Linear regression and generalizations

Erik Štrumbelj

1 Introduction

In these notes we will talk about generalizing the ideas of linear regression to target variables whose support is all of $\mathbb{R}$. For example, categorical data, count data, ordinal data, and bounded continuous data such as, for example, non-negative and unit interval data. The focus will be on general ideas, but we will also discuss some particular models, such as logistic regression, multinomial logistic regression, ordinal logistic regression, Poisson regression, and Gamma regression.

The motivation for learning learning linear regression and its generalizations is twofold. First, these models are, despite their simplicity (or due to their simplicity!) very useful in practice and still the workhorse of applied statistics. And second, the generalization part is preserved even if we go for non-linearity by replacing the linear term with a more complex model, such as a neural network, or if we transform the input space with kernel-based methods.

2 Linear regression

First, we briefly review linear regression. We use linear regression to model the relationship between continuous target (dependent) variable $y_i \in \mathbb{R}$ and corresponding input (independent) variables $x_i \in \mathbb{R}^k$. As the name suggests, the model assumes that the relationship between x_i and y_i is linear, governed by the vector of linear coefficients β.

Formally, we can write the model as $y_i = \beta^T x_i + \epsilon_i$, where the residuals ϵ_i are typically assumed to be independent draws from a normal distribution: $\epsilon_i \sim_{\text{iid}} N(0, \sigma^2)$. Or , in compact statistical modeling notation, $y_i | x_i, \beta, \sigma^2 \sim N(\beta^T x_i, \sigma^2)$.

2.1 Assumptions of linear regression

For future reference, let us explicitly repeat the assumptions: the relationship is assumed to be linear, the input variables are assumed to be known and measured without error, the residuals are assumed to be independent and distributed with the same variance. The latter is also known as homoskedasticity.

There is another assumption that is not immediately clear from the model's description, but reveals itself when we try to infer the parameters. If we choose ordinary least squares, which is also the maximum likelihood solution, the estimate is, in matrix form, $\hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y$, where X is the matrix with x_i as rows and therefore input variables as columns. This will only work if $X^T X$ is invertible. That is, if there is no multicolinearity in the input variables - no input variable can be expressed as a linear combination of one or more other input variables. If it can be, then there will be infinitely many solutions and our computation will fail.

The prototypical example of multicolinearity is when we have more input variables than we have observations. The most common approach to addressing this problem is regularization (penalizing the likelihood) or Bayesian inference with a suitable prior. Both of these topics will be discussed in later lectures.

2.2 Correlated input variables

Even if we do not have perfect multicolinearity, linear dependencies (correlations) between input features are problematic as as small change in the training observations can cause a large change in the estimated coefficients (but not in the predictions!). **For an illustration of the effects of correlation see Example 1 in accompanying dynamic report.** When coefficients are unstable, their interpretation is difficult.

The most common approaches to addressing this are removing highly correlated features or transforming the input variables into less correlated or uncorrelated features, for example, using PCA. However, transformations of the input space just replace one problem with another - instead of interpreting highly correlated features, we can now interpret uncorrelated features but each new feature is potentially a linear combination of many original features.

2.3 Influential observations

Often there will be outliers in our data - points that stand out from the other observations. Some models are robust to such points, while others are not. Linear regression, like most simple models, falls into the latter group and for such models the most typical approach is to remove the outliers. When it comes to dealing with outliers, there are two practical issues. First, there is no universal definition of what an outlier is. While it is simple to detect values that are impossible, we cannot designate a possible value as an outlier without specifying some arbitrary threshold on the probability of its occurrence. For example, in 100 normally distributed data points it is unusual to find a point that is more than 3 standard deviations from the mean. In 10000 data points it is not. And second, even if we do come up with some criteria, detecting an outlier in high dimensional data is very difficult and/or computationally expensive.

Fortunately, there is a way around this issue by focusing on a particular type of outliers - points that highly influence the model. In principle, we do not want any point to have a disproportionately high influence on the learning and we do not really care much if a point is an outlier if it does not have a disproportionately high influence. In practice we could even detect such points by re-learning the model without them and measuring the difference in coefficients or predictions.

Recall that the fitted values of linear regression are

$$\hat{y} = \hat{\beta}^T X = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy,$$

where $H = X(X^T X)^{-1} X^T$ is called the hat matrix and also the projection matrix because it projects the observed values of the target variable to the fitted values. We also have $\epsilon = (I - H)y$, where ϵ is the vector of residuals. The i-th diagonal element of H, h_{ii}, is called the *leverage* of the i-th observation (x_i, y_i) and we can compute it as $h_{ii} = x_i^T (X^T X)^{-1} x_i$.

The sum of all leverages is

$$\sum_{i=1}^{n} h_{ii} = trace(H) = trace(X(X^T X)^{-1} X^T) = trace(X^T X(X^T X)^{-1}) = trace(I) = k,$$

as $trace(AB) = trace(BA)$ for rectangular matrices of corresponding dimension. Therefore, the average leverage is $\frac{k}{n}$. And, $0 \leq h_{ii} \leq 1$, which follows from $Var(\epsilon_i) = (1 - h_{ii})\sigma^2 \geq 0$,

where we state the equality without proof. So, high leverage points (close to 1) will have very small residual variance - that is, they will pull the model to be close to them so that the residual is close to 0.

Leverage does not take into account the observed value y_i. Now we introduce Cook's distance, which takes into account both x_i and y_i. Cook's distance is defined in the spirit of our naive approach of retraining the model without the point and observing the difference. More precisely, as the normalized squared distance between the predictions $\hat{y}$ we obtain using all observations and prediction $\hat{y}_{-i}$ we obtain with the model that is trained on all but the i-th observation:

$$D_i = \frac{1}{k\hat{\sigma}^2}\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{-i,j})^2,$$

which can be shown to be equivalent to

$$= \frac{\epsilon_i^2}{k\hat{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2} = \frac{t_i^2}{k}\frac{h_{ii}}{(1-h_{ii})},$$

where $t_i = \frac{\epsilon_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ is called the Standardized residual or the (internally) Studentized residual[1] and $\hat{\sigma}$ is the mean squared error of the model. As a rule of thumb, points with Cook's distance above 1 can be considered very influential.

Note that it is surprising that we can compute the model's prediction without the i-th observation without retraining the model. Similarly, we can compute cross-validation estimates of model error for linear regression without retraining the model - another benefit of the simplicity of linear regression:

$$MSE_{\text{LOOCV}} = \frac{1}{n}\sum_{i=1}^{n}(\frac{(y_i - \hat{y}_i)^2}{1-h_{ii}})^2,$$

where $\hat{y}_i$ is the prediction of the model fit on all observations.

For an illustration of influential points Example 2 in accompanying dynamic report. The example also illustrates the use of Standardized coefficients, leverage, and Cook's distance to identify influential points.

2.4 Extrapolation with linear regression

Linear regression is primarily used for interpolating the value of the target variable between observed values. While there is no technical reason why we could not use it to extrapolate, that is, predict for input variables that are outside the range of observed values of the input variables, this is often a very bad idea and even when it is justifiable, it has to be done with caution.

To illustrate the issue, we'll use a figure from a 2004 Nature paper (see Figure 1). They used a linear regression model to extrapolate winning times for the Olympic 100m sprint for men and women. The extrapolation shows that in 2156 the women's winning time will be lower than the men's. Even more, around 2700, the Olympic 100m sprint is expected to finish before

[1] An externally Studentized residual does not use the current observation in the estimation of $\hat{\sigma}$.

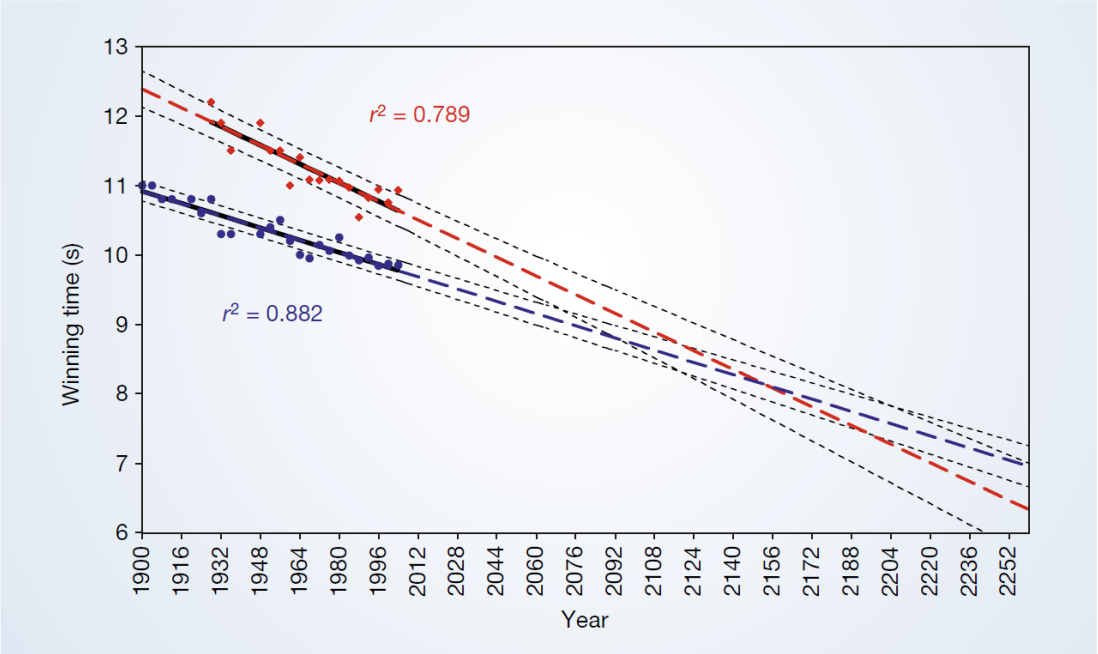

Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Figure 1: Figure 1 with caption from *Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 Olympics?. Nature, 431(7008), 525-525.*

it starts. It can be argued that the goal of the authors here was to illustrate the issues of extrapolation and model selection. Regardless, it is a good example of how linear regression is often (mis)used for extrapolation. Figure 2 shows a more humorous but no less relevant example.

Both examples feature time as the independent variable, but the issue applies to any variable that is not inherently bounded. Furthermore, the issue becomes more sever in higher dimensions, where it becomes difficult to even define when and determine if a point is outside the range of observations in the training set.

2.5 Putting it all together in practice

It follows from our discussion so far that a systematic application of linear regression is a process that involves several steps.

We should inspect the data for correlations and if the goal is interpretation, we should consider removing all but one input variable from every cluster of highly correlated variables. Alternatively, we can transform the variables in a way that de-correlates them.

We should check if the errors are correlated. If they are, we should consider using a model that takes potential correlation into account - if we do not, our error estimates will be biased. A prototypical example of (auto)correlated errors is time series data.

We should inspect the standard diagnostic plots: the plot of standardized residuals against fit-

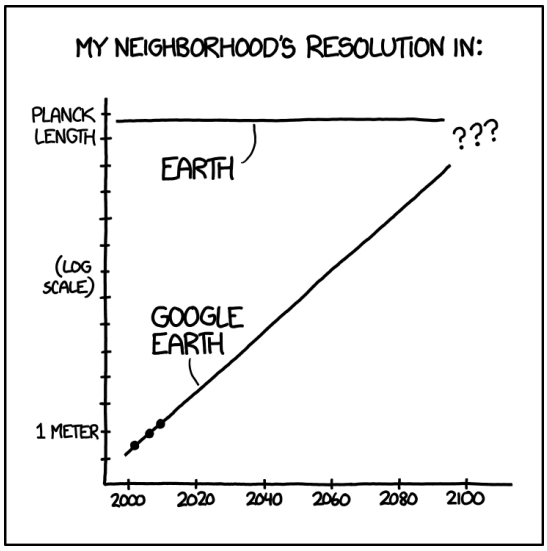

Figure 2: In less than a decade, the resolution of Google Earth will be better that the resolution of Earth. (https://xkcd.com/1204/).

ted values, a Q-Q plot comparison of the standardized residuals and the theoretical quantiles, a plot of fitted against actual values, and a plot of standardized residuals against leverage. These will help us identify deviations from linearity, heteroskedasticity, and influential points.

Also, as a rule, model parameters, not just in linear regression, but in all parametric models, should be interpreted together with some sort of quantification of uncertainty. If we use Bayesian inference, the uncertainty will represented by the posterior distribution. For linear regression and many of its generalizations, closed form expressions exist for standard errors of the parameters and confidence intervals or confidence intervals can be constructed using asymptotic normality arguments[2]. Alternatively, we could bootstrap the sampling distribution of the coefficients. While this is computationally intensive we can apply it to any model, even when a closed form expression is not available or we do not know how to derive it.

3 Generalized linear models

Before a more general treatment, we start our discussion of generalized linear models (GLMs) with a commonly used GLM - logistic regression. Indeed, logistic regression will be the focal point of our discussion of generalizations of linear regression, because it is a GLM, a categorical model, and even an ordinal model.

3.1 Logistic regression

If our target variable is dichotomous (0/1, no/yes) we can still apply linear regression by coding the values with 0/1 or -1/+1, but linear regression can give predictions outside that range, so at a minimum, we have to do some awkward post-processing of nonsensical predictions. We can avoid these issues by choosing a more appropriate distribution for our data, because the normal distribution is clearly not appropriate for dichotomous data.

Here the Bernoulli distribution is a natural choice. However, the main issue remains - the

[2]These quantifications are already built into popular modeling packages and libraries.

linear regression term is in $\mathbb{R}$ and the parameter of the Bernoulli needs to be between 0 and 1. So we need a function that maps $\mathbb{R} \to (0, 1)$. One such function is the inverse logit:

$$logit^{-1}(x) = \frac{1}{1 + e^{-x}}.$$

Putting it all together, we can define logistic regression. We have dichotomous target variable data $y_i \in 0, 1$, independent variables $x_i \in \mathbb{R}^k$, and the Logistic regression model is

$$y_i|\beta, x_i \sim Bernoulli(\theta_i),$$

where $\theta_i = logit^{-1}(\beta^T x_i)$. And if we explicitly write the log-likelihood:

$$\ell(\beta; y, x) = \sum_{i=1}^{n} (y_i \log \theta_i + (1 - y_i) \log (1 - \theta_i)).$$

The choice of the Bernoulli distribution for dichotomous data is a straightforward choice. The choice of the inverse logit, however, is not. Why this function and not some other functions? There many other functions that map from $\mathbb{R}$ to $(0, 1)$, including other cumulative distribution functions of continuous distributions [3]. The main reason is that the inverse logit is computationally convenient, which was historically very important. However, with the development of computers and numerical methods, we can easily use some other function. A common choice is the CDF of the normal distribution, which leads to a model called Probabilistic (or just Probit) regression.

3.2 Poisson regression

Now let our data be count data $y_i \in \{0, 1, 2, ...\}$. Can we apply the ideas of logistic regression to this new setting? First, we need to choose an appropriate distribution for our data. The classical first-attempt choice for count data is the Poisson distribution. Recall that the Poisson distribution is a discrete distribution with support on the non-negative integers. It has a single parameter $\theta > 0$, which is both its mean and its variance[4].

Now we need a function that maps from the real numbers to the positive real numbers - the support of λ. A sensible choice is e^x. Putting it all together, we get Poisson regression:

$$y_i|\beta, x_i \sim Poisson(\lambda_i),$$

where $\lambda_i = \exp(\beta^T x_i)$.

3.3 Exponential family

We will say that a distribution belongs to the natural exponential family of distributions if its PDF or PMF can be written in the form

[3] The inverse logit is the cumulative distribution function of the standard logistic distribution.

[4] Because of this, the Poisson distribution is not appropriate when the data are over or underdispersed. A more flexible distribution, such as the Negative Binomial is a better choice.

$$f(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right),$$

where θ_i is the parameter of the family, ϕ is a scale parameter, and a, b, and c are functions. A useful identity is $E[y_i] = \frac{d}{d\theta_i}b(\theta_i)$. Note that here we only deal with scalar parameter distributions and the natural exponential family, which contains most, but not all of the standard cases. For example, Beta, Multinomial, and multivariate Normal do not fit into this framework but are in the exponential family and have GLMs. The treatment of vector parameter distributions and the full exponential family is beyond the scope of this text. If interested in more details, Chapters 9.1-9.4 of *Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.* are a good starting point.

The Bernoulli/Binomial and Poisson distribution are in the exponential family.

Binomial: $P(Y = y) = \binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i} = \exp(y_i \log\frac{p_i}{1-p_i} + n_i\log(1-p_i) - \log\binom{n_i}{y_i})$, therefore, $\phi = 1$, $a(\phi) = 1$, $\theta_i = \log\frac{p_i}{1-p_i}$, $b(\theta_i) = n_i\log(1+e^{\theta_i})$, and $c(y_i,\phi) = -\log\binom{n_i}{y_i}$.

Poisson: $P(Y = y) = \frac{\lambda_i^{y_i}}{y_i!}e^{-\lambda_i} = \exp(y_i\log\lambda_i - \lambda_i - \log y_i!)$, therefore, $\phi = 1$, $a(\phi) = 1$, $\theta_i = \log\lambda_i$, $b(\theta_i) = e^{\theta_i}$, and $c(y_i,\phi) = -\log y_i!$.

And so are the Normal and Gamma distributions.

Normal: $p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} = \exp\left(\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} + (-1/2\log(2\pi\sigma^2) - y_i^2/2\sigma^2)\right)$, therefore, $a(\phi) = \phi$, $\phi = \sigma^2$,$\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, and $c(y_i,\phi) = -1/2\log 2\pi\phi - y_i^2/2\phi$.

Gamma: $p(y_i) = \frac{\beta_i^\alpha}{\Gamma(\alpha)}y_i^{\alpha-1}e^{-\beta_i y_i}$ (α is assumed to be known) $= \exp(\alpha\log y_i + \alpha\log\beta_i - \log\Gamma(\alpha) - \log y_i - \beta_i y_i)$, therefore, $a(\phi) = 1$, $\phi = 1$, $\theta_i = -\beta_i(\theta_i < 0)$, $b(\theta_i) = -\alpha\log(-\theta_i)$, and $c(y_i,\phi) = \alpha\log y_i - \log\Gamma(\alpha) - \log y_i$.

3.4 Defining GLMs

A GLM is a model with three components:

- A distribution from the exponential family that is assumed for our data y_i.
- A linear predictor $\eta_i = \beta^T x_i$.
- A link function g that connects the expected value $E[y_i] = \mu(\theta_i)$ with the linear predictor: $g(\mu(\theta_i) = \eta_i$.

The link function g is required to be monotonic and differentiable over the range of possible values of $\mu(\theta_i)$. When $\theta_i = \eta_i$ we arrive to the canonical link function:

- For the normal distribution we have $E[y_i] = \mu_i(\theta_i) = \theta_i$. So, the canonical link function is the identity $g(x) = x$. Therefore, linear regression is also a GLM.
- For the Poisson distribution we have $E[y_i] = \mu_i(\theta_i) = e^{\theta_i}$. So, the canonical link function is the logarithm $g(x) = \log x$.
- For the binomial distribution the canonical link is the log odds (or logit) function $g(x) = \log\frac{x}{1-x}$, whose inverse is the inverse logit function. Therefore, logistic regression is the Bernoulli GLM with canonical link function.

- For the Gamma distribution we have $E[y_i] = \mu_i(\theta_i) = \alpha/-\theta_i$, so the canonical link is the negative reciprocal $g(x) = -1/x$.

Except for linear regression, there are no closed form solutions for maximum likelihood estimation of GLMs, so numerical methods must be used. The use of canonical link functions makes computation and theory easier, but in practice other link functions might give better results. We already mentioned using the CDF of the normal distribution that leads to Probit regression (the link function is the inverse of the normal distribution). Another common link function for the [0,1] case is the complementary loglog function $g(x) = \log(-\log(1-x))$.

In the Bayesian framework, GLMs are typically inferred from using Markov Chain Monte Carlo Methods. In the non-Bayesian approaches, maximum likelihood is used. GLMs allow for a modified Newton-Rhapson approach called Iteratively Reweighted Least Squares (IRLS). See Chapter 4 of *Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.*

The diagnostics plots and procedures for linear regression from Section 2 can also be used for GLMs as long as appropriate transformations are made. However, for some distributions, logistic regression in particular, interpretation is more difficult.

Note that it is possible to extend the ideas of GLMs to distributions that are not in the exponential family. However, we lose some of the convenient statistical properties, so inference is more difficult. See, for example, Venter, G. G. (2007). Generalized linear models beyond the exponential family with loss reserve applications. ASTIN Bulletin: The Journal of the IAA, 37(2), 345-364.

3.5 Gamma regression with log link

The canonical negative reciprocal link is not the most popular link for the Gamma GLM. More often the log link is used.

Recall the pdf of the Gamma distribution $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ and that its mean and variance are $E[X] = \frac{\alpha}{\beta}$ and $Var[X] = \frac{\alpha}{\beta^2}$, respectively. Note that we use θ for the coefficients to avoid confusion with the β parameter of the Gamma distribution.

With the log link we gave $\log E[Y_i] = \theta^T x_i$ and $E[Y_i] = e^{\theta^T x_i}$. Let $\mu_i = e^{\theta^T x_i}$. We have $R[Y_i] = \mu_i = \frac{\alpha}{\beta}$ and $Var[X] = \frac{\mu_i}{\beta}$. The model is then

$$y_i|\theta, x_i, \beta \sim Gamma(\mu_i\beta, \beta).$$

We can see that in this model the variance increases with the mean. Such models are useful in settings with heteroskedastic errors, such as air pollution or insurance claims.

For an illustrative application and implementation of the Gamma GLM with log link see Example 3 in accompanying dynamic report.

4 Multinomial logistic regression

Logistic regression can also be interpreted as a latent strength (utility) model. Our observations are discrete (dichotomous) choices between, say, objects A and B, but we assume

that they are made based on some unobserved (latent) continuous strengths of the two objects being compared, $u_A, u_B \in \mathbb{R}$. The classical model for such pairwise comparison is the Bradley-Terry model, which relates the latent strengths to the probability with

$$P(\text{A is chosen}) = \frac{e^{u_A}}{e^{u_A} + e^{u_B}}.$$

The exponentiation here is necessary to ensure that the expression is between 0 and 1. An important observation here is that

$$\frac{e^{u_A}}{e^{u_A} + e^{u_B}} = \frac{e^{u_A+c}}{e^{u_A+c} + e^{u_B+c}},$$

for any $c \in \mathbb{R}$. That is, this model is not identifiable - for every solution there are infinitely many equally good solutions. In this model only the difference between strengths matters, not the absolute values, so we can, without loss of generality, set one of the latent strengths to 0, effectively choosing that object as the reference object. Let's do that for object B:

$$P(\text{A is chosen}) = \frac{e^{u_A}}{e^{u_A} + e^0} = \frac{e^{u_A}}{e^{u_A} + 1} = \frac{1}{1 + e^{-u_A}}.$$

If we make the latent strength depend on the context and assume a linear relationship $u_i = \beta^T x_i$, we arrive at the logistic regression model. So, logistic regression can be interpreted as a discrete choice model between two objects.

This latent strength approach can be generalized to more than 2 objects. Let m be the number of objects we are choosing from, let their latent strengths be $u_{ij} = \beta_j^T x_i, j = 1..(m-1)$ and $u_{im} = 0$ (let the last object be the reference category[5]). Note that here index i goes over the observations and index j over the objects. Analogous to the 2 object case, the probabilities are

$$P(\text{object j is chosen in i-th observation}) = \frac{e^{u_{ji}}}{\sum_{l=1}^{m} e^{u_{li}}} = \frac{e^{u_{ji}}}{1 + \sum_{l=1}^{m-1} e^{u_{li}}},$$

except for the reference category object, where we have

$$P(\text{object m is chosen in i-th observation}) = \frac{1}{1 + \sum_{l=1}^{m-1} e^{u_{li}}}.$$

This model has $(m-1) \times k$ parameters, one set of coefficients for each category, except the reference category. The model is also a GLM where we assume that the target variable follows a categorical distribution and the link function is the generalized inverse logit, also known as the Softmax function. In statistical modeling notation:

$$y_i|\beta, x_i \sim \text{Categorical}(\text{softmax}(u_{1i}, u_{2i}, ..., u_{(m-1)i}, 0)),$$

where

[5]Interpretation depends on which object is chosen as the reference object.

$$\text{softmax}(x_1, x_2, ..., x_m) = [\frac{e^{x_1}}{\sum_{i=1}^{m} e^{x_i}}, \frac{e^{x_2}}{\sum_{i=1}^{m} e^{x_i}}, ..., \frac{e^{x_m}}{\sum_{i=1}^{m} e^{x_i}}]^T.$$

5 Ordinal logistic regression

Ordinal data are categorical data where there is a natural ordering to the categories. For example, questionnaire answers (strongly disagree, disagree, neutral, agree, strongly agree).

Categorical models such as multinomial logistic regression can be used on ordinal data and they will typically perform well or even better in terms of predictive quality when enough data are available. The advantages of using an ordinal model is that the ordering assumption simplifies the model (the model has fewer parameters), which makes it more robust and easier to interpret.

We again turn to logistic regression, this time from the ordinal perspective. Similar to the discrete choice latent strength, we assume that there is an underlying real value u_i that determines whether the outcome is a 0 or a 1 and we can make that value depend on the linear term $u_i = \beta^T x_i$, so that we have regression. Let's say that if u_i is greater than some threshold, the outcome will be 1, and 0 otherwise. As long as we have an intercept term in $\beta^T x_i$, the threshold can be set to any constant, because the intercept will translate the solution accordingly[6]. A convenient choice of threshold is 0.

What remains is to add some noise ϵ_i to u_i - if we do not, the prediction will be deterministic. And if we want to arrive at logistic regression, we have to go with standard logistic noise[7]. Note that noise can be zero mean for the same reason that the threshold can be 0. Furthermore, the noise can have unit variance, because we can already scale everything by multiplying all coefficients with the same constant. If we allowed the noise variance to be a parameter, we would again have a nonidentifiable model.

Putting it all together, we have:

$$P(\text{outcome is 1}) = P(u_i + \epsilon_i > 0) = P(\epsilon_i > -u_i) = 1 - P(\epsilon_i \leq -u_i) = 1 - F_\epsilon(-u_i),$$

where in our case the CDF F_ϵ is that of the standard logistic distribution - the inverse logit. Inserting the CDF, we get

$$P(\text{i-th outcome is 1}) = 1 - F_\epsilon(-u_i) = 1 - \frac{1}{1+e^{u_i}} = \frac{e^{u_i}}{1+e^{u_i}} = \frac{1}{1+e^{-u_i}}.$$

So, in the case of two categories, logistic regression, multinomial logistic regression, and ordinal logistic regression are the same model.

Now we generalize this idea to ordinal data with more than 2 levels. Let m be the number of levels. Now we require $m-1$ thresholds to subdivide the latent real variable into m levels. Again, without loss of generality, let the first threshold (from level 1 to level 2) be $t_1 = 0$, and the remaining $m-2$ thresholds are t_2,...,t_{m-1}. For convenience, we add $t_0 = -\infty$ and

[6] If we have an intercept and we also allow the threshold to be a parameter, then adding a constant c to the threshold and the intercept would lead to an equivalent solution. That is, the model would not be identifiable.

[7] If we chose standard normal noise, we would arrive at Probit ordinal regression.

$t_m = +\infty$. Now, same as before, the probability of a level is the probability that $u_i + \epsilon_i$ falls in between the two thresholds that correspond to the level.

5.1 Model summary

Our target variable is ordinal with k levels $y_i \in \{1, 2, ..., k\}$ and our independent variables are real vectors $x_i \in \mathcal{R}^k$, for $i = 1..n$.

The ordinal regression model (or ordered logit) is defined as follows:

$$y_i|t, \beta, x_i \sim \text{Categorical}(p_i),$$

where β is the vector of size m of coefficients and t is a $k+1$ vector of thresholds 0-indexed for convenience $t_0 = -\infty < t_1 = 0 < t_2 < t_3 < \cdots < t_{k-1} < t_k = \infty$.

The size k probability vector of probabilities for each of the k categories for the $i-$th observation is defined (component-wise) as:

$$p_i(j) = F(t_j - u_i) - F(t_{j-1} - u_i), j = 1..k$$

where F is the CDF of the standard logistic distribution and $u_i = \beta^T x_i$.

Some practical considerations:

- The thresholds have to be ordered $t_1 < t_2 < \cdots < t_{k-1}$. This is a constraint that is not trivial to maintain during optimization. Instead, use the *stick breaking* parametrization $t_0 = -\infty$, $t_1 = 0$, $t_2 = t_1 + \Delta_1$, ..., $t_{k-1} = t_{k-2} + \Delta_{k-1}$, $t_k = \infty$. That is, let Δ_i be the parameters and derive t_i from them. A simple box constraint $\Delta_i > \epsilon = 0$ will suffice to keep the t_i ordered.
- In theory, $\epsilon = 0$, in practice, however, we might, with certain datasets and optimization algorithms run into problems where a Δ is so small that we get 0 probability and $-\infty$ likelihood. Starting values for Δ are also important (standardizing independent variables X makes this easier).
- Convergence issues may also arise if we have perfect separability (two or more categories can be perfectly fit). In such cases infinitely many thresholds are optimal. And, of course, if we have perfect colinearity in X, as is the case with any linear model.

Readings

Chapter 3 of *James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.*

Chapters 3, 4, 8, and 9.2 of *Dobson, A. J., & Barnett, A. G. (2008). An introduction to generalized linear models. Chapman and Hall/CRC.*

Chapters 9.1-9.4 of *Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.*