MY452: Applied Regression Analysis

Felix Wortmann Callejón, May 7, 2023

1 Hypothesis Testing & Inference

1.1 Null-Hypotheses

Null-hypothesis (H_0), usually $H_0 : \beta = 0$, where $\beta =$ population parameter of interest. Alternative hypothesis (H_a), can be two-sided ($H_a : \beta \neq 0$), one-sided ($H_a : \beta > 0$ or $H_a : \beta < 0$).

1.2 T-Test

The T-statistic compares $\hat{\beta}$ to it's estimated sampling variability, $\widehat{SE}(\hat{\beta})$.

$$\widehat{SE}(\hat{\beta}) = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{(n-k-1) \cdot \sum(x_i - \bar{x})^2}}$$

given a β_0 (under the null), the T-statistic is then (and always) calculated by

$$T = \frac{\hat{\beta} - \beta_0}{\widehat{SE}(\hat{\beta})}$$

1.3 p-values

In OLS, we compare T the T-distribution w/ $df = n-k-1$ to get a p-value, which gives us the $Pr(|\hat{\beta}| \geq \hat{\beta} \mid \beta = 0)$. For a two-sided test, $-|T| < T^{-1}(\frac{\alpha}{2}, df)$ for us to reject the null-hypothesis at α significance level.

1.4 Confidence Intervals

Confidence way of inverting p-values to get range of plausible values of β.

$$\hat{\beta} \pm t_{\alpha/2,df} \cdot \widehat{SE}(\hat{\beta})$$

Valid interpretations of 95% CIs: *"At the 95% confidence level we estimate that the population parameter lies between [...]"*, or *"Interval contains the parameter with 95% confidence."*

2 Ordinary Least Squares

With iid observations Y_i and $X_1, ..., X_k$ linear model is

$$Y_i = \alpha + \beta_1 X_{1i} + ... + \beta_1 X_{ki} + \epsilon_i$$

2.1 Model Assumptions

1. $Y_i \perp\!\!\!\perp Y_j \forall\ i \neq j$
2. $\epsilon_i \sim \mathcal{N}(0, \sigma)$ with $\sigma^2 < \infty$
3. $\mathbb{E}[Y_i]$ is indeed linear in all X
4. $Cov[\epsilon_i, \epsilon_j] = 0 \forall\ i \neq j$

2.2 Simple Linear Regression

In **SLR**, can estimate the t-statistic of $\hat{\beta}$ using R^2 by the following formula

$$t = \frac{\sqrt{R^2}}{\sqrt{\frac{1-R^2}{n-2}}}$$

In **SLR**, $R^2 = r_{XY}^2$, where r_{XY} correlation of X, Y.

2.3 Interaction Effects

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon$$

Partial association of X_1, Y has to be interpreted using $\beta_1 + \beta_3 X_2$. I.e., if $\beta_3 > 0$ as X_2 increases, partial relationship between X_1, Y gets **stronger**. If $\beta_3 < 0$ relationship gets **weaker**. Relationship between X_1, Y always has to be interpreted **conditional** on X_2.

2.4 R^2 and Adjusted R^2

Goodness of fit measure

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Adjusted R^2 is given by

$$R^2_{adj} = \frac{\frac{TSS}{n-1} - \frac{SSE}{n-(k+1)}}{\frac{TSS}{n-1}} = \frac{(n-1)R^2 - k}{n-(k+1)}$$

Residual standard deviation is estimated by

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-k-1}}$$

2.5 F-Test

To test whether a set of coefficient estimates are all zero, F-test. Null-hypothesis of F-test:

$$H_0 : \beta_{g+1} = \beta_{g+2} = ... = \beta_k = 0$$
$$H_a : \text{at least one of } \beta_{g+1}, \beta_{g+2}, ..., \beta_k \text{ is not } 0$$

In the context of OLS, usually a model M_0 with g independent variables (IVs), and M_a with $g+k$ IVs. M_0 is nested in M_a.

$$F = \frac{\frac{SSE_0 - SSE_a}{k_a - k_0}}{\frac{SSE_a}{n-(k_a+1)}} = \frac{\frac{R^2_a - R^2_0}{k_a - k_0}}{\frac{1-R^2_a}{n-k_a-1}} = \frac{\frac{\Delta R^2}{\Delta df}}{\frac{1-R^2_a}{n-(k_a+1)}}$$

Sampling distribution of F-statistic under the null hypothesis is the F-distribution with $k_a - k_0$ and $n-(k_a+1)$

degrees of freedom. If M_a has an additional coefficient, than $F = T^2$. If M_0 has no explanatory variables ($k_0 = 0$, $R_0^2 = 0$, $SSE_0 = TSS$), F-statistic becomes

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}} = \frac{\frac{SSM}{k}}{\frac{SSE}{n-(k+1)}}$$

3 Binary Logistic Regression

Because $0 \leq \pi_i \leq 1$, need to model transformation of π. To expand domain to $(0, \infty)$, we use Odds $= \frac{\pi}{1-\pi}$, then take the (natural) log, to expand domain to $(-\infty, \infty)$. With iid observations $Y_i \in \{0, 1\}$ and $X_1, ..., X_k$ binary logistic regression (BLR) model is

$$logit = \log_e(\frac{\pi_i}{1-\pi_1}) = \alpha + \beta_1 X_{1i} + ... + \beta_1 X_{ki}$$

where $\pi_i = P(Y_i = 1)$. Underlying distribution is Bernoulli (special case of binomial), $\mathbb{E}[Y] = \pi$, $Var[Y] = \pi(1-\pi)$. Fitted probabilities/odds

$$\begin{aligned} log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki} \\ \text{Odds} = \frac{\pi_i}{1-\pi_i} &= e^{\alpha+\beta_1 X_{1i}+...+\beta_k X_{ki}} \\ \pi_i &= \frac{e^{\alpha+\beta_1 X_{1i}+...+\beta_k X_{ki}}}{1+e^{\alpha+\beta_1 X_{1i}+...+\beta_k X_{ki}}} \end{aligned}$$

3.1 *p*-values and CIs for BLR

BLR coefficients tested for significance using

$$z = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}$$

p-values from the standard normal distribution (t-distribution with $df = \infty$). For 2-tailed test with $\alpha = 0.05$, $z_\alpha \approx 1.96$. Wald test statistic

$$W = z^2 = \left(\frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}\right)^2$$

is compared to χ^2 distribution with $df = 1$. Confidence intervals in log-odds are calculated by

$$\hat{\beta} \pm z_\alpha \cdot \widehat{SE}(\hat{\beta})$$

and in odds ratios by

$$e^{\hat{\beta} \pm z_\alpha \cdot \widehat{SE}(\hat{\beta})}$$

3.2 Likelihood Ratio Test

Likelihood $L \propto$ to probability of obtaining observed pattern of results in sample if, model were true. Likelihood ratio test compares models for improvements in fit. Consider L_1 of simpler M_1, L_2 of complex M_2 with additional β_k (M_1 nested in M_2). Likelihood ratio test $H_0 : \beta_k = 0$. Likelihood ratio test statistic is calculated by:

$$\begin{aligned} L^2 = \log\left(\frac{L_2}{L_1}\right)^2 &= -2\log\left(\frac{L_2}{L_1}\right) \\ &= -2\left(\log\,(L_1) - \log\,(L_2)\right) \\ &= 2\log\,(L_2) - 2\log\,(L_1) \end{aligned}$$

Compare to χ^2-distribution with df = extra parameters in M_2.

3.3 Fit Statistics

MacFadden's Pseudo-R^2 given by

$$\frac{-\log\,L_N - (-\log\,L_1)}{-\log\,L_1}$$

where L_N = likelihood of null (intercept only) model. Interpretation: proportional improvement in fit, **not** explained variance.

Deviance is "distance" between model of interest and "saturated" model with n parameters. L_S = likelihood of saturated model.

$$\begin{aligned} &\overbrace{2\log\,(L_S) - 2\log\,(L_N)}^{\text{Null deviance}} \text{ with } df = df_S - df_N \\ &\underbrace{2\log\,(L_S) - 2\log\,(L_1)}_{\text{Residual deviance}} \text{ with } df = df_S - df_1 \end{aligned}$$

Information criteria (IC), e.g. Akaike's IC (AIC) $AIC = -2\log\,L_1 + 2k$, with k = number of model parameters. Smaller AIC's are better.

4 Multinomial Logistic Regression

Odds generalise from something not happening/happening to something happening/something else happening.

$$\begin{aligned} \text{Odds}(Y=1) &= \frac{Pr(Y=1)}{Pr(Y=0)} = \frac{Pr(Y=1)}{1-Pr(Y=1)} \\ \text{Odds}_{k'}(k) &= \frac{Pr(Y=k)}{Pr(Y=k')} \end{aligned}$$

The multinomial logistic regression (MLR) model uses this property. Consider $Y_i, X_{1i}, X_{2i}, ..., X_{ki} \forall\ i \in \{1, 2, ..., n\}$, where $Y_i \in \{0, 1, ..., C-1\}$. Again Y_i's are iid from multinomial distribution with probabilites $\pi_i^{(0)}, \pi_i^{(1)}, ..., \pi_i^{(C-1)}$. MLR model is then defined as

$$\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(0)}}\right) = \alpha^{(j)} + \beta_1^{(j)} X_{1i} + ... + \beta_k^{(j)} X_{ki} \; \forall \, j \in \{1, 2, ..., C-1\}$$

Interpretation to the reference category are analogous to BLR. For non-reference categories, i.e. from j to 1, $e^{\beta^{(j)} - \beta^{(1)}}$.

$$\log\left(\frac{\pi_i^{(j)}}{\pi_i^{(1)}}\right) = (\alpha^{(j)} - \alpha^{(1)}) + (\beta_1^{(j)} - \beta_1^{(1)}) X_{1i} + ... + (\beta_k^{(j)} - \beta_k^{(1)}) X_{ki} \quad (1)$$

for each $j \in \{2, ..., C-1\}$.

4.1 Fitted Probabilities

Let $L(j) = \log\left(\frac{\pi_i^{(j)}}{\pi_i^{(0)}}\right)$, then

$$\pi^{(j)} = P(Y = j) = \frac{e^{L(j)}}{1 + \sum_{l=1}^{C-1} e^{L(l)}}$$
$$\pi^{(0)} = P(Y = 0) = \frac{1}{1 + \sum_{l=1}^{C-1} e^{L(l)}}$$

4.2 *p*-values and CIs

As with BLR, i.e. using z-statistic for individual coefficients, and LR for multiples coefs.

4.3 Independence of Irrelevant Alternatives

MLR has property called the Independence of Irrelevant Alternatives (IIA), implies that presence or absence of alternative has no effect on relative proportion of individuals choosing among remaining alternatives.

5 Ordinal Logistic Regression

Consider iid **ordinal** outcome variable Y_i with C categories, such that $j \in \{1, 2, ..., C\}$. Then

$$Pr(Y = j) = \pi^{(j)} \forall \, j \in \{1, 2, ..., C\}$$
$$Pr(Y \leq j) = \gamma^{(j)} = \pi^{(1)} + \pi^{(2)} + ... + \pi^{(j)}$$

$\forall \, j \in \{1, 2, ..., C-1\}$ and

$$Pr(Y \leq C) = \gamma^{(C)} = 1$$

The ordinal logistic regression (OLR) model considers a model for each comparison of all categories below a threshold to all categories above

$$\frac{Pr(Y \leq j)}{Pr(Y > j)} = \frac{\gamma^{(j)}}{1 - \gamma^{(j)}} \forall \, j \in \{1, 2, ..., C-1\}$$

The OLR model with $X_1, ..., X_k$ explanatory variables is then

$$log\left(\frac{Pr(Y_i \leq j)}{Pr(Y_i > j)}\right) = log\left(\frac{\gamma_i^{(j)}}{1 - \gamma_i^{(j)}}\right)$$
$$= \alpha^{(j)} - (\beta_1 X_{1i} + ... + \beta_k X_{ki})$$
$$Pr(Y_i \leq j) = \gamma_i^{(j)} = \frac{e^{\alpha^{(j)} - (\beta_1 X_{1i} + ... + \beta_k X_{ki})}}{1 + e^{\alpha^{(j)} - (\beta_1 X_{1i} + ... + \beta_k X_{ki})}}$$

Properties of coefficients are that $\alpha^{(1)} < \alpha^{(2)} < ... < \alpha^{(C-1)}$, to guarantee that $\gamma^{(1)} < \gamma^{(2)} < ... < \gamma^{(C-1)}$. $\beta_1, \beta_2, ..., \beta_k$ are the same $\forall \, j$. Finally,

$$Pr(Y_i = j) = Pr(Y_i \leq j) - Pr(Y_i \leq j-1) = \gamma^{(j)} - \gamma^{(j-1)}$$

5.1 Fitted Probabilities

Fitted probabilities for individual categories j are

$$P(Y = 1) = \gamma^{(1)} = \frac{e^{\alpha^{(1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}{1 + e^{\alpha^{(1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}$$
$$P(Y = j) = \gamma^{(j)} - \gamma^{(j-1)} = \frac{e^{\alpha^{(j)} - (\beta_1 X_1 + ... + \beta_k X_k)}}{1 + e^{\alpha^{(j)} - (\beta_1 X_1 + ... + \beta_k X_k)}} - \frac{e^{\alpha^{(j-1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}{1 + e^{\alpha^{(j-1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}$$

for $j \in \{2, ..., C-1\}$, and

$$P(Y = C) = 1 - \gamma^{(C-1)} = \frac{e^{\alpha^{(C-1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}{1 + e^{\alpha^{(C-1)} - (\beta_1 X_1 + ... + \beta_k X_k)}}$$

5.2 *p*-values and CIs

Same as with BLR

5.3 Proportional Odds Assumption

The OLR model assumes same coefficients $\beta_1, ..., \beta_k \; \forall \, j$. The increase increase in the odds of going from $Y \leq j$ to $Y > j$ associated with a increase in X, is the same $\forall \, j$.

6 Count Regression Models

6.1 Poisson Regression

Can use poisson distribution for count data, models probability of observing a number y of events per unit of observation.

$$p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$
$$\mathbb{E}[y|\lambda] = \lambda$$
$$Var[y|\lambda] = \lambda$$

with $y \in \{0, \mathbb{N}\}$, $\lambda \in (0, \infty)$. Poisson regression model is then

$$\log(\lambda_i) = \alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki}$$
$$\lambda_i = e^{\alpha + \beta_1 X_{1i} + ... + \beta_k X_{ki}}$$

Assumption that for given $\mathbb{E}[y] = \lambda$, $Var[y] = \lambda$ too is very strong. When there is overdispersion, Poisson will tend to yield SEs that are too small.

6.2 Negative-Binomial Regression

The negative-binomial relaxes this interpretation, allows for overdispersion (not for underdispersion). Following properties

$$\mathbb{E}[y|\lambda] = \lambda$$
$$Var[y|\lambda] = \lambda + \frac{\lambda^2}{\theta}$$

Estimates model of similar form as Poisson + θ.

6.3 *p*-values and CIs

Same as with BLR.

7 Properties of log and *e*

$$\log(e) = 1$$
$$\log(1) = 0$$
$$\log(x^r) = r\log(x)$$
$$\log(e^A) = A$$
$$e^{\log(A)} = A$$
$$\log(AB) = \log(A) + \log(B)$$
$$\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$$
$$e^{AB} = (e^A)^B$$
$$e^{A+B} = e^A e^B$$
$$e^{A-B} = \frac{e^A}{e^B}$$