Lecture 7. Statistical Models

Z. Tuba Suzer-Gurtekin/James Wagner

March 2025

Overview

Models

George E. P. Box:

*"All models are wrong,
but some models are useful."*

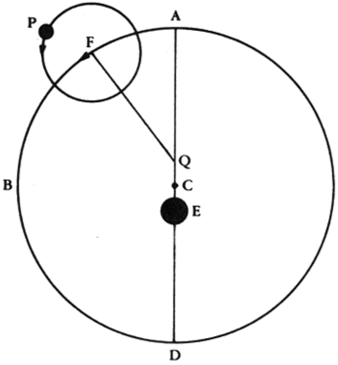
Models

Two Cautionary Examples

1. An example from physics
2. A large N problem

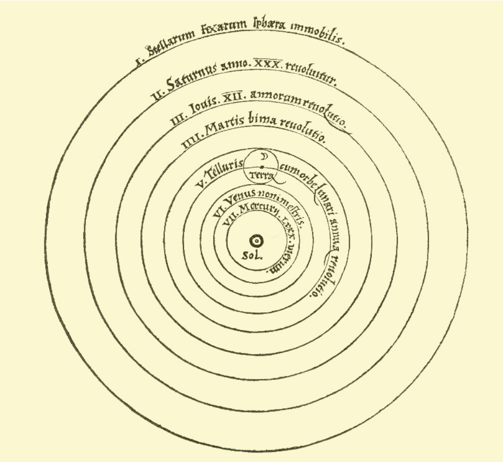
1. The Data Fit, But the Model is **Wrong**

- Analogy from physics.
- Ptolemaic view of the solar system fit the data at hand.



1. The Data Fit, But the Model is **Wrong**

- Copernican model was more parsi-monious.
- And eventually, fit new data better.



1. The Data Fit, But the Model is **Wrong**

- We are not doing physics
- Our models fit the data much less well
- Need to be aware of pitfalls

2. The Data Don't Fit, But the Model is Still **Useful**

An example from Large N.

10,000	100,500	80,125
10,150	100,475	80,150
10,450	100,450	80,175

2. The Data Don't Fit, But the Model is Still **Useful**

Using Loglinear model, the independence model:

```
> ind.model<-loglm( ~ 1 +2 , data=big)
> deviance(ind.model)
[1] 9.748811
> anova(ind.model)
Call:
loglm(formula = ~1 + 2, data = big)

Statistics:
                      X^2 df   P(> X^2)
Likelihood Ratio 9.748811  4 0.04487831
Pearson          9.768275  4 0.04451730
```

2. The Data Don't Fit, But the Model is Still **Useful**

Independence model not a good fit via significance test (Expected not "close" to Observed). But is the difference important?

Here are the fitted values from the independence model (left) and the observed data (right)

10,189.31	100,369.69	80,066.00
10,197.33	100,448.67	80,129.00
10,213.36	100,606.63	80,255.00

10,000	100,500	80,125
10,150	100,475	80,150
10,450	100,450	80,175

2. The Data Don't Fit, But the Model is Still **Useful**

Try a dataset with smaller n.

```
> small<-big/100
> ind.model2<-loglm( ~ 1 +2 , data=small)
> deviance(ind.model2)
[1] 0.09748811
> anova(ind.model2)
Call:
loglm(formula = ~1 + 2, data = small)

Statistics:
                        X^2 df  P(> X^2)
Likelihood Ratio 0.09748811  4 0.9988499
Pearson          0.09768275  4 0.9988454
```

2. The Data Don't Fit, But the Model is Still **Useful**

Smaller fitted values and observed data...

101.89	1003.70	800.66
101.97	1004.49	801.29
102.13	1006.07	802.55

100.00	1005.00	801.25
101.50	1004.75	801.50
104.50	1004.50	801.75

Model Selection

Multiple models are possible. How to select one?

From these examples, it should be clear that there aren't clear cut rules.

Need to consider two things for each model we seek to select:

1. What we already know about the problem, substantive expertise
2. The purpose of the analysis at hand

Model Selection

The **purpose** of the model is a key element of model selection.

Model Purpose

- Testing a theory
- Discovering relationships in the data
- Predicting new data

Model Selection

We have seen criteria for nested models.

For example, F-tests, Likelihood ratio tests.

These answer a specific kind of question:

- Is reduction in residuals due to additional parameters significant?

Two kinds of other situations:

1. What about models that are not nested?
2. What about models where statistical significance not important or useful?

Model Selection

From the “Large N” example above, need a useful summary, even if it doesn’t fit well (as defined by p-values, etc.).

In that example, the saturated model may not be a useful description.

The independence model may provide useful summaries.

Model Selection

Example: Grusky and Hauser (1984) look at a 3x3x16 table with large N (113,556).

Only the saturated model fits the data well.

Select a “quasi-symmetry” model since it fits reasonably well, but not by χ^2 test.

Model selection criterion: Quasi-Symmetry better than other non-saturated models.

Model Selection

Consider evaluation of an experiment. The purest form of testing a theory.

Treatment assignments are randomized, but unbalanced samples are possible.

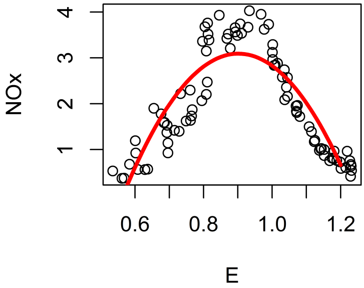
Using many covariates attempts to control these dimensions after the experiment is complete.

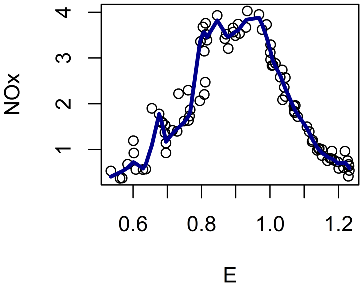
Many of these may not be *significant*, but controlling for them can be useful.

Model selection criterion: Robust results control for many observed covariates.

Model Selection

Consider another example. Here the goal is prediction for new data:





One produces precise estimates for these data. Another seems more plausible for a wide range of possible new data.

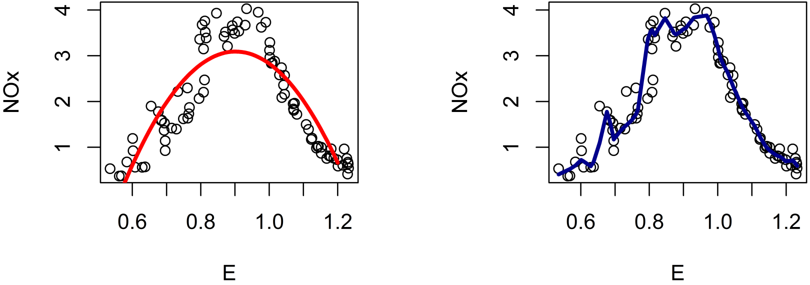
Model Selection

One method for selecting models based on their ability to predict for new data is **cross-validation**.

- K-fold validation.
  - Data are divided into K subsamples
  - Model selected and estimated with $K-1$ subsamples
  - Selected model tested with K^{th} subsample
  - Repeated across all K subsamples
- Random subsampling
  - Model may be fit to all data
  - Then tested across random subsamples
- Measure of error (residuals, misclassification) used to select final model
- Alternatively, average model estimates across subsamples

Model Selection

In this example, cross-validation may provide the means to evaluate the predictive power of these models.



Model selection criterion: Accuracy of predictions with new data

Modeling and P-Values

One approach to model selection is to remove insignificant predictors .

This may lead to problems.

- Involves the possibly strong assumption that $Pr(\beta = 0) = 1.0$.
- Problems the other direction as well: $Pr(\beta > 0) < 0.05$ when the true $\beta = 0$ (or irrelevant for current purpose.)

Modeling and P-Values

- Over many comparisons, we would expect some “false positives.”
- “Undisclosed modeling strategies” can lead to false positives.
- Ideal situation:
  1. State hypothesis before data collection
  2. Power analysis
  3. Collect data
  4. Test hypothesis
  5. Report results: positive or negative

P-Values

Ioannidis (2005) looks at the probability of false positives across studies within a field.

Produces seven corollaries of things that decrease the likelihood of true findings:

1. Smaller studies
2. Smaller effects
3. More relationships tested
4. Greater flexibility in design and analysis
5. Greater financial and other interests
6. “Hotter” the field

P-Values

Simmons, et al. (2011) looks at similar problem in psychology.

They propose some steps to limit the extent of the problem.

1. Stopping rules determined before data collection
2. Collect 20 cases per cell
3. List all variables collected in a study
4. Report all experimental conditions
5. Report results <u>with</u> deleted data
6. Report results with and without covariates

P-Values

Summary: More reporting needed of results, positive and negative. More disclosure of methods and modeling strategies.

P-values may be useful, but they aren't the only tool we have.

There isn't any magic in $p < 0.05$. Interesting things can happen above and below that line.

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

For evaluation the results of an experiment, we may be able to look at a two-way table (i.e. if the experiment is treatment-control). Recall this experiment from HW2:

Table: Evaluation of Impact of Including Plea for Help

Plea	Respond Yes	No
Yes	117	1,131
No	94	1,158

Evaluation of Experiment

Evaluation of Experiment

But, 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 earlier in the lecture, 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.

Model Purpose

Next time, we will look at examples of each of these purposes and how they impact model selection.

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

Preliminary Analyses

Step one: Inspect the data

- Frequencies, means, “5-number summary.”
  - Outliers?
  - Missing data?
- Bivariate analysis
  - For binary outcome, contingency tables or ANOVA.
  - Subgroup proportions
  - Empirical logit plots (linear on the logit scale?)
- Correlation matrix

Univariate Analyses

Example dataset: UMARA Impact Study

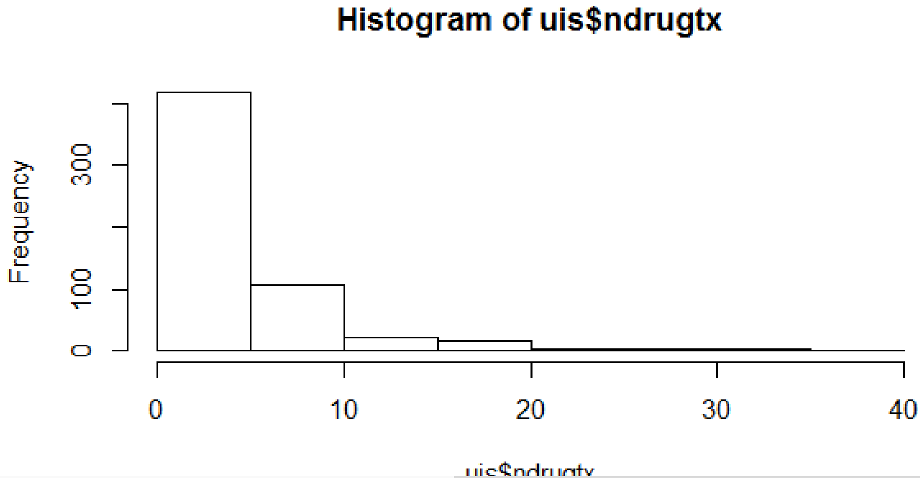
- 5-year evaluation of drug treatment programs
- Key question: Does duration of program effect outcomes?
- Two sites:
    1. A: 3- and 6-month programs
    2. B: 6- and 12-month programs

Univariate Analyses

```
age              beck              ivhx     ndrugtx          race
Min.   :20.00    Min.   : 0.00     1:223    Min.   : 0.000   0:430
1st Qu.:27.00    1st Qu.:10.00     2:109    1st Qu.: 1.000   1:145
Median :32.00    Median :17.00     3:243    Median : 3.000
Mean   :32.38    Mean   :17.37              Mean   : 4.543
3rd Qu.:37.00    3rd Qu.:23.00              3rd Qu.: 6.000
Max.   :56.00    Max.   :54.00              Max.   :40.000

treat             site              dfree
Min.   :0.0000    Min.   :0.0000    Min.   :0.0000
1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.0000
Median :0.0000    Median :0.0000    Median :0.0000
Mean   :0.4974    Mean   :0.3043    Mean   :0.2557
3rd Qu.:1.0000    3rd Qu.:1.0000    3rd Qu.:1.0000
Max.   :1.0000    Max.   :1.0000    Max.   :1.0000
```

Univariate Analyses



Feature Engineering

Data science and statistics are 80% data preparation. Need to make sure we have correct specification of the input variables.

- Imputing missing values
- Nonlinearities observed?
- Transformations may be suggested by scatter plots or logit plots
- Handling outliers – topcode, exclude, or leave?
- Binning of values
- Text variables – possible to code into categorical?
- Remove variables with near-zero variance
- Standardize the data (may be important depending upon the method)
- Data reduction techniques? PCA (later this semester)

Bivariate Analyses

Bivariate Analysis

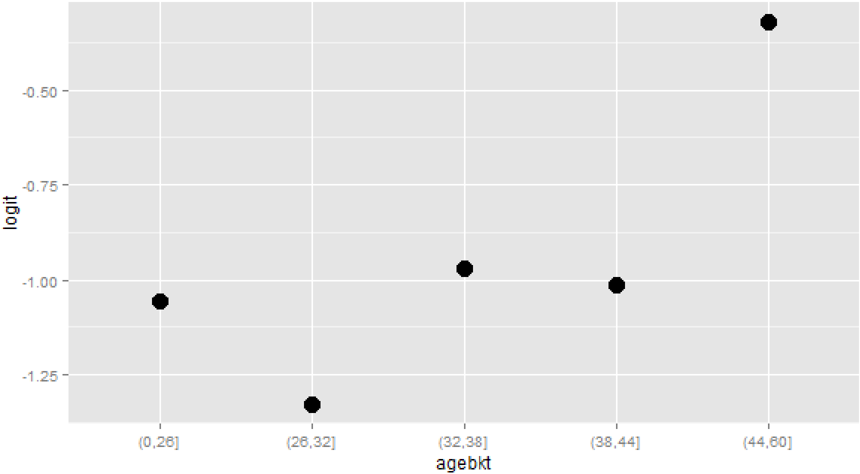
```
> glm.ivhx<-glm(dfree~ivhx,data=uis,family="binomial")
> summary(glm.ivhx)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6797     0.1417  -4.796 1.62e-06 ***
ivhx2        -0.4810     0.2657  -1.810 0.070242 .
ivhx3        -0.7748     0.2166  -3.578 0.000347 ***
---

Null deviance: 653.73  on 574  degrees of freedom
Residual deviance: 640.38  on 572  degrees of freedom
AIC: 646.38
```

Bivariate Analyses

Modeling Strategies



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.40541    0.55480  -4.336 1.45e-05 ***
age          0.05037    0.01732   2.908  0.00364 **
ivhx2       -0.60333    0.28725  -2.100  0.03570 *
ivhx3       -0.73272    0.25233  -2.904  0.00369 **
ndrugtx     -0.06151    0.02563  -2.400  0.01639 *
race1        0.22613    0.22334   1.012  0.31130
treat        0.44250    0.19929   2.220  0.02639 *
site         0.14858    0.21721   0.684  0.49394
```

Modeling Strategies

Consider transformation of some predictors.

Here, it seems that `age` and `ndrugtx` may need transformation.

```
uis$agebkt2<-as.factor(cut(uis$age,breaks=c(0,25,30,36,60)))
uis$ndrugtxbkt3[uis$ndrugtx< 4] <- 0
for (i in 1:575){
   if (uis$ndrugtx[i] >= 4) uis$ndrugtxbkt3[i] <- log(uis$ndrugtx[i])}
```

Modeling Strategies

Test models...

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.37151    0.32685  -4.196 2.71e-05 **
agebkt2(25,30]    0.32519    0.33974   0.957   0.3385
agebkt2(30,36]    0.80933    0.32511   2.489   0.0128 *
agebkt2(36,60]    0.60543    0.36503   1.659   0.0972 .
ivhx2            -0.50916    0.28619  -1.779   0.0752 .
ivhx3            -0.63949    0.25280  -2.530   0.0114 *
ndrugtx          -0.05654    0.02538  -2.228   0.0259 *
race1             0.25395    0.22460   1.131   0.2582
treat             0.43794    0.19931   2.197   0.0280 *
site              0.17036    0.21784   0.782   0.4342
```

Modeling Summary

Summary

- Models are useful summaries of data
- “Utility” is a function of the purpose
- Model selection is a process
- Judgments are made as part of the process