Wales Institute of Social & Economic Research, Data & Methods
Sefydliad Ymchwil Gymdeithasol ac Economaidd, Data a Dulliau Cymru

Disclosure control for regression outputs

WISERD DATA RESOURCES

WISERD/WDR/005

Felix Ritchie

December 2011

Authors
Felix Ritchie

Address for Correspondence:

Microdata Analysis and User Support
Office for National Statistics
Cardiff Road
Newport
South Wales
NP10 8XG

Email: felix.ritchie@ons.gsi.gov.uk or felix.ritchie@virgin.net

WISERD Hub Contact:

Cardiff University
46 Park Place
Cardiff
CF10 3BB

Tel: 02920879338

Email: wiserd@cardiff.ac.uk

Abstract

Disclosure detection and control for analytical outputs is an almost unexplored field. However, with the increase in access to detailed microdata, it is becoming increasingly important to be able to quantify exactly what the risks are from allowing, for example, regression coefficients to be released.

This paper looks in detail at the risks of linear regressions, and demonstrates that, even in the best-case scenario for an intruder, analytical results are fundamentally safe, and can be made utterly non-disclosive by the application of simple rules. Estimation of the risk of likely disclosure is also considered, and it is shown that the NSI can carry out its own safety tests easily, and can also prevent intruders generating meaningful fitted values by application of the same rules. Some comments on more general functional forms are provided.

Acknowledgements

The author is grateful for comments from colleagues at ONS, Statistics New Zealand and the US Census bureau; from participants at seminars and conferences in the UK; and from numerous academic researchers, particularly those that attend ONS' training courses for researchers; and Jonathan Haskel for the early discussions that sparked this work. All remaining errors are of course my own.

This paper originally circulated and was referenced as Ritchie F (2006) *Disclosure Control for Regression Outputs*, mimeo, Office for National Statistics, June.

1. Introduction

Most disclosure control techniques are concerned with providing safe microdatasets for research use, or for making aggregate statistics safe. In both cases, “safe” refers to combining, perturbing, removing or summarising the data in such a way that the confidentiality of the underlying data can be maintained. Almost no attention has been paid to the possible risks in analytical outputs, such as regressions, survival functions, factor analysis, and so on. A special edition of *The Journal of Official Statistics* (Feinberg and Willenborg, 1998) on confidentiality omitted the question of analytical outputs entirely. The two notable exceptions are Reznek (2004), who summarizes the literature on conditional explanatory variables and generalises this to the class of exponential general linear models; and Corscadden et al (2006) who derive expressions for the riskiness of regression results based upon summary statistics.

This is an important omission because in recent years a combination of increased computer power and changing policy regimes has led to a significant increase in access to confidential microdata for research purposes, particularly in national statistical institutes (NSIs). Whilst technological solutions vary across countries, a common feature is some form of laboratory, physical or virtual, where the researcher has freedom to operate but the NSI acts as a guardian of statistical outputs removed from the premises. This requires a different approach to disclosure control (see Ritchie, 2005). As outputs will often consist of analytical work, the NSI needs to have some way of evaluating the disclosiveness of outputs quickly and easily. With developments in model servers seen as one way to solve the issue of access to raw microdata (see Steel and Reznek (2006)), the need for guidelines which can be implemented automatically becomes even more important.

There is some confusion over the evaluation of analytical outputs. Disclosure control methodologists have suggested variations on rules designed for tables (for example, minimum frequencies, no influential or dominant points). Another key protecting factor is to put limits on the types of variables that are dangerous (outliers, “public” variables, extremely heterogeneous values etc). As these rules are typically designed for tabular outputs or anonymised data, the application of these rules can be at best inappropriate and at worst ineffective.

Researchers, on the other hand, will typically view analytical outputs as inherently safe because of the transformation of data, and will view attempts to control output of analytical results as needlessly bureaucratic. However, this is done without any formal proof. As a result of this difference in views, the international trend to wider access to restricted data runs the risk of being stymied due to confusion over what can be released.

The aim of this paper is to show that

- The researchers' view, that regressions are inherently safe, is generally correct
- There are a very small number of cases where problems could arise
- Even in these cases, the problem is the publication of summary statistics, not coefficients
- A simple rule is available to assess and ensure the safety of regression outputs
- Concern over the nature of variables and the validity of analysis is misplaced

We consider an extreme intruder scenario: that an intruder acquires a set of regression coefficients and standard summary statistics from repeated estimation on the same or a similar sample; that he/she has a large amount of information about the type and means of the variables and the sample; and that his/her only interest is in discovering something that should have been hidden – for example in order to embarrass an NSI. The purpose of this is to show that, even in the intruder's best-case scenario, that chances of being able to uncover information range from negligible to zero. Hence, in realistic applications, NSIs can feel confident about the application of the results here.

The next section describes the circumstances under which data points can be exactly identified, and how this can be prevented. Section three reviews approximate identification, and section four looks briefly at non-linear models. Section five discusses other aspects of analytical outputs which are relevant for disclosure control. Section six concludes.

2. Exact identification in a linear regression

In this section we consider a linear least-squares regression with N observations of the form

$$y_i = x_{i1}\beta_1 + ... + x_{iK}\beta_K + u_i \qquad i = 1..N \qquad u_i \sim \left(0, \sigma^2\right)$$

We deal only with "genuine" regressions; that is, where N>(K+1) and K>1. It is not necessary, for the purposes of disclosiveness, to specify that the sample distributions of the variables do not collapse. We also assume that a researcher does not "create" regressions solely for the purpose of disclosure by differencing. The issue of the trustworthiness of researchers is outside the scope of this paper.

We also do not make assumptions on the distribution of the disturbance term at this stage.

2.1 Direct disclosure

Direct disclosure from a genuine linear regression is not possible without an almost perfect knowledge of the data. We assert this without proof; the result will become clear in the following section, as this case of direct disclosure from a single regression is a reparameterisation of the problem of disclosure by differencing two regressions. However, intuitively this may be explained as follows.

A linear regression to determine K parameters implies K independent equations. These equations are linear in the coefficients but not in the explanatory variables. If the coefficients are known but the one or more of the variables is unknown, this can be calculated by unpicking the normal equations. This is feasible as long as the number of unknowns is not more than K. Therefore, for an intruder to be able to ascertain specific values he already needs to know NK values out of a possible (N+1)K. Conversely, a researcher can prevent a regression being disclosive by ensuring that at least K+1 variables are not known to the intruder.

The sole exception to this rule is where the explanatory variables are all binary. In this case the regression coefficients reflect table means, and few observations in particular categories can be disclosive. This holds for the class of exponential linear models: see Reznek(2004).

2.2 Disclosure by differencing

2.2.1 Two-variable case

In the two-variable case,

$$y_i = \alpha + x_i\beta + u_i \qquad i = 1..N \qquad u_i \square \left(0, \sigma^2\right)$$

The solutions for this model are given by

$$\hat{\alpha} = \sum \left(y_i - x_i \hat{\beta} \right) / N$$
$$\hat{\beta} = \left(\sum x_i^2 \right)^{-1} \left(\sum x_i y_i \right)$$

Consider the case where an intruder has two regression results. The difference between the two is that the second regression has one additional observation. Can anything be determined by the values of some variables and the estimated coefficients?

If the regression is re-run with the additional observation (x_0, y_0) to produce estimates

$$\hat{\alpha}_0 = \left(\sum\left(y_i - x_i\hat{\beta}_0\right) + y_0 - x_0\hat{\beta}_0\right)/\left(N+1\right)$$
$$\hat{\beta}_0 = \left(x_0^2 + \sum x_i^2\right)^{-1}\left(x_0 y_0 + \sum x_i y_i\right)$$

then the equations

$$\hat{\alpha} - \hat{\alpha}_0 = \left(\left(N+1\right)\sum\left(y_i - x_i\hat{\beta}\right) - N\left(\sum\left(y_i - x_i\hat{\beta}_0\right) + y_0 - x_0\hat{\beta}_0\right)\right)/N\left(N+1\right)$$
$$\hat{\beta} - \hat{\beta}_0 = \left(\sum x_i^2\right)^{-1}\left(\sum x_i y_i\right) - \left(x_0^2 + \sum x_i^2\right)^{-1}\left(x_0 y_0 + \sum x_i y_i\right)$$

contain two unknown values (x_0, y_0) but do not have a unique solution. It might be possible to impose one from economic knowledge (for example, wages must be positive) but this still requires that the other observation values ($x_1 \ldots x_N$, $y_1 \ldots y_N$) are all known. The non-linear interaction in the last term in the second equation mean that a complete knowledge of the N other observations is required in the general case.

It is possible to speculate that particular combinations could be both plausible and informative. We consider three cases which require less information than the whole dataset.

Case 1: known means of original variables and known values of additional variables

The property of the OLS estimator that estimated errors identically sum to zero implies that

$$\begin{aligned} y_0 &= \left(\sum_{i=1..N} y_i + y_0\right) - \sum_{i=1..N} y_i \\ &= (N+1)\hat{\alpha}_0 - N\hat{\alpha} + \left(\left(\sum_{i=1..N} x_i + x_0\right)\hat{\beta}_0 - \sum_{i=1..N} x_i\hat{\beta}\right) \\ &= (N+1)\hat{\alpha}_0 - N\hat{\alpha} + \left(N\overline{x}\left(\hat{\beta}_0 - \hat{\beta}\right) + x_0\hat{\beta}_0\right) \end{aligned}$$

In this case the additional value can be ascertained directly, irrespective of the individual values of x.

However, if more than one additional observation is included, then only the sum (or mean) of the additional dependent variables can be ascertained. This is because the above result rests upon the overall prediction error of the regression being zero, not the prediction error of the additional observations. This is developed further in the K-variable case, below.

Case 2: binary explanatory variables

Suppose the original x variables are 1/0 binaries with a n_1/n_0 split ($n_0+n_1 = N$). If a new pair of observations (y_0, x_0) is included in a new regression, then

$$x_0 = 0 \rightarrow y_0 = \left((N+1)\hat{\alpha}_0 - N\hat{\alpha}\right) + \left(\left(n_1+1\right)\hat{\beta}_0 - n_1\hat{\beta}\right)$$
$$x_0 = 1 \rightarrow y_0 = \hat{\alpha} + \left(\left(n_1+1\right)\hat{\beta}_0 - n_1\hat{\beta}\right)$$

Note that the intruder does not need to know in advance whether x_0 is 0 or 1; this can be determined easily by inspecting the constant term:

$$\hat{\alpha} = \hat{\alpha}_0 \rightarrow x_0 = 1$$
$$\hat{\alpha} \neq \hat{\alpha}_0 \rightarrow x_0 = 0$$

It is plausible that the sample proportions for the explanatory variables could have been published elsewhere, and therefore both values (y_0, x_0) can be inferred from published results only. However, if more than one observation is added then only the sum of unobserved values can be determined, even if the explanatory variables are known. This is because the above result depends on the zero-mean-error property of least-squares estimators.

This result is plausibly disclosive because the only explanatory variable is a binary variable: the estimates reflect set sizes not correlations, and so the frequency count is a sufficient statistic for the moments of x_i. This is not possible where an explanatory variable has more than two values.

Case 3: binary dependent variable, relative value of new observation known

This example is relevant for the linear probability model:

$$y_i^* = \alpha + x_i\beta + u_i \qquad y_i^* < 0.5 \rightarrow y_i = 0, \quad y_i^* \geq 0.5 \rightarrow y_i = 1$$

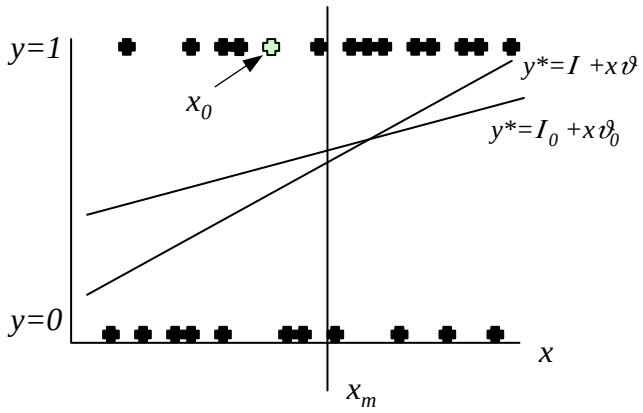
or for any model with a dichotomous outcome (as it can always be scaled to the above case). Define, using the above notation,

$$\tilde{y} \equiv (\hat{\alpha}_0 - \hat{\alpha}) + \frac{1}{N+1}\left(\left(N\hat{\beta}_0 - (N+1)\hat{\beta}\right)\bar{x} + x_0\hat{\beta}_0\right)$$

Then

$$\tilde{y} > 0 \quad \rightarrow \quad y_0 = 1 \qquad \tilde{y} < 0 \quad \rightarrow \quad y_0 = 0$$

In other words, a knowledge of the position of new observation relative to the original mean and the effect on the estimated coefficients can be used to determine whether the dependent variable has a positive or negative outcome. Diagrammatically this can be shown below:



The original mean is x_m. The new observation is below the mean but has flattened the slope, implying y_0 was a positive outcome against the predictions of the initial model.

In this case, the regression is potentially disclosive because

- the change in the slope can be unambiguously determined
- the additional dependent variable can have only two values
- the position of additional observation relative to the previous mean can be determined
- the monotonic function allows the change in slope to be unambiguously associated with the change in the dependent variables

Unlike the first case, the exact value of the mean is not required; only the relative value of the original mean and the new observation is required. If the mean is available, it is possible to

determine the dependent variables for two observations (distances from the mean act as relative weights and give the necessary second equation to solve the system).

These three examples illustrate cases where a linear regression is potentially disclosive without a complete knowledge of the other variables in the dataset. It may be possible to define other cases where plausible combination of known variables and functional form give rise to potentially disclosive results, but it should be clear by now that these are exceptional cases rather than the rule.

In addition, in each case we have specified that one single observation is the difference between the two regressions. If more observations are included, the individual values cannot be determined in the first two cases, and the binary dependent variables in the third case can only be ascertained for two additional observations if the exact values of the new explanatory variables and the means are known. In all cases if three or more observations is the difference between equations then the individual values cannot be identified.

In summary, in the two-variable case there are a limited set of conditions where it may be possible to ascertain exact values without a complete knowledge of the data; in general, however the regression is not disclosive in any meaningful way.

2.2.2 K-variable case

Extending this example to the general case of K variables we have, in matrix form,

$$y_i = x_i\beta + u_i \qquad i = 1..N \qquad x_i' = (x_{i1}...x_{iK}) \qquad \beta' = (\beta_1...\beta_K)$$

More compactly

$$y = X\beta + u \qquad y' = (y_1...y_N) \qquad X' = (x_1...x_N) \qquad u' = (u_1...u_N)$$

Define y_0, X_0, and u_0 as Sx1, SxK and Sx1 matrices of additional observations, and β_0 as the corresponding estimate:

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$
$$\hat{\beta}_0 = \left(X'X + X_0'X_0\right)^{-1}\left(X'y + X_0'y_0\right)$$

Following the same logic as above:

$$\hat{\beta} - \hat{\beta}_0 = \left(X'X\right)^{-1} X'y - \left(X'X + X_0'X_0\right)^{-1}\left(X'y + X_0'y_0\right)$$

This is a system of K equations. Therefore, it is directly solvable if there are K unknowns. To see this, consider the identification of y_0:

$$\begin{aligned} X_0'y_0 &= \left(X'X + X_0'X_0\right)\left(\left(X'X\right)^{-1} X'y - \hat{\beta} + \hat{\beta}_0\right) - X'y \\ &= X'X\left(\hat{\beta}_0 - \hat{\beta}\right) + X_0'X_0\hat{\beta}_0 \end{aligned}$$

Solving for y_0:

$$y_0 = \left(X_0X_0'\right)^{-1} X_0X'X\left(\hat{\beta}_0 - \hat{\beta}\right) + X_0\hat{\beta}_0$$

This equation has a solution if S≤K; in other words, as long as no more <u>new</u> observations are added than there are variables, an exact calculation of the value of y_0 is possible.
could be better explained…something about weighted average and identity in equations above?

In general this solution requires full knowledge of the explanatory variables. Again, are there are plausible situations for which less knowledge is required?

One candidate is the orthogonality of the variables. Suppose the explanatory variables are truly orthogonal ie X'X is diagonal; for example, X is composed of a single categorical variable with K or K-1 categories (allowing for the constant term). Then for each coefficient

$$\hat{\beta}_k = \left(\sum x_{ik}^2\right)^{-1} \sum x_{ik}y_i$$

Therefore, each coefficient can be assessed independently. However, the non-linear interactions of the explanatory variables mean that a full knowledge of the variables is still required, unless the sums were published for some reason. Orthogonality per se does not mean that a regression is disclosive.

It can be shown that, just as for the two-variable case, if the X matrix does consist exclusively of binary variables then a plausible problem can be identified. Define a tx1 unit vector $J_t=(1..1)'$. Using the same argument as before, that the mean error is identically zero,

$$\begin{aligned} J_S' y_0 &= \left(J_N' Y + J_S' y_0 \right) - J_N' Y \\ &= \left(J_N' X \hat{\beta}_0 + J_S' X_0 \hat{\beta}_0 \right) - J_N' X \hat{\beta} \\ &= J_N' X \left(\hat{\beta}_0 - \hat{\beta} \right) + J_S' X_0 \hat{\beta}_0 \end{aligned}$$

or in means

$$\bar{y}_0 = (N/S) \bar{X} \left(\hat{\beta}_0 - \hat{\beta} \right) + \bar{X}_0 \hat{\beta}_0$$

As for the two-variable model, only the total effect of the additional observations can be deduced (as JJ′ is not invertible). Only if a single observation is added can the dependent variable be deduced from just sample means and estimated coefficients.

As for the two-variable case, binary explanatory variables simplify the need to know means. Define N_1 as the Kx1 vector of frequencies in the matrix X. Then

$$\bar{y}_0 = (N/S) N_1' \left(\hat{\beta}_0 - \hat{\beta} \right) + \bar{X}_0 \hat{\beta}_0$$

It is plausible to assume that an intruder might have frequency tables, in which case N_1 is known. As for the two variable case, it is only possible to determine a value for Σy_{0i}. However, unlike the simpler case, X_0 cannot be inferred with K>2, even for a single additional observation; therefore, X_0 must be known.

It is often claimed that regressions containing only categorical variables are as disclosive as frequency tables, as the orthogonal nature of categorical variables means that coefficients reflect set sizes. The above discussion places the question of categorical variables within the context of regression results generally, and so a special rule is not required for these variables. The reason regressions with only binary variables cause concern is <u>not</u> because variables are categorical *per se* but because the sample proportion of positive responses is a sufficient statistic for Σx_{ik} (and thus is not relevant where other values are possible). It is quite conceivable that these frequencies may be available from other tables (whereas, for example, $\Sigma x_{ik} yi$ is not the sort of statistic usefully tabulated). This will become relevant when discussing possible responses in the next sub-section.

To summarise, in the K-variable case (K>2)

- orthogonality of regressors is not a sufficient condition for identification
- an incomplete knowledge of the matrix of explanatory variables is a sufficient condition for non-disclosiveness, unless
    - a sufficient statistic for Σx_{ik} exists, in which case an intruder can at best only determine Σy_{0i}

2.3 A simple rule to prevent direct identification

The above discussion shows that exact identification from a regression or combination of regressions is not easy and requires a specific set of conditions, such as solely binary variables or a complete knowledge of other variables. A simple rule can then be stated for use in research laboratories

> *In general, the exact values of variables underlying a regression cannot plausibly be determined unless the regression consists entirely of categorical variables or has a dependent binary variable; and disclosure by differencing is only possible route for identification.*

A simple addition can be devised that prevents even the extreme cases:

> *A linear regression is completely non-disclosive if (1) one or more coefficients is effectively suppressed (that is, the coefficient could not reasonably be determined from published information), and (2) the relevant variable is not orthogonal to all other variables*

By “could not reasonably be determined” we mean that no plausible information available to an intruder can be used to determine unknown values.

This covers all the cases above. Without a full set of coefficients it is clear that none of the equations above are solvable for additional observations. This also prevents disclosure by repeated differencing. Each new regression will create a new unknown variable, the new estimated mean, which in turn affects all other values. It is not therefore possible to build up a sequence of regression results to determine the unknown parameters. Nor is it possible to reconstruct the omitted constant and still determine other values. *check this last point but I’m pretty sure*.

The phrase “not orthogonal to all other variables” covers the case of estimation only on categorical variables. If estimation is carried out on a set of mutually exclusive variables, the values for any variable can be determined by differencing without reference to the others. However, where there is any non-zero correlation the missing coefficients cannot be re-estimated. Hence the “special case” of categorical variables can be dealt with in the same framework as other regressions.

This does not require that an estimated coefficient be statistically significant. The rule derives from the mathematical properties of the normal equations, not the statistical properties of the data. Suppressing a significant coefficient reinforces the rule but is not strictly necessary.

This suppressed-coefficient rule has the advantage of being clear, easy to implement and causing few problems for researchers. In business data a range of incidental parameters is often produced (such as industry or time dummies) in addition to the constant, any or all of which are commonly left out of published results. The rule has been in use at UK Office for National Statistics since early 2004 and has met little resistance.

One particularly useful effect of this rule is that a class of models which estimate incidental parameters become inherently safe. An important member of this is panel or longitudinal data (repeated measurement). A model such as

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}$$

will estimate individual-specific effects, even for random-effects models. These tend to be both numerous and of little interest and so are omitted from published results. These results will be non-disclosive without the need to omit estimates from the main coefficient vector.

3. Evaluating the likelihood of approximate disclosure

The previous section described how exact identification of values can be prevented. Exact identification is highly unlikely even without the above precautions in place, because it relies upon being able to difference regressions effectively, which in turn requires detailed information about how the regressions were constructed.

However, it may be sufficient for an intruder to have a rough idea of the value of a variable – for example, by taking coefficients and creating fitted values of the dependent variable. In this section we consider how we can quantify this risk and whether any additional rules are necessary. We

concentrate on created fitted values for dependent variables. A similar analysis could be carried out for trying to identify an explanatory variable.

3.1 Approximating values

Using the same matrix notation as before

$$y = X\beta + u$$

Estimated parameters are:

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$
$$\hat{\sigma}^2 = u'u/\left(N-K\right)$$

with the estimated variance typically being reported alongside coefficient estimates. The equation residuals are

$$e = y - \hat{y} = X\beta + u - X\hat{\beta} = u - X\left(\hat{\beta} - \beta\right)$$

Suppose an intruder wishes to find the exact value of a dependent variable y_1. The residual e_1 has the expected value

$$E(e_1) = E\left(u_1 - x_1'\left(\hat{\beta} - \beta\right)\right) = 0$$

and variance

$$\begin{aligned} V(e_1) &= E\left(\left(u_1 - x_1'\left(\hat{\beta} - \beta\right)\right)^2\right) \\ &= E\left(x_1'\left(\hat{\beta} - \beta\right)\left(\hat{\beta}' - \beta'\right)x_1 + u_1^2 - 2u_1x_1'\left(\hat{\beta} - \beta\right)\right) \\ &= \sigma^2 x_1'\left(X'X\right)^{-1}x_1 + \sigma^2 - 2E\left(u_1x_1'\left(\hat{\beta} - \beta\right)\right) \end{aligned}$$

The unknown error term u_1 is not independent of the estimated coefficients. Recalling that

$$\hat{\beta} - \beta = (X'X)^{-1} X'y - \beta$$
$$= (X'X)^{-1} X'(X\beta + u) - \beta$$
$$= (X'X)^{-1} X'u$$

Then

$$E\left(u_1 x_1' \left(\hat{\beta} - \beta\right)\right) = E\left(x_1' (X'X)^{-1} X'u.u_1\right)$$
$$= x_1' (X'X)^{-1} x_1 \sigma^2$$

and so

$$V(e_1) = \sigma^2 \left(1 - x_1' (X'X)^{-1} x_1\right)$$

This is smaller than the standard error of the regression, reflecting the fact that this observation contributed to the estimates. It reaches its minimum value when this observation contributes most to the regression ($X'X \rightarrow x_1 x_1'$), and approaches the standard error when the observation has a negligible impact ($x_1 \rightarrow 0$)[1].

If the published descriptive statistics are available, then an exact confidence interval can be calculated without the need for variable values. Using

$$\hat{\sigma}^2 = (TSS - ESS)/(N - K)$$
$$R^2 = ESS / TSS$$
$$ESS = \hat{\beta}' X'X \hat{\beta}$$

Then

$$x_1' (X'X)^{-1} x_1 = \left(x_1' \hat{\beta} \hat{\beta}' x_1\right) / ESS$$
$$= \left(\sum_k x_{1k}^2 \hat{\beta}_k^2\right) / ESS$$
$$= \left(\sum_k x_{1k}^2 \hat{\beta}_k^2\right) R^2 / \left(\hat{\sigma}^2 (N - K)\left(1 - R^2\right)\right)$$

[1] Rewrite as deviations from the mean…?

When evaluated at the largest vector in X, this enables the minimum predictive error on a dependent variable to be ascertained. In other words, this allows the NSI to determine whether an intruder, working with the published coefficients and descriptive statistics, would be able to derive a fitted value within a specified level of certainty.

Note that, although the above term contains ESS as a level not a ratio, and appears to be increasing in N, it cannot be stated that $N \to \infty$ leads to the error converging to the standard error of the regression. This ignores the dependency of the estimates of β on the current set of observations (X, y). For us to assert that the predictive error converges to the standard error would require some assumption on the distribution of variables, which we have avoided doing so far.

3.2 Approximation for new observations

If the published coefficients are used for prediction by the application of a new set of observations x_0, then a similar set of confidence limits can be derived. Without detailed proof (such a proof can be found in standard econometrics texts) we offer

$$E(e_0) = 0$$
$$V(e_0) = \sigma^2 \left(1 + x_0' \left(X'X\right)^{-1} x_0\right)$$

The intuition behind this is that the new error is assumed to be uncorrelated with the errors used to generate the coefficients. Therefore, the values of explanatory variables increase uncertainty as they move away from the mean values used in the regression.

In this case, the standard error of the regression is the minimum level of uncertainty, achieved when the new explanatory variables equal the mean of the variables used to calculate the coefficients. The predictive error cannot be reduced below this level.

3.3 Hiding the confidence interval

The above calculations provide an intruder with an indication of how likely his predictions are to be wide of the mark. They do not in themselves help with the identification of values. Nevertheless, it may be prudent to restrict an intruder's ability to define these confidence limits, to increase uncertainty surrounding any predictions.

The recommended solution is the same as for exact identification. Not publishing a coefficient means that neither a point prediction or a confidence interval can be determined. Again, it not necessary that the suppressed coefficient be statistically significant, as long as its insignificance is also not reported.

An alternative is to restrict the publication of descriptive statistics. This is not preferred. The statistics are published because they are useful. Suppression of descriptive statistics also cannot prevent exact disclosure as described above. This therefore requires two rules to be implemented instead of one.

3.4 Using R^2 directly as an estimate of riskiness

Corscadden et al (2006), using a similar analytical approach to functional form, develop an alternative measure where a direct relationship between R^2 and the required level of uncertainty in a regression can be quantified. This is a measure of the average riskiness, not the maximum, and, as in the above example, could be relatively easily coded to be a standard output from regressions.

4. Non-linear estimation

A non-linear estimate is inherently non-disclosive. Define a basic equation and the resulting estimate

$$y = f(X, \beta, u)$$
$$\hat{y} = f(X, \hat{\beta}, 0)$$

where y, X, β, and u are appropriate matrices or vectors. The characteristic of a non-linear equation is

$$\frac{\partial f(X)}{\partial X} \neq c$$

and that therefore

$$dy = \frac{\partial f(X, \beta, u)}{\partial X} dX \neq f(dX, \beta, u)$$
$$dy = \frac{\partial f(X, \beta, u)}{\partial \beta} d\beta \neq f(X, d\beta, u)$$

implying

$$y - \hat{y} = f(X, \beta, u) - f(X, \hat{\beta}, 0) \neq f(X, \beta - \hat{\beta}, u)$$

and

$$\overline{y} = \sum f(X, \beta, u) / N \neq f(\overline{X}, \hat{\beta}, 0)$$

This is in contrast to the linear case where the final equalities hold.

In general, there is no opportunity to differentiate two equations on the basis of summary statistics to identify the value of explanatory variables. Some exceptions can be derived; as noted in section 2.2.1 for the linear case, on a single explanatory variable and a binary dependent variable, the difference between the additional variable and the mean is sufficient to identify the qualitative features of the dependent variable. This does not hold in the general case (K>2) due to the interactions between explanatory variables. Reznek(2004) does however point out that where all the explanatory variables are binary some inferences can be drawn.

Because of the range of non-linear models, this paper does not investigate this issue further. This is an area for more work.

For non-exact identification, as for the linear case both fitted values and confidence intervals can be calculated. However, as for the linear case, hiding certain coefficients makes this completely non-disclosive. Hence the above rule still holds.

5. Discussion: the role of means and coefficients

In the preceding sections, regression models have been unpicked to generate special cases regressions might be disclosive by differencing. A solution proposed has been to hide certain coefficients, which solves both the problem of disclosure by differencing and the problem of calculating confidence intervals for fitted values.

There is however, an alternative. Many of the above results depend upon knowledge of the means of the variables (in the case of binary variables, these are frequencies). Without means, similar conclusions on the non-disclosiveness of regressions can be reached.

However, there are reasons for focusing on the hiding of coefficients:

- means are useful statistics and therefore being unable to publish means along with regressions would inconvenience researchers. This is particularly true in the case of binary frequencies
- many coefficients are “incidental” parameters; that is, they are included to improve the fit of the regression but are not of direct interest. Such parameters include time dummies, individual intercepts in panel models, sample conditioning variables, and even the constant in most cases.
- coefficients are specific to a regression and are therefore not easily reproduced by other researchers. Means, on the other hand, have an existence independent of any regressions, and so are more likely to be generated “by accident” in papers unrelated to the work in hand. It is quite possible that the means for variables would not all be published together, but could be split over several papers

In short, reducing the number of published coefficients is likely to meet less resistance from researchers and also offers more security that the omitted values are not going to be reproduced elsewhere.

6. Statistical quality and other issues

6.1 Quality of the regression

It has been suggested that certain features of data such as outliers or multicollinearity can increase the disclosiveness of regressions. These points can be addressed as follows:

Outliers: outliers are variables with large deviations from the regression line, but which are in themselves not significant in determining the relationship. It should be clear from the above commentary that this is not an issue for disclosure control. An outlier will have a large variance and poor fitted value. These make it less disclosive than any other variable, if anything. It does not add anything to the overall disclosiveness of the regression

Influential points: these differ from other outliers in that they do have a significant effect on the regression line – for example, by running regressions on several small companies and one very large one. This is a particular concern for SDC methodologists, as this is the situation in which differences between regressions are (a) most likely to be discernible and (b) most likely to be published – for example, a researcher is interested in demonstrating the impact of large companies. Section 3 gave the formula for calculating the confidence interval for fitted values.

The omitted-coefficient rule still deals with this issue. Without the coefficient, neither a fitted value or a confidence interval can be calculated. Although an intruder might be aware that there has been a large impact on the regression, this cannot be quantified.

Multicollinearity: multicollinearity raises the standard error and makes attribution of effects to particular variables more difficult. It therefore raises no new SDC issues.

Measurement error: as for multicollinearity, this is not an SDC issue. Measurement error increases variances as well as biasing the coefficients downwards. It does not add any new disclosure risk..

Estimation on public explanatory variables: in theory, estimation on public explanatory variables with an excellent fit allows a good approximation to actual values to be generated. Aside from the likelihood of such a model being fitted, the formula in section 3 allows the minimum prediction error to be assessed. Moreover, work by Corscadden et al (2006) seems to show that in practice this overstates the likelihood of making accurate predictions. In any case, removing coefficients prevents an intruder generating fitted values and confidence intervals.

In short, it should be clear from these examples that it is important to distinguish statistical quality from disclosiveness. The latter is not determined by whether a model is good or bad, but by the alternative information available. That said, on the whole poorly specified regressions would tend to cause fewer concerns for SDC.

An exception to the “bad is good” rule is where there are few observations. If a researcher estimates a model with no degrees of freedom, clearly coefficients relate directly to values of explanatory variables. However, this is an area where it is possible to identify quickly whether a regression is genuine or not. There is work to be done on determining whether there are disclosure issues in regressions with few degrees of freedom.

6.2 Transforming variables and relationships

Converting non-linear to linear equations (for example, via GEE, or log-linearisation) does not change the results of any sections. The emphasis in this paper is to see whether the form of estimated relationship is itself disclosive. Whether the variables themselves are useful is another issue. The linearised model has the same characteristics as the linear model described above, and hence should be treated as such (see Reznek(2004)).

Clearly, however, the discussion above has been taking place in an idealised world for intruders. In practice, data transformations, sample selection, treatment of missing values, simultaneous

equations, solution algorithms, method of estimation etc will all make the reproduction of the regression environment by intruders extremely difficult..

6.3 Recovering omitted coefficients

One potential flaw in the omission-of-coefficients argument is that it may be possible to recover coefficients. For example, if the estimated constant is omitted but the means of all variables left in, then the constant can be easily recalculated.

This is a red herring. With all the means and coefficients, a researcher can unpick the regression to determine the means of additional dependent variables. However, there is no incentive to do this. The same values can be derived entirely from the difference of the means – as can the means of the explanatory variables, which cannot be derived from the normal equations. The regression itself contributes nothing to increased disclosure risk.

6.4 Regressions on a single unit

The above discussions relate to regressions on several units, and assumes that an intruder is trying to get information on one unit. However, it is possible that a regression could be run on a single unit – for example, quarterly data on the performance of a company could provide sufficient observations to run regressions solely on that company. In this case, all coefficients are directly informative about the company, and hiding one or two does not reduce the disclosiveness of the others. This is a problem for the NSI which is not addressed here, as the solution requires the NSI to identify the units in regressions[2].

6.5 Releasing residuals

The above analysis assumes that the intruder does not have access to the residuals of the regression. If residuals were released, even if not identified with particular units, it could that scenarios could be constructed where even a reduced coefficient set would be informative. For example, it may be that, if most variables have a limited range (age, say, or categorical variables), then an intruder could try to identify units by looking at extreme values which could not be generated from known coefficients and acceptable variable range. At the moment, this is highly speculative, and one would suspect that the mean-reverting qualities of regression would make this outcome unlikely, but this clearly requires further work.

[2] In the UK this is addressed by having a blanket ban on regressions on individual companies. The author is grateful to Martin Weale for raising this possibility.

6.6 Releasing coefficients for prediction

One aim of modelling is to release a set of coefficients that can be used to predict values in other dataset (for example, using earnings information in one dataset to construct a model which can then be used to generate a predicted income variable in a second dataset). In this case, holding back coefficients is not a valid operation. However, as shown above and in Corscadden et al (2006), it is perfectly possible to assess the prediction risk for a full set of coefficients so that the risk of re-identification in the original dataset can be quantified. This is a maximum risk estimate, and would need to be adjusted to take account of, for example, the unavailability of the true explanatory variables.

7. Conclusion

This paper has discussed the opportunities for determining confidential information from regression outputs. This is an arcane but important topic: as increasing amounts of analysis are carried out on data in secure environment, there is little proof one way or the other to show whether there are any disclosure control issues for analytical results.

This paper has addressed one issue, that of regressions. In conditions conducive to intruders, it has shown that retrieving individual data points from estimated values and summary statistics is almost, but not quite, impossible. The exceptional cases can be identified in the linear case; for non-linear estimates, further work needs to be done.

Even for exceptional cases, a simple rule allows results to be made completely safe. This rule is simple, easily enforceable, classifies a group of models as inherently safe, and in practice has proved uncontroversial with researchers in the UK Virtual Microdata Laboratory since its introduction in 2004.

This has had a significant impact on the ability of the VML to process a large amount of requests for output with a very small number of staff: the target clearance time for results has dropped from two weeks to two days, with the median clearance time less than one day. This is therefore not a theoretical demonstration but a result which has a direct impact on the practices of NSIs and other guardians of confidential data.

This paper has presented the intruder with a near-ideal environment – the data is inherently interesting, has not been transformed or sampled in some way that would make it difficult to

identify the included observations, values of additional explanatory variables may be known. In practice, none of these conditions are likely to hold. Therefore, a linear regression can in general be treated as an extremely safe output, in that there is little practical chance of some of the access routes mentioned here to be exploited. The view of researchers, that regressions are inherently safe, is therefore upheld. Moreover, we have demonstrated here that a simple adjustment to outputs, one which is often done automatically when publishing results, makes them completely opaque.

References

Corscadden, L., Enright, J., Khoo, J., Krnsich, F., McDonald, S. and Zeng, I. (2006) *Disclosure assessment of analytical outputs*, mimeo, Statistics New Zealand, Wellington

Feinberg S.E., and L.C.R.J. Willenborg (1998), "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data", *Journal of Official Statistics*, v14:4 pp337-345

Reznek, A. (2004) *Disclosure risks in cross-section regression models*, mimeo, Center for Economic Studies, US Bureau of the Census, Washington

Ritchie, F.J. (2005) *Statistical disclosure control in a research environment*, mimeo, Office for National Statistics, London

Steel, P and Reznek, A. (2006) "Issues in designing a confidentiality-preserving model server", in *Monographs in Official Statistics: Work session on Statistical Data Confidentiality Geneva 2005*, UN/ECE, Geneva