STA 35C: Statistical Data Science III

Lecture 7: Qualitative Predictors & Potential Problems in Linear Regression

Dogyoon Song

Spring 2026, UC Davis

Announcements

Homework 2 is due tomorrow (Tue, Apr 14, 11:59 PM)

- Please submit on time and follow the submission instructions
  - To Gradescope via Canvas in a single PDF
  - Email submissions are not accepted

If you have questions or need help

- Discussion section tomorrow
- TA office hours: Tue, 12:30 PM – 2:30 PM (MSB 1143)
- Feel free to ask questions during lecture, in office hours or post them on Piazza

Agenda

Last time: Multiple linear regression

- Model, least squares, coefficient interpretation, model fit
- Polynomial regression as an extension

Today:

- More on inference
- Qualitative predictors via dummy variables
- Common pitfalls in linear regression

Recap: Linear regression

- Linear model: $Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{\text{signal}} + \underbrace{\varepsilon}_{\text{noise}}$
- Learn and use a model with estimated coefficient:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

via *least squares estimation* that minimizes

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(\hat{y}_i - y_i\right)^2 = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} - y_i\right)^2$$

- *Interpretation:* β_j is the average change in the mean response per unit increase in X_j, controlling the others
  - Association, not causation
- Model fit is often summarized by RSE and R^2

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \qquad \text{vs.} \qquad R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$$

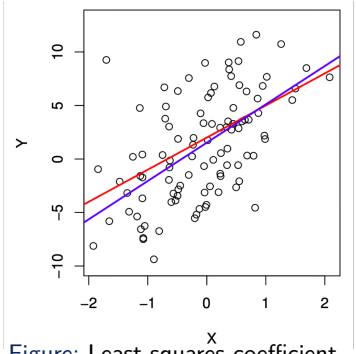


Figure: Least squares coefficient estimates from 100 data points. Red: population regression line, Blue: least squares line [JWHT21, Figure 3.3].

Estimation and inference

The coefficients $\beta_0, \ldots, \beta_p$ are fixed but unknown

- $\hat{\beta}_0, \ldots, \hat{\beta}_p$ depend on the random training sample $\rightarrow$ random variables

Estimation: Produce point estimates or interval estimates of unknown quantities

- Example: Use $\hat{\beta}_j$ to estimate β_j

Inference: Draw conclusion about the population based on data

- Construct a confidence interval for β_j
- Test whether $\beta_i = 0$ or not
- Form an interval for the response at a new predictor value x_{new}

All of these rely on the *sampling distributions* of the estimators

Recap: Confidence interval for β_j and hypothesis testing

Key fact: Under the standard linear-model assumptions,

$$t = \frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)}$$

has a t-distribution with $n - p - 1$ degrees of freedom

- $\mathbb{E}[\hat{\beta}_j] = \beta_j$
- $\mathrm{Var}(\hat{\beta}_j) = \mathrm{SE}(\hat{\beta}_j)^2$

$\mathrm{SE}(\hat{\beta}_j)$ involves $\sigma^2 = \mathrm{Var}(\varepsilon) \longrightarrow$ use $\widehat{\mathrm{SE}}(\hat{\beta}_j)$

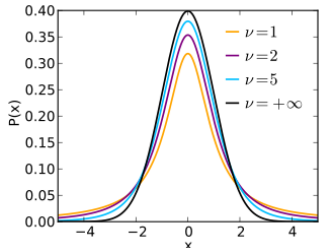


Figure: t-distribution (source: Wikipedia).

Confidence interval for β_j: Across repeated samples,

$$P\left(\beta_j \in \left[\hat{\beta}_j - 1.96\,\widehat{\mathrm{SE}}(\hat{\beta}_j),\ \hat{\beta}_j + 1.96\,\widehat{\mathrm{SE}}(\hat{\beta}_j)\right]\right) \approx 0.95$$

* For $n \gg 1$, t-distn is close to Gaussian

Hypothesis test: Assuming $H_0 : \{\beta_j = 0\}$,

$$t = \frac{\hat{\beta}_1 - \beta_j}{\widehat{\mathrm{SE}}(\hat{\beta}_1)} \text{ is likely near 0}$$

$\rightarrow$ If $|t|$ is large, then reject H_0

Predicting Y: Confidence and prediction intervals

Given a new test point $x_{\text{new}} = (x_{\text{new},1}, \ldots, x_{\text{new},p})$, we want to predict Y at $X = x_{\text{new}}$

$$\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new},1} + \cdots + \hat{\beta}_p x_{\text{new},p}$$

How uncertain is this prediction?

- $\hat{y} = \hat{f}_{\hat{\beta}}(x)$ only *estimates* $f_\beta(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$ $\rightarrow$ variability due to estimates $\hat{\beta}_j$
- $y = f(x) + \varepsilon$ has a noise term ε $\rightarrow$ *irreducible error* due to noise

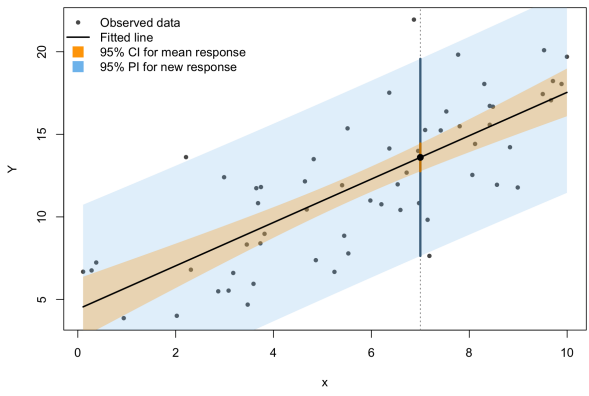
Confidence interval for the mean response $f(x) = \mathbb{E}[Y \mid X = x]$:

- Reflects uncertainty in the average response at predictor value x, due to estimating f
- Across repeated samples, 95% of CIs would contain the true *mean response* at x

Prediction interval for a new observation $Y_{\text{new}} \mid X = x$:

- Accounts for uncertainty in both $\hat{y} = \hat{f}(x)$ *and* the random noise ε
- Across repeated samples, 95% of PIs would contain a new *observed response* at x

Confidence and prediction intervals: Illustration



Fix a predictor value x_{new}

- Inner band: 95% CI for the mean response $E[Y \mid X = x_{\text{new}}]$
- Outer band: 95% PI for a new observation $Y_{\text{new}} \mid X = x_{\text{new}}$
- The PI is always wider than the CI because it includes both coefficient uncertainty and irreducible noise

Exact formulas are beyond our scope, but in **R**:

```
predict(model, newdata = x0_df, interval = "confidence", level = 0.95)
predict(model, newdata = x0_df, interval = "prediction", level = 0.95)
```

Pop-up quiz: Confidence interval vs. prediction interval

Question: At a fixed predictor value x_{new}, which statement is correct?

A) The confidence interval for the mean response is wider, because estimating a mean is harder than predicting a single value.

B) The prediction interval for a new observation is wider, because it includes both estimation uncertainty and random noise.

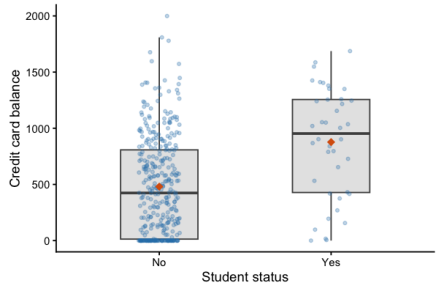
C) The two intervals should have the same width once n is moderately large.

D) Confidence intervals and prediction intervals answer exactly the same question.

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`" $- \ldots$?
- We cannot put labels like `Yes`/`No` or `East`/`West`/`South` directly into a linear model

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

Dummy variables

Idea: Represent each category by indicator (dummy) variables.

Two levels: For each binary variable, create a dummy variable:

$$D = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}$$

Then use D as a predictor in regression:

$$Y = \beta_0 + \beta_1 D + \cdots + \epsilon$$

- Example: Homeowner status `own` $\in \{$`Yes`, `No`$\}$

K levels ($K \geq 2$): Use $K - 1$ dummies (using one level setting as a baseline)

Dummy variables: Illustration

Example

We encode `region` $\in \{$`East`, `West`, `South`$\}$ using dummy variables

Region	D_{West}	D_{South}
East	0	0
West	1	0
South	0	1

- East: $\mathbb{E}[Y] = \beta_0$
- West: $\mathbb{E}[Y] = \beta_0 + \beta_1$
- South: $\mathbb{E}[Y] = \beta_0 + \beta_2$

$$Y = \beta_0 + \beta_1 D_{\text{West}} + \beta_2 D_{\text{South}} + \varepsilon$$

- The intercept $\beta_0$is the baseline-group mean
- Each dummy coefficient β_1, β_2 is a difference from the baseline

Why $K - 1$ dummies?

- If we also included D_{East}, then the intercept is redundant
- Naively coding `East`=1, `West`=2, `South`=3 would impose a spurious order/spacing

Interpretation of dummy variable 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:

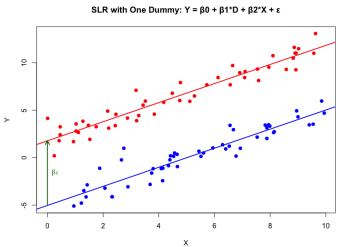
- $H_0 : \beta_1 = 0 \quad \Longrightarrow$ *no difference*
- $H_1 : \beta_1 \neq 0 \quad \Longrightarrow$ *significant difference*

Dummy variables with a quantitative predictor

No interaction:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon$$

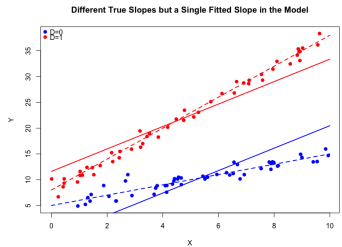
- If $D = 0$: $Y = \beta_0 + \beta_2 X$
- If $D = 1$: $Y = (\beta_0 + \beta_1) + \beta_2 X$
- β_1 is a constant shift between two groups



With interaction:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + \varepsilon$$

- If $D = 0$: $Y = \beta_0 + \beta_2 X$
- If $D = 1$: $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X$
- β_3 allows the slope to differ by group



$\rightarrow \beta_1$ reflects the *average* effect of D, *holding* X *fixed*; not necessarily a constant effect

Pop-up quiz: Dummy variables

Suppose that a predictor `region` has three levels:

$$\{\texttt{East}, \texttt{West}, \texttt{South}\}.$$

We fit the model

$$Y = \beta_0 + \beta_1 D_{\text{West}} + \beta_2 D_{\text{South}} + \varepsilon,$$

using `East` as the baseline, and obtain

$$\hat{\beta}_0 = 400, \qquad \hat{\beta}_1 = 120, \qquad \hat{\beta}_2 = -50.$$

Question: Which statement is correct?

A) The predicted mean is 400 for East, 520 for West, and 350 for South.

B) The predicted mean is 0 for East, 120 for West, and -50 for South.

C) West is predicted to have 70 less than South.

D) We should also include D_{East} to make the model complete.

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

Pitfall 1: Nonlinear relationship

Issue: The mean response is not linear in the predictors

- Example: $Y \approx \beta_0 + \beta_1 X^2 + \epsilon$

Why problematic: A linear fit is systematically biased (leading to large residuals)

How to diagnose: Residuals vs. fitted values show curvature or other systematic patterns (e.g., a systematic deviation from 0)

Possible remedies:

- Include nonlinear transformations of X
- Use a more flexible model

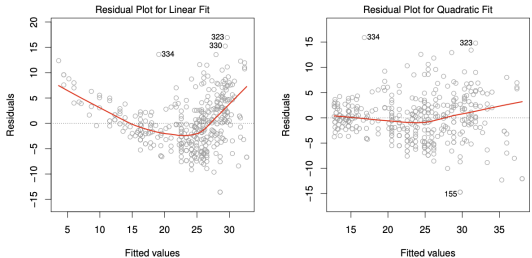


Figure: Plots of residuals vs. predicted values [JWHT21, Figure 3.9].

Pitfall 2: Correlated error terms

Issue: Errors $\{\epsilon_i\}$ are correlated, not 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

Why problematic: Coefficient estimates may still be okay, but standard errors (thus p-values and confidence intervals) can be *underestimated*

How to diagnose: Plot residuals in data order, by time, or by group; look for systematic patterns

Possible remedies:

- Use tailored methods for dependent data such as time series (ARIMA, etc.) or grouped data
- Carefully design experiments to avoid correlated errors

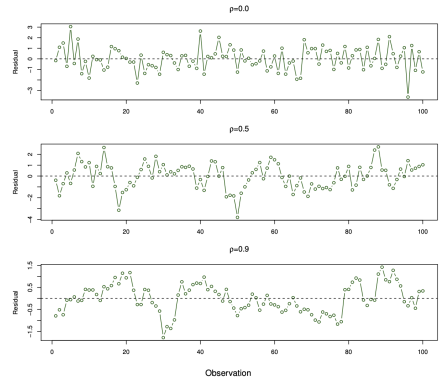


Figure: Plots of residuals from simulated time series data [JWHT21, Figure 3.10].

Pitfall 3: Non-constant variance of the error term

Issue: The error variance is not constant (heteroskedastic)

- Standard assumption is $\mathrm{Var}(\epsilon_i) = \sigma^2$ (constant)

Why problematic: OLS estimates are often still unbiased, but standard errors and inference can be unreliable

How to diagnose: Residual plots show a funnel shape or changing spread across fitted values

Possible remedies:

- Transform the response ($\log Y$, $\sqrt{Y}$, etc.) to stabilize variance
- Use weighted least squares, or use heteroskedasticity-robust standard errors

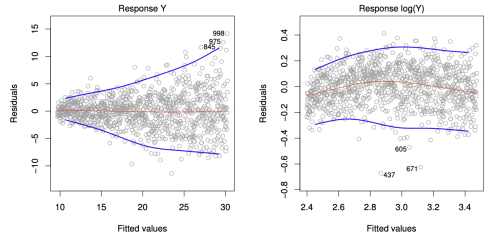


Figure: Residual plots with heteroskedastic error [JWHT21, Figure 3.11].

Pitfall 4: Outliers & high-leverage points

Issue: Influential points materially change the fitted model

- Outliers have unusual y-values
- High-leverage points have unusual x-values

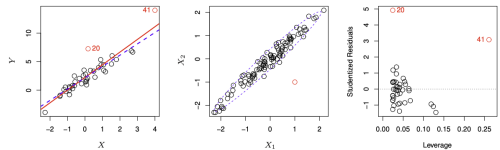


Figure: An illustration of outliers and high-leverage points [JWHT21, Figure 3.13].

Why problematic:

- 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

How to diagnose:

- 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

Pitfall 5: Collinearity

Issue: Two or more predictors are highly correlated

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

Why problematic:

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

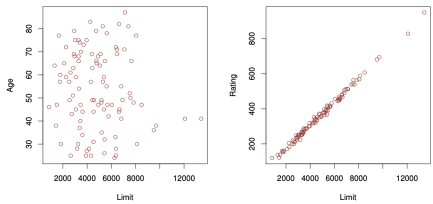


Figure: An illustration of high collinearity [JWHT21, Figure 3.14].

How to diagnose: Check the predictor correlation matrix or compute variance inflation factors (VIFs)

Possible 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

Wrap-up

- Inference in linear regression uses the sampling distributions of the estimated coefficients
- A confidence interval targets the mean response, while a prediction interval targets a new observation
- Qualitative predictors are incorporated through dummy variables; with an intercept and K levels, use $K - 1$ dummies
- Dummy coefficients are interpreted relative to the baseline group, and interactions allow group-specific slopes
- Linear regression can fail when assumptions break down, so residual plots and diagnostic tools are essential

Next lecture: Classification

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.