Lecture 14

Regression with Quantitative and Qualitative Predictors & Polynomial Regression

STAT 8020 Statistical Methods II
September 20, 2019

Whitney Huang
Clemson University

Notes

Agenda

1. Regression with Both Quantitative and Qualitative Predictors
2. Polynomial Regression

Notes

Regression with Both Quantitative and Qualitative Predictors

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

$X_1, X_2, \cdots, X_{p-1}$ are the predictors.

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

$\Rightarrow$ We will need to create **dummy (indicator) variables** for those categorical variables

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

Notes

Salaries for Professors Data Set

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
```


Notes

Predictors

```
> 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 are three categorical variables, namely, `rank`, `discipline`, and `sex`.

Notes

Dummy Variable

For binary categorical variables:

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

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

For categorical variable with more than two categories:

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

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

Notes

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 $Y = X\beta + \varepsilon$

Notes

Model Fit

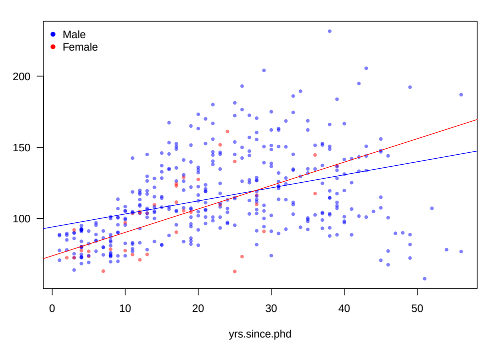
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     70738.7     3403.0  20.787  < 2e-16 ***
rankAssocProf   12907.6     4145.3   3.114  0.00198 ** 
rankProf        45066.0     4237.5  10.635  < 2e-16 ***
disciplineB     14417.6     2342.9   6.154 1.88e-09 ***
yrs.since.phd     535.1      241.0   2.220  0.02698 *  
yrs.service      -489.5      211.9  -2.310  0.02143 *  
sexFemale       -4783.5     3858.7  -1.240  0.21584    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22540 on 390 degrees of freedom
Multiple R-squared:  0.4547,	Adjusted R-squared:  0.4463 
F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

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

Notes

$\texttt{lm(salary} \sim \texttt{sex} * \texttt{yrs.since.phd)}$



Notes

Polynomial Regression

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

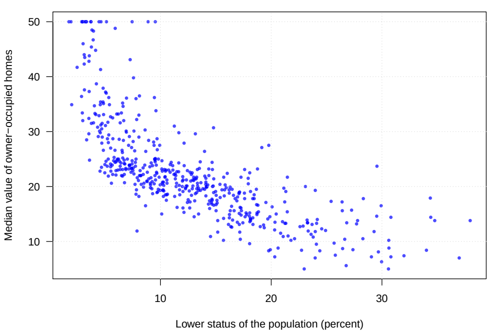
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \varepsilon$$

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

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^p \\ 1 & X_2 & X_2^2 & \cdots & X_2^p \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^p \end{pmatrix}$$

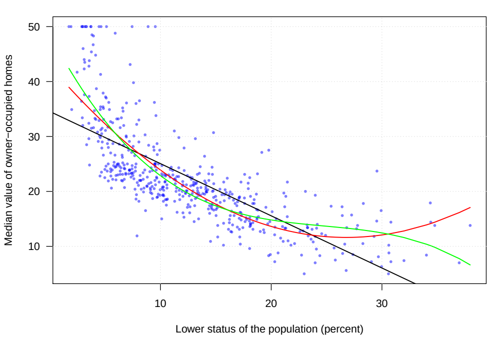
Notes

Housing Values in Suburbs of Boston Data Set



Notes

Polynomial Regression Fits



Notes

Potential Topics for Next Lecture

- Nonlinear Regression
- Non-Parametric Regression
- Ridge Regression
- Regression Tree
- Least Absolute Shrinkage and Selection Operator (LASSO)

Notes

Notes

Notes