Lecture 6: Other Considerations in Linear Regression

Friday, Oct 10

Upcoming + Reminders

Updates:

- Project groups have been assigned! **Please contact your group members** and decide on a method of communication ASAP.
  - If someone is unresponsive for ~1 week (or has dropped the class), please let me know.
- HW1 & Quiz 1 feedback is released—*please take a look on **Gradescope/Canvas!***
- Check out the exam bank (*on website, under Week 2: General*)

Assignments:

- **#FinAid Survey**—please complete it by today!

Friday's topic: *Other Considerations in Regression*

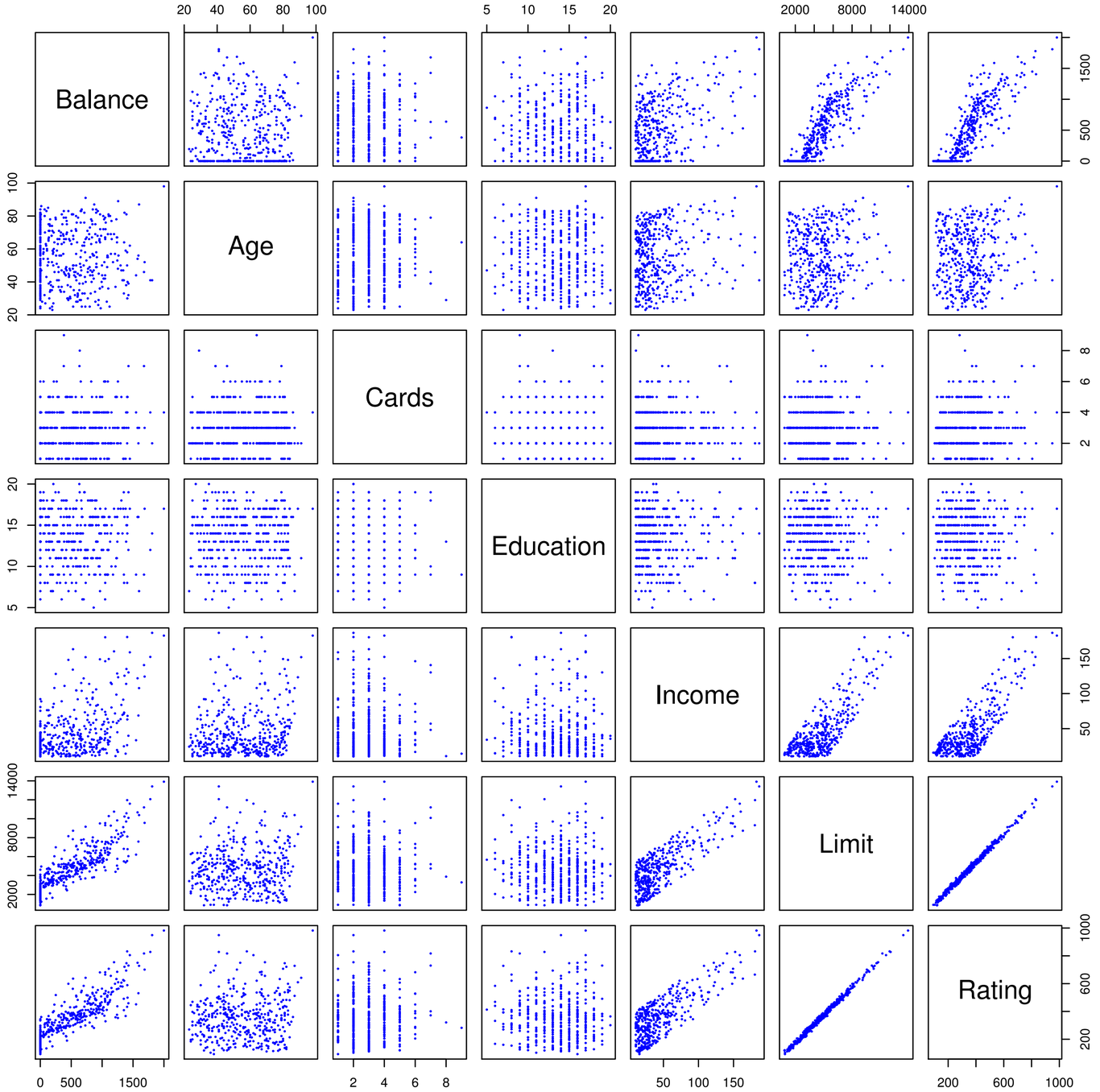
- **Read:** ISLP Ch. 3.3

Agenda

- Extensions of the linear model
  - Qualitative variables
  - Interaction effects (won't have time to cover this, but you can explore in your project!)
- Potential problems that you may encounter in linear regression
- Examples of MLR variations for group project
- Exam review (~20-25 mins)

Qualitative variables

- Some predictors are not quantitative but are qualitative (discrete, categorical).
- **Credit card data set** contains 7 quantitative and 4 qualitative variables:
  - own (home ownership, yes/no)
  - student (yes/no)
  - married (yes/no)
  - region (location, East/West/South)



Qualitative variables

- **Example:** We want to investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables for now.
- If qualitative predictor (or factor) only has two levels (possible values), we create an indicator or dummy variable for own:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

Qualitative variables

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

Interpretations:

- β_0 is the average credit card balance for those who **do not** own a house
- $\beta_0 + \beta_1$ is the average credit card balance for those who **do** own a house
- β_1 is the average **difference** in credit card balance between owners and non-owners.

*"Credit card balance is β_1 **higher** on average for owners compared to non-owners"*

Qualitative variables

- Predictors with more than two levels or factors (possible values), we can create more dummy variables! Take region (South, West, East) for example:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South} \\ 0 & \text{if } i\text{th person is not from the South} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases}$$

- Always have **one fewer** dummy variables than the **number of levels** in a predictor
- The level with no dummy variable is called the baseline (e.g., East)

Qualitative variables

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East.} \end{cases}$$

Interpretations:

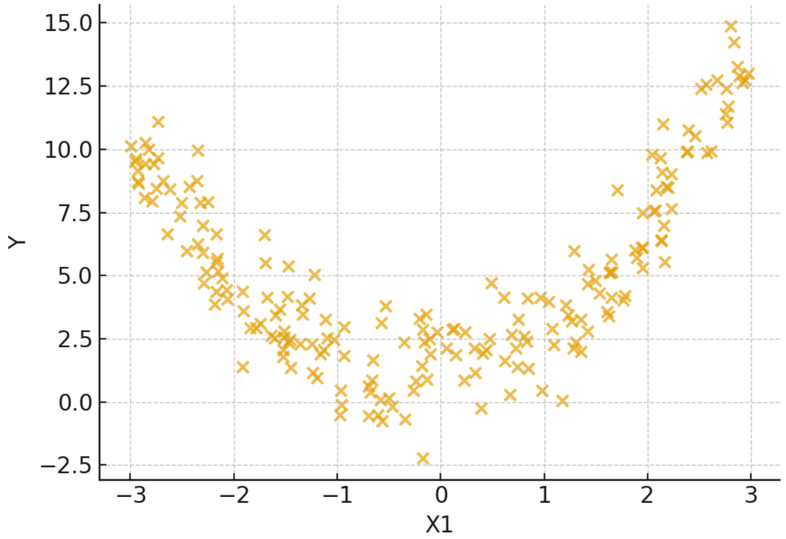
- β_0 is the average credit card balance for people who live in the East
- β_1 is the average **difference** in credit card balance between people in the South vs East
- β_2 average **difference** in credit card balance between people in the West vs East

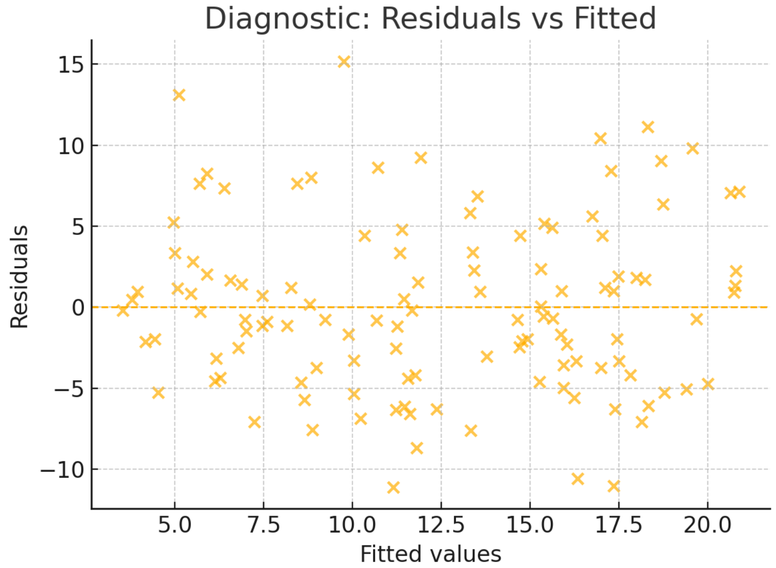
	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`region[South]`	−12.50	56.68	−0.221	0.8260
`region[West]`	−18.69	65.02	−0.287	0.7740

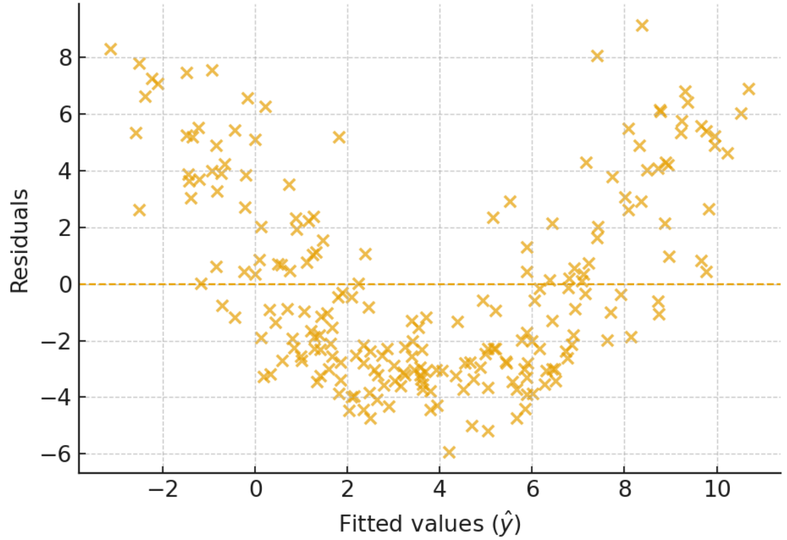
What is the average credit card balance for people who live in the…
South? West? East?

Potential problems

Non-linearity







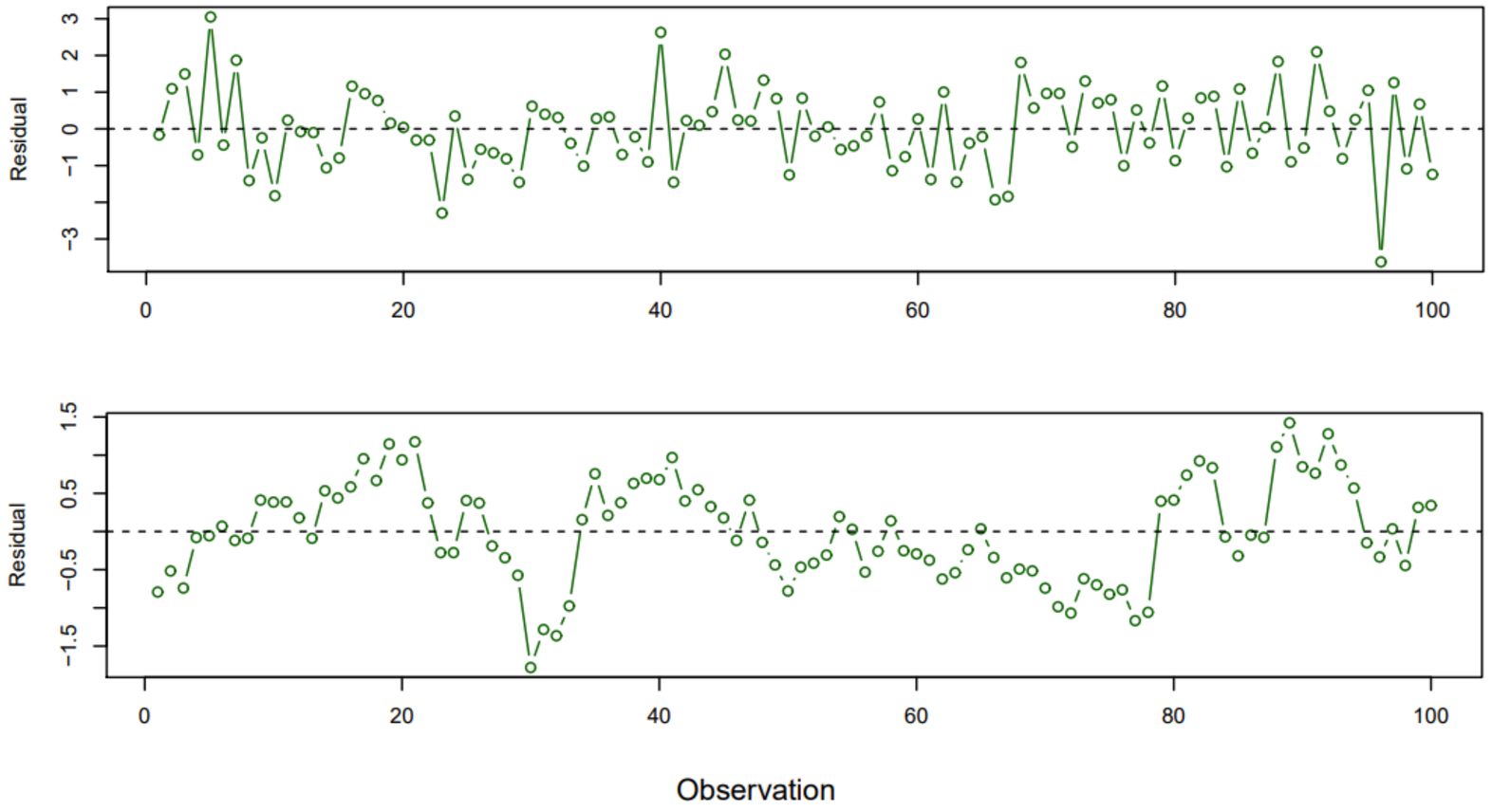
❌ ✅ ❌

Solutions:

- Use non-linear transformations of the predictors X's
- Choose a non-linear model

Potential problems

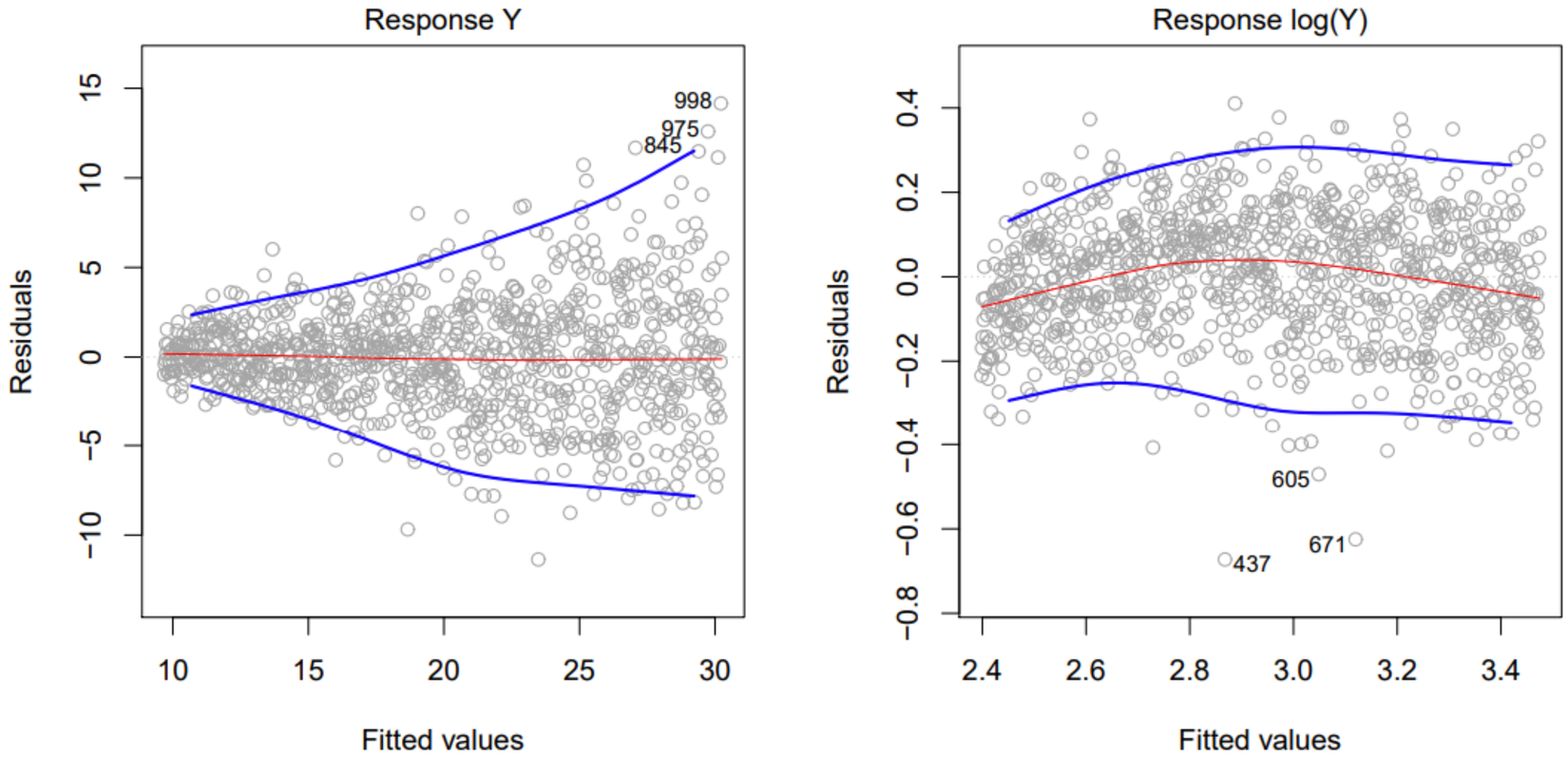
Correlation of residuals / errors



One solution: consider a time-series model instead

Potential problems

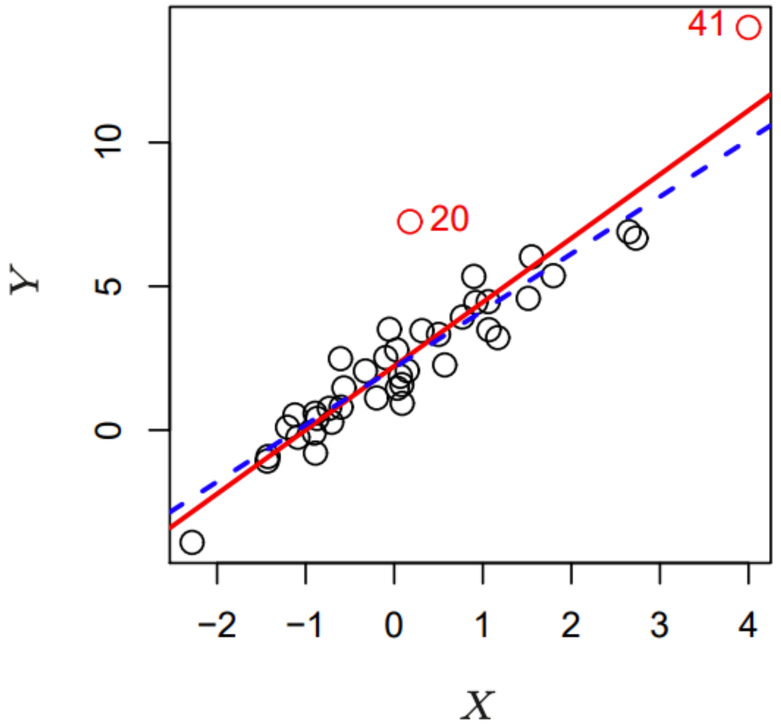
Heteroscedasticity: non-constant variance of errors



One solution: log-transform your Y's and fit new model: $\log(Y) = \ldots$

Potential problems

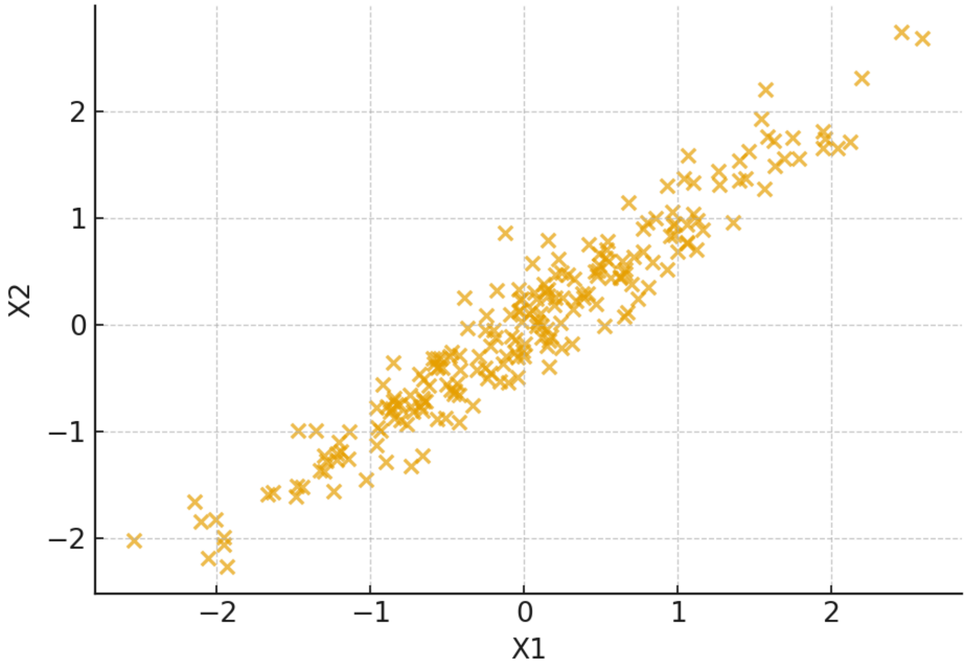
Outliers and high leverage points



Can remove these points, but must state your criteria

Potential problems

Collinearity



- Compute the *variance inflation factor* (VIF)
- Small VIF (≈1) indicates the complete absence of collinearity
- Large VIF (>5-10), drop the covarying variable

Investigating these potential problems in your dataset is one way to go *above and beyond* in your final project!

Linear regression project example

Example project contribution for <u>one student</u> within the group

Dataset: Student's GPA, StudyHours, Attendance, Sleep, Year, Major, etc.

Variant 1: Backward Selection

Identify the most relevant predictors by starting with all and removing one-by-one based on p-values.

$$\text{GPA} = \beta_0 + \beta_1(\text{StudyHours}) + \beta_2(\text{Attendance}) + \beta_3(\text{Sleep}) + \epsilon$$

Variant 2: Adding a Qualitative Variable

Include a categorical factor (e.g., Year, Major) to test if belonging to a certain group affects GPA.

$$\text{GPA} = \beta_0 + \beta_1(\text{StudyHours}) + \beta_2(\text{Attendance}) + \beta_3(\text{Sleep}) + \beta_4(\text{Year}) + \beta_5(\text{Major}) + \epsilon$$

Variant 3: Adding Interaction Terms

Does the effect of Sleep depend on Study Hours? Reveals whether there are combined effects between predictors.

$$\text{GPA} = \beta_0 + \beta_1(\text{StudyHours}) + \beta_2(\text{Attendance}) + \beta_3(\text{Sleep}) + \beta_4(\text{Sleep} \times \text{StudyHours}) + \epsilon$$

Exam 1 Details — Monday, October 13th

- **14 MC questions + 2 FR questions** (~4-6 sub-questions/FR, just like quiz)
- 50 mins (9-9:50am) — arrive early!
- Covers Week 1-2 material
  - **To study:** review HW1-2, quizzes, student-submitted exam Q bank
  - **Quiz 2** is “optional,” but will probably help you prep!

Exam 1 Details — Monday, October 13th

Week 1

- Reducible vs. Irreducible error
- Prediction vs. Inference
- Parametric vs. Non-parametric models
- Flexibility vs. Interpretability
- Supervised vs. Unsupervised learning
- Regression vs. Classification
- Training vs. Test accuracy (MSE and error rate)
- The bias-variance trade-off
- Overfitting and generalization
- Hypothesis vs. data-driven questions

Week 2

- Simple & multiple linear regression setup and assumptions
- Estimating coefficients
- Interpreting coefficients, confidence intervals, and measures of statistical significance
- Identifying important predictors and variables
- Model fitting and assessment metrics (RSS/RSE/RMSE, etc)
- Interpreting qualitative predictors

Some suggestions:

- Come to my / TA's office hours :)
- Go to section
- Start your homework early
- Read the textbook (HW should help with that!)
- Review your HW feedback
- Take the quiz multiple times / until you understand everything you got wrong
- Ask ChatGPT to tutor you!
  - "*I still don't understand ….*"
  - "*Can you help me figure out what I don't understand about…*"
  - "*Explain to me at the level of a…*"

Exam day reminders!

Bring:

- A pen/pencil
- (Optional) one 8.5x11” cheat-sheet, **one-sided**

You will use a bubble sheet for the multiple-choice questions

When you enter:

1. Put your phones on silent
2. Take out your cheat sheet and show us front/back
3. Pick up exam at the front

Exam practice