Econométrie TP - Solution Série 02

Giuseppe Gruttad'Auria
Institut de management, Université de Neuchâtel
giuseppe.gruttadauria@unine.ch

SA 2024

Exercice 2.1

On considère le modèle de régression suivant :

$$y = \beta_1 x_1 + \cdots + \beta_K x_K + u$$

où $x_1 = 1$ (le terme d'intercept), et u est un terme d'erreur aléatoire. Soit $\hat{\beta}$ l'estimateur des MCO (moindres carrés ordinaires) pour β, et soit r le vecteur des résidus défini par :

$$r := y - \hat{y} = y - X\hat{\beta}$$

1 : X' r = 0

Nous devons montrer que $X'r = 0$ et interpréter ce résultat.

Le modèle peut être écrit sous forme matricielle comme suit :

$$y = X\beta + u$$

où :

- y est le vecteur $n \times 1$ des observations de la variable dépendante,
- X est la matrice $n \times K$ des variables explicatives,
- β est le vecteur $K \times 1$ des paramètres à estimer,
- u est le vecteur $n \times 1$ des erreurs.

L'estimateur des MCO $\hat{\beta}$ est donné par :

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

Les valeurs ajustées (ou prédites) sont :

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

Ainsi, les résidus sont :

$$r = y - \hat{y} = y - X\hat{\beta}$$

Montrer que $X'r = 0$

Substituons $r = y - X\hat{\beta}$ dans l'expression $X'r$:

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

En remplaçant $\hat{\beta}$ par son expression $\hat{\beta} = (X'X)^{-1}X'y$, on obtient :

$$X'r = X'y - X'X(X'X)^{-1}X'y$$

Puisque $X'X(X'X)^{-1} = I$ (la matrice identité), cela se simplifie ainsi :

$$X'r = X'y - X'y = 0$$

Ainsi, nous avons montré que :

$$X'r = 0$$

Interprétation

Le résultat $X'r = 0$ signifie que les résidus r sont orthogonaux (non corrélés) aux colonnes de X (l'espace des régresseurs). Autrement dit, les résidus ne sont pas corrélés avec les variables explicatives (pas influencés par les régresseurs), ce qui signifie qu'il n'y a plus d'information dans les variables explicatives qui pourrait réduire davantage les résidus. C'est une propriété fondamentale de l'estimateur des MCO, assurant que le modèle ajusté minimise la somme des résidus au carré.

2: Covariance empirique

Nous devons montrer que la covariance empirique entre les résidus r et les régressions $\mathbf{x} = (x_1, \ldots, x_K)$ est nulle, c'est-à-dire :

$$\text{Cov}(r, \mathbf{x}) = 0$$

Définition de la covariance empirique

La covariance empirique entre deux variables a et b est définie par :

$$\text{Cov}(a, b) = \frac{1}{n} a'b$$

Dans notre cas, la covariance entre les résidus r et la matrice des régressions $\mathbf{x}$ est :

$$\text{Cov}(r, \mathbf{x}) = \frac{1}{n} r'X$$

où :

- $r = y - X\hat{\beta}$ est le vecteur des résidus,
- X est la matrice des régressions, et
- $\hat{\beta}$ est l'estimateur des MCO pour les paramètres.

Utilisation de $X'r = 0$

Nous savons que :

$$X'r = 0$$

Cela signifie que le produit de la transposée de X et des résidus r est nul, ce qui implique :

$$\frac{1}{n} r'X = 0$$

Ainsi, la covariance entre les résidus r et les régressions X est nulle :

$$\text{Cov}(r, \mathbf{x}) = 0$$

Interprétation

Le fait que la covariance empirique entre les résidus et les régressions soit nulle confirme qu'il n'y a aucune relation linéaire entre les résidus et les variables explicatives (les résidus ne sont pas corrélés avec les régressions), ce qui est une propriété fondamentale de l'estimateur des MCO.

3 : SST = SSR + SSE

Nous devons montrer la décomposition suivante de la somme des carrés :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (r_i - \bar{r})^2$$

Définitions

- y_i sont les valeurs observées de la variable dépendante.
- $\hat{y}_i = X\hat{\beta}$ sont les valeurs ajustées (prédites) du modèle de régression.
- $r_i = y_i - \hat{y}_i$ sont les résidus.
- $\bar{y} = \frac{1}{n}\sum_i y_i$ est la moyenne des valeurs observées de y.
- $\bar{r} = \frac{1}{n}\sum_i r_i$, qui est nulle dans un modèle de régression MCO (puisque la somme des résidus est nulle).

Somme totale des carrés (SST)

Le membre de gauche de l'équation représente la Somme totale des carrés (SST), qui mesure la variabilité totale dans les données observées :

$$\text{SST} = \sum_i (y_i - \bar{y})^2$$

Somme des carrés expliquée (SSR)

Le premier terme du membre de droite représente la Somme des carrés expliquée (SSR), qui mesure la variabilité expliquée par le modèle de régression :

$$\text{SSR} = \sum_i (\hat{y}_i - \bar{y})^2$$

Somme des carrés des résidus (SSE)

Le deuxième terme du membre de droite représente la Somme des carrés des résidus (SSE), qui mesure la variabilité restante dans les résidus :

$$\text{SSE} = \sum_i (r_i - \bar{r})^2 = \sum_i r_i^2$$

Puisque la moyenne des résidus $\bar{r}$ est nulle (comme propriété des MCO), ce terme se simplifie à la somme des carrés des résidus $\sum_i r_i^2$.

Décomposition de la variabilité

L'équation montre que la variabilité totale dans les valeurs observées y (SST) peut être décomposée en deux composantes :

- La variabilité expliquée par le modèle de régression (SSR).
- La variabilité non expliquée, c'est-à-dire les résidus (SSE).

Ainsi, nous avons :

$$\text{SST} = \text{SSR} + \text{SSE}$$

Interprétation

Cette décomposition montre comment la variabilité totale de la variable dépendante peut être répartie entre la partie expliquée par le modèle de régression et la partie non expliquée. Cela conduit au concept du coefficient de détermination R^2, qui mesure la proportion de la variabilité totale expliquée par le modèle :

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

L'équation suivante montre la décomposition de la variabilité totale des données observées en composantes expliquée et non expliquée :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i r_i^2$$

4 : R^2

Nous devons montrer que :

$$R^2 = \frac{\hat{\beta}_1 \text{Cov}(y, x_1) + \cdots + \hat{\beta}_K \text{Cov}(y, x_K)}{V(y)}$$

où R^2 est le coefficient de détermination, défini par :

$$R^2 = \frac{V(\hat{y})}{V(y)}$$

Définition de R^2

Le coefficient de détermination R^2 est défini comme le rapport entre la variance des valeurs ajustées $\hat{y}$ et la variance des valeurs observées y :

$$R^2 = \frac{V(\hat{y})}{V(y)}$$

Cette quantité mesure la proportion de la variance totale dans la variable dépendante y qui est expliquée par le modèle de régression.

Variance des valeurs ajustées $\hat{y}$

Les valeurs ajustées sont données par :

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

Substituons ceci dans la variance de $\hat{y}$:

$$V(\hat{y}) = V(X\hat{\beta})$$

Puisque $\hat{\beta}$ est un vecteur constant, la variance des valeurs ajustées est donnée par la matrice de covariance des variables explicatives X :

$$V(\hat{y}) = \hat{\beta}' \text{Cov}(X, X) \hat{\beta}$$

Covariance entre y et x_k

Nous savons que chaque $\hat{\beta}_k$ est lié à la covariance entre y et x_k par la formule de l'estimateur des MCO :

$$\hat{\beta}_k = \frac{\text{Cov}(y, x_k)}{V(x_k)}$$

Ainsi, la variance de $\hat{y}$ peut être exprimée comme une combinaison linéaire des covariances entre y et chaque x_k :

$$V(\hat{y}) = \hat{\beta}_1 \text{Cov}(y, x_1) + \cdots + \hat{\beta}_K \text{Cov}(y, x_K)$$

Expression de R^2

En utilisant l'expression de $V(\hat{y})$ dans la formule de R^2, nous obtenons :

$$R^2 = \frac{\hat{\beta}_1 \text{Cov}(y, x_1) + \cdots + \hat{\beta}_K \text{Cov}(y, x_K)}{V(y)}$$

5 : R^2 et la corrélation partielle

Nous devons montrer que :

$$(1 - R^2) = (1 - R_*^2)(1 - r^2_{yx_K \cdot x_1, \ldots, x_{K-1}})$$

où :

- R^2 est le coefficient de détermination pour le modèle complet avec $x_1, \ldots, x_K$,
- R_*^2 est le coefficient de détermination pour la régression avec seulement $x_1, \ldots, x_{K-1}$,
- $r_{yx_K \cdot x_1, \ldots, x_{K-1}}$ est le coefficient de corrélation partielle entre y et x_K, en contrôlant pour les autres variables.

Définition de la corrélation partielle

Le coefficient de corrélation partielle $r_{yx_K \cdot x_1, \ldots, x_{K-1}}$ mesure la relation linéaire entre y et x_K, en maintenant constantes les autres variables $x_1, \ldots, x_{K-1}$. Il est défini comme la corrélation entre les résidus des régressions suivantes :

1. y sur $x_1, \ldots, x_{K-1}$, donnant les résidus y_*,
2. x_K sur $x_1, \ldots, x_{K-1}$, donnant les résidus x_{K*}.

Ainsi, la corrélation partielle est définie par :

$$r_{yx_K \cdot x_1, \ldots, x_{K-1}} = \frac{\text{Cov}(y_*, x_{K*})}{\sqrt{V(y_*)V(x_{K*})}}$$

Relation entre R^2 et la corrélation partielle

Le coefficient de détermination R^2 mesure la proportion de la variance dans y expliquée par tous les régressions $x_1, \ldots, x_K$, tandis que R_*^2 mesure la proportion de la variance dans y expliquée par $x_1, \ldots, x_{K-1}$. L'ajout de la variable x_K augmente la variance expliquée, mais cette contribution supplémentaire dépend de la corrélation partielle entre y et x_K, en contrôlant pour les autres variables.

L'augmentation de la variance expliquée par l'ajout de x_K au modèle est donnée par :

$$R^2 - R_*^2 = r^2_{yx_K \cdot x_1, \ldots, x_{K-1}}(1 - R_*^2)$$

En réarrangeant cette équation, nous obtenons :

$$R^2 = R_*^2 + r^2_{yx_K \cdot x_1, \ldots, x_{K-1}}(1 - R_*^2)$$

Démonstration de l'équation

En partant de l'équation précédente :

$$R^2 = R_*^2 + r^2_{yx_K \cdot x_1, \ldots, x_{K-1}}(1 - R_*^2)$$

En soustrayant R^2 des deux côtés :

$$1 - R^2 = 1 - \left(R_*^2 + r^2_{yx_K \cdot x_1, \ldots, x_{K-1}}(1 - R_*^2)\right)$$

En simplifiant le membre de droite :

$$1 - R^2 = (1 - R_*^2) - r^2_{yx_K \cdot x_1, \ldots, x_{K-1}}(1 - R_*^2)$$

En factorisant $(1 - R_*^2)$:

$$1 - R^2 = (1 - R_*^2)(1 - r^2_{yx_K \cdot x_1, \ldots, x_{K-1}})$$

6 : Coefficients de détermination

Les trois équations de régression estimées sont les suivantes :

1. $C = -36.89 + 0.003927N$, avec $R^2 = 0.9757$ 2. $C = 7.4533 + 0.8933Y$, avec $R^2 = 0.9990$ 3. $C = -30.54 + 0.8070Y + 0.0003898N$, avec $R^2 = 0.9993$

où :

- C est la dépense totale de consommation en prix 1958,
- N est la population,
- Y est le revenu total disponible en prix 1958.

Interprétation des coefficients de détermination R^2

Le coefficient de détermination R^2 mesure la proportion de la variance dans la variable dépendante C qui est expliquée par les variables indépendantes dans le modèle.

- Dans l'**équation (1)**, $R^2 = 0.9757$, ce qui signifie que 97,57% de la variation dans C est expliquée par N seul.
- Dans l'**équation (2)**, $R^2 = 0.9990$, indiquant que 99,90% de la variation dans C est expliquée par Y seul.
- Dans l'**équation (3)**, $R^2 = 0.9993$, montrant que 99,93% de la variation dans C est expliquée par Y et N ensemble.

Comparaison entre les équations (1) et (3)

La comparaison entre l'**équation (1)** et l'**équation (3)** montre que l'ajout de la variable Y améliore considérablement le modèle. Alors que N explique 97,57% de la variation dans la consommation, l'ajout de Y augmente la variance expliquée à 99,93%. Cela indique que le revenu Y est un prédicteur très fort de la consommation, et son inclusion dans le modèle améliore considérablement le pouvoir explicatif du modèle.

Comparaison entre les équations (2) et (3)

En comparant l'**équation (2)** et l'**équation (3)**, nous constatons que R^2 augmente légèrement, passant de 0.9990 à 0.9993, lorsque N est ajouté au modèle qui inclut déjà Y. Cela suggère que la population N apporte une contribution supplémentaire très faible à l'explication de la consommation lorsque Y est déjà inclus. Ainsi, Y semble être le prédicteur principal de C, et N a un effet mineur lorsque Y est pris en compte.

Conclusion

- La population N explique une grande partie de la variation dans la consommation dans l'**équation (1)**, mais le revenu Y explique une proportion encore plus grande dans l'**équation (2)**.
- Lorsque les deux variables sont incluses dans l'**équation (3)**, le modèle explique 99,93% de la variation dans C, améliorant légèrement l'ajustement du modèle par rapport à l'**équation (2)**.
- La faible augmentation de R^2 entre l'**équation (2)** et l'**équation (3)** suggère que Y est le facteur dominant dans l'explication de la consommation, et N apporte peu d'explication supplémentaire une fois Y inclus.