People's Preferred News Source For Equine Disease

ST472 Individual Project Report

Michael Ingram

April 28, 2017

Abstract

The equine industry has a huge economic and social impact in Colorado, with its steady growth and growing numbers of horse owners. This study aims to determine what factors determine where and why individuals gather information on social media related to equine disease outbreaks. Specifically, we are interested in demographic factors (e.g. role, age), and features of news sources (availability and accuracy). An online sample (n= 256) was collected from different horse club members in Colorado. The questionnaire mostly consisted of categorical data, including age cohorts, roles of participants, the types of news sources accessed (e.g., daily), and Likert-scale questions on why individuals chose to access each source (availability and accuracy). Our first objective is to examine how age and role affect where and why individuals gather information. Our second and third objectives are to examine how the accuracy, availability, and accessibility of news sources affect where and the reason why individuals access information. We use multinomial and proportional odds models to analyze the first three objectives. The final objective is to examine the relationship between where individuals choose their preferred news source and their reason for choosing

it, using a formal test of association.

---

1. Introduction

1.1 Background

The client, Shelly McDaniel, is conducting a study about veterinary media in the equine science. She is interested in seeing where horse owners get there information and how reliable it is. She is especially interested in seeing how many people are getting information from online sources as opposed to a trained veterinarian. In the meeting, the client shared with us some of the history and reasons behind her study. She said that historically, horse owners would call an equine veterinarian any time their horses had a problem but usually there is a wait period because the veterinarian has to travel to their patient. So for those living in remote places or very populous areas where the veterinarian has a wait list this could mean up to a few weeks. Now with the internet, it is easy to obtain information through online sources, and she believes many people are turning to online sources first before the veterinarian can arrive and make their diagnoses.

1.2 Objectives

The client came to the meeting with 5 specific objectives to her study.

1. How do Age and Role affect where (Objective 1.1) and the reason why (Objective 1.2) individuals gather information?

2. How does the accuracy and availability of source affect where individuals access information?

3. How does the accuracy and availability of source affect the reason individuals access information?

4. To examine where individuals choose information first in relation to why individuals gather information.

1.3 Survey and Sampling Information

So as outlined in the client's objectives, the study is looking at the many different issues regarding how horse owners obtain information. During the meeting, we also discussed physically how the survey was conducted. The client said that she took her survey at local equine clubs in the Colorado area and at local equine competitions. Her survey consisted of 11 questions each with a specific relationship to one of her 5 objectives. She surveyed a total of 258 horse owners. The questionnaire mostly consists of categorical data, including age cohorts, professional roles of participants, the types of news sources accessed (e.g., daily), and Likert-scale questions on why individuals choose to access each source (availability and accuracy).Further refinement of her survey led to using only questions 1, 2, 3, 5, 6, 7, and 8 from the survey. This was due to many of the questions asking for similar information which was shown in the data through extremely high correlation.

1.4 Questionnaire Keywords

The questionnaire used the following specifications for role and news sources which will be referred to throughout this report.

1. Professional Role Specification: Pro (Equine professional), Comp (Competitive riders), Plea (Pleasure riders), Own (Horse owners), and Other.

2. Everyday News Sources: Social (Facebook and Twitter), Online (Online news sources and phone news app), and Traditional (Newspaper, TV, Radio).

3. Equine Disease News Sources: Social (as above), Online (Online and Google), TradProf (Traditional, for-profits sites), Vet (Veterinarian), and State (University sources, State veterinarian resources).

2. Methods

2.1 Multinomial Logistic Regression Model (Objectives 1.1, 2, 3)

The Multinomial model was used to fit our categorical response variables. It has a log-odds response that compares all categories with one category chosen as a baseline [1]. The baseline model works by relating π_i to covariates by taking a set of $r-1$ baseline category logits. If j^* is the baseline category then the model is:

$$log(\frac{\pi_{ij}}{\pi_{ij^*}}) = x_{ij}\beta_i^T \quad \text{where} \quad j \neq j^* \tag{1}$$

This can also be written in terms of π_{ij}:

$$\pi_{ij} = \frac{exp(x_i^T\beta_j)}{1+\sum(expx_i^T)} \tag{2}$$

For the Baseline Category, the numerator is just one [2].

2.2 Likelihood Ratio Test (Objectives 1, 2, 3)

Likelihood Ratio test is adopted to perform model comparisons. We used the likelihood ratio test to test a null model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 \tag{3}$$

against a full (saturated) model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{4}$$

to see if any of the coefficients were significant. Formally, we were testing $H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$ vs. $H_a : At\, least\, one\, \beta_k \neq 0$. If there was evidence to support the alternative hypothesis then we would continue on to test which specific beta coefficients were significant [3].

2.3 Linear Mixed Model (Objective 1.2)

Linear Mixed Model is fitted to evaluate fixed effects (e.g., age and sources of information) and random effects (e.g., subjects) [6].

Model:

$$Y = X\beta + Zu + \epsilon \tag{5}$$

where random effect $u \sim N(0, \sigma_u^2)$ and $\epsilon \sim N(0, \sigma^2)$.

2.4 Variance Inflation Factor (Objectives 1, 2, 3)

Variance inflation factors are used to examine whether collinearity exists among different predictors. Specifically we used this to make sure that there were no multicollinearity between the predictors Q3Online, Q3Social, Q3Traditional, Q5Online, Q5Social, Q5Traditional. Variance inflation factors are calculated by:

$$VIF_k = \frac{1}{1-R_k^2} \tag{6}$$

where R_k^2 is R^2 valued obtained regressing the k^{th} predictor. Variance inflation factors give a value where 1 is no correlation, between 1 to 5 is moderate correlation, greater than 5 to

10 is is highly correlated [4].

2.5 Chi-Square Test of Association (Objective 4)

Chi-Square Test of Association is adopted to determine if why individuals access information is associated with where individuals access information. The Chi-Square Test of Association compares the observed count to the expected count under the assumption of no association [5].

$$\chi^2 = \sum_{i=1}^{k} [\frac{(O_i - E_i)^2}{E_i}] \tag{7}$$

3. Results

3.1 Exploratory Data Analysis

In the initial phase of our data analysis we were able to find a lot of information through charts without even running a model. For Objective 1, we could see a clear trend in how people's preffered everyday news source differed among different age groups. See Figure 1.

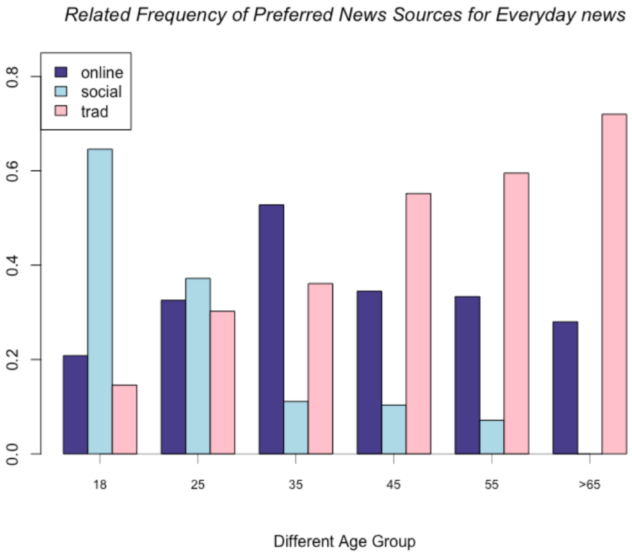
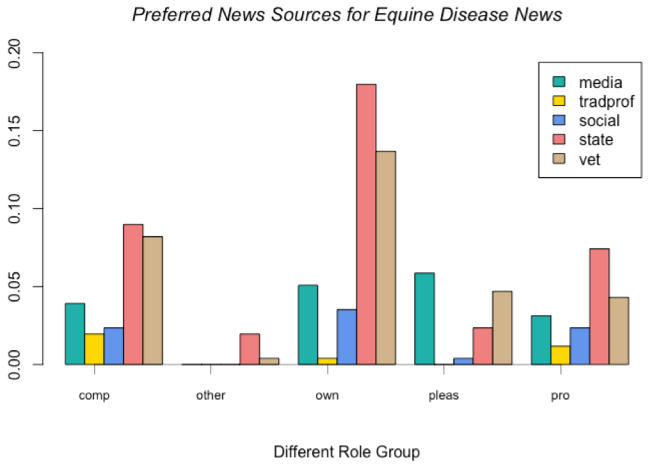


Figure 1: You can see a trend that the younger age group prefers social news sources, middle age prefer online and the oldest group prefers traditional

We also created another chart pertaining to Objective 1 that showed the distribution of role. Although it appears that the Other category could get thrown out, this is not true because it could further bias the data down the line in the data analysis.See Figure 2.

Figure 2: This chart is not as clear cut as Figure 1 however, it allows us to see that competitive riders, horse owners and professionals prefer state and vet sources while media sources only had a strong showing in pleasure riders and horse owners.



For Objective 2, we created a horizontal box plot. This plot gives a good overarching view of the data because it allows you to see how individuals ranked accuracy and availability of news sources based on where they get their news, side by side, for each Q3 and Q5 category. This chart allows us to draw some initial conclusions without even fitting a model such as online news sources had the highest availability rating and social news sources had the lowest accuracy rating. See Figure 3.

For Objective 3, the main difference from Objective 2 was the response changed. So we created another horizontal box plot with this one showing accuracy and availability ranking of news sources based on the reason why individuals sought equine health information. The four reasons or Q8 categories were a=How the disease is spread, b=Symptoms, c=Impact on Horse Shows and Events, d=Other. See Figure 4.

For Objective 4, we created a comparative bar chart because we are interested in where individuals choose information first and why they chose it. See Figure 5. You can see from Figure 5 that availability was heavily chosen for categories media, tradprof and social while

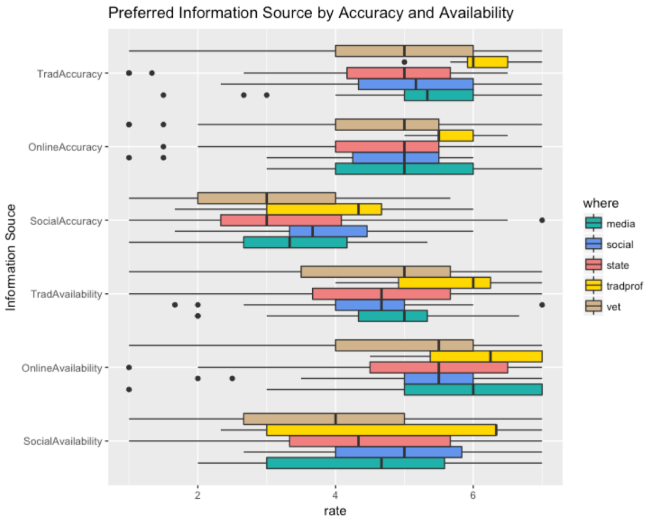


Figure 3: This chart shows how accuracy and availability ratings for different equine disease sources vary among different preferred news sources.

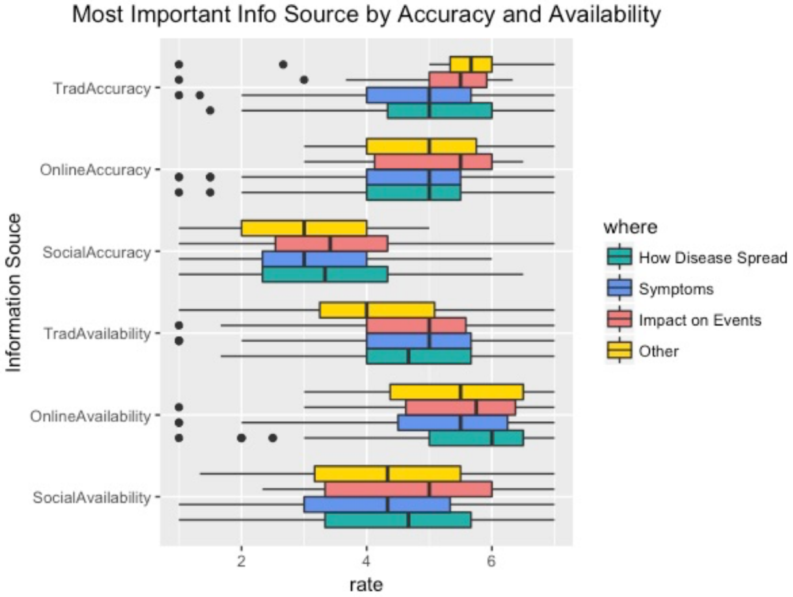


Figure 4: This chart shows accuracy and availability ratings of news sources based on the reason why individuals search for information from news sources.

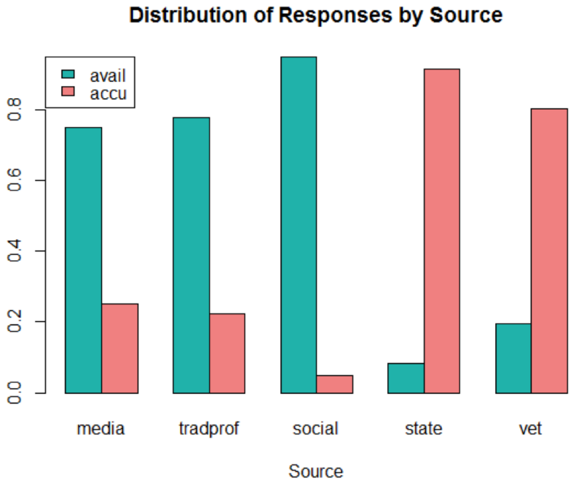


Figure 5: This chart shows relative frequency of accuracy and availability based on Q6 categories.

accuracy was heavily chosen for state and vet categories.

3.2 Objective 1.1: Age Does Affect Where Individuals Access Information

For Objective 1.1 we used a multinomial logit regression. See methods. Using a Type 3 Analysis of Effect we found that age and role are significant for where individuals access information. For age, $\chi^2(2, N = 256) = 47.8282$, $p < .0001$. For role, $\chi^2(8, N = 256) = 17.28$, $p < 0.0273$.We then took a Holm Correction for p-values of all roles and found none of them significant due to quasi-separation So in conclusion, age does affect where people gather information and role might affect however, under the current survey construction and sample size its hard to say. In Figure 6 we can see the relationship between age and news sources.

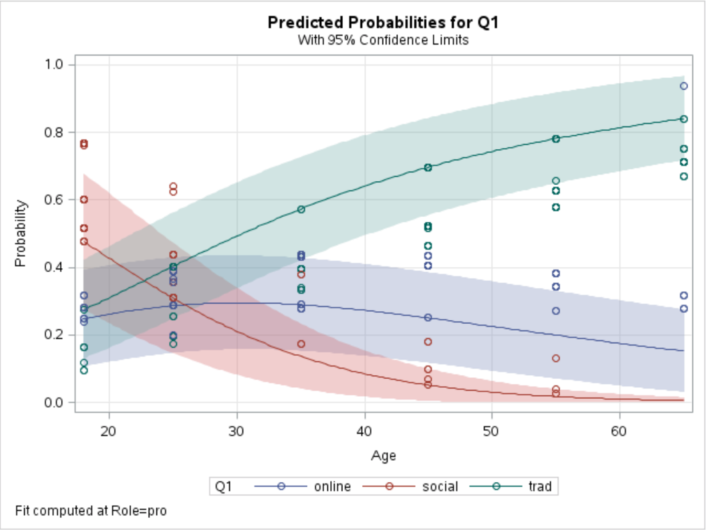


Figure 6: This Figure shows us that the older generation has a preference for traditional sources while the younger generation prefers social media sources.

3.3 Objective 1.2: Age Does Affect The Reason Why Individuals Access Information

In Objective 1.2 we used 2 linear mixed models. One for Question 3, with Q3 likert scale data as the response, age, source and intercept of participants as predictors where intercept of participants was a random effect. For Q5 the model was similar but with Q5 likert data as response and similar predictors but with an extra interaction term for age and source. Both models were selected by AIC. The surprising results we found was that age significantly affected accuracy ratings. See Figure 7.

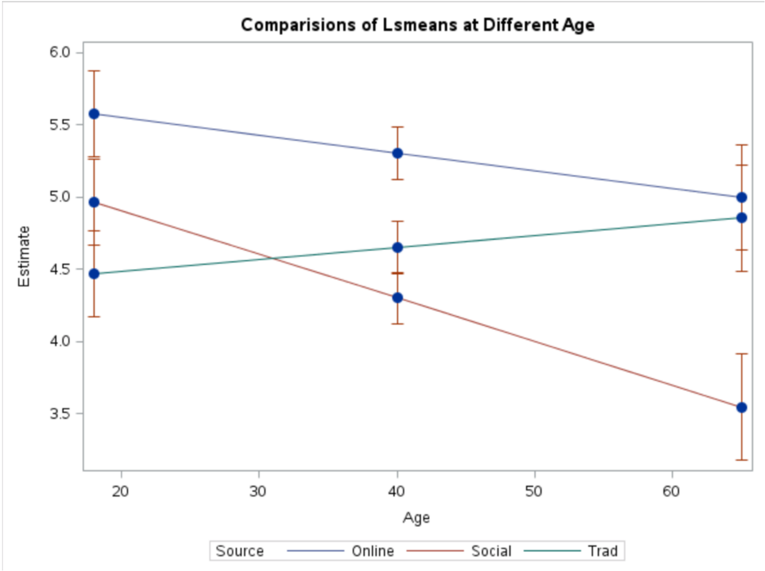


Figure 7: This Figure shows us that as age increases the ratings for accuracy, on average, decrease.

3.4 Objective 2: Accuracy Does Affect Where People Gather Information

For Objective 2, a multinomial model was also ran where the response was Q6 categories and predictors were the Q3 and Q5 categories. We also ran a Type 3 Analysis of Effects which showed that only Q3 Social and Q3 Traditional were significant. We find that accuracy ratings for social ($\chi^2(4) = 10.7492$, p=0.0295) relate to the probability that people will

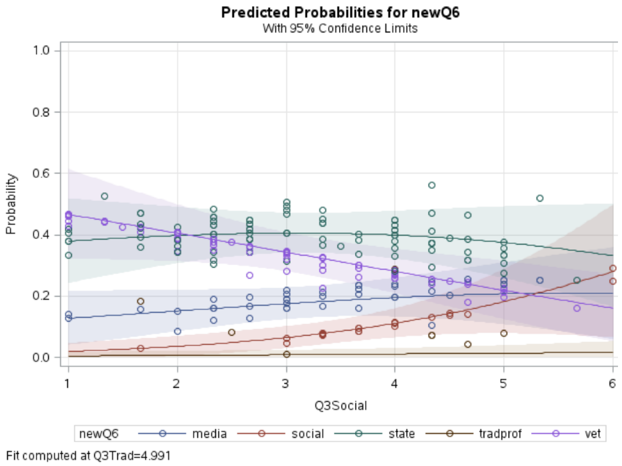


Figure 8: The Change in Information Sources Versus Accuracy Ratings for Social

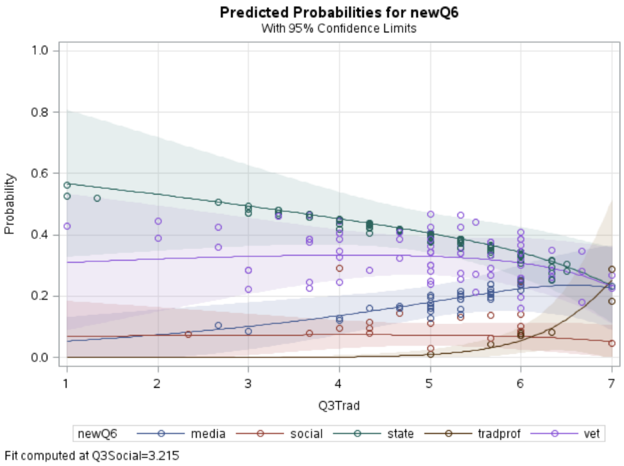


Figure 9: The Change in Information Sources Versus Accuracy Ratings for Trad

gather information from vet sources ($\chi^2(1)$ = 9.4140, p= 0.0176) versus social sources. See Figure 8. We also find accuracy ratings for traditional ($\chi^2(4)$ =11.93431, p= 0.0178) relate to the probability that people will gather information from Tradprof sources (χ^2 (1) = 7.4421, p= 0.0448) versus social sources. See Figure 9. The odds ratio estimate of Vet versus Social=0.733 with 95% CI [0.2980.766]. The odds ratio estimate of Tradprof versus Social= 4.501, 95% CI [1.65621.712]

3.5 Objective 3: Accuracy And Availability Do Not Affect The Reason Why Individuals Access Information

This objective focused on looking at survey Question 8 (Most Important Information Sought When an Equine Disease Outbreak Occurred) vs. Question 3 (Accuracy) and Question 5 (Availability). This objective used a multinomial logistic regression with the 4 categories of Question 8 (a=Spread of Disease, b=Symptoms, c=Impact on Shows/Events, d=Other) as the response and the three categories from both Question 3 and Question 5 (Social, Online Traditional) as predictors. However, in the early stages of the data analysis running a global likelihood ratio test comparing a null model against the full model resulted in $\chi^2(18) = 14.4751$ and $p = 0.6988$. This means none of the coefficients in the model are significant and therefore we conclude accuracy and availability do not affect the reason why individuals access information

3.6 Objective 4: Accuracy And Availability Of Sources Do Relate To Where Individuals Access Information

The comparative bar chart (Figure 5) answered a lot of the questions regarding this objective. We also ran a Chi-Square Test of Association. This test was significant with $\chi^2(4) = 110.9419$, $p < 0.0001$. This indicates there is a statistically significant results between where and why individuals choose an information source.

4. Conclusions

4.1 Results

Objective 1.1: Age significantly affects where and the reasons why individuals access everyday information, while professional Role does not.

Objective 1.2: As Age increase accuracy ratings decrease.

Objective 2: Accuracy significantly affects where individuals access equine disease information.

Objective 3: Accuracy and availability do not affect the reason why individuals access information.

Objective 4: There is an association between where individuals gather information first and what they consider to be more important between accuracy and availability for each source.

4.2 Future Improvement and Impact

Future Improvement

1. Better construction of the survey: to better examine the influence of professional roles and deal with quasi-separation

2. A better sampling of the population. Convenient sampling occured in this study due to where the sampling took place and also the large amount of females to males that took the survey.

Impact This research experiment explained what factors may influence peoples preferred news sources for equine disease information, which may help equine professionals deliver accurate equine disease information to the public in a more effective way.

References

[1] Rodriguez, German. n.d. Multinomial Response Models. Retrieved April 24, 2017 (http://data.princeton.edu/wws509/notes/c6.pdf)

[2] Anon. n.d. 8.2 - Baseline-Category Logit Model. 8.2 - Baseline-Category Logit Model — STAT 504. Retrieved April 25, 2017 (https://onlinecourses.science.psu.edu/stat504/node/173)

[3] Anon. n.d. 6.2.3 - More on Goodness-Of-Fit and Likelihood ratio tests. 6.2.3 - More on Goodness-of-Fit and Likelihood ratio tests — STAT 504. Retrieved April 25, 2017 (https://onlinecourses.science.psu.edu/stat504/node/220)

[4] Anon. n.d. What is a variance inflation factor (VIF)? What is a variance inflation factor (VIF)? - Minitab. Retrieved April 25, 2017 (http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/model-assumptions/what-is-a-variance-inflation-factor-vif/).

[5] Diener-West, Marie. 2008. Use of the Chi-Square Statistics. Retrieved April 24, 2017 (http://ocw.jhsph.edu/courses/FundEpiII/PDFs/Lecture17.pdf)

[6] West, Brady T., Kathleen B. Welch, and Andrzej T. Gaecki. 2015. Linear mixed models: a practical guide using statistical software. Boca Raton, Fla.: Chapman and Hall/CRC.