People's Preferred News Sources for Equine Disease Information

Abstract

The equine industry is a steadily growing and economically significant industry that is very prevalent within Colorado. Our client Shelly McDaniel, a graduate student in veterinary medicine at Colorado State University, is interested in determining what factors dictate where and why individuals gather information related to equine diseases. Our first objective is concerned with examining how age and role affect where and why individuals access information; after conducting our analysis age and role are significant in determining where individuals access information, however, only age is significant in why they chose the source. A linear mixed model was used for modeling why individuals chose the source. The second and third objectives examine how the availability and accuracy of sources affect where individuals seek out information and what the most important information they were looking for was in the face of a disease outbreak, respectively. Accuracy affected where individuals gathered information; availability and accuracy did not affect the reason why individuals gathered information. Objectives 1, 2, and 3 each used a global likelihood ratio test, a likelihood ratio test, measured collinearity with variance inflation factors, and fit a multinomial model. The fourth objective examined the relationship between where individuals chose their preferred news sources and why they chose the source first (accuracy, availability); after conducting the chi-square test of association, there was a significant association between where and why individuals chose to access information about equine diseases first.

1 INTRODUCTION

1.1 BACKGROUND

With nearly 10 million horses and close to 4.6 million Americans having some degree of involvement with horses, the equine industry has a very large economic and social impact. Due to the size of the industry, equine disease outbreaks can pose a very serious threat to all horse owners within the country. Our client Shelly McDaniel is a graduate student in veterinary medicine at Colorado State University and she is primarily interested in determining what factors affect where and why individuals choose to access information related to equine diseases. She is particularly concerned that people are gathering their information from unreliable sources; social media for example, might be considered unreliable and she wants to see if certain demographics are accessing information from certain sources more than others.

1.2 SURVEY DESIGN

Our data was provided by our client who sent out a survey to horse clubs across Colorado, with 256 responses and various questions concerning individuals' perception of several sources, in the event of an equine disease outbreak. The population of interest for this study is individuals involved in the equine industry in Colorado, however, because of the way the survey was sent out, inference can only be made on individuals involved in horse clubs within the state. It is also notable that most of the respondents to the survey were female. The survey divided participants by age cohorts and by roles. Age ranges consisted of 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and older. Roles of individuals consisted of professional, competitive, pleasure rider, horse owner, and other. Sources in the survey consisted of social, online, traditional, vet, state, profit, and other. Additionally, one question asked individuals where they first look for information in the event of an equine disease outbreak, traditional and profit were combined into the variable tradprof and online and other were combined into the variable media. This was done due to sparsity within the data, as there were few responses for these four sources regarding questions about accuracy and availability. Additionally, several questions asked participants to rate the accuracy and availability of sources, based on a Likert-scale. In the original survey, participants rated the accessibility of sources, but due to participants being unable to distinguish between availability and accessibility, the two categories were combined into availability.

1.3 OBJECTIVE

Objective 1

This objective was concerned with determining how age and role affect where and why individuals gather information related to equine disease information. The first half of this objective (Objective 1.1) deals with how age and role affect where individuals gather information.

For the second half of this objective (Objective 1.2), we determined how age and role affect why individuals gather information. Additionally, participants in the survey were asked to rate the accuracy and availability of information sources on a scale of 1-7, with 1 being the lowest and 7 being the highest. The sources that were ranked were Social, Online, and Traditional news sources.

Objective 2

This objective was concerned with determining how the accuracy and availability of sources affect where individuals gather information. In this case, individuals selected which information source they would choose first in the event of an equine disease outbreak and rated Online, Traditional, and Social sources based upon participants' perception of their accuracy and availability. These ratings of Online, Traditional, and Social for accuracy and availability were then compared to which source individuals chose first in the event of an outbreak (online, tradprof, media, social, state, vet), to see if ratings significantly affect which source individuals chose first.

Objective 3

In this objective, we are concerned with determining how accuracy and availability of sources affect why individuals access information. Individuals selected what the most important information was that they would look for in the event of an equine disease outbreak. Choices for this question were how the disease was spread, the symptoms and signs of the disease, impact on horse shows, traveling, or events, and other. Responses to this question were compared to individuals' ratings of accuracy and availability of Online, Traditional, and Social news sources, to see if these ratings affected why individuals accessed information.

Objective 4

Objective 4 is concerned with evaluating how the accuracy and availability of sources relate to where individuals gather information. One of the questions asked of participants is where do they first access information in the event of an equine disease outbreak (social, tradprof, media, vet, state). A follow up question asked them why they chose this resource first, for either its accuracy or availability.

2 METHOD

2.1. Global Likelihood Ratio Test – (Objectives 1,2,3)

H_0: $\beta_{i0} = \beta_{i1} = \ldots = \beta_{ij} = 0$ Ha: Not all $\beta_{i,j}=0$

This test is used to determine whether any of the predictor variables are associated with the response variable. If we reject the null hypothesis, then we can conclude that at least one of the independent variables has a significant impact on the dependent variable.

2.2. Multinomial Model – (Objectives 1.1,2,3)

Multinomial logistic regression is a method that generalizes logistic regression to problems with more than two possible outcomes. It is a model that is used to predict the probabilities of these different outcomes with a categorical dependent variable, given the independent variables. Regarding objective one, there was some sparsity within the data, so a penalized multinomial model was used because it can account for this data sparsity. This modeling assumes independence of irrelevant alternatives, or that the odds of preferring one class to another do not depend on the presence or absence of other "alternatives." Additionally, it is assumed that multicollinearity is low, which is tested with the variance inflation factors.

2.3. Likelihood Ratio Test – (Objectives 1,2,3)

H_0: $\beta_{i1} = \beta_{i2} = \ldots = \beta_{ij}$ H_a: Not all $\beta_{i,j} = 0$

This method compares the model for each specific objective against a null model, where nothing is included and sees which model is preferable, it is essentially testing the "goodness of fit" for the model.

2.4. Linear Mixed Model – (Objective 1.2)

$Y_{ij}=\mu+\alpha_i+\beta x_{ij}+a_i+\varepsilon_{ij}$
$a_j \sim N(0,\sigma^2)$ $\varepsilon_{ij} \sim N(0,\sigma^2)$

A linear mixed model combines fixed and random effects and is used for objective 1.2. The theoretical model above represents the combination of the fixed effects of age and source and the random effect of participants; all of which determine an individual's accuracy rating. Mu represents the theoretical mean, alpha is the source effect, beta represents the age effect, "a" represents the random effect of participant to participant variation, and epsilon is the error term of the model. Role was not included, as it is not significant. Additionally, "ai" and εij are normally distributed with a mean of zero and standard deviation of σ2.

$Y_{ij}=\mu+\alpha_i+\beta x_{ij}+ \gamma_i x_{ij}+a_i+\varepsilon_{ij}$
$a_j \sim N(0,\sigma^2)$ $\varepsilon_{ij} \sim N(0,\sigma^2)$

This linear mixed model is the mostly the same as the first, except it is concerning individuals' ratings of availability of sources and it includes an interaction term between age and source. The effect of the interaction term in this case is represented by gamma.

2.5. Variance Inflation Factors – (Objectives 1,2,3)

Variance inflation factors are used to check the assumption in the models we used, of no multicollinearity. It provides an index that measures how much the variance of the regression coefficients increase due to the multicollinearity.

2.6. Chi-Square Test of Association – (Objective 4)

The fourth objective uses the chi-square test of association to see whether there is a statistically significant relationship between two categorical variables. It is testing to see whether there is an association between where individuals choose to find information about equine diseases first and why they chose that source (accuracy or availability).

3 RESULT

3.1. Objective 1.1

For this objective, our values for the variance inflation factors were all 5 or less, which indicates that there is not much evidence for multicollinearity. After fulfilling this assumption, the likelihood ratio test was computed and this indicated that age and role were both significant predictors in determining where individuals seek out everyday information.

This is further backed up by the test statistics:

Age: $\chi^2(2) = 47.8282$, $p < .0001$ Role: $\chi^2(8) = 17.28$, $p < 0.0273$ (LRT)

Age and role are both significant, having p-values < 0.05.
As for the specific sources individuals choose to access information first (vet, state, tradprof, media, social), role was significant, while age was not:

Age: $\chi^2(4) = 6.0719$, $p = 0.1938$ Role: $\chi^2(16) = 36.461$, $p < 0.0025$ (LRT)

Despite role's significance with a p-value < 0.05, we cannot determine exactly in what way it is significant due to the problems involving quasi-separation in this objective.

3.2. Objective 1.2

Multicollinearity was also not an issue with this objective and the assumption of homogeneity of variance also held, as the residual diagnostics checked out. For this half of objective one, age and source were both important factors in determining how participants rated accuracy and availability. After running the Type-III test for fixed effects (for age and source) we obtained the following test statistics:

Age*: F*(1, 250) = 6.01, p=0.0149 Source*: F*(2, 482) = 16.952, p < 0.0001
Interaction*: F*(2, 487) = 16.64, p < 0.0001

Age and Source are significant for determining participants' ratings of accuracy, both having p-values below 0.05, however, the interaction between the two is significant. Thus, we cannot interpret the main effects of these two variables alone because the effects of each ar dependent on the levels of one another.

As for availability, age and source both significantly affect participants' ratings of this as both p-values were significant:

Age*: F*(1, 244) = 8.60, p=0.0037
The estimated slope for age is - 0.011 Source*: F*(2, 482) = 263.30, p < 0.0001.

It's of note that the slope for age is negative, so individuals trust all sources less overall, as they increase in age. Role was also not a significant predictor of either accuracy or availability ratings.

3.3. Objective 2

Similarly to the previous objective, multicollinearity was not an issue and the likelihood ratio test could be computed. After running this test, accuracy ratings proved to be important in two separate cases. The test statistic for social (χ^2 (4) = 10.7492 , p=0.0295) proved to be significant when looking at the probability people selected vet sources versus social sources (χ^2 (1) = 9.4140, p= 0.0176). Additionally, the test statistic for traditional (χ^2 (4) =11.93431, p= 0.0178) was significant when looking at the probability people selected tradprof sources versus social sources (χ^2 (1) = 7.4421, p= 0.0448). Availability was not significant in determining anything in this objective.

In addition to the test statistics gained from the likelihood ratio test, we could gain odds ratios that were interpretable, due to there not being quasi-separation issues with this objective:

The odds ratio estimate of vet versus social: 0.733, 95% CI [0.298，0.766].
The odds ratio estimate of tradprof versus social: 4.501, 95% CI [1.656，21.712]

These odds ratios can be interpreted as the odds that the first outcome will happen relative to the other. For example, in the second odds ratio, the odds of an individual choosing tradprof to social are 4.501 to 1, or that individuals are 4.501 times more likely to choose tradprof over social for every one unit increase in accuracy rating.

3.4. Objective 3

For objective three, participants were asked what the most important information they looked for in the event of an equine disease outbreak and this was compared to the ratings of

accuracy and availability to see if they affected individuals' choices of the most important information. After running the global likelihood ratio test we obtained a test statistic of:

χ^2 (18) = 14.4571, p= 0.6988

This tests to see if anything in the model is important and since the p-value is greater than 0.05, we can conclude that ratings for accuracy and availability of sources does not affect individuals' choices for what they consider the most important information is that they're looking for, in the event of an equine disease outbreak.

3.5. Objective 4

Regarding objective four, the chi-square test of association was significant, having a test statistic of 110.94 and an extremely small p-value of 4.582e-23. These results indicate that there is a statistically significant association between where individuals gathering information first in the event of an equine disease outbreak and why they choose that resource first (accuracy, availability). This can also be seen when referring to Figure 4.1.3, where the difference in accuracy and availability preference regarding the various sources, is quite evident. Individuals highly value accuracy for state and vet sources and they value availability for tradprof, social, and media sources.

4 cONCLUSION

4.1. Summary

This study was concerned with evaluating how accuracy and availability affect individuals' choices for news sources in the event of an equine disease outbreak and it resulted in several interesting observations about the relationship between these variables. Firstly, objective one shows that age significantly affects where individuals gather everyday information, as does role, however, it cannot be determined in what way due to the quasi-separation present. Additionally, objective one shows that age and source affect the ratings individuals give for accuracy and availability of sources. Objective two shows us that accuracy significantly affects where individuals gather equine disease information. The third objective was the only one that had nothing significant, indicating that accuracy and availability do not affect the reason why individuals gather information. The final objective showed that there is a significant association between where individuals decide to gather information first in the event of an equine disease outbreak and whether they consider the accuracy or availability of the source to be more important.

4.2. Discussion

Going forward with these results, we can address the concerns our client Shelly had concerning tendencies for certain demographics to access unreliable information concerning equine diseases. From our analysis, it is evident that younger people tend to use social sources for their everyday news source, where older individuals prefer traditional sources. Traditional sources may be more reliable, so concentrating efforts to either steer younger people toward traditional sources or put more reliable information on social media might be effective. Additionally, those who prefer social media as their preferred source of news tend to rate the accuracy for social sources higher than those who prefer other every day news sources. Despite younger people's preference for social sources, participants recognized that state and vet sources were where they should go for accuracy and social sources were more convenient for their availability. Our client should be relieved to find that despite younger people's preference for social media, they do recognize that the information presented to them may not necessarily be accurate and that they can access state and vet sources for much more accurate information.

5 BIBLIOGRAPHY

Allison, Paul D. "Convergence Failures in Logistic Regression." *SAS Global Forum 2008*. N.p., 2008. Web. 27 Apr. 2017.

"Conduct and Interpret a Multinomial Logistic Regression." *Statistics Solutions*. N.p., n.d. Web. 28 Apr. 2017.

G, Rodriguez. “Multinomial Response Models - Princeton University.” N.p., Sept. 2007. Web.15.2017

“SAS/STAT(R) 9.22 User’s Guide. ” SAS/STAR(R) 9.22 User’s Guide. SAS, n.d. Web. 15.2017

West, B. T., Welch, K. B., Gałecki, A. T., & Gillespie, B. W. (2015). Linear mixed models a practical guide using statistical software. Boca Raton, Fla: CRC Press.

6 APPENDIX

Exploratory Data Analysis

Conducting an exploratory data analysis (EDA) is a very useful way to notice trends in the data, before it is analyzed any further. In our EDA, the following plots showed various trends that we could use as a baseline for answering the questions posed in our four objectives.

Figure 4.1.1 shows us what news sources individuals prefer for their everyday news across different age groups. As is evident in the comparative bar chart, as age increases, people prefer traditional sources and do not prefer social news sources.

Figure 4.1.2 shows the distribution of preferred news sources across various roles in the equine industry. A constant across all roles is the preference for state and vet resources.

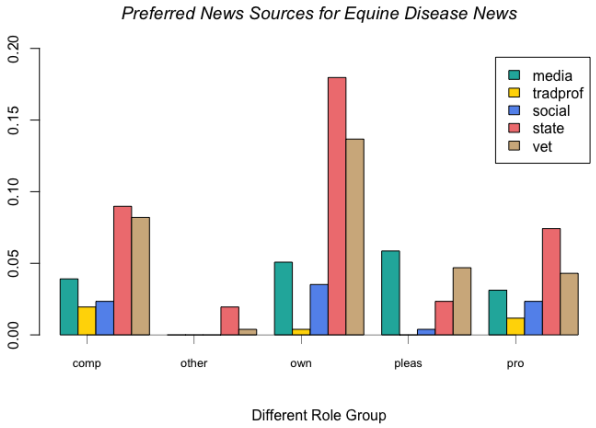

Figure 4.1.1

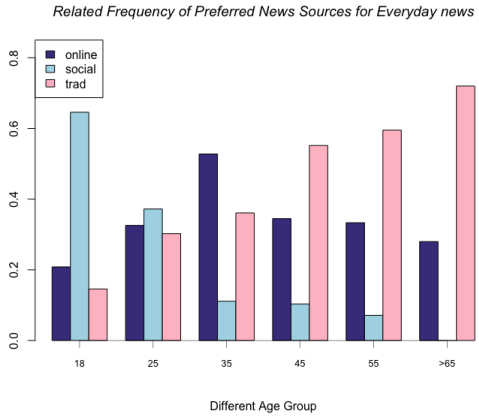

Figure 4.1.2

Figure 4.1.3 shows a comparative bar chart of the relative frequencies of individual's choice of accuracy or availability across all 5 sources. This chart shows that there's evidence of an association between where people first seek out information and why they seek it out. For the sources media, tradprof, and social, individuals seem to highly value the availability of information over the accuracy and for state and vet, individuals seek out these two sources for their accuracy.

Figure 4.1.4 shows comparative boxplots of individuals' ratings of accuracy and availability of five types of news sources, across online, social, and traditional sources. The boxplots show that there is not much of a difference in ratings for availability across all three sources, however, the accuracy ratings for social are much different than for online and traditional.

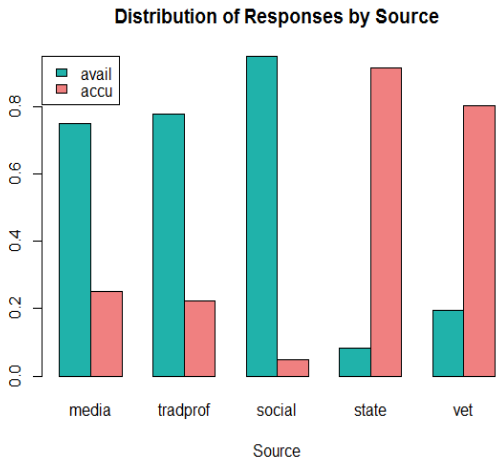

Figure 4.1.3

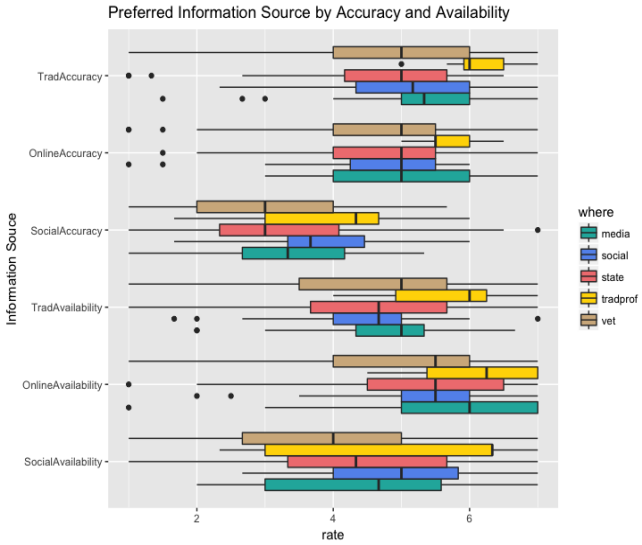

Figure 4.1.4

Result Graph Analysis

Figure 4.2.1 shows that as age increases, the probability of an individual choosing traditional for their source of every day news increases. Additionally, it shows that as age increases the possibility of them choosing social declines.

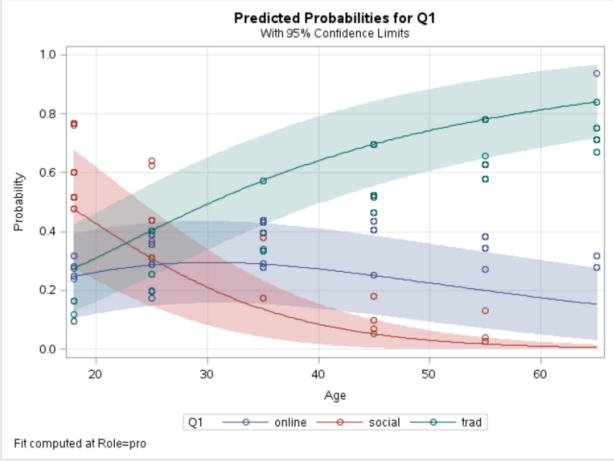

Figure 4.2.1

Figure 4.4.1 shows that the probability of an individual selecting social sources first in the event of an equine disease outbreak increases, as the accuracy rating for Social sources increases. Additionally, the probability of selecting vet goes down as individuals' accuracy scores for Social increases.

Figure 4.4.2 shows that the probability of an individual selecting tradprof sources first in the event of an equine disease outbreak increases, as the accuracy rating for Traditional sources increases. In contrast to this, the probability of an individual selecting social sources first decreases, as the accuracy rating for Traditional sources increases.

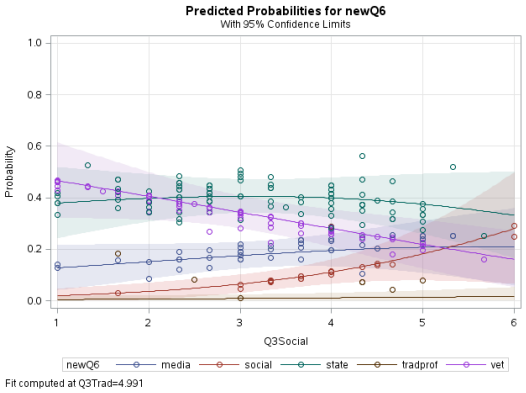

Figure 4.4.1

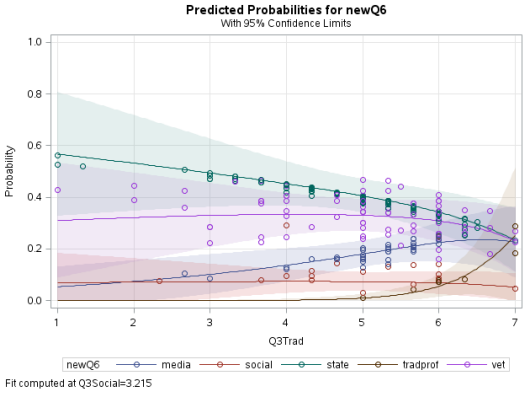

Figure 4.2.2