Lecture 10. Modeling Process and Model Diagnostics

Nan Ye

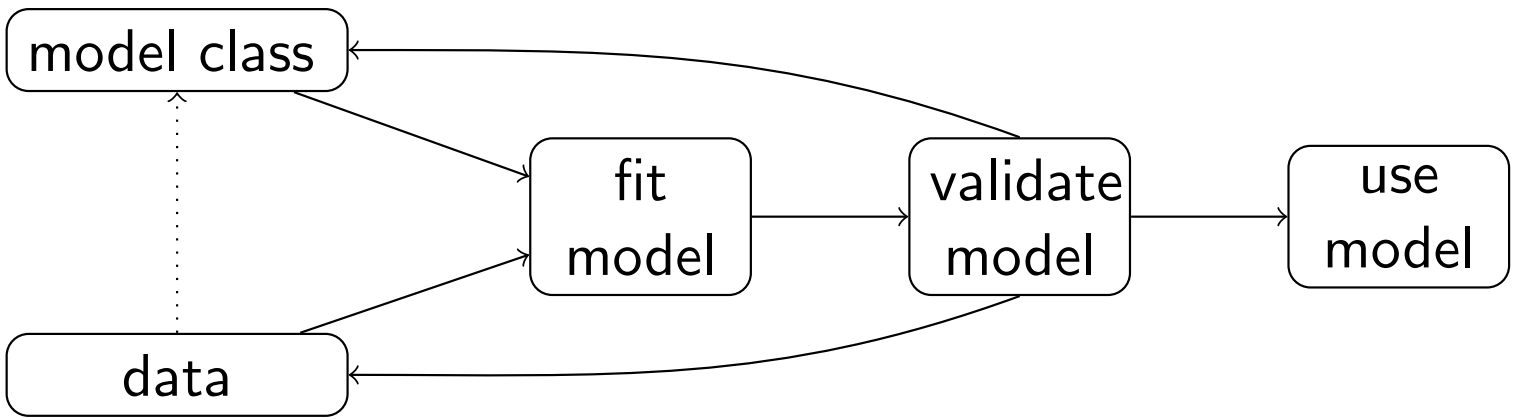
School of Mathematics and Physics
University of Queensland

This Lecture

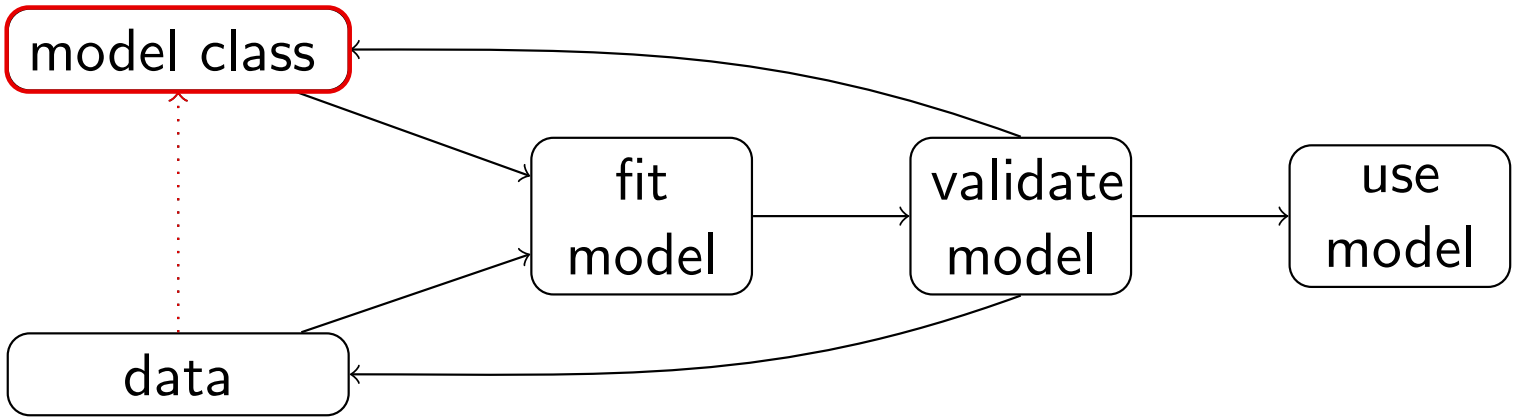
- Modeling process
- Goodness of fit
- Residuals

Modeling Process

Some key modeling activities

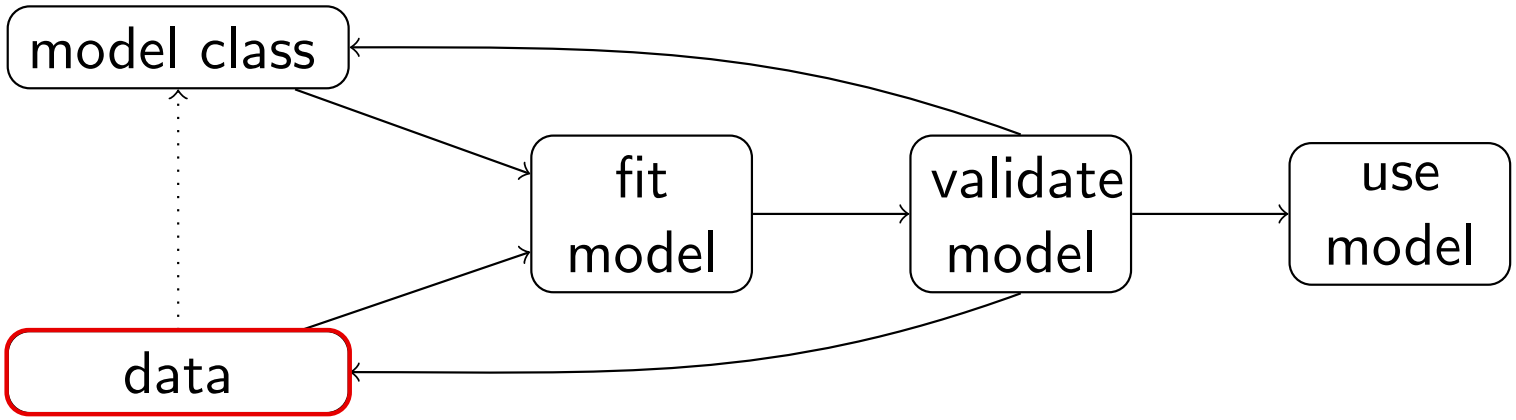


Some key modelling activities



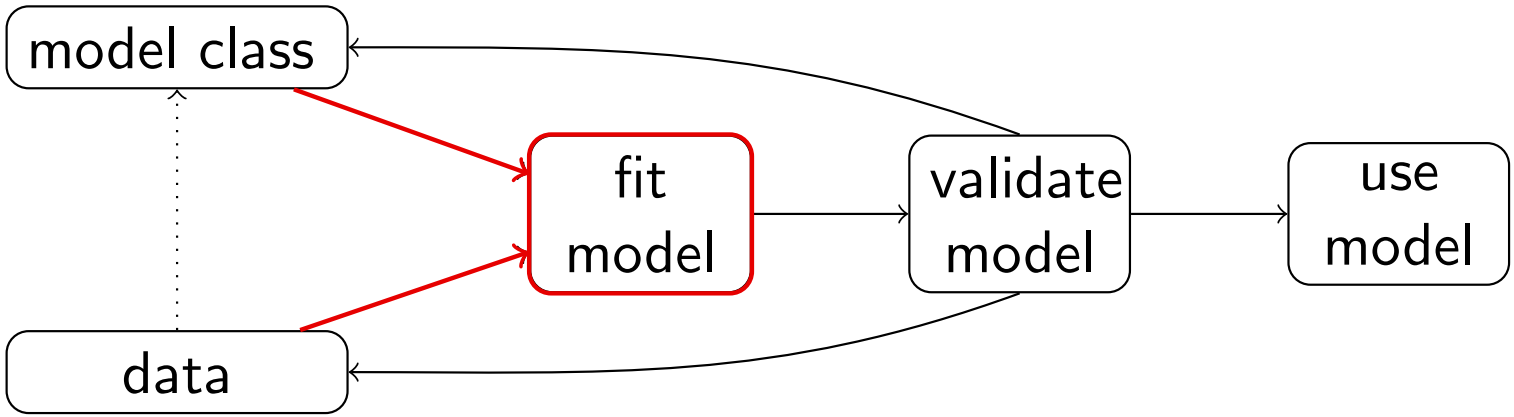
- The choice of a model class is often driven by many factors, including data characteristics, expressiveness, interpretability, computational efficiency...
- If predictive performance (expressiveness) is the main concern
  - try deep neural networks for image/text/speech data.
  - try random forests when high-level features are available.
- GLMs can be good in terms of interpretability.

Some key modelling activities



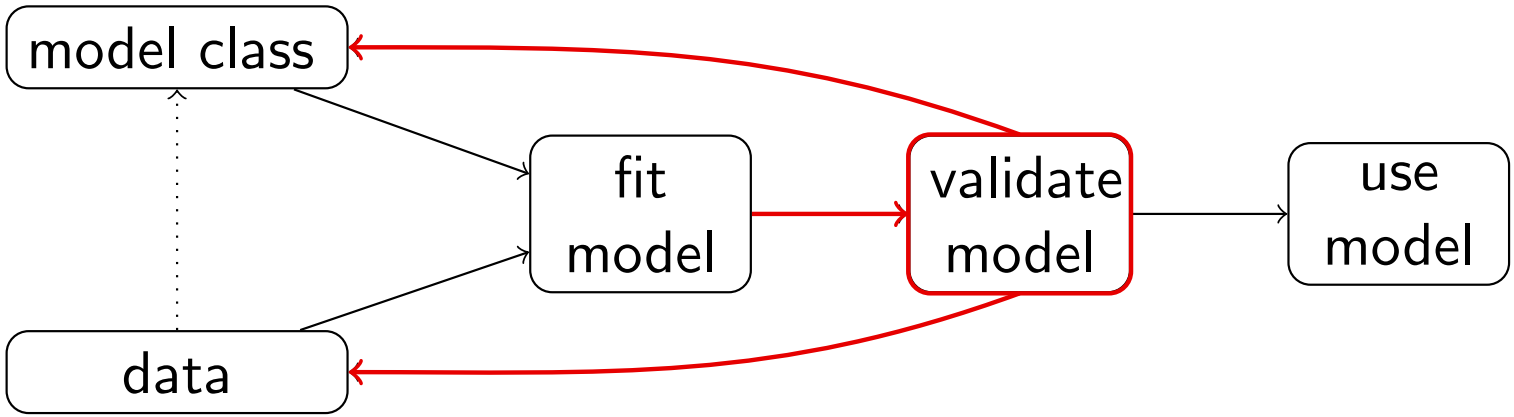
- More data is often better.
- With right features, even simple models can work well.
- Exploratory analysis can suggest useful features and models.

Some key modelling activities



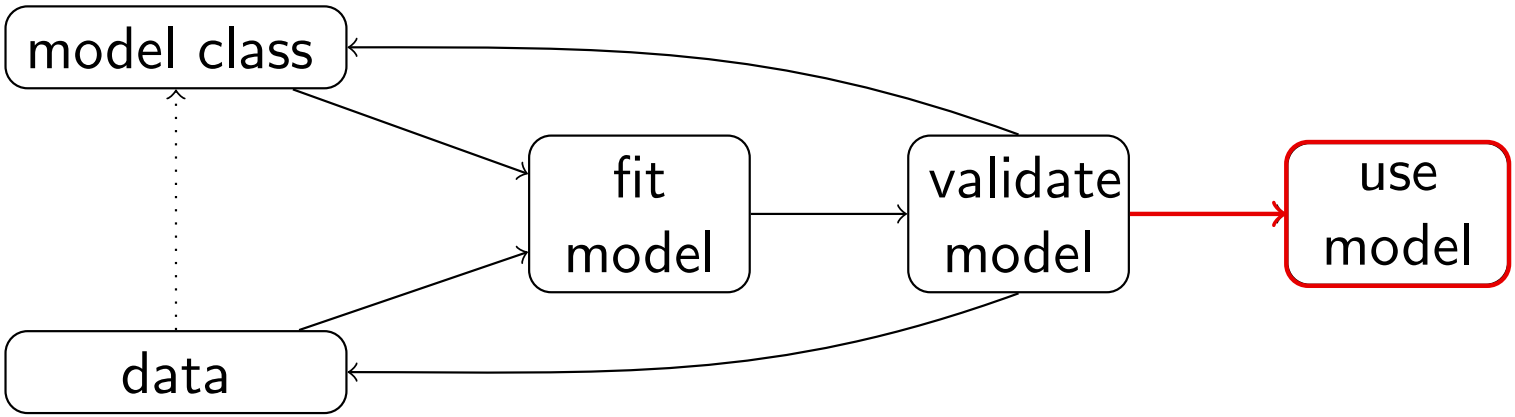
- Fitting is usually formulated as an optimization problem.
- MLE is often used to learn a statistical model.
- If predictive performance is the main concern, optimize the performance measure directly.
- Sophisticated optimization algorithms may be needed.
  - For GLM, Fisher scoring often works well for MLE.

Some key modelling activities



- Check model assumption
  - Check goodness of fit, residual plot et al on training set.
  - A good fit on the training set may mean overfitting.
- Check predictive performance
  - Check cross-validation score, validation set performance.
- Reconsider model class or data if checks are not satisfactory.

Some key modelling activities



- After checks on the model, the model can then be used to make predictions or draw conclusions (such as significance of variables, variable importance).

Goodness of Fit

Deviance

- Null model
  - Includes only the intercept term in the GLM.
  - Variation in y's comes from the random component only.
- Full model (saturated model)
  - Fit an exponential family distribution for each example.
  - The exponential family distribution for $(\mathbf{x}_i, y_i)$ is $f(y \mid \text{mean} = y_i)$.
  - Variation in y's comes from the systematic component only.
- GLM
  - Summarizes data with a few parameters.
  - The exponential family distribution for $(\mathbf{x}_i, y_i)$ is $f(y \mid \text{mean} = \hat{\mu}_i)$, where $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^\top \hat{\beta})$.

- Scaled deviance

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i \ln f(y_i \mid \text{mean} = y_i) - 2 \sum_i \ln f(y_i \mid \text{mean} = \hat{\mu}_i)$$

  This is twice the difference between log-likelihood of the full model and the maximum log-likelihood achievable for the GLM.

- Deviance

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = b(\phi) D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}).$$

  Deviance is thus scaled deviance with the nuisance parameter removed.

Example. Gaussian
The scaled deviance is

$$
\begin{aligned}
& D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) \\
&= 2 \sum_i \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - y_i)^2}{2\sigma^2} \right) - 2 \sum_i \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \hat{\mu}_i)^2}{2\sigma^2} \right) \\
&= \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}.
\end{aligned}
$$

The deviance is

$$
D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sigma^2 D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_i (y_i - \hat{\mu}_i)^2.
$$

distribution	deviance
normal	$\sum(y-\hat{\mu})^2$
Poisson	$2\sum(y \ln \frac{y}{\hat{\mu}} - (y-\hat{\mu}))$
binomial	$2\sum(y \ln \frac{y}{\hat{\mu}} + (m-y) \ln \frac{m-y}{m-\hat{\mu}})$
Gamma	$2\sum(-\ln \frac{y}{\hat{\mu}} + \frac{y-\hat{\mu}}{\hat{\mu}})$
inverse Gaussian	$\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y)$

Recall

```
> logLik(fit.ig.inv)
'log Lik.' -25.33805 (df=5)
> logLik(fit.ig.invquad)
'log Lik.' -50.26075 (df=5)
> logLik(fit.ig.log)
'log Lik.' -45.55859 (df=5)
```

Inverse Gaussian regression with inverse link has the best fit (much better than the other two).

```
> summary(fit.ig.inv)
    Null deviance: 0.24788404  on 17  degrees of freedom
Residual deviance: 0.00097459  on 14  degrees of freedom
> summary(fit.ig.invquad)
    Null deviance: 0.24788  on 17  degrees of freedom
Residual deviance: 0.01554  on 14  degrees of freedom
> summary(fit.ig.log)
    Null deviance: 0.2478840  on 17  degrees of freedom
Residual deviance: 0.0092164  on 14  degrees of freedom
```

- Inverse link has best fit.
- Same conclusion as obtained by looking at the log-likelihoods.
- summary function provides a comparison with the full model and null model.

Generalized Pearson X^2 statistic

- Recall: $\text{var}(Y) = b(\phi)A''(\eta)$ for a natural exponential family.
- $\text{var}(Y)/b(\phi)$ depends only on η, and thus only on μ.
- Often, $\text{var}(Y)/b(\phi)$ is called the variance function $V(\mu)$.
- Pearson X^2 statistic is

$$X^2 = \sum (y - \hat{\mu})^2 / V(\hat{\mu}),$$

  where $V(\hat{\mu})$ is the estimated variance function.
- The scaled version is $X^2/b(\phi)$.

distribution	X^2
normal	$\sum(y-\hat{\mu})^2$
Poisson	$\sum(y-\hat{\mu})^2/\hat{\mu}$
binomial	$\sum \frac{(y-\hat{\mu})^2}{\hat{\mu}(1-\hat{\mu})}$
Gamma	$\sum(y-\hat{\mu})^2)/\hat{\mu}^2$
inverse Gaussian	$\sum(y-\hat{\mu})^2/\hat{\mu}^3$

Asymptotic distribution

- If the model is true, then the scaled deviance or the scaled Pearson X^2 statistic asymptotically follows χ^2_{n-p}, where n is the number of examples, and p is the number of parameters estimated.
- In principle, this can be used to test goodness of fit, but this does not really work well.
- A test on the scaled deviance or the scaled Pearson X^2 statistic cannot be used to justify that the model is correct.

Residuals

Response residual

- This is the difference between the output and fitted mean

$$r_R = y - \hat{\mu}.$$

- Measures deviation from systematic effect on an absolute scale.

Pearson residuals

- This is the normalized response residual

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

- Constant variance and mean zero if model is correct.

distribution	Pearson residual
normal	$y - \hat{\mu}$
Poisson	$(y - \hat{\mu})/\sqrt{\hat{\mu}}$
binomial	$(y - \hat{\mu})/\sqrt{\hat{\mu}(1 - \hat{\mu})}$
Gamma	$(y - \hat{\mu})/\hat{\mu}$
inverse Gaussian	$(y - \hat{\mu})/\hat{\mu}^{3/2}$

Working residuals

- Recall: in the IRLS interpretation of Fisher scoring, at each iteration we try to fit the *adjusted response* vector

$$\mathbf{z} = \mathbf{G}\mathbf{y} - \mathbf{G}\mu + \mathbf{X}\beta,$$

  where $\mathbf{G} = \text{diag}(g'(\mu_1), \ldots, g'(\mu_n))$.
- The adjusted response for $(\mathbf{x}, y)$ is

$$z = g'(\mu)(y - \mu) + \mathbf{x}^\top \beta.$$

- The working residual is

$$r_W = z - \xi = (y - \hat{\mu})g'(\mu) = (y - \hat{\mu})\frac{\partial \xi}{\partial \mu}|_{\mu=\hat{\mu}},$$

  where $\xi = \mathbf{x}^\top \beta$.

Deviance residuals

- This is the signed contribution of each example to the deviance

$$r_D = \text{sign}(y - \hat{\mu})\sqrt{d},$$

  where $\sum_i d_i = D$.
- Closer to a normal distribution (less skewed) than Pearson residuals.
- Often better for spotting outliers.

distribution	deviance residual
normal	$y - \hat{\mu}$
Poisson	$\text{sign}(y - \hat{\mu})\sqrt{2(y \ln \frac{y}{\hat{\mu}} - (y - \hat{\mu}))}$
binomial	$\text{sign}(y - \hat{\mu})\sqrt{2(y \ln \frac{y}{\hat{\mu}} + (m - y) \ln \frac{m-y}{m-\hat{\mu}})}$
Gamma	$\text{sign}(y - \hat{\mu})\sqrt{2(-\ln \frac{y}{\hat{\mu}} + \frac{y-\hat{\mu}}{\hat{\mu}})}$
inverse Gaussian	$(y - \hat{\mu})/\hat{\mu}\sqrt{y}$

Computing residuals in R

```
> resid(fit.ig.inv, 'response')
> resid(fit.ig.inv, 'pearson')
> resid(fit.ig.inv, 'working')
> resid(fit.ig.inv, 'deviance')
```

What You Need to Know

- Modeling process
- Goodness of fit: deviance and Pearson X^2 statistic
- Response, working, Pearson, and deviance residuals