Stats744: HW4

Illustrating Statistical Inference

Data:

As part of my thesis I have information about substitutions and their variation across four bacterial genomes: *E. coli*, *B. subtilis*, *Streptomyces*, and *S. meliloti*. The bacteria *S. meliloti* is a multi-repliconic bacteria and therefore each of it's replicons is analyzed separately. The data is binary: determining at each site in the genome if there is a substitution (1), or there is not (0). This was also broken down into protein coding and non-protein coding sections of the genome. We expect that non-protein coding segments of the genome to be undergoing more substitutions because they are typically "less important" than protein-coding segments of the genome. I previously fit a logistic regression to the data and obtained a tables with the following results:

Protein Coding				
Bacteria and Replicon	Coefficient Estimate	Standard Error	z-value	P(>|z|)
E. coli Chromosome	-4.308×10^{-8}	2.584×10^{-9}	-16.67	$<2\times10^{-16}$
B. subtilis Chromosome	-4.971×10^{-8}	4.268×10^{-9}	-11.65	$<2\times10^{-16}$
Streptomyces Chromosome	1.989×10^{-8}	8.696×10^{-10}	57.37	$<2\times10^{-16}$
S. meliloti Chromosome	-1.903×10^{-7}	2.13×10^{-8}	-8.934	$<2\times10^{-16}$
S. meliloti pSymA	-6.642×10^{-7}	2.801×10^{-8}	-23.71	$<2\times10^{-16}$
S. meliloti pSymB	1.769×10^{-7}	2.33×10^{-8}	7.593	3.11×10^{-14}

Non-Protein Coding				
Bacteria and Replicon	Coefficient Estimate	Standard Error	z-value	P
E. coli Chromosome	9.896×10^{-9}	7.159×10^{-9}	1.382	0.167
B. subtilis Chromosome	-1.055×10^{-7}	1.25×10^{-8}	-8.436	$<2\times10^{-16}$
Streptomyces Chromosome	1.635×10^{-7}	2.893×10^{-9}	56.5	$<2\times10^{-16}$
S. meliloti Chromosome	-2.900×10^{-7}	3.852×10^{-8}	-7.527	5.18×10^{-14}
S. meliloti pSymA	-1.263×10^{-6}	5.595×10^{-8}	-22.57	$<2\times10^{-16}$
S. meliloti pSymB	4.771×10^{-7}	5.314×10^{-8}	8.978	$<2\times10^{-16}$

Read in the data and format: (I have over 160 million lines of data so it was just not feasible to run the logistic regressions. So I manually added the output from the logistic regression results)

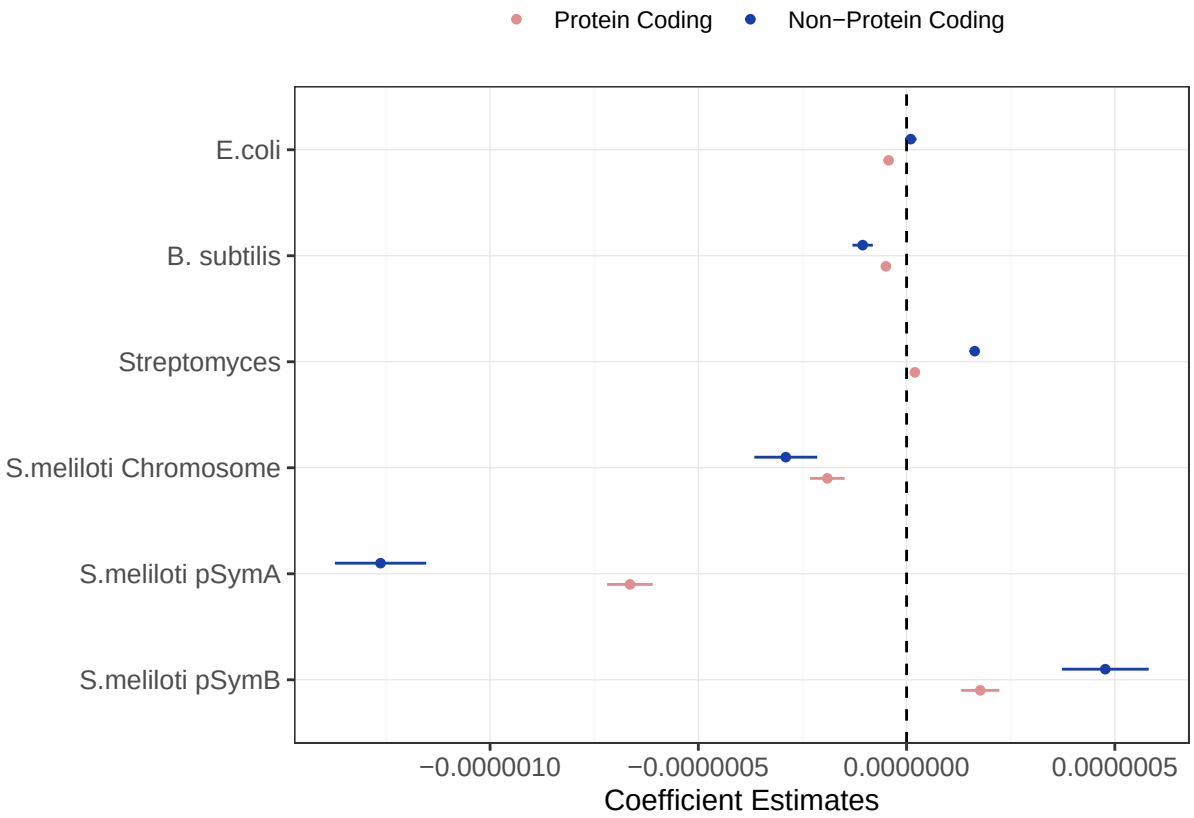
I think I may have cheated a bit, but I made each bacteria a "term" even though they were not actually "terms" in the model. Each logistic regression model was run separately on both the protein coding and non-protein coding sections of each bacterial replicon. I did this so the data would be displayed in the graph the way that I want it to be.

```
subs_dat <- read.table(header = TRUE, check.names = FALSE, text = "
term model  estimate  std.error statistic p.value
ecoli 1 -0.00000004308 0.000000002584  -16.67 0.0000000000000002
ecoli 0 0.000000009896  0.000000007159  1.382 0.167
bass  1 -0.00000004971  0.000000004268  -11.65  0.0000000000000002
bass  0 -0.0000001055 0.0000000125  -8.436  0.0000000000000002
strep  1 0.00000001989 0.0000000008696 57.37 0.0000000000000002
strep 0 0.0000001635  0.000000002893  56.5  0.0000000000000002
schrom  1 -0.0000001903 0.0000000213  -8.934  0.0000000000000002
schrom  0 -0.0000002900 0.00000003852 -7.527  0.0000000000000518
pa  1  -0.0000006642  0.00000002801 -23.71  0.0000000000000002
pa  0 -0.000001263  0.00000005595 -22.57  0.0000000000000002
pb  1 0.0000001769  0.0000000233  7.593 0.0000000000000311
pb  0 0.0000004771  0.00000005314 8.978 0.0000000000000002

")
```

Graph 1

```
dwplot(size = 3, subs_dat) %>% relabel_predictors(ecoli = "E.coli", bass = "B. subtilis"
    strep = "Streptomyces", schrom = "S.meliloti Chromosome", pa = "S.meliloti pSymA",
    pb = "S.meliloti pSymB", face = "italic") + xlab("Coefficient Estimates") +
    geom_vline(xintercept = 0, lty = 2) + scale_colour_manual(values = c("#E08F8F",
    "#173FAA"), name = "", labels = c("Protein Coding", "Non-Protein Coding"))
```



Discussion of Graph 1

In this graphic the coefficient estimates for each logistic regression are plotted. I have added in a reference line at zero to assist with determining statistical significance. Typically, any coefficient estimates (and their respective standard errors) that cross over the zero reference line are not considered significant. In this plot, it is clear that the only estimate that is not significant is the non-protein coding regions of *E. coli*. In the protein coding regions of *Streptomyces*, the coefficient estimate is very close to the zero reference line, but does not touch it, and is therefore significant.

It is also very easy to see which bacterial replicons have a positive or negative coefficient estimate. pSymB of *S. meliloti* and *Streptomyces* are the only replicons to have a significant positive coefficient estimate, while the rest of the bacterial replicons have a significant negative coefficient estimate.

Discussion on Biological Relevance

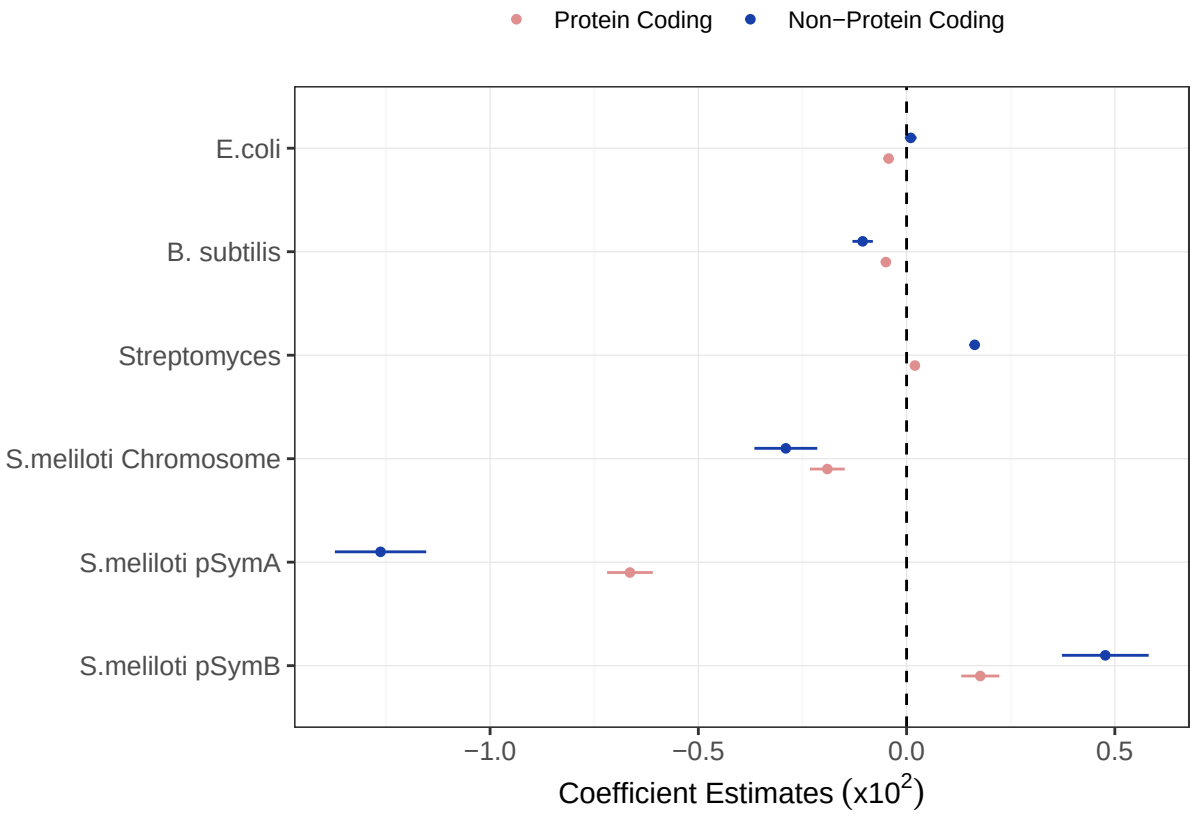
If the coefficient estimate is significantly positive, in this case it means that the probability of a substitution at a particular base increases with increasing distance from the origin of replication. This is to be expected because as replication proceeds from the origin towards the terminus, it is more prone to errors. Additionally, genes that are associated with core functions of the organism are typically found near the origin of replication. Since errors in

these genes could be detrimental, we expect less substitutions and mutations to be present. Therefore, we should observe more substitutions near the terminus of the genome, and less substitutions near the origin of replication.

If the coefficient estimate is significantly negative, in this case it means that the probability of a substitution at a particular base in the genome decreases with increasing distance from the origin of replication. This is counterintuitive to the typical organization of bacterial genomes. However, I failed to mention the most important part of my analysis, I take into account genomic rearrangements. Bacterial genomes are frequently re-arranging, re-organizing, and re-working their genomic information through a variety of processes such as duplications, translocations, and horizontal gene transfer. I account for these rearrangements by allowing genomic segments to be present in different locations in various bacterial strains. Previous studies have not accounted for rearrangements, which could explain why most of the bacterial replicons in this study have a higher probability of substitutions near the origin of replication.

Graph 2

```
subs_dat$estimate <- subs_dat$estimate * 1000000
subs_dat$std.error <- subs_dat$std.error * 1000000
dwplot(size = 3, subs_dat) %>% relabel_predictors(ecoli = "E.coli", bass = "B. subtilis
    strep = "Streptomyces", schrom = "S.meliloti Chromosome", pa = "S.meliloti pSymA",
    pb = "S.meliloti pSymB", face = "italic") + xlab(bquote("Coefficient Estimates" ~
    (x10^2))) + geom_vline(xintercept = 0, lty = 2) + scale_colour_manual(values = c("#
    "#173FAA"), name = "", labels = c("Protein Coding", "Non-Protein Coding"))
```



Discussion of Graph 2

I did not like the way that the coefficient estimate numbers looked on the x-axis so I wanted to try and see what they would look like as smaller numbers. I think it looks better, but I am not sure if this is maybe harder to read because people have to do some mental math to determine how small the coefficient estimate is. I am also unsure if scaling the coefficient estimates and standard errors this way is mathematically correct.

I was really excited for this assignment because I think it is a cool way to present multiple coefficient estimates focusing on their significance. It never occurred to me to present my regression results like this.

Other things...

```
# I tried many things to make the bacteria names italic (like the below
# code), but I just could not get it to work. dwplot() appears to not
# like expressions or paste...or substitute

# subs_dat %>% dwplot(relabel_predictors(c('ecoli' =
# expression(paste(italic('E.coli'), ' Chromosome')), 'bass' =
# expression(paste(italic('B. subtilis'), ' Chromosome')), 'strep' =
# expression(paste(italic('Streptomyces'), ' Chromosome')), 'sinoC' =
```

```
# expression(paste(italic('S.meliloti'), ' Chromosome')), 'pSymA' =
# expression(paste(italic('S.meliloti'), ' pSymA')), 'pSymB' =
# expression(paste(italic('S.meliloti'), ' pSymB'))))) + xlab('CHANGE
# ME (x10^2)') + geom_vline(xintercept=0,lty=2)+
# scale_colour_manual(values=c('#E08F8F','#173FAA'), name = '',
# breaks=c('0','1'), labels = c('Protein Coding', 'Non-Protein
# Coding'))

# I also tried this before I made the dwplot make sure bacteria names
# are italic bac_names <- c('ecoli' =
# expression(paste(italic('E.coli'), ' Chromosome')), 'bass' =
# expression(paste(italic('B. subtilis'), ' Chromosome')), 'strep' =
# expression(paste(italic('Streptomyces'), ' Chromosome')), 'sinoC' =
# expression(paste(italic('S.meliloti'), ' Chromosome')), 'pSymA' =
# expression(paste(italic('S.meliloti'), ' pSymA')), 'pSymB' =
# expression(paste(italic('S.meliloti'), ' pSymB'))) subs_dat$term <-
# bac_names
```