УДК 65.015.3:519.87 (045)

ОЦЕНКА ПЕРСОНАЛА С ИСПОЛЬЗОВАНИЕМ БИНАРНОЙ РЕГРЕССИИ

ЗИНЧЕНКО АЛЕКСЕЙ АЛЕКСЕЕВИЧ
аспирант кафедры «Прикладная математика», Финансовый университет, Москва, Россия
E-mail: *a_zinchenko@list.ru*

АННОТАЦИЯ

В статье показана возможность эффективной оценки кандидатов на должности при помощи моделей бинарного выбора. Применение математических методов в данной области способно повысить объективность принятия кадровых решений, а также упростить работу менеджеров по персоналу в случае осуществления массового подбора, что является обычной практикой для кадровых агентств.

Цель исследования — показать наличие статистической зависимости между информацией, указанной в резюме работника, и фактом прохождения испытательного срока. В качестве статистики автор располагал резюме работников, рекомендованных несколькими кадровыми агентствами своим клиентам. Среди этих данных присутствовали резюме людей, оставшихся работать в фирме и не прошедших испытательный срок. Для исследования были выбраны три типа моделей: пробит, логит и гомпит.

Для оценки параметров и качества построенных моделей разработана программа в среде Maple. Исходя из полученных результатов, была выбрана модель, наиболее удачно предсказывающая прохождение работником испытательного срока.

Ключевые слова: подбор персонала; бинарная регрессия; логит; пробит; гомпит.

STAFF APPRAISAL USING A BINARY REGRESSION

ALEX A. ZINCHENKO
graduate student of «Applied Mathematics», Financial University, Moscow, Russia
E-mail: *a_zinchenko@list.ru*

ABSTRACT

The paper shows the possibility of the efficient evaluation of candidates for positions with the help of the binary-regression. The absence of expertise in using math methods by personnel departments makes recruitment process modeling inefficient, so the results obtained via binary-regression is of great importance.

The purpose of the research is to show the relationship between the data in CVs and the fact of passing the probation period by employees. The author had at his disposal data of candidates' CVs provided by several HR-agencies to their clients. Some of employees had passed the probation, some of them had not passed. To carry out the research the author chose three types of models — logit, probit, gompit.

To estimate the parameters and the quality of the constructed models the author wrote the code in Maple computer algebra system. According to the results of the research the best predicting model was chosen.

Keywords: recruitment; binary regression; logit; probit; gompit.

Подбор высококвалифицированных кадров является если не первоочередной, то одной из главных задач в процессе организации деятельности любой компании. Поиск и отбор кандидатов, кадровый учет, аттестация кадров и их перестановка – функции кадровых служб в организациях. Самыми трудоемкими процессами, конечно же, остаются поиск и оценка профессионализма будущих работников. Это требует от кадровиков большого опыта, а если необходимо собрать большой штат за малое время, то понадобится еще и хорошая база соискателей. По этой причине многие компании предпочитают сотрудничать с кадровыми (рекрутинговыми) агентствами, основным видом деятельности

которых является подбор персонала, что позволяет осуществлять его на высоком профессиональном уровне. Поиском кандидатур и проведением собеседований занимается *HR*-отдел (*Human Resources*) самой компании. Для поиска кандидатов сотрудникам отдела предоставлена полная свобода, в отличие от других подразделений, где стараются перекрывать работникам доступ к социальным сетям на рабочих местах, в кадровом агентстве этот доступ может быть открыт по совершенно объективной причине — социальные сети предоставляют огромные возможности по поиску работников, сбору информации, проведению опросов.

Сотрудники *HR*-отдела должны прекрасно ориентироваться в потребностях клиента, а также обладать достаточным профессионализмом, чтобы оценить потенциального работника. Решение такой задачи требует всесторонних знаний, но каким бы широким ни был кругозор человека, фактор субъективности полностью исключить не удастся. Стоит ли говорить, что каждый «плохой» работник, предложенный рекрутинговым агентством и впоследствии уволенный клиентом, будет стоить агентству не только репутации. В случае если предложенный работник не проходит испытательный срок в компании, рекрутинговое агентство уменьшает сумму счета, выставленного за подбор. Для того чтобы повысить обоснованность принятия решений в процессе подбора персонала, автор предлагает использовать модели бинарного выбора.

Такие модели нашли применение в различных областях науки – социологии, физике, биологии. В эконометрике модели бинарного выбора используются для определения вероятности дефолта, в банковском деле – для принятия решений о выдаче кредита (кредитный скоринг).

Бинарная регрессия представляет собой зависимость эндогенной переменной, принимающей всего два значения – 0 и 1, от набора факторов [1]. Обычная линейная регрессия для таких переменных не применима, так как она допускает и отрицательные значения, и значения выше единицы. Поэтому обычно используются некоторые интегральные функции распределения, чаще всего функции нормального распределения (пробит), логистического распределения (логит) и распределения Гомперца (гомпит). От выбора функции распределения напрямую зависит соответствие прогнозов, полученных с помощью модели, реальным данным.

Предполагая, что зависимая переменная Y, которая представляет собой возможность или невозможность взять на работу кандидата (или, в случае с рекрутинговым агентством — рекомендовать его клиенту), принимает только два значения: $\{0;1\}$, вероятность того, что она примет соответствующее значение можно выразить, как функцию некоторых факторов:

$$Prob(Y=1|x)=F\left(x^T\beta\right);$$

$$Prob(Y=0|x)=1-F\left(x^T\beta\right).$$

Набор параметров β отражает влияние изменения каждого фактора на конечную вероятность. Задача состоит в том, чтобы подобрать адекватную функцию в правой части уравнения. Естественно предположить следующие условия:

$$\lim_{x^T\beta\to\infty} Prob(Y=1|x)=1;$$

$$\lim_{x^T\beta\to-\infty} Prob(Y=1|x)=0.$$

Оценка параметров β осуществляется методом максимального правдоподобия. Каждое наблюдение является схемой Бернулли, поэтому функция правдоподобия предстает в виде:

$$Prob(Y_1=y_1,\dots Y_n=y_n|\boldsymbol{X})=\prod_{y_i=0}\left[1-F\left(x_i^T\beta\right)\right]\prod_{y_i=1}F\left(x_i^T\beta\right).$$

Функцию правдоподобия для n наблюдений можно переписать в виде:

$$L(\beta|data)=\prod_{i=1}^{n}\left[F\left(x_i^T\beta\right)\right]^{y_i}\left[1-F\left(x_i^T\beta\right)\right]^{1-y_i}.$$

Задача поиска максимума функции правдоподобия равносильна задаче поиска нуля градиента $grad\,L(\beta)$.

Особенность применения бинарной регрессии для оценки работника заключается в необходимости дать количественную интерпретацию качественным переменным: образование, пол, навыки и др. Опыт работы может включать в себя также оценку самих организаций, в которых работал кандидат, оценку должностей, которые он занимал, и т.д.

Для оценки статистической значимости группы регрессоров модели используется так называемая статистика отношения правдоподобия [2], рассчитываемая по формуле

$$LR = 2 \cdot (L_{full} - L_{null}),$$

где L_{full} и L_{null} — значения логарифмической функции правдоподобия для оцененной модели, включающей все регрессоры, и для модели, состоящей из одной константы. Значение LR должно превышать критическое значение Хи-квадрат, с количеством степеней свободы, равным количеству факторов в модели.

Другим показателем качества модели является псевдокоэффициент детерминации. Существует несколько методик его расчета, в данной статье будет рассмотрен коэффициент детерминации МакФаддена [3], который рассчитывается по формуле

$$R^2_{pseudo} = 1 - \frac{L_{full}}{L_{null}}.$$

С помощью псевдокоэффициента детерминации так же можно оценивать необходимость включения каждого регрессора в модель. Для этого регрессоры последовательно включаются в модель и отслеживается изменение R^2_{pseudo}. Коэффициент детерминации МакФаддена не имеет абсолютной интерпретации, как классический R^2. С его помощью мы можем лишь сравнивать различные спецификации модели.

Итак, приступим к построению модели. В данной статье, в качестве функции $F\left(x^T \beta\right)$ будут последовательно выбраны функции нормального распределения, логистического и распределения Гомпертца.

В качестве факторов были выбраны следующие данные:

1) пол кандидата;

Таблица 1

Коэффициенты при переменных (пробит-модель)

β_0	−7,193
Пол кандидата	0,768
Возраст	−0,0502
Наличие высшего образования	3,870
Профиль	−1,051
Стаж работника (лет)	0,296
Количество организаций, в которых работал кандидат	−0,296
Перечисленные обязанности (количество)	−0,001
Знание английского языка	4,140
Другие иностранные языки, указанные в резюме	0,687
Уровень знания компьютера	0,360
Уровень запрашиваемой заработной платы	−0,0000062

2) возраст;

3) наличие высшего образования;

4) профиль;

5) стаж работника, лет;

6) количество организаций, в которых работал кандидат;

7) перечисленные обязанности, количество;

8) знание английского языка;

9) другие иностранные языки, указанные в резюме;

10) уровень знания компьютера [3-балльная шкала; 0 = не указан; 1 = знание *MS office*; 2 – знание специализированных пакетов анализа (*statistica*, *SAP* и т.д.), 3 = навыки программирования];

11) уровень запрашиваемой заработной платы; 0, если не указан.

Данные для статистики были взяты из резюме работников, рекомендованных

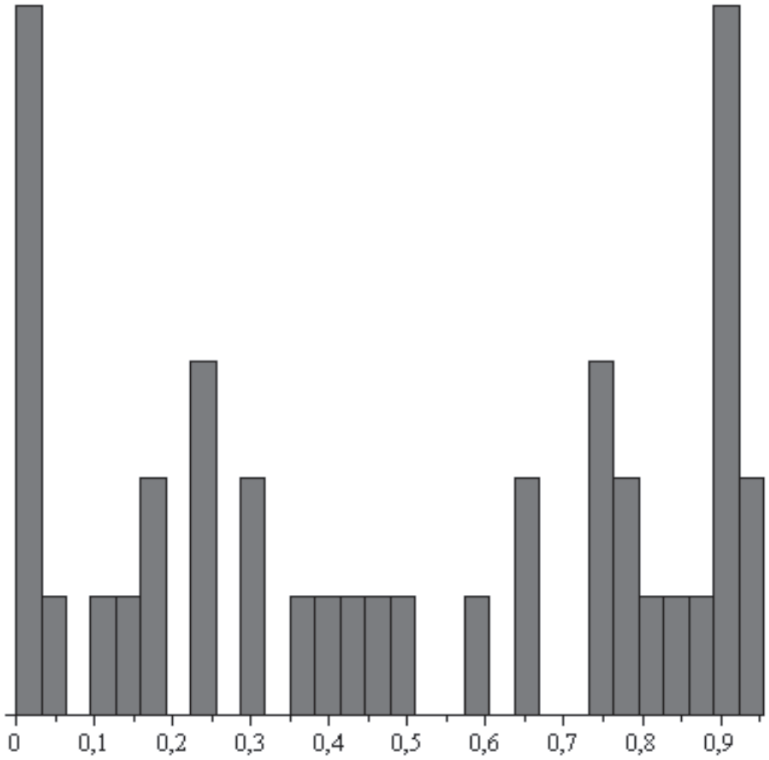

Рис 1. **Гистограмма предсказанных вероятностей, пробит-модель.**

несколькими кадровыми агентствами своим клиентам. Бинарная зависимая переменная принимает значение «1», в том случае, если человек продолжил работу в организации, по истечению испытательного срока, «0» в противоположном случае. Выбор в пользу этих факторов был сделан на основании расчета статистики LR и анализа изменений R^2_{pseudo} для различных спецификаций. Исключение любого из параметров из модели приведет к ухудшению ее качества.

Для оценки параметров модели, а также для расчета отношения правдоподобия и псевдокоэффициента детерминации автором написана программа в системе компьютерной алгебры *Maple*. Несмотря на то что существуют программные пакеты, в которых модели бинарного выбора уже реализованы, программный код данных пакетов является закрытым, и это уменьшает доверие к результатам вычислений. Другим аргументом в пользу пакета *Maple* является возможность проводить символьные вычисления, что крайне удобно для нахождения гессиана функции правдоподобия, необходимого для реализации многомерного метода Ньютона при поиске экстремума функции.

Начнем с пробит-модели:

$$Prob(Y=1|x)=\frac{1}{\sigma\sqrt{2\cdot\pi}}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}dx.$$

Оцененные коэффициенты при переменных представим в виде *табл. 1*.

$R^2_{pseudo}=0,42;$

$LR = 23,33 > \chi^2(11) = 19,68$;
при уровне значимости 95%.

Значение статистики *LR* показывает, что все параметры являются значимыми. Поскольку функция распределения всегда является монотонно возрастающей, по знаку параметра можно судить о том, какой вклад, отрицательный или положительный, он вносит в конечную вероятность. В нашем случае, поскольку значения всех факторов являются неотрицательными, можно утверждать, что знак коэффициента перед ним можно интерпретировать буквально. Так, например, знание иностранных языков оказывает положительное влияние на конечную вероятность,

Таблица 2

Коэффициенты при переменных (логит-модель)

β_0	$-16{,}867$
Пол кандидата	$0{,}956$
Возраст	$-0{,}094$
Наличие высшего образования	$9{,}472$
Профиль	$-1{,}8603$
Стаж работника (лет)	$0{,}436$
Количество организаций, в которых работал кандидат	$-0{,}588$
Перечисленные обязанности (количество)	$-0{,}009$
Знание английского языка	$9{,}859$
Другие иностранные языки, указанные в резюме	$0{,}937$
Уровень знания компьютера	$0{,}524$
Уровень запрашиваемой заработной платы	$-0{,}00001$

Таблица 3

Коэффициенты при переменных (гомпит-модель)

β_0	$-17{,}114$
Пол кандидата	$1{,}398$
Возраст	$-0{,}055$
Наличие высшего образования	$8{,}478$
Профиль	$-1{,}544$
Стаж работника (лет)	$0{,}211$
Количество организаций, в которых работал кандидат	$-0{,}291$
Перечисленные обязанности (количество)	$-0{,}007$
Знание английского языка	$8{,}5204$
Другие иностранные языки, указанные в резюме	$1{,}174$
Уровень знания компьютера	$0{,}502$
Уровень запрашиваемой заработной платы	$-0{,}000009$

что вполне очевидно. Столь же понятно и отрицательное влияние количества организаций, в которых работал кандидат. Уже на стадии обработки данных автор заметил тот факт, что люди, которые уволились в течение испытательного срока, могли сменить множество организаций за два-три года. Прошедшие испытательный срок, напротив, предпочитали не менять столь часто место работы.

Гистограмма предсказанных вероятностей (*рис. 1*) указывает на то, что большинство из них близки либо к 0, либо к 1. Это означает, что модель редко дает «неуверенную» оценку кандидатам.

Логит модель *(табл. 2)* дает несколько лучший результат:

$$Prob(Y=1|x)=\frac{e^x}{1+e^x};$$

$$R^2_{pseudo}=0{,}43;$$

$$LR = 24{,}046 > \chi^2(11) = 19{,}68;$$
при уровне значимости 95%.

На гистограмме (*рис. 2*) отчетливо видно, что логистическая функция реже предсказывает вероятность прохождения испытательного срока, близкую к 0,5, что, безусловно, дает большую уверенность в результате.

Гомпит-модель [4], сформированная на основе имеющихся данных (*табл. 3*, *рис. 3*), показала наихудший результат:

$$Prob(Y=1|x)=1-e^{-e^x};$$

$$R^2_{pseudo}=0{,}40;$$

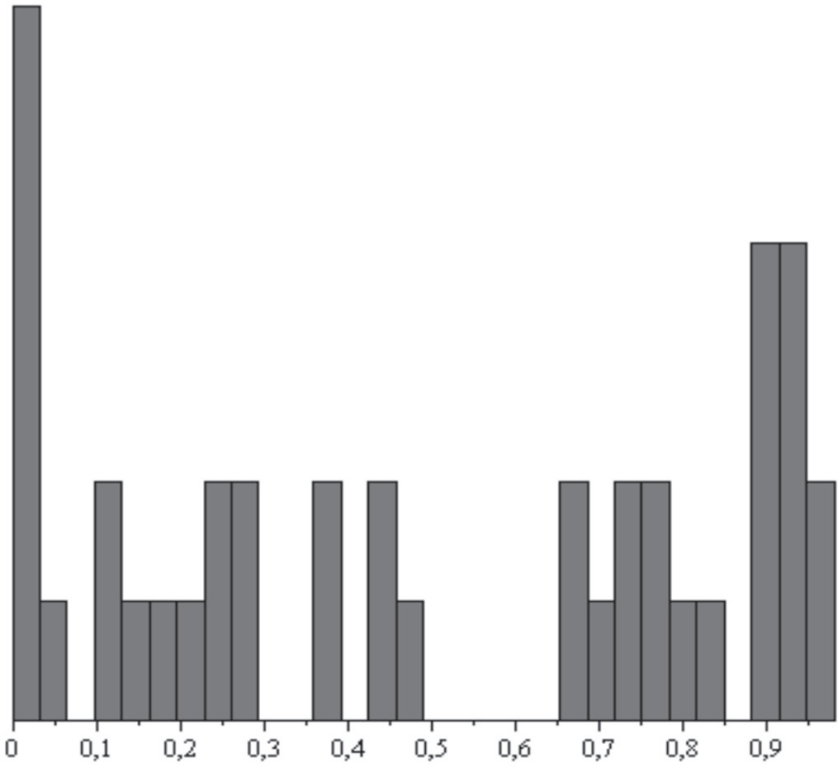

Рис. 2. **Гистограмма предсказанных вероятностей, логит-модель**

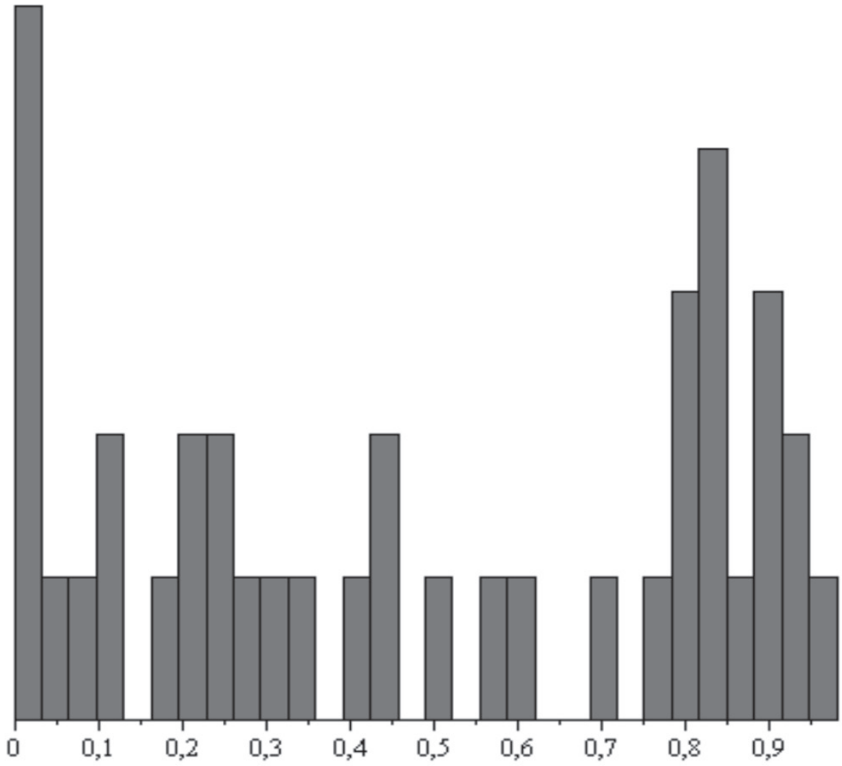

Рис. 3. **Гистограмма предсказанных вероятностей, гомпит-модель**

$LR = 24{,}046 > \chi^2(11) = 19{,}68$;
при уровне значимости 95%.

Выбор порогового значения вероятности для принятия окончательного решения зависит уже от руководителя кадровой службы самого предприятия. Очевидно, что чем выше пороговое значение, тем выше степень уверенности в будущем кандидате.

На сегодняшний день практически во всей литературе, посвященной подбору персонала, авторы делают акцент на применение *HR*-менеджерами методов, заимствованных преимущественно из психологии. Сотрудникам кадровых служб предлагается исследовать личность кандидата, опираясь на свой профессиональный опыт и интуицию [5]. Результаты, приведенные в статье, позволяют утверждать, что информация, которую люди указывают в резюме, несет в себе закономерности, поддающиеся статистическому анализу. Кроме того, модели бинарного выбора позволяют исследовать влияние отдельных факторов на конечную вероятность прохождения работником испытательного срока, благодаря чему рекрутинговые агентства и кадровые службы смогут, например, повысить эффективность корпоративных тренингов.

ЛИТЕРАТУРА

1. *Hand D.J.*, *Henley W.E.* Statistical classification methods in consumer credit // *Journal of the Royal Statistical Society*. Series A. 1997.
2. *Мхитарян В.С.*, *Архипова М.Ю.*, *Сиротин В.П.* Эконометрика: учебно-метод. комплекс. М., 2008.
3. *McFadden D. L.* Econometric analysis of qualitative response models // *Handbook of Econometrics*. Elsevier Science Publishers BV, 1984.
4. *Greene W.H.* Econometric Analysis. Prentict Hall, 2003.
5. *Tracy B.* Hire andKeep the Best People. Berrett-Koehler Publishers, 2001.

REFERENCES

1. *Hand D. J.*, *Henley W.E.* Statistical classification methods in consumer credit // *Journal of the Royal Statistical Society*. Series A. 1997.
2. *Mkhitaryan V.S., Arkhipov M.Y.*, *Sirotin V.P.* Econometrics: teaching method. complex. Moscow, 2008. (in Russ.)
3. *McFadden D. L.* Econometric analysis of qualitative response models // *Handbook of Econometrics*. Elsevier Science Publishers BV, 1984.
4. *Greene W.H.* Econometric Analysis. Prentict Hall, 2003.
5. *Tracy B.* Hire and Keep the Best People. Berrett-Koehler Publishers, 2001.