

# 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, stating at each site in the genome if there is a substitution, or there is not. This was also broken down into protein coding and non-protein coding sections of the genome. I previously fit a logistic regression to the data and obtained a table with the following results:

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

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

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.

```
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 0.00000000000000002
ecoli 0 0.000000009896 0.000000007159 1.382 0.167
bass 1 -0.00000004971 0.000000004268 -11.65 0.00000000000000002
bass 0 -0.0000001055 0.0000000125 -8.436 0.00000000000000002
```

```

strep 1 0.00000001989 0.0000000008696 57.37 0.00000000000000002
strep 0 0.0000001635 0.000000002893 56.5 0.00000000000000002
schrom 1 -0.0000001903 0.0000000213 -8.934 0.00000000000000002
schrom 0 -0.0000002900 0.00000003852 -7.527 0.00000000000000518
pa 1 -0.0000006642 0.00000002801 -23.71 0.00000000000000002
pa 0 -0.000001263 0.00000005595 -22.57 0.00000000000000002
pb 1 0.0000001769 0.0000000233 7.593 0.000000000000000311
pb 0 0.0000004771 0.00000005314 8.978 0.00000000000000002

")

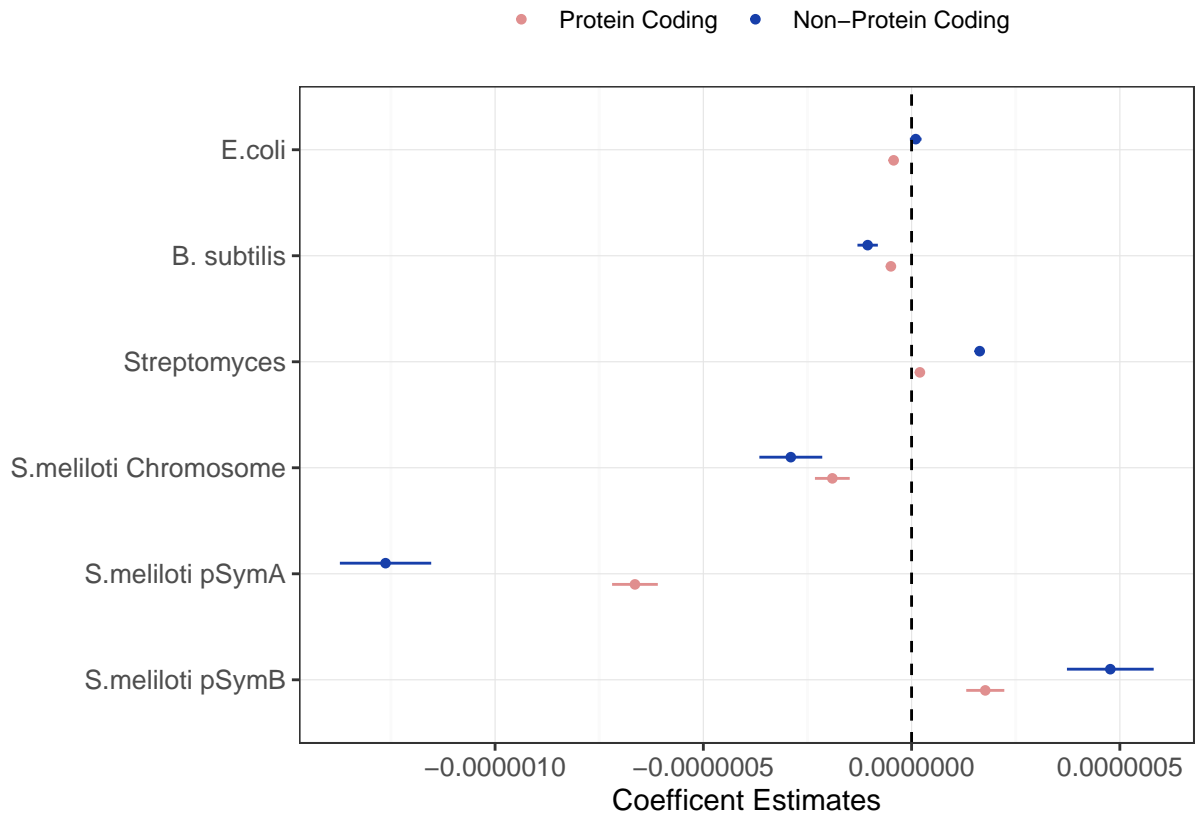
```

## Graph

```

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

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 logistic regression results like this.

```
#I tried many things to make the bacteria names italic (like the below code), but I ju
#dwplot() appears to not like expressions or paste...or substitute
#subs_dat %>% dwplot(relabel_predictors(c("ecoli" = expression(paste(italic("E.coli"),
#                                "bass" = expression(paste(italic("B. subtilis"), " Chromoso
#                                "strep" = expression(paste(italic("Streptomyces"), " Chromo
#                                "sinoC" = expression(paste(italic("S.meliloti"), " Chromoso
#                                "pSymA" = expression(paste(italic("S.meliloti"), " pSymA"))
#                                "pSymB" = expression(paste(italic("S.meliloti"), " pSymB")))
#  xlab("CHANGE ME (x102)") +
#  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
```