Survey of genetic technology knowledge and use among beef cattle producers

Harly Durbin

Contents

Overview/introduction to the data

This survey data comes from a collaboration between the MU Animal Genomics group and the Department of Agricultural Education & Leadership. The survey was designed to evaluate Missouri beef cattle producers' use of and attitude towards genetic (EPDs: "estimated progeny differences" for economically relevant traits calculated based on pedigree relatedness) and genomic (GE-EPDs: EPDs augmented with genomic relatedness data obtained from DNA genotyping) technologies when making selection and breeding decisions. In the first phase of the survey, beef cattle producers across the country who subscribe to BEEF Magazine were surveyed online. In the next phase of the survey, beef cattle producers at 11 sale barns were surveyed over the course of winter 2017. In 2017, Missouri ranked second in the nation for total calf crop and beef cows that have calved, and this survey and the results of these analyses will help to inform future university extension efforts. Questions were as follows:

---

1. Out of 100%, when choosing breeding stock how much do you use EPDs and how much do you use visual inspection?
2. How often do you use the following items when selecting breeding animals?
   - EPDs
   - Genomic tests
   - Visual inspections

- Breed association staff
- Breed/semen catalogs
- Semen distributors
- Trusted breeder from whom you've previously purchased stock
- Other producers

3. How often do you use consider the following EPD indexes when selecting breeding stock?
    - CED
    - BW
    - WW
    - YW
    - MILK
    - CEM
    - REA
    - MARB
    - Breed-specific value indexes ($W, $F, $TI, etc.)
    - ACC

Where:

- **CED:** calving ease direct
    - "…expressed as percentage of unassisted births, with a higher value indicating greater calving ease in first-calf heifers. It predicts the average difference in ease with which a sire's calves will be born when he is bred to first-calf heifers." (American Angus Association)
- **CEM:** calving ease maternal
    - "…expressed in percentage unassisted births with a higher value indicating greater calving ease in first-calf daughters. It predicts the average ease with which a sire's daughters will calve as first calf heifers when compared to daughters of other sires." (American Angus Association)
- **BW:** birth weight
- **WW:** weaning weight
- **YW:** yearling weight
- **MILK:** maternal milk yield
- **REA:** ribeye area
    - "Expressed in square inches, [REA] is a predictor of the difference in ribeye area of a sire's progeny compared to progeny of other sires." (American Angus Association)
- **MARB:** marbling
    - MARB is "expressed as a fraction of the difference in USDA marbling score of a sire's progeny compared to progeny of other sires" (American Angus Association), where marbling score is based upon degree of intramuscular fat marbling.
- **Breed specific value indexes:** allow for comparison of animals based upon multiple weighted traits
- **ACC:** accuracy, EPD reliability where 1 indicates higher reliability. "Accuracy is impacted by the number of progeny and ancestral records included in the analysis." (American Angus Association)

4. Which of the following would prevent you from using EPDs?
    - EPDs are difficult to read
    - Inconsistency between breed EPDs
    - Difficult to understand difference between breed baseline and animal reports
    - Too much overlap in composite data
    - Too many bull EPDs to comb through
    - EPDs are not available for the bulls I purchase
    - EPDs have not worked in my situation
    - EPDs don't accurately reflect genetic merit
    - EPDs don't reflect all important factors in selecting breeding animals
5. How important are the following factors in choosing breeding stock for your farm?
    - EPDs
    - Visual inspection

- Price
- Previous use of specific animal/genetic line
- Purebred breeder recommendation
- Other producers

6. Who makes the breeding decisions on your farm?
    - Me
    - My grandparents
    - My parents
    - My spouse
    - My siblings
    - My sons/daughters
    - My farm manager (not me)
7. How do you learn about new breeding information and new industry technologies?
    - Trade publications/magazines
    - Breed associations
    - Local extension office/agent
    - Local ag teacher
    - Online resources
    - Veterinarian
    - Semen Salesman (ABS, Select Sires, Genex, Cattle Visions, etc.)
    - Other producers
8. How big was your cattle operation (total cows, bulls, calves) in 2016?
9. How old are you?
10. Are you male or female?

Hypotheses

There are 3 main hypotheses which will be tested a number of different ways.

1. Older producers tend to be less progressive and therefore rely more heavily on visual appraisal when making breeding decisions.
2. Size of operation is predictive of EPD/GE-EPD use.
3. There is a relationship between reported EPD usage and reported barriers to EPD usage.

Data import and cleaning

- Surveys with reported age < 18 years were removed.
- Surveys with reported size of operation < 1 were removed.
- In cases where only one of the question 1 (visual appraisal vs. EPD usage) elements was completed, the other element was imputed by subtracting the completed element from 100.
    - Some respondents provided answers to question 1 that totaled to > 100. A new, standardized variable (`epd_usage_stand`) was created by dividing `usage_percent_epd` by the total of `usage_percent_epd` and `usage_percent_visual`.
- Responses that provided a range of numbers for `size_of_operation` were set to NA.

Data import and cleaning code below:

```
#Specify na = "N/A" to change "N/A" strings to NA
#trim leading and trailing whitespace in cells
#Skip first "question" header column
df1 <- read_excel(
  "~/Box Sync/ProducerSurvey/Copy of Tummons Questionaire.xlsx",
```

```
  na = "N/A",
  trim_ws = TRUE,
  skip = 1,
  col_types = "text"
) %>%
  #prefix each column with which question it pertains to
  rename_at(3:4, funs(paste0("usage.", .))) %>%
  rename_at(5:12, funs(paste0("resources.", .))) %>%
  rename_at(13:22, funs(paste0("consider_epd.", .))) %>%
  rename_at(23:31, funs(paste0("barrier_epd.", .))) %>%
  rename_at(32:37, funs(paste0("important_factors.", .))) %>%
  rename_at(38:44, funs(paste0("decision_makers.", .))) %>%
  rename_at(45:52, funs(paste0("learn.", .))) %>%
  #rename columns in R friendly way (no spaces etc)
  janitor::clean_names() %>%
  #remove completely empty rows and columns
  #(can sometimes result from Exel formatting)
  janitor::remove_empty() %>%
  #remove online entries since I got full dataset
  filter(sale_barn != "Online")

responses <- read_csv(
  "~/Box Sync/ProducerSurvey/181203_online_data.csv",
  trim_ws = TRUE,
  skip = 3,
  col_names = FALSE,
  col_types = cols(.default = "c"),
  na = c("N/A", "na", "n/a", "", " ", "NA")
) %>%
  select(9, 66:67, 18:65, 71:73) %>%
  mutate(sale_barn = "Online") %>%
  select(sale_barn, everything()) %>%
  setNames(object = ., colnames(df1)) %>%
  bind_rows(df1) %>%
  #janitor names things weird sometimes
  rename(important_factors_epds = important_factors_ep_ds) %>%
  #if they provided a size range for size, set to NA
  mutate(size_of_operation = if_else(str_detect(size_of_operation,
                                                "/ | -"),
                                     "0",
                                     size_of_operation)) %>%
  #remove words from size_of_operation
  mutate(size_of_operation = str_remove(size_of_operation, "[[:alpha:]]+")) %>%
  #make age and size of operation numeric
  mutate_at(.vars = vars(usage_percent_epd:age), as.numeric) %>%
  mutate(sex = if_else(sex == "1", "M", sex),
         sex = if_else(sex == "2", "F", sex)) %>%
  #Capitalize sex (all M or F)
  mutate(sex = R.utils::capitalize(sex)) %>%
  #Create a column for online vs. in-person
  mutate(medium = if_else(sale_barn == "Online" |
                            sale_barn == "Pilot",
                          "online",
```

```
                          "in person")) %>%
  #make sale barn, #, sex, medium, factors
  mutate_at(c("sale_barn",
              "survey_number",
              "sex",
              "medium"),
            as.factor) %>%
  #If one of % EPD vs visual is empty, fill in
  mutate(usage_percent_epd = if_else(
    is.na(usage_percent_epd) & !is.na(usage_percent_visual),
    100 - usage_percent_visual,
    usage_percent_epd
  )) %>%
  mutate(
    usage_percent_visual = if_else(
      is.na(usage_percent_visual) & !is.na(usage_percent_epd),
      100 - usage_percent_epd,
      usage_percent_visual
    )
  ) %>%
  #Some people gave EPD + visual percentages
  #that don't add up to 100: standardize
  mutate(epd_usage_stand =
           usage_percent_epd / (usage_percent_epd + usage_percent_visual)) %>%
  #One response with a 7 when only 6 options: remove
  mutate(learn_vet = replace(learn_vet, 7, NA)) %>%
  #Remove surveys where age is < 18 years
  #Remove surveys where size of operation = 0
  filter(age > 17,
         size_of_operation != 0,
         )
```

```
responses_long <- responses %>%
  melt(
    id = c(
      "sale_barn",
      "medium",
      "survey_number",
      "size_of_operation",
      "age",
      "sex",
      "usage_percent_epd",
      "usage_percent_visual",
      "epd_usage_stand"
    ),
    na.rm = FALSE
  ) %>%
  mutate(variable = as.character(variable)) %>%
  rename(response = value)
```

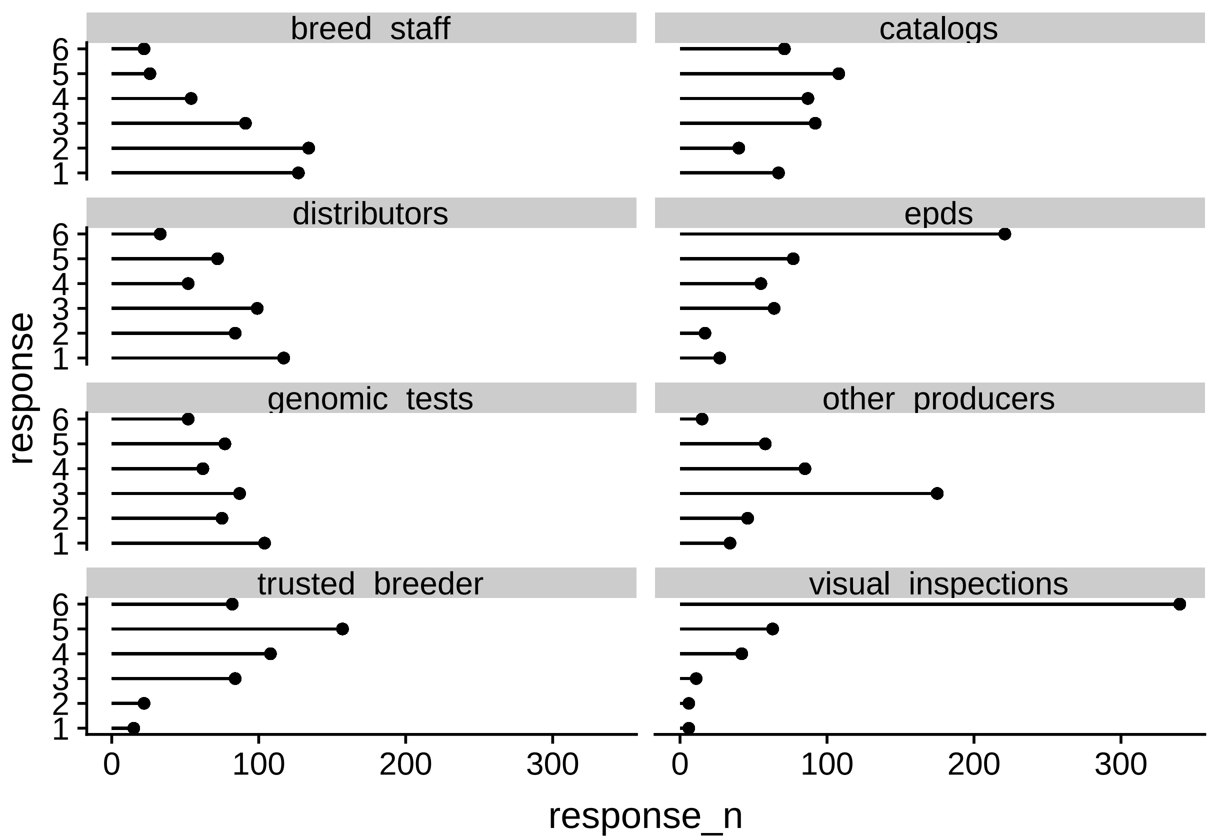
Exploratory data analysis and summarization

How often do you use the following items when selecting breeding animals?

- With possible answers:
  - 1: Never
  - 2: Rarely
  - 3: Occasionally
  - 4: Frequently
  - 5: Often
  - 6: Always
- Treat as continuous

Mean scores were as follows:

variable	mean_score
resources_visual_inspections	5.500000
resources_epds	4.737527
resources_trusted_breeder	4.316239
resources_catalogs	3.735484
resources_other_producers	3.319613
resources_genomic_tests	3.194748
resources_distributors	2.949672
resources_breed_staff	2.524229



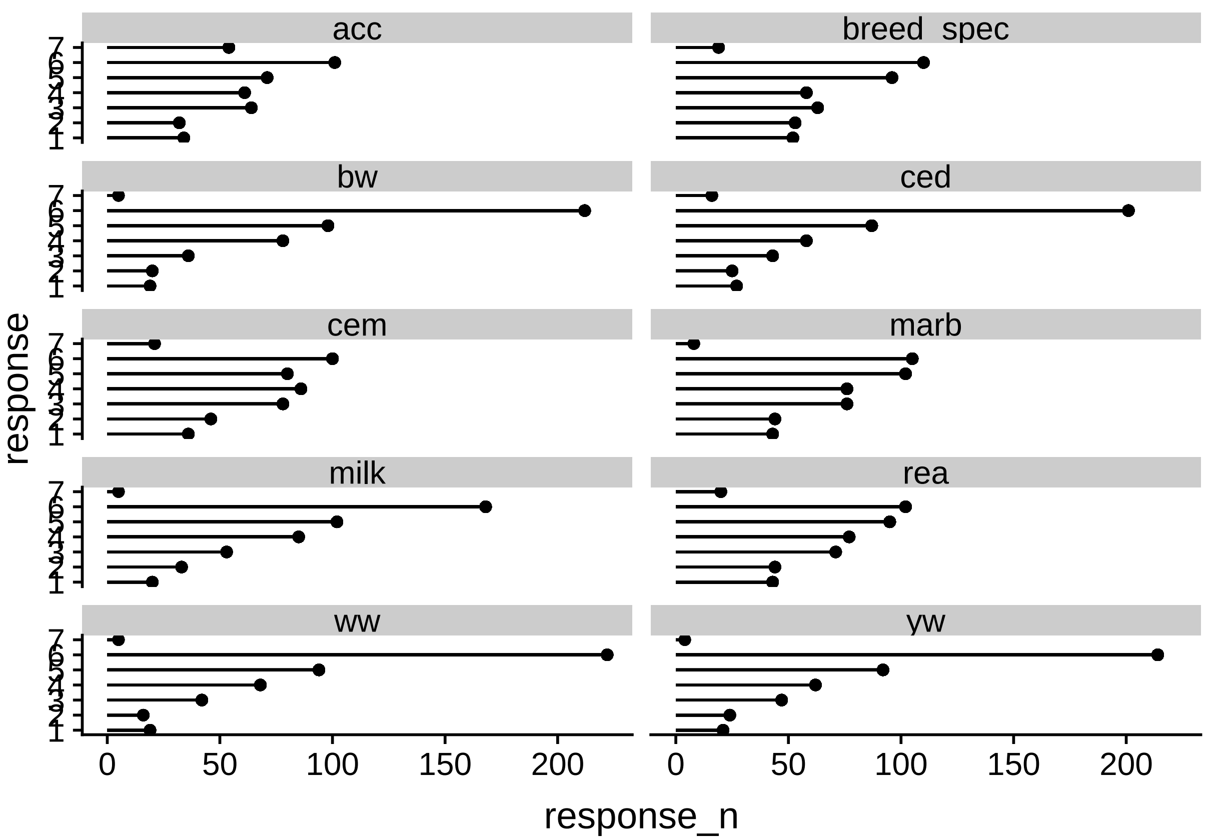
Unsurprisingly, surveyed producers rely most often on visual ispection with EPDs in second.

How often do you use consider the following EPD indexes when selecting breeding stock?

- With possible answers:
  - 1: Never
  - 2: Rarely
  - 3: Occasionally
  - 4: Frequently
  - 5: Often
  - 6: Always
  - 7: I do not know what this means

Excluding "7: I do not know what this means", mean scores for each EPD were as follows:

variable	mean_score
consider_epd_ww	4.882863
consider_epd_bw	4.840173
consider_epd_yw	4.786956
consider_epd_ced	4.714286
consider_epd_milk	4.561822
consider_epd_acc	4.118457
consider_epd_marb	4.042601
consider_epd_rea	4.025463
consider_epd_cem	4.004695
consider_epd_breed_spec	3.979167



Growth traits (weaning weight, birth weight, and yearling weight) are most often considered when making breeding decisions. This is unsurprising since most of the producers surveyed are likely cow/calf producers that sell cattle at weaning and are paid on the pound.

For each EPD index, what is the distribution of responses? How many responses are 7s ("I do not know what this means")? What percentage of the total responses for each variable are 7s?

response	variable	variable_n	response_n	freq
7	consider_epd_acc	417	54	0.1294964
7	consider_epd_cem	447	21	0.0469799
7	consider_epd_rea	452	20	0.0442478
7	consider_epd_breed_spec	451	19	0.0421286
7	consider_epd_ced	457	16	0.0350109
7	consider_epd_marb	454	8	0.0176211
7	consider_epd_ww	466	5	0.0107296
7	consider_epd_milk	466	5	0.0107296
7	consider_epd_bw	468	5	0.0106838
7	consider_epd_yw	464	4	0.0086207

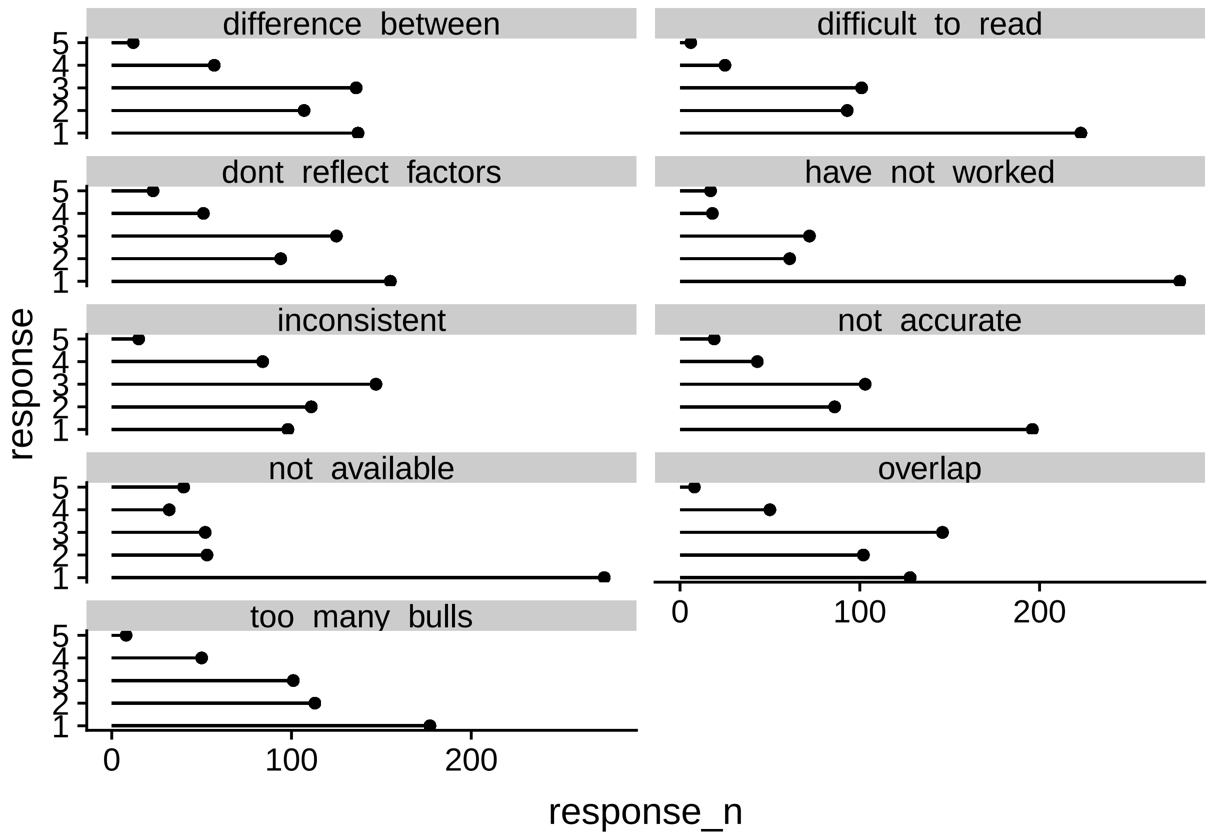
Of the 417 respondents that provided responses for the question about EPD accuracy, ~13% responded with "7: I do not know what this means". However, these results may be underestimated since many fewer people answered the question about accuracy than answered the questions about other EPD idexes. Since accuracy is not actually an EPD itself but a property of EPDs, this question may have been misleading. Still, correctly communicating how accuracy/increased observations affects EPD consistency is vital to dispelling producer mistrust of genetic technology, so this is an important insight.

Which of the following would prevent you from using EPDs?

- With possible answers:
  - 1: Not a barrier
  - 2: Small barrier
  - 3: Moderate barrier
  - 4: Large barrier
  - 5: Prohibitive barrier

Mean scores were as follows:

variable	mean_score
barrier_epd_inconsistent	2.575824
barrier_epd_difference_between	2.331849
barrier_epd_overlap	2.327189
barrier_epd_dont_reflect_factors	2.314732
barrier_epd_not_accurate	2.111857
barrier_epd_too_many_bulls	2.106904
barrier_epd_not_available	1.915743
barrier_epd_difficult_to_read	1.879464
barrier_epd_have_not_worked	1.733184



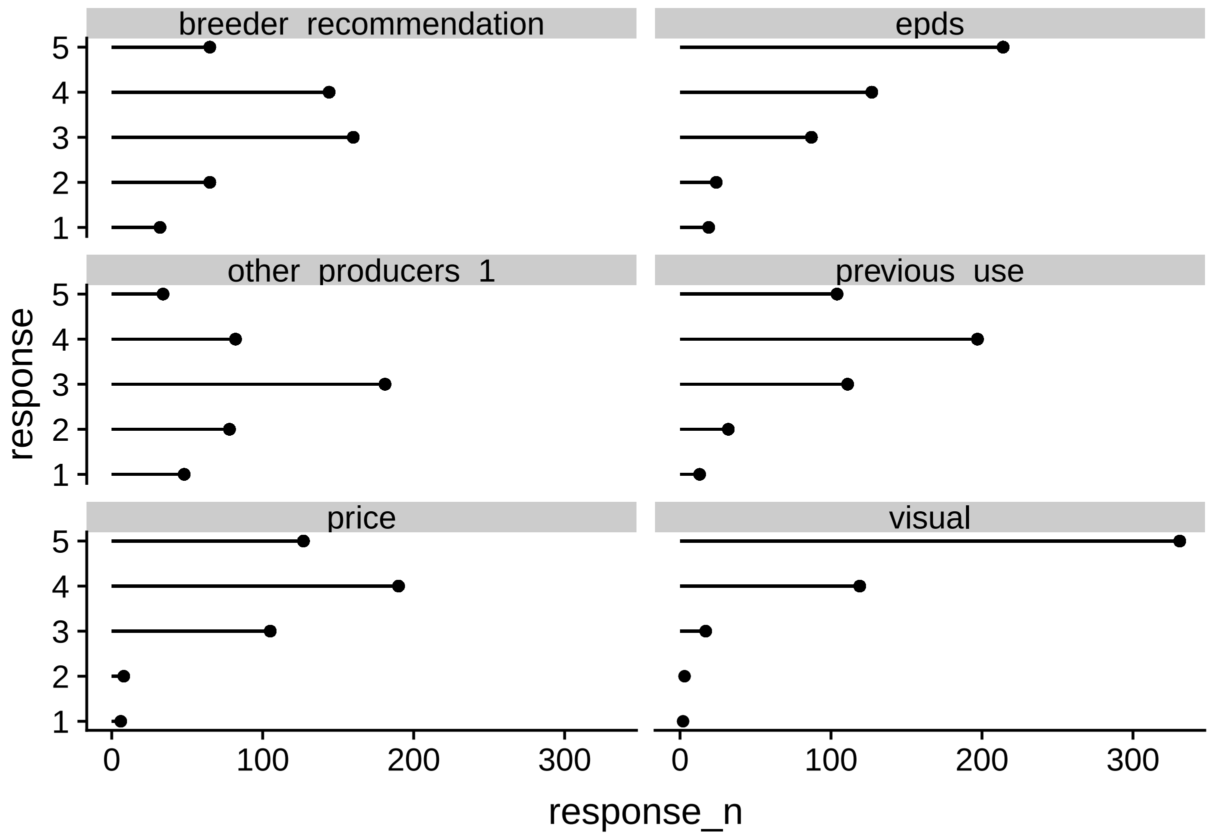
EPD inconsistency is reported as the largest barrier to EPD usage. However, the range of the mean scores for all barriers is quite small (0.84).

How important are the following factors in choosing breeding stock for your farm?

- With possible answers:
  - 1: Not important
  - 2: Of little importance
  - 3: Somewhat important
  - 4: Important
  - 5: Very important

Mean scores were as follows:

variable	mean_score
important_factors_visual	4.639831
important_factors_epds	4.046709
important_factors_price	3.972477
important_factors_previous_use	3.759300
important_factors_breeder_recommendation	3.311159
important_factors_other_producers_1	2.943262

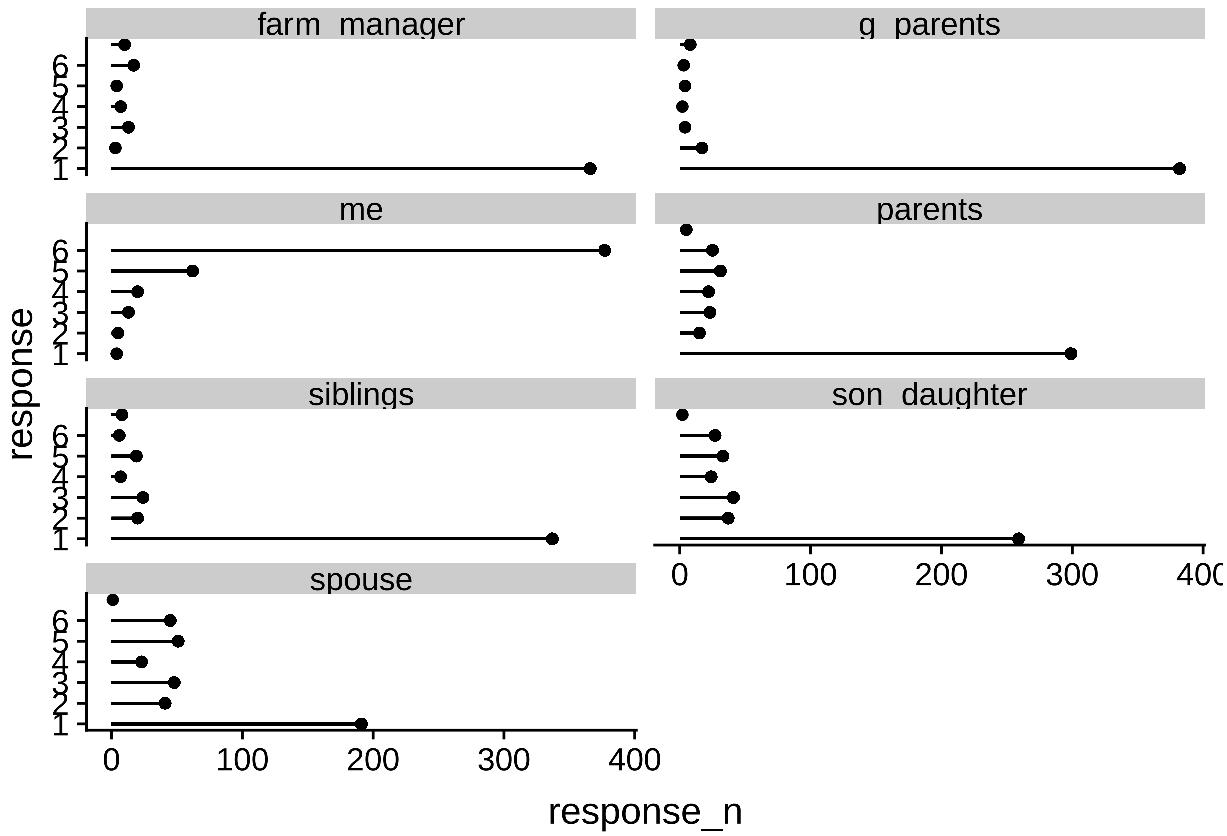


Who makes the breeding decisions on your farm?

- With possible answers:
  - 1: Never/doesn't apply
  - 2: Rarely
  - 3: Occasionally
  - 4: Frequently
  - 5: Often
  - 6: Always

Mean scores were as follows:

variable	mean_score
decision_makers_me	5.623701
decision_makers_spouse	2.602500
decision_makers_son_daughter	2.111111
decision_makers_parents	1.966667
decision_makers_siblings	1.577197
decision_makers_farm_manager	1.502381
decision_makers_g_parents	1.261905



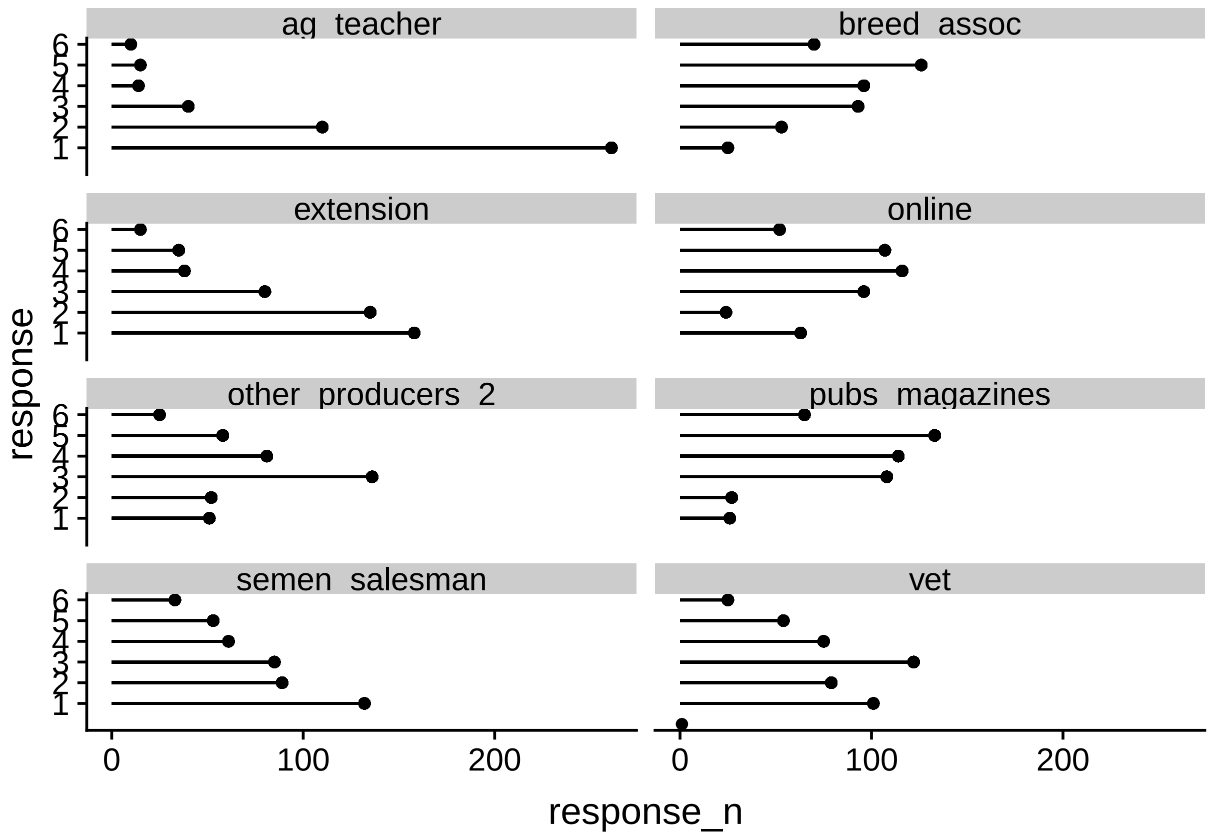
Overwhelmingly, surveyed producers list themselves as the main decision makers on their farms.

How do you learn about new breeding information and new industry technologies?

- With possible answers:
  - 1: Never
  - 2: Rarely
  - 3: Occasionally
  - 4: Frequently
  - 5: Often
  - 6: Always

Mean scores were as follows:

variable	mean_score
learn_pubs_magazines	4.048626
learn_breed_assoc	3.982721
learn_online	3.733624
learn_other_producers_2	3.292804
learn_vet	2.943107
learn_semen_salesman	2.807947
learn_extension	2.353579
learn_ag_teacher	1.760000



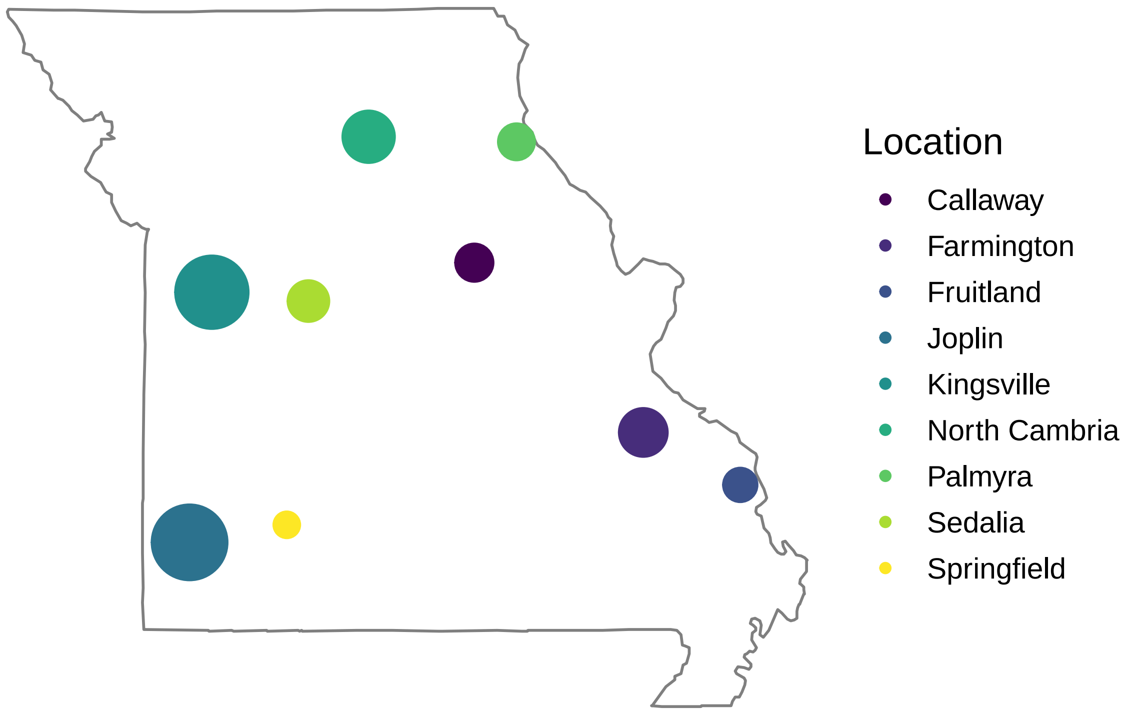
On average, producers rely the least on university extension and ag teachers to learn about technologies and the most on maggazines/publications. However, these results may be biased as the online phase of the survey was facilitated with the help of BEEF Magazine.

Demographics

Location and medium

After data cleaning, 257 surveys ere collected in person and 249 surveys were collected online. Of surveys collected in person at sale barns, Joplin Regional Stockyards and Kingsville Livestock Auction are the most highly represented.

medium	sale_barn	n
online	Online	224
in person	Joplin	62
in person	Kingsville	58
in person	North Cambria	29
in person	Farmington	25
online	Pilot	25
in person	Sedalia	18
in person	Callaway	15
in person	Palmyra	14
in person	Fruitland	12
in person	Springfield	7



Age and sex

The mean age of all respondents is ~ 51 years.

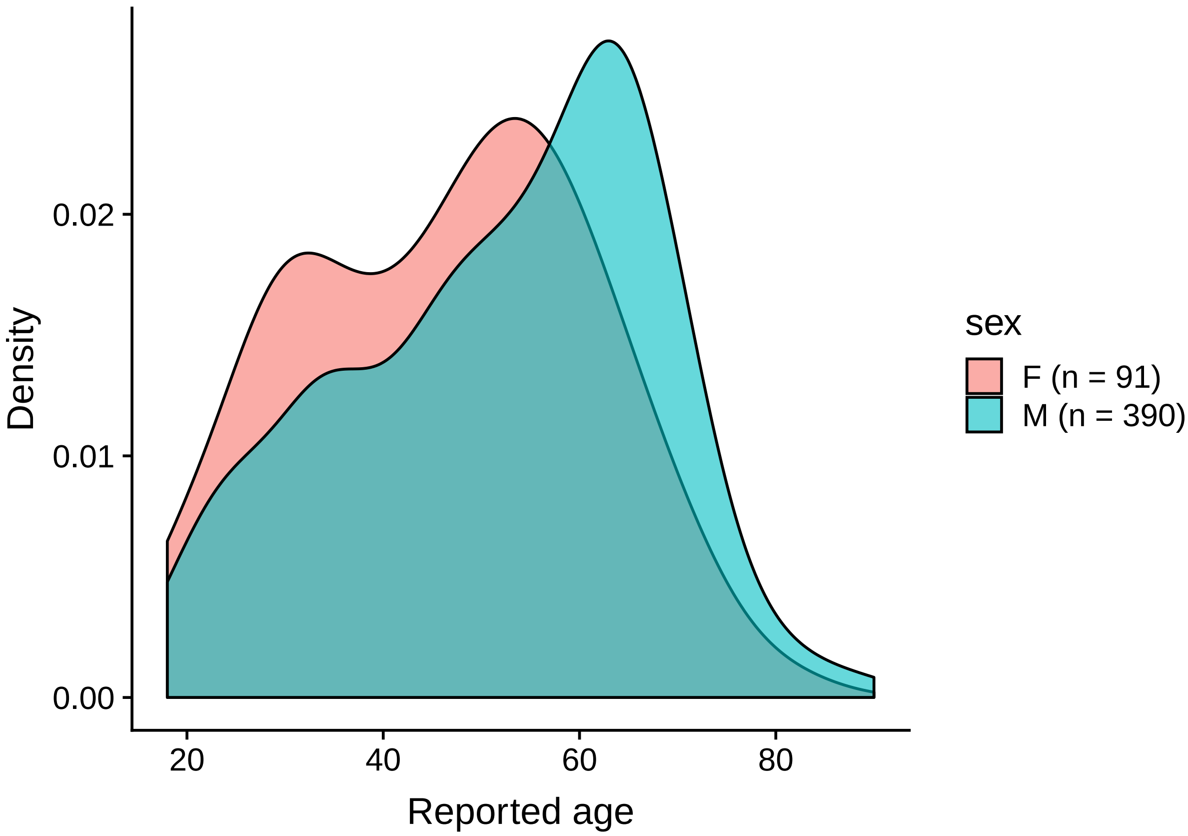
mean_age	median_age
50.98364	53

Online respondents were ~ 3 years younger than sale barn respondents on average.

medium	mean_age	median_age
in person	52.90833	56.5
online	49.12851	50.0

Overall, producers that identified as women tended to be slightly younger than producers that identified as men. However, only ~ 19% of total age-reporting respondents identified as women, the majority of which came from online responses. Of in person responses, 9% were women. Of online responses, 28% were women

sex	medium	n	mean_age	median_age
M	in person	211	52.84360	57.0
M	online	179	50.55866	52.0
F	online	70	45.47143	46.5
F	in person	21	49.09524	54.0



Size of operation

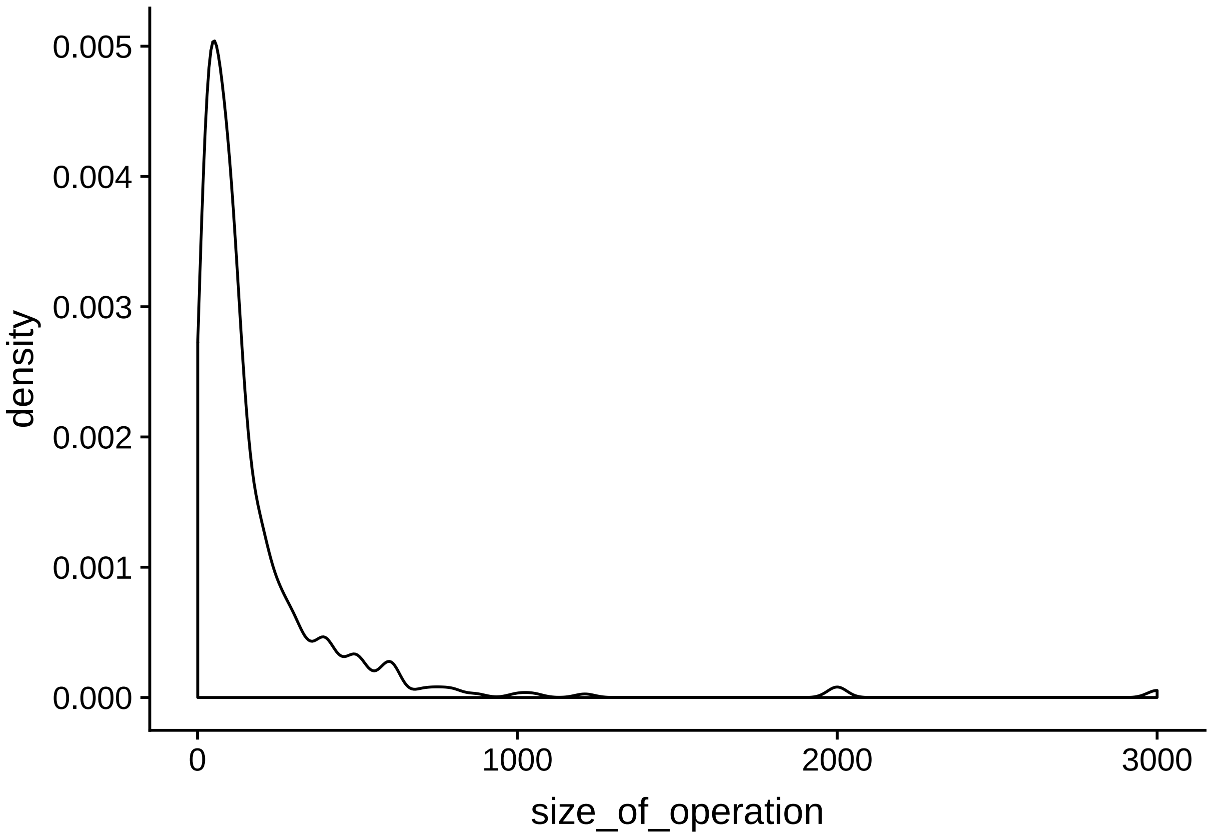
There is a huge spread in reported size of operation. Of all respondents, the median size of operation in 100. This is much higher than USDA ERS estimates of average beef cattle operation size, which may suggest some bias in our results.

mean_size	median_size	sd_size	min_size	max_size	n_reported
176.0777	100	287.9471	1	3000	489

Summary statistics broken down by medium and sale barn are as follows:

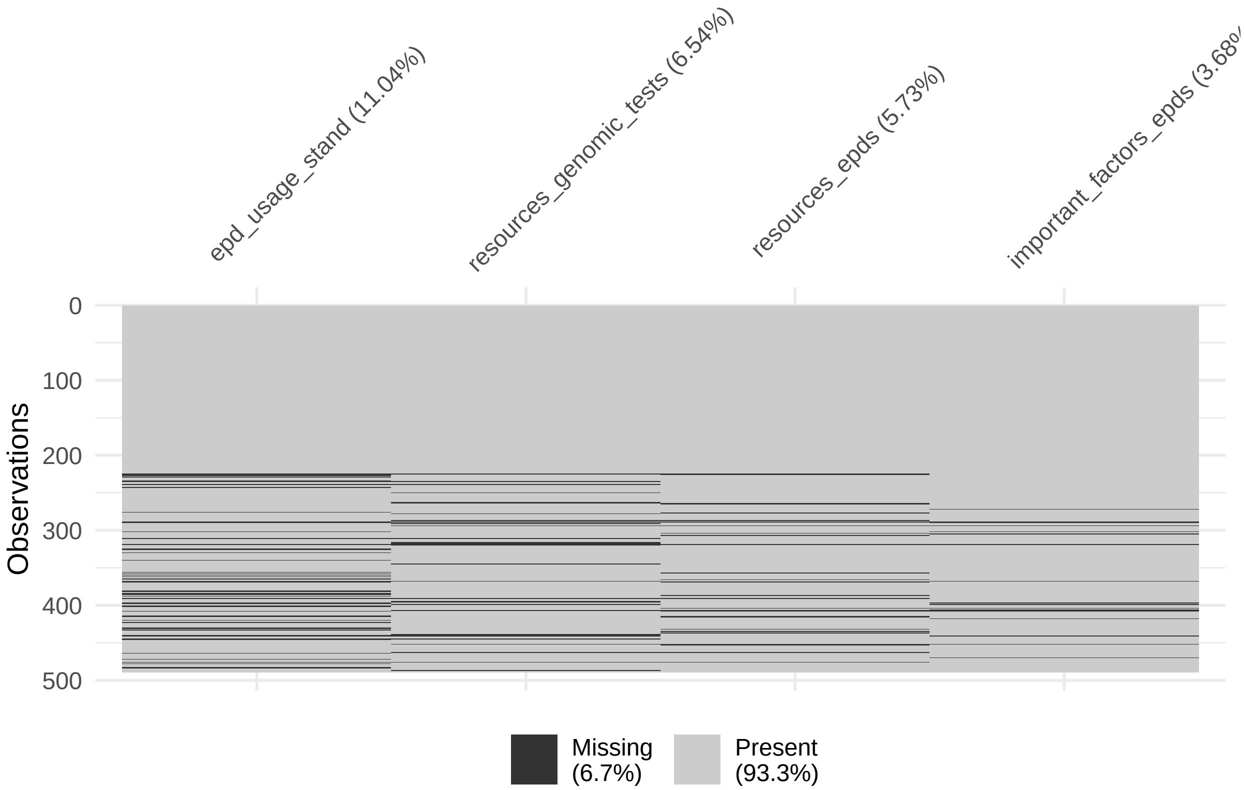
sale_barn	mean_size	median_size	sd_size	min_size	max_size	n_reported
Farmington	243.3200	125.0	284.33515	15	1211	25
Joplin	208.7097	122.5	382.27882	12	3000	62
North Cambria	203.1724	115.0	233.57089	15	750	29
Kingsville	162.5172	100.0	171.87391	1	800	58
Online	181.2143	100.0	327.76688	5	3000	224
Fruitland	114.0000	95.5	78.33146	16	280	12
Sedalia	121.9444	95.0	109.05340	20	400	18
Callaway	117.4000	80.0	122.74643	22	400	15
Palmyra	154.5714	70.0	188.32897	3	600	14
Pilot	86.6000	65.0	75.92924	2	300	25
Springfield	216.5714	60.0	358.90661	20	1000	7

Size of operation is heavily left-skewed:

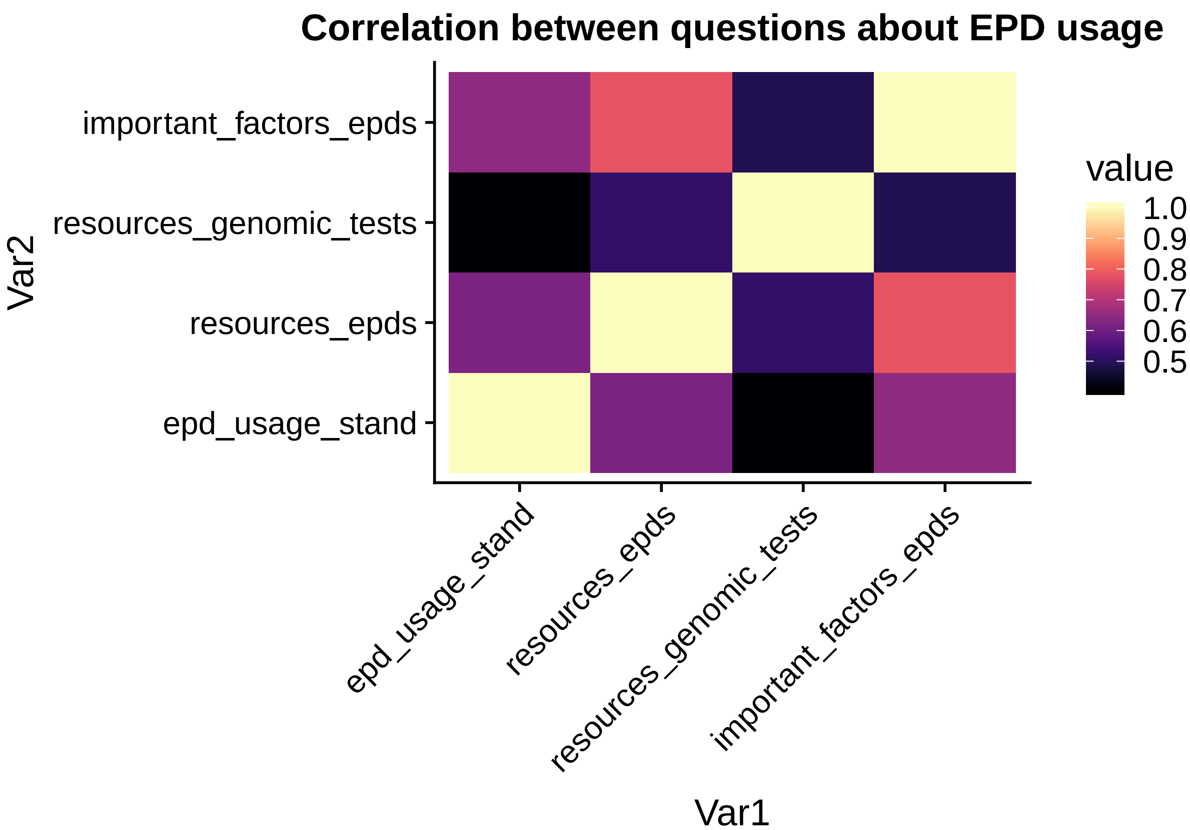


Modelling

In the survey, there are 4 questions that ask about frequency and value of EPD/GE-EPD usage. In the dataset, these questions have various degrees of missingness. Standardized EPD usage (% importance of visual evaluation vs. % importance of EPDs when making breeding decisions) by far has the most missing data (11.04%). This may be because the question was on the front of the survey in the paper version, and therefore overlooked or assumed to be part of the instructions.



Further, reporting a higher % EPD usage requires reporting a lower % visual appraisal usage since the answers can only add up to 100%. Visual appraisal is still a vital part of making breeding decisions, so the responses to this question may not accurately represent EPD usage among respondents. The correlation between the responses to questions about EPD/GE-EPD usage range from 0.406 to 0.782, with GE-EPD usage being the least correlated to other questions.



For this reason, some of the models were run iteratively using each of the four question responses as the Y variable.

1. Older producers tend to be less progressive and therefore rely more heavily on visual appraisal when making breeding decisions.

In order to determine the relationship between EPD usage and age, the following model was tested.

$$Y_{ijk} = \mu + \beta X_i + \gamma_{j(k)} + e_{ijk}$$

Where:

- Y is the scaled EPD usage from question 1

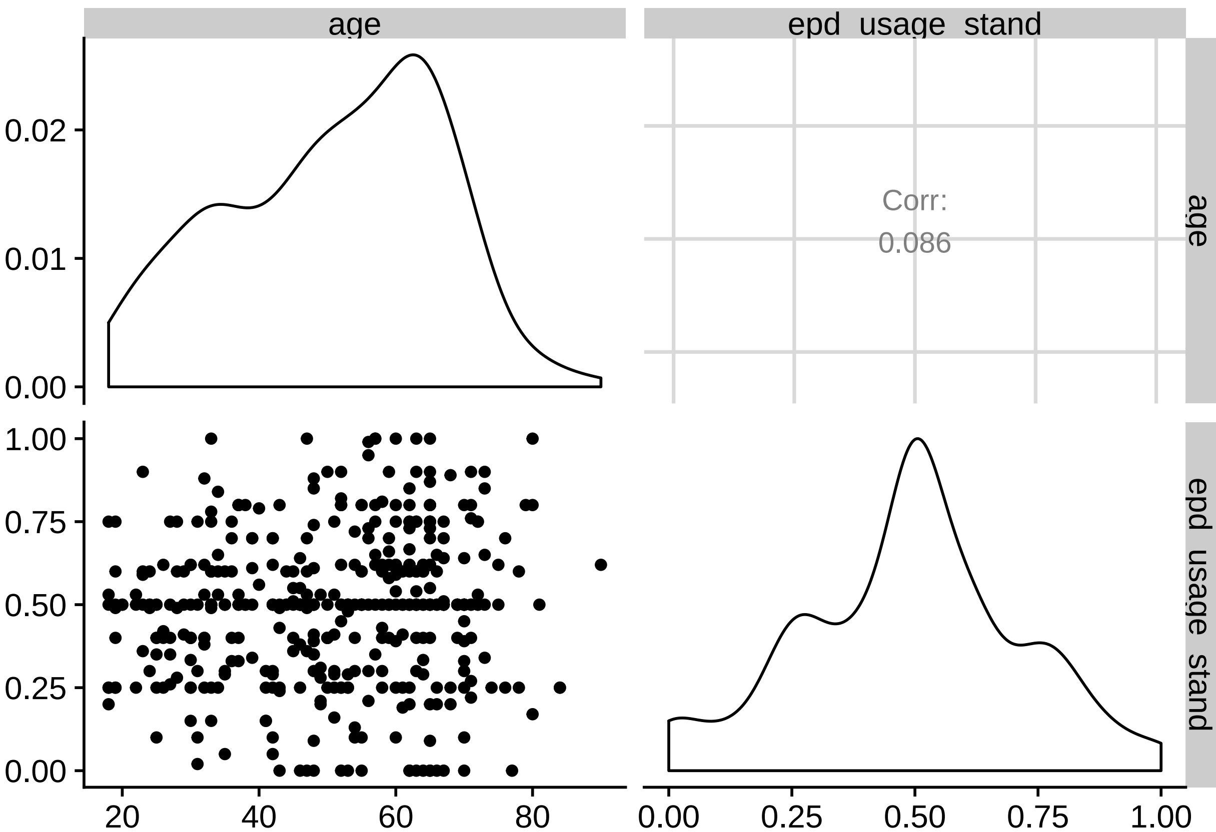
- μ is the mean
- X is the fixed effect of age (continuous)
- γ is the random effect of sale barn nested within medium (survey conducted online vs. in person) for $j = 1, 2$
- e is the residual

```
set.seed(88)

#Including random effects
car::Anova(lmer(epd_usage_stand ~
                  age +
#https://stats.stackexchange.com/questions/79360/mixed-effects-model-with-nesting
```

```
                 (1|medium/sale_barn),
     data = responses),
     type = "III")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: epd_usage_stand
##                Chisq Df Pr(>Chisq)
## (Intercept) 85.2237  1     <2e-16 ***
## age          3.7128  1      0.054 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model predicting scaled EPD usage from age is not significant at $\alpha = 0.05$ (p = 0.054). When the two variables are visualized, there doesn't appear to be a clear relationship.



2. Size of operation is predictive of EPD/GE-EPD use.

We hypothesized that producers who are more reliant on cattle productuon as their sole source of income (i.e., producers with reportedly larger operations) would be more data driven, and therefore more reliant on EPDs than visual appraisal. This hypothesis is supported by the USDA ERS assertion that "oerations with 40 or fewer head are largely part of multi-enterprises, or are supplemental to off-farm employment". In order to determine the relationship between size of operation and reported EPD usage, the following model was tested.

$$Y_{ijk} = \mu + \beta X_j + \gamma_{j(k)} + e_{ijk}$$

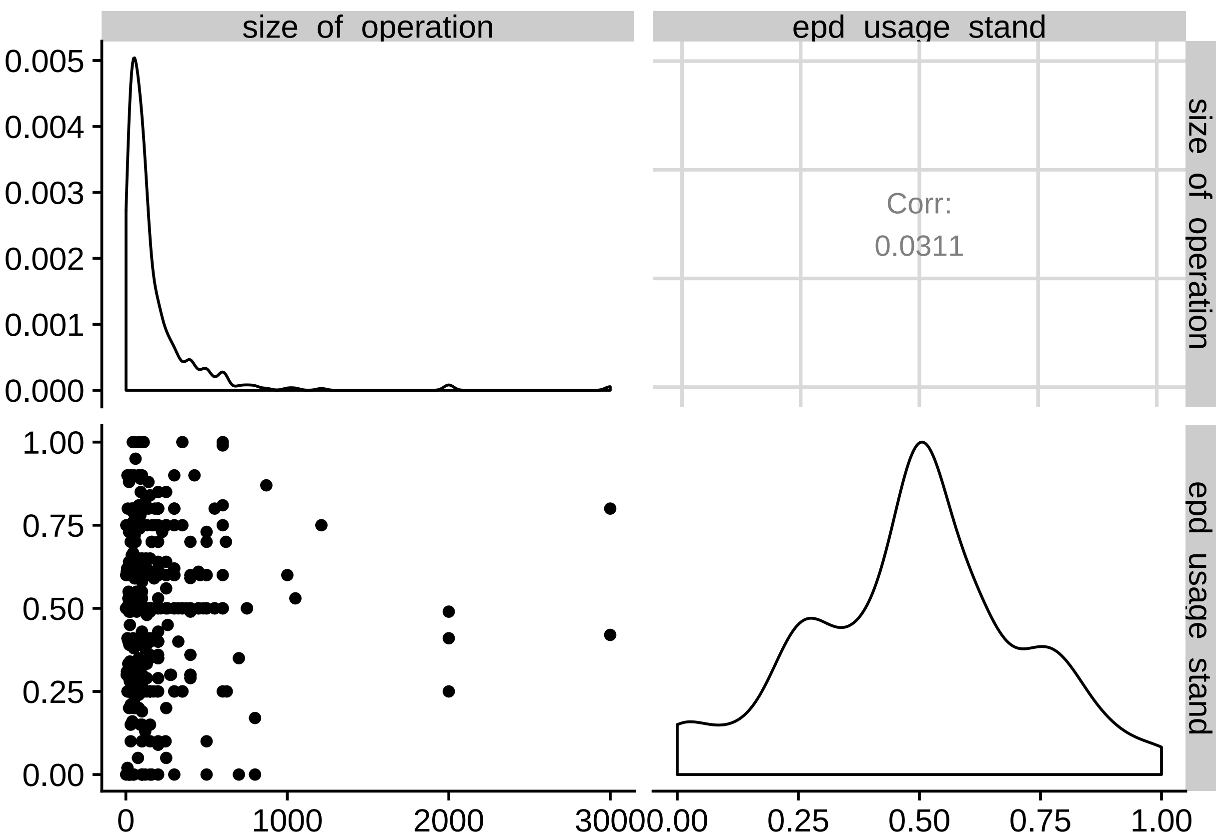
Where:

- Y is the scaled EPD usage from question 1
- μ is the mean
- X is the reported operation size reported in question 8 (continuous)
- γ is the random effect of sale barn nested within medium (survey conducted online vs. in person) for $j = 1, 2$
- e is the residual

```
#Including random effects
Anova(lmer(epd_usage_stand ~
                 size_of_operation +
                 #sale barn nested in medium
                 #https://stats.stackexchange.com/questions/79360/mixed-effects-model-with-nesting
                 (1|medium/sale_barn),
     data = responses))
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: epd_usage_stand
##                    Chisq Df Pr(>Chisq)
## size_of_operation 0.3465  1     0.5561
```

Surprisingly, size of operation is also not significant in predicting EPD usage than age (p = 0.5561). Again, there is no clear relationship between the two variables when both are visualized.



3. There is a relationship between reported EPD usage and reported barriers to EPD usage.

Simple linear regression

First, a simple linear regression was fit for each of the 4 response variables assaying EPD usage for each of the 9 barriers listed in question 4.

$$Y = \beta_o + \beta_{ij}X_{ij} + E$$

Where:

- Y is one of the four EPD usage responses described above (`epd_usage_stand`, `resources_epds`, `resources_genomic_tests`, or `important_factors_epds`)
- $/beta_o$ is the intercept
- $\beta_{ij}X_{ij}$ represents for the i^{th} response variable Y, the reported "prohibitiveness" on a scale of 1-5 where 1 is "not a barrier" and 5 is "prohibitive barrier" (see survey question 4) for one of the 9 potential barriers to EPD usage assayed
  - Continuous

```
barrier_data <- responses %>%
  melt(
    id = c(
      "sale_barn",
      "medium",
      "survey_number",
      "size_of_operation",
      "age",
      "sex",
      "epd_usage_stand",
      "resources_genomic_tests",
      "resources_epds",
      "important_factors_epds"
    ),
    na.rm = FALSE
  ) %>%
  mutate(variable = as.character(variable)) %>%
  #Pull out barrier responses
  filter(str_detect(variable,
                    "barrier_")) %>%
  rename(response = value,
         barrier = variable) %>%
  melt(
    id = c(
      "sale_barn",
      "medium",
      "survey_number",
      "size_of_operation",
      "age",
      "sex",
      "barrier",
      "response"
    ),
```

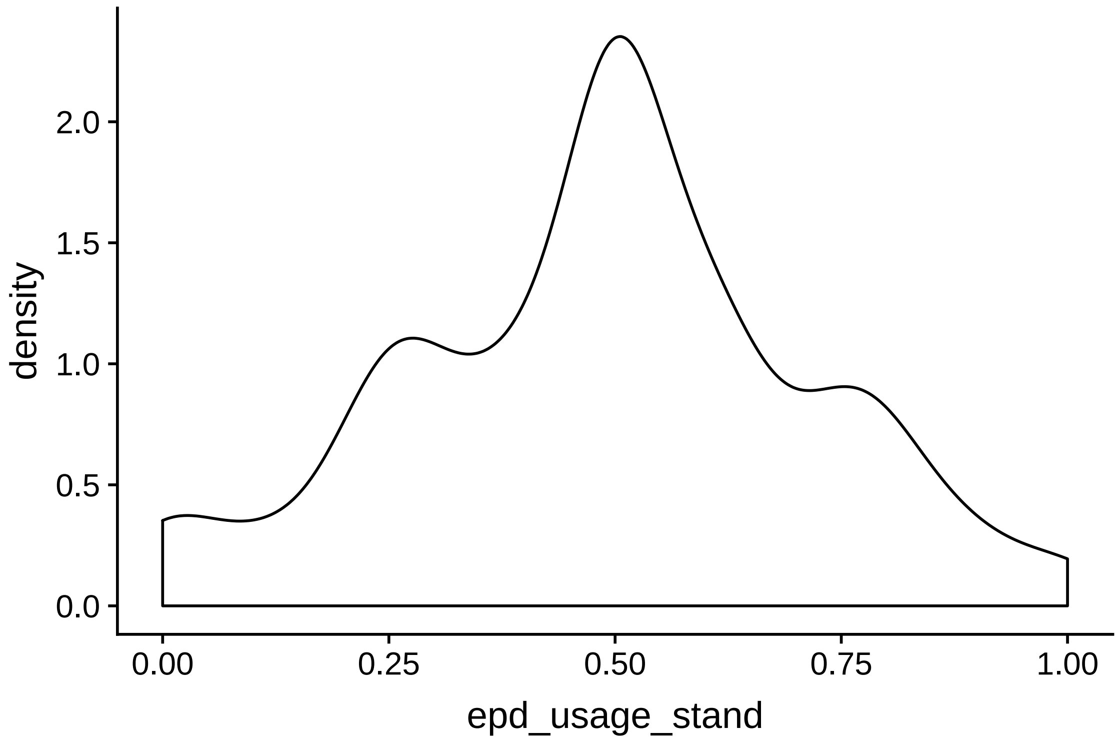
```r
    value.name = "yval",
    na.rm = FALSE
  ) %>%
  rename(yvar = variable) %>%
  #Group by each variable assayed
  group_by(yvar, barrier) %>%
  #For each variable assayed, simple linear model predicting
  #EPD usage from response
  nest()
```

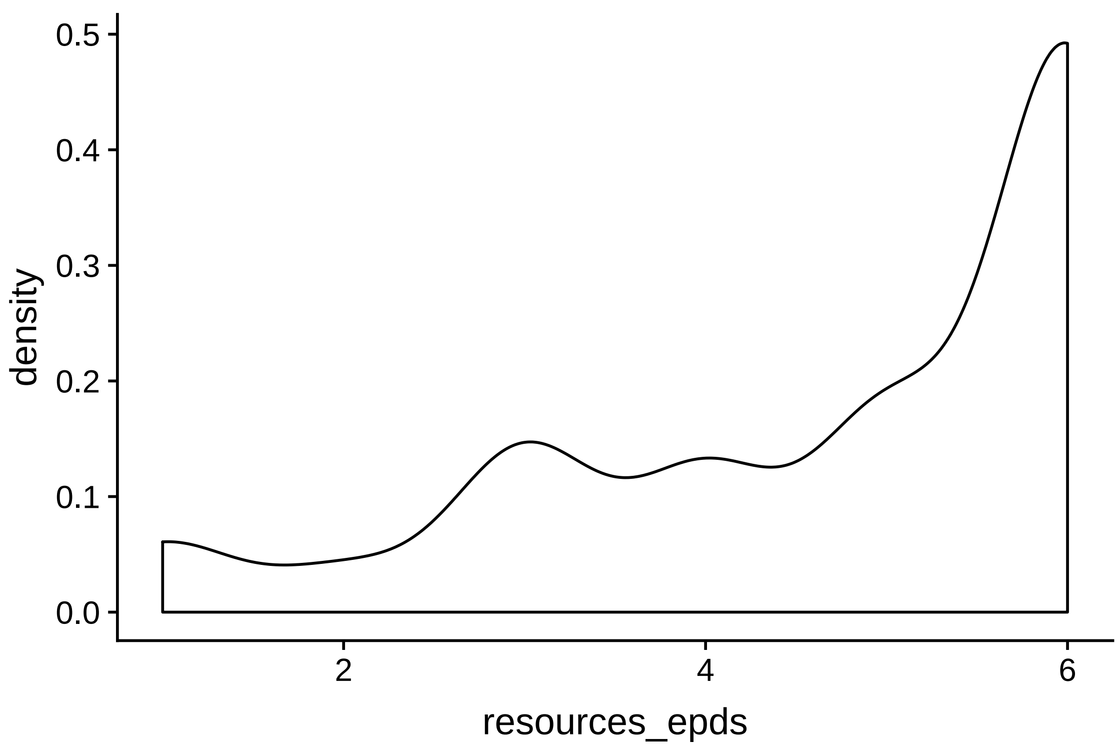
```r
barrier_data %>%
  mutate(
    mod = purrr::map(data,
                     ~ lm(
                       formula = yval ~ response,
                       data = .x
                     )),
    r_squared = purrr::map_dbl(mod,
                               ~ pluck(glance(.x),
                                       "r.squared")),
    pval = purrr::map_dbl(mod,
                          ~ pluck(glance(.x),
                                  "p.value")),
    aic = purrr::map_dbl(mod,
                         ~ pluck(glance(.x),
                                 "AIC"))
  ) %>%
  select(-mod, -data) %>%
  mutate(signif = if_else(pval < 0.05, "*", "Not significant")) %>%
  arrange(aic, pval) %>%
  kable("latex",
        caption = "Summary of all models with all possible response values",
        booktabs = TRUE) %>%
  kable_styling(position = "center")
```

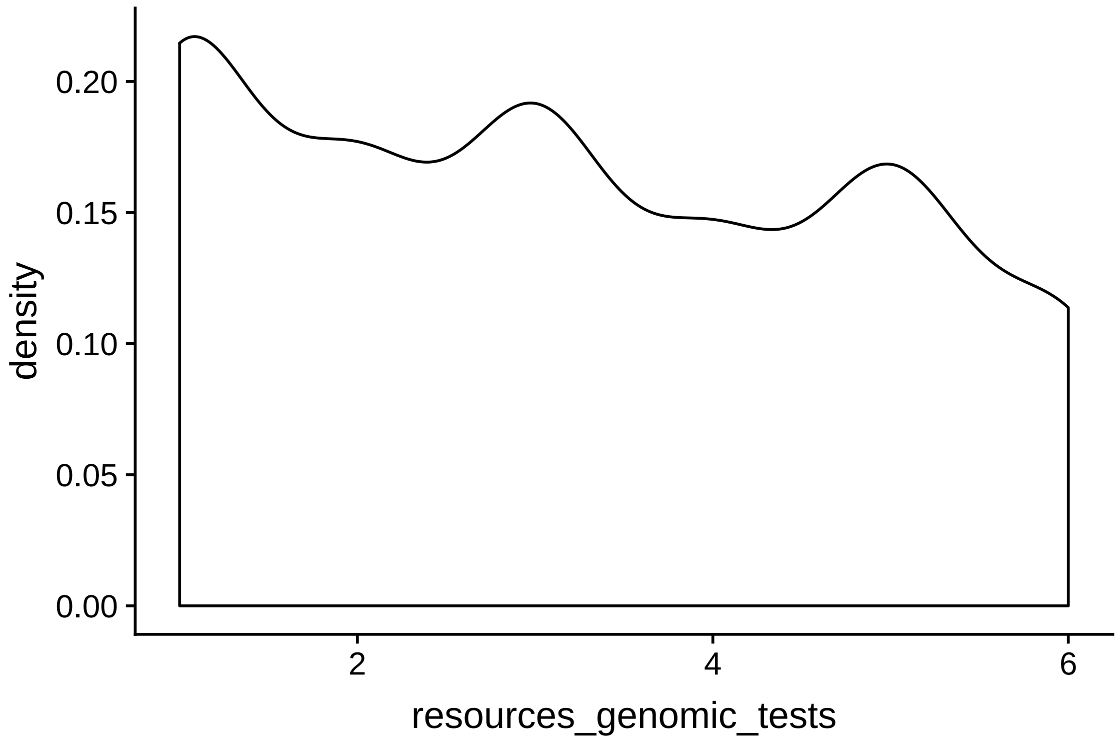
One should always check to ensure that regression assumptions (existence, independence, linearity, homoscedastity) are not violated prior to performing the analyses. It appears that `epd_usage_stand` is the only tested `Y` that is normally distributed.

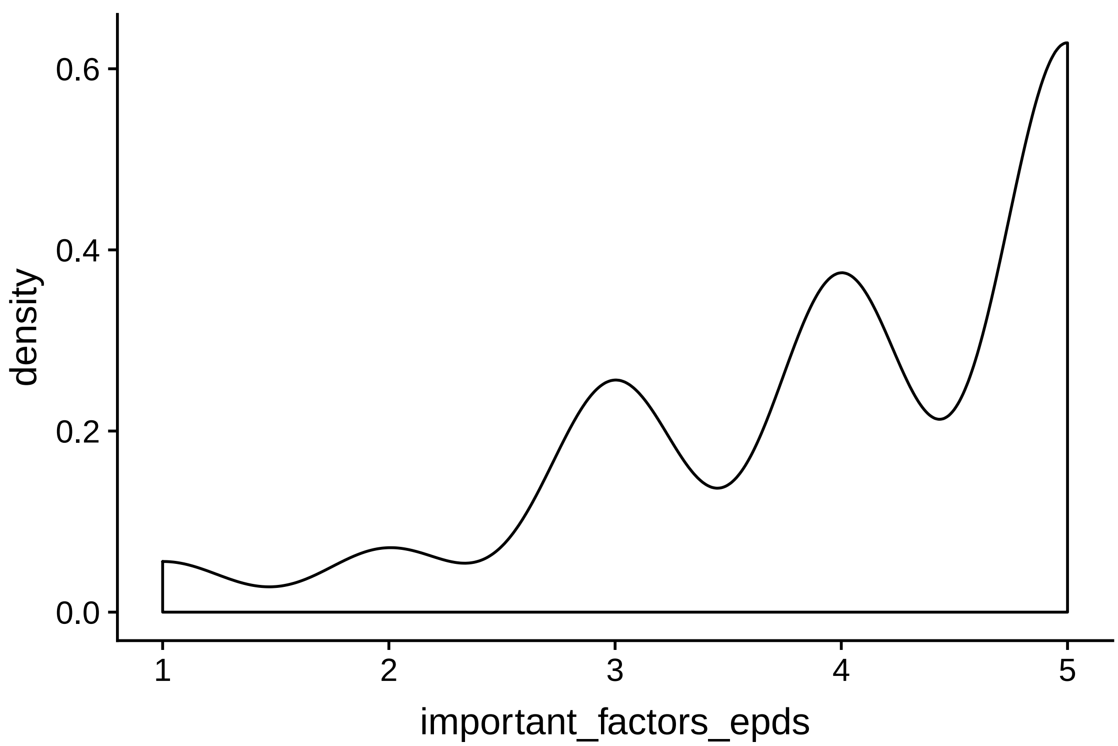
Table 1: Summary of all models with all possible response values

yvar	barrier	r_squared	pval	aic	signif
epd_usage_stand	barrier_epd_dont_reflect_factors	0.0662124	0.0000001	-100.26894	*
epd_usage_stand	barrier_epd_not_accurate	0.0620408	0.0000003	-93.91424	*
epd_usage_stand	barrier_epd_have_not_worked	0.0496418	0.0000057	-91.69472	*
epd_usage_stand	barrier_epd_too_many_bulls	0.0630820	0.0000003	-86.24392	*
epd_usage_stand	barrier_epd_difficult_to_read	0.0258348	0.0011225	-79.94505	*
epd_usage_stand	barrier_epd_difference_between	0.0165530	0.0093656	-74.61891	*
epd_usage_stand	barrier_epd_overlap	0.0177631	0.0079150	-70.15080	*
epd_usage_stand	barrier_epd_not_available	0.0003210	0.7165778	-65.26086	Not significant
epd_usage_stand	barrier_epd_inconsistent	0.0032134	0.2509534	-61.36950	Not significant
important_factors_epds	barrier_epd_overlap	0.0561707	0.0000007	1277.21453	*
important_factors_epds	barrier_epd_too_many_bulls	0.1072843	0.0000000	1295.07047	*
important_factors_epds	barrier_epd_difficult_to_read	0.0957229	0.0000000	1297.87592	*
important_factors_epds	barrier_epd_have_not_worked	0.0750547	0.0000000	1302.76258	*
important_factors_epds	barrier_epd_not_accurate	0.0605293	0.0000002	1314.88434	*
important_factors_epds	barrier_epd_difference_between	0.0528625	0.0000009	1315.54495	*
important_factors_epds	barrier_epd_dont_reflect_factors	0.0557115	0.0000005	1319.18272	*
important_factors_epds	barrier_epd_not_available	0.0039953	0.1822146	1343.44484	Not significant
important_factors_epds	barrier_epd_inconsistent	0.0131201	0.0150542	1356.95761	*
resources_epds	barrier_epd_overlap	0.0675132	0.0000001	1510.68964	*
resources_epds	barrier_epd_too_many_bulls	0.1255719	0.0000000	1534.75702	*
resources_epds	barrier_epd_have_not_worked	0.0874544	0.0000000	1540.60334	*
resources_epds	barrier_epd_difference_between	0.0680766	0.0000000	1557.13802	*
resources_epds	barrier_epd_not_accurate	0.0606957	0.0000002	1557.24736	*
resources_epds	barrier_epd_difficult_to_read	0.1083684	0.0000000	1562.66579	*
resources_epds	barrier_epd_dont_reflect_factors	0.0517250	0.0000018	1563.01964	*
resources_epds	barrier_epd_not_available	0.0175308	0.0057352	1583.40212	*
resources_epds	barrier_epd_inconsistent	0.0118763	0.0223921	1604.66184	*
resources_genomic_tests	barrier_epd_overlap	0.0268701	0.0007348	1620.25533	*
resources_genomic_tests	barrier_epd_difficult_to_read	0.0656788	0.0000001	1646.30779	*
resources_genomic_tests	barrier_epd_have_not_worked	0.0371452	0.0000564	1662.26731	*
resources_genomic_tests	barrier_epd_difference_between	0.0272492	0.0005717	1665.70615	*
resources_genomic_tests	barrier_epd_too_many_bulls	0.0400683	0.0000260	1680.31958	*
resources_genomic_tests	barrier_epd_dont_reflect_factors	0.0078891	0.0645023	1687.48544	Not significant
resources_genomic_tests	barrier_epd_not_available	0.0106751	0.0312021	1688.94295	*
resources_genomic_tests	barrier_epd_not_accurate	0.0082621	0.0584819	1689.03211	Not significant
resources_genomic_tests	barrier_epd_inconsistent	0.0020646	0.3427625	1712.68993	Not significant


density
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
epd_usage_stand


density
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
resources_epds





Of the models that used `epd_usage_stand` as the response variable, all barrier models excluding *"EPDs are not available for the bulls I purchase"* and *"Inconsistency between breed EPDs"* were significant. Of significant models, *"EPDs don't reflect all important factors in selecting breeding animals"* had the lowest AIC and lowest p-value. It would be interesting to see if there is a relationship between responses to this barrier and responses to question about selection index usage (which combine multiple selection traits into one EPD index).

Multiple linear regression

Next, all possible subsets selection was performed to choose a model relating EPD usage to reported barriers to EPD usage. The full model was as follows:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + E$$

Where:

- Y represents scaled EPD usage from question 1
- β_o represents the intercept
- $\beta_{1-9} X_{1-9}$ represents for each of the 9 potential barriers to EPD usage assayed, the reported "prohibitiveness" on a scale of 1-5 where 1 is "not a barrier" and 5 is "prohibitive barrier" (see survey question 4)
  - Continuous

All possible subset selection using the `leaps` package:

```
#https://github.com/alexpghayes/broom/blob/some_cleanup/R/leaps.R
tidy.regsubsets <- function(x, ...) {
  s <- summary(x)
  inclusions <- as_tibble(s$which)
  metrics <- with(
    s,
    tibble(
      r.squared = rsq,
      adj.r.squared = adjr2,
      BIC = bic,
      mallows_cp = cp
    )
  )
  bind_cols(inclusions, metrics)
}
```

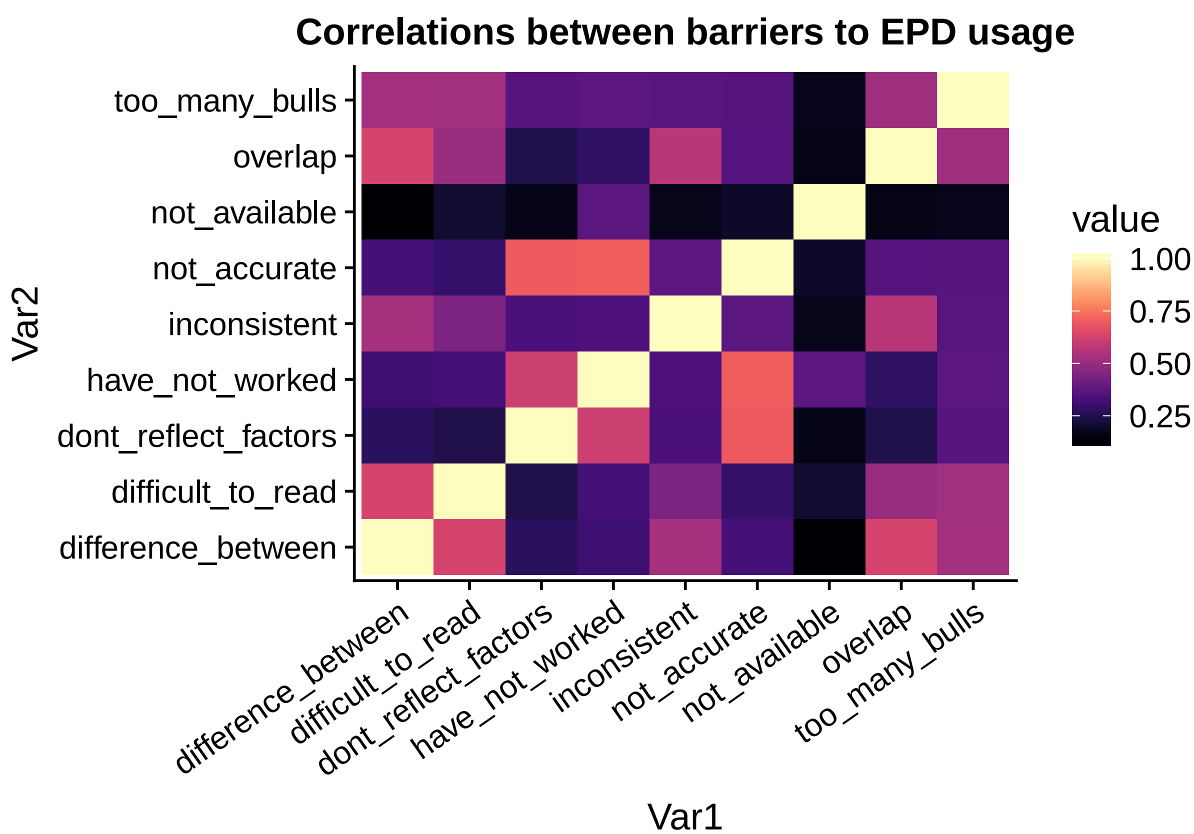
```
#https://rstudio-pubs-static.s3.amazonaws.com/2897_9220b21cfc0c43a396ff9abf122bb351.html
multi_barrier <- responses %>%
  select(epd_usage_stand, starts_with("barrier_"))

barrier_subsets <- regsubsets(epd_usage_stand ~ .,
           data = multi_barrier,
           method = "exhaustive"
           )

tidied_barr <- tidy.regsubsets(barrier_subsets) %>%
  mutate(n_variable = row_number())
```

All models have a very low R^2 and explain very little of the variablility in the data. `barrier_epd_not_accurate` (*"EPDs don't accurately reflect genetic merit"*) is included in all model iterations).

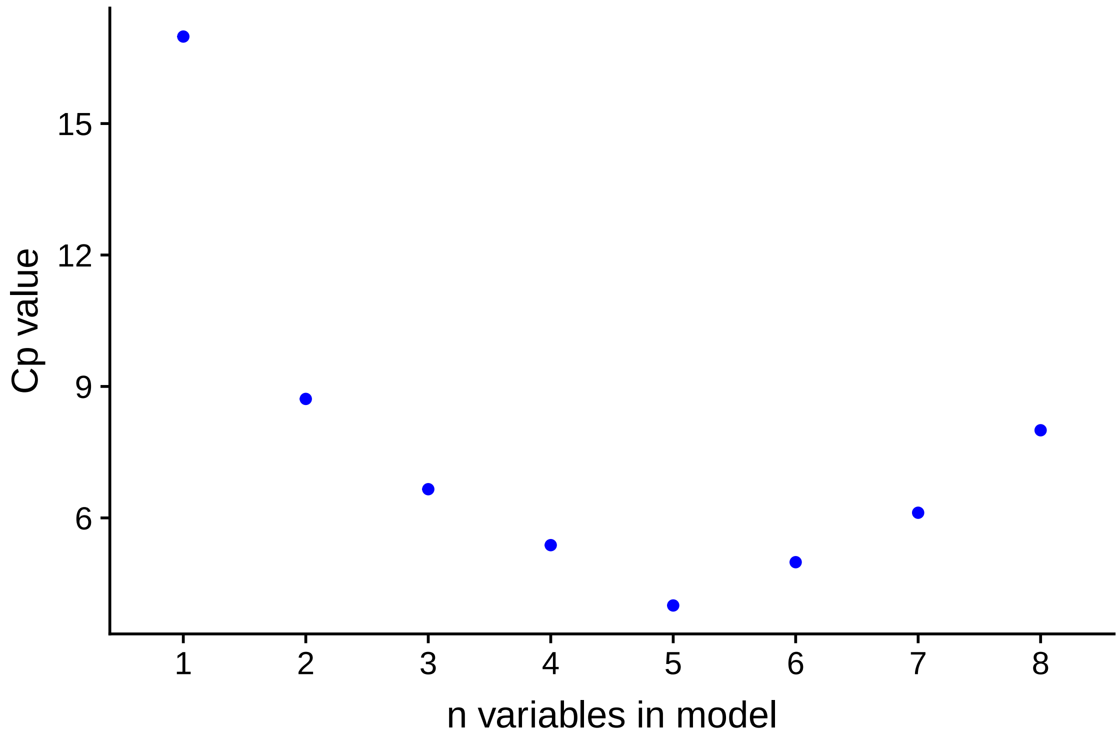
"EPDs don't reflect all important factors in selecting breeding animals" (the most significant model among the simple linear regressions) is not included until the 7 variable case. These 2 variables are probably highly collinear, as evidenced in the pairwise correlation matrix below.

Table 2: Summary of all possible subsets model selection results

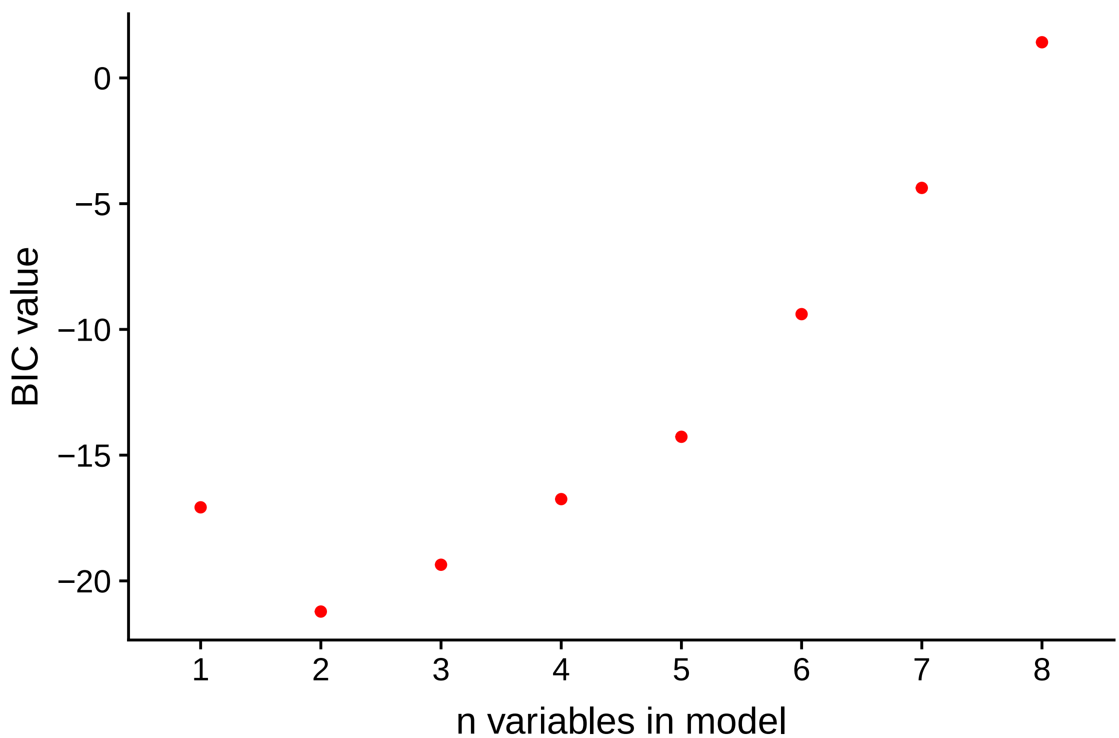
n_variable	r.squared	adj.r.squared	BIC	mallows_cp
1	0.0751411	0.0726279	-17.075209	16.985842
2	0.0999431	0.0950381	-21.219477	8.715300
3	0.1097471	0.1024500	-19.358395	6.655413
4	0.1176606	0.1079912	-16.748537	5.378426
5	0.1258145	0.1138064	-14.270169	4.001896
6	0.1282597	0.1138508	-9.393068	4.989316
7	0.1303639	0.1135477	-4.373731	6.117980
8	0.1306479	0.1113825	1.418903	8.000358



The model with 5 variables (`barrier_epd_inconsistent` + `barrier_epd_too_many_bulls` + `barrier_epd_not_available` + `barrier_epd_have_not_worked` + `barrier_epd_not_accurate`) had the lowest Cp value.



The model with 2 variables (`barrier_epd_too_many_bulls` + `barrier_epd_not_accurate`) had the lowest BIC.



Below are ANOVA results for the selected models with 2-5 variables.

- Five variables:

```
anova(lm(formula = epd_usage_stand ~ barrier_epd_inconsistent +
           barrier_epd_too_many_bulls +
```

```
          barrier_epd_not_available +
          barrier_epd_have_not_worked +
          barrier_epd_not_accurate,
   data = responses))
```

```
## Analysis of Variance Table
##
## Response: epd_usage_stand
##                              Df  Sum Sq Mean Sq F value    Pr(>F)
## barrier_epd_inconsistent      1  0.1140 0.11395  2.6670  0.103258
## barrier_epd_too_many_bulls    1  1.0730 1.07301 25.1131 8.215e-07 ***
## barrier_epd_not_available     1  0.0233 0.02327  0.5445  0.461015
## barrier_epd_have_not_worked   1  0.5933 0.59327 13.8851  0.000223 ***
## barrier_epd_not_accurate      1  0.4644 0.46443 10.8698  0.001067 **
## Residuals                   390 16.6635 0.04273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Four variables:

```
## Analysis of Variance Table
##
## Response: epd_usage_stand
##                              Df  Sum Sq Mean Sq F value    Pr(>F)
## barrier_epd_too_many_bulls    1  1.2049 1.20487 28.2761 1.766e-07 ***
## barrier_epd_not_available     1  0.0253 0.02533  0.5945 0.4411523
## barrier_epd_have_not_worked   1  0.5668 0.56676 13.3008 0.0003009 ***
## barrier_epd_not_accurate      1  0.4131 0.41309  9.6944 0.0019833 **
## Residuals                   394 16.7887 0.04261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Three variables:

```
anova(lm(formula = epd_usage_stand ~ barrier_epd_too_many_bulls +
          barrier_epd_not_available +
          barrier_epd_not_accurate,
   data = responses))
```

```
## Analysis of Variance Table
##
## Response: epd_usage_stand
##                             Df  Sum Sq Mean Sq F value    Pr(>F)
## barrier_epd_too_many_bulls   1  1.2543 1.25431 29.0832 1.193e-07 ***
## barrier_epd_not_available    1  0.0451 0.04512  1.0463     0.307
## barrier_epd_not_accurate     1  0.8356 0.83560 19.3746 1.383e-05 ***
## Residuals                  397 17.1220 0.04313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Twoo variables:

```r
anova(lm(formula = epd_usage_stand ~ barrier_epd_too_many_bulls +
           barrier_epd_not_accurate,
   data = responses))
```

```
## Analysis of Variance Table
##
## Response: epd_usage_stand
##                             Df  Sum Sq Mean Sq F value    Pr(>F)
## barrier_epd_too_many_bulls   1  1.2756 1.27559  28.533 1.546e-07 ***
## barrier_epd_not_accurate     1  0.5758 0.57575  12.879 0.0003734 ***
## Residuals                  402 17.9720 0.04471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions

Overall, there is no difference in EPD usage between ages or size of operation. Based on anecdotal evidence, the insignificance of age is unsurprising to me. The insignificance of size of operation is surprising to me, though. However, our data is heavily left skewed and may not be a representative sample of the population. This skewness may be due to medium to large scale producers purchasing cattle by other means than their local sale barn (where half of our data came from). Still, the median size of operation in our data is much larger than the national average.

Taking all results together:

- An overabundance of EPDs to choose evaluate
- A disconnect between the traits EPDs evaluate and the "whole picture" of animal worth

appear to be the biggest barriers to EPD usage. This may provide compelling evidence of the need for EPD indexes that evalueate health traits, structural traits, and environmental compatability.

One major caveat to take into consideration for these results is potential sample bias. Most of the in-person surveys were collected at Show-Me-Select replacement heifer sales, part of an added value marketing program facilitated by the University of Missouri. Therefore, our sample may represent more progressive producers that the population average.