Gaure, Simen

**Working Paper**

# OLS with multiple high dimensional category dummies

Memorandum, No. 2010,14

**Provided in Cooperation with:**
Department of Economics, University of Oslo

*Suggested Citation:* Gaure, Simen (2010) : OLS with multiple high dimensional category dummies, Memorandum, No. 2010,14, University of Oslo, Department of Economics, Oslo

This Version is available at:
https://hdl.handle.net/10419/47280

# MEMORANDUM

**No 14/2010**

# OLS with Multiple High Dimensional Category Dummies

**Simen Gaure**

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
**University of Oslo**
**Department of Economics**

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.oekonomi.uio.no
e-mail: econdep@econ.uio.no

In co-operation with
**The Frisch Centre for Economic Research**

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

## Last 10 Memoranda

| | |
|---|---|
| No 13/10 | Michael Hoel<br>*Is there a green paradox?* |
| No 12/10 | Michael Hoel<br>*Environmental R&D* |
| No 11/10 | Øystein Børsum<br>*Employee Stock Options* |
| No 10/10 | Øystein Børsum<br>*Contagious Mortgage Default* |
| No 09/10 | Derek J. Clark and Tore Nilssen<br>*The Number of Organizations in Heterogeneous Societies* |
| No 08/10 | Jo Thori Lind<br>*The Number of Organizations in Heterogeneous Societies* |
| No 07/10 | Olav Bjerkholt<br>*The "Meteorological" and the "Engineering" Type of Econometric Inference: a 1943 Exchange between Trygve Haavelmo and Jakob Marschak* |
| No 06/10 | Dag Kolsrud and Ragnar Nymoen<br>*Macroeconomic Stability or Cycles? The Role of the Wage-price Spiral* |
| No 05/10 | Olav Bjerkholt and Duo Qin<br>*Teaching Economics as a Science: The 1930 Yale Lectures of Ragnar Frisch* |
| No 04/10 | Michael Hoel<br>*Climate Change and Carbon Tax Expectations* |

Previous issues of the memo-series are available in a PDF® format at:
http://www.oekonomi.uio.no/forskning/publikasjoner/memo/

# OLS with Multiple High Dimensional Category Dummies

Simen Gaure

*The Ragnar Frisch Centre for Economic Research, Oslo, Norway*

**Summary.** We present some theoretical results which simplifies the estimation of linear models with multiple high-dimensional fixed effects. In particular, we show how to sweep out multiple fixed effects from the normal equations, in analogy with the common within-groups estimator.

*Keywords:* Method of Alternating Projections, Multiple Fixed Effects, OLS
*JEL Classification:* C13, C33, C60

## 1. Introduction

Several authors (e.g. Abowd et al. (1999), Carneiro et al. (2009)) have implemented procedures for estimation of linear models with two fixed effects. We present some results which may simplify this.

A common strategy if there is a single fixed effect (e.g. individual fixed effects only) is to centre the covariates and response on the group means, and do OLS on this projected system.

It seems to be common knowledge that sweeping out more than one category variable may not be done by centering on the group means, or by other simple transformations of the data, see e.g. (Andrews et al., 2008, p. 676) and (Cornelissen and Hubler, 2007, Section 5.2). Thus, even if one only wants to control for the fixed effects, elaborate estimation schemes are employed.

We present results that it is indeed possible to sweep out multiple fixed effects, due to theorems by von Neumann and Halperin. Moreover, the residual linear system will typically be sparse. In addition, we may often decompose it into smaller systems, depending on its structure.

We assume we have a linear model

$$Y = X\beta + D\alpha + \epsilon$$

where $X$ is a $(n \times k)$-matrix, and $D$ is a $(n \times g)$-matrix. $D$ is a set of dummies for $e$ category variables. I.e. $D = \begin{bmatrix} D_1 & D_2 & \dots & D_e \end{bmatrix}$. That is, the entries of each $D_i$ consists of 0 and 1, with only 1 non-zero entry per row. Hence, the columns of each $D_i$ are pairwise orthogonal. Though, in general, $D_i$ is not orthogonal to $D_j$ for $i \neq j$.

We further assume that $g$ is large (as in $10^6$), so that ordinary least squares solvers fall short. We may also assume that $k$ is reasonably small, so that the system without dummies is manageable.

In particular we look at the case $e = 2$, corresponding to two category variables, e.g. "firm" and "employee" as in Abowd et al. (1999) and Andrews et al. (2008).

We have the normal equations

$$\begin{bmatrix} X'X & X'D \\ D'X & D'D \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} X' \\ D' \end{bmatrix} Y$$

We recall some standard facts about these. We may write them as two rows

$$X'X\hat{\beta} + X'D\hat{\alpha} = X'Y \tag{1}$$
$$D'X\hat{\beta} + D'D\hat{\alpha} = D'Y \tag{2}$$

and do Gaussian elimination of the $\hat{\alpha}$-term in the last row to get

$$X'(I - D(D'D)^{-1}D')X\hat{\beta} = X'(I - D(D'D)^{-1}D')Y \tag{3}$$

provided $D'D$ has full rank.

Now, let $P = I - D(D'D)^{-1}D'$ and note that $P = P^2 = P'$, to get

$$(PX)'(PX)\hat{\beta} = (PX)'PY$$

which shows that $\hat{\beta}$ is the OLS-solution of the projected system

$$PY = PX\beta + \epsilon \tag{4}$$

Thus, we don't need $\hat{\alpha}$ to find $\hat{\beta}$. This is a standard way to eliminate the fixed effects from the equation in the case $e = 1$, i.e. with a single fixed effect. This result is also known as the Frisch-Waugh-Lovell theorem.

REMARK 1.1. *We do not cover the statistical properties of such models (this has been done in Abowd et al. (1999) and Andrews et al. (2008)), only some theory to simplify the solution of the normal equations.*

In practice, it may happen that $\hat{\beta}$ is not uniquely determined, this is the same problem as with ordinary fixed effects models with $e = 1$, (e.g. covariates which are constant for individuals). We do not treat that problem here. That is, we assume $PX$ is full rank.

Note that $P$ is the projection onto $R(D)^{\perp}$ where $R(D)$ denotes the range of $D$, i.e. the column space, and $^{\perp}$ denotes orthogonal complement. So $P$ exists even if $D$ is not full rank, and only depends on the column space of $D$. Thus, removing columns of $D$ participating in linear dependencies does not change $P$, hence not $\hat{\beta}$.

## 2. The structure of the projection $P$

Although (4) shows how to eliminate the fixed effects for arbitrary $e \geq 1$, it does not seem practical to compute the projection matrix $P$ unless $e = 1$.

However, we don't really need to find the matrix $P$ (which is $N \times N$ where $N$ may be of the order $10^7$), rather we need to compute $PY$ and $PX$.

To this end, for each $i = 1..e$ let $P_i$ be the projection onto the orthogonal complement of the range $R(D_i)^{\perp}$. We clearly have

$$R(D)^{\perp} = R(D_1)^{\perp} \cap R(D_2)^{\perp} \cap \cdots \cap R(D_e)^{\perp},$$

thus

$$P = P_1 \wedge P_2 \wedge \cdots \wedge P_e.$$

By (Halperin, 1962, Theorem 1), we have

$$P = \lim_{n\to\infty} (P_1 P_2 \cdots P_e)^n.$$

This shows that the following algorithm converges.

ALGORITHM 2.1 (METHOD OF ALTERNATING PROJECTIONS). *Let $v$ be a vector (typically a column of $X$ or $Y$). The following algorithm converges to $Pv$. It is a direct generalization of the familiar "mean deviations" transformation of the "within-groups" estimator (i.e. the case $e = 1$).*

*(1) Let $v_0 = v$, and $i = 0$.*

*(2) Let $z_0 = v_i$. Let $j = 0$.*

*(3) For $j = 1..e$, form $z_j$ by subtracting the group means of the groups in $D_i$ from $z_{j-1}$. I.e. $z_j = P_j z_{j-1}$.*

*(4) Let $v_{i+1} = z_e$. If $\|v_{i+1} - v_i\| < \epsilon$, terminate with the vector $v_{i+1}$ as an approximation to $Pv$. Otherwise, increase $i$ by 1. Go to step (2).*

REMARK 2.2. *This is known as the* Method of Alternating Projections. *The case $e = 2$ was first proved in (von Neumann, 1949, Lemma 22, p.475), and it's also in (von Neumann, 1950, Theorem 13.7). This really dates back to 1937, and even to lecture notes from von Neumann's lectures at Princeton in 1933–1934. It is also known as* The Kaczmarz Method, *from Kaczmarz (1937).*

EXAMPLE 2.3. *In the case with two fixed effects, like "firm" and "individual", Algorithm 2.1 amounts to repeating the process of centering on the firm means, followed by centering on the individual means, until the vector no longer changes.*

REMARK 2.4. *It is a standard fact from operator theory that for commuting projections $P_i$, we have $P_1 \wedge P_2 \wedge \cdots \wedge P_e = P_1 P_2 \cdots P_e$. (It is also a direct consequence of the Halperin theorem cited above.) Thus, for such commuting projections, Algorithm 2.1 converges after one step so that $Pv = v_1$. This is the case with nested categories.*

Recall from (2) that

$$D'D\hat{\alpha} = D'R \tag{5}$$

where $R = Y - X\hat{\beta}$ are the residuals for the original system with dummies omitted.

$D$ is typically quite sparse, and so is $D'D$, thus with a good sparse solver it ought to be possible to solve for $\hat{\alpha}$. Moreover, by Theorem 4.2, this system may in some cases be split into smaller systems.

In the special case with a single category ($e = 1$), the columns of $D$ are orthogonal, thus $D'D$ is diagonal; and $\hat{\alpha}$ is simply the group means of the residuals $R$. This is the within-groups estimator.

The residuals of the full system $Y = X\beta + D\alpha + \epsilon$ are easily shown to be $Y - (X\hat{\beta} + D\hat{\alpha}) = PY - PX\hat{\beta}$, the residuals of the centered system.

## 3. Identification in the case $e = 2$

Above, we assumed that $D$, and thus $D'D$ was full rank. If we construct it from dummies (with no references) we know that it is not.

Abowd et al Abowd et al. (1999) has analyzed this in the case where there are two dummy-groups (firms and individuals). In this case they construct an undirected graph $G$ where each vertex consists of a firm or an employee. A firm and an employee are adjacent

if and only if the employee has worked for the firm. There are no more edges in the graph. (I.e. $D'$ (with duplicate rows omitted) is the incidence matrix of the graph).

They then analyze identifiability in terms of the connection components of $G$ and show that it is sufficient to have a reference dummy in each of the connection components, see (Abowd et al., 2002, Appendix 1). They prove the theorem

THEOREM 3.1 (ABOWD ET AL). *If $e = 2$, the rank deficiency of $D$ equals the number of connection components of the graph $G$*

PROOF. We provide a different proof of this fact.

The matrix $D'$ may be viewed as the incidence matrix of a multigraph, then $D'D$ is the *signless Laplacian* of this graph. Moreover, the graph is bipartite (firms in one partition, employees in another). By (Cvetković et al., 2007, Corollary 2.2), the multiplicity of eigenvalue 0 is the number of components.

Since the above reference does not cover multigraphs explicitly, for convenience, the details are as follows, adapted from the proof of (Brouwer and Haemers, 2009, Proposition 1.4.2). Reorder the columns and rows of $D'D$ by connection component. The resulting matrix is a block-diagonal matrix. Its spectrum is the union (with multiplicities) of the spectra of the blocks (since the determinant of a diagonal block-matrix is the product of the determinants of the blocks). Thus, it is sufficient to show the result for connected graphs.

So, assume $G$ is connected. By construction from a complete set of dummy-variables with no references, we know that the rank-deficiency of $D'D$ is positive. We must show that it is 1.

Let $v = (v_1, v_2, \ldots, v_g)$ (indexed by the vertices of the graph) be a vector in the null-space of $D$, i.e. $Dv = 0$. This means that whenever vertices $i$ and $j$ are adjacent (so that a row in $D$ has a 1 in both column $i$ and $j$), then $v_i + v_j = 0$. Now, start out with a value for $v_1$, we may find $v_j = -v_1$ for all its adjacent vertices. We may continue this process for each of *them*; since the graph is connected we will eventually reach every $v_i$ in this way, thus the entire vector $v$ is determined by its first coordinate $v_1$, which proves that the rank-deficiency of $D$ (and thus $D'D$) is at most 1.

REMARK 3.2. *For $e = 3$, the graph is tripartite, and in general $e$-partite. We are not aware of similar general results for multipartite graphs. (Though there's more structure than that in our graphs, so it may still be possible.)*

*In this case, one may perform a pivoted Cholesky decomposition of $D'D + \epsilon I$ for some small $\epsilon$ and look for small pivots. Each of these corresponds to a column of $D'D$ which participates in a linear dependence. But the identification problem (thus, the interpretation of the resulting coefficients) is somewhat elusive.*

## 4. A commuting decomposition of $P$

The algorithm in Theorem 2.1 may take some time to finish. But if we can find commuting projections which has $P$ as their intersection, we know by Remark 2.4 that only one iteration is necessary.

To this end, assume the columns of $D$ are ordered by connection components. This makes $D'D$ block-diagonal. In other words, $D = \begin{bmatrix} B_1 & B_2 & \cdots & B_c \end{bmatrix}$ where the columns of each $B_i$ consists of a single connection component. We have $B_i'B_j = 0$ for every pair $i, j$ with $i \neq j$. That is, every column of $B_i$ is orthogonal to every column of $B_j$.

For $i = 1..c$, let $Q_i = I - B_i(B_i'B_i)^{-1}B_i'$ be the projection onto the complement of the column space of $B_i$.

Since the column spaces of $B_i$ and $B_j$ are orthogonal whenever $i \neq j$, we have $(I - Q_i)(I - Q_j) = 0$. This product may be expanded to yield $I - Q_i - Q_j + Q_iQ_j = 0$, thus $Q_iQ_j = Q_i + Q_j - I = Q_jQ_i$ So, the $Q_i$'s commute for $i \neq j$. Moreover, it is clear that $P = Q_1 \wedge Q_2 \wedge \cdots \wedge Q_c$.

By Remark 2.4, we have

$$P = Q_iQ_2 \cdots Q_c.$$

However, we have found no particularly easy way to compute $Q_iv$ for a vector $v$. We may partition each $B_i = \begin{bmatrix} B_{i1} & B_{i2} & \cdots & B_{ie} \end{bmatrix}$ into subsets of columns corresponding to single categories, let $Q_{ij}$ be the projection onto the $R(B_{ij})^\perp$ and compute $Q_iv$ by applying Algorithm 2.1 to the projections $\{Q_{ij}\}_{j=1}^e$.

Thus, we may use the component decomposition to split the problem of Algorithm 2.1 into several smaller problems.

REMARK 4.1. *By construction, for every $i, j$ we have $R(P_j)^\perp \subset R(Q_{ij})^\perp$. This is because for every $j = 1..e$ all the columns of $B_{ij}$ are columns of $D_j$. With a large $c$, there will be many runs of Algorithm 2.1, but since the $Q_{ij}$'s are in a sense smaller it may happen that the algorithm converges faster (both in terms of number of iterations and time for each iteration.).*

The rate of convergence has been analyzed in Deutsch and Hundal (1997).

For the case $e = 2$, Aronszajn, cited in (Deutsch and Hundal, 1997, Corollary 2.9), has an estimate

$$\|(P_1P_2)^n - P\| \leq \cos^{2n-1}(R(D_1)^\perp, R(D_2)^\perp)$$

where the function cos denotes the cosine of the (complementary) angle between subspaces. This was later shown by Kayalar and Weinert (Kayalar and Weinert, 1988, Theorem 2) to be an equality. This quantity is strictly smaller than 1 in finite dimensional spaces (Deutsch and Hundal, 1997, Lemma 2.3(3)). Thus, we have geometric convergence, but it may still be very slow. Moreover, the convergence is monotonous.

We have not succeeded in comparing the convergence rate of the $\{Q_{ij}\}_{j=1}^e$'s above to that of he $P_i$'s, so the status of Remark 4.1 remains open. Faster convergence could also possibly be achieved by the methods of Gearhart and Koshy (1989) or Salomon and Ur (2006).

The case $e > 2$ is also handled in Deutsch and Hundal (1997), but is more complicated.

We may also use this decomposition to simplify the system (5).

THEOREM 4.2. *With the above decomposition and reordering, we may write $D'D$ as a block matrix $\begin{bmatrix} B_i'B_j \end{bmatrix}$. By the above, it is a block diagonal matrix with $B_i'B_i$ on the diagonal.*

*Thus, our system* (5) *splits into $c$ separate systems*

$$B_i'B_i\hat{\alpha}_i = B_i'R \tag{6}$$

*where $\hat{\alpha}_i$ is the part of $\hat{\alpha}$ in component $i$.*

Thus, we may find the fixed effects for each connection component separately.

REMARK 4.3. *This matrix is the block-diagonal version of the matrix in (Abowd et al., 2002, Equation (4)) (i.e. with the covariates $X$ removed).*

## 5. Summary

Let us summarize this discussion. Let

$$Y = X\beta + D\alpha + \epsilon \tag{7}$$

be a linear system where $D$ is the dummies for a set of category variables. I.e. $D = \begin{bmatrix} D_1 & D_2 & \cdots & D_e \end{bmatrix}$, where each $D_i$ is the matrix of dummies for a single category.

Let $P_i$, for $i = 1..e$ be the operation of subtracting group means for category variable $i$. Formally, $P_i = I - D_i(D_i'D_i)^{-1}D_i'$, though, one doesn't need to find the matrices $P_i$ to perform this operation on a vector.

Let $P = I - D(D'D)^{-1}D'$ be the projection on the orthogonal complement of the range of $D$.

ALGORITHM 5.1. *To find OLS estimates $\hat{\beta}$ and $\hat{\alpha}$ for model (7) we proceed in the following manner.*

*(1) Compute $\bar{Y} = PY$ and $\bar{X} = PX$ according to Algorithm 2.1*

*(2) Perform an OLS on*

$$\bar{Y} = \bar{X}\beta + \epsilon.$$

*The result of this estimation is $\hat{\beta}$. The residuals $\bar{Y} - \bar{X}\hat{\beta}$ are the residuals of an OLS on (7). The standard errors need to be adjusted, taking into account the number of eliminated parameters in the degrees of freedom. (one for each connection component)*

*(3) Compute $R = Y - X\hat{\beta}$*

*(4) In case $e = 2$, find the $c$ connection components of the graph discussed in section 3. Reorder columns and partition $D = \begin{bmatrix} B_1 & B_2 & \cdots & B_c \end{bmatrix}$ into connection components. For each $i = 1..c$, find the column with the highest column sum in $B_i$. Remove the column. Solve $\hat{\alpha}_i$ from*

$$(B_i'B_i)\hat{\alpha}_i = B_i'R.$$

*with a sparse solver. $\hat{\alpha}_i$ is the estimate of the fixed effects for component $i$. Insert a 0 as the coordinate of $\hat{\alpha}_i$ corresponding to the removed column.*

*(5) In case $e > 2$, we may find the fixed effects, even though identification isn't fully understood. Do a pivoted Cholesky decomposition of $D'D + \epsilon I$ for some small $\epsilon > 0$. The small pivots correspond to row/columns of $D'D$ which participate in linear dependencies. Remove these row/columns from $D'D$. Solve $(D'D)\hat{\alpha} = R'Y$ and put back zeros for the removed entries.*

In principle, we could, in step (4), for each component, find the first fixed effect separately by sweeping out the others, then repeat the process with the remaining $c - 1$ fixed effects, and the residuals. However, keeping dummies for the first fixed effect and sweeping out the others will destroy the sparsity of the system, so we do not find such a procedure attractive.

## References

Abowd, J., R. Creecy, and F. Kramarz (2002). Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data. Technical Report TP-2002-06, U.S. Census Bureau.

Abowd, J., F. Kramarz, and D. Margolis (1999, March). High Wage Workers and High Wage Firms. *Econometrica 67*(2), 251–333.

Andrews, M., L. Gill, T. Schank, and R. Upward (2008). Hige wage workers and low wage firms: negative assortative matching or limited mobility bias? *J.R. Stat. Soc.(A) 171(3)*, 673–697.

Brouwer, A. and W. Haemers (2009?). *Spectra of Graphs*. `http://homepages.cwi.nl/~aeb/math/ipm.pdf`.

Carneiro, A., P. Guimaraes, and P. Portugal (2009, May). Real Wages and the Business Cycle: Accounting for Worker and Firm Heterogeneity. IZA Discussion Papers 4174, Institute for the Study of Labor (IZA).

Cornelissen, T. and O. Hubler (2007). Unobserved individual and firm heterogeneity in wage and tenure function: evidence from German linked employer-employee data. *IZA Discussion Papera* (2741).

Cvetković, D., P. Rowlinson, and S. Simić (2007). Signless Laplacians of finite graphs. *Linear Algebra and its applications 423*, 155–171.

Deutsch, F. and H. Hundal (1997). The Rate of Convergence for the Method of Alternating Projections, II. *J. Math. Anal. App. 205*(2), 381–405.

Gearhart, W. B. and M. Koshy (1989). Acceleration schemes for the method of alternating projections. *Journal of Computational and Applied Mathematics 26*(3), 235–249.

Halperin, I. (1962). The Product of Projection Operators. *Acta Sci. Math. (Szeged) 23*, 96–99.

Kaczmarz, A. (1937). Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres 35*, 355–357.

Kayalar, S. and H. Weinert (1988). Error Bounds for the Method of Alternating Projections. *Math. Control Signals Systems 1*, 43–59.

Salomon, B. and H. Ur (2006). Accelerating the Convergence of the von Neumann-Halperin Method of Alternating Projections. In *Digital Signal Processing Workshop, 12th - Signal Processing Education Workshop, 4th*, pp. 328 –331.

von Neumann, J. (1949). On Rings of Operators. Reduction Theory. *Ann. Math. 50*, 401–485.

von Neumann, J. (1950). Functional Operators. Vol II. In *Ann. Math. Stud.*, Volume 22, Chapter The Geometry of Orthogonal Spaces. Princeton Univ. Press.