Analyzing indonesian rice farms

Jonas Kernebeck, Alexander Flick, Felix Lehner

07/16/2021

Contents

1 Introduction

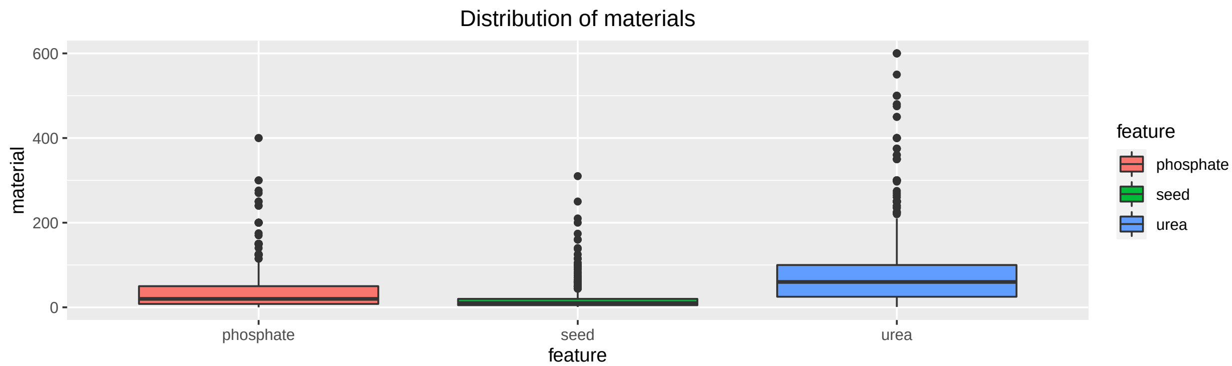
The present data set includes production data for 171 indonesian rice farms. The dataframe contains the following variables:

variable	description	expressions
id	unique identifier for a farm	unique id
time	unique identifier for a specific growing season	1 - 6
size	total production area in hectares	0.01 - 5.322
status	status of property rights	"owner", "share", "mixed"
varieties	rice seed varietes	"trad", "high", "mixed"
bimas	bimas-status of the farmers	"no", "yes", "mixed"
seed	seed in kilogram	1 - 1250 kg
urea	urea in kilogram	1 - 1250 kg
phosphate	phosphate in kilogram	0 - 700 kg
pesticide	pesticide cost in Rupiah	0 - 62600 r
pseed	price of seed in Rupiah per kg	40 - 375 r/kg
purea	price of urea in Rupiah per kg	50 - 100 r/kg
pphosph	price of phosphate in Rupiah per kg	60 - 120 r/kg
hiredlabor	hired labor in hours	1 - 4536 h
famlabor	family labor in hours	1 - 1526 h
totlabor	total labor (excluding harvest labor)	1 - 4774 h
wage	labor wage in Rupiah per hour	30 - 175.35 r/h
goutput	gross output of rice in kg	42 - 20960 kg
noutput	gross output minus harvesting cost	42 - 17610 kg
price	price of rough rice in Rupiah per kg	50 - 190 r/kg
region	region of the farm	

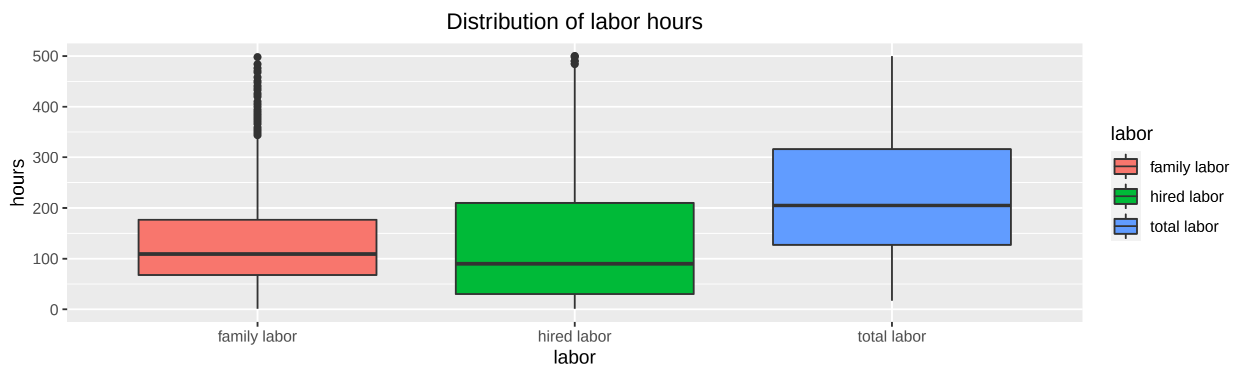
As present in the table, the data set consists of 16 numeric variables and 4 categorical variables. The target variable for the regression modeling will be *goutput*, what represents the gross output of rice in *kg* for the respective rice farm. The variable *noutput* is a linear transformation of *goutput* as it represents *goutput* decreased by the harvesting costs. Therefore it is not used for the modeling because it would violate the multicollinearity assumption. In the following some explorative data analysis will be made to get to get a first impression of the distribution of the individual variables.

1.1 Numerical Variables

The following figure shows boxplots for the used materials and the prices paid for the materials of the respective rice farms. The boxplots for the materials show, that the distribution of all materials is right-skewed. The spread width of seed is the lowest, followed by phosphate and urea. Therefore *urea* also has the highest variance with 16166 followed by *phosphate* with 2264 and *seed* with 2048. The distribution of *urea* indicates that rice farms in Indonesia may use urea very different, caused by e.g. the bimas-status. The bimas program is a rice intensification program by the government to support local rice production by providing high-yield rice seeds as well as technical assistance. If we look at the prices for phosphate *pphosph* and urea *purea*, we can see a slight left-skewed distribution with low variance (75 for *purea* and 86 for *pphosph*). In contrast to that, the prices for seeds scatter much. The distribution of *pseed* is strongly right-skewed as well as the distribution for the rice price *price*. The price for the rice also scatters, but less than *pseed*. The two prices have a correlation of 0.67. Of course, the price of seeds affects the selling price of rice. The prices may fluctuate due to seasonal or regional factors and have an impact on each other. The distribution of labor hours is also slightly skewed to the right. Overall, the dispersion is lowest for the *famlabor*. For *hiredlabor* and *totlabor* we have a similar spread, but *totlabor* has a higher level overall. This is caused by the *hiredlabor* which is a subset of *totlabor*.

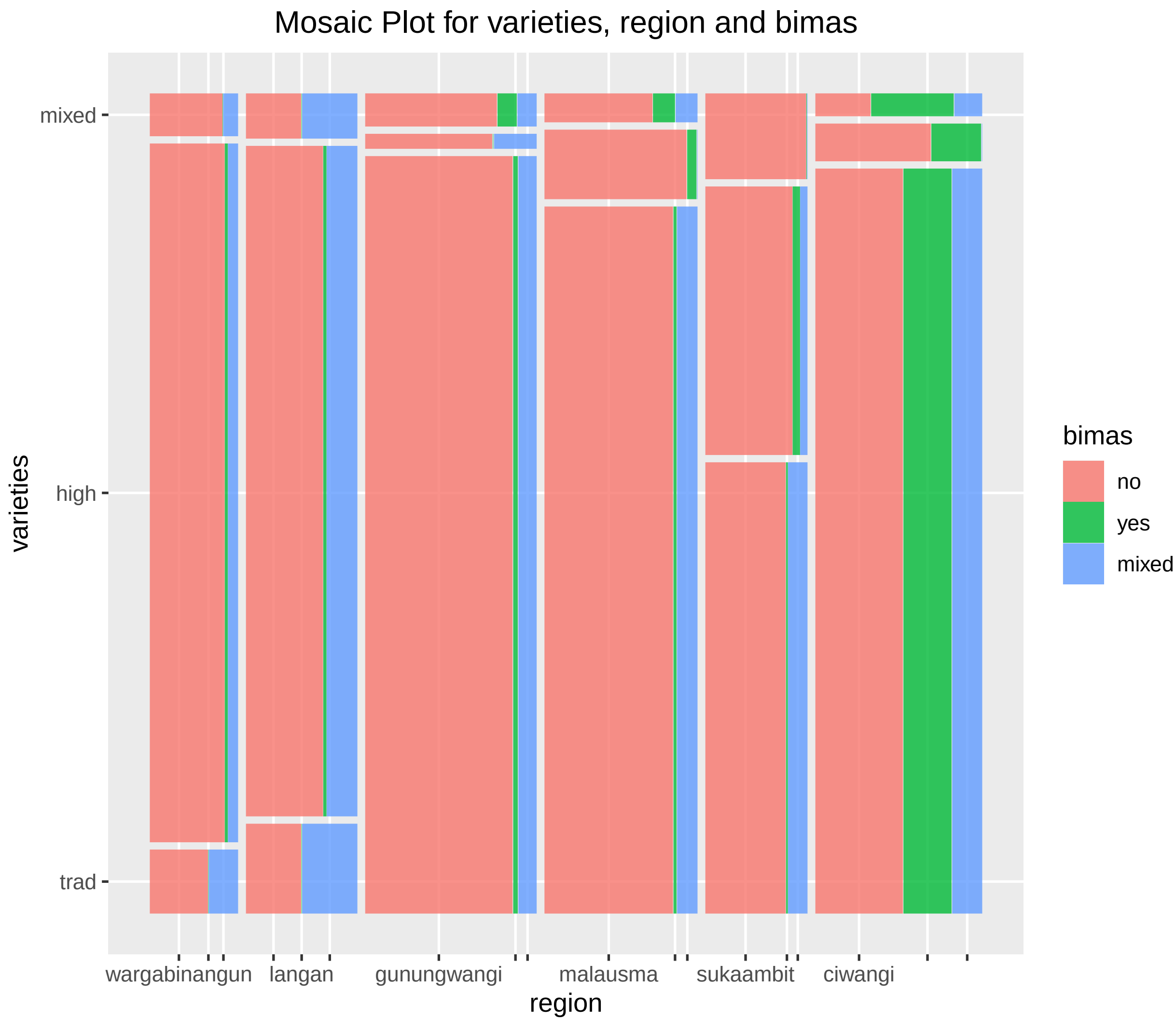

Distribution of materials
600
400
200
0
material
phosphate
seed
urea
feature
feature
phosphate
seed
urea


Distribution of material prices
300
200
100
rupiah_per_kg
price phosphate
price rice
price seed
price urea
feature
feature
price phosphate
price rice
price seed
price urea


Distribution of labor hours
500
400
300
200
100
0
hours
family labor
hired labor
total labor
labor
labor
family labor
hired labor
total labor

1.1 Categorical Variables

The following mosaic plot shows the distribution of of the categorical variables *varietes*, *region* and *bimas*. Overall, all regions are roughly equally represented in the data set. We can detect, that most of the farmers with the bimas status *yes* and *mixed* are located in the region *ciwangi*. The distribution of the different varieties is strongly dependent on the region. While the *high* varieties have the biggest share in the regions *wargabinangun* and *langan*, the *traditional* varieties are dominating the regions *gunungwangi*, *malausma* and *ciwangi*. The *mixed* varities are only used slightly in all regions.



To test wheter the categorical variables have impact on our target variable *goutput*, one- and two-sided anovas are performed. The results of these are summarized in the following table:

formula	F-value	p-value	significant
region	22.981	$< 2e\text{-}16$	yes
varieties	11.764	8.94e-06	yes
bimas	14.817	4.57e-07	yes
region+varietes	3.847	3.96e-05	yes
region+bimas	5.651	2.94e-08	yes
varieties+bimas	0.791	0.531	no
region+varieties+bimas	0.860	0.580	no

The anova outputs show, that all of the categorical variables have a significant effect on *goutput*. The null hypothesis, that the mean of *goutput* is the same across the groups is rejected. The results of the two-sided anovas also show a significant interaction effect on *goutput*. While the interaction effect from the *region* with *varieties* and *bimas* is significant, the interaction effect of *varieties* and *bimas* and the interaction effect of all three variables is not.

Data Preparation

Factor levels

We reset factor levels for later visualization purposes for variable bimas, varieties, etc.

Outliers

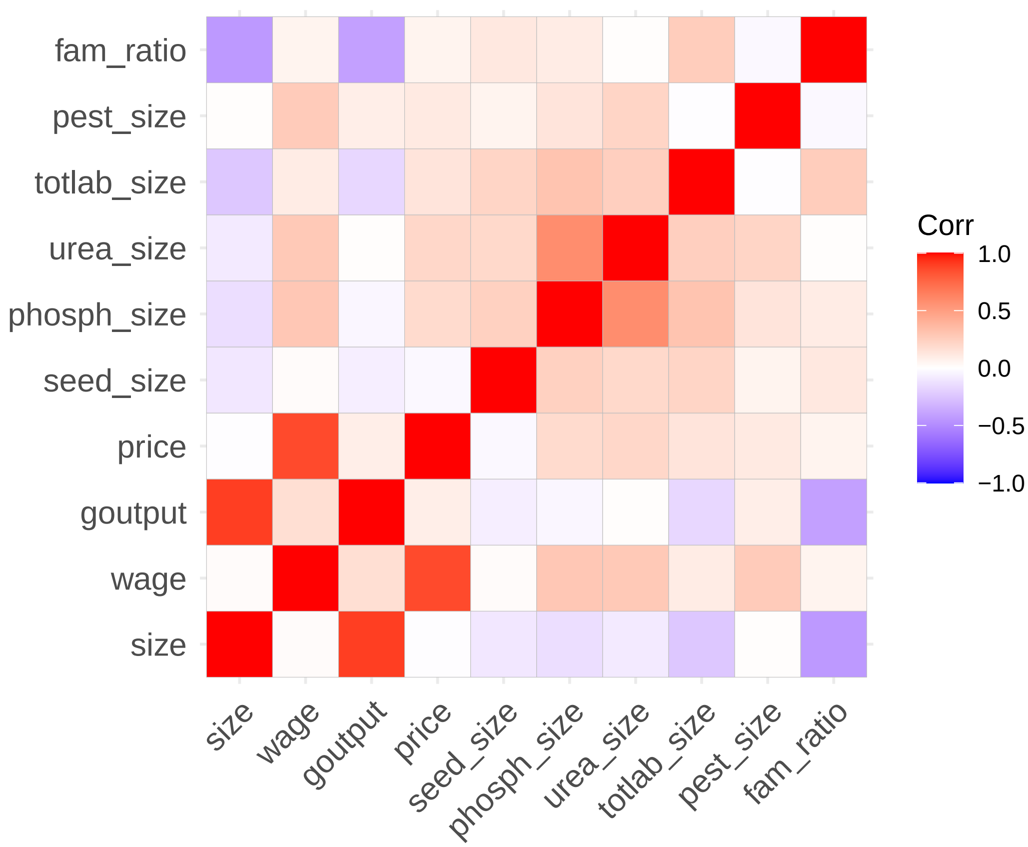
3 Outliers most likely due to typos for observation 110, 947 and 1004

standardizing variables due to correlation

correlation with size the amount of seed, phosphor, urea, totlabour and pesticide is highly dependent on the size of the cultivated area. Thus we will calculate a ratio for each of these variables so that they are becoming independent.

correlation with labour create variable fam_ratio based on the amount of family labour from total labour

Correlation plot



Feature Engineering

wage variable shows a bimodal distribution. We will use this to create a categorical variable.

Model split

60-20-20 split using seed to obtain always the same split

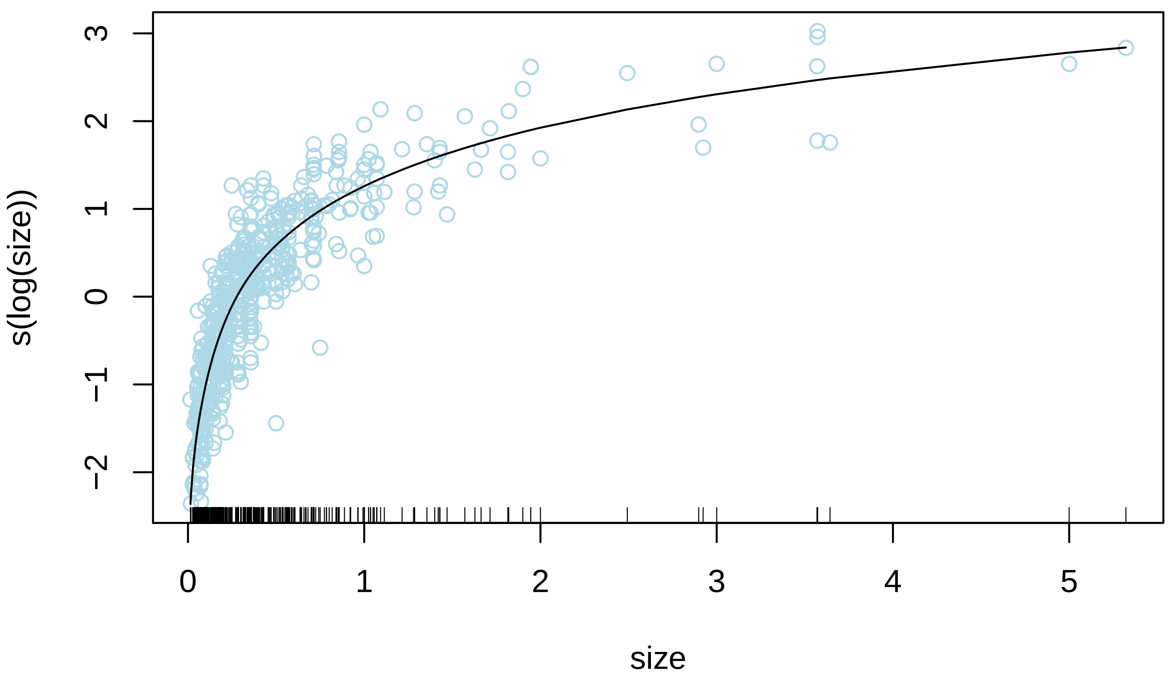
Model metrics

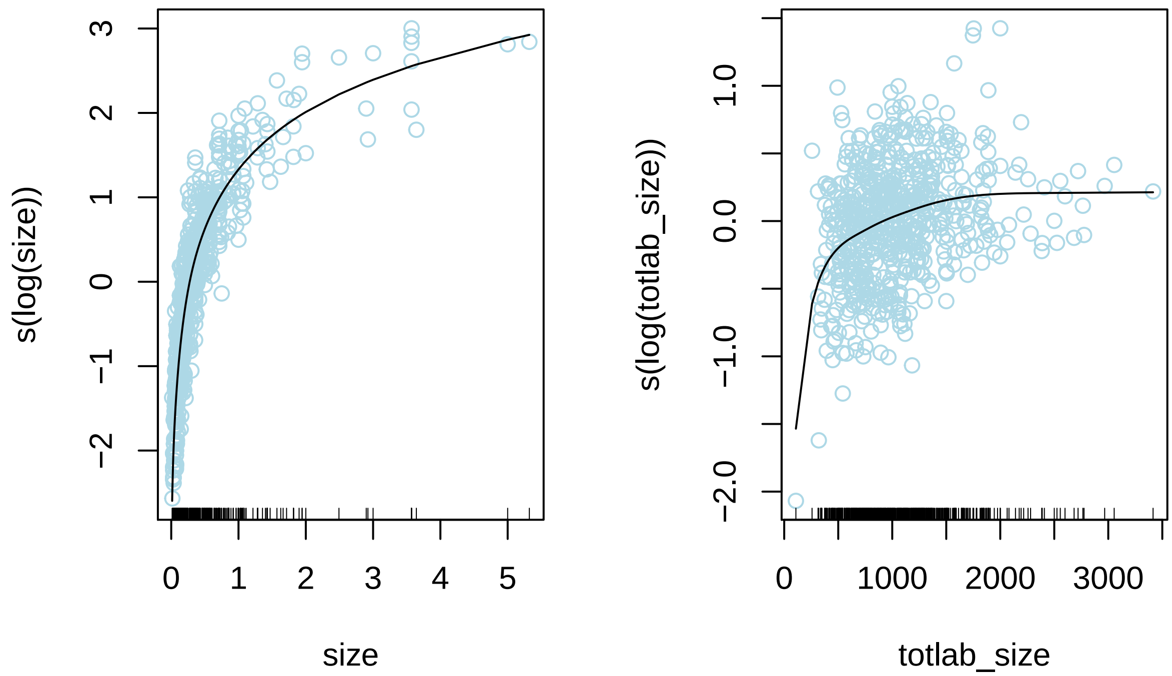
MSE Mean Squared Error: comparing true values with predicted ones

2 First Model

3 Second Model

GAM





Explanation: The left-hand panel indicates that holding totlab_size fixed, goutput increases with size. The right-hand panel indicates that holding size fixed, goutput increases drasticall with the increased proportion of labour per size up to a proportion of 500 h / hectar and then flattens out.

GAM Comparison

table for model selection

Table 3: gam comparison

var	df	MSE.train	MSE.val	dev	aic	p_val_p	p_val_np	df_np
size								
s(size)	609	819023	992780	131.0	806.2	0.0000	0.0000	3
s(log(size))	609	823909	928486	111.3	706.1	0.0000	0.0070	3
labour								
s(log(totlab_size))	605	730930	774765	98.9	641.5	0.0000	0.0084	3
s(totlab_size)	605	732302	766662	100.6	651.8	0.0000	0.0001	3
urea								
s(log(urea_size))	601	765281	734252	78.2	505.5	0.0000	0.0000	3
s(urea_size)	601	769145	735393	78.3	506.3	0.0000	0.0001	3
s(log(urea_size), df = 2)	603	754823	744796	79.2	509.2	0.0000	0.0000	1
phosphor								
s(log(phosph_size + 1))	599	672024	549454	74.6	480.3	0.0000	0.0000	3
s(log(phosph_size + 1), df = 2)	602	655449	570382	76.5	489.5	0.0000	0.0000	1
s(phosph_size)	599	664319	558066	74.9	482.4	0.0000	0.2493	3
seed								
s(log(seed_size))	595	666719	541371	72.4	469.6	0.0000	0.0532	3
s(seed_size)	595	671676	550941	72.2	468.0	0.0002	0.0213	3
pesticide								
s(pest_size)	591	632672	509940	70.2	459.2	0.0001	0.4334	3
s(pest_size, df = 2)	593	637740	512192	70.3	456.0	0.0001	0.1711	1
price								
s(price)	587	524477	409469	66.2	431.3	0.0006	0.0000	3
wage								
s(wage)	583	481786	417154	63.9	416.9	0.0000	0.1172	3
wage_cat>100	586	506621	381141	64.1	412.7	0.0000	NA	NA
family labour								
s(fam_ratio)	582	480148	411697	63.0	410.1	0.0074	0.3896	3
categorical variables								
bimasyes	580	482069	438276	61.9	403.2	0.0026	NA	NA
varietieshigh	578	474691	415467	61.4	402.4	0.1031	NA	NA
statusmixed	578	464596	456488	61.5	404.0	0.2121	NA	NA
regionciwangi	575	473663	528887	60.0	395.0	0.0003	NA	NA

gam2 is better, so we will use logarithm of size

plot.gam for urea: slope at beginning of urea size => reduce df?

phosphate: this variable has lots of zeros as values. So we can not use directly log transformation because log of 0 is -inf. According to : https://discuss.analyticsvidhya.com/t/methods-to-deal-with-zero-values-while-performing-log-transformation-of-variable/2431 log(x+1) transformation is the best way to avoid errors created by log transformation and is widely used among data scientists. So we will use this approach.

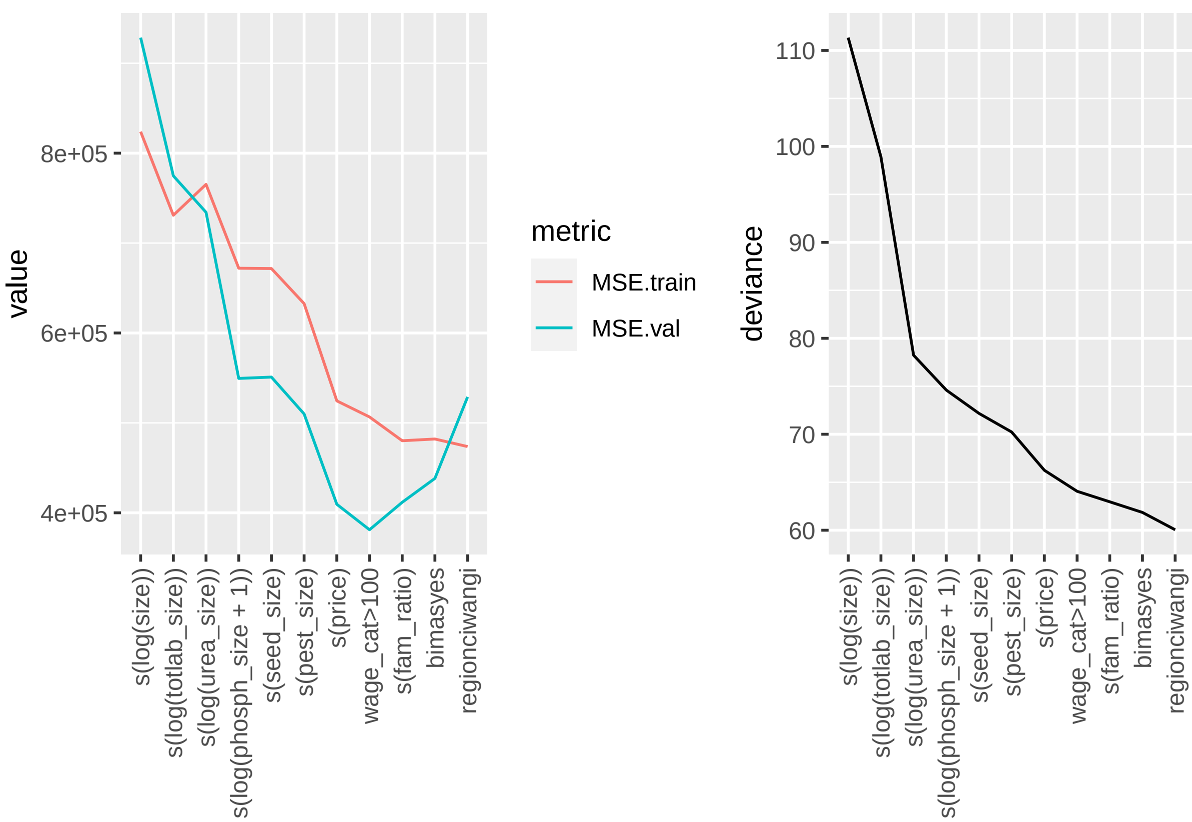
pesticide: Anova for Non-Parametric Effects p-value = 0.0001351187 => keep variable in model Anova for Parametric Effects p-value=0.433418 => reduce df of pest_size

check categorical variables

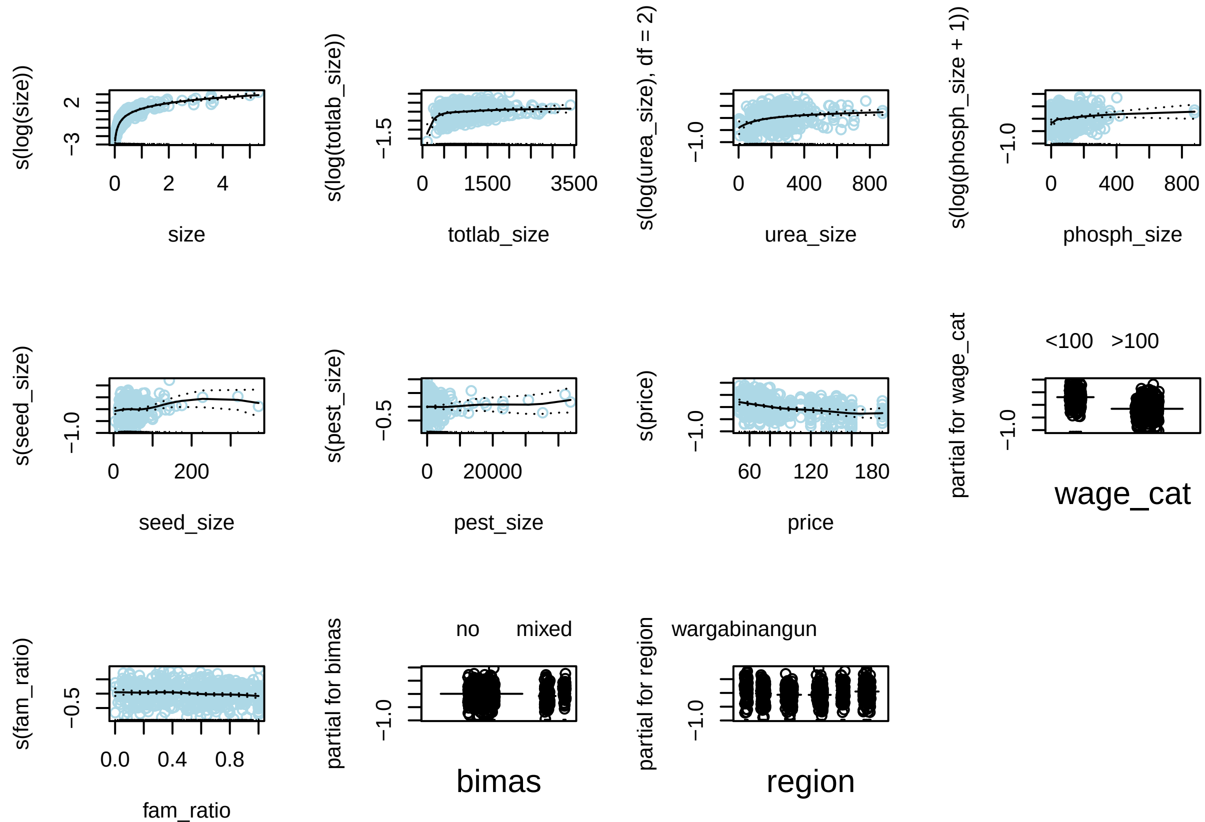
varieties: p-value 0.103104 => not enough evidence that this variable is significantly important based on a 5% significance level. BUT MSE decreases!

status: p-value 0.212109 => not important

MSE and deviance visualization



Final model visualization



4 Comparision

5 Conclusion