

Statistics 572 Midterm 1 Solutions

Target, Spring 2008

True/False and Explain, Problems 1–8, 4 points each

For each statement, circle TRUE or FALSE. If TRUE, you may add a brief clarification, but a correct answer without explanation will receive full credit. (The clarification might be worth some partial credit if the correct response was FALSE.) If FALSE, provide a very brief explanation.

1. Circle TRUE or FALSE.

In a simple linear regression model with explanatory variable x and outcome variable y , we have these summary statistics for sample means and standard deviations: $\bar{x} = 10$, $s_x = 3$, $\bar{y} = 20$, and $s_y = 5$.

For a new data point with $x = 13$, it is possible that the predicted value $\hat{y} = 26$.

Solution: False. The new x value is $z = (13 - 10)/3 = 1$ standard deviation above the mean. The predicted \hat{y} is $(26 - 20)/5 = 1.2$ standard deviations above its mean. This is supposed to be rz standard deviations which implies that $r = 1.2$; however, this is impossible as $-1 \leq r \leq 1$.

2. Circle TRUE or FALSE.

A standard multiple regression model with quantitative predictors x_1 and x_2 , a factor predictor T with four levels, an interaction between x_1 and T , and an intercept has for its model coefficients an 11×1 vector β .

Solution: False. The number of coefficients in this model is $2 + 3 + 3 + 1 = 9$, not 11. There are 2 coefficients for the quantitative variables, 3 for the factor with four levels, 1×3 for the interaction between the quantitative variable and the factor, and one for the intercept.

3. Circle TRUE or FALSE.

In a standard multiple regression model, if a plot of residuals versus fitted values shows a fan-shaped pattern with residuals becoming more spread out as fitted values increase, a log transformation of the response variable may result in data more consistent with model assumptions.

Solution: True. You would need to check that the outcome variable was positive before using the log transformation.

4. Circle TRUE or FALSE.

If the outcome variable is quantitative and all explanatory variables take values 0 or 1, a logistic regression model is most appropriate.

Solution: False. Logistic regression is only appropriate when the *outcome variable* is 0/1.

5. Circle TRUE or FALSE.

In a greenhouse experiment with several predictors, the response variable is the number of seeds that germinate out of 60 planted with each treatment combination. A Poisson regression model is most appropriate for this data.

Solution: False. Each outcome is a count from 0 to 60 that can be treated as a number of independent trials with some success probability. A binomial distribution is appropriate, so this is better modeled by logistic regression.

6. Circle TRUE or FALSE.

In a greenhouse experiment with several predictors, the response variable is the number of seeds produced for each plant with a sample size of 60 plants. A Poisson regression model is most appropriate for this data.

Solution: True. We have count data with no fixed range. A Poisson regression is the place to start.

7. Circle TRUE or FALSE.

The same data is fit with two models using exactly the same predictors. The first model uses standard logistic regression (with `glm(...,family=binomial)`) while the second model accounts for overdispersion (with `glm(...,family=quasibinomial)`). The estimated coefficients for the predictors in the two models will be identical.

Solution: True. Adding an overdispersion parameter to a logistic or Poisson regression simply inflates the standard errors.

8. Circle TRUE or FALSE.

In a fitted quasi-Poisson regression model, the overdispersion parameter is estimated to be 4.0. This means that the residuals have a standard deviation that is about 4 times larger than a standard Poisson regression model would predict.

Solution: False. The variance is four times larger, so that standard deviation is only two times as large.

Short Answer, Problems 9–11, 6 points each

Provide *very brief* answers to the questions. Correct responses can be one or a few sentences long.

9. Given that the fitted values and goodness-of-fit measures do not change when transforming predictors by centering (say, subtracting the mean), what is the benefit of making such a transformation?

Solution: This transformation may make the coefficients easier to interpret in a biological context.

10. After fitting a multiple regression model, explain how you might detect that the linearity assumption of the model, $E[y_i] = X_i\beta = \beta_1 + \beta_2x_2 + \dots + \beta_kx_k$ is violated.

Solution: A plot of residuals versus fitted values may show a pattern.

11. The generalized linear models we have seen, logistic regression and Poisson regression, have variants that include an overdispersion parameter. Why do we not need to model overdispersion in a standard multiple regression model with a normally distributed response?

Solution: The normal distribution has two parameters that can be estimated separately, the mean and the variance. In contrast, for the Poisson distribution or the binomial distribution (with n known), there is only one parameter. When you use the data to estimate the mean, you automatically get a variance estimate as well, which may not fit the data.

Data Analysis, Problems 12–13, 25 points each

12. (a) (12 points) Each model describes the relationship between UUN and CP with a line for each lactation group. Fill in the table indicating the slope and intercepts for each model for each group.

Solution:

Lactation	Slope	Model 1	Slope	Model 2
		Intercept		Intercept
Early	33.8	-419.3	34.2	-427.6
Mid	33.8	$-419.3 + 4.6 = -414.7$	$34.2 + 4.8 = 39.0$	$-427.6 - 80.9 = -508.5$
Late	33.8	$-419.3 - 11.6 = -430.9$	$34.2 - 6.1 = 28.1$	$-427.6 + 98.9 = -328.7$

- (b) (4 points) For each model, predict the UUN for a cow in the mid lactation stage with a diet CP of 20 percent.

Solution: Model 1: $-414.7 + 20(33.8) = 261.3$; Model 2: $-508.5 + 20(39.0) = 271.5$

- (c) (4 points) For Model 1, provide an interpretation of each estimated coefficient.

Solution:

- i. The intercept -419.3 is the predicted UUN for an early lactation cow with 0 percent crude protein in the diet. This is not a relevant biological interpretation.
- ii. The slope 33.8 is the predicted increase in grams per day of UUN for every increase of one percent crude protein in the diet for cows at any stage of lactation.
- iii. The coefficient 4.6 is the amount in grams per day that the UUN of cows in mid lactation is predicted to be higher than cows in early lactation when given the same crude protein content in the diet.

- iv. Cows in late lactation are expected to have *UUN* measurements in grams per day that are 11.6 lower than those in early lactation cows when given the same crude protein content in the diet.
- (d) (5 points) The R^2 statistics for both models are high, over 0.9. Is this alone sufficient evidence to assess that the linear models are adequate fits to the data? Briefly explain.

Solution: The R^2 statistic alone is insufficient to assess if a linear model fits well. A plot of residuals versus fitted values may show that a nonlinear relationship could fit the data better.

13. (a) (5 points) Briefly explain why logistic regression is an appropriate model for this data.

Solution: The outcome variable is 0/1.

- (b) (12 points) Calculate and record the probability of spawning for various mating trial classes under each model indicated in the table. In each case, show sufficient work below the table to indicate how you arrived at each numerical answer.

Solution: There are three parent types and two choices for the lake factor for $3 \times 2 = 6$ possible experimental conditions. Each will have its own probability of spawning calculated for each model. The problem asks us to compute three of these six probabilities for each model.

The inverse logit function is

$$\text{logit}^{-1}(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

<i>Types Lakes</i>	Class 1 <i>LL Same</i>	Class 2 <i>LL Different</i>	Class 3 <i>BL Same</i>
Model 1	$\text{logit}^{-1}(-0.36 - 0.02 + 0.13)$ = 0.438	$\text{logit}^{-1}(-0.36 + 0.13)$ = 0.443	$\text{logit}^{-1}(-0.36 + 1.15 - 0.02)$ = 0.178
Model 2	$\text{logit}^{-1}(-0.38 + 0.02 + 0.22)$ = 0.465	$\text{logit}^{-1}(-0.38)$ = 0.406	$\text{logit}^{-1}(-0.38 + 0.02 - 0.94 - 0.51)$ = 0.141

- (c) (4 points) Using Model 1, compare the spawning probabilities between Class 1 and Class 2 and also between Class 1 and Class 3. Does it appear that the types of fish or the lakes of origin is a more important predictor of spawning probability?

Solution: The two Limnetic fish spawn with nearly equal probability (0.438 versus 0.443) whether or not they are from the same lake. However, Limnetic and Benthic from the same lake spawn much less frequently (0.178) than two Limnetics from the same lake (0.438).

It appears that types of the fish is a more important factor than the lake origins of the fish in predicting spawning probability.

- (d) (4 points) If you expected that the probability of spawning for Benthic/Benthic trials would be about the same whether the fish were from the same lake or not, but thought that the probability of spawning might differ for Limnetic/Limnetic trials depending on whether or not they were from the same lake, would it be better to use Model 1 or Model 2? Briefly explain.

Solution: Model 2 would be preferable. If you expect there to be an interaction, it is better to use a model that includes an interaction.