Statistics Simplified

Relationships between categorical dependent variables and other variables and between waiting times and other variables

Categorical dependent variables and censored data are frequently encountered in veterinary research. Statistical methods for analyzing these types of data will be discussed here. Some of the statistics, such as relative risks, odds ratios, and hazard ratios, may seem rather technical. The effort required to understand them is well spent, however, because they provide important information in many veterinary studies.

Relationships Between Categorical Dependent Variables and Other Variables

The relationship between a categorical dependent variable and another variable is often of interest in veterinary research. For example, a study[1] was conducted to determine prognostic indicators for 46 horses that were treated medically for colic. In that study, the outcome of interest, also known as the dependent variable, is euthanasia (euthanized vs discharged from hospital). Packed cell volume at hospital admission is a potential prognostic indicator and was thus treated as an independent variable.

Correlation and least squares regression cannot be used because the dependent variable is categorical. Instead, logistic regression should be considered. *Logistic regression* is a statistical procedure for investigating the relationship between a categorical dependent variable and 1 or more independent variables. In *binary logistic regression*, the dependent variable has 2 categories, which are coded 0 (typically for characteristic absent) or 1 (for characteristic present). In *multinomial logistic regression* (also called *polytomous logistic regression*), the dependent variable has more than 2 categories.

Logistic regression with 1 independent variable is called *univariable* or *univariate logistic regression*, and that with 2 or more independent variables is called *multivariable* or *multivariate logistic regression*. Univariate logistic regression will be discussed here, and the multivariate procedure will be discussed in a future article. Although independent variables can be categorical or noncategorical, special numeric coding similar to that used for categorical dependent variables is needed for categorical independent variables.

The goal of logistic regression is to determine whether the dependent variable is related to the independent variable. Sometimes, prediction is also a goal. When the dependent and independent variables are strongly related, the independent variable can sometimes be used to predict the dependent variable. For example, suppose the dependent variable is the outcome (success or failure) of tumor excision in parrots. A preoperative independent variable that could accurately predict the outcome would be quite useful. However, significant relationships do not guarantee accurate prediction and independent variables that predict well enough to be clinically useful are hard to find.

Logistic regression is used to investigate whether an independent variable is related to the logarithm of the odds, called the log odds or logit, for the dependent variable. The *odds* of an event is the probability that the event will occur divided by the probability that the event will not occur. For example, if a cat has a 75% chance of vomiting on the couch when visitors arrive, its odds of vomiting are 75%/25% = 3/1, or 3 to 1. For a dependent variable with 2 categories, the odds equal the probability of the category coded 1 divided by the probability of the category coded 0, such as the probability of wound infection divided by the probability of no wound infection.

The univariate logistic regression equation expresses the log odds as follows:

$$\text{Estimated log odds} = \text{constant} + (\text{coefficient} \times \text{independent variable})$$

The *coefficient*, which is the number by which the independent variable is multiplied, provides information about how the dependent and independent variables are related. When the coefficient is positive, the probability of the dependent variable category coded 1 increases as the value of the independent variable increases. When the coefficient is negative, the probability of the dependent variable category coded 1 decreases as the value of the independent variable increases. When the coefficient is 0, there is no linear relationship between the independent variable and the log odds for the dependent variable. This does not mean that there is no relationship. The size of the logistic regression coefficient indicates nothing about the strength of the relationship between the dependent and independent variables because it is affected by the unit of measurement of the independent variable.

For the colic data, euthanasia was coded as 1 = euthanized and 0 = discharged from hospital. The independent variable PCV has a positive logistic regression coefficient. If the relationship between euthanasia and PCV is significant, one can conclude that the probability of euthanasia increases as PCV increases.

To determine whether the dependent and independent variables have a significant relationship, we test the null hypothesis that the population logistic regression coefficient for the independent variable is 0. The alternative hypothesis states that the population logistic regression coefficient is not 0. Two commonly used methods to test this hypothesis are the *Wald test* and the *likelihood ratio test*. Because the Wald test has substantial drawbacks, the likelihood ratio test is preferred.[2]

This report was submitted by Susan Shott, PhD; from Statistical Communications, PO Box 671, Harvard, IL 60033.
Address correspondence to Dr. Shott (stattwit@aol.com).

To test the hypothesis that a population logistic regression coefficient is 0, the data do not need a normal distribution or any other distribution. However, the following assumptions are involved:

- *Random sampling*. If the sample is not biased, random sampling is not essential.
- *Noncensored observations*. None of the observations can be based on censored data.
- *Independent observations*. All of the observations for the dependent variable must be independent, and all of the observations for the independent variable must be independent. The observations for the dependent variable are not necessarily independent of those for the independent variable because these variables may be related.
- *Linear relationship*. When a relationship exists between the independent variable and the log odds for the dependent variable, that relationship should be linear. Checking this assumption requires advanced statistical methods described elsewhere.[2–4] If there is reason to suspect a nonlinear relationship, the independent variable can sometimes be mathematically transformed to satisfy the linearity assumption.

For the colic data, to determine whether we can test the hypothesis that PCV has a population logistic regression coefficient of 0, the assumptions for logistic regression need to be checked. The data have not been obtained from a random sample of horses, but this is not critical. Although the data concern survival, none of the observations are censored. The outcome of interest to the researchers was survival during the hospital stay, not survival after hospital discharge. For each horse, the researchers knew that outcome. The euthanasia data are independent because the euthanasia outcome for one horse tells us nothing about the euthanasia outcome for another horse. The PCV data are independent because one horse's PCV tells us nothing about another horse's PCV, and only 1 PCV measurement/horse was used in the analysis. The linearity assumption was checked and found to be reasonable.

Because the assumptions appear to hold, the hypothesis can be tested that the population logistic regression coefficient for PCV is 0. A significance level of 0.05 was chosen. Because the *P* value obtained for this test was < 0.05, we can conclude that the population logistic regression coefficient for PCV is not 0.

Understanding the results—Logistic regression coefficients are commonly converted into odds ratios, which are used to describe the relationship between the dependent and independent variables. The *odds ratio* is the odds for one group divided by the odds for another group.

In logistic regression, the odds ratio tells us how many times larger or smaller the odds for the dependent variable become when the independent variable increases 1 unit. For the colic data, the odds ratio associated with PCV is 1.22. This means that the odds of euthanasia become about 1.22 times as large when the PCV goes up 1%. For example, the odds of euthanasia when the PCV is 60% is about 1.22 times as large as the odds of euthanasia when the PCV is 59%.

The odds ratio is sometimes used to estimate the relative risk or relative probability of an event. The *relative risk* of an event is the risk of the event for one group divided by the risk of the event for another group. For example, if cats are 4 times as likely as dogs to vomit on the couch when visitors arrive, the relative risk of vomiting for cats, compared with that for dogs, is 4. When the relative risk for 2 groups is 1, the groups have the same risk. When it is > 1, the group in the relative risk numerator has a higher risk than the other group, and when it is < 1, the group in the relative risk numerator has a lower risk. The odds ratio can be used as an estimate of relative risk when the probability of the event is small and other requirements, which are described elsewhere, are met.[5]

In the colic study, PCV was used as a noncategorical independent variable. In a study[6] of antibody titers against canine distemper virus in 431 dogs at an animal shelter, logistic regression was performed with categorical independent variables. The dependent variable in that study[6] is protective antibody titer at shelter admission, coded 1 = present and 0 = absent.

Univariate logistic regression found a significant relationship between protective antibody titer and the categorical independent variable neuter status, coded 1 = neutered and 0 = unneutered. Because the regression coefficient is positive, this indicates that neutered dogs are more likely than unneutered dogs to have protective antibody titers. The coefficient can be used to calculate the odds ratio for neuter status, which is 8.3, meaning the odds of having a protective antibody titer are about 8.3 times as large for neutered dogs as for unneutered dogs.

The independent variable category that is coded 0 is called the *reference category* or *referent* because the other category (the *nonreference category* or *nonreferent*) is compared with it. For the shelter dog study, the reference category is the unneutered category. In general, the odds ratio for a categorical independent variable indicates how many times larger or smaller the dependent variable odds are for the nonreference category than for the reference category.

When the odds ratio for a categorical independent variable is 1, the nonreference and reference categories have the same odds for the dependent variable. When the odds ratio is > 1, the nonreference category has larger odds for the dependent variable than does the reference category, and when the odds ratio is < 1, the nonreference category has smaller odds. Interpreted as a relative risk estimate, the odds ratio tells us how many times more or less likely the dependent variable category coded 1 is for the nonreference category than for the reference category.

To interpret logistic regression odds ratios, the reader needs to know the coding of the dependent variable and the categorical independent variables. However, this is not always clearly described in veterinary reports. In some studies, the reference category is indicated by an odds ratio of 1.

Paired data—Sometimes, a logistic regression analysis is desired for a dependent variable that is based on paired data. For example, a case-control study[7] of risk factors for endotoxin-induced mastitis involved

cows with mastitis and control cows without mastitis that were individually matched with mastitis cows. Matching ensured that each control cow came from the same herd and had a parturition date similar to that of the matched mastitis cow. The dependent variable was endotoxin-induced mastitis (present or absent).

Because the mastitis categories were obtained from matched cows, the data are paired. The independence assumption is violated, so the usual logistic regression methods cannot be used. However, an adjusted logistic regression procedure called *paired logistic regression* or *conditional logistic regression* can be used to investigate relationships between the dependent variable and other variables. Additional information about this procedure and other aspects of logistic regression can be found elsewhere.[2,3]

Relationships Between Waiting Times and Other Variables

When relationships between survival times or other waiting times and other variables are investigated, the waiting times are often censored. Censored data rule out the possibility of using correlation coefficients, least squares regression, and logistic regression to assess relationships between variables. However, statistical procedures called *survival analysis methods* have been developed to handle waiting times with censored data. Although the term survival analysis suggests that these methods only apply to survival times, they can be used for other waiting times as well. They can also be used to analyze waiting times that do not include censored data.

In studies of waiting times, a graph is useful to show how a group of animals experiences the event of interest over time. The *Kaplan-Meier curve* provides this information by plotting the waiting time on the horizontal axis and the percentage or proportion of animals that have not experienced the event on the vertical axis. Censored waiting times are represented by tick marks or other symbols that indicate when the waiting times were censored. However, many Kaplan-Meier curves in veterinary reports do not include censored data symbols. These symbols convey important information. For example, a study with many animals that were censored early is generally not as useful as a study with animals that mostly had long follow-up times before censoring.

For example, a Kaplan-Meier curve was constructed to show survival time for 50 dogs with osteosarcoma treated with amputation and chemotherapy (**Figure 1**).[8] Here, the event of interest is death. When the curve drops, at least 1 dog died at that time point. The curve quickly begins to fall very steeply, indicating that many dogs died rapidly. In this curve, dogs with censored survival times are represented by circles. For example, 1 dog was censored after > 2,000 days of follow-up.

Kaplan-Meier curves involve the following assumption: *equivalent waiting time experience for animals with censored waiting times and animals without censored waiting times*. If animals have censored waiting times, their waiting time experience should be the same as that for animals without censored waiting times. For example, suppose sheep were withdrawn from a fertility drug study before they became pregnant. These sheep have censored times to pregnancy. Their exact times to pregnancy are unknown because they did not stay in the study long enough to become pregnant. All that is known is that it would have taken at least as long as the time they remained in the study to become pregnant. If they were withdrawn because of adverse effects that also delayed pregnancy, their time-to-pregnancy experience differs from that of sheep without censored times to pregnancy. The assumption of equivalent waiting time experience is violated.

When this assumption does not hold, the results may be biased. However, checking this assumption is difficult. The only way to discover whether animals with censored waiting times have the same waiting time experience as other animals is to obtain complete follow-up information for them. However, clinical differences between animals with and without censored waiting times may indicate that this assumption is violated. For example, suppose the time from canine total hip replacement to failure of the implant is investigated. If dogs with censored data have significantly worse degenerative joint disease than do dogs without censored data, their times to failure may be shorter. For the canine osteosarcoma data, we do not have enough information to determine whether this assumption is reasonable. For this reason, we cannot rule out the possibility that the results are biased.

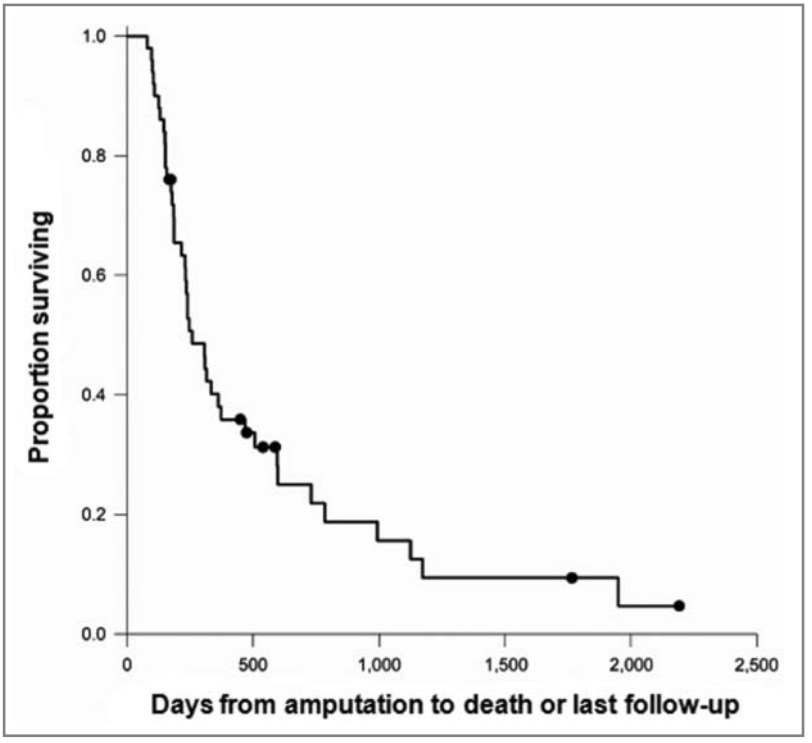

Figure 1—Kaplan-Meier survival curve for 50 dogs with osteosarcoma. Censored survival times are indicated by circles. (Adapted from Bacon NJ, Ehrhart NP, Dernell WS, et al. Use of alternating administration of carboplatin and doxorubicin in dogs with microscopic metastases after amputation for appendicular osteosarcoma: 50 cases (1999–2006). *J Am Vet Med Assoc* 2008;232:1504–1510. Reprinted with permission.)

The *log-rank test* is commonly used to determine whether a waiting time is related to a categorical variable. Each category defines a group of animals. These groups are compared to determine whether there is a significant difference with respect to their waiting times. The log-rank test is a test of the null hypothesis that all of the groups have the same population waiting time curves. The alternative hypothesis states that at least 2 groups have different population waiting time curves. Note that it is differences in the waiting time curves (ie, the pattern over time) and not in mean waiting times that pertain here.

When > 2 groups are compared and a significant difference is found, one cannot conclude that all groups have different population waiting time curves. Additional log-rank tests need to be performed to compare the groups 2 at a time to identify the groups that differ. Other statistical tests can also be used to compare groups with respect to waiting times. A description of these tests can be found elsewhere.[9,10]

The log-rank test is based on the following assumptions:

- *Random sampling*. A random sample is not essential as long as the sample is not biased.
- *Independent observations*. All of the waiting times must be independent.
- *Equivalent waiting time experience for animals with censored waiting times and animals without censored waiting times*. The waiting time experience should be the same for animals with censored waiting times and animals without censored waiting times.

In a study[11] of lymphoma in cats, the duration of remission was investigated for 22 cats with a complete response to chemotherapy and 15 cats with a partial response. Here, the event of interest is end of remission (progression or recurrence). The remission curve for cats with a complete response is higher than the curve for cats with a partial response (**Figure 2**). Remission duration is better for the higher curve because the proportion of cats in remission is larger for this curve than for the lower curve at each time point.

The difference between the curves is substantial; however, we still need to know whether this difference is significant. To answer this question, the log-rank test assumptions need to be checked. Random sampling was not done, but this is not essential. The remission duration for one cat tells us nothing about the remission duration for another cat, so the remission durations are independent. Sufficient information is not available to check the assumption of equivalent remission duration experience for cats with and without censored durations. Because the risk of bias is unknown, the results of the log-rank test must be interpreted with caution.

The log-rank test was used to test the hypothesis that the population remission curves are the same for the complete-response and partial-response groups. A significance level of 0.05 is chosen. Because the test yielded a *P* value of 0.013, this hypothesis is rejected. The conclusion is that remission duration is related to treatment response.

To determine whether a waiting time is related to a noncategorical independent variable, *Cox proportional hazards regression* or *Cox regression* is often used. Categorical independent variables can also be used in Cox regression, but they must be coded in a special way (usually 0 and 1). For example, in a study[12] of 275 horses, Cox regression was used to identify independent variables that were related to long-term survival after surgery for large intestinal disease. Here, the event of interest is death.

In *univariable* or *univariate Cox regression*, only 1 independent variable is analyzed. In *multivariable* or *multivariate Cox regression*, 2 or more independent variables are analyzed. Univariate Cox regression will be discussed here, and multivariate Cox regression will be discussed in a future article.

Cox regression is used to investigate whether the hazard function for the waiting time is related to the independent variable. A precise definition of the *hazard function* requires calculus, but it can be thought of as the instantaneous potential per unit of time for the event of interest to occur, given that the animal has not yet experienced the event. The univariate Cox regression equation expresses the logarithm of the hazard function as follows:

Estimated log hazard function =
constant + (coefficient × independent variable)

Similar to the situation in logistic regression, the coefficient can be used to determine the nature of the relationship between the event of interest and the independent variable. A positive coefficient means that the probability of the event increases as the value of the independent variable increases, and a negative coefficient means that the probability decreases as the value of the independent vari-

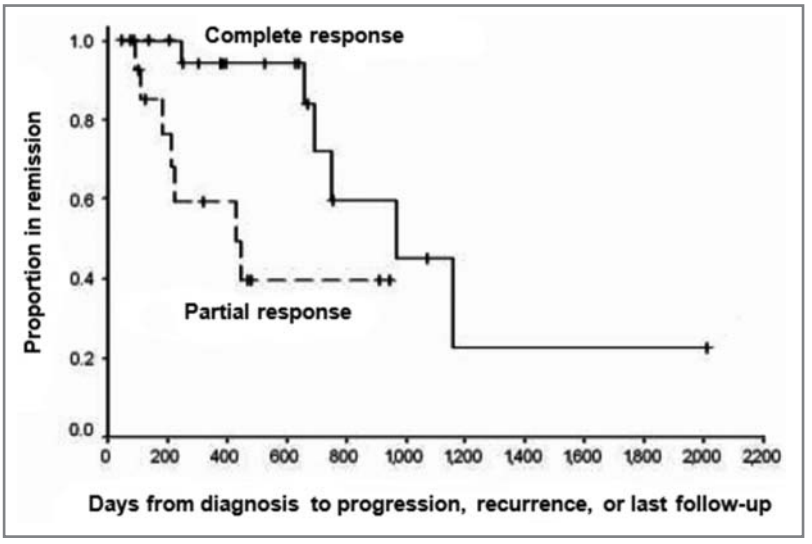

Figure 2—Kaplan-Meier remission curves for 22 cats with a complete response and 15 cats with a partial response to lymphoma treatment. Censored remission durations are indicated by tick marks. (Adapted from Kiselow MA, Rassnick KM, McDonough SP, et al. Outcome of cats with low-grade lymphocytic lymphoma: 41 cases (1995–2005). *J Am Vet Med Assoc* 2008;232:405–410. Reprinted with permision.)

able increases. A coefficient of 0 means there is no linear relationship between the logarithm of the hazard function for the event and the independent variable; however, it does not mean that there is no relationship. The size of the Cox regression coefficient provides no information about the strength of the relationship between the dependent and independent variables because it is affected by the independent variable's unit of measurement.

For the horse surgery data, Cox regression yielded a positive coefficient for the independent variable age. If the relationship between the hazard function and age is significant, one can conclude that the risk of death increases as age increases. This relationship is significant if we can reject the null hypothesis that the population Cox regression coefficient for the independent variable is 0. The alternative hypothesis states that the population Cox regression coefficient is not 0. Although the Wald test and the likelihood ratio test can both be used to test this hypothesis, the likelihood ratio test is preferred.[9]

To test this hypothesis, normality or any other data distribution is not needed. However, as for the log-rank test, the assumptions of random sampling, independent observations, and equivalent waiting time experience for animals with and without censored waiting times are involved. In addition, the following assumptions must also be met:

- *Linear relationship*. Any relationship between the logarithm of the hazard function and the independent variable should be linear. The methods used to check this assumption are complicated and described elsewhere.[9] When the relationship is not linear, the independent variable can sometimes be transformed to satisfy the linearity assumption.
- *Proportional hazards*. For any 2 animals with different values for the independent variable, the ratio of their hazard functions should be constant over time. For example, if Cox regression is performed with weight as the independent variable, the ratio of the hazard functions for any 2 animals with different weights should remain the same for all time points. A description of the methods for checking this assumption, which are fairly complex, can be found elsewhere.[9,10]

For the horse surgery data, we need to determine whether the hypothesis can be tested that the population Cox regression coefficient for age is 0. Although a random sample was not obtained, this is not necessary. Because the survival time for one horse tells us nothing about the survival time for another horse, the survival times are independent. The linearity and proportional hazards assumptions were checked and found to be reasonable. Insufficient information is available to check the assumption of equivalent survival time experience, so the possibility of bias cannot be ruled out. A significance level of 0.01 is chosen. Because Cox regression yielded a P value < 0.01, we can conclude that the population Cox regression coefficient for age is not 0.

Understanding the results—Hazard ratios are commonly calculated from Cox regression coefficients and used to describe the relationship between the hazard function and the independent variable. The *hazard ratio* tells us how many times larger or smaller the hazard function becomes when the independent variable increases 1 unit. It also estimates the relative risk of the event when the independent variable increases 1 unit. For the horse surgery data, the hazard ratio for age is 1.07. This means that the hazard function becomes about 1.07 times as large when age increases 1 year. Interpreted as a relative risk estimate, the hazard ratio means that the risk of death becomes about 1.07 times as large when age increases 1 year.

The independent variable age is noncategorical. In a study[13] of bone fractures in dogs and cats, Cox regression was performed to investigate relationships between categorical independent variables and the time from fracture surgery to bone union in 24 dogs and cats. Here the event of interest is union. A significant relationship was found between time to union and fracture comminution (1 = major and 0 = none or minor).

The Cox regression coefficient for comminution in that study is negative, which indicates the chance of union is smaller for animals with major comminution. The hazard ratio for comminution is 0.26, indicating that the hazard function for animals with major comminution is about 0.26 times as small as the hazard function for animals without major comminution. Interpreted as a relative risk estimate, the hazard ratio means that the chance of union is about 0.26 times as small for animals with major comminution as for animals without major comminution.

As in logistic regression, the reference category or referent for an independent variable is the category coded 0. The category coded 1 is the nonreference category or nonreferent. In general, the hazard ratio for a categorical independent variable tells us how many times larger or smaller the hazard function is for the nonreference category versus the reference category.

When the hazard ratio for a categorical independent variable is 1, the nonreference and reference categories have the same hazard function. When the hazard ratio is > 1, the nonreference category has a larger hazard function than the reference category, and when it is < 1, the nonreference category has a smaller hazard function. Interpreted as a relative risk estimate, the hazard ratio tells us how many times more or less likely the event of interest is for the nonreference category versus the reference category.

To interpret the hazard ratio when the independent variable is categorical, the reader needs to know which category is the reference category. This information is sometimes difficult to decipher in veterinary reports. In some studies, a hazard ratio of 1 is used to identify the reference category.

Although Cox regression analyses are very useful in veterinary research, they can be quite complicated. Further discussion of Cox regression can be found elsewhere.[9,10]

Research Myths

Misleading research myths about associations between variables are widespread. One popular research myth holds that 2 variables are not related if they are not correlated. This is incorrect for several reasons. First, 2 variables with a Spearman or Pearson correlation of 0 can have a strong nonlinear relationship because correlations

measure only linear relationships. Second, 2 variables may appear unrelated when other variables are not taken into account but may be strongly related in a model that includes other variables. Finally, failure to find a significant correlation may be attributable to low statistical power rather than a population correlation coefficient of 0.

Another common research myth is that if 2 methods for measuring the same quantity are highly correlated, they produce the same measurements. In fact, 2 methods that always produce different measurements can be perfectly correlated. For example, suppose 2 observers rate the degree of osteoarthritis on radiographs of equine knee joints. If one observer's score is always exactly 6 points higher than the other observer's score, the Pearson and Spearman correlation coefficients for the 2 scores will be 1. When 2 methods are evaluated for equivalence, correlations are important but not sufficient. Additional statistical analyses (eg, the Friedman test) must also be performed to determine whether a given method yields higher values than another method.

A persistent research myth is that variables are causally related if they are associated. This myth is sometimes perpetuated in the discussion and conclusion sections of veterinary reports. The wording shifts from statements about association to claims about a supposed causal relationship between variables. Although association is necessary to establish causation, it does not, by itself, imply causation.

Statistically literate readers of the veterinary literature can detect these research myths and many other statistical errors. Because these mistakes are widespread, critical statistical evaluation of research reports is an essential part of staying informed about new developments in veterinary medicine.

References

1. Ihler CF, Venger JL, Skjerve E. Evaluation of clinical and laboratory variables as prognostic indicators in hospitalised gastrointestinal colic horses. *Acta Vet Scand* 2004;45:109–118.
2. Hosmer DW, Lemeshow S. *Applied logistic regression*. 2nd ed. New York: Wiley, 2000.
3. Kleinbaum DG, Klein M. *Logistic regression: a self-learning text*. 3rd ed. New York: Springer, 2010.
4. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001.
5. Gordis L. *Epidemiology*. 4th ed. Philadelphia: Elsevier-Saunders, 2008.
6. Lechner ES, Crawford PC, Levy JK, et al. Prevalence of protective antibody titers for canine distemper virus and canine parvovirus in dogs entering a Florida animal shelter. *J Am Vet Med Assoc* 2010;236:1317–1321.
7. Menzies FD, Gordon AW, McBride SH, et al. Risk factors for toxic mastitis in cows. *Vet Rec* 2003;152:319–322.
8. Bacon NJ, Ehrhart NP, Dernell WS, et al. Use of alternating administration of carboplatin and doxorubicin in dogs with microscopic metastases after amputation for appendicular osteosarcoma: 50 cases (1999–2006). *J Am Vet Med Assoc* 2008;232:1504–1510.
9. Hosmer DW, Lemeshow S, May S. *Applied survival analysis: regression modeling of time-to-event data*. 2nd ed. New York: Wiley, 2008.
10. Kleinbaum DG, Klein M. *Survival analysis: a self-learning text*. 2nd ed. New York: Springer, 2005.
11. Kiselow MA, Rassnick KM, McDonough SP, et al. Outcome of cats with low-grade lymphocytic lymphoma: 41 cases (1995–2005). *J Am Vet Med Assoc* 2008;232:405–410.
12. Proudman CJ, Edwards GB, Barnes J, et al. Modelling long-term survival of horses following surgery for large intestinal disease. *Equine Vet J* 2005;37:366–370.
13. Kirkby KA, Lewis DD, Lafuente MP, et al. Management of humeral and femoral fractures in dogs and cats with linear-circular hybrid external skeletal fixators. *J Am Anim Hosp Assoc* 2008;44:180–197.