Если у нас не выполняются предполож. модели (прежде всего, $E\varepsilon=0$, $D\varepsilon=\sigma^2 I_n$, ε – независимы, $\varepsilon\sim N$, X^TX – обратима), то оценки коэфф. регрессии могут оказаться смещёнными и неустойч., доверит. интервалы будут иметь другие уровни доверия, т. е. всё нарушится. Поэтому надо уметь проверять эти предположения.

1) $E\varepsilon=0$ – это к выбору адекватной модели относится, уже рассмотрели

2) $D\varepsilon=\sigma^2 I_n$. Такая ситуация нарушается гомоскедастичностью, обратная, т.е. когда не все дисперсии равны, – гетероскедастичностью.

Рассмотрим $e_i=y_i-\hat y_i$, $\hat y=(X(X^TX)^{-1}X^Ty)_i$ – i-тая коорд. оценки отклика.

Если $E\varepsilon=0$, то $Ee_i=E(y-\hat y)=E(y-X\hat\theta)=X\theta-X\theta=0$

$e=y-\hat y=(I-H)y$, где $H=X(X^TX)^{-1}X^T$ (эта часть лекции называется и анализ остатков)

$\operatorname{Var} e=\operatorname{Var}(y-X\hat\theta)=\operatorname{Var}((I-H)y)=(I-H)\operatorname{Var}y(I-H)^T=\sigma^2(I-H)(I-H)^T$

$=\sigma^2(I-H)$, т.к. $H^2=H$ и $H^T=H\Rightarrow De_i=\sigma^2(1-H_{ii})$.

Чтобы унифицировать все остатки перейдём к нормир. остаткам: $\frac{e_i}{\sqrt{De_i}}=\frac{e_i}{\sigma\sqrt{1-h_{ii}}}$

Но σ^2 неизвестно $\Rightarrow$ заменим σ^2 на её оц. $\frac{1}{n-k}\|y-X\hat\theta\|^2=\frac{RSS}{n-k}$

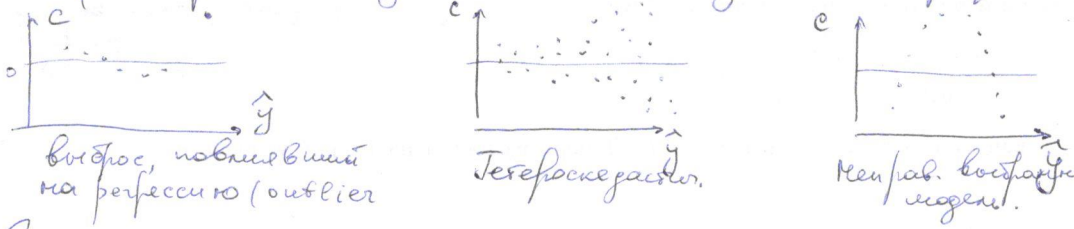
получаем $d_i=\frac{e_i}{\sqrt{\frac{RSS}{n-k}}\sqrt{1-h_{ii}}}$ – студентизированные остатки.

Замечание $\sum_{i=1}^n h_{ii}=k$, т.к. $\operatorname{tr}H=\operatorname{tr}(X(X^TX)^{-1}X^T)=\operatorname{tr}(X^TX(X^TX)^{-1})=$

$=\operatorname{tr}(I_k)=k$, поэтому если $k\ll n$, то h_{ii} можно считать ≈ 0 и

рассм. $c_i=\frac{e_i}{\sqrt{\frac{RSS}{n-k}}}$ (стандартизир. остатки). Хотя стандартиз. остатки зав. зависимыми, но при больших n их поведение похоже на незав. $N(0,1)$, (если предпол. $\varepsilon\sim N(0,\sigma^2 I_n)$ верно).

Можно при анализе остатков пользоваться графическим анализом:



выброс, повлиявший на регрессию (outlier

Гетероскедастич.

Неправ. выбранная модель

Итак, проверка гетероскедастичности: проверим гипотезу H_0: $\varepsilon\sim N(0,\sigma^2 I_n)$ (на самом деле, распред. ε может отходить от нормального)

Критерий Голдфелда-Квандта

Упорядочим наблюд. по предполагаемому возрастанию дисперсий случ. ошибок, отбросим c центральных наблюдений. Строим 2 регрессионные модели: по первым $\frac{n-c}{2}$ наблюдениям и по $\frac{n-c}{2}$ последним, вычисляем по ним RSS_1 и RSS_2 – остат. суммы квадратов по каждой из моделей. Если H_0 верна, то $F=\frac{RSS_2}{RSS_1}\sim F_{\frac{n-c}{2}-k,\ \frac{n-c}{2}-k}$,

отвергаем, если $F>u_{1-\alpha}$, т.к. $RSS_2>RSS_1$, по выбору модели

МСПС | Регрессионный анализ

Критерий Уайта Расам. вспомогат. модель $e_i^2 = \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij} + \sum_{j=1}^{k} \beta_j x_{ij}^2 + v_i$,
где e_i – обычные остатки модели и гип. $H_0: \alpha_j = \beta_j = 0 \ \forall j$.
При верной H_0: $nR^2 \sim \chi^2_{2k}$ (при большом кол-ве наблюд.),
где R^2 – коэфф. детерминации данной модели.
Есть ещё крит. **Бреша-Пагана**, где β_j сразу $= 0$, и $nR^2 \sim \chi^2_k$ в этой модели.

3) $\varepsilon \sim N$, проверка остатков на нормальность – см. крит. проверки нормальности, чаще всего (ну-а зависимости остатков) люди удовлетв. хорошим видом QQ-plot и критерием Харка-Бера. Если остатки не нормальны, то можно применить преобраз. Бокса-Кокса

4) **Замечание** Удаление выбросов с помощью расст. Кука: если $D_i := \frac{e_i^2}{RSS(k+1)} \frac{H_{ii}}{(1-H_{ii})^2}$ больше порога (напр., 1, $3\bar{D}$, $\frac{4}{n}$), то удаляем это наблюдение.

4) Независимость – проверяется корреляционным анализом.
Если вдруг кажется, что остатки образуют временной ряд, то можно проверить на автокоррелированность (см. крит. Дарбина-Уотсона, а также критерии ну авторегрессий, это будет далее в курсе)

Замечание. Если от гетероскедастичности избавиться не удаётся (напр., с помощью взвешенного МНК – т.е. одни наблюдения на что-то увелич., другие – уменьшаем (имеются в виду отрезки)), то можно использовать для оценки значимости признаков критерии, основанные на устойчивой оц. дисперсии: White's heteroscedasticity-consistent estimator (HCE): $\hat{\Omega} = n(X^TX)^{-1}(X^T \operatorname{diag}(e_i^2) X)(X^TX)^{-1}$.
Если вместо e_i^2 поставить обычную оценку дисперсии, то получится стандарт. оценка дисперсии θ.
Модификации: вместо e_i^2 ставить $e_i^2 \cdot \frac{n}{n-k}$, $\frac{e_i^2}{1-H_{ii}}$, $\frac{e_i^2}{(1-H_{ii})^2}$ (поправки Маккиннона Уайта)

5) X^TX – плохо обратима, это случается, когда столбцы X близки к линейной зависимости, т.е. есть сильная совместная корреляция между признаками. Такая ситуация называется мультиколлинеарностью.

а) В чём, собственно, проблема? Помним, что $\hat{\theta} \sim N(\theta, \sigma^2(X^TX)^{-1})$.
Если X^TX плохо обратима, то $\det((X^TX)^{-1})$ очень большой, т.е. оценка $\hat{\theta}$ будет крайне неустойчива. Критерием плохой обуслов. матрицы X^TX явл. высокое знач. отношения $\frac{\lambda_{max}}{\lambda_{min}}$ максимал. и минимал. собств. значений матрицы X^TX. Если $\frac{\lambda_{max}}{\lambda_{min}}$ от 10 до 100, то уже стоит задуматься, а если > 1000, то принимать меры.

б) Бывает так, что вектор-признак X_j имеет небольшой разброс значений и, как результат, будет коррелировать с $X_0 = (1 \dots 1)^T$.
$\Rightarrow$ следует перейти к нормированным данным $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j^2}}$, ну и центрировать отклик $y' = y - \bar{y}$.

Добавление к п. 4 – преобразование Бокса-Кокса.

Хотим, чтобы наши данные (например, отклик) были похожи на нормальные. Пусть все комп. вектора Y – положит. Иначе $Y \to Y - \min_i Y_i$

Делаем замену переменных: $V = \begin{cases} \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases}$ $\dot{y} = (y_1 \dots y_n)^{1/n}$ – среднее геометр. отклика.

Находить опт. λ можно с помощью
а) критериев проверки нормальности
б) метода макс. пр-я
в) метода Бокса-Кокса: пусть $\lambda \in (-a, a)$, строим для каждого λ регрессию V по признакам X, получаем $RSS(\lambda)$, строим по нему график $R^2(\lambda)$, выбираем таким образом оптимал. λ.
См. функцию BoxCoxTrans() пакета caret в R.

~~Вернемся, против ослика~~ Возвращаемся к проблеме мультиколлинеар.

в) Удаление признаков с помощью VIF.

$VIF(\theta_j) = \frac{1}{1 - R_j^2}$, где R_j^2 – коэфф. детерминации модели

$$X_j = c_0 + c_1 X_1 + \dots c_{j-1} X_{j-1} + c_{j+1} X_{j+1} + \dots c_k X_k.$$

Логично, что если R_j^2 высок, то X_j хорошо объясняется ~~оставшимися~~ остальными признаками, и его надо удалить. Обычно удаляют признак, если $VIF_j > 4$.

г) Регуляризация
- гребневая регрессия (ridge-regression): $\hat{\theta}_{ridge} = \arg\min_\theta \left(\|Y - X\theta\|^2 + \lambda \|\theta\|^2 \right)$,
$\hat{\theta}_{ridge} = (X^T X + \lambda I_n)^{-1} X^T Y$, т.е. если $\lambda \to 0$, то это будет обычная о.н.к.
Проблема в том, что $E\hat{\theta}_{ridge} = (I_k - \lambda W)\theta$, где $W = (X^T X + \lambda I_n)^{-1}$, т.е. оценка смещённая.
Но зато $Var \hat{\theta}_{ridge} = \sigma^2 W X^T X W$, и с ростом λ она уменьшается, но смещение увеличивается.
Теорема $\exists \lambda > 0$ такое, что $E\|\hat{\theta}_{ridge} - \theta\|^2 < E\|\hat{\theta}_{MNK} - \theta\|^2$, но вот на практике способа находить такое λ не придумали пока.
- лассо $\hat{\theta}_{LASSO} = \arg\min_\theta \left(\|Y - X\theta\|^2 + \lambda \sum_{j=1}^{k} |\theta_j| \right)$, и ещё куча других регуляризаций, см. Википедию или Mastie, Tibshirani

Регуляризации тесно связаны с байесовской постановкой ~~пред~~ задачи регрессии, где в кач. априорного распределения параметров модели θ берётся гауссовское (в ridge-regression) и распред. Лапласа (LASSO).

б) Ещё мы не обсудили проблему выбросов.
Выбросы очень сильно влияют на оценку м.н.к., см. квартет Энскомба, напр.
Ранее для поиска оценки $\hat{\theta}$ мы пользовались методом наим. квадратов:
$\hat{\theta} = \arg\min_\theta \rho(Y - X\theta)$, где $\rho(\vec{v}) = \|\vec{v}\|^2$
Так ведь можно взять другие ф-ции ρ! Требуем от них, чтобы они были 1) симметрич., 2) неотриц.; 3) монотонно ~~не~~ неубывающими.

Регрессионный анализ | МСПС |

(Обсуждаем робастные регрессионные модели). ~~Пусть~~.
Если ρ - дифф., то $\psi = \rho'$ называют ф-цией влияния.
С пом. неё ~~робастная регрессия~~ поиск оценки θ осуществл. из такого ур-ния: $\sum_{i=1}^{n} \psi(y_i - X_i^T\theta) \cdot X_i = 0$

Примеры ф-ций ρ:

Тип	L_2	L_1	Huber	L_p	Cauchy	German-Mcclure	Металькин
$\rho(x)$	$x^2/2$	$\lvert x\rvert$	$\frac{x^2}{2}\cdot I(\lvert x\rvert\le k) + k(\lvert x\rvert - \frac{k}{2})\cdot I(\lvert x\rvert > k)$	$\lvert x\rvert^p/p$	$\frac{c^2}{2}\log\left(1+\left(\frac{x}{c}\right)^2\right)$	$\frac{x^2/2}{1+x^2}$	$1-\exp(-\frac{x^2}{2b})$

Ну, требования: огранич. ф-ции влияния (для устойчивости относ. выбросов) и единств. точка минимума.
L_2 - ф. влияния не ограничена, L_1, Коши, Маккуэр - необход. единств. решение.
Для Хьюбера обычно выбирают $k = 1.345$, кроме того, оценки методом Хьюбера ~~похожи~~ близки к асимпт. эффективным.
Ещё метод - метод скользящей медианы. Пусть m - определённое число выбросов. Тогда заменяем y_i на $\tilde{y}_i = med\{y_{i-m}, y_{i-m+1}, \dots y_{i+m}\}$.
Есть ещё метод, который, увы, плохо работает в ситуации, когда признаки зависимы - метод Тейла:
есть регресс. модель $y_i = \theta_0 + \sum_{j=1}^{k}\theta_j x_{ij} + \varepsilon_i$, тогда берём
$\hat{\theta}_j = med\left\{\frac{y_i - y_k}{x_{ij} - x_{kj}},\ i \ne k \in \overline{1..n}\right\}$, $\hat{\theta}_0 = med\left\{y_i - \sum_{j=1}^{k}\hat{\theta}_j x_{ij}\right\}$. Зато устойчив к выбросам.