

Genetics 210
Problem Set 2
Due: May 2, 2013
Email to gene210.stanford@gmail.com

In all of the previous GWAS examples we have explored, the phenotype has been a discrete variable. For example, either you have wet earwax or dry, bitter taster or not, brown eyes or blue (or green). However, phenotypes are not always black and white. For example, people are not just tall and short but many gradations in between -- this is called a quantitative variable. This is true for many clinically relevant phenotypes as well. For example, the severity of Type 2 Diabetes is often assessed using fasting glucose levels in the blood. Higher levels indicate a more severe case of diabetes, while lower levels (but still high) may only indicate a risk of diabetes. Another case where quantitative phenotypes are important is in drug response. The necessary dose of warfarin (a common anticoagulant and rat poison) is highly variable across the population. Finding the correct stable dose is important to mitigate the chance of severe adverse events associated with warfarin use (e.g. internal bleeding or excessive clotting). Here we explore a quantitative GWAS, compare it to a traditional case/control GWAS, and also learn a little about covariates and regression analysis.

To complete this part of the problem set you will need to download some data from the website. You can download the data file here:

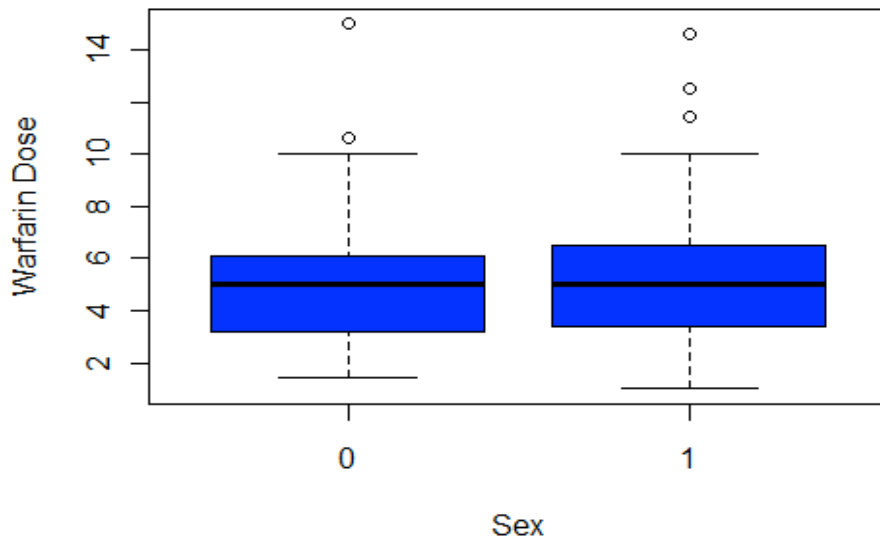
<http://stanford.edu/class/gene210/files/problem-sets/2011/warfarin-gwas.csv>

The data you downloaded is from a quantitative GWAS exploring the genetic determinants of warfarin dosing (<http://bloodjournal.hematologylibrary.org/content/112/4/1022.full.pdf>). The genetics of warfarin dosing is of special interest since it is difficult to predict from clinical variables alone. For example, your little grandmother may require a huge dose of warfarin while Stanford Linebacker, Shane Skov (6' 3", 251lbs) may require a very small dose. This variance can be partially explained using genetics.

The following instructions assume you are using Microsoft Excel to perform this analysis. Experts may use the data analysis software of their choice.

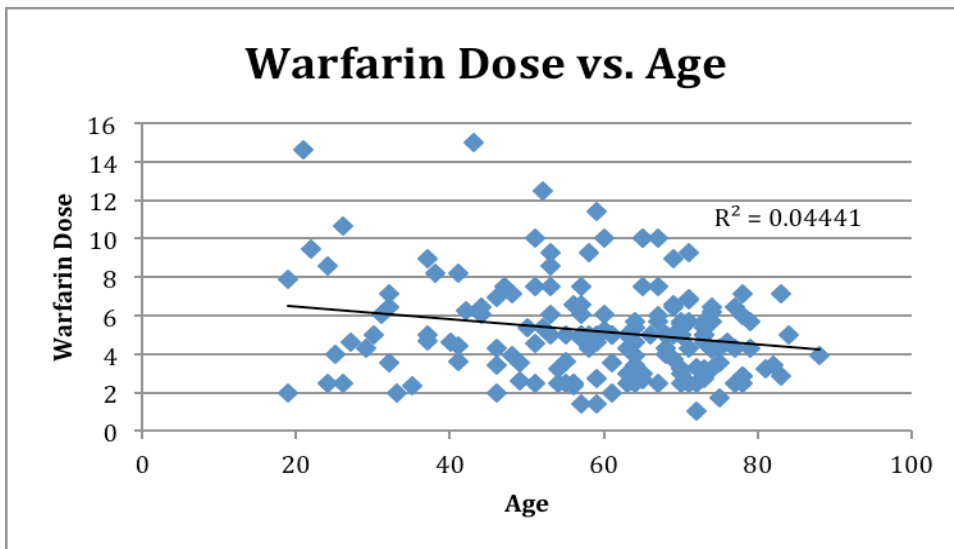
- 1) We will now make 4 scatter plots of the data. For each of the following clinical variables make a scatter plot with warfarin dose on the y-axis and the clinical variable on the x-axis. For each plot, describe whether the data appear correlated, anti-correlated or unrelated.

Warfarin Dose vs. Sex

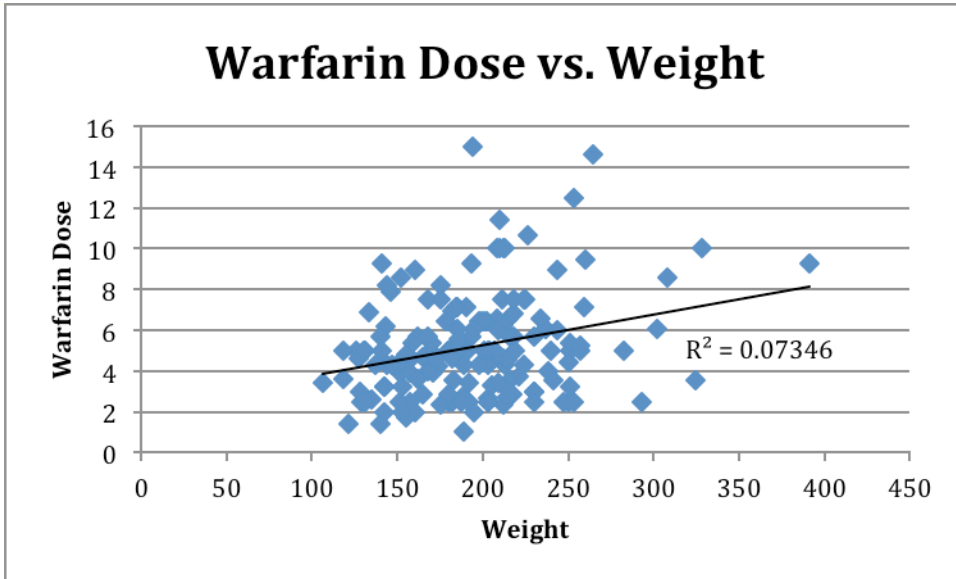


Sex does not appear to be correlated with warfarin dose.

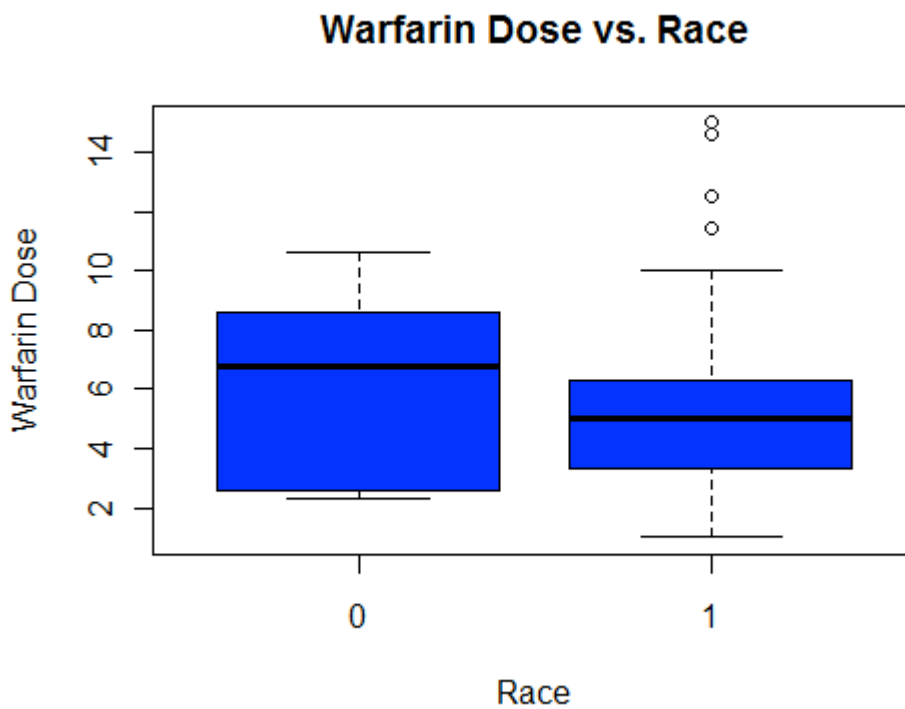
Warfarin Dose vs. Age



Age does not appear to be correlated with warfarin dose.

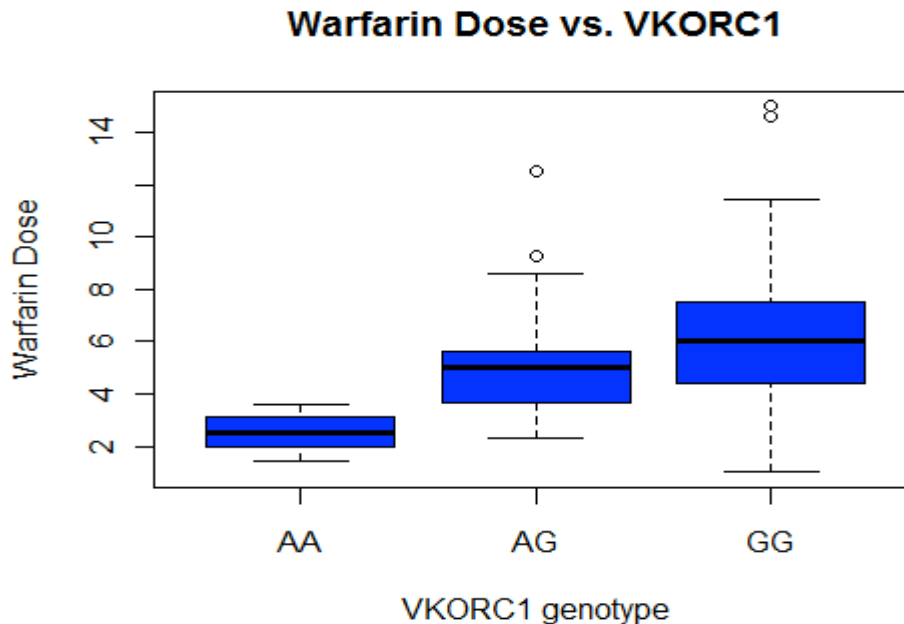


Weight appears to be slightly correlated with warfarin dose. Individuals with higher weight generally require a higher warfarin dose.



Warfarin dose does not appear to be correlated with race.

- 2) One SNP in the gene, VKORC1, is the most significant genetic covariate known for warfarin dose. Here is a bar graph showing warfarin dose versus VKORC1 genotype. Can you say anything about whether the A allele of VKORC1 is recessive, semi-dominant, or dominant to the G allele for this trait? Explain.



Overall individuals with an A allele seem to require less warfarin, while individuals with G alleles require more. Looking at the graph it appears that heterozygotes resemble GG homozygotes more closely than AA homozygotes, suggesting that the A allele is recessive to the G allele. It is hard to tell if the G allele is truly dominant or semi-dominant to the A allele given the range of the phenotype.

Regression analysis allows you to combine multiple independent variables together in order to predict an outcome variable. This outcome variable is often called the dependent variable. In our case the dependent variable is the warfarin dose and the independent variables are the clinical variables (i.e. sex, weight, age, and race) and the genetic variable (i.e. VKORC1 genotype). What is important about regression is that it allows you to combine the independent variables in proportion to how much of the phenotypic variance they explain. For example, if sex explains more of the variance of warfarin dose than age then sex will receive more weight. We will now perform a regression analysis to predict warfarin dose.

In this class we are focused on the interpretation and implications of the results of genetic analysis. Therefore the majority of the computational tasks have been completed for you.

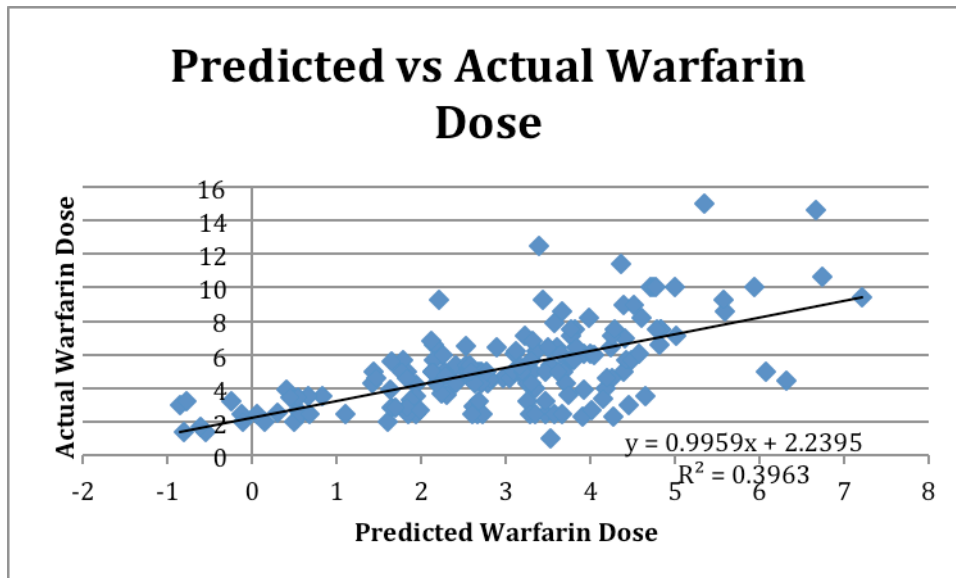
3) A regression analysis was performed using sex, age, weight, race, and VKORC1 genotypes as the independent variables (these data are listed in your downloaded file). Each variable's coefficient is listed in the table below. The coefficient is the "weight" that's assigned to that variable. Note that when a dependent variable is discrete, as it is for sex, then you must use "indicator" variables, which transform them into numerical values. For example, the indicator variable for sex says that if the patient is female the value is "0" and if the patient is male, the value is "1." The fact that the coefficient of the sex variable is negative means that males, on average, require a lower dose of warfarin than females.

Independent Variable	Coefficient
Sex (0 = female, 1 = male)	-0.639
Age	-0.038
Weight	0.016
Race (0 = Hispanic, 1 = white)	-0.239
VKORC1 (0 = AA, 1 = AG)	2.203
VKORC1 (0 = AA, 1 = GG)	4.117

- a. Examine the coefficient values in the table above. What do they tell you about the relationship between each independent variable and warfarin dose? Does this make sense considering the plots you made in (1)? (Hint: If you are stuck, read through the paragraph that precedes the table again ;)

These coefficients make sense. They tell us the relative effect of each variable on warfarin dose prediction. The variables age, weight, and race all have a tiny effect on warfarin dose, whereas VKORC1 genotypes at this SNP have a large (~2X per G allele present) effect on warfarin dose. This is consistent with the relationships seen in previous plots.

- b. Create a new column in the spreadsheet named "predicted_dose." Enter a formula into each cell of this column that multiplies the value of each independent variable by its coefficient. For example, the formula for just the first two independent variables is " $=-0.639*B2+0.038*C2$ ". Make sure you use all the independent variables listed in the table above in your formula.
- c. Make a plot of the actual warfarin dose to the predicted warfarin dose. Paste the plot below and describe the relationship between the actual dose and predicted dose.



This plot predicts actual warfarin dose much better than the previous plots analyzed in part 1. The R^2 value for VKORC1 genotype alone was 0.2388, compared to this correlation value of 0.3963, indicating that including other clinical measures is helpful for predicting warfarin dose, despite the fact that alone they are not very useful. Overall the predictive dose appears to underestimate the actual dose.

- d. Compare and contrast the plot you made in (c) to those you made in (1) and (2).

The (C) graph has by far the best correlation between values, suggesting that overall when the data is taken together we have a relatively good predictive power of warfarin dose. This is because the predicted dose takes each variable into account while also weighing the effect. Taking all the variables into account produces the most complete picture. Using multiple variables allows a much greater range of predicted doses, matching the wide range of actual doses.

- 4) We have just completed our first quantitative GWAS*! We did not go through how to compute the p-values for this analysis, but you'll need to know that the p-value for the association between VKORC1 and warfarin dose in the multivariate linear regression is $8.45e-14$. Now we are going to compare quantitative GWAS to a case/control GWAS.
- a. Create a new column called "discrete_dose" which contains a "TRUE" if the warfarin dose is greater than 5 and "FALSE" if the warfarin dose is less than or equal to 5.
 - b. Using this new column complete the following contingency table (note the similarity to the tables you've made in previous GWAS analysis).

Observed:	AA	AG	GG
TRUE	0	24	40
FALSE	20	51	26

Expected:

	AA	AG	GG
TRUE (dose > 5)	7.95031	29.81366	26.23602
FALSE (dose <= 5)	12.04969	45.18634	39.76398

- c. Compute the chi-squared statistic using the same model (recessive, semi-dominant or dominant) that you used in 2b. Report the p-value (you can use the same websites we used in class or R). How does this compare to the p-value for VKORC1 in the quantitative GWAS? What can you say about quantitative GWAS versus case/control GWAS? Explain.

For this question, you could use either a recessive model (combining AG and GG genotypes) or a semi-dominant model.

Semi-dominant model:

The chi-squared value is 27.063 with 2 degrees of freedom and the p-value is 1.329e-06. [Used R: `chisq.test(matrix(c(0,24,40,20,51,26), nrow=3), correct=FALSE)`].

Recessive model:

Observed:	AA	AG/GG
TRUE	0	64
FALSE	20	77

The chi-squared value is 15.068 with 1 degree of freedom and the p-value is 1.037e-04. [Used R: `chisq.test(matrix(c(0,64,20,77), nrow=2), correct=FALSE)`].

This is a bigger p-value than the p-value from the quantitative GWAS (8.45e-14), which suggests that the data is less significant. This is expected since a warfarin dose can be very close to 5 or very far away from 5. In contrast, when we are predicting the warfarin dose with the quantitative regression model, we expect that our prediction is closer to the actual dose.

*The astute observer will notice that there was nothing “genome-wide” about this. We actually performed 1/500,000th of a typical GWAS. :)