Notes on Lesson 1

I like to start from the simplest possible situation and then start building things up a little at a time. So, imagine that we have a collection of data with zero covariates. We would like to summarize those data and possibly calculate some properties about those data.

In most cases, we are interested in the mean of the data (represented by y). We often talk about "on average", so having a good estimate of the average is a good thing. We understand that our sample is a representation of an underlying population. Our goal is to make inference about the population. In our very first class of statistics, we calculated an estimate ($\hat{\mu}$) of the population average (μ) using the sample average:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Where we assume that our sample has n observations which we refer to as y_i for $i = 1, \cdots, n$.

This is a *non-parametric* estimate. That is, we simply used the sample mean to estimate the population mean. Using this approach, we have a good estimate of the sample mean, but really nothing more than that. We could also estimate the population variance (V) using the sample variance

$$\frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

Nice. Now, we have estimated two things about our population. We could even estimate the probability that outcomes exceed some particular value Y by looking at the proportion of times our sample values exceed that value

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i > Y)$$

If we are willing to assume that our data follow a particular distribution, we also can derive estimates of the mean, variance, and interesting probabilities of certain outcomes. This *parametric* approach is even more useful. In addition, we can assume that our outcome follows a particular distribution for each set of values of related covariates (predictors). That is, we can model a relationship between the values of the predictors and the mean of the outcomes.

We do this not knowing which distribution is best for our data. We choose distributions based on the nature of the outcome, and then we can compare different assumed distributions in terms of how well each model reproduces the data.

In count data models, the simplest distribution used to build regression models is the Poisson. We assume that for each set of values of the predictors, there is a population of possible response values that follows a Poisson distribution. We assume that the predictors are related to the mean of this distribution via the relationship $\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$. That is, the mean of the distribution of outcomes is conditional on the values of the predictors; it is the *conditional mean*. The variance of the outcomes for a particular set of values of the covariates (the *conditional variance*) is equal to the conditional mean; this is a property of the Poisson distribution.

Often, it is the case for observed data that the conditional variance is greater than the conditional mean. Such data are said to be overdispersed relative to data from a Poisson distribution. In such cases, we can consider alternative distributions that allow for greater variance. Here is a handy table to show the conditional variance (as a function of the conditional mean) allowed by various distributions

Distribution	**Conditional Variance**
Poisson	μ
Negative binomial 1	$\mu + \alpha\mu$
Negative binomial 2	$\mu + \alpha\mu^2$
Generalized Poisson	$\mu/(1-\alpha)^2$
Negative binomial P	$\mu + \ \alpha\mu^P$

Except for the Poisson distribution, all of the other distributions include an extra dispersion parameter α that dictates the distance away from the Poisson assumption of equidispersion. The dispersion parameter is required to be non-negative for the negative binomial 1, negative binomial 2, and negative binomial P distributions so that they all allow the conditional variance to be greater than the conditional mean (dispersion greater than allowed by the Poisson – a condition known as overdispersion relative to the Poisson). For each of these distributions, the conditional variance is equal to the conditional mean (equidispersion assumed by the Poisson) in the limit as $\alpha \to 0^+$. For the generalized Poisson distribution, the α parameter is allowed to be negative, 0, or positive (but less than 1). Respectively, the generalized Poisson distribution can have conditional variance that is smaller than the conditional mean (underdispersion relative to the Poisson), equal to the conditional mean (equidispersion), or greater than the conditional mean (overdispersion relative to the Poisson).

This list is far from exhaustive as there are many more distributions for which we could derive regression models. The Poisson-inverse Gaussian distribution has conditional variance like the negative binomial 2 distribution, but is more "heavy-tailed".

Note that any regression model for any of these distributions can be estimated for a collection of data. Whether that choice is the correct choice depends on the conditional variance of the data. That is, if the data are underdispersed, then estimating a negative binomial 2 regression model is going to assume variance that is greater than implied by the data. This will translate into larger estimated standard errors, lower test statistics, and p-values indicating no significant relationships where such relationships may actually be present (increased Type II error).

All count data models assume a linear predictor $\eta = X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ is related to the conditional mean via the log-link function. That is, the relationship is given by

$$\log(\mu) = \eta$$

or, equivalently, we can use the inverse of the log-link function to note that

$$\mu = \exp(\eta)$$

When you consider that the mean for all of these count data models assumes that all outcomes are non-negative, it makes sense that we utilize a function of the linear predictor that ensures positive values for the conditional mean.

Because count data regression models use the log-link, we must be careful how we interpret estimated coefficients. Recall in linear regression, we could make the following illustration to interpret β_1

Linear regression model cond. mean	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Conditional mean for $x_1 = k + 1$	$\beta_0 + \beta_1(k + 1) + \beta_2 x_2$
Conditional mean for $x_1 = k$	$\beta_0 + \beta_1(k) + \beta_2 x_2$
Difference	β_1

Thus, we can interpret β_1 as the change in the conditional mean of the outcomes when the x_1 covariate is 1 greater than a baseline value (given the data; equal values of x_2).

In count data models using the log-link (or exp-inverse-link), the interpretation is described by

Count regression model cond. mean	$y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$
Conditional mean for $x_1 = k + 1$	$\exp(\beta_0 + \beta_1(k + 1) + \beta_2 x_2)$
Conditional mean for $x_1 = k$	$\exp(\beta_0 + \beta_1(k) + \beta_2 x_2)$
Ratio	$\exp(\beta_1)$

That is, $\exp(\beta_1)$ is the ratio of (1) the incidence rate (conditional mean) when the x_1 covariate is 1 greater than a baseline value (given the data; equal values of x_2) to (2) the incidence rate (conditional mean) when the x_1 covariate is at its baseline value. Thus, $\exp(\beta_1)$ is the incidence rate ratio because it is the ratio of two incidence rates. It is the incidence rate ratio associated when the x_1 covariate is 1 greater than a baseline value (given the data; equal values of x_2).

Note that had we considered taking the difference like we did in linear regression, then there would be no simplification, and no easy interpretation of any of the parameters.

Remember: $\hat{\beta}$ is the estimated coefficient; $\exp(\hat{\beta})$ is the estimated incidence rate ratio.

Some researchers leave out the word incidence and refer to rate ratios.

Some researchers leave out the word ratio and refer to incidence rates.

Don't let either of these researchers confuse you!