Homework 2

Denis Ostroushko

Introduction

Handling missing data

Since we are not given explicit instructions on how to handle missing data, and what imputation techniques we need to use, I will preform a complete case analysis. There are 151 observations in the full data set. Dropping observations with at least one missing value results in a data set with 145 observations.

Question 1

1- A

Figure 1 shows the distribution of outcome variable on two scales. We added one unit, `dmfs + 1`, when performing natural logarithm transformation. All 0-values on the log scale represent 0-values on the original scale.

1 - B

We measure baseline mutans streptococci in log(cfu/ml), for the rest of this assignment I will use `log(cfu/ml)` to refer to the units, or measurement of this predictor variable.

We measure `dmfs` as the count of diseased/missing/filled surfaces, allowing one tooth to contribute up to 5 to a total count for each individual, which is a positive integer random variable.

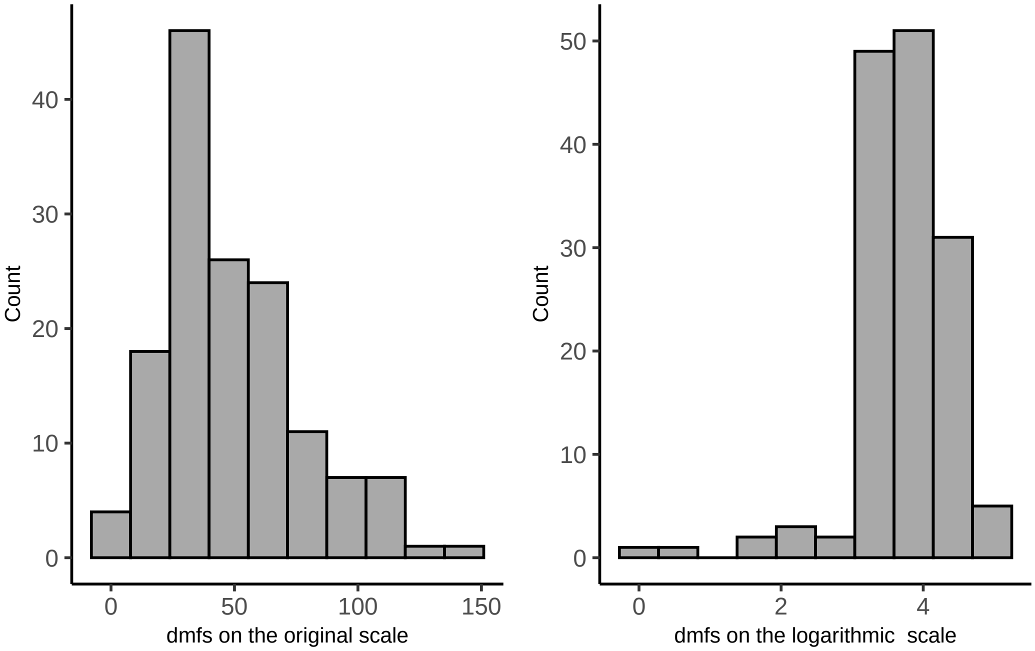

Figure 1: Distribtion of dmfs outcome variable on the natural and logarithmic scales

(i)

Model: Gaussian GLM with an identity link. Also known as a linear regression model.

(1)

We assume that mean for each random variable Y_i can be expressed as a linear combination of predictors X_{ji} and some model parameters β_j. We also assume that Y_i are independent random variables with constant variance.

(2)

Variance is a constant estimated from the model, and has no relationship with the mean, and therefore has no relationship with model parameters β_j

(3)

For a gaussian general linear model with one predictor and an identity link we write mean of Y_i as

$$\hat{E}[Y_i] = \hat{\beta}_0 + \hat{\beta}_1 * X_{1i} \tag{1}$$

Where Y_i is the condition of the teeth, *dmfs*, and X_{1i} is the baseline mutans streptococci.

Table 1: Gaussian GLM with identity link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.06	17.12	1.11	0.27
mutpre	5.46	3.18	1.72	0.09

[a] mutpre represents the effect of baseline mutans streptococci

Table 1 shows estimated model parameters.

We interpret that one unit increase in the baseline log(cfu/ml) measurement results in the average 5.46 incresase in the number of diseased/missing/filled surfaces. Effect is bounded by a (-0.77, 11.69) 95% confidence interval, interval includes 0, there is no evidence of a statistically significant association between baseline log(cfu/ml) and the number of diseased/missing/filled surfaces.

(ii)

Model: Gaussian GLM with a log link.

(1)

We assume that logarithm of a mean for each random variable Y_i can be expressed as a linear combination of predictors X_{ji} and some model parameters β_j. We also assume that $E[Y_i]$, and thus $log(E[Y_i])$ are independent random variables with constant variance.

(2)

We know that for a Gaussian, i.e. normal, distribution, mean and variance have no functional relationship, therefore, parameters β_j have no effect on variance. We estimate variance to be a constant parameter.

(3)

Table 2: Gaussian GLM with log link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.27	0.35	9.41	0.00
mutpre	0.11	0.06	1.78	0.08

[a] mutpre represents the effect of baseline mutans streptococci

We interpret that one unit increase in the baseline log(cfu/ml) measurement results in approximately 12 percent increase in the average number of diseased/missing/filled surfaces. Effect is bounded by a (-1.5, 26.51) 95% confidence interval, interval includes 1, there is no evidence of a statistically significant association between baseline log(cfu/ml) and the average number of diseased/missing/filled surfaces.

(iii)

Model: Gaussian GLM with identity link, using log-transformed values of dmfs. I add +1 to the number of DMF teeth in order to perform a logarithmic transformation and create a linear model.

(1)

Assume that we can write $E[log(Y_i + 1)]$ can be written as a linear combination of predictors. Independence assumption applies here as well.

(2)

Same as **(i)** and **(ii)**

(3)

Table 3: Gaussian GLM with identity link and log transformation of response: model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2	0.44	7.29	0.00
mutpre	0.1	0.08	1.18	0.24

[a] mutpre represents the effect of baseline mutans streptococci

We interpret that one unit increase in the baseline log(cfu/ml) measurement results in the 0.1 increase in the average number of diseased/missing/filled surfaces on a logarithmic scale. Effect is bounded by a (-0.06, 0.26) 95% confidence interval, interval includes 0, there is no evidence of a statistically significant association between baseline log(cfu/ml) and the average number of diseased/missing/filled surfaces on a natural logarithmic scale.

(iv)

Poisson regression model with a log link function

(1)

We assume that we can model $log(E[Y_i])$ as a linear combination of estimated model parameters and observed predictors. Independence assumption holds here as well.

(2)

We assume that $Var[Y_i|\mathbf{X}_i] = E[Y_i]$, so we expect that as our predicted mean $E[Y_i]$ increases, variance increases as well.

Specifying conditions to our case, variance can be written in terms of model parameters:

$$\widehat{Var[Y_i|\mathbf{X}_i]} = \widehat{E[Y_i|\mathbf{X}_i]} = exp[\hat{\beta}_0 + \hat{\beta}_1 * X_{1i}] \tag{2}$$

(3)

Table 4: Poisson GLM with log link model estiamtes

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.28	0.09	36.96	0
mutpre	0.11	0.02	6.81	0

[a] mutpre represents the effect of baseline mutans streptococci

We interpret that one unit increase in the baseline log(cfu/ml) measurement results in the 12 percent increase in the average number of diseased/missing/filled surfaces. Effect is bounded by a (8.22, 15.37) 95% confidence interval, interval does not include 1, there is strong evidence of a statistically significant association between baseline log(cfu/ml) levels and the average number of diseased/missing/filled surfaces.

1 - C

- Immediately we observe that Poisson regression with log link and Gaussian regression with log link produce the most similar estimates, and result in similar estimates. The only difference is that we allow variance of predicted means to be dependent on the value of the mean when using a poisson regression model. It is evident that such model fits the data better, and results in a slope estimate that is statistically significant, giving us evidence that baseline MS values re associated with the condition of teeth.
- No Gaussian model was able to detect statistically significant associations. I suspect that is due to the fact that a Gaussian model imposes a strict assumption of constant variance. Perhaps, for smaller values of damaged surfaces the true variance is smaller than that of higher counts of damaged surfaces. Such data fits Poisson distribution, and regression model, better, which is what we saw right now.

1 - D

(i)

Figure 2 shows residual plot and normality of residuals for a Gaussian regression model with an identity link. We display Pearson residuals, which were studentized using estimated variance of residuals. Smooth trend line fluctuates around zero line for all estimated values, suggesting that the mean of residuals is zero. There are some data points for which we underestimate the value of damaged surfaces, however, they do not appear to be severe outliers.

It appears that residuals deviate from a normal distribution, as evidenced by the QQ-normal plot. Such shape of observed residuals, when compared with the standard normal distribution, suggests that the right tail is quite heavy, while the left end of distribution is not “stretched” enough. Overall, there is some evidence to suggests that the residuals are not normally distributed .

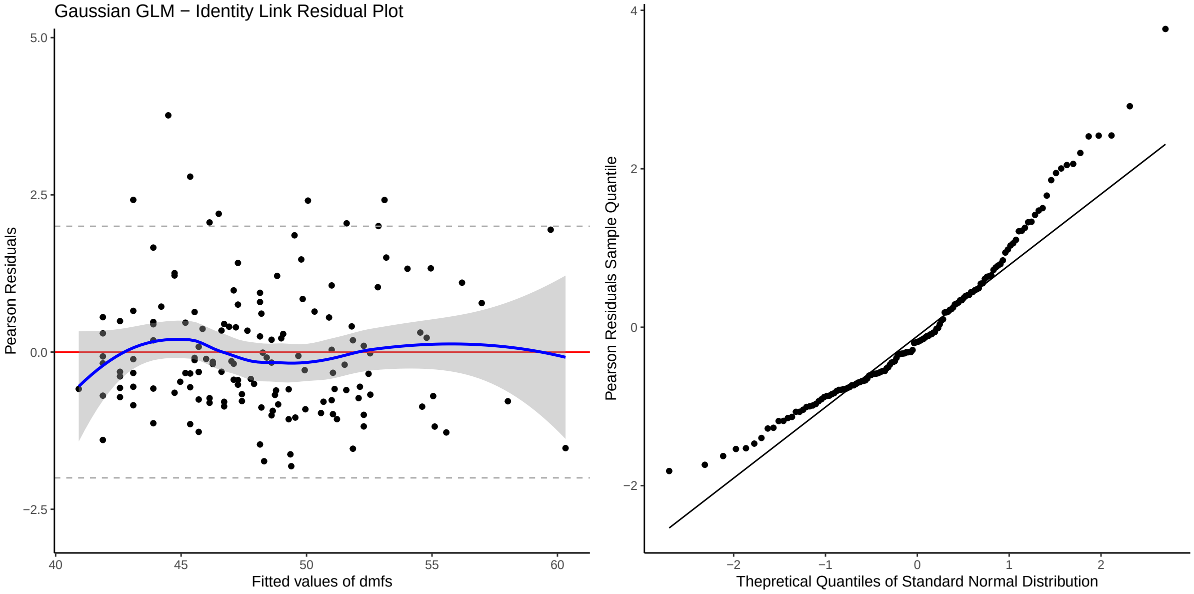

Figure 2: Residuals do not fit normal distribution. No evidence of serious outliers, or issues with linear fit.

(ii)

Figure 3 shows residual plot and normality of residuals for a Gaussian regression model with a log link. We display Pearson residuals, which were studentized using estimated variance of residuals. Smooth trend line fluctuates around zero line for all estimated values, suggesting that the mean of residuals is zero. It also confirms that there are no issues with a linear trend, or evidence of omitted predictors. There are some data points for which we underestimate the value of damaged surfaces, however, they do not appear to be severe outliers.

It appears that residuals deviate from a normal distribution, as evidenced by the QQ-normal plot. Such shape of observed residuals, when compared with the standard normal distribution, suggests that the right tail is quite heavy, while the left end of distribution is not "stretched" enough. Overall, there is some evidence to suggests that the residuals are not normally distributed .

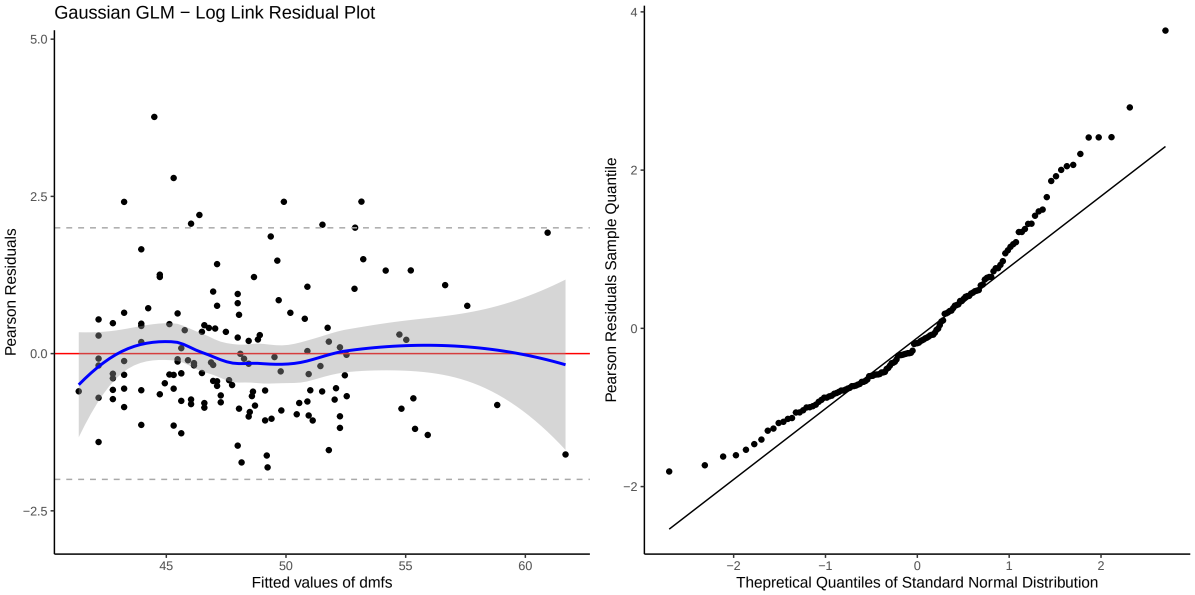

Figure 3: Residuals do not fit normal distribution. No evidence of serious outliers, or issues with linear fit.

(iii)

Figure 4 shows residual plots for a linear model with a log transformed response. These are the most serious issues with a linear model out of the three we evaluated so far. First, there are some serious residual outliers. Some residuals are 5 standard deviations below their predicted value, implying that the true value is much smaller than what the linear model estimates. These outliers are all in the center of predicted values. This may appear because log transformation helps us compare values from different orders of magnitude, however, similar values, especially smaller, are sensitive to mishandling. Moreover, QQ-lot suggests that the left tail is not “long” enough to fit the standard normal distribution.

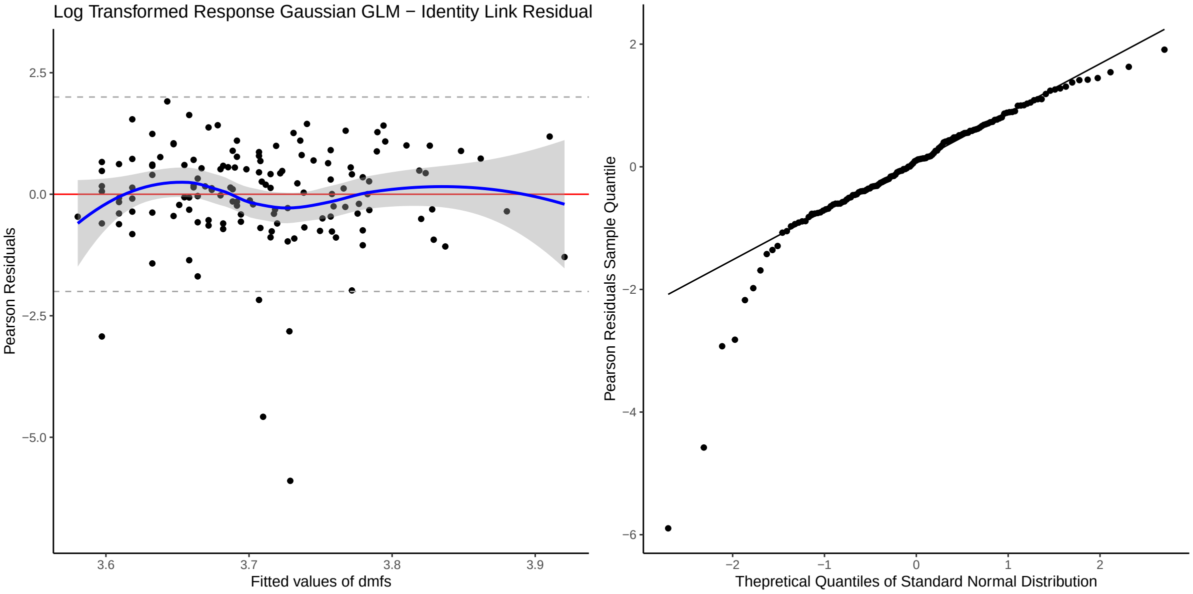

Figure 4: Residuals have a much heavier lower tail than a theoretical standard normal distribution. Evidence of serious residual outliers.

(iv)

We evaluate residuals of poisson regression by studying residual plots and spread of residuals, as well as fit of predicted values vs observed values. Figure 5 shows that we have an issue with the poisson regression model. We know that the assumption that mean must equal variance is highly restrictive. Poisson model underestimates true variance. This is evidenced by Pearson residuals, where some residuals are 10 standard deviations away from the predicted value. Since studentized residuals rely on parameter ϕ, we cam see that estimated parameter $\phi = 1$ is too small, and the true parameter must be much larger.

Therefore, mean variance relationship must be misspecified in this case.

We can see that the fitted values and confidence bands produced from the assumed mean-variance relationship are too tight, and fail to capture the true variation degree in the data.

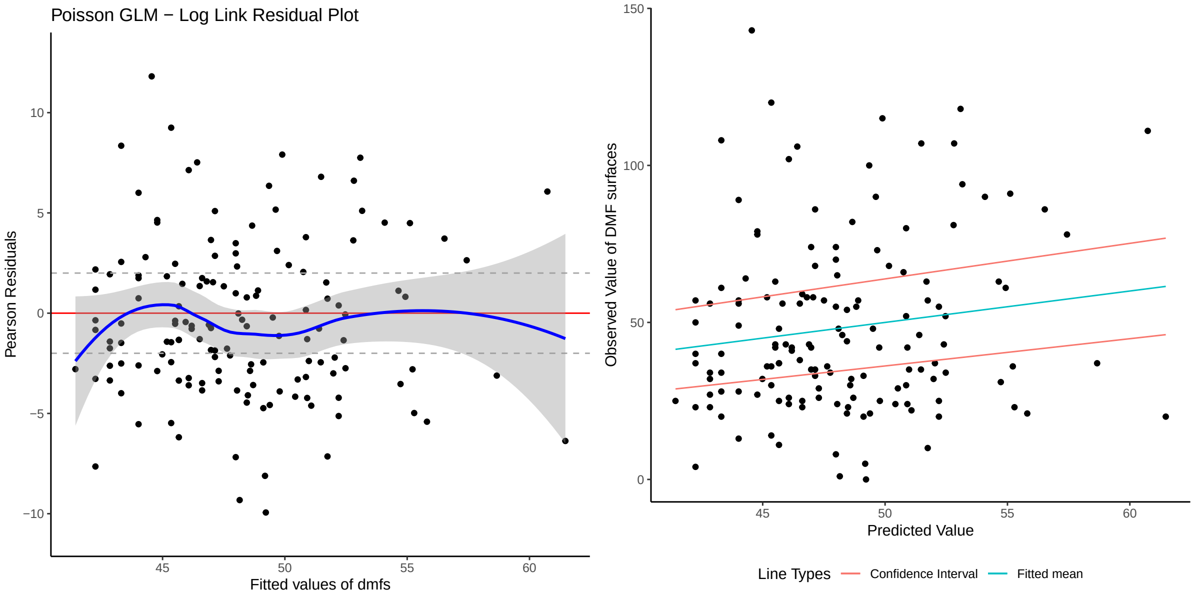

Figure 5: Clear issue with Variance Fit: Need to use quasipoisson model to estiamte correct variance

Figure 6 shows residual plot for a quasipoisson regression model with a log link. Overdispersion parameter is 15.64, which is much larger than 1. Estimating variance function with a correct ϕ parameter produces confidence bound for fitted means that captures variation of data points much more appropriately.

Residual plot also shows that studentized residuals are now mostly within -2/2 bounds, with mean fluctuating around 0. Quasipoisson regression is a more appropriate fit for the sample of data we have.

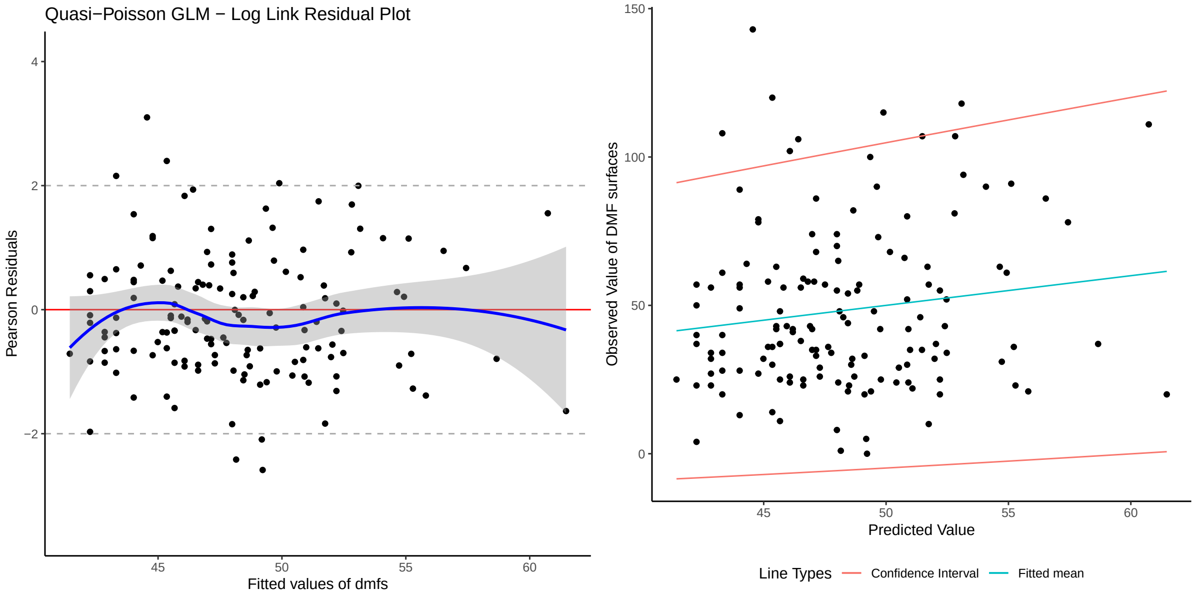

Figure 6: Demonstration that Quasipoisson regression is a better fit to the data

Question 2

There are 49 people in the control group who received no gum treatment, in the complete case data set. After removing these individuals, we are left with 96 observations for the comparison of the effect of gum type on the mutans streptococci levels three months after randomization.

2 - A

Table 5: Gaussin GLM with identity link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.66	0.23	20.25	0
trtgroupDrug	-1.33	0.32	-4.13	0

[a] trtgroupDrug represents the effect of treatment on the outcome

[a] Placebo is the reference for ‘Drug‘ group

Table 5 shows contrast between the average levels of MS between treated and placebo group three months after stating treatment. Average MS levels for the treatment group were 1.33 log(cfu/ml) lower than MS levels for placebo group. The estimate is bounded by the (-1.96, -0.7) 95% confidence interval. The interval does not include 0, there is strong evidence of a statistically significant association between treatment assignment and log(cfu/ml) MS levels.

2 - B

Table 6: Gaussian GLM with identity link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.15	1.10	-0.14	0.89
trtgroupDrug	-1.32	0.29	-4.48	0.00
mutpre	0.89	0.20	4.47	0.00

[a] trtgroupDrug represents the effect of treatment on the outcome

[a] Placebo is the reference for ‘Drug‘ group

Table 6 shows contrast between the average levels of MS between treated and placebo group three months after stating treatment, after adjusting for other predictors. Average MS levels for the treatment group were 1.32 log(cfu/ml) lower than MS levels for placebo group, after accounting for other predictors. The estimate is bounded by the (-1.89, -0.74) 95% confidence interval. The interval does not include 0, there is strong evidence of a statistically significant association between treatment assignment and log(cfu/ml) MS levels.

2 - C

Table 7: Gaussian GLM with identity link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.27	1.44	-0.19	0.85
trtgroupDrug	-1.03	2.21	-0.47	0.64
mutpre	0.91	0.26	3.48	0.00
trtgroupDrug:mutpre	-0.05	0.41	-0.13	0.90

[a] trtgroupDrug represents the effect of treatment on the outcome

[a] Placebo is the reference for ‘Drug‘ group

[a] trtgroupDrug:mutpr represents treatment effect modification

Estimated effect modification of baseline MS log(cfu/ml) measurements for different levels of treatment is NA, after adjusting for other predictors. The interval is bounded by (-0.85, 0.74) the 95% confidence interval, which shows no evidence that the effect of baseline MS levels on the 3 month-post treatment MS levels differs for the two groups.

It appears that neither the main effect nor the interaction terms are statistically significant. This may be due to the fact the the effect of treatment is diluted, or shared, among the two terms.

I also notice that the standard error for the treatment increased drastically, while the point estimate changed a little. This may suggest that with the inclusion of an interaction term we introduce more collinearity among predictors, inflate standard errors, and reduce statistical significance of the two effects.

We may suggest centering the baseline MS levels, and other methods to address these issues. For example, these are model estimates using the same glm family and link function, but centering the baseline MS levels around the sample mean. Table 8 shows that the main effect of treatment aligns closely with the previous results.

Table 8: Effect of centering of baseline MS levels on model estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.66	0.21	22.05	0.0
trtgroupDrug	-1.32	0.30	-4.46	0.0
mutpre	0.91	0.26	3.48	0.0
trtgroupDrug:mutpre	-0.05	0.41	-0.13	0.9

[a] trtgroupDrug represents the effect of treatment on the outcome

[a] Placebo is the reference for 'Drug' group

[a] trtgroupDrug:mutpr represents treatment effect modification

2 - D

Table 9: Gaussian GLM with identity link model estiamtes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.38	1.08	0.35	0.73
trtgroupPlacebo	0.32	0.30	1.05	0.30
mutpre	0.73	0.20	3.58	0.00

[a] trtgroupPlacebo represents the effect of placebo on the outcome

[a] 'No Gum' is the reference for Placebo group

Table 9 shows contrast between the average levels of MS between placebo and "no-intervention" group three months after stating treatment, after adjusting for other predictors. Average MS levels for the placebo group were 0.32 log(cfu/ml) higher than MS levels for no treatment group, after accounting for other predictors. The estimate is bounded by the (-0.28, 0.91) 95%

confidence interval. The interval includes 0, there is no evidence of a statistically significant association between placebo/no-treatment assignment and log(cfu/ml) MS levels. There is no evidence that Placebo improves MS levels, while active treatment does.

Question 3

3 - A

Assumptions

Assume that $f(E[Y_i])$ can be modeled as a linear combination of predictors.

Assume that Y_i are independent Binomial random variables with parameters p_i and $n = 1$. herefore, variance for each Y_i is given by $p_i \times (1 - p_i)$.

We are fitting a logistic regression model, which uses a logit link function. In context of our model we can write assumed model as

$$log[\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}] = \beta_0 + \beta_1 * I(Group = "Placebo") + \beta_2 * I(Group = "Drug") = \mathbf{X}_i\beta \quad (3)$$

In terms of parameters β_j, variance can be written as

$$Var(Y_i|\mathbf{X}_i) = \frac{exp(\mathbf{X}_i\beta)}{1 + exp(\mathbf{X}_i\beta)} \times \frac{1}{1 + exp(\mathbf{X}_i\beta)} \quad (4)$$

Table 10: Binary Outcome GLM with Logit link model estiamtes

	Exp. Estimate	Log-standard Error	Z-statstic	Pr(>|t|)
(Intercept)	0.68	0.30	-1.3	0.19
trtgroupNoGum	0.65	0.43	-1.0	0.32
trtgroupDrug	0.17	0.56	-3.2	0.00

[a] 'NoGum' shows odds ratio for NoGum vs Placebo

[a] 'trtgroupDrug' shows odds ratio for Drug vs Placebo

Table 10 shows contrast between the odds of having elevated MS levels one month post treatment for those in the treatment group when compared to the placebo group. Contrast is expressed as odds ratio. The odds of having elevated MS levels for the treatment group were 0.17 times the odds of the placebo group. Alternatively, the odds of having elevated MS levels in the treatment group were approximately 83% lower than the odds of having elevated MS levels in the placebo group. The estimate is bounded by the (0.05, 0.47) 95% confidence interval. The interval does not include 1, there is strong evidence of a statistically significant association between treatment assignment and likelihood of elevated MS levels.

3 - B

To compare relative risks, of risk ratios, we need to fit a binomial family glm with a log link function.

We assume that Bernoulli random variables Y_i are independent.

Let $E[Y_i] = P(Y_i = 1) = p_i$, then we can write mean of each Y_i as exponentiation linear combination of predictors, $p_i = exp(\mathbf{X}_i\beta)$.

Using this model, variance of Bernoulli random variable can be written as $exp(\mathbf{X}_i\beta) * (1 - exp(\mathbf{X}_i\beta))$

Table 11: Binary Outcome GLM with log link model estiamtes

	Exp. Estimate	Log-standard Error	Z-statstic	Pr(>|t|)
(Intercept)	0.40	0.18	-5.11	0.00
trtgroupNoGum	0.76	0.28	-1.00	0.32
trtgroupDrug	0.25	0.46	-3.00	0.00

[a] ‘NoGum‘ shows odds ratio for NoGum vs Placebo

[a] ‘trtgroupDrug‘ shows odds ratio for Drug vs Placebo

Table 11 shows contrast between the risks of having elevated MS levels one month post treatment for those in the treatment group when compared to the placebo group. Contrast is expressed a risk ratio. The risks of having elevated MS levels for the treatment group were 0.25 times the risks of the placebo group. Alternatively, the risks of having elevated MS levels in the treatment group were approximately 75% lower than the risks of having elevated MS levels in the placebo group. The estimate is bounded by the (0.09, 0.57) 95% confidence interval. The interval does not include 1, there is strong evidence of a statistically significant association between treatment assignment and likelihood of elevated MS levels.

3 - C

We expect that the direction of Risk Ratio and Odds Ratio contrast should be the same. Both are a function of an estimated model parameter p_i. However, odds are a non-linear function that takes in this parameter. Therefore, due to a non-linear nature, the ratio of odds is no longer similar to the ratio of risks.