Week2 ANOVA And GLM Part 1

ANOVA as a test

Group Mean: $\hat{\mu_j} = \text{mean}\{\hat{y}_j, j = 1 \ldots J_i\}$, j represents index of some group, i represents the index of every sample in the group.

Grand Mean: $\hat{\mu}_0 = \text{mean}\{y_{ij}, i = \ldots \hat{k}, j = 1 \ldots J_i\}$, i, j represent the same thing as above, grand mean is the mean of all samples regardless of group.

$\frac{\sum_j n_j (\hat{\mu}_j - \hat{\mu}_0)^2 / (k-1)}{\sum_{ij} (\hat{y}_{ij} - \hat{\mu}_i)^2 / (N-k)} \sim F_{N-1, N-K}$

$\frac{\text{variability between groups}}{\text{variability within groups}} \sim F_{N-1, N-k}$

$N - k$ is the number of observation groups.

Notes

ANOVA will reject H_0 for any large datasets. Large dataset $\rightarrow$ large numerator $\rightarrow$ large F value $\rightarrow$ small p-value $\rightarrow$ reject H_0.

ANOVA can be useful when a dataset is small, for large datasets, fit a random effects model.

Likelihood-ratio Test

> A different way to check if there is a difference between models.

> The **likelihood-ratio test** lets us compare the goodness of fit of two competing models based on the ratio of their likelihoods.

```r
# Intercept only model, the intercept is equal to the grand mean of the data
lm0 <- lm(av_rating ~ 1, data=my_genre_data)
summary(lm0)
```

```
##
## Call:
## lm(formula = av_rating ~ 1, data = my_genre_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6522 -0.2521  0.0771  0.3953  1.7302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.95217    0.04902   162.2   <2e-16
##
## Residual standard error: 0.6685 on 185 degrees of freedom
```

```r
# Linear model with decade coefficient for decade
lm1 <- lm(av_rating~decade, data=my_genre_data)
summary(lm1)
```

```
##
## Call:
## lm(formula = av_rating ~ decade, data = my_genre_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4439 -0.2456  0.0913  0.3624  1.9385
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.74388    0.08305  93.244   <2e-16
## decade2000   0.36818    0.11745   3.135   0.0020
## decade2010   0.25671    0.11745   2.186   0.0301
##
## Residual standard error: 0.6539 on 183 degrees of freedom
## Multiple R-squared:  0.05346,    Adjusted R-squared:  0.04312
## F-statistic: 5.168 on 2 and 183 DF,  p-value: 0.006554
```

```
# install.packages("lmtest")
lmtest::lrtest(lm0, lm1)
```

```
## Likelihood ratio test
##
## Model 1: av_rating ~ 1
## Model 2: av_rating ~ decade
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   2 -188.52
## 2   4 -183.41  2 10.22   0.006036
```

Hypotheses

- $H_0 : Y_{ij} \sim N(\mu_0, \sigma^2)$
- $H_a : Y_{ij} \sim N(\mu_j, \sigma^2)$, prediction does depend on different groups

Likelihood under H_a is always larger than H_0. More feature is better. But if H_0 is true, H_a's likelihood shouldn't be much larger.

2 $\times$ the difference in log likelihoods will follow a chi-square distribution if H_0 is true. (all hypothesis testing is done as if the null is true)

$2[\log L(\hat{\beta}, y) - \log L(\hat{\beta}^{(C)}, y)] \sim \chi^2_P$

P is the number of parameters in β. 是2个相比较的model中相差几个β.

GLM Generalised Linear Model

Assumptions of the GLM

- $Y's$ are independently distributed, as well as the errors.
- The dependent variable Y_i does not need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, poisson, multinomial, normal, ...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables. e.g. for binary logistic regression $logit(\pi) = \beta_0 + \beta X$

  π represents the probability. $\log \frac{\pi}{1-\pi} = X\beta$
- Explanatory variables can be even the power terms or some other non-linear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied.

- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

Components of a Generalised Linear Model

1. **random component**: the response and an associated probability distribution
2. **systematic component**: explanatory variables and relationships among them (e.g., interaction terms)
3. **link function**, which tell us about the relationship between the systematic component (or linear predictor) and the mean of the response

It is the **link function** that allows us to generalise the linear models for count, binomial and percent data. It ensures the linearity and constrains the predictions to be within a range of possible values.

GLM

$$\begin{aligned} Y_i &\sim G(\mu_i, \theta) \\ h(\mu_i) &= X_i^T \beta \end{aligned}$$

OLS

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= X_i^T \beta \end{aligned}$$

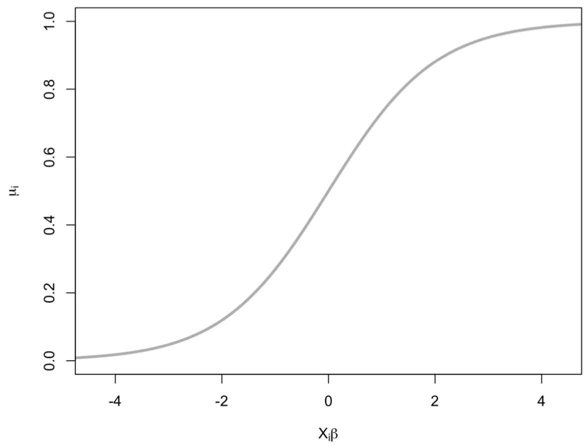
OLS is just a flavour of GLM when:

- G is a Normal distribution
- θ is the variance parameter, denoted σ^2
- h is the identity function

Binomial (or logistic) Regression

$$
\begin{aligned}
Y_i \sim & \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) = & X_i \beta
\end{aligned}
$$

- G is a Binomial distribution
- ... or a Bernoulli if $N_i = 1$
- h is the logit link



- $X_i^T \beta$ can be negative
- μ_i is between 0 and 1.

Notes

- No closed form MLEs for GLMs
- Derivatives are easy so maximization is quick

Interpreting Logistic Model

$$
\begin{aligned}
Y_i \sim & \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) = & \sum_{p=1}^{P} X_{ip}\beta_p \\
\left(\frac{\mu_i}{1-\mu_i}\right) = & \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}
\end{aligned}
$$

- μ_i is a probability
- $\log[\mu_i/(1-\mu_i)]$ is a log-odds
- $\mu_i/(1-\mu_i)$ is an odds
- If $\mu_i \approx 0$, then $\mu_i \approx \mu_i/(1-\mu_i)$

Suppose $X_{1p} = X_{2p}$ for all p except $X_{2q} = X_{1q} + 1$

$$
\begin{aligned}
\beta_q = & \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right) \\
\exp(\beta_q) = & \left(\frac{\mu_2}{1-\mu_2}\right) \Big/ \left(\frac{\mu_1}{1-\mu_1}\right)
\end{aligned}
$$

- β_q is the log-odds ratio
- $\exp(\beta_q)$ is the odds ratio
- $\exp($intercept$)$ is baseline odds, when $X_{i2} \ldots X_{iP} = 0$.