Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2021)
Exam #1 Key: Median=88 (IQR=[83-94])

Instructions: Please adhere to the following guidelines:

- Your responses to this exam are due by email at **1:00p** on **Friday, March 26**. I intend for this cutoff to be strict. You should not need anywhere near the full amount of time to complete the exam, but I am building in a generous cushion to allow for, e.g., technology issues. For your convenience—and to minimize the time and effort you need to spend on formatting—I have provided a template for your responses on which you should word-process your solutions and send to andrew.spieker@vumc.org by the deadline. Please label your solutions according to the format "LASTNAME-EXAM.docx." I will confirm receipt of your exam. If you want to be cautious, you are welcome to e-mail exam drafts along the way (I will not start grading until after 1:00p on March 26, so there is no risk in sending drafts. Don't worry about spamming my inbox; it won't bother me at all—I'll just grab the most recent copy at the deadline). I intend to score the exams, post the key, and provide detailed feedback to you by Monday, March 29.
- There are five required problems (each with multiple sub-questions of varying length and difficulty) and two optional problems; there are five pages of appendix material. You are *not* expected to utilize software for any of these problems—all software output provided in this exam is sufficient to answer the questions.
- Please read the instructions very carefully; answer no more and no less than what you are being asked to answer. You'll notice throughout the exam that I repeatedly implore you to be concise. There are a lot of sub-questions on this exam, many of which should go relatively quickly and should only require very brief responses. My strong recommendation is to provide your responses to the problems you find easiest first, and then return to the more challenging ones.
- This exam is open book/notes/calculator, but is an **individual effort**. You are not permitted to collaborate with individuals inside or outside the class, in-person, electronically, telepathically, or otherwise. You of course *may* consult any course materials including those on the course webpage. I cannot stop you from consulting other online resources, although doing so should not be necessary.
- All questions regarding the exam should be directed to Andrew (andrew.spieker@vumc.org). If I am able to respond to your question, I will provide my response to the whole group (and I will of course anonymize your question). If I am unable to respond, I will let you know.
- Upon completion of your exam, please indicate on the first page of the template whether you agree with the following statement: "On my honor, I have neither given nor received unauthorized aid on this exam." If you have concerns about your ability to answer this in the affirmative, please turn in your exam anyway, and send me an email so we can discuss.
- Please round any final calculations to a reasonable number of significant digits!
- As is always the case in this class, any reference to logarithmic transformations are based on the *natural* logarithm (i.e., having base e).
- **Importantly**: Take a deep breath — you've got this! This is an opportunity to showcase all of the hard work you've done so far this semester.

---

1. [20 pts] A study was conducted of $N = 975$ independently sampled children between ages 5 and 10 years with asthma. The study involved multiple goals, two of which we will discuss and explore in this problem.

The primary study objective was to investigate whether average time participating in physical activities (measured by average hours spent involved in physical activities per week over a six-week period; `physact`) varied across levels of asthma severity (measured by baseline forced expiratory volume; `fev`, in L). Consider the following simple linear regression model, used to address this goal:

$$\mathbf{E}[\texttt{physact}|\texttt{fev}] = \alpha_0 + \alpha_1 \texttt{fev} \qquad \textbf{(Model 1)}$$

One of the secondary study goals was to establish "normal ranges" of physical activity for children—that is, to formulate prediction intervals. To that end, consider the following simple linear regression model, which differs from that of Model 1 in that the predictor has undergone a log-transformation:

$$\mathbf{E}[\texttt{physact}|\texttt{fev}] = \beta_0 + \beta_1 \log(\texttt{fev}) \qquad \textbf{(Model 2)}$$

Stata output for Models 1 and 2 Are presented in Appendices I and II, respectively; each includes two diagnostic plots (a predictor-versus-residual plot and a normal quantile-quantile plot).

Parts (a) and (b) pertain specifically to the *primary* study objective.

(a) State a literal interpretation for α_0 in plain language. To what degree is this interpretation meaningful in the real world?

Ans: α_0 denotes the mean hours of physical activities per week among those with an FEV of zero liters. This is *not* a real-world interpretation but instead a literal one implied by the model.

(b) State an interpretation for α_1 in plain language.

Ans: α_1 denotes the difference in mean hours spent involved in physical activities per week between subgroups differing in their FEV by one liter.

i. State any assumptions necessary in order to estimate and conduct inference on α_1 in a way that is valid in large samples.

Ans: We would need the errors to be pairwise independent (or uncorrelated), and for linearity (of mean physical activity in FEV) to be (at least approximately) satisfied. You may also note that we are formally assuming finite error variance. It is already given to us that the number of model coefficients is far lower than the number of independently sampled subjects.

ii. Briefly discuss (maximum of three sentences) the extent to which the information provided in Appendix I demonstrates (or fails to demonstrate) evidence of violations to these assumptions.

Ans: The LOWESS smoother in the predictor-versus-residual plot shows a non-negligible departure from the line $y = 0$, meaning that we have graphical evidence of a departure from the linearity assumption. As discussed in class, independence/uncorrelatedness of the error terms is something that is given to us by the context of the study design; the diagnostic plots provide no insights into the degree to which this assumption is satisfied or violated. The problem description told us that the subjects were sampled independently.

iii. Briefly discuss (maximum of three sentences) the degree to which you believe the evidence of assumption violations you've noted in part (b) *severely* compromises your ability to address the study's primary goal.

Ans: Recall that the "slope" coefficient of a simple linear regression can be understood as a weighted average of the slopes of the segments that connect each pair of points. This characterization does not rely upon linearity. Even in light of the evidence of a departure from linearity, it is reasonable to utilize simple linear regression to obtain insights into whether there is a "first-order trend" relating FEV and mean physical activity.

Parts (c) and (d) pertain specifically to the *secondary* study objective.

(c) Compare the information provided in Appendices I and II; identify and briefly describe what information in these appendices makes it evident that Model 2 appears better equipped to accomplish the secondary goal as compared to Model 1.

Ans: The LOWESS smoother in the predictor-versus-residual plot provided in Appendix II lies much closer to the line $y = 0$, suggesting that the linearity assumption—which, incidentally, is much more important for achieving the secondary study objective—is better satisfied in the model that log-transforms the exposure.

(d) Using the output from Model 2, form a naive prediction interval that would be estimated to encompass 95% of the distribution of the physical activity outcome among individuals with an FEV of 2.00.

Ans: The estimated model is given by:

$$\widehat{Y}(x) = \widehat{\mathbf{E}}[Y|X = x] = 2.031183 + 1.041884 \times \log(x).$$

plugging in $x = 2$ allows us to obtain our predicted value of $\widehat{y}(2) = 2.75336$. The RMSE is given by $\widehat{\sigma} = 0.40189$, and hence the naive prediction interval is given by:

$$\widehat{y} \pm 1.96 \times \widehat{\sigma} = 2.75336 \pm 1.96 \times 0.40189 = [1.97, 3.54].$$

Now, suppose that the study data also included information on each child's age.

(e) Briefly discuss (maximum of two sentences) *possible* advantages that could be gained in addressing the primary study objective from a model that *adjusts* for age.

Ans: I imagine that age could be a possible confounder (e.g., a common cause of the predictor and the outcome). Adjustment for confounding is one approach that allows us to uncover causal relationships between variables.

---

2. [10 pts] **Background**: One way in which vaccine candidates are screened in early stages of research is through comparisons of serologic immunogenicity markers (i.e., measures that pertain to identification of antibodies in the blood serum). The *antibody titer* is one such serologic measure, and is obtained by sequentially diluting a serum sample and testing each dilution for the antibody of interest; a participant's titer is defined to be the relative concentration of the *final* dilution that responds to an antibody test (higher titer values are indicative of a greater concentration of specific antibodies in the blood and are therefore more desirable). A vaccine is considered a good candidate if, after a certain period of time post-vaccination, subjects receiving the vaccine tend to have higher titers as compared to a control condition. Owing to the fact that titers are defined multiplicatively, it is standard to log-transform titer values and compare groups on the basis of the geometric mean titer.

A double-blind, placebo-controlled randomized phase-II trial of $N = 184$ subjects is conducted to evaluate the immunogenicity associated with a booster vaccine for protection against SARS-CoV-2, the virus responsible for COVID-19, among previously vaccinated subjects. In this study, the initial dilution of the original serum sample occurs at a ratio of 1:10, and the sample is diluted by a factor of two until response (or until a maximum of seven subsequent dilutions). Anyone whose serum does not respond to an antibody test at the initial concentration of 1:10 is given a titer value of 5, and anyone responding to a relative concentration of 1:1280 is given a titer value of 1280. Therefore, to be totally clear, the titer values in this study can take on the values 5, 10, 20, 40, 80, 160, 320, 640, and 1280. You may assume that titers are correlated within subjects over time. The variables collected in this study are provided below:

`grp`	treatment assignment (0 = control; 1 = SARS-CoV-2 candidate vaccine)
`logpretiter`	log-transformed pre-booster titer value
`logtiter`	log-transformed titer value three weeks following the booster

Consider the following two linear regression models, each of which you may freely assume to be correctly specified for the purposes of this problem.

$$\begin{aligned} \mathbf{E}[\texttt{logtiter}|\texttt{grp}] &= \beta_0^* + \beta_1^*\texttt{grp} && \text{(unadjusted)} \\ \mathbf{E}[\texttt{logtiter}|\texttt{grp},\texttt{logpretiter}] &= \beta_0 + \beta_1\texttt{grp} + \beta_2\texttt{logpretiter} && \text{(adjusted)} \end{aligned}$$

(a) Briefly justifying your response, what can you say about how the true values of the model parameters β_1^* and β_1 relate to one another?

Ans: Because this is a randomized trial, pre-booster titer value is not systematically associated with randomization group. Mean differences, characterized by parameters in linear regression models, are *collapsible*; therefore, β_1 and β_1^* should be equal.

(b) If you had to choose only *one* of the two models above to fit *a priori* (meaning, before actually conducting the analysis and having the opportunity to view the results), which would you choose? Briefly justify your response.

Ans: We are given that titer values are correlated within subjects; thus, the log-transformed pre-booster titer value would be associated with the post-booster value. Since this is a randomized trial, pre-booster titer value is not systematically associated with randomization group (see response to part (a)). This fits the classic description of a precision variable. Adjustment for pre-booster titer should improve efficiency/precision of estimation for the vaccine's effect (meaning, adjustment should reduce the variance). I would therefore choose the adjusted model.

(c) In Appendix III, you are provided with Stata output for each of the above models. Based on your response to part (b), write a short (3-5 sentence) summary of the study results in the usual way that we have done in this class. Remember that for the purposes of this exam, it is totally legitimate to use homework keys and course notes as resources.

Ans: This study provides sufficient evidence of an association between vaccination group and post-booster titers (p = 0.041). Considering subgroups of the same pre-booster titer values but differing in their vaccine group, we estimate the geometric mean titer post-booster to be 25.3% higher in those receiving the SARS-CoV-2 vaccine booster as compared to the control group. Based on a 95% confidence interval, this estimate would not be considered surprising if in truth the geometric mean were between 0.938% and 55.5% higher in the active vaccine group.

---

3. [20 pts] This problem is a continuation of Problem 2. Titers suffer from inherent measurement error in the way they are derived. The easiest way to realize this is by example: take, for instance, someone whose serum *would* respond until a relative concentration of 1:159. Because of the sequential dilution procedure, that individual would be observed to have a titer value of 80 (because 1:80 reflects the maximum dilution *performed* at which the serum would be observed to respond), even though his or her *true* titer value is very clearly better reflected by 160. For better or for worse, clinical researchers have mostly responded to this dilemma by dichotomizing outcomes in one of several ways. One such way is to evaluate whether subjects have at least a four-fold rise in their post-booster titer from their pre-booster titer (this is typically referred to as *seroconversion*). For this problem, the variables under consideration are as follows:

`grp`	treatment assignment (0 = control; 1 = SARS-CoV-2 candidate vaccine)
`logpretiter`	log-transformed pre-booster titer value
`seroconversion`	indicator of at least four-fold rise in titer (0 = no; 1 = yes)

Consider the following two logistic regression models, each of which you may freely assume to be correctly specified for the purposes of this problem.

$$\text{logit}(\text{P}(\texttt{seroconversion} = 1|\texttt{grp})) = \alpha_0^* + \alpha_1^*\texttt{grp} \quad \text{(unadjusted)}$$

$$\text{logit}(\text{P}(\texttt{seroconversion} = 1|\texttt{grp}, \texttt{logpretiter})) = \alpha_0 + \alpha_1\texttt{grp} + \alpha_2\texttt{logpretiter} \quad \text{(adjusted)}$$

(a) If you had to choose only *one* of the above models to fit *a priori* (meaning, before actually conducting the analysis and having the opportunity to view the results), which would you choose? Briefly justify your response.

Ans: I would still prefer to use the adjusted model, but for a different reason than described in the previous problem. Here, we note that because of non-collapsibility of the logit-link, the value of α_1 will be further away from the null value of zero as compared to α_1^*.

(b) In Appendix IV, you are provided with Stata output for each model. Based on your response to part (a), write a short (3-5 sentence) summary of the study results in the usual way that we have done in this class. Remember that for the purposes of this exam, it is totally legitimate to use homework keys and course notes as resources.

Ans: This study does not provide sufficient evidence that the SARS-CoV-2 booster is associated with odds of subsequent seroconversion ($p = 0.162$). Comparing subgroups of the same pre-booster titer values, but differing in their vaccination group, we estimate the odds of seroconversion to be 96.8% higher in the group receiving the SARS-CoV-2 vaccine booster as compared to the control group. Based on a 95% confidence interval, these data would not be judged unusual if in truth the the odds were between 23.8% lower and 409% higher in the active vaccine group.

(c) Regardless of your answers to parts (a)-(b), consider the 2×2 table below:

	Seroconversion	No seroconversion
Vaccine	a	b
Control	c	d

Despite the fact that you do not know the specific values of a, b, c, or d, use information provided in Appendix IV to state the value of the estiamted odds ratio, $\widehat{\text{OR}} = (a \times d)/(b \times c)$.

Ans: The estimated crude/unadjusted odds ratio should be 1.19, which was obtained by exponentiating the coefficient from the unadjusted model. We know this because simple logistic models with binary predictors are *saturated*, in that the linearity structure on the log-odds scale is trivially satisfied.

(d) Regardless of your answers to parts (a)-(b), is it possible to use the information in Appendix IV to estimate the odds of seroconversion among control subjects with a pre-booster titer value of 80? If so, do so; if not, briefly explain why not.

Ans: Yes, and this value is given by $\exp(-3.577783 + 1.511318 \times \log(80)) \approx 21.0$.

(e) Regardless of your answers to parts (a)-(b), is it possible to use the information in Appendix IV to estimate the probability of seroconversion among vaccine-candidate subjects with a pre-booster titer value of 160? If so, do so; if not, briefly explain why not.

Ans: Yes, and this value is given by $\text{expit}(-3.577783 + 0.677033 + 1.511318 \times \log(160)) \approx 0.9916$.

(f) Regardless of your answers to parts (a)-(b), is it possible to use the information in Appendix IV to determine a point estimate and 95% confidence interval for the the odds ratio that compares the odds of seroconversion between subgroups belonging to the same vaccination group but differing in their pre-booster titer value by a factor of two (or, to put it another way, by 100%)? If so, do so; if not, briefly explain why not.

Ans: Yes; the estimate is given by $2^{1.511318} = 2.85$, and the confidence interval should be given by $[2^{1.007523}, 2^{2.015112}] = [2.01, 4.04]$.

---

4. [20 pts] A case-control study was conducted among $N = 362$ prior or current smokers between the ages of 50 and 75 previously diagnosed with lung cancer to evaluate whether reducing or quitting smoking after diagnosis is associated with a lower odds of cancer recurrence. In particular, $n_0 = 180$ subjects were sampled whose cancer *did not* recur within a year after having been diagnosed and having gone into remission; $n_1 = 182$ subjects were sampled whose cancer *did* recur within a year after having been diagnosed and gone into remission. Enrolled subjects responded to a survey question regarding their smoking habits following cancer diagnosis. You may assume that in this study population, recurrence within a year is rare. The investigators of this study decided to adjust for age and allow an interaction between age and smoking status. The variables evaluated in this study are as follows:

`rec`	cancer recurrence within one year of remission (0 = no; 1 = yes)
`smk`	(0 = no change after diagnosis; 1 = reduced after diagnosis; 2 = quit after diagnosis)
`age`	age (years)

The investigators propose the following model:

$$\text{logit}(\text{P}(\texttt{rec}|\texttt{smk}, \texttt{age})) = \beta_0 + \beta_1 \texttt{age} + \beta_2 1(\texttt{smk=1}) + \beta_3 1(\texttt{smk=2}) + \beta_4 1(\texttt{smk=1}) \times \texttt{age} + \beta_5 1(\texttt{smk=2}) \times \texttt{age}.$$

The output for this model is presented in Appendix V, along with results of two hypothesis tests.

(a) Four scientific goals related to the study are listed below, and each could be addressed by testing some subset or combination of parameters of the above model (β_0 through β_5). For each of the four goals, first express a corresponding hypothesis test in terms of the model parameters. Then, examine the output provided in Appendix V; if you can use any of the provided output to summarize the statistical strength of evidence pertaining to the goal, do so (maximum: one sentence for each goal). If this cannot be done, simply say that not enough information is provided in Appendix V to learn about the result of the hypothesis test.

i. **Goal A**: To evaluate whether there is evidence of an overall association between post-diagnosis change in smoking habits and odds of cancer recurrence.

Ans: The principle is to conduct a joint test of parameters involving smoking. The hypothesis test is given by $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs. H_1 : not H_0. Not enough information is provided in Appendix V to learn about the result of this test.

ii. **Goal B**: To evaluate whether there is evidence that age modifies the association between change in smoking habits and odds of cancer recurrence.

Ans: The principle is to conduct a joint test of parameters involving smoking interactions. The hypothesis test is given by $H_0 : \beta_4 = \beta_5 = 0$ vs H_1 : not H_0. There is not sufficient evidence that age modifies the association between change in smoking and odds of cancer ($p = 0.38$).

iii. **Goal C**: To evaluate whether there is evidence of an association between age and odds of cancer recurrence among those reporting not changing their smoking habits after diagnosis.

Ans: The principle is to find the reduced model among those reporting not changing smoking status, and then see which parameters correspond to a test of the association of interest. The reduced model is given by $\text{logit}(\text{P}(\texttt{rec}|\texttt{smk=0}, \texttt{age})) = \beta_0 + \beta_1 \texttt{age}$. The hypothesis test is given by $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. There is not sufficient evidence of an association between age and odds of cancer recurrence among those reporting not changing their smoking habits after diagnosis ($p = 0.17$).

iv. **Goal D**: To evaluate whether there is evidence of an association between age and odds of cancer recurrence among those reporting reducing their smoking habits after diagnosis.

Ans: The principle is to find the reduced model among those reporting reducing their smoking habits and then see which parameters correspond to a test of the association of interest. The reduced model is given by $\text{logit}(\text{P}(\texttt{rec}|\texttt{smk=1}, \texttt{age}) = \beta_0 + \beta_1\texttt{age} + \beta_2 + \beta_4\texttt{age} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)\texttt{age}$. The hypothesis test is given by $H_0 : \beta_1 + \beta_4 = 0$. There is not sufficient evidence of an association between age and odds of cancer recurrence among those reporting reducing their smoking habits after diagnosis ($p = 0.36$).

(b) Is it possible to use the information in Appendix V to estimate the odds of cancer recurrence among individuals 68 years old who had reported quitting smoking after diagnosis? If so, do so; if not, briefly explain why not.

Ans: No, because this is a case-control study. The outcome-dependent sample scheme does not allow us to estimate subgroup-specific odds (or risks) without external information.

(c) Is it possible to use the information in Appendix V to obtain a point estimate and 95% confidence interval for the odds ratio comparing the odds of cancer recurrence between subgroups differing in age by *two years* and reporting not changing their smoking habits after diagnosis? If so, do so; if not, briefly explain why not.

Ans: Yes; the estimate is given by $\exp(2 \times 0.0685194) = 1.147$, and the confidence interval should be given by $[\exp(2 \times (-0.0290891)), \exp(2 \times 0.1661278)] = [0.9435, 1.394]$.

(d) Is it possible to use the information in Appendix V to approximate the risk ratio comparing the risk of cancer recurrence between subgroups differing in age by one year and reporting reducing smoking after diagnosis? If so, do so *and state what information allows you to do this*; if not, briefly explain why not.

Ans: Yes, and the reason is because the outcome is stated to be rare—hence, the odds ratio should approximate the risk ratio, and is estimated as $\exp(0.0685194 - 0.0403019) = 1.0286$.

---

5. [30 pts] This question is about translating "biostatistics language" into concrete terms. For each subquestion, you are provided a variable description and a model, and are asked to use *plain language* to interpret: (i) a specific (possibly transformed) parameter, and (ii) a formal null hypothesis. You are *not* to include notation of regression models (e.g., $\mathbf{E}[Y|X = x]$); however, you *are* permitted to use numerals/percentages, or refer to variables that clarify subgroups (statements such as "among those with $X = 5$," or "between subgroups differing in X by 10%" are acceptable). See the sample question/response below.

SAMPLE: Consider a linear regression model with binary (0/1) X and continuous Y:

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

i. Provide a plain-language interpretation for β_0.
Ans: This is the mean value of Y among the subgroup defined by $X = 0$.

ii. Provide a plain-language interpretation for the null hypothesis $H_0 : \beta_1 = 0$.
Ans: The null hypothesis is that the difference in mean Y between groups ($X = 0$ and $X = 1$) is zero.

(a) Consider a linear regression model with continuous X, Z, and Y:

$$\mathbf{E}[Y|X = x, Z = z] \quad = \quad \beta_0 + \beta_1(x - \overline{x}) + \beta_2 z + \beta_3(x - \overline{x})z.$$

i. Provide a plain-language interpretation for β_2.

Ans: This is the difference in the mean value of Y that compares subgroups differing in Z by one unit but with $X = \overline{x}$.

ii. Provide a plain-language interpretation for the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$.

Ans: The null hypothesis is that Z does not have an overall (linear) association with mean Y in any subgroup of X.

(b) Consider a logistic regression model with a continuous, positive-valued X and a binary (0/1) Y:

$$\text{logit}(\text{P}(Y = 1|X = x)) \quad = \quad \beta_0 + \beta_1 \log(x).$$

i. Provide a plain-language interpretation for $e^{\log(1.05)\beta_1} = 1.05^{\beta_1}$.

Ans: This is the odds ratio that compares the odds of $Y = 1$ between subgroups differing in their value of X by 5%.

ii. Provide a plain-language interpretation for the null hypothesis $H_0 : \beta_0 = 0$.

Ans: The null hypothesis is that the odds of $Y = 1$ would be equal to one among the subgroup with $X = 1$ (alternatively, that the proportion with $Y = 1$ would be 0.5 among the subgroup with $X = 1$).

(c) Consider a multinomial model with three-category values for both X and Y (0, 1, or 2).

$$\log\left(\frac{\mathrm{P}(Y=j|X=x)}{\mathrm{P}(Y=0|X=x)}\right) = \beta_{0j} + \beta_{1j}1(x=1) + \beta_{2j}1(x=2),\ j=1,2.$$

i. Provide a plain-language interpretation for $e^{\beta_{12}}$.

Ans: This can be interpreted as a ratio of risk ratios; each of the risk ratios is comparing the proportion having the outcome $Y=2$ to the reference category of $Y=0$ in specific subgroups of X; the specific subgroups of X being compared are $X=1$ and $X=0$.

ii. Provide a plain-language interpretation for the null hypothesis $H_0: \beta_{21} - \beta_{11} = 0$.

Ans: The null hypothesis is that the risk ratios that compares the proportion having the outcome of $Y=1$ to the proportion having the outcome $Y=0$ are the same in the subgroups defined by $X=1$ and $X=2$.

(d) Consider a proportional odds model with a discrete, ordinal Y and continuous X, Z, and W:

$$\log\left(\frac{\mathrm{P}(Y\le j|X=x,Z=z,W=w)}{\mathrm{P}(Y>j|X=x,Z=z,W=w)}\right) = \beta_{0j} - \beta_1 x - \beta_2 z - \beta_3 w - \beta_4 xz - \beta_5 xw - \beta_6 zw - \beta_7 xzw,\ j=0,1.$$

i. Provide a plain-language interpretation for e^{β_2}.

Ans: This is the odds ratio that compares, for all j, the odds of $Y>j$ between subgroups differing in Z by one unit but with $W=0$ and $X=0$.

ii. Provide a plain-language interpretation for the null hypothesis $H_0: \beta_3 = \beta_5 = \beta_6 = \beta_7 = 0$.

Ans: The null hypothesis is that, for all j, the odds of $Y>j$ does not vary (multiplicatively) across the values of W for any subgroups defined by their values of X and Z.

(e) Consider a Poisson model with non-negative discrete Y and binary (0/1) X and Z:

$$\log\left(\mathbf{E}[Y|X=x,Z=z]\right) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz.$$

i. Provide a plain-language interpretation for e^{β_1}.

Ans: This is the incidence rate ratio that compares the subgroup $X=1$ to the subgroup $X=0$, both having $Z=0$.

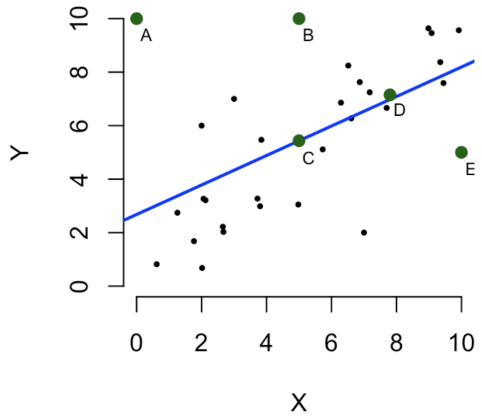
ii. Provide a plain-language interpretation for the null hypothesis $H_0: \beta_2 + \beta_3 = 0$.

Ans: The null hypothesis is that the incidence rate ratio that compares the subgroup $Z=1$ to the subgroup $Z=0$, both having $X=1$, is one.

---

6. ***Optional problem 1***: This is an optional problem — I recommend not attempting it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for attempting the problem, and an additional small amount of credit can be earned for a correct response.

Consider the following scatterplot of some variables X and Y, along with a line obtained from fitting an ordinary least squares linear regression model:



Each of the data points receives equal weight in the analysis, but I have labeled, shaded, and enlarged five points on the graph (A, B, C, D, and E) to bring them to your attention; the questions in this problem pertain to the five points in the figure. Here are some useful facts for the problem:

- Typically, the word *influence* implicitly refers to the parameter β_1. When I refer to *influence* in this problem, I am doing so in accordance with this practice.
- The sample mean of X is 5.
- The estimated regression model parameters are $\widehat{\beta}_0 = 2.8$ and $\widehat{\beta}_1 = 0.54$.
- The emphasized points have coordinates A(0, 10); B(5, 10); C(5, ?); D(7.8, 7.2); E(10, 5). I deliberately left out the y-coordinate of point C.

Respond to each of the following parts, justifying each response in no more than a single sentence:

(a) Determine the sample mean of Y.

Ans: $\overline{y} = \widehat{\beta}_0 + \overline{x}\widehat{\beta}_1 = 2.8 + 5 \times 0.54 = 5.5$.

(b) State which two of the five points are best described as "high influence (on β_1) and high leverage."

Ans: Point A and Point E.

(c) State which of the five points is best described as "very low influence (on β_1) and modest leverage."

Ans: Point D.

(d) State which two of the five points are best described as "no influence (on β_1) and minimal leverage."

Ans: Point B and Point C.

---

7. ***Optional problem 2***: This is an optional problem — I recommend not attempting it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for attempting the problem, and an additional small amount of credit can be earned for a correct response.

Consider the following log-linear model (as per, for instance, Poisson regression):

$$\log(\mathbf{E}[Y|X = x]) = \beta_0 + \beta_1(x - \overline{x}) + \beta_2(x - \overline{x})^2.$$

State (in words, or mathematical notation) an interpretation for the value of $\beta_1 e^{\beta_0}$.

Hint: If you need a refresher about the in-class discussion on quadratic terms, see Slides 330 - 334 in Set 3 and/or refer back to the recording on February 23, about 20 minutes into the lecture. First try to tackle this problem in the un-centered case, following the process we used to handle interpretation of β_1 in a linear regression model with a quadratic term—this requires a small amount of differential calculus, like in the course notes. If you took calculus but have forgotten some of the differentiation rules, I will provide you with all the ones you'll need below. If you do not want to do the calculus, you're welcome to just take a guess! :)

$$\frac{d}{dx} c \times x = c.$$

$$\frac{d}{dx} x^p = p x^{p-1}.$$

$$\frac{d}{dx} e^x = e^x.$$

$$\frac{d}{dx} f(g(x)) = f'(g(x))g'(x).$$

$$\frac{d}{dx} f(x - x_0)\Big|_{x=x_0} = \frac{d}{dx} f(x)\Big|_{x=0}.$$

Ans: This is the rate of change of $\mathbf{E}[Y|X = x]$ at $X = \overline{x}$. To see this, exponentiate both sides and follow the calculus:

$$\begin{aligned} \frac{\partial \mathbf{E}[Y|X = x]}{\partial x} &= \frac{\partial}{\partial x} \exp\left(\beta_0 + \beta_1(x - \overline{x}) + \beta_2(x - \overline{x})^2\right) \\ &= \exp\left(\beta_0 + \beta_1(x - \overline{x}) + \beta_2(x - \overline{x})^2\right) \times \left(\beta_1 + 2\beta_2(x - \overline{x})\right). \end{aligned}$$

Now, plugging in $x = \overline{x}$, we achieve the desired result:

$$\frac{\partial \mathbf{E}[Y|X = x]}{\partial x}\Big|_{x=\overline{x}} = \exp\left(\beta_0 + \beta_1(\overline{x} - \overline{x}) + \beta_2(\overline{x} - \overline{x})^2\right) \times \left(\beta_1 + 2\beta_2(\overline{x} - \overline{x})\right) = \beta_1 e^{\beta_0}.$$

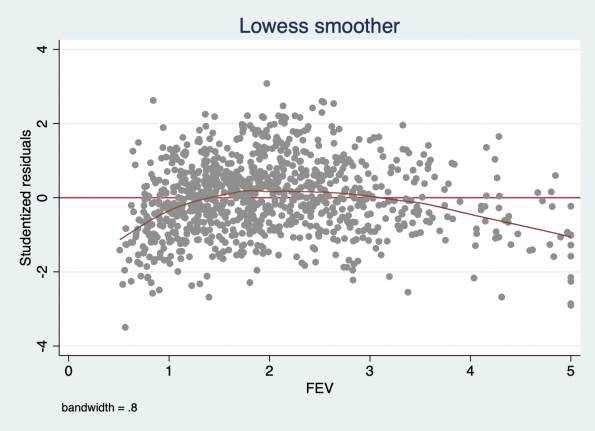
Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2021)
Exam #1

Appendix Material for Exam 1

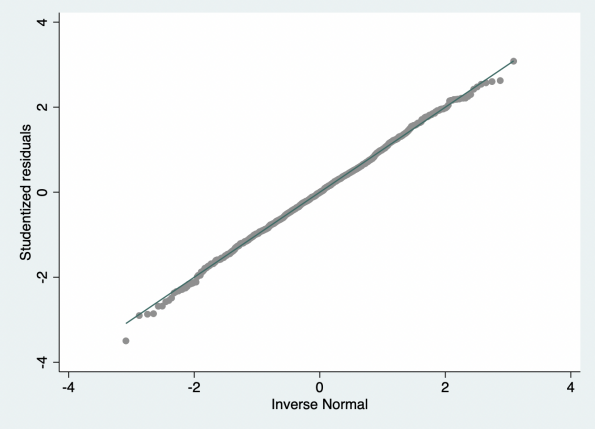
APPENDIX I: Stata output for Problem 1

```
*** ANALYSIS OF UNTRANSFORMED FEV ***

. regress physact fev, robust

Linear regression                               Number of obs     =        975
                                                F(1, 973)         =     500.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3869
                                                Root MSE          =     .41126

------------------------------------------------------------------------------
             |               Robust
     physact |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fev |   .4956053   .0173399    28.58   0.000     .4615774    .5296331
       _cons |    1.65395   .0366488    45.13   0.000      1.58203     1.72587
------------------------------------------------------------------------------
```

Diagnostic Plot #1: Predictor vs. studentized residuals.



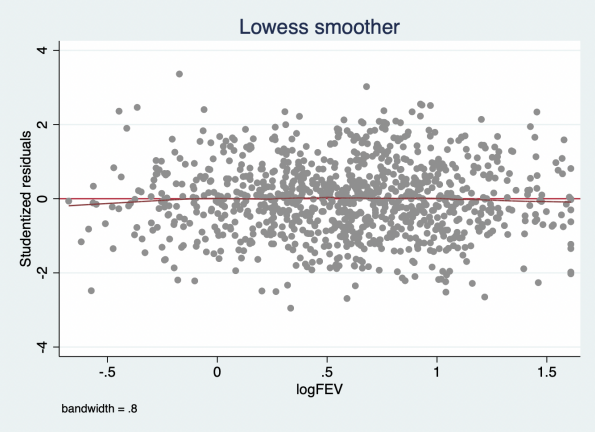
Diagnostic Plot #2: Quantile-quantile plot.



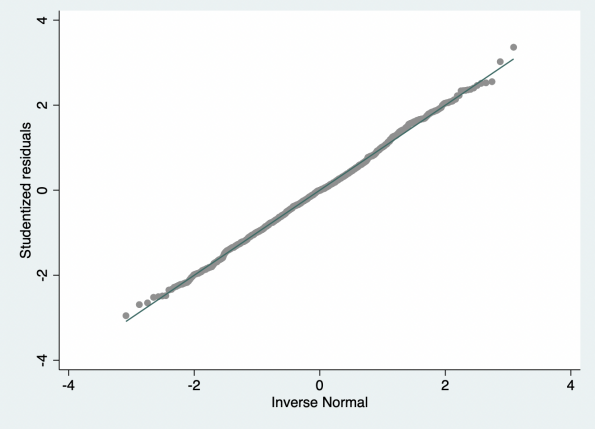
APPENDIX II: Stata output for Problem 1 (continued)

```
*** ANALYSIS OF LOG-TRANSFORMED FEV ***

. regress physact logfev, robust

Linear regression                               Number of obs     =        975
                                                F(1, 973)         =     679.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4145
                                                Root MSE          =     .40189

------------------------------------------------------------------------------
             |               Robust
     physact |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logfev |   1.041884   .0280954    37.08   0.000     .9867493    1.097018
       _cons |   2.031183   .0206557    98.34   0.000     1.990648    2.071718
------------------------------------------------------------------------------
```

Diagnostic Plot #1: Predictor vs. studentized residuals.



Diagnostic Plot #2: Quantile-quantile plot.



APPENDIX III: Stata output for Problem 2

```
*** UNADJUSTED ANALYSIS ***

. regress logtiter grp, robust

Linear regression                               Number of obs     =        184
                                                F(1, 182)         =       3.03
                                                Prob > F          =     0.0832
                                                R-squared         =     0.0164
                                                Root MSE          =     .99738

------------------------------------------------------------------------------
             |               Robust
    logtiter |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         grp |   .2561631   .1470555     1.74   0.083    -.0339898     .546316
       _cons |   4.020385   .1035813    38.81   0.000      3.81601    4.224759
------------------------------------------------------------------------------



*** ADJUSTED ANALYSIS ***

. regress logtiter grp logpretiter, robust

Linear regression                               Number of obs     =        184
                                                F(2, 181)         =      75.84
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4564
                                                Root MSE          =     .74352

------------------------------------------------------------------------------
             |               Robust
    logtiter |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         grp |   .2253742    .109487     2.06   0.041     .0093392    .4414092
 logpretiter |   .4540608   .0371566    12.22   0.000      .380745    .5273766
       _cons |   2.249621   .1670856    13.46   0.000     1.919935    2.579307
------------------------------------------------------------------------------
```

APPENDIX IV: Stata output for Problem 3

```
*** UNADJUSTED ANALYSIS ***

. logit seroconversion grp, robust nolog

Logistic regression                             Number of obs     =        184
                                                Wald chi2(1)      =       0.26
                                                Prob > chi2       =     0.6081
Log pseudolikelihood = -102.22404               Pseudo R2         =     0.0013

------------------------------------------------------------------------------
             |               Robust
seroconversion |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
         grp |   .1767036   .3445596     0.51   0.608    -.4986209     .852028
       _cons |  -1.218157   .2490798    -4.89   0.000    -1.706345   -.7299701
------------------------------------------------------------------------------




*** ADJUSTED ANALYSIS ***

. logit seroconversion grp logpt, robust nolog

Logistic regression                             Number of obs     =        184
                                                Wald chi2(2)      =      34.83
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -56.502458               Pseudo R2         =     0.3472

------------------------------------------------------------------------------
             |               Robust
 seroconversion |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
         grp |    .677033   .4843872     1.40   0.162    -.2723485    1.626415
       logpt |   1.511318   .2570426     5.88   0.000     1.007523    2.015112
       _cons |  -3.577783   .7648304    -4.68   0.000    -5.076823   -2.078743
------------------------------------------------------------------------------
```

APPENDIX V: Stata output for Problem 4

```
. logit rec c.age##i.smk, robust nolog

Logistic regression                             Number of obs     =        362
                                                Wald chi2(5)      =       7.30
                                                Prob > chi2       =     0.1995
Log pseudolikelihood = -151.05378               Pseudo R2         =     0.0313

------------------------------------------------------------------------------
             |               Robust
         rec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0685194   .0498011     1.38   0.169    -.0290891    .1661278
             |
         smk |
          1  |   2.805594   4.604232     0.61   0.542    -6.218536    11.82972
          2  |  -7.793166   8.879124    -0.88   0.380    -25.19593    9.609596
             |
   smk#c.age |
          1  |  -.0403019   .0585656    -0.69   0.491    -.1550883    .0744846
          2  |    .106017   .1199353     0.88   0.377    -.1290519    .3410858
             |
       _cons |  -3.536572   3.934267    -0.90   0.369    -11.24759     4.17445
------------------------------------------------------------------------------



. lincom age + 1.smk#c.age

 ( 1)  [rec]age + [rec]1.smk#c.age = 0

------------------------------------------------------------------------------
         rec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0282175   .0308184     0.92   0.360    -.0321855    .0886205
------------------------------------------------------------------------------



. testparm smk#c.age

 ( 1)  [rec]1.smk#c.age = 0
 ( 2)  [rec]2.smk#c.age = 0

           chi2(  2) =    1.92
         Prob > chi2 =    0.3820
```