Линейная и логистическая регрессии

Регрессия (от англ. “обратное развитие”) – метод изучения зависимости между откликом Y и регрессором X (независимая переменная, признак)

Один из “теоретических” способов оценить зависимость – подсчитать

$$r(x) = \mathbb{E}(Y|X = x) = \int y\, f(y|x)dy$$

Задача состоит в том, чтобы оценить функцию $r(x)$ по данным

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y},$$

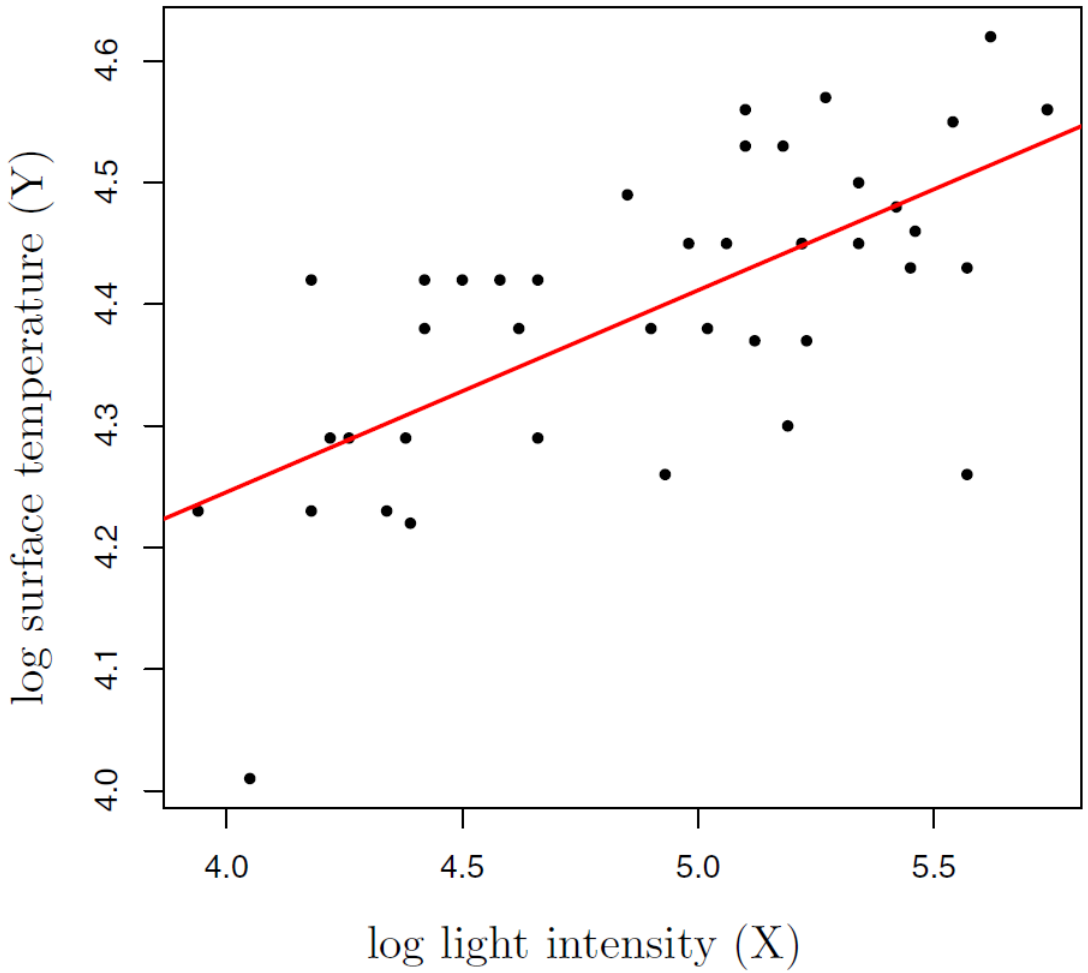
$F_{X,Y}$ **- совместное распределение** X **и** Y

а) Стандартная линейная регрессия

$$r(x) = \beta_0 + \beta_1 x$$

<u>Определение</u>. Пусть $\mathbb{E}(\epsilon_i|X_i) = 0$ **и** $\mathbb{V}(\epsilon_i|X_i) = \sigma^2$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



(данные о близлежащих звездах)

Пусть $\widehat{\beta}_0$ **и** $\widehat{\beta}_1$ **- оценки неизвестных параметров**

Подгонка регрессии

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Подогнанные (предсказанные) значения

$$\widehat{Y}_i = \widehat{r}(X_i)$$

3 **Остатки регрессии**

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)$$

Сумма квадратов остатков (RSS)

$$\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

$\widehat{\beta}_0$ **и** $\widehat{\beta}_1$ **- оценки неизвестных параметров с помощью метода наименьших квадратов, если** $\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2$ **для этих оценок минимальна**

Теорема. Оценки параметров β_0 и β_1 с помощью метода наименьших квадратов имеют вид

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n.$$

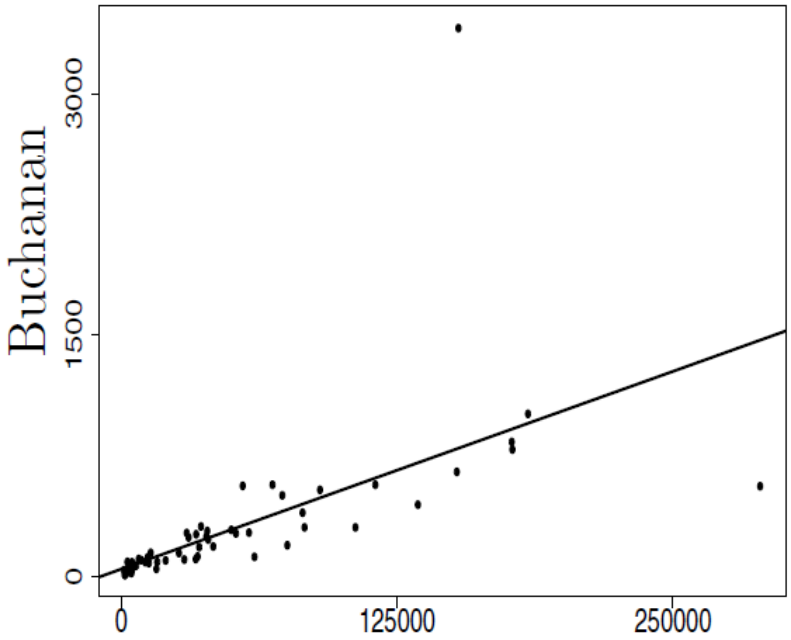
При это несмещенная оценка дисперсии шума σ^2 равна

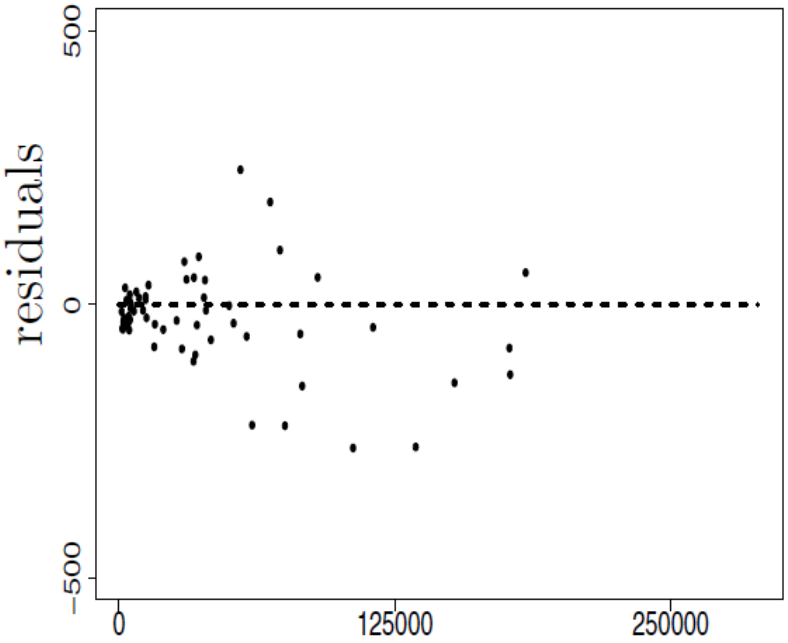
$$\widehat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2$$

Пример. (Звезды) Оценки равны: $\widehat{\beta}_0 = 3.58$ **и** $\widehat{\beta}_1 = 0.166$,

$\widehat{r}(x) = 3.58 + 0.166\, x$

<u>Пример</u>. (Выборы) Голоса за Buchanan (Y) vs. голоса за Bush (X) во Флориде.





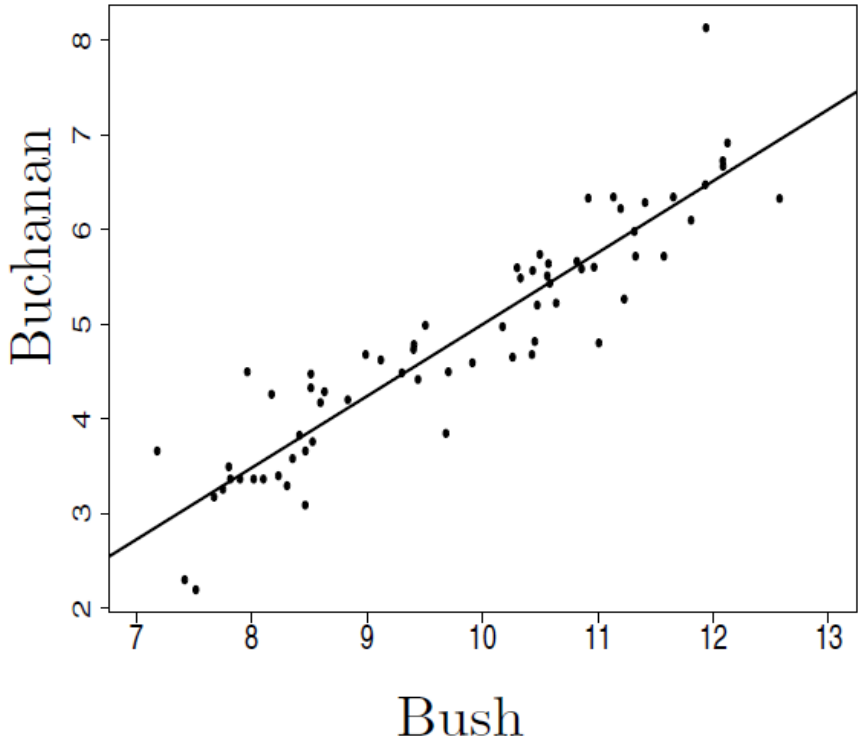
$$\widehat{\beta}_0 = 66.0991 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 17.2926$$

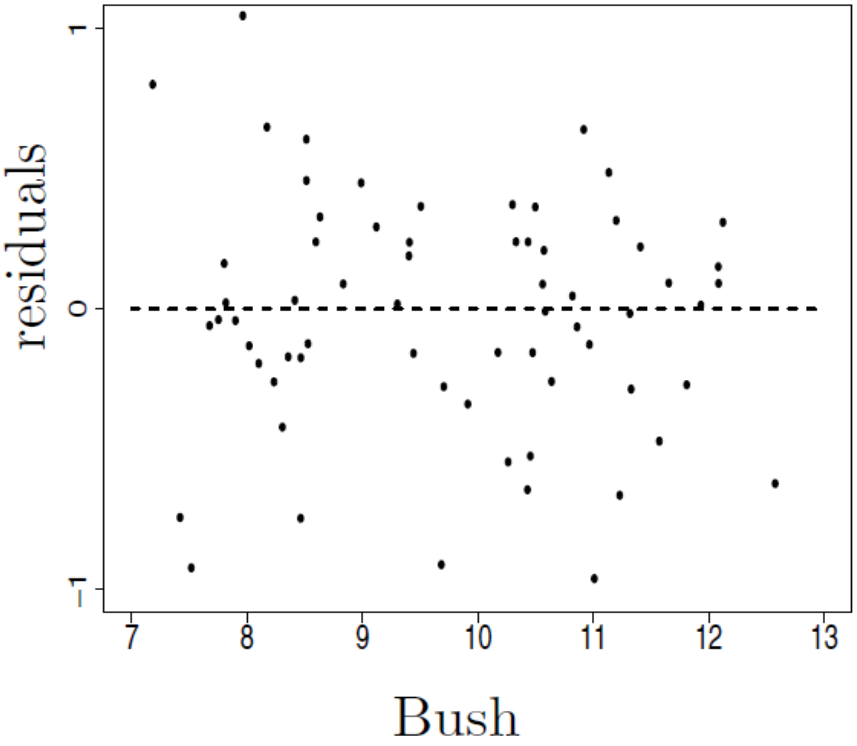
$$\widehat{\beta}_1 = 0.0035 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0002.$$

$$\text{Buchanan} = 66.0991 + 0.0035\,\text{Bush}$$

Если прологарифмировать данные, то остатки сильнее будут “напоминать” случайные числа

$$\widehat{\beta}_0 = -2.3298 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 0.3529$$
$$\widehat{\beta}_1 = 0.730300 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0358.$$
$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \ \log(\text{Bush})$$





b) Метод оценивания на основе минимизации невязок/максимизации правдоподобия

Предположим, что $\epsilon_i|X_i \sim N(0, \sigma^2)$

$$Y_i|X_i \sim N(\mu_i, \sigma^2)$$ **, где** $\mu_i = \beta_0 + \beta_1 X_i$

Правдоподобие имеет вид

$$\begin{aligned}
\prod_{i=1}^{n} f(X_i, Y_i) &= \prod_{i=1}^{n} f_X(X_i) f_{Y|X}(Y_i|X_i) \\
&= \prod_{i=1}^{n} f_X(X_i) \times \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \\
&= \mathcal{L}_1 \times \mathcal{L}_2
\end{aligned}$$

$$\mathcal{L}_1 = \prod_{i=1}^{n} f_X(X_i)$$

$$\mathcal{L}_2 = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)$$

Функция $\mathcal{L}_1$ не содержит параметры β_0 и β_1

Рассмотрим $\mathcal{L}_2$ - условную функцию правдоподобия

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2\right\}$$

$$\ell(\beta_0, \beta_1, \sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2 \quad \textbf{(*)}$$

ОМП (β_0, β_1) <=> максимизации (*) <=> минимизации

$$\text{RSS} \sum_{i=1}^{n}\Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2$$

<u>Теорема</u>. В предположении нормальности ОМП оценка совпадает с оценкой метода наименьших квадратов

Максимизируя $\ell(\beta_0, \beta_1, \sigma)$ **по** σ**, получаем ОМП оценку**

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\epsilon}_i^2$$

Пусть $X^n = (X_1, \ldots, X_n)$ **- значения регрессоров из обучающей выборки**

9 **с) Свойства оценок МНК**

<u>Теорема</u>. Пусть $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ **- оценка метода наименьших квадратов**

$$\mathbb{E}(\widehat{\beta}|X^n) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbb{V}(\widehat{\beta}|X^n) = \frac{\sigma^2}{n\, s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}$$

при $s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$

Таким образом,

$$\widehat{\mathsf{se}}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}.$$

10 **В дальнейшем мы будет опускать обозначения типа** $\widehat{\mathsf{se}}(\widehat{\beta}_0|X^n)$

<u>Теорема</u>. При выполнении условий регулярности (см. Bilodeau, Brenner, Theory of multivariate statistics)

1. $\widehat{\beta}_0 \xrightarrow{\mathrm{P}} \beta_0$, $\widehat{\beta}_1 \xrightarrow{\mathrm{P}} \beta_1$

2. $\dfrac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathsf{se}}(\widehat{\beta}_0)} \rightsquigarrow N(0,1)$ **и** $\dfrac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}(\widehat{\beta}_1)} \rightsquigarrow N(0,1)$

3. Приближенные доверительные интервалы размера $1-\alpha$ **для параметров**

$$\widehat{\beta}_0 \pm z_{\alpha/2}\,\widehat{\mathsf{se}}(\widehat{\beta}_0) \text{ и } \widehat{\beta}_1 \pm z_{\alpha/2}\,\widehat{\mathsf{se}}(\widehat{\beta}_1)$$

4. Тест Вальда для проверки $H_0 : \beta_1 = 0$ **vs.** $H_1 : \beta_1 \neq 0$ **имеет вид:**

H_0 **отклоняется, если** $|W| > z_{\alpha/2}$**, где** $W = \widehat{\beta}_1/\widehat{\mathsf{se}}(\widehat{\beta}_1)$

<u>Замечание</u>. Критерий Вальда для проверки $H_0 : \beta = \beta_0$ **vs.** $H_1 : \beta \neq \beta_0$ **имеет вид** $W = (\widehat{\beta} - \beta_0)/\widehat{\mathsf{se}}(\widehat{\beta})$

<u>Пример</u>. (Выборы) Для регрессии (в логарифмическом масштабе) 95% доверительный интервал имеет вид 0.7303±2*0.0358 = (0.66, 0.80). Статистика Вальда для проверки $H_0 : \beta_1 = 0$ **против альтернативы** $H_1 : \beta_1 \neq 0$ **равна** $|W| = |.7303 - 0|/.0358 = 20.40$**, причем p-value равно** $\mathbb{P}(|Z| > 20.40) \approx 0$ **=> действительно есть зависимость**

d) Прогнозирование

Модель - $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$**, построенная по выборке данных** $(X_1, Y_1), \ldots, (X_n, Y_n)$**. Необходимо предсказать значение отклика** Y_* **при** $X = x_*$

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$$

$$\mathbb{V}(\widehat{Y}_*) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = \mathbb{V}(\widehat{\beta}_0) + x_*^2 \mathbb{V}(\widehat{\beta}_1) + 2x_* \mathsf{Cov}(\widehat{\beta}_0, \widehat{\beta}_1)$$

=> можно подсчитать $\widehat{\mathsf{se}}(\widehat{Y}_*)$**, используя в качестве оценки** σ^2 **величину** $\widehat{\sigma}^2$**. Тем не менее, доверительный интервал для** Y_* **имеет вид, отличный от**

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{\mathsf{se}}$$

Теорема. Пусть

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_i (X_i - \overline{X})^2} + 1 \right)$$

Приблизительный prediction interval для Y_* **размера** $1 - \alpha$ **имеет вид**

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{\xi}_n$$

Пример. (Выборы)

$$\log(\text{Buchanan}) = -2.3298 + 0.7303 \log(\text{Bush})$$

В Palm Beach за Bush отдали 152 954 голосов, а за Buchanan – 3 476. В логарифмической шкале это составляет 11.93789 и 8.151045 соответственно. Насколько вероятен этот исход в предположении, что модель верна? Предсказание для Buchanan равно -2.3298 + 0.7303*11.93789 = 6.388441.

Существенно ли это больше, чем мы наблюдаем на практике? $\widehat{\xi}_n$ **= 0.093775 и 95% доверительный интервал имеет вид (6.2, 6.578), или, в исходных единицах – (493,717), что мало в сравнении с 3 476**

е) Множественная регрессия

В это случае данные имеют вид

$$(Y_1, X_1), \ldots, (Y_i, X_i), \ldots, (Y_n, X_n)$$

$$X_i = (X_{i1}, \ldots, X_{ik})$$

Модель имеет вид ($i = 1,\ldots,n$)

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i$$

$$\mathbb{E}(\epsilon_i | X_{1i}, \ldots, X_{ki}) = 0$$

Чтобы включить нулевой коэффициент, обычно полагают

$$X_{i1} = 1 \text{ при } i = 1, \ldots, n$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}$$

15

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

Теорема. Предположим, что матрица $X^T X$ размера $(k \times k)$ невырожденная, тогда

$$\begin{aligned} \widehat{\beta} &= (X^T X)^{-1} X^T Y \\ \mathbb{V}(\widehat{\beta}|X^n) &= \sigma^2 (X^T X)^{-1} \\ \widehat{\beta} &\approx N(\beta, \sigma^2 (X^T X)^{-1}) \end{aligned}$$

Оценка функции регрессии имеет вид

$$\widehat{r}(x) = \sum_{j=1}^{k} \widehat{\beta}_j x_j$$

$$\widehat{\sigma}^2 = \left(\frac{1}{n-k}\right) \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

$\widehat{\epsilon} = X\widehat{\beta} - Y$ **- вектор остатков**

Приближенный доверительный интервал размера $1-\alpha$ **для** β_j **равен**

$$\widehat{\beta}_j \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\beta}_j),$$

где $\widehat{\mathsf{se}}^2(\widehat{\beta}_j)$ **- j-ый диагональный элемент матрицы** $\widehat{\sigma}^2 (X^T X)^{-1}$

Пример. (Данные о преступлениях по 47 штатам США в 1960г. http://lib.stat.cmu.edu/DASL/Stories/USCrime.html)

Подгонка значений уровня преступности по 10-факторной модели

Регрессор	$\widehat{\beta}_j$	$\widehat{\mathrm{se}}(\widehat{\beta}_j)$	t-value	p-value
Нулевой коэффициент	-589.39	167.59	-3.51	0.001 **
Возраст	1.04	0.45	2.33	0.025 *
Южный штат(да/нет)	11.29	13.24	0.85	0.399
Образование	1.18	0.68	1.7	0.093
Расходы	0.96	0.25	3.86	0.000 ***
Труд	0.11	0.15	0.69	0.493
Количество мужчин	0.30	0.22	1.36	0.181
Численность населения	0.09	0.14	0.65	0.518
Безработные (14-24)	-0.68	0.48	-1.4	0.165
Безработные (25-39)	2.15	0.95	2.26	0.030
Доход	-0.08	0.09	-0.91	0.367

Возникают вопросы: (1) надо ли удалить незначимые переменные из модели? (2) Можно ли расценивать полученную зависимость как причинную?

f) Выбор модели
Бритва Оккама – не надо “плодить” сущности

Много переменных – большая дисперсия прогноза и маленькое смещение и наоборот

Возникают две задачи при выборке подходящей модели: (1) выбор значения целевой функции для характеризации качества используемой модели; (2) поиск оптимальной модели согласно выбранному критерию качества

Пусть $S \subset \{1, \ldots, k\}$ **и** $\mathcal{X}_S = \{X_j : \ j \in S\}$

β_S **- коэффициенты при соответствующих регрессорах,** $\widehat{\beta}_S$ **- оценки этих коэффициентов**

X_S **- подматрица матрицы плана** X **в соответствии с выбранным подмножеством регрессоров**

20 $\widehat{r}_S(x)$ **- оцененная функция регрессии,** $\widehat{Y}_i(S) = \widehat{r}_S(X_i)$ **- предсказанные значения**

Риск прогноза

$$R(S) = \sum_{i=1}^{n} \mathbb{E}(\widehat{Y}_i(S) - Y_i^*)^2$$

Y_i^* **- реальное значение выхода для** X_i

Задача состоит в том, чтобы выбрать такое S, что $R(S)$ принимает минимальное значение

Оценка риска прогноза (ошибка подгонки; ошибка на обучающей выборке)

$$\widehat{R}_{\mathrm{tr}}(S) = \sum_{i=1}^{n} (\widehat{Y}_i(S) - Y_i)^2$$

21 **<u>Теорема</u>. Оценка риска прогноза смещена "вниз" по сравнению с реальным значением риска прогноза**

$$\mathbb{E}(\widehat{R}_{\mathrm{tr}}(S)) < R(S)$$

$$\mathsf{bias}(\widehat{R}_{\mathrm{tr}}(S)) = \mathbb{E}(\widehat{R}_{\mathrm{tr}}(S)) - R(S) = -2 \sum_{i=1} \mathsf{Cov}(\widehat{Y}_i, Y_i)$$

Причина в том, что данные использовались дважды – для оценки параметров и для оценки риска прогноза

Если параметров много, то $\mathsf{Cov}(\widehat{Y}_i, Y_i)$ принимает большое значение, при этом прогноз на данных, отличных от данных в обучающей выборке, может оказаться существенно хуже!!!

Статистика C_p Mallow

$$\widehat{R}(S) = \widehat{R}_{\mathrm{tr}}(S) + 2|S|\widehat{\sigma}^2$$

$|S|$ - количество регрессоров

$\widehat{\sigma}^2$ - оценки дисперсии шума σ^2, полученная по полной модели (с включением всех возможных регрессоров)

Критерий = качество подгонки к обучающей выборке + “сложность” модели (регуляризация)

AIC(Akaike information criterion): выбрать S, для которого

$$\ell_S - |S|$$

принимает максимальное значение; ℓ_S - логарифм правдоподобия модели, где в качестве неизвестных параметров были подставлены их оценки, полученные с помощью максимизации ℓ_S

В линейной регрессии в случае нормальных ошибок (σ берется равным оценке, полученной по модели с максимальным количеством регрессоров) максимизации AIC эквивалента минимизации C_p

Риск можно оценить и с помощью кросс-проверки (cross-validation; leave-one-out)

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{(i)})^2$$

$\widehat{Y}_{(i)}$ **- предсказание значения** Y_i**, полученное с помощью модели, параметры которой были оценены по урезанной обучающей выборке без** Y_i

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2$$

$$U(S) = X_S (X_S^T X_S)^{-1} X_S^T$$

K-кратная кросс-проверка

1. Данные случайным образом делятся на k непересекающихся подвыборок (часто берут 10 подвыборок)

2. По одной подвыборке за раз удаляется (с возвращением), по остальным происходит оценка параметров. Данные в удаленной подвыборке оцениваются с помощью оцененной модели

25 **3. Риск полагается равным $\sum_i (Y_i - \widehat{Y}_i)^2$ (сумма берется по наблюдениям из удаленной подвыборки)**

4. Процесс повторяется для остальных подвыборок, после чего полученная оценка риска усредняется

Для линейной регрессии оценка на основе коэффициента C_p Mallow и оценка на основе K-кратной кросс-проверки зачастую совпадают. В более сложных случаях кросс-проверка работает лучше

BIC (байесовский информационный критерий): выбирается та модель, для которой

$$\mathrm{BIC}(S) = \ell_S - \frac{|S|}{2}\log n$$

принимает максимальное значение

Этот функционал имеет байесовскую интерпретацию. Пусть $\mathcal{S} = \{S_1, \ldots, S_m\}$ **обозначает множество возможных моделей. Допустим, что априорное распределение имеет вид** $\mathbb{P}(S_j) = 1/m$**. Также предположим, что параметры внутри каждой модели имеют некоторое "гладкое" априорное распределение. Можно показать, что апостериорная вероятность модели примерно равна**

$$\mathbf{P}\left(S_j \mid \textit{выборка}\right) \approx \frac{e^{BIC(S_j)}}{\sum_r e^{BIC(S_r)}}$$

Выбор модели с наибольшим BIC <=> выбор модели с наибольшей апостериорной вероятностью

BIC также можно интерпретировать с точки зрения теории минимальной длины описания информации
BIC обычно “выбирает” модели с меньшим числом параметров

Если в модели максимальное количество регрессоров равно k, то существует 2^k всевозможных моделей

В идеале необходимо “просмотреть” все модели, поставить каждой в соответствие значение критерия качества и выбрать наилучшую согласно этому критерию качества

Количество регрессоров большое => регрессия методом включений, исключения, включений-исключений

Включения: на первом шаге регрессоров нет вообще; далее добавляется регрессор, для которого критерий качества максимальный и т.д.

Выключения: на первом шаге количество регрессоров максимальное; на каждом шаге удаляется регрессор, приводящий к наименьшему уменьшению критерия качества

<u>Пример</u>. (Данные о преступлениях) Используем критерия AIC (в данном случае в силу определения происходит минимизация, а не максимизация AIC) <=> минимизации C_p Mallow

В модели с полным набором регрессоров AIC = 310.37. В порядке убывания AIC при удалении каждой из переменных равен
Численность населения (AIC = 308), Труд (AIC = 309), Южный штат (AIC = 309), Доход (AIC = 309), Количество мужчин (AIC = 310), Безработные I (AIC = 310), Образование (AIC = 312), Безработные II (AIC = 314), Возраст (AIC = 315), Расходы (AIC = 324)

Таким образом, имеет смысл удалить переменную “Население”

Южный штат (AIC = 308), Труд (AIC = 308), Доход (AIC = 308), Количество мужчин (AIC = 309), Безработные I (AIC = 39), Образование (AIC = 310), Безработные II (AIC = 313), Возраст (AIC = 313), Расходы (AIC = 329)

И т.д.

Уровень преступности = 1.2 Возраст + 0.75 Образование + 0.87 Расходы + 0.34 Количество мужчин – 0.86 Безработные I + 2.31 Безработные II.

Не дан ответ на то, какие переменные вызывают рост уровня преступности!

Метод выборка подмножества регрессоров Zheng/Loh

В основе метода предположения о том, что часть коэффициентов β_j в точности равна 0 => ищется подмодель, состоящая из ненулевых коэффициентов β_j

30 **1. Погоняется модель с полным набором из k регрессоров. Пусть $W_j = \widehat{\beta}_j / \widehat{\mathsf{se}}(\widehat{\beta}_j)$ обозначает статистику Вальда для $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$**

2. Строится вариационный ряд абсолютных значений статистик Вальда

$$|W_{(1)}| \geq |W_{(2)}| \geq \cdots \geq |W_{(k)}|$$

3. Пусть $\widehat{j}$ - значение индекса j, на котором достигается минимум функционала

$$\mathrm{RSS}(j) + j\,\widehat{\sigma}^2\,\log n$$

$\mathrm{RSS}(j)$ **- сумма квадратов остатков для модели, в которой были использованы j регрессоров с наибольшими значениями статистики Вальда**

4. В качестве окончательной модели выбирается модель с $\widehat{j}$ регрессорами, имеющими наибольшие абсолютные значения статистики Вальда

При некоторых достаточно общих условиях этот метод позволяет с вероятностью единица определить точную модель при все увеличивающемся объеме выборки

g) Логистическая регрессия

До сих пор предполагалось, что Y_i принимает действительные значения

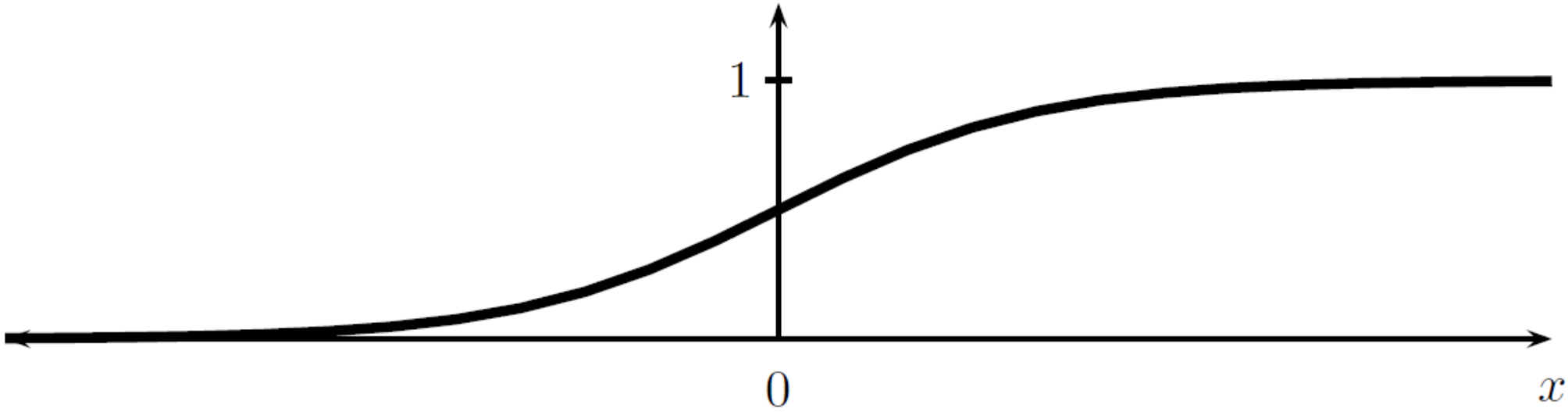
Логистическая регрессия – параметрический метод регрессии для случая, когда $Y_i \in \{0, 1\}$. Для k-мерного регрессора модель имеет вид

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}}}$$

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

“Логистическая” регрессия <=> $e^x/(1+e^x)$ **- логистическая кривая**



Так как $Y_i \in \{0,1\}$ **, то** $Y_i|X_i = x_i \sim \text{Bernoulli}(p_i)$

Значит, условная функция правдоподобия имеет вид

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i}(1-p_i(\beta))^{1-Y_i}$$

Оценка параметров $\widehat{\beta}$ получается за счет максимизации $\mathcal{L}(\beta)$ численным образом

EM-алгоритм

Выбирается начальная точка $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$, по формуле

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^k \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^k \beta_j x_{ij}}}$$

подсчитываются p_i^0 при $i = 1, \ldots, n$. Пусть $s = 0$. Выполняются итерации:

1. **Подсчитываются**

$$Z_i = \text{logit}(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1 - p_i^s)}, \quad i = 1, \ldots, n$$

2. Пусть W **- диагональная матрица, в которой** (i,i) **элемент равен** $p_i^s(1-p_i^s)$

3. Подсчитываются

$$\widehat{\beta}^s = (X^T W X)^{-1} X^T W Y$$

(взвешенная линейная регрессия Z на Y)

35

4. $s = s + 1$**. GOTO 1**

Оценку матрицы Фишера также можно получить численным методом. В таком случае оценка стандартной ошибки $\widehat{\beta}_j$ **это** (j,j) **элемент матрицы** $J = I^{-1}$**. Выбор количества регрессоров обычно делается с помощью AIC** $\ell_S - |S|$

Пример. (коронарная болезнь сердца) 462 мужчины, 15 - 64 лет, 3 сельских местности в Южной Африке. Выход: наличие ($Y = 1$) / отсутствие ($Y = 0$) коронарной болезни сердца. 9 регрессоров: систолическое давление, потребление табака (kg), ldl (low density lipoprotein cholesterol), ожирение, склонность родственников к сердечным заболеваниям, typea (поведение типа А), тучность, потребление алкоголя, возраст

Регрессор	$\widehat{\beta}_j$	$\widehat{\text{se}}$	W_j	p-значение
нулевой коэффициет	-6.145	1.300	-4.738	0.000
сист. кров. дав.	0.007	0.006	1.138	0.255
табак	0.079	0.027	2.991	0.003
ldl	0.174	0.059	2.925	0.003
ожирение	0.019	0.029	0.637	0.524
склонность	0.925	0.227	4.078	0.000
тип А	0.040	0.012	3.233	0.001
тучность	-0.063	0.044	-1.427	0.153
алкоголь	0.000	0.004	0.027	0.979
возраст	0.045	0.012	3.754	0.000

Систолическое давление не оказывает влияние на коронарную болезнь сердца? Перед регрессором “тучность” стоит знак минус? – эффекты, вызванные взаимозависимостью регрессоров.

Тот факт, что кровяное давления не является значимым, вовсе не означает, что кровяное давление не является существенной причиной появления болезни. Это означает, что регрессор “кровяное давление” не является “существенным” по сравнению с другими переменными в модели

h) AIC

Рассмотрим множество моделей $\{M_1, M_2, \ldots\}$**. Пусть** $\widehat{f}_j(x)$ **обозначает оценка плотности, полученную с помощь максимизации функции правдоподобия для модели** M_j**, то есть** $\widehat{f}_j(x) = \widehat{f}(x; \widehat{\beta}_j)$**, где** $\widehat{\beta}_j$ **- ОМП параметров** β_j **для модели** M_j

$$D(f,g) = \sum_x f(x) \log \left(\frac{f(x)}{g(x)} \right)$$

Соответствующая функция риска имеет вид

$$R(f, \widehat{f}) = \mathbb{E}(D(f, \widehat{f}))$$

Отметим, что $D(f, \widehat{f}) = c - A(f, \widehat{f})$ **, где** $c = \sum_x f(x) \log f(x)$ **не зависит от** $\widehat{f}$ **, и**

$$A(f, \widehat{f}) = \sum_x f(x) \log \widehat{f}(x)$$

Минимизация риска эквивалентна максимизации $a(f, \widehat{f}) \equiv \mathbb{E}(A(f, \widehat{f}))$

Оценка величины $a(f, \widehat{f})$ **с помощью** $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ **имеет существенное смещение, причем смещение пропорционально** $|M_j|$

Таким образом, получаем, что $AIC(M_j)$ **примерно является несмещенной оценкой величины** $a(f, \widehat{f})$