
Statistics 572 Semester Review

Final Exam Information: The final exam is Friday, May 16, 10:05 - 12:05, in Social Science 6104. The format will be 8 True/False and explains questions (3 pts. each/ 24 pts. total), and three data applications in the areas of classical regression, generalized linear models, and multilevel models (25 pts. each/ 75 pts. total), with 1 point for free. You may the textbook and any notes from the semester (your or mine). You will need a calculator, but cannot use a laptop computer.

During this semester we have studied several types of models:

1. simple and multiple regression with quantitative and categorical explanatory variables;
2. logistic regression;
3. Poisson regression;
4. multilevel models;

You should know when each type of model is appropriate. In particular,

standard regression assumes that the outcome variable is *normally distributed* with a mean that is a linear combination of the predictors; sometimes transformation of the outcome variable or inputs is needed for the model to fit;

logistic regression assumes that each outcome is a *zero or one*; the probability of success is modeled as the *inverse logit function*, $1/(1 + \exp(-x))$, of a linear combination of the predictors;

Poisson regression assumes that the outcome variable has a *Poisson distribution* which models counts without a fixed maximum; the mean of the distribution is the exponential of a linear combination of predictors.

Multilevel models are useful when there are grouping variables and there is a desire to partially pool information within the group and outside of the group. Examples include multiple observations on single individual subjects (repeated measures), or grouping variables such as sites, blocks, or plots.

Multilevel generalized linear models You can also have multilevel generalized linear models where the response is binomial or Poisson, but there are explanatory grouping variables.

Both logistic regression and Poisson regression are examples of *generalized linear models*. In both cases, they can be extended (with family quasibinomial or quasipoisson) to account for *overdispersion*.

You should be prepared for the following:

1. from a description of data and a biological scenario, be able to identify which type of model is appropriate;
 2. given R output for each type of model, be able to make a numerical prediction of an outcome based on new input variable values;
 3. be able to interpret the meaning of estimated coefficients;
-

4. be able to make confidence statements;
5. understand how to interpret simulation output for statistical inference;
6. understand how to interpret simulation output to examine goodness of fit.

There are also a number of concepts that we have discussed that you should be clear about.

- **Simple linear regression model assumptions:** (1) linear relationship between response and explanatory variables; (2) errors are independent of explanatory variables and each other; (3) constant variance; and (4) errors are normally distributed. Know how to identify possible violations, for example from residual plots.
- **Alternative viewpoint of simple linear regression:** Understand that if the input x is z standard deviations above its mean, then the predicted y is rz standard deviations above its mean where r is the correlation coefficient between x and y .
- **Correlation:** The correlation coefficient r measures the strength of the linear relationship between x and y on a scale between -1 and 1 . Strong nonlinear relationships can have any correlation coefficient strictly in the range. The sign of the correlation coefficient is the same as the direction of the association. For a perfectly linear relationship between x and y , r is either -1 or 1 exactly.
- **Transformations:**
 - In regression problems, transformations of the response or explanatory variables can make the linear model fit the data better.
 - *Linear transformations of the explanatory variables* do not change the *goodness of fit of the model*, but do change the *interpretation of regression coefficients*. Common linear transformations include *centering* by subtracting a mean or some other central value and *standardizing* which subtracts the mean and divides by the standard deviation (or some other statistic related to variability).
 - Log transformations do change the model and are often useful for addressing violations of the linear model assumptions.
- **Model matrix representation for multiple regression:** You are not responsible for the matrix algebra, but you should understand that a multiple regression model depends on a matrix where: (1) There is a column of ones for the intercept. (2) Each continuous variable is represented by a single column. (3) Each factor with k levels is represented by $k - 1$ columns. There is no unique way to do this, but the typical parameterization is for each column to be filled with 0s and 1s where one level is treated as a reference and each other level is associated with a single column that indicates if the observation is in that level. (4) Interactions between quantitative variables add a single column that contains the product of the two variables. (5) Interactions between a quantitative variable and a factor with k levels add $k - 1$ columns to the matrix, each of which is the product of the column for the quantitative variable and one of the $k - 1$ columns for the factor. (6) In general, interactions between any number of variables add columns for each product of columns associated with the main variables.

Know that there is a single coefficient β for each column in the model matrix.

- **Least squares versus maximum likelihood for model criteria:** Simple and multiple regression models find the parameters that *minimize the sum of squared residuals*. This least squares criteria is equivalent to the *maximum likelihood criteria* which estimates parameters to make the likelihood, or probability of the observed data, as large as possible. This equivalence follows from the normal distribution. Generalized linear models based on other distributions (like binomial for logistic regression or the Poisson) estimate parameters by maximum likelihood which is different from least squares.
- **Overdispersion:**
 - Both generalized linear models we have seen can be generalized with *overdispersion*.
 - These models allow for more variability than the binomial or Poisson model would predict.
 - These models can be understood as a two-stage hierarchical model where the parameter (p for binomial, μ for Poisson) is drawn first from some distribution and the response y depends on this parameter.
- **Simulation for Prediction and Estimation:**
 - Simulation can be used to find appropriate prediction and confidence intervals which account for uncertainty in parameter estimates.
 - The standard errors in the summary of a fitted model are only appropriate for some comparisons of potential interest.
 - *Predictions* are inferences about *new outcomes*.
 - *Estimation* is inference about *an average or expected value*.
 - For models fit with `lm()` and `glm()` in R, the logic and mechanics of making prediction intervals is the following:
 1. Fit the model to find *maximum likelihood* parameter values (or, perhaps, parameter values that are optimal under some different criteria).
 2. Use `sim()` to simulate a sample of parameter values that are *probable versions of the true parameter values, given the data*.
 3. For the given inputs (explanatory variables), find the true expected value of the new outcome for each sampled version of truth.
 4. For each expected value, simulate a single observation including individual random variation using the appropriate probability distribution for the outcome (here, normal, Poisson, or binomial). Examples:
 - * **normal:**
 - > # N is the simulation size
 - > # N = nrow(sim())\$beta = length(sim())\$sigma
 - > # mu is a vector of probable true expected values
 - > # computed from sim())\$beta
 - > # sigma is a vector of probable true standard deviations
 - > # computed from from sim())\$sigma
 - > new.y = rnorm(N,mu,sigma)

* **Poisson:**

```
> # N is the simulation size
> # N = nrow(sim())$beta = length(sim())$sigma
> # mu = exp(eta) is a vector of probable expected values
> #      computed from sim()$beta
> new.y = rpois(N,mu)
```

* **Binomial (logistic):**

```
> # N is the simulation size
> # N = nrow(sim())$beta = length(sim())$sigma
> # n are the observed sample sizes
> #      (possibly one for standard logistic regression)
> # p = invlogit(eta) is a vector of probable success probabilities
> #      computed from sim()$beta
> new.y = rbinom(N,n,p)
```

5. Use `quantile()` to find the prediction interval.

- For a *confidence interval*, we do not need the individual level variation.
- Simply find quantiles from the appropriate expected value (μ for a normal or Poisson model and p for a binomial/logistic regression model).

• **Fake data simulation for goodness of fit**

- When we use simulation for *prediction or estimation with confidence*, we are assuming that the model is adequate and we are assessing uncertainty *in the context of the model*.
- To test goodness of fit, we want to compare *simulated data* from a best-fitting model with *the real data* to see if the simulated data has characteristics we see in the real data.
- An important example is the proportion of zeros predicted by a Poisson model.

• **Causal Inference**

- Causal inference requires stricter assumptions than predictive inference.
- Randomized experiments are better than observational studies for justified interpretations of causal inference.
- Important variables not included in a model (lurking variables) can lead to misleading causal inferences.
- Typically, justifiable causal inference depends on background scientific understanding and not statistics alone.

• **Multilevel Models**

- Multilevel models include *multiple sources of variation*.
- Multilevel models include *grouping variables* that cluster multiple outcomes together, and the effects of these grouping variables are modeled *as the unobserved outcomes from some random distribution* instead of as *unrelated and unknown fixed values*.

- Multilevel models effectively *partially pool* information from within and outside the group.
- Multilevel models are most effective when the number of groups is fairly large, but even the effects of small numbers of groups can be modeled with a distribution without much statistical loss.
- Fitting large numbers of fixed and unrelated grouping effects can lead to poor statistical estimation.
- Simulation-based inference for multilevel models is similar to that for classical and generalized linear models.
- The R syntax is different.
- Use `mcmcscamp()` instead of `sim()`.
- The output of `mcmcscamp()` is a single matrix that combines β parameters for the mean with transformations of σ parameters for the standard deviations.

Exam format:

- The exam will be worth 100 points.
 - There will be eight True/False and explain questions worth 3 points each (24 points total).
 - These questions will each be designed to test your understanding of a single concept discussed above.
 - If the statement is true, there is no need to add explanation (but if the correct answer is false, explanation can be worth partial credit if it shows understanding).
 - If the answer is false, you should *briefly* explain why.
 - There will be three data analysis questions worth 25 point each (75 points total). Each of these data analysis questions will have several parts.
 - Each problem will have a biological description of data similar to those in your homeworks.
 - Each problem will include summaries of a data analysis including possibly R output of numerical coefficient estimates as well as graphs of data.
 - Many parts will involve doing calculations associated with the model, such as making predictions or confidence intervals. Some parts will ask you to interpret model summaries or summaries of simulation output in the context of the biological problem.
-