# STAT 535 HW 5 Example Solutions

Exercise 12.3)

a) Poisson regression forces the mean of the response variable to be equal to the variance of the response (at a fixed set of predictor values). If in an actual data set, the variance greatly exceeds the mean, then the Poisson regression will likely be a poor fit.

b) In negative binomial regression, the variance is allowed to exceed the mean, so it is a good model for "overdispersed" data.

c) When the mean and the variance are similar, the Poisson regression model will likely fit well, and it will be a simpler model than the negative binomial.

Exercise 12.4)

a) This is the mean number of likes for a person with zero followers, when there is no emoji in the tweet.

b) This is the factor by which the mean number of likes changes for a one-person increase in number of followers, holding fixed whether or not there is an emoji.

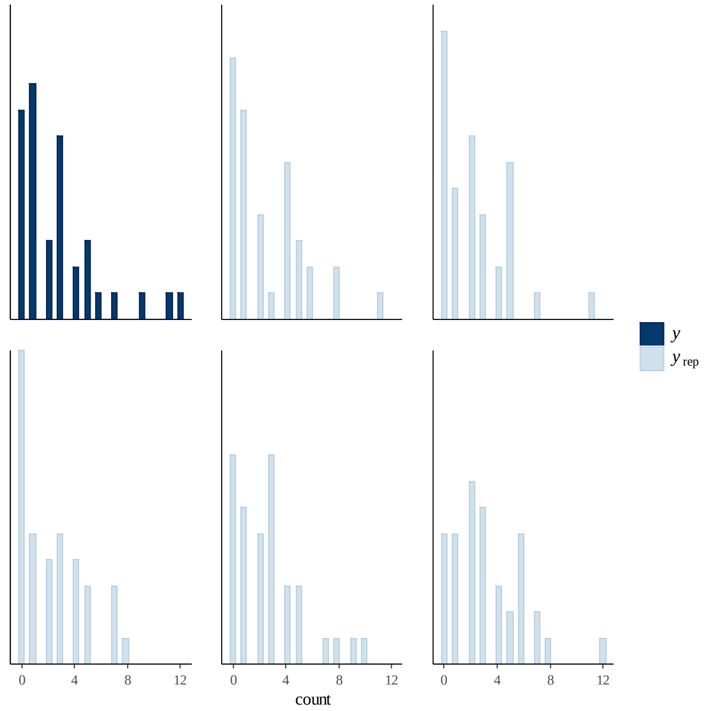
c) This is the factor by which the mean number of likes changes for having an emoji compared to not having an emoji, holding fixed the number of followers.

d) $\exp(\beta_0 + 300\beta_1)$

Bald eagle analysis:

See course webpage for some example R code. Below is an analysis using stan_glm:

Bar graphs of generated posterior predictive distributions show a good fit for the Poisson model:



```
> tidy(eagles_model, conf.int = TRUE, conf.level = 0.95)
# A tibble: 3 x 5
  term          estimate std.error   conf.low conf.high
  <chr>            <dbl>     <dbl>      <dbl>     <dbl>
1 (Intercept) -164.       26.1     -216.      -114.
2 year           0.0822    0.0132     0.0569     0.108
3 hours          0.00319   0.00394   -0.00476    0.0111
```

The ‘Year’ predictor looks important, but "number of hours" may not be an important predictor (the credible interval for $\beta_2$ includes 0)

Based on the within_50 and within+95 measures, the predictive accuracy looks good:

```
> prediction_summary(model = eagles_model, data = bald_eagles)
      mae mae_scaled within_50 within_95
1 1.02245   0.649225 0.7027027         1
> poisson_cv <- prediction_summary_cv(model = eagles_model,
+                                    data = bald_eagles, k = 5)
> poisson_cv$cv
     mae mae_scaled within_50 within_95
1 1.0343  0.8331364 0.7285714     0.975
```

After fitting the interaction model (see R code), we see that the credible interval for the coefficient of the interaction term includes 0:

```
> tidy(eagles_model_int, conf.int = TRUE, conf.level = 0.95)
# A tibble: 4 x 5
  term        estimate  std.error     conf.low    conf.high
  <chr>          <dbl>      <dbl>        <dbl>        <dbl>
1 (Intercept) -1.65e+2  27.5      -220.        -111.
2 year         8.24e-2  0.0138       0.0553       0.110
3 hours        5.20e-3  0.0554      -0.105        0.113
4 year:hours  -7.91e-7  0.0000277   -0.0000551    0.0000541
```

The OUT-OF-SAMPLE (CV) prediction accuracy measures for the interaction model are worse than for the no-interaction model:

```
> prediction_summary(model = eagles_model_int, data = bald_eagles)
      mae mae_scaled within_50 within_95
1 1.02655  0.6454268 0.7027027         1
> poisson_cv_int <- prediction_summary_cv(model = eagles_model_int,
+                                    data = bald_eagles, k = 5)
> poisson_cv_int$cv
       mae mae_scaled within_50 within_95
1 1.249765   0.872032 0.5928571 0.9428571
```

Based on the ELPD measure, the interaction model is VERY SLIGHTLY preferred, but they are almost identical:

```
> # Calculate ELPD for the models
> loo_1 <- loo(eagles_model)
> loo_2 <- loo(eagles_model_int)
> loo_1$estimates
            Estimate        SE
elpd_loo -66.787195  5.895014
p_loo      3.723189  1.037626
looic    133.574389 11.790029
> loo_2$estimates
            Estimate        SE
elpd_loo -66.747720  5.891948
p_loo      3.685873  1.027018
looic    133.495441 11.783895
```

Exercise 13.1)
a) Logistic (the response is binary)
b) Normal (the response is numerical/approximately continuous depending on how the times are recorded)
c) Normal (the response is numerical)

Exercise 13.4(b,c,d)
Let Y=1 if subject believes in climate change
NOT ASSIGNED: a) Odds of belief = P(Y=1)/[1-P(Y=1)] = exp(1.43 - 0.02age)
Probability of belief = P(Y=1) = exp(1.43 - 0.02age) / [1 + exp(1.43 - 0.02age)]

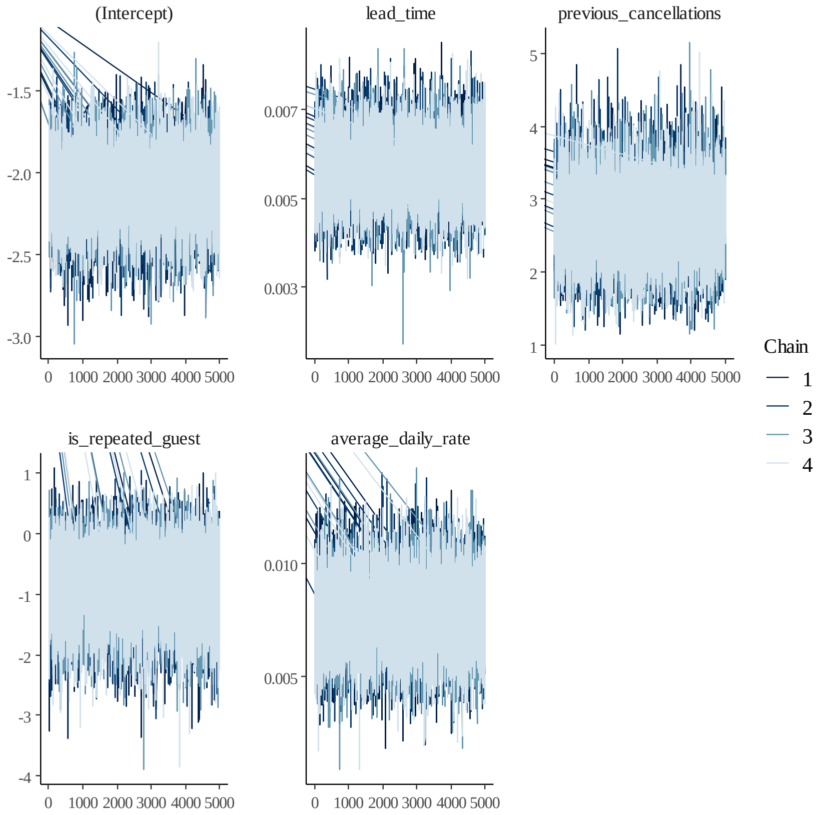
b) The estimated odds of climate change belief change/decrease by a factor of exp(-0.02)=0.98 (a 2% decrease) for each one-year increase in the person's age.
c) P(Y=1 | age=60) = exp(1.43 - 0.02*60) / [1 + exp(1.43 - 0.02*60)] = 0.557.
d) P(Y=1 | age=20) = exp(1.43 - 0.02*20) / [1 + exp(1.43 - 0.02*20)] = 0.737.

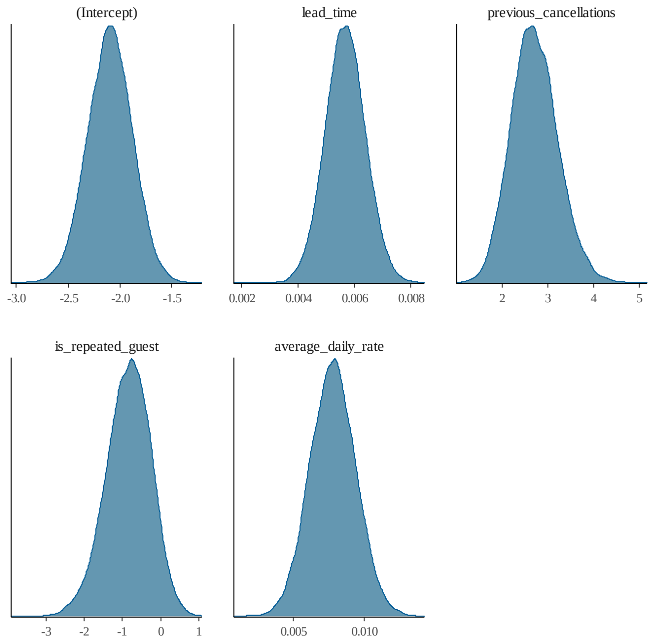
Exercise 13.5)
a) (50+620)/1000 = 0.67
b) 620/(620+30) = 0.954
c) 50/(50+300) = 0.143
d) We should increase c, so that the model will predict "0" for more subjects. This will increase the chance of predicting "0" when the true Y really is 0. (However, it will also increase the chance of wrongly predicting "0" when the true Y really is 1.)

Problem 13.7:

The estimated model coefficients (based on stan_glm, and this is slightly dependent on the prior specification):

```
> tidy(hotel_model, effects = "fixed", conf.int = TRUE, conf.level = 0.80)
# A tibble: 5 x 5
  term                   estimate std.error conf.low conf.high
  <chr>                     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)            -2.09     0.216    -2.37     -1.82
2 lead_time               0.00568  0.000703  0.00478   0.00660
3 previous_cancellations  2.69     0.500     2.08      3.37
4 is_repeated_guest      -0.818    0.584    -1.61     -0.117
5 average_daily_rate      0.00781  0.00160   0.00576   0.00986
```





b) On the probability scale, the estimated model is:
P(Y = 1 | x1,x2,x3,x4) =
exp(-2.09+0.00568x1+2.69x2-0.818x3+0.00781x4)/[1 + exp(-2.09+0.00568x1+2.69x2-0.818x3+0.00781x4)]

c) For $\beta_1$: (0.0048,0.0066)
For $\beta_2$: (2.08, 3.37)
For $\beta_3$: (-1.61, -0.117)
For $\beta_4$: (0.00576, 0.00986)

The estimated odds that the booking is canceled increases by a factor of between exp(2.08)= 8.00 and exp(3.37)=29.1 for a one-unit increase in the number of times the guest has previously canceled, holding the other predictors constant.

The estimated odds that the booking is canceled for people who were previous guests is **less**, i.e., between exp(-1.61)= 0.2 and exp(-0.117)=0.89 times the odds of cancellation for people who were not previous guests, holding the other predictors constant.

d) All of the predictors seem to be somewhat significant predictors statistically since the 80% credible intervals for each excludes zero. However, only “previous cancellations” and “is repeated guest” seem to be important meaningfully. For the other two, the change in odds corresponding to a unit increase in the predictor is small (although for those predictors, perhaps it is more meaningful to consider an increase of more than one unit…).