Comments logistic regression exercise

Remember, the notation for odds, $odds(Y = 1 | X = x)$, means $P(Y = 1 | X = x) / P(Y = 0 | X = x)$.

Models for the M3-marker

You used both a multiplicative model (a dose-effect type of model) and a genotype-model for SNP 3, coded with the variable M3.

Multiplicative model

Model:

$$\log\left(\frac{P(Y = 1 \mid M3)}{P(Y = 0 \mid M3)}\right) = \alpha + \beta M3,$$

where M3 is treated as a continuous (numeric) variable.

Output from R:

```
                Estimate Std.Error   z value Pr(>|z|)
(Intercept)     2.2063   0.3623      6.09    1.13e-09 ***
M3              -0.6986  0.2085      -3.35   0.000808 ***
```

The p-values shown in the logistic regression outputs are based on test-variables of the form

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})} \text{ which is approximately N(0,1) under H0.}$$

$H_0 : \beta = 0$, is tested against the two-sided alternative hypothesis $H_1 : \beta \neq 0$. For the M3 effect parameter we have that the p-value= P(Z<-3.35)+ P(Z>3.35) = 2*P(Z>3.35) = 0.0008.

An approximate 95% confidence interval for β is given by: $\hat{\beta} \pm 1.96 * SE(\hat{\beta})$. Here we get $\hat{\beta} \in$ (-1.1 , -0.28), and a 95% confidence interval of the OR estimate is then $(\exp(-1.1), \exp(-0.28))$ = (0.33 , 0.75).

Note that you can draw conclusions about parameter significance directly from confidence interval information (without performing any test): For instance, $\widehat{OR} \in (0.33 , 0.75)$, tells us that OR is significantly different from 1, with significance level = 1 - confidence level, =0.05.

OR interpretation: The estimated OR is $\exp(\hat{\beta}) = 0.5$. This means that the odds for being a case decreases with a factor 0.5 for every '1'-allele you have, $\Longleftrightarrow$ the odds for being a case increases with a factor 2 for every '2'-allele you have.

The multiplicative model assumes that

$$\exp(\beta) = \frac{odds(Y = 1|M3 = 1)}{odds(Y = 1|M3 = 0)} = \frac{odds(Y = 1|M3 = 2)}{odds(Y = 1|M3 = 1)},$$

$$\text{and therefore,} \quad \frac{odds(Y = 1|M3 = 2)}{odds(Y = 1|M3 = 0)} = \left(\frac{odds(Y = 1|M3 = 2)}{odds(Y = 1|M3 = 1)}\right)^2 = \exp(2\beta).$$

Genotype model

In a genotype model M3 is regarded as a categorical variable, where the three categories represent the genotypes 22, 12, and 11, which here are denoted 0, 1 and 2 (but could just as well been denoted A, B, and C, for instance). Using R, the lowest of the M3 levels will be used as reference category by default, and dummy variables are automatically constructed when using the as.factor()- function.

Model:

$$\log\left(\frac{P(Y = 1 \mid M3)}{P(Y = 0 \mid M3)}\right) = \alpha + \beta_1(M3)1 + \beta_2(M3)2,$$

where (M3)1 and (M3)2 denote the dummy variables.

Output from R:

```
                   Estimate Std.Error   z value Pr(>|z|)
(Intercept)        1.8458   0.6213      2.971   0.00297 **
as.factor(M3)2     -0.2715  0.6539      -0.415  0.67802
as.factor(M3)3     -1.0509  0.6350      -1.655  0.09794

....
anova()-output:

Resid.       Df Resid.  Dev Df Deviance P(>|Chi|)
1            460520.11
2            458507.26      2  12.85    0.001618
```

The p-values in the first part of output above refer to the tests of $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$, respectively, while the hypothesis you tested by using the anova()-function is $H_0 : \beta_1 = \beta_2 = 0$. The latter test is the one to use to investigate the marker association. (We are usually not interessted in the separate genotype associations at that stage of a fine mapping study).

The interpretation of the parameters in the genotype-model:

$$\exp(\beta_1) = \frac{odds(Y = 1|M3 = 1)}{odds(Y = 1|M3 = 0)}, \text{ and } \exp(\beta_2) = \frac{odds(Y = 1|M3 = 2)}{odds(Y = 1|M3 = 0)}.$$

Hence, using the genotype model the odds for being a case is estimated to be a factor $e^{-0.27} = 0.76$ lower for people with genotype 12, and a factor $e^{-1.05} = 0.35$ lower for people with genotype 11, as compared to genotype 22.

Which model is the best?

The multiplicative model uses only one parameter to model the relationship between SNP and disease, but relies on the assumption that the alleles act multiplicatively on the association (as measured by OR). The genotype model spends two parameters on the disease association, which makes it more flexible than the multiplicative model, but also less powerful. Therefore, if the multiplicative effect-assumption is reasonable, the multiplicative model is the one to use.

Controlling for known risk factor

The marker called DQDR is really made up of three different genes. It is situated in the HLA (Human Leucocyte Antigene) complex, which is a region about 3.5 Mb long on the short arm of chromosome 6. This complex contains numerous genes and many of them are involved in the immune system. The three genes DQDR is constructed from are HLA-DRB1, -DQA1 and -DQB1, three genes in extremely tight LD - almost always inherited in one piece. The alleles of DQDR are thus really haplotypes, different combinations of the alleles at DRB1, DQA1, and DQB1.

In type 1 diabetes the genes HLA-DRB1, -DQA1 and -DQB1 are well known susceptibility genes. This means that when you are searching for other disease genes in the HLA complex, the known risk genes must be taken into consideration. Markers that happen to be associated with the DQDR will tend to show association with diabetes, weather or not they are causal themselves (or in LD with other causal genes).

Logistic regression provides a mean to test the effect of a certain marker while controlling for the effect of others. When you included the M3-marker into the model containing the D-variables (representing the DQDR-alleles) its effect on Y almost completely vanished:

```
             Estimate Std.Error z value Pr(>|z|)
(Intercept)  0.51807  0.89092   0.582   0.560903
M3           -0.07429 0.29202   -0.254  0.799175
D6           -0.92618 0.41412   -2.236  0.025320 *
D7           -0.84171 0.45122   -1.865  0.062125 .
D9           1.02792  0.36012   2.854   0.004312 **
D10          0.14615  0.49634   0.294   0.768407
D11          0.99073  0.28432   3.485   0.000493 ***
D12          0.50879  0.42322   1.202   0.229288
D13          -0.01108 0.70226   -0.016  0.987411
D14          -0.26926 0.58860   -0.457  0.647343
D15          -1.82205 0.52797   -3.451  0.000558 ***
```

The estimate of the OR for the M3 marker changes from 0.5, (95% CI =(0.33, 0.75)), based on the model containing only M3, to exp(−0.0743) =0.93, (95% CI=(0.52, 1.65)), when the effect of the DQDR on Y is taken into account.