STAT 8561: Assignment 4

Fall 2026 | Linear Statistical Analysis I

Chi-Kuang Yeh

STAT 8561 | Fall 2026
Due date: Friday, November 13, 2026
Total: 40 points
Coverage: Chapter 5: Multiple Regression and Categorical Predictors, Chapter 6 through Section 6.3: Outliers, and a guided extension to Section 6.4: Influence. Chapter 4 model comparisons are used throughout.

Questions 1-3 [34 points] provide practice within the Exam 2 coverage through Section 6.3. Question 4 [6 points] is a guided extension to influence; it is not part of the announced Exam 2 coverage.

Question	Topic / format	Points
1	Interaction models and coefficient interpretation	12
2	Nested comparisons: improvement and evidence	10
3	Leverage and residual diagnostics	12
4	Guided extension: influence and Cook's distance	6
Total		**40**

Assignment PDF | Response template (.qmd)

Instruction

Use **Quarto** or **R Markdown**. Upload **both** files to the **Assignment 4** folder on **iCollege**:

1. Your rendered **PDF**, named `A4_LastName_FirstName.pdf`.
2. Its matching source, named `A4_LastName_FirstName.qmd` or `A4_LastName_FirstName.Rmd`.

The PDF must be generated from the submitted source. **HTML alone does not satisfy the submission requirement.**

- Put your name and course number at the beginning, and label every question and subpart. Use the response template or follow its labeling format; you do not need to copy a table.
- For multiple-choice items, enter exactly one letter. For True/False items, enter T or F. Give numerical answers in the requested order. Where requested, add only 1-2 sentences or a short counterexample; lengthy derivations are unnecessary. Equivalent correct reasoning is accepted.
- Round numerical answers to **three decimal places**, retaining full precision during calculations. Counts and degrees of freedom should be integers.
- R code and plots are optional. If you use R, include the code in executable chunks after your answers. Before generating random values, use `set.seed(8561)`.
- Your source must render from beginning to end without errors. Use LaTeX for mathematics; do not submit screenshots of answers or code.

Question 1: Interaction Models and Interpretation [12]

A hypothetical packaging study records processing time Y (seconds), temperature x (°C, observed from 40 to 60), and machine type A or B. Let $z = 0$ for A and $z = 1$ for B. Consider

$$\mu(x, z) = E[Y \mid x, z] = \beta_0 + \beta_x(x - 50) + \beta_g z + \beta_{xg}(x - 50)z,$$

with fitted mean

$$\hat{\mu}(x, z) = 40 - 0.4(x - 50) + 6z - 0.3(x - 50)z.$$

Reading guide. Substitute the group indicator before interpreting a slope. The group difference is obtained by subtracting the two conditional means. See interactions and conditional comparisons.

(a) A population coefficient [2]

What does β_x represent in this interaction model?

A. The B-minus-A mean difference at every temperature.
B. The change in population mean time per 1°C increase for machine A.
C. The common temperature slope for both machines.
D. The exact change in every individual processing time per 1°C.

(b) Machine B [2]

Report the fitted **B slope** (seconds per °C) and the fitted **B mean time at** $60°$**C** (seconds), in that order (**1 point each**).

(c) The group coefficient [2]

Interpret the fitted group coefficient **6** in one sentence. State both the comparison and the temperature at which it applies.

(d) Challenge a claim [2]

A student says, "The coefficient 6 means that B always has a fitted mean time 6 seconds higher than A throughout the observed temperature range." Use **one temperature other than** $50°$**C** to construct a numerical counterexample [**1**], and give a one-sentence explanation [**1**].

(e) Change the centering point [2]

Rewrite the **same fitted model** using $w = x - 60$, retaining the intercept, w, z, and wz. Report the new **intercept** and the new **coefficient of** z, in that order (**1 point each**).

(f) What changes? [2]

Enter `T` or `F` (**1 point each**).

1. The recentering in part (e), with all terms retained, changes the fitted values and SSE.
2. The fitted association alone does not establish that switching machine type causes the corresponding change in processing time.

Question 2: Nested Comparisons: Improvement and Evidence [10]

In a **separate hypothetical study** with the same variable definitions, three OLS models are fitted to the **same 36 observations**. All designs have full column rank; p includes the intercept.

Model	Mean structure	p	SSE
Common line	$\alpha_0 + \alpha_x(x-50)$	2	168
Additive	$\gamma_0 + \gamma_x(x-50) + \gamma_g z$	3	152
Interaction	$\beta_0 + \beta_x(x-50) + \beta_g z + \beta_{xg}(x-50)z$	4	128

For nested models use

$$F = \frac{(\mathrm{SSE}_R - \mathrm{SSE}_F)/(p_F - p_R)}{\mathrm{SSE}_F/(n - p_F)}.$$

Assume independent $N(0, \sigma^2)$ errors under the full model, conditional on the fixed design. Use $\alpha = 0.05$. The 95th percentile of $F_{1,32}$ is **4.149**. See Chapter 4: partial F tests.

(a) Equal slopes [2]

Which comparison tests whether the two machine types have equal temperature slopes?

A. Common line versus additive; $H_0 : \gamma_x = 0$.
B. Common line versus interaction; $H_0 : \beta_0 = 0$.
C. Additive versus interaction; $H_0 : \beta_{xg} = 0$.
D. Additive versus interaction; $H_0 : \beta_g = 0$.

(b) Carry out the comparison [3]

For the equal-slopes test, report the **F statistic** [1], the ordered pair of **numerator and denominator degrees of freedom** [1], and **reject** or **do not reject** [1].

(c) Assess a claim [3]

A student says, "The interaction fit has lower SSE, so that fact alone proves the population slopes differ."

In 1-2 sentences:

- Explain why the lower SSE alone is insufficient [1].
- State why the models must use the same response and observations for this comparison [1].
- State the error assumption that justifies the exact F reference distribution here [1].

(d) A different question [2]

Which null hypothesis tests whether the two population mean lines are identical within the interaction model?

A. $H_0 : \beta_g = \beta_{xg} = 0$, comparing common line with interaction.
B. $H_0 : \beta_{xg} = 0$, comparing additive with interaction.
C. $H_0 : \beta_x = 0$, deleting only the temperature main effect.
D. $H_0 : \beta_0 = \beta_x = 0$, deleting the reference-group line.

Question 3: Leverage and Residual Diagnostics [12]

A different full-rank OLS fit has an intercept, $n = 25$, $p = 3$, and full-fit residual SD $s = 2$. Two observations have:

Observation	Y_i	Full-fit $\hat{Y}_i$	h_{ii}	Deleted-fit residual SD $s_{(-i)}$
A	36	30	0.10	$\sqrt{16/7}$
B	61	60	0.60	$\sqrt{57/14}$

Use these definitions from Section 6.3:

$$e_i = Y_i - \hat{Y}_i, \qquad r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \qquad t_i = \frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}.$$

Here r_i is the standardized (internally studentized) residual and t_i is the externally studentized residual. In both formulas, e_i and h_{ii} come from the **full fit**. Only $s_{(-i)}$ is estimated after deleting observation i. These formulas are supplied; no supplementary derivation is required. Average leverage is p/n.

(a) Raw residuals [2]

Report the raw residuals for **A** and **B**, in that order (**1 point each**).

(b) Two scaled residuals [2]

For **observation A**, calculate r_A and t_A, in that order (**1 point each**).

(c) Predictor position [2]

Which observation, A or B, has greater leverage [**1**]? In one sentence, explain why this conclusion does not depend on the observed response values [**1**].

(d) An unusually large response? [2]

A student labels B a response outlier solely because its observed response is 61, larger than A's 36. Give a 1-2 sentence correction using the fitted means and residuals.

(e) Does external studentization always enlarge a residual? [2]

Without further fitting, determine whether $|t_B|$ is **larger or smaller** than $|r_B|$ [**1**] by comparing $s_{(-B)}$ with s. Use this case to assess the claim that external studentization **always increases** the absolute standardized residual, giving a one-sentence reason [**1**].

(f) A shift in measurement origin [2]

Add 100 to every response, retain the same predictor matrix and intercept, and refit. Which statement is correct?

A. All leverages increase by 100.
B. All raw residuals increase by 100.
C. Only the externally studentized residuals change.
D. Leverages and both types of scaled residual remain unchanged.

Question 4: Guided Extension: Influence [6]

Beyond the announced Exam 2 scope. This part previews Section 6.4: influence and Cook's distance.

Reading guide. Influence concerns how much the fitted result changes when an observation is omitted. Cook's distance combines response discrepancy and leverage:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}.$$

Here r_i is the **internally** studentized residual, p counts all design columns, and $h_{ii} < 1$. A large value is a reason to investigate the observation and compare fits, not an automatic deletion rule or a p-value.

In a full-rank fit with $p = 3$, two candidate observations have:

Observation	r_i	h_{ii}
C	3	0.08
D	1	0.60

(a) Calculate Cook's distance [2]

Report D_C and D_D, in that order (**1 point each**).

(b) Compare the diagnostics [2]

Which interpretation agrees with these results?

A. C is more influential because its absolute standardized residual is larger.
B. D has greater Cook's distance because its leverage more than offsets its smaller standardized residual.
C. Both observations have the same Cook's distance.
D. D must be a recording error.

(c) What to do next [2]

A student proposes deleting D immediately. Give **two appropriate next steps**, in 1-2 sentences (**1 point each**).