*Topics for these notes*:

- *Ordinal logistic regression*
- *Mixture models*
- *Clump at zero data*
- *2-part models*
- *PROC NLMIXED*

*Associated reading*: *related topics in course notes.*

Ordinal logistic regression

Case study:

“Since 2001, 3 million soldiers have deployed to Southwest Asia (SWA), with exposure to inhalants that cause respiratory disease. Department of Defense uses standard occupational codes, termed Military Occupational Specialty (MOS), to classify military personnel by job/training. We characterized Marine MOS by estimated exposure to inhalational hazards. We developed an MOS-exposure matrix containing five major deployment inhalational hazards--sandstorms, burn pits, exhaust fumes, combat dust, occupational VDGF (vapor, dust, gas, fumes)--plus time worked outdoors. A 5 member expert panel of two physician deployment veterans and three occupational pulmonologists independently ranked 38 Marine MOS codes for estimated exposure intensity (3=high, 2=medium, 1=low) to each hazard.” From Pepper et al., 2017.

The MOS occupational codes (or MOS_num) are numbered 1 through 38, for convenience, but they relate to specific job types. For example, 1=personnel and administration, 2=intelligence, 3=infantry, etc.

Our data follows this form, for a given inhalation hazard:

| | MOS_num | | | | | |
|---|---|---|---|---|---|---|
| Rater | #1 | #2 | #3 | #4 | #5 | … |
| 1 | 1 | 1 | 3 | 2 | 1 | |
| 2 | 1 | 1 | 3 | 1 | 1 | |
| 3 | 1 | 2 | 3 | 2 | 1 | |
| … | | | | | | |

The outcome is ordinal and given that there are only 3 levels (3 is high exposure, 2 is medium, 1 is low), we consider a model that is specialized for this type of outcome.

A GzLMM that can be used to fit our data has the form

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid b_i, b_j)}{1 - P(Y_{ij} \le k \mid b_i, b_j)}\right] = \alpha_k + b_i + b_j \quad ,$$

where $i$=MOS_num, $j$=rater, and $k$ is outcome level; $\alpha_k$, $k$=1,…, $K$–1 are strictly increasing intercepts; $b_i$ and $b_j$ are random intercepts for MOS_num and rater, respectively.

In order to get estimates that are commensurate with increasing levels of the outcome, we can reverse the inequalities to obtain

$$\lambda^{c}_{ijk} = \log\left[\frac{P(Y_{ij} \ge k \mid b_i, b_j)}{1 - P(Y_{ij} \ge k \mid b_i, b_j)}\right] = \alpha_k + b_i + b_j \quad .$$

This is the model we will fit for the application. We achieve this model using a ‘descending’ option, discussed shortly.

Some questions of interest for our data:

(1) How do variances for raters compare with the variances over MOS types?
(2) Are there any raters that significantly differ from the group average?
(3) After adjusting for crossed random effects of MOS type and rater, what are the cumulative odds of low, medium, high exposure for a given inhalation hazard?
(4) What is the probability of a particular job of having a high exposure to a given exposure type?

To answer these questions, we can fit the ordinal logistic regression model shown on the last slide that accounts for multiple measures per MOS type (called *MOS_num* below), which is the experimental unit here (instead of subjects).

# SAS Code for one inhalation exposure source, burn pits:

```
proc glimmix data=all2 method=laplace;
 class mos_num rater;
 model burn_pits(desc) = / solution dist=multinomial
link=cumlogit;
   random mos_num rater / solution;  run;
```

The 'desc' option is added so that the direction of estimates and outcome levels are consistent.

The GLIMMIX Procedure

**Model Information**

| | |
|---|---|
| **Data Set** | WORK.ALL2 |
| **Response Variable** | Burn_Pits |
| **Response Distribution** | Multinomial (ordered) |
| **Link Function** | Cumulative Logit |
| **Variance Function** | Default |
| **Estimation Technique** | Maximum Likelihood |
| **Likelihood Approximation** | Laplace |
| **Degrees of Freedom Method** | Containment |

The Laplace method approximates the true likelihood, and hence considered ML estimation.

**Number of Observations Used** 184

**Response Profile**

| Ordered Value | Burn_Pits | Total Frequency |
|---|---|---|
| 1 | 3 | 13 |
| 2 | 2 | 56 |
| 3 | 1 | 115 |

**The GLIMMIX procedure is modeling the probabilities of levels of Burn_Pits having lower Ordered Values in the Response Profile table.**

This part of the output is a little confusing, but stems from our previous 'descending' choice. A lower 'ordered value' means a higher outcome value, so an intercept for Burn_Pits=2 means that the associated odds ratio will be for levels 2 or 3, relative to 1; the intercept for Burn_Pits=3 compares 3 versus 1 and 2.

**Covariance Parameter Estimates**

| Cov Parm | Estimate | Standard Error |
|---|---|---|
| MOS_num | 2.9181 | 1.2889 |
| rater | 0.7259 | 0.6157 |

The variance estimates indicate that the variability of the exposure estimates among job types (MOS_num) is 4 times greater than for the raters, which is probably reassuring to the raters.

**Solutions for Fixed Effects**

| Effect | Burn_Pits | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | 3 | -3.8512 | 0.6760 | 4 | -5.70 | 0.0047 |
| Intercept | 2 | -0.7868 | 0.5219 | 4 | -1.51 | 0.2062 |

The odds of a rater ascribing a job type as having medium or high exposure (relative to low) is exp(-0.7868)=0.46; the odds of high versus medium or low is exp(-3.8512)=0.02. Even with the Total Frequency table above, we see that 3's (i.e., 'High' exposure) are more rare.

**Solution for Random Effects**

| Effect | rater | MOS_num | Estimate | Std Err Pred | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| **MOS_num** | | **1** | -0.5945 | 0.9462 | 142 | -0.63 | 0.5308 |
| **MOS_num** | | **2** | 0.1359 | 0.8699 | 142 | 0.16 | 0.8761 |
| **MOS_num** | | **3** | 3.2523 | 1.0217 | 142 | 3.18 | 0.0018 |
| … | | | | | | | |
| **MOS_num** | | **73** | 0.09849 | 0.8591 | 142 | 0.11 | 0.9089 |
| **rater** | **Gottschall** | | -1.1538 | 0.5745 | 142 | -2.01 | 0.0465 |
| **rater** | **Kreft** | | 0.3367 | 0.4953 | 142 | 0.68 | 0.4977 |
| **rater** | **Meehan** | | 0.4260 | 0.4993 | 142 | 0.85 | 0.3951 |
| **rater** | **Pepper** | | 0.9793 | 0.5202 | 142 | 1.88 | 0.0618 |
| **rater** | **Rose** | | -0.08430 | 0.4930 | 142 | -0.17 | 0.8645 |

The random effect estimates make sense. For example, 1 is administrative, and the random effect estimate is below average…we would not expect administrative personnel to have high exposure to burn pits. However, we might expect Infantry (Mos_num=3) to have higher exposure to burn pits, which the raters also conclude.

- We see that Pepper scores job types higher, on average, with respect to Burn pit exposure, compared with the average rater; similarly, Gottschall scores lower. These both occur with marginal significance.

- From our ordinal logistic regression model, we note that $P(Y_{ij} \geq k \mid b_i, b_j) = \frac{1}{1+e^{-\lambda_{ijk}}}$ and $P(Y_{ij} \geq k \mid b_i = 0, b_j = 0) = \frac{1}{1+e^{-\alpha_k}}$. From the latter, we can estimate that for an average rater and MOS_num, the probability of 'high' classification is $1/(1+e^{3.8512}) = 0.02$.

- Job and rater-specific probability estimates can be obtained by using the first formula. We can also compute for specific MOS_num or raters, holding the other at its mean, since random effects are crossed. For example, for an average MOS_num the probability of a high classification for Gottschall is $1/(1+e^{-(-3.8512-1.15)}) = 0.7\%$, while for Pepper it is $1/(1+e^{-(-3.85+0.98)}) = 5.4\%$.

- We can get probabilities for any given level by computing the cumulative probabilities, and then taking differences [e.g., $P(Y=2)=P(Y\geq 2) - P(Y\geq 3)$.]

*Using the mixed-effects ordinal logistic regression for longitudinal data*

- We can generalize the formula for the mixed-effects ordinal logistic regression model so that it can be used for clustered / longitudinal data and include covariates. One such model that is useful for repeated measures within subjects (or subjects within clusters) is

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \le k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^r\boldsymbol{\beta} + \mathbf{z}_{ij}^r\mathbf{b}_i$$

where $i$ denotes subject, with measure $j$ (or subject $j$ in cluster $i$). Here, we have hierarchical data and so the random effects (as is usually done) are defined for the level 2 data (subjects).

- The previous model can be used for longitudinal ordinal logistic regression, although we only account for repeated measures via random effects. (Using pseudo-likelihood methods, you could consider models that account for random effects or serial correlation, or both.)

- Now we have what is called a proportional odds model (see McCullagh, 1980) that results from the fact that the relationship between the cumulative logit and the predictors does not depend on $k$.

- For example, say that the previous case study also had measurements over time (x=time). If we added this as a predictor, then the cumulative logits (and hence probabilities) would not change over time.

- We can generalize the model slightly so that for certain predictors, we do not require the proportional odds assumption.

- For example, Hedeker and Mermelstein (1998, 200) suggest the model

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \le k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^r\boldsymbol{\beta} + \mathbf{s}_{ij}^r\boldsymbol{\gamma}_k + \mathbf{z}_{ij}^r\mathbf{b}_i$$

  where the additional term involving $\boldsymbol{\gamma}_k$ allows the effects for the associated covariates to vary across the cumulative logits.

- For more detail, see the above references or Hedeker and Gibbons (2006). Hedeker does warn about use of this partial proportional odds model, with respect to inference for certain values of the covariates. For more detail, see Hedeker and Gibbons (2006).

*Mixture distributions*

- Some distributions are more complex and cannot be modeled well using standard methods. For example, some distributions have possibility of 0 but where positive values can be well-modeled as continuous. Some examples:
  - Health care costs.
  - Precipitation amounts.

- Such a distribution is a discrete and continuous mixture, so both aspects need to be accounted for properly.

- Some possible distributions for the continuous part would be: lognormal, gamma, Weibull, truncated normal.

- With the cost example, some potential interesting questions are:
  - What is the chance someone will incur a cost?
  - What is the mean of costs for those who do?
  - What is the overall mean, taking into account both sources (those who have some costs, and those who don't)?

- With the precipitation example, equivalent questions are:
  - What is the probability of rain in a given city?
  - What is the mean rainfall on days when it did rain?
  - What is the overall mean rainfall in each city?

- We can use the Theorem on Total Probability to derive the complete '0+continuous' distribution. Let $R$ denote an indicator variable for positive values of $Y$ and let $p$ denote the probability of a positive value. Then we have

$$\begin{aligned} F_Y(y) &= P(Y \le y \mid r = 0)P(R = 0) + P(Y \le y \mid r > 0)P(R = 1) \\ &= P(Y \le y \mid r = 0) \cdot (1 - p) + P(Y \le y \mid r = 1) \cdot p \\ &= (1 - p) \cdot I_{\{y=0\}} + p \cdot F_{Y|r=1}(y \mid r = 1) \end{aligned}$$

where $F_{Y|r=1}(y \mid r = 1)$ is the CDF of a random variable with positive density for positive values of $Y$ (e.g., Weibull, gamma, log Normal).

- Although the CDF is easier to work with for mixed distributions, we need the PDF for the likelihood. The form can be defined mathematically as

$$f_Y(y) = (1-p)\cdot\delta_0(y) + p\cdot f_{Y|r=1}(y \mid r=1)$$

where $\delta_0(y)$ is the Dirac delta function, defined to be 0 when $y \neq 0$ but integrates to 1 over all $y$ on the real line. For practical purposes (e.g., in the likelihood function) we set this term to 1 so that

$$f_Y(y) = (1-p)\cdot I_{\{r=0\}} + p\cdot f_{Y|r=1}(y \mid r=1)$$

- In our example with rainfall, we'll consider the gamma distribution for positive amounts (i.e., given $R$=1), which has density

$$f_Y(y) = \frac{x^{\theta-1}e^{-x/\lambda}}{\lambda^{\theta}\Gamma(\theta)} \qquad \text{for } y>0,$$

  where $\theta>0$ is a shape parameter and $\lambda>0$ is a scale parameter.

- The mean of this distribution is $\theta\lambda$ and the variance is $\theta\lambda^2$.

- We can define a mixed model for this mixed distribution as follows.
  - Occurrence model: $\text{Logit}\left(p_{ij}\right) = \alpha_0 + b_{0i}$, where $p_{ij} = P(Y_{ij} = 1 \mid b_{0i})$.
  - Intensity model: $Y_{ij} \mid r = 1, b_{1i}, b_{2i} \sim \text{Gamma}\left(\theta + b_{1i}, \lambda + b_{2i}\right)$
  - Covariance structure of random effects: $\mathbf{b} = (b_{0i}, b_{1i}, b_{2i})^t \sim N(\mathbf{0}, \mathbf{G})$, where **G** is a 3x3 unstructured matrix.
- The addition of random intercepts to both shape and scale parameters for each city allows unique gamma rainfall distribution by city.

*Deriving the mean, variance and covariance*

- For our model, $E(Y_{ij} \mid \mathbf{b}) = E[E(Y_{ij} \mid \mathbf{b}, r)] = 0(1 - p_{ij}) + \mu_{ij+} p_{ij} = p_{ij}\mu_{ij+}$ where $\mu_{ij+}$ is the mean of the positive $Y$ values and $p_{ij}$ is the probability of a positive value. Now

$$p_{ij} = \frac{1}{1 + \exp(-\alpha_0 - b_{0i}^{p})} \text{ and } \mu_{ij+} = (\theta + b_{0i}^{shape})(\lambda + b_{0i}^{scale})$$

(where conditioning on random effects is implied). Thus the mean that puts the 0's and positive data together is

$$E(Y_{ij} \mid \mathbf{b}) = \frac{(\theta + b_{0i}^{shape})(\lambda + b_{0i}^{scale})}{1 + \exp(-\alpha_0 - b_{0i}^{p})}.$$

- It may be just as meaningful to jointly report $p_{ij}$ and $\mu_{ij+}$, which represent the probability of rain or snow on day $j$ for city $i$ and the average precipitation over days when it did rain/snow.

- We can also derive $Var(Y_{ij} \mid \mathbf{b}) = p_{ij}(\sigma^2_{ij+} + \mu^2_{ij+}(1 - p_{ij}))$, where $\sigma^2_{ij+}$ is the variance of the positive $Y$ values (show for homework).

- Covariance and correlation:

$$Cov(Y_{ij}, Y_{ik}) = E(Cov(Y_{ij}, Y_{ik} \mid \mathbf{b})) + Cov(E(Y_{ij} \mid \mathbf{b}), E(Y_{ik} \mid \mathbf{b}))$$

  - The second term is straightforward to determine since we have already defined the model in terms of mean responses given the random effects.
  - The first term is more difficult since $Cov(Y_{ij}, Y_{ik} \mid \mathbf{b}))$ does not come directly from the defined model. Specifically, no error term is defined for the model. (For an LMM, it would be the $(i,j)^{th}$ element of the error covariance matrix, $\mathbf{R}_i$.) We could employ residuals to estimate the quantity. Check.

- For the (straightforward) term on the right side,

$$Cov(E(Y_{ij} \mid \mathbf{b}), E(Y_{ik} \mid \mathbf{b})) = Cov\left(\mu_{ij+} p_{ij}, \mu_{ik+} p_{ik}\right)$$

$$= Cov\left(\frac{(\theta + b_{1i})(\lambda + b_{2i})}{1 + \exp(-\alpha_0 - b_{0i})}, \frac{(\theta + b_{1i})(\lambda + b_{2i})}{1 + \exp(-\alpha_0 - b_{0i})}\right)$$

where $\mathbf{b} = (b_{0i}, b_{1i}, b_{2i})^t \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{G} = \begin{pmatrix} \sigma^2_{b_0^p} & & \\ \phi_{21}\sigma_{b_0^p}\sigma_{b_0^{shape}} & \sigma^2_{b_0^{shape}} & \\ \phi_{31}\sigma_{b_0^p}\sigma_{b_0^{scale}} & \phi_{32}\sigma_{b_0^{shape}}\sigma_{b_0^{scale}} & \sigma^2_{b_0^{scale}} \end{pmatrix}$

- Here we use an unstructured G matrix and formulate the model so that the correlation parameters are directly estimated. Note that covariances depend on city *i* but not day *j* since we only have random intercepts in the model (but we keep both subscripts on parameters for potential generalizations). To make a time-dependent structure, we could add fixed and random effects for day in the model (in the intensity and/or occurrence parts). For models I tried it did not seem to help.

- Here is the analysis of rainfall data in 6 cities selected from across the U.S. Note that this is more of a demonstration of methods; cities were not randomly selected and more advanced time-series models might be used for actual analysis. But it is real data and the modeled distribution appears to fit the data well.
  - The cities: Atlanta, Aurora, Chicago, Houston, New York, Phoenix, Sacramento, Seattle. These are the 'subjects'.
  - Data collection time frame: first 100 days of 2017.
  - Outcome variable: precipitation, measured in inches.

- SAS code and output follow. Note that the names given in the SAS code are consistent with the quantities shown above, just written out instead of in Greek symbols.

```
PROC NLMIXED DATA=precip_data2 qpoints=5 absfconv=0.0000001;

PARMS ALPHA0=-0.8 SHAPE_MEAN=1 SCALE_MEAN=0.58
  VARB0_P=0.5 VARB0_SHAPE=0.05 VARB0_SCALE=0.05 PHI21=0.1 PHI31=-0.1 PHI32=-0.4;
BOUNDS VARB0_P VARB0_SHAPE VARB0_SCALE >=0;

SHAPE=SHAPE_MEAN+B0_SHAPE;
SCALE=SCALE_MEAN+B0_SCALE;
MULOGIT=ALPHA0+B0_P;
P=1/(1+EXP(-MULOGIT));

IF PRECIP=0 THEN LOGLIKE=LOG((1-P));
ELSE LOGLIKE=LOG(P)+(SHAPE-1)*LOG(PRECIP)-SHAPE*LOG(SCALE)
             -LOG(GAMMA(SHAPE))-(PRECIP/SCALE);

MODEL precip~GENERAL(LOGLIKE);
RANDOM B0_P B0_SHAPE B0_SCALE ~ NORMAL([0,0,0],
 [VARB0_P, PHI21*(VARB0_P*VARB0_SHAPE)**.5, VARB0_SHAPE,
  PHI31*(VARB0_P*VARB0_SC)**.5, PHI32*(VARB0_SHAPE*VARB0_SC)**.5, VARB0_SC])
 SUBJECT=city out=randout;

predict p out=p;
predict SHAPE out=SHAPE;
predict SCALE out=SCALE;run;
```

The NLMIXED Procedure

| Specifications | | Dimensions | |
|---|---|---|---|
| Random Effects | B0_P B0_SHAPE B0_SC | Observations Used | 800 |
| Distribution for Random Effects | Normal | Subjects | 8 |
| Subject Variable | city | Max Obs Per Subject | 100 |
| Optimization Technique | Dual Quasi-Newton | Parameters | 9 |
| Integration Method | Adaptive Gaussian Quadrature | Quadrature Points | 5 |

Parameter Estimates

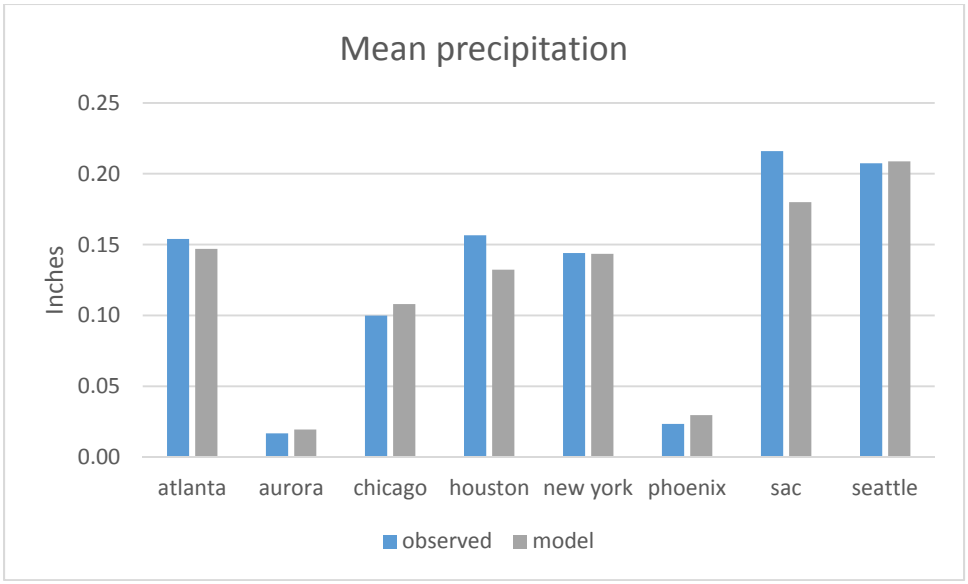
| Parameter | Estimate | SE | DF | t Value | Pr>\|t\| | Lower | Upper | Gradient |
|---|---|---|---|---|---|---|---|---|
| ALPHA0 | -0.8370 | 0.2810 | 5 | -2.98 | 0.0308 | -1.5593 | -0.1147 | -0.00115 |
| SHAPE_MEAN | 0.7085 | 0.07268 | 5 | 9.75 | 0.0002 | 0.5216 | 0.8953 | 0.007636 |
| SCALE_MEAN | 0.5458 | 0.1008 | 5 | 5.41 | 0.0029 | 0.2866 | 0.8050 | 0.002785 |
| VARB0_P | 0.5749 | 0.3250 | 5 | 1.77 | 0.1371 | -0.2605 | 1.4103 | -0.0006 |
| VARB0_SHAPE | 0.008 | 0.01025 | 5 | 0.78 | 0.4682 | -0.01832 | 0.03441 | 0.03419 |
| VARB0_SCALE | 0.0474 | 0.03798 | 5 | 1.25 | 0.2676 | -0.05027 | 0.1450 | 0.000624 |
| PHI21 | 0.1917 | 0.9727 | 5 | 0.20 | 0.8515 | -2.3086 | 2.6920 | 0.001991 |
| PHI31 | 0.2855 | 0.4463 | 5 | 0.64 | 0.5506 | -0.8618 | 1.4327 | 0.001384 |
| PHI32 | -0.8628 | 0.3944 | 5 | -2.19 | 0.0803 | -1.8767 | 0.1511 | -0.00495 |

- We could also add a fixed or fixed and random effect for *day* in either the Occurrence or Intensity models (or both), which would induce a slightly more time-sensitive correlation structure. However for the data at hand, such additions did not improve the model fit.

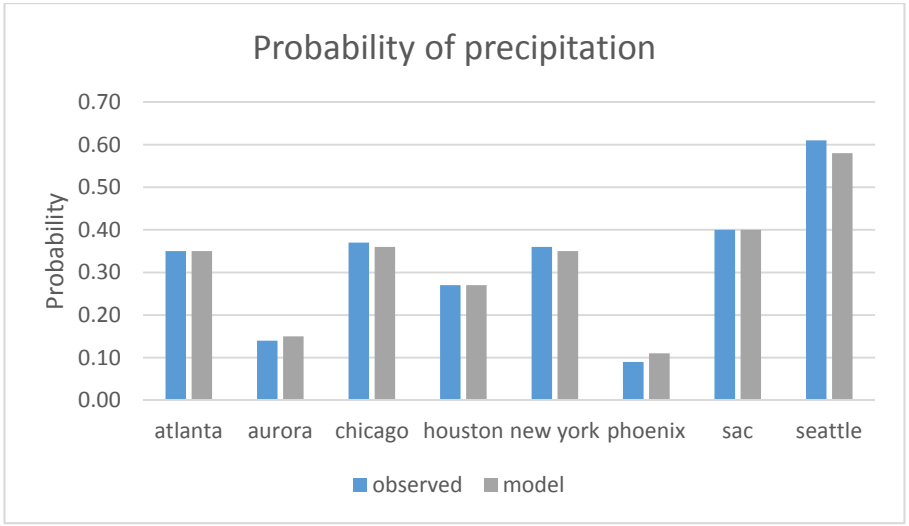
- Predicted values that include random effect variations are obtained by the 'predict' statements given at the end of the SAS code. Here is some additional code that gets quantities of interest ( $p_{ij}$ , $\mu_{ij+}$ , and $\mu_{ij}$ )

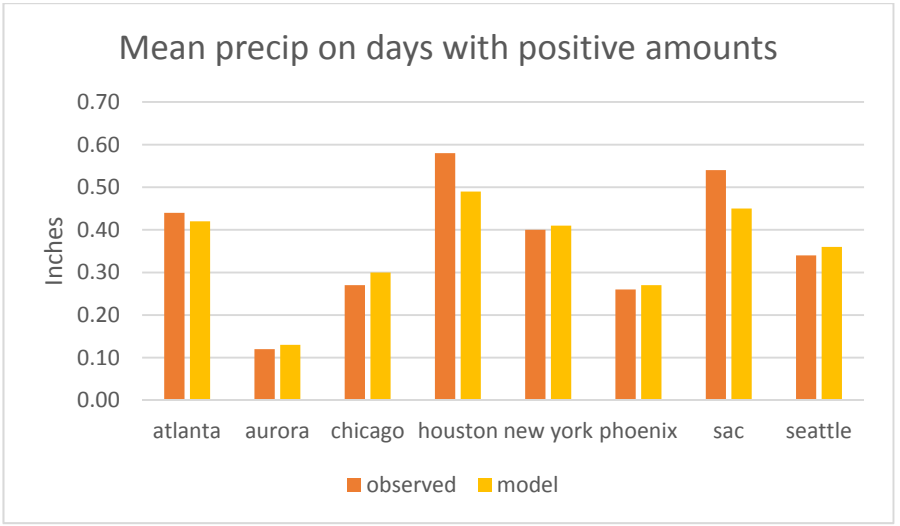
```
data p; set p; rename pred=pred_p;
data SHAPE; set SHAPE; rename pred=pred_SHAPE;
data SCALE; set SCALE; rename pred=pred_SCALE;
data all; merge p SHAPE SCALE; mean=pred_SHAPE*pred_SCALE; run;
proc means data=all; var pred_p mean; by city; run;
```

- The graph below shows overall means for each city using both descriptive statistics and the model-based approach. They are generally in agreement.



- The graphs below show how modeled values of $\mu_{ij} \mid b_{1i}, b_{2i}$ and $p_{ij} \mid b_{0i}$ versus descriptive quantities. These graphs demonstrate how information is lost when we only consider mean precipitation as in the last graph. For example, Seattle has greater likelihood of rain on any given day, but when we restrict to days where it did rain, Sacramento and Houston had higher mean daily precipitation amounts. Note that modeled amounts by city tend to exhibit shrinkage towards to overall mean (a bit higher for drier cities; less for wetter cities) which is expected for empirical Bayes estimates.



- Important note: historical data might yield somewhat different results. Here we only used the early part of 2017 to build the model, so inference should be restricted to 'around this time or times that are similar in climate' and during winter.

- One might wonder why the model approach is of any use, since we can estimate these quantities directly. Remember that we additionally can address correlation in our model. Also, the modeling approaches allows for the addition of other covariates and random effects, if we so wish.

- Given that probabilities and means were not time sensitive in our given model, the correlation between responses should be somewhat like the compound symmetric structure.

Additional thoughts.

- Mixture distributions may be useful even if the distribution is completely discrete or continuous. For example, a zero-inflated Poisson distribution takes a standard Poisson and then adds a binomial random variable such that the probability that the mixed random variable takes on a value of 0 is increased.

- We can also define mixture models based on how values in the mixture can be distinguished with respect to structure and sampling.

- Hurdle models versus zero-inflated models

  - From McDowell (2003): A hurdle model is "a modified count model in which the two processes generating the zeros and the positives are not constrained to be the same" (Cameron and Trivedi 1998). Mullahy (1986) states, "The idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome of whether a count variate has a zero or a positive realization. If the realization is positive, the "hurdle is crossed", and the conditional distribution of the positives is governed by a truncated-at-zero count data model.

  - A zero-inflated model is one where the 0's could come from 2 different types of processes (structural and sampling), and the 0's versus nonzero's are not governed by one overlying Bernoulli process. So, for example, we have a Poisson process, which could include 0's and positive integers, but then is also a structural source for the 0's.

- As an example, consider a type of number of packs of cigarettes smoked in the last week. For smokers, most will likely smoke, but there is the chance that some will not; these will be ‘sampling 0’s’; but if the cohort also includes non-smokers, then those would be structural 0’s since, by definition, they do not smoke. This would be an example of a zero-inflated model.

- However, say that the time frame considered is much longer, like 3 months. In this case, it may be reasonable to assume that 0’s only come from non-smokers and positive values come from smokers. We might use a hurdle model in this case.

- Consider a model that needs to account for added 0's (either zero inflated, or via hurdle model). For simplicity of notation, let $p = P(Y = 0 \mid z, \gamma)$. Also *f* may represent either a pdf or pmf, depending on whether the distribution of positive values is continuous or discrete.
  - For a zero-inflated model, we have

$$f_{ZIP}(x, z, \beta, \gamma) = p_{z,\gamma} I_{\{0\}}(y) + (1 - p_{z,\gamma}) f_{count}(y \mid x, \beta) I_{\{0,1,2,\ldots\}}(y)$$

  - For a hurdle model, we have

$$f_{hurdle}(x, z, \beta, \gamma) = \begin{cases} p_{z,\gamma} & y = 0 \\ (1 - p_{z,\gamma}) \dfrac{f_{count}(y \mid x, \beta)}{(1 - f_{count}(0 \mid x, \beta))} & y > 0 \end{cases}$$

- The primary difference between models is that for the ZIP model, we have a standard distribution ($f_{count}$), such as a Poisson distribution and add some 0's to it, while for the hurdle model, we distinguish modeling of the 0's versus modeling of the positive values based on their structural differences. In order to model the positive values, we take a standard distribution like the Poisson and truncate it so that a value of 0 has no positive probability/mass.

- Going back to the rainfall application, we combined a discrete and continuous model, the latter of which already does not have any probability mass on 0 (no need to truncate it). In this sense we intrinsically have a hurdle model. It may also make sense theoretically if there are not 'structural' and 'sampling' 0's.

- However a zero-inflated model might make sense theoretically if there is some condition considered. For example, clouds must be present for rain or snow. But precipitation is not guaranteed when clouds are present. Thus, 0's could be distinguished by those on sunny (structural) and cloudy (sampling) days. One model governs rainfall when it is cloudy, and one whether it is cloudy or sunny. For the 'cloudy' model, we'd need some distribution that allows positive probability for 0 but also for positive values. A count-type model might work if we categorize the precipitation levels.

In some cases we may not need to consider the theoretical constructs of zero and nonzero values. We may use a model and be more concerned with how accurate the distribution is, and not estimate parameters based on distinguishing sampling versus structural-based zeroes.