PM 592
Regression Analysis for Public Health Data Science

Week 7

Complex Coding Schemes

1

Complex Coding Schemes

Polynomial Terms

Fractional Polynomials

Splines

Dose-Response Coding

Overfitting

Adjusted R-Squared

2

Lecture Objectives

- Implement and interpret a polynomial independent variable.
- Implement and interpret a spline.
- Implement and interpret dose-response coding.
- Explain and diagnose over-fitting.
- Explain the meaning and utility of adjusted R-squared.

3

1. Review

4

- ✓ What is confounding?
- ✓ How to detect confounding
- ✓ What is effect modification?
- ✓ How to detect effect modification

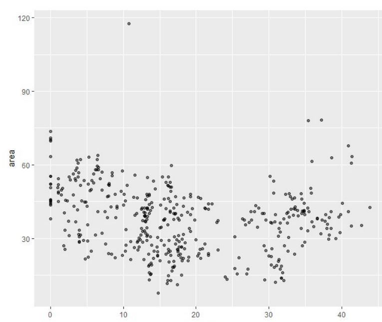
4

2. Polynomial Terms

5

Example

Let's examine the real estate data—namely, how selling price is related to age of the house.



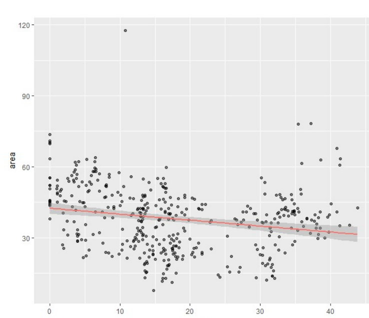
5

2. Polynomial Terms

6

A linear relationship clearly does not seem sufficient to describe this relationship.

Furthermore, there does not appear to be a good transformation for Y to conform to linearity.



6

2. Polynomial Terms 7

A **polynomial regression** model includes higher-order polynomial terms for X, such that:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_h X^h + e$$

Where *h* is the **degree** of the polynomial.

7

2. Polynomial Terms 8

A **polynomial regression** model includes higher-order polynomial terms for X, such that:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_h X^h + e$$

Where *h* is the **degree** of the polynomial.

This is still considered a form of “linear” regression as the model is linear in the regression coefficients β.

8

2. Polynomial Terms 9

Let’s examine our regression relationship by considering the following models:

1) Linear

$$\hat{Y} = \beta_0 + \beta_1 X$$

2) Quadratic

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

3) Cubic

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

9

2. Polynomial Terms 10

Let's examine our regression relationship by examining the following models:

1) Linear

$$\hat{Y} = \beta_0 + \beta_1 X$$

2) Quadratic

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$

3) Cubic

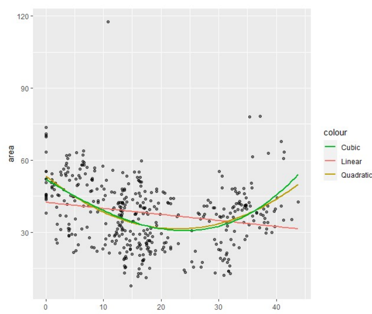
$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

10

2. Polynomial Terms 11

It looks like the quadratic model provides us better fit to the data compared to the linear model.

However, the cubic model doesn't seem to add much beyond the quadratic.



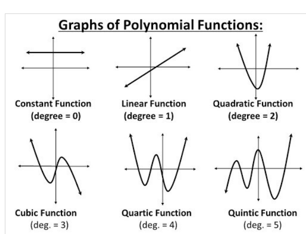
11

2. Polynomial Terms 12

What Degree Polynomial to Choose?

There are a couple of ways to determine the degree of the polynomial terms to include in your model.

1) You can "eyeball" it and examine what you believe to be appropriate.



12

2. Polynomial Terms 13

What Degree Polynomial to Choose?

2) You can continue to add higher degree terms until they are no longer significant in the model.

What method do we use to check for the significance of adding an additional variable into the regression model?

1. We can perform the **extra sums of squares** F-test
2. We can examine the **Type I sums of squares** in an ANOVA table

13

2. Polynomial Terms 14

Recall:

1) The Type I Sums of Squares tells us the additional sums of squares that are explained by each *additional* variable that is added to the model.
2) The Type I Sums of Squares is reported by the anova() function.
3) The car::Anova() function can provide Type III sums of squares.

14

2. Polynomial Terms 15

What do the Type I (sequential) Sums of Squares tell us here?

```
> lm(area ~ age + I(age^2) + I(age^3), data = re) %>% anova()
Analysis of Variance Table

Response: area
           Df Sum Sq Mean Sq F value    Pr(>F)    
age         1   3390  3390.2 22.8513 2.444e-06 ***
I(age^2)    1  12020 12019.9 81.0194 < 2.2e-16 ***
I(age^3)    1    224   224.4  1.5126    0.2194    
Residuals 410  60827   148.4                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A quadratic term significantly improves model fit compared to the linear term (p<.001).
$\hat{Y} = \beta_0 + \beta_1 X + \boldsymbol{\beta_2 X^2}$

A cubic term does <u>not</u> improve the model fit compared to the model with the linear and quadratic terms (p=.22).
$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \boldsymbol{\beta_3 X^3}$

15

2. Polynomial Terms

16

Just for kicks, what would happen if we examine the Type III (simultaneous) Sums of Squares?

```
> lm(area ~ age + I(age^2) + I(age^3), data = re) %>% car::Anova()
Anova Table (Type II tests)

Response: area
          Sum Sq  Df F value    Pr(>F)
age         1787   1 12.0424 0.0005752 ***
I(age^2)      27   1  0.1807 0.6710037
I(age^3)     224   1  1.5126 0.2194417
Residuals  60827 410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We know a higher-order term is necessary, but this output shows us that the quadratic and cubic term are redundant when in the model together.

16

2. Polynomial Terms

17

Therefore our best-fit model should include up to the quadratic polynomial term:

$$\hat{Y} = 53.4 - 1.93X + 0.04X^2$$

```
> lm(area ~ age + I(age^2), data = re) %>% summary()

Call:
lm(formula = area ~ age + I(age^2), data = re)

Residuals:
    Min      1Q  Median      3Q     Max
-26.542  -9.085  -0.445   8.260  79.961

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.450539   1.651624  32.362   <2e-16 ***
age         -1.928877   0.193759  -9.955   <2e-16 ***
I(age^2)     0.042181   0.004689   8.995   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.19 on 411 degrees of freedom
Multiple R-squared:  0.2015,     Adjusted R-squared:  0.1977
F-statistic: 51.87 on 2 and 411 DF,  p-value: < 2.2e-16
```

17

2. Polynomial Terms

18

Interpreting Beta Coefficients

$$\hat{Y} = 53.4 - 1.93X + 0.04X^2$$

- The β coefficients are less interpretable for variables represented by a polynomial set.
- This is because the effect of a one-unit increase in X on Y depends on the value of X.

Example

$$\Delta\hat{Y}_{X=2\ vs\ X=1} = \left(53.4 - 1.93(2) + 0.04(2^2)\right) - \left(53.4 - 1.93(1) + 0.04(1^2)\right) = -1.81$$

$$\Delta\hat{Y}_{X=1\ vs\ X=0} = \left(53.4 - 1.93(1) + 0.04(1^2)\right) - \left(53.4 - 1.93(0) + 0.04(0^2)\right) = -1.89$$

18

2. Polynomial Terms 19

Interpreting Beta Coefficients

$$\hat{Y} = 53.4 - 1.93X + 0.04X^2$$

- The intercept is still interpreted the same way.

Example

"The expected selling price is $53.4 per unit area for a house of age = 0."

19

2. Polynomial Terms 20

The Hierarchy Principle

In general, if your model includes X^h as a statistically significant predictor of Y, then your model should include X^j for all $j<h$, regardless of whether the lower-degree terms are significant in the model.

20

2. Polynomial Terms 21

The Hierarchy Principle

Are there any exceptions to this? Let's look at an equation with a quadratic term.

Note that a quadratic equation can be written the following two ways:

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 = \beta_2 (X - \gamma_1)^2 + \gamma_2$$

Where,

$$-2\beta_2\gamma_1 = \beta_1$$
$$\beta_2\gamma_1^2 + \gamma_2 = \beta_0$$

21

The Hierarchy Principle

In this equation, the value $x = \gamma_1$ reflects the extremum/vertex of the quadratic relationship (the "dip" or "peak" point of the parabola).

$$\hat{Y} = \beta_2(X - \gamma_1)^2 + \gamma_2$$

So if we don't include a linear term, our equation becomes:

$$\hat{Y} = \beta_0 + \beta_2 X^2 = \beta_2(X - 0)^2 + \gamma_2$$

In other words, if we exclude the linear term then we essentially force $\gamma_1 = 0$, thereby forcing the vertex of the parabola at X=0.

22

The Hierarchy Principle

What are the implications of this?

- We do not need the linear term if we are certain the vertex of the parabola is at X=0.
- If we are unsure of where the vertex is located (which is usually the case) then we need to include the linear term.
- This reasoning extends to polynomial terms of higher degree as well.

23

Centering X with Polynomials

- It is generally a good idea to center your X variables on their means
- One criticism of polynomial equations is that the higher- and lower-order terms are strongly related.
- Mean-centering reduces the amount of correlation among polynomial terms
- Recall, high correlation (collinearity) among independent variables increases the chance of numerical/estimation problems

24

2. Polynomial Terms


```
> re %>%
+   select(age) %>%
+   mutate(age2 = age*age,
+          age3 = age2*age) %>%
+   cor()
           age      age2      age3
age  1.0000000 0.9623851 0.9094008
age2 0.9623851 1.0000000 0.9863162
age3 0.9094008 0.9863162 1.0000000
```

For the non-centered variable, age, age^2, and age^3 are all very highly related.

```
> re %>%
+   select(age.c) %>%
+   mutate(age2.c = age.c*age.c,
+          age3.c = age2.c*age.c) %>%
+   cor()
           age.c    age2.c    age3.c
age.c  1.0000000 0.3606833 0.9015645
age2.c 0.3606833 1.0000000 0.4904115
age3.c 0.9015645 0.4904115 1.0000000
```

By mean-centering these variables the correlations are reduced.

25

2. Polynomial Terms

Conclusion Statement

Upon visual inspection we found that age was not linearly related to house selling price. We found that a quadratic model fit the data well (p<.001) and a cubic term did not improve model fit (p=0.22). Our best-fit equation for selling price was $\hat{Y} = 53.4 - 1.93X_{AGE} + 0.04X_{AGE}^2$, as selling price decreased until house age of approximately 22 years, and then subsequently began to rebound.

Overall F-test for the model with the quadratic relationship.

No need to interpret the beta coefficients directly, but a general description of the quadratic effect should be included.

26

2. Polynomial Terms

Recap

- By adding polynomial terms, you can fit models where X is not linearly related to Y.
- Polynomial terms have a tendency to be highly correlated; be wary of this when adding many of them to a model.
- It is a good idea to center polynomial terms on their means.

27

2. Polynomial Terms 28

Recap

- Determine if a polynomial term is necessary in a model
- Determine the order of polynomials to be included
- Describe the effect of X on Y when polynomial terms are included

28

3. Fractional Polynomials 29

Fractional Polynomials

The Fractional Polynomials (FP) approach provides a more flexible way to parameterize variables.

Strategy: Find a transformation of X (e.g., log(X), X^2) that fits the data best.

$$g(x, \beta) = \beta_0 + \sum_{j=1}^{J} F_j(x)\beta_j$$

This just means that we'll have some combination of transformations of our X variable in our model.

$$F_j(x) = x^{p_j},\ if\ p_j \neq p_{j-1}$$

$$F_j(x) = F_{j-1}\ln(x),\ if\ p_j = p_{j-1}$$

This means that if we specify two identical values of p, the first term will be x^p and the second will be $x^p\ln(x)$

29

3. Fractional Polynomials 30

Notice that this is similar to a Box-Cox transformation, but:

- It is for the X variables
- Up to two terms are chosen
- In addition to choosing the polynomial term, the algorithm will scale your variables to find an appropriate transformation
- Powers are chosen only from the following discrete set of transformations:

p	-2	-1	-1/2	0	1/2	1	2	3
x*	$1/x^2$	1/x	$1/x^{1/2}$	log(x)	$x^{1/2}$	x	x^2	x^3

30

3. Fractional Polynomials

31

Some Examples:

Model (P's)	g(x)
-2, -1	$\beta_0 + \beta_1\left(\frac{1}{x^2}\right) + \beta_2\left(\frac{1}{x}\right)$
-2, 0	$\beta_0 + \beta_1\left(\frac{1}{x^2}\right) + \beta_2(\ln(x))$
1, 1	$\beta_0 + \beta_1(x) + \beta_2(x)\ln(x)$

31

3. Fractional Polynomials

32

The **Multiple Fractional Polynomial** "mfp" package allows for this testing.

It will assess:

Null/unconditional model (no x)

Linear model (linear x, or $\beta_0+\beta_1 x$)

Best fitting J=1 (1-term) model ($\beta_0+\beta_1 x^{p1}$)

Best fitting J=2 (2-term) model ($\beta_0+\beta_1 x^{p1}+\beta_2 x^{p2}$)

32

3. Fractional Polynomials

33

```
> mfp(area ~ fp(age), data = re)
Call:
mfp(formula = area ~ fp(age), data = re)


Deviance table:
                      Resid. Dev
Null model            76461.38
Linear model          73071.2
Final model           60852.65

Fractional polynomials:
    df.initial select alpha df.final power1 power2
age          4      1  0.05        4      1      3


Transformations of covariates:
                                    formula
age I(((age+0.1)/10)^1)+I(((age+0.1)/10)^3)

Re-Scaling:
Non-positive values in some of the covariates. No re-scaling was performed.

Coefficients:
Intercept      age.1      age.2
  51.6829   -12.9878     0.7182

Degrees of Freedom: 413 Total (i.e. Null);  411 Residual
Null Deviance:          76460
Residual Deviance: 60850        AIC: 3249
```

The "deviance" is another name for SSE.

This indicates the inclusion of a (1) linear term and (3) cubic term.

Note that age was scaled so that there were no negative or zero values, and divided by 10.

The final "best" model is:

$$\hat{Y} = \beta_0 + \beta_1 \frac{X_{AGE}+0.1}{10} + \beta_2\left(\frac{X_{AGE}+0.1}{10}\right)^3$$

33

Recap

- Fractional polynomials finds the best-fitting transformations of X variables
- Because the transformations can be complex (and less interpretable), this approach is better suited for prediction models (vs. models of association)
- By examining several possible flexible transformations, the FP approach can sometimes "overfit" the data
- This approach can also be used to generally test for a departure from linearity

34

Recap

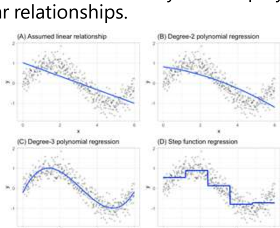
➢Implement the fractional polynomials method in analysis

➢Describe the strengths and drawbacks of using fractional polynomials in practice

35

Accounting for Non-Linearity

In the previous section we examined ways to use polynomial terms to account for nonlinear relationships.

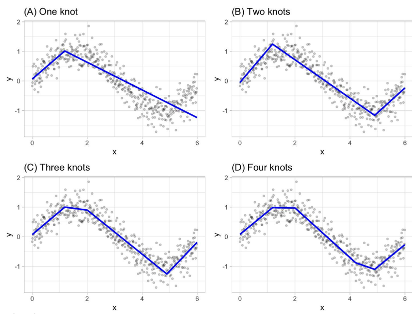


https://bradleyboehmke.github.io/HOML/mars.html

36

4. Splines 37

Splines are a modeling approach where the regression equation between Y and X is broken into "chunks" based on values of X.



https://bradleyboehmke.github.io/HOML/mars.html

37

4. Splines 38

The choice of where to create these spline "knot" points can depend on:

1. **Data-Driven Approach**: choose the knot points that fit the data the best
   - Good for machine learning
   - Good for prediction modeling
2. **Theory-Driven Approach**: choose the knot points a priori to answer a specific research question
   - Good for hypothesis testing

We'll focus on the theory-driven approach here.

38

4. Splines 39

Examples

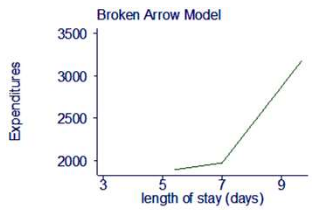
1. You're modeling out-of-pocket medical expenditures (Y). You're told that HMOs typically only pay for the first week of a hospital stay, after which out-of-pocket expenditures will likely increase dramatically.
2. In the US, older adults are eligible for Medicare, a national health insurance program, once they reach the age of 65. Therefore their medical expenditure patterns (Y) may be drastically different after they reach age 65.
3. Children in the CHS are followed into early adulthood. We expect a non-linear relationship between FEV1 and age during adolescence.

39

4. Splines

This graph depicts the medical expenditures we expect to see in Example 2. This type of model is called a:

- Broken arrow
- Hockey stick
- Piecewise linear spline



40

4. Splines

Suppose our regression model is of the form:

We can introduce a spline into this model by including another term:

$$X_{LOS.C7^+} = \begin{cases} X_{LOS} - 7, if\ X_{LOS} > 7 \\ 0, if\ X_{LOS} \leq 7 \end{cases}$$

Our model then becomes:

$$\hat{Y} = \beta_0 + \beta_1 X_{LOS} + \beta_2 X_{LOS.C7^+}$$

41

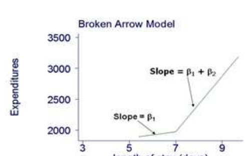
4. Splines

So when X ≤ 7,

$$\hat{Y} = \beta_0 + \beta_1 X_{LOS}$$

When X > 7,

$$\begin{aligned} \hat{Y} &= \beta_0 + \beta_1 X_{LOS} + \beta_2 (X_{LOS} - 7) \\ &= \beta_0 + \beta_1 X_{LOS} + \beta_2 X_{LOS} - 7\beta_2 \\ &= (\beta_0 - 7\beta_2) + (\beta_1 + \beta_2) X_{LOS} \\ &= \beta_0^* + \beta_1^* X_{LOS} \end{aligned}$$



This allows for a different slope after 7 days.

42

Notes on Splines

$$\hat{Y} = \beta_0 + \beta_1 X_{LOS} + \beta_2 X_{LOS.C7^+}$$

- The value of the spline term β_2 quantifies the difference in slopes before the given X value and after.
- The test of $\mathrm{H}_0\text{: } \beta_2 = 0$ is a test of whether there is a significant change in slope after the specified value of X.

43

Example

The Affordable Care Act (aka “Obamacare”) was initiated in March 2010. Data on the state-level per capita healthcare expenditure was obtained (kff.org) and this data was plotted for the years 2000 through 2014.

While health expenditure per capita (HEPC) tended to increase over time, we want to know if the yearly increase was attenuated after the year 2010.

44

Examine the linear effect of HEPC over time.

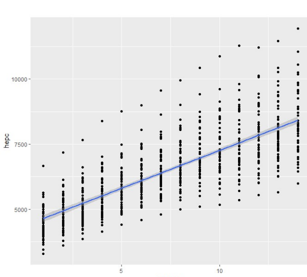
```
> lm(hepc ~ yearnum, data = hccp) %>% summary()

Call:
lm(formula = hepc ~ yearnum, data = hccp)

Residuals:
    Min      1Q  Median      3Q     Max 
-2473.2  -634.5  -144.5   513.3  3731.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4353.158     76.655   56.79   <2e-16 ***
yearnum      290.618      9.003   32.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 969.7 on 712 degrees of freedom
Multiple R-squared:  0.5941,	Adjusted R-squared:  0.5935 
F-statistic:  1042 on 1 and 712 DF,  p-value: < 2.2e-16
```

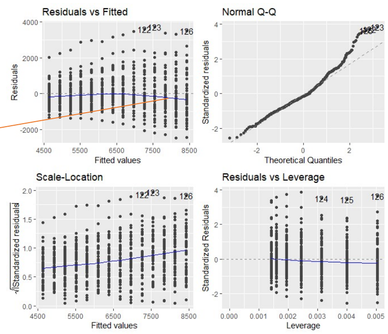


45

4. Splines

46

Interestingly, linearity does seem to be slightly violated in the 4 highest fitted values (corresponding to years 2011 to 2014).



46

4. Splines

47

Examine the spline model.

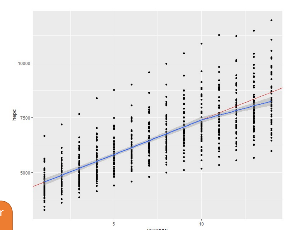
```
> lm(hepc ~ yearnum + I((yearnum-10)*(yearnum>10)), data = hccp) %>% summary()

Call:
lm(formula = hepc ~ yearnum + I((yearnum - 10) * (yearnum > 10)),
    data = hccp)

Residuals:
    Min      1Q  Median      3Q     Max
-2377.2  -643.1  -157.6   514.4  3694.3

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         4242.85      90.31  46.980   <2e-16 ***
yearnum                              314.89      13.88  22.689   <2e-16 ***
I((yearnum - 10) * (yearnum > 10))  -100.38      43.79  -2.293   0.0222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 966.8 on 711 degrees of freedom
Multiple R-squared:  0.5971,    Adjusted R-squared:  0.5959
F-statistic: 526.8 on 2 and 711 DF,  p-value: < 2.2e-16
```



The predicted per capita health expenditure increase is \$314.89 for years 2000-2010. After year 2010, this increase was significantly attenuated by \$100.38/year (p=.02).
The annual increase in HEPC was (314.89-100.38) = \$214.51.

47

4. Splines

48

Reparametrizing the spline model.

Note: we can reparametrize the spline model to instead obtain the slope for year *after* 2010 and the change in slope for years 2010 and prior:

```
> lm(hepc ~ yearnum + I((yearnum-10)*(yearnum<=10)), data = hccp) %>% summary()

Call:
lm(formula = hepc ~ yearnum + I((yearnum - 10) * (yearnum <=
    10)), data = hccp)

Residuals:
    Min      1Q  Median      3Q     Max
-2377.2  -643.1  -157.6   514.4  3694.3

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          5246.69     397.16  13.211  < 2e-16 ***
yearnum                               214.50      34.39   6.237 7.65e-10 ***
I((yearnum - 10) * (yearnum <= 10))   100.38      43.79   2.293   0.0222 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 966.8 on 711 degrees of freedom
Multiple R-squared:  0.5971,    Adjusted R-squared:  0.5959
F-statistic: 526.8 on 2 and 711 DF,  p-value: < 2.2e-16
```

48

Recap

- Splines allow for a different β coefficient across a certain range of X
- Splines can be useful for:
  - 1) Helping to model a certain effect
  - 2) Testing hypotheses about when a certain effect differs as a function of X

49

Recap

➢Create the variables necessary to implement a linear splines approach

➢Correctly interpret the slope coefficients from a linear splines analysis

50

Applications

We often find variables that are only applicable when the participant has the existence of some health behavior.

Examples:

- For those who smoke, how many packs per week do you smoke?
- For those who exercise, how many hours per week do you exercise?
- What is the percent of your named friendships that are reciprocated (reciprocation is undefined for those naming no friends)?

51

5. Dose-Response Coding 52

Example

Children in high schools were surveyed about the peers at school they considered a friend. They were also given a survey asking them about their "maladaptive coping" habits (e.g., turning to drugs/alcohol when under duress).

Two network measures were computed:

Reciprocity: The proportion of friendships named by the student that were also named in return (i.e., reciprocated).

Out-degree: The number of friendship nominations made.

The outcome is "maladaptive coping" z-score.

52

5. Dose-Response Coding 53

Problem: reciprocity is undefined for those who did not name friends.

How do we include these two concepts in a model?

Consider the following coding scheme:

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

53

5. Dose-Response Coding 54

What is the baseline category?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

54

5. Dose-Response Coding

What is the baseline category?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

Named 0 friends.

55

5. Dose-Response Coding

What does $\beta_{any\_friends}$ represent?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

56

5. Dose-Response Coding

What does $\beta_{any\_friends}$ represent?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

The effect on Y associated with naming any friends (vs. naming no friends), holding reciprocity constant (by necessity, at 0).

i.e., the difference in Y for a student with zero reciprocity who named any friends, compared to a student who named no friends.

57

5. Dose-Response Coding

What does β_{recip} represent?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

58

5. Dose-Response Coding

What does β_{recip} represent?

$$X_{any\_friends} = \begin{cases} 1, named \geq 1 friends \\ 0, named\ 0\ friends \end{cases}$$

$$X_{recip} = \begin{cases} reciprocity, named \geq 1\ friends \\ 0, named\ 0\ friends \end{cases}$$

The predicted change in Y for a 1-unit increase in reciprocity, holding $X_{any\_friends}$ constant (by necessity, at 1).

i.e., the difference in Y for a one-unit increase in reciprocity among those who named any friends.

59

5. Dose-Response Coding


```
> lm(m_cope ~ recip + any_friends,
+    data = sos %>%
+      mutate(recip = ego1_or,
+             any_friends = (ego_odg>0)
+             )
+    ) %>%
+   summary()

Call:
lm(formula = m_cope ~ recip + any_friends, data = sos %>% mutate(recip = ego1_or,
    any_friends = (ego_odg > 0)))

Residuals:
    Min      1Q  Median      3Q     Max
-1.2025 -0.6537 -0.2270  0.4917  3.5776

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.31839    0.02584  12.319  < 2e-16 ***
recip           -0.24011    0.03053  -7.865 4.06e-15 ***
any_friendsTRUE -0.29826    0.03006  -9.922  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.845 on 10253 degrees of freedom
  (813 observations deleted due to missingness)
Multiple R-squared:  0.02597,	Adjusted R-squared:  0.02578
F-statistic: 136.7 on 2 and 10253 DF,  p-value: < 2.2e-16
```

The predicted maladaptive coping score for someone who named no friends is 0.32.

The decrease in maladaptive coping score associated with a 1-unit increase in reciprocity, for those who named any friends.

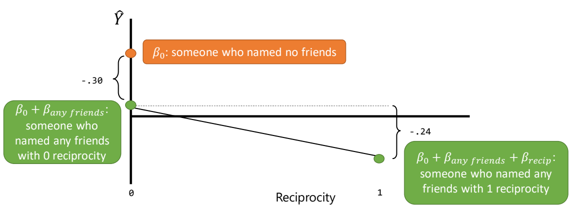
The decrease in maladaptive coping score associated with naming any friends but having 0 reciprocity, compared to somebody who named no friends.

In other words, reciprocity decreases maladaptive coping scores. Furthermore, naming any friends additionally decreases maladaptive coping.

60

5. Dose-Response Coding

$$\hat{Y} = 0.32 - 0.30X_{any\ friends} - 0.24X_{recip}$$



61

5. Dose-Response Coding

Conclusion

We examined the effect of reciprocity (continuous, between 0 and 1) and naming any friends (1 vs. 0) on maladaptive coping attitudes. These variables were significantly related to maladaptive coping score ($p<.001$) but only explained 2.6% of the variance in outcome. We found that a one-unit increase in reciprocity was associated with a decrease in maladaptive coping score of 0.24 ($p<.001$). Furthermore, when compared to somebody who named no friends, naming any friends but having 0 reciprocity was associated with a decrease in maladaptive coping score of 0.30 ($p<.001$).

62

5. Dose-Response Coding

Recap

- This particular coding scheme can be used when one variable is an indicator of the presence of a phenomenon, and the second variable indicates an additional effect that is contingent on the first variable.
- This can be extended to processes such as:
  - Smoking: variable1=yes/no, variable2=packs per day
  - Exercise: variable1=yes/no, variable2=intensity
  - Disease: variable1=presence/absence, variable2=severity
  - Dose-response: variable1=medication/placebo, variable2=dose

63

5. Dose-Response Coding

64

Recap

➢Create the variables necessary to implement dose-response coding

➢Correctly interpret the slope coefficients from an analysis with dose-response coding

64

6. Overfitting

65

Explaining Too Well

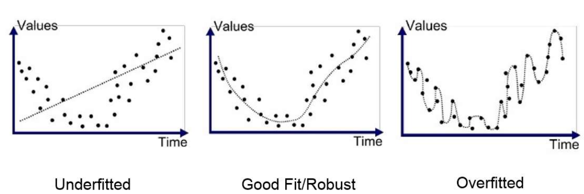
- One gut desire when creating a model is to try to explain as much about our Y variable as possible
- When we try to model phenomena, there is always some random component *e* associated with that phenomenon that cannot be explained
- **Overfitting** occurs when our model is fit too well to the data and no longer is an accurate representation of the underlying process

65

6. Overfitting

66

An **overfit** model may explain your current data set quite well, but will not be generalizable to other data sets with the same variables.



66

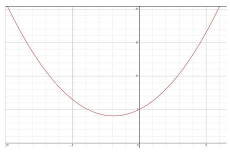
6. Overfitting
67

An Overfitting Experiment

Let's examine the output from a randomly generated sequence of X and Y variables. The underlying (true) relationship is:

$$Y = 5 + X + 0.25X^2 + e$$

I actually randomly generated 31 Y values to reflect this relationship; you'll see this next.



67

6. Overfitting
68

True relationship:

$$Y = 5 + X + 0.25X^2 + e$$

Model Prediction:

$$\hat{Y} = 7.25 + 0.82X + 0.24X^2$$

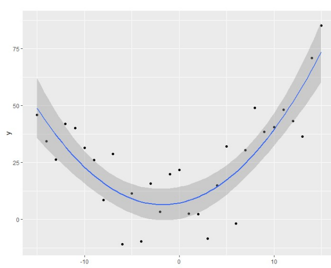
```
> lm(y ~ x + I(x^2), data = ex1) %>% summary()

Call:
lm(formula = y ~ x + I(x^2), data = ex1)

Residuals:
    Min      1Q  Median      3Q     Max
-22.523  -7.912   2.333   9.514  19.773

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.25440    3.43310   2.113  0.04364 *
x            0.82082    0.25567   3.211  0.00332 **
I(x^2)       0.23987    0.03201   7.494 3.67e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.73 on 28 degrees of freedom
Multiple R-squared:  0.7036,	Adjusted R-squared:  0.6824
F-statistic: 33.23 on 2 and 28 DF,  p-value: 4.04e-08
```



68

6. Overfitting
69

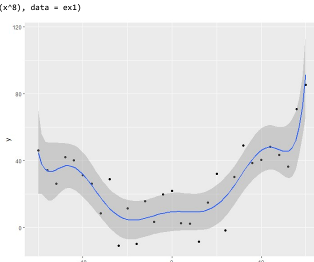
What happens if we include several polynomial terms to try to increase the fit of the model?

```
> lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) + I(x^7) + I(x^8), data = ex1)
Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) +
    I(x^7) + I(x^8), data = ex1)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1790  -6.8967   0.3464   6.0284  18.5761

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.300e+00  5.513e+00   1.687   0.1058
x            4.499e-01  1.660e+00   0.271   0.7889
I(x^2)      -3.434e-01  4.652e-01  -0.738   0.4682
I(x^3)       2.447e-02  5.084e-02   0.481   0.6351
I(x^4)       1.656e-02  9.634e-03   1.719   0.0996 .
I(x^5)      -3.227e-04  4.472e-04  -0.722   0.4782
I(x^6)      -1.337e-04  6.833e-05  -1.957   0.0631 .
I(x^7)       1.047e-06  1.157e-06   0.905   0.3751
I(x^8)       3.203e-07  1.542e-07   2.078   0.0496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.38 on 22 degrees of freedom
Multiple R-squared:   0.78,	Adjusted R-squared:    0.7
F-statistic: 9.748 on 8 and 22 DF,  p-value: 1.085e-05
```



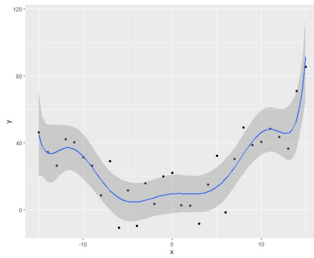
69

6. Overfitting

70

This model appears to fit "better" than the quadratic model, but:

- It has a ridiculous amount of complexity
- Its parameter estimates (8th degree polynomial) will be incredibly difficult to interpret
- It is not likely to generalize to other situations

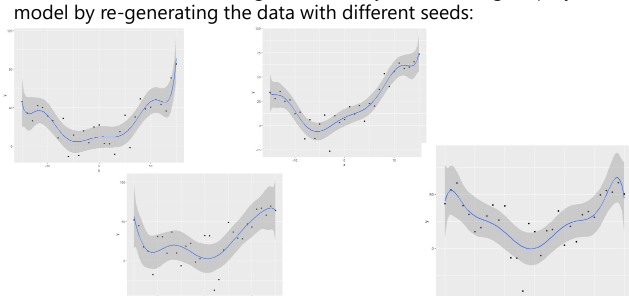


70

6. Overfitting

71

We can examine the lack of generalizability in the 8th degree polynomial model by re-generating the data with different seeds:

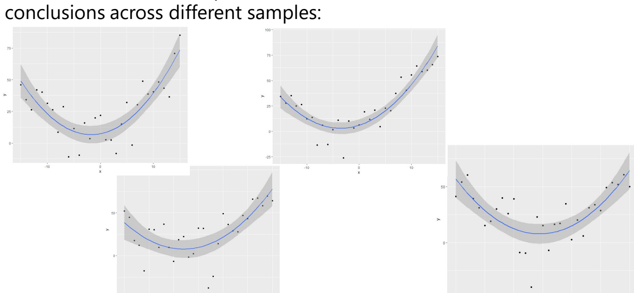


71

6. Overfitting

72

On the other hand, the quadratic model would lead to similar conclusions across different samples:



72

6. Overfitting

Recap

- Error is inherent in life, and in models ($Y = \beta_0 + \beta_1 X + e$)
- Instead of explaining away inherent error with overly-complicated models, be realistic about what your model can and cannot explain
- **Parsimony**—having the simplest model that still does well explaining an outcome—is desirable

73

6. Overfitting

Recap

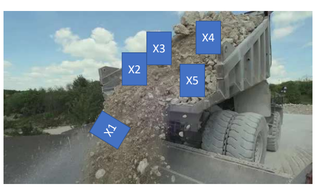
- Explain why a model that fits "too well" can be a bad thing
- Discuss the tradeoffs between model simplicity and complexity

74

7. Adjusted R-Squared

Too Many Variables

- In multiple regression, each additional X variable should improve model fit, even if just due to chance.
- Models with several variables tend to have better fit simply because they have more terms.
- The **adjusted R-squared** metric is a version of R-squared that adjusts for the number of predictors in the model.
- This is a great way to prevent overfitting!



75

7. Adjusted R-Squared

76

Adjusted R-Squared

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1}\right)$$

Where n is the sample size and k is the number of predictors.

76

7. Adjusted R-Squared

77

Let's revisit the models from the last section.

```
> lm(y ~ x + I(x^2), data = ex1) %>% summary()

Call:
lm(formula = y ~ x + I(x^2), data = ex1)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.523  -7.912   2.333   9.514  19.773 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.25440    3.43310   2.113  0.04364 *  
x            0.82082    0.25567   3.211  0.00332 ** 
I(x^2)       0.23987    0.03201   7.494 3.67e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.73 on 28 degrees of freedom
Multiple R-squared:  0.7036,	Adjusted R-squared:  0.6824 
F-statistic: 33.23 on 2 and 28 DF,  p-value: 4.04e-08
```

The quadratic model's R^2 is modestly lower, but the adjusted R^2 is practically the same as the 8-degree polynomial model.

```
> lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) + I(x^7) + 
I(x^8), data = ex1) %>%
+   summary()

Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) + 
    I(x^7) + I(x^8), data = ex1)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.1790  -6.8967   0.3464   6.0284  18.5761 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  9.300e+00  5.513e+00   1.687   0.1058  
x            4.499e-01  1.660e+00   0.271   0.7889  
I(x^2)      -3.434e-01  4.652e-01  -0.738   0.4682  
I(x^3)       2.447e-02  5.084e-02   0.481   0.6351  
I(x^4)       1.656e-02  9.634e-03   1.719   0.0996 .
I(x^5)      -3.227e-04  4.472e-04  -0.722   0.4782  
I(x^6)      -1.337e-04  6.833e-05  -1.957   0.0631 .
I(x^7)       1.047e-06  1.157e-06   0.905   0.3751  
I(x^8)       3.203e-07  1.542e-07   2.078   0.0496 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.38 on 22 degrees of freedom
Multiple R-squared:   0.78,	Adjusted R-squared:    0.7 
F-statistic: 9.748 on 8 and 22 DF,  p-value: 1.085e-05
```

77

7. Adjusted R-Squared

78

Predicted R-Squared

The predicted R-squared is a measure of how good a particular model is at estimating new values.

That is, what percent of the variation in *new values* can be explained by the model?

This is an even better measure of overfitting.

78

7. Adjusted R-Squared

79

Here, we see that the amount of variation in new values is much higher for the quadratic model compared to the 8-degree polynomial model.

```
> lm(y ~ x + I(x^2), data = ex1) %>% pred_r_squared()
[1] 0.6410801

> lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6) + I(x^7) + I(x^8), data = ex1) %>%
+   pred_r_squared()
[1] 0.07810942
```

https://rpubs.com/RatherBit/102428

79

7. Adjusted R-Squared

80

Recap

- Alternate R-squared metrics can help determine if the model's explanatory power is simply due to overfitting

80

7. Adjusted R-Squared

81

Recap

➢Use adjusted and predicted R-squared values to assess a model's performance

81

8. Recap

82

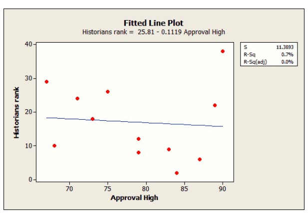
- **Plan ahead when modeling.** Knowing the type of data you have and how you want to model it will make your job easier. Think about the terms you want to use in your model and why you are using them.
- **Use coding to test hypotheses.** In addition to being more flexible, spline coding can be used to test hypotheses about differences in slopes for a particular range of X.
- **Don't be too good.** As in: don't over-fit. Good models explain things as simply as possible but as comprehensively as possible. Remember—all models are just a simplification of reality.
- **Sanity check.** Whatever your model estimates are, think about them to make sure they are sensible.

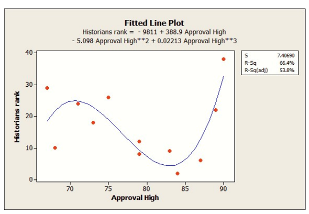
82

8. Recap

83

Relationship between President's highest approval rating and their ranking by historians





83

8. Recap

84

Additional Reading

- More on Predicted R-Squared (and function) https://rpubs.com/RatherBit/102428
- More on splines, with applications to machine learning https://bradleyboehmke.github.io/HOML/mars.html

84

8. Recap

85

Packages and Functions

- `mfp::fp()`

85