Understanding the Value of R^2 in Regression Analysis

NABIGH Mohamed

(*Received 23 July 2024*)

Abstract

Ce rapport vise à clarifier le concept de la valeur de R^2 dans l'analyse de régression, en particulier en abordant les cas où R^2 peut être négatif. Il explique le calcul de R^2 en utilisant la somme des carrés des résidus (RSS), la somme totale des carrés (TSS) et la somme expliquée des carrés (ESS), et fournit un exemple pratique avec des visualisations pour démontrer les concepts.

1 Introduction

La valeur de R^2, ou coefficient de détermination, est un outil essentiel dans l'analyse de régression. Dans certains cas, R^2 peut être négatif. Ce rapport explique pourquoi et comment cela peut se produire, avec une explication détaillée des calculs de R^2 et une illustration pratique.

2 Calcul de R^2

Pour comprendre pourquoi R^2 peut être négatif, nous devons examiner comment il est calculé. Nous utilisons trois variables clés dans ce calcul : RSS (Residual Sum of Squares), TSS (Total Sum of Squares), et ESS (Explained Sum of Squares).

2.1 Calcul de RSS

Pour chaque variable indépendante x, nous avons une variable dépendante y. Nous traçons une ligne de régression linéaire qui prédit les valeurs de y pour chaque valeur de x. Appelons les valeurs prédites $\hat{y}$. L'erreur entre ce que la ligne prédit et les valeurs réelles de y est calculée par soustraction. Toutes ces différences sont mises au carré et additionnées, ce qui donne la somme des carrés des résidus, RSS.

$$RSS = \sum(y - \hat{y})^2$$

2.2 Calcul de TSS

Nous pouvons calculer la valeur moyenne de y, appelée $\bar{y}$. Si nous traçons $\bar{y}$, c'est simplement une ligne horizontale à travers les données. En soustrayant $\bar{y}$ de chaque valeur réelle de y, nous obtenons la somme totale des carrés, TSS.

$$TSS = \sum (y - \bar{y})^2$$

2.3 Calcul de ESS

Les différences entre les valeurs prédites $\hat{y}$ et la valeur moyenne $\bar{y}$ sont mises au carré et additionnées. Ceci est la somme expliquée des carrés, ESS.

$$ESS = \sum (\hat{y} - \bar{y})^2$$

3 Relation entre TSS, RSS et ESS

Les différences entre les valeurs prédites $\hat{y}$ (les valeurs de y prédites par la ligne) et la valeur moyenne $\bar{y}$ sont mises au carré et additionnées. Ceci est la somme expliquée des carrés, qui est égale à :

$$ESS = \sum (\hat{y} - \bar{y})^2$$

Rappelons que :

$$TSS = \sum (y - \bar{y})^2$$

Mais nous pouvons ajouter un $+\hat{y} - \hat{y}$ dedans, car cela s'annule. Donc,

$$TSS = \sum (y - \hat{y} + \hat{y} - \bar{y})^2$$

En développant ces parenthèses, nous obtenons :

$$TSS = \sum (y - \hat{y})^2 + 2 \sum (y - \hat{y})(\hat{y} - \bar{y}) + \sum (\hat{y} - \bar{y})^2$$

Lorsque, et seulement lorsque la ligne est tracée avec une interception, ce qui suit est toujours vrai :

$$2 \sum (y - \hat{y})(\hat{y} - \bar{y}) = 0$$

Par conséquent,

$$TSS = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

Ce qui signifie simplement que :

$$TSS = RSS + ESS$$

Si nous divisons tous les termes par TSS et réarrangeons, nous obtenons :

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

4 Cas où R^2 peut être négatif

Cependant, sans interception, la relation ci-dessus change. La formule devient :

$$TSS = RSS + ESS + 2\sum(y-\hat{y})(\hat{y}-\bar{y})$$

En divisant tous les termes par TSS, nous obtenons :

$$R^2 = \frac{ESS + 2\sum(y-\hat{y})(\hat{y}-\bar{y})}{TSS}$$

Le terme supplémentaire peut rendre le numérateur négatif, ce qui fait que R^2 peut être négatif. Cela se produit lorsque la ligne horizontale $\bar{y}$ explique mieux les données que la ligne de régression.

5 Quand le terme $2\sum(y-\hat{y})(\hat{y}-\bar{y})$ est nul ou non nul

- **Nul :** Ce terme est nul lorsque la ligne de régression passe par le point moyen des données $(\bar{x}, \bar{y})$, ce qui se produit lorsque le modèle inclut une interception. Dans ce cas, les erreurs $(y-\hat{y})$ et les différences $(\hat{y}-\bar{y})$ sont orthogonales, ce qui signifie que leur produit est en moyenne nul.
- **Non nul :** Ce terme n'est pas nécessairement nul lorsque la régression ne comprend pas d'interception. Dans ce cas, la ligne de régression peut ne pas passer par le point moyen des données, et les erreurs et les différences peuvent avoir une covariance non nulle, conduisant ainsi à un terme supplémentaire qui peut être positif ou négatif.

6 Exemple

Considérons un ensemble de données et un modèle sans interception :

6.1 Données

- $X = [1, 2, 3, 4, 5]$
- $Y = [2, 4, 6, 8, 10]$
- Moyenne de Y: $\bar{Y} = 6$

6.2 Calcul de TSS

$$TSS = (2-6)^2 + (4-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2 = 40$$

6.3 Prédiction du Modèle (sans interception, en supposant un mauvais ajustement)

$$\hat{Y} = [3, 6, 9, 12, 15]$$

6.4 Calcul de RSS

$$RSS = (2-3)^2 + (4-6)^2 + (6-9)^2 + (8-12)^2 + (10-15)^2 = 55$$

6.5 Calcul de R^2

$$R^2 = 1 - \frac{55}{40} = -0.375$$

7 Exemple Python

Voici un code Python qui illustre cet exemple :

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split

# Generate non-linear data
np.random.seed(0)
X = np.sort(5 * np.random.rand(80, 1), axis=0)
y = np.sin(X).ravel() + np.random.randn(80) * 0.5

# Split data into training and testing sets
‘‘‘latex
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.model_selection import train_test_split

# Generate non-linear data
np.random.seed(0)
X = np.sort(5 * np.random.rand(80, 1), axis=0)
y = np.sin(X).ravel() + np.random.randn(80) * 0.5

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# Fit linear model with intercept
model_with_intercept = LinearRegression()
model_with_intercept.fit(X_train, y_train)
y_train_pred_with_intercept = model_with_intercept.predict(X_train)
y_test_pred_with_intercept = model_with_intercept.predict(X_test)
```

```
# Calculate R^2 with intercept
r2_train_with_intercept = r2_score(y_train, y_train_pred_with_intercept)
r2_test_with_intercept = r2_score(y_test, y_test_pred_with_intercept)

print(f"R^2 with intercept on training data: {r2_train_with_intercept}")
print(f"R^2 with intercept on test data: {r2_test_with_intercept}")

# Fit linear model without intercept
model_without_intercept = LinearRegression(fit_intercept=False)
model_without_intercept.fit(X_train, y_train)
y_train_pred_without_intercept = model_without_intercept.predict(X_train)
y_test_pred_without_intercept = model_without_intercept.predict(X_test)

# Calculate R^2 without intercept
r2_train_without_intercept = r2_score(y_train, y_train_pred_without_intercept)
r2_test_without_intercept = r2_score(y_test, y_test_pred_without_intercept)

print(f"R^2 without intercept on training data: {r2_train_without_intercept}")
print(f"R^2 without intercept on test data: {r2_test_without_intercept}")

# Plot results
plt.figure(figsize=(12, 6))

# Plot with intercept
plt.subplot(1, 2, 1)
plt.scatter(X, y, color='black')
plt.plot(X, model_with_intercept.predict(X), color='blue', linewidth=3)
plt.title(f'Linear Regression with Intercept\nR^2 (train) = {r2_train_with_intercept:.2f}, R^2 (

# Plot without intercept
plt.subplot(1, 2, 2)
plt.scatter(X, y, color='black')
plt.plot(X, model_without_intercept.predict(X), color='red', linewidth=3)
plt.title(f'Linear Regression without Intercept\nR^2 (train) = {r2_train_without_intercept:.2f},

plt.tight_layout()
plt.show()
```

8 Commentaire sur le graphique

Le graphique compare deux modèles de régression linéaire ajustés sur le même ensemble de données. L'un des modèles inclut une interception, et l'autre non. Voici un commentaire détaillé sur le graphique :

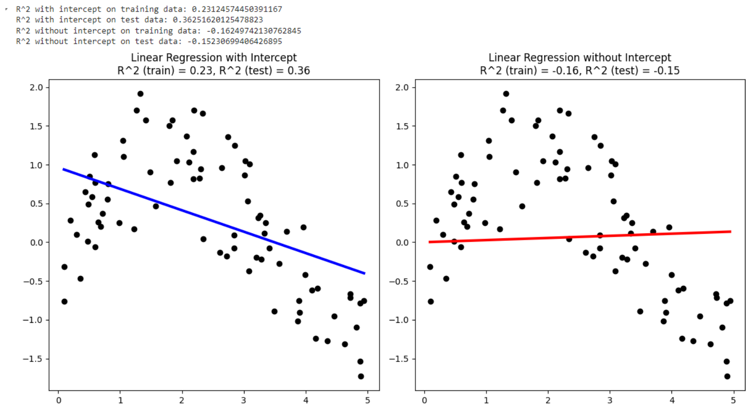

Fig. 1. Regression with and without Intercept

8.1 Graphique de gauche : Régression linéaire avec interception

- **Modèle** : Ce modèle inclut un terme d'interception.
- **$\mathbf{R}^2$ (entrainement)** : La valeur de R^2 sur les données d'entraînement est de 0.23, ce qui indique qu'environ 23% de la variance de la variable dépendante y est expliquée par la variable indépendante x dans l'ensemble de données d'entraînement.
- **$\mathbf{R}^2$ (test)** : La valeur de R^2 sur les données de test est de 0.36, ce qui suggère que le modèle performe mieux sur l'ensemble de données de test qu'il ne le fait sur l'ensemble de données d'entraînement, expliquant 36% de la variance.
- **Ligne** : La ligne de régression bleue est inclinée vers le bas, montrant une relation négative entre x et y. Cela indique qu'à mesure que x augmente, y a tendance à diminuer.

8.2 Graphique de droite : Régression linéaire sans interception

- **Modèle** : Ce modèle n'inclut pas de terme d'interception.
- **$\mathbf{R}^2$ (entrainement)** : La valeur de R^2 sur les données d'entraînement est de -0.16, indiquant un mauvais ajustement. Une valeur négative de R^2 signifie que le modèle performe moins bien qu'une simple ligne horizontale à la moyenne de y.
- **$\mathbf{R}^2$ (test)** : La valeur de R^2 sur les données de test est de -0.15, indiquant également un mauvais ajustement et que le modèle ne parvient pas à expliquer la variance dans les données de test.
- **Ligne** : La ligne de régression rouge est presque horizontale, suggérant qu'il n'y a pas de relation significative entre x et y. Cette ligne n'explique pas adéquatement la variation de y.

Ce graphique illustre efficacement l'impact significatif de l'inclusion d'un terme d'interception dans les modèles de régression linéaire, renforçant l'importance de la spécification correcte du modèle.

9 Conclusion

La valeur de R^2 peut être négative si le modèle de régression s'adapte très mal aux données. Dans ce cas, une ligne horizontale pourrait fournir une meilleure explication des variations dans les données que le modèle lui-même. Comprendre cette nuance est crucial pour une interprétation correcte des résultats de régression.

En résumé, la formule correcte pour calculer R^2 dans tous les cas est :

$$R^2 = 1 - \frac{RSS}{TSS}$$

Ce graphique montre les données avec une ligne de régression sans interception (en rouge), ainsi qu'une ligne horizontale représentant la moyenne de y (en bleu). La ligne rouge ne passe pas par le point moyen des données, ce qui explique pourquoi R^2 peut être négatif dans ce cas.

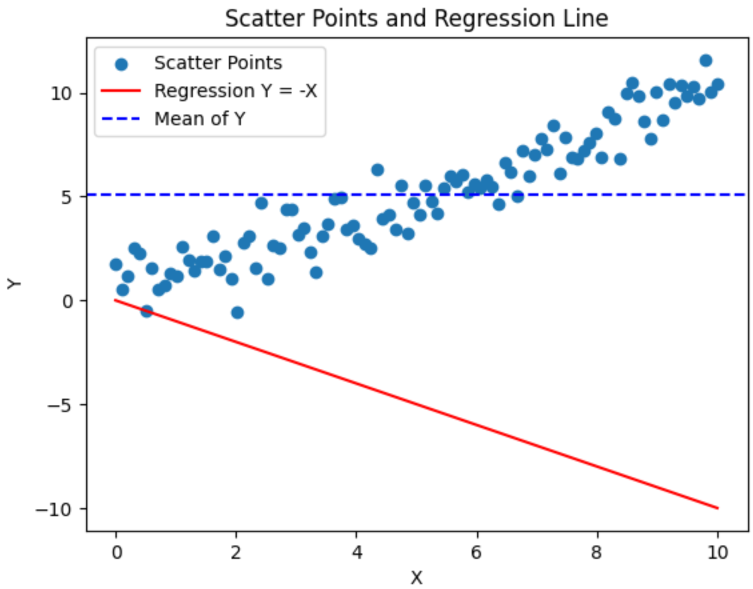

Fig. 2. Regression without Intercept

```
RSS: 13521.118637555743
TSS: 901.1954340996754
ESS: 10970.142081109538
2 * ∑(y - y_hat)(y_hat - y_bar): -23590.06528456561
R^2: -14.003536553715229
```

Fig. 3. metrics