Influential Observations in Stochastic Frontier Analysis

Arne Henningsen

March 25, 2020

1 Introduction

In statistical data analysis, it if often relevant to indentify influential observations, i.e., individual observations that have a noticeable effect on the results of the analysis. The vast majority of theoretical and empirical literature about influential observations is based linear regression with the ordinary least squares (OLS) method. However, it is also relevant to identify influential observations in other types of regression analysis, for instance in stochastic frontier analysis (Aigner et al., 1977; Meeusen and van den Broeck, 1977). Identifying influential observations is particularly important when the results of the statistical analysis can have substantial effects on the 'real world', e.g., if the results of a stochastic frontier analysis are used for regulating prices and revenues in natural monopolies, which is done in several countries.

2 Cook's distance in linear regression models

The most frequently used method for detecting influential observations is the so-called Cook's Distance, which was introduced by Cook (1977, 1979). Cook's Distance is based on a linear regression model:

$$y_i = \beta' x_i + \epsilon_i \quad \forall\, i = 1, \ldots, N, \tag{1}$$

where y_i is the the dependent variable, x_i is a vector of k explanatory variables (including a constant), β is a vector of k regression coefficients, ϵ_i is an error term, subscript i indicates the observation, and N is the number of observations used in the analysis.

The Cook's Distance of the ith observation is defined as:

$$D_i = \frac{\sum_{j=1}^{N} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{k\; \hat{\sigma}^2}, \tag{2}$$

where $\hat{y}_j \equiv \hat{\beta}' x_j$ is the predicted value of the dependent variable at the jth observation based on the estimated coefficients $\hat{\beta}$ that are obtained from a regression that includes all N observations, $\hat{y}_{j(i)} \equiv \hat{\beta}'_{(i)} x_j$ is the predicted value of the dependent variable at the jth observation based on the estimated coefficients $\hat{\beta}_{(i)}$ that are obtained from a regression that excludes the ith observation (and, thus, only includes $N-1$ observations), k is the number of explanatory variables (including a constant), and $\hat{\sigma}^2 = \widehat{\sum_{i=1}^N \hat{\epsilon}_i^2}/(N-k)$ is the estimated variance of the error term ϵ_i with $\hat{\epsilon}_i = y_i - \hat{\beta}' x_i$ being the residual of the regression analysis that includes all N observations.

There is no consensus on a cut-off point that indicates whether the influence of an observation is “large” not not. Several different cut-off points have been suggested, e.g., 1, 0.5, $4/N$, and the 50% quantile of an F-distribution with p and $N-p$ degrees of freedom (see, e.g., Cook and Weisberg, 1982; Bollen and Jackman, 1990).

Besides having no clear cut-off point, Cook’s Distance has been criticised for that under certain circumstances, an observation can have a substantial influence on the estimated coefficients but only a minor influence on the predicted values of the dependent variable so that this observation has a low Cook’s Distance in spite of its large influence on the estimated coefficients (Kim, 2017). As in most empirical applications, analysts are more interested in the coefficients than in the predicted values of the dependent variable, this potential weakness of Cook’s distance can be a major problem.

3 Pseudo-Cook’s distance in stochastic frontier analysis

A stochastic frontier model is generally specified as:

$$y_i = \beta' x_i - u_i + v_i \quad \forall\, i = 1, \ldots, N, \tag{3}$$

where y_i is the logarithmic output quantity, x_i is a vector of k explanatory variables (e.g., a constant, logarithmic input quantities, $\ldots$), β is a vector of k regression coefficients, u_i is a non-negative inefficiency term, v_i is a random error term with an expected value of zero that captures statistical noise, subscript i indicates the firm, and N is the number of firms used in the analysis. Most empirical analyses assume that the non-negative inefficiency term u_i follows a half-normal, truncated-normal, or exponential distribution, while almost all empirical analyses assume that the random error term v_i follows a normal distribution. This specification is usually estimated by the maximum-likelihood method based on the distributional assumptions of the inefficiency term u_i and the statistical noise term v_i.

Strictly speaking, Cook’s Distance is not applicable to stochastic frontier analysis because it is based on the assumption that the error term ϵ_i is normally distributed, while stochastic frontier analysis assumes a composed error term $\epsilon_i = -u_i + v_i$ that is not normally distributed unless there is no inefficiency, i.e., $u_i = 0 \;\forall\; i = 1, \ldots, N$ (Wheat et al., 2019, p. 23). However, the concept of Cook’s distance can be applied to stochastic frontier analysis and a pseudo-Cook’s distance can be calculated by equation (2). When this equation is applied to stochastic frontier analysis, $\hat{y}_j \equiv \hat{\beta}' x_j$ and $\hat{y}_{j(i)} \equiv \hat{\beta}'_{(i)} x_j$ are the predicted *frontier* values of the logarithmic output quantity of firm j based on the maximum-likelihood estimates of the coefficients based on all N firms and based on all firms except for firm i, respectively. As in the original Cook’s distance measure, k is the number of explanatory variables. The choice of $\hat{\sigma}^2$ is the only tricky part of applying Cook’s distance to stochastic frontier models. One could choose the variance of the random error term v_i because it indicates to which extent the frontier values of individual firms fluctuate around the ‘true’ population frontier function $\beta' x_i$ (similar to $\hat{\sigma}^2$ in linear regression models that indicates to which extent the observed values of the dependent variable fluctuate around the true population regression line $\beta' x_i$). However, the estimated variance of the random error term v_i can become zero so that the pseudo-Cook's distance based on this variance would be undefined. Furthermore, removing a firm from the sample could substantially change the maximum likelihood estimate of the division of the total error variance between statistical noise (v_i) and inefficiency (u_i) so that the estimated variance of the random error term v_i is much less stable than the estimated variance of the error term in linear regression models. Therefore, I suggest to define $\hat{\sigma}^2$ as the estimated variance of the composed error term $\epsilon_i = -u_i + v_i$, i.e., $\hat{\sigma}^2 = \sum_{i=1}^N \left(\hat{\epsilon}_i - \bar{\hat{\epsilon}}\right)^2 / (N-k)$ with $\hat{\epsilon}_i = y_i - \hat{\beta}' x_i$ being the residual of the regression analysis that includes all N observations and $\bar{\hat{\epsilon}} \equiv N^{-1} \sum_{i=1}^N \hat{\epsilon}_i$.

Given that the influence of an observation is defined in relative terms, i.e., by comparing it to the influence of the other observations (Belsley et al., 1980, p. 11) and that there is no consensus on an absolute cut-off value of the Cook's distance for indicating a large influence, the choice of a value for $\hat{\sigma}^2$ is only of minor importance, because the value of $\hat{\sigma}^2$ affects the (pseudo-)Cook's distances of all observations in a proportional way so that the choice of $\hat{\sigma}^2$ does not affect the (pseudo-)Cook's distance of one observation relative to the (pseudo-)Cook's distances of all observations.

If one is mainly interested in the estimated coefficients ($\hat{\beta}$) of a stochastic frontier model, the critique of Kim (2017) also applies to the pseudo-Cook's distance of stochastic frontier models. However, if the primary interest of a stochastic frontier analysis is to obtain the frontier, e.g., for regulating natural monopolies, it is an advantage rather than a potential weakness that the pseudo-Cook's distance (as the Cook's distance in linear regression analysis) evaluates the influence on the predicted values rather than on the estimated coefficients.

4 Pseudo-Cook's distance for efficiency estimates

In many stochastic frontier analyses, the primary interest are not the frontier values but the efficiency estimates, e.g., when the efficiency estimates are used to regulate prices and revenues in natural monopolies. In these cases, it could be relevant to assess the influence of each firm in the data set on the efficiency estimates, e.g., by a modified pseudo-Cook's distance measure:

$$D_i^{eff} = \frac{\sum_{j=1}^N \left(\widehat{eff}_j - \widehat{eff}_{j(i)}\right)^2}{k \; \hat{\sigma}^2_{eff}}, \tag{4}$$

where $\widehat{eff}_j$ is the efficiency estimate of the jth firm derived from a stochastic frontier analysis that is based on all N observations, $\widehat{eff}_{j(i)}$ is the efficiency estimate of the jth firm derived from a stochastic frontier analysis that excludes the ith observation (and, thus, only includes $N-1$ observations), k is the number of explanatory variables (including a constant), and $\hat{\sigma}^2_{eff} = \sum_{i=1}^N \left(\widehat{eff}_i - \overline{\widehat{eff}}\right)^2 / (N-1)$ is the estimated variance of the efficiency estimates with $\overline{\widehat{eff}} = N^{-1} \sum_{i=1}^N \widehat{eff}_i$ being the average efficiency estimate (based on all N observations).

For given variances of u_i and v_i, the influence of an observation on the predicted values and on the efficiency estimates is very similar. However, the efficiency estimates are very sensitive to the estimated variances of u_i and v_i. Hence, an observation that has a notable influence on the estimated variances of u_i and v_i usually has a notable influence on the efficiency estimates while it does not necessarily have a notable influence on the predicted values.

5 Calculating pseudo-Cook's distances with the "R" package "frontier"

The following command loads the "frontier" package:

```
library( "frontier" )
```

The following command loads a data set:

```
data( "front41Data" )
```

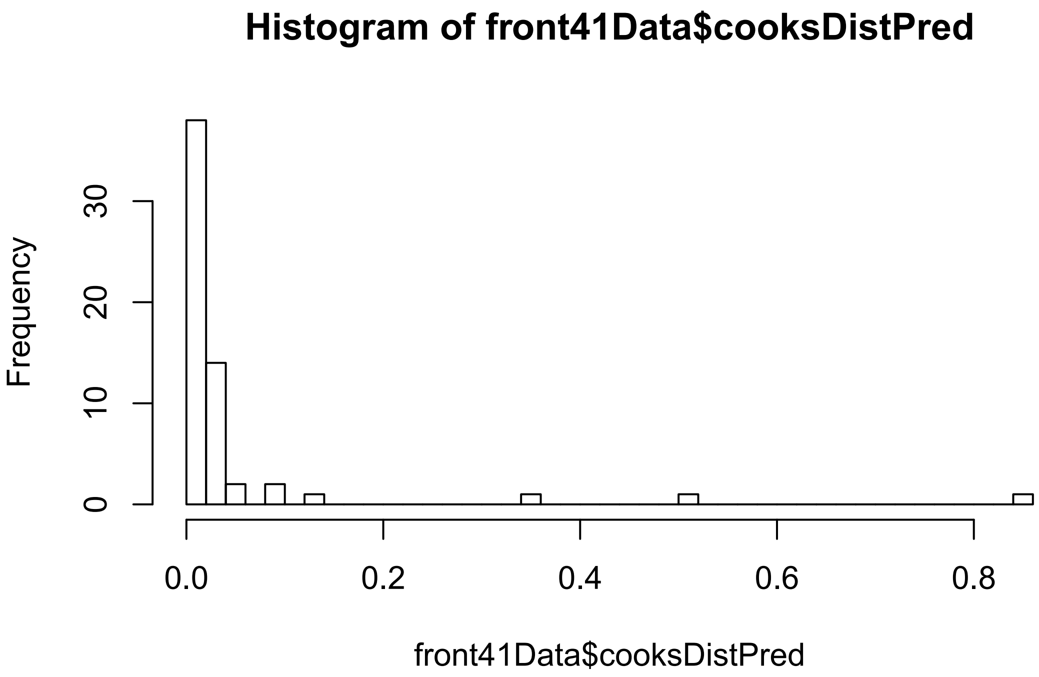
The following commands estimate a Cobb-Douglas production frontier and display the estimation results:

```
cobbDouglas <- sfa( log( output ) ~ log( capital ) + log( labour ),
  data = front41Data )
summary( cobbDouglas )

## Error Components Frontier (see Battese & Coelli 1992)
## Inefficiency decreases the endogenous variable (as in a production function)
## The dependent variable is logged
## Iterative ML estimation terminated after 7 iterations:
## log likelihood values and parameters of two successive iterations
## are within the tolerance limit
##
## final maximum likelihood estimates
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  0.561619   0.202617  2.7718 0.0055742 **
## log(capital) 0.281102   0.047643  5.9001 3.632e-09 ***
## log(labour)  0.536480   0.045252 11.8555 < 2.2e-16 ***
## sigmaSq      0.217000   0.063909  3.3955 0.0006851 ***
## gamma        0.797207   0.136424  5.8436 5.109e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## log likelihood value: -17.02722
##
## cross-sectional data
## total number of observations = 60
##
## mean efficiency: 0.7405678
```

The following commands obtain the pseudo-Cook's distances for the predicted values, visualize them with a histogram, and identify the three observations with the largest pseudo-Cook's distance:

```
front41Data$cooksDistPred <- cooks.distance( cobbDouglas, asInData = TRUE,
  progressBar = FALSE )
hist( front41Data$cooksDistPred, 50 )
```

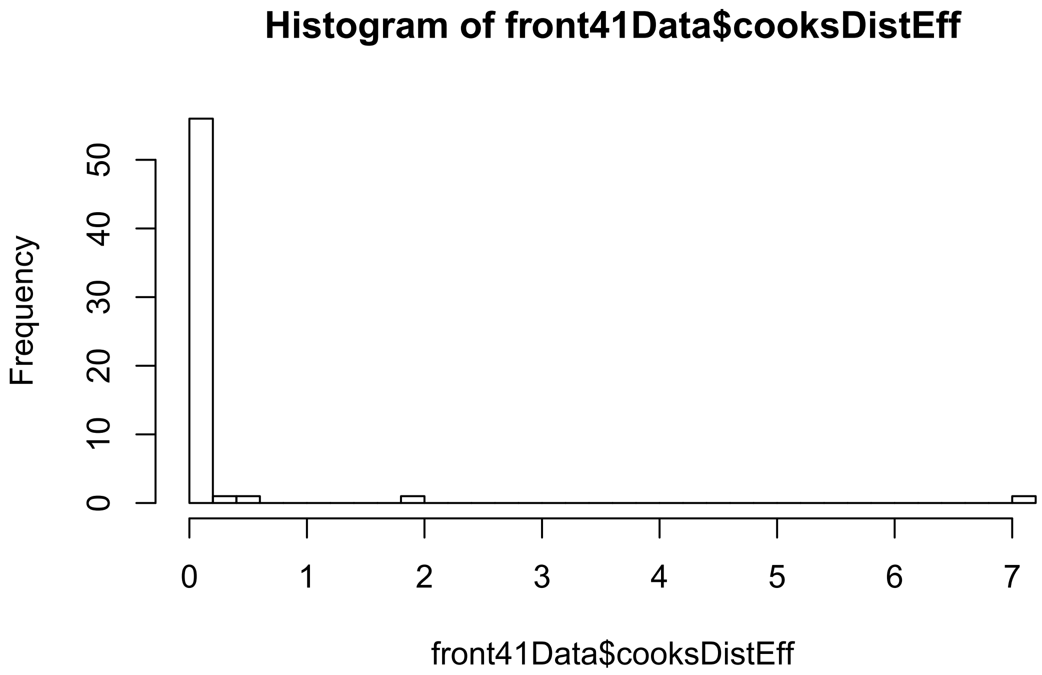


```
front41Data[ front41Data$cooksDistPred > 0.3, c( "firm", "cooksDistPred" ) ]

##    firm cooksDistPred
## 12   12     0.8579408
## 35   35     0.3528187
## 57   57     0.5049758
```

The following commands obtain the pseudo-Cook's distances for the efficiency estimates, visualize them with a histogram, and identify the three observations with the largest pseudo-Cook's distance:

```
front41Data$cooksDistEff <- cooks.distance( cobbDouglas,
  target = "efficiencies", asInData = TRUE, progressBar = FALSE  )
hist( front41Data$cooksDistEff, 50 )
```

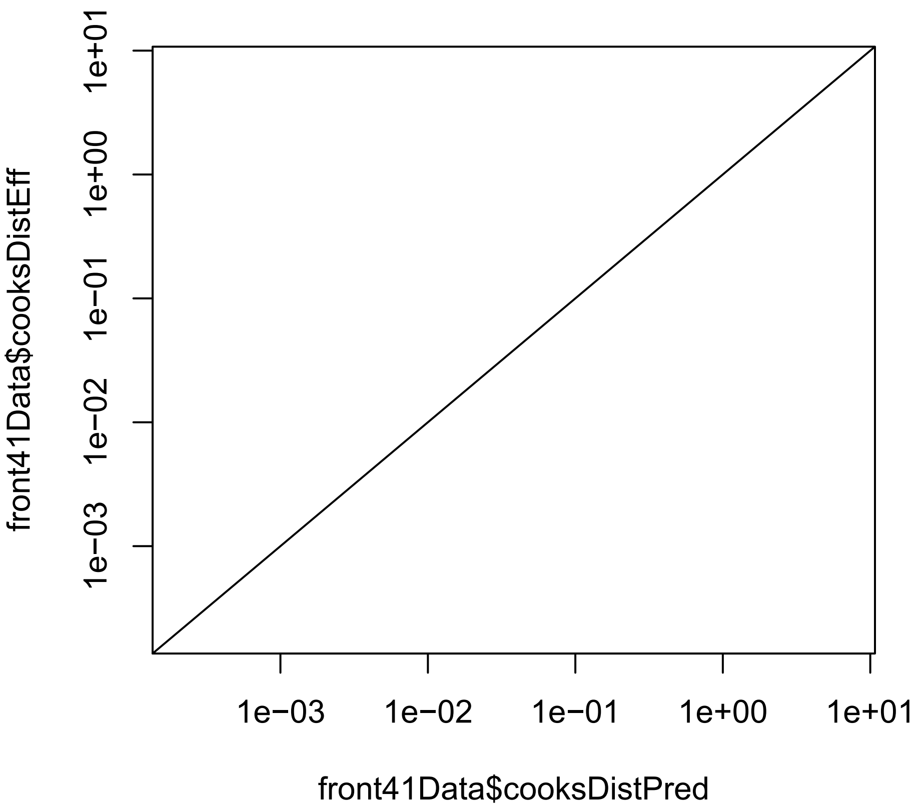


```
front41Data[ front41Data$cooksDistEff > 0.3, c( "firm", "cooksDistEff" ) ]

##    firm cooksDistEff
## 12   12    7.0858707
## 35   35    1.8668939
## 57   57    0.4409058
```

The following commands visualize the relationship between the pseudo-Cooks distances for the predicted values and the pseudo-Cooks distances for the efficiency estimates (on a logarithmic scale):

```
library( "miscTools" )
compPlot( front41Data$cooksDistPred, front41Data$cooksDistEff, log = "xy" )
```



The pseudo-Cooks distances for the predicted values and the pseudo-Cooks distances for the efficiency estimates are highly correlated but the pseudo-Cooks distances for the efficiency estimates are generally slightly larger than the pseudo-Cooks distances for the predicted values.

6 Discussion

It is important to note that observations that are identified to be influential should not be automatically excluded from the analysis, because the influence of the observation on the regression results is not necessarily 'bad' (i.e., diverting the regression results away from the 'true' values) but can also be 'good' (i.e., directing the regression results towards the 'true' values). Therefore, observations that are identified to be influential should be carefully checked for data errors and potential unobserved heterogeneties between these observations and the other observations. Only if data errors are found and cannot be corrected or if unobserved heterogeneities are found and cannot be addressed (e.g., by adding further explanatory variables), influential observations should be excluded from the analysis.

References

Aigner, D., Lovell, C. A. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6:21–37.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics - Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Bollen, K. A. and Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In Fox, J. and Long, J. S., editors, *Modern Methods of Data Analysis*. Sage, Newbury Park, CA, USA.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Kim, M. G. (2017). A cautionary note on the use of Cook's distance. *Communications for Statistical Applications and Methods*, 24(3):317–324.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2):435–444.

Wheat, P., Stead, A. D., and Greene, W. H. (2019). Robust stochastic frontier analysis: A Student's t-half normal model with application to highway maintenance costs in England. *Journal of Productivity Analysis*, 51(1):21–38.