Regression with R

Chris Johnson

12/3/22

Table of contents

Preface

This book is a work in progress. It is primarily intended for myself. References will be formalized soon.

1 Likelihood

Probability function vs. likelihood function

Same function, but different knowns and unknowns.

Let X be a random variable, and x be an observed set from X. Similarly, let θ be a parameter set.

Consider the Binomial probability function which generates a probability provided the number of trials, the probability of observing a 1, and the number of successes.

For probability, n and p are known, and k is unknown. $k \in \{0, 1, ..., n\}$. Plugging in all values of k generates a probability distribution.

$$\Pr(X = k) = \binom{8}{k} 0.5^k (1 - 0.5)^{8-k}$$

The interpretation for this distribution is the probability for the number of 1s (successes) given we observed k ($k = 1$, $k = 2$, etc.).

We see the probability function is a function of k—what's observed (generally $X = x$, but commonly $X = k$ for Binomial).

For likelihood, n and k are known, and p is unknown. Recall p is the probability of success, and $p \in \{0, 1\}$. This situation is now the following: We did an experiment: n trials, and we observed a single value of k (the number of successess that occurred during the n trials).

The likelihood function is a function of p; the RHS of the function is the same as the probability function:

$$\mathrm{L}(p) = \binom{8}{4} p^4 (1 - p)^{8-4}$$

For the Binomial distribution, the probability function is discrete whereas the likelihood function is continuous. In fact, the likelihood function is always continuous. It is $\mathrm{L}(p)$ (likelihood) vs. p.

Likelihood: What's the likelihood that the parameter is θ? Probability: What's the probability the event $X = x$ occurs?

Likelihood: You've observed data, determined a candidate distribution, and want to determine the parameter value that is most likely (i.e. maximizes the likelihood).

Note:

$$\Pr(X = 4) = \binom{8}{4} 0.5^4 (1 - 0.5)^{8-4}$$
$$\mathrm{L}(p = 0.5) = \binom{8}{4} 0.5^4 (1 - 0.5)^{8-4}$$

Probability: You've observed data, and have determined a candidate distribution and determined a set of parameter values (possibly by likelihood, maximum likelihood, etc.), and want to determine how probable an event is.

Likelihood: distribution parameter Probability: event

Maximum likelihood estimator (MLE)

2 Multiple linear regression

Interaction term: `x1 * x2`

3 Logistic regression

3.1 Prerequisite knowledge

The *odds* of an event is a function of probability, defined as

$$\frac{p}{(1-p)}$$

where p is the probability of the event occurring, and $1-p$ is the probability of the event *not* occurring. The ratio of the odds for two events is the *odds ratio* (OR). The odds in the denominator is the *reference event*. *Even odds* occurs when the OR is 1, and means both events are equally likely. A positive OR means the event is more likely to occur, whereas a negative OR means the event is less likely to occur.

3.2 Interpreting coefficients

For categorical variables, the OR compares the odds of the non-reference category to the odds of the reference category, with all other variables held constant.

Lucy Dickinson published an article on Medium titled "How to Interpret the Odds Ratio with Categorical Variables in Logistic Regression". In this article, she fit a logistic regression model to attempt to study the effects of customer type, gender, day type, age category, and day of week on length of bike rides. She defined long bike rides as those being greater than 20 minutes and created a dichotomous response variable based on recorded length of bike rides.

Reference levels were subscribers (`user_type`), males (`gender`), weekdays (`daytype`), over-30s (`under30years`), and Wednesdays (`weekday`).

The results were

predictor	estimate
`const`	-1.5459
`user_type_customer`	1.6643
`gender_female`	0.3207
`daytype_Weekend`	0.0027

predictor	estimate
`under30years_under30`	-0.0780
`weekday_Friday`	-0.0041
`weekday_Monday`	0.0551
`weekday_Saturday`	0.0398
`weekday_Sunday`	-0.0432
`weekday_Thursday`	-0.0647
`weekday_Tuesday`	-0.0149

The coefficients are interpreted relative to the reference levels and are natural-log-odds.

The odds of a ride exceeding 20 minutes is 37% higher if you are female ($e^{0.3207} \approx 1.37$). To make this clear, the log-odds ratio of females to males for a long bike ride is

$$\ln\left(\frac{\text{odds}_{\text{female}}}{\text{odds}_{\text{male}}}\right) = 0.3207$$

So

$$\begin{aligned} e^{\ln\left(\frac{\text{odds}_{\text{female}}}{\text{odds}_{\text{male}}}\right)} &= e^{0.3207} \\ \frac{\text{odds}_{\text{female}}}{\text{odds}_{\text{male}}} &\approx \frac{1.37}{1} \end{aligned}$$

This odds ratio translates to females' odds being 37% higher than the odds of males for taking a long bike ride.

The odds of a ride exceeding 20 minutes is 8% lower if you are under 30 years of age:

$$\begin{aligned} \ln\left(\frac{\text{odds}_{\text{under 30}}}{\text{odds}_{\text{over 30}}}\right) &= -0.0780 \\ \frac{\text{odds}_{\text{under 30}}}{\text{odds}_{\text{over 30}}} &= e^{-0.0780} \\ &\approx \frac{0.92}{1} \end{aligned}$$

The odds of a ride exceeding 20 minutes are approximately even for weekdays and weekend days: $e^{0.0027} \approx \frac{1.003}{1}$. (Technically, the odds of a long bike ride are 0.3% higher for weekdays, which is in contrast to Lucy's hypothesis that longer bike rides would occur on weekends.)

3.3 Converting odds ratio to probability

To convert the odds ratio to a probability is straightforward:

$$\begin{aligned} \text{OR} &= \frac{p}{(1-p)} \\ \text{OR} \cdot (1-p) &= p \\ \text{OR} - \text{OR} \cdot p &= p \\ \text{OR} &= p + \text{OR} \cdot p \\ \text{OR} &= p(1+\text{OR}) \\ \frac{\text{OR}}{(1+\text{OR})} &= p \end{aligned}$$

So to convert the odds ratio to a probability, we use $p = \frac{\text{OR}}{(1+\text{OR})}$. The probability is still interpreted with respect to the reference level. Recall that

> the odds of a ride exceeding 20 minutes is 8% lower if you are under 30 years of age

was based on $\frac{\text{odds}_{\text{under 30}}}{\text{odds}_{\text{over 30}}} \approx \frac{0.92}{1}$. To put this in terms of probability:

$$\begin{aligned} p &= \frac{0.92}{1+0.92} \\ &\approx 0.48 \end{aligned}$$

which translates to the probability of a ride exceeding 20 minutes is 48% lower if you are under 30 years of age.

3.4 Logistic regression in R

4 Poisson regression

Poisson regression is based on the exponential function $y = e^{b_0+b_1\times x}$. The exponent of the exponential function is linear, and linear regression requires the RHS to be linear, so a simple log transformation will

$$y = e^{b_0+b_1\times x}$$
$$\ln(y) = b_0 + b_1 \times x$$

Poisson regression: Modeling rate data and the offset

Pt. 1 notes are on the whiteboard. Notes adapted from TileStats

Data: Number of births of rabbits in spring, weekly, in a certain area.

General trend: Birth rate is higher in early spring and lower in late spring. Counts for this dataset were observed using a fixed time interval.

Amount of time	Count
0.85	90
1.10	76
0.85	37
1.20	27
0.95	19
1.10	13
0.90	9

Note: 0.85 weeks is approximately 142 hours (168 hours in a 7-day week).

Rates are counts per time. In a situation where the time intervals differ, calculate rates.

To use rates in the Poisson regression framework, we model

$$\ln\left(\frac{y}{t}\right) = b_0 + b_1 \times \text{time}$$

where y is the count and t is the time interval. The Poisson regression framework is expecting counts, so use one of the laws of logarithms:

$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)$

So our model becomes

$$\begin{aligned} \ln\left(\frac{y}{t}\right) &= b_0 + b_1 \times \text{time} \\ \ln(y) - \ln(t) &= b_0 + b_1 \times \text{time} \\ \ln(y) &= b_0 + b_1 \times \text{time} + \ln(t) \end{aligned}$$

Now the counts are isolated on the LHS. $\ln(t)$ on the RHS is the *offset*. Note: The offset is *not a parameter*, so it doesn't require estimation. It is simply an additive adjustmen tto account for unequal time intervals.

Let's switch to a new example: Cancer cases over a certain time period for a given population.

Age range	Number of cases
40–59	30
60–79	31
80+	29

Subject matter expertise might suggest more cancer cases in the older age groups. In other words, we might expect moer cancer cases in the elderly populations. An explanation for the observed data where the number of cases are similar for each age group (contrary to SME expectations) might be population.

There are more individuals in the 40–59 population, so more opportunities for cancer to occur, and in turn, more opportunities for cancer to be observed.

To adjust, we compute *cases per population* (i.e. compute a rate).

A model for this dataset would be

$$\ln(y) = b_0 + b_1 \times \text{Age}_{\text{60-79}} + b_2 \times \text{Age}_{\text{80+}} + \ln(n)$$

where y is the number of cases, n is the population size. This model uses the age group 45–59 as the reference level. In this example, $n = (\text{pop}_{\text{40-59}}, \text{pop}_{\text{60-79}}, \text{pop}_{\text{80+}})$.

Important: We can't model rates directly because the information about the number of counts and population size is lost in the process of computing rates. We must model the counts, and if the unit (time, area, population) isn't equal for each group, we must include an offset.

Notes about the explanatory variables on a categorical scale require discussion as well.

Dataset: Lymph nodes. The human body has approximately 500 lymph nodes. The number of metastatic lymph nodes is a prognostic factor because it is associated with the progression of cancer. Let A and B be cancer treatments. 8 patients, four per treatment group.

Counts A	Counts B
7	3
4	1
4	1
2	0

or in long form

Treatment	ID	Count
A	1	7
A	2	4
A	3	4
A	4	2
B	5	3
B	6	1
B	7	1
B	8	0

When a variable is a factor, it must be recoded. Let Treatment A be our baseline. Then the model becomes

$$\ln(y) = b_0 + b_1 \times \text{Treatment}_\text{B}$$

where $\text{Treatment}_\text{B}$ can be 0 (for no) or 1 (for yes). When $\text{Treatment}_\text{B} = 0$, the model returns the expected counts for $\text{Treatment}_\text{A}$. When $\text{Treatment}_\text{B} = 1$, the model returns the expected counts for $\text{Treatment}_\text{B}$.

Treatment A:

$$\begin{aligned}\ln(y) &= b_0 + b_1 \times 0 \\ &= b_0\end{aligned}$$

Treatment B:

$$\begin{aligned}\ln(y) &= b_0 + b_1 \times 1 \\ &= b_0 + b_1\end{aligned}$$

So the expected log-counts for Treatment A is just the intercept, $ln(y) = b_0$, and the expected log-counts for Treatment B is $ln(y) = b_0 + b_1$.

Let's assume R returns the following:

parameter	estimate
b_0	1.447
b_1	-1.224

$\ln(y) = 1.447 - 1.224 \times \text{Treatment}_\text{B}$.

Let $\text{Treatment}_\text{B} = 0$ so we can interpret b_0. $\ln(y) = 1.447$ is interpreted as "the log-count for Treatment A is 1.447".

$e^{\ln(y)} = e^{1.447}$ becomes $y = e^{1.447} = 4.25$, which is interpreted as "the expected count of metastatic lymph nodes for patients in Treatment A is 4.25".

Let $\text{Treatment}_\text{B} = 1$ so we can interpret $b_1 = -1.224$.

$ln(y) = 1.447 - 1.224 = 0.223$. The expected log-count of MLNs for Treatment B patients i 0.223.

$y = e^{0.223} = 1.25$. The expected count of MLNs for Treatment B patients is 1.25.

Imporant: The *incident rate ratio* (IRR) is $e^{b_1} = e^{-1.224} = 0.294$ (the multiplicative factor). For this scenario, the expected count for Treatment A multiplied by the IRR returns the expected count for Treatment B: $4.25 \times 0.294 = 1.25$.

On average, there are 70.6% ($1 - \text{IRR} = 1 - 0.294 = 0.706$) fewer metastatic lymph nodes for patients on Treatment B than those on Treatment A.

The reason we don't just compute the means directly is because the regression output provides additional information, specifically if b_1 is significantly different than 0, and if so, tha there is an effect, or in other words, Treatment B has an effect. In this case, Treatment B reduces the number of metastatic lymph nodes more than Treatment A:

parameter	estimate	p-value
b_0	1.447	< 0.001
b_1	-1.224	0.016

To plot, do the following:

$y = e^{b_0 + b_1 \times \text{Treatment}_\text{B}}$

Plug in $\text{Treatment}_\text{B} = 0$ to get Treatment A. $Plugin \text{Treatment}_\text{B} = 1$ to get Treatment B.

Extremely important: Do not use an unmatched t-test for count data! Counts are Poisson-distributed, and a t-test assumes data are normally distributed. Instead, use Poisson regression to detect differences between groups!

Recall: Poisson regression assumes independence, fixed unit (time, space, etc.), and that the mean = variance. If these aren't met, Poisson regression is not appropriate.

Since Poisson regression returns the expected log-counts, and the expected counts can be derived, we can check the assumption mean = variance by computing the variance from the observed data:

The expected count for Treatment A is $\bar{A} \approx 4.25$; the expected count for Treatment B is $\bar{B} \approx 1.25$

$\text{Var}(A) = s_A^2 = 4.25 \approx \bar{A}$

$\text{Var}(B) = s_B^2 = 1.58 \neq \bar{B}$

The variance of B is not approximately equal to the mean of B, but it is possible this is due to sampling variability. The rule of thumb is that as long as the variance is not $2\times$ the mean, they can be considered approximately equal.

Important: If the rule of thumb suggests the mean and variance are unequal, use Negative Binomial regression instead.

Comparing models with likelihood ratio tests (LRT) and Akaike's Information Criterion (AIC)

Continuing with the dataset about metastatic lymph nodes. Two treatments, A and B. Four patients per treatment. Each observed for number of metastatic lymph nodes.

The model: $\ln(y) = b_0 + b_1 \times \text{Treatment}_\text{B}$

Treatment A: $\text{Treatment}_\text{B} = 0$

Treatment B: $\text{Treatment}_\text{B} = 1$

parameter	estimate
b_0	1.447
b_1	-1.224

Expected count of metastatic lymph nodes for Treatment A patients: $e^{1.447} = 4.25$

Expected count of metastatic lymph nodes for Treatment B patients: $e^{1.447+(-1.224)} = 1.25$

Deviance

$\text{Deviance}_{\text{null}} = 2 \times (\text{LL(saturated model)} - \text{LL(null model)})$

$\text{Deviance}_{\text{residual}} = 2 \times (\text{LL(saturated model)} - \text{LL(proposed model)})$

Deviance is a measure of how well GLM fits to data. It is analogous to sum of squared residuals (SSR) for SLR.

Terminology:

The null model is the model that includes no explanatory variables. It only includes an intercept. One expected count, considers all data points a single group.

The proposed model is the model that includes the explanatory variables of interest. The number of expected counts is equal to the number of groups.

The saturated model is the model where each data point is considered a group, and the observed count consequently is the expected count.

Recall the PMF of the Poisson distribution is $\Pr(k) = \frac{\lambda^k e^{-\lambda}}{k!}$, where k is the observed count and λ is both the expected count and variance.

In the example model, our fit produced an expected count $\lambda_A = 4.25$ and an expected count $\lambda_B = 1.25$. For each data point, we use its associated λ to compute its probability of occurring.

Recall the dataset

Treatment	ID	Count
A	1	7
A	2	4
A	3	4
A	4	2
B	5	3
B	6	1
B	7	1
B	8	0

The probability of observing 7 metastatic lymph nodes given the patient is receiving Treatment A is

$$\Pr(k = 7) = \frac{4.25^7\, e^{-}4.25}{7!} \approx 0.0709$$

We do this calculation for each data point with its respective λ.

The likelihood is the product of the probabilities (because they are independent):

$$\mathrm{L}(\lambda) \prod_{i=1}^{n} \frac{\lambda_i^{k_i}\, e^{-\lambda_i}}{k_i!}$$

In this example, $\mathrm{L}(\lambda) = 0.0709 \cdot 0.1939 \cdot 0.1939 \cdot 0.1288 \cdot 0.0933 \cdot 0.3581 \cdot 0.3581 \cdot 0.2865 \approx 0.0000012$

Because the likelihood is typically an extremely small value, we instead compute the log-likelihood:

$$\mathrm{LL}(\lambda) = \ln(\mathrm{L}(\ ))$$

.

For our data, $\mathrm{LL}(\theta) = \ln(0.0709) \approx -13.65$.

Poisson regression uses the method of maximum likelihood to estimate parameters. It can be visualized as

The candidate b_0 that maximizes LL is chosen as the estimate. The b_0 and b_1 chosen based on maximum likelihood means the $\Pr(k)$ for each data point is optimized to have the greatest set of $\Pr(k)$ for each group. The above was an explanation of the calculation of the LL for the proposed model.

The log-likelihood of the null model is computed using a single value for λ, the expected count irrespective of group, i.e. for all data points. The null model would be $\ln(y) = 0$ and the fit result is

parameter	estimate
b_0	1.012

$e^{1.012} \approx 2.75$, the expected count if all the data were treated as one group.

Computing the log-likelihood for the null model is the same as the proposed model, but with a single value of λ.

$$\begin{aligned} \Pr(k=7) &= \frac{2.75^7\, e^{-2.75}}{7!} && \approx 0.0151 \\ &\vdots = \vdots \\ \Pr(k=0) &= \frac{2.75^0\, e^{-2.75}}{0!} && \approx 0.0639 \end{aligned}$$

$\mathrm{LL}(\lambda) = -17.11$

Finally, the log-likelihood of the saturated model is computed by setting $\lambda = k$ for each data point:

$$\begin{aligned} \Pr(k=7) &= \frac{7^7\, e^{-7}}{7!} && \approx 0.1490 \\ &\vdots = \vdots \\ \Pr(k=0) &= \frac{0^0\, e^{-0}}{0!} && \approx 1 \end{aligned}$$

$\mathrm{LL}(\lambda) = -9.97$.

Compute deviance.

$\mathrm{LL}_{\text{proposed}} = -13.65$, $\mathrm{LL}_{\text{null}} = -17.11$, $\mathrm{LL}_{\text{saturated}} = -9.97$

$\text{Deviance}_{\text{null}} = 2 \times (-9.97 - (-17.11)) \approx 14.28$

$\text{Deviance}_{\text{residual}} = 2 \times (-9.97 - (-13.65)) \approx 7.36$

In summary, deviance compares both the null and proposed models to the saturated model, specifically the log-likelihood of the null model to the log-likelihood of the saturated model, and similarly for the proposed model, according to the formulas

$\text{Deviance}_{\text{null}} = 2 \times (\mathrm{LL}(\text{saturated model}) - \mathrm{LL}(\text{null model}))$

$\text{Deviance}_{\text{residual}} = 2 \times (\mathrm{LL}(\text{saturated model}) - \mathrm{LL}(\text{proposed model}))$

To use these to assess the proposed model, compare the residual deviance to the null deviance. The residual deviance for a good model will be low relative to the null deviance.

In this example, the null deviance is 14.28 and the residual deviance is 7.36—approximately half of the null deviance.

The likelihood ratio test (LRT) can be used to compare nested models. What is considered the "null model" is relative. The process usually starts with considering an intercept-only model as the null model, but once a proposed model that significantly improves upon the null model, it can then become the new null model to search for additional improvements. This is hypothesis testing for models.

The LRT test statistic (TS) is

$$D = -2 \ln \left(\frac{\mathrm{L}_{\text{null}}}{\mathrm{L}_{\text{proposed}}} \right)$$

Recall that $\ln(1) = 0$, $\ln(x)$ when $x \in (0, 1)$ is negative, and $\ln(x)$ when $x \in (1, \infty)$ is positive. (True for any logarithm, base-10, base-e, base-2, etc.)

The likelihood (assuming independence of observations) is the product of probabilities, each being in $[0, 1]$, so will be a negative number.

Note: D can be re-written using one of the laws of logarithms:

$$\begin{aligned} D &= -2 \ln \left(\frac{\mathrm{L}_{\text{null}}}{\mathrm{L}_{\text{proposed}}} \right) \\ D &= -2 \times \left[\ln(\mathrm{L}_{\text{null}}) - \ln(\mathrm{L}_{\text{proposed}}) \right] \end{aligned}$$

For our example where $\mathrm{LL}_{\text{null}} = -17.11$ and $\mathrm{LL}_{\text{proposed}} = -13.65$, $D \approx 6.9$.

Equivalently, the deviances can be used to calculate D:

$$D = \text{Deviance}_{\text{null}} - \text{Deviance}_{\text{residual}}$$

For our example where $\text{Deviance}_{\text{null}} = 14.28$ and $\text{Deviance}_{\text{residual}} = 7.36$, we can see that $D \approx 6.9$.

Proof: Recall that

$\text{Deviance}_{\text{null}} = 2 \times (\text{LL(saturated model)} - \text{LL(null model)})$

$\text{Deviance}_{\text{residual}} = 2 \times (\text{LL(saturated model)} - \text{LL(proposed model)})$

Then

$$\begin{aligned} D &= \text{Deviance}_{\text{null}} - \text{Deviance}_{\text{residual}} \\ &= 2\left[\text{LL(saturated model)} - \text{LL(null model)}\right] - 2\left[\text{LL(saturated model)} - \text{LL(proposed model)}\right] \\ &= -2\left[(-\text{LL(saturated model)} + \text{LL(null model)}) + (\text{LL(saturated model)} - \text{LL(proposed model)})\right] \\ &= -2\left[\text{LL(null model)} - \text{LL(proposed model)}\right] \end{aligned}$$

$D \sim \chi^2$ with degrees of freedom equal to the number of *additional* estimated parameters in the proposed model. In our example,

model	formula
null model	$\ln(y) = b_0$
proposed model	$\ln(y) = b_0 + b_1 \text{Treatment}_\text{B}$

The number of *additional* estimated parameters is 1 ($b_1 for \text{Treatment}_\text{B}$), so use $\chi^2(\text{df} = 1)$. $D \approx 6.9$, and the p-value is $0.0086 \ll 0.05$.

H_0: Proposed model doesn't significantly improve upon null model H_a: Proposed model significantly improves upon null model

At $\alpha = 0.05$, reject H_0: Go with the proposed model.

If we seek further improvement, the proposed model becomes the new null model, and a new proposed model that extends the new null model is established and tested in the same way.

The LRT is one way to compare models, but the Akaike's Information Criterion (AIC) is a method that has a penalty built in for oversaturated (a.k.a. overspecified models). Additionally, it can be used to compare models that are not nested:

$$\text{AIC} = 2p - 2\text{LL}$$

where p is the number of parameters and LL is the log-likelihood of the model.

The optimum AIC is achieved when p is its lowest while LL is its greatest. AIC chooses the model that is simple as possible while still fitting the data well. Using the example, where $\text{LL}_\text{null} = -17.11$ and $\text{LL}_\text{proposed} = -13.65$,

$$\begin{aligned} \text{AIC}_\text{null} &= 2 \times 1 - 2 \times (-17.11) \approx 36.22 \\ \text{AIC}_\text{proposed} &= 2 \times 2 - 2 \times (-13.65) \approx 31.30 \end{aligned}$$

The model with the lowest AIC is preferred, which is the proposed model.

If the model is over-dispersed or under-dispersed, one might need to use *quasi-Poisson regression* or *negative binomial regression*. Over-dispersion is generally when the variance is much larger than the mean. Similarly, under-dispersion is when the variance is much smaller than the mean. So when the variance can't be assumed equal to the mean, this is a problem for Poisson regression as it is a major assumption. Quasi-Poisson and negative binomial regression frameworks solve problems related to over- and under-dispersion.

The previous sections were focused on categorical explanatory variables (e.g. Treatment B). With categorical variables, there is a natural grouping of data points. The variance requires more than one data point to compute. For example, Treatment A had a variance of 4.25 (which was equal to the mean of Treatment A), while the variance of Treatment B was 1.58 (slightly larger than the Treatment B mean of 1.25).

With a continuous explanatory variable such as age (not age category, but age number), there may not be more than one observation. For example, a 72-year-old participant who has volunteered to have their MSLNs counted may be the only 72 year old in the study. The variance can't be computed for 72 year olds because there is only one. And since the variance can't be computed, it can't be compared to the mean to check the assumption of variance.

The solution: Set bins on the explanatory variable and compute the mean and variance within the bins. Plot variance vs. mean with a reference line. For Poisson, since the assumption is that the mean and variance are equal, set the line to have a slope of 1. This strategy can work for other distributions as well; just use a reference line that describes the theoretical relationship between the mean and variance.

Another diagnostic is a familiar one: residuals:

$$r_i = y_i - \widehat{y}_i$$

Plot the residuals vs. the explanatory variable(s) and try to visually detect patterns. A band pattern means the variance is equal as the explanatory variable increases. For Poisson regression, this is *not* desired. What *is desired* is a cone that opens to the right, which indicates that as the explantory variable increases, so does the variance.

A better version of this plot is residuals vs. fitted values. Yet another diagnostic is the Pearson residuals:

$$r_i^P = \frac{y_i - \widehat{y}_i}{\sqrt{\widehat{y}_i}}$$

Plotting the Pearson residuals vs. the fitted values should yield a plot with a band pattern, in contrast to the previous residual plots. In this case, the desired patter *is* a band.

Dispersion is the spread of the data. The variance and standard deviation are measures of dispersion.

Formally, over-dispersion is the presence of more variability in the data than expected. This could occur when explanatory variables are missing from the model.

Underdispersion is the presence of less variability than expected.

In Poisson regression, if overdispersion is an issue, the standard errors will be underestimated. The consequence is that p-values will be small, increasing the risk for a Type-I error.

Quasi-Poisson regression: A model framework for under- or over-dispersed data. The quasi-Poisson model has one additional parameter: how much the variance changes *linearly* in relation to the mean.

The parameter estimates will be the same as Poisson regression, but the SEs and p-values will differ as the model framework is designed to adjust for over- or underdispersion.

Quasi-Poisson models are not fit with maximum likelihood, and consequently, AIC can't be computed (as it is a function of likelihood).

The dispersion parameter is estimated using the equation var $= \phi$mean, where

$$\hat{\phi} = \frac{1}{N-k} \sum \frac{(y_i - \hat{y_i})^2}{\hat{y_i}}$$

Note: $\frac{y_i - \hat{y_i}}{\sqrt{\hat{y_i}}}$ appears in $\hat{\phi}$, but squared. It is the Pearson residual formula:

$$\begin{aligned} \left(\frac{y_i - \hat{y_i}}{\sqrt{\hat{y_i}}} \right)^2 &= \frac{(y_i - \hat{y_i})}{\hat{y_i}} \\ &= (r_i^P)^2 \end{aligned}$$

N is the number of data points; k is the number of estimated parameters.

In our example, $N = 71$ and $k = 2$ (slope and intercept).

Suppose $(r_i^P)^2 = 65.6$. Then $\hat{\phi} = \frac{1}{69} \times 65.6 = 0.95$.

To determine if there is under- or over-dispersion, compare $\hat{\phi}$ to 1:

$\hat{phi} < 1$ implies underdispersion; $\hat{phi} > 1$ implies overdispersion.

Recall the method of arbitrarily grouping the data to compute the mean and variance for each group. Recall those points could be plotted as variance vs. mean, and the line of perfectly equal mean and variance had a slope of 1. In this example, we computed $\hat{\phi} = 0.95$, so this

value could be used in place of 1 on that plot. When the slope is 1, this implies the data are neither under- or over-dispersed.

This line adjusts the estimate (or rule of thumb) for what the variance should be for a given mean. When the slope was 1, a mean of 60 should have a variance of 60.

Our equation var = ϕmean would be

$$\begin{aligned} \text{var} &= \phi\text{mean} \\ &= 1 \times 60 \\ &= 60 \end{aligned}$$

Since $\hat{\phi} = 0.95$,

$$\begin{aligned} \text{var} &= \phi\text{mean} \\ &= 0.95 \times 60 \\ &= 57 \end{aligned}$$

For a mean of 60, we'd expect a variance of 57.

Hypothesis tests for $\hat{\phi} = 1$ exist and may be present in R or even appear in the default output. This is an objective way to dtermine if $\hat{\phi}$ is significantly different than 1.

If the relationship between the variance and the mean is linear, and over- or under-dispersion is present, then quasi-Poisson is the go-to solution.

If the relationship is *non-linear*, the go-to solution is Negative Binomial (NB) regression.

Going back to the familiar diagnostic of Pearson residuals, over-dispered situation appears as a cone. Recall, if mean = variance, the r_i^P vs. $\hat{y}$ plot will have a band pattern.

The relationship between the variance and the mean for the NB model is

$$\text{var} = \mu + \frac{\mu^2}{\theta}$$

where μ is the mean (x-axis in the line plot) and θ, an estimated parameter which may account for overdispersion. If θ gets sufficiently large, then $\frac{\mu^2}{\theta} \to 0$, and the variance equals the mean. If not, this will produce the approximate non-linear relationship between the variance and the mean. (The variance is a function of the mean.)

Important: The parameter θ is always positive, and consequently the variance can't be smaller than the mean. Therefore, NB regression can only account for *overdispersion*. This is in contrast to quasi-Poisson regression, which can handle both over- and under-dispersion.

Recall: The coefficients produced by Poisson and quasi-Poisson are identical. This is not the case for NB regression. They may be similar, but not equal.

The main difference between the three frameworks is the estimated SEs (and consequently p-values).

The last special case is zero-inflation. Zero-inflation occurs when there is a subgroup in the sample which can only produce zeros.

The most straightforward example is asking people how much TV they watched yesterday in hours. Maybe the assumption is that every respondent owns a TV, so an answer of "none" (i.e. 0 hours) would imply they didn't watch TV by chance. However, it is also possible that they don't own a TV at all, so they certainly didn't watch TV. The respondents who don't own a TV are examples of the aforementioned subgroup. The zeros produced by this subgroup can't be explained by the Poisson distribution. The technical term for these zeros is *structural zeros*.

Let's compare a non-zero-inflated situation with a zero-inflated situation to highlight the issue.

Suppose 300 cancer patients were included in a study where they each had the number of MSLNs counted. For each outcome (1 MSLN, 2 MSLNs, etc.), the number of patients was counted.

Let the mean of the sample be 5, so $\lambda = 5$. For each outcome, the expected number of patients can be computed as

$$300 \times \Pr(X = 0) = 300 \times \frac{5^0\, e^{-5}}{0!} = 300 \times 0.0067 \approx 2.01$$

So we expect (assuming $\lambda = 5$) that about $\frac{2}{300}$ patients should have 0 MSLNs. We actually observed 4, but this isn't much different. Continue for each outcome to produce expected counts. the result can be visualized as

5 Linear mixed effects models

TileStats notes: Mixed effects models

Fixed effect example: population mean (fixed because it doesn't vary), usually population parameters Random effects: parameters that vary between groups of dependent data points. Example: Measurements on the same individual will have a mean. Each individual will have a unique mean. In other words, each individual has a parameter to be estimated.

person	before diet	after 1 week	after 2 weeks
1	102	97	95
2	96	93	87
3	83	79	78
4	79	77	75

Regular linear regression would estimate the overall intercept of 89.875 (the mean weight before starting the diet). The estimated slope is -3.125 (the average weight change—in this case loss—per week).

A hypothesis test of whether the slope is significantly different than zero results in FTR ($p \approx 0.372$). This is in conflict with the data because each individual reduces their weight over time.

A linear mixed effects model does determine the slope is significantly different than 0, i.e. rejects $H_0 : b_1 = 0$. One reason the SLR failed as a model is because these four individuals were randomly sampled from the population, and their measurements are far from the fitted values (i.e. large residuals). Additionally, $n = 4$, which is a small sample size.

The aim of the study is to assess whether the diet reduces the weights, and not so much about each individual's weight at the start of the study. So there is variation in body weights between individuals at a given time point, and there is variation between the cases (weight over time for a given individual). The variation in body weights between individuals is irrelevant to the research question; they only want to know if the diet works. To solve this problem, use a LME model. The LME model can estimate the intercept and slope for each individual. The residuals are computed using each individual's model, not a "global" one like SLR.

How to interpret configurations:

Fixed slope, random intercept: All individuals in the population are assumed to have the same slope (i.e. lose weight at a similar rate). The population slope is estimated using the available data (from the four subjects). Each individual has their own intercept to account for their different starting weights (102, 96, 83, 79). A model with random intercepts but fixed slopes would be written as $\text{weight}_i = a_i + b \times \text{weeks}$, where i indexes the subjects. Note: a_i, the intercept, differs based on i, whereas b (the slope) is fixed.

The intercepts for each individual are computed and interpreted in the following way:

$$\begin{aligned} a_1 &= 89.875 + 11.2 = 101 \\ a_2 &= 89.875 + 5.2 = 95 \\ a_3 &= 89.875 - 6.7 = 83 \\ a_4 &= 89.875 - 9.7 = 80 \end{aligned}$$

where 89.875 comes from the fixed effects model. $\text{weight} = 89.875 - 3.125\text{weeks}$.

89.875 is the global intercept (not considering individuals) so each random effect (individual intercept) can be thought of as a deviation from it ($89.875 \pm$ some number): 11.2, 5.2, -6.7, and -9.7 aer random effects.

Special note: The mean of the estimated intercepts (101, 95, 83, 80) is equal to the estimated overall intercept (89.875).

$\text{weight}_1 = 101 - 3.125\text{weeks}$ $\text{weight}_2 = 95 - 3.125\text{weeks}$ $\text{weight}_1 = 89.875 - 3.125\text{weeks}$ (global)
$\text{weight}_1 = 83 - 3.125\text{weeks}$ $\text{weight}_1 = 80 - 3.125\text{weeks}$

These equations allow the "individual regression lines" to be drawn. (The "global" line can be plotted too.)

Each individual's data points have their own line, so residuals are smaller, resulting in lower standard errors, and in turn a smaller p-value.

SSR for SLR: 896.0 SSR for LME: 11.9

$$\begin{aligned} weight_1 &= 101 - 3.125 \times \text{weeks} \\ weight_2 &= 95 - 3.125 \times \text{weeks} \\ weight &= 89.875 - 3.125 \times \text{weeks} \\ weight_3 &= 83 - 3.125 \times \text{weeks} \\ weight_4 &= 80 - 3.125 \times \text{weeks} \end{aligned}$$

These equations allow the "individual regression lines" to be drawn. (The "global" line can be plotted too.)

Each individual's data points have their own line, so residuals are smaller, resulting in lower standard errors, and in turn a smaller p-value.

SSR for SLR: 896.0 SSR for LME: 11.9

In other words, the variance around the lines from the LME model is much smaller than the variance of the SLR model.

Important: To further clarify "random", the differences (11.2, 5.2, -6.7, -9.7) from the overall intercept can be thought of as a random variable with mean = 0 and variance estimated by the model.

$$\frac{11.2 + 5.2 + (-6.7) + (-9.7)}{4} = 0$$

We assume a random sample of subjects (in our case $n = 4$, but imagine $n \gg 4$) would have weights that follow a Normal distribution (`family = "gaussian"`) though counts would follow a Poisson distribution, etc.

How LME model differs from MLR where subject is a factor:

LME: Subjects as a random factor MLR: Subjects as a factor

MLR: `weight = b_0 + b_1 \times weeks + b_j \times subject` LME: `weight = b_0 + b_1 \times weeks + (1 | subject)`

In MLR, subject is a fixed effect, so this framework would compare these four individuals only (i.e. the subjects are *not* a random sample from the population).

Term	LM	LMM
Intercept	101.125	89.875
Weeks	-3.125	-3.125
Subject 1		11.174
Subject 2	-6.0	5.215
Subject 3	-18.0	-6.705
Subject 4	-21.0	-9.685

Both models have same slope. In LM, the intercept is the baseline category (in this case, Subject 1). In LMM, the intercept is the *overall intercept*. So in LM, interpretation of the coefficients—-6.0, -18.0, and -21.0—is relative to Subject 1.

This weight dataset, paired with the research question, violates the assumption of independence required by MLR. Each individual is measured over time, and those measurements are dependent.

LME model vs. repeated measures ANOVA

LME model pros:

- can estimate parameters
- can work even when data is missing
- doesn't require the dependent variable to be continuous; works fine with binary outcomes or count data
- independent variable can be on continuous scale
  - measurement of each subject doesn't have to occur on a fixed, constant schedule; they can be measured at any time

RM ANOVA cons:

- removes *all* data points for an individual if one or more data points are missing, resulting in a reduced sample size which in turn decreases statistical power
- dependent variable *must be* continuous
- the repeated measures must be categorical (all observations across subjects must line up, e.g. everyone has a $t = 1$ measurement, a $t = 2$ measurement, etc.)

Let's consider a different dataset:

Person	Diet	Before	Week 1	Week 2	Week 3
1	A	102	97	95	93
2	A	96	93	87	85
3	B	83	79	78	74
4	B	79	77	75	72

Random intercepts model: `weight ~ weeks + (1 | subjects)`

Term	Intercept	Slope	Note
Overall	89.775	-2.975	fixed effects
Subject 1	11.391		random effect
Subject 2	4.918		random effect
Subject 3	-6.785		random effect
Subject 4	9.524		random effect

Random slopes and intercepts model: `weight ~ weeks + (1 + weeks | subjects)`

Term	Intercept	Slope
Overall	89.775	-2.975
Subject 1	11.910	-0.333
Subject 2	5.429	-0.319
Subject 3	-7.160	0.249
Subject 4	-10.179	0.423

R provides $r = -0.88$ which is the correlation for the random slopes and intercepts. Higher initial weights are correlated with more negative slopes (i.e. weight loss).

This checks out with reality as it is easier for fatter people to lose weight than people who are less fat.

Random slopes only: `weight ~ weeks + (0 + weeks | subject)`

Term	Intercept	Slope
Overall	89.775	-2.975
Subject 1		3.816
Subject 2		1.382
Subject 3		-2.212
Subject 4		-2.986

Important note: The fixed-intercept-random-slope model is usually only appropriate for modeling changes. Initially, all observations (e.g. subjects) have a change of 0, so share an intercept. The changes over time would be $t_1 - t_0$, $t_2 - t_1$, etc.

Additional fixed effect: `weight ~ weeks + diet + (1 | subjects)`

Term	Estimate	p-value
Intercept	97.962	< 0.001
Diet B	-16.375	0.003
slope	-2.975	< 0.001
Subject 1	3.095	
Subject 2	-3.095	
Subject 3	1.309	
Subject 4	-1.309	

Factors become intercepts; non-factors become slopes.

Diet is a factor: A or B. Week is not a factor: 0, 1, 2, or 3 (continuous?)

Therefore, "Intercept" is the intercept for Diet A (the reference category, 97.962), whereas Diet B is the delta between Diet B and Diet A (in this case $97.962\,\text{kg} - 16.372\,\text{kg} \approx 81\,\text{kg}$.

Subject 1—a Diet A member—has an intercept that is $97.962 + 3.095$ (the intercept for their subject-specific "line"). To get the intercept for Subject 3 (a Diet B member), start at 97.962, subtract 16.375 to get to Diet B's overall intercept, then add 1.309 to adjust for Subject 3 ($97.962 - 16.375 + 1.309 \approx 82.896$).

The p-value for Diet B subjects ($p = 0.003$) means that Diet B is significantly different than Diet A. Because the estimate for Diet B is -16.375 (or in general, negative), Diet B subjects weigh less than Diet A subjects.

6 Mixed effects models

6.1 Re-reference

"Random effects in regressions with School Data" on DataCamp. Chapter 2 "Linear Mixed Effects Models" covers interpreting mixed effects models.

6.2 Facts

- Hierarchical model allows parameter sharing across groups
- Outliers in groups with small sample sizes have less of an effect if treated as a random effect
- Random effect parameters assume data share a common error distribution and can produce different estimates when there are small amounts of data or outliers.
- Random effects are usually not plotted
- A random effect intercept comes from a shared distribution of all random effect intercepts (?)
- A random effect slope comes from a shared distribution of all random effect slopes. (?)
- Small populations are subject to random, stochastic variability.
- Restricted Maximum Likelihood (REML) is a method to fit the model when maximum likelihood fails to fit a mixed model. Mixed models are numerically difficult.
- There is no need to use arcsine transforms; use generalized linear mixed effects models.
- Linear mixed models assume the residuals are normally distributed.
- You can always aggregate data to answer new questions, then fit new models. Primary terms must be included, while additinoal terms make the story more full. Models can explain data. Adding terms helps explain variability. Hence "explanatory variables". Models can predict the future. Adding terms improves the accuracy of predictions. Hence "predictor variables". predictor $\rightarrow$ response, explanatory $\rightarrow$ outcome.
- Models can answer groups of related questions. But different questions might require different models.
- "Trend" = "slope coefficient" that differs from 0, i.e. is statistically significant.
- Power refers to the ability to detect statistically significant differences or trends.

6.3 Fixed effect or random effect?

Fixed effects answer core questions; random effects "correct for".

If planning to use random intercepts (or to determine you should), plot lines by group with `geom_line()` and plot trend lines with `geom_smooth()`. If all points have similar ranges and means, don't use random intercepts. If the trends look consistent across groups, don't use random slopes. Note: These are guidelines, not rules.

Random effects allow one to say "corrected for".

6.4 Plot the data

Before fitting a model, plot the data. This allows trends to be uncovered, data points and outliers to be discovered, or other aspects to be noted and considered later in the process.

GLMs can be plotted with `stat_smooth(method = "glm", method.args = list(family = ""))`.

It seems permissable to use `stat_smooth()` to produce trend lines for each group as an approximation for `glmer()` outputs.

6.5 MEMs in R

The `lme4` package

`(1 | group)` random intercept `(var | group)` random slope

A `|` requests that the random effects are correlated; a `||` requests that the random effects are uncorrelated. Uncorrelated random effects can be easier to interpret. Additionally, if a model with *correlated* random effects is failing to fit, specifying uncorrelated random effects can solve this problem. SMEs might request uncorrelated random effects. Remember: "Uncorrelated" is not equivalent to "independent".

`(continuous predictor | random effect group)` = random slope

`y ~ 1 + (1 | group)` is identical to `y ~ (1 | group)` which estimates a global intercept and a random intercept.

The reference group is always the first level of the factor.

`lme4::lmer()`

The package `broom` no longer supports models fit with `{lme4}`. Use the `broom.mixed` package.

It is perfectly fine to include a predictor as both a fixed and random effect. For an intercept, the fixed intercept is estimated for all data (no groups) while the random intercept adjusts by group. For a slope, the same: overall slope + slope by group. If specifying both in a model, the fixed effect should go before the random effect.

Possibilities of `lmer` models: random intercept; fixed mean; nested intercepts; multiple intercepts; correlated random slopes; correlated random intercepts; uncorrelated random slopes; uncorrelated random intercepts.

`broom.mixed::tidy()` can be handy, but is complex.

`family` is the error distribution, how the error distribution is linked to the observed data.

6.5.1 Analyzing output

```
tidy(model) %>% filter(term = x1)
```

```
tidy(model, conf.int = TRUE)
```

If adding a term to the model reduces the standard error for an existing term, this means inclusion of the new term explains a source of variability in the data.

`print()` prints REML, input formula, REML criterion, SD for random effect and residuals, number of observations and groups, fixed effect (similar to `lm()`). `summary()` has everything from `print()`, but also summary details of residuals, standard errors and t-values for fixed effects, and correlations of estimators.

`fixef()` extracts fixed effects and `ranef()` extracts random effects. `confint()` extracts confidence intervals for the fixed effects. (There are no confidence intervals for random effects.)

The package `lmerTest` is a package for ad-hoc estimation of p-values for random effects. See the American Statistical Association's statement about p-values. `lmerTest::lmer()` has a similar (if not the same) syntax as `lme4::lmer()`.

ANOVA is used to compare models; it tells which model explains more variability. Additional model selection methods exist (e.g. AIC), but were beyond the scope of the DataCamp course. Example with ANOVA: build null model with random intercept only vs. a model with random intercept and slope predictor. The null model can be anything. `anova()`: `Pr(>ChiSq)` is the output from the null hypothesis of both models explaining the same amount of variability.

Plot point estimates with `geom_point()` and confidence intervals with `geom_linerange()`. Add a reference line at 0 with `geom_hline()` and finally use `coord_flip()`. This is a plot that shows if 0 is in the confidence interval.

6.5.2 Troubleshooting

Use `scale()` to make model more numerically stable. `mutate(var_scaled = scale(var))`

If the model isn't fitting, look at the REML criterion at convergence.

`scale()` can be used when either of the following warning messages are produced:

- unable to evaluated scaled gradient
- Hessian with X negative eigenvalues

Scale so that the first value or middle value is 0. The former can yield a coefficient that is easier to explain.

6.5.3 Predictions

Let `lmer_out` be a model.

```
original_dataset %>% mutate(lmer_predict = predict(lmer_out))
```

6.6 Logistic regression

An assumption of binomial regression is monotonic.

GLM with binomial family. Also called binomial regression.

In R, binomial data can be in three formats: binary, Wilkinson–Rogers, or weighted. Coefficient produced are the same, but the degrees of freedom and deviance will differ. Binary format will yield the most degrees of freedom because dataset is in long format. The W–R and weighted will have degrees of freedom equal to the number of treatments because dataset is in wide format.

If using W–R format, the LHS of the formula has to be `cbind(fail, pass)`.

6.7 Poisson regression

GLM with Poisson family. The number of events per unit (time, area, etc.). Discrete values, positive values, mean equal to variance. Appropriate if the number of observations is less than 30.

Intro stat courses cover χ^2 test to compare count data. `glmer()` can be used as an alternative. To do this, estimate an intercept for each treatment group. An ANOVA can then be run to determine if the coefficients differ from zero.

7 R

`y ~ x` is a shortcut for `y ~ x + 1`.

`summary()`, `plot()`

References