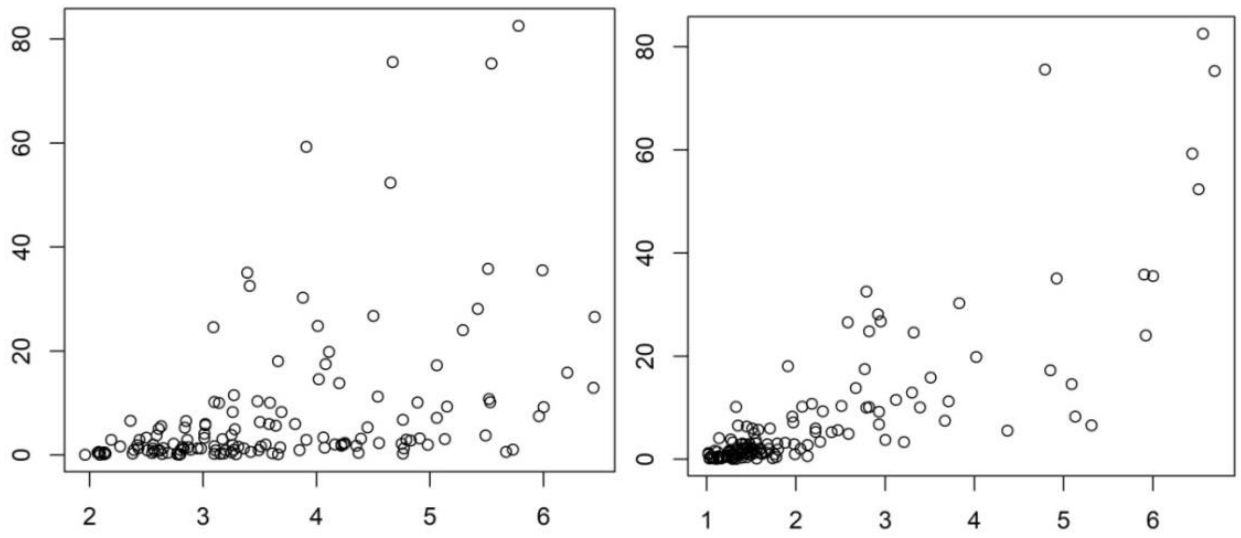
Heteroscedasticity test	***p*-value for LS**	**Asymptotic *p*-value for LWS**
Breusch-Pagan on 14 variables	$4 \cdot 10^{-6}$	$3 \cdot 10^{-5}$
Breusch-Pagan on the set of regressors {14, 11, 12}	$7 \cdot 10^{-10}$	$2 \cdot 10^{-8}$
White test	0.0000	0.0000

Source: Own processing

The White test performed in the Aitken model (2) in the given particular form yields $p=0.96$. Thus, it may seem that the analysis of the considered dataset should exploit the Aitken model, which is particularly popular in tourism modeling applications (Assaf and Tsionas, 2020). However, the graphical visualization in Figure 2 reveals that there is basically no trend of the response against the regressors in the Aitken model. This remains true even if potential outliers are trimmed away. Therefore, we conclude that the auxiliary regression model (2) is not suitable and the Aitken estimator, which would represent a standard ("vanilla") method, cannot be recommended.

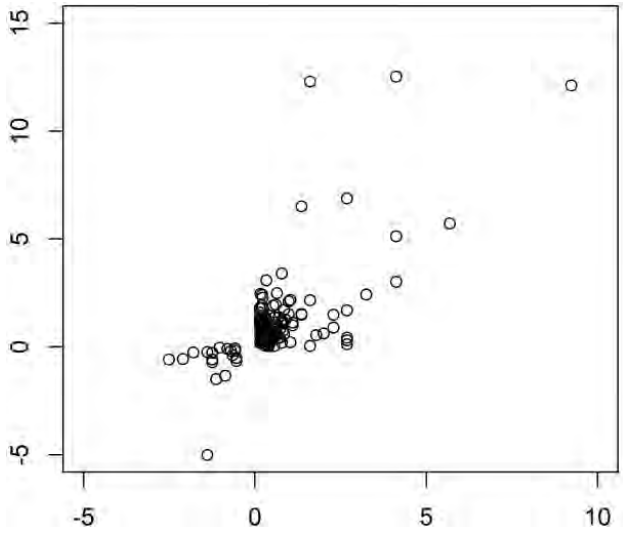
Econometricians acknowledge heteroscedasticity as a serious issue to be solved in modeling economic data, but it still remains unclear how to practically proceed with modeling particular heteroscedastic data. Regression quantiles represent a promising methodology especially for heteroscedastic data. For the tourism dataset, Figure 3 shows the relationship of Y on pillar 5, i.e. Information and Communication Technologies Readiness. In this simple situation with a single regressor, the plots show that LWS-quantiles do not much differ from standard regression quantiles.

Figure 1 TTCI dataset: the plots of the response Y against pillar 11 (left) and against pillar 14 (right) reveal a strong heteroscedasticity in the model.



Source: Own processing

Figure 2 The Aitken model (2) computed for the TTCI data, shown for the relationship of Y (vertical axis) on the regressor 14 (the most relevant regressor found by t-tests computed for the least squares estimator).



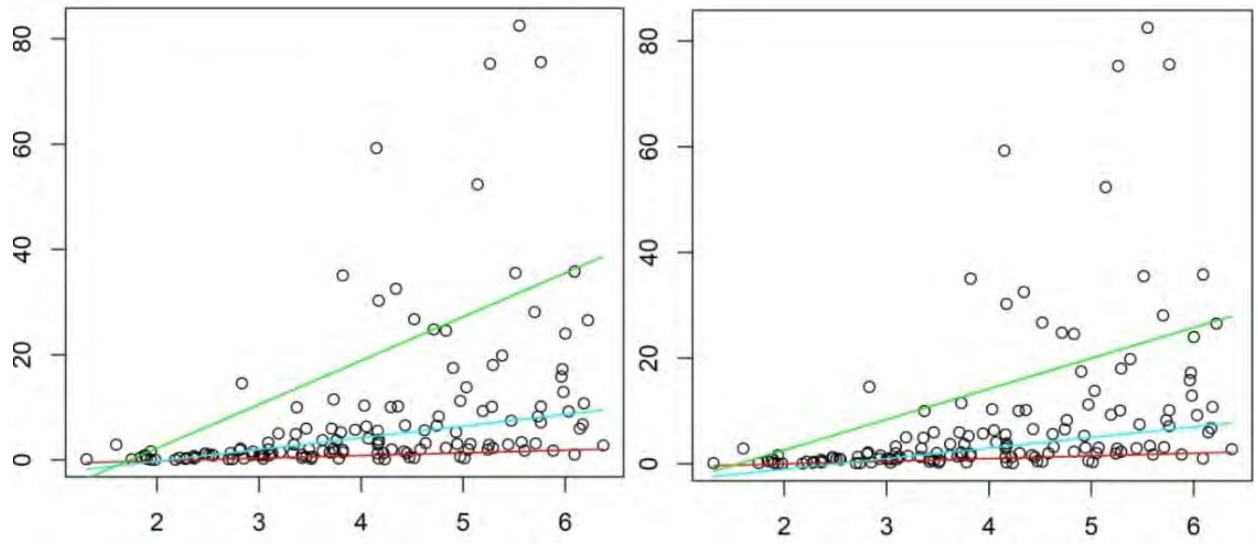
Source: Own processing

Table 2 The artificial dataset of Figure 1: estimated values of the parameters.

Value of τ	Standard regression quantiles		Robust regression quantiles	
	Intercept	Slope	Intercept	Slope
0.1	0.97	–0.40	1.32	–0.46
0.5	1.63	–0.40	2.09	–0.48
0.9	1.56	0.17	2.94	–0.53

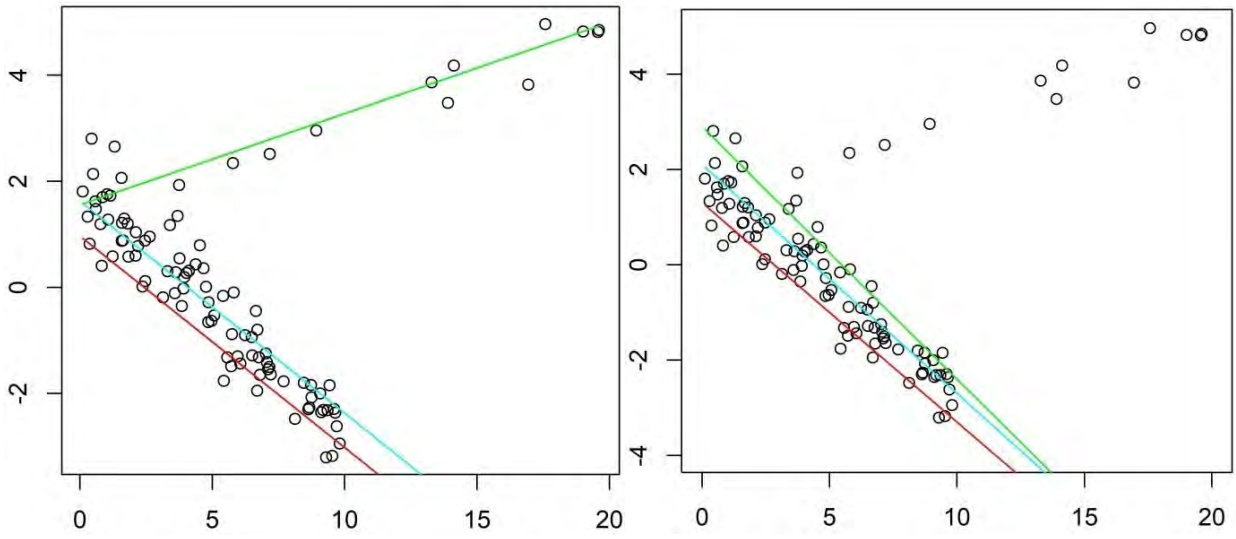
Source: Own processing

Figure 3 TTCI dataset: two different approaches to estimating regression quantiles for the model with the response variable fitted against pillar 5. Left: regression quantiles. Right: LWS-quantiles.



Source: Own processing

Figure 4 The artificial dataset of Section 3: two different approaches to estimating regression quantiles. Left: standard regression quantiles. Right: LWS-quantiles.



Source: Own processing

3 Standard versus robust regression quantiles

Regression quantiles represent a popular methodology for modeling economic data especially under heteroscedasticity (Koenker, 2017). A regression τ-quantile represents a generalization of the concept of a τ-quantile (i.e. quantile at τ) to the linear regression context and is an estimate of the conditional τ-quantile of the response given the values of the regressors. It may be beneficial to replace using a single regression hyperplane by a whole set of several hyperplanes corresponding to regression quantiles with different levels of τ. The concept of regression quantiles is very useful as acknowledged e.g. in the econometric textbook by Greene (2018) and hypothesis tests are available for the significance of regression quantiles (Liu et al., 2023). The (highly robust) LWS-quantiles may be very different from (standard) regression quantiles and the aim of this section is to illustrate this on an artificial dataset.

Still, it is important to consider their robust version instead of the much more popular standard (non-robust) regression quantiles. LWS-quantiles as highly robust regression quantiles were as tools directly inspired by the LWS estimator. They were proposed in Kalina and Vidnerová (2022), where they were at the same time illustrated on various real data. The idea of LWS-quantiles is to assign weights to individual observations so that outlying data obtain small (possibly zero) weights.

Regression quantiles are very vulnerable with respect to outliers in the data, which will be illustrated on an artificial dataset shown in Figure 4. For this dataset with a single regressor, standard regression quantiles and LWS-quantiles were computed for values of τ equal to 0.1, 0.5, and 0.9. The regression lines are shown in Figure 4 and the estimates values of the intercept and slope are given in Table 2. The results of standard regression quantiles are misleading due to the contamination of the data by a small percentage of observations that actually follow a different linear trend than the majority of data points. This breakdown of the top regression quantile is shown very clearly in Figure 4 (left). The figure on the right is apparently much more appropriate compared to the left one. Because the LWS-quantiles possess a high robustness to extreme outliers, we may consider their results to be reliable.

4 Conclusions

Linear regression represents the most commonly used statistical model in econometric practice. In general, it is recommended to check the assumptions whenever a linear regression model is fitted. Such checking requires to use one of the available diagnostic tools (Fox, 2019). Standard significance tests about regression parameters require to assume normality of the random errors and homoscedasticity. This paper is aimed to persuade the readers that standard methods commonly used for regression modeling may be misleading for a particular dataset.

There is a widely spread misunderstanding (or perhaps desinformation) among non-statisticians that standard methods for economic data analysis are sufficient in all situations. The task to find an appropriate statistical method for a particular task is non-trivial also for quite simple (low-dimensional) data and is even more intricate for high-dimensional data (Kalina and Rensová, 2015). For data contaminated by outliers, we may recommend to use a (possibly highly) robust method. In this paper, new arguments in favor of LWS-quantiles are illustrated in the numerical experiments. Another possible complication of regression modeling is heteroscedasticity; the presented experiments show a situation with

regression quantiles being much more suitable compared to the Aitken estimator. Practitioners may appreciate that the non-standard methods used here (robust regression, robust regression quantiles) allow a clear interpretation of the results.

Some additional important topics not mentioned in this paper include a need for proposing and investigating new statistical methods suitable for data with outliers as well as with heteroscedasticity. Particularly, there remains a need for a further study of diagnostic tools for robust regression beyond the results obtained within our research project on modern nonparametric methods in econometrics. Other new methodology still to be proposed includes a lasso version of LWS-quantiles or a heteroscedastic version of multivariate quantiles of Hlubinka and Šiman (2013).

Acknowledgement

The research was supported by the Czech Science Foundation project 21-05325S ("Modern nonparametric methods in econometrics"). The author would like to thank Michal Jelínek for helpful suggestions.

References

Assaf, A.G., & Tsionas, M. (2020). Correcting for Endogeneity in Hospitality and Tourism Research. *International Journal of Contemporary Hospitality Management,* 32, 2657-2675.

Calderwood, L.U., & Soshkin, M. (2019). The Travel & Tourism Competitiveness Report 2019: Travel and Tourism at a Tipping Point. World Economic Forum, Geneva, https://www3.weforum.org/docs/WEF_TTCR_2019.pdf.

Fox, J. (2019). *Regression Diagnostics: An Introduction*. 2. ed. Thousand Oaks: SAGE Publications.

Gneiting, T., Wolffram, D., Resin, J., Kraus, K., Bracher, J., et al. (2023). Model Diagnostics and Forecast Evaluation for Quantiles. *Annual Review of Statistics and Its Application,* 10, 597-621.

Greene, W.H. (2012). *Econometric Analysis*. 7. ed. New York: Pearson.

Hlubinka, D., & Šiman M. (2013). On Elliptical Quantiles in the Quantile Regression Setup. *Journal of Multivariate Analysis,* 116, 163-171.

Jurečková, J., Picek, J., & Schindler, M. (2019). *Robust Statistical Methods with R*. 2. ed. Boca Raton: CRC Press.

Kalina, J., & Rensová, D. (2015). How to Reduce Dimensionality of Data: Robustness Point of View. *Serbian Journal of Management,* 10, 131-140.

Kalina, J., & Tichavský, J. (2020). On Robust Estimation of Error Variance in (Highly) Robust Regression. *Measurement Science Review,* 20, 6-14.

Kalina, J., Vašaničová, P., & Litavcová, P. (2019). Regression Quantiles under Heteroscedasticity and Multi-collinearity: Analysis of Travel and Tourism Competitiveness. *Ekonomický časopis/Journal of Economics,* 67, 69-85.

Kalina, J., & Vidnerová P. (2022). Least Weighted Squares Quantiles Reveal how Competitiveness Contributes to Tourism Performance. *Czech Journal of Economics and Finance,* 72, 150-171.

Koenker, R. (2017). Quantile Regression: 40 Years on. *Annual Review of Economics,* 9, 155-176.

Li, Z., & Yao, J. (2019). Testing for Heteroscedasticity in High-Dimensional Regressions. *Econometrics and Statistics,* 9, 122-139.

Liu, X., Long, W., Peng, L., & Yang, B. (2023). A Unified Inference for Predictive Quantile Regression. *Journal of the American Statistical Association*. In press.

Long, J.D., & Teetor, P. (2019). *R Cookbook. Proven Reciples for Data Analysis, Statistics & Graphics*. 2. ed. Sebastopol: O'Reilly.

Roser, M. (2020). https://ourworldindata.org/tourism.

Saleh, A., Picek, J., & Kalina, J. (2012). R-estimation of the Parameters of a Multiple Regression Model with Measurement Errors. *Metrika*, 75, 311-328.

Sasaki, Y., & Wang, Y. (2023). Diagnostic Testing of Finite Moment Conditions for the Consistency and Root-n Asymptotic Normality of the GMM and M Estimators. *Journal of Business & Economic Statistics,* 41, 339-348.

Víšek, J.Á. (2011). Consistency of the Least Weighted Squares under Heteroscedasticity. *Kybernetika,* 47, 179-206.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica,* 48, 817-838.