

# Statistics 572 Midterm 2 Solutions

## Problem 1:

**Summary:** The study examines the effect of planting six different species of trees on forest regeneration as measured by the count of stems of woody individuals in the understory at three different sites in Costa Rica. We model the counts with a Poisson regression model using overstory treatment (species), site, the interaction between site and treatment, and spacing as predictors. We find that there is a significant interaction between treatment and site, meaning that the relative effects of treatment differ substantially from site to site. Consequently, any predictions about the effects of planting these species at other sites must be taken as speculative. We also estimate that a one square meter increase in the space per tree is associated with about a 3 percent decline in the woody stem count in the understory.

Due to the very poor study design (spacing very different at each site, no spacing variation for the same species within a site, and so on), it is very difficult to determine which factors associated with each site are most important for the response.

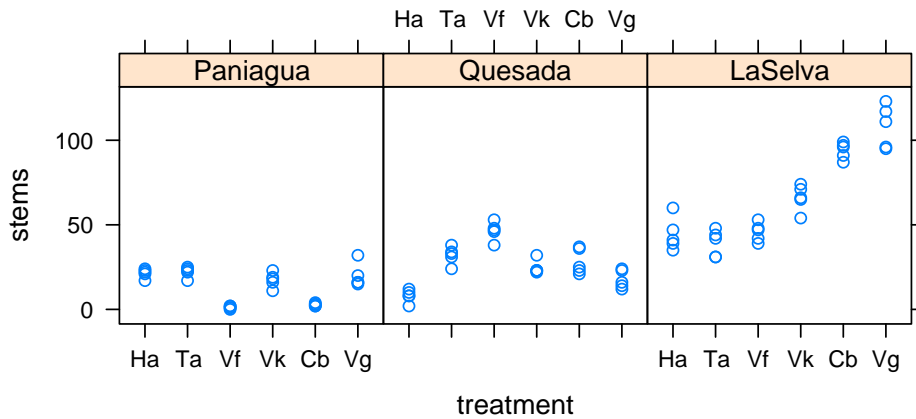
- (a) The primary question of interest is to compare the regeneration measures of the six different species of planted trees. Prepare graphs that compare the regeneration measure for the treatments, indicating the sites as well. Do the different treatments behave similarly at each site?

Solution: To better see the patterns for the response `stems` based on treatment overstory and site, I reordered each of these variables from smallest to largest based on mean number of stems.

```
> regenerate = read.table("regenerate.txt", header = T)
> regenerate$treatment = with(regenerate, reorder(regenerate$treatment,
+ stems))
> regenerate$site = with(regenerate, reorder(regenerate$site,
+ stems))
```

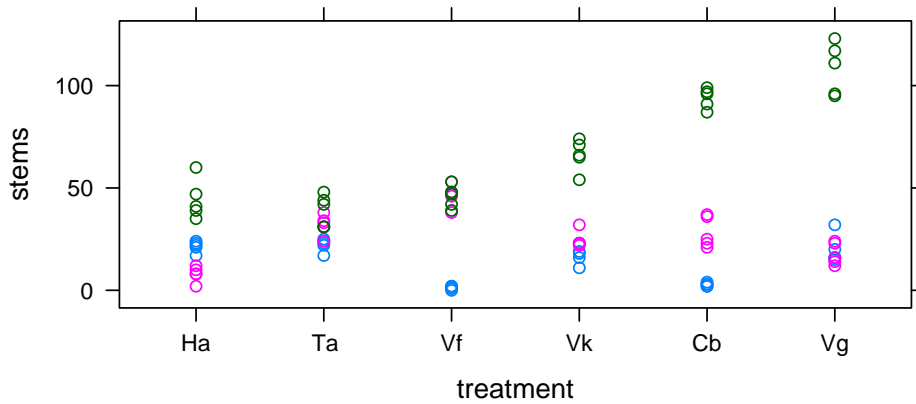
To present the number of stems versus overstory treatment by site, I made the following graph.

```
> print(xyplot(stems ~ treatment | site, data = regenerate))
```



An alternative would be to use a single graph with separate colors for each site.

```
> print(xyplot(stems ~ treatment, groups = site, data = regenerate))
```



I see that there is more regeneration at the former research station La Selva than at the two former private plantations. It also appears that there is variation from site to site as to which treatments do better. The lack of a consistent pattern across sites is indicative of an interaction between site and overstory treatment. There may, however, be other explanatory variables not shown in this graph that can explain some of the visible variation.

- (b) Create a model to predict the woody stem counts in plots using any important inputs. Justify your selection of the model.

Solution: Since the response is a count and many counts are small, a Poisson response is most appropriate. It is acceptable to consider the effects of grouping variables such as site and overstory treatment to be either fixed (regular GLM) or random (a multilevel model).

There is a strong interaction between the site and the treatment. (The change in deviance is over 500 for a model with and without the interaction, the interaction adding 9 parameters

to the model.)

```
> fit.0 = glm(stems ~ treatment + site + spacing, family = "poisson",
+ data = regenerate)
> fit.1b = glm(stems ~ treatment * site + spacing, family = "poisson",
+ data = regenerate)
> anova(fit.0, fit.1b)
```

Analysis of Deviance Table

```
Model 1: stems ~ treatment + site + spacing
Model 2: stems ~ treatment * site + spacing
  Resid. Df Resid. Dev Df Deviance
1      81    619.71
2      72     80.70  9   539.02
```

Some treatments do much better at some sites and much worse at others. A good-fitting model will allow for some kind of interaction between site and treatment.

I decided to keep spacing in the model. There is little reason to keep drainage. There is some justification for keeping canopy (possibly with a quadratic term) and slope.

A model with an overdispersion parameter fits only slightly better (parameter is 1.1). Without a formal test, I kept the simpler Poisson model.

The fitted model I selected has these parameters:

#	Parameter	Estimate	SE
1	(Intercept)	4.03	0.23
2	treatmentTa	0.04	0.14
3	treatmentVf	-2.88	0.42
4	treatmentVk	-0.45	0.16
5	treatmentCb	-2.88	0.34
6	treatmentVg	0.893	0.080
7	siteQuesada	-1.47	0.22
8	siteLaSelva	0.00	0.22
9	spacing	-0.0303	0.0050
10	treatmentTa:siteQuesada	1.35	0.22
11	treatmentVf:siteQuesada	4.88	0.46
12	treatmentVk:siteQuesada	1.82	0.26
13	treatmentCb:siteQuesada	4.15	0.39
14	treatmentVg:siteQuesada	0.51	0.21
15	treatmentTa:siteLaSelva	-0.16	0.17
16	treatmentVf:siteLaSelva	3.15	0.44
17	treatmentVk:siteLaSelva	0.85	0.19
18	treatmentCb:siteLaSelva	3.63	0.36

Several other choices are acceptable, but a Poisson or quasi-Poisson model that includes an interaction between site and treatment seems to fit substantially better than any alternative.

- (c) Use your model to create a confidence interval for the difference in regeneration measures between *Calophyllum brasiliense* (Cb) and *Terminalia amazonia* (Ta) at each site. Explain how you find these intervals.

Solution: For each of the three sites, I report a 95% confidence interval for the difference in the mean number of stems between Cb and Ta. A positive number indicates that the mean is expected to be larger for Cb than for Ta.

To answer the question for my model, I need to assume something for the spacing. These treatments occurred at different spacings for each of these sites: both treatments at spacing 8 at site La Selva, both at 16 at site Quesada, and Cb at 4 and Ta at 32 at site Paniagua. I will use spacing 8 for the comparison at LaSelva, 16 for the comparison at Quesada, and try both 4 and 32 for site Paniagua.

Because of the log link function, a confidence interval for a ratio would have been easier than a confidence interval for a difference as the intercept and any other parameters would have canceled. The simulation-based approach means that we do not need to derive a formula for the SE if we really want the difference and not the ratio.

Confidence intervals are based on the quantiles of a simulation.

The following table shows the expected number of stems for each of the two treatments. As an example, the number for La Selva with treatment Cb and spacing 8 is calculated as

$$\exp(\beta_1 + \beta_5 + \beta_8 + 8\beta_9 + \beta_{18})$$

and the same mean for treatment Ta is

$$\exp(\beta_1 + \beta_2 + \beta_8 + 8\beta_9 + \beta_{15}) .$$

Note that it is critical to take differences *after* the exponential transformation. Note especially that

$$\exp(\beta_1 + \beta_5 + \beta_8 + 8\beta_9 + \beta_{18}) - \exp(\beta_1 + \beta_2 + \beta_8 + 8\beta_9 + \beta_{15}) \neq \exp(\beta_5 + \beta_{18} - \beta_2 - \beta_{15}) .$$

Had the question been about a change in *the ratio of means*, then taking differences first would have been okay since

$$\exp(\beta_1 + \beta_5 + \beta_8 + 8\beta_9 + \beta_{18}) / \exp(\beta_1 + \beta_2 + \beta_8 + 8\beta_9 + \beta_{15}) = \exp(\beta_5 + \beta_{18} - \beta_2 - \beta_{15}) .$$

Site	Spacing	Cb	Ta	Difference	95% Interval
La Selva	8	94.0	39.2	54.8	(44.7, 65.2)
Paniagua	4	2.8	51.9	-49.1	(-70.4, -34.4)
Paniagua	32	1.2	22.2	-21.0	(-25.6, -17.2)
Quesada	16	28.4	32.0	-3.6	(-10.6, 3.2)

- (d) The site of a new former plantation will be planted with one of the six species with the hope of maximizing the forest regeneration. The new site is flat and has good drainage. Which tree would you recommend planting, and at what spacing? You will need to make some additional assumptions to answer the question. (For example, the canopy openness in the future will depend on the growth of the trees planted.) Briefly justify your recommendation.

Solution: The slope and drainage of the site do not appear to make much difference. Apparently, unmeasured characteristics at each site help to determine the relative success of the various choices of species for the overstory. If the new site is similar to one of the existing sites, I would select the species that seemed to do the best at that particular site. The species that resulted in the largest counts overall and did not do relatively poorly at any site was *Vochysia guatemalensis* (Vg). Since the smaller spacing (higher density) seems to be associated with the highest woody stem counts in the understory, I would use a spacing of 8 meters square per tree, the lowest used with Vg in the original data.

A more formal prediction using a fitted model seems speculative given the substantial site/treatment interaction.

## Problem 2:

**Summary:** In an experiment with randomized assignment of ten hamsters into each of two treatment groups, the concentration of an enzyme in the heart was significantly higher in hamsters exposed to 16 hours of daylight per day (long days) than in the group exposed to 8 hours of light each day (short days). The mean concentration for the long day treatment is  $1.632 \pm 0.022$  and the mean concentration for the short day treatment is  $1.377 \pm 0.022$ . An approximate 95% confidence interval for the difference (long minus short) is  $0.255 \pm 0.060$  or (0.195, 0.315). These inferences are based on a multilevel model that with an unmodeled treatment effect and hamster effects that were modeled as normally distributed. The standard deviation of hamster effects was 0.065 and the standard deviation of variation in measurements from the same hamster was 0.041.

- (a) Display a graph that highlights the primary research question.

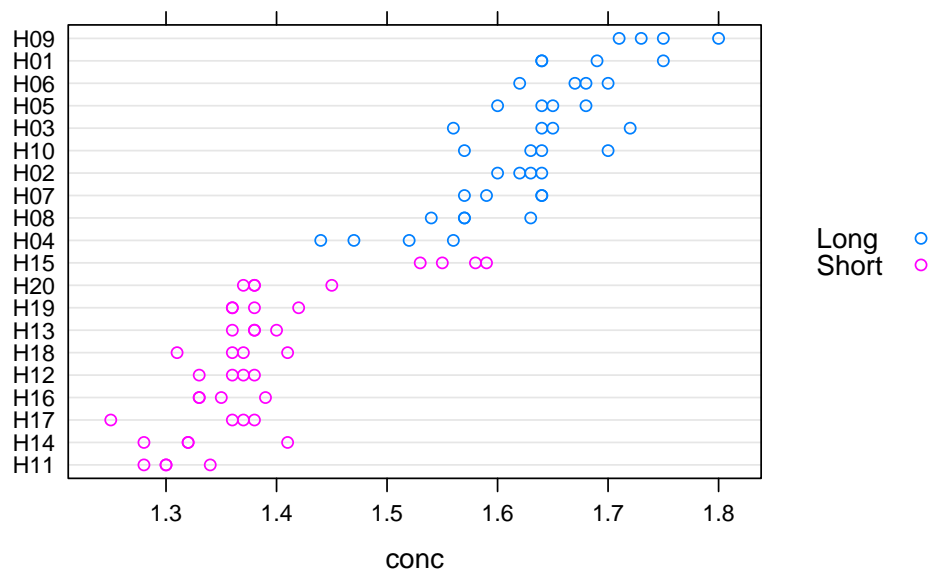
Solution: As the hamster ids are arbitrary, I reordered these levels so that the short day would come before long and within each group, the mean concentration level would increase. This is not a necessary step, but makes it easier to see patterns in the plot.

As an side, before reading in the data I changed the single digit ids from H1 to H01, for example, so that they would be in order alphabetically and numerically.

```
> hamster = read.table("hamster.txt", header = T)
> hamster$id = with(hamster, reorder(hamster$id, 10 * (day ==
+   "Long") + conc))
```

I decided to display the data with a dotplot so that I could easily see the variation of measurements within each hamster as well as between groups.

```
> print(dotplot(id ~ conc, groups = day, data = hamster,
+   auto.key = list(space = "right")))
```



It is readily apparent that the long day group is associated with a much higher average concentration of the heart enzyme. There is substantial variation measurement to measurement within each hamster, but this variation is much smaller than the treatment effect. It looks like the measurement variation within each hamster is of comparable size.

- (b) Fit and interpret a model that addresses the primary research question. Summarize your results.

Solution: I elected to fit a simple multilevel model, modeling the mean measure for each hamster as a normally distributed random effect. Here is a summary of the fitted model. I fit it with and without an intercept to make it easy to see both the comparison between groups and the means of each group.

```
> fit.2b = lmer(conc ~ day + (1 | id), data = hamster)
> display(fit.2b, digits = 3)
```

```
lmer(formula = conc ~ day + (1 | id), data = hamster)
      coef.est coef.se
(Intercept)  1.632   0.022
dayShort     -0.255   0.030
```

Error terms:

```
Groups   Name Std.Dev.
id       0.065
Residual 0.041
```

---

number of obs: 80, groups: id, 20

```
AIC = -221.7, DIC = -251.1
deviance = -239.4
```

```
> fit.2bi = lmer(conc ~ day - 1 + (1 | id), data = hamster)
> display(fit.2bi, digits = 3)
```

```
lmer(formula = conc ~ day - 1 + (1 | id), data = hamster)
      coef.est coef.se
dayLong  1.632    0.022
dayShort 1.377    0.022
```

Error terms:

Groups	Name	Std.Dev.
	id	0.065
	Residual	0.041

---

```
number of obs: 80, groups: id, 20
AIC = -221.7, DIC = -251.1
deviance = -239.4
```

The mean concentration for the long day treatment is  $1.632 \pm 0.022$  and the mean concentration for the short day treatment is  $1.377 \pm 0.022$ . An approximate 95% confidence interval for the difference (long minus short) is  $0.255 \pm 0.060$  or  $(0.195, 0.315)$ . As this is an experiment and hamsters were placed in treatment groups at random, we are justified to conclude that the treatment caused this increase in the concentration of the heart enzyme.

The hamster effects are modeled as  $N(\mu = 0, \sigma^2 = 0.065^2)$  random variables. The standard deviation of measurements within hamsters is 0.041, so much of the variation we see is among hamsters.

- (c) Find prediction intervals for the difference in concentrations of two heart samples from the same hamster in the long day treatment group and from the same hamster in the short day treatment group. Show how you arrived at the intervals.

Solution: A quick solution that ignores uncertainty in the variance components estimates is that the standard error for the difference will be  $\sqrt{2 \times 0.41^2} = 0.58$  so that we predict with 95% confidence that the difference between a pair of measurements from the same hamster will be within about 1.16 of each other.

Alternatively, we can use simulation to make a prediction interval and account for the uncertainty in the estimation of the variance components. The interval is the same for both treatment groups.

```
> set.seed(34323)
> sim.2b = mcmcSamp(fit.2b, 10000)
> sim.2b[1, ]
```

```
(Intercept)      dayShort log(sigma^2) log(id.(In))
      1.6297336    -0.2218506    -6.2322020    -6.0882387
```

```
> sigma = sqrt(exp(sim.2b[, 3]))
> sigmaA = sqrt(exp(sim.2b[, 4]))
> hamEff1 = rnorm(10000, 0, sigmaA)
> meas1 = sim.2b[, 1] + hamEff1 + rnorm(10000, 0, sigma)
> meas2 = sim.2b[, 1] + hamEff1 + rnorm(10000, 0, sigma)
> quantile(meas1 - meas2, c(0.025, 0.975))
```

```
      2.5%      97.5%
-0.1168420  0.1146080
```

The simulation-based estimate is nearly identical.

- (d) Find a prediction interval for the difference in mean concentrations of the four heart samples from two different hamsters, one in the long day and one in the short day treatment group. Show how you arrived at the interval.

Solution: The standard error of this difference is

$$SE = \sqrt{2(\sigma_\alpha^2 + \sigma^2/4)}$$

which is  $\sqrt{2(0.065^2 + 0.041^2/4)} = 0.096$ .

Here, the quick solution is the estimated difference (long minus short) plus and minus twice the SE, or

$$0.255 \pm 2(0.096) \text{ or } (0.063, 0.447)$$

Again, we can account for uncertainty in the parameter and variance component estimates using simulation.

```
> muLong = rnorm(10000, sim.2b[, 1], sigmaA)
> muShort = rnorm(10000, sim.2b[, 1] + sim.2b[, 2], sigmaA)
> xbarLong = rnorm(10000, muLong, sigma/2)
> xbarShort = rnorm(10000, muShort, sigma/2)
> round(quantile(xbarLong - xbarShort, c(0.025, 0.975)),
+      3)
```

```
      2.5% 97.5%
0.047 0.465
```

This interval is a bit longer than the one that fails to account for uncertainty in the estimated difference in mean treatment effects