
Review 2

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;

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.

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

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.

Exam format:

- The exam will be worth 100 points.
- There will be eight True/False and explain questions worth 4 points each (32 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 *short answer questions* worth 8 points each (18 points total).
 - The answer to each of these questions should be *brief*, consisting of no more than a few sentences.
 - These questions will test conceptual understanding of major topics seen in the course.
- There will be two data analysis questions worth 25 point each (50 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 model, such as making predictions. Some parts will ask you to interpret the coefficients or your prediction calculations in the context of the biological problem.

Sample Problems

1. Circle TRUE or FALSE. If FALSE briefly explain why.

After fitting a logistic regression model, a plot of residuals versus fitted values is useful for seeing if model assumptions are violated.

2. Circle TRUE or FALSE. If FALSE briefly explain why.

In a multiple regression problem, an quantitative input variable x is replaced by $x - \text{mean}(x)$. The R^2 statistic for the fitted model will be the same.

3. Circle TRUE or FALSE. If FALSE briefly explain why.

In a multiple regression problem, an quantitative input variable x is replaced by $x - \text{mean}(x)$. The coefficient β associated with x will have the same numerical value after the transformation that it had before.

4. **SHORT ANSWER:** In multiple regression model that predicts the weight of a dairy cow one year after birth with inputs (1) the weight of its mother in kg, and (2) a factor with k diets, state what the intercept term measures in the model, *briefly* explain why the model should include the intercept term, even though it has no useful biological interpretation.
-

5. PREDICTION:

A regression model predicts the urine urea nitrogen (UUN) concentration (mg/dL) of Holstein dairy cattle on the basis of the the weight of the cow and diet, which is one of four treatments, A, B, C, and D.

```
> display(fit1, digits = 3)

lm(formula = UUN ~ weight + diet)
      coef.est coef.se
(Intercept)  -2.652  226.073
weight         0.192   0.150
dietB         178.087  40.553
dietC         426.424  42.661
dietD         627.502  42.401
---
n = 20, k = 5
residual sd = 63.937, R-Squared = 0.95
```

- Predict the UUN concentration for a 1600 pound cow in each of the four treatment groups.
- A second model includes an interaction between weight and diet.

```
> display(fit2, digits = 3)

lm(formula = UUN ~ weight * diet)
      coef.est coef.se
(Intercept)   212.943  388.186
weight         0.048   0.259
dietB        -434.496  607.750
dietC        -278.767  754.457
dietD         742.005  583.112
weight:dietB   0.406   0.402
weight:dietC   0.453   0.483
weight:dietD  -0.090   0.402
---
n = 20, k = 8
residual sd = 65.746, R-Squared = 0.96
```

Predict the UUN concentration for a 1600 pound cow in each of the four treatment groups.

- For each model, you could graph four lines showing the predicted UUN versus weight with separate lines for each treatment group. Briefly describe how to distinguish between the graphs for each model.
- For each model, what change in UUN concentration is predicted from an increase of 100 pounds in a cow in treatment group B?

Sample Problem Solutions

1. FALSE. A plot of residuals versus fitted values in logistic regression will show two curves, one for the 0 outcomes and one for the 1 outcomes. This plot will not be helpful in assessing the quality of fit of the model.
2. TRUE. Linear transformations do not affect goodness of fit of linear models.
3. FALSE. Linear transformations do affect the values of coefficients.
4. The intercept is the predicted weight of a dairy cow whose mother weighed 0 kg when given the first diet. There is no biological relevance to this interpretation as the mother will not have weighed 0 kg. However, the model should include an intercept so that the fitted line is not constrained to pass through the origin. We want to fit the best possible line within the range of the observed data.
5. (a) For the first model, the predicted UUN values are: 305 for A, 483 for B, 731 for C, and 932 for D. These are found from $-2.652 + 1600 * (0.192) + x$ where $x = 0, 178, 426.4, 627.5$ for the four treatment groups.

$$A \quad 212.943 + 0.048 * 1600 = 290$$

(b) The predictions are:

$$B \quad 212.943 - 434.5 + 0.048 * 1600 + 0.406 * 1600 = 505$$

$$C \quad 212.943 - 278.8 + 0.048 * 1600 + 0.453 * 1600 = 736$$

$$D \quad 212.943 + 742 + 0.048 * 1600 - 0.090 * 1600 = 888$$

(c) The model in (a) will show four parallel lines. The model in part (b) will show four lines, but they will not be parallel.

(d) In model (a), the slope is 0.192 mg/dL per pound, so the predicted increase in concentration is 19.2 mg/dL.

In model (b), the slope is $0.048 + 0.406 = 0.454$ so the predicted increase in UUN concentration would be 45.4 mg/dL.
