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 analized separatly. 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	Coefficent 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}

Bacteria and Replicon	Coefficent 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	$<2\times10^{-16}$

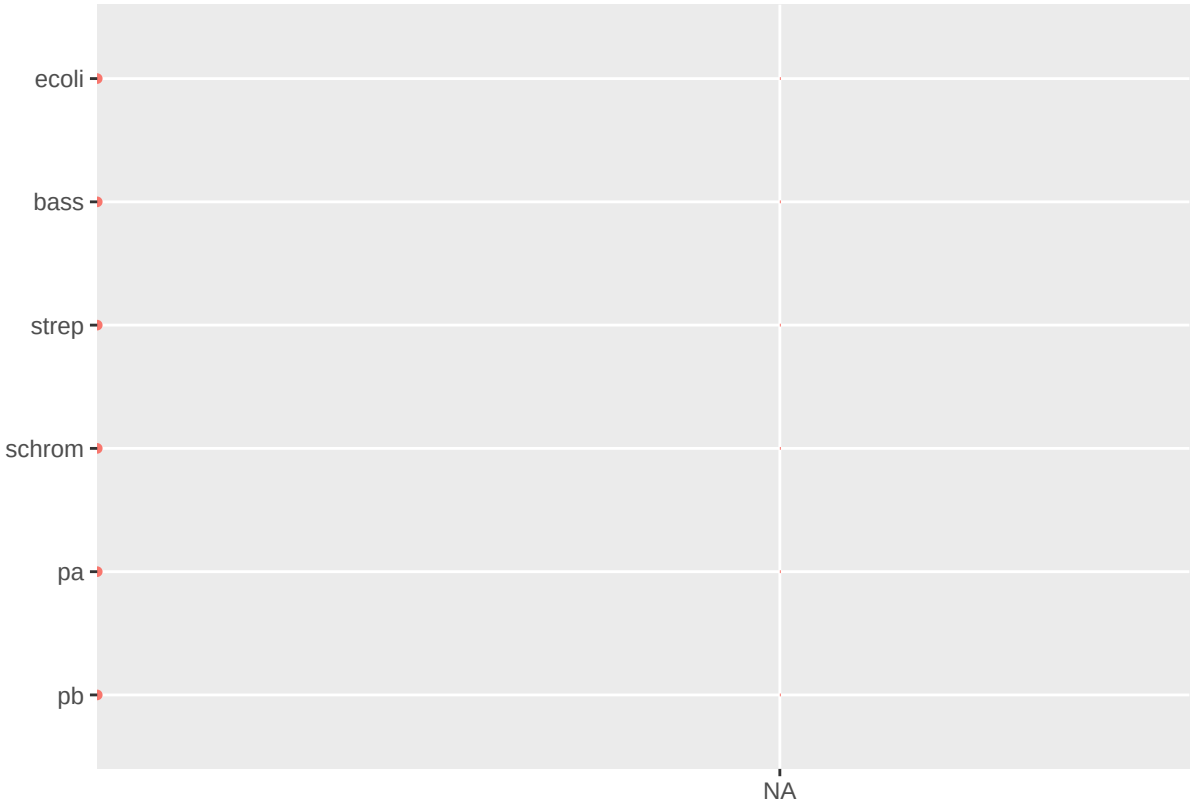
Read in the data and format: (I have over 160 million lines of data so it was just not feisable to run the logistic regressions. So I manually added the output from the logistic regression results)

```
subs_dat <- read.table(header=TRUE,check.names=FALSE,text="
model bac cod_class term  estimate  std.err statistic p.value
logit ecoli 1 pos -0.00000004308 0.000000002584  -16.67 0.0000000000000002
logit ecoli 0 pos 0.000000009896  0.000000007159  1.382 0.167
logit bass  1 pos -0.00000004971  0.000000004268  -11.65  0.0000000000000002
logit bass  0 pos -0.0000001055 0.0000000125  -8.436  0.0000000000000002
logit strep  1 pos 0.00000001989 0.0000000008696 57.37 0.0000000000000002
logit strep 0 pos 0.0000001635  0.000000002893  56.5  0.0000000000000002
logit schrom  1 pos -0.0000001903 0.0000000213  -8.934  0.0000000000000002
```

```
logit schrom  0 pos -0.0000002900 0.00000003852 -7.527  0.0000000000000518
logit pa  1 pos  -0.0000006642  0.00000002801 -23.71  0.0000000000000002
logit pa  0 pos -0.000001263  0.00000005595 -22.57  0.0000000000000002
logit pb  1 pos 0.0000001769  0.0000000233  7.593 0.0000000000000311
logit pb  0 pos 0.0000004771  0.00000005314 8.978 0.0000000000000002

")

#separated this into coding
subs_dat_cod <- read.table(header=TRUE,check.names=FALSE,text="
term  estimate  std.err statistic p.value
ecoli -0.00000004308 0.000000002584  -16.67 0.0000000000000002
bass -0.00000004971  0.000000004268  -11.65  0.0000000000000002
strep 0.00000001989 0.0000000008696 57.37 0.0000000000000002
schrom -0.0000001903 0.0000000213  -8.934  0.0000000000000002
pa  -0.0000006642  0.00000002801 -23.71  0.0000000000000002
pb 0.0000001769  0.0000000233  7.593 0.0000000000000311

")

#separated this into non coding
subs_dat_noncod <- read.table(header=TRUE,check.names=FALSE,text="
term  estimate  std.err statistic p.value
ecoli 0.000000009896  0.000000007159  1.382 0.167
bass -0.0000001055 0.0000000125  -8.436  0.0000000000000002
strep 0.0000001635  0.000000002893  56.5  0.0000000000000002
schrom -0.0000002900 0.00000003852 -7.527  0.0000000000000518
pa -0.000001263  0.00000005595 -22.57  0.0000000000000002
pb 0.0000004771  0.00000005314 8.978 0.0000000000000002

")
```

Graphs

```
#why does this have NA on the x-axis??? is it because things are too small??
dotwisk <- dwplot(subs_dat_noncod)
print(dotwisk)
```




```
#my other failed attempt
#plot_model(subs_dat_noncod)
```