Andrew J. Spieker, PhD
BIOS 6312 - Modern Regression Analysis (Spring 2023)
Exam #1 Key: Mean=90.2; Median=90.5 (IQR=[88, 94]; Max=98)

Name (Printed): _____

Instructions: Please adhere to the following guidelines:

- There are four required problems (each with multiple sub-questions of varying length and difficulty), and two optional problems that are optional for everyone. There is one appendix.

- 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 below 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	30
3	30
4	30
Total:	100
Optional 1	
Optional 2	

Signature for integrity statement: _____

1. ☐ 10 pts ☐ Below are six statements regarding simple linear regression that are either misleading or inaccurate:

   (I) The "slope" marks the change in outcome caused by a one-unit change in the value of the predictor.

   (II) The robust standard error serves to produce an estimate with greater efficiency (i.e., lower variance).

   (III) A regression model's validity always depends upon the errors following a normal distribution.

   (IV) A high-leverage observation must also be highly influential.

   (V) The mean squared error estimates the outcome variance in any subgroup defined by its predictor value.

   (VI) The predictor should be log-transformed if it exhibits any signs of skewness.

   Please select any **two** of the above statements (your choice) and write a brief response in which you explain/elaborate on their inaccuracies. This can mean pointing out logical flaws; providing specific corrections, clarifications, or qualifications to the statement; or even sketching a figure as a counterexample if you think it helps your case. Your responses should be specific, although they can be concise (one-two sentences).

   Ans: Though you were only expected to select two, I am placing answers to each for your reference. Some of the responses below are a little longer than what I would have expected you to write.

   (I) The slope parameter from a simple linear regression model characterizes the difference in the mean of the outcome between subgroups differing in the value of their predictor by one unit. So, there are a lot of things wrong with this characterization: for instance, the predictor may not be "changeable," the slope parameter need not possess a causal interpretation, and the regression model does not tell you how all outcomes compare between groups—but rather how their *means* compare.

   (II) Employing a robust standard error does not have any impact whatsoever on the point estimate obtained from, e.g., ordinary least squares. Therefore, the robust standard error cannot possibly serve to improve the efficiency of estimation (rather, it serves to protect against heteroscedasticity in the case of ordinary least squares).

   (III) The Lindeberg-Feller central limit theorem (no, you did not need to remember the name of this :)) asserts that the sampling distribution of the ordinary least squares estimate will be approximately normal in sufficiently large samples; as such, we may feel comfortable forming confidence intervals and obtaining p-values from these models even when error normality does not hold (provided we think our sample size is large enough).

   (IV) An observation can be high-leverage by having a very extreme value for the predictor (i.e., a value very far away from the sample mean); however—as a counter-example to this claim—note that if it lies *on* the fitted regression line, it has zero influence. A highly influential observation is one that, when removed, would change the estimated regression line substantially.

   (V) This is only true if the assumption of error homoscedasticity is satisfied (otherwise, it's estimating the *average* within-group variance across groups defined by their predictor).

   (VI) Not all skewed variables are strictly positive, so this may not even be possible; what's more, log-transformation would only serve to address *right*-skewness in particular. The decision to log-transform a predictor must be based on arguments that are far more nuanced (for instance, it should be related to the scientific question and the purpose for which the model is being used).

2. $\boxed{\text{30 pts}}$ A cross-sectional observational study was conducted with the goal of evaluating the association between smoke exposure and lung function (measured by forced expiratory volume, or FEV) in children 10 to 18 years of age. A subset of variables measured in this study include:

$$X = \begin{cases} 0 & \text{no smoke exposure} \\ 1 & \text{smoke exposure} \end{cases}, \qquad Z = \text{age (years)}, \qquad Y = \text{FEV (L)}$$

Consider the following two regression models (one unadjusted and one with a linear adjustment for age):

$$\mathbf{E}[Y|X = x] = \alpha_0 + \alpha_1 x \qquad\qquad \textbf{(MODEL 1)}$$

$$\mathbf{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z \qquad\qquad \textbf{(MODEL 2)}$$

(a) $\boxed{\text{4 pts}}$ State a plain-language interpretation for α_0 from **MODEL 1**.

Ans: The mean FEV among those with no smoke exposure.

(b) $\boxed{\text{4 pts}}$ State a plain-language interpretation for α_1 from **MODEL 1**.

Ans: The difference in mean FEV between those with and without smoke exposure.

(c) $\boxed{\text{2 pts}}$ Let Y^* denote FEV measured in mL instead of L. Consider the model $\mathbf{E}[Y^*|X = x] = \alpha_0^* + \alpha_1^* x$. How does the value of α_1^* compare to the value of α_1 from **MODEL 1**? *Note*: 1 L = 1000 mL.

Ans: $\alpha_1^* = 1000 \times \alpha_1$.

(d) $\boxed{\text{4 pts}}$ Suppose you have reason to believe that the standard deviation of FEV among the exposed is twice that of the unexposed. Complete the following paragraph by filling in the blanks and circling the correct word choice where noted:

*"In a weighted least squares fit to the unadjusted model, we could reasonably pre-specify a plan to give the observations from the unexposed group __4__ times the weight of the observations in the exposed group. Assuming that our hypothesis regarding the standard deviations as stated above is true, the Gauss-Markov theorem asserts that the resulting weighted least squares estimator of α_1 will have the **__smallest__** variance among all **__linear__** **__unbiased__** estimators of α_1."*

(e) $\boxed{\text{4 pts}}$ State a plain-language interpretation for β_0 from **MODEL 2**. In one sentence, why might you be skeptical about the ability of data from this study to reliably estimate this quantity?

Ans: This is the mean FEV among newborns with no smoke exposure. This study only samples individuals from age 10-18, so relying on this model to estimate this quantity would involve severe extrapolation.

(f) $\boxed{\text{4 pts}}$ Provide an example of a simple transformation on age (Z) that, if applied to **MODEL 2**, could result in a more trustworthy estimate of the intercept. Simply state or describe the transformation.

Ans: Center age about some value between 10 and 18; one such choice would be the sample mean.

2

(g) $\boxed{\text{4 pts}}$ State a plain-language interpretation for β_1 from **MODEL 2**. For the purposes of capturing the deleterious effects of smoke exposure on lung function, why might β_1 from **MODEL 2** be more informative as compared α_1 from **MODEL 1**?

Ans: This is the difference in mean FEV between subgroups of the same age but differing in their smoke exposure status. Adjustment for age in this way serves to control for the likely contaminating impact of confounding.

(h) $\boxed{\text{4 pts}}$ State a scenario in which you would expect to benefit from placing a natural cubic spline with three unique knots on age instead of adjusting for it with a linear term (make clear what the benefit would be). How many *additional* degrees of freedom would this modification to **MODEL 2** cost?

Ans: If the relationship between age and mean FEV is nonlinear, then this sort of adjustment will do a better job of addressing confounding. This modification only requires *one* additional degree of freedom.

3. 30 pts Following the emergence of SARS-CoV-2, the virus responsible for COVID-19, a group of investigators sought to develop and compare assays to quantify antibody levels in the blood. In one study, they compared an existing immunofluorescence measure to their novel thermal measure in $N = 220$ independently sampled individuals with SARS-CoV-2 infection documented in the last year. The immunofluorescence assay produces values between 0 and 1 immunofluorescence units (IU), and the thermal assay produces values between 1 and 150 thermal units (TU). For both measures, higher values signify higher antibody levels, although the thermal assay is defined on a multiplicative scale such that it is typically log-transformed for analysis. Therefore, consider a simple linear regression model with the log-transformed thermal assay as the outcome and the immunofluorescence assay as the predictor. The Stata output from an ordinary least squares fit to this model is shown in the appendix, along with three associated diagnostic plots. **Please use the appendix material to support your responses as appropriate. Further, it will help you in this problem to know that the mean of the immunofluorescence assay is 0.30 IU in this sample.**

(a) 3 pts What proportion of variation in the log-transformed thermal assay is estimated to be (linearly) explained by the immunofluorescence assay?

Ans: $R^2 = 0.7411$, so about 74.1%.

(b) 3 pts Determine the geometric mean of the thermal assay in this sample.

Ans: $\overline{\log(Y)} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} = 0.4757339 + 3.540891 \times 0.30 = 1.538 \Rightarrow \widehat{\text{GeoMean}}(Y) = \exp(1.538) = 4.66$ TU.

(c) 4 pts I believe this model could reasonably be used to estimate the geometric mean thermal measure among individuals with an immunofluorescence measure of 0.1 IU. Identify **one** key feature of the study design and **one** key observation from the diagnostic plots that likely led me to this conclusion.

Ans: (1) The patients were sampled independently. (2) The predictor vs. residual plot does not seem to indicate a serious departure from linearity (as evidenced by the fact that the LOWESS smoother is fairly close to the x-axis). You could have also commented on the fact that 0.1 IU is well within the range of data available.

(d) 3 pts Is it possible to use the output provided to determine a point estimate for the quantity described in part (c)? If so, do so. If not, briefly explain why not.

Ans: Yes: $\widehat{\mathbf{E}}[\log(Y)|X = 0.1] = \widehat{\beta}_0 + \widehat{\beta}_1 \times 0.1 = 0.829823 \Rightarrow \widehat{\text{GeoMean}}(Y|X = 0.1) = 2.29$ TU.

(e) 3 pts Is it possible to use the output provided to construct a a 95% confidence interval for the quantity described in part (c)? If so, do so. If not, briefly explain why not.

Ans: No, because we are not given information on $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)$. We would need a command like `lincom` to accomplish this in Stata.

(f) 3 pts I believe that this model could reasonably be used to construct a 95% prediction interval for the thermal measure among individuals with an immunofluorescence measure of 0.1 IU. Identify **three** key observations from the diagnostic plots that likely led me to this conclusion.

Ans: The predictor vs. residual plot does not suggest a serious departure from either (1) linearity or (2) error homoscedasticity, and (3) the Q-Q plot does not suggest a serious departure from error normality.

(g) ☐ 3 pts ☐ Construct a naive 95% prediction interval for the subgroup described in part (f).

Ans: From the output, $\widehat{\sigma}$ = 0.33564 (RMSE). Further leveraging information from part (d), this is given by: $[\exp(0.829823 - 1.96 \times 0.33564), \exp(0.829823 + 1.96 \times 0.33564)] = [1.19, 4.43]$.

(h) ☐ 2 pts ☐ Complete the following paragraph by circling the correct word choices where noted:

*"A properly formed prediction interval would account for the fact that the coefficients are estimated with random variation. The 95% prediction interval formed in part (g) is likely too **narrow** as compared to a properly formed 95% prediction interval for this subgroup. The discrepancy between a naively formed and properly formed 95% prediction interval would be **smaller** for the subgroup of individuals with an immunofluorescence measure of 0.3 IU."*

(i) ☐ 3 pts ☐ On the immunofluorescence scale, a difference of 0.2 IU between groups is considered the smallest that is clinically meaningful. Based on the model, estimate a quantity signifying how groups differing in their immunofluorescence measure by this amount compare on the thermal scale. Please state the quantity you're estimating; report a corresponding point estimate and a 95% confidence interval.

Ans: The quantity of interest is the geometric mean ratio comparing the geometric mean thermal assay between groups differing in their immunofluorescence assay by 0.2 IU. The point estimate is $\exp(0.2 \times 3.540891) = 2.03$ IU; the 95% CI is $[\exp(0.2 \times 3.261368), \exp(0.2 \times 3.820414)] = [1.92, 2.15]$.

(j) ☐ 3 pts ☐ Suppose it is discovered that values obtained from the thermal assay are extraordinarily sensitive to ambient temperature. If the thermal tests were conducted gradually on the samples over a period of a few months without regard to this limitation, what concerns might you have about the validity of the ordinary least squares fit as it was implemented? *Note*: this question is hypothetical and stands alone; do not go back and alter your responses to parts (a)-(i) to accommodate this scenario.

Ans: If ambient temperature is not well controlled in the lab setting, it could conceivably change over the course of a few months (e.g., due to season changes). This would produce a violation to the assumption of independent errors, as observations that are temporally closer to one another will tend to be correlated.

4. 30 pts A team sought to study reactions associated with SARS-CoV-2 and influenza vaccines. They conducted a randomized study of simultaneous vaccinations via a 2×3 factorial design:

$$X = \begin{cases} 0 & \text{SARS-CoV-2 placebo} \\ 1 & \text{SARS-CoV-2 mRNA vaccine} \end{cases} \quad \text{and} \quad Z = \begin{cases} 0 & \text{influenza placebo} \\ 1 & \text{standard-dose influenza vaccine} \\ 2 & \text{high-dose influenza vaccine} \end{cases}$$

Patients were randomly and evenly allocated to each of the six groups. The outcome, Y, was a continuously measured reaction score taking into account an aggregate of self-reported injection-site and systemic reactions (e.g., pain, tenderness, nausea, fever). The questions in this problem pertain to the following two models:

$$\mathbf{E}[Y|X = x, Z = z] = \alpha_0 + \alpha_1 1(x = 1) + \alpha_2 1(z = 1) + \alpha_3 1(z = 2) \quad \text{(MODEL 1)}$$

$$\mathbf{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 1(x = 1) + \beta_2 1(z = 1) + \beta_3 1(z = 2) + \beta_4 1(x = 1, z = 1) + \beta_5 1(x = 1, z = 2) \quad \text{(MODEL 2)}$$

(a) 4 pts Many factors could go into an informed *a priori* choice between these two models. Below are arguments from four different researchers pertaining to their personal preferences for one model over another. Irrespective of your *own* personal preferences, indicate below whether each of their arguments is valid or invalid (simply circle your response; no further justification is required).

(I) **Valid** or **Invalid** Researcher 1 argues that **MODEL 2** is automatically better than **MODEL 1** by virtue of the fact that **MODEL 2** is saturated.

Note: A model that is saturated is not necessarily better or worse than a model that is not.

(II) **Valid** or **Invalid** Researcher 2 points out that **MODEL 2** is equipped to determine the extent of interaction between vaccines (whereas **MODEL 1** is not), so **MODEL 2** may be preferable if part of the study's goal is to characterize that interaction.

Note: This seems reasonable to me.

(III) **Valid** or **Invalid** Researcher 3 argues that the best justification for **MODEL 1** is that it has fewer parameters as compared to **MODEL 2**.

Note: A model is not necessarily better or worse because it uses fewer degrees of freedom than another. I'll also note that two models using the *same* number of degrees of freedom can involve very different assumptions and therefore have very different behavior.

(IV) **Valid** or **Invalid** Researcher 4 is *only* interested in using this study to compare the high-dose influenza vaccine to the standard-dose influenza vaccine and therefore would prefer to use **MODEL 1**.

Note: This seems reasonable to me, though I'll note that someone made a convincing argument to me that this is not valid logic; hence, I accepted either answer for this one.

6

(b) ⌊3 pts⌋ In regards to **MODEL 1**, what does it mean scientifically if $\alpha_2 = \alpha_3 = 0$?

Ans: The mean reaction score does not differ across the three influenza vaccine doses.

(c) ⌊3 pts⌋ In regards to **MODEL 1**, what does it mean scientifically if $\alpha_2 = \alpha_3$?

Ans: The mean reaction score does not differ between those receiving the standard-dose influenza vaccine and those receiving the high-dose influenza vaccine.

(d) ⌊15 pts⌋ The table below depicts five hypotheses, (A)-(E), stated in plain scientific language on the left-hand side. The hypotheses are stated in terms of the parameters of **MODEL 2** on the right-hand side. However, the hypotheses are not in the correct order (for instance, statement (A) does not correspond to hypothesis (1)). Your task is to match each statement on the left-hand side to the correct hypothesis on the right-hand side. Please indicate your selections in the spaces below the table. You are not required to show the "regression math," though you may want to confirm your answers on scratch paper.

	Plain-language statement		Hypothesis
(A)	The vaccines do not interact in terms of their effects on the mean reaction score.	(1)	$H_0 : \beta_1 = 0$
(B)	The SARS-CoV-2 mRNA vaccine does not affect the mean reaction score among patients receiving the influenza placebo.	(2)	$H_0 : \beta_4 = \beta_5 = 0$
(C)	The SARS-CoV-2 mRNA vaccine does not affect the mean reaction score among patients receiving the high-dose influenza vaccine.	(3)	$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
(D)	The mean reaction score does not differ between groups receiving the standard-dose influenza vaccine and the high-dose influenza vaccine.	(4)	$H_0 : \beta_1 + \beta_5 = 0$
(E)	The mean reaction score does not differ across the three influenza vaccine doses.	(5)	$H_0 : \beta_3 - \beta_2 = \beta_5 - \beta_4 = 0$

Ans: The reasoning is shown for your reference, though you weren't required or expected to present it:

(A) The shortcut is to jointly test the interaction terms: $H_0 : \beta_4 = \beta_5 = 0$.
(B) The reduced model is given by $\mathbf{E}[Y|X = x, Z = 0] = \beta_0 + \beta_1 1(x = 1)$; $H_0 : \beta_1 = 0$.
(C) The reduced model is given by $\mathbf{E}[Y|X = x, Z = 2] = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)1(x = 1)$; $H_0 : \beta_1 + \beta_5 = 0$.
(D) The placebo model is given by $\mathbf{E}[Y|X = 0, Z = z] = \beta_0 + \beta_2 1(z = 1) + \beta_3 1(z = 2)$ so that $\beta_3 - \beta_2 = 0$. The vaccine model is $\mathbf{E}[Y|X = 1, Z = z] = (\beta_0 + \beta_1) + (\beta_2 + \beta_4)1(z = 1) + (\beta_3 + \beta_5)1(z = 2)$, so that $(\beta_3 + \beta_5) - (\beta_2 + \beta_4) = 0$. Removing the redundancy, $H_0 : \beta_3 - \beta_2 = \beta_5 - \beta_4 = 0$.
(E) The shortcut is to jointly test terms involving influenza dose: $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

(A) __(2)__   (B) __(1)__   (C) __(4)__   (D) __(5)__   (E) __(3)__

(e) ☐ 3 pts ☐ In regards to *either* model, identify a circumstance under which you might expect including base-line age as a covariate to provide an advantage. Make clear what that advantage would be.

Ans: If baseline age were associated with mean vaccine reaction score, we would expect to estimate the model's parameters more precisely (i.e., with greater efficiency, or less variability) by adjusting for baseline age as compared to a model that does not.

(f) ☐ 2 pts ☐ Suppose you seek to add a low-dose influenza vaccine group to the study. How many parameters would the corresponding saturated model possess in this case?

Ans: The saturated model would involve $2 \times 4 = 8$ parameters.

5. **_Optional problem 1_**: 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.

We have covered a range of topics this semester thus far, with many connected by common themes. Here are a few examples of themes that we've touched on:

(I) A regression model's assumptions and how the relative importance of those assumptions depends on the purpose for which that model is used.

(II) Informed/intelligent spending of degrees of freedom in a regression model.

(III) The trade-off between building a regression model that closely reflects the data and the ease with which you can interpret its coefficients.

Choose _one_ of these themes and provide _one_ specific example of a topic we've covered in this class that fits in with this theme (make clear how that topic falls under the theme you've selected). Your response should be fairly brief (a maximum of five sentences).

Ans: There are many correct answers!

---

6. **Optional problem 2**: 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.

Suppose that ten independent observations are randomly sampled from a population of interest with the goal of evaluating the association between a predictor, X, and an outcome, Y. Consider the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$ (which you may assume to be correctly specified), and consider ordinary least squares estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. As usual, let $\mathbf{X}$ denote the design matrix and $\mathbf{y}$ the outcome vector. Recall that the ordinary least squares estimator is given by:

$$\widehat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

(a) Derive an expression for $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ that assumes homoscedasticity (that is, an error variance σ^2 that does not depend upon X) but does not rely on the values of X being fixed by design. Make clear where in your derivation key assumptions are being invoked.

Ans: The derivation is as follows:

$$
\begin{aligned}
\mathrm{Var}[\widehat{\boldsymbol{\beta}}] &= \mathrm{Var}[\mathbf{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \mathbf{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] && \text{(law of total variance)} \\
&= \mathrm{Var}[\boldsymbol{\beta}] + \mathbf{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] && \text{(unbiasedness)} \\
&= \mathbf{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] && \text{(the variance of a constant is zero)} \\
&= \mathbf{E}[\mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}]] && \text{(definition of } \widehat{\boldsymbol{\beta}}) \\
&= \mathbf{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathrm{Var}[\mathbf{y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] && \text{(property of variance)} \\
&= \mathbf{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\sigma^2\mathbf{I})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] && \text{(homoscedasticity and independence)} \\
&= \mathbf{E}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}] = \sigma^2\mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}] && \text{(simplification)}
\end{aligned}
$$

(b) Write an expression for an estimate of $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$. Note that your response should make it clear how you're estimating σ^2—that is, don't simply include $\widehat{\sigma}^2$ as part of your response without further elaboration.

Ans: This can be estimated as:

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] \;=\; \widehat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1},$$

where:

$$\widehat{\sigma}^2 \;=\; \frac{1}{8}\sum_{i=1}^{10}(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2.$$

(c) Although normality of errors was not required in order to obtain answers to parts (a) and (b), state at least one example in which it might it be nice if this additional assumption were in fact true. *Hint*: You could appeal to the most common uses of an estimated standard error, or alternatively you could appeal to a specific use of the regression model.

Ans: If we seek to form confidence intervals or obtain p-values involving, say, β_1, then error normality would give us valid confidence intervals without having to rely on something like the central limit theorem (often tenuous in the setting of only *ten* observations). We could also argue that error normality would be nice in order to form valid prediction intervals.

Appendix: Software output and diagnostic plots for problem 3

Stata output

```
. regress logthermal immunofluorescence, robust

Linear regression                               Number of obs   =        220
                                                F(1, 218)       =     623.34
                                                Prob > F        =     0.0000
                                                R-squared       =     0.7411
                                                Root MSE        =     .33564


------------------------------------------------------------------------------
                    |               Robust
        logthermal | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------------+----------------------------------------------------------
immunofluorescence |   3.540891   .1418245    24.97   0.000     3.261368    3.820414
             _cons |   .4757339   .0444449    10.70   0.000     .3881372    .5633307
------------------------------------------------------------------------------
```
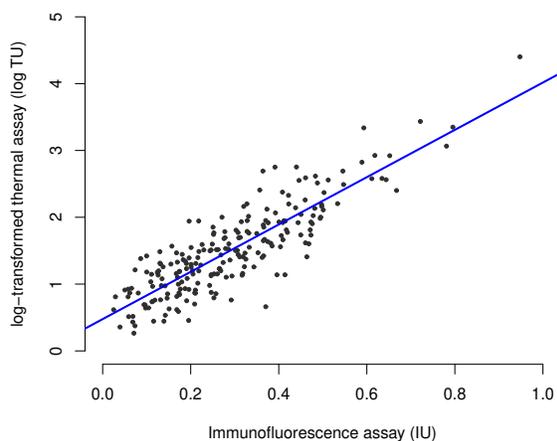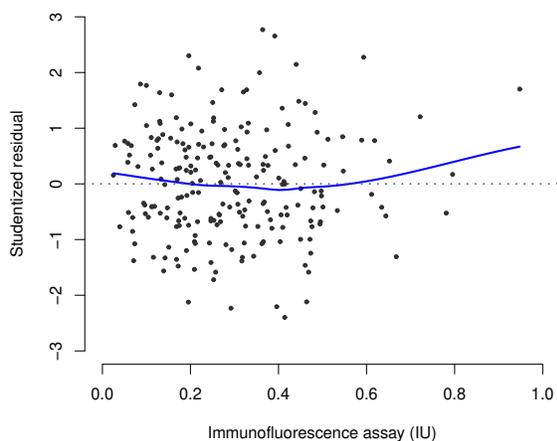
Diagnostic plots

(A) Scatter plot with fitted regression line

(B) Predictor vs. residual plot with LOWESS smoother

(C) Normal Q–Q plot