Lecture 11

Finish Section 3.5 from last Monday (on board)
Section 3.6, Continued: Nested F tests
Section 3.4:
Polynomial Regression
Especially quadratic
Second-order models (including interaction)

Example: 1982 State SAT Scores
(First year state by state data available)

Unit = A state in the United States
Response Variable:
\( Y = \) Average combined SAT Score
Potential Predictors:
\( X_1 = \) Takers = % taking the exam out of all eligible students in that state
\( X_2 = \) Expend = amount spent by the state for public secondary schools, per student ($100's)

Is \( Y \) related to one or both of these \( X \) variables?

Example: State SAT with \( X_1 \) only

\[ Y = \text{Combined SAT} \]
\[ X = \% \text{Taking SAT} \]

Things to notice:
- Two clusters in the \( X \) range. Why?
- Possible curved relationship
- As %Takers goes up, average SAT goes down.

Example: State SAT with \( X_1 \) only

Would a “curved” line work better?

\[ Y = \text{Combined SAT} \]
\[ X = \% \text{Taking SAT} \]

Residuals vs fitted values

Example: State SAT with \( X_2 \) only

\[ Y = \text{Combined SAT} \]
\[ X = \text{Expenditure} \]

Not clear what pattern is; Alaska is very influential.

Polynomial Regression
For a single predictor \( X \):
\[
Y = \beta_0 + \beta_1X + \beta_2X^2 + \cdots + \beta_pX^p + \epsilon
\]
\[
Y = \beta_0 + \beta_1X + \epsilon \quad \text{(Linear)}
\]
\[
Y = \beta_0 + \beta_1X + \beta_2X^2 + \epsilon \quad \text{(Quadratic; curve)}
\]
\[
Y = \beta_0 + \beta_1X + \beta_2X^2 + \beta_3X^3 + \epsilon \quad \text{(Cubic)}
\]
Polynomial Regression in R

Method #1: Create new columns with powers of the predictor.
To avoid creating a new column…

Method #2: Use `I( )` in the `lm( )`
```r
quadmod=lm(SAT~Takers+I(Takers^2))
```

Method #3: Use `poly`
```r
quadmod=lm(SAT~poly(Takers,degree=2,raw=TRUE))
```

Quadratic Model
```r
> Quad<-lm(sat~takers+I(takers^2), data=StateSAT)
> summary(Quad)
```

```
Call: lm(formula = sat ~ takers + I(takers^2), data = StateSAT)
Coefficients:

               Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1053.13112    9.27372 113.561  < 2e-16 ***
takers       -7.16159    0.89220  -8.027 2.32e-10 ***
I(takers^2)   0.07102    0.01405   5.055 6.99e-06 ***

Residual standard error: 29.93 on 47 degrees of freedom
Multiple R-squared:  0.8289, Adjusted R-squared:  0.8216
F-statistic: 113.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

Residual Plot Looks Good
(Two clusters still obvious)

How to Choose the Polynomial Degree?
- Use the minimum degree needed to capture the structure of the data.
- Check the t-test for the highest power.
- (Generally) keep lower powers—even if not “significant.”

Interaction
Recall:

\[ Active = \beta_0 + \beta_1 \text{Rest} + \beta_2 \text{Gender} + \beta_3 \text{Rest} \times \text{Gender} + \epsilon \]

Product allows for different Active/Rest slopes for different genders

In General: Interaction is present if the relationship between two variables (e.g. Y and X) changes depending on a third variable (e.g. X).

Modeling tip: Include a product term to account for interaction.

Complete Second-order Models
Definition: A complete second-order model for two predictors would be:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_1 X_1 X_2 + \epsilon \]

First order Quadratic Interaction
Second-order Model for State SAT

Example: Try a full second-order model for $Y = \text{SAT}$ using $X_1 = \text{Takers}$ and $X_2 = \text{Expend}$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \epsilon$$

Second-order Model for State SAT

$$\text{secondorder} = \text{lm}(\text{SAT} \sim \text{Takers} + I(\text{Takers}^2) + \text{Expend} + I(\text{Expend}^2) + \text{Takers}:\text{Expend}, \text{data} = \text{StateSAT})$$

Do we really need the quadratic terms? Nested F-test

Simultaneously test all three terms involving “Expend” in the second order model with “Takers” to predict SAT scores.

Strong evidence that the quadratic terms are significant as a pair as well as individually.

Three "new" predictors reduce the SSE by 17432, a sig. amount.
Second-order Model for State SAT

summary(secondorder)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 893.6628 | 36.1409    | 24.727  | < 2e-16  *** |
| takers         | -7.0556  | 0.8374     | -8.426  | 9.96e-11 *** |
| I(takers^2)    | 0.0773   | 0.0133     | 5.816   | 6.28e-07 *** |
| expend         | 10.3333  | 2.4960     | 4.140   | 0.000155 *** |
| I(expend^2)    | -0.1178  | 0.0443     | -2.660  | 0.010851 *  |
| takers:expend  | -0.0334  | 0.0372     | -0.900  | 0.373087  |

Residual standard error: 23.68 on 44 degrees of freedom
Multiple R-squared:  0.8997, Adjusted R-squared:  0.8883
F-statistic: 78.96 on 5 and 44 DF,  p-value: < 2.2e-16

Do we really need the interaction? T-test for takers:expend

SUMMARY

- Full second-order model is better than the model with no quadratic terms
- Full second-order model is better than the quadratic model with “Takers” only
- Model with no interaction seems acceptable
- Comparing Adjusted R-squared:
  - Full model: 88.83%; No interaction: 88.88%
  - No quadratic terms: 76.38%
  - Takers only, quadratic: 82.16%
  - Expend only, quadratic: -3.59%, partly because of the extreme outlier for Alaska!