

# Methods of Economic Investigation II

## Lecture 1

Oliver Linton

October 4, 2002

## 0.1 Linear Model

- The first part of the course will be concerned with estimating and testing in the linear model. We will suggest procedures and derive their properties under certain assumptions. The linear model is the basis for most of econometrics and a firm grounding in this theory is essential for future work.
- We observe the following data

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} ;$$
$$X = \begin{pmatrix} x_{11} & \cdots & x_{K1} \\ \vdots & & \vdots \\ x_{1n} & & x_{Kn} \end{pmatrix} = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix},$$

where  $\text{rank}(X) = K$ . Note that this is an assumption, but it is immediately verifiable from the data in contrast to some other assumptions we will make.

- It is desirable for statistical analysis to specify a model of how these data were generated. We suppose that there is a random mechanism which is behind everything - the data we have is one realisation of an infinity of such potential outcomes. We shall make the following assumptions regarding the way  $y$ ,  $X$  were generated:

- Fixed Design Linear Model

- (A1)  $X$  is fixed in repeated samples
- (A2)  $\exists \beta = (\beta_1, \dots, \beta_K)'$  such that  $E(y) = X\beta$ .
- (A3)  $\text{Var}(y) = \sigma^2 I_{n \times n}$ .

- We stick with fixed design for most of the linear regression section. Fixed design is perhaps unconvincing for most economic data sets, because of the asymmetry between  $y$  and  $x$ . That is, in economic datasets we have no reason to think that some data were randomly generated while others were fixed. This is especially so in time series when one regressor might be a lagged value of the dependent variable.
  
- A slightly different specification is Random Design Linear Model
  - (A1r)  $X$  is random with respect to repeated samples
  - (A2r)  $\exists \beta$  s.t.  $E(y|X) = X\beta$
  - (A3r)  $\text{Var}(y|X) = \sigma^2 I_{n \times n}$ ,

where formally A2r and A3r hold with probability one.

- However, one can believe in a random design model, but want to conduct inference in the conditional distribution [given  $X$ ]. This is sensible at least in the cross-section case where there are no lagged dependent variables. In this case, we are effectively working in a fixed design model. So the real distinction in this case is whether one evaluates quantities in the conditional or unconditional distribution.
- Finally, we write the regression model in the more familiar form. Define  $\varepsilon = y - X\beta = (\varepsilon_1, \dots, \varepsilon_n)'$ , then

$$y = X\beta + \varepsilon,$$

where [in the fixed design]

$$\begin{aligned} E(\varepsilon) &= \mathbf{0} \\ E(\varepsilon\varepsilon') &= \sigma^2 I_n. \end{aligned}$$

The linear regression model is more commonly stated like this with statistical assumptions made about

the unobservable  $\varepsilon$  rather than directly on the observable  $y$ . The assumptions about the vector  $\varepsilon$  are quite weak in some respects - the observations need not be independent and identically distributed, since only the first two moments of the vector are specified - but strong in regard to the second moments themselves.

- It is worth discussing here some alternative assumptions made about the error terms. For this purpose we shall assume a random design, and moreover suppose that  $(x_i, \varepsilon_i)$  are i.i.d. In this case, we can further assume that

- $E(\varepsilon_i x_i) = 0$

- $E(\varepsilon_i | x_i) = 0$ , denoted  $\varepsilon_i \perp x_i$

- $\varepsilon_i$  are i.i.d. and independent of  $x_i$ , denoted  $\varepsilon_i \perp\!\!\!\perp x_i$ .

- $\varepsilon_i \sim N(0, \sigma^2)$ .

- The first assumption, called an unconditional moment condition, is the weakest assumption needed to ‘identify’ the parameter  $\beta$ .
- The second assumption, called a conditional moment restriction, is a little bit stronger. It is really just a rewriting of the definition of conditional expectation.
- The third assumption is much stronger and is not strictly necessary for estimation purposes although it does have implications about efficiency and choice of estimator.
- The fourth assumption we will sometimes make in connection with hypothesis testing and for establishing optimality of least squares.

## 0.2 The OLS Procedure

- In practice we don't know the parameter  $\beta$  and seek to estimate it from the data.

- For any  $b$ , define  $Xb$  and

$$u(b) = y - Xb.$$

Then  $u(b)$  is the vector of discrepancies between the observed  $y$  from the predicted by  $Xb$ .

- The Ordinary Least Squares (OLS) procedure chooses  $\hat{\beta}$  to minimize the quadratic form

$$S(b) = u(b)'u(b) = \sum_{i=1}^n u_i^2(b) = (y - Xb)'(y - Xb)$$

with respect to  $b \in R^k$ . This is perhaps the main estimator of  $\beta$ , and we shall study its properties at length.

- The first question is whether a minimum exists. Since the criterion is a continuous function of  $b$ , a minimum over any compact subset always exists.
- A necessary condition for the uniqueness of a solution is that  $n \geq K$ . If  $n = K$ , the solution essentially involves interpolating the data, i.e., the fitted value of  $y$  will be equal to the actual value.
- When the assumption that  $\text{rank}(X) = K$  is made,  $\hat{\beta}$  is uniquely defined for any  $y$  and  $X$  independently of model; so there is no need for assumptions A1-A3 when it comes to computing the estimator.
- We now give two derivations of the well-known result that

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

- We suppose the answer is given by this formula and demonstrate that  $\hat{\beta}$  minimizes  $S(b)$  with respect to  $b$ . Write

$$u(b) = y - X\hat{\beta} + X\hat{\beta} - Xb,$$

so that

$$\begin{aligned} S(b) &= (y - X\hat{\beta} + X\hat{\beta} - Xb)'(y - X\hat{\beta} + X\hat{\beta} - Xb) \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &\quad + (\hat{\beta} - b)'X'X(\hat{\beta} - b) \\ &\quad + (y - X\hat{\beta})'X(\hat{\beta} - b) + (\hat{\beta} - b)'X'(y - X\hat{\beta}) \\ &= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - b)'X'X(\hat{\beta} - b), \end{aligned}$$

because

$$X'(y - X\hat{\beta}) = X'y - X'X\hat{\beta} = 0.$$

But

$$(\hat{\beta} - b)'X'X(\hat{\beta} - b) \geq 0,$$

and equality holds only when  $b = \hat{\beta}$ .

- A minimizer of  $S(b)$  must satisfy the vector of first order conditions:

$$\frac{\partial S}{\partial b} = 2X'(y - X\hat{\beta}) = 0.$$

Therefore,

$$X'y = X'X\hat{\beta}.$$

Now we use the assumption that  $X$  is of full rank. This ensures that  $X'X$  is invertible, and

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

as required. To verify that we have found a local minimum rather than maximum it is necessary to calculate the second derivatives

$$\frac{\partial^2 S}{\partial b \partial b'} = 2X'X > 0.$$

- The vector derivatives follow by straightforward calculus

$$\frac{\partial}{\partial b_j} \sum_{i=1}^n u_i(b)^2 = 2 \sum_{i=1}^n u_i(b) \frac{\partial u_i}{\partial b_j}$$

$$= -2 \sum_{i=1}^n u_i(b) x_{ij} ,$$

since

$$\frac{\partial u_i}{\partial b_j} = -x_{ij}.$$

- Characterization of the Solution. Define the fitted value  $\hat{y} = X\hat{\beta}$  and the OLS residuals

$$\hat{u} = y - \hat{y} = y - X\hat{\beta}.$$

- The OLSE  $\hat{\beta}$  solves the normal equations  $X'\hat{u} = 0$ , i.e.,

$$\begin{aligned} \sum_{i=1}^n x_{1i} \hat{u}_i &= 0 \\ \sum_{i=1}^n x_{2i} \hat{u}_i &= 0 \\ &\vdots \\ \sum_{i=1}^n x_{Ki} \hat{u}_i &= 0. \end{aligned}$$

- We say that  $X$  is orthogonal to  $\hat{u}$ , denoted  $X \perp \hat{u}$ . Note that if, as usual  $X_{1i} = \mathbf{1}$ , then, we have  $\sum_{i=1}^n \hat{u}_i = 0$ .

## 0.3 Some Alternative Estimation Paradigms

- We briefly mention some alternative estimation methods which actually lead to the same estimator as the OLS estimator in some special cases, but which are more broadly applicable.
- Maximum Likelihood. Suppose we also assume that  $y \sim N(X\beta, \sigma^2 I)$ . Then the density function of  $y$  [conditional on  $X$ ] is

$$f_{y|X}(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2}(y - X\beta)'(y - X\beta)\right).$$

- The density function depends on the unknown parameters  $\beta, \sigma^2$ , which we want to estimate. We therefore switch the emphasis and call the following quantity the log likelihood function for the observed data

$$\begin{aligned} \ell(b, \omega^2 | y, X) = & -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \omega^2 \\ & - \frac{1}{2\omega^2} (y - Xb)'(y - Xb), \end{aligned}$$

where  $b$  and  $\omega$  are unknown parameters.

- The maximum likelihood estimator  $\hat{\beta}_{mle}, \hat{\sigma}_{mle}^2$  maximizes  $\ell(b, \omega^2)$  with respect to  $b$  and  $\omega^2$ . It is easy to see that

$$\hat{\beta}_{mle} = \hat{\beta}$$

$$\hat{\sigma}_{mle}^2 = \frac{1}{n}(y - X\hat{\beta}_{mle})'(y - X\hat{\beta}_{mle}).$$

Basically, the criterion function is the least squares criterion apart from an affine transformation involving only  $\omega$ .

- Note however, that if we had a different assumption about the errors than A4, e.g., they were from a t-distribution, then we would have a different likelihood and a different estimator than  $\hat{\beta}$ . In particular, the estimator may not be explicitly defined and may be a nonlinear function of  $y$ .

- Method of Moments. Suppose that we define parameters through some population moment conditions; this can arise from an economic optimization problem, see below.
- For example, suppose that we say that  $\beta$  is defined as the unique parameter that satisfies the  $K$  moment conditions [we need as many moment conditions as parameters]

$$E[x_i(y_i - x_i'\beta)] = 0.$$

Note that this is the natural consequence of our assumption that  $E(\varepsilon_i x_i) = 0$ .

- Replacing the population by the sample average we must find  $b$  such that

$$\frac{1}{n} \sum_{i=1}^n x_i(y_i - x_i'b) = 0.$$

The solution to this is of course

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

i.e., the MOM estimator is equal to OLS in this case. Thus, for the moment conditions above we are lead to the least squares estimator.

- However, if we chose some other conditions, then a different estimator results. For example, suppose that we assume that

$$E[x_i(y_i - x_i'\beta)^3] = 0,$$

we would be lead to a different estimator - any solution of

$$\frac{1}{n} \sum_{i=1}^n x_i(y_i - x_i'b)^3 = 0.$$

In general, this would be more complicated to analyze.

- We emphasize here that the above estimation methods are all suggested or motivated by our assumptions, but of course we can always carry out the procedure without regard to underlying model - that is, the procedures only require data, not assumptions.

## 0.4 Geometry of OLS and Partitioned Regression

- We want to give a geometric interpretation to the OLS procedure.
- The data:  $y, x_1, \dots, x_K$ , can all be viewed as elements of the vector space  $R^n$ . Define the set

$$\begin{aligned}\mathcal{C}(X) &= \{\alpha_1 x_1 + \dots + \alpha_K x_K\} \\ &= \{X\alpha : \alpha \in R^K\} \subseteq R^n,\end{aligned}$$

otherwise known as the column span of  $X$ .

- Then,  $\mathcal{C}(X)$  is a linear subspace of  $R^n$  of dimension  $K$  assuming that the matrix  $X$  is of full rank. If it is only of rank  $K^*$  with  $K^* < K$  then  $\mathcal{C}(X)$  is still a linear subspace of  $R^n$  but of dimension  $K^*$ .

- The OLS procedure can equivalently be defined as finding the point in  $\mathcal{C}(X)$  closest to  $y$ , where closeness is measured in terms of Euclidean distance, i.e.,

$$d(y, Xb) = \|y - Xb\|^2 = (y - Xb)'(y - Xb)$$

is the Euclidean distance of  $y$  to the point  $Xb \in \mathcal{C}(X)$ .

- This is an old problem in geometry, which is now given a key role in abstract mathematics.
- The projection theorem [Hilbert] says that there is a unique solution to the minimization problem, call it  $\hat{y}$ , which is characterized by the fact that

$$\hat{u} = y - \hat{y}$$

is orthogonal to  $\mathcal{C}(X)$ .

- Equivalently we can write uniquely

$$y = \hat{y} + \hat{u},$$

where  $\hat{y} \in \mathcal{C}(X)$  and  $\hat{u} \in \mathcal{C}^\perp(X)$  [the space  $\mathcal{C}^\perp(X)$  is called the orthocomplement of  $\mathcal{C}(X)$ , and consists of all vectors orthogonal to  $\mathcal{C}(X)$ ]. Essentially, one is dropping a perpendicular, and the procedure should be familiar from high school geometry.

- For example, let  $n = 3$  and

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then  $\mathcal{C}(X)$  is the set of all vectors in  $R^3$  with third component zero. What is the closest point for the example above with  $y = (1, 1, 1)'$ ? This is  $(1, 1, 0)' = X\hat{\beta}$ ,  $\hat{u} = (0, 0, 1)'$ . In fact  $\hat{u}$  is orthogonal to  $\mathcal{C}(X)$ , i.e.,  $\hat{u} \in \mathcal{C}^\perp(X) = (0, 0, \alpha)'$ .

- In general how do we find  $\hat{y}$ ? When  $X$  is of full rank we can give a simple explicit solution

$$\hat{y} = P_X y,$$

where the Projector matrix

$$P_X = X(X'X)^{-1}X'$$

projects onto  $\mathcal{C}(X)$ .

- Let  $\hat{u} = y - \hat{y} = M_X y$ , where the Projector matrix

$$M_X = I - X(X'X)^{-1}X'$$

projects onto  $\mathcal{C}^\perp(X)$ . Thus for any  $y$ , we can write

$$y = \hat{y} + \hat{u} = P_X y + M_X y.$$

The matrices  $P_X$  and  $M_X$  are symmetric and idempotent, i.e.,

$$P_X = P_X' \text{ and } P_X^2 = P_X.$$

After applying  $P_X$  once you are ready in  $\mathcal{C}(X)$ . This implies that

$$P_X X = X \text{ and } M_X X = 0,$$

so that

$$P_X M_X y = 0 \text{ for all } y.$$

- Since  $\hat{y} \in \mathcal{C}(X)$ , we can rewrite it as  $\hat{y} = X\hat{\beta}$ , so that  $\hat{\beta} = (X'X)^{-1}X'y$ .
- The space  $\mathcal{C}(X)$  is invariant to nonsingular linear transforms

$$X \longmapsto XA_{K \times K}, \text{ where } \det A \neq 0.$$

Let  $z \in \mathcal{C}(X)$ . Then there exists an  $\alpha \in R^K$  such that  $z = X\alpha$ . Therefore,

$$z = XAA^{-1}\alpha = XA\gamma,$$

where  $\gamma = A^{-1}\alpha \in R^K$ , and vice-versa.

- Since  $\mathcal{C}(X)$  is invariant to linear transformations, so are  $\hat{y}$  and  $\hat{u}$  (but not  $\hat{\beta}$ ). For example, rescaling

of the components of  $X$  does not affect the values of  $\hat{y}$  and  $\hat{u}$ .

$$y \text{ on } (x_1, x_2, x_3)(1)$$

$$y \text{ on } (x_1 + x_2, 2x_2 - x_3, 3x_1 - 2x_2 + 5x_3)(2)$$

in which case the transformation is

$$A = \begin{pmatrix} 1 & 0 & 3 \\ 1 & 2 & -2 \\ 0 & -1 & 5 \end{pmatrix},$$

which is of full rank. Therefore, (1) and (2) yield the same  $\hat{y}$ ,  $\hat{u}$ .

- Emphasizing  $\mathcal{C}(X)$  rather than  $X$  itself is called the coordinate free approach. Some aspects of model/estimate are properties of  $\mathcal{C}(X)$  choice of coordinates is irrelevant.
- When  $X$  is not of full rank

- the space  $\mathcal{C}(X)$  is still well defined, as is the projection from  $y$  onto  $\mathcal{C}(X)$ .
  - The fitted value  $\hat{y}$  and residual  $\hat{u}$  are uniquely defined in this case,
  - but there is no unique coefficient vector  $\hat{\beta}$ .
  - This is the case commonly called multicollinearity.
- We next consider an important application of the projection idea. Partition

$$X = (X_{1n \times K_1}, X_{2n \times k_2}), \quad K_1 + K_2 = K,$$

and suppose we are interested in obtaining the coefficient  $\hat{\beta}_1$  in the projection of  $y$  onto  $\mathcal{C}(X)$ .

- A key property of projection is that if  $X_1$  and  $X_2$  are orthogonal, i.e., if  $X_1'X_2 = 0$ , then

$$P_X = P_{X_1} + P_{X_2}.$$

This can be verified algebraically, but also should be obvious geometrically. In this case, write

$$\hat{y} = X\hat{\beta} = P_X y = P_{X_1} y + P_{X_2} y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2.$$

This just says that if  $X_1$  and  $X_2$  were orthogonal, then we could get  $\hat{\beta}_1$  by regressing  $y$  on  $X_1$  only, and  $\hat{\beta}_2$  by regressing  $y$  on  $X_2$  only.

- Very rarely are  $X_1$  and  $X_2$  orthogonal, but we can construct equivalent regressors that are orthogonal. Suppose we have general  $X_1$  and  $X_2$ , whose dimensions satisfy  $K_1 + K_2 = K$ . We make the following observations:
  - $(X_1, X_2)$  and  $(M_2 X_1, X_2)$  span the same space. This follows because

$$X_1 = M_2 X_1 + P_2 X_1,$$

where  $\mathcal{C}(P_2 X_1) \subset \mathcal{C}(X_2)$ . Therefore,

$$\mathcal{C}(M_2 X_1, X_2) = \mathcal{C}(X_1, X_2).$$

- $M_2X_1$  and  $X_2$  are orthogonal.
- This says that if we regress  $y$  on  $(X_1, X_2)$  or  $y$  on  $(M_2X_1, X_2)$  we get the same  $\hat{y}$  and  $\hat{u}$ , and that if we wanted the coefficients on  $M_2X_1$  from the second regression we could in fact just regress  $y$  on  $M_2X_1$  only.
- What are the coefficients on  $M_2X_1$ ? Recall that

$$\begin{aligned}
 \hat{y} &= X_1\hat{\beta}_1 + X_2\hat{\beta}_2 \\
 &= (M_2 + P_2)X_1\hat{\beta}_1 + X_2\hat{\beta}_2 \\
 &= M_2X_1\hat{\beta}_1 + X_2[\hat{\beta}_2 + (X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1] \\
 &= M_2X_1\hat{\beta}_1 + X_2\hat{C},
 \end{aligned}$$

where

$$\hat{C} = \hat{\beta}_2 + (X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1.$$

- So the coefficient on  $M_2X_1$  is the original  $\hat{\beta}_1$ , while that on  $X_2$  is some combination of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ . Note

that  $M_2X_1$  are the residuals from a regression of  $X_1$  on  $X_2$ .

- Practical Implication. If  $K$  is large and primarily interested in first  $K_1$  variables, then we can get  $\hat{\beta}_1$  by regressing  $y$  [or  $M_2y$  equivalently] on  $M_2X_1$  only, i.e.,

$$\begin{aligned}\hat{\beta}_1 &= (X_1' M_2 X_1)^{-1} X_1' M_2 y \\ &= (X_1' M_2 M_2 X_1)^{-1} X_1' M_2 M_2 y.\end{aligned}$$

This involves inversion of only  $K_1 \times K_1$  and  $K_2 \times K_2$  matrices, which involves less computing time than inverting  $K \times K$  matrices, especially when  $K$  is large [this computation can be as bad as  $O(K^3)$ ].

- Suppose that  $X_2 = (1, 1, \dots, 1)'$  =  $i$ , then

$$M_2 = I_n - i(i'i)^{-1}i' = I_n - \frac{ii'}{n}$$

and

$$M_2 x_{1_{n \times 1}} = x_1 - \frac{1}{n} \sum_{i=1}^n x_{1i} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_{1i} - \bar{x}_1 \\ \vdots \\ x_{1n} - \bar{x}_1 \end{pmatrix}.$$

When regression includes an intercept, can first demean the  $X$  variables (and the  $y$ 's) then do regression on the demeaned variables.

## 0.5 Goodness of Fit

- How well does the model explain the data? One possibility is to measure the fit by the residual sum of squares

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

In general, the smaller the  $RSS$  the better. However, the numerical value of  $RSS$  depends on the units used to measure  $y$  in so that one cannot compare across models.

- Generally used measure of goodness of fit is the  $R^2$ . In actuality, there are three alternative definitions in general.
  - One minus the ratio of the residual sum of squares to total sum of squares,

$$R_1^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

- The sample correlation squared between  $y$  and  $\hat{y}$ ,

$$R_2^2 = \frac{[\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}.$$

- The ratio of explained sum of squares to total sum of squares

$$R_3^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Here,  $\bar{y} = \sum_{i=1}^n y_i/n$  and  $\bar{\hat{y}} = \sum_{i=1}^n \hat{y}_i/n$ .

- Theorem. When an intercept is included, all three measures are the same.
- Proof of  $R_1^2 = R_2^2$ . Since an intercept is included, we have

$$\sum_{i=1}^n \hat{u}_i = 0,$$

which implies that  $\overline{\hat{y}} = \bar{y}$ . Therefore,

$$\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \overline{\hat{y}}) = \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

because

$$\sum_{i=1}^n \hat{u}_i(\hat{y}_i - \bar{y}) = 0.$$

- Proof of  $R_1^2 = R_3^2$ . Similarly,

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

■

- If an intercept is included, then  $0 \leq R^2 \leq 1$ . If not, then  $0 \leq R_2^2 \leq 1$ , but  $R_3^2$  could be greater than one, and  $R_1^2$  could be less than zero.
- If  $y = \alpha + \beta x + u$ , then  $R^2$  is the squared sample correlation between  $y$  and  $x$ .

- The  $R^2$  is invariant to some changes of units.
- If  $y \mapsto ay + b$  for any constants  $a, b$ , then
  - $\hat{y}_i \mapsto a\hat{y}_i + b$  and
  - $\bar{y} \mapsto a\bar{y} + b$ ,
  - so  $R^2$  is the same in this case.
  - Clearly, if  $X \mapsto XA$  for a nonsingular matrix  $A$ , then  $\hat{y}$  is unchanged, as is  $y$  and  $\bar{y}$ .
- $R^2$  always increases with addition of variables. With  $K = n$  we can make  $R^2 = 1$ .
- Theil's adjusted  $R^2$  is defined as follows

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2).$$

This amounts to dividing the sum of squares by the appropriate degrees of freedom, so that

$$1 - \bar{R}^2 = \frac{\frac{1}{n-K} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

It follows that

$$\frac{\Delta \bar{R}^2}{\Delta K} = \underbrace{\frac{n-1}{n-K}}_{+} \frac{\Delta R^2}{\Delta K} - \underbrace{\frac{n-1}{(n-K)^2}}_{-} (1 - R^2).$$

This measure allows some trade-off between fit and parsimony.