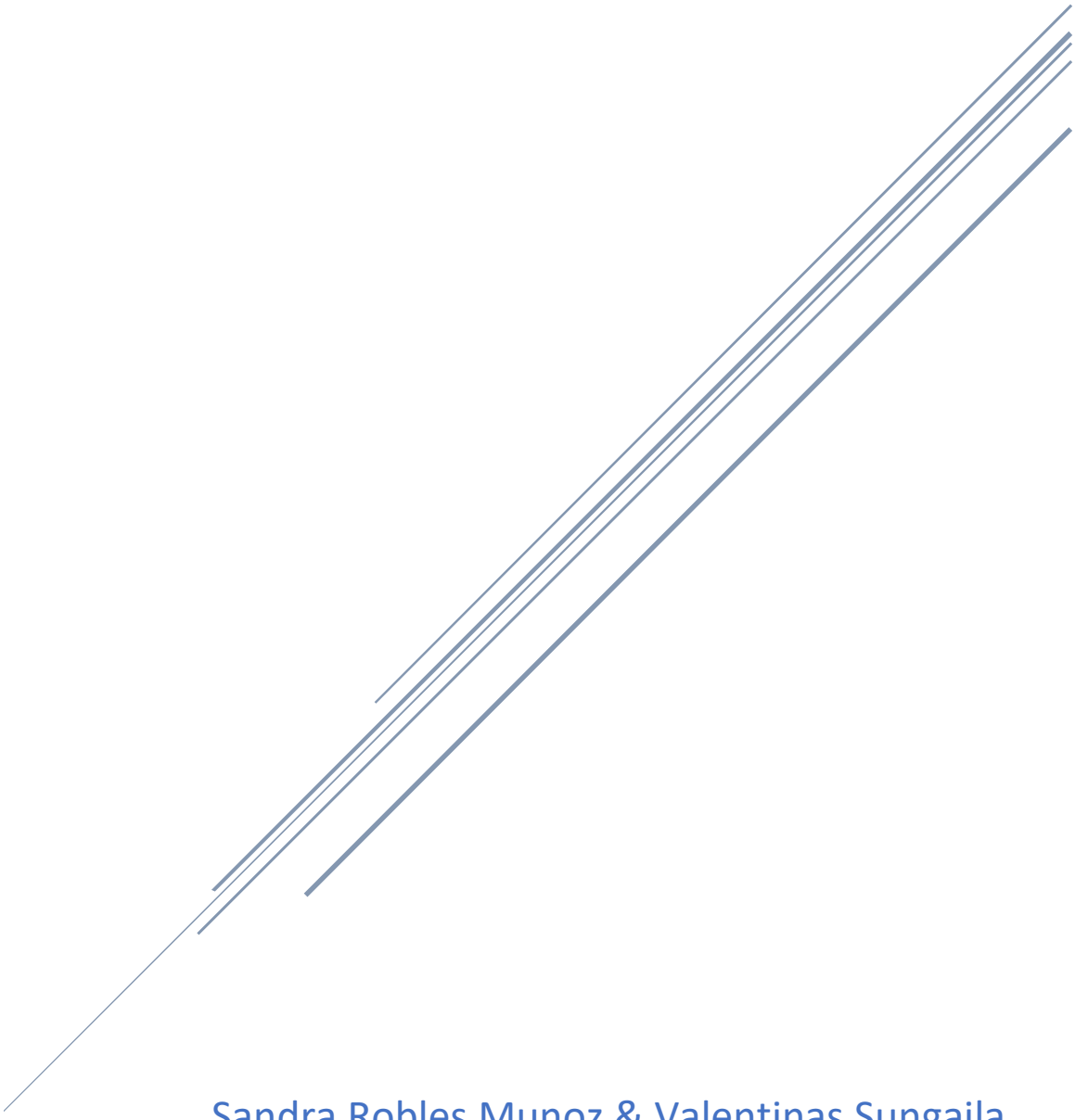
VITILIGO DISEASE CHARACTERISTICS AND GENETIC VARIANTS



Sandra Robles Munoz & Valentinas Sungaila

MATH 6330 – Spring 2020

I. Introduction

Vitiligo is a complex chronic autoimmune disease that causes patches of skin and hair to lose pigmentation, with a population prevalence of anywhere from 0.001 to 0.002 (Santorico, 2018). The disease occurs when melanocytes stop producing melanin, with the degree of coverage varying from covering a few areas of the body to possibly covering all skin surfaces (MayoClinic, 2020). People suffering from vitiligo can be susceptible to social or psychological stress due to the disease, as well as other complications (MayoClinic, 2020), so studying how the disease ties to genetics can help the medical community come up with better treatments for the disease.

The University of Colorado organized the *International VitGene Consortium* to try to understand the genetic makeup of vitiligo, with "the long-term goal of developing novel treatments that suppress or re-regulate the autoimmune process, enhancing treatments that stimulate skin re-pigmentation by melanocyte re-population" (Spritz, 2008). Through this consortium, a multi-stage genetic-wide association study was carried out, gathering approximately 2,800 cases and 37,000 controls (Santorico, 2018). The first set of analyses carried out with this data showed 48 genetic variants, also called SNPs, being statistically associated with the risk of developing vitiligo, with individual effect sizes being fairly small.

Our client, Doctor Stephanie Santorico, approached us seeking to know whether these pre-identified genetic variants are also associated with the disease characteristics of vitiligo. The main characteristics of note where total percentage of skin depigmentation, and age of onset of the disease. These two secondary phenotypes were gathered during the consortium and provided to us alongside 46 of the 48 identified genetic variants and four risk scores from the first set of analyses. A further question posed was to identify the power of the associations found, with the

end goal of defining what effect sizes are necessary to ascertain 80% power. Finally, the client also stated interest in knowing whether it was possible to derive any detectable structure between the disease characteristics and the genetic variants or score values. However, this last question was placed in last priority, to be done if time permitted it.

II. Analysis and Results

DATASET

The dataset provided to us contains 2,268 total observations of people with vitiligo gathered across three study waves from the consortium. Information included in the dataset spanned a whole array of disease characteristics beyond the two to be studied, including auto immune disease information, presence of the Koebner phenomenon, halo nevi, acrofacial patterns, gender, disease risk scores, as well as genetic principal components. The individual genetic variants were measured as the total sum of minor allele counts ranging from 0 to 2, since as per our client, modeling loci like this has been shown to be the most precise and widely accepted methodology to model genetic variants. For the two disease characteristics being studied, age of onset was measured numerically, and skin coverage was measured in quantiles, with their breakdown below. Totals do not add up to the reported 2,268 names in the dataset due to missingness of some values.

		Age of onset, banded									
		0-10	10-21	21-30	31-40	41-50	51-60	61-70	71-80	80+	Totals
Skin coverage, in quantiles	up to 25%	251	283	273	250	136	84	30	2	0	1,309
	26-50%	103	97	66	62	31	17	7	0	0	383
	51-75%	37	53	29	21	11	6	2	0	0	159
	76-100%	39	38	15	13	11	6	0	0	0	122
	Totals	430	471	383	346	189	113	39	2	0	1,973

Figure 1

ACCOUNTING FOR COVARIATES

Prior to the analysis of the disease characteristics, we needed to determine if accounting for genetic covariates was necessary. The possible covariates came to us in the form of 45 distinct genetic principal components. These principal components were included by the client in the dataset to allow us to account for any confounding effects. We were also provided with biological sex, but it was stated by the client that this was known to not be a covariate, leaving us with the 45 principal components to work with. To determine if a given genetic principal component needed to be included in the models, each one was added as the sole predictor in a regression for the two disease characteristics. When building a model for age of onset, a linear regression was performed, when doing so for skin coverage, an ordinal multinomial logistic regression model was performed. Through this process, 90 regressions were constructed and tested against the following hypotheses:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

For age of onset, this entailed carrying out an ANOVA test; meanwhile skin coverage used a chi-squared test. In layman's terms, this kind of testing is trying to determine whether a given predictor is contributing any significant information to the model by being included – meaning if it has a non-zero effect in explaining the outcome.

A problem with iterating through these kinds of regressions – 45 times per chosen disease characteristic – is that Type I errors are likely. A Type I error occurs when the null hypothesis H_o is rejected when it should not be, and can be expected to occur with a frequency defined by the

threshold alpha, α, which is normally set at 0.05 (Mcleod, 2019). In the context of our data, we would expect to see, on average, by pure chance that two of the genetic principal components needed to be included as covariates when that is not the case. In order to correct for this possibility while doing multiple testing, a family-wise error correction (FWER) was applied, changing the threshold to reject the null hypothesis to $\alpha = 0.10/45$. Using this threshold, zero of the 45 genetic principal components proved statistically significant for either disease characteristic, and thus none were used in the analysis.

REGRESSIONS

The core of the analysis resulted in creating a series of linear and multinomial logistic regressions that tried to discern the relationship between the two desired disease characteristics and the 46 distinct genetic variants. This meant creating a total of 92 different regressions, half using age of onset as the outcome, and half using total skin coverage. Each genetic variant was narrowed down to non-missing observations and tested individually, as per the literature on genetics, "the de facto analysis […] is a series of single-locus [genetic variant] statistic tests, examining each SNP independently for association to the phenotype" (Bush & Moore, 2012). Since none of the genetic principal components proved to be necessary covariates, the regressions had only one predictor, generally having the equation

$$E(age\_onset) = \beta_0 + \beta_1 * genetic\_variant \quad (1)$$

$$logit[P(skincoverage \leq k)] = \log\left(\frac{P(skincoverage \leq k)}{1 - P(skincoverage \leq k)}\right) = \beta_{0,k} + \beta_1 * genetic\_variant \quad (2)$$

Tables A, B, C in the appendix provide a summary of the regressions built for the two disease characteristics. The question posed to us entailed figuring out if any of the genetic

variants could be used to explain the disease characteristics. To do this, the following hypothesis were defined for both age of onset and skin coverage:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

As with the covariate testing in the previous section, they have the same basic interpretation of, does the desired predictor explain with statistical significance the two disease characteristics? The kind of test done depends on the distribution of the alternative hypothesis. In the case of age of onset, an ANOVA test was performed, in the case of skin coverage, a likelihood ratio test was performed. The results of these tests can also be seen in Tables A, B, C in the Appendix, with tests results based on the unadjusted p-value, Bonferroni adjustment, and genome-wide adjustment. For skin coverage, as built with a multinomial logistic regression, none of the predictors were significant with a genome-wide adjustment, with '**RS1126809'** and '**RS10986311'** being significant with a Bonferroni adjustment. For age of onset, as built with a linear regression, none of the predictors were significant with a genome-wide adjustment, with '**RS114448410'** being significant with a Bonferroni adjustment.

POWER TESTING

The second question posed to us was to determine which effect sizes for the disease characteristics are needed to detect with 80% power. Power is defined as the probability of not making a Type II error, meaning of rejecting the null hypothesis when it is false (Walmsley & Brown). The way to measure effect sizes is highly dependent on the kind of question being posed as well as the distribution of the data. In the case of the vitiligo disease characteristics, we are

looking at the variance in these two characteristics explained by the SNPs, and as such, Cohen's f^2 was used to determine effect size.

	Inputs			Cohen's f2			R-squared Effect Size		
	Complete Observations	Power Desired	Effective DF	Unadjusted	Bonferroni Adjustment	Genome-wide adjustment	Unadjusted	Bonferroni Adjustment	Genome-wide adjustment
Age of onset	2,190	80%	1	0.0035889	0.00774619	0.01821105	0.0035761	0.00768665	0.01788534
Skin coverage**	1,988	80%	1	0.0039673	0.00850737	0.02007796	0.0039517	0.0084356	0.01968277

***Computed as a linear regression*

Figure 2

The power analysis computed for skin coverage was done by approximating the model as a linear regression, with skin coverage recalculated as numeric with values from 1 to 4 corresponding to each of the quantiles of total coverage. This was done as no good literature resources were found that discussed the assessment of power for categorical outcomes when considering variance explained as the effect size.

DISEASE CHARACTERISTICS STRUCTURE

In order to determine if there was any structure between the observations, the clustering algorithm *k-means* was used. We iterated through various number of centers from two thru 20. This kind of analysis lends itself to visual interpretation, by plotting the clusters created against the first two principal components. The results from our analysis can be seen in Figure 3 below.

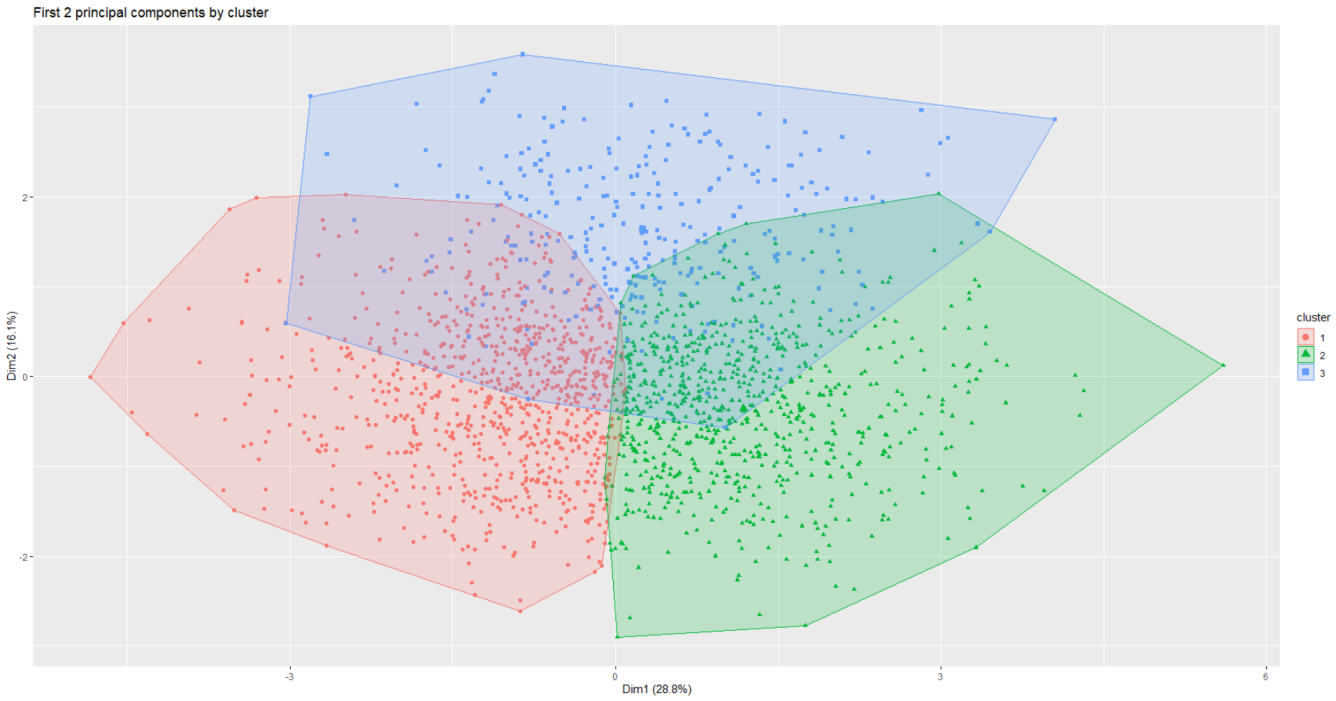


Figure 3

This plot shows that there is no strong discernible structure found among the observations when using the two disease characteristics and the risk scores as the function space.

The reason why only the risk scores were included in the clustering analysis is because they were numerical. Most clustering algorithms cannot handle categorical data, which eliminated the binary flags of associated diseases such as diabetes. Further, they also cannot handle missing values, with most of the other variables having a non-negligible proportion of its data missing.

However, it should be noted that there are other algorithms to be explored, as well as missing data imputation techniques, so there is hope still that the data might show some structure under a more thorough analysis including more of the variables available in the dataset.

III. Discussion

In prior analyses, the client had been able to see an inverse relationship between age of onset and risk scores, with no significant association between these risk scores and skin coverage (Santorico, 2018). During our analysis, we established what effect size would be needed given our sample size and a desired power of 80% at different adjustment levels, using an F-distribution to carry out the power calculations.

Based on these thresholds, when looking at age of onset, the SNP '**RS114448410'** possesses a sufficiently high effect size to have at least 80% power, with the p-value having a Bonferroni adjustment. Unfortunately, this level of power does not hold when a genome-wide adjustment is made to the p-value threshold.

For skin coverage, when treated as a numeric variable, saw the SNP '**RS10986311**' showed a sufficiently strong effect to have 80% power, however, this was only with an unadjusted p-value. Both the Bonferroni adjustment and the genome-wide adjustment did not produce any significant genetic variants in relations to skin coverage. The genetic variant '**RS1126809**' also proved statistically significant with a Bonferroni adjustment but did not have at least 80% power. It should be noted that effect sizes in relations to power were only calculated because the disease characteristic was transformed into a numeric variable. We suspect there is likely loss of information that occurs by treating skin coverage as numeric instead of as a set of ordinal categories, however there is not strong literature supporting the use of power calculations for multinomial logistic regressions. This is because the use of power analysis is dependent on the distribution of the alternate hypothesis and the availability of a variance explained calculation.

Whether these two genetic variants prove to be true determinants of the secondary phenotypes of age of onset and skin coverage is beyond the scope of this project, but based on our very primitive understanding of genomics, we do not see a strong association including in relations to data structure. However, if a bigger sample size of genetic information for people with vitiligo was obtained, it would be possible to have more power in detecting the effect of these genetic variants on the secondary phenotypes.

IV. Model and methods

When the client presented us with the data, we had observations of people with vitiligo that were explicitly of European descent. It is important to note that different nationalistic or geographical backgrounds have different genetic effects on different diseases. As stated above, the client was interested in two response variables, age of onset and skin coverage, one being numeric and the other being categorical.

REGRESSION

Given that age of onset is a continuous variable ranging from zero to 83 years of age, linear regression was chosen for the analysis of this variable, with the general equation seen in equation 3.

$$\hat{y}_i = \sum_{j=1}^{K} \hat{\beta}_j X_{ij} + \epsilon_i \quad \forall\, i = 1, \ldots, n \tag{3}$$

Linear regression models can be estimated using the R statistical language with the *lm* function from the *base* package. Linear regression models normally have clear assumptions in relations to the error term, with the Gauss-Markov Theorem requiring the error term to be uncorrelated, have a mean of 0, and a constant variance (Gauss–Markov theorem 2020). While

these assumptions are stringent, linear regressions can be applied to most data without, as was done for this project.

Skin coverage, on the other hand, was divided into four distinct categories: up to 25%, 26-50%, 51-75%, and 76-100% of skin coverage. Thus, a cumulative multinomial logistic regression was needed to model the relationship between the genetic variants and skin coverage, with the general form as seen in equation 4 below.

$$logit[P(K \leq k)] = \log\left(\frac{P(K \leq k)}{1-P(K \leq k)}\right) = \log\left(\frac{\pi_1+\cdots+\pi_k}{\pi_{k+1}+\cdots+\pi_K}\right) = \beta_{o,j} + \beta x$$

$$\forall\ k = 1,2,\ldots,K-1 \qquad (4)$$

It should be noted that this model type makes the assumption that the outcome categories are ordinal. This in turn creates a logit model that is nested, whereby the probability of *k* category occurring includes the probabilities of classes *k-1* occurring as well. This kind of assumption hinges on the fact that there is value in respecting the ordinality of the variables. A cumulative multinomial logistic model can be estimated using the *orm* function in the *rms* R package.

POWER & EFFECT SIZES

Power is defined as the ability to reject a given null hypothesis H_0 when said hypothesis should be rejected (UCLA, Intro to Power). In order words, it is the ability to detect an effect when an effect *is* present. Normally, a power level of 80% is desired, with it being interpreted as, if there truly was a connection between our disease characteristics and the 46 genetic variants, we should be able to capture said effect at least 80% of the time.

To calculate power, three other factors besides power are needed. One is effect size which is the strength of association between two variables that are being studied (Effect size 2020). In this study it would be the association between age of onset and each genetic variant or skin coverage and each genetic variant. Second is sample size, and third is chosen significance level, denoted as α. Having any three of the four parameters allows us to calculate the fourth one. Meaning that if we have power, α level, and sample size, we can calculate effect size. In the case of our current study, this effect size was the desired metric to be derived.

The first step to calculate power or any of its variants is to determine the null and alternative hypothesis, H_0 and H_a. Based on these hypotheses, a distribution can be chosen, such as t when looking at difference in means, χ^2 when looking at binary or categorical data, or F when comparing two models. The chosen distribution will be used to calculate the associated power probabilities, as well as determining how the effect size will be measured.

Many of the effect size definitions were established by the 20th century statistician Jacob Cohen in his seminal paper '*A Power Primer*' and other earlier published texts. This is the case with the effect size metric we chose for our analysis, called *Cohen's* f^2, chosen in order to calculate the variance explained by the genetic variants. An F-distribution is used when calculating power using Cohen's f^2, which can be calculated from the equation below.

$$f^2 = \frac{R^2}{(1-R^2)} \quad (5)$$

Where R^2 is the variation explained by the genetic variant in each of the 46 regression for age of onset and each of the 46 regression for skin coverage.

Now to calculate power by hand the first thing is to calculate the F-distribution which can be done using the equation below.

$$F = \frac{R^2/k}{1-R^2/(n-k-1)} \quad (6)$$

Where *k* is the number of predictors and *n* is the sample size. Degrees of freedom can be calculated as $df = n - k - 1$. Then we can calculate the non-centrality parameter λ with the equation below. The non-centrality parameter λ can be defined as the degree to which a null hypothesis is false (Non Centrality Parameter (NCP) 2018).

$$\lambda = f^2 * (df) \quad (7)$$

Finally to calculate power all the needs to happen is first determine the critical value (*c*) of the test's statistic using a desired α for $F_{K,\ n-k-1,\ \alpha}$. This means that power can be defined as $Pr[F(t, v, \lambda) \geq c]$, where $F(d, v, \lambda)$ denotes a random variable with a non-central F distribution. Or this can be done easier using R statistical software and the *pwr.f2.test* function from the *pwr* package. A snippet from the source code of the above R function can be found in the appendix.

K-MEANS CLUSTERING

K-means clustering is a type of unsupervised algorithm that seeks to take unlabeled data and position them into *k* clusters or subgroups. Normally these subgroups will have a similar structure, which in layman's terms can be thought of grouping like with like. These subgroupings will be calculated from a dissimilarity matrix in such a way that the within-cluster variance is minimized (James, Witten, Hastie, & Tibshirani). The function to be minimized can be seen below, where $W(C_k)$ is the amount of dissimilarity between observations in the *k*th cluster.

$$\min_{C_1,\ldots,C_k} \{\sum_{k=1}^{K} W(C_k)\} \quad (8)$$

The dissimilarity is normally measured with the squared Euclidean distance, defined in equation 9 below. It should be noted that this measure of dissimilarity makes the assumption of numerical variables, lending itself poorly if not outright erroneously to categorical variables.

$$W(C_k) = \frac{1}{|C_k|}\sum_{i,i'\epsilon C_k}\sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2 \quad (9)$$

Another drawback of the k-means cluster algorithm is its inability to handle missing values. They must be removed or handled in some other way prior to model execution.

A.References

Power Vignette. (n.d.). Retrieved May 15, 2020, from https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html

Kabacoff, R. (n.d.). Power Analysis. Retrieved May 15, 2020, from https://www.statmethods.net/stats/power.html

Effect Size. (n.d.). Retrieved May 15, 2020, from https://www.statisticssolutions.com/statistical-analyses-effect-size/

Multiple logistic regression power analysis. (1965, March 01). Retrieved May 15, 2020, from https://stats.stackexchange.com/questions/162294/multiple-logistic-regression-power-analysis

Newsom. Sample Size and Power for Regression. 2020, web.pdx.edu/~newsomj/mvclass/ho_sample%20size.pdf.

Vitiligo. (2020, April 10). Retrieved May 15, 2020, from https://www.mayoclinic.org/diseases-conditions/vitiligo/symptoms-causes/syc-20355912

Spritz, R. (2008, September 11). VitGene International Consortium to Identify Susceptibility Genes for Generalized. Retrieved May 15, 2020, from https://grantome.com/grant/NIH/R01-AR056292-01

Mcleod, S. (2019, July 04). What are Type I and Type II Errors? Retrieved May 15, 2020, from https://www.simplypsychology.org/type_I_and_type_II_errors.html

Walmsley, A., & Brown, M. (n.d.). What Is Power? Retrieved May 15, 2020, from https://www.statisticsteacher.org/2017/09/15/what-is-power/

Gauss–Markov theorem. (2020, March 01). Retrieved May 15, 2020, from https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem

Intro to Power. (n.d.). Retrieved May 15, 2020, from https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/

Effect size. (2020, May 09). Retrieved May 15, 2020, from https://en.wikipedia.org/wiki/Effect_size

Stephanie. (2018, July 24). Non Centrality Parameter (NCP). Retrieved May 15, 2020, from https://www.statisticshowto.com/non-centrality-parameter-ncp/

Cohen, J. (1991). A Power Primer. *Psychological Bulletin*.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology, 8*(12). doi:10.1371/journal.pcbi.1002822

Santorico, S., Dr. (2018, July). *Genetic risk prediction for complex traits and its relationship to sub-phenotypes in vitiligo*. Lecture.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An introduction to statistical learning: With applications in R*. Springer.

B. Appendix

Table A: Results for Age of Onset

	Age of onset								
genetic_variant	Coefficient	P(>|t|)	s.e.	R^2	R^2 Adj	P(>anova)	Effect Unadj	Effect Bonferroni Adj	Effect Genone Adj
RS114448410	-2.624977297	0.00	0.49	0.0129108	0.012453	0.00	Sufficient effect	Sufficient effect	Insufficient effect
RS12771452	1.475806296	0.01	0.60	0.0027971	0.002339	0.01	Insufficient effect	Insufficient effect	Insufficient effect
RS2687812	0.851474251	0.09	0.49	0.0013558	0.000898	0.09	Insufficient effect	Insufficient effect	Insufficient effect
RS60131261	-0.838736251	0.09	0.50	0.0013002	0.000841	0.09	Insufficient effect	Insufficient effect	Insufficient effect
RS4308124	-0.818425965	0.10	0.49	0.0012851	0.000822	0.10	Insufficient effect	Insufficient effect	Insufficient effect
RS2111485	0.86662474	0.09	0.52	0.0012809	0.000824	0.09	Insufficient effect	Insufficient effect	Insufficient effect
RS2247314	0.870545258	0.11	0.54	0.0011822	0.000720	0.11	Insufficient effect	Insufficient effect	Insufficient effect
RS4268748	0.975865394	0.13	0.64	0.0011037	0.000625	0.13	Insufficient effect	Insufficient effect	Insufficient effect
RS11079035	-0.962032129	0.13	0.64	0.0010357	0.000578	0.13	Insufficient effect	Insufficient effect	Insufficient effect
RS78037977	-1.002416907	0.15	0.70	0.0010159	0.000525	0.15	Insufficient effect	Insufficient effect	Insufficient effect
RS301807	-0.694944575	0.15	0.49	0.0009402	0.000479	0.15	Insufficient effect	Insufficient effect	Insufficient effect
RS706779	0.711317796	0.16	0.51	0.0009079	0.000450	0.16	Insufficient effect	Insufficient effect	Insufficient effect
RS9611565	-0.73075471	0.23	0.61	0.0006806	0.000213	0.23	Insufficient effect	Insufficient effect	Insufficient effect
RS2017445	-0.598671214	0.23	0.50	0.0006576	0.000199	0.23	Insufficient effect	Insufficient effect	Insufficient effect
RS78521699	-1.032651085	0.25	0.90	0.0006062	0.000143	0.25	Insufficient effect	Insufficient effect	Insufficient effect
RS11021232	-0.606862967	0.30	0.59	0.0005059	0.000030	0.30	Insufficient effect	Insufficient effect	Insufficient effect
RS12203592	0.940741007	0.38	1.06	0.0005025	(0.000139)	0.38	Insufficient effect	Insufficient effect	Insufficient effect
RS4807000	-0.512688345	0.31	0.50	0.0004914	0.000022	0.31	Insufficient effect	Insufficient effect	Insufficient effect
RS6012953	0.483952586	0.33	0.50	0.0004381	(0.000021)	0.33	Insufficient effect	Insufficient effect	Insufficient effect
RS1126809	0.584481287	0.35	0.62	0.0004080	(0.000055)	0.35	Insufficient effect	Insufficient effect	Insufficient effect
RS41342147	-0.753414653	0.35	0.80	0.0004034	(0.000057)	0.35	Insufficient effect	Insufficient effect	Insufficient effect
RS10774624	0.440215721	0.38	0.50	0.0003710	(0.000111)	0.38	Insufficient effect	Insufficient effect	Insufficient effect

RS10986311	-0.434179786	0.40	0.51	0.0003335	(0.000132)	0.40	Insufficient effect	Insufficient effect	Insufficient effect
RS12421615	-0.449228708	0.40	0.53	0.0003303	(0.000137)	0.40	Insufficient effect	Insufficient effect	Insufficient effect
RS148136154	-0.521687623	0.46	0.70	0.0002547	(0.000208)	0.46	Insufficient effect	Insufficient effect	Insufficient effect
RS6059655	0.84766171	0.46	1.15	0.0002504	(0.000208)	0.46	Insufficient effect	Insufficient effect	Insufficient effect
RS13076312	-0.303951358	0.56	0.52	0.0001670	(0.000328)	0.56	Insufficient effect	Insufficient effect	Insufficient effect
RS34346645	0.296785397	0.56	0.51	0.0001561	(0.000305)	0.56	Insufficient effect	Insufficient effect	Insufficient effect
RS229527	0.258881152	0.60	0.49	0.0001261	(0.000331)	0.60	Insufficient effect	Insufficient effect	Insufficient effect
RS16843742	-0.319347603	0.61	0.62	0.0001257	(0.000346)	0.61	Insufficient effect	Insufficient effect	Insufficient effect
RS71508903	0.304802025	0.63	0.62	0.0001210	(0.000386)	0.63	Insufficient effect	Insufficient effect	Insufficient effect
RS2476601	-0.310978724	0.67	0.74	0.0000817	(0.000381)	0.67	Insufficient effect	Insufficient effect	Insufficient effect
RS8192917	0.22438798	0.68	0.54	0.0000776	(0.000379)	0.68	Insufficient effect	Insufficient effect	Insufficient effect
RS35161626	-0.194220122	0.69	0.49	0.0000726	(0.000386)	0.69	Insufficient effect	Insufficient effect	Insufficient effect
RS10200159	0.291287962	0.75	0.93	0.0000452	(0.000412)	0.75	Insufficient effect	Insufficient effect	Insufficient effect
RS231725	0.15012198	0.77	0.52	0.0000381	(0.000420)	0.77	Insufficient effect	Insufficient effect	Insufficient effect
RS1635168	0.2133329	0.80	0.83	0.0000322	(0.000451)	0.80	Insufficient effect	Insufficient effect	Insufficient effect
RS2304206	0.161694862	0.80	0.64	0.0000315	(0.000463)	0.80	Insufficient effect	Insufficient effect	Insufficient effect
RS8083511	-0.151495035	0.80	0.59	0.0000299	(0.000427)	0.80	Insufficient effect	Insufficient effect	Insufficient effect
RS1043101	0.109054428	0.82	0.49	0.0000224	(0.000435)	0.82	Insufficient effect	Insufficient effect	Insufficient effect
RS117744081	-0.242361927	0.86	1.39	0.0000147	(0.000467)	0.86	Insufficient effect	Insufficient effect	Insufficient effect
RS35860234	-0.034159542	0.95	0.55	0.0000018	(0.000457)	0.95	Insufficient effect	Insufficient effect	Insufficient effect
RS6583331	0.020668797	0.97	0.50	0.0000008	(0.000458)	0.97	Insufficient effect	Insufficient effect	Insufficient effect
RS1031034	0.022172301	0.97	0.57	0.0000007	(0.000480)	0.97	Insufficient effect	Insufficient effect	Insufficient effect
RS72928038	0.019303523	0.98	0.68	0.0000004	(0.000526)	0.98	Insufficient effect	Insufficient effect	Insufficient effect

RS12482904	0.008858358	0.99	0.58	0.0000001	(0.000484)	0.99	Insufficient effect	Insufficient effect	Insufficient effect

Table B: Results for Skin Coverage, logistic model

	Skin Coverage, logistic model				
genetic_variant	**Likelihood Ratio**	**df**	**P(>|likelihood|)**	**Chi-squared Value**	**P(>|chisqrd|)**
RS1126809	7.88142800	1	0.00	7.729274	0.01
RS10986311	7.87971100	1	0.00	7.909896	0.00
RS6059655	5.59697800	1	0.02	5.841717	0.02
RS41342147	5.26836700	1	0.02	5.401491	0.02
RS1043101	4.33543700	1	0.04	4.339732	0.04
RS78037977	3.74509700	1	0.05	3.812090	0.05
RS12482904	3.45776400	1	0.06	3.485963	0.06
RS72928038	3.41973900	1	0.06	3.465436	0.06
RS2476601	2.85273900	1	0.09	2.899991	0.09
RS60131261	2.40831100	1	0.12	2.415479	0.12
RS6583331	2.35457500	1	0.12	2.348238	0.13
RS12771452	2.08136600	1	0.15	2.061985	0.15
RS229527	1.70439800	1	0.19	1.704003	0.19
RS34346645	1.26639800	1	0.26	1.269035	0.26
RS13076312	1.14011500	1	0.29	1.140355	0.29
RS114448410	1.04321300	1	0.31	1.041764	0.31
RS12203592	0.93560680	1	0.33	0.950589	0.33

RS2017445	0.87464890	1	0.35	0.873046	0.35
RS4807000	0.74667880	1	0.39	0.745704	0.39
RS10200159	0.73806370	1	0.39	0.746987	0.39
RS35161626	0.73154670	1	0.39	0.730611	0.39
RS4308124	0.67650350	1	0.41	0.676997	0.41
RS231725	0.67186680	1	0.41	0.673150	0.41
RS2247314	0.62386700	1	0.43	0.621663	0.43
RS2111485	0.61036700	1	0.43	0.608776	0.44
RS2304206	0.50544180	1	0.48	0.503164	0.48
RS12421615	0.45430000	1	0.50	0.455379	0.50
RS6012953	0.45276070	1	0.50	0.452699	0.50
RS11079035	0.42353490	1	0.52	0.421586	0.52
RS78521699	0.42236670	1	0.52	0.426235	0.51
RS1635168	0.40212120	1	0.53	0.398936	0.53
RS8192917	0.33500370	1	0.56	0.334160	0.56
RS10774624	0.32050230	1	0.57	0.320540	0.57
RS11021232	0.29123940	1	0.59	0.290313	0.59
RS9611565	0.28871320	1	0.59	0.289709	0.59
RS4268748	0.26698220	1	0.61	0.267882	0.60
RS1031034	0.25676120	1	0.61	0.257494	0.61
RS2687812	0.16222960	1	0.69	0.162240	0.69
RS148136154	0.13922390	1	0.71	0.138824	0.71
RS8083511	0.11163380	1	0.74	0.111404	0.74

RS16843742	0.11089280	1	0.74	0.110615	0.74
RS35860234	0.00415405	1	0.95	0.004155	0.95
RS301807	0.00312239	1	0.96	0.003122	0.96
RS71508903	0.00167269	1	0.97	0.001672	0.97
RS117744081	0.00041639	1	0.98	0.000417	0.98
RS706779	0.00005224	1	0.99	0.000052	0.99

Table C: Results for Power Analysis for Skin Coverage transformed to discrete numerical values

	Skin Coverage, as a discrete continuous variable								
genetic_variant	**Coefficient**	**P(>|t|)**	**s.e.**	**R^2**	**R^2 Adj**	**P(>anova)**	**Effect Unadj**	**Effect Bonferroni Adj**	**Effect Genone Adj**
RS10986311	0.08578399	0.00	0.03	0.00	0.00	0.003185159	Sufficient effect	Insufficient effect	Insufficient effect
RS301807	0.008995756	0.75	0.03	0.00	(0.00)	0.745202972	Insufficient effect	Insufficient effect	Insufficient effect
RS2476601	0.058958262	0.16	0.04	0.00	0.00	0.16120651	Insufficient effect	Insufficient effect	Insufficient effect
RS78037977	0.090169716	0.02	0.04	0.00	0.00	0.021289586	Insufficient effect	Insufficient effect	Insufficient effect
RS16843742	-0.028632099	0.41	0.04	0.00	(0.00)	0.414068241	Insufficient effect	Insufficient effect	Insufficient effect
RS10200159	0.048051883	0.36	0.05	0.00	(0.00)	0.360235844	Insufficient effect	Insufficient effect	Insufficient effect
RS4308124	0.015598954	0.58	0.03	0.00	(0.00)	0.576488494	Insufficient effect	Insufficient effect	Insufficient effect
RS2111485	-0.021415153	0.47	0.03	0.00	(0.00)	0.471416566	Insufficient effect	Insufficient effect	Insufficient effect
RS231725	0.021717429	0.47	0.03	0.00	(0.00)	0.465347446	Insufficient effect	Insufficient effect	Insufficient effect
RS41342147	0.063203803	0.17	0.05	0.00	0.00	0.171780757	Insufficient effect	Insufficient effect	Insufficient effect

RS35161626	-0.016962528	0.54	0.03	0.00	(0.00)	0.543244167	Insufficient effect	Insufficient effect	Insufficient effect
RS34346645	0.01683259	0.56	0.03	0.00	(0.00)	0.562545642	Insufficient effect	Insufficient effect	Insufficient effect
RS148136154	-0.030365116	0.44	0.04	0.00	(0.00)	0.442453857	Insufficient effect	Insufficient effect	Insufficient effect
RS13076312	-0.030227744	0.31	0.03	0.00	0.00	0.312996444	Insufficient effect	Insufficient effect	Insufficient effect
RS6583331	-0.048744242	0.09	0.03	0.00	0.00	0.089401665	Insufficient effect	Insufficient effect	Insufficient effect
RS1031034	0.031765959	0.33	0.03	0.00	(0.00)	0.333385844	Insufficient effect	Insufficient effect	Insufficient effect
RS12203592	0.027028305	0.65	0.06	0.00	(0.00)	0.651872063	Insufficient effect	Insufficient effect	Insufficient effect
RS78521699	0.032275797	0.54	0.05	0.00	(0.00)	0.535910533	Insufficient effect	Insufficient effect	Insufficient effect
RS60131261	0.055834535	0.05	0.03	0.00	0.00	0.049132067	Insufficient effect	Insufficient effect	Insufficient effect
RS114448410	0.043079306	0.13	0.03	0.00	0.00	0.129849575	Insufficient effect	Insufficient effect	Insufficient effect
RS72928038	0.0516978	0.19	0.04	0.00	0.00	0.190815947	Insufficient effect	Insufficient effect	Insufficient effect
RS2247314	-0.015706994	0.61	0.03	0.00	(0.00)	0.61280818	Insufficient effect	Insufficient effect	Insufficient effect
RS117744081	-0.017297763	0.82	0.08	0.00	(0.00)	0.823048517	Insufficient effect	Insufficient effect	Insufficient effect
RS2687812	-0.001133884	0.97	0.03	0.00	(0.00)	0.967788115	Insufficient effect	Insufficient effect	Insufficient effect
RS706779	0.007394403	0.80	0.03	0.00	(0.00)	0.796985785	Insufficient effect	Insufficient effect	Insufficient effect
RS71508903	-0.000679361	0.98	0.04	0.00	(0.00)	0.984993004	Insufficient effect	Insufficient effect	Insufficient effect
RS12771452	-0.028735328	0.40	0.03	0.00	(0.00)	0.399779838	Insufficient effect	Insufficient effect	Insufficient effect
RS1043101	0.052548578	0.06	0.03	0.00	0.00	0.064378766	Insufficient effect	Insufficient effect	Insufficient effect
RS12421615	0.03416644	0.26	0.03	0.00	0.00	0.258225555	Insufficient effect	Insufficient effect	Insufficient effect
RS1126809	-0.079818307	0.02	0.04	0.00	0.00	0.023937597	Insufficient effect	Insufficient effect	Insufficient effect
RS11021232	-0.023017922	0.49	0.03	0.00	(0.00)	0.491165731	Insufficient effect	Insufficient effect	Insufficient effect
RS2017445	-0.03579956	0.21	0.03	0.00	0.00	0.206511051	Insufficient effect	Insufficient effect	Insufficient effect
RS10774624	-0.019876863	0.49	0.03	0.00	(0.00)	0.488296713	Insufficient effect	Insufficient effect	Insufficient effect

RS35860234	-0.009230517	0.77	0.03	0.00	(0.00)	0.76727565	Insufficient effect	Insufficient effect	Insufficient effect
RS8192917	-0.022651581	0.46	0.03	0.00	(0.00)	0.464717456	Insufficient effect	Insufficient effect	Insufficient effect
RS1635168	-0.027260972	0.56	0.05	0.00	(0.00)	0.560052501	Insufficient effect	Insufficient effect	Insufficient effect
RS4268748	0.011336936	0.76	0.04	0.00	(0.00)	0.757998138	Insufficient effect	Insufficient effect	Insufficient effect
RS11079035	-0.003347475	0.93	0.04	0.00	(0.00)	0.926449887	Insufficient effect	Insufficient effect	Insufficient effect
RS8083511	-0.02492736	0.46	0.03	0.00	(0.00)	0.463214843	Insufficient effect	Insufficient effect	Insufficient effect
RS4807000	-0.00794876	0.78	0.03	0.00	(0.00)	0.780399025	Insufficient effect	Insufficient effect	Insufficient effect
RS2304206	-0.019648561	0.59	0.04	0.00	(0.00)	0.592580353	Insufficient effect	Insufficient effect	Insufficient effect
RS6059655	0.157031282	0.02	0.07	0.00	0.00	0.018438074	Insufficient effect	Insufficient effect	Insufficient effect
RS6012953	-0.006735166	0.81	0.03	0.00	(0.00)	0.811084647	Insufficient effect	Insufficient effect	Insufficient effect
RS12482904	0.062586821	0.06	0.03	0.00	0.00	0.056378265	Insufficient effect	Insufficient effect	Insufficient effect
RS229527	-0.038477902	0.17	0.03	0.00	0.00	0.17131107	Insufficient effect	Insufficient effect	Insufficient effect
RS9611565	0.024153041	0.49	0.03	0.00	(0.00)	0.486249571	Insufficient effect	Insufficient effect	Insufficient effect

Source code snippet for *pwr.f2.test* function:

```
[…]
p.body <- quote({
        lambda <- f2 * (u + v + 1)
        pf(qf(sig.level, u, v, lower = FALSE), u, v, lambda,
            lower = FALSE)
    })
[…]


f2 <- uniroot(function(f2) eval(p.body) - power, c(1e-07,
            1e+07))$root
```

Code:

```
# Genetic variants and vitiligo disease characteristics
# Authors: Sandra Robles Munoz & Valentinas Sungaila
# Spring 2020, MATH 6330


######### Libraries #########
library(data.table)
library(nnet)
library(rms)
library(lmtest)
library(pwr)
library(factoextra)



######### Data #########
dir_path <- "~/Grad School/MATH-6330_S20/Project/" #WALDO, change me
dat <- as.data.frame(fread(paste0(dir_path,"Data_6330_Project.csv")))

######### Data Hygiene #########
colnames(dat) <- gsub("bcsnp.vitiligo_assoc","vit_ass",colnames(dat))
colnames(dat) <- gsub("bcsnp.vitiligo","vit_ass",colnames(dat))
colnames(dat) <- gsub("bcsnp.patient","patient",colnames(dat))

dat$vit_ass.rheumatoid_arthritis <- as.integer(ifelse(dat$vit_ass.rheumatoid_arthritis
=='yes',1,0))
dat$vit_ass.systemic_lupus_erythematosus <- as.integer(ifelse(dat$vit_ass.systemic_lup
us_erythematosus=='yes',1,0))
dat$vit_ass.pernicious_anemia <- as.integer(ifelse(dat$vit_ass.pernicious_anemia=='yes
',1,0))
dat$vit_ass_onset.acrofacial <- as.integer(ifelse(dat$vit_ass_onset.acrofacial=='yes',
1,0))
dat$vit_ass.halo_nevi <- as.integer(ifelse(dat$vit_ass.halo_nevi=='yes',1,0))
dat$vit_ass_onset.koebner_phenomenon <- as.integer(ifelse(dat$vit_ass_onset.koebner_ph
enomenon=='yes',1,0))

genotype_cols <- grep('^RS',colnames(dat),value=T)
covariates_cols <- grep('^GWAS',colnames(dat),value=T)
overlay_cols <- grep('patient.|vit_ass',colnames(dat),value=T)
dependent_cols <- grep('SKIN|DURATION',colnames(dat),value=T)
score_cols <- grep("^score",colnames(dat),value = T)

dat$SKIN <- factor(dat$SKIN,ordered=TRUE,levels=c('up to 25%','26-50%','51-75%','76-10
0%'))

######### PRE-PROCESSING: Picking top Ancestry Principal Components #########
```

```
# All gwas are populated otherwise we'd have to filter down to full observations
fwer_value <- 0.10/length(covariates_cols) # Using family-wise error adjustment (alpha
/number of tests)
ageonset_keep <- c()
skin_keep <- c()

# Step 1 - Duration
mean_model_age <- lm(patient.vitiligo_age_of_onset ~ 1, data = dat)
for(count in covariates_cols){
  temp_model <- lm(paste0("patient.vitiligo_age_of_onset ~ ",count),data = dat)
  fwer_test <- anova(mean_model_age,temp_model)$`Pr(>F)`[2] < fwer_value
  if(fwer_test){ageonset_keep <- c(ageonset_keep,count)}
}

# Step 2 - SKIN

# No need to create a mean model, the 'orm' function natively does a loglikelihood tes
t
for(count in covariates_cols){
  temp_model <- orm(as.formula(paste0("SKIN ~ ",count)),data = dat)
  fwer_test <- as.numeric(temp_model$stats[7]) < fwer_value
  print(as.numeric(temp_model$stats[7]))
  if(fwer_test){skin_keep <- c(skin_keep,count)}
  # Note: no PCs kept using ordinal model, GWAS1.EV1 kept if nominal.However, ordinali
ty is respected, so orm() is used instead of multinom()
  # Note 2: POLR fails on GWAS1.V1 due to 'initial value in 'vmmin' is not finite', go
ogle search indicates that using the package RMS bypasses this problem

}


######### QUESTION 1 - Duration #########
# Filters down to complete observations
age_model_list <- list()
age_df <- data.frame(genetic_variant=character(),
                     Coefficient=double(),
                    var_prob=double(),
                    std_err=double(),
                    Rsqd=double(),
                    Rsqrd_adj=double(),
                    pval_anova=double(),stringsAsFactors = F)
if(length(ageonset_keep>0)){ #PC covariates will be applied
 for(count in genotype_cols){
   age_model_list[[count]] <- lm(paste0("patient.vitiligo_age_of_onset ~ ",count," + "
,paste0(ageonset_keep,collapse = " + ")),data = dat[!is.na(colnames(dat) %in% c(count)
)])
   tmp_summary <- summary(age_model_list[[count]])
   tmp_coeff <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,"E
stimate"]
   tmp_stderr <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,"
Std. Error"]
   tmp_valprob <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,
"Pr(>|t|)"]
   tmp_rsqrd <- tmp_summary$r.squared
   tmp_rsqrdadj <- tmp_summary$adj.r.squared
   tmp_fstatprob <- as.double(pf(tmp_summary$fstatistic[1],tmp_summary$fstatistic[2],t
mp_summary$fstatistic[3],lower.tail=FALSE))
   age_df[nrow(age_df)+1,]<- c(count,tmp_coeff,
                               tmp_valprob,
                               tmp_stderr,
                               tmp_rsqrd,
                               tmp_rsqrdadj,
                               tmp_fstatprob)
 }
}else{
  for(count in genotype_cols){
    age_model_list[[count]] <- lm(paste0("patient.vitiligo_age_of_onset ~ ",count),dat
a = dat[!is.na(colnames(dat) %in% c(count))])
    tmp_summary <- summary(age_model_list[[count]])
    tmp_coeff <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,"
Estimate"]
```

```
    tmp_stderr <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,
"Std. Error"]
    tmp_valprob <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count
,"Pr(>|t|)"]
    tmp_rsqrd <- as.double(tmp_summary$r.squared)
    tmp_rsqrdadj <- as.double(tmp_summary$adj.r.squared)
    tmp_fstatprob <- as.double(pf(tmp_summary$fstatistic[1],tmp_summary$fstatistic[2],
tmp_summary$fstatistic[3],lower.tail=FALSE))
    age_df[nrow(age_df)+1,]<- c(count,tmp_coeff,
                                tmp_valprob,
                                tmp_stderr,
                                tmp_rsqrd,
                                tmp_rsqrdadj,
                                tmp_fstatprob)
  }

}
age_df[!colnames(age_df) %in% c("genetic_variant")] <- lapply(age_df[!colnames(age_df)
%in% c("genetic_variant")], as.numeric)

######### QUESTION 1 - Skin surface #########
# The model can be built with skin treated as ordinal or nominal. There's often a lot
of information to be gained from leaving alone the ordinal nature of the outcome

skin_model_list <- list()
skin_df <- data.frame(genetic_variant=character(),
                      likelihoodration=double(),
                      df=integer(),
                      likelihood_pval=double(),
                      chisqrd_val=double(),
                      chisqrd_pval=double(),
                      stringsAsFactors = F)
if(length(skin_keep>0)){ #PC covariates will be applied
  for(count in genotype_cols){
    skin_model_list[[count]] <- orm(as.formula(paste0("SKIN ~ ",count," + ",paste0(ski
n_keep,collapse = " + "))),data=dat[!is.na(colnames(dat) %in% c(count))])
    tmp_summary <- skin_model_list[[count]]$stats
    tmp_likelihoodratio = as.double(tmp_summary[5])
    tmp_df = as.integer(tmp_summary[6])
    tmp_LR_pval = as.double(tmp_summary[7])
    tmp_chisqrd = as.double(tmp_summary[8])
    tmp_chisqrd_pval = as.double(tmp_summary[9])
    #pchisq(as.double(tmp_summary[8]),df = 2,lower.tail = F)
    skin_df[nrow(skin_df)+1,]<- c(count,tmp_likelihoodratio,
                                  tmp_df,
                                  tmp_LR_pval,
                                  tmp_chisqrd,tmp_chisqrd_pval)
  }

}else{
  for(count in genotype_cols){
    skin_model_list[[count]] <- orm(as.formula(paste0("SKIN ~ ",count)),data=dat[!is.n
a(colnames(dat) %in% c(count))])
    tmp_summary <- skin_model_list[[count]]$stats
    tmp_likelihoodratio = as.double(tmp_summary[5])
    tmp_df = as.integer(tmp_summary[6])
    tmp_LR_pval = as.double(tmp_summary[7])
    tmp_chisqrd = as.double(tmp_summary[8])
    tmp_chisqrd_pval = as.double(tmp_summary[9])
    #pchisq(as.double(tmp_summary[8]),df = 2,lower.tail = F)
    skin_df[nrow(skin_df)+1,]<- c(count,tmp_likelihoodratio,
                                  tmp_df,
                                  tmp_LR_pval,
                                  tmp_chisqrd,tmp_chisqrd_pval)
  }

}
skin_df[!colnames(skin_df) %in% c("genetic_variant")] <- lapply(skin_df[!colnames(skin
_df) %in% c("genetic_variant")], as.numeric)
```

```
######### QUESTION 2 - Duration #########
# Q2: What effect sizes for disease characteristics are detectable with >= 80% power
        # without adjustment for multiple testing,
        # with adjust for the number of tests in Q1,
        # and with adjustment genome-wide?

n_ageonset <- nrow(dat[!is.na(dat$patient.vitiligo_age_of_onset),])
df_num = 2-1 # Predictors minus intercept. We don't have any covariates to account for
, otherwise that'd go here
effect_unadj <- pwr.f2.test(u = df_num, v=n_ageonset-1-1, sig.level = 0.05, power = 0.
8)
effect_bonferroniadj <- pwr.f2.test(u = df_num, v=n_ageonset-1-1, sig.level = 0.05/46,
power = 0.8)
effect_genomewide <- pwr.f2.test(u = df_num, v=n_ageonset-1-1, sig.level = 5e-8, power
= 0.8)

# Effect (R^2) needed for 80% power with a 2190 sample size:
eff_unadj <- effect_unadj$f2/(1+effect_unadj$f2)  #0.003576067
eff_adj <- effect_bonferroniadj$f2/(1+effect_bonferroniadj$f2) #0.007686651
eff_genadj <- effect_genomewide$f2/(1+effect_genomewide$f2) #0.01788534

age_df$effect_largeenough_unadj <- ifelse(age_df$Rsqd>=eff_unadj,'Sufficient effect','
Insufficient effect')
age_df$effect_largeenough_bonfadj <- ifelse(age_df$Rsqd>=eff_adj,'Sufficient effect','
Insufficient effect')
age_df$effect_largeenough_genomeadj<- ifelse(age_df$Rsqd>=eff_genadj,'Sufficient effec
t','Insufficient effect')
# Only RS114448410 has a big enough effect size for 80% power at the unadjusted and bo
nferroni adjustment level. None at genone-wide


######### QUESTION 2 - Skin surface #########
#NOTE: Power is not readily available for Multinomial Logistic Regression and its assu
med distribution
# So this step will requite recreating the 46 regressions with SKIN recoded as numeric
al, and then run as linear models
#dat_skin <- dat[!is.na(dat$SKIN),]
dat$SKIN_numeric <- ifelse(dat$SKIN=='up to 25%',1,
      ifelse(dat$SKIN=='26-50%',2,
             ifelse(dat$SKIN=='51-75%',3,
                    ifelse(dat$SKIN=='76-100%',4,dat$SKIN
       ))))

# First figure out any covariates
skin_numerical_keep <- c()
mean_model_ageNumerical <- lm(SKIN_numeric ~ 1, data = dat)
for(count in covariates_cols){
  temp_model <- lm(paste0("SKIN_numeric ~ ",count),data = dat)
  fwer_test <- anova(mean_model_ageNumerical,temp_model)$`Pr(>F)`[2] < fwer_value
  if(fwer_test){skin_numerical_keep <- c(skin_numerical_keep,count)}
}
# Now we construct the linear models to get the R^2 values
skinnumerical_model_list <- list()
skinnumerical_df <- data.frame(genetic_variant=character(),
                               Coefficient=double(),
                               var_prob=double(),
                               std_err=double(),
                               Rsqd=double(),
                               Rsqrd_adj=double(),
                               pval_anova=double(),stringsAsFactors = F)
if(length(ageonset_keep>0)){ #PC covariates will be applied
  for(count in genotype_cols){
    skinnumerical_model_list[[count]] <- lm(paste0("SKIN_numeric ~ ",count," + ",paste
0(ageonset_keep,collapse = " + ")),data = dat[!is.na(colnames(dat) %in% c(count))])
    tmp_summary <- summary(skinnumerical_model_list[[count]])
    tmp_coeff <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,"
Estimate"]
    tmp_stderr <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,
"Std. Error"]
```

```
    tmp_valprob <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count
,"Pr(>|t|)"]
    tmp_rsqrd <- tmp_summary$r.squared
    tmp_rsqrdadj <- tmp_summary$adj.r.squared
    tmp_fstatprob <- as.double(pf(tmp_summary$fstatistic[1],tmp_summary$fstatistic[2],
tmp_summary$fstatistic[3],lower.tail=FALSE))
    skinnumerical_df[nrow(skinnumerical_df)+1,]<- c(count,tmp_coeff,
                                                    tmp_valprob,
                                                    tmp_stderr,
                                                    tmp_rsqrd,
                                                    tmp_rsqrdadj,
                                                    tmp_fstatprob)
  }
}else{
  for(count in genotype_cols){
    skinnumerical_model_list[[count]] <- lm(paste0("SKIN_numeric ~ ",count),data = dat
[!is.na(colnames(dat) %in% c(count))])
    tmp_summary <- summary(skinnumerical_model_list[[count]])
    tmp_coeff <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,"
Estimate"]
    tmp_stderr <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count,
"Std. Error"]
    tmp_valprob <- tmp_summary$coefficients[row.names(tmp_summary$coefficients)==count
,"Pr(>|t|)"]
    tmp_rsqrd <- as.double(tmp_summary$r.squared)
    tmp_rsqrdadj <- as.double(tmp_summary$adj.r.squared)
    tmp_fstatprob <- as.double(pf(tmp_summary$fstatistic[1],tmp_summary$fstatistic[2],
tmp_summary$fstatistic[3],lower.tail=FALSE))
    skinnumerical_df[nrow(skinnumerical_df)+1,]<- c(count,tmp_coeff,
                                                    tmp_valprob,
                                                    tmp_stderr,
                                                    tmp_rsqrd,
                                                    tmp_rsqrdadj,
                                                    tmp_fstatprob)
  }

}
skinnumerical_df[!colnames(skinnumerical_df) %in% c("genetic_variant")] <- lapply(skin
numerical_df[!colnames(skinnumerical_df) %in% c("genetic_variant")], as.numeric)


# And now the Power calculations:
n_skin <- nrow(dat[!is.na(dat$SKIN_numeric),])
df_num = 2-1 # Predictors minus intercept. We don't have any covariates to account for
, otherwise that'd go here
effect_unadj_skin <- pwr.f2.test(u = df_num, v=n_skin-1-1, sig.level = 0.05, power = 0
.8)
effect_bonferroniadj_skin <- pwr.f2.test(u = df_num, v=n_skin-1-1, sig.level = 0.05/46
, power = 0.8)
effect_genomewide_skin <- pwr.f2.test(u = df_num, v=n_skin-1-1, sig.level = 5e-8, powe
r = 0.8)

# Effect (R^2) needed for 80% power with a 1988 sample size:
eff_unadj_skin <- effect_unadj_skin$f2/(1+effect_unadj_skin$f2)  # 0.00395165
eff_adj_skin <- effect_bonferroniadj_skin$f2/(1+effect_bonferroniadj_skin$f2) # 0.0084
35601
eff_genadj_skin <- effect_genomewide_skin$f2/(1+effect_genomewide_skin$f2) # 0.0200827
8

skinnumerical_df$effect_largeenough_unadj <- ifelse(skinnumerical_df$Rsqd>=eff_unadj_s
kin,'Sufficient effect','Insufficient effect')
skinnumerical_df$effect_largeenough_bonfadj <- ifelse(skinnumerical_df$Rsqd>=eff_adj_s
kin,'Sufficient effect','Insufficient effect')
skinnumerical_df$effect_largeenough_genomeadj<- ifelse(skinnumerical_df$Rsqd>=eff_gena
dj_skin,'Sufficient effect','Insufficient effect')
# Only RS10986311 has a big enough effect size for 80% power at the unadjusted level.
None at bonferroni or genone-wide level.



######### QUESTION 3 - Structure of cases based on disease characteristics? #########
```

```
# structure of cases can be derived using clustering algorithms, however,  they cannot handle categorical data or missing data
# For categorical: clustering depends on distances (dissimilarities) which is almost ways done with eucledian distances
# For missing: same reason of distance calculations. There are imputation techniques available

cluster_data <- dat[,c(score_cols,'SKIN_numeric','patient.vitiligo_age_of_onset')]
cluster_data <- na.omit(cluster_data)
cluster_data <- scale(cluster_data)

clustersss <-kmeans(x=as.matrix(cluster_data),centers = 3,nstart = 150, iter.max=30)
table(clustersss$cluster)
fviz_cluster(clustersss, geom = "point", data = cluster_data) + ggtitle("First 2 principal components by cluster")



######### PRESENTATION/PAPER CODE #########
dat$temp_ageonset = ifelse(between(dat$patient.vitiligo_age_of_onset,0,10),'0-10',
                           ifelse(between(dat$patient.vitiligo_age_of_onset,11,20),'11-20',
                                  ifelse(between(dat$patient.vitiligo_age_of_onset,21,30),'21-30',
                                         ifelse(between(dat$patient.vitiligo_age_of_onset,31,40),'31-40',
                                                ifelse(between(dat$patient.vitiligo_age_of_onset,41,50),'41-50',
                                                       ifelse(between(dat$patient.vitiligo_age_of_onset,51,60),'51-60',
                                                              ifelse(between(dat$patient.vitiligo_age_of_onset,61,70),'61-70',
                                                                     ifelse(between(dat$patient.vitiligo_age_of_onset,71,80),'71-80',
                                                                            ifelse(dat$patient.vitiligo_age_of_onset>80,'80+',dat$patient.vitiligo_age_of_onset)))))))))
table(dat$SKIN,dat$temp_ageonset)
```