Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2022)
Exam #1

Name (Printed): ______________________________

Instructions: Please adhere to the following guidelines:

- There are five required problems (each with multiple sub-questions of varying length and difficulty), and one optional problem that is optional for all students. There are no appendices.
- Please read the questions carefully and answer no more or less than what you are being asked to answer.
- My recommendation is to provide your responses to the problems you find easiest first, and then return to the more challenging ones.
- This exam is closed-everything, and is an **individual effort**. You will, however, be permitted the use of a scientific calculator.
- 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!
- **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.

Further information: You may find the following information helpful.

- Any reference to logarithmic transformations are based on the *natural* logarithm (i.e., having base e).
- The approximate 97.5th percentile of the standard normal distribution is given by $z_{0.975} \approx 1.96$.
- A linear regression model of a continuous outcome (Y) that places a natural cubic spline on a continuous exposure (X) having K knots uses K degrees of freedom (including the intercept).

#	Score	Points
1		10
2		25
3		25
4		25
5		15
Total:		100
Optional		

Signature for integrity statement: ______________________________

1. $\boxed{\text{10 pts}}$ Below are ten true-or-false questions (1 pt. each), all of which pertain to the simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon, \text{ with } \mathbf{E}[\epsilon | X = x] = 0 \text{ for all } x.$$

Note that this model can also be expressed as $\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x$. In this problem, assume that $\widehat{\beta}_1$ refers to the ordinary least squares estimator. Circle your choice (**TRUE** or **FALSE**) for each question. Please read the statements *carefully*. **There is no need to provide a written justification for your response**.

(a) **TRUE** or **FALSE** The parameter β_1 quantifies the change in Y that will occur when X is increased by a single unit.

(b) **TRUE** or **FALSE** The parameter β_1 denotes the difference in mean Y between subgroups differing in their value of X by one unit.

(c) **TRUE** or **FALSE** The intercept, β_0, should be excluded from the model (that is, the model should be reduced to the simpler model $Y = \beta x + \epsilon$) if it possesses no real-world, clinically meaningful interpretation and/or if it cannot be reliably estimated in your data.

(d) **TRUE** or **FALSE** Unbiased estimation of β_1 requires the errors to have constant variance (i.e., satisfy the assumption of homoscedasticity).

(e) **TRUE** or **FALSE** The only way a 95% CI of the form $\widehat{\beta}_1 \pm z_{0.975} \times \widehat{\text{SE}}(\widehat{\beta}_1)$ can be theoretically justified is if the errors are exactly normally distributed.

(f) **TRUE** or **FALSE** It is critically important for the linearity assumption to hold exactly for this model to be of any use whatsoever.

(g) **TRUE** or **FALSE** The linearity assumption is important for this model to be able to establish valid 95% prediction intervals for Y.

(h) **TRUE** or **FALSE** If homoscedasticity is not satisfied, the sandwich standard error can be used for efficiency gains.

(i) **TRUE** or **FALSE** If both linearity and homoscedasticity hold, the Gauss-Markov theorem asserts that $\widehat{\beta}_1$ has the smallest variance of *all* unbiased estimators of β_1.

(j) **TRUE** or **FALSE** Conceptually, a pair of observations that are far apart on a scatter plot provides "more information" about β_1 as compared to a pair of observations that are closer together.

2. 25 pts Rapid Education/Encouragement and Communications for Health (REACH) is a text message intervention that was evaluated in a randomized trial of independently sampled adults with uncontrolled type 2 diabetes. The goal was to evaluate whether REACH could improve hemoglobin A1c (HbA1c). Suppose we use linear regression to understand the relationship between baseline and six-month HbA1c (HbA1c$_0$ and HbA1c$_6$, respectively) among REACH patients. Below is the corresponding Stata output.

```
. regress a1c6 a1c0 if reach == 1, robust

Linear regression                                   Number of obs     =        217
                                                    F(1, 215)         =      48.49
                                                    Prob > F          =     0.0000
                                                    R-squared         =     0.2343
                                                    Root MSE          =     1.5367

------------------------------------------------------------------------------
             |               Robust
        a1c6 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        a1c0 |   .4419786   .0634714     6.96   0.000     .3168726    .5670846
       _cons |   4.205857   .5618046     7.49   0.000     3.098507    5.313207
------------------------------------------------------------------------------
```

Four commonly used diagnostic plots are presented on Page 4 (two points have been labeled as Points 1 and 2). You should use information from these plots to support your responses as appropriate.

(a) 4 pts Appropriately supporting your response using the diagnostic plots, briefly comment on the degree to which you would trust a model-based point estimate and 95% CI for the mean HbA1c$_6$ among individuals with an HbA1c$_0$ of 10%.

(b) 4 pts If it is possible to use the Stata output above to determine a point estimate and/or 95% CI for the mean described in part (a), do what you can; if something is not possible, briefly summarize why not.

(c) $\boxed{\text{4 pts}}$ I believe that this model could reasonably be expected to produce a naive 95% prediction interval for HbA1c$_6$ among individuals with an HbA1c$_0$ of 9% that is approximately valid. Briefly summarize the key features of the diagnostic plots on Page 4 that led me to this conclusion.

(d) $\boxed{\text{4 pts}}$ If it is possible to use the Stata output on the previous page to determine the naive 95% prediction interval described in part (c) of the form $\widehat{y} \pm z_{0.975}\widehat{\sigma}$, do so; if not, briefly summarize why not.

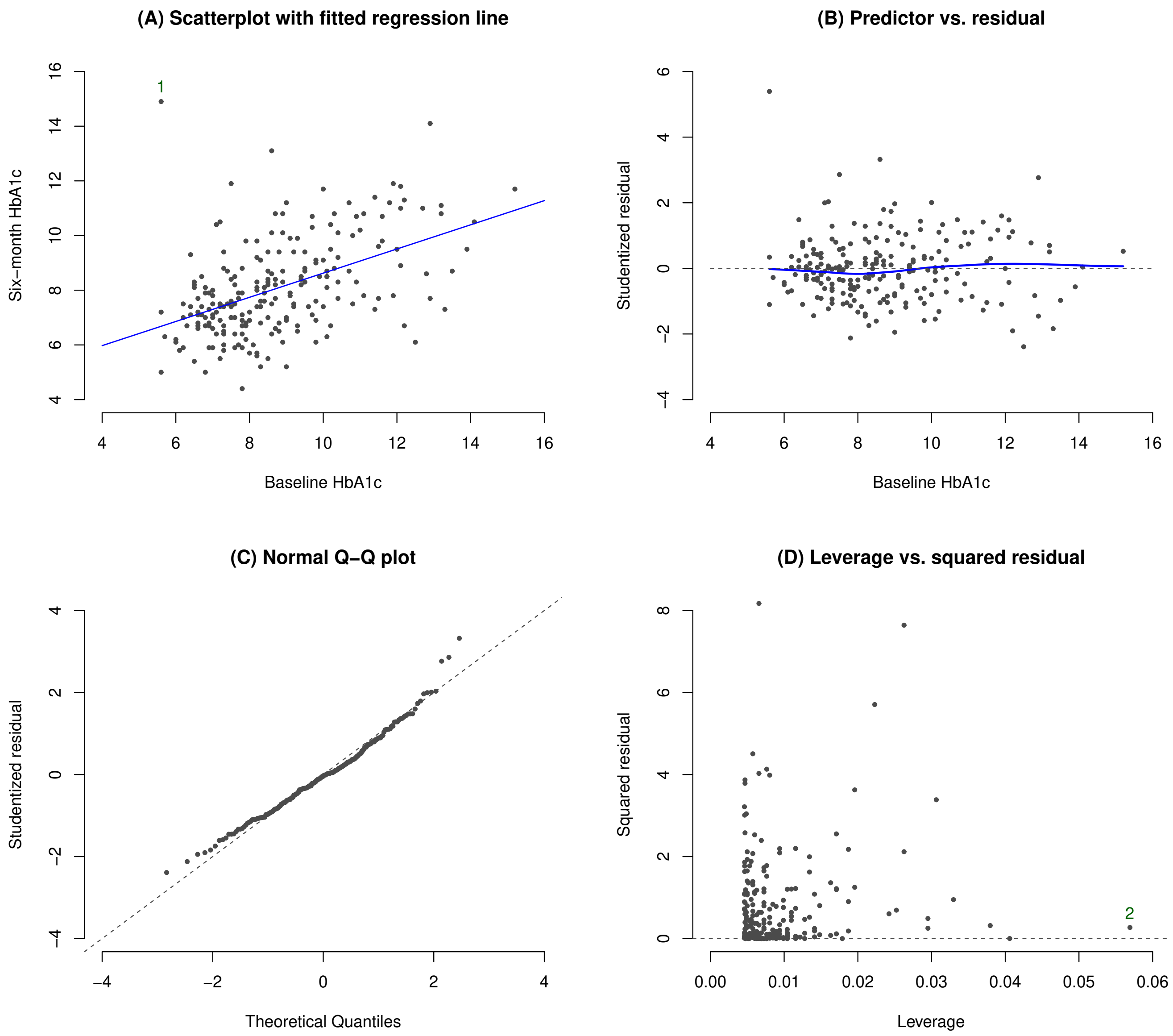
(e) $\boxed{\text{3 pts}}$ Determine a point estimate and 95% CI for the difference in mean HbA1c$_6$ between those with an HbA1c$_0$ of 8% and those with an HbA1c$_0$ of 12%. Please do this "the easy way," as we have discussed in class.

(f) $\boxed{\text{2 pts}}$ Circle the point on Plot (D) to which Point 1 on Plot (A) corresponds (no justification required).

(g) $\boxed{\text{2 pts}}$ Circle the point on Plot (A) to which Point 2 on Plot (D) corresponds (no justification required).

(h) $\boxed{\text{2 pts}}$ It is clear that Point 2 has the highest leverage of all points. Briefly explaining your answer (heuristically, not mathematically), do you believe it to also be a highly influential point?

Diagnostic plots for problem 2



3. $\boxed{\text{25 pts}}$ **Background**: One way in which vaccines are evaluated is through comparisons of antibody response. The *antibody titer* 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 immunogenic if subjects receiving the vaccine tend to have higher titers as compared to control. Owing to the fact that titers are defined multiplicatively, it is standard to log-transform titers.

A large double-blind, placebo-controlled randomized trial was 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. 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 dilution, whichever comes first (typically, 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). The variables collected in this study are provided in the table below. **You may freely assume that titers are correlated within subjects over time**.

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

Consider the following two linear regression models.

$$\begin{aligned} \mathbf{E}[\texttt{logtiter}|\texttt{grp}] &= \beta_0^* + \beta_1^*\texttt{grp} & \textbf{(MODEL 1)} \\ \mathbf{E}[\texttt{logtiter}|\texttt{grp}, \texttt{logpretiter}] &= \beta_0 + \beta_1\texttt{grp} + \beta_2\texttt{logpretiter} & \textbf{(MODEL 2)} \end{aligned}$$

(a) $\boxed{\text{3 pts}}$ Circle the number corresponding to the correct statement, though you need not justify your response.

(i.) $\beta_1^* = \beta_1$.
(ii.) $\beta_1^* < \beta_1$.
(iii.) $\beta_1^* > \beta_1$.
(iv.) There is not enough information to determine which of the above is true.

(b) $\boxed{\text{3 pts}}$ Suppose you estimate β_1^* and β_1 by ordinary least squares. Circle the number corresponding to the correct statement, though you need not justify your response.

(i.) $\widehat{\beta}_1^* = \widehat{\beta}_1$.
(ii.) $\widehat{\beta}_1^* < \widehat{\beta}_1$.
(iii.) $\widehat{\beta}_1^* > \widehat{\beta}_1$.
(iv.) There is not enough information to determine which of the above is true.

(c) $\boxed{\text{3 pts}}$ Suppose you estimate β_1^* and β_1 by ordinary least squares. Circle the number corresponding to the correct statement, though you need not justify your response.

(i.) $\text{Var}[\widehat{\beta}_1^*] = \text{Var}[\widehat{\beta}_1]$.
(ii.) $\text{Var}[\widehat{\beta}_1^*] < \text{Var}[\widehat{\beta}_1]$.
(iii.) $\text{Var}[\widehat{\beta}_1^*] > \text{Var}[\widehat{\beta}_1]$.
(iv.) There is not enough information to determine which of the above is true.

(d) $\boxed{\text{4 pts}}$ State a plain-language interpretation of $e^{\beta_1^*}$.

(e) $\boxed{\text{4 pts}}$ State a plain-language interpretation of 2^{β_2}. *Hint*: recall that $2^{\beta_2} = \exp(\beta_2 \times \log(2))$.

(f) $\boxed{\text{4 pts}}$ Suppose that in the context of **MODEL 1**, you had a compelling reason to suspect *a priori* that the standard deviation of the error would be *approximately* twice as high for the vaccine group as compared to the control group. Briefly describe a principled modeling strategy that would allow you to leverage this knowledge.

(g) $\boxed{\text{4 pts}}$ In the context of **MODEL 2**, consider replacing the *linear term* on log-transformed pre-booster titer with a *natural cubic spline* on log-transformed pre-booster titer having three knots. How many *additional* degrees of freedom would this modification require?

4. $\boxed{\text{25 pts}}$ A group of investigators seeks to research possible treatment strategies for patients with substance abuse. They propose treatments to be randomized in a 2×3 factorial design. Specifically, let X denote the specific type of treatment (0 = behavior modification therapy, 1 = psychotherapy), and let Z denote the setting in which the treatment is provided (0 = outpatient, 1 = day-treatment, 2 = inpatient). You may assume that patients are randomly and evenly allocated to each of the six groups. The outcome, Y, is given by a substance abuse severity score that is measured continuously on a scale of 0-100, with higher values indicating greater severity. The questions in this problem pertain to the following two models:

$$\begin{aligned}
\mathbf{E}[Y|X = x, X = z] &= \alpha_0 + \alpha_1 1(x = 1) + \alpha_2 1(z = 1) + \alpha_3 1(z = 2) && \textbf{(MODEL 1)} \\
\mathbf{E}[Y|X = x, X = z] &= \beta_0 + \beta_1 1(x = 1) + \beta_2 1(z = 1) + \beta_3 1(z = 2) + \beta_4 1(x = 1, z = 1) + \beta_5 1(x = 1, z = 2) && \textbf{(MODEL 2)}
\end{aligned}$$

(a) $\boxed{\text{3 pts}}$ Provide a plain-language interpretation for the parameter α_0.

(b) $\boxed{\text{4 pts}}$ Provide a plain-language interpretation for the parameter β_1.

(c) $\boxed{\text{4 pts}}$ Based on **MODEL 1**, what parameter or combination of parameters should be tested to evaluate whether the mean post-treatment severity score differs by treatment setting? Specifically write the null hypothesis, H_0, though you need not show the "regression math."

(d) $\boxed{\text{3 pts}}$ In regards to **MODEL 2**, what does it mean in practical terms if $\beta_4 = \beta_5 = 0$?

(e) $\boxed{\text{4 pts}}$ Based on **MODEL 1**, what parameter or combination of parameters should be tested to evaluate whether the mean post-treatment severity score differs between patients receiving outpatient psychotherapy and patients receiving inpatient behavior modification therapy? Specifically write the null hypothesis, H_0, and **please show the "regression math" in your response.**

(f) $\boxed{\text{4 pts}}$ Based on **MODEL 2**, what parameter or combination of parameters should be tested to evaluate whether mean post-treatment severity score differs between patients receiving day-treatment or inpatient treatment? Specifically write the null hypothesis, H_0, and **please show the "regression math" in your response.**

(g) $\boxed{\text{3 pts}}$ Suppose that once the study is conducted, it is determined that the sample mean severity score in the group receiving day-treatment behavior modification therapy is 70. To what extent, if any, does this inform you about any of the ordinary least squares estimates of the coefficients of **MODEL 1** and/or **MODEL 2**? Briefly justify your response.

5. $\boxed{\text{15 pts}}$ A study was conducted of independently sampled children with asthma. The primary study objective was to investigate whether asthma severity (measured by baseline forced expiratory volume [FEV], in L) was associated with mean physical activity (PA, measured as hours spent involved in physical activities over a six-week period—**you may of course freely use the abbreviation PA throughout this problem to save time**). Consider the following two linear regression models:

$$\begin{aligned}\mathbf{E}[\texttt{PA}|\texttt{fev}] &= \beta_0^* + \beta_1^*\texttt{fev} && \textbf{(MODEL 1)}\\ \mathbf{E}[\texttt{PA}|\texttt{fev},\texttt{age}] &= \beta_0 + \beta_1(\texttt{fev} - 3.5) + \beta_2(\texttt{age} - 7) && \textbf{(MODEL 2)}\end{aligned}$$

In this study, the FEV ranged from 1 to 6 L and age ranged from 5 to 15 years.

(a) $\boxed{\text{3 pts}}$ State an interpretation for β_0^* in plain language. To what degree is this interpretation meaningful in the real world?

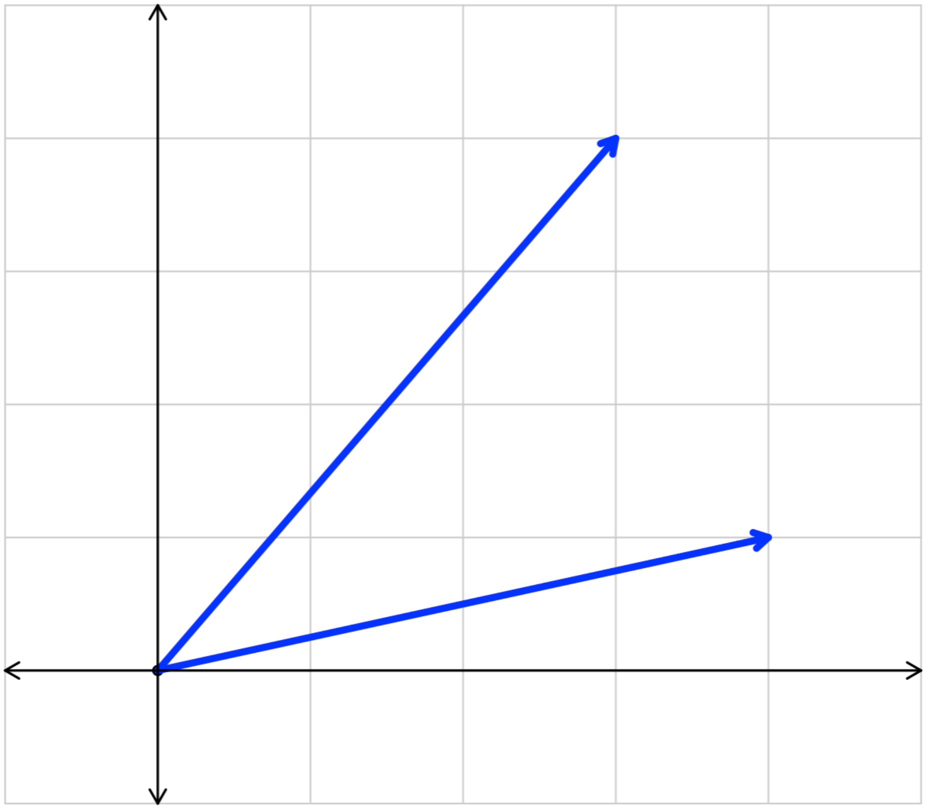
(b) $\boxed{\text{4 pts}}$ State an interpretation for β_0 in plain language.

(c) $\boxed{\text{4 pts}}$ State an interpretation for β_1 in plain language. Then, briefly summarize why this parameter may better capture the association about which the investigators are seeking to learn than β_1^*.

(d) $\boxed{\text{4 pts}}$ Suppose that it was determined that 20% of children in the study had a sibling that was also included in the study. In a maximum of one sentence, comment on why you might not trust conclusions from an ordinary least squares regression fit to **MODEL 2**.

6. ***Optional problem***: This is an optional problem — please do not attempt it until you have completed and are satisfied with your answers to the required problems. A small amount of credit can be earned for correct responses.

Consider the "through-the-origin" simple linear regression model $Y = \beta X + \epsilon$ that would be appropriate, for instance, if both X and Y were centered to have mean zero. You obtain the following two observations: $(x_1 = 4, y_1 = 3)$ and $(x_2 = 1, y_2 = 4)$. In this very simple case, the ordinary least squares process can be visualized on a two-dimensional graph. The design matrix and the outcome are both represented as vectors in the plot below.



(a) On the plot, label the design matrix as $\mathbf{x}$ and the outcome vector as $\mathbf{y}$.

(b) On the plot, draw the fitted value, $\widehat{\mathbf{y}}$ (making it clear how you know this by marking any known angles accordingly).

(c) On the plot, represent and label the residual vector, $\widehat{\boldsymbol{\epsilon}}$.

(d) Determine the ordinary least squares estimate, $\widehat{\beta}$, which in this case is given by:

$$\widehat{\beta} = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}.$$