STA 35C: Statistical Data Science III

Lecture 6: Qualitative Predictors & Potential Problems in Linear Regression

Dogyoon Song

Spring 2025, UC Davis

Agenda

Last time: Multiple linear regression

- Model
- Estimation via least squares
- Some key statistical questions
- Incorporating non-linear relationships

Today:

- Qualitative predictors
- Potential problems in linear regression
- Comparison: linear regression vs. k-NN

Qualitative predictors: Motivation

Motivating example: `Credit` dataset

- Response: `balance`
- Quantitative predictors: `age`, `cards`, `education`, `income`, `limit`, `rating`
- Qualitative (categorical) predictors: `own`, `student`, `status`, `region`
  - These do not have a natural numeric scale

Question: How do we incorporate categorical variables into a linear regression model?

- `balance`$= -0.4 \times$ "`own` a house" $+ 2.33 \times$ "not a `student`" $- \ldots$?

Answer: Use a "dummy variable" to numerically encode each categorical level

Dummy variables

Idea: Convert a qualitative (categorical) predictor into *dummy* (*indicator*) variables

Case 1: Two-level factor

- Example: Homeowner status `own` $\in \{\texttt{Yes}, \texttt{No}\}$
- Create a dummy variable: $D = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}$
- In regression: $Y = \beta_0 + \beta_1 D + \cdots + \epsilon$

Case 2: More than two levels

- Example: `region` $\in \{\texttt{East}, \texttt{West}, \texttt{South}\}$
- Create $K - 1$ dummies if there are K categories (with one level setting a baseline):

$$\texttt{East}, \texttt{West}, \texttt{South} \quad \rightarrow \quad D_{\text{West}}, \; D_{\text{South}}$$

Interpretation of the regression coefficient

Simple linear regression setup (with a dummy):

$$Y = \beta_0 + \beta_1 D + \epsilon, \quad \text{where } D \in \{0, 1\}.$$

- If $D = 0$: $Y = \beta_0 + \epsilon$.
- If $D = 1$: $Y = (\beta_0 + \beta_1) + \epsilon$.
- β_1: The *difference* between the two group means ($D = 1$ vs. $D = 0$)

Again, we can use standard errors to compute t-stats, and p-values for hypothesis testing:

- $H_0 : \beta_1 = 0 \quad \Longrightarrow$ *no difference*
- $H_1 : \beta_1 \neq 0 \quad \Longrightarrow$ *significant difference*

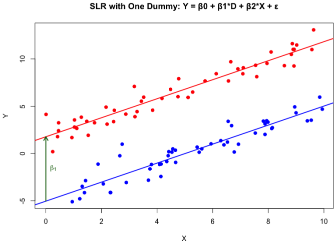
Interpretation of the regression coefficient (continued)

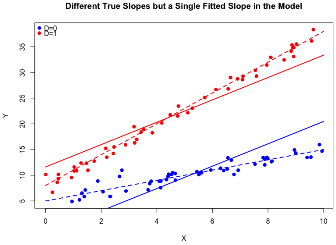
Potential complications:

- When additional X are present:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \epsilon, \quad \text{where } D \in \{0, 1\}$$

  - β_1 reflects the *average* effect of D, *holding X fixed*
  - It may not represent a *constant* difference if other interactions are present





- Using different coding schemes ({0,2} or {-1,1}, etc.) changes the interpretation of β_0 and β_1, but not the *predictions*

Pop-up quiz: Linear regression with a dummy variable

Model:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \epsilon,$$

where $D = 1$ for treatment and $D = 0$ for control.

Question: Which choice *most correctly* interprets β_0, β_1, β_2, and a large p-value for β_1?

A) β_0 is mean outcome for treatment at $X = 0$; β_1 is difference in slope; β_2 is slope for control; a large p-value means X has no effect.

B) β_0 is mean outcome for control at $X = 0$; β_1 is difference in intercept (treatment vs. control); β_2 is the common slope; a large p-value means no evidence of an intercept difference.

C) β_0 is a shared intercept; β_1 is the slope for $D = 1$; β_2 is slope for $D = 0$; a large p-value means no effect of X.

D) β_0 is the intercept at $X = 1$; β_1 is slope for control; β_2 is slope for treatment; a large p-value means the treatment group has a zero slope.

Potential pitfalls in linear regression

Linear regression is powerful, but it can fail if certain assumptions are not met

Possible issues:

- Validity of model assumptions
  - Is the Y-X relationship truly linear?
  - Are the errors ϵ_i truly uncorrelated?
  - Is the variance of ϵ constant?
- Outliers & High-leverage points
  - What if there are extremely unusual points in the training data?
- Collinearity among predictors
  - What if some predictors are highly correlated?

Let's examine what these problems entail, how to diagnose and possibly address them

Problem 1: Nonlinear relationship

Problem: The response–predictor relationship may not be linear

- Example: $Y \approx \beta_0 + \beta_1 X^2 + \epsilon$
- A purely linear model would systematically misfit (leading to large residuals)

Diagnosis: Residual plots often reveal a pattern (e.g., a systematic deviation from 0)

Remedies: (1) Include nonlinear transformations of X; (2) Use more flexible models

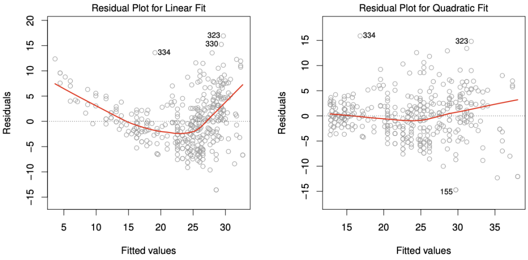


Figure: Plots of residuals vs. predicted values [JWHT21, Figure 3.9].

Problem 2: Correlated error terms

Problem: Errors $\{\epsilon_i\}$ correlated rather than independent

- Common in time series or grouped data (e.g., repeated measurements)
- If data is artificially duplicated or has a temporal pattern, errors can "track" each other

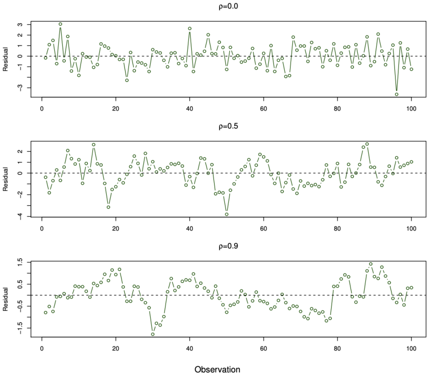


Figure: Plots of residuals from simulated time series data [JWHT21, Figure 3.10].

Problem 2: Correlated error terms

Problem: Errors $\{\epsilon_i\}$ correlated rather than independent

- Common in time series or grouped data (e.g., repeated measurements)
- If data is artificially duplicated or has a temporal pattern, errors can "track" each other

Issue: Standard errors (thus *p*-values and confidence intervals) can be *underestimated*

Diagnosis: Examine residuals vs. time or group for systematic patterns

Possible remedies:

- Tailored techniques in time series (ARIMA, etc.) or grouped data
- Generically, careful experimental design to avoid correlated errors

Problem 3: Non-constant variance of the error term

Problem: Heteroskedasticity (non-constant variance) of the errors

- $\mathrm{Var}(\epsilon_i)$ not constant for each data point
- Classic OLS assumption is $\mathrm{Var}(\epsilon_i) = \sigma^2$ (constant)

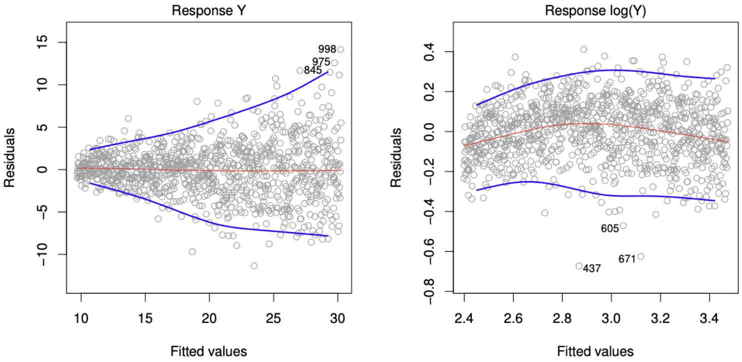


Figure: Residual plots with heteroskedastic error [JWHT21, Figure 3.11].

Problem 3: Non-constant variance of the error term

Problem: Heteroskedasticity (non-constant variance) of the errors

- $\mathrm{Var}(\epsilon_i)$ not constant for each data point
- Classic OLS assumption is $\mathrm{Var}(\epsilon_i) = \sigma^2$ (constant)

Issue: Distorts standard errors and inference; *RSE* may be biased

Diagnosis: Check residual plots to detect a "funnel" shape

Possible remedies:

- Transform the response ($\log Y$, $\sqrt{Y}$, etc.) to stabilize variance
- Use weighted least squares to downweight high-variance points

Problem 4: Outliers & high-leverage points

Definitions:

- *Outlier:* An observation where y_i is "very far" from its predicted value $\hat{y}_i$.
- *High-leverage point:* A point with unusual x_i; it can strongly influence the fit
- *Leverage score* $h_i = [X(X^\top X)^{-1}X^\top]_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$ takes value between $\frac{1}{n}$ and 1

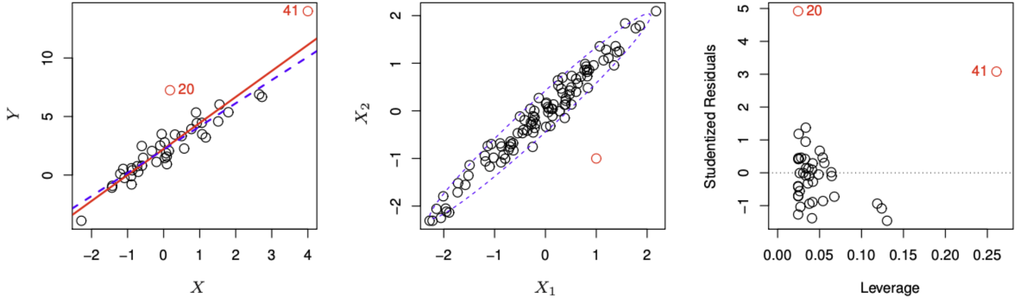


Figure: An illustration of outliers and high-leverage points [JWHT21, Figure 3.13].

Problem 4: Outliers & high-leverage points

Definitions:

- *Outlier:* An observation where y_i is "very far" from its predicted value $\hat{y}_i$.
- *High-leverage point:* A point with unusual x_i; it can strongly influence the fit
- *Leverage score* $h_i = [X(X^\top X)^{-1}X^\top]_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$ takes value between $\frac{1}{n}$ and 1

Why worry?

- Outliers can lead to a misfit, inflate RSE, and degrade R^2.
- A small change in high-leverage points can pull the regression line substantially

Diagnosis:

- Residual plots, especially *studentized residuals*, can help identify outliers
- Plot leverages or Cook's distance to find high-leverage points.

Possible remedies:

- Inspect and possibly remove or adjust suspicious observations
- Use a "robust" statistical method

Problem 5: Collinearity

Definition: Two (or more) predictors are highly correlated

- Example: $X_2 = X_1 + \text{small noise}$, or $X_3 = -2X_1 + 3X_2$, etc.

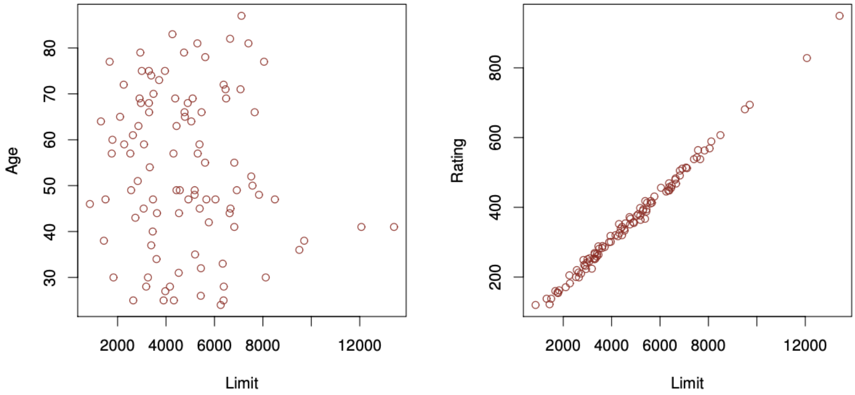


Figure: An illustration of high collinearity [JWHT21, Figure 3.14].

Problem 5: Collinearity

Definition: Two (or more) predictors are highly correlated

- Example: $X_2 = X_1 +$ small noise, or $X_3 = -2X_1 + 3X_2$, etc.

Problem:

- Difficult to separate individual effects
- Coefficients may become *unstable*, with large standard errors

Diagnosis:

- Correlation matrix among predictors
- Variance Inflation Factor (VIF): $\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}$

Simple remedies:

- Drop one of the correlated predictors
- Combine or merge them (e.g., sum, average, or principal components)
- Use regularization techniques (e.g., ridge, lasso) to reduce variance

Pop-up quiz: Spot the problem and suggest a remedy

Scenario: You fit a linear regression model and notice the residual plot has a distinct "funnel" shape, where the spread of residuals grows wider as the fitted values increase.

Question 1: Which problem does this indicate, and what is one possible remedy?

A) *Correlated errors*; consider using mixed models or time-series methods.
B) *Non-constant variance*; stabilize variance by transforming the response or using weighted least squares.
C) *Outliers*; remove data points with excessively large studentized residuals.
D) *Collinearity*; drop or combine highly correlated predictors, or use regularization.

Question 2: If we ignore this issue and proceed with standard OLS, which is most likely?

A) Coefficient estimates could be heavily biased.
B) The data becomes unusable for any regression method.
C) All predictors will appear perfectly correlated.
D) The standard error is misestimated, leading to misleading inference.

Comparison: Linear regression vs. k-NN

Linear regression (parametric):

- Assumes $f(X)$ is approximately linear in X
- Fits a small number of parameters $(\beta_0, \ldots, \beta_p)$
- Inference is straightforward (confidence intervals, p-values, etc.)

k-nearest neighbors (kNN) (non-parametric)

- Predicts y at a new point x_0 by averaging y_i of its k nearest neighbors

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_k(x_0)} y_i, \quad \mathcal{N}_k(x_0) : k\text{-neighborhood of } x_0$$

- No explicit model assumption such as linearity
- Instead, the complexity lies in defining "closeness" and choosing k

Visualization of k-NN

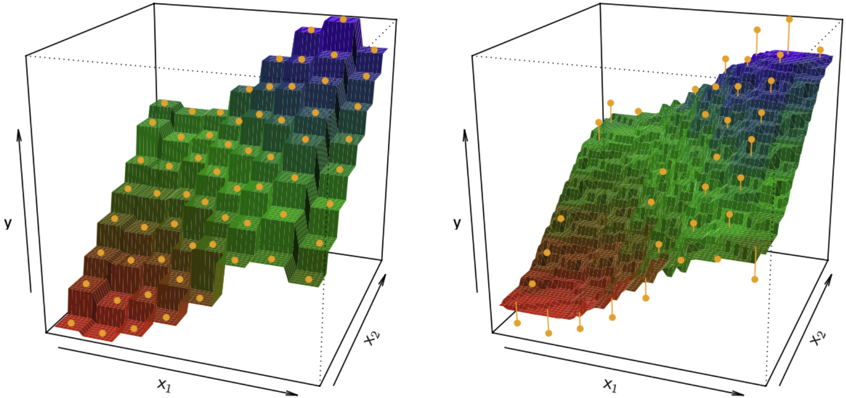


Figure: An illustration of kNN method (Left: $k = 1$; Right: $k = 9$) [JWHT21, Figure 3.16].

Comparison: Parametric vs. nonparametric

When linear regression shines:

- The linear model is a good approximation to reality
- The number of predictors is large, but sample size is moderate
- We need interpretable coefficients for inference (CIs, p-values)

When nonparametric methods (like k-NN) outperforms:

- Fewer assumptions, can capture more complex relationships
- Perform well in low-dimensional settings ("curse of dimensionality" if p is large)
- Often better for pure prediction if plenty of data is available

Wrap-up

- **Qualitative (categorical) predictors**:
  - Represented using dummy variables (indicator)
  - Interpretation as a "shift" across groups
- **Pitfalls in linear regression**:
  - *Model assumptions:* Non-linearity, correlated errors, heteroskedasticity
  - *Unusual data points:* Outliers & high-leverage points
  - Collinearity among predictors
- **Comparison:** Linear regression vs. k-NN
  - Parametric vs. nonparametric trade-offs

Next lecture: Assessing model accuracy & the bias-variance tradeoff

References

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An Introduction to Statistical Learning: with Applications in R, volume 112 of *Springer Texts in Statistics*.
Springer, New York, NY, 2nd edition, 2021.