STA303 notes

July 25, 2022

Contents

Lecture notes for STA303 in Summer 2022 with Justin Slater.

July 6: Lecture 1

In simple linear regression, we model variable Y to have the relationship

$$E(Y \mid X) = \beta_0 + \beta_1 X,\ Y = \beta_0 + \beta_1 X + \epsilon$$

where for paired data $\{(x_i, y_i)\}_{1 \le i \le n}$ each realization has $y_i = \beta_0 + \beta_1 x_i + e_i$ such that $E(\epsilon_i) = 0$. We assume $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ and ϵ_i are i.i.d. In order to estimate β_0, β_1 we minimize

$$RSS = \sum_i e_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

and get that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \hat{x}$, $\hat{\beta}_1 = \dfrac{SS_{X,Y}}{SS_{X,X}} = \dfrac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$. The estimators are unbiased. As well, $\hat{\sigma}^2 = \dfrac{RSS}{n-2}$.

$$V(\beta_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}}{SS_{X,X}} \right),\ V(\beta_1) = \frac{\sigma^2}{SS_{X,X}}$$

For more, please see STA302 notes.

July 6: Lecture 2

Maximum Likelihood

Given data, we quantify the how likely it is given model parameters using maximum likelihood estimation (STA261).

Example 1. Suppose $f(y_i, \theta) = \theta \exp(-y_i \theta)$. Then given the vector of realizations y, then

$$L(\theta, y) = \prod_i f(y_i, \theta) = \theta^n \exp(-n \overline{y} \theta)$$

We maximize L with respect to θ. We take logarithm of the function to simplify computation. Maximizing $\ell = \log L$ also maximizes L. Maximize:

$$\ell = n \log \theta - n \overline{y} \theta, \quad \frac{\partial \ell}{\partial \theta} = \frac{n}{\theta} - n \overline{y} = 0 \implies \hat{\theta} = \frac{1}{\overline{y}}$$

Moreover, $\dfrac{\partial^2 \ell}{\partial \theta^2} = -\dfrac{n}{\theta^2}$ is negative at $\hat{\theta}$, and is a maximum.

In simply linear aggression, we assume $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, so the log likelihood given y is

$$\ell(\beta_0, \beta_1, \sigma^2, y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} RSS$$

This is maximized when RSS is minimzed, so MLE estimates match RSS estimates.

Logistic Regression

In logistic regression we model when X is a continuous predictor of a binary outcome Y. We write $y_i \sim \text{Bern}(p_i)$, and we assume

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

Since $E(Y \mid X) = p_i$ since it is Bernoulli distributed, we see that $g(E(Y \mid X)) = \beta_0 + \beta_1 x_i$. g is an example of a link function: it links the conditional expectation to a linear function of the parameters. Then

$$p_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) = \frac{\exp(X\beta)}{1+\exp(X\beta)}$$

where X is row vector, β column vector. Namely, p is a function of the predictor,

$$p = \text{logit}^{-1}(x)$$

in the single variable case. The interpretation of coefficients is more difficult than in linear regression. Consider the single x case. There are two main interpretations:

1. $\dfrac{\partial \,\text{logit}^{-1}(\beta_0 + \beta_1 x_i)}{\partial x} = \dfrac{\beta_1 \exp(\beta_0 + \beta_1 x)}{(1+\exp(\beta_0 + \beta_1 x))^2}$ plugging in $\overline{x}$, we get the change in the probability near the mean. Very similarly in the multivariable case.

2. Denoting odds as $\Omega_i = \frac{p_i}{1-p_i}$, $\log(\Omega_1) = \beta_0 + \beta_1 x$, $\log(\Omega_2) = \beta_0 + \beta_1(x+1)$, then

   $$\frac{\Omega_2}{\Omega_1} = \exp(\beta_1)$$

   So $\exp(\beta_1)$ is the multiplicative increase in odds ratio for a one unit increase in x.

We use the likelihood method to estimate the betas. When $y_i \sim Bern(p_i)$ and we have one continuous predictor, we begin with the likelihood function

$$L(\beta_0, \beta_1, y) = \prod_i^n p_i^{y_i}(1-p_i)^{1-y_i}$$

Taking logarithms and doing arithmetic,

$$\ell(\beta_0, \beta_1, y) = \sum_i y_i \log p_i + (1-y_i)\log(1-p_i) = \sum_i^n -\log(1+\exp(\beta_0 + \beta_1 x_i)) + \sum_i^n y_i(\beta_0 + \beta_1 x_i)$$

The equation $\dfrac{\partial \ell}{\partial \beta_1} = 0$ requires a *numerical solution*. In the general setting, for a parameter θ of some score ℓ', the numerical method we use to find the root is Newton-Raphson. Iteratively update θ using

$$\theta_{t+1} = \theta_t - \frac{\ell'(\theta_t)}{\ell''(\theta_t)}$$

until the θ corresponding to $\ell'(\theta) = 0$ is reached. With multiple parameters, we used a similar algorithm.

Once we obtain maximum likelihood estimates for two models, we can decide which model is better with a **likelihood ratio test**. Suppose we have

$$M_1 : \text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad M_2 : \text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Then if $L_1(\hat{\beta}_1), L_2(\hat{\beta}_2)$ are the full and nested likelihood of estimates from each model,

$$-2\log\left(\frac{L_2(\hat{\beta}_2)}{L_1(\hat{\beta}_1)}\right) \sim \chi^2_1$$

where the degrees of freedom is the difference in number of parameters. A significant test statistic shows the model with interactions explains much more variance. $D = -2\ell(\beta)$ is called the **deviance** and can be seen as a measure of model fit when considering nested models. The above is just the difference

$$D_{\text{nested}} - D_{\text{full}} \sim \chi^2_{(\text{\# diff parameters})}$$

If the change in deviance is too large, then nested model is much better. Deviance is analogue of residual sum square, but for models fit with MLE. The lower the deviance, the better the model fit.

Residuals for logistic regression are defined similarly as linear regression:

$$r_i = y_i - \text{logit}^{-1}(\beta_0 + \sum_i \beta_i x_i)$$

July 11: Lecture 3

Inference of Parameters

In linear regression, we often assume normality for the errors, which allows us to build Z, t confidence for the $\hat{\beta}$ estimators. In logistic regression, no equivalent assumption exists, so we use maximum likelihood to compute the variance. For a sample $y_1, \ldots, y_n$,

$$L(y, p) = \prod_i p^{y_i}(1-p)^{1-y_i} \implies \ell(y, p) = \sum_i y_i \log p + (1-y_i)\log(1-p)$$

Differentiating,

$$\frac{\partial \ell}{\partial p} = \frac{n(\overline{y} - p)}{p(1-p)}, \quad \frac{\partial \ell}{\partial p} = 0 \implies \hat{p} = \overline{y}$$

and can be verified as maximum with second derivative test. The **curvature** of ℓ contains information about the variance of our estimator $\hat{p}$. A high curvature $\frac{\partial^2 \ell}{\partial p^2}$ means low variance, while a flatter likelihood means higher variance. *The more concave the likelihood the lower the variance.* We can use this to find the standard error.

$$I(p) = E\left(\frac{\partial^2 \ell}{\partial p^2}\right) \implies SE(\hat{\beta}) \approx \frac{1}{\sqrt{I(\hat{p})}}$$

We know

$$\text{logit}(\hat{p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

$I(p)$ is the **Fisher information**. In the multivariable case, we get a matrix of second derivaties and an information matrix $I(\beta)$, where

$$V(\beta_j) = \frac{1}{\sqrt{I_{j,j}(\beta)}}$$

For any MLE, it is invariant, asymptotically unbiased, asymptotically normal.

Proposition 1. Invariance: If f is a one to one function, MLE of $f(\theta)$ is $f(\theta^{MLE})$.

Proposition 2. Asymptotically unbiased: As $n \rightarrow \inf$, $E(\theta^{MLE}) = \theta$

Proposition 3. Asymptotically normal: As $n \rightarrow \inf$, $\theta^{MLE} \sim N\left(\theta, \frac{1}{I(\theta)}\right)$

We may now test hypotheses for β. In the case logit$(P(Y_i = 1 \mid X)) = \beta_0 + \beta_1 x$, we test $H_0, \beta_1 = 0$ and $H_1, \beta_1 \neq 0$ with the **Wald test**:

$$W = \frac{(\beta_1^{MLE} - \beta_1^{H_0})^2}{\hat{V}(\beta_1^{MLE})} \sim \chi_1^2$$

Where U is the score, and I is the information, we test **slope to curvature ratio**. Slope at MLE is 0, but if slope at H_0 relative to curvature is far from 0, then $\beta_1^{H_0}$ is likely far from the MLE. **Slope test:**

$$S = \frac{U(\beta_1^{H_0})^2}{I(\beta_1^{H_0})} \sim \chi_1^2$$

We can also do **Likelihood ratio test** as discussed before.

$$LRT = 2\log(\ell(\beta_1^{MLE}) - \ell(\beta_1^{H_0})) \sim \chi_1^2$$

If $\beta_1^{H_0}$ is different from MLE, this is evidence against the null. In the limit, $n \longrightarrow \inf$, all three tests are actually the same. The literature recommends likelihood ratio aka Wilks test.

Wald test intervals can be built with z score, but score and LR intervals require numerical solutions.

Validation of Logistic Regression Models

Definition 1. The **Brier score** of a model is defined as

$$B = \frac{1}{n} \sum_i^n (\hat{p}_i - y_i)^2$$

The score assess both **calibration** and **discrimination**.

Definition 2. Calibration is the ability to accurately infer p_i given x_i

Definition 3. Discrimination is the ability to accurately choose $0/1$.

A model that is good at discrimination is *not necessarily* well calibrated. The analog of R^2 is also used in logistic regression:

Definition 4.

$$R_N^2 = \frac{1 - \exp(LR/n)}{1 - \exp(L^0/n)}$$

where LR is the likelihood ratio test statistic, and L_0 is $-2\ell_0$, the log likelihood of the null model.

Similar to R^2, it is a measure of predictive performance: how much deviance can be explained. Another measure of discrimination is D_{xy}.

Definition 5. The **Somers** D_{xy} measures the ability of the model to distinguish between high and low values.

$$D_{xy} = 2(c - 0.5)$$

where c is the **probability of concordance**: for every 0, 1 pair it is the proportion of correct $\hat{p}_0 < \hat{p}_1$

R_N^2, B, improve as the number of predictors increases, but more complicated models do not generalize well. A β closer to 0 slightly closer to 0 gives better predictive performance since it reduces overfitting to training data. Choosing this β is called shrinkage. We perform shrinkage by fitting

$$\text{logit}(p) = \gamma_0 + \gamma_1 X\beta$$

and finding best γ_0, γ_1 on test data. We do this through resampling via bootstrap, then taking means of sampling distributions of γ_0, γ_1. We can also view the difference between measure on the entire data, and on sample data, to see the deviation, which is referred to as **optimism**.

July 13: Lecture 4

Review: The curvature of the likelihood tells us how certain our MLE estimates are. Standard errors that rely on MLE are based on **asymptotic results**. We have three tests: Wald, score, likelihood, which are asymptotically equivalent. We discussed model validation.

Generalized Linear Models

Definition 6. A **generalized linear model** is a model of a relationship between X, Y which has three components

1. A random distribution of $Y_i \sim$ Norm, Bern,...
2. A systematic component, $\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
3. A link function so that $g(E(Y \mid X)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

Y can be distributed as anything, but in this class we understand the *exponential family* of distributions.

Definition 7. The distribution of y belongs to the **exponential family** if

$$p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

θ is the **canonical parameter**, ϕ is the **scale parameter**, $b(\theta)$ is the **cumulant function**, and $c(y, \phi)$ makes the integral $\int_{\mathbb{R}} p(y \mid \theta, \phi) dy = 1$

We show Bernoulli, Poisson, and Normal distributions are in the exponential family.

Example 2. When y is normally distributed,

$$p(y, \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right)$$

$\theta = \mu$ is the **canonical parameter**, $\phi = \sigma^2$ is the **scale parameter**, $b = \frac{\mu^2}{2}$ is the **cumulant function**, $c = -\left(\frac{y^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)$ is a function that makes the pdf have volume 1.

Example 3. The Bernoulli distribution

$$p(y \mid p) = \exp\left(y \log \frac{p}{1-p} + \log(1-p) \right)$$

where $\theta = \text{logit}(p)$, $\phi = 1$, $b = -\log(1-p)$, $c = 0$.

For Bernoulli, the canonical parameter is $\log(\frac{p}{1-p})$ which is the link function in logistic regression. This is the **logit link**. For the normal model, the canonical parameter is μ, which appears on left side of linear regression: $y \sim N(\mu, \sigma^2)$,

$$\mu = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_K$$

In this case, we call this the **identity link**. In general, the canonical parameter tells us the most natural link to use in different types of regression.

The mean and variance of random variables form the exponential family can eb determined using

$$E(Y \mid \theta, \phi) = b'(\theta),\ V(Y \mid \theta, \phi) = \phi b''(\theta)$$

Example 4. In the normal distribution,

$$b(\theta) = \frac{\theta^2}{2} \implies b'(\theta) = \theta = \mu, \quad \phi b''(\theta) = \phi = \sigma^2$$

The link function $\eta_i = g(E(Y_i \mid X)$ could be any smooth and monotonic function that relates the expectation of the outcome to the linear predictor. In general, it can be defined on any domain.

The canonical link $h()$ is *the most natural one*. It is usually defined on all of $\mathbb{R}$.

$$\eta_i = h(E(Y_i \mid X)) = \theta_i = \beta_0 + \beta_1 x_1 + \ldots$$

All we are doing is relating the expectation of Y to some linear function of the predictors.

Example 5. The Poisson distribution is part of the exponential family:

$$p(y, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp(y \log \lambda - \lambda - \log y!)$$

$\theta = \log \lambda$, $\phi = 1$. In this case, the canonical link function is given by $\theta = \log(\lambda)$. By differentiating $b(\theta)\exp(\theta)$, $b'(\theta) = \lambda$, $\phi b''(\theta) = \lambda$.

Poisson Regression

$y_i \sim \text{Pois}(E_i \lambda_i)$, then $\log(\lambda_i) = \beta_0 + \beta_1 x_1 + \ldots$. Here E_i is the **exposure** or **offset**. This lets us consider lambda as a rate.

Example 6. If we want to comapre the risk of covid infection between provinces, and the ys are the cases in each province, then Ontario and Quebec will be large relative to other provinces due to larger populations. We use the offest E_i in this case.

In Poisson regression, we have the link function $h = \log$. Since $V(Y) = E(Y) = \lambda$, then we may standardize residuals via

$$z_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

since $\sqrt{\hat{y}_i}$ is standard error in Poisson model. In many cases, we observe **overdispersion**, when many residuals are $> |2|$. By Chebyshev's inequality, 95% of data points should lie within these bounds. A statistical test for these is

$$\sum_i^n z_i^2 \sim \chi^2_{n-k}$$

where k is the number of parameters. In the null, overdispersion does not occur. When it occurs, we estimate the overdispersion factor

$$\omega = \frac{1}{n-k} \sum_i^n z_i^2$$

and model $y_i \sim \text{QuasiPoiss}(\lambda, \omega\lambda)$ where $V(Y_i) = \omega E(Y) = \omega\lambda$. Here we do not know the likelihood function, and we need to use quasi likelihood methods. Instead, we can use negative binomial with

$$E(Y) = \lambda,\ V(Y) = \lambda + \frac{\lambda^2}{\omega}$$

July 18: Lecture 5

Recal that we can build a GLM when Y is an exponential family distribution. For example, Normal, Poisson, and Bernoulli. All material this week is testable on midterm next week.

GLMs Continued

Last time we learned Poisson was useful for count data. Sometimes it is not appropriate, since count data can have a different structure.

Example 7. We are interested on the effect of lymphotic infiltration on osteoid pathology. Other predictors include gender, osteoid pathology. We use group size E_i as an offset for the number of successes: $y \sim \text{Poiss}(E_i\lambda_i)$. None of the predictors are significant.

$$\log(\lambda_i) = \log(E_i) + \beta_0 + \beta_1 x_1 + \dots$$

In the current example, the number of the outcome is capped by the number of people in each group. This means λ_i is capped at 1, which isnt consistent with Poisson, where $\lambda_i \in \mathbb{R}^+$. Instead, we could treat the successes y_i in a group of size n through a binomial model.

In other count data, $y_i \sim \text{Bin}(n_i, p_i)$. Y_i is also in the exponential family, and the canonical parameter is $\log\left(\frac{p_i}{1-p_i}\right)$. The interpretation for a one unit increase in predictors is the same as in logistic regression. When assessing the significance of predictors, Wald tests do not do well with a small number of successes and failures. This is because the estimated β is close to $\pm\infty$, so the standard error is also big. This the **Hauck-Donner effect**. We do the likelihood ratio test, since it does not rely on $\hat{\beta}$, it just uses their likelihood.

With binomial models, we run into the same issues with over dispersion:

$$E(y_i) = np_i,\ V(Y_i) = np_i(1-p_i)$$

so the same parameter p_i describes both. We can check for overdispersion using standardized residuals:

$$z_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$$

We can do the test $\sum z_i^2 \sim \chi^2_{n-k}$ where there is no overdispersion under the null. Use $\alpha = 0.05$. For dispersion factor ω, we need to fit an overdispersed binomial where

$$V(Y) = \omega np(1-p)$$

Fitting `family=quasibinomial` in R, we see that predictors are no longer significant.

What if we wanted to use a likelihood ratio test to compare the model with an interaction to the model without interaction. Since quasi-Poisson and quasi-binomial models are not exponential, we cannot do χ^2 test. In the Poisson case, we can solve this problem with negative binomial. In binomial case, we cannot. Deviances are also not χ^2 distributed, they are $\omega\chi^2$ distributed where ω is unknown. The statistic

$$F = \frac{(D_{simple} - D_{complex})/(k_{simple} - k_{complex})}{(\sum_i z_i^2)/(n-k)} = \frac{(D_{simple} - D_{complex})/(k_{simple} - k_{complex})}{\hat{\omega}^2} \sim F_{(1,n-k)}$$

The statistic $\hat{\omega} = \frac{\sum_i z_i^2}{n-k}$ is based on the more complex model. The first degree of freedom in $F_{1,n-k}$ comes from the difference in number of predictors.

Example 8. If we consider the Lymphotic infiltration, but separate each row into distinct trial with individual outcomes and do logistic regression, we are not able to assess over dispersion. We cannot assess the variance of `1,0,0,1,1,0,0,1`: it is silently overdispersed. High variance is easily seen in the binomial case.

Goodness of Fit Tests

In an ideal world, we would be able to perfectly explain data with the model, and have $\hat{y}_i = y_i$. If we have n parameters, we can do this easily: this is a **saturated model**. In the saturated model, the deviance is 0. We cannot do the usual likelihood ratio test where $D_{model} - D_{sat} \sim \chi^2_{n-k}$, and test the null that they explain the same amount of deviance. We did not consider the case where our number of parameters grows with sample size. Then we can approximate via **saddle point approximation** the statistic with

$$\frac{D_{model}}{\phi} \sim \chi^2_{n-k}$$

where D_{model} is the residual deviance, ϕ is the dispersion parameter, and k is the number of predictors. Recall in Poisson and binomial, $\phi = 1$. We can use this to test goodness of fit of the model, but our sample must satisfy $\min y_i \geq 3, \min n_i - y_i \geq 3$: each trial has at least 3 successes and failures. There is a second goodness of fit test called the **Pearson test**:

$$\frac{\sum_i z_i^2}{\phi} \sim \chi^2_{n-k}$$

which checks overdispersion. Under the null, the model is a good fit, but significant overdispersion shows that it is not.

July 20: Lecture 6

July 25: Lecture 7

Multilevel Models

Some call these hierarchical, mixed, random effect, or multilevel models. The idea is more general.

Example 9. Consider the example of trees. They take up air and lose water through their stomata. Sometimes they are limited by the amount of carbon dioxide they can absorb since the stomata are too small. Consider an experiment where trees are grown under 2 levels of carbon dioxide concentrations with 3 trees assigned to each experiment. Our question: if there is more carbon dioxide, is the area of the stomata greater?

A normal model is a reasonable place to start. $y_i \sim N(\mu_i, \sigma^2)$ and

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_6 x_{6i}$$

which is a simple linear regression. Let $\beta_1 x_{1i} = 1$ for high CO_2 and 0 otherwise. Let $x_{ji} = 0/1$ be indicator variables for the trees, with comparison being first tree.
A different way to write this model:

$$\mu_i = \alpha_{j[i]} + \beta_{k[i]}$$

where

- $\alpha_{j[i]}$ is the tree effect: the influence of the tree j that effect i was taken from
- $\beta_{k[i]}$ is the CO_2 effect: the influence of the tree j that effect i was taken from

In the linear model described earlier, the effect of tree number absorbs the effect of high CO_2. The CO_2 variable and tree numbers are highly related.

We write $\mu_i = \alpha_{j[i]} + \beta_{k[i]}$ and let tree effects be given by probability distribution $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$.