

Lecture 5

Multiple Linear Regression: Analysis of covariance; Non-linear Regression

Reading: Faraway (2014) Chapters 14, 9.4 or Faraway (2002)
Chapters 15, 8.2.2

DSA 8020 Statistical Methods II
February 1-5, 2021

Whitney Huang
Clemson University

Agenda

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression

1 **Analysis of Covariance**

2 **Polynomial Regression**

3 **Nonlinear Regression**

Regression with Both Quantitative and Qualitative Predictors

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression

Multiple Linear Regression

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

x_1, x_2, \dots, x_{p-1} are the predictors.

Question: What if some of the predictors are qualitative (categorical) variables?

⇒ We will need to create **dummy (indicator) variables** for those categorical variables

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

CLEMSON
UNIVERSITY

Analysis of Covariance

Polynomial Regression

Nonlinear Regression

Salaries for Professors Data Set

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

```
> 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

```
> summary(Salaries)
```

rank	discipline	yrs.since.phd	yrs.service
AsstProf : 67	A:181	Min. : 1.00	Min. : 0.00
AssocProf: 64	B:216	1st Qu.:12.00	1st Qu.: 7.00
Prof :266		Median :21.00	Median :16.00
		Mean :22.31	Mean :17.61
		3rd Qu.:32.00	3rd Qu.:27.00
		Max. :56.00	Max. :60.00

sex	salary
Female: 39	Min. : 57800
Male :358	1st Qu.: 91000
	Median :107300
	Mean :113706
	3rd Qu.:134185
	Max. :231545

We have three categorical variables, namely, rank, discipline, and sex.

For binary categorical variables:

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

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

For categorical variable with more than two categories:

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

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

Design Matrix

```
> 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
```

With the design matrix X , we can now use method of least squares to fit the model $Y = X\beta + \varepsilon$

Model Fit:

```
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

Question: Interpretation of the slopes of these dummy variables (e.g. $\hat{\beta}_{\text{rankAssocProf}}$)? Interpretation of the intercept?

Model Fit for Assistant Professors

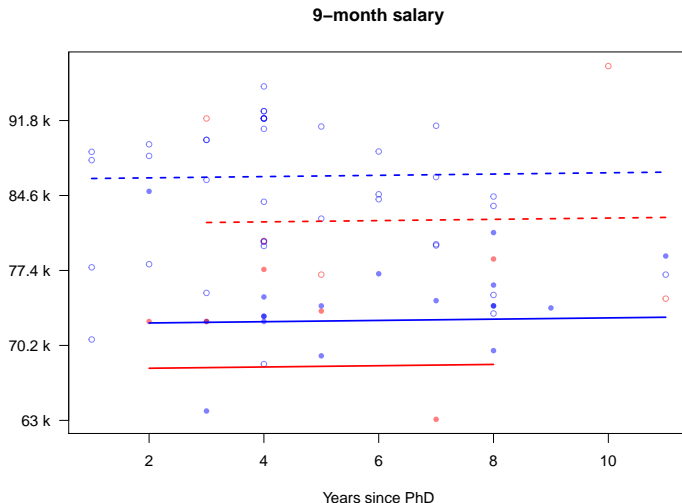
Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression



Model Fit for Associate Professors

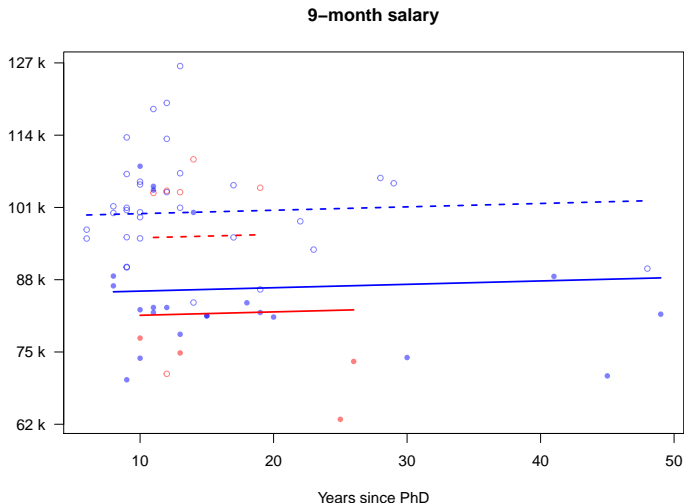
Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression



Model Fit for Full Professors

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression



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

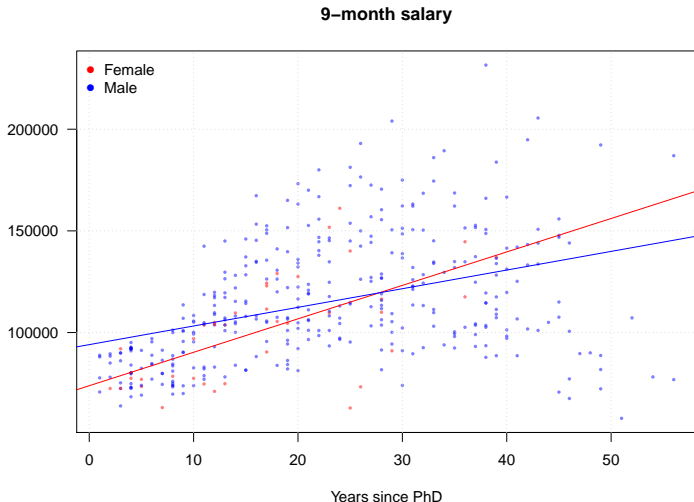
Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

Nonlinear Regression



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

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression

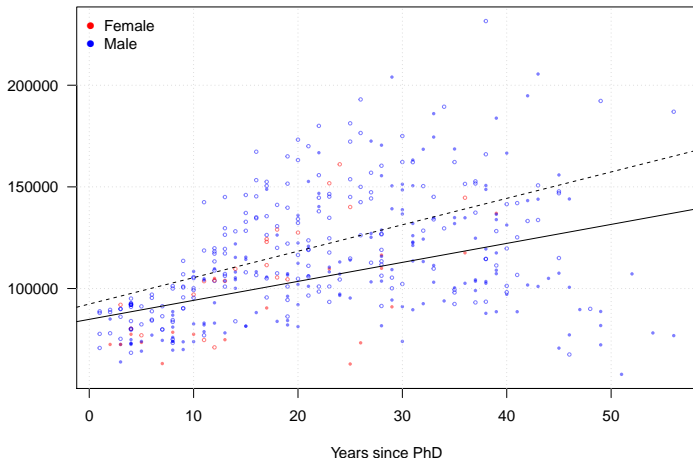


Analysis of Covariance

Polynomial Regression

Nonlinear Regression

9-month salary



Suppose we would like to model the relationship between response Y and a predictor X as a p_{th} degree polynomial in x :

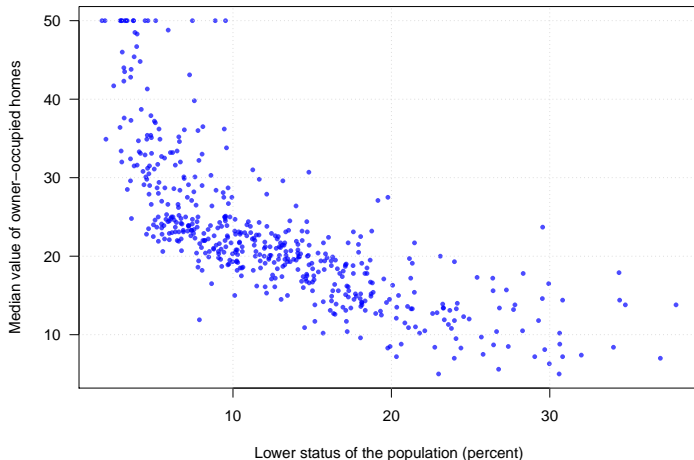
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

We can treat polynomial regression as a special case of multiple linear regression. In specific, the design matrix takes the following form:

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

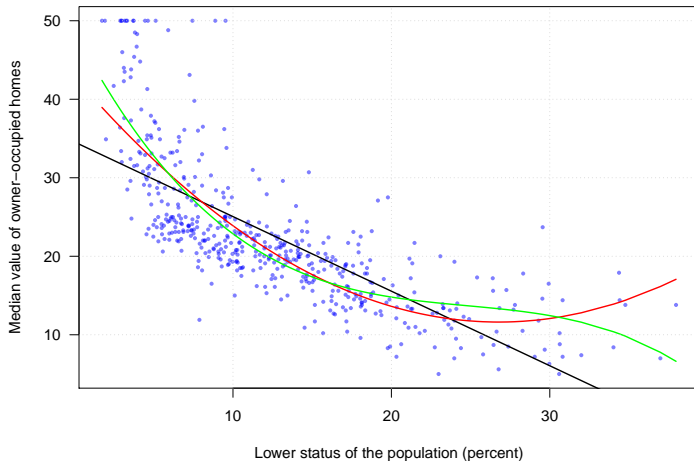
Housing Values in Suburbs of Boston Data Set

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



Polynomial Regression Fits

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



Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression

CLEMSON
UNIVERSITY

Analysis of Covariance

Polynomial Regression

Nonlinear Regression

Moving Away From Linear Regression

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression



Analysis of Covariance

Polynomial Regression

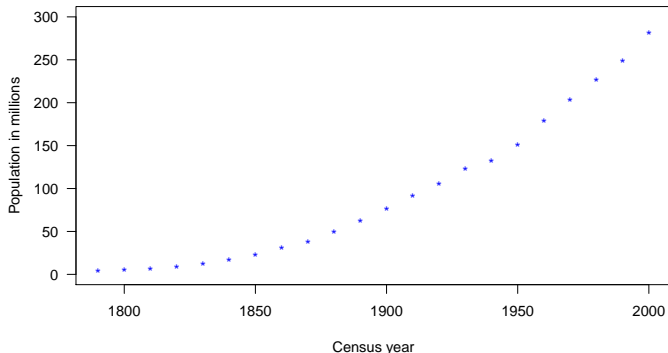
Nonlinear Regression

- We have mainly focused on **linear regression** so far
- The class of **polynomial regression** can be thought as a starting point for relaxing the linear assumption
- In the next few slides we are going to discuss **non-linear** regression modeling

Population of the United States

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

U.S. population

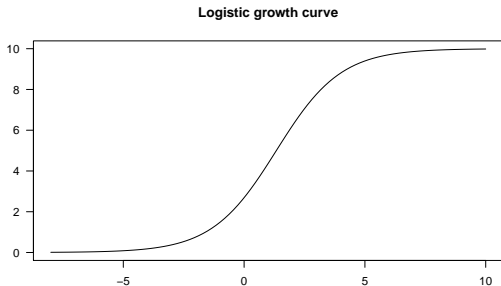


Logistic Growth Curve

A simple model for population growth is the **logistic growth model**,

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

where ϕ_1 is the curve's maximum value; ϕ_2 is the curve's midpoint in x ; and ϕ_3 is the "range" (or the inverse growth rate) of the curve.



Fitting logistic growth curve to the U.S. population

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression

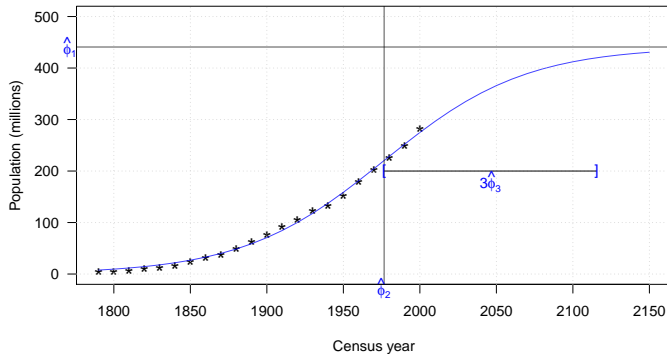


Analysis of Covariance

Polynomial Regression

Nonlinear Regression

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



Comparing the Logistic Growth Curve Fit and Cubic Polynomial Fit

Multiple Linear
Regression: Analysis
of covariance;
Non-linear
Regression

CLEMSON
UNIVERSITY

Analysis of Covariance

Polynomial Regression

Nonlinear Regression

