Prediction and model assessment

(MATH10093)

Some notation

- For a random variable Y, we write $p_Y(y)$, $y \in D$ for the probability mass function for a discrete random variable or the probability density function for a continuous random variable, taking values in some set or region D.
- In this part of the course, we will often have more than one distribution relating to the same outcome y. To keep track of which probability measure is involved in each expression, we will use letters F and G, and sometimes F' to denote different distributions.
- The full notation for an expectation of $h(Y)$ when Y has distribution F is denoted $\mathsf{E}_{Y\sim F}[h(Y)]$, and the probability mass/density function (pmf/pdf) is $p_F(y)$.
- When it is clear which variable is involved, we may abbreviate to $\mathsf{E}_F[h(Y)]$, and especially in Bayesian contexts, we won't always distinguish between a random variable Y and outcome y, but instead use lower case y, as in $\mathsf{E}_{y\sim F}[h(y)]$.
- For convenience, the letter used to identify a distribution will also be used to denote the cumulative distribution function (cdf), so that $F(y) = \mathsf{P}_{Y\sim F}(Y \leq y)$.

Expectation and variance

$$\mathsf{E}_F[h(Y)] = \sum_{y \in D} h(y)\, p_F(y), \quad \text{discrete outcomes } y \in D,$$

$$\mathsf{E}_F[h(Y)] = \int_D h(y) p_F(y)\, \mathrm{d}y, \quad \text{continuous outcomes } y \in D$$

$$\mathsf{Var}_F(Y) = \mathsf{E}_F\left\{[Y - \mathsf{E}_F(Y)]^2\right\} = \mathsf{E}_F(Y^2) - \mathsf{E}_F(Y)^2$$

(Often, we use the tower property trick for variances instead of the plain definition)

Prediction and proper scoring rules

- When using numerical estimation methods subject to both numerical precision errors, random data collection variation, and methodological approximations, it's useful to be able to assess the end result in a way that's not tied to a particular method or model.
- One approach: *split* the data into *observations for estimation* and *test* data: $\mathcal{Y}_{\text{obs}}$ and $\mathcal{Y}_{\text{test}}$.
- Estimate the model parameters using the *estimation* data $\mathcal{Y}_{\text{obs}}$.
- Assess how good the estimated model is at predicting the values of the *test* data $\mathcal{Y}_{\text{test}}$.
- The most common assessment is to compare *point predictions* with their corresponding actual value in the test data:
  Squared error $= (y_{\text{test}} - \widehat{y})^2$.
- The family of *proper scores* or *proper scoring rules* helps to keep comparisons fair; they ensure that predictions that don't match true variability do not get a lower score expectation than the true distribution.
- To assess methodology, it's often useful to use *simulated* data, so that we know the true model. If the method is able to come close to the true values, we are more confident that it will also work on data where we don't know the true model.

Forecasting and prediction

- The term *forecasting* typically means doing a *prediction* of future events or values, e.g. for weather forecasts
- Based on some statistical model and observed data, we typically construct a *point estimate* that is our best guess of the future value. Ideally, we also compute some measure of *uncertainty* about how large we expect the error to be, i.e. the difference between the point estimate and the true future value.
- In statistical terminology, this process is called *prediction*, and we seek useful *predictive distributions* that encode our knowledge from a statistical model and previously observed data into a representative distribution of possible future data values.
- Note: In Bayesian statistics, *prediction* (distributions and prediction intervals) can apply to *any* quantity that has not yet been observed. In frequentist statistics, fixed but unknown parameter values are instead associated with *confidence intervals*, and *prediction intervals* are reserved for observable random quantities.
- We will gloss over the differences between frequentist and Bayesian approaches, and focus on prediction of observable data values.

Bayesian prediction distributions

Consider a (Bayesian) hierarchical model structure

$$\begin{aligned} \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}) \\ \boldsymbol{y}|\boldsymbol{\theta} &\sim p(\boldsymbol{y} \mid \boldsymbol{\theta}) \end{aligned}$$

with parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p\}$ and observations/outcomes $\boldsymbol{y} = \{y_1, \ldots, y_n\}$.

Posterior prediction

Given observations $\boldsymbol{y}$, we want to *predict* new outcomes y'.
The posterior uncertainty about $\boldsymbol{\theta}$ is captured by the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{y}$, with density $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$.
The predictive uncertainty for a single outcome y' can be obtained from the predictive density

$$p_{y'|\boldsymbol{y}}(y') = \int_D p_{y'|\boldsymbol{\theta}}(y') p_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

We will identify the predictive distribution for y' with the CDF F,
$F(x) = \mathsf{P}_{y' \sim F}(y' \leq x) = \int_{-\infty}^{x} p_{y'|y}(u) \, \mathrm{d}u$.
When we're just considering a given prediction F, we'll drop the $\cdot'$ from y'.

Approximate frequentist prediction distributions

- From asymptotic likelihood theory we know that, when $\boldsymbol{\theta}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, then approximately, $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{true}} \sim \mathsf{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_\theta)$, where $\widehat{\boldsymbol{\Sigma}}_\theta^{-1}$ can be estimated by $H(\widehat{\boldsymbol{\theta}})$ (log-likelihood Hessian from Lecture 2, the *observed Fisher information*)
- With a slight abuse of frequentist notation, we will represent the estimation uncertainty by a distribution of *potential* parameter values: $\boldsymbol{\theta} \sim \mathsf{N}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}_\theta)$, where $\widehat{\theta}$ is now treated as a fixed value, determined by the oberved $\boldsymbol{y}$ values. We write the density for the potential parameter values as $p_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})$, just as in the Bayesian case.
- We can then use the same notation as in the Bayesian case to define an approximate predictive distribution $p_F(y') = p_{y'|\boldsymbol{y}}(y')$
- A common further approximation is to ignore the uncertainty about $\boldsymbol{\theta}$, and instead define $p_F(y') = p_{y'|\widehat{\boldsymbol{\theta}}}(y')$, the *plug-in* estimator of the observation distribution.

Example: Non-constant regression model variance

- Let $y \sim \mathsf{N}\left[\boldsymbol{z}_E^\top \boldsymbol{\theta}, \exp(\boldsymbol{z}_V^\top \boldsymbol{\theta})\right]$, i.e. the expectation has a linear model, and the variance has a log-linear model, where the same parameters could potentially influence both the expectation and the variance.
- With the $\boldsymbol{z}_{\cdot}$ vectors for each observation stored as rows in two matrices $\boldsymbol{Z}_E$ and $\boldsymbol{Z}_V$, the vector of observation expectations can be written as $\mathsf{E}_{y|\theta}(\boldsymbol{y}) = \boldsymbol{Z}_E \boldsymbol{\theta}$, and similarly for the log-variances.
- From numerical optimisation of the log-likelihood and asymptotic likelihood theory we obtain $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}_\theta$ for the uncertainty distribution $\boldsymbol{\theta}|\boldsymbol{y} \sim \mathsf{N}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}_\theta)$
- We want the predictive expectation and variance for given $\boldsymbol{z}_E$ and $\boldsymbol{z}_V$ for a new observation y'.
- Numerical examples will use the model $\boldsymbol{z}_E = \begin{bmatrix} 1 & x & 0 & 0 \end{bmatrix}$, $\boldsymbol{z}_V = \begin{bmatrix} 0 & 0 & 1 & x \end{bmatrix}$, where x is a covariate with different value for each observation.

Example: Predictive distribution

- Recall the *tower property* ($\mathsf{E}_A(A) = \mathsf{E}_B[\mathsf{E}_{A|B}(A)]$), which gives

  $$\mu_F = \mathsf{E}_F(y') = \boldsymbol{z}_E^\top \widehat{\boldsymbol{\theta}}, \qquad \sigma_F^2 = \mathsf{Var}_F(y) = \mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\exp\left(\boldsymbol{z}_V \boldsymbol{\theta}\right)\right] + \mathsf{Var}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(\boldsymbol{z}_E^\top \boldsymbol{\theta}\right).$$

  The second term of the variance is

  $$\mathsf{Var}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(\boldsymbol{z}_E^\top \boldsymbol{\theta}\right) = \mathsf{Cov}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(\boldsymbol{z}_E^\top \boldsymbol{\theta}, \boldsymbol{z}_E^\top \boldsymbol{\theta}\right) = \boldsymbol{z}_E^\top \mathsf{Cov}_{\boldsymbol{\theta}|\boldsymbol{y}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}\right) \boldsymbol{z}_E = \boldsymbol{z}_E^\top \widehat{\boldsymbol{\Sigma}}_\theta \boldsymbol{z}_E.$$

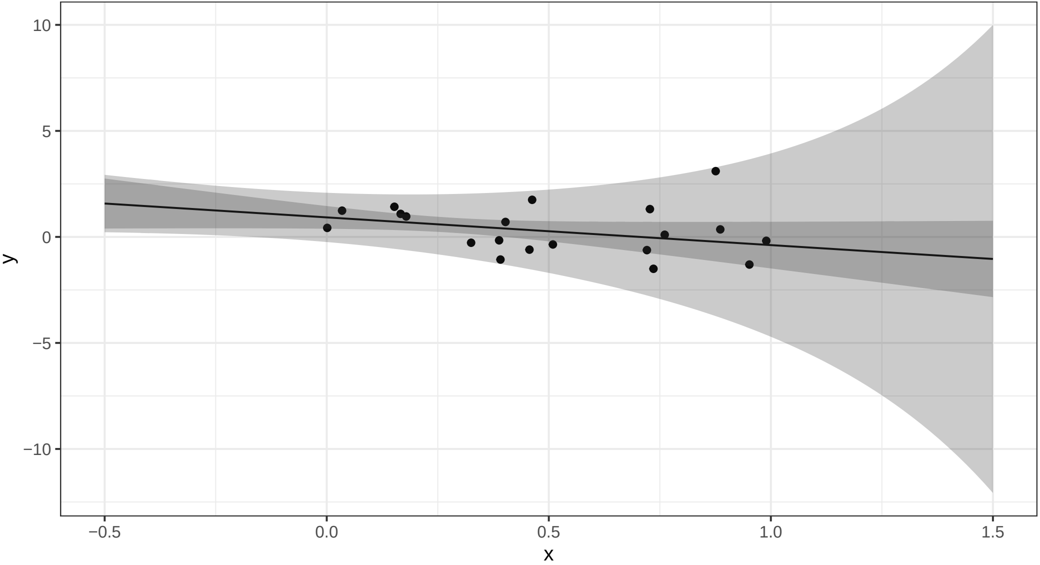
  For the first term, we use a result that says that if $x \sim \mathsf{N}(\mu, \sigma^2)$, then $\mathsf{E}(\mathrm{e}^x) = \mathrm{e}^{\mu + \sigma^2/2}$:

  $$\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}[\exp(\boldsymbol{z}_V^\top \boldsymbol{\theta})] = \exp\left(\boldsymbol{z}_V^\top \widehat{\boldsymbol{\theta}} + \boldsymbol{z}_V^\top \widehat{\boldsymbol{\Sigma}}_\theta \boldsymbol{z}_V / 2\right).$$

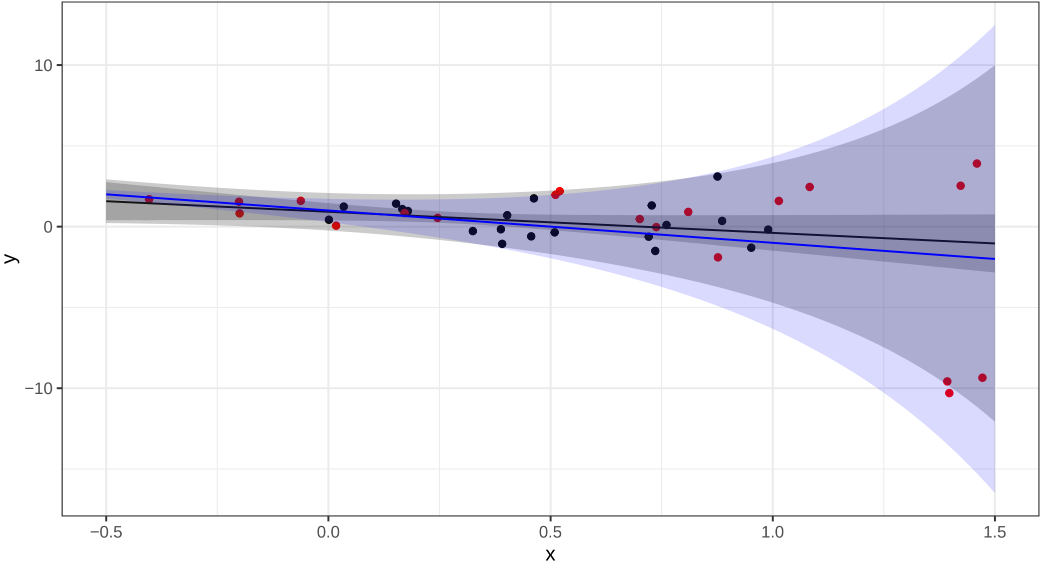
  Combining the results, we get the predictive variance as

  $$\sigma_F^2 = \mathsf{Var}_F(y) = \exp\left(\boldsymbol{z}_V^\top \widehat{\boldsymbol{\theta}} + \boldsymbol{z}_V^\top \widehat{\boldsymbol{\Sigma}}_\theta \boldsymbol{z}_V / 2\right) + \boldsymbol{z}_E^\top \widehat{\boldsymbol{\Sigma}}_\theta \boldsymbol{z}_E.$$

Estimation data, point estimates, confidence and prediction intervals



Add the true model&predictions, and some test data



Scores

- We want to quantify how well our predictions represent the test data.
- We define *scores* $S(F, y)$ that in some way measure how well the prediction F matched the actual value, y.
- The scores defined here are *negatively oriented*, meaning that the *lower the score, the better*.

Squared errors and log-likelihood scores

- Squared Error (SE): $S_{\mathsf{SE}}(F, y) = (y - \widehat{y}_F)^2$,
  where $\widehat{y}_F$ is a point estimate under F, e.g. the expectation μ_F.
- Logarithmic/Ignorance score (LOG/IGN): $S_{\mathsf{LOG}}(F, y) = -\log p_F(y)$,
  where $p_F(\cdot)$ is the predictive probability density function.
- Dawid-Sebastiani (DS): $S_{\mathsf{DS}}(F, y) = \frac{(y-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)$.

Proper scoring rules

- What functions of the predictive distributions are useful scores?
- We want to reward accurate (unbiased) and precise (small variance) predictions, but not at the expense of understating true uncertainty.
- First, we define the expectation of a score under a true distribution G as

$$S(F,G) = \mathsf{E}_{y\sim G}[S(F,y)]$$

Proper scores/scoring rules

A negatively oriented score is *proper* if it fulfils

$$S(F,G) \geq S(G,G).$$

A proper score that has equality of the expectations *only* when F and G are the same, $F(\cdot) \equiv G(\cdot)$, is said to be *strictly proper*.

The practical interpretation of this is that a proper score does not reward cheating; stating a lower (or higher) forecast/prediction uncertainty will not, on average, give a better score than matching the true variability.

Proper scores

$$\begin{aligned}
S_{\mathsf{SE}}(F,G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{SE}}(F,y)] = \mathsf{E}_{y\sim G}[(y-\mu_F)^2] = \mathsf{E}_{y\sim G}[(y-\mu_G+\mu_G-\mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y-\mu_G)^2 + 2(y-\mu_G)(\mu_G-\mu_F) + (\mu_G-\mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y-\mu_G)^2] + 2(\mu_G-\mu_F)\mathsf{E}_{y\sim G}[y-\mu_G] + (\mu_G-\mu_F)^2 \\
&= \sigma_G^2 + (\mu_G-\mu_F)^2
\end{aligned}$$

This is minimised when $\mu_F = \mu_G$. Therefore $S_{\mathsf{SE}}(F,G) \geq S_{\mathsf{SE}}(G,G) = \sigma_G^2$, so the score is proper. Is it strictly proper?

$$\begin{aligned}
S_{\mathsf{DS}}(F,G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{DS}}(F,y)] = \frac{\mathsf{E}_{y\sim G}[(y-\mu_F)^2]}{\sigma_F^2} + \log(\sigma_F^2) \\
&= \frac{\sigma_G^2 + (\mu_G-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)
\end{aligned}$$

This is minimised when $\mu_F = \mu_G$ and $\sigma_F = \sigma_G$. Therefore $S_{\mathsf{DS}}(F,G) \geq S_{\mathsf{DS}}(G,G) = 1 + \log(\sigma_G^2)$, so the score is proper. Is it strictly proper?

Absolute error and CRPS

Absolute error and Continuous Ranked Probability Score

- Absolute Error (AE): $S_{\mathsf{AE}}(F, y) = |y - \widehat{y}_F|$, where $\widehat{y}_F$ is a point estimate under F, e.g. the *median* $F^{-1}(1/2)$.
- CRPS: $S_{\mathsf{CRPS}}(F, y) = \int_{-\infty}^{\infty} \left[\mathbb{I}(y \leq x) - F(x)\right]^2 \, \mathrm{d}x$

How good are prediction interval procedures?

Tradeoffs for prediction intervals

Desired properties for methods generating prediction intervals (PIs) for a quantity Y:

1. Appropriate *coverage* under the true distribution, G: $\mathsf{P}_{Y\sim G}(Y \in \mathsf{PI}_F) \geq 1-\alpha$
2. Narrow intervals

Note: For confidence intervals similar properties are desired, but then the F are random.

- A wide prediction F helps with 1 but makes 2 difficult
- A narrow prediction F helps with 2 but makes 1 difficult

A proper score for interval predictions

The *Interval Score* For a PI (L_F, U_F) is defined by

$$S_{\mathsf{INT}}(F, y) = U_F - L_F + \frac{2}{\alpha}(L_F - y)\mathbb{I}(y < L_F) + \frac{2}{\alpha}(y - U_F)\mathbb{I}(y > U_F)$$

It is a proper scoring rule, consistent for equal-tail error probability intervals:
$S(F, G)$ is minimised for the narrowest PI that has expected coverage $1-\alpha$.

Assessing event prediction probabilities

Often, specific events are of particular interest. We can use an estimated general model to compute predictive probabilities for several events, or construct a model that is only aimed at predicting a specific event. The indicator variable $Z = \text{The event occurred}$ has a Binomial prediction distribution, $\text{Bin}(1, p_F)$. We can therefore use the Log-score, $S_{\text{LOG}}(F, z) = -z \log(p_F) - (1-z)\log(1-p_F)$, to assess predictive performance.

Brier score

The *Brier Score* is a popular event prediction assessment score that for single events can be defined as the Squared Error score with respect to the Binomial expectation:

$$S_{\text{Brier}}(F, z) = (z - p_F)^2$$

For multi-option outcomes, where an observation is classified into one of several categories (instead of just "no/yes"), the score can be generalised to

$$S_{\text{Brier}}(F, z) = \sum_{k=1}^{K} \left[\mathbb{I}(z = k) - \mathsf{P}_F(Z = k)\right]^2$$

The Brier score is proper, but only strictly proper w.r.t. a specific set of event categories.

Average scores

Average score

Given a collection of prediction/truth pairs, $\{(F_i, y_i), i = 1, \dots, n\}$, define the *average* or *mean* score:

$$\overline{S}(\{(F_i, y_i), i = 1, \dots, n\}) = \frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i)$$

- When comparing prediction quality, we often look at the difference in average scores across the test data set.
- For modern, complex models with explicit spatial and temporal model components, and for constructing formal tests, the *pairwise* differences are more useful: For two prediction methods, F and F',

  $$S_i^{\Delta}(F_i, F_i', y_i) = S(F_i, y_i) - S(F_i', y_i)$$

  We can have $\overline{S}^{\Delta} \approx 0$ at the same time as all $|S_i^{\Delta}| \gg 0$, if the two models/methods are both good, but e.g. at different spatial locations.
- Later: How can we assess whether the score differences are indistinguishable?