Quadratsummenzerlegung im einfachen Linearen Modell

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST(otal)} = \underbrace{\sum_{i=1}^{n}(\overbrace{y_i-\hat{y}}^{\hat{\varepsilon}_i})^2}_{SSE(rror)\ RSS} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSM(odel)}$$

Gesamtstreuung von y = Residuen-Streuung + Modell-Streuung
erklärt durch Prädiktor-Variable X

Beweis: $\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}((y_i-\hat{y}_i)+(\hat{y}_i-\bar{y}))^2$

$$= \sum_{i=1}^{n}(y_i-\hat{y})^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \underbrace{2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})}_{=0}$$

$\sum \hat{\varepsilon}_i = 0,\ \sum \hat{\varepsilon}_i x_i = 0$

↑ Normalgleichungen

$$= \sum_{i=1}^{n} \hat{\varepsilon}_i(\hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y}) = \hat{\beta}_0\sum_{i=1}^{n}\hat{\varepsilon}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{\varepsilon}_i x_i - \bar{y}\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$$

Freiheitsgrade der Quadratsummen ≙ Anzahl der frei bestimmbaren Summanden der obigen Quadratsummen

SST: $\sum(y_i-\bar{y})^2$ **$df=n-1$** Wir verlieren einen Freiheitsgrad durch Schätzung von $\bar{y}$ (genauer gesamt muss $\bar{y}$ so gesch. werden, dass $\sum(y_i-\bar{y})=0$) $n-1=(n-2)+1$

SSE: $\sum_{i=1}^{n}(y_i-\hat{y})^2$ Diese minimieren wir bei KQ-Schätzung von $\hat{\beta}_0, \hat{\beta}_1$ => **$df=n-2$**, da wir zwei Parameter daraus schätzen (zwei Restriktion durch Normalengleichungen)

SSM: $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum(\hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y})^2 = \sum(\hat{\beta}_1(x_i-\bar{x}))^2$ **$df=1$**, da es nur einen Regressor gibt 1 FG verbraucht für $\bar{y}$, (n-2) FG verbraucht für $\hat{\beta}$

$\hat{\beta}_0 = \bar{y}-\hat{\beta}_1\bar{x}$ durch Wählen von einem y-Wert liegt $\hat{\beta}_1$ fest

Anpassungsgüte "fit"

Bestimmtheitsmaß R^2 $R^2 = \frac{SSM}{SST} = 1-\frac{SSE}{SST}$ Durch X erklärter Anteil der Varianz von y

(bzw. schlechtestens bleibt es konstant)

Auch R^2 ↑ mit mehr Regressoren im Modell

$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \dots \frac{\sum(\hat{\beta}_1(x_i-\bar{x}))^2}{\sum(y_i-\bar{y})^2}$ R^2 hängt von wahrem β_1 ab, für das $\hat{\beta}_1$ ein erwartungstreuer Schätzer ist

Je mehr Streuung der x_i, umso höher R^2. SSE wird immer gleich bleiben und so steigen SSM u. R^2
es ist also Vorteilhaft Regressoren mit großer Varianz zu wählen (zudem dann $Var(\hat{\beta}_1) = \frac{\sigma^2}{n s_x^2}$ ↓)

$$R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i-\bar{y})^2}{n s_y^2} = \frac{\sum(\hat{\beta}_1(x_i-\bar{x}))^2}{n s_y^2} = \frac{\hat{\beta}_1^2\sum(x_i-\bar{x})^2}{n s_y^2} = \frac{\hat{\beta}_1^2 \cdot n s_x^2}{n s_y^2} \ (\hat{\beta}_1 = s_{xy}/s_x^2) = \frac{n(s_{xy})^2}{n s_y^2 s_x^2} = \frac{(s_{xy})^2}{s_y^2 s_x^2} = \left(\frac{s_{xy}}{s_y s_x}\right)^2 = (r_{xy})^2$$

nach Schätzung können wir $\hat{\beta}_1$ als feste Zahl betrachten, also aus Reihe ziehen ✓

gilt so nur im einfachen Modell

Für GLMs und ML-Modell gibt es diese Zerlegung gar nicht bzw. schwer allgemein zu definieren

Quadratsummenzerlegung im multiplen Linearen Modell

Im Prinzip alternative Varianzberechnung ($\cdot 1/n$ ergäbe Var( ) statt Quadratsumme)

$$\underbrace{(y-\bar{y})^T(y-\bar{y})}_{SST} = \underbrace{\overbrace{(y-\hat{y})^T(y-\hat{y})}^{\varepsilon^T\varepsilon}}_{SSE} + \underbrace{(\hat{y}-\bar{y})^T(\hat{y}-\bar{y})}_{SSM}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y})^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

Freiheitsgrade der Quadratsummen

SST: $n-1$

SSE: $n-(p+1)$

SSM: $(p+1)-1$

Geometrische Intuition: Quadratsummen als Längen im Raum im Sinne der euklidischen 2-Norm, Zerlegung folgt aus Satz d. Pythagoras da $(y-\bar{y})\perp \hat{\varepsilon}=(y-\hat{y})$

$SST = \|y-\bar{y}\cdot e\|^2$ misst Gesamtabweichung von y zu seinem MW, völlig unabh. vom Regressionsmodell (braucht nur Beob. und $\bar{y}$) im Modell mit Intercept

$SSE = \|y-\hat{y}\|^2$ ist Länge der orthogonalen Residualkomponente $\hat{\varepsilon}$, also das was die Regression nicht "erklären kann (gilt auch generell)

$SSM = \|\hat{y}-\bar{y}\cdot e\|^2$ misst Länge der orth. Projektion von $(y-\bar{y}\cdot e)$ auf col([1, X])

ohne intercept nur: $\|y\|^2 = \|y-\hat{y}\|^2 + \|\hat{y}\|^2$

nutzen das nicht im Intercept-Modell, weil wir dann nicht Variation von y zerlegen

Entsprechend R^2: Erklärt ⊥ Projektion auf col([1, X]) etwas an y (zentriert) oder ist es rein zufällig?

Hohes R^2 bedeutet: Gute Erklärungskraft des Modells (bezüglich Varianz (Summe) von y, SST) also im Vergleich mit Vorhersage nur mit Mittelwert von y!

- R^2 hoch => Modell gut, R^2 niedrig => Modell nutzlos ~~stimmt nicht allgemein~~ R^2 kann sehr niedrig sein, wenn linearer Zhang sinnig aber sehr schwach
  bzw. niedriges R^2 kann wertvolle Info sein, dass linearer Effekt schwächer als gedacht
- R^2 kann "künstlich" vergrößert werden, indem man ganz viele Regressoren ins Modell aufnimmt. Aber ist das die zusätzliche Komplexität wert?!

  R^2↑ manipulieren: Overfitting Risiko (grobe Idee hinter R^2_{adj} ist Penalisierung von mehr Kovariaten in exzessiver Weise)

Merke: $R^2<0$ => Modell performt schlechter als einfache Vorhersage mit Mittelwert von y

bzw. $SSM=0 \to \frac{SSM}{SST}=0$

$R_2>0$ geht nur, wenn $SSE<SST$

dann $\hat{y}=\bar{y}$ und $SST=SSE$ => $R^2 = 1-\frac{SSE}{SST} = 1-1=0$

Wie kann R^2 negativ werden? i) Modell ohne Intercept (da ist R^2 auch kein sinnvolles Maß)

, garantiert mindestens besser als $\hat{y}=\bar{y}$ denn das wäre zulässiger Schätzer von SSE-Minimierung, also $R^2_{min}\geq 0$ (aber gilt nur für Trainings-Datensatz!)

ii) Modell mit Intercept wird evaluiert auf (fremden) Testdaten, auf denen es sehr schlecht abschneidet evtl. auch aufgrund von Overfitting durch R^2-Optimierung

$R^2<0$ dann möglich, wenn lineare Beziehung im Test-Set sehr schwach ist oder sich im Vgl. zu Trainingsmodell sogar umkehrt, sodass $SSE_{test} > SST_{test}$

kann schnell passieren bei ML bzw. allgemein Extrapolation bei Trend-Vorhersagen, wo sich Trend im Testdatenset umkehrt

=> Neben Anpassungsgüte R^2 gibt es noch **Prognosegüte RMSE** zur Modell-Evaluierung und Vergleich von (Prognose-)Qualität verschiedener Modelle

Maßzahl für durchschnittl. Abweichung von Prädiktion

"Root Mean Squared Error" **RMSE** $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$ sollte immer auf Basis von Testdaten bestimmt werden, die nicht in Modellschätzung einbezogen wurden

Wurzel von mittlerem quadr. Fehler