ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

KHOA TOÁN - TIN
Faculty of Mathematics and Informatics

Suy luận thống kê

Phân tích các yếu tố ảnh hưởng đến sản lượng lúa bằng mô hình hồi quy tuyến tính bội ở Malausma.

Họ tên: Nguyễn Đức Hùng

MSSV: 20237441

Lời mở đầu

Trong bối cảnh khoa học dữ liệu và các phương pháp phân tích định lượng ngày càng được ứng dụng rộng rãi, môn *Suy luận thống kê* đóng vai trò quan trọng trong việc trang bị cho sinh viên những kiến thức nền tảng về ước lượng, kiểm định giả thuyết và phân tích dữ liệu. Thông qua những bài giảng của cô Nguyễn Thị Thu Thuỷ, em đã có cơ hội tiếp cận hệ thống các khái niệm và kỹ thuật cốt lõi, từ đó hiểu rõ hơn về cách vận dụng thống kê vào các vấn đề thực tiễn.

Bài báo cáo giữa kỳ này được thực hiện với mục tiêu củng cố kiến thức, rèn luyện kỹ năng phân tích số liệu và áp dụng các phương pháp suy luận thống kê vào một bộ dữ liệu cụ thể. Em hy vọng rằng nội dung bài báo cáo sẽ thể hiện được sự nghiêm túc, tinh thần học hỏi và những hiểu biết mà em đã tích luỹ trong quá trình học tập môn học.

Trước hết, em xin gửi lời cảm ơn chân thành và sâu sắc tới cô Nguyễn Thị Thu Thuỷ, người đã tận tình giảng dạy và truyền đạt những kiến thức quý báu trong suốt học phần *Suy luận thống kê*. Những chỉ dẫn và sự hỗ trợ của cô đã giúp em rất nhiều trong việc định hướng tư duy và hoàn thành bài báo cáo này.

Em cũng xin cảm ơn các thầy cô trong khoa, cùng bạn bè đã đồng hành, chia sẻ tài liệu và hỗ trợ em trong quá trình tìm hiểu và phân tích dữ liệu. Bên cạnh đó, em xin cảm ơn gia đình đã luôn động viên và tạo điều kiện thuận lợi để em yên tâm học tập.

Mặc dù đã cố gắng hoàn thiện bài báo cáo một cách tốt nhất, nhưng do kiến thức và kinh nghiệm của bản thân còn hạn chế, bài làm của em khó tránh khỏi thiếu sót. Em rất mong nhận được những ý kiến đóng góp của cô để em có thể cải thiện và hoàn thiện hơn trong tương lai.

Nguyễn Đức Hùng

Bảng ký hiệu

Ký hiệu	Diễn giải
n	Số quan sát (kích thước mẫu)
k	Số biến giải thích trong mô hình hồi quy
p	Số tham số của mô hình, với $p = k + 1$
Y, y_i	Biến phụ thuộc và giá trị quan sát thứ i
x_j, x_{ij}	Biến giải thích thứ j và giá trị quan sát thứ i
β_0	Hệ số chặn của mô hình hồi quy
$\beta_1, \ldots, \beta_k$	Các hệ số hồi quy
$\boldsymbol{\beta}$	Vectơ hệ số hồi quy
ε, ε_i	Sai số ngẫu nhiên
$\mathbf{X}$	Ma trận biến giải thích
$\mathbf{y}$	Vectơ giá trị quan sát
$\widehat{\boldsymbol{\beta}}$	Ước lượng OLS của vectơ hệ số hồi quy
$\widehat{y}_i$	Giá trị ước lượng của y_i
e_i	Phần dư tại quan sát thứ i
$\boldsymbol{e}$	Vectơ phần dư
σ^2	Phương sai của sai số ngẫu nhiên
s^2	Ước lượng phương sai của sai số
SS_T	Tổng bình phương chung
SS_R	Tổng bình phương hồi quy
SS_E	Tổng bình phương sai số
R^2	Hệ số xác định
R^2_{adj}	Hệ số xác định hiệu chỉnh
F_0	Thống kê kiểm định F
t_0	Thống kê kiểm định t
α	Mức ý nghĩa thống kê
W_α	Miền bác bỏ giả thuyết

Bảng 1: Danh mục ký hiệu hồi quy tuyến tính bội

Mục lục

Danh sách bảng

Danh sách hình vẽ

1 Lý thuyết

1.1 Hồi quy tuyến tính bội

Trong hồi quy tuyến tính bội, biến phụ thuộc Y được mô tả như một tổ hợp tuyến tính của các biến giải thích x_j, với $j = 1, 2, \ldots, k$, và được biểu diễn dưới dạng:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

Ta có thể viết mô hình trên thành

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon.$$

Trong đó:

- x_j $(j = 1, 2, \ldots, k)$ là biến độc lập hoặc biến hồi quy;
- β_0 là hệ số chặn;
- $\beta_1, \ldots, \beta_k$ là các hệ số góc chưa biết và được gọi chung là hệ số hồi quy;
- ε là sai số ngẫu nhiên.

Giá trị kỳ vọng có điều kiện của biến phụ thuộc Y được biểu diễn bởi:

$$E(Y|x_1, \ldots, x_k) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j.$$

1.2 Phương pháp tiếp cận ma trận

Giả sử, có n quan sát, k số biến giải thích với $j = 1, 2, ..., k$. Mô hình hồi quy tuyến tính bội được viết dưới dạng ma trận, như sau:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Ta có:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1j} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2j} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nj} & \ldots & x_{nk} \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{pmatrix}; \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}.$$

1.3 Giả thiết

- Việc ước lượng dựa trên cơ sở mẫu ngẫu nhiên $(\mathbf{X}, \mathbf{y})$.
- $E(\varepsilon|\mathbf{X}) = 0$ (ma trận 0).
- $E(\varepsilon^\top \varepsilon|\mathbf{X}) = \sigma^2 I_n (I_n$ là ma trận đơn vị cấp $n)$.
- Tồn tại ma trận nghịch đảo $(\mathbf{X}^\top \mathbf{X})^{-1}$.

1.4 Ước lượng hệ số hồi quy

Phương pháp bình phương nhỏ nhất (Ordinary Least Squares – OLS)
Được sử dụng để ước lượng vectơ hệ số hồi quy $\boldsymbol{\beta}$ bằng cách tối thiểu hóa tổng bình phương phần dư, cụ thể:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Ước lượng bình phương nhỏ nhất của $\boldsymbol{\beta}$ là:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Ước lượng phương sai là:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-k-1}.$$

Trong hồi quy tuyến tính bội với $p = k + 1$ tham số, và mẫu số $n - p$ được gọi là bậc tự do phần dư.
Ma trận hiệp phương sai của ước lượng $\widehat{\boldsymbol{\beta}}$ là:

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 \mathbf{C}.$$

Trong đó, Các phần tử trên đường chéo chính của $(\mathbf{X}^\top \mathbf{X})^{-1}$ là các phương sai của $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$.
Với $k = 2$ biến hồi quy:

$$\mathbf{C} = (\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{pmatrix}$$

là ma trận đối xứng vì ma trận $(\mathbf{X}^\top \mathbf{X})^{-1}$ là ma trận đối xứng và

$$V(\widehat{\beta}_j) = \sigma^2 C_{jj}, \ j = 0, 1, 2 \text{ và } \text{cov}(\widehat{\beta}_i, \widehat{\beta}_j) = \sigma^2 C_{ij}, \ i \neq j.$$

Khi thay thế σ^2 bằng s^2, sai số chuẩn của $\widehat{\beta}_j$ được xác định bởi:

$$se(\widehat{\beta}_j) = \sqrt{s^2 C_{jj}}, \ j = 0, 1, \ldots, k.$$

1.5 Phân tích phần dư

Sự khác nhau giữa quan sát y_i và giá trị ước lượng $\widehat{y}_i$ là một phần dư, ký hiệu là $e_i = y_i - \widehat{y}_i$. Vectơ $n \times 1$ của phần dư được biểu thị bằng:

$$\boldsymbol{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

Trong đó:

- $\mathbf{y}$ là vectơ giá trị quan sát;
- $\widehat{\mathbf{y}}$ là vectơ giá trị ước lượng.

Phân tích phần dư dùng để kiểm tra mức độ phù hợp của mô hình và tính hợp lệ của các giả thiết thống kê.

1.6 Kiểm định ý nghĩa của hồi quy

Kiểm định ý nghĩa của hồi quy là kiểm định để xác định xem có tồn tại mối quan hệ tuyến tính giữa biến phụ thuộc Y và các biến giải thích $x_1, x_2, \ldots, x_k$ hay không. Giả thuyết:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0;$$
$$H_1 : \beta_j \neq 0 \text{ với ít nhất một } j,\ j = 1, 2, \ldots, k.$$

Việc bác bỏ $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ ngụ ý rằng ít nhất một trong các biến hồi quy $x_1, x_2, \ldots, x_k$ có ý nghĩa thống kê vào mô hình.

Khi mô hình có chứa hệ số chặn thì:

$$SS_T = SS_R + SS_E.$$

Ta xác định các đại lượng sau:

- SS_T là tổng bình phương chung:

$$SS_T = \mathbf{y}^\top \mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n},$$

- SS_R Tổng bình phương hồi quy:

$$SS_R = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n},$$

- Tổng bình phương sai số:

$$SS_E = \mathbf{y}^\top \mathbf{y} - \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y}.$$

Chọn tiêu chuẩn kiểm định

$$\boxed{F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E}.}$$

Nếu giả thuyết $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ là đúng thì F_0 có phân phối F với $\nu_1 = k$ và $\nu_2 = n - p$ bậc tự do.

Dựa trên dữ liệu quan sát được $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i),\ i = 1, \ldots, n\}$, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$f_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E}.$$

Tìm miền bác bỏ giả thuyết H_0:

$$W_\alpha = \{f_0 \mid f_0 > f_\alpha(k, n-p)\},$$

ở đây, $f_\alpha(k, n-p)$ được xác định từ bảng giá trị tới hạn phân phối F với $\nu_1 = k$ và $\nu_2 = n - p$ bậc tự do.

Xét xem f_0 có thuộc W_α hay không để kết luận.

- Nếu $f_0 \in W_\alpha$ thì bác bỏ giả thuyết H_0.
- Nếu $f_0 \notin W_\alpha$ thì chưa có cơ sở để bác bỏ giả thuyết H_0.

1.7 Hệ số xác định

Hệ số xác định

Hệ số xác định bội của mô hình hồi quy tuyến tính bội được ký hiệu và định nghĩa là

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Hệ số xác định hiệu chỉnh

Hệ số xác định đã hiệu chỉnh được ký hiệu và định nghĩa bởi

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}, \qquad p = k + 1$$

1.8 Kiểm định giả thuyết về hệ số hồi quy riêng

Kiểm định giả thuyết này được gọi là kiểm định từng phần hoặc kiểm định biên vì hệ số hồi quy phụ thuộc vào tất cả các biến hồi quy x_i khác $(i \neq j)$ có trong mô hình.

- Xác định dạng cụ thể của cặp giả thuyết cần kiểm định.

$$H_0 : \beta_j = \beta_{j0} = 0,\ H_1 : \beta_j \neq 0.$$

- Chọn tiêu chuẩn kiểm định:

$$T_0 = \frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}}.$$

Nếu giả thuyết $H_0 : \beta_j = \beta_{j0}$ là đúng thì $T_0 \sim t(n-p)$.

- Dựa trên dữ liệu quan sát được $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i),\ i = 1, \ldots, n\}$, tính giá trị quan sát của tiêu chuẩn kiểm định:

$$t_0 = \frac{\widehat{\beta}_j - \beta_{j0}}{se(\widehat{\beta}_j)}.$$

- Tìm miền bác bỏ giả thuyết

$$H_0 : W_\alpha = \{t_0 \mid |t_0| > t_{\alpha/2}(n-p)\},$$

Ở đây, $t_{\alpha/2}(n-p)$ được xác định từ bảng giá trị tới hạn phân phối t với $n-p$ bậc tự do.
(5) Xét xem t_0 có thuộc W_α hay không để kết luận.

- Ta biết rằng, nếu x_j không tác động đến Y thì $\beta_j = 0$ và ngược lại nếu x_j tác động đến Y thì $\beta_j \neq 0$.
- Nếu giả thuyết $H_0 : \beta_j = 0$ không bị bác bỏ, thì biến hồi quy x_j có thể bị xóa khỏi mô hình.

2 Dữ liệu

2.1 Mô tả bộ dữ liệu

Bộ dữ liệu *RiceFarms* phản ánh tình hình sản xuất lúa tại Indonesia, được thu thập dưới dạng dữ liệu bảng (panel data) với tổng cộng 1.026 quan sát, trong đó mỗi quan sát tương ứng với một nông hộ trồng lúa. Dữ liệu bao gồm thông tin về đặc điểm nông hộ, điều kiện canh tác, các yếu tố đầu vào trong sản xuất, lao động, giá cả cũng như kết quả đầu ra của quá trình sản xuất lúa.

Các nông hộ trong mẫu được khảo sát tại nhiều khu vực khác nhau của Indonesia, bao gồm Wargabinangun, Langan, Gunungwangi, Malausma, Sukaambit và Ciwangi. Với bài báo cáo sau, ta sẽ tập trung phân tích ở Malausma.

2.2 Giải thích các biến trong bộ dữ liệu

Tên biến	Mô tả	Đơn vị / Giá trị
id	Mã định danh của nông hộ trồng lúa	–
size	Tổng diện tích canh tác lúa của nông hộ	Hecta
status	Tình trạng sở hữu đất canh tác	owner / share / mixed
varieties	Loại giống lúa được sử dụng	trad / high / mixed
bimas	Mức độ tham gia chương trình thâm canh BIMAS	no / yes / mixed
seed	Lượng giống lúa sử dụng	Kilogram
urea	Lượng phân đạm urê sử dụng	Kilogram
phosphate	Lượng phân lân sử dụng	Kilogram
pesticide	Chi phí thuốc bảo vệ thực vật	Rupiah
pseed	Giá giống lúa	Rupiah/kg
purea	Giá phân urê	Rupiah/kg
pphosph	Giá phân lân	Rupiah/kg
hiredlabor	Số giờ lao động thuê ngoài	Giờ
famlabor	Số giờ lao động gia đình	Giờ
totlabor	Tổng số giờ lao động (không bao gồm lao động thu hoạch)	Giờ
wage	Tiền công lao động	Rupiah/giờ
goutput	Sản lượng lúa thu hoạch (tổng)	Kilogram
noutput	Sản lượng lúa ròng (đã trừ chi phí thu hoạch)	Kilogram
price	Giá lúa thóc trên thị trường	Rupiah/kg
region	Khu vực địa lý của nông hộ	Tên vùng

Bảng 2: Mô tả các biến trong bộ dữ liệu RiceFarms

2.3 Nội dung vấn đề nghiên cứu

Trong bối cảnh nông nghiệp vẫn giữ vai trò quan trọng đối với nền kinh tế Indonesia, việc phân tích các yếu tố ảnh hưởng đến sản lượng lúa có ý nghĩa thiết thực cả về mặt kinh tế lẫn hoạch định chính sách. Sản lượng lúa không chỉ phản ánh hiệu quả sản xuất của từng nông hộ mà còn chịu tác động đồng thời của nhiều yếu tố như quy mô canh tác, lượng đầu vào sản xuất, lao động, giá cả và điều kiện canh tác.

Xuất phát từ thực tế đó, nghiên cứu này sử dụng mô hình hồi quy tuyến tính bội nhằm phân tích mối quan hệ giữa sản lượng lúa, được đo lường thông qua biến *goutput*, và các biến giải thích đại diện cho các yếu tố đầu vào trong quá trình sản xuất nông nghiệp. Việc lựa chọn *goutput* làm biến phụ thuộc cho phép đánh giá trực tiếp mức độ tác động của từng yếu tố đến tổng sản lượng lúa thu hoạch của nông hộ.

Thông qua mô hình hồi quy, nghiên cứu hướng đến việc kiểm định các giả thuyết thống kê liên quan đến ý nghĩa và chiều hướng tác động của các biến độc lập, đồng thời đánh giá mức độ phù hợp của mô hình trong việc giải thích biến thiên của sản lượng lúa. Kết quả ước lượng kỳ vọng sẽ cung cấp bằng chứng thực nghiệm về vai trò của các yếu tố sản xuất, từ đó làm cơ sở khoa học cho việc đề xuất các giải pháp nhằm nâng cao hiệu quả sản xuất lúa trong thực tiễn.

3 Phân tích kết quả

Các kết quả nghiên cứu được thực hiện bằng ngôn ngữ lập trình *Python* và được triển khai trên máy tính cá nhân *MacBook Air M2*. Các biểu đồ minh họa được xây dựng bằng mã nguồn Python.

3.1 Thống kê mô tả

Bảng 3 trình bày các thống kê mô tả cơ bản của các biến định lượng trong mô hình nghiên cứu với cỡ mẫu gồm 198 quan sát. Các chỉ tiêu thống kê bao gồm số quan sát, giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất, các phân vị (25%, 50%, 75%) và giá trị lớn nhất.

Kết quả cho thấy diện tích canh tác lúa trung bình của các nông hộ tương đối nhỏ, phản ánh đặc điểm sản xuất nông nghiệp quy mô hộ gia đình. Các yếu tố đầu vào như lượng giống, phân bón và lao động có mức độ phân tán khá lớn, thể hiện sự khác biệt đáng kể trong phương thức canh tác giữa các nông hộ. Sản lượng lúa có độ biến động cao, cho thấy sản xuất lúa chịu ảnh hưởng mạnh của nhiều yếu tố đầu vào cũng như điều kiện canh tác khác nhau.

Chỉ tiêu	Diện tích canh tác	Lượng giống sử dụng	Lượng phân urê
Số quan sát (count)	198	198	198
Giá trị trung bình (mean)	0,224	8,687	50,444
Độ lệch chuẩn (std)	0,179	6,603	45,285
Giá trị nhỏ nhất (min)	0,014	1	1
Phân vị 25%	0,1	4	20
Trung vị (50%)	0,157	6	35
Phân vị 75%	0,287	10	70
Giá trị lớn nhất (max)	1,12	33	270

Bảng 3: Thống kê mô tả các biến định lượng trong mô hình

Chỉ tiêu	Lượng phân lân	Chi phí thuốc BVTV	Giá giống	Giá phân urê
Số quan sát (count)	198	198	198	198
Giá trị trung bình (mean)	27,308	35,530	95,790	81,020
Độ lệch chuẩn (std)	24,798	133,754	48,855	10,514
Giá trị nhỏ nhất (min)	0	0	40	65
Phân vị 25%	10	0	60	70
Trung vị (50%)	20	0	80	80
Phân vị 75%	35	0	140	90
Giá trị lớn nhất (max)	170	1000	300	100

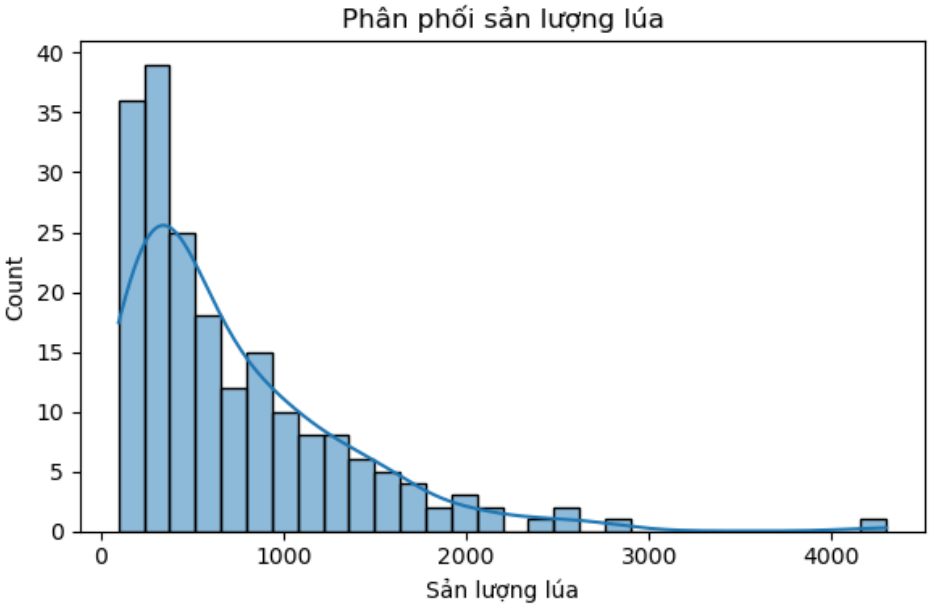
Thống kê mô tả các biến định lượng trong mô hình

Chỉ tiêu	Giá phân lân	Lao động thuê	Lao động gia đình
Số quan sát (count)	198	198	198
Giá trị trung bình (mean)	81,49	100,222	150,621
Độ lệch chuẩn (std)	10,847	131,303	119,099
Giá trị nhỏ nhất (min)	65	1	12
Phân vị 25%	70	19,5	78
Trung vị (50%)	80	52	117,5
Phân vị 75%	90	126,75	181,5
Giá trị lớn nhất (max)	110	825	855

Thống kê mô tả các biến định lượng trong mô hình

Chỉ tiêu	Tổng lao động	Tiền công lao động	Giá lúa	Sản lượng lúa
Số quan sát (count)	198	198	198	198
Giá trị trung bình (mean)	250,692	78,381	92,167	717,02
Độ lệch chuẩn (std)	207,204	37,701	41,095	610,766
Giá trị nhỏ nhất (min)	28	30,260	50	95
Phân vị 25%	123	49,225	60	270
Trung vị (50%)	183,5	59,285	71,5	511
Phân vị 75%	316,25	123,275	140	1.000
Giá trị lớn nhất (max)	1.680	152,500	160	4.300

Thống kê mô tả các biến định lượng trong mô hình



Hình 3.1: Biểu đồ histogram

3.2 Loại bỏ biến ngoại lai

Chỉ tiêu	Giá trị
Biến xem xét ngoại lai	Sản lượng lúa
Ngưỡng dưới	−825
Ngưỡng trên	2095
Số quan sát ban đầu	198
Số quan sát sau khi làm sạch	191
Số quan sát bị loại	7

Bảng 4: Kết quả loại bỏ quan sát ngoại lai đối với biến sản lượng lúa

3.3 Kết quả trước khi loại bỏ biến

Chỉ tiêu	R	R^2	R^2_{adj}	s
Giá trị	0,9130	0,8335	0,8213	197,8999

Bảng 5: Các chỉ tiêu đánh giá độ phù hợp của mô hình hồi quy

Thành phần	Bậc tự do	SS	MS	F	Prob $> F$
Hồi quy	13	$3,4711 \times 10^7$	$2,6701 \times 10^6$	68,1757	0,0000
Phần dư	177	$6,9321 \times 10^6$	$3,9164 \times 10^4$		
Chung	190	$4,1643 \times 10^7$			

Bảng 6: Bảng phân tích phương sai (ANOVA) của mô hình hồi quy

3.4 Lựa chọn biến bằng phương pháp loại bỏ lùi

Phương pháp loại bỏ lùi được sử dụng nhằm lựa chọn tập biến giải thích phù hợp cho mô hình hồi quy tuyến tính bội. Quy trình bắt đầu với mô hình đầy đủ bao gồm tất cả các biến độc lập, sau đó lần lượt loại bỏ các biến có mức ý nghĩa thống kê thấp nhất (tức là có p-value lớn hơn mức ý nghĩa lựa chọn), cho đến khi tất cả các biến còn lại đều có ý nghĩa thống kê.

Quá trình thực hiện Dựa trên kết quả ước lượng từ dữ liệu, quá trình loại bỏ lùi được thực hiện qua 6 vòng. Ở mỗi vòng, biến có p-value lớn nhất trong mô hình được loại bỏ. Cụ thể:

- Vòng 1: Loại bỏ biến lượng phân lân (kg) với p-value $= 0,8980$
- Vòng 2: Loại bỏ biến lao động thuê (công) với p-value $= 0,6112$
- Vòng 3: Loại bỏ biến giá giống với p-value $= 0,5138$
- Vòng 4: Loại bỏ biến giá phân lân với p-value $= 0,4445$
- Vòng 5: Loại bỏ biến lao động gia đình (công) với p-value $= 0,4104$
- Vòng 6: Loại bỏ biến chi phí thuốc BVTV với p-value $= 0,4231$

Kết quả cuối cùng

Kết thúc quá trình loại bỏ lùi, mô hình hồi quy cuối cùng bao gồm 8 biến, trong đó có 1 hệ số chặn và 7 biến giải thích. Cụ thể:

- Các biến bị loại: Lượng phân lân, lao động thuê ngoài, giá giống lúa, giá phân lân, lao động gia đình và chi phí thuốc bảo vệ thực vật.
- Các biến được giữ lại: Hệ số chặn, diện tích canh tác, lượng giống sử dụng, lượng phân urê, giá phân urê, tổng lao động, tiền công lao động và giá lúa.

Nhận xét Kết quả lựa chọn biến cho thấy sản lượng lúa chịu ảnh hưởng chủ yếu bởi *diện tích canh tác*, *lượng phân urê*, *chi phí thuốc bảo vệ thực vật*, cùng với các yếu tố về *giá đầu vào* và *giá đầu ra*. Trong khi đó, các biến về lao động và một số đầu vào khác không thể hiện vai trò đáng kể về mặt thống kê trong mô hình, do đó được loại bỏ nhằm nâng cao tính gọn nhẹ và hiệu quả giải thích của mô hình hồi quy.

3.5 Kết quả phân tích hồi quy

Chỉ tiêu	R	R^2	R^2_{adj}	s
Giá trị	0,9117	0,8311	0,8246	196,0433

Bảng 7: Các chỉ tiêu đánh giá độ phù hợp của mô hình hồi quy

Thành phần	**Bậc tự do**	**SS**	**MS**	F	**Prob** $> F$
Hồi quy	7	$3,4610 \times 10^7$	$4,9442 \times 10^6$	128,6455	0,0000
Phần dư	183	$7,0332 \times 10^6$	$3,8428 \times 10^4$		
Chung	190	$4,1643 \times 10^7$			

Bảng 8: Bảng phân tích phương sai (ANOVA) của mô hình hồi quy

Mô hình	**Số biến**	R^2	R^2_{adj}
Ban đầu	13	0,8335	0,8213
Sau khi loại bỏ biến	7	0,8311	0,8246

Bảng 9: So sánh mức độ phù hợp của mô hình trước và sau khi loại bỏ biến

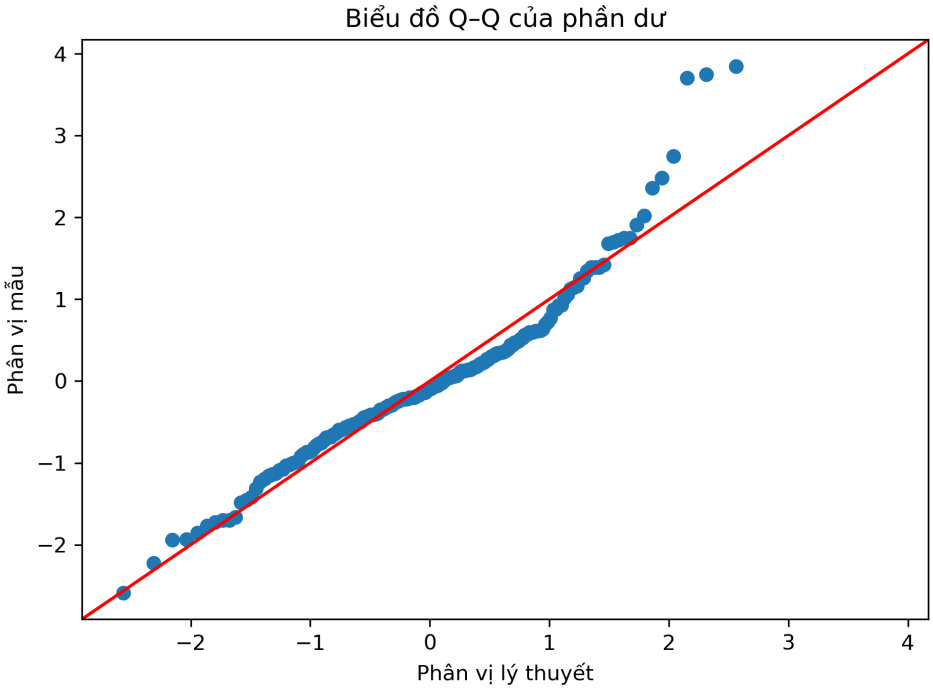
Nhận xét Kết quả ước lượng cho thấy cả mô hình hồi quy ban đầu và mô hình sau khi loại bỏ biến đều có ý nghĩa thống kê theo kiểm định F thông qua bảng ANOVA, chứng tỏ tập hợp các biến độc lập có ảnh hưởng đến biến phụ thuộc. Tuy nhiên, mô hình ban đầu bao gồm nhiều biến không có ý nghĩa thống kê riêng lẻ theo kiểm định t, làm cho mô hình trở nên cồng kềnh và hạn chế khả năng diễn giải.

Sau khi loại bỏ các biến không có ý nghĩa thống kê, mô hình hồi quy cuối cùng vẫn giữ được mức độ phù hợp cao với hệ số R^2 hiệu chỉnh không giảm mà còn tăng nhẹ, trong khi giá trị F tăng mạnh, cho thấy sức mạnh giải thích của mô hình được cải thiện. Đồng thời, sai số chuẩn của mô hình giảm nhẹ, phản ánh chất lượng ước lượng tốt hơn.

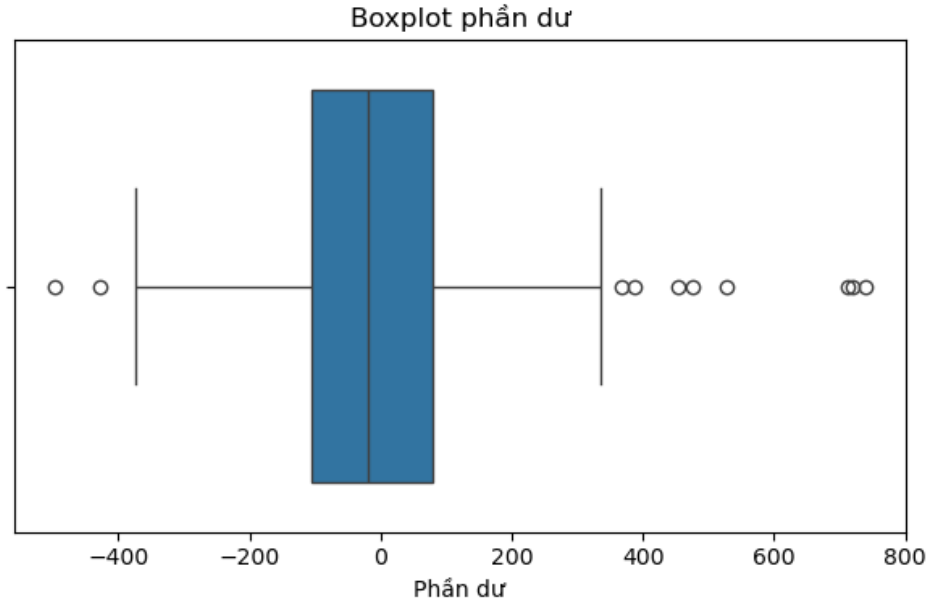
Kết luận Như vậy, mô hình hồi quy sau khi loại bỏ biến được xem là mô hình tối ưu để phân tích và diễn giải, đáp ứng tốt cả về mặt thống kê lẫn tính thực tiễn, và có thể được sử dụng làm cơ sở cho các phân tích và khuyến nghị tiếp theo.

3.6 Phân tích phần dư

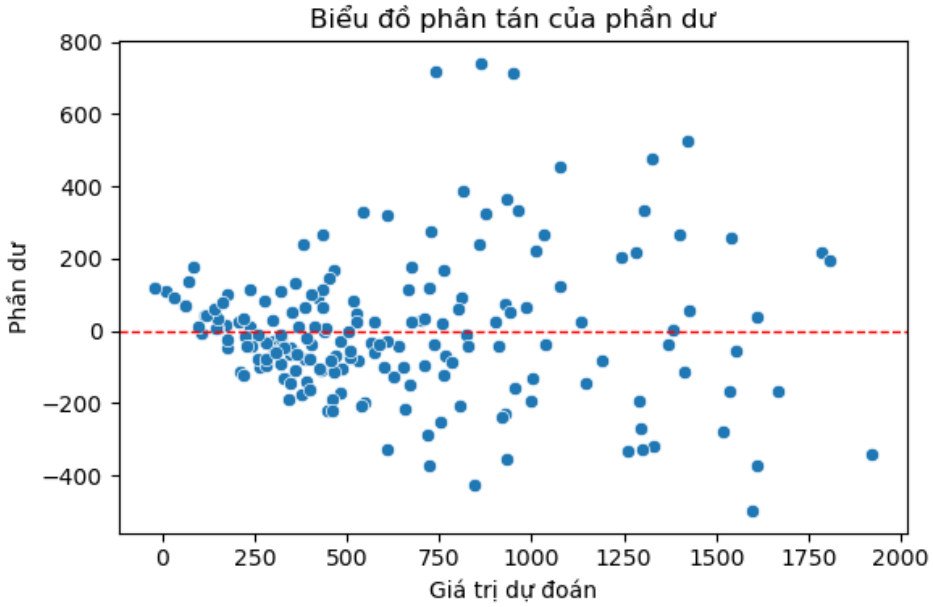
Nhằm kiểm tra các giả định cơ bản của mô hình hồi quy OLS, nghiên cứu tiến hành phân tích phần dư thông qua các biểu đồ Q–Q, boxplot, biểu đồ phân tán và histogram. Các công cụ này giúp đánh giá giả định phân phối chuẩn, phát hiện ngoại lệ và xem xét hiện tượng phương sai không đổi của sai số, từ đó khẳng định mức độ phù hợp và độ tin cậy của mô hình.



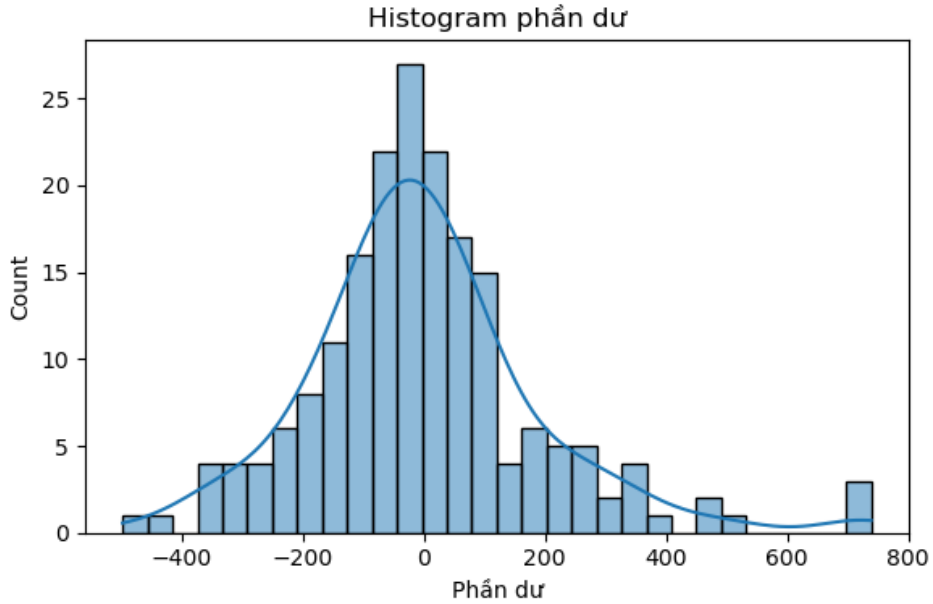
Hình 3.2: Biểu đồ Q-Q



Hình 3.3: Biểu đồ Boxplot phần dư



Hình 3.4: Biểu đồ phân tán phần dư



Hình 3.5: Biểu đồ histogram phần dư

3.7 Mô hình cuối cùng

$$\begin{aligned}\hat{y} = & -409,7006 \\ & + 1605,5712\,\text{Diện tích canh tác (ha)} \\ & + 11,7922\,\text{Lượng giống sử dụng (kg)} \\ & + 1,8763\,\text{Lượng phân urê (kg)} \\ & + 6,0026\,\text{Giá phân urê} \\ & + 0,3950\,\text{Tổng lao động (công)} \\ & + 5,6443\,\text{Tiền công lao động} \\ & - 5,1692\,\text{Giá lúa} + \varepsilon.\end{aligned}$$

3.8 Hàm ý thực tiễn

Từ kết quả ước lượng mô hình hồi quy tuyến tính bội, có thể rút ra một số hàm ý thực tiễn như sau:

- Diện tích canh tác có tác động dương và có ý nghĩa thống kê đến sản lượng lúa, cho thấy việc sử dụng hiệu quả và mở rộng quy mô đất canh tác là yếu tố quan trọng nhằm nâng cao sản lượng.
- Các yếu tố đầu vào như phân bón và thuốc bảo vệ thực vật có ảnh hưởng tích cực đến sản lượng, hàm ý rằng đầu tư hợp lý vào vật tư nông nghiệp góp phần cải thiện năng suất, song cần được thực hiện theo hướng tối ưu để tránh lãng phí và tác động tiêu cực đến môi trường.
- Tiền công lao động và giá các yếu tố đầu vào có ảnh hưởng đáng kể đến sản lượng, cho thấy vai trò của chi phí sản xuất trong quyết định canh tác của nông hộ, từ đó nhấn mạnh sự cần thiết của các chính sách ổn định thị trường lao động và vật tư nông nghiệp.
- Giá lúa có tác động đến sản lượng, phản ánh vai trò của tín hiệu thị trường trong việc định hướng sản xuất, gợi ý rằng các chính sách bình ổn giá và hỗ trợ tiêu thụ sẽ góp phần khuyến khích đầu tư sản xuất lúa.

3.9 Hạn chế của nghiên cứu

Mặc dù đạt được một số kết quả đáng chú ý, nghiên cứu vẫn tồn tại một số hạn chế nhất định như sau:

- Dữ liệu được thu thập tại một thời điểm và trong một phạm vi không gian nhất định, do đó kết quả ước lượng có thể chưa phản ánh đầy đủ sự biến động của sản xuất lúa theo thời gian và giữa các vùng khác nhau.
- Mô hình hồi quy tuyến tính giả định mối quan hệ tuyến tính giữa sản lượng lúa và các biến giải thích, trong khi trên thực tế mối quan hệ này có thể mang tính phi tuyến hoặc chịu tác động của các yếu tố tương tác chưa được xem xét.

- Một số biến quan trọng như điều kiện thời tiết, chất lượng đất, trình độ canh tác hoặc mức độ cơ giới hóa chưa được đưa vào mô hình do hạn chế về dữ liệu, có thể dẫn đến sai lệch trong ước lượng.
- Việc lựa chọn biến dựa trên phương pháp backward elimination phụ thuộc vào tiêu chí thống kê, do đó kết quả có thể nhạy cảm với mức ý nghĩa lựa chọn và chưa phản ánh đầy đủ các cân nhắc về mặt kinh tế học.

3.10 Hướng phát triển nghiên cứu tiếp theo

Dựa trên các kết quả đạt được và những hạn chế còn tồn tại, nghiên cứu trong tương lai có thể được mở rộng theo một số hướng sau:

- Mở rộng tập dữ liệu theo chiều thời gian hoặc không gian nhằm phân tích sự biến động của sản lượng lúa và cải thiện khả năng khái quát hóa của mô hình.
- Xem xét các dạng mô hình phi tuyến hoặc mô hình có biến tương tác để phản ánh tốt hơn mối quan hệ phức tạp giữa sản lượng lúa và các yếu tố đầu vào trong sản xuất.
- Bổ sung các biến giải thích quan trọng như điều kiện thời tiết, chất lượng đất, mức độ cơ giới hóa hoặc trình độ kỹ thuật canh tác nhằm giảm thiểu sai lệch do thiếu biến.
- So sánh hiệu quả của phương pháp backward elimination với các kỹ thuật lựa chọn biến khác như LASSO, Ridge hoặc Elastic Net để lựa chọn mô hình tối ưu hơn.
- Kết hợp phân tích hiệu quả kỹ thuật và hiệu quả chi phí nhằm cung cấp cái nhìn toàn diện hơn về hiệu quả sản xuất lúa của các nông hộ.

Tài liệu tham khảo

1. Tô Văn Ban, *Xác suất thống kê (Dành cho sinh viên các trường kỹ thuật và công nghệ)*, Nhà xuất bản Giáo Dục Việt Nam, 2014.
2. Nguyễn Thị Thu Thủy, *Slide giảng dạy Suy luận thống kê*, Đại học Bách Khoa Hà Nội, 2025.
3. Tống Đình Quỳ, *Giáo trình Xác suất thống kê*, Nhà xuất bản Bách Khoa Hà Nội, 2002.
4. Feng, Q. and Horrace, W. C. (2012). Alternative technical efficiency measures: Skew, bias and scale. *Journal of Applied Econometrics*.