MTH2302: La régression linéaire Simple

Wissem Maazoun

École Polytechnique de Montréal

Département de Mathématiques et de génie industriel

Introduction

La régression est une méthode d'analyse statistique dont l'objet est d'établir le lien (fonction) entre une variable dite dépendante, y, et k variables $x_1, x_2, \ldots, x_k$ dites indépendantes. Le but principal est de pouvoir faire des prévisions sur la variable y lorsque les variables indépendantes sont mesurées.

Contexte

On dispose de n points expérimentaux (ou nuage de points) $(x_1, y_1), \ldots, (x_n, y_n)$ sur les deux variables x et y. Lorsque le diagramme de dispersion indique une tendance linéaire, on peut supposer que le modèle est de la forme

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \text{ou encore} \quad E(Y|x) = \beta_0 + \beta_1 x,$$

où β_0 est l'ordonnée à l'origine (un paramètre), β_1 est la pente de la droite (un paramètre), x est une variable que l'on peut mesurer sans erreur, Y est la variable dépendante (une v.a), et ϵ est une erreur aléatoire telle que $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$.

MTH2302: La régression linéaire Simple 3 / 28
Contexte

Wissem Maazoun MTH2302: La régression linéaire Simple 3 / 28

Contexte (suite)

En utilisant les n points expérimentaux, le but visé est :

1. Estimer les paramètres β_0, β_1 et σ^2.
2. Vérifier si le modèle est adéquat.

Pour cela on suppose que :

1. Pour chaque valeur de x, $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$.
2. Les erreurs ϵ sont non corrélées (i.e. indépendantes), ie. "abscence d'autocorrélation des erreurs".
3. Les erreurs sont de loi normale, i.e., $\epsilon \sim N(0, \sigma^2)$.

Estimation des paramètres

La méthode d'estimation est celle des moin dres carrés ordinaires qui consiste à déterminer les valeurs de β_0 et β_1, qu'on notera $\hat{\beta}_0$ (ou b_0) et $\hat{\beta}_1$ (ou b_1), qui minimisent la somme des carrés des distances verticales $L(\beta_0, \beta_1)$ définie par

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

$$\Longrightarrow \begin{cases} \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \end{cases} \quad \text{avec}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \text{ et } S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Remarques

- On dit que la droite des moindres carrés (ou de la droite de régression) est

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{ou encore} \quad \hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x}).$$

- On note parfois b_0 et b_1 au lieu de $\hat{\beta}_0$ et $\hat{\beta}_1$. On a alors

$$\hat{y} = b_0 + b_1 x \quad \text{ou encore} \quad \hat{y} = \bar{y} + b_1 (x - \bar{x}).$$

MTH2302: La régression linéaire Simple 6 / 28
Estimation des paramètres

Wissem Maazoun MTH2302: La régression linéaire Simple 6 / 28

Exemple

Lors d'une étude sur la dureté Brinell d'un certain alliage, les données suivantes ont été obtenues sur :

- La température (x en $^\circ F/100$).
- La dureté Brinell (y en N/mm^2).

x	10,2	11,8	12,1	12,5	12,8	13,4	13,7
y	80,9	67,2	62,2	57,4	55,2	49,9	50,3

On considère le modèle d'équation $y = \beta_0 + \beta_1 x + \epsilon$.

1. Estimer β_0 et β_1 et déterminer l'équation de la droite des moindres carrés.
2. Estimer la dureté moyenne de l'alliage pour une température de $1250^\circ F$.

Propriétés des estimateurs de β_0 et β_1

$\hat{\beta}_0$ et $\hat{\beta}_1$ constituent les meilleurs estimateurs de β_0 et β_1 (parmi les estimateurs de la forme $\sum_{i=1}^{n} a_i Y_i$). Il est démontré que ces estimateurs sont sans biais et précis. On a en effet :

- $E(\hat{\beta}_0) = \beta_0$ et $V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$.
- $E(\hat{\beta}_1) = \beta_1$ et $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.

Estimation de σ^2

L'estimation de σ^2 est basée sur la somme des carrés des résidus

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

où $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$. Le calcul direct de SS_E est long surtout si n est grand. On utilise plutôt l'égalité fondamentale suivante :

$$\begin{array}{llll} & \sum_{i=1}^{n}(y_i - \bar{y})^2 & = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 & + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \Longrightarrow & S_{yy} & = SS_E & + SS_R \\ \text{d.d.l.} & n-1 & = (n-2) & +1. \end{array}$$

MTH2302: La régression linéaire Simple 9 / 28
Estimation de σ^2

Wissem Maazoun MTH2302: La régression linéaire Simple 9 / 28

Estimation de σ^2 (suite)

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$
$$SS_R = \hat{\beta}_1^2 \times S_{xx} = \frac{(S_{xy})^2}{S_{xx}}$$
$$\Longrightarrow \hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2} = \frac{S_{yy} - SS_R}{n-2}.$$

Le meilleur estimateur de σ^2 est $\hat{\sigma}^2 = MS_E$. Il est démontré que MS_E est un estimateur sans biais de σ^2.

Exemple

En utilisant les données de l'exemple précédent donner une estimation ponctuelle de σ^2.

Les tests sur les paramètres

Lorsque l'hypothèse de normalité des erreurs est valide, ie., $\epsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$, on a alors

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim N(0,1) \quad \text{et} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

En remplaçant σ^2 par MS_E, on obtient

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim T_\nu \quad \text{et} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_E}{S_{xx}}}} \sim T_\nu \text{ avec } \nu = n-2.$$

Les tests sur β_1

Pour une valeur de $\beta_{1,0}$ donnée, on peut tester les hypothèses $H_0 : \beta_1 = \beta_{1,0}$ contre $H_1 : \beta_1 \neq \beta_{1,0}$, la statistique du test est

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{MS_E}{S_{xx}}}}$$

La règle de décision (au seuil critique α) est de rejeter H_0 si $|t_0| > t_{\alpha/2;n-2}$ ou de manière équivalente si la valeur de $p - value = 2P(T > |t_0|)$ est petite

MTH2302: La régression linéaire Simple 12 / 28
Les tests sur β_1

Wissem Maazoun MTH2302: La régression linéaire Simple 12 / 28

Remarque

En particulier, lorsque $\beta_{1,0} = 0$, le test revient à vérifier si le modèle est significatif. On teste $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$, la statistique du test est

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{MS_E}{S_{xx}}}}.$$

Exemple

En utilisant les données de l'exemple précédent, tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ au seuil critique $\alpha = 0,05$.

Tableau d'analyse de la variance

Dans une analyse de régression, les résultats sur les sommes de carrés S_{yy}, SS_R et SS_E sont présentés dans un tableau appelé tableau d'analyse de la variance du modèle de régression. Dans un modèle de régression simple, ce tableau est de la forme :

Source de variation	Somme des carrés	Nombre de degrés de lib.	Moyenne des carrés	F	$p-value$
Régression (modèle)	$SS_R = \hat{\beta}_1 S_{xy}$	1	$MS_R = \frac{SS_R}{1}$	$F_0 = \frac{MS_R}{MS_E}$	$P(F \geq F_0)$
Résidus(Erreur)	$SS_E = S_{yy} - SS_R$	$n-2$	$MS_E = \frac{SS_E}{n-2}$		
Totale	$S_{yy} = \sum_{i=1}^{n} y_i^2 - n(\overline{y})^2$	$n-1$			

Tableau d'analyse de la variance : modèle linéaire simple

Remarques

- Ce tableau sert à tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$. La statistique de test est $f_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}$. Au seuil critique α, la règle de décision est de rejeter H_0 si $f_0 > F_{\alpha;1,n-2}$ ou de manière équivalente si la valeur $p - value = P(F \geq f_0)$ est petite.
- Le test avec f_0 est équivalent au test avec t_0 pour $H_1 : \beta_1 \neq 0$. En effet, $t_0^2 = \left(\frac{\hat{\beta}_1\sqrt{S_{xx}}}{\sqrt{MS_E}}\right)^2 = \frac{MS_R}{MS_E}$, et $|t_0| > t_{\alpha/2;n-2} \iff f_0 > f_{\alpha;1,n-2}$.

MTH2302: La régression linéaire Simple 15 / 28
Tableau d'analyse de la variance

Wissem Maazoun MTH2302: La régression linéaire Simple 15 / 28

Exemple

En utilisant les données de l'exemple précédent, tester $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ au seuil critique $\alpha = 0,05$.

Intervalles de confiance pour β_0 et β_1

Lorsque les erreurs ϵ_i sont de loi normale, il est démontré que pour un niveau de confiance $1 - \alpha$ donné les I.C pour β_0 et β_1 sont :

$$\beta_1 \in \hat{\beta}_1 \pm t_{\alpha/2;n-2}\sqrt{\frac{MS_E}{S_{xx}}}$$

et

$$\beta_0 \in \hat{\beta}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

Exemple

En utilisant les données de l'exemple précédent, calculer un I.C pour β_1 au niveau de confiance 95%.

Intervalle de confiance pour $E(Y|x_0)$

Pour une valeur donnée x_0 de x, la moyenne correspondante de Y est $E(Y|x_0) = \beta_0 + \beta_1 x_0$.
On l'estime ponctuellement par $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
Pour un niveau de confiance $1 - \alpha$ donné, l'intervalle de confiance pour $E(Y|x_0)$ est

$$E(Y|x_0) \in \hat{y}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Intervalle de prévision pour $Y|x_0$

Pour une valeur donnée x_0 de x, la valeur correspondante de Y est $Y_0 = Y|x_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$.
On l'estime ponctuellement par $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
Pour un niveau de confiance $1-\alpha$ donné, l'intervalle de confiance pour $Y|x = x_0$ est

$$Y_0 \in \hat{y}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}.$$

Remarque

Dans le cas de la moyenne de k observations, $\bar{Y}_0$, au point $x + x_0$, on a $\bar{Y}_0 \in \hat{y}_0 \pm t_{\alpha/2;n-2}\sqrt{MS_E\left(\frac{1}{k}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$.

Exemple

- Donner un intervalle de confiance pour la dureté moyenne de l'alliage lorsque la température est de $1250°F$, au niveau de confiance 95%.
- Donner un intervalle de prévision pour la dureté de l'alliage lorsque la température est de $1250°F$ au niveau de confiance 95%.

Validation d'un modèle de régression

Lors d'une analyse de régression, plusieurs modèles peuvent être envisagés. Les modèles sont ensuite évalués en fonction de leur aptitude à expliquer la variabilité observée dans les données. Les résidus jouent un rôle important dans cette évaluation ; ils permettent de vérifier jusqu'à quel point le modèle utilisé est adéquat.
Lorsqu'un modèle est choisi, son ajustement se fait par la méthode des moindres carrés ordinaires. Cette méthode suppose que :

MTH2302: La régression linéaire Simple 21 / 28
Validation d'un modèle de régression

Wissem Maazoun MTH2302: La régression linéaire Simple 21 / 28

Validation d'un modèle de régression

1. Pour chaque valeur de x, $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$.
2. Les erreurs ϵ sont non corrélés (ie. indépendantes). Cette propriété est qualifiée "d'abscence d'autocorrélation des erreur".
3. Les erreurs sont de loi normale i.e $\epsilon_i \sim N(0, \sigma^2)$.

Une fois que les paramètres du modèle sont estimés, on procède à la vérification de ces conditions en examinant les résidus.
Les résidus sont définis par $e_i = y_i - \hat{y}_i$ où $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \dots, n$. Et ils satisfont $E(\epsilon_i) = 0$, puisque $\sum_{i=1}^{n} e_i = 0$.

Analyse graphique des résidus

Il est important d'effectuer une analyse graphique des résidus. Le plus courant de ces graphes sont constitués des points $(\hat{y}_i, e_i), i = 1, \dots, n$; ou bien $(x_i, e_i), i = 1, \dots, n$, dans le plan.
Ces graphiques doivent être soigneusement analysés afin de détecter la présence possible de tendances particulières ou de points atypiques (aberrants). Une tendance particulière dans un de ces graphiques indique la présence d'une anomalie dans le modèle utilisé.
Afin de vérifier la normalité, on peut construire un histogramme des résidus (si on dispose de beaucoup d'observations) ; un diagramme de Tukey ; un diagramme quantile-quantile. On peut aussi effectuer un test de normalité sur les résidus.

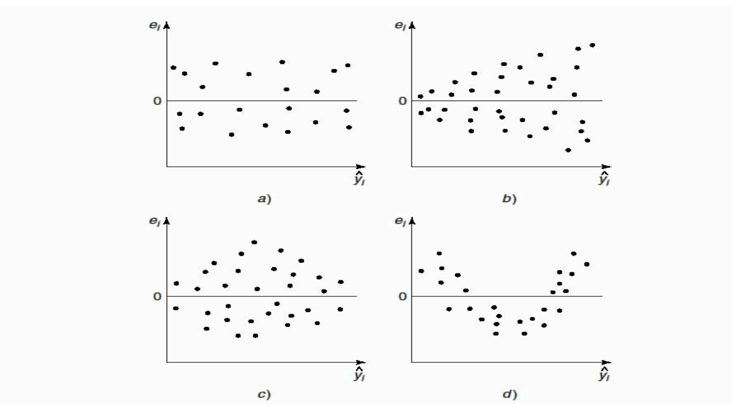

Figure 13.5 Différentes configurations des représentations graphiques des résidus : *a)* satisfaisante ; *b)* en entonnoir ; *c)* à deux arcs ; *d)* non linéaire.

MTH2302: La régression linéaire Simple 24 / 28
Validation d'un modèle de régression

Wissem Maazoun MTH2302: La régression linéaire Simple 24 / 28

Le coefficient de détermination R^2

Une des mesures de vérification d'un modèle de régression est le coefficient de détermination qui est défini par $R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$.
R^2 mesure le pourcentage de la variabilité totale S_{yy} qui est expliquée par le modèle de régression.

Exemple

Calculer le coefficient de détermination R^2 et interpréter le résultat.

Définition

Lorsque l'examen du diagramme de dispersion des points $(x_1, y_1), \ldots, (x_n, y_n)$ montre que les variables x et y sont liés par une relation non linéaire, on peut se ramener à une relation linéaire en transformant les variables (et parfois les paramètres).

Équation initiale	Équation transformée	Modèle
$y = \beta_0 e^{\beta_1 x}$	$\ln(y) = \ln(\beta_0) + \beta_1 x$	$y^* = \beta_0^* + \beta_1 x + \epsilon$ avec $y^* = \ln(y)$; $\beta_0^* = \ln(\beta_0)$.
$y = \beta_0 x^{\beta_1}$	$\ln(y) = \ln(\beta_0) + \beta_1 \ln(x)$	$y^* = \beta_0^* + \beta_1^* x^* + \epsilon$ $y^* = \ln(y)$; $x^* = \ln(x)$; $\beta_0^* = \ln(\beta_0)$.
$y = \beta_0 + \frac{\beta_1}{x}$		$y = \beta_0 + \beta_1 x^* + \epsilon$ avec $x^* = \frac{1}{x}$.
$y = \frac{1}{\beta_0 + \beta_1 x}$	$\frac{1}{y} = \beta_0 + \beta_1 x$	$y^* = \beta_0 + \beta_1 x + \epsilon$ avec $y^* = \frac{1}{y}$.
$y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$	$\ln(\frac{y}{1-y}) = \beta_0 + \beta_1 x$	$y^* = \beta_0 + \beta_1 x + \epsilon$ avec $y^* = \frac{y}{1-y}$.

MTH2302: La régression linéaire Simple 27 / 28
Les modèles non linéaires et transformations

Wissem Maazoun MTH2302: La régression linéaire Simple 27 / 28

Définition

Le lien entre deux v.a. X et Y est généralement mesuré par un paramètre, ρ, appelé le coefficient de corréa lation linéaire entre X et Y. Ce paramètre, tout comme μ, σ^2, se calcule à partir de la distribution des deux v.a et est d efini par $\rho = \frac{E(XY)-E(X)E(Y)}{\sqrt{V(X)V(Y)}}$.

On a toujours $-1 \leq \rho \leq 1$.
L'estimation du paramètre *rho* se fait à partir de n couples d'observations sur X et Y. On estime ponctuellement ρ par le coefficient de corrélation échantilonnal $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. r est un nombre sans unité et $-1 \leq r \leq 1$.