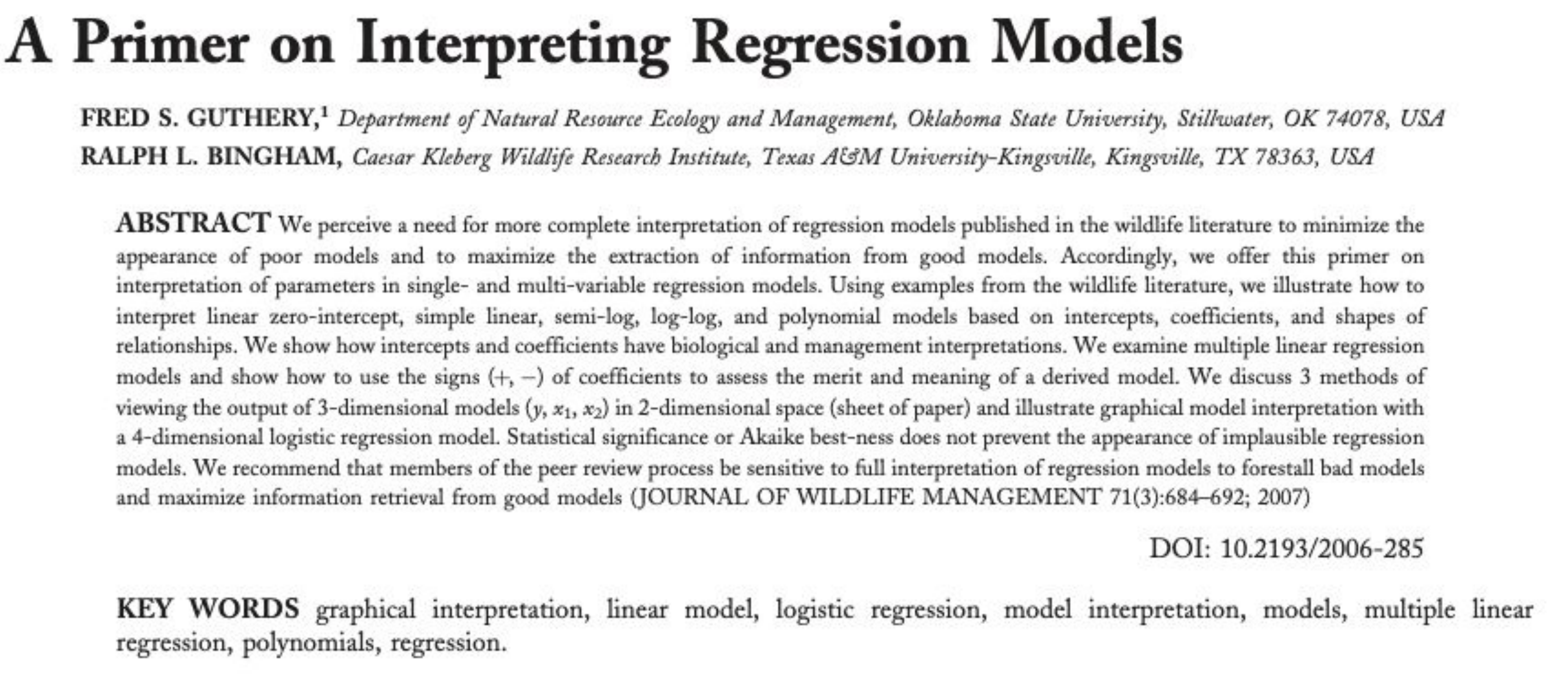
Ethics 2 - Live Session 1

Session will run 09:00-09:50



A Primer on Interpreting Regression Models

FRED S. GUTHERY,[1] *Department of Natural Resource Ecology and Management, Oklahoma State University, Stillwater, OK 74078, USA*

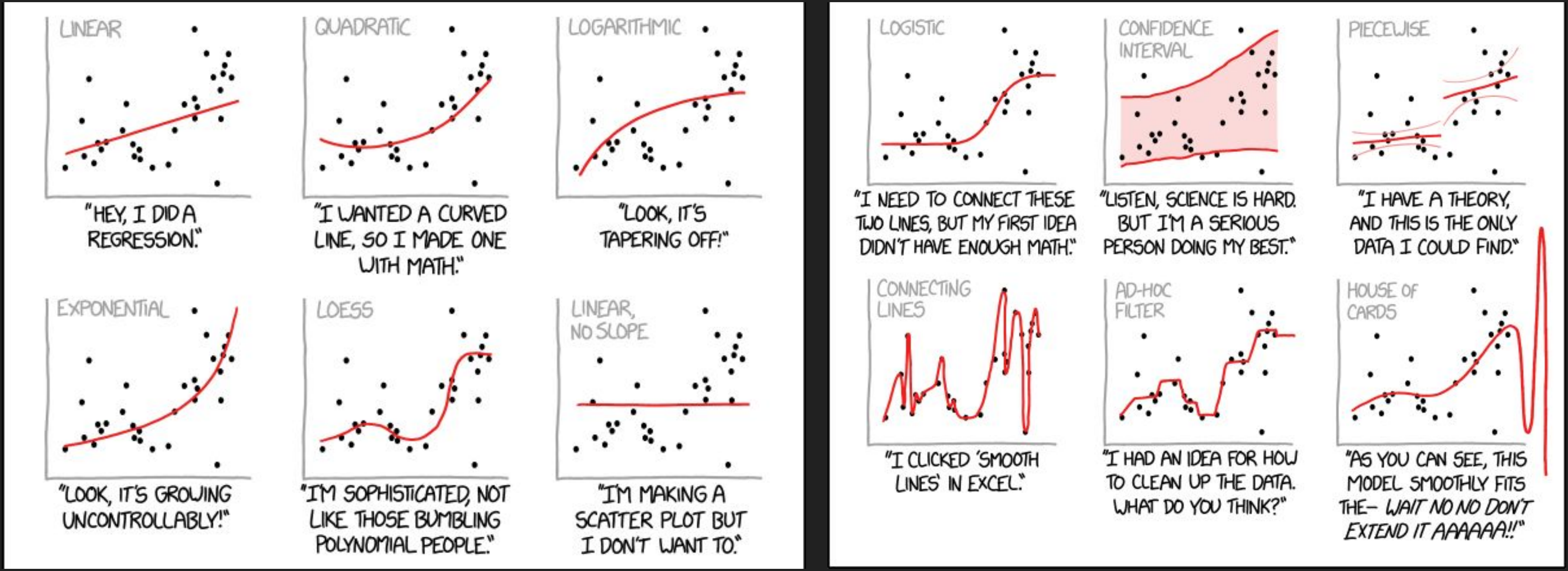
RALPH L. BINGHAM, *Caesar Kleberg Wildlife Research Institute, Texas A&M University-Kingsville, Kingsville, TX 78363, USA*

ABSTRACT We perceive a need for more complete interpretation of regression models published in the wildlife literature to minimize the appearance of poor models and to maximize the extraction of information from good models. Accordingly, we offer this primer on interpretation of parameters in single- and multi-variable regression models. Using examples from the wildlife literature, we illustrate how to interpret linear zero-intercept, simple linear, semi-log, log-log, and polynomial models based on intercepts, coefficients, and shapes of relationships. We show how intercepts and coefficients have biological and management interpretations. We examine multiple linear regression models and show how to use the signs (+, −) of coefficients to assess the merit and meaning of a derived model. We discuss 3 methods of viewing the output of 3-dimensional models (y, x_1, x_2) in 2-dimensional space (sheet of paper) and illustrate graphical model interpretation with a 4-dimensional logistic regression model. Statistical significance or Akaike best-ness does not prevent the appearance of implausible regression models. We recommend that members of the peer review process be sensitive to full interpretation of regression models to forestall bad models and maximize information retrieval from good models (JOURNAL OF WILDLIFE MANAGEMENT 71(3):684–692; 2007)

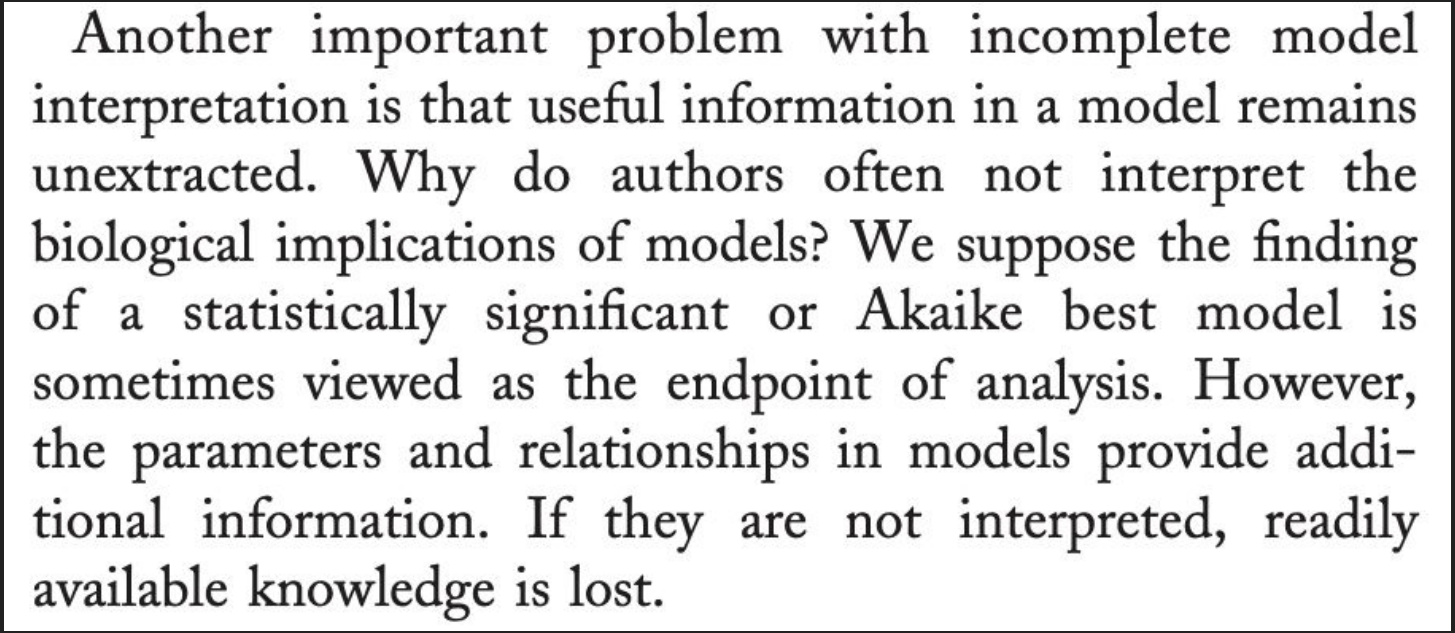
DOI: 10.2193/2006-285

KEY WORDS graphical interpretation, linear model, logistic regression, model interpretation, models, multiple linear regression, polynomials, regression.

CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND

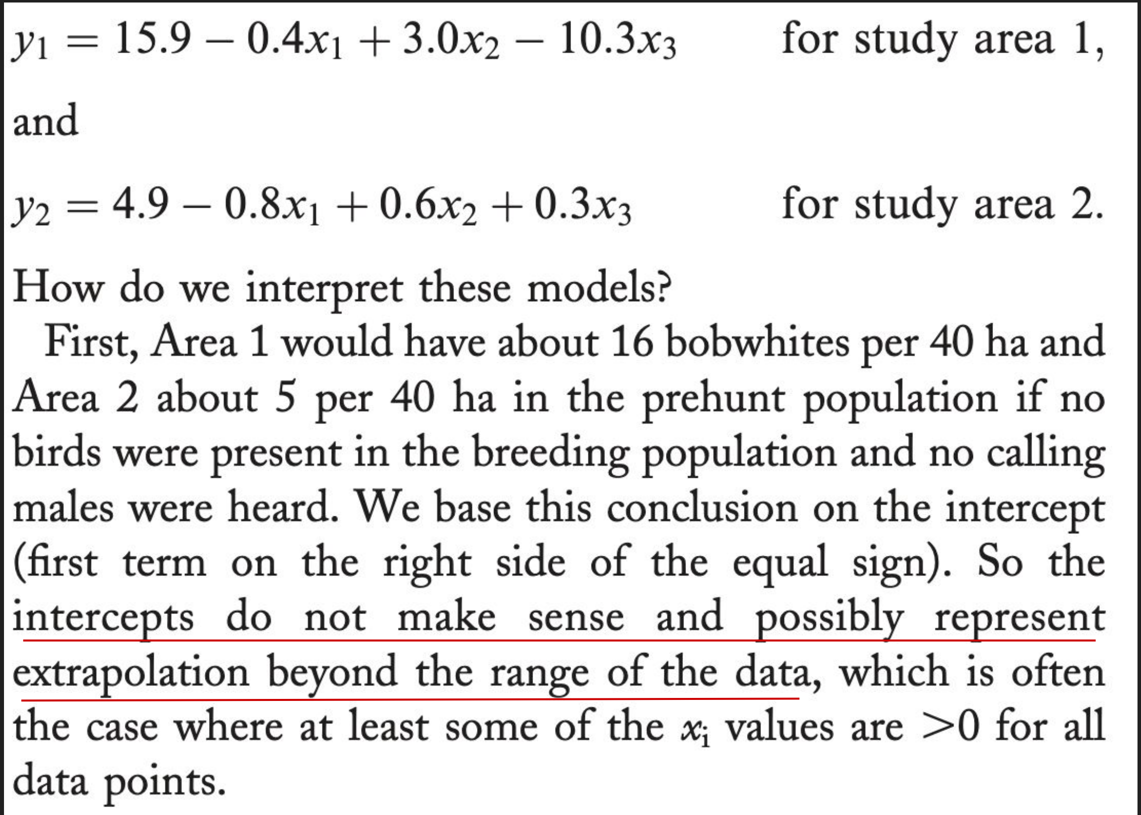

LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

From page 1: Does this sound familiar?



Another important problem with incomplete model interpretation is that useful information in a model remains unextracted. Why do authors often not interpret the biological implications of models? We suppose the finding of a statistically significant or Akaike best model is sometimes viewed as the endpoint of analysis. However, the parameters and relationships in models provide additional information. If they are not interpreted, readily available knowledge is lost.

From page 5: Multiple Linear Regression



$$y_1 = 15.9 - 0.4x_1 + 3.0x_2 - 10.3x_3 \qquad \text{for study area 1,}$$

and

$$y_2 = 4.9 - 0.8x_1 + 0.6x_2 + 0.3x_3 \qquad \text{for study area 2.}$$

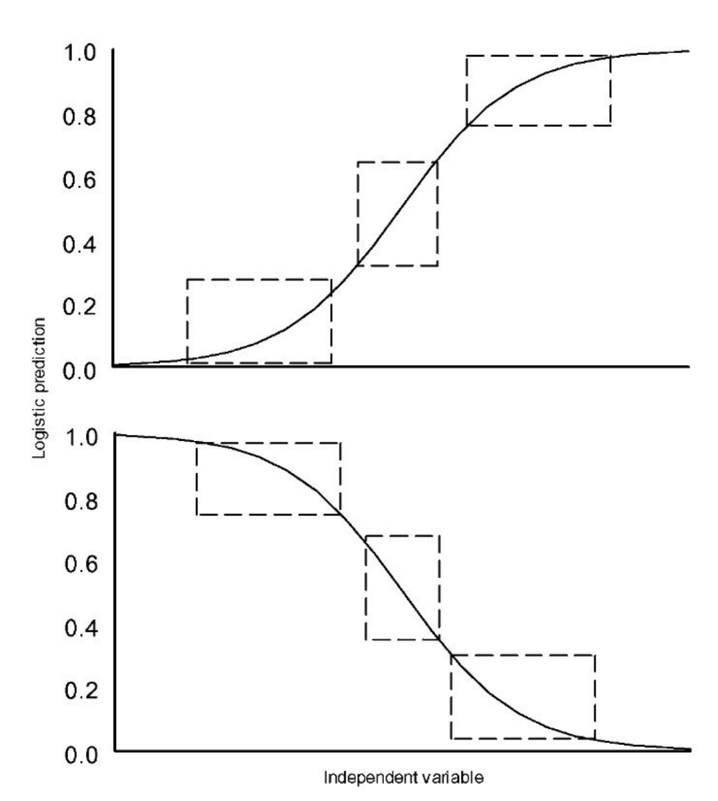
How do we interpret these models?

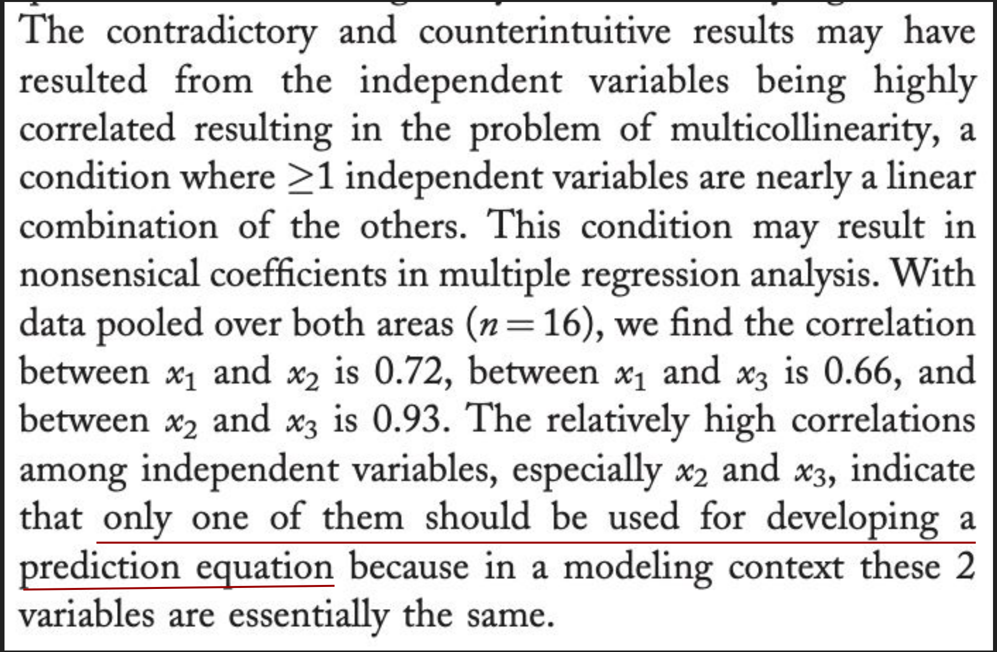
First, Area 1 would have about 16 bobwhites per 40 ha and Area 2 about 5 per 40 ha in the prehunt population if no birds were present in the breeding population and no calling males were heard. We base this conclusion on the intercept (first term on the right side of the equal sign). So the intercepts do not make sense and possibly represent extrapolation beyond the range of the data, which is often the case where at least some of the x_i values are >0 for all data points.

Any other examples where this has happened?

What are a couple of ways that we could fix this problem?

From page 6: logistic regression





The contradictory and counterintuitive results may have resulted from the independent variables being highly correlated resulting in the problem of multicollinearity, a condition where ≥1 independent variables are nearly a linear combination of the others. This condition may result in nonsensical coefficients in multiple regression analysis. With data pooled over both areas ($n = 16$), we find the correlation between x_1 and x_2 is 0.72, between x_1 and x_3 is 0.66, and between x_2 and x_3 is 0.93. The relatively high correlations among independent variables, especially x_2 and x_3, indicate that only one of them should be used for developing a prediction equation because in a modeling context these 2 variables are essentially the same.

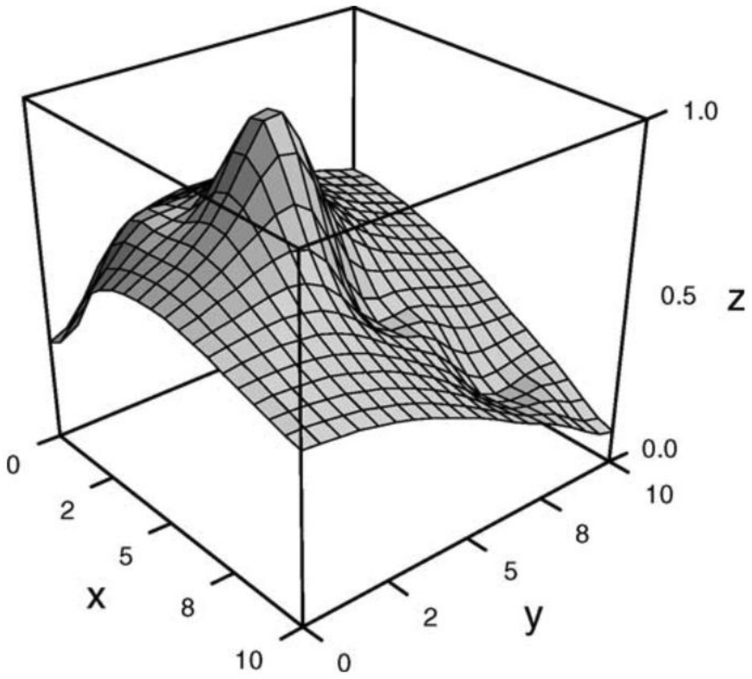
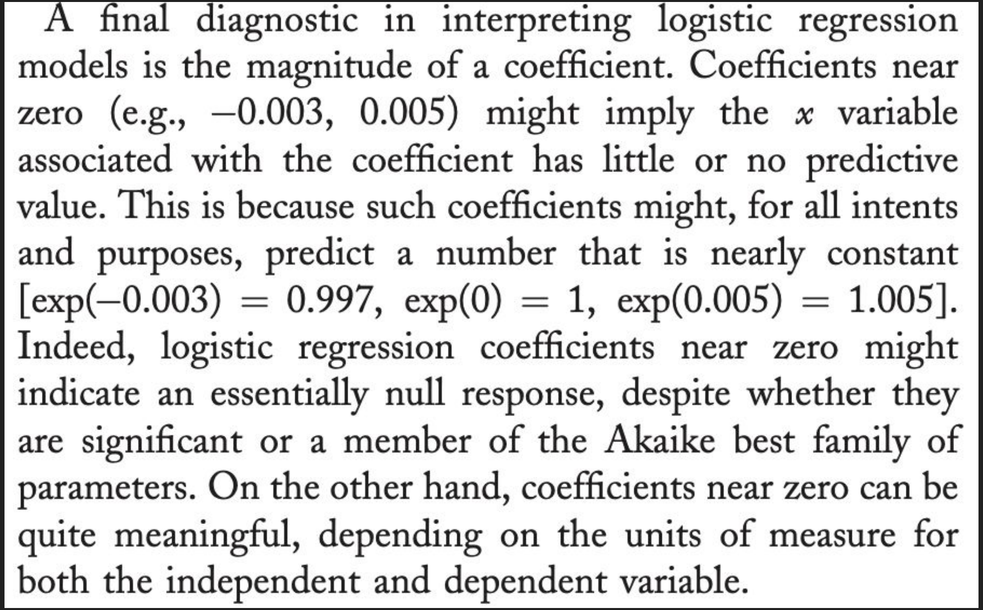


Figure 7. Three-dimensional graphic in 2 dimensions and an example of a response surface.

What are some issues with this type of plot?

What is an alternative and why is it not used here?

From page 8:



A final diagnostic in interpreting logistic regression models is the magnitude of a coefficient. Coefficients near zero (e.g., –0.003, 0.005) might imply the *x* variable associated with the coefficient has little or no predictive value. This is because such coefficients might, for all intents and purposes, predict a number that is nearly constant [exp(–0.003) = 0.997, exp(0) = 1, exp(0.005) = 1.005]. Indeed, logistic regression coefficients near zero might indicate an essentially null response, despite whether they are significant or a member of the Akaike best family of parameters. On the other hand, coefficients near zero can be quite meaningful, depending on the units of measure for both the independent and dependent variable.

Significant ⇎ Meaningful

Take care with the scale of predictors

(0.000001, 0.00004)

Any other thoughts or comments?

$$Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \sigma^2)$$

How to explain this model?

$$Y_i \sim$$

$$N(\beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i1} x_{i2}, \sigma^2)$$

How to explain this model?

$$Y_i \sim N(\beta_0 + \beta_1\, t_i + \beta_2\, t_i^2, \sigma^2\,)$$

How to explain this model?

$$Y_i \sim N(\beta_0 + \beta_1\ t_i + \beta_2\ t_i^2, \sigma^2\ )$$

How to explain this model?