MATH 2625: Biostatistical Methods
Homework 3, due Thursday, February 27

Please submit a PDF or .doc version of your homework to Canvas by 3:30pm on the due date. Please type *all* responses. You are encouraged to use `R` for all calculations.

Yu Fan Mei, Bailey Coughlan

Theory

1. Let $y_1, \ldots, y_n$ be realizations from the RVs $Y_i \overset{iid}{\sim} Pois(\mu)$. Using each of the following test constructions, derive a test of the null and alternative hypotheses $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$:

   (a) Score Test

   (b) Likelihood Ratio Test

   The likelihood function L for the poisson distribution and the parameter μ is

   $$L(\mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}.$$

   When we take the log-likelihood and simplify the function, it becomes

   $$\ell(\mu) = \left(\sum y_i\right) \log(\mu) - n\mu - \sum \log(y_i!).$$

   The first derivative of $\ell(\mu)$ (with respect to μ) is

   $$U(\mu) = \frac{\sum y_i}{\mu} - n.$$

   The Fisher information is

   $$I(\mu) = \frac{n}{\mu}$$

   Then the score test statistic S is

   $$S = \frac{\frac{\sum y_i}{\mu} - n}{\sqrt{\frac{n}{\mu}}}.$$

   When we simplify, we get

   $$S = \frac{\sqrt{\mu}(\sum y_i - n)}{\sqrt{n}}.$$

   We know that S follows a standard normal distribution as n tends towards infinity. Additionally, under the null hypothesis, this means that

   $$S = \frac{\sqrt{\mu_0}(\sum y_i - n)}{\sqrt{n}} \sim N(0, 1).$$

The likelihood function L for the poisson distribution and the parameter μ is

$$L(\mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}.$$

When we substitute in our null and alternative hypotheses, $\mu = \mu_0$ and $\mu = \mu_1$ respectively, we can construct the likelihood ratio test statistic. This is defined as

$$\lambda = \frac{\prod_{i=1}^{n} \frac{\mu_0^{y_i} e^{-\mu_0}}{y_i!}}{\prod_{i=1}^{n} \frac{\hat{\mu}^{y_i} e^{-\hat{\mu}}}{y_i!}}.$$

We take the log likelihood and simplify the equation further. Finally, we multiply the equation by -2 to align it with the χ^2 distribution, and we are left with

$$\lambda = 2 \sum_{i=1}^{n} \left(y_i \log \left(\frac{y_i}{\mu_0} \right) - y_i - \mu_0 \right) \sim \chi^2(df = 1).$$

2. Suppose we have a sample, $x_1, \ldots, x_n$, from the RVs $X_i \overset{iid}{\sim} Bern(p)$. We wish to build a confidence interval for p but we know, a priori, that the event of interest is rare. For rare outcomes, it is common to use the complimentary log-log transformation: $g(p) = \log[-\log(1-p)]$. First find a $100(1-\alpha)\%$ CI for $g(p)$. Then, find a $100(1-\alpha)\%$ CI for p itself. State any necessary assumptions.

   By the chain rule, the derivative $g'(p)$ of $g(p)$ is

   $$\frac{1}{-\log(1-p)(1-p)}(-1).$$

   Since X is a bernoulli random variable, we know that

   $$Var(X) = p(1-p).$$

   Then we have the variance of this transformed random variable, which is

   $$\frac{p}{\log(1-p)(1-p)}.$$

   Then the confidence interval for $g(p)$ is

   $$z_{\alpha/2} \sqrt{\frac{p}{\log(1-p)^2(1-p)n}}.$$

   Define this term as G (because I don't feel like rewriting that square root so many times). When we transform this function back, we get a confidence interval constructed by

   $$p = 1 - \exp^{\exp^{g(p) \pm G}}$$

   This works asymptotically for large n because of the law of large numbers, but we assume that n is large, and that all of the observations are iid.

Case Studies

For the following case studies, create a structured abstract no longer than 4 pages in length (including figures, tables, and references). The Background section is provided for each and should be included in your write-up. You must write the Methods, Results, and Conclusion sections. Code should be included in an appendix as well.

Background

An evolving understanding of the immunopathogenesis of multiple sclerosis suggests that depleting B cells, a type of white blood cell that produces antibodies, could be useful for treatment. We studied ocrelizumab, a humanized monoclonal antibody treatment that selectively depletes CD20-expressing B cells, in the primary progressive form of the disease. In this trial, we randomly assigned 732 patients with primary progressive multiple sclerosis in a 2:1 ratio to receive intravenous ocrelizumab (600 mg) or placebo every 24 weeks for at least 120 weeks and until a prespecified number of confirmed disability progression events had occurred. Disability progression was defined as an increase in the Expanded Disability Status Scale (EDSS). The primary outcome of interest was the percentage of patients with confirmed disability progression at 12 weeks. As secondary analyses, we are interested in 24-week percent confirmed disability progression as well as the 120-week percent worsened performance on the timed 25-foot walk—-a quantitative score of mobility and leg function.

Methods

732 patients who were diagnosed with primary progressive multiple sclerosis were randomly selected for participation in this randomized clinical trial. The study was double-blind and placebo-controlled. Subjects were randomly assigned to the treatment or placebo group in a 2:1 ratio. Every 24 weeks, eligible participants were either administered 600 mg of intravenous ocrelizumab or a placebo drug. In total, 488 of the participants received the treatment while 244 received the placebo.

A check-in was conducted at the 12-week mark as well as the 24-week mark to determine whether subjects had undergone disability progression. As a secondary analysis, we also measured subjects' performance on the timed 25-foot walk at the end of the study (120 weeks). No subjects were lost during the 120 week-long trial.

To examine possible association between the ocrelizumab treatment and disability progression, we used a Pearson's χ^2 Test. We also examined the odds ratios to examine the magnitude of association. Progression results were collected at the 12 and 24-week mark, and at 120 weeks, subjects' performance on the 25-foot walk was also measured. All analyses were performed at the nominal level.

Results

12 weeks into the study, we observed that 33.0% of the 488 subjects who were administered the B cell depletion treatment underwent disability progression, while 39.3% of the 244 subjects who received the placebo also underwent disability progression. A more detailed summary of the findings is available in Table 1. At the 12-week mark, there was no significant difference in disability progression between those who received the ocrelizumab treatment and those who received the placebo ($p = 0.106, \chi^2 = 2.609, df = 1, N = 732$). While we observed within the trial that the odds of disability progression for patients who received ocrelizumab was 24.1% lower than patients who only received a placebo, these results were not statistically significant (OR = 0.76, 95% CI: [0.552, 1.044]).

Table 1: **Disability Progression After 12 Weeks by Treatment**

Treatment	Disability Progression	No Progression	Total
Ocrelizumab	161	327	488
Placebo	96	148	244
Totals	257	475	732

At the 24-week mark, we observed a disability progression in just 24.5% of the 488 subjects who received ocrelizumab. This is in comparison to 55.4% of the 157 receiving a placebo, as shown in Table 2. However, we did not find a significant association between receiving the ocrelizumab treatment and a lowered disability progression ($p = 0.109, \chi^2 = 2.569, df = 1, N = 732$). Despite the observed odds in the study of having a disability progression being 24.5% lower in patients receiving immunosuppression than in patients receiving the placebo, this observation was not statistically significant (OR = 0.755, 95% CI: [0.545, 1.047]).

Table 2: **Disability Progression After 24 Weeks by Treatment**

Treatment	Disability Progression	No Progression	Total
Ocrelizumab	144	344	488
Placebo	87	157	244
Totals	231	501	732

At the end of the study period, we found worsened 25-foot walk performance in 38.9% of the participants receiving ocrelizumab, while 55.3% of the control group had a worsened 25-foot walk. Further information is also available in Table 3. We found that in 120 weeks, patients who received ocrelizumab were significantly less likely to have worsened mobility. The subjects who were administered ocrelizumab were less likely to have a worsened 25-foot walk performance ($p < 0.0001, \chi^2 = 17.051, df = 1, N = 732$). Among the participants, the odds of having a worsened 25-foot walk performance were 48.5% lower among subjects receiving ocrelizumab than subjects receiving the placebo (OR = 0.515, 95% CI: [0.377, 0.703]).

Table 3: **Worsened Performance on 25-Foot Walk by Treatment**

Treatment	Worsened	Did not Worsen	Total
Ocrelizumab	190	298	488
Placebo	135	109	244
Totals	325	407	732

Discussion

We found that at the 12 and 24-week mark, the ocrelizumab did not make a significant impact in reducing disability progression. However, at the conclusion of the study, patients who received ocrelizumab during the 120-week duration were less likely to have worsened mobility. We conclude that ocrelizumab works best as a long-term treatment for multiple sclerosis. While short-term findings may differ, this is consistent with findings in other studies and confirms that ocrelizumab works as a treatment for primary progressive multiple sclerosis (Lamb, 2022).

Follow-up research should seek to look more closely at the impacts over time of administering ocrelizumab. Additionally, the dose-response relationship of ocrelizumab and disability progression should also be explored. Since all subjects were administered a 600 mg dosage, future studies should examine the impacts of different dosages of ocrelizumab on disability progression. One limitation of our study was the aggregation of all participants. Stratification of subjects into different groups based on age, gender, social class, and levels of physical activity may be able to account for the impacts of these potential cofounders. Epidemiologic reviews of the disease suggest that exposure to smoking, low levels of vitamin D, or a history of childhood obesity— all risk factors associated with multiple sclerosis— are worth looking into as confounding variables as well (Dobson & Giovanni, 2019).

Finally, multiple sclerosis is known to be a chronic disease with no known cure. Its damaging impacts on the central nervous system include much more than mobility and motor-related functions, which are also worth exploring separately (Goldenberg, 2012). Future studies with longer observation periods could help to examine the long-term impacts of ocrelizumab, which remain unexplored in our short study.

For the rs8034191 2 SNP, we found 724 people with the homozygous A genotype. 939 participants were heterozygous (AB), while 290 had the homozygous B genotype. We found that there was an association between a genotype and the frequency of cases and controls. The post-hoc odds ratios showed that having even one B allele significantly increased the odds of having small cell lung carcinoma. The odds of having SCLC for those with the homozygous A genotype was 0.639 times the odds of having SCLC for those with the heterozygous genotype (97.5% CI: [0.511, 0.799]). The odds of having SCLC for those with the homozygous A genotype was even lower compared to the homozygous B genotype, being 0.571 times the odds when a person had both B alleles (97.5% CI: [0.416, 0.781]).

Appendix

- Dobson, R., & Giovannoni, G. (2019). Multiple sclerosis–a review. *European Journal of Neurology*, 26(1), 27-40.
- Goldenberg, M. M. (2012). Multiple sclerosis review. *Pharmacy and Therapeutics*, 37(3), 175.
- Lamb, Y. N. (2022). Ocrelizumab: a review in multiple sclerosis. *Drugs*, 82(3), 323-334.

R Code

Listing 1: R code for analyzing progression and walk performance in Ocrelizumab treatment

```
#install.packages('epitools')
library(epitools)

#### Case Study 1 ####
##### 12 week progression #####
prog12 <- matrix(c(161, 96, 327, 148), nrow = 2,
                 dimnames = list(Treatment = c('Ocrelizumab', 'Placebo'),
                                 Progression = c('Yes', 'No')))

chisq.test(prog12)
oddsratio.wald(prog12)

##### 24 week progression #####
prog24 <- matrix(c(144, 87, 344, 157), nrow = 2,
                 dimnames = list(Treatment = c('Ocrelizumab', 'Placebo'),
                                 Progression = c('Yes', 'No')))

chisq.test(prog24)
oddsratio.wald(prog24)

##### 120 week worsened 25-foot walk performance #####
foot25 <- matrix(c(190, 135, 298, 109), nrow = 2,
                 dimnames = list(Treatment = c('Ocrelizumab', 'Placebo'),
                                 Worsened = c('Yes', 'No')))

chisq.test(foot25)
oddsratio.wald(foot25)
```

Background

Small cell lung carcinoma, or SCLC, is a highly malignant type of cancer that typically presents in the lungs. Given its tendency to metastasize early, the prognosis is poor with only 10 to 15% of patients surviving three years after diagnosis. Early diagnosis could potentially improve patient prognosis if treatment can be administered before the cancer metastasizes as SCLC does respond well to chemotherapy and radiotherapy in the early stages. The Harvard Lung Cancer Susceptibility Study was a case-control study conducted to identify possible risk factors associated with SCLC. In total, 1000 cases and 1000 controls were recruited, however after drop-out, there 984 cases and 969 controls available for analysis. While a large number of variables were selected, we are primarily interested in assessing the association between case status and two different genetic markers as well as smoking status. The genetic markers are the single nucleotide polymorphisms (SNPs) rs8034191_2 and rs1051730_1 which are found on Chromosome 15 in a region that had previously been identified to contain genes associated with SCLC. We're interested in seeing if either of the two SNPs are associated with SCLC using the appropriate measure of association. At each SNP, there are two possible alleles, A or B. The B allele is considered the risk allele. Subjects can have at most two risk alleles and it is thought that subjects with two risk alleles are at even more risk than those with just one (and those with one are at more risk than those with none). We might also want to see if the environmental factor, smoking status (levels: never, former, current), is associated with presence of SCLC.

Methods

The initial selection for this case-control study involved 2,000 participants, 1,000 of whom had SCLC ("cases") and 1,000 who did not ("controls"). Neighborhood control methods were used to match cases and controls as much as possible. After dropout, 984 cases and 969 controls remained in the study. We sampled DNA from the participants, with a specific interest in Chromosome 15 and the two SNPs— rs8034191 2 and rs1051730 1— the SNPs associated with SCLC.

We used a Pearson's χ^2 test to examine the potential association between SCLC and the 3 genotypes: homozygous A, homozygous B, and heterozygous. We checked for this association at both SNP sites. We ran a similar analysis with smoking status to check if it was a strong confounder. We planned to conduct post-hoc association tests by observing odds ratios and their respective confidence intervals. For these odds ratios, we used either homozygous A or nonsmokers as reference groups. Initial tests for association were conducted at the nominal level. For post-hoc tests, a Bonferroni correction was applied to control for type I error, resulting in an adjusted p-value of $\alpha^* = 0.025$.

Results

For the rs8034191 2 SNP, we found 724 people with the homozygous A genotype. 939 participants were heterozygous (AB), while 290 had the homozygous B genotype. A more detailed breakdown is available in Table 4. We found that there was an association between a genotype and the frequency of cases and controls ($\chi^2 = 26.088, df = 2, p < 0.0001$). The post-hoc odds ratios showed that having even one B allele significantly increased the odds of having small cell lung carcinoma. The odds of having SCLC for those with the homozygous A genotype was 0.639 times the odds of having SCLC for those with the heterozygous genotype (97.5% CI: [0.511, 0.799]). The odds of having SCLC for those with the homozygous A genotype was even lower compared to the homozygous B genotype, being 0.571 times the odds when a person had both B alleles (97.5% CI: [0.416, 0.781]).

SCLC Status	**AA**	**AB**	**BB**	**Total**
Has SCLC	311	508	165	984
No SCLC	413	431	125	969
Total	724	939	290	1953

Table 4: Breakdown of rs8034191 2 Genotype Frequency

Comparison	**Odds Ratio**	**97.5% Confidence Interval**
AA Odds vs. AB Odds	0.639	[0.511, 0.799]
AA Odds vs. BB Odds	0.571	[0.416, 0.781]

Table 5: Pairwise Odds Ratios Show Increased SCLC Risk in B Allele rs8034191 2

For the second SNP, rs1051730 1, we found that 711 of the 1953 subjects had the homozygous A genotype, while 935 of them were heterozygous. 307 participants had the homozygous B genotype. We found that there was a significant association between whether or not a person had SCLC and their genotype ($\chi^2 = 26.938, df = 2, p < 0.0001$). For this allele, the odds of having SCLC were significantly lower for subjects with the homozygous A genotype when compared directly to both the heterozygous genotype and the homozygous B genotype (AA-AB OR: 0.676, 97.5% CI: [0.540, 0.845]; AA-BB OR: 0.522, 97.5% CI: [0.382, 0.712]).

SCLC Status	**AA**	**AB**	**BB**	**Total**
Has SCLC	307	495	182	984
No SCLC	404	440	125	969
Total	711	935	307	1953

Table 6: Breakdown of rs1051730 1 Genotype Frequency

Comparison	**Odds Ratio**	**97.5% Confidence Interval**
AA Odds vs. AB Odds	0.676	[0.540, 0.845]
AA Odds vs. BB Odds	0.522	[0.382, 0.712]

Table 7: Pairwise Odds Ratios Show Increased SCLC Risk in B Allele rs1051730 1

Among the 1953 participants in the study, we found that 253 had never smoked, 1,056 were former smokers, and 644 were current smokers. A more detailed breakdown of the participants' smoking status is available in Table 8. We found that a person's SCLC status was also associated with their smoking status ($\chi^2 = 49.987, df = 2, p < 0.0001$). We found that the odds of a person having SCLC were significantly lower if they didn't smoke directly compared to if they smoked or even if they had quit smoking (vs. former smokers OR: 0.631, 97.5% CI: [0.455, 0.871]; vs. current smokers OR: 0.373, 07.5% CI: [0.455, 0.871]).

SCLC Status	**Never Smoked**	**Former Smoker**	**Current Smoker**	**Total**
Has SCLC	92	502	390	984
No SCLC	161	554	254	969
Total	253	1056	644	1953

Table 8: Breakdown of Participants' Smoking Status

Smoker Comparison	Odds Ratio	97.5% Confidence Interval
Nonsmoker vs. Former Smoker	0.631	[0.455, 0.871]
Nonsmoker vs. Smoker	0.373	[0.263, 0.525]

Table 9: Pairwise Odds Ratios Show Increased SCLC Risk in Smokers & Former Smokers

Discussion

Our findings suggest that both the rs1051730 1 and the rs8034191 2 SNPs are significant genetic contributors to the odds of developing SCLC. This is consistent with findings from other case-control studies for both SNPs (Amos et al., 2009; Hung et al., 2008; Shiraishi et al., 2008). More precisely, we found an association between the risk alleles of each SNP and greater odds of SCLC in what appeared to be an additive model, such that the odds ratio of the homozygous A genotype vs. the heterozygous genotype was below 1, and the odds ratio of the homozygous A genotype vs. the homozygous B genotype was less than the previous AA/AB odds ratio. With this robust evidence base for the rs1051730 1 and the rs8034191 2 SNPs as risk factors, biomedical researchers should direct more attention to the physiological mechanisms of these mutations which give rise to SCLC. More importantly, though, future therapeutic development should use these findings to develop sensitive diagnostic tools to detect these SNPs in the early stages of their proliferation.

Furthermore, we found evidence to support smoking as a likewise tiered risk factor for developing SCLC, also corroborated by other studies like Shiraishi & colleagues (2008). Further research should seek to disentangle whether smoking status is associated with the prevalence rs1051730 1 and the rs8034191 2 SNP mutations, and future SCLC prevention campaigns could emphasize smoking cessation for the general public as well as in more targeted interventions for the genetically predisposed.

Appendix

- Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y.-Y., Chen, W. V., Shete, S., Spitz, M. R., & Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics*, 40, 616-622. https://doi.org/10.1038/ng.109

- Hung, R. J., McKay, J. D., Gaborieau, V., et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, 452, 633-637. https://doi.org/10.1038/nature06885

- Shiraishi, K., Kohno, T., Kunitoh, H., Watanabe, S., Goto, K., Nishiwaki, Y., Shimada, Y., Hirose, H., Saito, I., Kuchiba, A., Yamamoto, S., & Yokota, J. (2008). Contribution of nicotine acetylcholine receptor polymorphisms to lung cancer risk in a smoking-independent manner in the Japanese. *Carcinogenesis*, 30(1), 65-70. https://doi.org/10.1093/carcin/bgn257

R Code

R Code for Case Study 2 Analysis

Listing 2: R code for analyzing case study 2 with genetic markers and smoking status

```
#install.packages('epitools')
library(epitools)

#### Case Study 2 ####
##### Case vs First Genetic Marker, rs8034191_2 #####
first   <- matrix(c(311, 508, 165, 413, 431, 125), nrow = 2, byrow = TRUE,
                  dimnames = list(Case = c('Case', 'Control'), Allele = c('AA', 'AB', 'BB')))
chisq.test(first)

# AB AA comparison (AA in first col)
ab_aa_comp = matrix(c(first[1,1], first[1,2], first[2,1], first[2,2]), nrow = 2)

# BB AA comp (AA in first col)
bb_aa_comp = matrix(c(first[1,1], first[1,3], first[2,1], first[2,3]), nrow = 2)

ab_bb_comp = matrix(c(first[2,1], first[2,3], first[2,1], first[2,3]), nrow = 2)

oddsratio(ab_aa_comp, conf.level = 0.975)

oddsratio(bb_aa_comp, conf.level = 0.975)

##### Case vs Second Genetic Marker, rs1051730_1 #####
secon   <- matrix(c(307, 495, 182, 404, 440, 125), nrow = 2, byrow = TRUE,
                  dimnames = list(Case = c('Case', 'Control'), Allele = c('AA', 'AB', 'BB')))
chisq.test(secon)

# AB AA comparison (AA in first col)
ab_aa_comp2nd = matrix(c(secon[1,1], secon[1,2], secon[2,1], secon[2,2]), nrow = 2)

# BB AA comp (AA in first col)
bb_aa_comp2nd = matrix(c(secon[1,1], secon[1,3], secon[2,1], secon[2,3]), nrow = 2)

oddsratio(ab_aa_comp2nd, conf.level = 0.975)

oddsratio(bb_aa_comp2nd, conf.level = 0.975)

##### Case vs Smoking Status #####
smoke   <- matrix(c(92, 502, 390, 161, 554, 254), nrow = 2, byrow = TRUE,
```

```
                    dimnames = list(Case = c('Case', 'Control'), SmokeStatus = c('Never', 'Form
chisq.test(smoke)

# nonsmoker quitter comparison (nonsmoker in first col)
nonsmoke_quit = matrix(c(smoke[1,1], smoke[1,2], smoke[2,1], smoke[2,2]), nrow = 2)

# nonsmoker smoker comparison (nonsmoker in first col)
nonsmoke_smoke = matrix(c(smoke[1,1], smoke[1,3], smoke[2,1], smoke[2,3]), nrow = 2)

oddsratio(nonsmoke_quit, conf.level = 0.975)
oddsratio(nonsmoke_smoke, conf.level = 0.975)
```