Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Exam #2 Key

Instructions: Please adhere to the following guidelines:

- This take-home exam is an individual effort, though it is open book/notes/calculator. In particular, you are not permitted to collaborate with any other individuals, either online or in-person (although you may consult any materials on the course webpage or on our Brightspace page). All questions regarding the exam should be directed to Andrew (andrew.spieker.vumc.org).
- I cannot physically stop you from consulting other online resources, although doing so is unnecessary and could lead to you incorrect information. If you have a question, you are better off asking me. If I am able to answer your question, I will answer it to the whole group (I will anonymize).
- Please read the questions carefully and answer only what you are being asked to answer.
- This looks like a long exam at first glance, but many of the problems should go quickly. Be concise. The *vast* majority of questions on this exam can be answered in one to three sentences; you'll notice throughout the exam that I repeatedly implore you to respond in no more than a sentence or two.
- Please round any final calculations to **three** significant digits! If reporting an odds/risk/rate/hazard ratio, please provide three digits beyond the decimal.
- There are six required problems and two optional problems; there are six pages of appendix material.
- The exam will be made available by 10:00am on Wednesday, March 25, and responses are due by email at 3:00pm on Thursday, March 26. For convenience and to minimize the amount of effort you need to spend with formatting, I have provided a template for your responses that you can use. You should word-process your solutions on the template and send them to andrew.spieker@vumc.org by the stated deadline. I expect this exam should take approximately two hours, but please spend no more than four hours on the required problems. Feel free to spend as much or as little time on the optional problems as you would like. I think they're kind of fun, but that's just me :).
- At the end of your exam solutions, please include whether you agree with the following statement: "On my honor, I have neither given nor received unauthorized aid on this exam." If you are unable to, please send me an email and we can discuss.

Score Range	**Frequency**
95+	5
90-94	5
85-89	3
80-84	3
75-79	2
Statistics	
Mean	88.5%
Median	90.0%
Corr($\overline{\text{HW}}_{1:5}$, Exam 2)	0.681
Corr(Exam 1, Exam 2)	0.791

These scores *do not* reflect the optional points, which will be accounted for at the end of the semester.

1. A case-control study of $N = 1{,}175$ men and women between the ages of 20 and 90 years is conducted to answer a clinical question of whether there is an (age-adjusted) association between history of tobacco consumption and esophageal cancer. The variables considered in this study are as follows:

$$X = \begin{cases} 0 & \text{if no history of tobacco consumption} \\ 1 & \text{if any history of tobacco consumption} \end{cases},$$

$$Z = \text{Age (years)},$$

$$Y = \begin{cases} 0 & \text{if not diagnosed with esophageal cancer} \\ 1 & \text{if diagnosed with esophageal cancer} \end{cases}.$$

Michael and Rebecca are at it again—will they ever agree on anything? Michael performs an analysis based on the following model:

$$\text{logit}\left(\text{P}(Y = 1|X = x, Z = z)\right) = \beta_0 + \beta_1 x + \beta_2 z. \tag{1}$$

Rebecca instead performs an analysis based on the following model:

$$\log\left(\text{P}(Y = 1|X = x, Z = z)\right) = \beta_0 + \beta_1 x + \beta_2 z. \tag{2}$$

(a) Which of the two models, (1) or (2), is better equipped to answer the clinical question in this study? In a maximum of two sentences, explain your response.

Ans: Model (1) is better equipped to answer the clinical question posed in this study. Recall that because of the outcome-dependent nature of the study, risk ratios cannot be modeled without external information, while analogous odds ratios of interest can. Since Model (1) is linear on the log-odds scale, $\exp(\widehat{\beta}_1)$ is consistent for the odds ratio that answers the clinical question.

(b) The Stata output from each of these models can be found in Appendix I. Provide a write-up of the results based on the output corresponding to the model you chose in part (a). Be sure to include a point estimate, a 95% confidence interval, a measure of strength of evidence, and a summary of your conclusions, minding both directionality and interpretation.

Ans: This study provides strong evidence of an age-adjusted association between tobacco consumption and subsequent esophageal cancer ($p < 0.001$). We estimate that those with any history of tobacco consumption have a 93.6% higher odds of esophageal cancer as compared to those with no such history of the same age. Based on a 95% confidence interval, these data would not be deemed unusual if in truth the odds were between 40.8% and 166% higher for the subgroup with a history of tobacco consumption.

OR: This study provides strong evidence of an age-adjusted association between tobacco consumption and subsequent esophageal cancer ($p < 0.001$). We estimate the odds ratio comparing the odds of esophageal cancer among those with any history of tobacco to the odds of esophageal cancer among those with no such history, but of the same age, to be 1.936. Based on a 95% confidence interval, these data would not be deemed unusual if in truth the odds ratio were between 1.408 and 2.663.

(c) Is it possible to use the results of the model you chose in part (a) to predict either the odds or risk of esophageal cancer among fifty-year old smokers? If it is possible to predict both, do so. If it is possible to predict one but not the other, predict the one you *can* predict and briefly explain why the other one cannot be predicted. If it is not possible to predict either, briefly explain why.

Ans: Neither is possible. Though we *can* estimate the odds *ratio* of interest under this study design out of mathematical convenience, the subgroup-specific risks or odds cannot themselves be identified. Note: If you can identify one, you can identify the other; there is a one-to-one correspondence between the risk and the odds.

(d) Michael and Rebecca both adjusted for age. In at most two sentences, describe the most likely reason why this was a good idea (irrespective of any other shortcomings of either of their models).

Ans: Age is likely associated with both tobacco consumption and with esophageal cancer (and does not lie on the causal pathway of interest—because age cannot be caused by smoking). Thus, having adjusted for age serves to account for its confounding impact.

(e) A genetic variant, W, is thought to be strongly associated with esophageal cancer (but not associated with tobacco consumption). If W were measured as part of the study, state whether you would prefer to adjust for it; in a maximum of two sentences, explain your response.

Ans: We would prefer to adjust for W owing to issues surrounding non-collapsibility of odds ratios. Failure to account for variables associated with Y tends to attenuate an association even when such variables are completely unrelated to the exposure of interest.

---

2. Consider the MRI study, a cohort study of $N = 735$ men and women over the age of 65, which includes the following variables:

$$X = \text{smoking history (in pack years);}$$

$$Z = \text{number of years since quitting smoking (0 if never smoked or if current smoker);}$$

$$Y = \begin{cases} 0 & \text{if no myocardial infarction} \\ 1 & \text{if myocardial infarction} \end{cases}.$$

The number of pack years ranges from 0 to 240. The following log-linear model is fit using Stata:

$$\log\left(\mathrm{P}(Y = 1 | X = x, Z = z)\right) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz;$$

the output is provided in Appendix II.

(a) Using plain language, provide the literal interpretations for $\exp(\beta_1)$, $\exp(\beta_2)$, and $\exp(\beta_3)$ in the context of this problem.

Ans: The interpretations are as follows:

- $\exp(\beta_1)$ corresponds to a risk ratio that compares the risk of myocardial infarction (MI) between current smokers differing in smoking history by one pack year (and it also corresponds to the risk ratio that compares current smokers with a history of one pack year to never-smokers).
- $\exp(\beta_2)$ corresponds to a risk ratio that compares the risk of MI between never-smokers differing in years since quitting smoking by one year. *Comment*: You may be scratching your head at this one; I only asked for the literal interpretations—I didn't ask if they all made sense! :) Because theoretically β_2 should be zero, some would remove it from the model. This is not a good idea, for the same reason that removing an intercept is not always a good idea even if it is thought to be theoretically zero.
- $\exp(\beta_3)$ corresponds to a ratio of risk ratios; the numerator corresponds to a risk ratio that compares the risk of MI between subgroups differing in their smoking history by one pack year but of a common time-since quitting smoking; the denominator is analogous but compares those with one fewer year since quitting smoking than the comparison in the numerator.

(b) In terms of the model parameters, provide an expression for the risk ratio that compares subgroups of current smokers that differ in their smoking history by five pack years, and then estimate it based on the output from Appendix II.

Ans: This risk ratio can be characterized by the expression $\exp(5\beta_1) = [\exp(\beta_1)]^5$. Based on the output, we have $[\exp(\widehat{\beta}_1)]^5 = 1.005545^5 = 1.028$.

(c) Citing appropriate output from Appendix II, state your conclusions regarding whether there is sufficient evidence that time since quitting smoking modifies the association between smoking history and myocardial infarction (maximum: one sentence).

Ans: This study does not provide sufficient evidence of such effect modification ($p = 0.414$).

(d) Citing appropriate output from Appendix II, state your conclusions regarding whether there is sufficient evidence of an overall association between smoking history and mycardial infarction (maximum: one sentence).

Ans: This study provides sufficient evidence of such an association ($p = 0.0073$).

---

3. A cohort study of $N = 87$ men and women is conducted to understand how the (gender-adjusted) prevalence of kidney stones varies by age. Let X denote age in years and Z denote gender (0: female; 1: male). Because people with a history of two kidney stone episodes are extraordinarily more likely to have recurring events as compared to people with a history of only one episode, the kidney stone outcome was considered categorically (rather than as a binary variable) as follows:

$$Y = \begin{cases} 0 & \text{if no kidney stone history} \\ 1 & \text{if only one prior kidney stone episode} \\ 2 & \text{if at least two prior kidney stone episodes} \end{cases}.$$

In these data, the proportions of subjects in each kidney stone category are 55.2%, 34.5%, and 10.3%, respectively. A multinomial logistic regression model is fit with $Y = 0$ chosen as the reference group; the output is provided in Appendix III (note that the untransformed coefficients are reported).

(a) Citing appropriate output from Appendix III, state your conclusions regarding the evidence of a gender-adjusted association between age and kidney stone history (maximum: one sentence).

Ans: This study provides sufficient evidence of such an association ($p = 0.0189$).

(b) Consider the following odds ratios, comparing the following quantities between subgroups differing in age by one year but of the same gender:

- Odds of only one prior kidney stone episode relative to no kidney stone history.
- Odds of at least two prior kidney stone episodes relative to no kidney stone history.
- Odds of at least two prior kidney stone episodes relative to only one prior kidney stone episode.

Estimate each of these odds ratios based on the reported coefficients provided in Appendix III.

Ans: The three odds ratios we are looking for are given by:

- $\exp(0.0627077) = 1.065$,
- $\exp(0.0748245) = 1.078$, and
- $\exp(0.0748245 - 0.0627077) = 1.012$.

(c) Can the first odds ratio in part (b) be used to approximate the *risk* ratio (comparing the risk of only one prior kidney stone episode between subgroups differing in age by one year but of the same gender)? Briefly justify your response.

Ans: No, it cannot. The odds ratio would approximate the risk ratio if *both* the comparator group were rare *and* the reference category were very common (the latter actually implies the former, but not the other way around). Neither of these things is true in this case, and so the risk ratio is not well approximated by the odds ratio.

(d) Based on the output provided, are you able state any conclusions regarding whether there is sufficient evidence of an age-adjusted association between gender and kidney stone history? If so, state those conclusions and cite appropriate output from Appendix III. Otherwise, explain why this is not possible in a maximum of two sentences.

Ans: No, we are not. We would need results from a joint test on all coefficients corresponding to gender, which are not provided to us.

(e) A collaborator suggests performing an analysis with Y treated ordinally rather than nominally. Ignoring potential challenges of doing multiple analyses of the same data set, state one potential advantage and one potential disadvantage of following your collaborator's advice.

Ans: One advantage is that it would take the ordered nature of the outcome into account; a disadvantage, however, is that the way the ordered nature is taken into account is through an assumption that may not be satisfied: namely, the proportional odds assumption, in which comparisons of odds across subgroups of the exposure are assumed to be the same whether you're comparing $\mathrm{P}(Y = 1)$ to $\mathrm{P}(Y = 0)$ or $\mathrm{P}(Y = 2)$ to $\mathrm{P}(Y = 1)$.

---

4. A preliminary laboratory study was conducted to evaluate doxorubicin as a potential chemotherapy agent, whereby doxorubicin was applied to $N = 282$ independent cell cultures of the same size at one of the following concentrations, X: 0.00, 0.05, 0.10, 0.50, 1.00, and 5.00 μmol/L. After a common incubation period, the total number of colonies, Y was measured. The total number of colonies ranged from about 0 to 250. Consider the following Poisson regression model:

$$\log \mathbf{E}[Y|X = x] = \beta_0 + \beta_1 \log x.$$

Note the log-transformation of doxorubicin in this problem.

(a) Briefly explaining your response, what key pieces of information are provided in the problem description that tell you that we do *not* require an offset term in the regression model?

Ans: We are told that the independent cell cultures are of the same size and measured after a common incubation period. Offsets are used to account for measurements that occur with variation in space-time, and no such variation is claimed to exist in this problem.

(b) State a meaningful interpretation for the quantity $\exp(\log(2)\beta_1)$. (*Hint*: Take this one step at a time. You know how to interpret coefficients corresponding to log-transformed predictors in linear models, and you know how to interpret exponentiated coefficients in Poisson models).

Ans: The expression $\exp(\log(2)\beta_1) = 2^{\beta_1}$ corresponds to the incidence rate ratio that compares the expected number of colonies between subgroups that differ by 100% (or a factor of two) in the applied doxorubicin concentration.

(c) Appendix IV contains Stata output from the regression model (note that the untransformed coefficients are reported). Based on the output, determine the concentration of doxorubicin at which a mean of 90 colonies are predicted to form post-incubation (*Hint*: If your answer does not lie between 0.00 and 5.00 μmol/L, check again).

Ans: We can solve backwards for the appropriate value of x:

$$\begin{aligned}
\log 90 &= \widehat{\beta}_0 + \widehat{\beta}_0 \log(x) \\
4.49980967 &= 3.747289 - 0.3661423 \log(x) \\
\log(x) &= (4.49980967 - 3.747289)/(-0.3661423) \\
\widehat{x}_{90} &= \exp((4.49980967 - 3.747289)/(-0.3661423)) = 0.128 \ \mu\text{mol/L}.
\end{aligned}$$

(d) Briefly state *two* advantages of having employed the "robust" option in this problem.

Ans: Employing the "robust" option allows us to obtain asymptotically valid standard errors even under possible misspecification of (1) the log-linear relationship between $\mathbf{E}[Y|X = x]$ and $\log(x)$, and/or (2) the presumed mean-variance relationship (namely, $\text{Var}(Y|X = x) = \mathbf{E}[Y|X = x]$).

---

5. Amoxicillin is a common treatment for acute otitis media (middle ear infection) in children. A cause for concern in the over-prescription of antibiotics is that it can result in resistant bacteria. A small study of $N = 20$ children was conducted to understand differences in time to symptoms resolving between individuals receiving antibiotics (`txgrp` $= 1$) and not receiving antibiotics (`txgrp` $= 0$). Ten children were randomized to each group. In the group receiving antibiotics, four children reported symptoms resolving on the fourth day, five reported having symptoms resolve on the sixth day, and one child still experienced symptoms at the ten-day mark (censored). In the group receiving no antibiotics, two children reported having symptoms resolve on the sixth day, six reported having symptoms resolve on the eighth day, and two children still experienced symptoms at the ten-day mark (censored). Kaplan-Meier curves are depicted for each treatment group in Appendix V, along with results from a log-rank test for equality for survival distributions.

(a) Estimate the proportion of children still experiencing symptoms at the one-week mark in each group.

Ans: $\widehat{S}_0(7) = 0.800$ and $\widehat{S}_1(7) = 0.100$.

(b) Estimate the proportion of children whose symptoms resolve within five days in each group.

Ans: $\widehat{F}_0(5) = 1 - \widehat{S}_0(5) = 1 - 1.000 = 0.000$ and $\widehat{F}_1(5) = 1 - \widehat{S}_1(5) = 1 - 0.600 = 0.400$.

(c) Estimate the median time to symptoms resolving in each group.

Ans: $\widehat{S}_0(t) = 0.5 \Rightarrow \widehat{t}_0^{\text{mdn}} = 8.00$ days, and $\widehat{S}_1(t) = 0.5 \Rightarrow \widehat{t}_1^{\text{mdn}} = 6.00$ days.

(d) Estimate the ten-day restricted mean time to symptoms resolving in each group.

Ans: We determine the area under the curve by dividing the respective regions into rectangles. $\widehat{\text{AUC}}_0(10) = 6 \times 1 + 2 \times 0.8 + 2 \times 0.2 = 8.00$ days, and $\widehat{\text{AUC}}_1(10) = 4 \times 1 + 2 \times 0.6 + 4 \times 0.1 = 5.60$ days.

(e) Estimate the nine-day cumulative hazard of symptoms resolving in each group.

Ans: $\widehat{\Lambda}_0(9) = -\log(\widehat{S}_0(9)) = -\log(0.2) = 1.61$, and $\widehat{\Lambda}_1(9) = -\log(\widehat{S}_1(9)) = -\log(0.1) = 2.30$.

(f) Using output from Appendix V, state (in one sentence) your conclusions regarding the association between amoxicillin and time to symptoms resolving.

Ans: Based on a log-rank test for equality of survival distributions, there is sufficient evidence of an association between use of amoxicillin and time to symptoms resolving ($p = 0.0096$).

(g) Despite your response to part (f), briefly explain how the Kaplan-Meier plot provides clinical evidence *against* prescribing antibiotics for acute otitis media in children.

Ans: Despite the strong statistical evidence of an an association, note that 80% of people in the control group report symptoms resolving by eight days, only two days later than the time at which 80% of people in the amoxicillin group report symptoms resolving. This suggests that most children may not require treatment with antibiotics for symptoms to resolve fairly quickly.

(h) Suppose it is later revealed that not all subjects started their randomized treatment on day zero, but some waited as long as four days to commence treatment. With which of the two statements do you most closely agree? No explanation is required.

(I) If we are most interested in conducting an intention-to-treat analysis, we can safely ignore this challenge altogether.

(II) It would be preferable to accommodate the time-varying nature of treatment so as not to potentially overstate the treatment effect by acting as if subjects were treated for longer than they actually were.

Ans: I agree more closely with Statement (II). Although no elaboration is required, the justification of the response is as follows. The idea of an intention-to-treat analysis is to compare people based on the treatment received at baseline. But what we're seeing in this problem is that the treatment randomized isn't even necessarily received at baseline but only some amount of time later. They're two separate issues that cannot/should not be conflated.

---

6. A chemotherapy agent is proposed in the hopes of improving survival time in a population of advanced-stage cancer patients. $N = 600$ subjects are randomized to receive either a placebo (`grp` $= 0$) or an experimental chemotherapy (`grp` $= 1$); the outcome of interest is time-to-death. The earliest censoring event occurs at four years. After ten years, the study ends and all remaining subjects are administratively censored. Appendix VI contains Kaplan-Meier curves for each group, a diagnostic plot, as well as a statistical test of the Schoenfeld residuals to gain insights into whether the proportional hazards assumption is met. Also included is output from an adjusted Cox model, adjusted for a binary indicator of overall baseline health (`basehealth`).

(a) In one sentence, describe how the Kaplan-Meier plot provides graphical evidence of a potential violation to the proportional hazards assumption.

Ans: The estimated survival curves grow pretty far apart before coming closer together, providing graphical evidence of a potential violation to the proportional hazards assumption.

(b) In one sentence, describe how the diagnostic plot adjacent to the Kaplan-Meier plot provides graphical evidence of a potential violation to the proportional hazards assumption.

Ans: Ignoring the noise at the earlier analysis times, the group-specific curves, which represent the negative log cumulative hazards in each group, are clearly *not* parallel, providing graphical evidence of a potential violation to the proportional hazards assumption.

(c) In one sentence, summarize the statistical evidence of a departure from proportional hazards.

Ans: A test of the Schoenfeld residuals yields a p-value of $p < 0.001$, providing strong statistical evidence of a departure from the proportional hazards assumption.

(d) Provide a write-up of the results based on the output from the Cox model. Be sure to include a point estimate, a 95% confidence interval, a measure of strength of evidence, and a summary of your conclusions, minding both directionality and interpretation.

Ans: This study provides strong evidence that the hazard of death differs between the control and experimental treatment groups ($p < 0.001$). We estimate that patients receiving the experimental chemotherapy have a 54.4% lower hazard of death as compared to control patients of the same baseline health status. Based on a 95% confidence interval, this estimate would not be deemed atypical if in truth the hazard were between 41.9% and 64.2% lower for the chemotherapy group.

OR: This study provides strong evidence that the hazard of death differs between the control and experimental treatment groups ($p < 0.001$). We estimate the hazard ratio, comparing the hazard of death between patients receiving the experimental chemotherapy and patients of the same baseline health status receiving the control treatment, to be 0.456. Based on a 95% confidence interval, this estimate would not be deemed atypical if in truth the hazard ratio were between 0.358 and 0.581.

(e) Your collaborator states that the evidence of a violation to proportional hazards you've alluded to in parts (a)-(c) discredits the evidence of an association between treatment and improved survival suggested by the Cox proportional hazards model. In a maximum of one sentence, explain why this argument is not sound.

Ans: This is not a sound argument because if there were no association whatsoever, the proportional hazards would be trivially satisfied; that is to say that evidence against the proportional hazards assumption is in fact evidence of some kind of association (and indeed, given that the Kaplan-Meier curve for the treatment group dominates that of the control group, the evidence in favor of a survival benefit for the experimental treatment is not particularly subtle).

(f) A question regarding whether the experimental chemotherapy improves survival can be answered using logistic regression (with the outcome being the indicator of death by a certain time, T_m). Briefly explain why $T_m = 3$ years could be used in these data but $T_m = 8$ years could not.

Ans: The first censoring event occurs at the four years, allowing us to use $T_m = 3 < 4$ but preventing us from using $T_m = 9 > 4$.

(g) Ignoring the challenges raised in (f), explain in a sentence why from a *clinical* perspective you might prefer to use $T_m = 8$ over $T_m = 3$.

Ans: Being able to claim evidence of a longer-term survival benefit is often more clinically compelling than being able to claim evidence of shorter-term survival benefit, particularly since eight-year the survival rate in the control group is still relatively high.

(h) Ignoring the challenges raised in (f), explain in a sentence why from a *statistical* perspective you might prefer to use $T_m = 8$ over $T_m = 3$.

Ans: All else being equal, maximal power would be achieved in a logistic model when about half the subjects experience an event; an analysis based on a cut-off of 8 years comes much closer to this benchmark than an analysis based on a cut-off of 3 years.

(i) Another collaborator suggests performing an analysis with progression-free survival as the outcome (i.e., time to either death or cancer progression, whichever comes first). Ignoring potential challenges of doing multiple analyses of the same data set, state one potential advantage and one potential disadvantage to following your collaborator's advice.

Ans: One advantage is the increase in power associated with having more events; one disadvantage is that when you have a composite outcome, it's hard to untangle where the effect is coming from—is it coming from the progression, or is it coming from survival differences?

(j) Suppose several subjects on the chemotherapy arm withdrew from the study at eight years because they were doing so well that they decided to go travel the world. In a sentence, describe how this insight might impact your ability to trust your results.

Ans: This appears to violate the assumption of non-informative censoring; the uncensored patients in the chemotherapy arm are not a representative sample cross-section at the time those subjects withdraw.

---

7. ***Optional problem 1***: This is an optional problem — do not attempt it until you have completed the rest of the exam. A small bonus can be earned from a correct response.

Despite appearances, this problem is actually not *that* bad. Give it a Recall that logistic, relative-risk, and Poisson regression all fall under the category of *generalized linear models* (GLMs), which entail two components: (1) a distribution for Y that follows a particular form, and (2) a linear relationship between $\boldsymbol{x}$ and some transformation of $\mathbf{E}[Y|\mathbf{X} = \boldsymbol{x}]$.

A simplified class of single-predictor GLMs supposes that Y has density function:

$$f(y;\theta) = h(y)\exp(y\theta - c(\theta)),$$

having mean $c'(\theta)$. We choose a function, g, to model the (conditional) mean Y, $\mu(x) = \mathbf{E}[Y|X = x]$:

$$g(\mu(x)) = \beta_0 + \beta_1 x.$$

Note: θ is referred to as the *natural parameter*, and g is referred to as the *link function*. Take, for instance, logistic regression, in which we model the conditional distribution of Y given $X = x$ as Bernoulli($p = \text{expit}(\beta_0 + \beta_1 x)$). I went on a minor tangent (what else is new?) about the fact that $g(\mu(x)) = \text{logit}(\mu(x))$ was a good choice for binary outcomes, but I didn't elaborate on *how* we would have arrived at that conclusion.

If you'd like to humor me, allow me to elaborate on that procedure now. Suppose you factor the density for Y into the form above. Then, it turns out that modeling θ linearly in x—or, equivalently, choosing $g^{-1}(\cdot) = c'(\cdot)$—results in a lot of mathematical simplifications that in turn lead to very good theoretical properties. This special choice of $g(\mu(x))$ is referred to as the *canonical link function*.

Let's see how this works. Suppose Y follows a Bernoulli(p) distribution so that its density is given by:

$$\begin{aligned} f(y;p) = p^y(1-p)^{1-y} &= \exp(y\log p + (1-y)\log(1-p)) = \exp(y(\log p - \log(1-p)) + \log(1-p)) \\ &= \exp(y\log(p/(1-p)) + \log(1-p)). \end{aligned}$$

Look how nicely that factored! Based on the factorization, we have $h(y) = 1$, and the natural parameter is given by $\theta = \log[p/(1-p)]$, the log odds! As such, $p = \text{expit}(\theta)$, and so with a small amount of algebra, we see that $c(\theta) = \log(1+\exp(\theta))$. In turn, $c'(\theta) = \text{expit}(\theta)$; hence, choosing $g(\mu(x)) = \text{logit}(\mu(x))$ turns out to be the choice that simplifies life!

If you've gotten this far, great! Now, assuming you're feeling sufficiently ambitious (and/or are sufficiently bored due to coronavirus-related isolation), follow a similar line of logic to learn why Poisson models are usually log-linear by default. Suppose Y follows a Poisson distribution with density:

$$f(y;\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!},$$

where λ is the rate parameter. Factor the density into the form $f(y;\theta) = h(y)\exp(y\theta - c(\theta))$. Based on your factorization, state $h(y)$, $\theta(\lambda)$, and $c(\theta)$; in turn, determine $c'(\theta)$. What do you notice? If you're stuck, you can always try to work backwards, since you know what the final answer should be.

Ans: First, we factor:

$$\begin{aligned} f(y;\lambda) &= \frac{1}{y!}\exp(\log(\lambda^y) - \lambda) \\ &= \frac{1}{y!}\exp(y\log(\lambda) - \lambda). \end{aligned}$$

Based on this factorization, we have that $h(y) = (y!)^{-1}$, $\theta(\lambda) = \log(\lambda)$, and $c(\theta) = \exp(\theta)$. We see, then, that $c'(\theta) = \exp(\theta)$, from which we derive the conclusion that $g(\mu(x)) = \log(\mu(x))$ is the canonical link.

---

8. ***Optional problem 2***: This is an optional problem — do not attempt it until you have completed the rest of the exam. A small bonus can be earned from a correct response.

Consider the REACH data set, `reach.csv`. Define a *controlled* A1c as an A1c of at most 7.0%. Perform an analysis to determine whether the REACH treatment is associated with a higher odds of a controlled six-month A1c as compared to control, adjusting for baseline A1c. Provide the Stata code you use to accomplish this. Your written response to this question should be **no more than four sentences**, and you should be able to perform this analysis in **at most ten lines of code**.

Ans: This study provides sufficient evidence that patients receiving REACH have a higher odds of a controlled A1c by six months as compared to patients receiving the control of the same baseline A1c ($p = 0.003$). We estimate the odds of a controlled A1c to be 103% higher in the REACH patients as compared to control patients of the same baseline A1c. Based on a 95% confidence interval, these data would not be judged unusual if in truth the odds of controlled A1c were between 27.4 and 224% higher in the REACH group.

Apart from reading in the data, this analysis could be accomplished with the following code:

```
gen ca1c6 = .
replace ca1c6 = 1 if a1c6 <= 7 & a1c6 != .
replace ca1c6 = 0 if a1c6 > 7 & a1c6 != .
logistic ca1c6 reach a1c0, robust
```

Many of you attempted this problem but did not properly account for missingness. When in doubt, you can always check the data to make sure that the way you coded it resulted in what you intended. In programming world, there's not always a concordance between intention and result.

Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis
Spring 2020
Exam #2

Appendix Material for Exam 2

APPENDIX I: Stata output for Problem 1

```
* MICHAEL'S ANALYSIS

. logistic esophcancer tobacco age, nolog robust

Logistic regression                             Number of obs     =      1,175
                                                Wald chi2(2)      =      88.39
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -500.01079               Pseudo R2         =     0.0673

------------------------------------------------------------------------------
             |               Robust
 esophcancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |   1.936062   .3148197     4.06   0.000      1.40769    2.662758
         age |   1.046899     .00569     8.43   0.000     1.035806    1.058111
       _cons |   .0108976   .0036411   -13.53   0.000     .0056614    .0209765
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.



* REBECCA'S ANALYSIS

. glm esophcancer tobacco age, family(binomial) link(log) nolog robust eform

Generalized linear models                         Number of obs   =      1,175
Optimization     : ML                             Residual df     =      1,172
                                                  Scale parameter =          1
Deviance         =  1005.286813                   (1/df) Deviance =   .8577533
Pearson          =  1101.169111                   (1/df) Pearson  =   .9395641

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   .8606696
Log pseudolikelihood = -502.6434067               BIC             =  -7279.609

------------------------------------------------------------------------------
             |               Robust
 esophcancer | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |   1.636773   .2118367     3.81   0.000     1.270056    2.109375
         age |   1.033968   .0040028     8.63   0.000     1.026152    1.041843
       _cons |   .0194104   .0047964   -15.95   0.000     .0119591    .0315042
------------------------------------------------------------------------------
Note: _cons estimates baseline risk.
```

APPENDIX II: Stata output for Problem 2

```
. glm myo c.yrsquit##c.packyrs, family(binomial) link(log) nolog robust eform

Generalized linear models                         Number of obs   =        734
Optimization     : ML                             Residual df     =        730
                                                  Scale parameter =          1
Deviance         =  536.7011719                   (1/df) Deviance =   .7352071
Pearson          =  729.7460431                   (1/df) Pearson  =   .9996521

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   .7420997
Log pseudolikelihood =  -268.350586               BIC             =   -4280.21

------------------------------------------------------------------------------
                     |               Robust
                 myo | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------------+--------------------------------------------------------------
            packyrs |   1.005545   .0030536     1.82   0.069     .9995783    1.011548
            yrsquit |   1.010497   .0070044     1.51   0.132     .9968618    1.024319
                    |
c.packyrs#c.yrsquit |   1.000121   .0001481     0.82   0.414     .9998306    1.000411
                    |
              _cons |   .0930322   .0136154   -16.23   0.000     .0698327    .1239389
------------------------------------------------------------------------------
Note: _cons estimates baseline risk.



. testparm packyrs c.packyrs#c.yrsquit

 ( 1)  [myo]packyrs = 0
 ( 2)  [myo]c.packyrs#c.yrsquit = 0

           chi2(  2) =    9.83
         Prob > chi2 =    0.0073
```

APPENDIX III: Stata output for Problem 3

```
. mlogit kidney age gender, nolog robust

Multinomial logistic regression                 Number of obs     =         87
                                                Wald chi2(4)      =      10.42
                                                Prob > chi2       =     0.0340
Log pseudolikelihood = -76.439054               Pseudo R2         =     0.0604

------------------------------------------------------------------------------
             |               Robust
      kidney |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |  (base outcome)
-------------+----------------------------------------------------------------
1            |
         age |   .0627077   .0272573     2.30   0.021     .0092843     .116131
      gender |  -.8255368   .4917756    -1.68   0.093    -1.789399    .1383256
       _cons |  -4.616693   2.025362    -2.28   0.023    -8.586329   -.6470565
-------------+----------------------------------------------------------------
2            |
         age |   .0748245   .0321393     2.33   0.020     .0118326    .1378164
      gender |  -.6018081   .7538408    -0.80   0.425    -2.079309    .8756928
       _cons |  -6.869637   2.428045    -2.83   0.005    -11.62852   -2.110756
------------------------------------------------------------------------------



.  testparm age

 ( 1)  [0]o.age = 0
 ( 2)  [1]age = 0
 ( 3)  [2]age = 0
       Constraint 1 dropped

           chi2(  2) =    7.94
         Prob > chi2 =    0.0189
```

APPENDIX IV: Stata output for Problem 4

```
. poisson count logdox, nolog robust

Poisson regression                              Number of obs     =        282
                                                Wald chi2(1)      =     708.10
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -4431.8274               Pseudo R2         =     0.6242

------------------------------------------------------------------------------
             |               Robust
       count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logdox |  -.3661423   .0137595   -26.61   0.000    -.3931104   -.3391742
       _cons |   3.747289   .0553973    67.64   0.000     3.638713    3.855866
------------------------------------------------------------------------------
```

APPENDIX V: Stata output for Problem 5

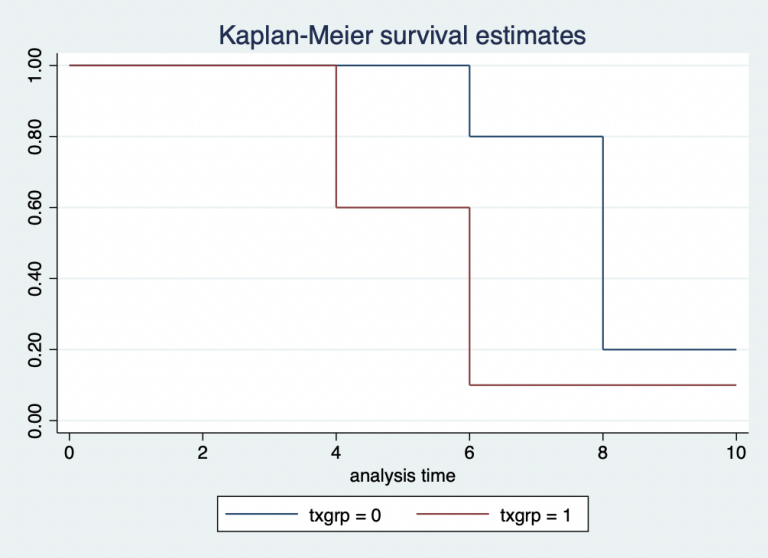

Figure 1: Kaplan-Meier curves for time to symptoms resolving in each treatment group (0: no antibiotics; 1: antibiotics).

```
. sts test grp

         failure _d:  death
   analysis time _t:  tte

Log-rank test for equality of survivor functions

      |   Events         Events
txgrp |  observed       expected
------+-------------------------
0     |         8          11.71
1     |         9           5.29
------+-------------------------
Total |        17          17.00

            chi2(1) =       6.71
            Pr>chi2 =     0.0096
```

APPENDIX VI: Stata output for Problem 6

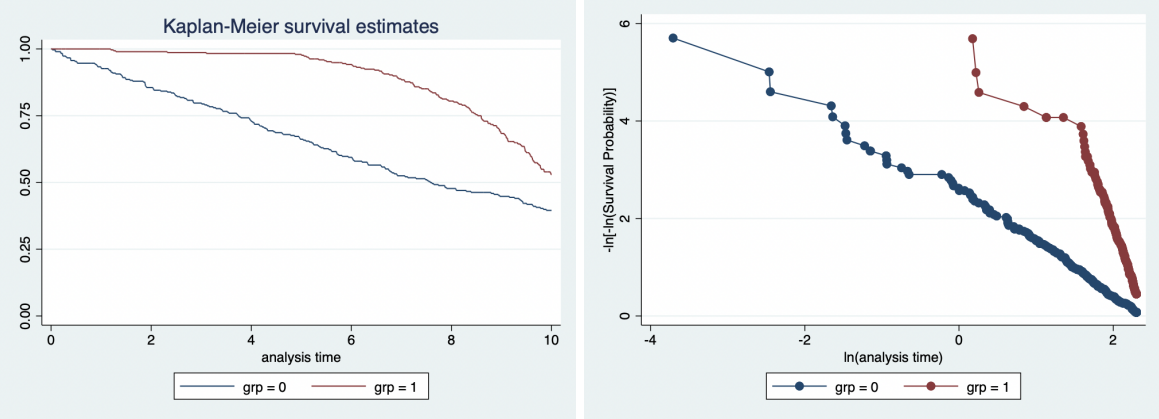

Figure 2: Left: Kaplan-Meier curves for each group. Right: $\log(t)$ against $-\log(-\log(\widehat{S}(t|X = x)))$ for each group.

```
. estat phtest

      Test of proportional-hazards assumption

      Time:  Time
      ----------------------------------------------------------------
                  |                   chi2       df       Prob>chi2
      ------------+---------------------------------------------------
      global test |                 100.90        2          0.0000
      ----------------------------------------------------------------



. stcox grp basehealth, nolog robust

        failure _d:  death
  analysis time _t:  tte

Cox regression -- no ties

No. of subjects      =          600            Number of obs    =         600
No. of failures      =          289
Time at risk         =  4456.669702
                                                Wald chi2(2)     =       70.39
Log pseudolikelihood =   -1682.6913            Prob > chi2      =      0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         grp |    .456256   .0561764    -6.37   0.000     .3584303    .5807811
  basehealth |   .3503929   .0525054    -7.00   0.000     .2612192    .4700082
------------------------------------------------------------------------------
```