

## MA5771: Applied Generalized Linear Models

### Week 5 Homework

**Homework Problem 5.1 (25 points)** A Chromosome aberration assays are used to determine whether or not a substance induces structural changes in chromosomes. One study compared the results of two substances at various doses. A large number of cells were sampled at each dose to see how many were aberrant.

Substance	Dose (in mg/gl)	No. cell Samples	No. Cells Aberrant	Substance	Dose (in mg/gl)	No. cell Samples	No. Cells Aberrant
A	0	400	3	B	0.0	400	5
A	20	200	5	B	62.5	200	2
A	100	200	14	B	125.0	200	2
A	200	200	4	B	250.0	200	4
				B	500.0	200	7

- (1) Fit a binomial GLM to determine if there is evidence of a difference between the two substances.
- (2) Use the dose and the logarithm of dose as an explanatory variable in separate GLMs, and compare. Which is better, and why?
- (3) Compute the 95% confidence interval for the dose regression parameter, and interpret.
- (4) Why would estimation of the ED50 be inappropriate?

**Homework Problem 5.2 (30 points)** A study of the habitats of the noisy miner (a small but aggressive native Australian bird; data set: `nminer`) recorded whether noisy miners were present in various two hectare transects in buloke woodland patches (Miners), and considered the following potential explanatory variables: the number of eucalypt trees (`Eucs`); the number of buloke trees (`Bulokes`); the area of contiguous remnant patch vegetation in which each site was located (`Area`); whether the area was grazed (`Grazed`: 1 means yes); whether shrubs were present in the transect (`Shrubs`: 1 means yes); and the number of pieces of fallen timber (`Timber`). Fit a suitable logistic regression model for predicting the presence of noisy miners in two hectare transects in buloke woodland patches. You do not need to present the detailed calculations.

- (1) Find  $\hat{\beta}_j$ ,  $se(\hat{\beta}_j)$  and 95% Wald confidence interval of  $\beta_j$ . Interpret your finding about  $\beta_j$ .

- (2) Perform the diagnostic analysis. You should include: plot of the quantile residuals against the fitted values transformed to the constant-information scale; plot of the working responses against the linear predictors; the Q-Q plot of the quantile residuals; plot of the Cook's distance  $D$  and determine if there are any outliers and/or influential observations.

**Homework Problem 5.3 (25 points)** A study of seed germination used two types of seeds and two types of root stocks. Use the proportion of seeds germinating as  $y$  and `Seeds` and `Extract` as the explanatory variables. Fit a logistic regression model using the data `germ`. Then fit a Bernoulli GLM with the logit link using data `germBin` which contains record whether or not each individual seed germinates.

- (1) Show that both the Bernoulli and binomial glms produce the same values for the parameter estimates and standard errors.
- (2) Show that the two models produce different values for the residual deviance, but the same values for the null deviance.
- (3) Show that the two models produce similar results from the sequential likelihood-ratio tests.
- (4) Compare the log-likelihoods for the binomial and Bernoulli distributions. Comment.
- (5) Explain why overdispersion cannot be detected in the Bernoulli model.

**Homework Problem 5.4 (30 points)** The times to death (in weeks) of two groups of leukemia patients whose white blood cell counts were measured (data set: `leukwbc`) were grouped according to a morphological variable called the `ag` factor.

- (1) Plot the survival time against white blood cell count (WBC), distinguishing between `ag`-positive and `ag`-negative patients. Comment on the relationship between WBC and survival time, and the `ag` factor.
- (2) Plot the survival time against  $\log_{10}(WBC)$ , and argue that using  $\log_{10}(WBC)$  is likely to be a better choice as an explanatory variable.
- (3) Fit a GLM(Gamma; log) model to the data, including the interaction term between the `ag` factor and  $\log_{10}(WBC)$ , and show that the interaction term is not necessary.
- (4) Refit the GLM without the interaction term, and evaluate the model using diagnostic tools.
- (5) Plot the fitted lines for each `ag`-factor on a plot of the observations.
- (6) The original source uses an exponential distribution, which is a gamma distribution with  $\phi = 1$ . Does this seem reasonable?

**Homework Problem 5.5 (40 points)** An experiment to investigate the initial rate of benzene oxidation over a vanadium oxide catalyst used three different reaction temperatures and varied oxygen and benzene concentrations. A subset of the data is presented in the data set `rrates` for a benzene concentration near  $2 \times 10^{-3}$  gmoles/L.

- (1) Plot the reaction rate against oxygen concentration, distinguishing different temperatures. What important features of the data are obvious?
- (2) Compare the previous plot to the following plots. Suggest two functional relationships between oxygen concentration and reaction rate that could be compared.
- (3) Fit an inverse Gaussian GLMs identified above, and separately plot the fitted systematic components on the data. Select a model, explaining your choice.
- (4) For your chosen model, perform a diagnostic analysis, identifying potential problems with the model.
- (5) By looking at the data for each temperature separately, is it reasonable to assume the dispersion parameter  $\phi$  is approximately constant? Explain.

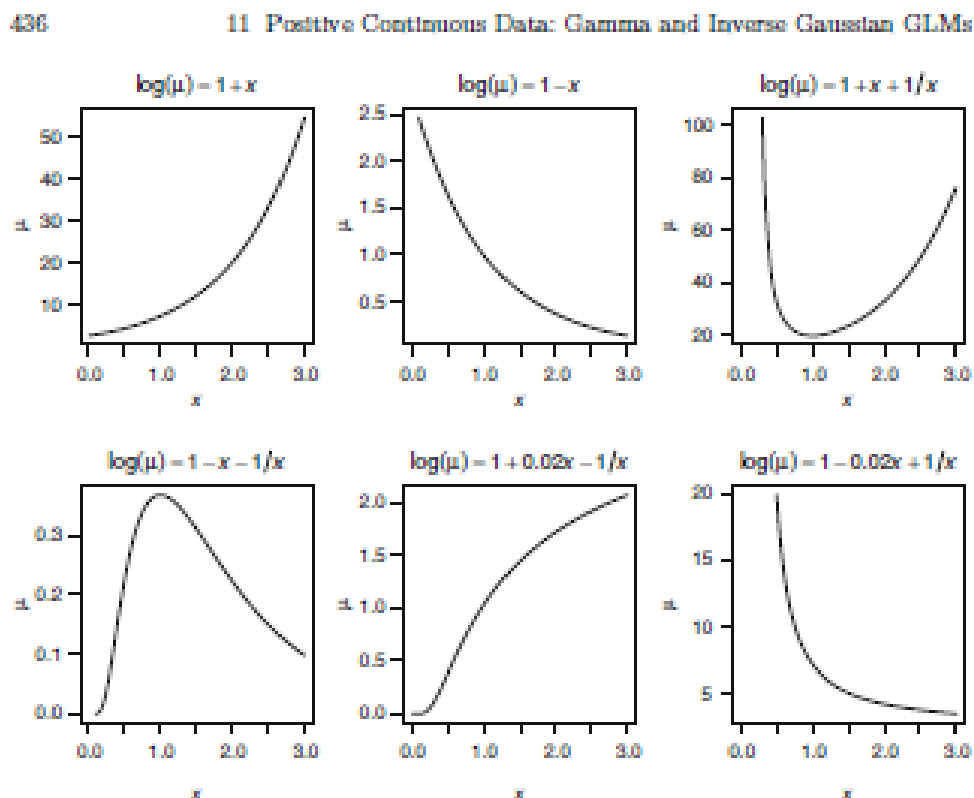


Figure: Various logarithmic link function relationships.

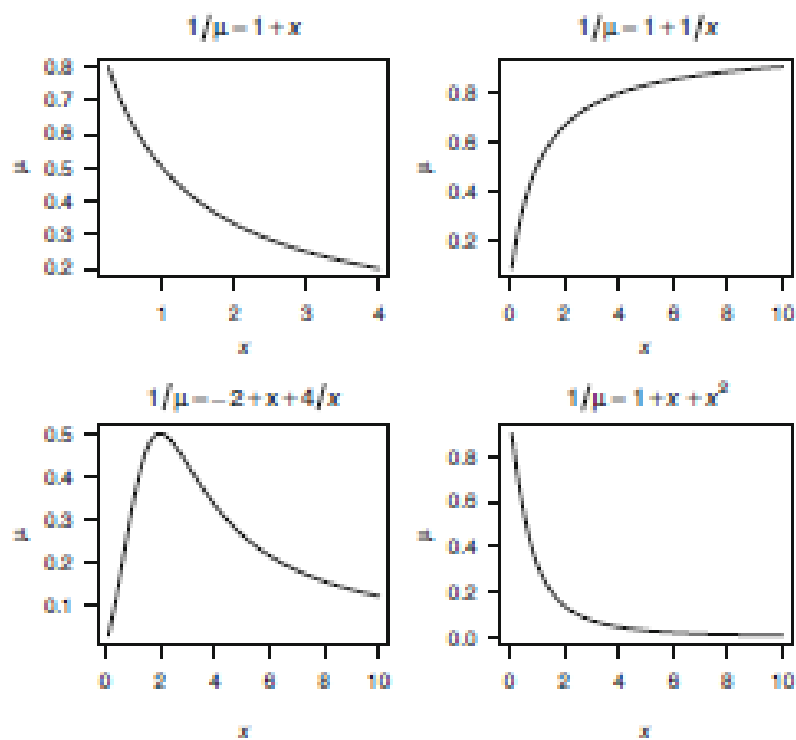


Figure: Various inverse link function relationships.