Lecture 10: R-squared

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

Fitted values

- Consider the multiple regression model with k regressors: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + U_i$.
- Let $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ be the OLS estimators.
- The **fitted (or predicted)** value of Y is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + ... + \hat{\beta}_k X_{k,i}$.
- The **residual** is: $\hat{U}_i = Y_i - \hat{Y}_i$.
- Consider the average of $\hat{Y}$:

$$\begin{aligned}\overline{\hat{Y}} &= \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \hat{U}_i\right) \\ &= \bar{Y} - \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i = \bar{Y},\end{aligned}$$

because **when there is an intercept**, $\sum_{i=1}^{n} \hat{U}_i = 0$.

Sum-of-Squares

- The **total** variation of Y in the **sample** is:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \text{ (Total Sum-of-Squares).}$$

- The **explained** variation of Y in the **sample** is:

$$SSE = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \text{ (Explained or Model Sum-of-Squares).}$$

- The **residual** (unexplained or error) variation of Y in the **sample** is:

$$SSR = \sum_{i=1}^{n} \hat{U}_i^2 \text{ (Residual Sum-of-Squares).}$$

- If the regression contains an intercept:

$$SST = SSE + SSR.$$

Proof of SST=SSE+SSR

- First,

$$\begin{aligned} SST &= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n} \left(\hat{Y}_i + \hat{U}_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n} \left(\left(\hat{Y}_i - \bar{Y}\right) + \hat{U}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \hat{U}_i^2 + 2\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\hat{U}_i \\ &= SSE + SSR + 2\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\hat{U}_i. \end{aligned}$$

- Next, we will show that $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\hat{U}_i = 0$.

Proof of SST=SSE+SSR

- Since $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + ... + \hat{\beta}_k X_{k,i}$,

$$\begin{aligned} &\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\hat{U}_i \\ &= \sum_{i=1}^{n} \left(\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + ... + \hat{\beta}_k X_{k,i}\right) - \bar{Y}\right)\hat{U}_i \\ &= \hat{\beta}_0 \sum_{i=1}^{n} \hat{U}_i + \hat{\beta}_1 \sum_{i=1}^{n} X_{1,i}\hat{U}_i + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{k,i}\hat{U}_i - \bar{Y}\sum_{i=1}^{n} \hat{U}_i. \end{aligned}$$

- The OLS normal equations for a model with an intercept:

$$\sum_{i=1}^{n} \hat{U}_i = \sum_{i=1}^{n} X_{1,i}\hat{U}_i = ... = \sum_{i=1}^{n} X_{k,i}\hat{U}_i = 0.$$

- It follows that $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)\hat{U}_i = 0$.

R^2

- Consider the following measure of goodness of fit:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \\ &= \frac{SSE}{SST} \\ &= 1 - \frac{SSR}{SST} \\ &= 1 - \frac{\sum_{i=1}^{n} \hat{U}_i^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}. \end{aligned}$$

- $0 \leq R^2 \leq 1$.
- R^2 measures the proportion of variation in Y **in the sample** explained by the X's.

R^2 is non-decreasing in regressors

- Consider two models:

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1,i} + \tilde{U}_i,$$
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i} + \hat{\beta}_2 X_{2,i} + \hat{U}_i.$$

- We will show that

$$\sum_{i=1}^{n} \tilde{U}_i^2 \geq \sum_{i=1}^{n} \hat{U}_i^2$$

  and therefore the R^2 from the regression with one regressor is less than or equal to the R^2 from the regression with two regressors.
- This can be generalized to the case of k and $k+1$ regressors.

Proof

- Consider

$$\sum_{i=1}^{n} \left(\tilde{U}_i - \hat{U}_i\right)^2 = \sum_{i=1}^{n} \tilde{U}_i^2 + \sum_{i=1}^{n} \hat{U}_i^2 - 2\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i.$$

- We will show that

$$\sum_{i=1}^{n} \tilde{U}_i \hat{U}_i = \sum_{i=1}^{n} \hat{U}_i^2.$$

- Then,

$$0 \leq \sum_{i=1}^{n} \left(\tilde{U}_i - \hat{U}_i\right)^2 = \sum_{i=1}^{n} \tilde{U}_i^2 - \sum_{i=1}^{n} \hat{U}_i^2,$$

  or

$$\sum_{i=1}^{n} \tilde{U}_i^2 \geq \sum_{i=1}^{n} \hat{U}_i^2.$$

Proof

$$\begin{aligned}\sum_{i=1}^{n}\tilde{U}_i\hat{U}_i &= \sum_{i=1}^{n}\left(Y_i-\tilde{\beta}_0-\tilde{\beta}_1X_{1,i}\right)\hat{U}_i\\ &= \sum_{i=1}^{n}Y_i\hat{U}_i-\tilde{\beta}_0\sum_{i=1}^{n}\hat{U}_i-\tilde{\beta}_1\sum_{i=1}^{n}X_{1,i}\hat{U}_i\\ &= \sum_{i=1}^{n}Y_i\hat{U}_i-\tilde{\beta}_0\cdot 0-\tilde{\beta}_1\cdot 0\\ &= \sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1X_{1,i}+\hat{\beta}_2X_{2,i}+\hat{U}_i\right)\hat{U}_i\\ &= \sum_{i=1}^{n}\hat{U}_i^2.\end{aligned}$$

Adjusted R^2

- Since R^2 cannot decrease when more regressors are added, **even if the additional regressors are irrelevant**, an alternative measure of goodness-of-fit has been developed.
- **Adjusted** R^2: the idea is to adjust SSR and SST for degrees of freedom:

$$\bar{R}^2 = 1-\frac{SSR/\left(n-k-1\right)}{SST/\left(n-1\right)}.$$

- $\bar{R}^2 < R^2$.
- $\bar{R}^2$ can decrease when more regressors are added.

Estimation of σ^2

- In the multiple linear regression model, we can estimate $\sigma^2 = \mathrm{E}\left[U_i^2 \mid \mathbf{X}\right]$ as follows:

  Let

$$\hat{U}_i = Y_i-\hat{\beta}_0-\hat{\beta}_1X_{1,i}-\hat{\beta}_2X_{2,i}-\ldots-\hat{\beta}_kX_{k,i}.$$

- An estimator for σ^2 is

$$\begin{aligned}s^2 &= \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{U}_i^2\\ &= \frac{SSR}{n-k-1}.\end{aligned}$$

- The adjustment $k+1$ is for the number of parameters we have to estimate in order to construct the $\hat{U}$'s:

$$\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k.$$

Unbiasedness of s^2

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{U}_i^2.$$

- s^2 is an unbiased estimator of σ^2 (i.e., $\mathrm{E}\left[s^2 \mid \mathbf{X}\right] = \sigma^2$) if the following conditions hold:
  1. $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + U_i$.
  2. Conditional on $\mathbf{X}$, $\mathrm{E}\left[U_i \mid \mathbf{X}\right] = 0$ for all i's.
  3. Conditional on $\mathbf{X}$, $\mathrm{E}\left[U_i^2 \mid \mathbf{X}\right] = \sigma^2$ for all i's (homoskedasticity).
  4. Conditional on $\mathbf{X}$, $\mathrm{E}\left[U_i U_j \mid \mathbf{X}\right] = 0$ for all $i \neq j$.

R example

- Using the `hprice1` dataset from the `wooldridge` package, we regress house price on square footage, number of bedrooms, and lot size ($n = 88$, $k = 3$):

```
library(wooldridge)
m <- lm(price ~ sqrft + bdrms + lotsize, data = hprice1)
summary(m)
```

- The `summary()` output reports:

```
Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724, Adjusted R-squared:  0.6607
```

- From here we can read off $R^2 = 0.6724$, $\bar{R}^2 = 0.6607$, and $s = 59.83$.
- The **residual degrees of freedom** is $n - k - 1 = 88 - 3 - 1 = 84$.

Recovering SSR, SST, SSE from R output

- Since $s = 59.83$, we have $s^2 = 59.83^2 \approx 3{,}580$.
- $SSR = s^2 \cdot (n - k - 1) \approx 3{,}580 \times 84 \approx 300{,}720$.
- From $R^2 = 1 - SSR/SST$:

$$SST = \frac{SSR}{1 - R^2} \approx \frac{300{,}720}{1 - 0.6724} = \frac{300{,}720}{0.3276} \approx 918{,}100.$$

- $SSE = SST - SSR \approx 918{,}100 - 300{,}720 = 617{,}380$.