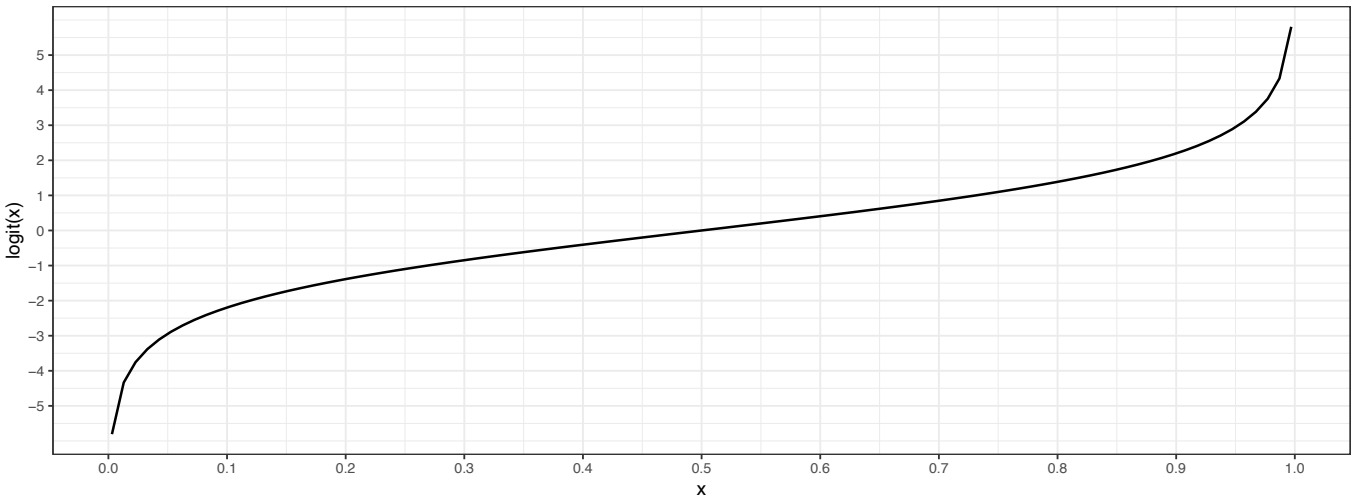
In-Class Problems

2 Below is a plot of the logit function.



a) Use this graph to estimate `ilogit(-2)` (the inverse logit function evaluated at -2) as accurately as the graph allows. Show on the graph how you are getting your estimate.

b) What is the mathematical formula for the logit function?

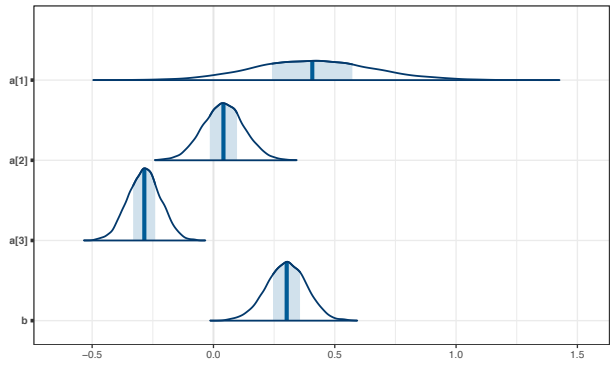
c) How is the logit function used in logistic regression?

d) What property/properties of the logit function make it useable/useful for this purpose?

Some Models and Output

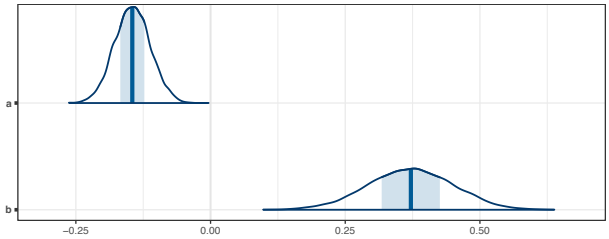
Model u1

```
Hamiltonian Monte Carlo approximation
8000 samples from 4 chains

Sampling durations (seconds):
        warmup sample total
chain:1   0.03   0.05  0.08
chain:2   0.03   0.05  0.08
chain:3   0.03   0.05  0.08
chain:4   0.03   0.05  0.08

Formula:
cases ~ dbinom(n, p)
logit(p) <- a[geno_idx] + b * male
a[geno_idx] ~ dnorm(0, 1.5)
b ~ dnorm(0, 0.5)
      mean   sd  5.5% 94.5% n_eff Rhat4
a[1]  0.41 0.24  0.03  0.81  6296     1
a[2]  0.04 0.08 -0.09  0.18  4899     1
a[3] -0.29 0.07 -0.39 -0.18  3890     1
b     0.30 0.08  0.17  0.43  3864     1
```



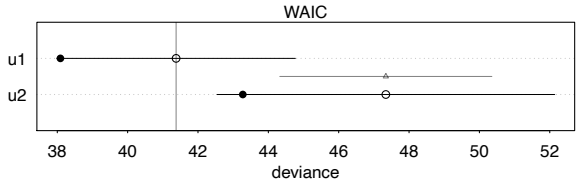
Model u2

```
Hamiltonian Monte Carlo approximation
8000 samples from 4 chains

Sampling durations (seconds):
        warmup sample total
chain:1   0.02   0.05  0.07
chain:2   0.02   0.04  0.06
chain:3   0.02   0.04  0.07
chain:4   0.02   0.04  0.06

Formula:
cases ~ dbinom(n, p)
logit(p) <- a * Gdose + b * male
a ~ dnorm(0, 0.5)
b ~ dnorm(0, 1.5)
   mean   sd    5%   95% n_eff Rhat4
a -0.15 0.03 -0.20 -0.09  3080     1
b  0.37 0.08  0.24  0.50  3175     1
```




```
compare(u1, u2)

   WAIC   SE dWAIC  dSE pWAIC weight
u1 41.4 3.39     0   NA   1.6   0.95
u2 47.3 4.80     6 3.01   2.0   0.05
```

```
compare(u1, u2) %>% plot()
```



3 The FUSION study is a study of genetic factors that may predispose to type 2 diabetes (T2D). One part of this study looked at the effect of SNPs (single nucleotide polymorphisms), a type of genetic marker. The aggregated data below shows the sex, T2D status, and genotype (GG, GT, or TT) for the RS12255372 SNP. Cases are subjects who have T2D. So the first row indicates that of the 733 women with the GG genotype, 314 (42.84%) have T2D. Gdose is 2 because the genotype includes 2 G's; Tdose is 0.

```
# A tibble: 6 x 9
  sex    male genotype geno_idx Tdose Gdose     n cases prop_cases
  <fct> <dbl> <fct>       <dbl> <dbl> <int> <int> <dbl>      <dbl>
1 F         0 GG              3     0     2   733   314      0.428
2 F         0 GT              2     1     1   339   174      0.513
3 F         0 TT              1     2     0    34    19      0.559
4 M         1 GG              3     0     2   839   423      0.504
5 M         1 GT              2     1     1   345   201      0.583
6 M         1 TT              1     2     0    41    29      0.707
```

a) In Model `u2`, most of the posterior for `a` is negative and for `b` it is positive. What does this tell us about what `u2` thinks about type 2 diabetes? Does Model 1 agree (at least qualitatively)?

b) What proportion of men with genotype TT does each model predict will get T2D?

i. Estimate this as well as you can from the output given.

ii. Explain what you would do if you had access to the data and model in R that would be better.

c) Explain the primary difference between the two models.

d) According to WAIC, is one model preferable to the other? If so, which one?

e) Explain briefly what WAIC (or PSIS) attempts to estimate and how it performs this estimation.

f) In model 1, the posterior distribution for `a[1]` is wider than the posterior distributions of `a[2]` and `a[3]`. Why do you think this is?

g) Why is it not possible to tell from this output how confident Model `u1` is that there is a difference between the TT and GT genotypes? What would you need to do to answer that question?

h) **Either, neither, or both?** Are either of these models multilevel (hierarchical) models? Explain.