Lecture 5

Analysis of Covariance, Polynomial Regression and Non-linear Regression

Reading: Faraway 2014 Chapters 9.4, 14.2-14.4; ISLR 2021 Chapter 3.3

STAT 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

MLR with Categorical Predictors

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

$x_1, x_2, \cdots, x_{p-1}$ are the numerical predictors.

Problem: Categorical variables cannot enter model directly

$\Rightarrow$ Encode categorical variables using **dummy variables**

Example: We can encode `Gender` into 1 (Female) and 0 (Male)

Running Example: Salaries for Professors Data Set

Faculty salary data from a U.S. college are used to study how salary is associated with rank, experience, discipline, and gender.

```
> head(Salaries)
      rank discipline yrs.since.phd yrs.service  sex salary
1     Prof          B            19          18 Male 139750
2     Prof          B            20          16 Male 173200
3 AsstProf          B             4           3 Male  79750
4     Prof          B            45          39 Male 115000
5     Prof          B            40          41 Male 141500
6 AssocProf         B             6           6 Male  97000
```

Rank, discipline, and gender are categorical predictors in MLR $\Rightarrow$ require dummy variable encoding. How?

Dummy Variables

Binary variable:

$$x = \begin{cases} 1 & \text{Group A} \\ 0 & \text{Group B} \end{cases}$$

Multiple categories (k levels):

- Use $k - 1$ dummy variables
- One level serves as the reference group

Next: A concrete example using faculty salary data

Examples of Dummy Variable Encoding

For binary categorical variables:

$$x_{\texttt{sex}} = \begin{cases} 1 & \text{if } \texttt{sex} = \text{male}, \\ 0 & \text{if } \texttt{sex} = \text{female}. \end{cases}$$

$$x_{\texttt{discip}} = \begin{cases} 0 & \text{if } \texttt{discip} = \text{A}, \\ 1 & \text{if } \texttt{discip} = \text{B}. \end{cases}$$

For categorical variable with more than two categories ($k = 3$ here):

$$x_{\texttt{rank1}} = \begin{cases} 0 & \text{if } \texttt{rank} = \text{Assistant Prof}, \\ 1 & \text{if } \texttt{rank} = \text{Associated Prof}. \end{cases}$$

$$x_{\texttt{rank2}} = \begin{cases} 0 & \text{if } \texttt{rank} = \text{Associated Prof}, \\ 1 & \text{if } \texttt{rank} = \text{Full Prof}. \end{cases}$$

Design Matrix: Bringing Numerical and Categorical Predictors Together

```
> head(X)
  (Intercept) rankAssocProf rankProf disciplineB yrs.since.phd
1           1             0        1           1            19
2           1             0        1           1            20
3           1             0        0           1             4
4           1             0        1           1            45
5           1             0        1           1            40
6           1             1        0           1             6
  yrs.service sexMale
1          18       1
2          16       1
3           3       1
4          39       1
5          41       1
6           6       1
```

Punchline: After encoding, fit the standard linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Interpreting Dummy Variables:

`lm(salary ~ rank + sex + discipline + yrs.since.phd)`

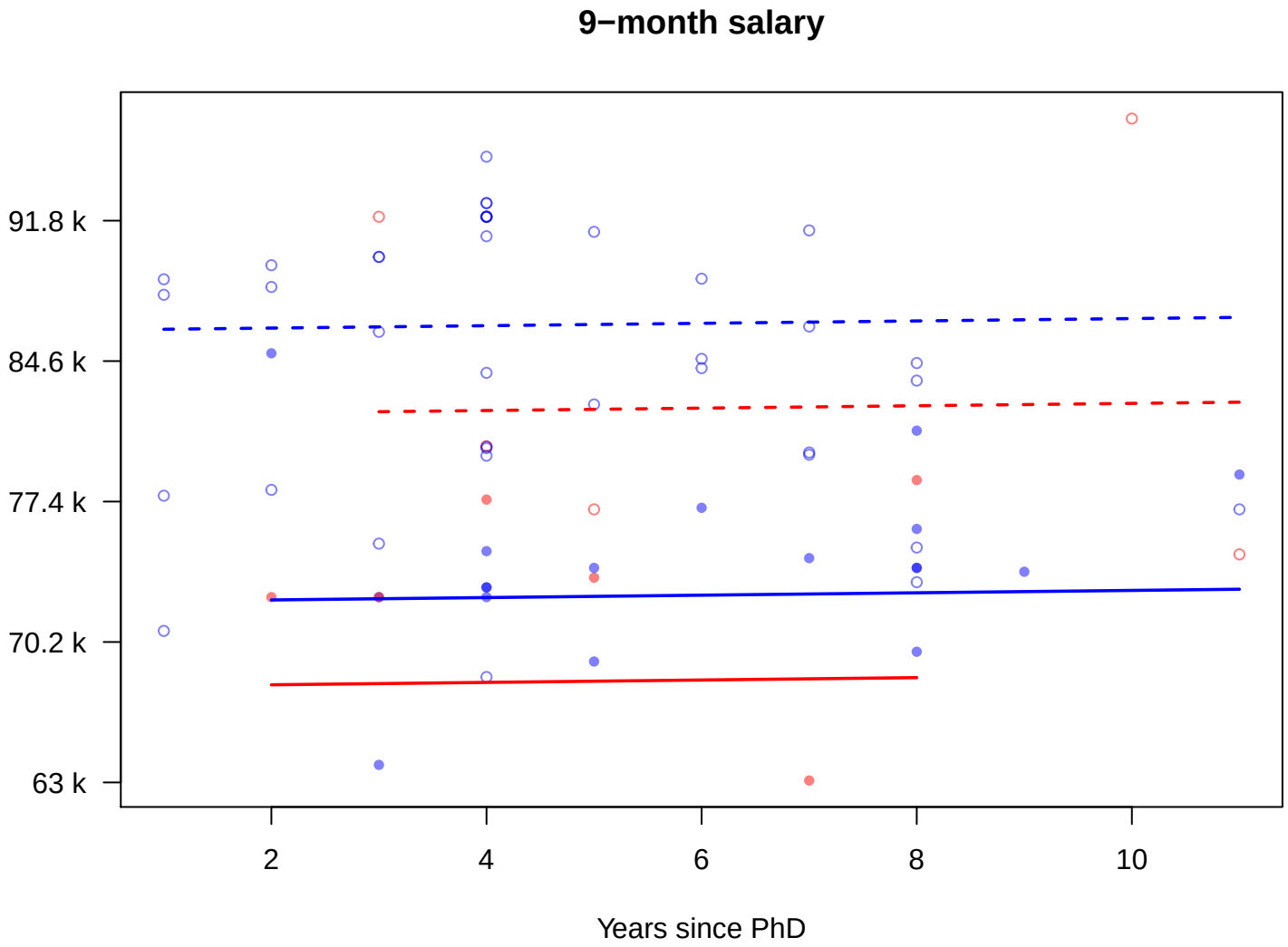
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  67884.32    4536.89  14.963  < 2e-16 ***
disciplineB  13937.47    2346.53   5.940 6.32e-09 ***
rankAssocProf 13104.15    4167.31   3.145  0.00179 **
rankProf     46032.55    4240.12  10.856  < 2e-16 ***
sexMale       4349.37    3875.39   1.122  0.26242
yrs.since.phd   61.01     127.01   0.480  0.63124
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22660 on 391 degrees of freedom
Multiple R-squared:  0.4472,    Adjusted R-squared:  0.4401
F-statistic: 63.27 on 5 and 391 DF,  p-value: < 2.2e-16
```

Key idea: Coefficients = difference from reference group

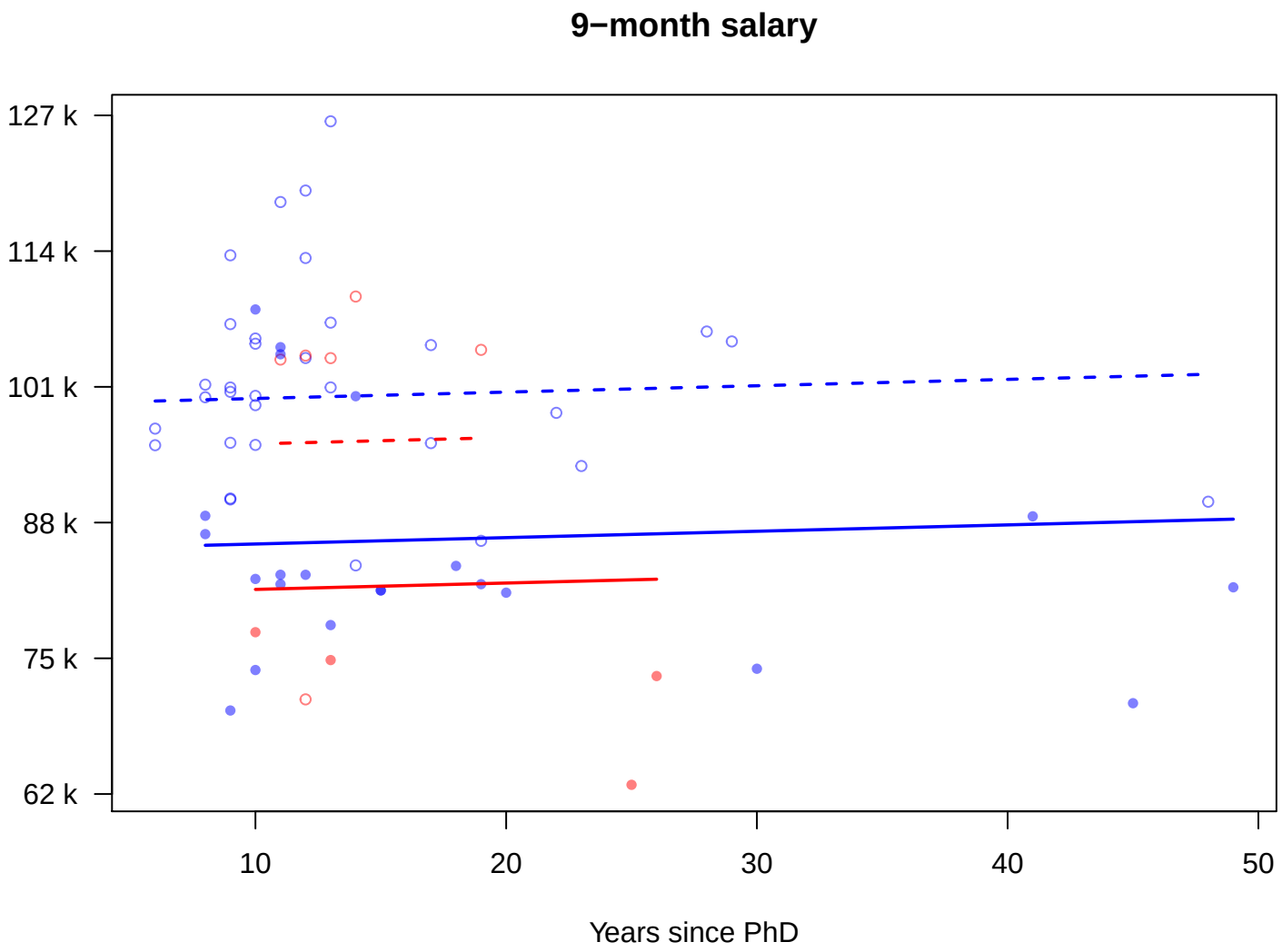
- Intercept: baseline (reference group)
- Dummy coefficients: shift from baseline
- Slope: same interpretation (common across groups)

Model Fit for Assistant Professors

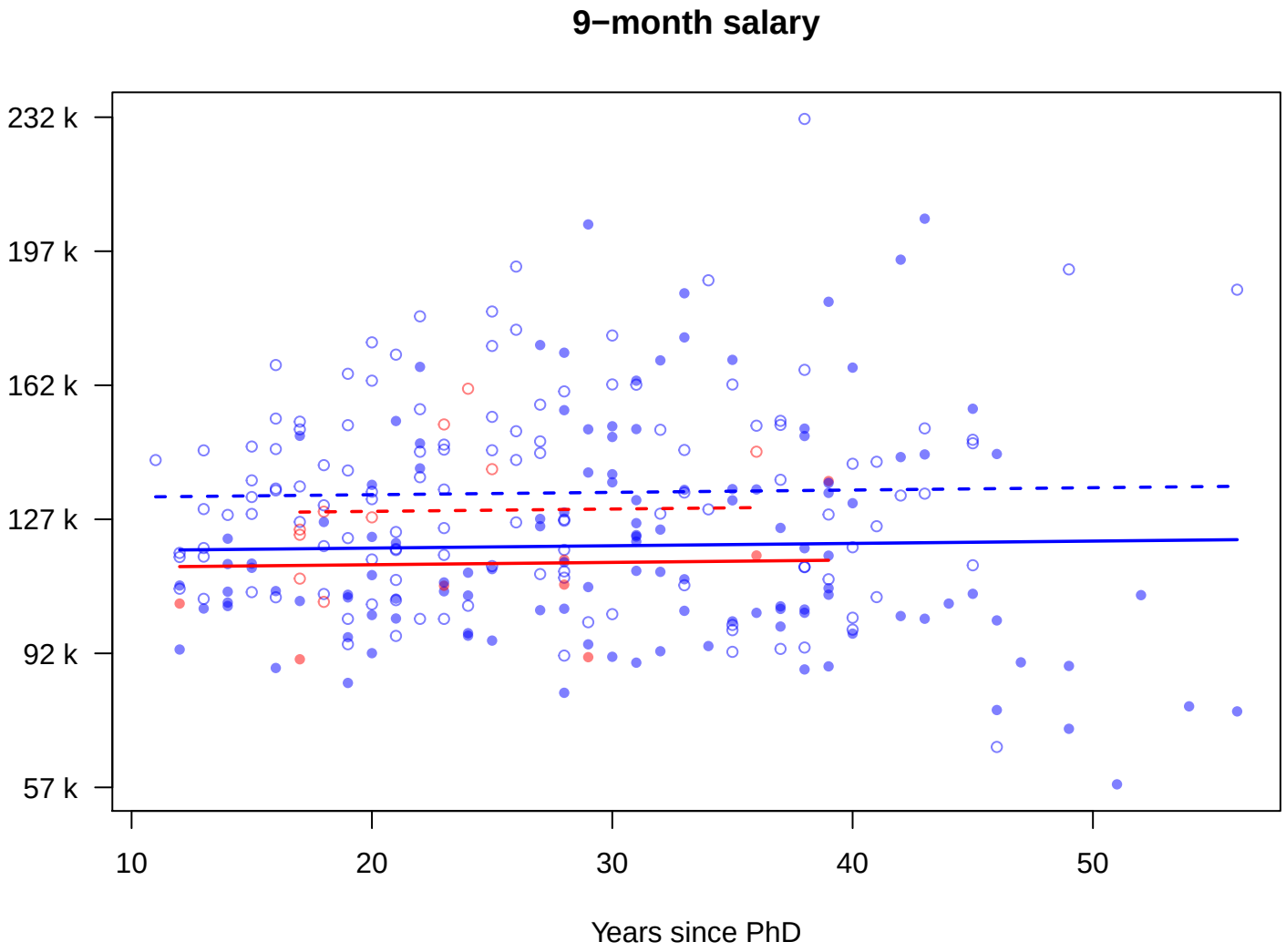
Color	Line Type
Red: Female	—-: Applied (discipline B)
Blue: Male	- - -: Theoretical (discipline A)



Model Fit for Associate Professors



Model Fit for Full Professors



Incorporating Interactions: Different Slopes Across Groups

```
lm(salary ~ sex * yrs.since.phd)
```

Interaction: Effect of x (numerical predictors) depends on group

Interpretation: Different slopes across groups

`lm(salary ~ disp * yrs.since.phd)`

Punchline: Years since PhD has a stronger effect on salary in applied sciences (- - - , steeper slope) than in theoretical sciences (—)

Beyond ANCOVA: Numerical Interactions

Idea: Interactions are not limited to categorical variables

Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Interpretation: Effect of x_1 depends on x_2

Why use it? Capture non-additive relationships between predictors

Polynomial Regression

Idea: Add powers of x to capture curvature

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

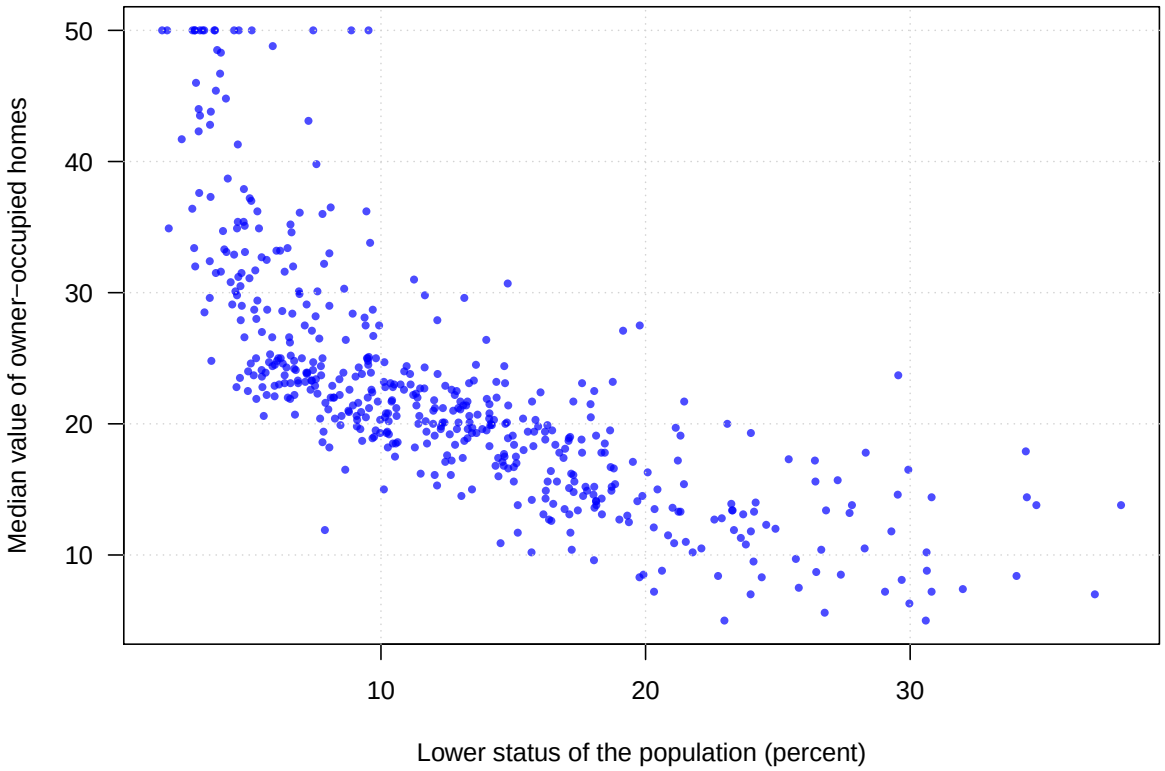
Design matrix:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{pmatrix}$$

Key point: Still linear in parameters $\Rightarrow$ use MLR

Housing Values in Suburbs of Boston Data Set

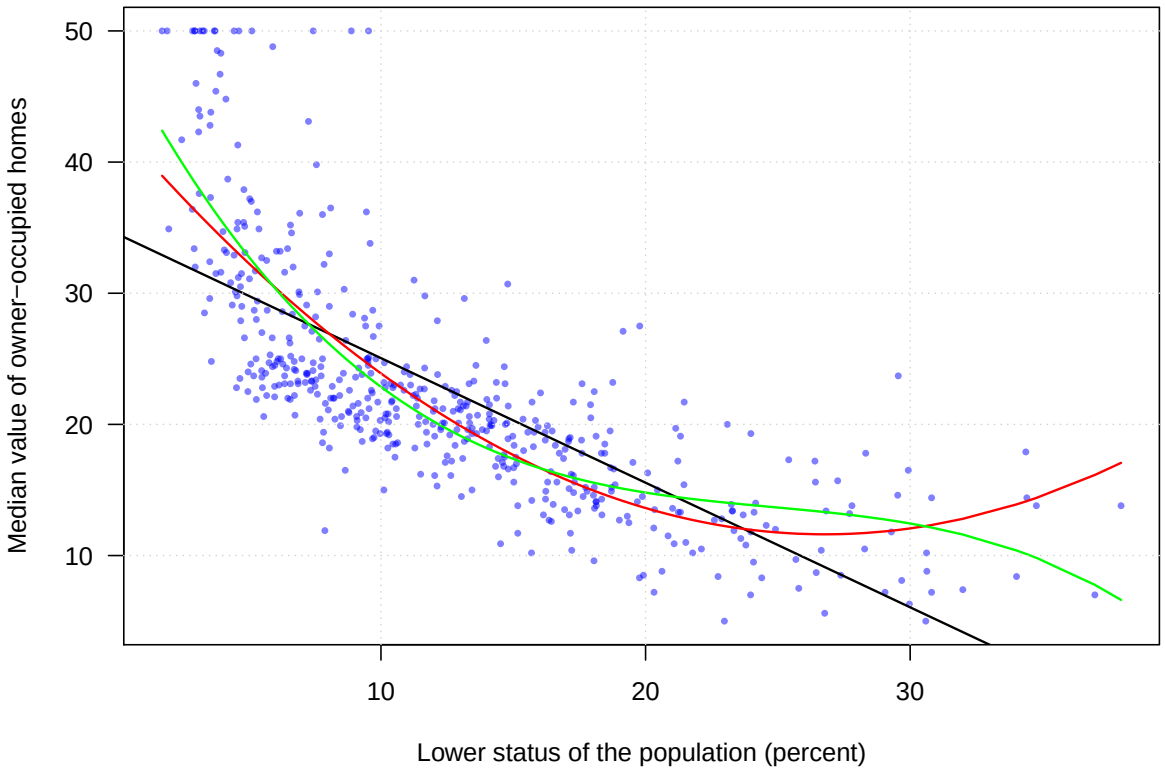
- y: the median value of owner-occupied homes (in thousands of dollars)
- x: percent of lower status of the population



Evidence of curvature ⇒ consider polynomial regression

Polynomial Regression Fits

1st, 2nd, and 3rd polynomial regression fits



Takeaway:

- Linear → underfit
- Higher degree → more flexible but may overfit
- Can use AIC/BIC to select the degree

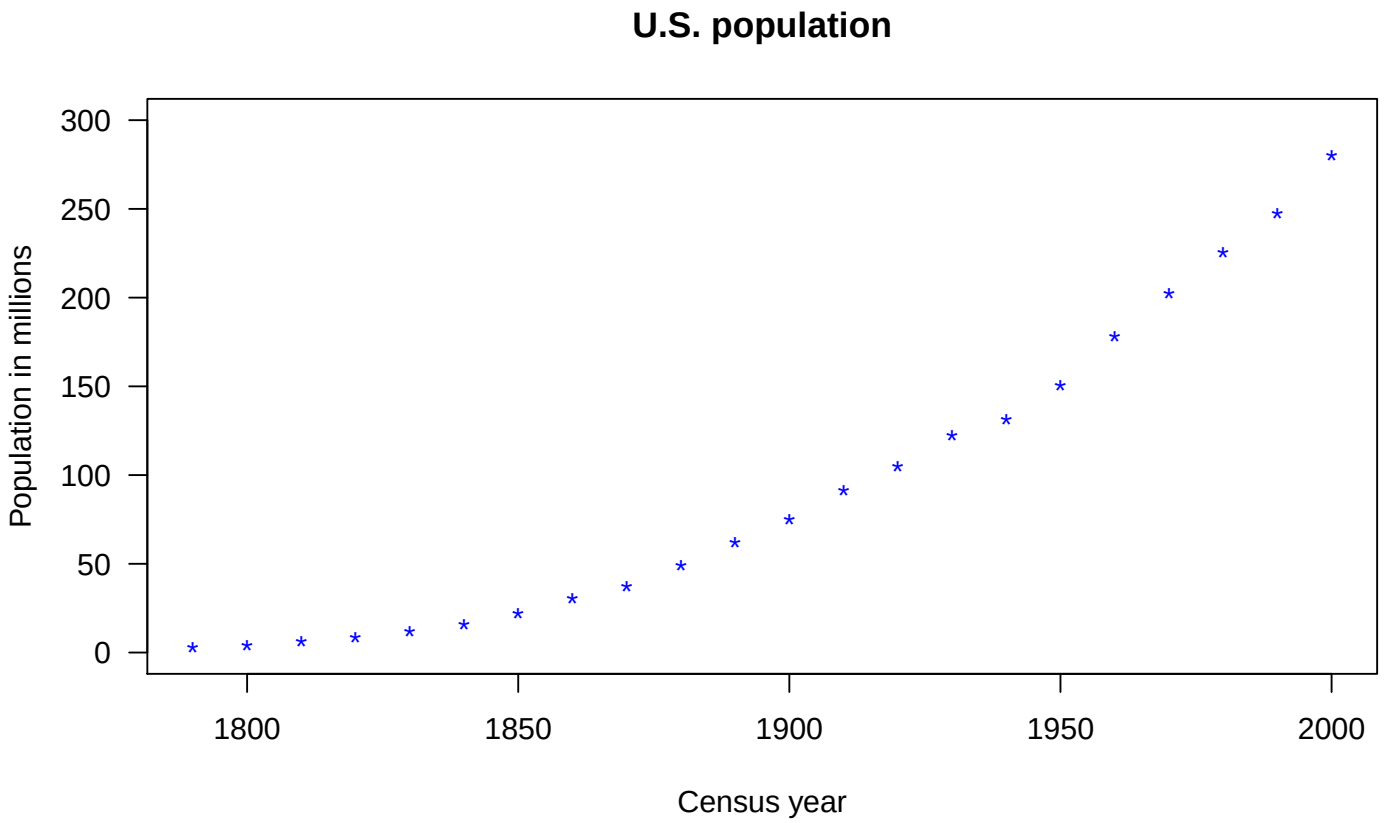
Moving Beyond Linear Regression

- So far, we have focused on linear regression
- Polynomial regression can introduce curvature and improve fit
- However, global polynomials may still lack flexibility:
  - High-degree polynomials may overfit
  - Poor local fit in some regions
- In practice, relationships may follow specific nonlinear forms (e.g., exponential growth, saturation)

> **Next:** Nonlinear models with specified functional forms (guided by domain knowledge)

Population of the United States

Let's look at the `USPop` data set, a bulit-in data set in `R`. This is a decennial time-series from 1790 to 2000.



Logistic Growth Model

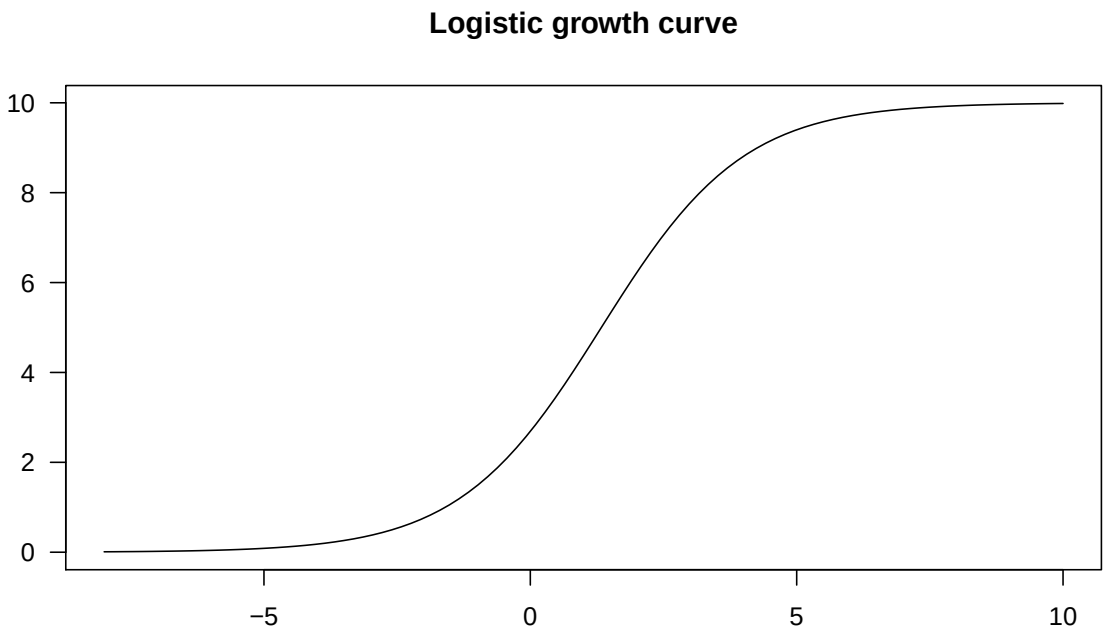
Use: S-shaped growth (e.g., population)

Model:

$$y = \frac{\phi_1}{1 + \exp\left[-(x - \phi_2)/\phi_3\right]} + \varepsilon,$$

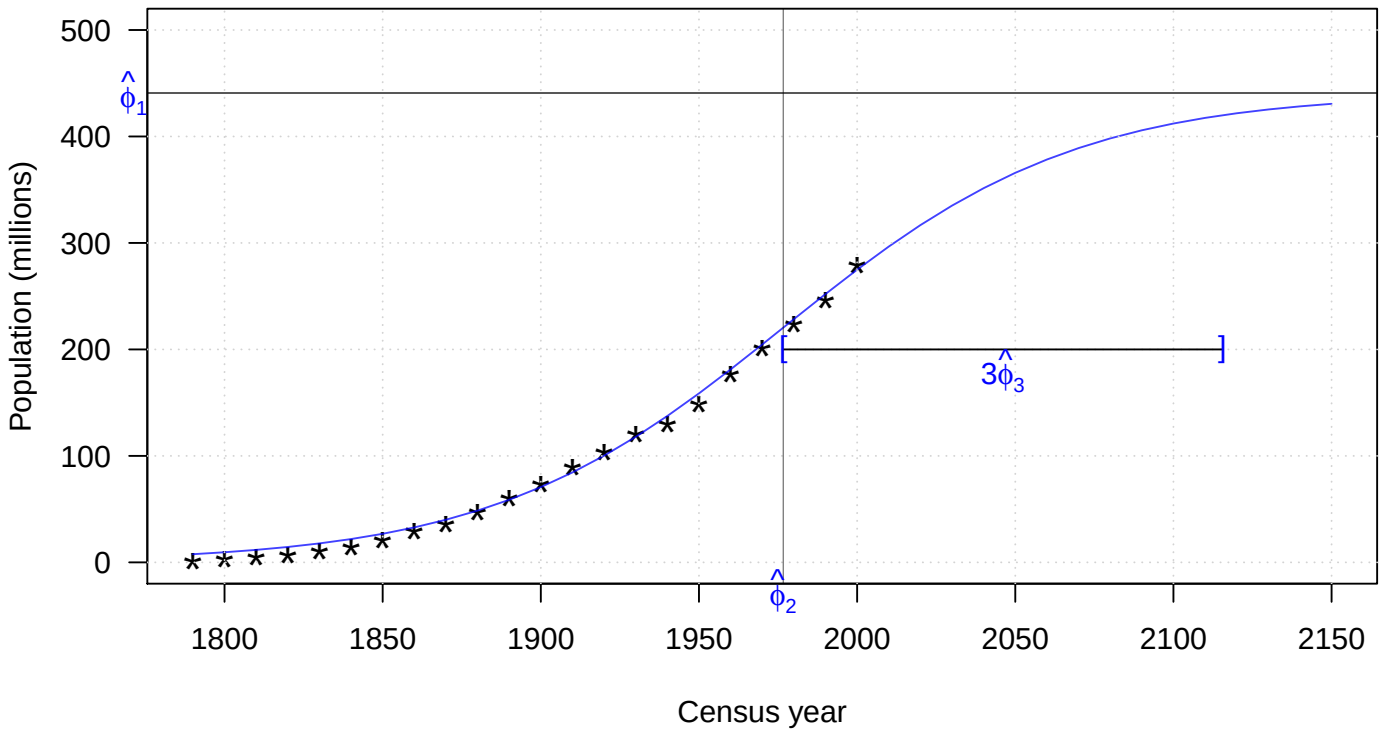
Parameters:

- ϕ_1: maximum level
- ϕ_2: midpoint
- ϕ_3: "range" (or the inverse growth rate) of the curve



Logistic Growth Curve Fit to the U.S. Population

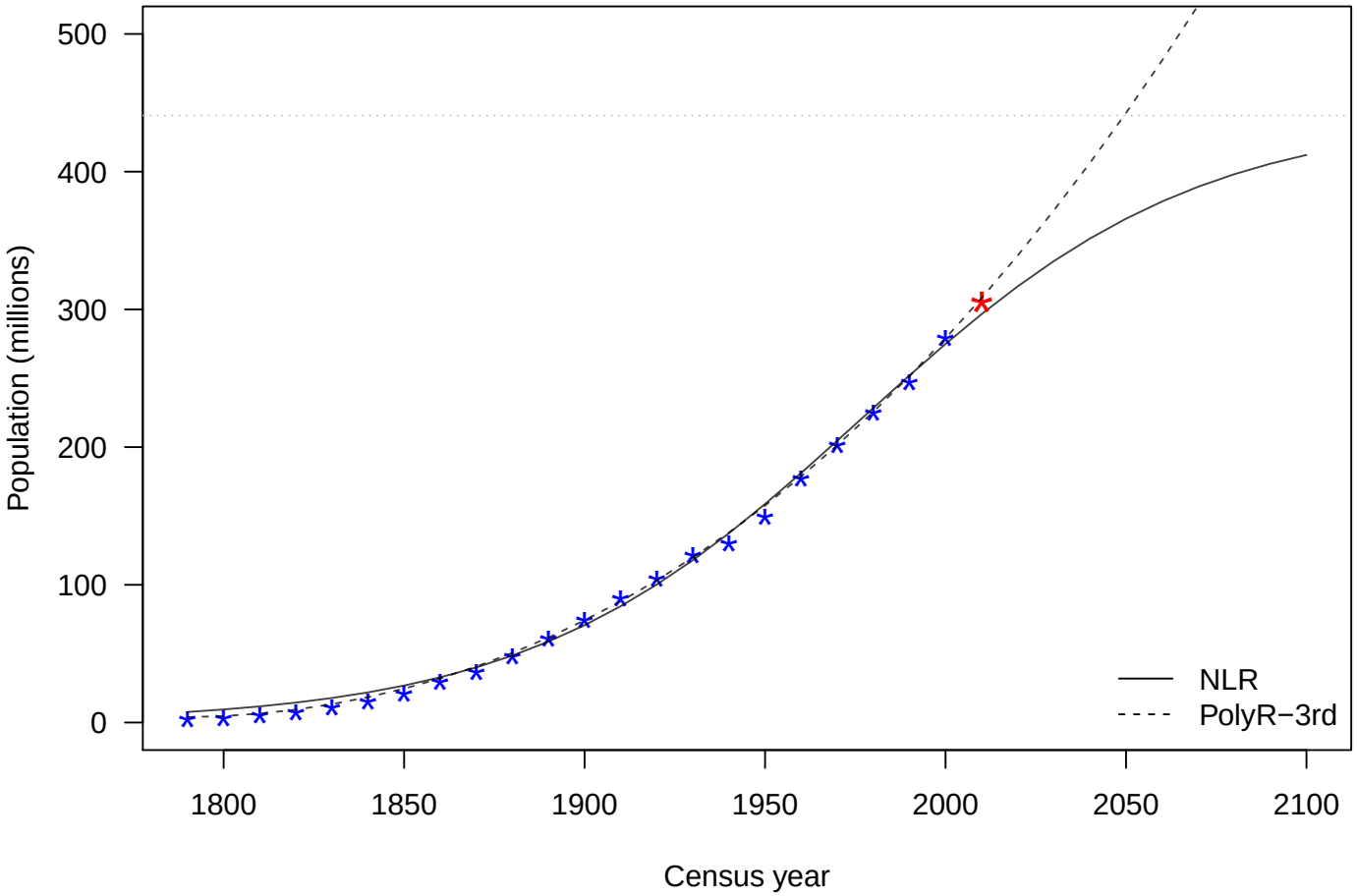
$\hat{\phi}_1 = 440.83$, $\hat{\phi}_2 = 1976.63$, $\hat{\phi}_3 = 46.29$



Notes:

- Requires numerical optimization
- More interpretable than polynomial fits (if form is known)
- Better extrapolation

Logistic vs. Polynomial Fit: U.S. Population



Takeaway:

- Logistic: theory-driven, realistic long-term behavior
- Polynomial: flexible fit, but poor extrapolation

Summary

These slides cover:

- Analysis of Covariance to handle the situations where there both some of the predictors are categorical variables
- Polynomial Regression, where polynomial terms are added to increase the model flexibility
- Nonlinear Regression

`R` functions to know:

- Use `*` to create interaction terms in `lm`
- Use `I(x)` or `poly(x, df)` to create polynomial terms
- Use `nls` to perform nonlinear least squares for nonlinear regression