Prediction in MLM

Model comparisons and regularization
PSYC 575

October 13, 2020 (updated: 31 October 2021)

Learning Objectives

- Describe the role of **prediction** in data analysis
- Describe the problem of **overfitting** when fitting complex models
- Use **information criteria** to compare models

Prediction

Yarkoni & Westfall (2017)[1]

- “Psychology’s near-total focus on explaining the causes of behavior has led [to] … theories of psychological mechanism but … little ability to predict future behaviors with any appreciable accuracy” (p. 1100)

[1]: https://doi.org/10.1177/17456916176933

Prediction in Data Analysis

- Explanation: Students with higher SES receive higher quality of education prior to high school, so schools with higher MEANSES tends to perform better in math achievement
- Prediction: Based on the model, a student with an SES of 1 in a school with MEANSES = 1 is expected to score 18.5 on math achievement, with a prediction error of 2.5

Can We Do Explanation Without Prediction?

- “People in a negative mood were more aware of their physical symptoms, so they reported more symptoms.”
- And then . . .
- “Knowing that a person has a mood level of 2 on a given day, the person can report anywhere between 0 to 10 symptoms”
- Is this useful?

Can We Do Explanation Without Prediction?

- "CO_2 emission is a cause of warmer global temperature."
- And then . . .
- "Assuming that the global CO_2 emission level in 2021 is 12 Bt, the global temperature in 2022 can change anywhere between -100 to 100 degrees"
- Is this useful?

Predictions in Quantitative Sciences

- It may not be the only goal of science, but it does play a role
  - Perhaps the most important goal in some research
- A theory that leads to no, poor, or imprecise predictions may not be useful
- Prediction does not require knowing the causal mechanism, but it requires more than binary decision of significance/non-significance

Example (M1)

- A subsample of 30 participants

Level 1:

$$\text{symptoms}_{ti} = \beta_{0i} + \beta_{1i}\text{mood1\_pmc}_{ti} + e_{ti}$$

Level 2:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\text{mood1\_pm}_i + \gamma_{02}\text{women}_i + \gamma_{03}\text{mood1\_pm}_i \times \text{women}_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}\text{women}_i + u_{1i}$$

Two Types of Predictions

- Cluster-specific: For a person (cluster) in the data set, what is the predicted symptom level when given the predictors (e.g., mood1, women) and the person- (cluster-)specific random effects (i.e., the *u*'s)?

```
> (obs1 <- stress_data[1, c("PersonID", "mood1_pm", "mood1_pmc",
"women")])
  PersonID mood1_pm mood1_pmc women
1      103         0          0 women
> pred1 <- predict(m1, newdata = obs1)
[1] 0.3191718
```

For person with ID 103, on a day with mood = 0, she is predicted to have 0.32 symptoms

Two Types of Predictions

- Unconditional/marginal: for a new person not in the data, given the predictors but not the *u*'s

```
> predict(m1, newdata = obs1, re.form = NA)
[1] 0.9219858
```

For a random person who's a female and with an average mood = 0, on a day with mood = 0, she is predicted to report 0.92 symptom

Prediction Errors

- Prediction error = Predicted Y ($\tilde{Y}$) – Actual Y
- For our observation:
  $\tilde{e}_{ti} = \tilde{Y}_{ti} - Y = 0.32 - 0 = 0.32$

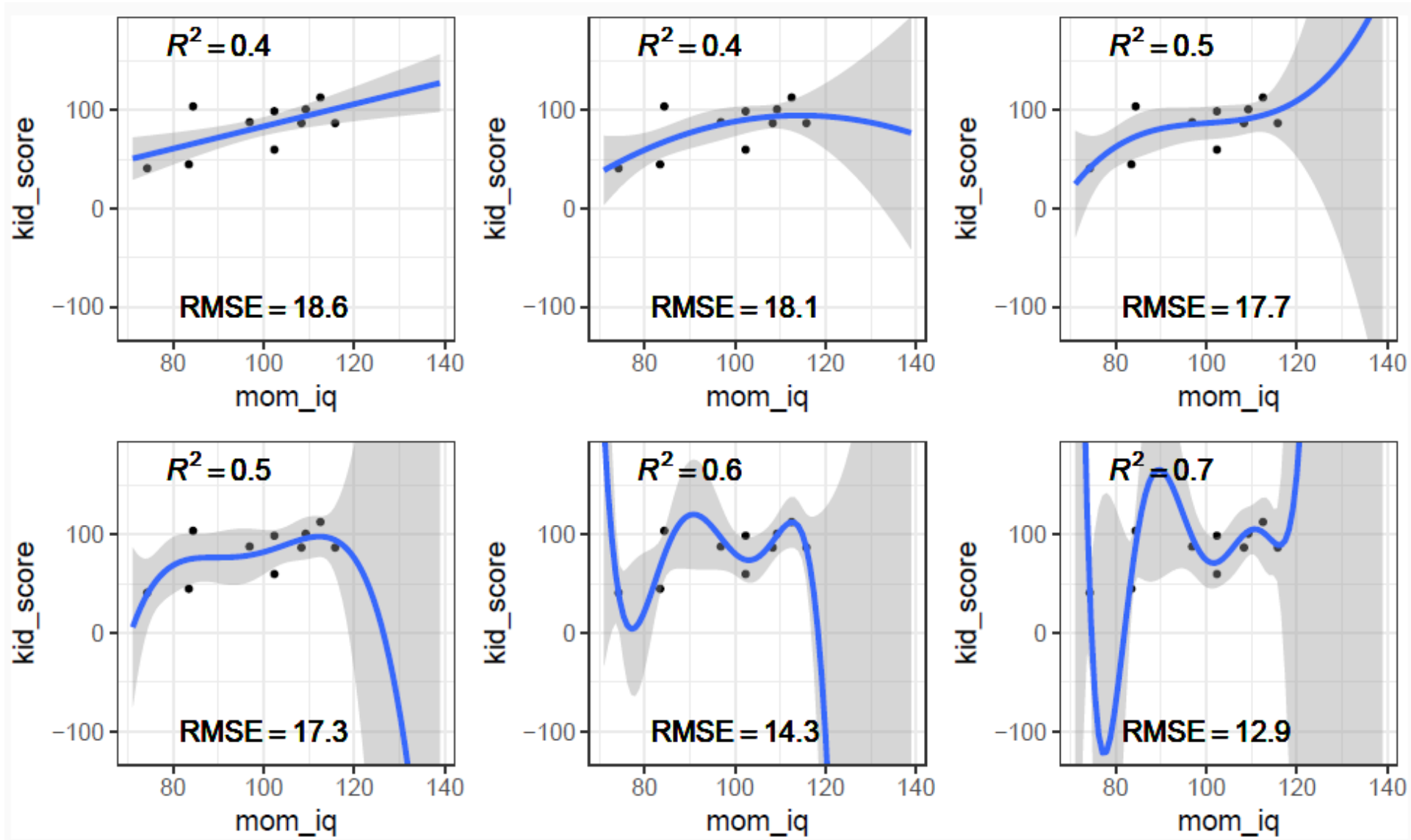
Average In-Sample Prediction Error

- Mean squared error (MSE): $\sum\sum \tilde{e}_{ti}^2 \, / N$
- In-sample MSE: average squared prediction error when using the same data to build the model and compute prediction
- Here we have in-sample MSE = 1.04
  - The average squared prediction error is 1.04 symptoms

Overfitting

Overfitting

- When a model is complex enough, it will reproduce the data perfectly (i.e., in-sample MSE)
- It does so by capturing all idiosyncrasy (noise) of the data


$R^2 = 0.4$
RMSE = 18.6
$R^2 = 0.4$
RMSE = 18.1
$R^2 = 0.5$
RMSE = 17.7
$R^2 = 0.5$
RMSE = 17.3
$R^2 = 0.6$
RMSE = 14.3
$R^2 = 0.7$
RMSE = 12.9
kid_score
mom_iq
100
0
−100
80
100
120
140

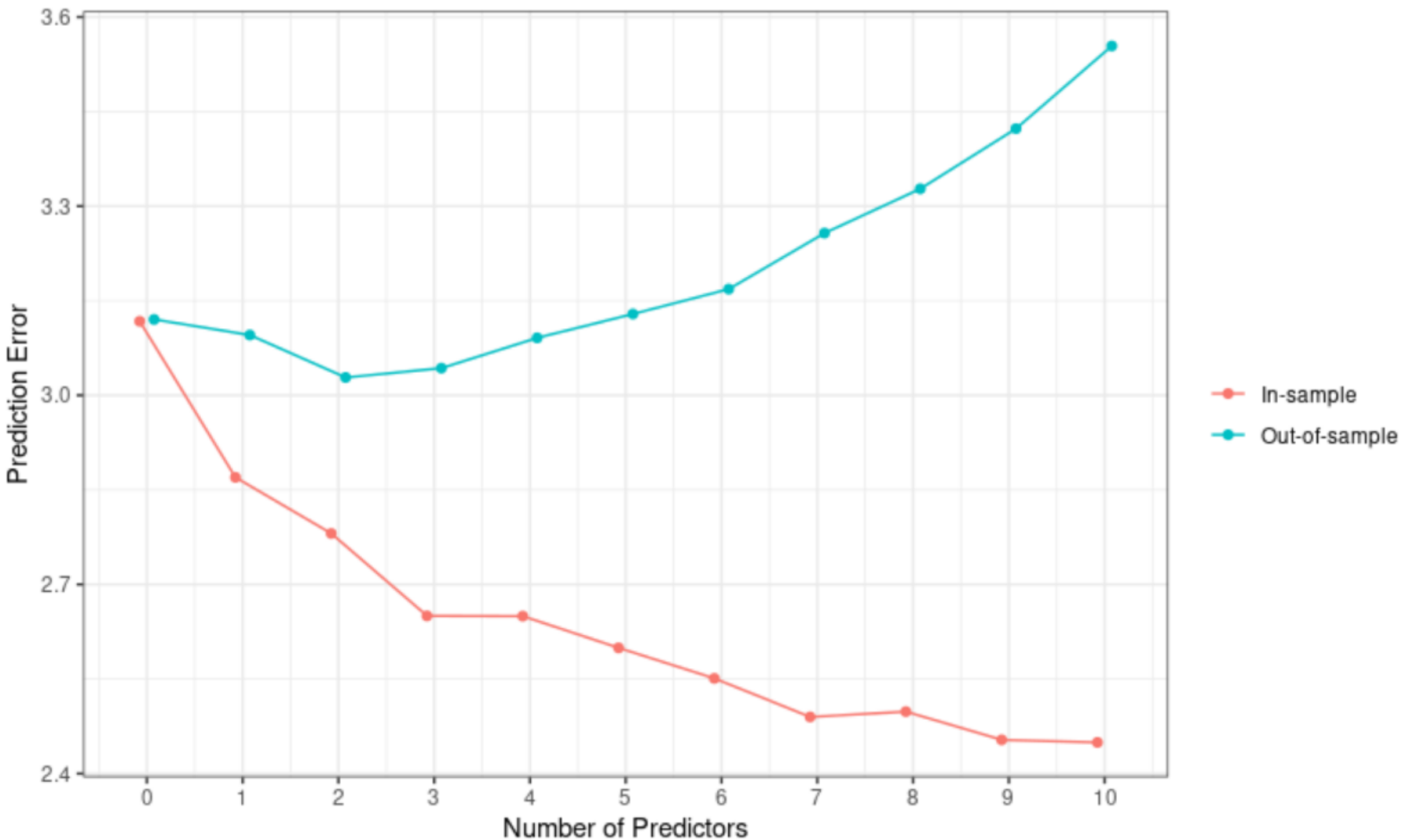
Example (M2)

```
symptoms ~ (mood1_pm + mood1_pmc) * (stressor_pm + stressor) *
            (women + baseage + weekend) +
            (mood1_pmc + stressor | PersonID)
```

- 35 fixed effects
- In-sample MSE = 0.76
  - Reduction of 27%
- Some of the coefficient estimates were extremely large

Out-Of-Sample Prediction Error

- A complex model tends to overfit as it captures the noise of a sample
  - But we’re interested in something generalizable in science
- A better way is to predict another sample not used for building the model (i.e., the remaining 75 participants)
- Out-of-sample MSE:
  - M1: 1.84
  - M2: 2.47
- So M1 is more generalizable, and should be preferred


3.6
3.3
3.0
2.7
2.4
Prediction Error
0
1
2
3
4
5
6
7
8
9
10
Number of Predictors
In-sample
Out-of-sample

Estimating Out-of-Sample Prediction Error

Approximating Out-Of-Sample Prediction Error

- But we usually don’t have the luxury of a validation sample
- Possible solutions
  - Cross-validation
  - Information criteria
- They are basically the same thing; just with different approaches (brute-force and analytical)

K-fold Cross-Validation (CV)

- E.g., 5-fold
- Splitting the data at hand
- M1:
  5-fold MSE = 1.18
- M2:
  5-fold MSE = 2.95

Data	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
1st Fold 110, 125, 518, 526, 559, 564	Prediction Error	Model Building	Model Building	Model Building	Model Building
2nd Fold 130, 133, 154, 517, 523, 533	Model Building	Prediction Error			
3rd Fold 103, 143, 507, 519, 535, 557		Model Building	Prediction Error		
4th Fold 106, 111, 136, 137, 509, 547			Model Building	Prediction Error	
5th Fold 131, 147, 522, 530, 539, 543				Model Building	Prediction Error

Leave-One-Out (LOO) Cross Validation

- LOO, or *N*-fold CV, is very computationally intensive
  - Fitting the model *N* times
- M1: LOO MSE = 1.23
- M2: LOO MSE = 3.37
- So M1 should be preferred

Information Criteria

- AIC: An Information Criterion
  - Or Akaike information criterion (Akaike, 1974)
- Under some assumptions,
  - Prediction error = deviance + $2p$
  - where p is the number of parameters in the model

Two AICs in MLM

- (Marginal) AIC: predicting a new cluster
  - Most software reports this
- (Conditional) AIC: predicting a new observation of an existing cluster
  - Available in the “cAIC4” package in R

mAIC vs cAIC

mAIC

- Sensitive to differences at level 2

```
>#         df      AIC
># fit_m1 10 399.4346
># fit_m2 43 399.9684
```

cAIC

- Sensitive to differences at level 1

```
>#    df       caic
># m1 29.06952 367.2822
># m2 51.5599  388.6662
```

Summary

- More complex models are more prone to **overfitting** when the sample size is small
- A model with smaller **out-of-sample prediction error** should be preferred
- Out-of-sample prediction error can be estimated by
  - Cross-validation
  - Information criteria (e.g., AIC)

Model Comparison

Comparing Models

- Previously, we talked about using the likelihood ratio test (LRT) to test two nested models, such as
  - M_0: X_1, X_2
  - M_1: X_1, X_2, X_3, X_4
  - Significant LRT suggested non-zero coefficients for at least one of X_3 and X_4, holding constant X_1 and X_2
- However, LRT should not be used for non-nested models
  - M_0: X_1, X_3
  - M_1: X_1, X_2, X_4

Comparing Models

- Also, LRT is not very useful for selecting the best models in a set of candidate models
- Instead, AIC is more useful

Model Comparison Example

1. `mood1` and `stressor`, no random slopes
2. `mood1_pm`, `mood1_pmc`, `stressor_pm`, `stressor`, no random slopes
3. `mood1_pm`, `mood1_pmc`, `stressor_pm`, `stressor`, random slopes
4. Model 3 + interaction terms: `mood1_pm:stressor_pm`, `mood1_pmc:stressor`

Model Comparison Example

- Marginal AIC

```
>#     df      AIC
># m_1  5 404.2322
># m_2  7 403.5750
># m_3 12 410.2415
># m_4 13 414.3175
```

- Conditional AIC

```
>#     df       caic
># m_1 28.93959 369.3262
># m_2 28.75855 369.2557
># m_3 28.75855 369.2557
># m_4 30.80038 373.5176
```

Models 1 and 2 should be preferred

Using AIC

- In practice, model comparison should be based on both statistical performance and substantive considerations
  - E.g., some variables may be included due to theoretical importance
- Instead of just selecting one model with the best AIC, identify a few models with similar AICs
- Difference in AIC > 10 usually considered big

Topics Not Covered

- Other information criteria (e.g., BIC; there are hundreds of them)
- Classical regularization techniques (e.g., Lasso, ridge regression)
  - See bonus R code for Lasso with MLM
- Variable selection methods
- Model averaging
  - See also bonus R code