BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Homework 4 (version 3) - posted October 3

Assigned October 2; Due 11:59PM October 9

Problem 1 (Easy)

a. For the genetic regression model:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \tag{1}$$

$$\epsilon \sim N(0, \sigma^2_\epsilon) \tag{2}$$

prove that the $\mathrm{E}(Y|X_a = x_a, X_d = x_d)$ (i.e. the expected value of the phenotype for an individual with a specific genotype: x_a, x_d) is entirely a function of the genotype and the β parameters.

$\mathrm{E}(Y|X_a = x_a, X_d = x_d)$=$\mathrm{E}(\beta_\mu + x_a\beta_a + x_d\beta_d + \epsilon)$

=$\mathrm{E}(\beta_\mu)$ + $\mathrm{E}(x_a\beta_a)$ + $\mathrm{E}(x_d\beta_d)$ + $\mathrm{E}(\epsilon)$

Note that $\mathrm{E}(\epsilon)$ is the expected value of a normal distribution with parameter $\mu = 0$ and since $\mathrm{E}(\epsilon) = \mu$:

=$\mathrm{E}(\beta_\mu)$ + $\mathrm{E}(x_a\beta_a)$ + $\mathrm{E}(x_d\beta_d)$

Also note that with the x_a and x_d fixed, these are constants, as are all the β's and since the expectation of a constant is equal to the constant, i.e. $\mathrm{E}(c) = c$:

$= \beta_\mu + x_a\beta_a + x_d\beta_d$

and since the x_a and x_d are a function of the genotype, this proves the assertion.

b. Write down parameter values for the genetic regression model that would produce a case of 'pure' dominance with Mendelian genetic inheritance (=a single locus where genotype exactly predicts phenotypic value).

Note that the dominance case asked for here was a case where the expected value of the phenotype for the heterozygote and one of the homozygotes is the same i.e. $\mathrm{E}(Y|A_1A_1) = \mathrm{E}(Y|A_1A_2)$ or $\mathrm{E}(Y|A_2A_2) = \mathrm{E}(Y|A_1A_2)$ which I have called as 'pure dominance' in class, however, from discussions with students I have realized that 'classic dominance' is the better

term for such a case.

For a Mendelian, it needs to be the case that the observed phenotype Y for every individual with a given genotype A_iA_j needs to be the same value, i.e. $Y|A_iA_j = \mathrm{E}(Y|A_iA_j)$. Starting with the model

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \tag{3}$$

$$\epsilon \sim N(0, \sigma^2_\epsilon) \tag{4}$$

this means that $\sigma^2_\epsilon = 0$, such that $\epsilon = 0$ for every individual. To make this a 'pure' (='classic') case, the value of β_μ and the other two β parameters must have the following relationship $\beta_a = \frac{1}{2}\beta_d$ or $\beta_a = -\frac{1}{2}\beta_d$.

Problem 2 (Medium)

For the following question, you are going to simulate and then analyze GWAS data, where these data will include one normally distributed phenotype and $N = 200$ diploid genotypes measured for each of $n = 100$ samples. For many parts of this question, the answer will be your R code presented in a text file that is easy for us to run to produce the requested output (NOTE THAT FOR FULL CREDIT = easy to run code in a txt file, name your files appropriately, send a zip file, and do not copy someone else's code!! etc.).

a. Write R code to create a 100 x 400 matrix of sample genotypes (i.e. simulate 200 genotypes for each of the 100 individuals, where each genotype is represented by two columns). For each genotype, there should be two alleles (you must represent these with two of 'a', 'c', 't', 'g' OR as 'A1' and 'A2') where all three genotype states (two homozygotes and a heterozygote) should be present for each genotype. The minor allele frequency for each genotype is up to you but should not be < 0.2 for any of the genotypes. Also, each genotype should not be highly correlated with other genotypes, (e.g. don't just repeat the same genotype 200 times!).

b. Write R code to convert your genotype matrix into two new matrices, the first a 100 x 200 matrix where each genotype is converted to the appropriate X_a value and the second a 100 x 200 matrix where each genotype is converted to the appropriate X_d value.

c. Write R code to simulate a 100 x 1 vector of Y values for the 100 individuals where we are assuming there are only two causal genotypes (genotypes number 25 and 150 of your simulated set). Use the following equation to simulate the phenotypes:

$$y_i = \beta_\mu + x_{i,a,25}\beta_{a,1} + x_{i,d,25}\beta_{d,1} + x_{i,a,150}\beta_{a,2} + x_{i,d,150}\beta_{d,2} + \epsilon_i \tag{5}$$

$$\epsilon \sim N(0, \sigma^2_\epsilon) \tag{6}$$

where the $x_{i,a,25}$ and $x_{i,d,25}$ are the random variable codings for the 25th genotype of individual i and the $x_{i,a,150}$ and $x_{i,d,150}$ are the random variable codings for the 150th genotype. Assume the following true parameter values (in real situations, these will be unknown to you!!): $\beta_\mu = 1, \beta_{a,1} = 1.0, \beta_{d,1} = 0, \beta_{a,2} = 0.5, \beta_{d,2} = 0.5, \sigma^2_\epsilon = 1$.

d. Using one sentence, explain why can we simulate the phenotypes in c. by considering only the two causal genotypes and ignoring all others?

e. For each genotype individually, write R code to calculate the $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$. Plot three x-y graphs with the genotypes in order on the x-axis and the parameter estimates on the y-axis = one graph for each of the three parameters.

f. For each genotype individually, write R code to calculate an F-statistic. Plot a histogram of the F-statistic values you obtained.

g. For each genotype individually, calculate a p-value based on your F-statistic and produce a Manhattan plot.

h. Your Manhattan plot will not look like the empirical Manhattan plots we have discussed in class. Using no more than two sentences, describe what is different about your Manhattan plot and what explains why it looks different?

i. Unless you got incredibly unlucky... the two causal genotypes (25 and 150) resulted in the lowest p-values overall (=the highest -log p-values on your Manhattan plot). If you had used a type 1 error $\alpha = 0.05$ to decide whether these were significant or not, would you have rejected the null for these two genotypes? How many other genotypes would you have rejected the null in your analysis? Using no more than two sentences, explain why it is not that surprising that you had additional (incorrect!) cases where you rejected the null at this type 1 error level?

j. For your two causal SNPs (25 and 150) produce two x-y plots for each (=four plots total) with the X_a values on the x-axis and the phenotype Y on the y-axis for the first, and X_d values on the x-axis and the phenotype Y on the y-axis for the second.

Problem 3 (Difficult)

When using a genetic linear regression to model the family of probability distributions that could describe the relationship between phenotypes and genotypes $Pr(Y, X_a, X_d)$, the following are three equivalent statements of the null and alternative hypotheses when assessing whether a genotype is causal:

$$H_0 : Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d) \tag{7}$$

$$H_A : Pr(Y, X_a, X_d) \neq Pr(Y)Pr(X_a, X_d) \tag{8}$$

$$H_0 : Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0 \tag{9}$$

$$H_A : Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \tag{10}$$

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \tag{11}$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \tag{12}$$

Prove that these are equivalent. Note that to get full credit, you must show the steps of each derivation and provide a justification for each step (i.e. what rule(s) you are using).

The following may be helpful (!!): note that for arbitrary random variables X_1, X_2, and X_3 that $Cov(X_1, X_2 + X_3) = Cov(X_1, X_2) + Cov(X_1, X_3)$, that for a linear regression, the random variable ϵ is assumed to be independent of the variable(s) X such that $Pr(X_a, \epsilon) = Pr(X_a)Pr(\epsilon)$ and $Pr(X_d, \epsilon) = Pr(X_d)Pr(\epsilon)$, that for the genetic linear regression $Pr(Y|X) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d$, $\sigma^2_\epsilon)$, and that for $Pr(X_a, X_d)$, only three combinations of X_a and X_d have non-zero probabilities.

The underlying objective of this question is to make the following point about conditional probabilities and covariances in general and the relationship of these two concepts for the model we use in genetics. For any arbitrary random variables (in any study!) X_1, X_2 (assume any probability distribution is possible!) it is the case that if $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2) \Rightarrow Cov(X_1, X_2) = 0$, i.e. if the random variables are conditionally independent, this implies / means that the random variables will have a covariance of zero. However, if $Cov(X_1, X_2) = 0$ this does not necessarily mean that $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2)$, i.e. it may but it depends on the true probability distribution! Interestingly, it IS the case that if $Cov(X_1, X_2) \neq 0 \Rightarrow Pr(X_1 \cap X_2) \neq Pr(X_1)Pr(X_2)$ but again, not vice versa!

Now, it should be clear that it would be more ideal if $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2) \Leftrightarrow Cov(X_1, X_2) = 0$ AND $Pr(X_1 \cap X_2) \neq Pr(X_1)Pr(X_2) \Leftrightarrow Cov(X_1, X_2) \neq 0$, i.e. one implies the other. The reason is that while we are actually interested in the true $Pr(X_1 \cap X_2)$ this is difficult to estimate directly from data. We therefore employ a summary function (like covariance!) that captures the critical information that we would like to know about $Pr(X_1 \cap X_2)$ pretty well. Now covariance is a function that captures this pretty well, but not perfectly, since a covariance of zero does not necessarily mean that $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2)$ (what we want to know!) although it does map perfectly for many distributions and cases, i.e. the mapping is not perfect, but it is pretty good for many cases (note that there are functions that do map perfectly such as 'distance correlation' but these have other downsides = there are always trade-offs!).

Now, here is the good news: for the genetic regression model it is the case that $Pr(X_1 \cap X_2) = Pr(X_1)Pr(X_2) \Leftrightarrow Cov(X_1, X_2) = 0$ and with one realistic exception (which I remembered while writing out the proof - we'll see this example again in our lecture on additive genetic variances) it is the case that $Pr(X_1 \cap X_2) \neq Pr(X_1)Pr(X_2) \Leftrightarrow Cov(X_1, X_2) \neq 0$ (that is, this relationship is almost always true for the genetic model, but not always!!). The goal of the question was to have you prove that this is the case.

To start, I'm going to outline how I originally envisioned proving this, and then I will outline ideas for slightly simpler proofs, some of which depend on information that I may have given you / said you can assume. That is, what I will outline first provides the general rigorous proof, where the following will be acceptable answers under various levels of rigor / assumptions I said you could make. Note that since I have given out different amounts of information on what you can assume, since one of the proofs did not have a particularly simple version (I thought there was one and realized later there was not...), and since there is actually an exception for the alternative hypothesis (!!) in the case of covariance for the genetic model (=one of the assertions I made within the question was not true), you will get full credit for this problem if you successfully proved any of the relationships at any reasonable level of rigor.

RIGOROUS COMPLETE PROOF: To start, it will actually be easiest to start with the genetic regression model and show when the other two relationships hold, separately for the null and alternative:

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \tag{13}$$

$$H_A : \beta_a \neq 0 \cup \beta_d \neq 0 \tag{14}$$

Let's therefore start with H_0 for this model and prove that $[\beta_a = 0 \cap \beta_d = 0] \Rightarrow [Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0]$. We will start by doing this for X_a alone, where the same approach is used for X_d. We therefore want to show $Cov(X_a, Y) = 0$, which we can do using the following steps

$$Cov(X_a, Y) = Cov(X_a, \beta_\mu + \epsilon) \tag{15}$$

substitution of the formula for Y with $\beta_a = 0$ and $\beta_d = 0$

$$Cov(X_a, Y) = Cov(X_a, \beta_\mu) + Cov(X_a, \epsilon) \tag{16}$$

rule for algebra of covariances, given to you in the hint,

$$Cov(X_a, Y) = Cov(X_a, \beta_\mu) \tag{17}$$

since the random variable ϵ is uncorrelated with the X_a (or X_d) by the assumptions of the linear regression model and the covariance of uncorrelated random variables is zero, i.e. $Cov(X_a, \epsilon) = 0$

$$Cov(X_a, Y) = \mathrm{E}X_a\beta_\mu - \mathrm{E}X_a\mathrm{E}\beta_\mu \tag{18}$$

by the definition of covariance

$$Cov(X_a, Y) = \beta_\mu \mathrm{E}X_a - \beta_\mu \mathrm{E}X_a \tag{19}$$

rule for the algebra of expectations where the expectation of a constant times a random variable is constant times expected value of random variable, i.e. the covariance of a random variable and a constant is zero.

$$Cov(X_a, Y) = 0 \tag{20}$$

and since we have the same for $Cov(X_d, Y) = 0$ this completes the proof for the null.

For the relationships of the alternative, use the same approach:

$$Cov(X_a, Y) = Cov(X_a, \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon) \tag{21}$$

substitution of the formula for Y,

$$Cov(X_a, Y) = Cov(X_a, \beta_\mu) + Cov(X_a, X_a\beta_a) + Cov(X_a, X_d\beta_d) + Cov(X_a, \epsilon) \tag{22}$$

rule for algebra of covariances,

$$Cov(X_a, Y) = Cov(X_a, X_a\beta_a) + Cov(X_a, X_d\beta_d) \tag{23}$$

since the random variable ϵ is uncorrelated with the X_a (or X_d) and the covariance of a random variable and a constant is zero,

$$Cov(X_a, Y) = \beta_a Cov(X_a, X_a) + \beta_d Cov(X_a, X_d) \tag{24}$$

rule for the algebra of covariances where $Cov(c_1X_1, c_2X_2) = c_1c_2Cov(X_1, X_2)$ for constants c_1, c_2,

$$Cov(X_a, Y) = \beta_a Var(X_a) + \beta_d Cov(X_a, X_d) \tag{25}$$

since the covariance of a random variable with itself is equal to the variance. Now, note that the only way for the variance of X_a to be zero is the genotype in the sample is the same for every individual (e.g. every genotype would have to be A_1A_1), which we can assume is not the case, since we are considering segregating causal polymorphisms. Since we get the same result for $Cov(X_d, Y)$, under almost every assignment of β_a and β_d, it will be the case that:

$$Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0 \tag{26}$$

However, there is one exception. If it is the case that both β_a and β_d are not zero and that $Cov(X_a, X_d)$ is not zero, if the following relationships holds, then the $Cov(X_a, Y) = 0$ and $Cov(X_d, Y) = 0$ even when $\beta_a \neq 0$ and $\beta_d \neq 0$:

$$Cov(X_a, X_d)\beta_d = Var(X_a)\beta_a \cap Cov(X_a, X_d)\beta_d = Var(X_d)\beta_d \tag{27}$$

That is, under a very specific combination of genotype frequencies and parameter values, what we are trying to prove does not hold! Now, this does require a very specific set of conditions but it could happen (!!) at least in theory. Note that the conditions are so restrictive, the chances that it would ever apply in the system you are working is so small, it can be effectively ignored.

To complete this section, note that by the relationships:

$$Cov(X_a, Y) = \beta_a Var(X_a) + \beta_d Cov(X_a, X_d) \tag{28}$$

$$Cov(X_d, Y) = \beta_a Cov(X_a, X_d) + \beta_d Var(X_d) \tag{29}$$

for the relationship $Cov(X_a, Y) \neq 0 \cup Cov(X_d, Y) \neq 0$ to hold, it must be the case that $\beta_a \neq 0 \cup \beta_d \neq 0$ since each of the terms in equations (28-29) include a β_a or β_d term.

To summarize, we have now shown the following relationships:

$$[\beta_a = 0 \cap \beta_d = 0] \Rightarrow [Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0] \tag{30}$$

$$[\beta_a \neq 0 \cup \beta_d \neq 0] \nRightarrow [Cov(X_a, Y) = 0 \cup Cov(X_d, Y) \neq 0] \tag{31}$$

although very strict conditions have to be true for $\nRightarrow$ under the genetic model (and when these do not hold the relationship is $\Rightarrow$).

$$[Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0] \nRightarrow [\beta_a = 0 \cap \beta_d = 0] \tag{32}$$

$$[Cov(X_a, Y) = 0 \cup Cov(X_d, Y) \neq 0] \Rightarrow [\beta_a \neq 0 \cup \beta_d \neq 0] \tag{33}$$

where again, very strict conditions have to be true for $\nRightarrow$ under the genetic model (and when these do not hold the relationship is $\Rightarrow$).

Next, again let's start with H_0 for this model and prove that $[\beta_a = 0 \cap \beta_d = 0] \Rightarrow [Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d)]$. In this case, we would like to use the following definition of independence $Pr(Y|X_a, X_d) = Pr(Y)$, such that we are going to make use of the formula for conditional probability:

$$Pr(Y|X_a, X_d) = Pr(Y|X_a \cap X_d) = \frac{Pr(Y \cap X_a \cap X_d)}{Pr(X_a \cap X_d)} \tag{34}$$

To make use of this relationship, we therefore need the joint probability distributions for $Pr(Y \cap X_a \cap X_d)$ and $Pr(X_a \cap X_d)$. For the latter, note that there are only three possible genotypes A_1A_1, A_1A_2, A_2A_2 such that there are only three possible non-zero probabilities for $Pr(X_a \cap X_d)$, which are $Pr(X_a = -1 \cap X_d = -1) \neq 0, Pr(X_a = 0 \cap X_d = 1) \neq 0, Pr(X_a = 1 \cap X_d = -1) \neq 0$. We could therefore represent this probability distribution as a multinomial or with by using indicator functions, where for the latter approach, we have:

$$Pr(X_a \cap X_d) \sim p_{I_{A_1A_1}(X_a,X_d)} + p_{I_{A_1A_2}(X_a,X_d)} + p_{I_{A_2A_2}(X_a,X_d)} \tag{35}$$

where the $I_{A_iA_j}(X_a, X_d)$ are indicator functions that take in the X_a and X_d values of an individual and the output would take the value one when these values matched the genotype in the subscript and zero if not, e.g. for an individual with genotype A_1A_1 which therefore has $X_a = -1, X_d = -1$, the input to will be $(-1, -1)$ so we have $I_{A_1A_1}(-1, -1) = 1$ and $I_{A_1A_2}(-1, -1) = 0$ and $I_{A_2A_2}(-1, -1) = 0$. Note that $p_{I_{A_2A_2}(X_a,X_d)} = 1 - p_{I_{A_1A_1}(X_a,X_d)} + p_{I_{A_1A_2}(X_a,X_d)}$ and that this is a probability mass function.

For the joint probability distribution $Pr(Y \cap X_a \cap X_d)$, remember that $Pr(Y) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d, \sigma^2_\epsilon)$. So, the probability will depend on what three possible (X_a, X_d) categories, so we have the following writing $p_{I_{A_iA_j}(X_a,X_d)}$ as $p_{I_{A_iA_j}}$:

$$Pr(Y \cap X_a \cap X_d) \sim p_{I_{A_1A_1}} N(\beta_\mu + -\beta_a - \beta_d, \sigma^2_\epsilon) + p_{I_{A_1A_2}} N(\beta_\mu + \beta_d, \sigma^2_\epsilon) + p_{I_{A_2A_2}} N(\beta_\mu + \beta_a - \beta_d, \sigma^2_\epsilon) \tag{36}$$

where note that $N()$ is shorthand for the normal distribution, where we have written out this equation previously. With these two equations, we can now write out the formula for the conditional probability:

$$\frac{Pr(Y \cap X_a \cap X_d)}{Pr(X_a \cap X_d)} = \frac{p_{I_{A_1A_1}} N(\beta_\mu + -\beta_a - \beta_d, \sigma^2_\epsilon) + p_{I_{A_1A_2}} N(\beta_\mu + \beta_d, \sigma^2_\epsilon) + p_{I_{A_2A_2}} N(\beta_\mu + \beta_a - \beta_d, \sigma^2_\epsilon)}{p_{I_{A_1A_1}} + p_{I_{A_1A_2}} + p_{I_{A_2A_2}}} \tag{37}$$

With this equation, we are now ready to show the relationships, starting with the H_0 for the genetic regression model. Remember, our goal is to show

$$Pr(Y|X_a, X_d) = \frac{Pr(Y \cap X_a \cap X_d)}{Pr(X_a \cap X_d)} = Pr(Y) \tag{38}$$

which will show the independence relationship $Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d)$. To start, note that for the null $\beta_a = 0$ and $\beta_d = 0$ the distribution of the variable Y regardless of genotype is:

$$Pr(Y) \sim N(\beta_\mu, \sigma^2_\epsilon) \tag{39}$$

Given that this is true regardless of genotype, we have the following joint distribution:

$$Pr(Y|X_a, X_d) = \frac{p_{I_{A_1A_1}} N(\beta_\mu, \sigma^2_\epsilon) + p_{I_{A_1A_2}} N(\beta_\mu, \sigma^2_\epsilon) + p_{I_{A_2A_2}} N(\beta_\mu, \sigma^2_\epsilon)}{p_{I_{A_1A_1}} + p_{I_{A_1A_2}} + p_{I_{A_2A_2}}} \tag{40}$$

$$Pr(Y|X_a, X_d) = \frac{(p_{I_{A_1A_1}} + p_{I_{A_1A_2}} + p_{I_{A_2A_2}}) N(\beta_\mu, \sigma^2_\epsilon)}{p_{I_{A_1A_1}} + p_{I_{A_1A_2}} + p_{I_{A_2A_2}}} \tag{41}$$

$$Pr(Y|X_a, X_d) = N(\beta_\mu, \sigma^2_\epsilon) = Pr(Y) \tag{42}$$

so we have proved the result for the null. To prove the alternative, note that we cannot reduce:

$$Pr(Y|X_a, X_d) = \frac{p_{I_{A_1A_1}} N(\beta_\mu + -\beta_a - \beta_d, \sigma^2_\epsilon) + p_{I_{A_1A_2}} N(\beta_\mu + \beta_d, \sigma^2_\epsilon) + p_{I_{A_2A_2}} N(\beta_\mu + \beta_a - \beta_d, \sigma^2_\epsilon)}{p_{I_{A_1A_1}} + p_{I_{A_1A_2}} + p_{I_{A_2A_2}}} \tag{43}$$

any further, where taking taking the conditional probability of any of the three genotypes does not return $Pr(Y)$. To see this, note that the distribution of Y under the alternative is:

$$Pr(Y) \sim N(\beta_\mu + X_a\beta_a + X_d\beta_d, \sigma^2_\epsilon) \tag{44}$$

so if we were consider one of the genotypes, say A_1A_1, we have:

$$Pr(Y|X_a = -1, X_d = 1) = \frac{p_{I_{A_1A_1}} N(\beta_\mu + -\beta_a - \beta_d, \sigma^2_\epsilon)}{p_{I_{A_1A_1}}} = N(\beta_\mu + -\beta_a - \beta_d, \sigma^2_\epsilon) \tag{45}$$

which does not have the same form as equation (44). Note that in this case (not usually the case!!) the marginal distribution $Pr(Y)$ has the same formula as the joint distribution $Pr(Y \cap X_a \cap X_d)$ in equation (36) (i.e. equation (36) is a way to write out equation (44)), where this is a 'mixture' distribution, i.e. it combines discrete and continuous distributions.

Also, note in contrast to the proof of the alternative in the covariances, we do have the following relationships (!!): $[\beta_a = 0 \cap \beta_d = 0] \Leftrightarrow [Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d)]$ and $[\beta_a \neq 0 \cap \beta_d \neq 0] \Leftrightarrow [Pr(Y, X_a, X_d) \neq Pr(Y)Pr(X_a, X_d)]$, where you can show the reverse using the same equations that we used to show the forward relationships in the same way.

OTHER ACCEPTABLE PROOFS (given what I have said you can assume):

Given that I effectively asserted that EACH of the 3 sets of relationships stated in the beginning implied the other (not correct for the covariance cases! see above), if you used the fact that $Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d) \Rightarrow [Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0]$ and proved that $[\beta_a = 0 \cap \beta_d = 0] \Rightarrow [Cov(X_a, Y) = 0 \cap Cov(X_d, Y) = 0]$ under my incorrect assertion your proof for the null would be complete.

Another simpler proof of $[\beta_a = 0 \cap \beta_d = 0] \Rightarrow [Pr(Y, X_a, X_d) = Pr(Y)Pr(X_a, X_d)]$ is simply to show that the formula for the pdf of Y under the null $Pr(Y) \sim N(\beta_\mu, \sigma^2_\epsilon)$, is not a function of X_a or X_d, such that the $Pr(Y)$ must be independent of $Pr(X_a, X_d)$. Although slightly less rigorous, you could also show that the $Pr(Y)$ under the alternative changed depending on the state of X_a or X_d, which by definition, would mean that there is a conditional relationship, such that $[\beta_a \neq 0 \cup \beta_d \neq 0] \Rightarrow [Pr(Y, X_a, X_d) \neq Pr(Y)Pr(X_a, X_d)]$

Many other acceptable answers...