General Regression Model

Scott S. Emerson, M.D., Ph.D.

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

January 5, 2015

Abstract

Regression analysis can be viewed as an extension of two sample statistical analysis models to an "infinite sample" setting in which groups being compared may be defined by each combination of values for multiple variables, including continuous variables. The most commonly used regression models (linear, logistic, Poisson, proportional hazards, and parametric accelerated failure time models) differ primarily in the functional (summary measure) of the distribution that is used as the basis of inference. In this document I describe the general notation shared by all of these models, and then I highlight a few of the issues unique to the individual methods. In all cases I restrict attention to the use of univariate (i.e., one response variable) regression models to investigate associations among multiple variables.

1 Notation for the Statistical Analysis Model

The general regression model involves:

1. Random variable Y is the "response" variable. We will estimate some summary measure of the distribution of Y within groups.
2. Random variable X is the "predictor of interest" (POI) variable. We will divide the population into subpopulations based on this variable. Each individual in a given subpopulation will have the same value of X as all other individuals in that subpopulation.
3. Random variables $W_1, W_2, \ldots, W_p$ are additional covariates that are used to adjust for confounding or to provide additional precision for inference. We will also use these variables to divide the population into subpopulations. Each individual in a given subpopulation will have the same value of each of the W_j's as all other individuals in that subpopulation.
4. θ is some summary measure of the distribution of Y. (The technical statistical term is "functional". It is something that can be computed from the probability distribution, but in the most general case it may be itself a function.) This will have to depend on the type of variable (e.g., binary, unordered categorical, ordered categorical, continuous). Common choices for continuous random variables include
   (a) Mean
   (b) Geometric mean (providing Y is always positive)
   (c) Median (or, less commonly, some other quantile such as 25th or 75th percentile)
   (d) Probability that $Y > c$ for some specified, scientifically relevant value of c
   (e) Odds that $Y > c$ for some specified, scientifically relevant value of c

(f) Hazard function (instantaneous rate of failure at a specified time, conditional on being at risk of failure at that time)

5. $\vec{\beta}$ is a $p+2$ dimensional vector of "regression parameters" (or borrowing the algebraic term, "regression coefficients").

6. η is a "linear predictor" that describes the combined effect of all regression predictors (X, W_1, W_2, $\ldots$, W_p) on θ. For specified values of those predictors, the exact form of η is taken to be

$$\eta = \beta_0 + \beta_1 X + \beta_2 W_1 + \beta_3 W_2 + \cdots + \beta_{p+1} W_p.$$

7. $g(\ )$ is some link function that "links" the linear predictor to θ. That is, we have "regression model"

$$g(\theta) = \eta = \beta_0 + \beta_1 X + \beta_2 W_1 + \beta_3 W_2 + \cdots + \beta_{p+1} W_p.$$

Although more varied choices of link functions are sometimes considered, our typical use of the link function is either

(a) to describe an additive model using an "identity link" ($g(\theta) = \theta$),

(b) the describe a multiplicative model using the "log link" ($g(\theta) = \log(\theta)$).

Interpretation of the regression parameters depends on the link function.

With an identity link:

1. β_0 is the value of θ within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, $\ldots$, $W_p = 0$ (i.e., all modeled covariates are zero).

   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.

2. β_1 is the value of the difference in θ between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, $\ldots$, W_p.

   - Depending on how the variables W_1, W_2, $\ldots$, W_p are defined, it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

3. β_j for $j > 1$ has interpretations analogous to that for $beta_1$: we consider differences in θ across groups that differ by one unit in the variable corresponding to β_j, but agreeing in their values for all other variables.

With a log link:

1. e^{β_0} is the value of θ within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, $\ldots$, $W_p = 0$ (i.e., all modeled covariates are zero).

   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.

2. e^{β_1} is the value of the ratio of θ between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, $\ldots$, W_p.

- Depending on how the variables W_1, W_2, ..., W_p are defined, it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
- When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

3. e^{β_j} for $j > 1$ has interpretations analogous to that for $beta_1$: we consider ratios of θ across groups that differ by one unit in the variable corresponding to β_j, but agreeing in their values for all other variables.

Note that in the above interpretations, we were pretending that the differences in θ (for an identity link) or the ratios of θ (for the log link) would be the same for every two subpopulations that differed by 1 unit in the corresponding variable. If this "linearity" assumption does not hold, we then interpret the slope parameters as some sort of average difference or average ratio across subpopulations differing by 1 unit in the corresponding variable.

1.1 Linear Regression

The linear regression model uses our general regression model with

1. θ is the mean of the distribution of Y.
2. $g(\ )$ is the "identity link" ($g(\theta) = \theta$).

Interpretation of parameters;

1. β_0 is the mean of Y within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, ..., $W_p = 0$ (i.e., all modeled covariates are zero).
   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.
2. β_1 is the value of the difference in the mean of Y between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, ..., W_p. Hence for some arbitrary x, w_1, w_2, ..., w_p

$$\beta_1 = E[Y \mid X = x + 1, W_1 = w_1, \ldots, W_p = w_p] - E[Y \mid X = x, W_1 = w_1, \ldots, W_p = w_p].$$

   - Depending on how the variables W_1, W_2, ..., W_p are defined (and the particular values of x, w_1, w_2, ..., w_p), it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

Estimates of the regression parameters uses "least squares". This corresponds to "maximum likelihood estimation" when the distribution of Y is Gaussian within subpopulations.

Optimality properties of linear regression include

- In the presence of independent observations, "homoscedasticity" (variances of Y within subpopulations are equal to each other) and the correct linear model for the means, the ordinary least squares estimates (OLSE) are "best linear unbiased estimates (BLUE)". This means that among all unbiased estimators that use linear combinations of the observed values of Y, the OLSE have the greatest precision. (In the presence of "heteroscedasticity", we should use weighted least squares estimates (WLSE).)
- In the presence of independent observations, homoscedasticity, the correct linear model for the means, and a Gaussian distribution of Y within subpopulations, the OLSE are MLE and are the efficient estimators of the means. Because we already know the OLSE are BLUE, this added fact just tells us that it is best to use linear combinations of the observed values of Y– nonlinear combinations of Y will not be better estimates.

The probability distribution of OLSE can be shown to be "asymptotically normal" under reasonable assumptions on the distribution of the predictors. That is, as the sample sizes get larger, the probability distribution for the OLSE gets closer and closer to a Gaussian distribution. If Y is normally distributed within subpopulations, the OLSE are exactly normally distributed no matter how small the sample size (so long as their is enough data to make the estimates).

1.2 Linear Regression on log Transformed Response Variables

This regression model is relevant only when Y is always positive. In this case, the linear regression model uses our general regression model with

1. θ is the geometric mean of the distribution of Y.
2. $g(\ )$ is the "log link" ($g(\theta) = \log(\theta)$).

Interpretation of parameters;

1. e^{β_0} is the geometric mean of Y within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, ..., $W_p = 0$ (i.e., all modeled covariates are zero).
   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.
2. e^{β_1} is the value of the ratio of geometric means of Y between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, ..., W_p. Hence for some arbitrary x, w_1, w_2, ..., w_p
$$e^{\beta_1} = \frac{GM[Y \mid X = x+1, W_1 = w_1, \ldots, W_p = w_p]}{GM[Y \mid X = x, W_1 = w_1, \ldots, W_p = w_p]}.$$
   - Depending on how the variables W_1, W_2, ..., W_p are defined (and the particular values of x, w_1, w_2, ..., w_p), it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

Estimates of the regression parameters uses "least squares". This corresponds to "maximum likelihood estimation" when the distribution of Y is lognormal (so the distribution of $\log(Y)$ is Gaussian) within subpopulations.

Optimality properties and probability distributions for the regression parameters of linear regression on log transformed response variables derive directly from those of linear regression.

1.3 Logistic Regression

This regression model is relevant only when Y is a binary 0-1 variable. In this case, the logistic regression model uses our general regression model with

1. θ is the odds that $Y = 1$.
2. $g(\ )$ is the "log link" $(g(\theta) = \log(\theta))$.

Interpretation of parameters;

1. e^{β_0} is the odds of $Y = 1$ within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, ..., $W_p = 0$ (i.e., all modeled covariates are zero).
   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.
2. e^{β_1} is the value of the ratio of the odds that $Y = 1$ between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, ..., W_p. Hence for some arbitrary x, w_1, w_2, ..., w_p

$$\begin{aligned} e^{\beta_1} &= \frac{odds[Y = 1 \,|\, X = x + 1, W_1 = w_1, \ldots, W_p = w_p]}{odds[Y = 1 \,|\, X = x, W_1 = w_1, \ldots, W_p = w_p]} \\ &= \frac{Pr[Y = 1 \,|\, X = x + 1, W_1 = w_1, \ldots, W_p = w_p]}{Pr[Y = 0 \,|\, X = x + 1, W_1 = w_1, \ldots, W_p = w_p]} \frac{Pr[Y = 0 \,|\, X = x, W_1 = w_1, \ldots, W_p = w_p]}{Pr[Y = 1 \,|\, X = x, W_1 = w_1, \ldots, W_p = w_p]}. \end{aligned}$$

   - Depending on how the variables W_1, W_2, ..., W_p are defined (and the particular values of x, w_1, w_2, ..., w_p), it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

Estimates of the regression parameters uses "maximum likelihood estimation" (MLE) based on the Bernoulli (or binomial) distribution for Y. As such, the variability of the regression parameter estimates is classically based on the mean-variance relationship dictated by the Bernoulli distribution: if $E[Y] = p$, then $Var(Y) = p(1 - p)$.

If the observations of Y are independent, and if the log odds across subpopulations adhere to the linear model, the MLE are "asymptotically efficient".

1.4 Poisson Regression

This regression model is relevant only when Y is nonnegative (it is classically defined for count data that has Poisson distributions within subpopulations).. In this case, the Poisson regression model uses our general regression model with

1. θ is the mean value of Y (if the data is truly Poisson count data, that mean is also an event rate).
2. $g(\ )$ is the "log link" ($g(\theta) = \log(\theta)$).

Interpretation of parameters;

1. e^{β_0} is the mean (event rate?) of Y within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, ..., $W_p = 0$ (i.e., all modeled covariates are zero).
   - This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the intercept.
2. e^{β_1} is the value of the ratio of the mean value of Y between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, ..., W_p. Hence for some arbitrary x, w_1, w_2, ..., w_p
$$e^{\beta_1} = \frac{E[Y \mid X = x+1, W_1 = w_1, \ldots, W_p = w_p]}{E[Y \mid X = x, W_1 = w_1, \ldots, W_p = w_p]}.$$
   - Depending on how the variables W_1, W_2, ..., W_p are defined (and the particular values of x, w_1, w_2, ..., w_p), it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

Estimates of the regression parameters uses "maximum likelihood estimation" (MLE) based on the Poisson distribution for Y. As such, the variability of the regression parameter estimates is classically based on the mean-variance relationship dictated by the Poisson distribution: if $E[Y] = \lambda$, then $Var(Y) = \lambda$. When used with the "robust" Huber-White sandwich estimator of the standard errors, this model can be viewed as a regression model for ratios of means across groups.

If the observations of Y are independent Poisson random variables, and if the log mean across subpopulations adhere to the linear model, the MLE are "asymptotically efficient".

1.5 Proportional Hazards Regression

This regression model is classically relevant only when Y is nonnegative (it is classically defined for time-to-event data that exhibits "proportional hazards" across subpopulations). In this case, the proportional hazards regression model uses our general regression model with

1. θ is the hazard function (instantaneous rate of failure) for Y. (This is a function that might vary over time.)
2. $g(\ )$ is the "log link" ($g(\theta) = \log(\theta)$).

Interpretation of parameters;

1. The intercept in this model is a hazard function $\lambda_Y(t)$ of Y within a subpopulation having $X = 0$, $W_1 = 0$, $W_2 = 0$, ..., $W_p = 0$ (i.e., all modeled covariates are zero). Standard statistical software does not automatically return this estimated function, because the slope parameters can be estimated without ever estimating this "baseline" hazard function.

- This is quite often outside the range of our data and sometimes not even scientifically possible. Most often we are not really interested in the baseline hazard function.

2. e^{β_1} is the value of the ratio of the hazard funtion of Y between two subpopulations that differ by 1 unit in their value of X, but agree in their values of W_1, W_2, ..., W_p. Under the proportional hazards assumption, the same ratio would be obtained at every point in time. Hence for some arbitrary x, w_1, w_2, ..., w_p

$$e^{\beta_1} = \frac{\lambda_Y(t \,|\, X = x+1, W_1 = w_1, \ldots, W_p = w_p)}{\lambda_Y(t \,|\, X = x, W_1 = w_1, \ldots, W_p = w_p)}.$$

   - Depending on how the variables W_1, W_2, ..., W_p are defined (and the particular values of x, w_1, w_2, ..., w_p), it may not be scientifically possible to have groups differing in their value of X and not differing in the value of some W_j. In that case we will have to find other interpretations of the slope parameter.
   - When it is scientifically possible and relevant to think about groups differing in their X value but agreeing in the value of all other modeled variables, this regression parameter is most often our measure of an association between Y and X.

Estimates of the regression parameters uses "maximum partial likelihood estimation" (MLE) based on the proportional hazards assumption for Y. This is somewhat related to logistic regression, and as such, the variability of the regression parameter estimates is classically based on the mean-variance relationship dictated in part by the Bernoulli distribution.