1. Omitted variables. The influence of the soil quality is omitted. The better the quality of soil, the better the yield. So, the **coefficient** of soil quality regarding yield should be **positive**. Besides, the better the soil quality, the easier it is to rent out the corresponding land. So, the **covariance** between rental\_out\_share and soil\_quality **is positive**. Thus, β1-hat is overestimated, that is, the bias direction is **upwards**.
2. Measurement error. When measuring the area of land, it is easy for us to get a **lower measurement** than the real area, because generally the shape of land is not regular. So, rental\_out\_share could be measured smaller. Then the regressor we use is rental\_out\_share – v, where v is a positive error. Hence, β1-hat becomes β1Var(X)/(Var(X)-Var(v)), which is larger than the real β1.Then, holding other things unchanged, the estimated β1-hat will have an **upwards** bias.
3. Simultaneity. While rental\_out\_share has a **positive** effect on the yield, **in turn** the yield has a **postive** effect on rental\_out\_share, because the higher the yield, the more people who want to rent the land and the land will be easier to rent out at a good price. In this case, β1-hat will be underestimated, that is, it has an **upwards** bias direction.

Relevance Assumption: Corr(d31, rental\_out\_sharei) ≠ 0. d31 should be relative to rental\_out\_sharei.

If a household owns a lot of land, there might be some idle. Thus, the land will be more likely to be rented out, hence d31 and rental\_out\_share have a positive relation.

Exogeneity Assumption: Corr(d31, ui) = 0. d31 should not directly influence the yield. It can only do this thourgh influencing rental\_out\_sharei.

Total land area cannot directly affect the output per unit of land, which is the yield, because they have nothing to do with each other directly.

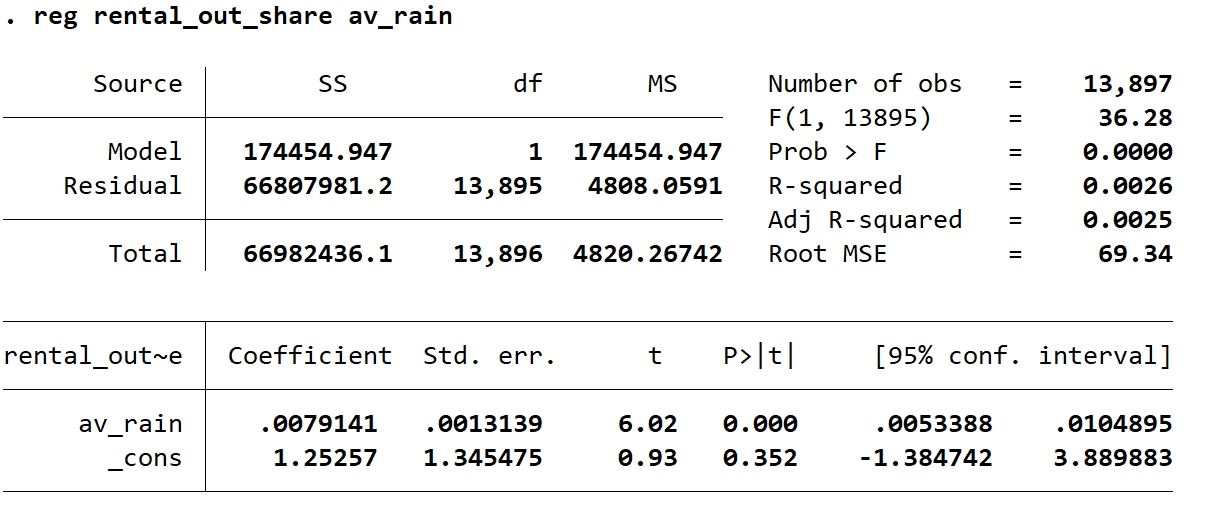
Therefore, the instrumental variable is true.

**rental\_out\_sharei = π0+π1av\_rain+vi**

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | rental\_out\_share |
|  |  |
| av\_rain | 0.00791\*\*\* |
|  | (0.00131) |
| Constant | 1.253 |
|  | (1.345) |
|  |  |
| Observations | 13,897 |
| R-squared | 0.003 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1



The result can be interpreted as: holding other things unchanged, when average rainfall increases by one unit, rental\_out\_sharei increases by 0.0079141 unit (the proportion of the area of land rented out increases by 0.79141%).

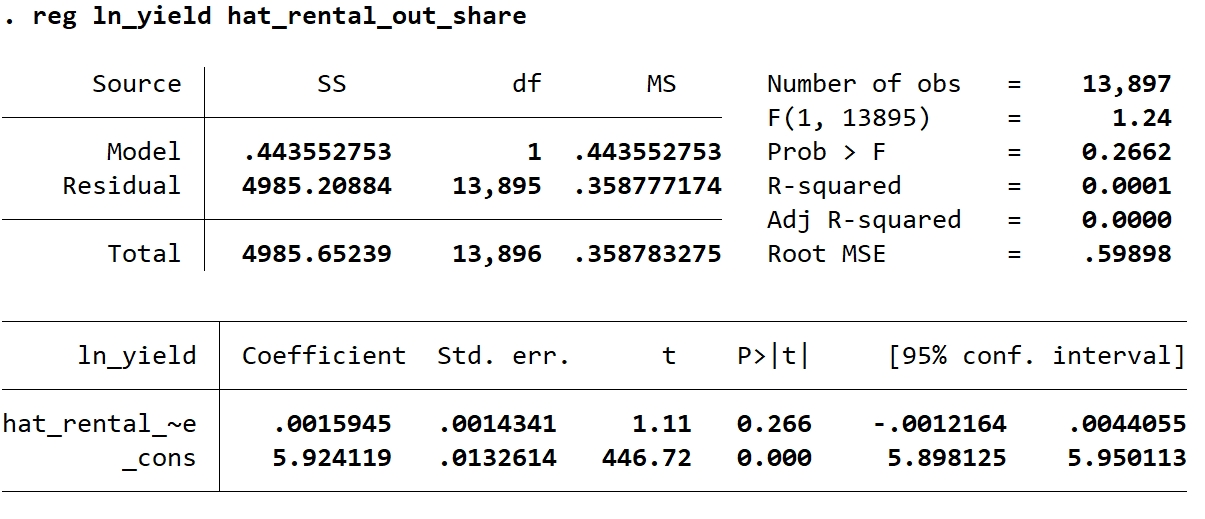
The p-value is 0.000, less than 0.01. Thus, the result is statistically significant, and we can reject the null hypothesis under 1% significance level.

**logyield*­*i = β0+β1rental\_out\_sharei-hat+ui**

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | ln\_yield |
|  |  |
| hat\_rental\_out\_share | 0.00159 |
|  | (0.00143) |
| Constant | 5.924\*\*\* |
|  | (0.0133) |
|  |  |
| Observations | 13,897 |
| R-squared | 0.000 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

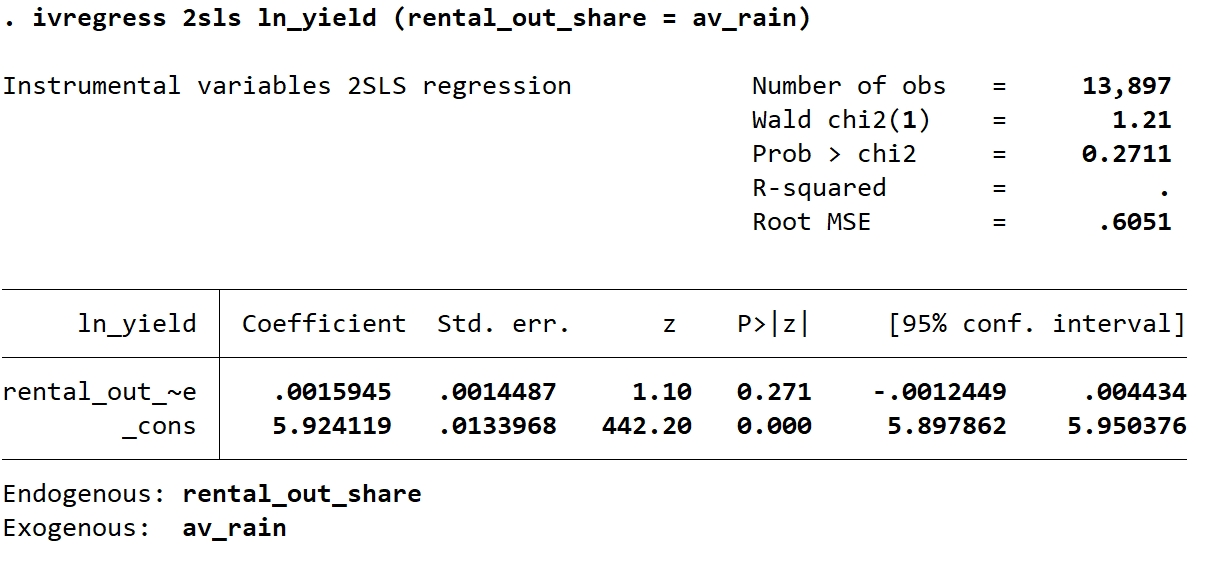
****

The result can be interpreted as: after eliminating the endogenous bias, holding other things unchanged, when rental\_out\_share (predicted) increases by 1 unit, the yield will increase by 0.0015945 units. However, the p-value is 0.266, larger than 0.1, which means that our estimate is not statistically significant under 10% significance level, so we cannot reject the null hypothesis.

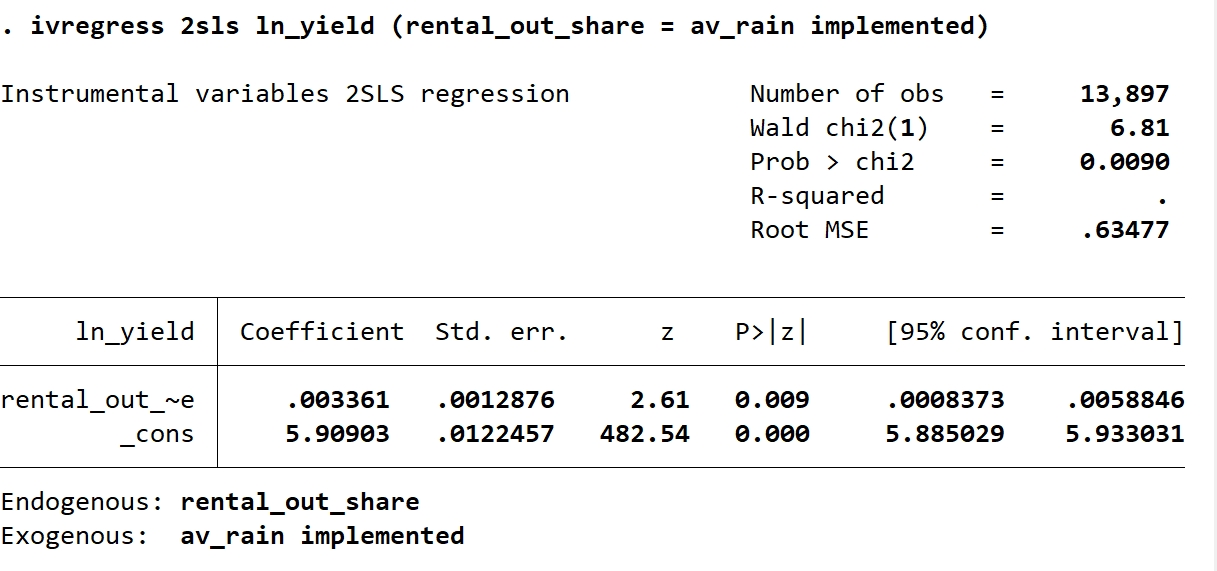
|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | ln\_yield |
|  |  |
| rental\_out\_share | 0.00159 |
|  | (0.00145) |
| Constant | 5.924\*\*\* |
|  | (0.0134) |
|  |  |
| Observations | 13,897 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1



The two estimated coefficients **look the same**, while the standard errors have **a little difference**.



Relevance Assumption:

**rental\_out\_sharei = π0 + π1Z1i ­+ π2Z2i + vi**

H0: π1 = 0 and π2 = 0

H1: either π1 or π2 is nonzero.

Then we do **F-test** to look at the strength of the instruments.

Exogenous Assumption:

In this context, the number of instrumental variables, which is 2, is larger than that of the endogenous regressor, which is 1. Thus, we can use **J-test**.

First we run the regression to get logyieldi-hat

**logyieldi = β0 + β1rental\_out\_sharei + ui**

ui-hat = logyieldi – logyieldi-hat

Then we regress ui-hat on Z1i and Z2i

**ui-hat = π0 + π1Z1i + π2Z2i + vi**

H0: π1 = 0 and π2 = 0

H1: either π1 or π2 is nonzero.

Then we figure out the F statistic, and get the J statistic: J=2F

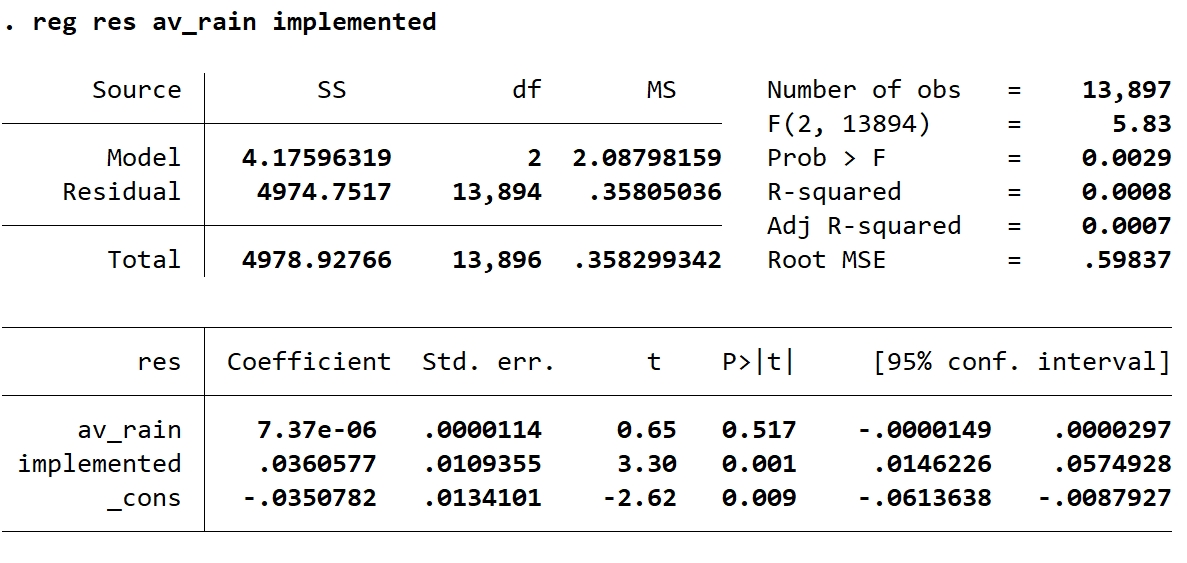
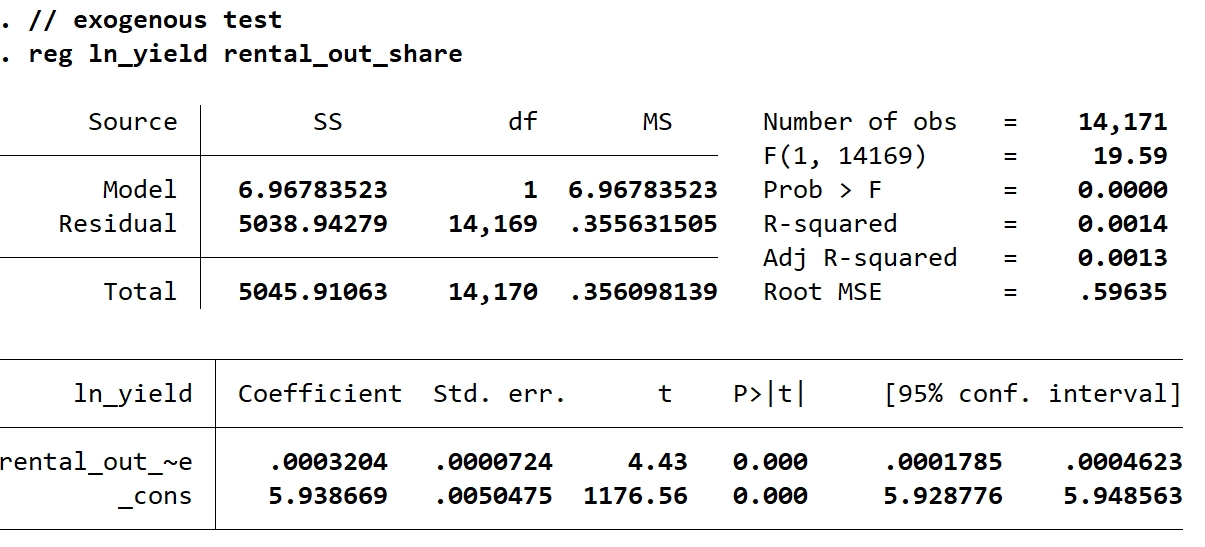
For relevance test:

表格

Description automatically generated

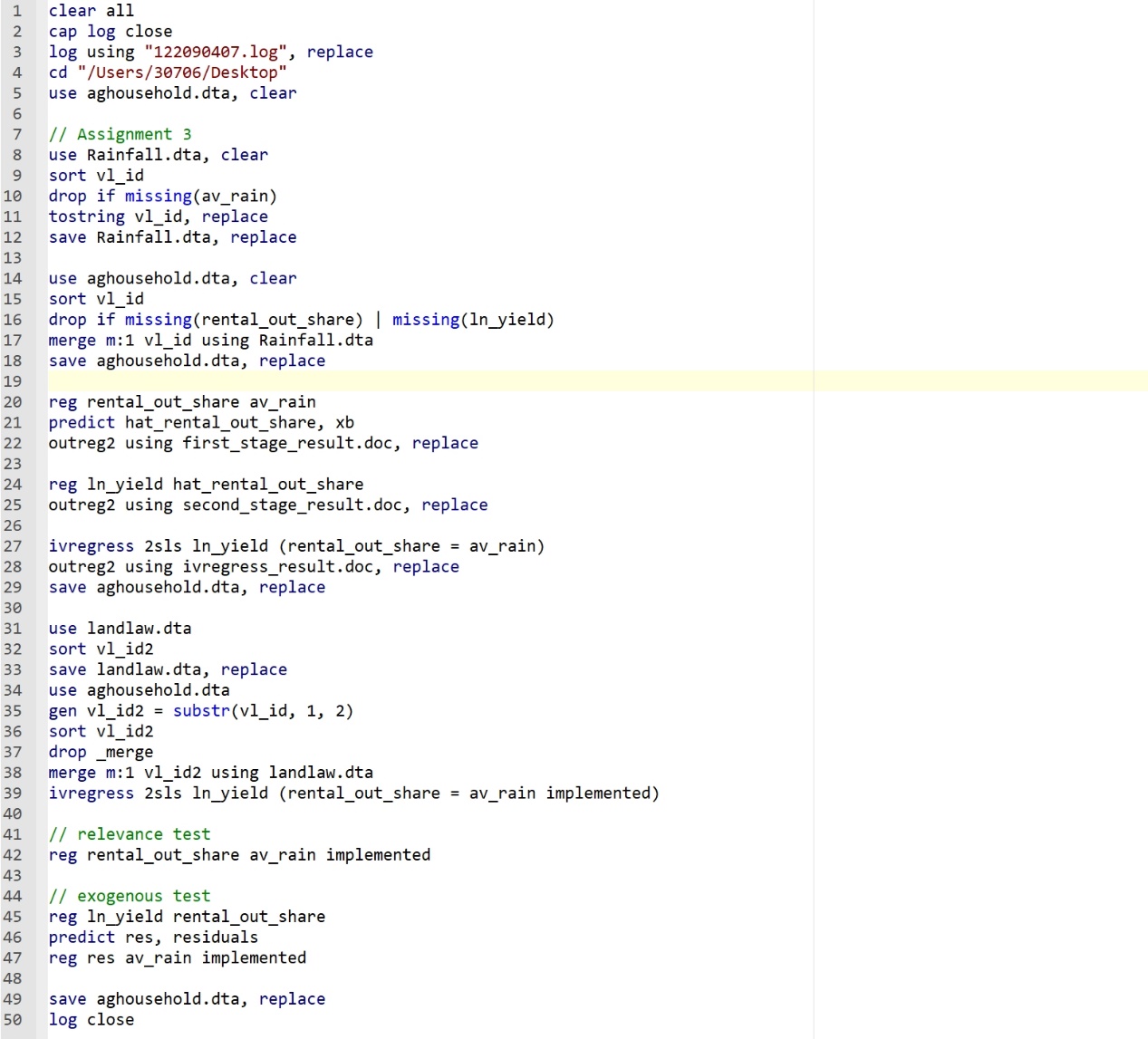
The F statistic is 25.30 > 2.9964 and both p-values of the IVs are 0. So, we can reject the null hypothesis under 5% significance level. Thus, the relevance condition is likely to be true.

For exogenous test:



J = 2F = 11.66 > 3.84. Thus, we may reject the null hypothesis under 5% significance level, so at least one iv is endogenous.

Code



**Log**

文本

Description automatically generated文本

Description automatically generated文本

Description automatically generated文本

Description automatically generated电脑屏幕的照片上有文字

Description automatically generated with medium confidence文本

Description automatically generated