## Постановка задачи

Есть объекты $1 \dots n$, есть признаки $Y, X_1 \dots X_k$, меняющиеся на объектах. $X_1 \dots X_k$ – объясняющие переменные (факторы, признаки)
$Y$ – отклик. Задача: объяснить $Y$ через $X_1 \dots X_k$, т.е. найти такую ф-цию $f$, что $Y \approx f(X_1 \dots X_k)$.
Ц.т. о наилучшем квадратич. прогнозе, $\arg\min_f E(Y - f(X_1 \dots X_k))^2 = E(Y | X_1 \dots X_k)$
Если ищем $f$ в произв. случае, то просто модель регрессии
(будем изучать эту модель в рамках непараметрич. регрессии)
Если ищем $f$ как $\theta_0 + \sum_{i=1}^{k} \theta_i X_i$, то модель линейной регрессии
(для гаусс. случая оптимальна, т.е. $E(Y | X_1 \dots X_k)$ есть лин. ф-ция от $X_1 \dots X_k$)
Т.е. задача (линейной регрессии) найти по $Y$ и $X$ оценку неизвест. вектора пар-ров $\theta$ в такой модели: $Y = X\theta + \varepsilon$,
где $X = (X_0 X_1 \dots X_k)$ – матрица $n \times (k+1)$, $X_0 = (1 \dots 1)^T$, $\varepsilon$ – об этом позже.
Знаем из метода наименьших квадратов, что (МНК-метод)
$\hat\theta = \arg\min_\theta \|Y - X\theta\|^2 = (X^TX)^{-1}X^TY$ и сама оц. для $Y$: $\hat Y = X\hat\theta$.

## Предположения Гаусса–Маркова

$E\varepsilon = 0$, $D\varepsilon = \sigma^2 I_n$, где $I_n = \mathrm{diag}\{1, \dots 1\}$ – единич. матрица разм. $n \times n$,
$\sigma^2$ – неизвестно. В этом случае $\hat\theta$ – оптимальная (с наим. дисперсией) оценка $\theta$ в классе линейных оценок вида $BY$, где $B$ – матрица (т. Гаусса–Маркова)
Знаем, кроме того, что $\hat\theta$ – несмещ. и состоят. оц. $\theta$.
Несмещ. и состоят. оц. для $\sigma^2$ служит $\hat\sigma^2 = \frac{1}{n-k}\|Y - X\hat\theta\|^2$
В случае же, если $\varepsilon \sim N(0, \sigma^2 I_n)$, то $\hat\theta$ – оптимальная (в классе всех несмещённых оценок, а не только линейных)

Теперь, собственно, к вопросам модели:
1) Все ли признаки значимы? (т.е. проверить гипотезу о том, что $\theta_j = 0$)
2) Предсказать $Y$ на новом объекте (т.е. дов. инт. построить)
3) Проверить, адекватна ли модель.

Итак, к значимости признаков. Проверим гипотезу $H_0: \theta_j = 0$.
Мы знаем, что $\hat\theta \sim N(\theta, \sigma^2(X^TX)^{-1})$, отсюда вытекает
$\frac{c^T(\hat\theta - \theta)}{\hat\sigma\sqrt{c^T(X^TX)^{-1}c}} \sim t_{n-k-1}$, где $c$ – вектор из $\mathbb{R}^k$. Если хотим проверить гипотезу $H_0: \theta_j = 0$, берем $c_j = (0 \dots \underset{j}{1} \dots 0)$ и получаем $\frac{\hat\theta_j - \theta_j}{\hat\sigma\sqrt{a_{jj}}} \sim t_{n-k-1}$, где $a_{jj} = \left((X^TX)^{-1}\right)_{jj}$, отсюда получаем доверит. интервал для $\theta_j$, откуда и будет следовать критерий.
ДИ: $\left(\hat\theta_j - u_{1-\frac{\alpha}{2}}\hat\sigma\sqrt{a_{jj}}, \hat\theta_j + u_{1-\frac{\alpha}{2}}\hat\sigma\sqrt{a_{jj}}\right)$.
где $u_{1-\frac{\alpha}{2}}$ – $(1-\frac{\alpha}{2})$-кв. $t_{n-k-1}$
($t$-критерий Стьюдента)
Критерий: если 0 не попал в этот ДИ, то отвергаем $H_0$ (в пользу значимости признака $j$)

Теперь проверим гипотезу о том, что несколько признаков не явл. значимыми. Для этого введём коэфф. детерминации $R^2$:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \text{ где } RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Логично, что чем меньше RSS, тем лучше мы приблизили наши $y_i$, т.е. чем ближе $R^2$ к 1, тем лучше качество регрессии (подробнее об этом ещё поговорим). Кроме того, $R^2 = (\hat{\rho}(y, \hat{y}))^2$, где $\hat{\rho}(y, \hat{y})$ - коэфф. корр. Пирсона между $y$ и $\hat{y}$.

**Критерий Фишера** Проверим гипотезу $H_0: \beta_{k_1} = \dots = \beta_{k_q} = 0$.

Пусть $RSS_{H_0}$ построена по модели, где у нас отсут. признаки $k_1 \dots k_q$, т.е. $\hat{y}_{H_0} = X_{H_0}(X_{H_0}^T X_{H_0})^{-1} X_{H_0}^T y$, где $X_{H_0}$ - матрица с отсут. столбцами $k_1, k_2, \dots k_q$, т.е. $(X_1 X_2 \dots X_{k_1-1} X_{k_1+1} \dots)$

Тогда если $H_0$ верна, то $\frac{(RSS - RSS_{H_0})/q}{RSS/(n-k-1)} \sim F_{q, n-k-1}$

Для проверки гипотезы $H_0: \beta_1 = \dots = \beta_k = 0$ используем такой крит. Фишера:

$$F = \frac{R^2/k}{(1-R^2)/n-k-1}, \quad F \sim F(k, n-k-1) \text{ при верной } H_0.$$

2) И для значения отклика на новом наблюд. $x_0$: $y(x_0) = x_0^T\theta + \varepsilon(x_0)$ строится так: $\left(x_0^T\hat{\theta} - u_{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0 + 1},\ x_0^T\hat{\theta} + u_{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}\right)$ (вместо с на предыд. странице подст. $x_0$ и учитываем, что дисп. $\varepsilon(x_0)$ равна $\sigma^2$)

3) Адекватность модели и model selection.

Видели (по графику, по критериям, по $R^2$), что выбранная модель плохо данное описывает. Возможные (↑ позднее) варианты действий:

1) Замена переменных. Т.е. применяем к $y$ и $X$ функции. Возмож. варианты ф-ций: $\ln x$, $e^x$, $e^{-x}$, $\frac{1}{x}$, и т.д. Здесь важно проверить, что при переходе к новым переменным $\varepsilon_i$ продолжали оставаться независимыми и, по возможности, были нормальны распред.
2) Добавление переменных вида $X_i^k$, т.е. были просто $y = \sum \theta_i x_i$, а тут ещё степени $x_i$ добавили. Может привести к переобучению, т.е. нестабильности решения на новых объектах.
3) Убрать перемен. с помощью критериев Стьюдента и Фишера.

Как выяснить, какая модель лучше?

а) Информационные "критерии" Акаике и Шварца (нормальность тут не требуется) Акаике: $AIC = 2k + n\left[\ln\frac{RSS}{n} + 1\right]$;

Шварц: $BIC = 2k\ln n + n\left[\ln\frac{RSS}{n} + 1\right]$. Выбираем ту модель, у которой наименьший AIC и/или BIC.

Инф. критерии не явл. стат. критериями проверки гипотез, к ним не применимо понятие "значимости критерия". Просто числовые характеристики.

б) Критерий RESET Рэмси

В рамках мн. гаусс. модели проверяет предположение $E\varepsilon_i = 0$ (мы ж хотим, чтобы наша оценка отклика была несмещённой!)
Можно выявить с помощью этого крит.:
1) наличие пропущенных, значимых для регрессии факторов
2) неправильную функц. форму регрессии
3) наличие коррелиции между факторами (мультиколлинеарность)
Правда, заранее неясно, что именно выявим.
Как устроен критерий: с помощью МНК получаем $\hat{y}_i$ – оценки отклика, потом рассм. вспомогат. модель $y_i = \sum_{j=1}^{k} \theta_j x_{ij} + \gamma_1 (\hat{y}_i)^2 + \eta_i$, $i = 1..n$
и проверяется гипотеза $H_0: \gamma_1 = 0$. Для этого вычисляем МНК новой модели, $\tilde{y}_i$ получаем, и пишем такую статистику:
при верной $H_0$ $F = \dfrac{RSS_{(1)} - RSS_{(2)}}{RSS_{(2)}/(n-k-1)} \xrightarrow{d} F_{1,\,n-k-1}$.

Замечание 1) Почему важно убрать незначимые признаки? Потому что при их убирании оценка $\hat{\theta}$ остаётся несмещённой, а дисперсия снижается.
в) Стандартный коэфф. детерминации $R^2$ всегда увелич. при добавлении признаков в модель, поэтому для отбора признаков его нельзя использовать. Можно брать приведённый коэфф. детерминации:
$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}.$$
г) Пошаговая регрессия. Вообще по регрессионному анализу хорошая книга Сэбера „Линейный регрессионный анализ"
Шаг 0 Для опис. у берём модели либо только с константой, либо с 1 перемен. Выбираем 2 порога вер. $p_{IN}$, $p_{out}$ ($p_{IN} = 0.05$, $p_{out} = 0.1$). Перем. включ. в модель, если у неё наимен. p-значение по крит. Фишера (см. стр. 2, $\dfrac{(RSS - RSS_{H_0})/q}{RSS/(n-k-1)}$ – вот этот) уровня $p_{IN}$ и $p_{out}$
Шаг 2N+1 Если есть перемен., при удалении которых крит. Фишера даёт p-знач. > $p_{out}$, то удаляем их.
Шаг 2N Если есть перем., при добавл. которых крит. Фишера даёт p-знач. < $p_{IN}$, то добавляем их.
Мысфесть. проверку гипотез тут не устроить, потому что процедура пошаговая.

---

Что делать с категориал. переменными?
1) Кодировать, присвоив каждому призн. номер. Двоич. кодом номера. Но мне сказал программист, что много места ест.
2) Заполнять частотами встречания категор. признака. Надо это делать аккуратно, чтобы не переобучиться (касается задач прогнозирования)
3) Хэширование. См. курс маш. обучения