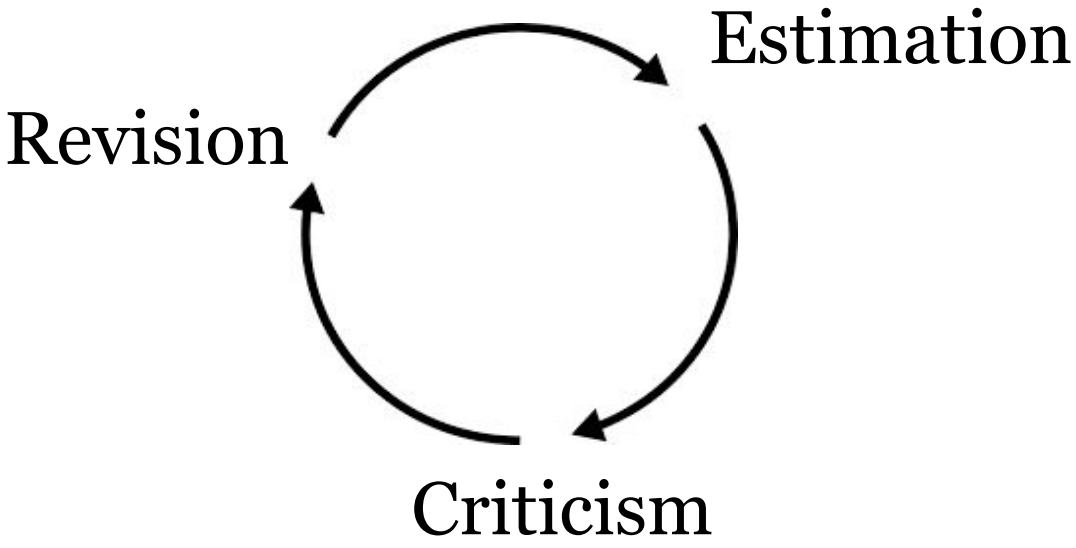
Data 102 Lecture 11:

Model checking for GLMs


Estimation
Revision
Criticism

Lecture overview

- GLMs from 10,000 feet
  - Review of supervised learning
  - GLMs vs. black-box models
- Model interpretations
  - Interpretations of GLMs when the model is correct
  - Interpretations of GLMs when the model is “wrong”
- What makes a good model?
  - Goodness-of-fit and generalization
  - Expanding and contracting GLMs
  - Goodness-of-fit checks for frequentist GLMs
  - Posterior predictive checks for Bayesian GLMs

GLMs from 10,000 feet

The 3 components of a GLM

A GLM comprises

- (**Systematic** component) A design matrix X and a coefficient vector β.
- (**Random** component) A noise distribution family *p(-|mean, other params)*
- A **link function** g so that $g(\mathbb{E}[Y_i|X_i]) = X_i^T\beta$

Another way to write:

$$y_i = g^{-1}(X_i^T\beta) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|X_i] = 0$$

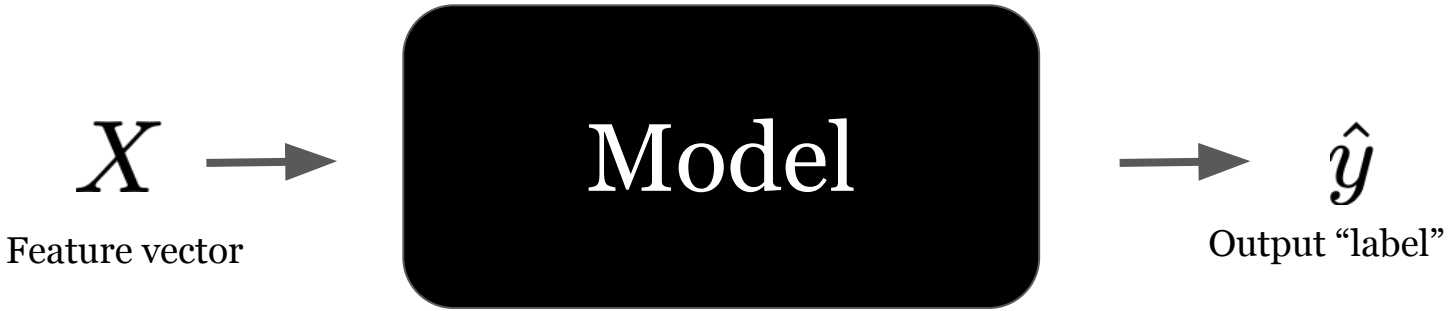
Using a GLM to predict

This is a **data generating model**, but we can fit it to any collection of data points by estimating the parameter β.

For any new data point with an unseen label, we may then predict the label via

$$\hat{y}_i = g^{-1}(X_i^T \hat{\beta})$$

GLM is a “grey box” supervised learning model

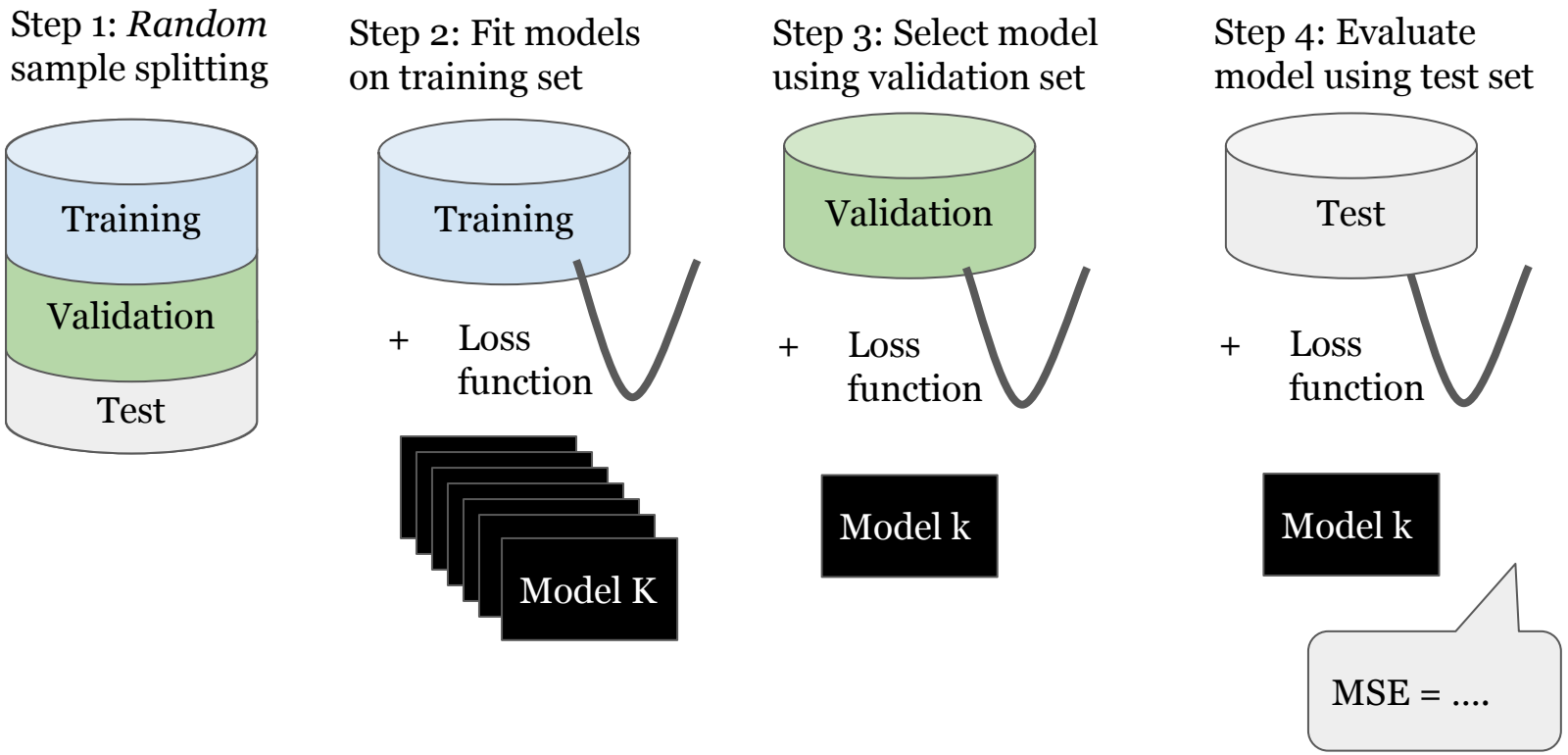


Neural network

Random forest

Gradient-boosted trees

The supervised learning pipeline



GLMs vs. black box models

Pros

- Can make use of subject matter knowledge to increase sample **efficiency** / extrapolate
- Models are naturally **interpretable** (in terms of the fitted vector of coefficients)
- **"Easier"** uncertainty quantification

Cons

- **Less flexible** than black box models, so may not fit the data well enough
- Requires **more trial and error** to fit a good model

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Notebook Demo 1: RF vs GLM

GLM interpretations

Interpreting the fitted coefficients $\hat{\beta}$

```
negbin_model = sm.GLM(
    ok_turbines.totals, sm.add_constant(ok_turbines.year),
    family=sm.families.NegativeBinomial()
)
negbin_results = negbin_model.fit()
print(negbin_results.summary())
```

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	NegativeBinomial	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-134.14
Date:	Wed, 17 Feb 2021	Deviance:	7.1483
Time:	12:51:51	Pearson chi2:	1.90
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.2059	0.544	7.725	0.000	3.139	5.273
year	0.2389	0.043	5.514	0.000	0.154	0.324

$$\log(\text{turbines\_built}) \approx 0.24 \cdot (\text{year} - 2000) + 4.2$$

Interpreting the fitted coefficients $\hat{\beta}$

$$\log(\text{turbines\_built}) \approx 0.24 \cdot (\text{year} - 2000) + 4.2$$

"Number of turbines built increases by roughly 24% each year on average"

In order to make sense of the interpretation, need to understand the approximation.

3 parts to the approximation

- Rounding error
- Noise in the response
- Estimation error

Classical frequentist view: A correct model

Assume that the model specification is **correct**, i.e. there is a true β_0 such that the data is generated from

$$y_i = g^{-1}(X_i^T \beta_0) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i | X_i] = 0$$

Then we have statistical theory that controls the **estimation error** $\hat{\beta} - \beta_0$

Asymptotically,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \Rightarrow \mathcal{N}(0, I_n(\beta_0)^{-1})$$

Classical frequentist view: A correct model

```
negbin_model = sm.GLM(
    ok_turbines.totals, sm.add_constant(ok_turbines.year),
    family=sm.families.NegativeBinomial()
)
negbin_results = negbin_model.fit()
print(negbin_results.summary())
```

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	NegativeBinomial	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-134.14
Date:	Wed, 17 Feb 2021	Deviance:	7.1483
Time:	12:51:51	Pearson chi2:	1.90
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.2059	0.544	7.725	0.000	3.139	5.273
year	0.2389	0.043	5.514	0.000	0.154	0.324

$$\beta_{0,year} = 0.24 \pm 0.08$$

Interpreting the fitted coefficients $\hat{\beta}$

$$\log(\text{turbines\_built}) \approx 0.24 \cdot (\text{year} - 2000) + 4.2$$

"Number of turbines built increases by roughly 16% to 32% each year on average"

In order to make sense of the interpretation, need to understand the approximation.

3 parts to the approximation

- Rounding error
- Noise in the response
- Estimation error

Model misspecification

Notebook demo 2

Model misspecification

Model misspecification means that the data generating distribution q(-) does not actually lie in the GLM family we are trying to fit

Under some assumptions, the fitted coefficients are an estimate of the “projected model”, i.e. the closest $p(-|\beta)$ to the the data generating distribution q(-)

Projected model may be meaningful, or not…

Bayesian view: Philosophically different, practically similar

Philosophical differences

- Don’t assume a true model β_0
- Instead, fitted model expresses our posterior belief, contingent on our assumptions
- Allows for assumptions not revealed in the data
- A “poor fit” does not necessarily mean that the model is useless

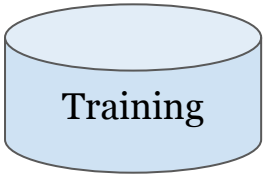
Practical similarities

- If model does not fit the data well, then we should question it
- Need to be able to diagnose whether the model is “good”

What makes a good model?

Goodness-of-fit and generalization

Goodness-of-fit: How well the model fits the training data



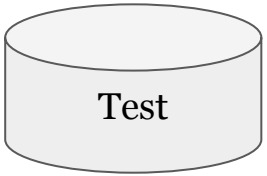
Training loss

Visual inspection

Other statistics:
deviance, chi-squared…

Model

Generalization: How well the model fits the test data



Test loss

Expanding and contracting models

Two ways to expand the model

- Adding new features:

$$x_1, x_2 \rightarrow x_1, x_2, x_1^2, x_2^2, x_1 x_2, \ldots$$

- Making the noise model more flexible:
  - Gaussian -> t-distribution
  - Poisson -> negative binomial

Conversely, can contract the model by dropping features, etc.

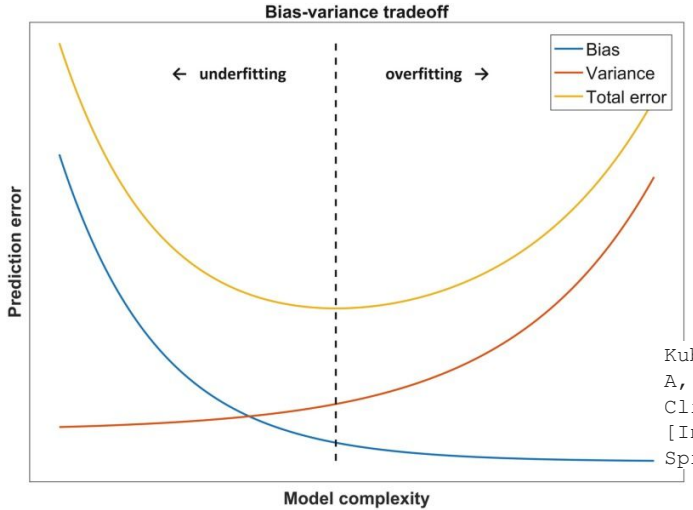
Need to balance underfitting vs overfitting

The more we expand the model…

… the better it fits the data (less bias)

… the better it fits noise in the data (more variance)

Get the bias-variance tradeoff



Kubben P, Dumontier M, Dekker A, editors. Fundamentals of Clinical Data Science [Internet]. Cham (CH): Springer; 2019.

Goodness-of-fit for frequentist GLM

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	NegativeBinomial	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-134.14
Date:	Fri, 19 Feb 2021	Deviance:	7.1483
Time:	13:45:05	Pearson chi2:	1.90
No. Iterations:	11		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	4.2059	0.544	7.725	0.000	3.139	5.273
year	0.2389	0.043	5.514	0.000	0.154	0.324

Neg of training loss

Statistics for testing hypothesis that "model fits data well"

Goodness-of-fit for frequentist GLM

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	NegativeBinomial	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-134.14
Date:	Fri, 19 Feb 2021	Deviance:	7.1483
Time:	13:45:05	Pearson chi2:	1.90
No. Iterations:	11		
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	Poisson	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-755.42
Date:	Fri, 19 Feb 2021	Deviance:	1366.3
Time:	13:45:05	Pearson chi2:	1.20e+03
No. Iterations:	5		
Covariance Type:	nonrobust		

Goodness-of-fit for frequentist GLM

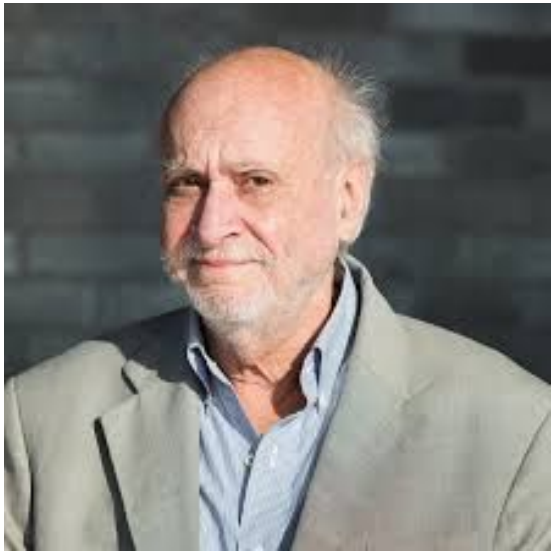
Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	NegativeBinomial	Df Model:	1
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-134.14
Date:	Fri, 19 Feb 2021	Deviance:	7.1483
Time:	13:45:05	Pearson chi2:	1.90
No. Iterations:	11		
Covariance Type:	nonrobust		

Generalized Linear Model Regression Results

Dep. Variable:	totals	No. Observations:	17
Model:	GLM	Df Residuals:	15
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	1.1810
Method:	IRLS	Log-Likelihood:	-24.472
Date:	Fri, 19 Feb 2021	Deviance:	17.716
Time:	13:45:05	Pearson chi2:	17.7
No. Iterations:	3		
Covariance Type:	nonrobust		

Goodness-of-fit for Bayesian GLM: Posterior predictive checks



Donald Rubin

Andrew Gelman

Goodness-of-fit for Bayesian GLM: Posterior predictive checks

Basic principle:

"Given observed data, X_{obs}, what would we expect to see in hypothetical replications of the study that generated X_{obs}? Intuitively, if the model specifications are appropriate, we would expect to see something similar to what we saw this time, at least similar in 'relevant ways'."

Donald Rubin (1984)

Because we are being Bayesian, we do replications conditioned on having seen the data. I.e. we use the posterior predictive distribution

$$p(y_{rep}|X, y_{obs}) = \int p(y_{rep}|X, \beta)p(\beta|X, y_{obs})d\beta$$

Goodness-of-fit for Bayesian GLM: Posterior predictive checks

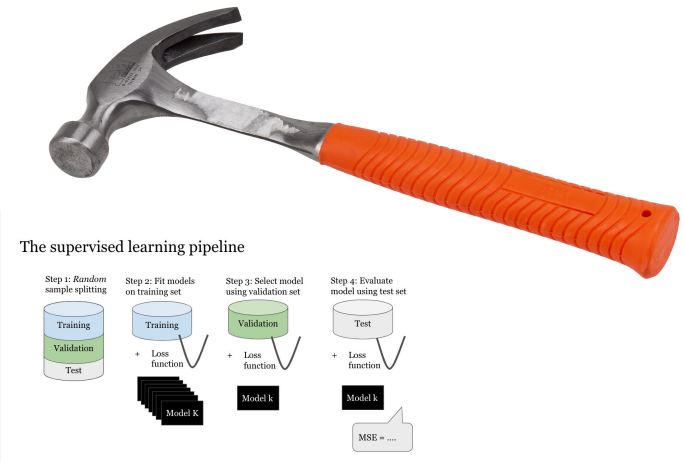
Algorithm:

1. Simulate $\beta_{pos} \sim p(\beta|X,y)$
2. Simulate $y_{rep} \sim p(y|X,\beta_{pos})$
3. Repeat B times

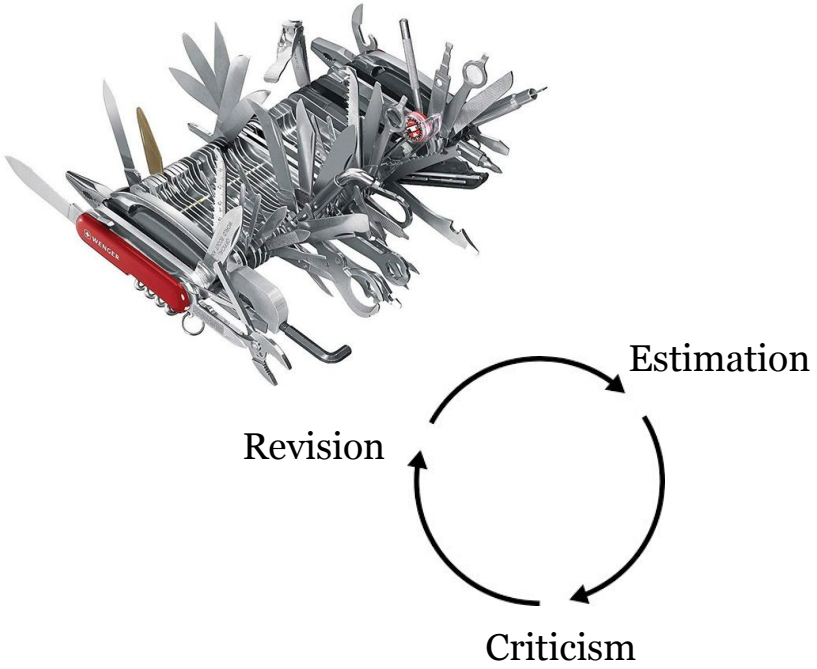
Notebook demo

Summary

Beginner data scientist



Intermediate data scientist



Resources

GLMs

- Eduardo García Portugués's notes: https://bookdown.org/egarpor/PM-UC3M/

Posterior predictive checks:

- Jeffrey B. Arnold's notes: https://jrnold.github.io/bayesian_notes/
- David Blei's notes: https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/ppc.pdf