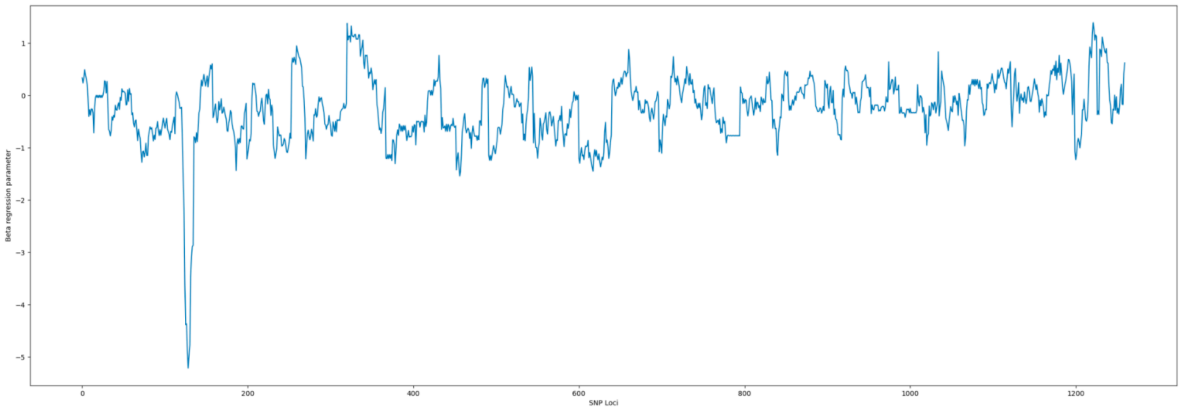
(a) Univariate regression

- [10 pts] Perform a univariate regression, using the $LEU2$ expression levels as output and each SNP as input. Obtain $\beta = [\beta_1, \ldots, \beta_J]$, where $J = 1,260$ and β_i is the regression parameter estimated by regressing $LEU2$ expression levels on SNP i. Plot β_i's across the genome.

5.(a) See python file for code. Plot attached below:



(b) Multivariate regression

- [10 pts] Can you obtain an estimate of the regression coefficients for a multivariate regression model using all SNPs as inputs and $LEU2$ expression as output? Explain why this is not possible. [Hint: you do not need to write a code to answer this question.]

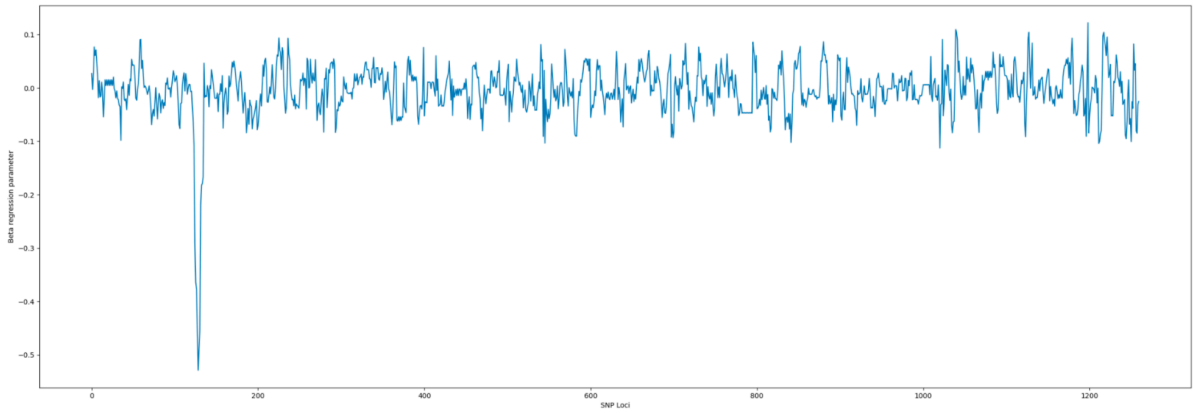
5. (b) It is not possible because the number of samples is way less than the number of SNPs in each sample and so the X^TX term will not be invertible which is required for finding the regression parameters.

(c) Ridge regression.

- [10 pts] Perform a ridge regression, assuming a prior distribution $\beta_i \sim N(0, \sigma_0^2)$ for $i = 1, \ldots, J$, $J = 1,260$, with $\sigma_0^2 = 5.0$. Your multivariate regression model should predict $LEU2$ expression levels given all SNPs. Plot your estimated regression coefficients $\beta = [\beta_1, \ldots, \beta_J]$.

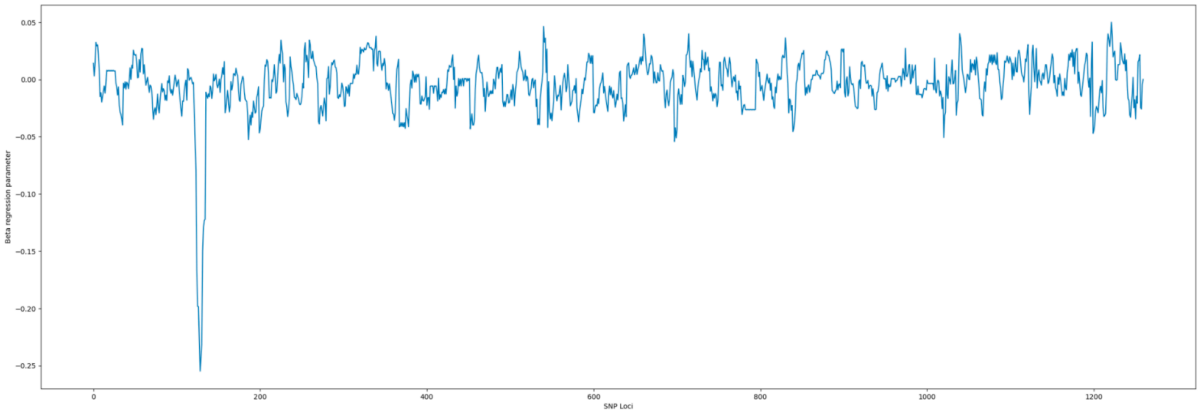
5.(c) See python file for code. Plots attached below:

σ_o^2= 5.0



- [10 pts] Repeat the analysis above with a prior distribution $\beta_i \sim N(0, \sigma_0^2)$ for $i = 1, \ldots, J$, $J = 1,260$, with $\sigma_0^2 = 0.005$. Plot your estimated regression coefficients $\beta = [\beta_1, \ldots, \beta_J]$.

σ_o^2= 0.005



- [10 pts] Explain the different effects of the two prior distributions above on the regression parameter estimates. Which SNP has the strongest influence on the $LEU2$ expression?

The resulting slope using a lesser σ_0^2 (0.005) is less steep than using the higher σ_0^2 (5.0). This means that the predictions for the parameters becomes less sensitive to the SNP loci as the σ_0^2 decreases and s increases. Small σ_0^2 means large s and a stronger prior belief and it means stronger pull of estimates toward zero.

SNP no. 128 (0-based index) has highest influence. It is called YDR085C.