# Lecture 8. Statistical Models

Z. Tuba Suzer-Gurtekin/James Wagner

March 2025

# Overview

1. Model Purpose
2. Evaluating an Experiment
3. Discovering Relationships
4. Making Predictions

# Model Purpose

The choices we make in model selection depend upon the purpose of the model. Consider the following purposes:

- Evaluating the results of an experiment
- Discovering relationships
- Prediction for new cases

# Evaluation of Experiment

Randomization doesn't always "work." That is, randomization is meant to balance covariates (observed and unobserved) across treatment groups.

In practice, sometimes this doesn't happen. For example, person 50+ might be half the sample, but we end up with 46% 50+ in one treatment group and 54% in another.

We might want to **control** for many or even all observed covariates as a way to address these imbalances.

One way to control for observed covariates is to build a regression model. With a binary outcome, logistic regression is one way to implement that.

# Discovering Relationships

Another purpose for building models is to discover relationships in existing data.

Need to be cautious – we don’t want to ransack a data set in order to identify all significant models.

But many large surveys collect many variables. This allows multiple investigators to test hypotheses.

Try to follow suggestions from last week, e.g. Simmons (2011).

# Discovering Relationships

Useful relationships can be discovered from the data this way.

It would be good to replicate such results.

As with an experiment that has poor randomization, **observational studies**, i.e. those that do not involve random assignment of treatments, may be prone to confounding.

May need to use regression techniques to control for imbalances across “treatments.”

More on this next semester.

# Prediction for New Cases

Predicting values for new cases is a different problem.

We want a model that isn’t fit to features that are ‘specific‘ to the training data. That is, we do not want estimate a model that reliably predicts for the same dataset that is used to estimate the model, but fails with new data.

Therefore, we often are less concerned with coefficients, p-values, or measures of fit.

# Evaluating an Experiment

We will use Groves, et al. (2000) as an example.

Initial face-to-face survey, the Detroit Area Study, conducted in 1996.

The study asked four measures of “political and community involvement.”

# Evaluating an Experiment

Then, 15 months later, another survey was conducted among respondents to the DAS. The new survey was on assisted suicide.

Hypothesis: “...we would expect the effect of the incentive to be diminished for those with heightened levels of community involvement.”

# Evaluating an Experiment

Table: Percentage Participating in Second Survey

| Low Community Involvement | | High Community Involvement | |
|---|---|---|---|
| No Incentive | Incentive | No Incentive | Incentive |
| 21.4 | 63.3 | 50 | 65.9 |
| 56 | 60 | 130 | 132 |

# Evaluating an Experiment

Table: Logistic Regression Predicting Participating in Second Survey

# Evaluating an Experiment

From the paper:

“Given that this test of the differing incentive effects was performed on a set of respondents to the first-phase measurement (the DAS interview), we verified that the result was not an artifact of the response rate patterns of the first phase. We tested the same interaction hypothesis with more rigorous controls on various attributes of respondents that might differ between the two incentive groups (and be related to their cooperation propensity). *The first model in table 2 merely imposes the logistic regression framework on the contingency table analysis of table 1*, with the coefficients of the incentive and community involvement variables showing positive influences and an interaction effect between the two. This reflects the diminished effects of incentives for those high on community involvement. *The second model, with controls for other variables related to cooperation (race, gender, age, and education), yields the same conclusion*-those with high community involvement display diminished positive effects of incentives (approximately 22 percentage points for the typical respondent) relative to those low on involvement (approximately 47 points).”(pp. 305-306)

# Evaluating an Experiment

Some useful things from this example:

1. We can restate contingency tables as logistic models (at least with binary outcomes)
2. It may be useful to verify that the experimental results don't change when available covariates are included
3. Here, a key characteristic, community involvement, is interacted with the treatment. The significant coefficient indicates that incentives have different effects across subgroups in the population.

# Discovering Relationships

Here, we will look at a unique study, the Relationship Dynamics and Social Life study.

The study following approximately 1,000 women 18-19 years of age for 30 months. It included a baseline face-to-face survey and weekly journals.

The study was designed to identify prospectively antecedents of unintended pregnancy.

We will look at one paper using these data: Hall, et al, (2014). “The Risk of Unintended Pregnancy among Young Women with Mental Health Symptoms.”

# Discovering Relationships

Table: Predictors of Unintended Pregnancy: Part 1

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | Univariate | | Sociodemos | | Full | | Final Reduced | |
| **Predictor** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** |
| **Depression Symptoms** | | | | | | | | |
| No (<4pts CESD) | 1 | | 1 | | 1 | | 1 | |
| Yes (>=4pts CESD) | 1.6 | 1.0,2.7 | 1.2 | 0.7,2.0 | 1.1 | 0.6,1.7 | 1.2 | 0.7,1.9 |
| **Age (18=REF)** | | | | | | | | |
| 19 years | | | 1.0 | 0.6,1.6 | 1.0 | 0.6,1.6 | | |
| 20 years | | | 0.5 | 0.1,1.4 | 0.4 | 0.1,1.3 | | |
| **Race/Ethnicity** | | | | | | | | |
| Non-Black | | | 1 | | 1 | | | |
| Black | | | 1.5 | 0.8,2.6 | 1.2 | 0.7,2.0 | | |
| **Educational Enrollment (Not Enrolled=REF)** | | | | | | | | |
| High School | | | 1.0 | 0.5,2.0 | 1.1 | 0.6,2.2 | | |
| 2-year College | | | 0.7 | 0.4,1.3 | 0.7 | 0.4,1.3 | | |
| 4-year College | | | 0.5 | 0.3,1.1 | 0.8 | 0.4,1.5 | | |
| High School Drop-out | | | 0.4 | 0.2,1.0 | 0.4 | 0.2,0.9 | | |
| **Employment Status** | | | | | | | | |
| Unemployed | | | 1 | | 1 | | | |
| Employed | | | 0.7 | 0.4,1.2 | 0.7 | 0.4,1.1 | | |
| **Receiving Public Assistance** | | | | | | | | |
| No | | | 1 | | 1 | | 1 | |
| Yes | | | 1.8 | 1.1,3.0 | 1.3 | 0.8,2.3 | 1.9 | 1.2,3.0 |

# Discovering Relationships

Table: Predictors of Unintended Pregnancy: Part 2

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | Univariate | | Sociodemos | | Full | | Final Reduced | |
| **Predictor** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** |
| **Childhood Family Structure** | | | | | | | | |
| 2 Parents (biological/step) | | | 1 | | 1 | | | |
| 1 Parent Only | | | 1.5 | 0.9,2.7 | 1.5 | 0.9,2.5 | | |
| Other | | | 1.2 | 0.5,2.7 | 1.1 | 0.5,2.4 | | |
| **Mother's Age at First Birth** | | | | | | | | |
| >=20 years | | | 1 | | 1 | | 1 | |
| <20 years | | | 1.9 | 1.2,3.1 | 1.7 | 1.1,2.7 | 2.0 | 1.3,3.2 |
| **Religious Service Attendance** | | | | | | | | |
| Never | | | 1 | | 1 | | | |
| < Weekly | | | 1.5 | 0.8,2.8 | 1.5 | 0.8,2.7 | | |
| >= Weekly | | | 1.3 | 0.6,2.9 | 1.7 | 0.7,3.7 | | |
| **Relationship Status** | | | | | | | | |
| Married | | | 1 | | 1 | | | |
| Engaged | | | 2.6 | 0.3,26.0 | 2.8 | 0.3,25.0 | | |
| Romantic Relationship | | | 1.6 | 0.2,14.0 | 1.7 | 0.2,14.0 | | |
| Physical/Emotional | | | 1.2 | 0.1,11.0 | 1.4 | 0.2,12.0 | | |
| None | | | 0.6 | 0.1,6.3 | 1.2 | 0.1,10.8 | | |
| **Cohabitation Status** | | | | | | | | |
| Not Cohabiting | | | 1 | | 1 | | 1 | |
| Cohabiting | | | 2.3 | 1.2,4.2 | 1.8 | 1.0,3.1 | 2.2 | 1.3,3.7 |

# Discovering Relationships

Table: Predictors of Unintended Pregnancy: Part 3

| | **Model 1** | | **Model 2** | | **Model 3** | | **Model 4** | |
|---|---|---|---|---|---|---|---|---|
| | **Univariate** | | **Sociodemos** | | **Full** | | **Final Reduced** | |
| **Predictor** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** | **RR** | **CI** |
| **Age at Coitarche** | | | | | | | | |
| >=16 years | | | | | 1 | | 1 | |
| < 16 years | | | | | 2.2 | 1.1,4.4 | 4.2 | 2.2,7.8 |
| **History of Pregnancy** | | | | | | | | |
| No | | | | | 1 | | | |
| Yes | | | | | 1.5 | 0.9,2.5 | | |
| **Ever Had Unprotected Sex** | | | | | | | | |
| No | | | | | 1 | | | |
| Yes | | | | | 1.1 | 0.6,1.9 | | |

# Discovering Relationships

The authors use the following process:

1. Examine the two-way table results (Model 1)
2. Estimate a logistic model including Depressions Symptoms and Sociodemographics (Model 2)
3. Estimate a “full” model adding variables regarding sex and pregnancy (Model 3)
4. Select a “final” model including significant predictors and depression (Model 4)

# Discovering Relationships

The authors' interpretation is nuanced and not all results are presented here.

They focus on the two-way table analysis, and note that the other predictors in models 2, 3, and 4 are strong predictors of depression.

They also look at aspects of depression symptoms that may be important in explaining unintended pregnancy for important subgroups (i.e. those who have had a previous pregnancy and those who have not)

# Discovering Relationships

Some takeaways:

- Although the “final” model might use statistical significance as a selection criterion for most variables, it is not the only criterion.
- The interpretation of the results may require further analyses. Some of the additional analyses could have been incorporated into the models as interaction terms.
- Should we be concerned that the authors searched for significant relationships and only reported those? What else can we do?

# Making Predictions

Making predictions is a different problem. The goal isn't to observe significant relationships in the data at hand.

The goal is to predict the values of **new** data.

Therefore, measures of predictive accuracy are more important than statistical significance of estimates.

One strategy is use some of the data for **training** and some for **testing** the model.

# K-fold Cross Validation

A common approach to train-test is called **k-fold validation**.

In this approach, the data is split into $k$ subsets.

Then, one of the $k$ subsets is held out for testing and $k - 1$ are used for training.

This step is repeated for each of the $k$ subsets.

Performance metrics (e.g. MSE, Accuracy) are recorded for each of the $k$ iterations.

# K-fold Cross Validation

The number of folds, $k$, can be a function of the sample size.

A common or even default option is $k = 10$.

One concern is “**leakage**” – is the outcome somehow signalled via variables not available when making predictions?

# Other Methods of Cross Validation

Other methods of cross validation:

1. **Leave *p* out.** Repeat all samples leaving *p* out.
2. **Leave one out.** Leave out each element one at a time.
3. **Holdout.** Designate a subset for **training** and another subset for **testing**.
4. **Time series.** Train on time period(s) $t$ and test on time period(s) $t+1$.
5. **Nested.** In one step, *hyperparameters* are chosen. In a second step, the resulting *predictions* are evaluated.

We will use *k-fold* validation for selection of hyperparameters and *holdout* validation for testing resulting predictions.

# Regularization

**Regularization** is a general description for approaches to estimation and prediction that result in “simpler” solutions.

For prediction problems, this is a tool to avoid **overfitting**.

Regularization works by adding a penalty or constraint that give “simpler” solutions an advantage.

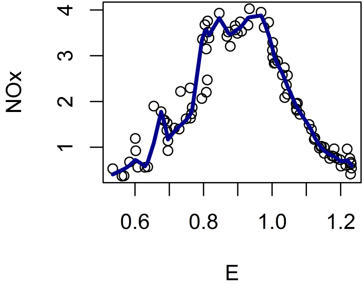
Two common types of regularization penalty functions:

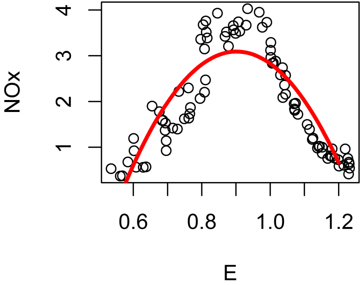
- $L_1$, based on the absolute value of parameters
- $L_2$, based on the squared value of parameters

We will look at LASSO regression as an example of an $L_1$ regularization method. An example of an $L_2$ regularization method is ridge regression.

# Overfitting

From a previous lecture. Two models – one produces **precise** estimates for these data. But may be **overfit** when concerned with making predictions for another dataset. Another seems more plausible for a wide range of possible new data.





# Model Selection for Prediction: LASSO

## LASSO

- Least Absolute Shrinkage and Selection Operator
- OLS minimizes $\sum_{i=1}^{n}(y_i - \hat{\beta}X_i)^2$
- Lasso specifies a constraint on $\hat{\beta}$: minimize $\sum_{i=1}^{n}(y_i - \hat{\beta}X_i)^2$ such that $\sum_{j=1}^{p} \mid \hat{\beta} \mid < t$
- This is often rewritten as: minimize $\sum_{i=1}^{n}(y_i - \hat{\beta}X_i)^2 + \lambda\sum_{j=1}^{p} \mid \hat{\beta} \mid$
- This approach has been extended to other GLMs (e.g. Binomial outcome)

# Model Selection for Prediction: LASSO

Under LASSO, coefficients shrink toward zero.

Some coefficients even go to 0 – therefore, this is a model selection technique.

But, need to choose the hyperparameter $\lambda$.

# Model Selection for Prediction: LASSO

LASSO is combined with cross-validation to determine which $\lambda$ leads to selecting the model that has best prediction.

**K-fold cross validation** is used to select $\lambda$.

Model with lowest misclassification error is selected. Other loss functions are possible - mean absolute error, deviance, auc, etc.

Some preference for model with 1 SE above the model with lowest misclassification error.

Then, use a **holdout sample** to validate accuracy given the selected $\lambda$.

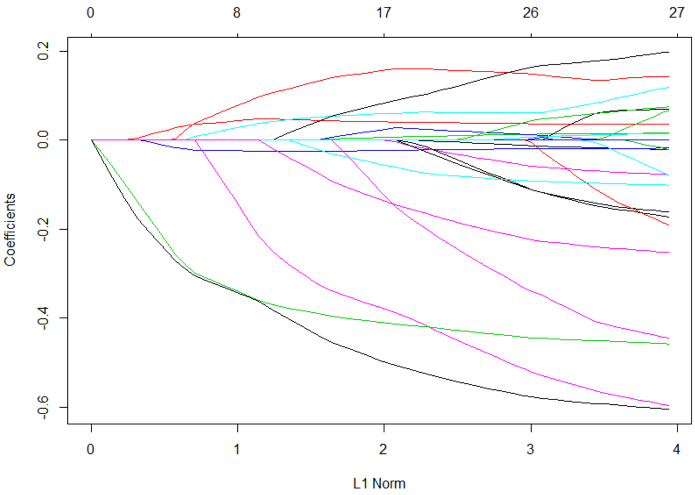
# Model Selection for Prediction: LASSO

Example Data: GPS survey data included with book *Handbook on Nonresponse in Household Surveys.*

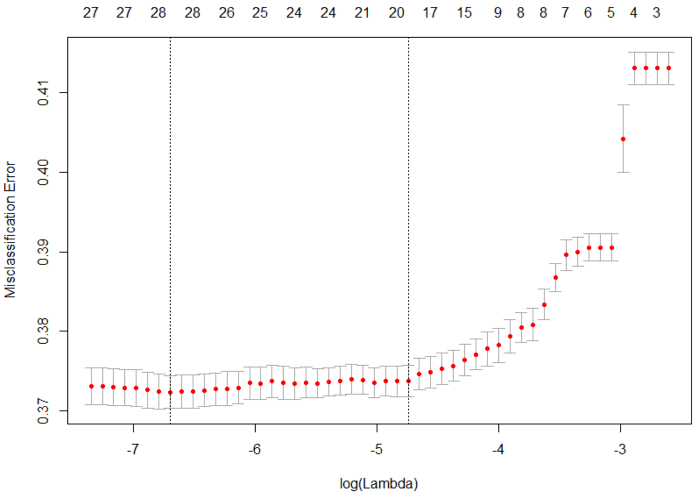
Table: Predictors in GPS Dataset

| | |
|---|---|
| GENDER | Gender |
| MARSTAT | Marital status |
| AGE13 | Age in 13 categories |
| NONNATIV | Is non-native |
| ETHNIC | Type of non-native |
| HHSIZE | Household size |
| HHTYPE | household type |
| CHILDREN | Children in household |
| PHONE | Has listed phone number |
| HASJOB | Has a job |
| SOCALL | Has social allowance |
| DISABALL | Has disability allowance |
| REGION | Region of the country |
| URBAN | Degree of urbanization |
| HOUSEVAL | Average house value in neighborhood |
| PNONNAT1 | Percentage non-natives in neighborhood |

# Model Selection for Prediction: LASSO



# Model Selection for Prediction: LASSO



# Model Selection for Prediction: LASSO

Table: Comparison of LASSO to Stepwise, Part 1

| Variable | Category | Stepwise | LASSO |
|---|---|---|---|
| Intercept | | -0.50 | 0.13 |
| AGE13 | | -0.02 | |
| CHILDREN | 1 | | |
| CHILDREN | 0 | | |
| DISABALL | 0 | 0.04 | |
| DISABALL | 1 | | |
| ETHNIC | 0 | 0.20 | |
| ETHNIC | 1 | -0.41 | -0.39 |
| ETHNIC | 2 | 0.03 | |
| ETHNIC | 3 | -0.04 | |
| ETHNIC | 4 | | |
| GENDER | 1 | -0.04 | -0.004 |
| GENDER | 2 | | |
| HASJOB | 0 | -0.05 | -0.06 |
| HASJOB | 1 | | 0.00 |
| HHSIZE | | 0.15 | 0.06 |
| HHTYPE | 1 | 0.24 | |
| HHTYPE | 2 | 0.13 | |
| HHTYPE | 3 | -0.08 | |
| HHTYPE | 4 | 0.05 | |
| HHTYPE | 5 | | -0.15 |

# Model Selection for Prediction: LASSO

Table: Comparison of LASSO to Stepwise, Part 2

| Variable | Category | Stepwise | LASSO |
|---|---|---|---|
| HOUSEVAL | | 0.02 | 0.01 |
| MARSTAT | 1 | -0.17 | |
| MARSTAT | 2 | 0.17 | 0.16 |
| MARSTAT | 3 | -0.02 | |
| MARSTAT | 4 | | |
| NONNATIV1 | 0 | | 0.03 |
| NONNATIV1 | 1 | | |
| PHONE | 0 | -0.23 | -0.41 |
| PHONE | 1 | | 0.00 |
| PNONNAT1 | | -0.02 | -0.02 |
| REGION | 21 | 0.18 | |
| REGION | 22 | 0.19 | |
| REGION | 23 | -0.09 | -0.15 |
| REGION | 24 | 0.17 | |
| REGION | 25 | | -0.51 |
| SOCALL | 0 | 0.11 | 0.09 |
| SOCALL | 1 | | 0.00 |
| URBAN | | 0.04 | 0.04 |

# Model Selection for Prediction

Some takeaways:

- Prediction is different from other reasons to select a model
- Checking predictive accuracy is the key criterion
- To avoid “overfitting”, a train-test approach is often used
- The coefficients in the model are not of interest, but reviewing them is still helpful and important

# Model Selection

We have looked at several different uses for regression models.

Each of these required different model selection strategies.

Last week, we talked about preliminary steps that are important no matter which approach we take to model selection.