Лекция 3 по микроэконометрике

Buchko Daniil, BEC175

16 октября 2020 г.

Пусть есть модель:

$$Y_i^* = x_i'\beta + z_i'\alpha + \varepsilon_i$$

Представим, что мы хотим проверить гипотезу о том, что переменные при α не влияют на модель:

$$H_0\colon \alpha = 0$$

Запишем функцию правдоподобия

$$\mathcal{L} = \prod_{i=1}^{n} \left(F(x_i'\beta + z_i'\alpha)\right)^{y_i} \left(1 - F(x_i'\beta + z_i'\alpha)\right)^{1-y_i}$$

$$\begin{aligned}
\ln \mathcal{L} &= \sum_{i=1}^{n} \left[y_i \ln F(x_i'\beta + z_i'\alpha) + (1-y_i)\left(1 - \ln F(x_i'\beta + z_i'\alpha)\right)\right] \\
\frac{\partial \ln \mathcal{L}}{\partial \beta} &= \sum_{i=1}^{n} \left[y_i \frac{f(x_i'\beta + z_i'\alpha)}{F(x_i'\beta + z_i'\alpha)} x_i - (1-y_i)\frac{f(x_i'\beta + z_i'\alpha)}{1 - F(x_i'\beta + z_i'\alpha)} x_i\right]_{\beta=\hat{\beta}} = 0 \\
&= \sum_{i=1}^{n} \underbrace{\frac{(y_i - F(x_i'\beta + z_i'\alpha))\, f(x_i'\beta + z_i'\alpha)}{F(x_i'\beta + z_i'\alpha)\,(1 - F(x_i'\beta + z_i'\alpha))}}_{\hat{\varepsilon}_i} x_i = 0 \\
\frac{\partial \ln \mathcal{L}}{\partial \alpha} &= \sum_{i=1}^{n} \frac{(y_i - F(x_i'\beta + z_i'\alpha))\, f(x_i'\beta + z_i'\alpha)}{F(x_i'\beta + z_i'\alpha)\,(1 - F(x_i'\beta + z_i'\alpha))} z_i = 0
\end{aligned}$$

Оказывается, что:

$$n\mathbb{R}^2 \sim \chi^2_{dim(r)}, \quad \text{где } \mathbb{R}^2\colon \underbrace{(1, \ldots, 1)}_{\text{n штук}} \text{ на } \underbrace{\hat{\varepsilon}_i x_i}_{\text{k штук}}, \ \underbrace{\hat{\varepsilon}_i z_i}_{\text{m штук}}, \ i \in \{1, \ldots, n\}$$

Какие проблемы есть в бинарных моделях?

Мультиколлинеарность нас не тревожит, потому что в линейных моделях она возникала в оценке коэффициентов регрессии из-за больших значений, возникающих в произведении $X^T X$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \quad Cov(\hat{\beta}) = (X^T X)^{-1}\sigma^2$$

Все-таки влияние есть, но не такое сильное как в обычных регрессиях.

Гетероскедастичность. Очень серьезная проблема. Для оценки параметров модели должно быть большое количество данных, из-за ассимптотичности оценок. А поскольку данных много, то можно ли считать их однородными (то есть одинаков ли разброс случайной ошибки во всей обучающей совокупности)? Если у ε своя дисперсия в каждом наблюдении, то это означает, что функцию правдоподобия мы записали неверно, так как мы её записывали в предположении, что дисперсия ошибок у всех наблюдений равна. А если мы неверно записали функцию правдоподобия, то получается, что

оценки могут быть несостоятельными.

Что делать с ней?
Пусть есть модель

$$Y_i^* = x_i'\beta + \varepsilon_i$$

и предположим, что у этой модели дисперсия случайной ошибки – функция от неизвестных переменных z_i. Возможно, что это те же самые переменные x_i, может быть это другие переменные:

$$Var(\varepsilon_i) = h(z_i'\alpha)K$$

$$K = \begin{cases} \frac{\pi^2}{3}, & \text{если logit} \\ 1, & \text{если probit} \end{cases}$$

Будем так же предполагать, что функция $h(.)$ гладкая, $h'(0) \neq 0$ и $h(0) = 1$. Теперь будем проверять гипотезу об отсутствии гетероскедастичности, то есть:

$$H_0\text{: } \alpha = 0 \iff Var(\varepsilon_i) = K$$

Используем тест множителей Лагранжа.

$$P(Y_i = 1) = P(Y_i^* > 0) = P(\varepsilon_i > -x_i'\beta)$$

Если дисперсия $Var(\varepsilon_i)$ является описанного выше вида, то чтобы привести к нормальному стандартному распределению, необходимо поделить аргументы на корень из дисперсии, а именно:

$$P(Y_i = 1) = P(Y_i^* > 0) = P\left(\frac{\varepsilon_i}{\sqrt{h_i}} > \frac{-x_i'\beta}{\sqrt{h_i}}\right) = F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)$$

Как после введения такой штуки поменяется функция правдоподобия?

$$\mathcal{L} = \prod_{i=1}^{n} F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)^{y_i} \left(1 - F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)\right)^{1-y_i} \to \max_{\beta}$$

Воспользуемся вычислениями с предыдущего пункта, а именно тем, что при взятии производной по бэте изменится только домножение на константу: $F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)$

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} \frac{\left(y_i - F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)\right) f\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)}{\left(1 - F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)\right) F\left(\frac{x_i'\beta}{\sqrt{h_i}}\right)} \frac{x_i}{\sqrt{h_i}} = 0 \tag{1}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} \hat{\varepsilon}_i \frac{x_i'\beta}{(\sqrt{h_i})^3} x_i'\beta \frac{h_i'}{(\sqrt{h_i})^3} z_i = 0 \tag{2}$$

Верхние равенства выполняются в точках $\beta = \hat{\beta}$ и $\alpha = \hat{\alpha}$. Соответственно, если у нас нет гетероскедастичности, то уравнения (1) и (2) совпадают с условиями первого порядка для обычной модели. При справделивости основной гипотезы $\alpha = 0$, то имеем следующее:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \hat{\varepsilon}_i x_i$$

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \hat{\varepsilon}_i (x_i'\beta) z_i$$

Соответственно нам нужно регрессировать на вектор единиц следующие вещи:

$$\hat{\varepsilon}_i x_i, \quad \hat{\varepsilon}_i (x_i'\beta) z_i$$

И если у нас нет гетероскедастичности, то

$$n\mathcal{R}^2 \sim \chi^2_m$$

Проверка нормальности остатков
Пусть мы предполагаем, что случайная ошибка имеет распределение кривых Пирсона. Если $\gamma_1 \neq 0$, то нормальное распределение перестает быть симметричным, а если $\gamma_2 \neq 0$, то случается островершинность.

$$F_{\varepsilon_i}(t) = \Phi(t + \gamma_1 t^2 + \gamma_2 t^3)$$

Учитывая, что ошибка имеет такую функцию распределения, а вероятность единицы в бинарных моделях соответствует распределению ε_i, можно записать, что:

$$P(Y_i = 1) = \Phi(x_i'\beta + \gamma_1 (x_i'\beta)^2 + \gamma_2 (x_i'\beta)^3)$$

Теперь, нам нужно доказать, что $\gamma_1 = \gamma_2 = 0$. Для этого вектор единиц регрессируем на $\hat{\varepsilon}_i x_i$, $\hat{\varepsilon}_i z_1$ и $\hat{\varepsilon}_i z_2$. И в случае выполнении нулевой гипотезы:

$$n\mathcal{R}^2 \sim \chi^2_2$$

Автокорреляции нет.

Метрики качества бинарной классификации
Обычно пользуется статистикой:

$$\mathcal{R}^2_{MF} = 1 - \frac{\ln L}{\ln L_0}$$

где l_0 - значение правдоподобия для модели без объясняющих переменных.