Fundamentals

Exponential Dispersion Family

The EDF is a family of distributions which have the following mass/density function:

$$P_Y(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right), \text{ where:}$$

- θ is the natural/location parameter, relating to the location or mean of Y values.
- ϕ is the dispersion parameter, relating to the spread of Y values.

The EDF family enjoys the following properties:

1. $\text{E}(Y) = \mu = b'(\theta)$.
2. $\text{Var}(Y) = \phi b''(\theta)$.
3. The variances function $\text{V}(\mu)$ is defined as $b''(\theta)$ rewritten in terms of the mean μ. This formulation is a reminder that variance depends on the mean of a variable, and it uniquely identifies probability models under the EDF umbrella.

A subgroup of this family is the **Natural Exponential Family** (NEF), where we assume $\phi = 1$.

Generalized Linear Models

All models have a set number of y_i response/outcome variables, and associated lists of $x_{i1}, \cdot x_{iJ}$ covariance/features (namely J of them). The basic framework of a generalized linear model is in its three base components:

- Random Component: The response variables y_i follow a probability distribution, usually an EDF member, with mean μ_i.
- Systematic Component: The predictor variables combine with unknown coefficients in a linear manner to create a coalesced term η_i.

  $$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ}$$

  To create our models, we try to estimate the coefficients β using Maximum Likelihood Estimation. Additionally, it is sometimes useful to estimate the dispersion parameter ϕ of the underlying EDF distribution. The maximum likelihood estimate is unfortunately unreliable and biased for this purpose, so we usually use the following estimator:

  $$\hat{\phi} = \frac{\chi^2}{n-J-1}, \text{ where } \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\text{V}(\hat{\mu}_i)}$$

- Link Component: The link function $g(\cdot)$ connects the coalesced term η_i to the mean of the response variable μ_i.

  $$\eta_i = g(\mu_i)$$

  The inverse link function connects/transforms the mean of the response μ_i back to the coalesced term η_i.
  Usually we set the following three **requirements** for the link function $g(\cdot)$:

  1. It is differentiable.
  2. It is monotonically increasing over the range of μ.
  3. It expands the range of μ to the entire real line.

  A simple choice for the link function is called the **canonical link** function, defined as $g(\cdot) = b'^{-1}(\cdot)$, enjoying the following two properties:

  – It constrains the data to functions of the data called sufficient statistics.
  – It results in the Newton-Raphson and Fisher Scoring algorithms in coinciding.

Likelihood Maximization

The likelihood function is the probability of the observed data as a function of the unknown parameters. Typically, we need to maximize likelihood functions to estimate unknown parameters of interest. There are three main approaches to maximization:

(a) Analytical Maximization: Maximize using calculus:

1. Find the log-likelihood $\ell(p)$ for some unknown parameter p.
2. Take the derivative of the log-likelihood function with respect to p; this is the score function.
3. Set the score function equal to zero & solve for parameter p.
4. Check that the derivative of the score function with respect to p is negative over the relevant region of p to make sure we found a maximum point.

- **Advantage**: Gives a correct result right away.
- **Disadvantage**: Isn't always feasible, especially when dealing with more complex scenarios with many parameters.

(b) Newton-Raphson Algorithm:

1. Find the first & second derivatives of the log-likelihood:

   $$\ell'(p) = \frac{d\ell(p)}{dp}, \quad \ell''(p) = \frac{d^2\ell(p)}{dp^2}$$

2. Pick a starting point $p = p^{(0)}$.
3. Iterate according to the following update formula until the difference between successive guesses is negligible:

   $$p^{(k+1)} = p^{(k)} - \frac{\ell'(p^{(k)})}{\ell''(p^{(k)})}$$

- **Advantage**: Fastest iterative approach at arriving at a solution.
- **Disadvantage**: Breaks down with poor starting value choices.

(c) Fisher Scoring Algorithm: The same as the NR algorith, but replaces the second derivative with its espectation.

- **Advantage**: Approximates the second derivative, making it more robust to divergence issues.
- **Disadvantage**: Converges slower than alternatives like the Newton-Raphson algorithm.

Binary Response Models

We typically use the Bernoulli distribution to represent the underlying statistical distribution of binary outcome data. For a Bernoulli model, we can solve for its EDF parameters:

Natural Parameter	$b(\theta) = \log(1+e^{\theta})$
Natural Parameter Function	$\theta = \log\left(\frac{p}{1-p}\right), p = \frac{e^{\theta}}{1+e^{\theta}}$
Other Parameters	$\phi = 1, c(y,\phi) = 0$

As for the creation of the generalized linear models for binary data, we have multiple options, depending on how we think of underlying error in a latent variable formulation. Namely, we can think of binary response models as ones where we have a latent variable z_i, and associated response variable y_i:

$$z_i = \mathbf{x}_i'\beta + \epsilon_i, \quad \text{and} \quad y_i = \begin{cases} 0 & \text{if } z_i \le 0 \\ 1 & \text{if } z_i > 0 \end{cases}$$

Different distribution for the error term ϵ_i imply different link/modeling choices, which are:

- The error ϵ_i follows a **logistic distribution**, giving rise to **logistic regression**. Logistic regression uses the logit function as its link function – the canonical link function of the Bernoulli distribution. It has the following link and inverse link functions:

  $$g(\mu_i) = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$$

  $$g^{-1}(\mathbf{x}_i'\beta) = \mu_i = \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)} = \frac{1}{1+\exp(-\mathbf{x}_i'\beta)}$$

- The error ϵ_i follows a **normal distribution**, giving rise to **probit regression**. Probit regression predicts more extreme probabilities (ones closer to 0 and 1) than logistic regression. It has the following link and inverse link functions:

  $$g(\mu_i) = \text{probit}(\mu_i) = \Phi^{-1}(\mu_i), \quad g^{-1}(\mathbf{x}_i'\beta) = \mu_i = \Phi(\mathbf{x}_i'\beta)$$

- The error ϵ_i follows a **gumbel distribution**, giving rise to **complementary log-log regression**. This type of regression predicts extreme low probabilities but conservative high probabilities. It has the following link and inverse link functions:

  $$g(\mu_i) = \log(-\log(1-\mu)), \quad g^{-1}(\mathbf{x}_i'\beta) = \mu_i = 1-\exp(-\exp(\mathbf{x}_i'\beta))$$

Factors

Categorical predictors can be modeled as factors, with each value termed a *level.* Some factors are ordered (e.g., weather condition) while others are unordered (e.g., ethnicity). Instead of using M one-hot-encoded predictors to represent factors with M levels (which would introduce aliasing of the intercept term), we typically employ treatment contrast coding.This approach involves creating $M-1$ new predictor variables $(u_1, \dots, u_{M-1})$ for a factor with M levels. When level 0 of the factor is observed, all u's are set to 0. For other levels ($i > 0$), u_i is set to 1, while the rest remain 0. This setup allows model coefficients to represent the comparative effects of belonging to level $i > 0$ compared to level 0 (the reference category). Treatment contrast coding can be applied to any number of factors, with the procedure performed separately for each.

We can asses whether including a factor is statistically significant by performing a **Likelihood Ratio Test** (LRT). This implies two models: H_0 excluding the factor predictors, and H_a including the k added terms of the factor. The steps of the LRT are:

1. Calculate the maximum likelihood estimates of the coefficients of both models, then calculate the associated log-likelihoods.
2. Create the test statistic:

   $$T = -2(\ell(\hat{\mu}_0|y) - \ell(\hat{\mu}_a|y)) \sim \chi^2_k$$

3. Reject the simpler model (meaning the factor is significant) if the p-value of the associated test statistic is $\le \alpha$.

We can use a variant of the LRT, the method of profile likelihoods, to find model confidence intervals:

1. Calculate the maximum likelihood estimates of the coefficients and the associated log-likelihood of the resulting model.
2. Fix the coefficient $\beta_j = \tilde{\beta}_j$ at a certain value, and then do maximum likelihood estimation to find the MLE of β_{-j} and the associated "profile log-likelihood".
3. Reject the null hypothesis if:

   $$-2(\tilde{\ell}(\tilde{\beta}_j) - \ell(\hat{\beta})) \ge z^2_{1-\alpha/2}$$

4. Find all values of $\tilde{\beta}_j$ that do not result in rejecting the null, these are the confidence interval values.

Coefficient Interpretation

Odds

We define the odds of an event A as:

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)}$$

Which is a function mapping probabilities to the scale $[0, \infty)$. This is a natural framework to interpreting the logit of a probability:

$$\text{logit}(P(Y = 1)) = \log(\text{odds}(Y = 1))$$

We define the odds ratio as the effect of increasing a covariate by 1 on the odds of an event:

$$\text{OR}_j = \frac{\text{odds}(Y = 1|x_j = a + 1)}{\text{odds}(Y = 1|x_j = a)} = \exp(\beta_j)$$

Example

Consider a model predicting whether or not a streaming platform subscriber will renew their subscription. The model has the following expression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \alpha_1 u_1 + \gamma_1 v_1 + \zeta_1 v_1 x_{i1} + \delta_1 u_1 v_1, \text{ where:}$$

- x_1 is the number of hours spent on the platform.
- u_1 is an indicator for whether a person tunes in through their tablet (reference) or TV.
- v_1 is an indicator for whether a person watches Action (reference) or Drama shows.

The coefficients of the model have the below interpretations. It is important to note which variables (if any) need to be held constant for the interpretation of a particular coefficient to hold:

1. β_0 is the log-odds that a person that uses a tablet and watches Action shows renews their subscription.
   Alternatively, $\text{logit}^{-1}(\beta_0)$ is the probability that a tablet-user who watches Action shows renews their subscription.
2. A 1 hour increase in the hours spent on the platform is associated with a β_1 increase in the log-odds (logit of the probability) of a user renewing their subscription.
3. The log-odds of renewing their subscription for a user that tunes in through their TV is α_1 higher/lower than the log-odds for a user that tunes in through their tablet.
   Alternatively, the odds of a TV-user renewing their subscription are $\exp(\alpha_1)$ times as large/small as the odds of a tablet-user renewing their subscription.
4. The log-odds of renewing their subscription for a user that watches Drama is γ_1 higher/lower than the log-odds for a user that watches Action.
   Alternatively, the odds of a Drama-viewer renewing their subscription are $\exp(\gamma_1)$ times as large/small as the odds of an Action-viewer renewing their subscription.
5. The effect of a 1 hour increase in the number of hours spent on the platform on the log-odds of a viewer renewing their subscription is higher/lower by ζ_1 for Drama-viewers relative to Action-viewers.
6. The effect on the log-odds of renewing their subscription of being a Drama-viewer relative to being an Action-viewer is δ_1 larger/smaller for TV-viewers relative to tablet-viewers.
   Alternatively, the effect on the log-odds of renewing their subscription of being a TV-viewer relative to being a tablet-viewer is δ_1 larger/smaller for Drama-viewers relative to Action-viewers.

Model Issues

Collinearity

Collinearity refers to the case where predictors are linear combinations of one another or nearly linear combinations of one another. This issue can result in coefficient signs that don't make much sense, and large coefficient values and standard errors. However, many diagnostics exist to detect and fix this issue:

- Pairwise Correlations: Calculate pairwise correlations, checking to see if any are close to 1.
- Condition Number: The condition number of a matrix is:

  $$\kappa = \sqrt{\lambda_1 / \lambda_J}$$

  Where the λs are the largest and smallest eigenvalues, respectively. Condition numbers > 30 mean we are dealing with a nearly singular matrix, meaning collinearity is likely an issue. It is useful to note that this metric tends to overstate collinearity issues, however.
- Variance Inflation Factor: The vif is a quantity we calculate for each predictor according to the following formula:

  $$\text{vif}_j = \frac{1}{1 - R_j^2}$$

  Where R_j^2 is the R^2 from regressing predictor j on the other predictors. Values > 10 indicate collinearity issues.
- Generalize Variance Inflation Factor: A modified version of the vif that allows for the inclusion of factor variables in analysis, with the following modified formula:

  $$\text{GVIF}_j = \frac{\det R_j \times \det R_{-j}}{\det R}$$

  Where R_j is the correlation matrix of p_j quantitative variables. We use the $1/p_j$th root as our stand-in for vif.

Separability

Exact or quasi separability is when binary responses can be separated or nearly separated according to a particular predictor, resulting in large coefficients and general model instability.

Goodness of Fit

Hosmer-Lemeshow Test

The Hosmer-Lemeshow test examines the deviation of estimated probabilities of success from a model and the observed fraction of successes, checking if probabilities are poorly calibrated. It is used in cases where we have quantitative variables in our dataset, as that can result in a spectrum of different predicted probability values for our observations. The steps of the test are as follows:

1. Fit a model, and compute the predicted probabilities for observations, $\hat{p}_i$.
2. Sort the predicted probabilities in order, and divide them into G groups (conventionally $G = 10$, but we usually repeat the test for different values of G).
3. Compute the average predicted probability per group, $\bar{p}_g$.
4. Compute the expected number of success in each group, $E_g = n_g \bar{p}_j$, and observed number of successes in the group, O_g. Do the same for the probabilities/counts of failure.
5. Calculate the test statistic:

   $$\sum_{g=1}^{G} \frac{n_g (O_g - E_g)^2}{E_g (n_g - E_g)} \sim \chi^2_{G-2}$$

Deviance

Understanding deviance requires understanding what a saturated model is. We can create a saturated model by maximizing each term in its log-likelihood individually (i.e. setting $\mu_i^* = y_i$), instead of pooling everything together. This allows us to create a scaled deviance statistic:

$$D_\phi(\mu|y) = \frac{1}{\phi} D(\mu|y) = 2\ell(\mu^*|y) - 2\ell(\mu|y) \sim \chi^2_{n-J-1}$$

So, when performing the Likelihood Ratio Test, we can simply use the difference of the scaled deviances of the models as our test statistic, as it boils doing to the same statistic.
We can use the quotient of the unscaled deviances and $n - J - 1$ to judge lack of fit (either version of deviance is equivalent here since for binary outcome models $\phi = 1$). We know that:

$$\text{If } \frac{D(\mu|y)}{n - J - 1} \gg 1, \text{ then the model might be of poor fit.}$$

It is important to note that this test has a low power: if we reject the null then we definitely have an issue in terms of lack of fit, and if we don't we still might!

Pearson Residual

$$r_i^{(p)} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i (1 - \hat{\mu}_i)}} > 2 \text{ indicates outliers.}$$

Deviance Residual

$$r_i^{(d)} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i} > \text{ 2 or 3 indicates outliers,}$$
$$\text{given, } d_i = -2(y_i \log(\hat{\mu}_i) + (1 - y_i)\log(1 - \hat{\mu}_i))$$

Jackknifed Residual

$$r^*_{(i)} = \frac{y_i - \hat{\mu}_{(i)}}{\sqrt{\text{Var}(y_i - \hat{\mu}_{(i)})}} \approx r_i^* \sqrt{\frac{n - J - 1}{n - J - r_i^{*2}}}$$

Diagnostic Plots

Deviance residuals as well as jackknifed residuals can be plotted to detect non-linearities or outliers. However, these plots are ineffective for binary response models as we get two streaks, one per response. Instead, we can plot the average fitted residual over a bin against the average fitted probability and examine the plot for non-linearities.

Cook's Distances

These measure how much the coefficients of a model change due to the inclusion of a particular observation.

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})' \text{Var}(\hat{\beta})^{-1} (\hat{\beta}_{(i)} - \hat{\beta})/(J + 1) \geq 1 \text{ indicates an outlier.}$$

Probability Confidence Intervals

Less Precise Method: Use the Delta method to find the standard error of an estimated probability:

$$\text{S.E.}(\hat{\mu}) = \left|\frac{\partial \mu}{\partial \eta}\right| (x' \text{Var}(\hat{\beta}) x)^{1/2} \approx \left|\frac{\partial \mu}{\partial \eta}\right|_{\mu = \hat{\mu}} (x' \text{Var}(\hat{\beta}) x)^{1/2}$$
$$= \left|\frac{1}{\partial g(\mu)/\partial \mu}\right|_{\mu = \hat{\mu}} (x' \text{Var}(\hat{\beta}) x)^{1/2}$$

More Precise Method: Find the S.E. on the logit scale and then invert the bounds of the C.I. using the inverse logit function:

$$\text{S.E.}(\text{logit } p_i) = \text{S.E.}(\mathbf{x}_i' \hat{\beta}) = \sqrt{x_i' \text{Var}(\hat{\beta}) x_i}$$

Missingness

Assumption: Missingness mechanism is ignorable for all features exhibiting missingness; the reason values are missing does not depend on their values.
Solution: Create a missingness indicator for the missing data, and then using the mean of the observed values for the feature as the impute value. The coefficient value is the effect of an observation having missingness in a feature over not having missingness.

Poisson Models

Consider a model predicting how many customers will enter an ice cream shop on a given day. The model has the following expression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \alpha_1 u_1, \text{ where:}$$

- x_1 is the temperature outside that day (in Fahrenheit).
- u_1 is an indicator for whether it rained that day.

The coefficients of the model have the below interpretations. It is important to note which variables (if any) need to be held constant for the interpretation of a particular coefficient to hold:

1. A 1 degree increase in the outside temperature corresponds to an increase of β_1 in the mean number of customers at the ice cream shop on the log scale.

   Alternatively, a 1 degree increase in the outside temperature corresponds to a multiplicative increase in the mean number of customers at the ice cream shop by $\exp \beta_1$.
2. The mean number of customers at the ice cream shop on a rainy day is $\exp \alpha_1$ times as large as the mean number of customers on a non-rainy day.

The Poisson deviance function is:

$$D(\hat{\mu}|y) = -2(\ell(\hat{\mu}|y) - \ell(\mu^*|y)) = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{\mu_i}}\right) - (y_i - \hat{\mu_i})\right)$$

Since $\phi = 1$ for Poisson models, a good measure of fit is to check if:

$$\frac{D(\hat{\mu}|y)}{(n-J-1)} \approx 1$$

Data is not always collected at a uniform exposure (area, duration of time), so adjusting for that is really useful. We model each observation as:

$$\log(T_i\lambda_i) = \log(T_i) + x_i'\beta$$

Where T_i is a measurement offset without its own coefficient.

Multinomial Models

Multinomial models provide a framework for analyzing units with the same set of features which fall into different categories. The multinomial logit model utilizes the logistic link function and computes a unique set of coefficients for each category, excluding the reference category whose coefficients are set to zero. As a result, the model possesses $(K-1)(n-J-1)$ degrees of freedom. The model takes on the following form:

$$p_{ik} = \begin{cases} \frac{1}{1+\exp(x_i'\beta_2)+\cdots+\exp(x_i'\beta_K)} & \text{if } k = 1 \\ \frac{\exp(x_i'\beta_k)}{1+\exp(x_i'\beta_2)+\cdots+\exp(x_i'\beta_K)} & \text{if } k = 2, 3, \ldots K \end{cases}$$

The formulaic expression for the model is:

$$\log(p_{ik}/p_{i1}) = \log\left(\frac{\exp(x_i'\beta_k)/\left(1+\sum_{\ell=2}^{K}\exp(x_i'\beta_\ell)\right)}{1/\left(1+\sum_{\ell=2}^{K}\exp(x_i'\beta_\ell)\right)}\right)$$

Consider a model, which, given a customer's characteristics, outputs the probabilities of the customer buying vanilla, chocolate, or strawberry ice cream. The model has the following features:

- x_1 is the annual income of the customer, in 1000's of dollars.
- u_1 is categorical variable for whether the customer is a Democrat (reference category), a Republican, or an Independent.

The coefficients of the model have the below interpretations:

1. The log of the ratio of the probability of getting chocolate ice cream over vanilla ice cream increases by β_2 for every extra 1000 dollars in annual income.

   Alternatively, the ratio of the probability of getting chocolate ice cream over vanilla ice cream increases by a factor of $\exp(\beta_2)$ for every extra 1000 dollars in annual income.
2. The log of the ratio of the probability of getting chocolate ice cream to the probability of getting vanilla ice cream is higher by α_2 for Republicans relative to Democrats.

   Alternatively, the ratio of probabilities that a Republican gets chocolate ice cream versus vanilla ice cream is $\exp(\alpha_2)$ the ratio of probabilities that a Democrat gets chocolate ice cream versus vanilla ice cream.

After modeling, we can determine the probability of an individual belonging to each class based on its features. This involves computing class-specific regressions and applying a formula for probabilities.
The multinomial logit model quantifies the appeal of each class to an individual unit. Selection probabilities are determined by the ratio of class appeal to the total appeal across all classes. This ensures that eliminating options only redistributes probabilities among the remaining options, maintaining consistent choice odds.
The deviance function can be expressed as (depending on the multinomial observation counts):

$$D(\hat{\mu}|y) = \begin{cases} -2\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log(p_{ik}) & \text{if } N_i = 1 \\ -2\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log\left(\frac{N_i p_{ik}}{y_{ik}}\right) & \text{otherwise.} \end{cases}$$

In multinomial models we can't check feature significance with usual methods, because one may get different results in different categories; what we should do is perform a likelihood ratio test.

Model Selection

Typically, a model's quality is evaluated using its log-likelihood. However, when comparing different models, log-likelihood tends to rise with model complexity, which doesn't guarantee good generalization. Employing a measure that balances goodness of fit with modeling complexity can mitigate overfitting compared to relying solely on log-likelihood. Two widely used metrics for this purpose are the Akaike Information Criterion and Bayesian Information Criterion:

$$\text{AIC}_m = -2\ell(\hat{\beta}_m|y, X_m) + 2d_m$$

$$\text{BIC}_m = -2\ell(\hat{\beta}_m|y, X_m) + d_m \cdot \log n$$

The model with the lowest AIC aims to minimize the Kullback-Leibler divergence from the true model. AIC is efficient for non-parametric frameworks and suited for prediction, but it can converge to the wrong model in parametric settings with large sample sizes.
BIC is consistent and efficient for parametric models, suitable for inference. However, it lacks efficiency in non-parametric settings.
To find the best model, a stepwise approach adjusts features to minimize AIC or BIC. Alternatively, cross-validation partitions data into development and test sets, with models selected based on lowest total deviance. One kind of cross-validation, called Leave-One-Out-Cross-Validation (LOOCV), involves setting the fold size in the development set to equal 1. In high dimensional settings, it is also popular to use LASSO penalization for the number of parameters during training, which automatically zeroes out some coefficients.
After model selection, inferences should be made only on the test set to avoid data dredging. Models must be refitted on the test set for accurate inferences.

Ordinal Response Regression

Proportional response or ordered logit models are a way to model the relationship between features and an ordered set of categories. The proportional odds model is set up to measure the probability of being in lower categories, while the logistic model is set up to measure the probability of being in the higher category.
If we let $\gamma_k(x) = \text{P}(Y \le k|x)$, the proportional odds model assumes:

$$\text{logit } \gamma_k(x) = \log\left(\frac{\gamma_k(x)}{1-\gamma_k(x)}\right) = \theta_k - x'\beta$$

$$\frac{\text{odds}(Y \le k|x_1)}{\text{odds}(Y \le k|x_2)} = \left(\frac{\gamma_k(x)}{1-\gamma_k(x)}\right) \Big/ \left(\frac{\gamma_k(x)}{1-\gamma_k(x)}\right) = \exp(-(x_1 - x_2)'\beta)$$

And if the difference between the two sets of predictors is that the first has a feature value greater by 1, then this becomes:

$$\frac{\text{odds}(Y \le k|x_1)}{\text{odds}(Y \le k|x_2)} = \exp(-\beta_j), \text{or } \frac{\text{odds}(Y > k|x_1)}{\text{odds}(Y > k|x_2)} = \exp(\beta_j)$$

Consider a model, which, given a reviewer's characteristics, outputs whether they would rate a movie as "poor", "okay", or "good".
The model has the following features:

- x_1 is the number of movies the reviewer watched before.
- u_1 is categorical variable for whether the reviewer maintains a blog of their reviews.

The coefficients of the model have the below interpretations:

1. For each additional movie watched, the odds that a reviewer would rate a movie a higher category increases by a factor of $\exp(\beta_1)$.
2. Having a maintained blog increases a reviewer's odds of rating a movie a higher category increases by a factor of $\exp(\alpha_1)$.

Given that the θ's are the boundaries between categories on the $x'\beta$ scale, we can write the probabilities of belonging in each class as follows:

$$\text{Cat. 1. } \text{P}(Y_i = 1|x_i) = \frac{1}{1+\exp(-\hat{\theta}_1 + x_i'\hat{\beta})}$$

$$\text{Cat. 2. } \text{P}(Y_i = k|x_i) = \frac{1}{1+\exp(-\hat{\theta}_k + x_i'\hat{\beta})} - \frac{1}{1+\exp(-\hat{\theta}_{k-1} + x_i'\hat{\beta})}$$

$$\text{Cat. 3. } \text{P}(Y_i = K|x_i) = 1 - \frac{1}{1+\exp(-\hat{\theta}_{K-1} + x_i'\hat{\beta})}$$

The hazard for Y evaluated at level k given predictors x is defined as:

$$h_Y(k|x) = \mathrm{P}(Y = k|Y \geq k, x) = \frac{\mathrm{P}(Y = k)}{\mathrm{P}(Y \geq k)}$$
$$= 1 - \exp\left(- e^{-x'\beta}(e^{\theta_k} - e^{\theta_{k-1}})\right)$$

In latent variable representation, the noise in the model can be modeled using the logistic distribution, the normal distribution (resulting in an ordered probit model), or extreme value distribution (resulting in a proportional hazards model), which has the following CDF:

$$P(Y \leq y) = 1 - \exp(-e^y)$$

Overdispersion

We usually use the quotient of the residual deviance and the number of degrees of freedom as a measure of goodness of fit. When the value is greater than 1, we know the lack of fit may be poor for any of the following reasons:

- Incorrect functional form due to non-linearities.
- Dependence across observations results in observations clustering together.
- Important predictors are omitted from the model.

One potential solution is to fit GLMs with ϕ treated as unknown and needing to be estimated.

The other alternative is to use non-GLMs that account for overdispersion.

The beta-binomial model is an alternative to a binomial model with two unknown parameters instead of one (p). The model can be expressed using the following expression:

$$\text{logit } p_i = \text{logit}\left(\frac{\alpha_{1i}}{\alpha_{1i} + \alpha_{2i}}\right) = x_i'\beta$$
$$\varphi = \frac{1}{\alpha_{1i} + \alpha_{2i} + 1}$$

The model estimates the values of β and φ, and those can be used to solve for the two parameters of the beta-binomial distribution.

Negative binomial model is an alternative to a Poisson model, also with two unknown parameters instead of one (μ). The model can be expressed as follows:

$$\log \mu_i = \log\left(\frac{\alpha_1}{\alpha_{2i}}\right) = x_i'\beta$$

The model estimates the values of β and α_1, and as a result we can solve for α_{2i} using:

$$\hat{\alpha}_{2i} = \frac{\hat{\alpha}_1}{\exp(x_i'\beta)} = \hat{\alpha}_1 \exp(-x_i'\beta)$$

To compare the performance of these models in relation to their typical counterparts, we typically look at the log-likelihoods. Residual deviances cannot be compared as the saturated model log-likelihoods are not the same for these groups of models. The Chi-Squared Likelihood Ratio Test can be used to compare models.

Skewed Models

For continuous output models with a skewed distribution, Poisson modeling might not be the best option as the expected value grows linearly with the variance; we sometimes want the variance to grow more rapidly. One model that allows us to do so is the Gamma model, where the variance equals ϕ times the expected value squared.

The canonical link function for the Gamma model is $g(\mu) = -\frac{1}{\mu}$, which is not a good choice for the link function. As a result, we default to using the log link.

If a Gamma model has J predictors, and a second model adds an additional K new predictors, then:

$$F = \frac{D(\hat{\mu}_1|y) - D(\hat{\mu}_2|y)}{K\hat{\phi}}$$

Where F has approximately an F-distribution with K and $n - J - K - 1$ degrees of freedom under the null hypothesis that the first model is the true model ($\hat{\phi}$ is computed from the second model). This can be used (as opposed to a Chi-Squared LRT) to compare models.

Another model that allows us to account for skew is the Inverse Gaussian model, where the variance equals the ϕ times expected value cubed. The Inverse Gaussian distribution is derived as the time until a Brownian motion (Gaussian process) with positive drift reaches a fixed value.

The canonical link function is $g(\mu) = 1/\mu^2$, which is not a good choice, and so we default to using the log link as well.

The Tweedie distributions are a generalized family of distributions such that:

$$V(\mu) = \mu^p, \text{ where } p \geq 0 \text{ except for } p \in (0, 1)$$

When $p = 0$, we have the Normal distribution, when $p = 1$, we have the Poisson distribution, when $p = 2$, we have the Gamma distribution, and when we have $p = 3$, we have the Inverse Gaussian distribution. Using Tweedie GLMs allows us to estimate p without assuming its value, which is useful but results in interpretability concerns. For values $1 < p < 2$, there is a positive probability at $Y = 0$, so these sets of distributions can be used to model inflated-zero datasets.

Smoothers

Here are the general types of smoothers:

- Running-Mean (Moving Average): In a running-mean smoother, you start by choosing a window length. For each data point x, you calculate the mean of the y values within the window centered around x. This mean is then assigned as the smoothed value for the data point x.
- Running-Line Smoother: In a running-line smoother, you also choose a window length. For each data point x, you fit a line (usually using least squares regression) to the data points within the window centered around x. The fitted line's value at x is then assigned as the smoothed value for that data point.
- Kernel Smoother: Kernel smoothers compute a weighted average of all y values, where the weights are determined by a predefined function that assigns higher weights to points closer to a target x and lower weights to points further away. The Gaussian density function is a common choice for this weighting function, creating a bell-shaped curve of influence.
- Lowess: Lowess, or locally-weighted running-line smoother, fits a smooth curve to data by performing weighted least-squares regression. The weights for the regression are derived from a kernel function, ensuring that nearby points have more influence on the fit than distant ones.

A smoother can be represented as an optimization problem. Among all functions f that are twice-differentiable, find the one that minimizes the penalized sum of squares:

$$\text{PSS}(f|y, x) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 \, dt$$

for a given value of $\lambda \geq 0$, where $a \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq b$. The first term measures the lack of fit having chosen f. The second term is a penalty for the wiggliness of f.

For a given λ, the unique f that minimizes the penalized sum of squares is referred to as a "cubic smoothing spline" with "knots" positioned at each x_i. This means that a different cubic polynomial is fitted between each x_i and x_{i+1}, though these polynomials do not necessarily intersect the adjacent data points (x_i, y_i) or (x_{i+1}, y_{i+1}). The smoothing spline maintains continuity at every x_i, ensuring a seamless transition between adjacent segments. Moreover, not only do the cubic polynomials connect at each x_i, but they also exhibit equal first and second derivatives at these points.

Generalized additive models (GAMs) assume that the Y_i have a distribution that belongs to the EDF (Exponential Dispersion Family) family, with $\mu_i = E(Y_i)$. The additive predictor is defined as:

$$\eta_i = g(\mu_i) = \beta_0 + S_1(x_{i1}) + \cdots + S_J(x_{iJ})$$

where each S is a separate smoother. Cross-validation is used to find the optimal λ parameter for each smoother.

Trees

Partioning process walkthrough:

- **Partitioning Data:** Consider splitting each of the k partitions according to a cutoff on each individual predictor variable.
- **Evaluate Deviance for Each Candidate Split:** For each candidate split, which results in a total of $k + 1$ partitions of the data set:
  - For quantitative (normal) responses:
  $$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$
  - For binary responses:
  $$-2\sum_{i=1}^{n}(y_i \log \hat{p}_i + (1 - y_i)\log(1 - \hat{p}_i))$$
- **Choose the Optimal Partitioning:** Select the partitioning into $k + 1$ subsets of the data from Step 2 that produces the smallest deviance.

Since deviance always increases as you make more partitions, including a penalty term for complexity is useful. We do this using different candidate penalty values α:

$$D_\alpha = D + \alpha \cdot (\text{number of partitions})$$

We then select the optimal value for α using cross-validation. Full grown trees are usually pruned afterwards based on the selected complexity value.

We can also tune complexity using the complexity parameter:

$$\texttt{cp} = \hat{\alpha}/D^{(0)}$$

We then either select the parameter than minimizes the 10-fold cross-validated deviance statistic, or use the 1-SE rule: select the parameter with the largest value whose 10-fold cross-validated deviance statistic is less than the sum of the lowest 10-fold cross-validated deviance statistic plus its associated standard error.