Genetics 210
Problem Set 2
Due: April 18, 2011

Part I. 1) To put our GWAS results in a personal genome context, it is typically much more convenient to consider the effect of diploid genotypes rather than alleles. Using this approach, our Type 2 Diabetes GWAS finds a new variant, rs9914, found in the *SWEET* gene. You would like to find out if this SNP is significantly associated with the disease, which has G and T alleles. The all-knowing Francis Collins has told you that G is the dominant allele. Given 6000 controls (600 GG, 2580 GT, 2820 TT) and 2000 diabetics (270 GG, 930 GT, 800 TT), calculate the χ^2 statistic (and p-value) for the association (using a model in which G dominant).

2) Calculate the odds ratio for this SNP for Type 2 Diabetes in GG/GT individuals, compared to TT individuals.

3) Calculate the odds ratio* (assuming prevalence of 25% as in the individuals in the GWAS) for GG/GT individuals and then for TT individuals. Why is this a more relevant statistic?

4) Calculate the likelihood ratio for T2D conferred by this SNP for GG/GT individuals.

5) Calculate the percent variance of the trait explained by rs9914.

Part II. We have learned a lot about GWAS and even performed quite a few of them on our own. In each of the 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). But phenotypes are not always so 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 if 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/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 Fullback, Owen Marecic (6' 1", 244lbs) 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 of plot with warfarin_dose on the y axis and the clinical variable on the x axis. Paste in the plot and under each bullet point and describe the relationship (i.e. do they appear correlated?). Note: You may use a different type of plot if you believe it will be more informative/appropriate.
    a. warfarin dose versus sex:
    b. warfarin dose versus age:
    c. warfarin dose versus weight:
    d. warfarin dose versus race:
2) One SNP in the gene, VKORC1, is the most significant genetic covariate known for warfarin dose. Make another scatter plot, except now plot warfarin dose versus VKORC1 genotype (Note: you may have to transform the genotypes to arbitrary numerical values depending on your version of Excel).
    a. Paste in the plot and describe the relationship.
    b. Assume this SNP is the causative variant in VKORC1. Can you say anything about whether this trait is recessive or dominant? Explain.

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 preformed 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 ;)

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.

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

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 **(dose > 5)**			
FALSE **(dose <= 5)**			

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 and report the p-value (you can use the same websites we used in class). 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.

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