FlatPPL 0.1 / REGRESSION FORMULAS

The formula API

PYTHON ENTRY POINT

```
from flatppl import from_formula
from_formula("y ~ x", data)
```

POPULATION TERMS

η_i is the linear predictor for row i. Here α is the intercept on the original covariate scale.

Intercept alone

$\eta_i = \alpha$

```
y ~ 1
```

A line with an intercept

$\eta_i = \alpha + \beta x_i$

```
y ~ x
```

A line through the origin

$\eta_i = \beta x_i$

```
y ~ 0 + x
```

The product alone

$\eta_i = \alpha + \delta x_i z_i$

```
y ~ x:z
```

Both slopes and their interaction

$\eta_i = \alpha + \beta_x x_i + \beta_z z_i + \delta x_i z_i$

```
y ~ x * z
```

A quadratic in one variable

$\eta_i = \alpha + \beta_1 x_i + \beta_2 x_i^2$

```
y ~ x + I(x^2)
```

`I(...)` evaluates the arithmetic inside the parentheses.

All two-way interactions

$\eta_i = \alpha + \sum_j \beta_j u_j + \sum_{j<k} \gamma_{jk} u_j u_k$

```
y ~ (x + z + w)^2
```

At row i, $u = (x_i, z_i, w_i)$.

A factor with a reference level

$\eta_A = \alpha, \quad \eta_B = \alpha + \delta$

```
y ~ f
```

The unordered factor `f` has levels A, B, in that order.

One coefficient per level

$\eta_A = \beta_A, \quad \eta_B = \beta_B$

```
y ~ 0 + f
```

GROUP EFFECTS AND CURVES

Write $j = g_i$ for the group of row i. Conditional on their scales and correlations, the group effects are zero-mean Gaussian.

Varying intercept

$\eta_i = \alpha + u_j + \beta x_i$

```
y ~ x + (1 | g)
```

Varying intercept and slope

$\eta_i = \alpha + u_j + (\beta + v_j) x_i$

```
y ~ x + (1 + x | g)
```

The group covariance includes a correlation parameter ρ.

Independent intercept and slope

$\eta_i = \alpha + u_j + (\beta + v_j) x_i, \quad \rho = 0$

```
y ~ x + (1 + x || g)
```

Crossed group intercepts

$\eta_i = \alpha + u_{g_i} + r_{h_i}$

```
y ~ (1 | g) + (1 | h)
```

Each row belongs to one level of each grouping factor.

Monotonic dose effect

$\eta_i = \alpha + bD \sum_{k=1}^{d_i} \zeta_k$

```
y ~ mo(d)
```

Integer doses run from 0 to D. The nonnegative steps ζ_k sum to one, so b is the mean change per step.

A smooth curve

$\eta_i = \alpha + f(x_i)$

```
y ~ s(x, k=6)
```

`k=6` sets the thin-plate basis dimension. The fitted curve shares one smoothing scale.

An exponential curve with two parameters

$\eta_i = a \exp(b x_i), \quad a, b \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$

```
y ~ a * exp(b * x)
  nlpars={"a": "a ~ 1", "b": "b ~ 1"},
  priors=[{"class": "b", "nlpar": p,
           "prior": "normal(0,1)"}
          for p in ("a", "b")]
```

With `nlpars`, `*` multiplies in the response expression. Each named parameter requires a coefficient prior.

RESPONSE LAWS AND OBSERVATION TERMS

In this column $\eta_i = \alpha + \beta x_i$. The default family is Gaussian. Indented keywords are Python call arguments.

Successes in a known number of trials

$Y_i \sim \text{Binomial}(n_i, p_i), \quad \text{logit}(p_i) = \eta_i$

```
hits | trials(n) ~ x
  family="binomial"
```

Counts per unit exposure

$Y_i \sim \text{Poisson}(t_i \exp(\eta_i))$

```
count ~ x + offset(log(t))
  family="poisson"
```

The coefficient of $\log t_i$ is fixed at one.

A Poisson count with extra zeros

$\Pr(Y_i = 0) = \pi + (1 - \pi) e^{-\lambda_i}, \quad \lambda_i = e^{\eta_i}$

```
count ~ x
  family="zero_inflated_poisson"
```

`zi` is the extra-zero probability π.

Known error and residual variance

$Y_i \sim \mathcal{N}(\eta_i, s_i^2 + \sigma^2)$

```
y | se(s, sigma=TRUE) ~ x
```

A predictor for the residual scale

$Y_i \sim \mathcal{N}(\eta_i, \sigma_i^2), \quad \log \sigma_i = \gamma_0 + \gamma_1 z_i$

```
y ~ x
  dpars={"sigma": {"formula": "sigma ~ z"}}
```

An interval-censored reading

$\Pr(y_i < Y_i \le u_i) = F_i(u_i) - F_i(y_i)$

```
y | cens(status, upper) ~ x
```

For `status=2`, `y` and `upper` bound the interval. F_i is the Gaussian CDF.

A response restricted to a range

$p(y_i \mid 0 < Y_i < 3) = \frac{f_i(y_i)}{F_i(3) - F_i(0)}$

```
y | trunc(lb=0, ub=3) ~ x
```

This density applies inside the range and is zero outside it.

Weighted observations

$p(\theta \mid y) \propto p(\theta) \prod_i p(y_i \mid \theta)^{w_i}$

```
y | weights(w) ~ x
```

Nonnegative weights multiply the row log likelihoods in the posterior.

Jessica Cox · bcox@mpp.mpg.de