Don't Use Heteroscedastic Corrected Covariance Estimates When Predicting or Testing for Outliers

John Fox and Sanford Weisberg

October 12, 2022

Abstract

When an assumption of constant residual variance in a linear regression model is in doubt, a heteroscedastic adjusted covariance matrix (sandwich estimator) is commonly used to adjust for possibly failed assumption of constant variance. We show that this adjustment is inappropriate when testing for outliers.

Suppose we fit a linear model for which we have a full rank $n \times p$ matrix $\mathbf{X}$ of predictors (Weisberg, 2014), and $n \times 1$ vector $\mathbf{Y}$ of corresponding responses, and assume

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{1}$$

$$\mathrm{var}(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I} \tag{2}$$

Standard methodology is to fit via ordinary least squares, ols, which produces the ols estimates

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{3}$$

$$\hat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \tag{4}$$

$$\hat{\sigma}^2 = (\mathbf{Y} = \hat{\mathbf{Y}})'(\mathbf{Y} = \hat{\mathbf{Y}})/(n-p) \tag{5}$$

Inferences concerning $\boldsymbol{\beta}$ and other aspects of the regression problem are generally based on normality (asymptotic, approximate or exact, depending on assumptions) of $\widehat{\boldsymbol{\beta}}$.

Suppose the variance given by (2) is misspecified, and the true population covariance matrix is given by $\sigma^2\mathbf{W}$ for some diagonal matrix $\mathbf{W}$ of unspecified positive numbers. It is easy to show that the variance of the least squares estimate is then (Weisberg, 2014, Sec. 7.2.1)

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\mathbf{W}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \tag{6}$$

of an interesting “sandwich" form. The work of White (1980) and others suggested estimating $\mathbf{W}$ by a diagonal matrix with entries $\hat{w}_i = k_i e_i^2$, where the k_i are a set of known constants and the e_i^2 are the squared ols residuals. Then

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \tag{7}$$

could be used in place of (5) in inference about coefficient estimates. Several choices of k_i have been proposed (Long and Ervin, 2000), including setting $k_i = n/(n-p)$, often called HC1, or setting $k_i = 1/(1-h_i)^2$, where h_i is the leverage (Fox and Weisberg, 2019, Sec. 8.3.2) for the ith observation, often called HC3. HC1 is available in Stata linear regression routines simply by adding a keyword to the call to the regression procedure. Since using the sandwich estimator loses very little efficiency in the homoscedastic case, this choice may be used as a matter of course. In R, there are at least two functions that compute these sandwich estimators, `sandwich::vcovHC` and `car::hccm`. Both of these functions use HC3 by default but permit use of other choices as options.

While the use of the corrected covariance matrices is benign for most aspects of inference in regression, there are two related areas where this method can lead to incorrect inferences: testing for outliers, and predicting the response for new units.

For example, suppose we have a population of n stores whose sales receipts are a mixture of cash sales and credit card sales. Suppose it is suspected that in m particular stores cash sales are under-reported. We can model the response reported cash sales, or a transformation of it, with regressors derived from relevant predictors, such as reported credit card sales and demographic and income data on the customers served by each store. Imagine adding m additional regressors to the regression model (1) that are indicators for the m focal stores at issue. Assuming the data on these stores are the last m rows of data, the revised model is

$$\begin{aligned}\mathrm{E}(\mathbf{Y}|(\mathbf{X},\mathbf{Z}) &= (\mathbf{X},\mathbf{Z})\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{X}_2 & \mathbf{I}_m\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{pmatrix} \end{aligned}\tag{8}$$

Both $\mathbf{X}$ and $\mathbf{Z}$ have been partitioned into the first $n-m$ and last m rows; partition $\mathbf{Y}' = (\mathbf{Y}_1', \mathbf{Y}_2')$ similarly. This fits one parameter for each of the m suspect stores. The coefficients $\boldsymbol{\gamma}$ correspond to the mean-shift outlier model (Cook and Weisberg, 1982, Sec. 2.2.2), and provide the basis for testing if stores require an additional parameter to model their mean. Any problem with m cases each with leverage equal to 1 can be shown to be equivalent to this example by a reparameterization.

The ols estimator for the expanded model is

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{\gamma}}\end{pmatrix} = [(\mathbf{X},\mathbf{Z})'(\mathbf{X},\mathbf{Z})]^{-1}(\mathbf{X},\mathbf{Z})'\mathbf{Y}$$

Substituting the partitioned forms into this last equation,

$$[(\mathbf{X},\mathbf{Z})'(\mathbf{X},\mathbf{Z})]^{-1} = \begin{pmatrix}\mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2'\\ \mathbf{X}_2 & \mathbf{I}_m\end{pmatrix}^{-1}$$

From (Schott, 1997, Theorem 7.1),

$$((\mathbf{X},\mathbf{Z})'(\mathbf{X},\mathbf{Z}))^{-1} = \begin{pmatrix} (\mathbf{X}_1\mathbf{X}_1)^{-1} & -(\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_2' \\ -\mathbf{X}_2(\mathbf{X}_1\mathbf{X}_1)^{-1} & \mathbf{I}_m + \mathbf{X}_2'(\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_2 \end{pmatrix} \tag{9}$$

and the ols estimator is

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1 \\ \mathbf{Y}_2 - \mathbf{X}_2\widehat{\boldsymbol{\beta}} \end{pmatrix} \tag{10}$$

We recognize that $\hat{\boldsymbol{\gamma}}$ is equal the observed values of the response minus the prediction for them from the non-focal cases, sometimes called predicted residuals. Individual tests concerning each of the m focal cases are obtained using t tests with estimated $\hat{\gamma}$ divided by $\hat{\sigma}$ times the square root of the corresponding diagonal entry in (9). Unlike the standard errors for the $\hat{\beta}$s, the standard errors for the γs have an term that does not decrease with sample size.

1 Can We Use Hetroscadasticity with Outlier Testing?

Consider using a hetroscedasticity corrected standard error. For the model with the potential outliers, the matrix $\widehat{\mathbf{W}}$ is given by

$$\widehat{\mathbf{W}} = \mathrm{diag}([k_i(\mathbf{Y}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}})^2]', \mathbf{0}') \tag{11}$$

The last m elements correspond to the focal cases and are exact zeros because observed and fitted values are identical. Letting $\widehat{\mathbf{W}}_1$ be the first $n-m$ rows and columns of $\widehat{\mathbf{W}}$,

$$\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\widehat{\mathbf{W}}_1\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{12}$$

This matrix is of rank $n-m$, as the last m rows and columns are all zeros, and so the rank of the sandwich estimator is no greater than $n-m$ (Gruber, 2014, Theorem 6.2) and it is not consistent as an estimate of the covariance of the regression coefficients. Thus the sandwich estimator should not be used in this setting.

It is instructive to write out the sandwich estimator.

$$\begin{aligned} \widehat{\mathrm{var}}\left(\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} |\mathbf{X}\right) &= \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} \\ &= \hat{\sigma}^2 \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}[\mathbf{X}'\widehat{\mathbf{W}}_1\mathbf{X}_1](\mathbf{X}_1'\mathbf{X}_1)^{-1} & (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \\ \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \end{pmatrix} \end{aligned} \tag{13}$$

1. As previously shown, the sandwich estimator has rank of at most $n-m$.

2. The sandwich estimator for $\boldsymbol{\beta}$, the upper left element of (13) is correct, in the sense that it is the same estimate that would have been obtained by fitting only to the first $n-m$ cases.

3. The covariance matrix for the $\hat{\gamma}$s is wrong, as it is just the covariance of the estimated fitted values from the regression ignoring the last m observations. This covariance matrix will tend to zero as sample size increases, even though each of the $\hat{\gamma}$s is determined by a single observation. Any inference based on these covariance will be incorrect.

4. The heteroscedastic model that the true variances are $\sigma^2\mathbf{W}$ for an unknown $\mathbf{W}$ actually precludes any outlier testing because elements of $\mathbf{W}$ for the focal units are unknown, so no meaningful variance can be computed for the predictions.

2 Moral

Don't use the sandwich type estimates in prediction problems or to identify outliers. Do the work to find a get a model like (1-2) that fulfills assumptions.

References

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression.* Boca Raton FL: Chapman & Hall/CRC. [Online; accessed 12-October-2022].

Fox, J. and S. Weisberg (2019). *An R Companion to Applied Regression* (Third ed.). Sage.

Gruber, M. H. J. (2014). *Matrix Algebra for Linear Models.* Hoboken NJ: Wiley.

Long, J. S. and L. H. Ervin (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician 54*(3), 217–224.

Schott, J. (1997). *Matrix Analysis for Statistics.* Hoboken NJ: Wiley.

Weisberg, S. (2014). *Applied Linear Regression* (Fourth ed.). Wiley.

White, H. (1980). A heteroskedastic consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica 48*, 817–838.