Задание 3:Выбор модели, линейная регрессия.

1. Скачать данные о пробеге автомобиля (в милях) на единицу расхода горючего

   $$http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html$$

   (a) Подогнать простую линейную регрессию к данным, чтобы предсказать значение переменной MPG (miles per gallon) от значений переменной HP (horsepower). Проанализировать полученные результаты, снабдив их графиком, на котором изображена выборка и оцененная регрессионная зависимость

   (b) Повторить эксперимент из предыдущего пункта, но при этом использовать $\log(\text{MPG})$ в качестве отклика регрессии. Сравнить качество подгонки полученной зависимости с качеством подгонки зависимости из предыдущего пункта (по сумме квадратов остатков подгонки исходных значений MPG).

   (c) Подогнать к данным множественную линейную регрессию, чтобы предсказать значение переменной MPG от всех остальных переменных. Проанализировать полученные результаты.

   (d) Использовать коэффициент C_p (см. слайды 22-28 лекции "Линейная и логистическая регрессии") для того, чтобы выбрать наилучшее подмножество регрессоров. Использовать и прямой и обратный варианты пошагового выбора. Проанализировать полученные результаты.

2. Допустим, что в регрессионной модели $y = \sum_{j=1}^{k} \beta_j x_j + \varepsilon$ шум $\varepsilon \sim N(0, \sigma^2)$ и дисперсия σ^2 – известна. Показать, что модель с наибольшим значением AIC является моделью с наименьшим значением статистики Mallow C_p.

3. Пусть $X_1, \ldots, X_n$ – i.i.d. наблюдения. Рассмотрим две модели – M_0 и M_1.

   $$M_0 : X_1, \ldots, X_n \sim N(0, 1),$$
   $$M_1 : X_1, \ldots, X_n \sim N(\theta, 1),\ \theta \in \mathbb{R}.$$

   По сути, критерии типа AIC (см. слайды 22-26 лекции "Линейная и логистическая регрессии") позволяют рассмотреть проблему выбора между двумя гипотезами $H_0 : \theta = 0$ и $H_1 : \theta \neq 0$ с точки зрения выбора наилучшей модели. Пусть $l_n(\theta)$ – логарифм функции правдоподобия. Значение AIC для модели M_0 составляет $AIC_0 = l_n(0)$, а значение AIC для модели M_1 составляет $AIC_1 = l_n\left(\hat{\theta}\right) - 1$. Допустим, что выбирается модель с наибольшим значением AIC. Пусть J_n обозначает номер выбранной модели

   $$J_n = \begin{cases} 0, & \text{если } AIC_0 > AIC_1; \\ 1, & \text{если } AIC_1 > AIC_0. \end{cases}$$

(a) Допустим, что модель M_0 – верная. Необходимо найти

$$\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right).$$

Также найдите $\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right)$ при $\theta \neq 0$.

(b) Пусть $\phi_\theta(x)$ обозначает плотность нормального распределения, среднее значение которого равно θ, а дисперсия равна 1. Определим

$$\hat{f}_n(x) = \begin{cases} \phi_0(x), & \text{если } J_n = 0; \\ \phi_{\hat{\theta}}(x), & \text{если } J_n = 1. \end{cases}$$

Если $\theta = 0$, то показать, что $D\left(\phi_0, \hat{f}_n\right) \to 0$ по вероятности при $n \to \infty$, где

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

является расстоянием Кульбака. Показать также, что $D\left(\phi_\theta, \hat{f}_n\right) \to 0$ по вероятности при $n \to \infty$, если $\theta \neq 0$. Таким образом, AIC состоятельно "оценивает" настоящую плотность распределения несмотря на то, что $\lim_{n\to\infty} \mathrm{P}\left(J_n = 0\right) \neq 1$ при $\theta = 0$.