Intervalles de confiance : le bootstrap

Julien Parfait BIDIAS

2024-08-28

Background

Le Bootstrap

Le Bootstrap est une méthode statistique très utile pour l'estimation d'un intervalle de confiance des paramètres d'un modèle de régression. En effet, dans les modèles de régression la distribution de l'erreur est supposée normale et l'une des conséquences de cette hypothèse est la possibilité de construire un intervalle de confiance pour les paramètres du modèle or en réalité rien ne garantit que les erreurs suivent une loi normale. L'idée du boostrap est de pouvoir estimer cette loi via un ré-échantillonnage.

Principe

- Le modèle utilisée est $Y = X\beta + \varepsilon$ où la loi de l'erreur ε est inconnue et d'espérance nulle ;
- On commence par estimé β par les MCO pour récupérer les résidus $\hat{\varepsilon} = \hat{Y} - Y$;
- On tire au hasard et avec remise n résidus estimés $\hat{\varepsilon}$ qu'on note $\hat{\varepsilon}^*$
- A partir de ces n résidus on construit un nouvel échantillon (échantillon bootstrapé)

$$Y^* = X\hat{\beta} + \hat{\varepsilon}^*$$

Principe

- A partir de cet échantillon (X, Y^*) on estime le vecteur des paramètres et on obtient $\hat{\beta}^* = (X'X)^{-1}X'Y^*$. L'idée étant d'approcher la distribution de $\sqrt{n}(\hat{\beta} - \beta)$, inconnue à cause de β qui est inconnu, par $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$;
- Enfin pour calculer cette distribution, on calcule B-échantillons bootstrapé afin d'obtenir B-estimateurs $\hat{\beta}^*$ de $\hat{\beta}$. Il faut donc répéter B fois les ces étapes.

Principe

En somme, il faut donc :

1. tirer au hasard et avec remise n résidus estimés $\hat{\varepsilon}_i$ notés $\hat{\varepsilon}_i^{(k)}$;
2. à partir de ces n résidus, construire $\hat{y_i}^{(k)} = x_i\hat{\beta} + \hat{\varepsilon}_i^{(k)}$;
3. à partir de cet échantillon, estimer $\hat{\beta}^{(k)}$

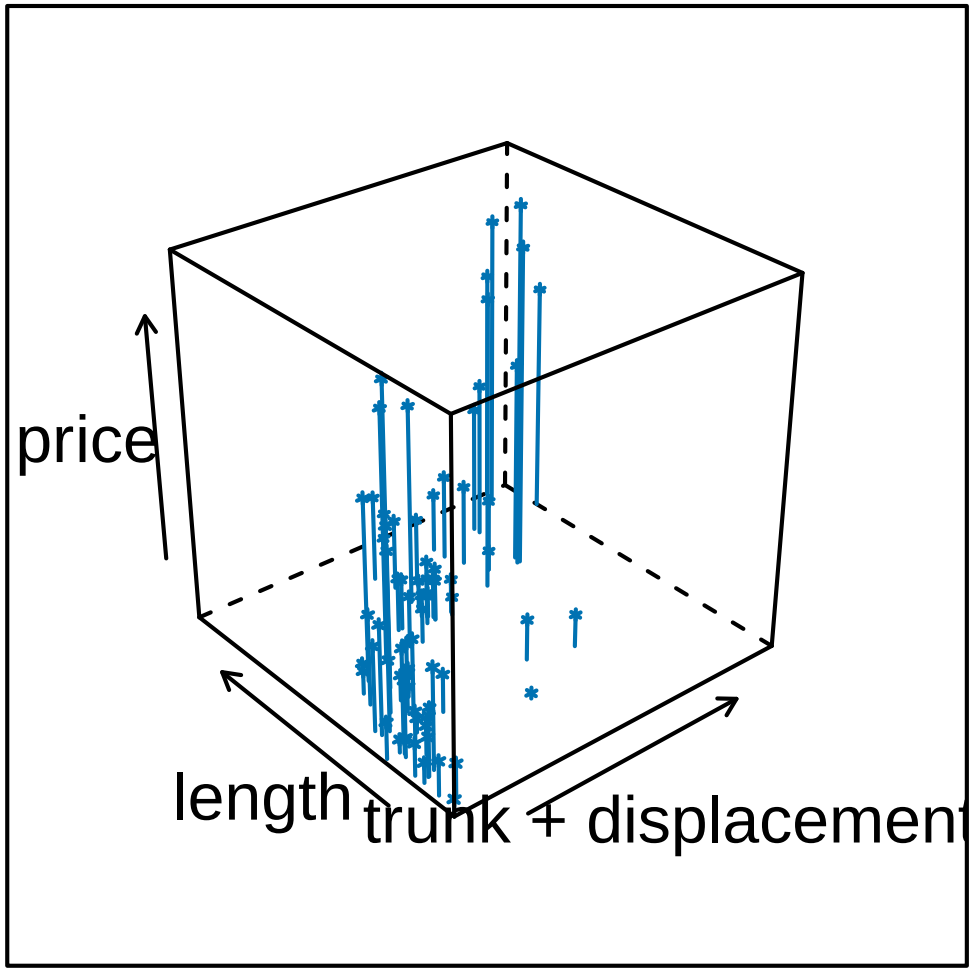
Exemple d'application

Pour l'exemple, on dispose d'une base de données de 74 observations et 4 variables. Nous présentons quelques lignes de la base.

	price	trunk	displacement	length
1	4099	11	121	186
2	4749	11	258	173
3	3799	12	121	168
4	4816	16	196	196
5	7827	20	350	222
6	5788	21	231	218

Pour appliquer l'algorithme du Bootstrap nous allons effectuer une régression linéaire en considérant le prix comme variable dépendante et les autres variables comme explicatives.

Représentation graphique



Estimation des paramètres du modèle

Table 1: Estimation

	(1)
trunk	-15.792
	(98.024)
displacement	12.358**
	(4.400)
length	20.931**
	(7.445)
Num.Obs.	74
R2	0.861
R2 Adj.	0.855
AIC	1378.7
BIC	1388.0

Calcul des résidus

Dans le bout de code ci-dessous la première ligne correspond au calcul des valeurs prédites $\hat{Y}$, la deuxième au calcul des résidus $\hat{\varepsilon} = \hat{Y} - Y$. La troisème ligne permet d'avoir une matrice de trois colonnes qui contiendra les estimateurs bootstrapés, la suivante attribue les noms des coefficients du modèles aux scoefficients bootstrapés.

```
ychap   <- predict(modele)
residus <- residuals(modele)
coeff   <- matrix(0,ncol=3,nrow=1000)
colnames(coeff) <- names(coef(modele))
auto.bootstrap <- auto
```

Ensuite on applique la procédure de bootstrap avec B $= 1000$ échantillons bootstrapés.

Algorithme de Bootstrap

Le programme ci-dessous est une application directe des éléments de la Slide numéro 5.

```
set.seed(123)
for(i in 1:nrow(coeff)){
  residus.etoile <- sample(residus,length(residus),
                           replace=T)
  Y.etoile       <- ychap + residus.etoile
  auto.bootstrap[,"price"] <- Y.etoile
  regress.boot <- lm(formula(modele),data=auto.bootstrap)
  coeff[i,] <- coef(regress.boot)
}
```

Intervalles de confiance bootstrapés

On affiche l'intervalle de confiance obtenu par bootstrap pour les paramètres de deux variables.

```
interval_conf_bootstrap <- apply(coeff,2,
                                 quantile,
                                 probs=c(0.025,0.975))
t(interval_conf_bootstrap)[-1,]
```

```
                2.5%    97.5%
displacement 3.690576 21.23149
length       7.605782 36.22049
```

Intervalles de confiance en supposant une loi normale

Un IC à 95% pour le coefficient associé à **displacement** est donc donné par $[3,69; 21.24]$. En supposant que les erreurs suivent une loi normale, nous avions $[3,59; 21,13]$.

```
                2.5 %   97.5 %
displacement 3.584336 21.13124
length       6.086434 35.77646
```

Histogramme des estimateurs bootstrapés

L'histogramme semble indiquer que la loi des estimateurs bootstrapés est proche d'une loi normale.