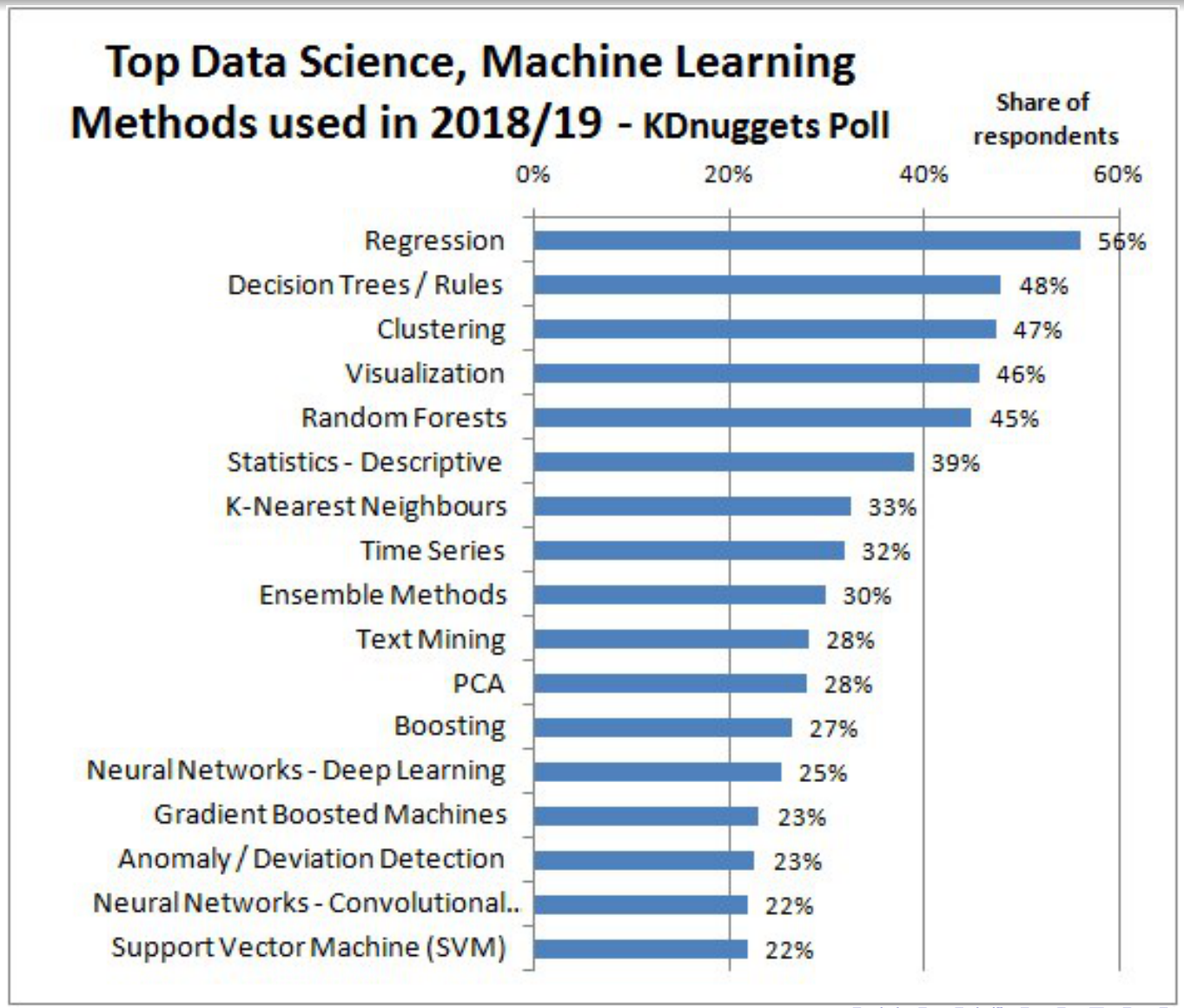
Analiza danych

prof. UAM dr hab. Tomasz Górecki

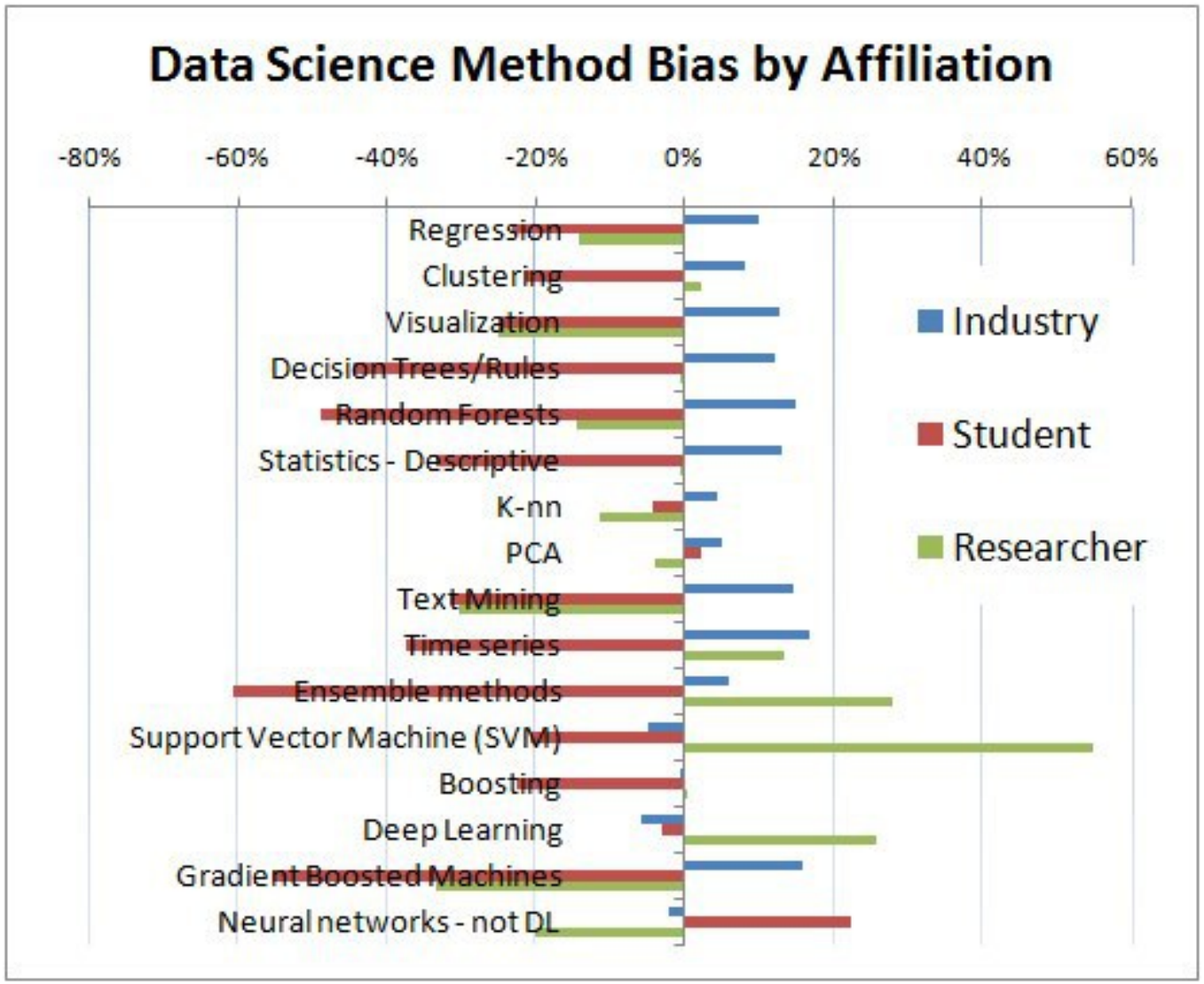
tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Najpopularniejsze metody ML (2019)...



... i kto ich używa



Idea

Termin regresja oznacza metodę pozwalającą na zbadanie związku pomiędzy zmiennymi i wykorzystanie tej wiedzy do przewidywania nieznanych wartości jednych wielkości na podstawie innych. W praktyce poszukuje się związku między domniemaną jedną (lub więcej) zmienną objaśniającą (niezależną), a zmienną objaśnianą (zależną) Y. Związek ten może być dalej wykorzystywany do prognozowania wartości Y w zależności od X. Jeżeli badać będziemy zależność zmiennej Y od wartości innej zmiennej, to wartości zmiennej objaśniającej będziemy oznaczać przez x i traktować jako wartości deterministyczne zmiennej X, które wybieramy w celu obserwacji zmiennej losowej Y. Jak widać zmienne X oraz Y traktowane są odmiennie. Mianowicie zmienna X uważana jest za w pełni kontrolowaną przez eksperymentatora, a co za tym idzie pozbawiona jest ona elementu losowości (de facto traktowana jest jako liczba).

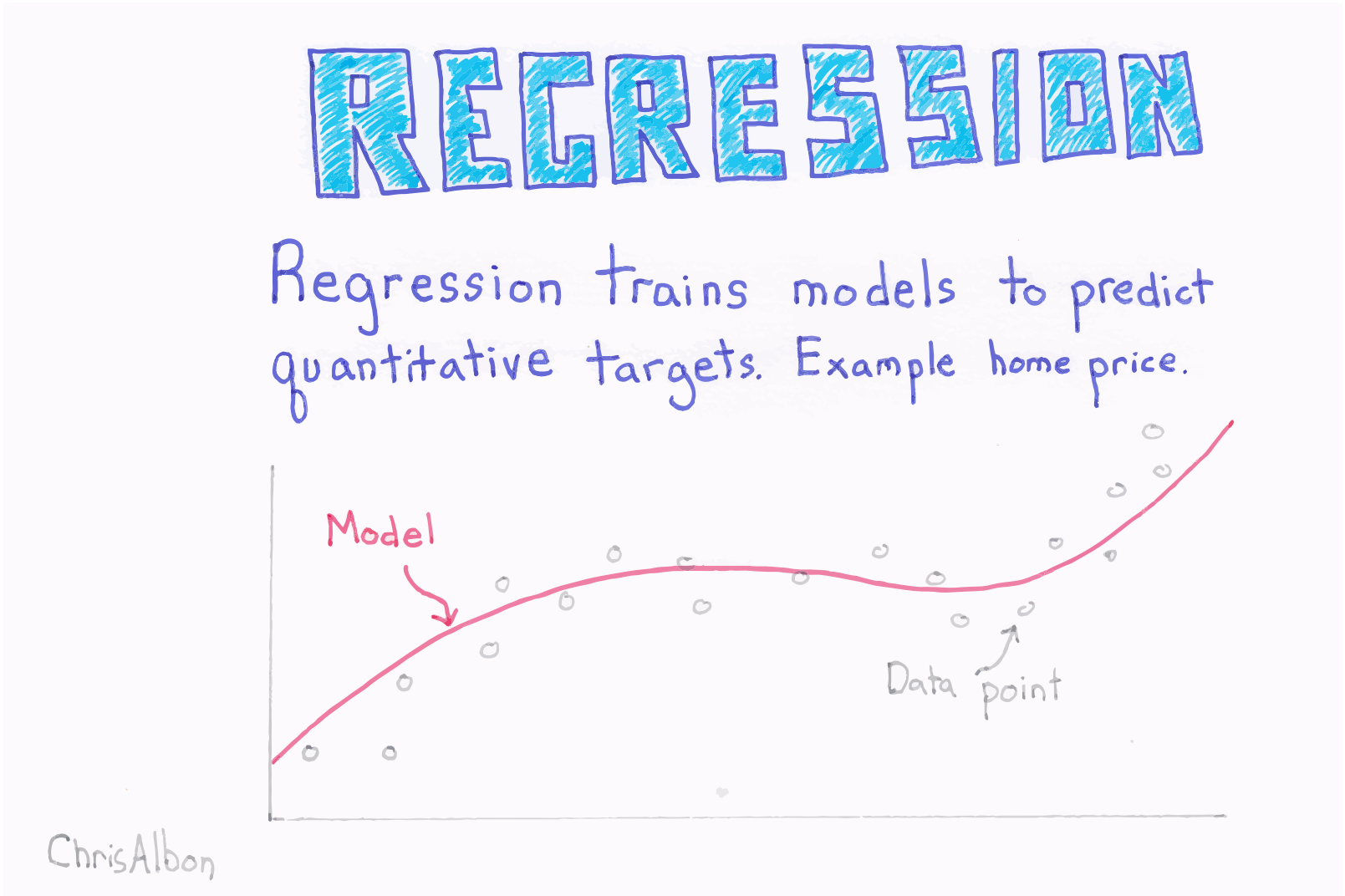
Idea

Chcemy zatem odpowiedzieć na pytanie jak zmienia się wartość oczekiwana zmiennej Y w zależności od wartości x zmiennej X, czyli:

$$\mathbb{E}(Y|X) = g(X),$$

gdzie $g(X)$ jest funkcją regresji opisującą poszukiwany związek. Zwyczajowo zakłada się dodatkowo, że $\text{Var}(Y|X)$ jest dla wszystkich wartości $X = x$ stała i równa σ^2 (jednorodność wariancji). Z matematycznego punktu widzenia regresją nazywana jest każda metoda, która umożliwia oszacowanie tego równania.

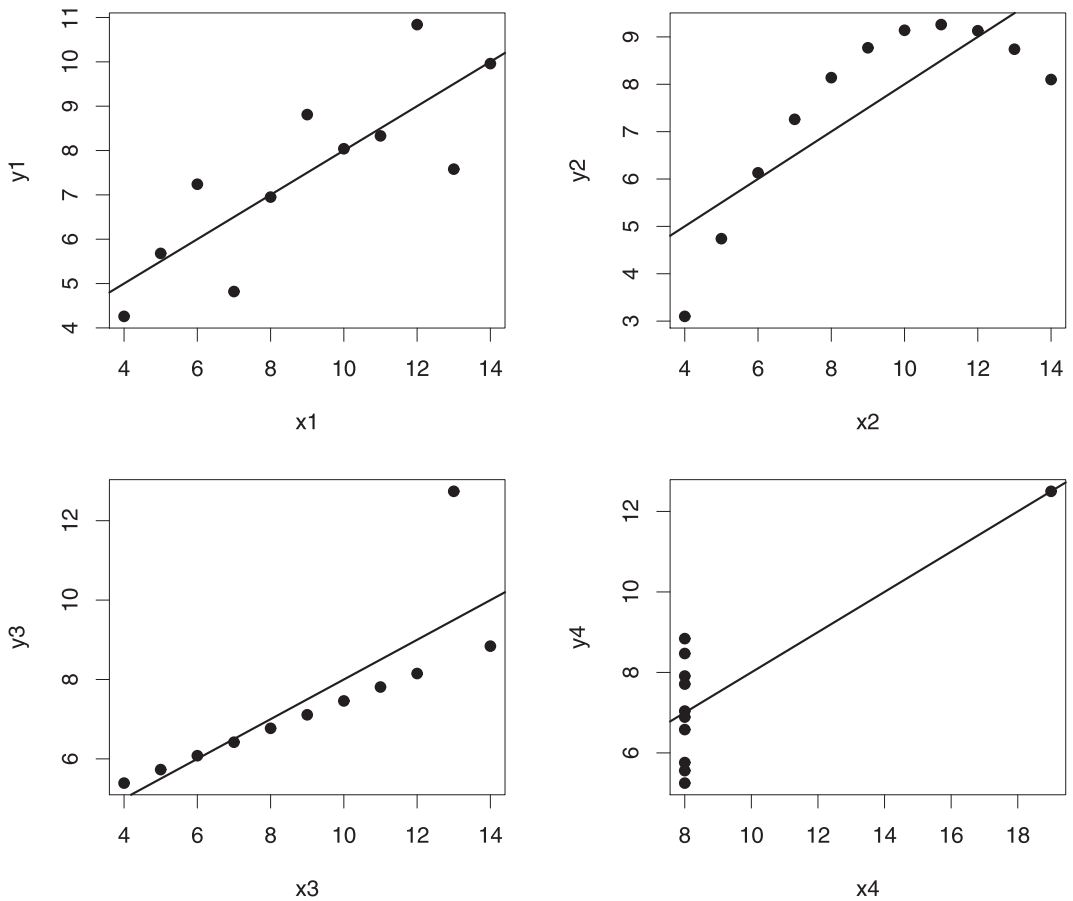
Idea


REGRESSION
Regression trains models to predict quantitative targets. Example home price.
Model
Data point
ChrisAlbon

Diagramy korelacyjne

W celu wstępnej oceny zależności najczęściej konstruuje się diagramy korelacyjne. Ich wagę doskonale uwypuklił `Anscombe` (1973), który skonstruował 4 zbiory danych, mające identyczne podstawowe charakterystyki, ale ich diagramy korelacyjne diametralnie się różniły. Średnia dla każdej zmiennej x_i wynosiła 9, zmiennej $y_i = 10$; wariancja dla $x_i = 7{,}5$, dla $y_i = 2{,}75$; współczynnik korelacji liniowej wynosił 0,816 dla każdego zbioru oraz prosta regresji miała postać $y = 3 + 0{,}5x$.

Diagramy korelacyjne



Francis John Anscombe
(1918-2001)

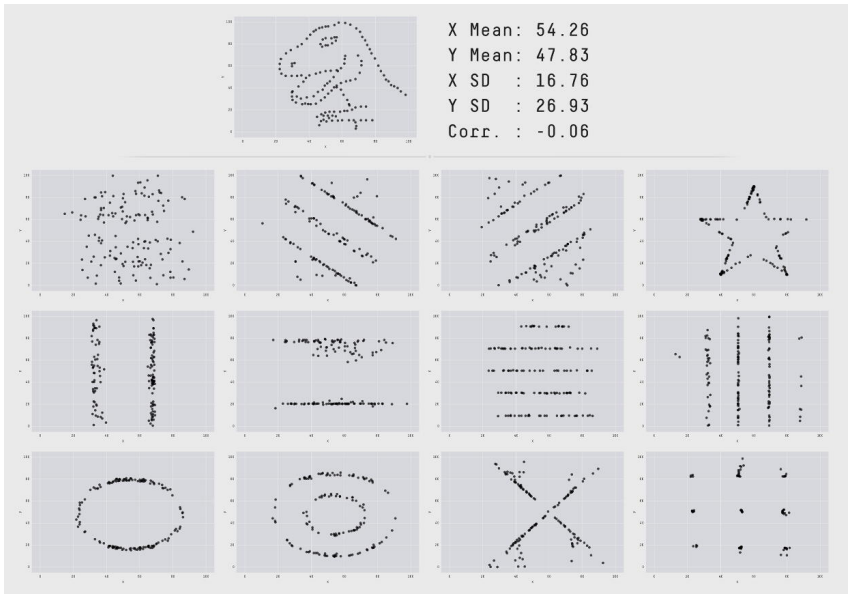
Anscombe, F.J. (1973). Graphs in statistical analysis. *American Statistician* 27(1):17–21.

Diagramy korelacyjne

Różnią się one w sposób bardzo wyraźny. Pierwszy wykres (górny lewy róg) sugeruje, że dane mają rozkład normalny i prosta regresji oraz współczynnik korelacji są poprawne. Drugi wykres (górny prawy róg) pokazuje nieliniowy charakter zależności, a zatem i brak uzasadnienia dla prostej regresji oraz współczynnika korelacji. Wykres trzeci (dolny lewy róg) wskazuje na wagę obserwacji odstającej, która jest powodem zaniżenia współczynnika korelacji. Ostatni wykres (dolny prawy róg) pokazuje inne zjawisko, mianowicie tzw. obserwacją wpływową, która tutaj spowodowała, że współczynnik korelacji jest wysoki, mimo, że taki być nie powinien.

Diagramy korelacyjne

The Datasaurus Dozen. Chociaż różnią się wyglądem, każdy zbiór danych ma te same statystyki zbiorcze (średnia, odchylenie standardowe i współczynnik korelacji `Pearsona`) z dokładnością do dwóch miejsc po przecinku.



Matejka, J., Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. *CHI 2017 Conference proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems.*

Regresja liniowa

Zależności regresyjnej poszukuje się w pewnej zadanej z góry klasie funkcji, na ogół klasie funkcji wielomianowych. Np. gdy za $g(X)$ przyjmiemy funkcję liniową, otrzymamy równanie regresji liniowej:

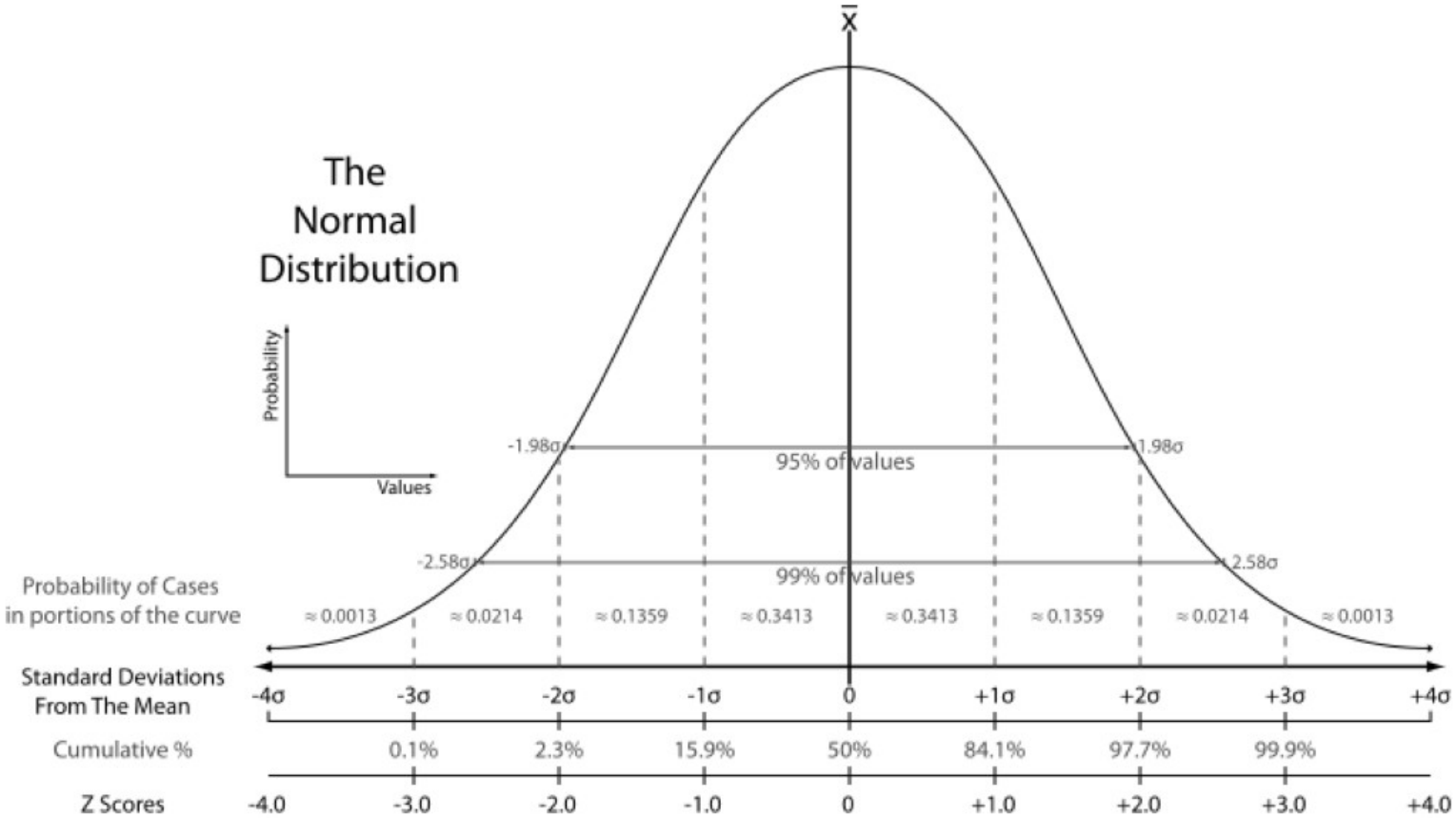
$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X,$$

w którym β_0 oraz β_1 są nieznanymi parametrami. W praktyce wygodniej jest posługiwać się następującym modelem regresji liniowej:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Występujące w równaniu zmienne losowe ε_i nazywane są składnikami losowymi. Zakładamy, że mają one wartość oczekiwaną 0, stałą wariancję równą σ^2 (homoskedastyczność) oraz są nieskorelowane między sobą. Zauważmy, że nie jest wymagane określenie rozkładu składnika losowego (zwykle zakłada się, że jest to rozkład normalny).

Regresja liniowa



Regresja liniowa

W praktyce nie dysponujemy pełną informacją o populacji. Musimy zatem oszacować parametry funkcji regresji na podstawie próby. Odpowiednie oszacowanie ma postać:

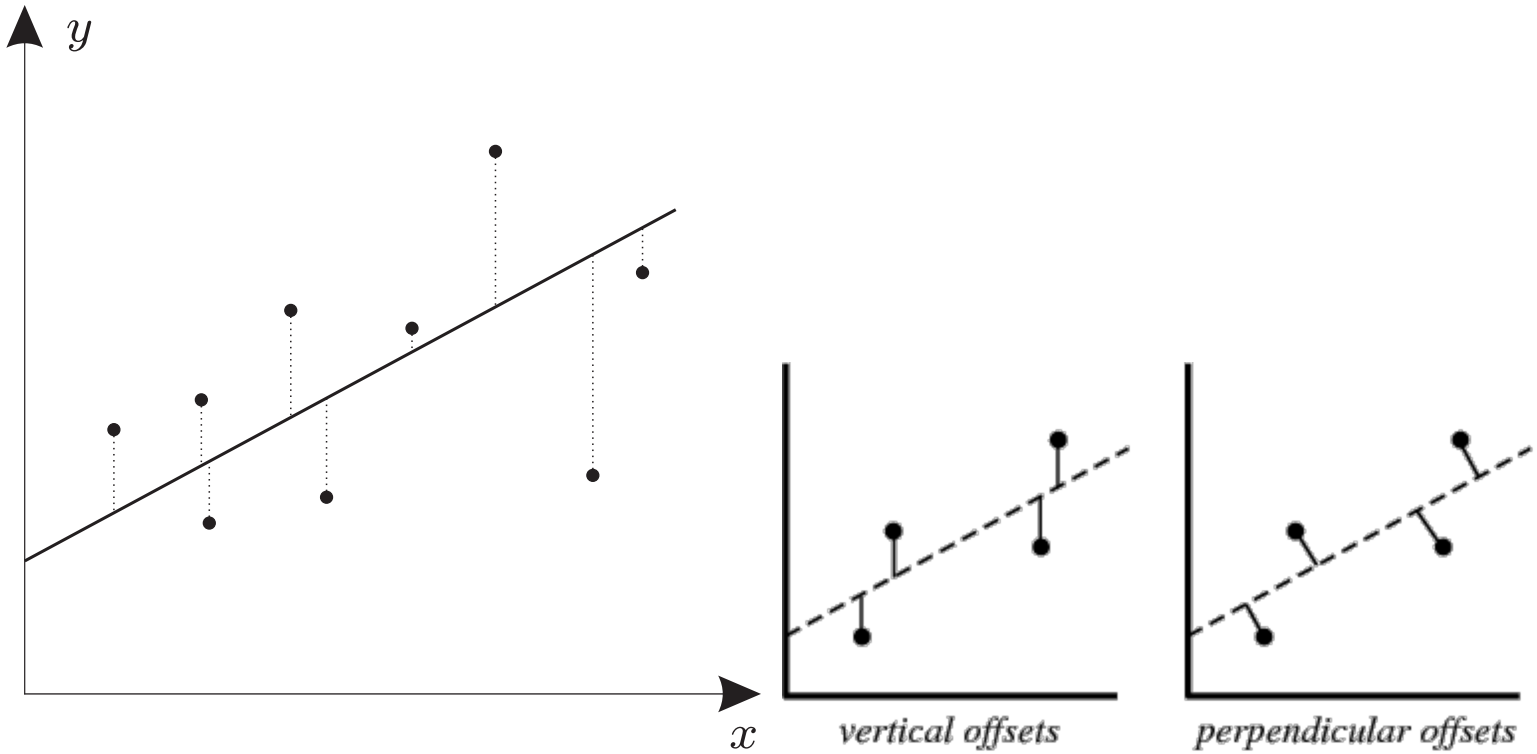
$$\hat{y}_i = b_0 + b_1 x_i.$$

Element

$$e_i = y_i - \hat{y}_i$$

nazywany jest składnikiem resztowym (resztą, residuum). Jak jednak znaleźć taką „dobrze dopasowaną" linię prostą? Punktem wyjścia jest suma kwadratów reszt, opisująca rozbieżność pomiędzy wartościami empirycznymi zmiennej zależnej, a jej wartościami teoretycznymi, obliczonymi na podstawie wybranej funkcji. Oszacowania parametrów dobieramy tak, aby suma kwadratów reszt osiągnęła minimum. Metoda ta nosi nazwę metody najmniejszych kwadratów (MNK) – *ang. Least Squares Method (LS)*.

Regresja liniowa



Regresja liniowa

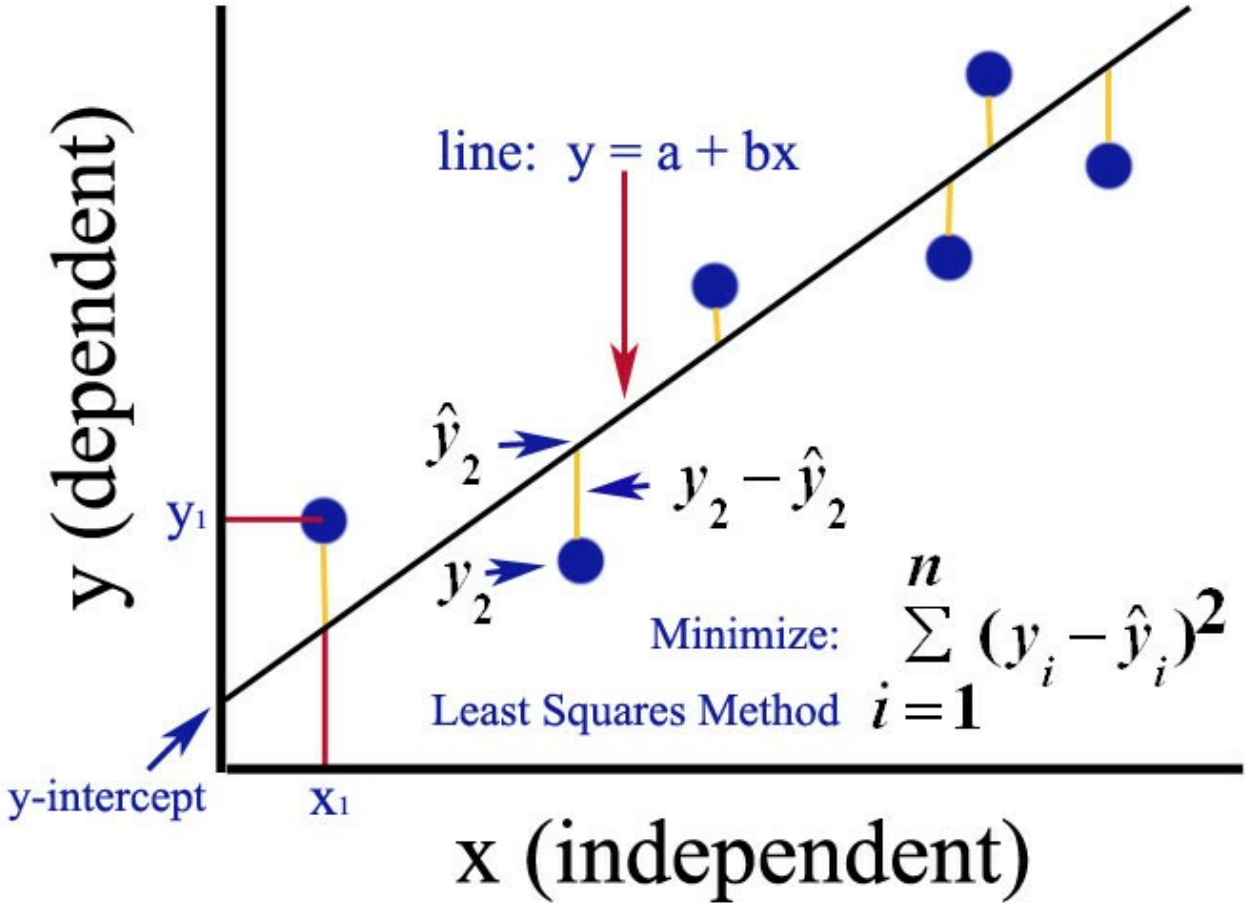
Estymatory parametrów otrzymane za pomocą MNK mają postać:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

Tak otrzymane estymatory są najefektywniejszymi i równocześnie nieobciążonymi estymatorami parametrów regresji liniowej. Współczynnik kierunkowy b_1 (*ang. slope*) nazywamy współczynnikiem regresji liniowej. Odpowiada on na pytanie, jaki jest przeciętny przyrost wartości zmiennej zależnej na jednostkę przyrostu zmiennej niezależnej.

Regresja liniowa



Regresja liniowa

Dokładność oszacowania oceniamy za pomocą współczynnika determinacji (*ang. coefficient of determination*) R^2. Mierzy on jaka część ogólnej zmienności zmiennej zależnej jest wyjaśniona przez regresję liniową (współczynnik determinacji nie ma sensu, jeśli w modelu brak wyrazu wolnego). Dołączenie nowej zmiennej do modelu zawsze zwiększa R^2. Został zaproponowany przez amerykańskiego biologa `Sewalla Wrighta` w 1921 roku.

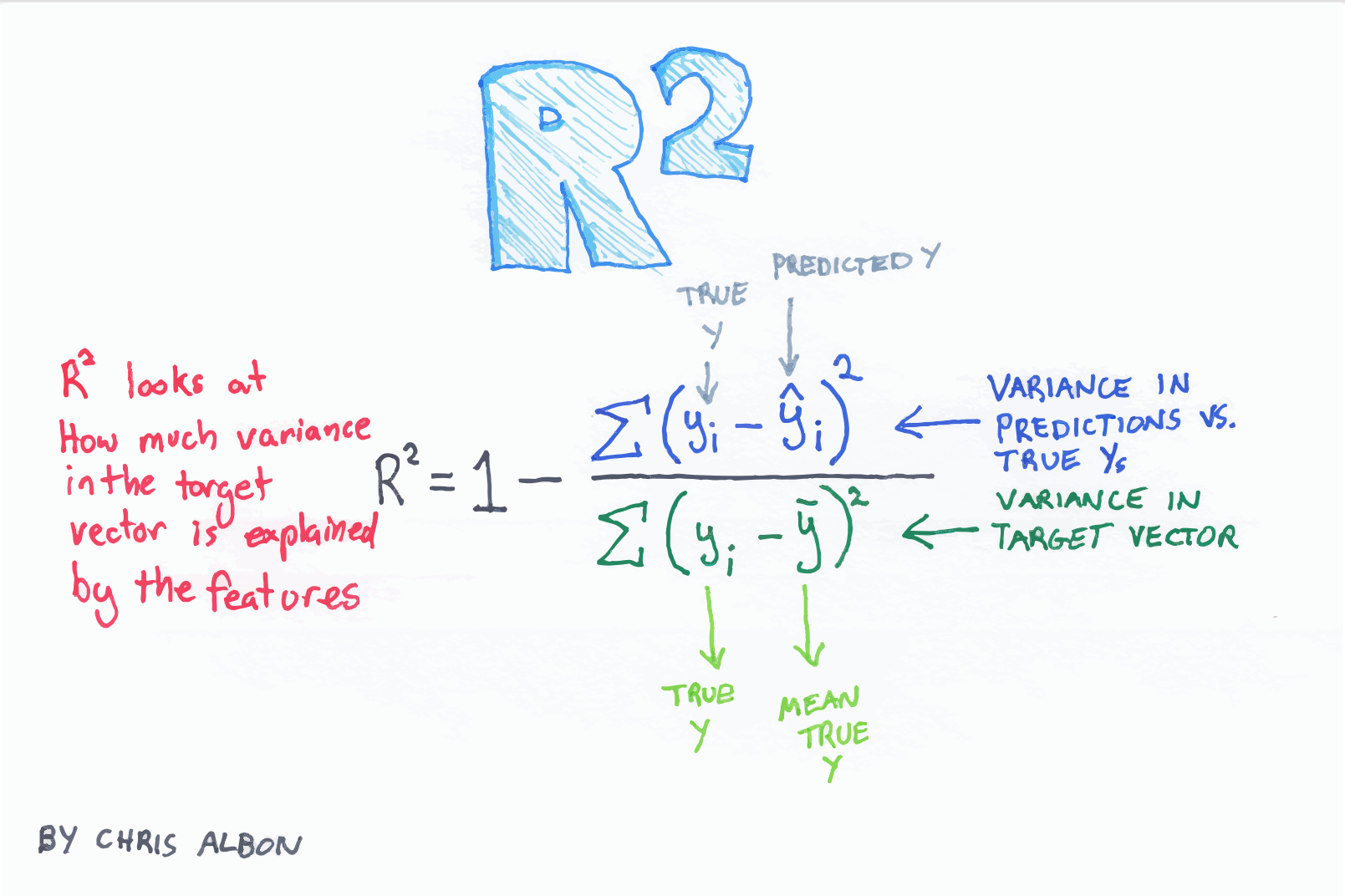
Sewall Wright
(1889-1988)

Wright, S. (1921). Correlation and causation. Journal of Agricultural Research, 20:557-585.

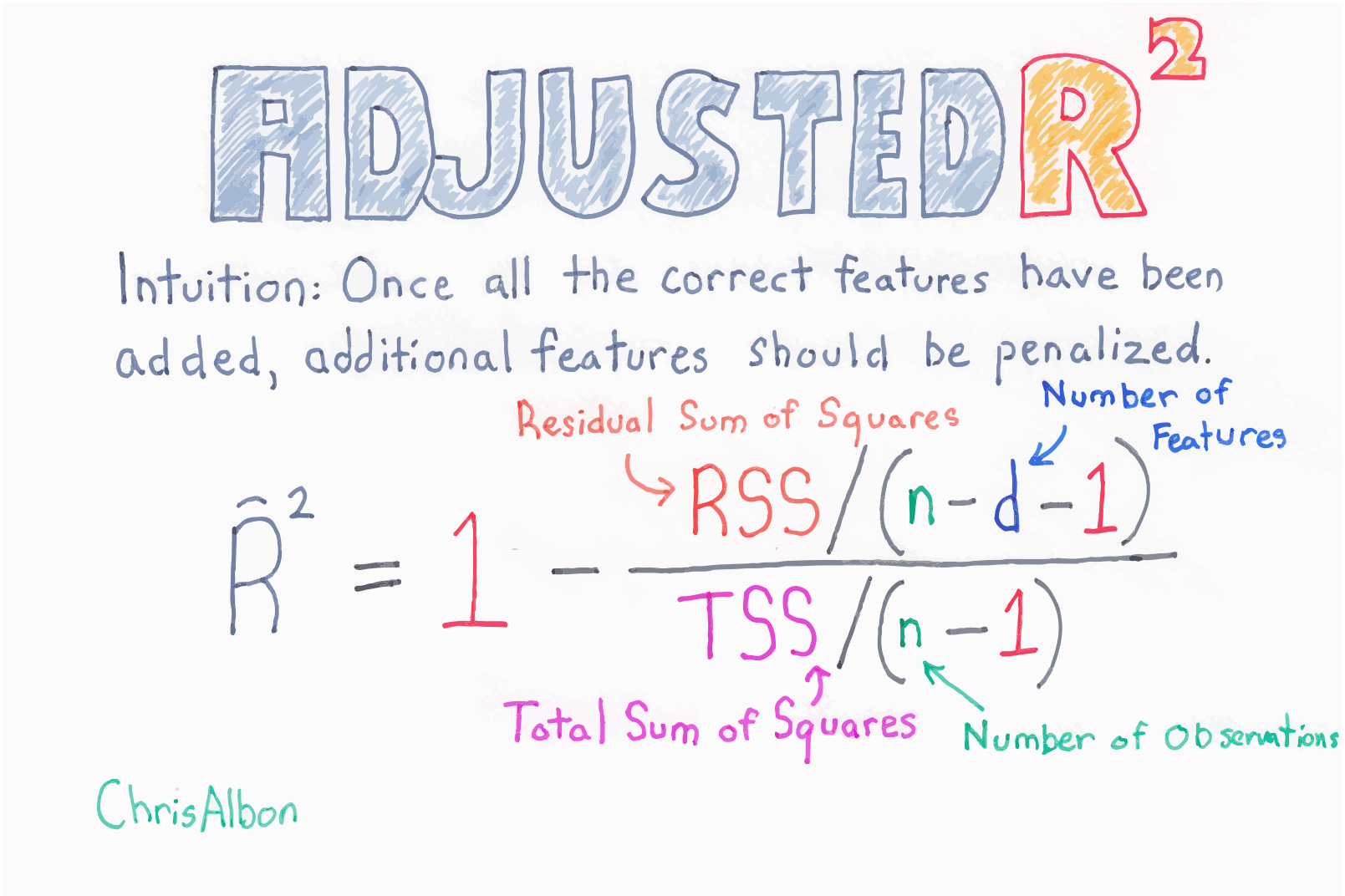
Regresja liniowa

Celem nie jest uzyskanie jak największej wartości tego współczynnika, lecz znalezienie związku między X i Y z rzetelnymi ocenami parametrów. Dlatego w praktyce dla więcej niż jednej cechy objaśniającej używamy raczej tzw. poprawionego R^2 (*adjusted* R^2). Uwzględnia on, że R^2 jest obliczony z próby i jest „za dobry", jeśli uogólniamy nasze wyniki na populację. Poprawiony R^2 jest zawsze mniejszy od R^2. Przyjmuje się, że aby pozytywnie zweryfikować model $R^2 > 60\%$. Należy również pamiętać, że taka ocena jakości modelu jest poprawna wtedy i tylko wtedy gdy gdy spełnione są założenia modelu (model adekwatny).

Regresja liniowa

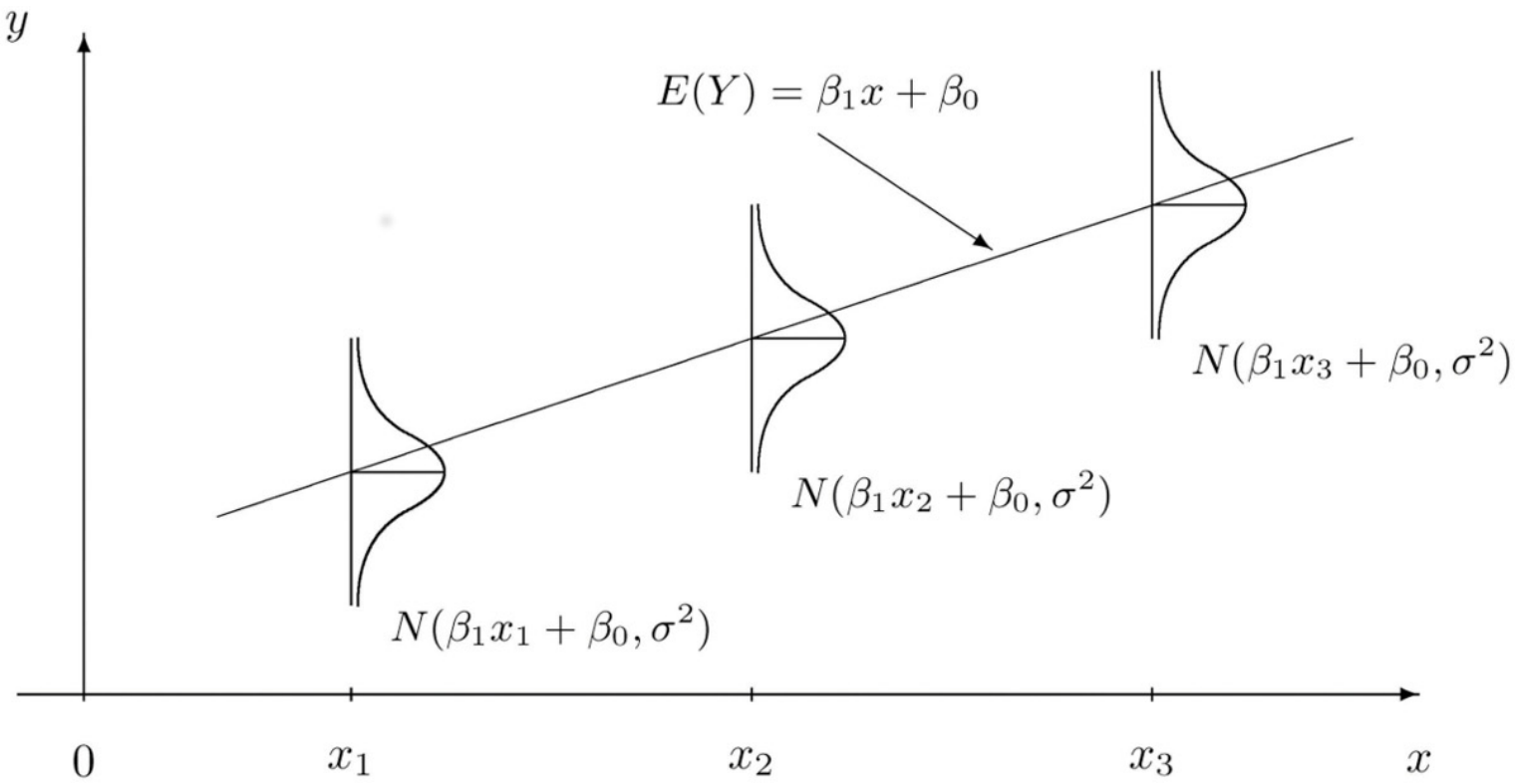


Regresja liniowa



Regresja liniowa

Założenie stałej wariancji przedstawia rysunek poniżej.



Regresja liniowa

Regresja liniowa

Miary oceny jakości modelu – podejście biznesowe

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \text{ (}\textit{ang. Mean Absolute Error}\text{)},$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{|y_i|} \times 100 \text{ (}\textit{ang. Mean Absolute Percentage Error}\text{)},$$

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{|y_i| + |\hat{y}_i|} \times 100 \text{ (}\textit{ang. Symmetric Mean Absolute Percentage Error}\text{)},$$

$$\text{WAPE} = \frac{\sum_{i=1}^{n}|e_i|}{\sum_{i=1}^{n}|y_i|} \times 100 \text{ (}\textit{ang. Weighted Absolute Percentage Error}\text{)},$$

$$\text{WMAPE} = \frac{\sum_{i=1}^{n} w_i|e_i|}{\sum_{i=1}^{n} w_i|y_i|} \times 100 \text{ (ang. }\textit{Weighted Mean Absolute Percentage Error}\text{)},$$

gdzie $w_i > 0$ są wagami.

Regresja liniowa

Miary oceny jakości modelu – podejście statystyczne

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \text{ (ang. Mean Squared Error)},$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \text{ (ang. Root Mean Squared Error)},$$

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \ln^2 \frac{\hat{y}_i + 1}{y_i + 1}} \text{ (ang. Root Mean Squared Logarithmic Error)},$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

$$R_0^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} \text{ (model bez wyrazu wolnego)},$$

$$R^2_{\text{adj.}} = 1 - \frac{(1 - R^2)(n-1)}{n-k}.$$

Regresja liniowa

Mediana vs. średnia – optymalizacja matematyczna

$$\frac{\partial \text{MSE}}{\partial \hat{y}_i} = \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0,$$

$$\sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n} y_i.$$

Zatem aby zoptymalizować MSE (RMSE), model będzie dążył do tego, aby całkowita prognoza była równa całkowitym wartościom obserwowanym. Oznacza to, że optymalizacja taka ma na celu uzyskanie prognozy, która jest średnio poprawna, a zatem nieobciążona.

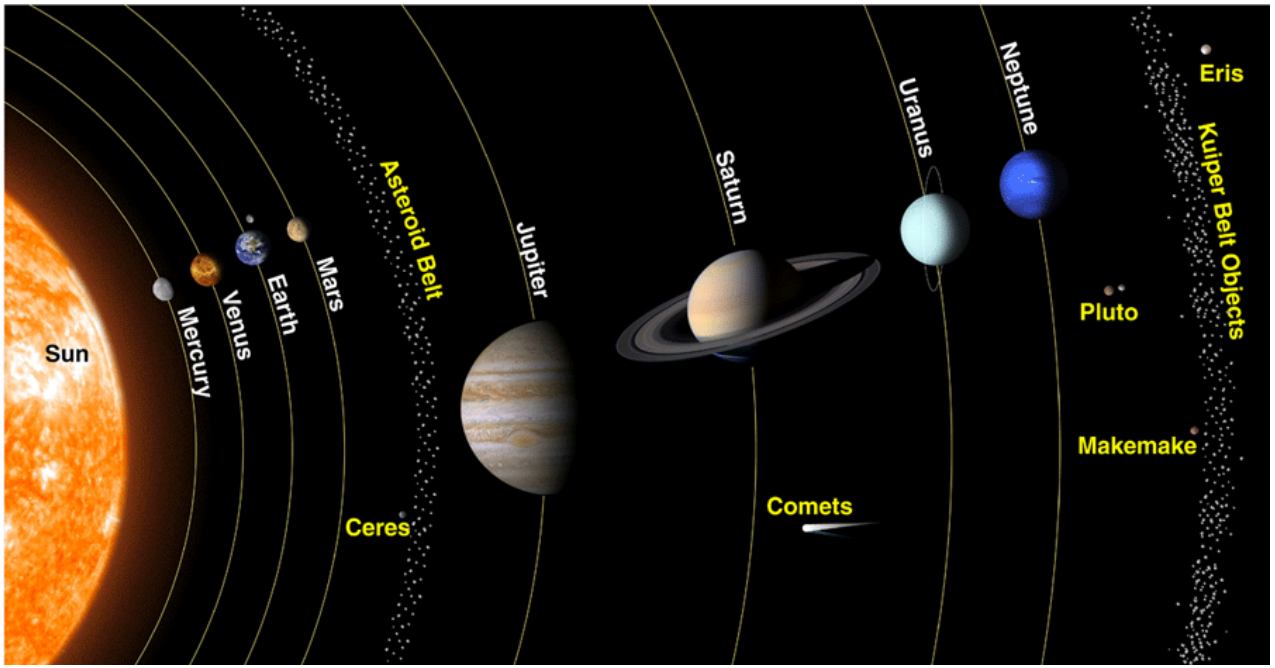
$$\frac{\partial \text{MAE}}{\partial \hat{y}_i} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1 & \text{for } y_i < \hat{y}_i \\ -1 & \text{for } y_i > \hat{y}_i \end{cases}.$$

Innymi słowy, szukamy wartości, która dzieli nasz zbiór danych na dwie równe części. To jest właśnie dokładna definicja mediany.

Regresja liniowa

MNK została wymyślona przez `Gaussa`, który uważał ją jednak za trywialną i był przekonany, że już ktoś ją wcześniej używał. Pierwszą pracę na jej temat opublikował `Legendre`. Obaj używali tej techniki do wyjaśnienia przyszłych orbit komet na podstawie wcześniejszych obserwacji (asteroida `Ceres`). Nie używali jednak pojęcia regresja, które zostało wprowadzone przez `Galtona`. Współczesna analiza regresji jest dziełem `Pearsona` i `Fishera`.

Regresja liniowa



Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15:246–263.

Gauss, K.F. (1809). Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections.

Legendre, A.M. (1805). New Methods for Determination of the Orbits of Comets.

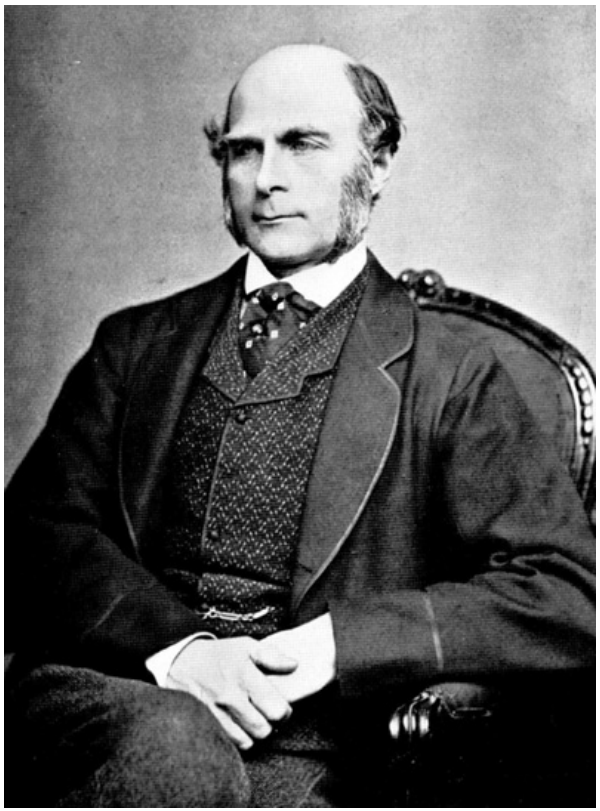
Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia, Philosophical Transactions of the Royal Society of London 187:253–318.

Regresja liniowa

Carl Friedrich Gauss
(1777-1855)

Adrien-Marie Legendre
(1752-1833)

Regresja liniowa



Francis Galton
(1822-1911)

Karl Pearson
(1857-1936)

Regresja liniowa

Sir Ronald Aylmer Fisher
(1890-1962)

Wykresy diagnostyczne – wykres dźwigni

Wykorzystywany do zbadania, czy występują obserwacje odstające. Dla każdego residuum obliczana jest siła dźwigni zwana również wpływem (miara wpływu obserwacji na oceny). W modelu adekwatnym siła dźwigni nie powinna być zbyt duża, gdyż oznacza, to że pojedyncza obserwacja ma duży wpływ na oceny parametrów. Przyjmuje się, że obserwacja jest wpływowa jeśli przekracza dwie średnie siły dźwigni. Inną podobną miarą wpływu obserwacji na model jest odległość `Cooka` (*ang. Cook's distance*). Wykazuje ona różnicę między wyznaczonymi wartościami współczynników, a wartościami obliczonymi przy wyłączeniu danego przypadku z obliczeń. Wszystkie odległości powinny być tego samego rzędu. Jeśli nie są, to można przypuszczać, że dany przypadek miał istotny wpływ na obciążenie współczynników równania regresji.

Wykresy diagnostyczne – wykres dźwigni

$$D_i = \frac{e_i^2}{s^2 k}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right],$$

gdzie

$$s^2 = \frac{1}{n-k}\boldsymbol{e}^\top \boldsymbol{e}$$

jest błędem średniokwadratowym,

$$h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i$$

jest siłą dźwigni, a k oznacza liczbę estymowanych parametrów.

Ralph Dennis Cook
(1944-)

Wykresy diagnostyczne – wykres dźwigni

Duże wartości (zwykle większe niż 1; przy czym wartości powyżej 0,5 oznaczają przypadek budzący wątpliwości) wskazują na znaczny wpływ pojedynczej obserwacji na oszacowane współczynniki regresji. Można się również spotkać z progiem odcięcia

$$\frac{4}{n-k-1}$$

lub

$$\frac{4}{n}.$$

Zaleca się spojrzenie na wartości D i sprawdzenie czy dla pewnej obserwacji są one znacznie większe niż dla pozostałych.

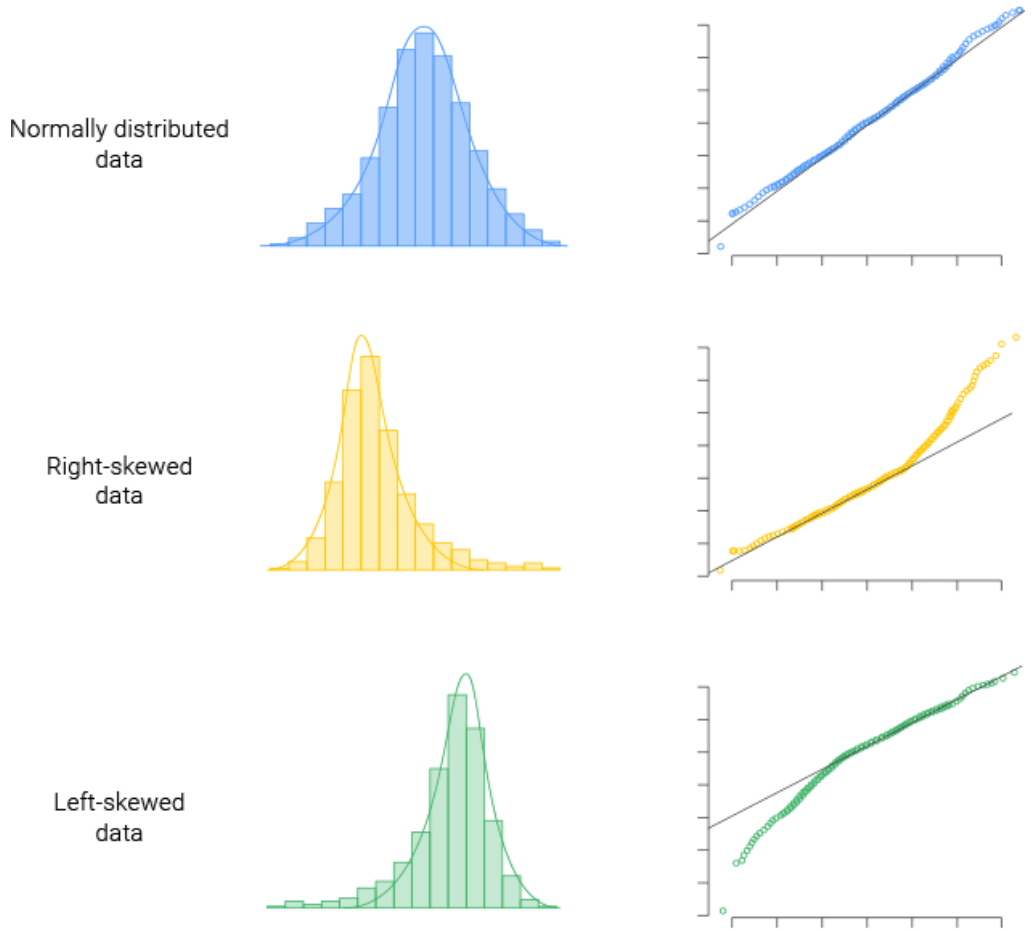
Cox, R.D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics* 19(1):15–18.

Wykresy diagnostyczne – wykres residuów

Wykres przedstawiający na jednej osi wartości dopasowane przez model, a na drugiej residua lub standaryzowane residua. Powszechną praktyką jest uznawanie, że obserwacja jest odstająca jeżeli jej residuum standaryzowane jest większe co do wartości bezwzględnej od 2. Dla modelu adekwatnego średnia wartość residuum nie powinna zależeć od wartości dopasowania (powinniśmy w wyniku dostać pas punktów losowo rozmieszczonych wokół prostej $y = 0$).

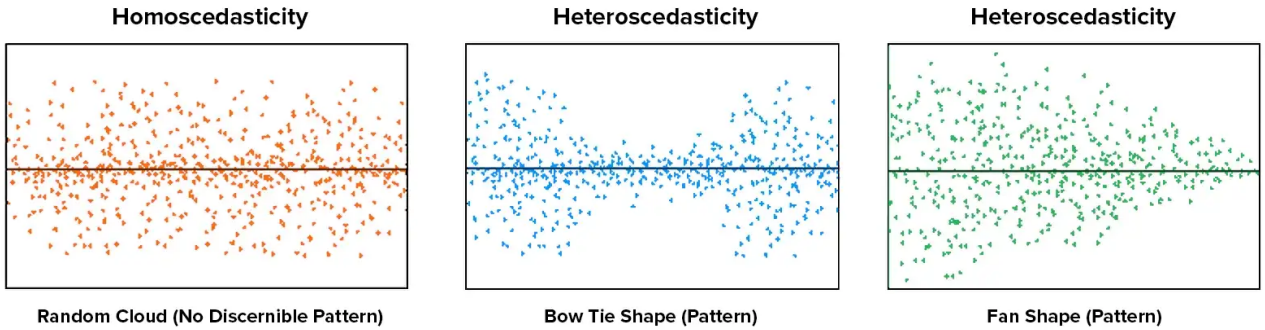
Wykresy diagnostyczne – wykres kwantylowy

Wykresy kwantylowe dla standaryzowanych residuów – powinny wskazać na ich normalność.



Wykresy diagnostyczne – wykres pierwiastków

Wykres, na którym dla każdej wartości y_i wyznaczono pierwiastek z wartości bezwzględnej jej residuum standaryzowanego. Nie powinno być żadnego trendu (jeśli jest, to wariancja błędu nie jest stała). Można również wykonać jeden z wielu testów. Najczęściej używany jest test `Breuscha-Pagana` lub test `White'a`. Hipoteza zerowa zakłada, że zachodzi homoskedastyczność.



Breusch, T.S., Pagan, A.R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 47(5):1287–1294.

White, H. (1980). Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4):817–838.

Przekształcenie Boxa-Coxa

W przypadku problemów z normalnością reszt można spróbować wykorzystać przekształcenie Boxa-Coxa:

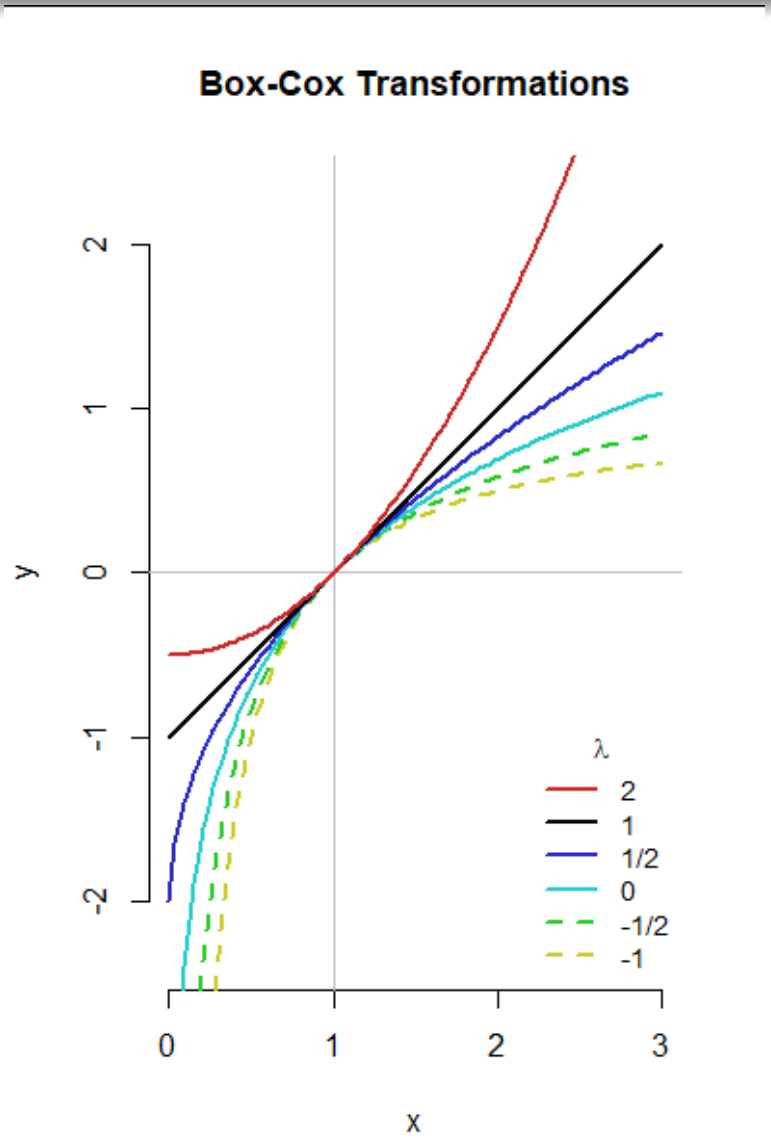
$$y_i = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \text{dla } \lambda \neq 0 \\ \ln(y_i) & \text{dla } \lambda = 0. \end{cases}$$

Transformacja oprócz poprawy normalności wyrównuje również wariancję. Powyższa postać transformacji jest poprawna jedynie dla nieujemnych wartości y. Dla ujemnych y mamy:

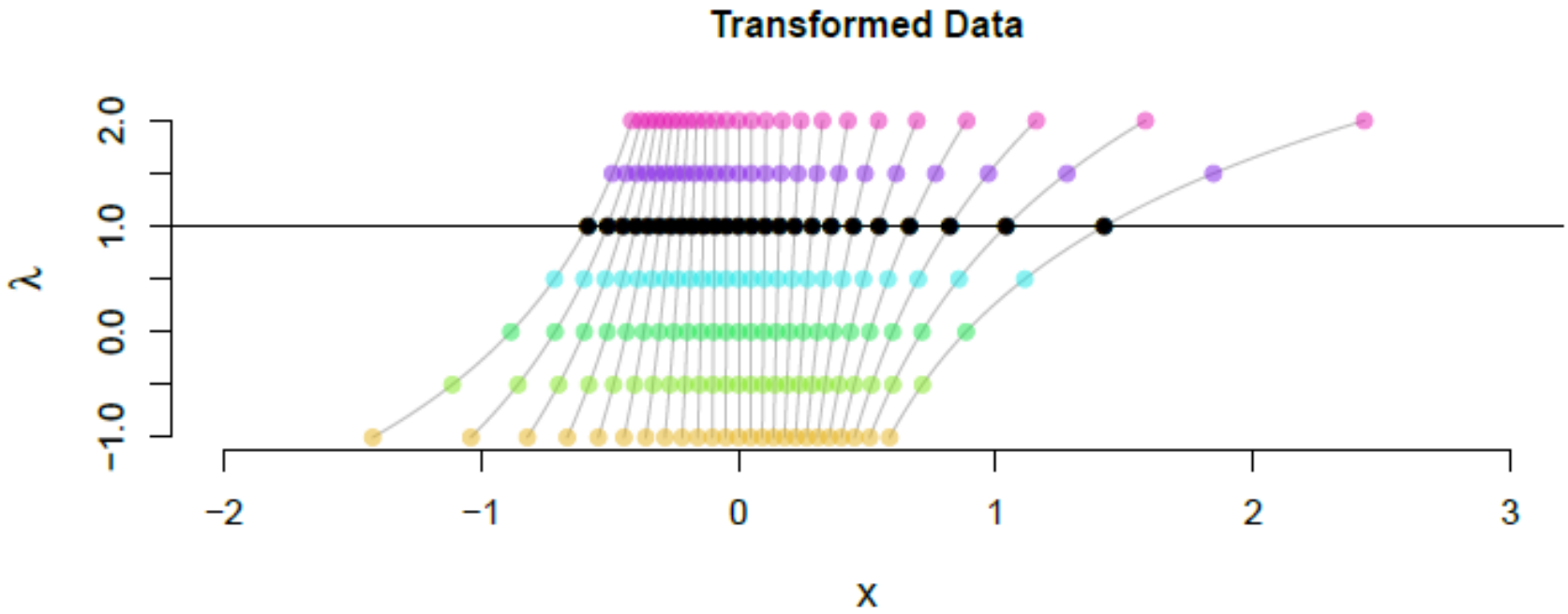
$$y_i = \begin{cases} \frac{(y_i+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{dla } \lambda_1 \neq 0 \\ \ln(y_i+\lambda_2) & \text{dla } \lambda_1 = 0. \end{cases}$$

Box, G.E.P., Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26(2):211–252.

Przekształcenie Boxa-Coxa

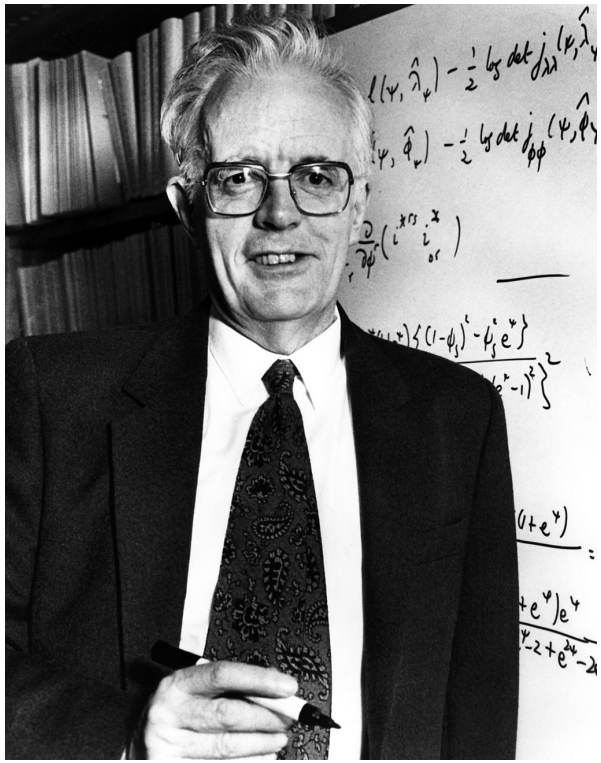


Przekształcenie Boxa-Coxa



Przekształcenie Boxa-Coxa

George Edward Pelham Box
(1919-2013)



Sir David Roxbee Cox
(1924-2022)

Modelowanie w R

Odpowiednie sformułowanie modelu w R odbywa się przy pomocy specjalnych formuł opisujących zależności zmiennych. Postać formuły jest następująca:

$$\text{zmienna objaśniana} \sim \text{zmienna(e) objaśniająca(e)},$$

gdzie symbol $\sim$ oznacza „jest modelowana jako funkcja" (zależy od).

Modelowanie w R

W formułach można używać wielu specjalnych symboli takich jak:

- $+$ dodanie zmiennej do modelu (nie suma zmiennych),
- $-$ usunięcie zmiennej z modelu (nie różnica zmiennych),
- -1 usunięcie wyrazu wolnego z modelu,
- $*$ dodanie wszystkich zmiennych oraz interakcji między nimi (nie mnożenie zmiennych),
- $\hat{}\, n$ wszystkie zmienne oraz interakcje pomiędzy nimi aż do rzędu n,
- $:$ interakcja pomiędzy zmiennymi,
- $.$ zależność od wszystkich zmiennych w podanej ramce danych.

Modelowanie w R

Można również używać funkcji arytmetycznych. Jeśli jednak chcemy skorzystać z operatorów arytmetycznych, które mają specjalne znaczenie w formułach powinniśmy skorzystać z funkcji I. Może się również zdarzyć sytuacja, w której chcemy jedynie poprawić istniejący już model, służy do tego funkcja update, w której kluczową rolę odgrywa „.". W zależności po której stronie znaku „∼" się znajduje, zastępuje prawą lub lewą stronę oryginalnej formuły.

```
model <- lm(y ~ x)
update(model, . ~ . -1) # y ~ x - 1
update(model, log(.) ~ .) # log(y) ~ x
```

Przykładowe formuły w R

Formuła	Opis
$y \sim 1$	Model pusty (średnia)
$y \sim x$	Regresja liniowa
$y \sim x - 1$	Regresja bez wyrazu wolnego (również $y \sim x + 0$)
$y \sim x + z$	Regresja wielokrotna
$y \sim x * z$	Regresja z interakcją, inaczej $y \sim x + z + x : z$
$y \sim x + \mathbf{I}(x\hat{}2)$	Regresja kwadratowa
$y \sim x + \mathbf{I}(x\hat{}2) + \mathbf{I}(x\hat{}3)$	Regresja sześcienna
$y \sim (x + z + w)\hat{}2$	$y \sim x + z + w + x : z + x : w + z : w$
$y \sim x * z - x$	$y \sim z + x : z$
$y \sim x/z$	$y \sim x + x : z$
$\log(y) \sim \mathbf{I}(1/x) + \mathbf{sqrt}(z)$	Użycie funkcji arytmetycznych

Regresja liniowa w R

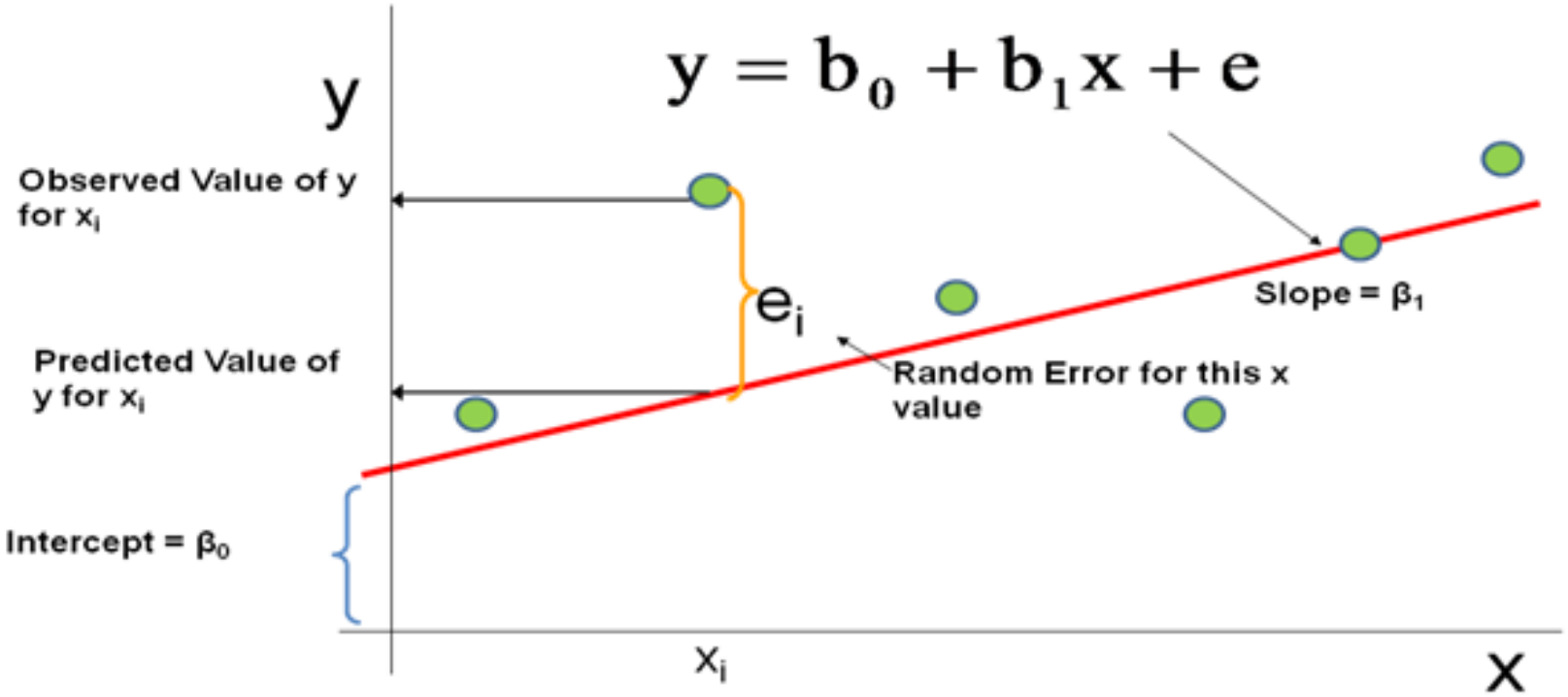
Do wykonania analizy regresji służy funkcja lm, w której podajemy jako argument formułę opisującą model. Jako wynik otrzymujemy oszacowany model regresyjny. Wywołanie na nim funkcji summary przedstawia kolejno wartości reszt (lub, w przypadku większej ich liczby, wartości skrajne, medianę i kwartyle), estymatory nachylenia prostej i przecięcia z osią y. Dla każdego z estymatorów podany jest błąd standardowy oraz odpowiadające mu wartości statystyki t i p-wartości dla jego istotności, otrzymujemy również współczynnik R^2 oraz R^2_{popr}. Na skonstruowanym modelu można również wywołać funkcje: coef (współczynniki modelu), confint (przedziały ufności dla parametrów), fitted (wartości dopasowane przez model), residuals (wartości reszt). Przeciążona funkcja plot rysuje wykresy diagnostyczne (domyślnie cztery opisane wcześniej).

Przykład – zachorowania na gruźlicę

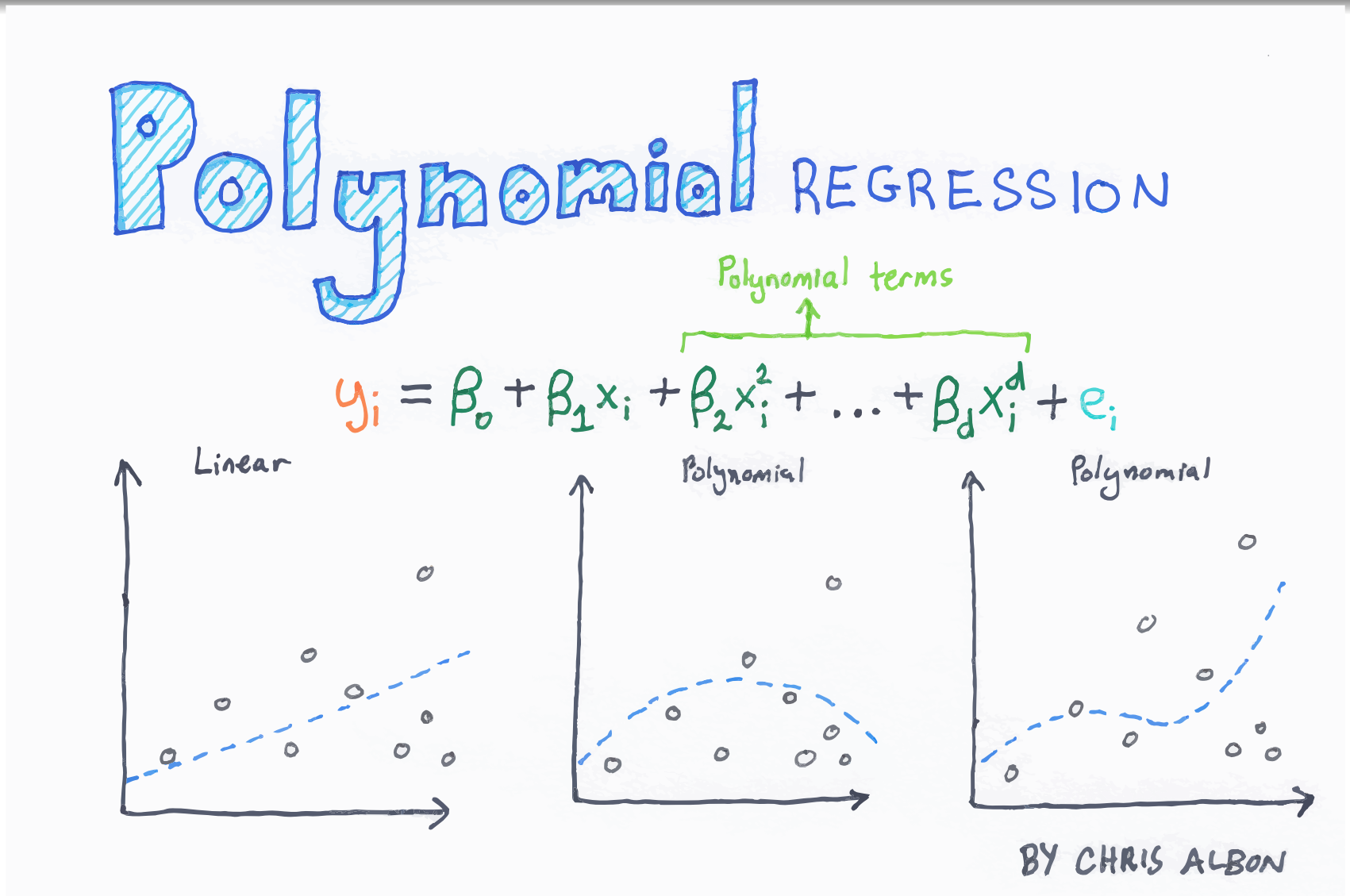
Poniższa tabela przedstawia liczbę zachorowań na gruźlicę układu oddechowego w latach 1995-2002. Liczba zachorowań została podana w przeliczeniu na 100 tys. ludności.

Rok (x_i)	1995	1996	1997	1998	1999	2000	2001	2002
Zachorowania (y_i)	39,7	38,2	34,7	33,1	30,1	28,4	26,3	24,7

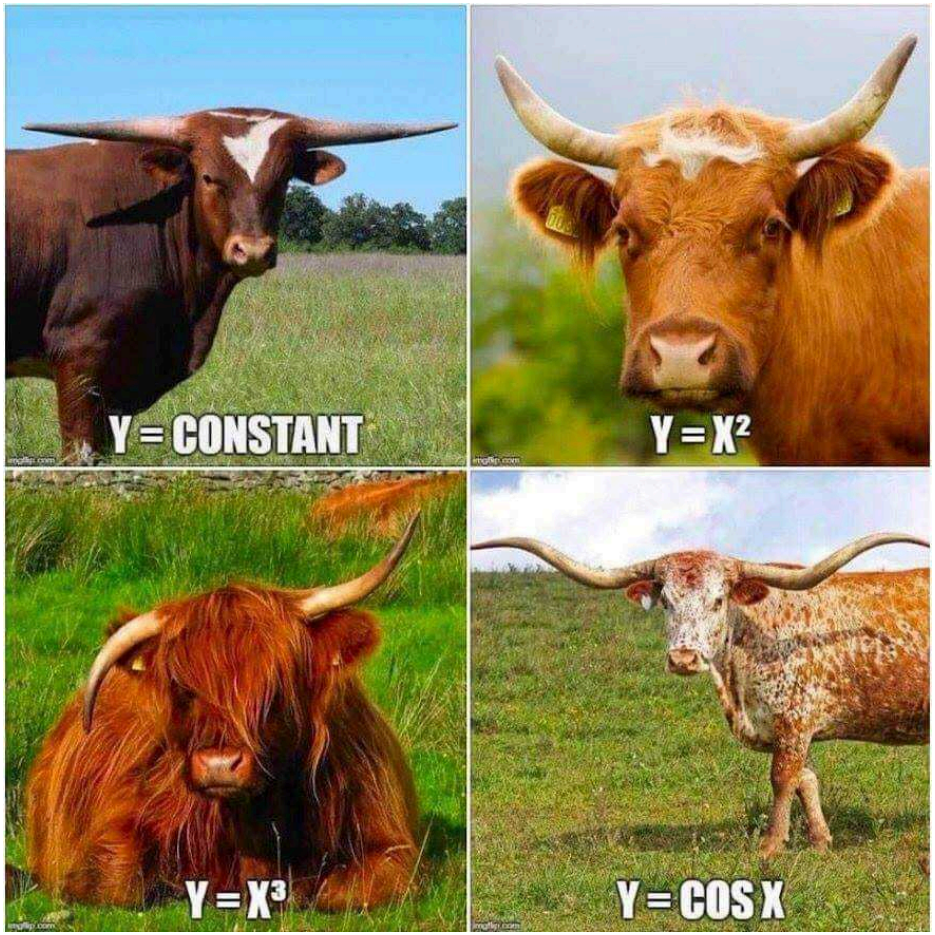
Przykład – zachorowania na gruźlicę



Regresja wielomianowa



Regresja wielomianowa



Joseph Diez Gergonne
(1771-1859)

Gergonne, J.D. (1815). Application de la méthode des moindre quarrés a l'interpolation des suites (ang. The application of the method of least squares to the interpolation of sequences) *Annales des Math Pures et Appl* 6:242–252.

Regresja wielokrotna

Wcześniej założyliśmy, że zmienna objaśniana zależy jedynie od jednej zmiennej objaśniającej. Jest to duże uproszczenie. Zdarza się, że badane zjawisko zależy nie tylko od jednego czynnika, ale od wielu. Uogólnieniem regresji prostej jest regresja wielokrotna (ang. *multiple regression*) lub wieloraka (termin wprowadzony przez `K. Pearsona` w 1908 roku), w której uwzględnia się wpływ wielu cech niezależnych na jedną cechę zależną. Załóżmy, że dysponujemy teraz układem k cech $X_1, X_2, \ldots, X_k$. Model regresji wielokrotnej można zapisać w postaci:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + \beta_k \boldsymbol{X}_k + \boldsymbol{\varepsilon},$$

gdzie $\boldsymbol{Y}$ jest wektorem obserwacji zmiennej objaśnianej, $\boldsymbol{X}$ macierzą z pomiarami zmiennych objaśniających (pierwsza kolumna to kolumna jedynek odpowiadająca za wyraz wolny w modelu) a $\boldsymbol{\beta}$ jest wektorem parametrów.

Regresja wielokrotna

W celu estymacji parametrów modelu ponownie używamy MNK otrzymując (oprócz poprzednich założeń, musimy jeszcze przyjąć, że nie istnieje liniowa zależność pomiędzy zmiennymi objaśniającymi):

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Regresja wielokrotna

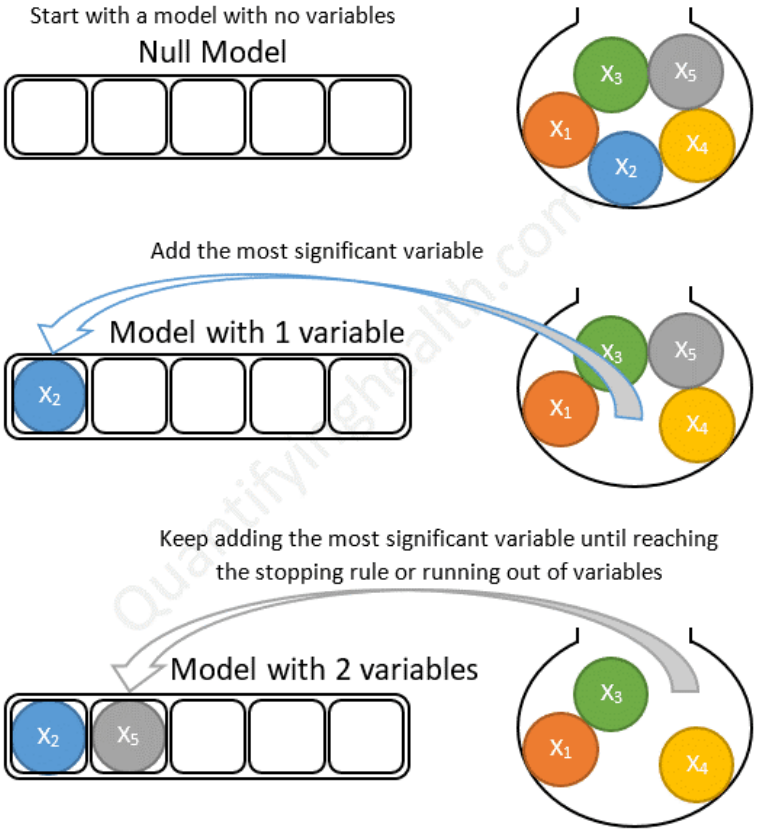
Częstokroć w przypadku wykorzystania regresji wielorakiej bardziej od prognozy interesuje nas, które zmienne wpływają na badane zjawisko w sposób pobudzający, a które je hamują. Pierwsze z tych czynników nazywamy stymulantami, a drugie destymulantami. Oczywiście stymulantami są zmienne, które w oszacowanym modelu regresji mają dodatnie wartości parametrów regresji. Destymulanty to zmienne o ujemnych parametrach. Można jeszcze określić zmienne neutralne (nieistotne), czyli takie, które nie mają wpływu na badane zjawisko.

Regresja krokowa

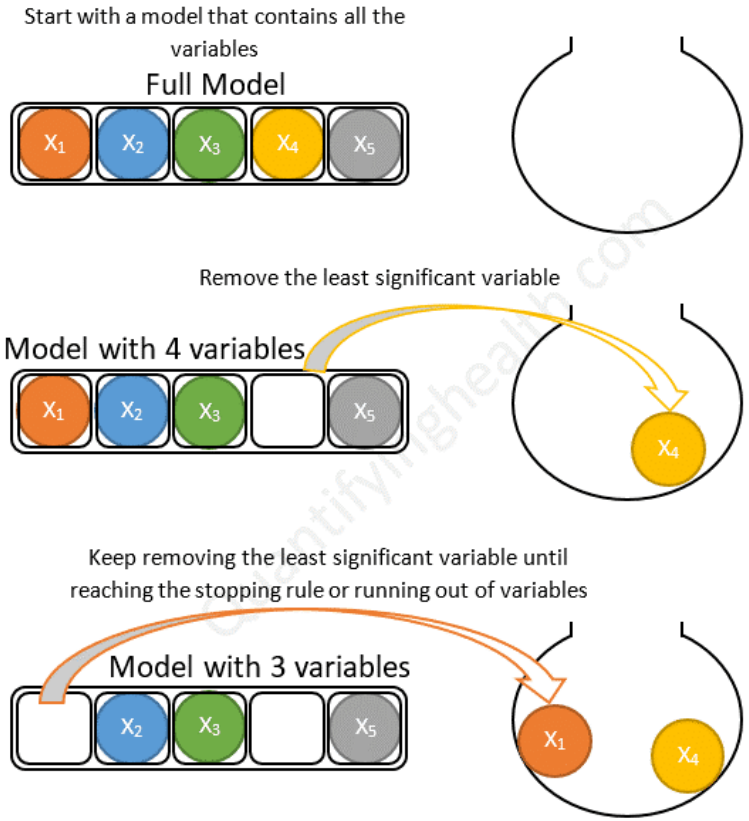
Istnieje również inna metoda budowania modeli z dużą liczbą zmiennych objaśniających niż konstrukcja pełnego modelu i oszacowanie jego parametrów. Jest to procedura regresji krokowej, w której na każdym kroku możemy odrzucić lub dodać zmienną. Powiedzmy, że zaczynamy od modelu zawierającego tylko stałą (można zacząć również od modelu pełnego). W kolejnym kroku dodajemy najlepszą w sensie jakiegoś kryterium (np. test t) zmienną. W kolejnym dodajemy znowu, ale sprawdzamy również co się dzieje jak byśmy z tego modelu usunęli dodaną w poprzednim kroku zmienną itd.

Regresja krokowa

Forward stepwise selection example with 5 variables:



Backward stepwise selection example with 5 variables:



Regresja krokowa

Jakość modelu oceniana jest za pomocą współczynnika informacyjnego `Akaike` (AIC) – *ang. Akaike information criterion*. Wartość tego współczynnika zależy nie tylko od sumy kwadratów reszt, ale również od ilości zmiennych w modelu. Zatem zwiększając liczbę parametrów w modelu, pomimo iż suma kwadratów reszt zawsze maleje, od pewnego momentu współczynnik AIC zacznie rosnąć. Kryterium AIC ma tendencję do wybierania modelu ze zbyt dużą liczbie parametrów. Jeśli bardziej zależy nam na jakości predykcji powinniśmy użyć kryterium AIC, jeśli natomiast priorytetem jest jakość dopasowania modelu należy wybrać bayesowski współczynnik informacyjny (BIC) – *ang. Bayesian information criterion or Schwarz information criterion*.

Regresja krokowa

$$\text{AIC} = -2\ln(\mathcal{L}) + 2k,$$
$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1},$$
$$\text{BIC} = -2\ln(\mathcal{L}) + \ln(n)k,$$

gdzie $\mathcal{L}$ jest funkcją największej wiarogodności, a k oznacza liczbę estymowanych parametrów. Dla małych prób ($n/k < 40$) zalecany jest współczynnik AICc.

Hirotugu Akaike
(1927-2009)

Regresja krokowa

W przypadku estymacji parametrów metodą najmniejszych kwadratów i założeniu normalności błędów formuła ma następującą postać:

$$\mathrm{AIC} = n \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + n\ln(2\pi) + n + 2 + 2k.$$

Bezwzględne wartości współczynnika AIC nie podlegają interpretacji, ponieważ zawierają w sobie stałe zależne od wielkości próby. Z tego powodu wylicza się

$$\Delta_i = \mathrm{AIC} - \mathrm{AIC}_{\min}.$$

Teraz najlepszy model ma $\Delta_i = 0$. Zatem Δ_i mierzy stratę informacji jakiej doznamy jeśli użyjemy modelu i-tego zamiast modelu z najmniejszą wartością współczynnika AIC.

Regresja krokowa

Została zaproponowana pewna skala według jakiej można interpretować Δ_i:

- $\Delta_i \leq 2$ – model jest porównywalny z modelem z AIC_{min}.
- $2 < \Delta_i \leq 4$ – model ma dużą szansę na bycie porównywalnym z modelem z AIC_{min}.
- $4 < \Delta_i \leq 7$ – model ma niewielką szansę na bycie porównywalnym z modelem z AIC_{min}.
- $\Delta_i \geq 10$ – model jest gorszy od modelu z AIC_{min}.

 Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716–723.

 Burnham K.P., Anderson D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods Research* 33(2):261–304.

 Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2):461–464.

Regresja odporna

Podobnie jak średnia czy odchylenie standardowe współczynniki regresji są wrażliwe na obserwacje odstające. I podobnie jak dla nich możemy poszukiwać tzw. regresji odpornej. Jedna z nich to tzw. metoda najmniejszych przyciętych kwadratów (*ang. least trimmed squares (LTS)*), w której zamiast zwykłej sumy używamy sumy przyciętej (wykonujemy regresję liniową, liczymy residua, usuwamy największe residua i ponownie estymujemy parametry minimalizując sumę kwadratów $m = \lfloor n/2 \rfloor + \lfloor (k+2)/2 \rfloor$ najmniejszych residuów.). Nieco inne podejście oferuje metoda regresji odpornej używająca M estymatorów (*ang. robust regression using an M estimator*), która jest najbardziej polecana w przypadku istnienia obserwacji odstających.

Huber, P.J. (1981). *Robust Statistics*. Wiley.

Rousseeuw, P.J., Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. Wiley.

Regresja kwantylowa

Regresja kwantylowa (*ang. quantile regression*) została zaproponowana przez Koenkera i Bassetta (1978). Szczególny przypadek regresji kwantylowej dla kwantyla rzędu 0,5 (czyli mediany) jest równoważny estymatorowi LAD (*ang. Least Absolute Deviation*) – minimalizuje sumę bezwzględnych błędów. Wprowadzenie różnych kwantyli regresji daje pełniejszy opis rozkładów warunkowych zwłaszcza w przypadku rozkładów asymetrycznych lub uciętych.

Koenker R., Bassett G. (1978). Regression Quantiles. *Econometrica* 46(1):33–50.

Regresja kwantylowa

Jeżeli u_i jest błędem predykcji, to OLS (*ang. Ordinary Least Squares*) minimalizuje $\sum_i u_i^2$, natomiast LAD minimalizuje $\sum_i |u_i|$. Regresja kwantylowa minimalizuje sumę, która daje asymetryczne wagi: $(1-\tau)$ dla zbyt wielkich predykcji oraz τ dla zbyt małych, czyli minimalizuje funkcję:

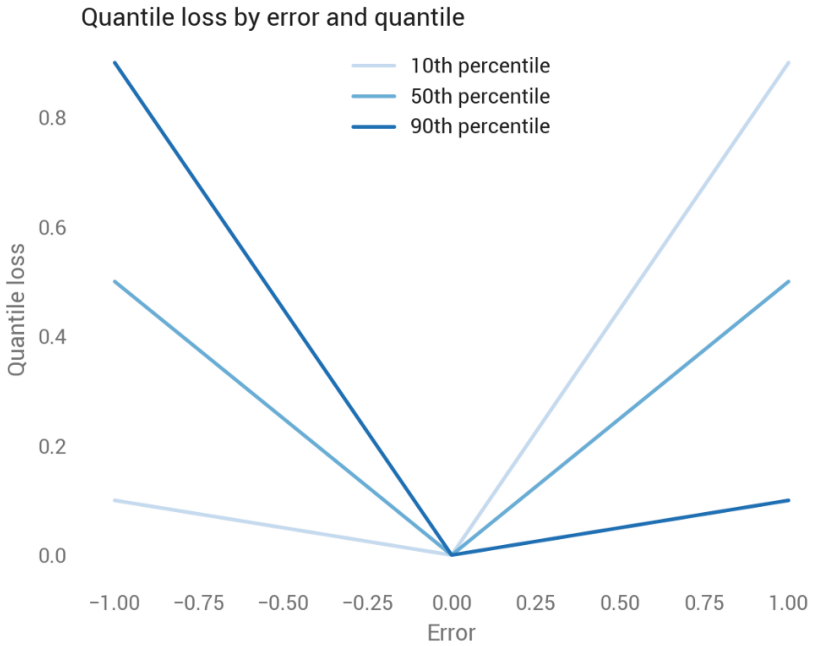
$$Q(\boldsymbol{\beta}_\tau) = \sum_{i=1}^{n} \rho_\tau |y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau| =$$

$$= \sum_{i: y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau \geq 0} \tau |y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau| + \sum_{i: y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau < 0} (1-\tau) |y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau|,$$

gdzie

$$\rho_\tau(u_i) = \tau \cdot \max(u_i, 0) + (1-\tau) \cdot \max(-u_i, 0).$$

Regresja kwantylowa



Funkcja Q jest nieróżniczkowalna i jej minimum znajduje się za pomocą metody programowania liniowego. Tak znalezione estymatory są asymptotycznie normalne. Regresja kwantylowa jest bardziej odporna na obserwacje odstające oraz unikamy założeń co do rozkładów błędów.

Regresja segmentowa

Modele segmentowe (*ang. segmented, broken-line, piecewise*) to modele regresji, w których relacje między odpowiedzią a jedną lub więcej zmiennymi objaśniającymi są liniowe w sposób kawałkowy, a mianowicie reprezentowane przez dwie lub więcej prostych linii połączonych w nieznanych wartościach: te wartości zwykle określa się jako punkty zmiany (*ang. breakpoints, changepoints, joinpoints*). Dla jednego punktu zmiany mamy:

$$Y_i = \beta_1 X_i + \beta_2 (X_i - \psi)_+ + \varepsilon_i,$$

gdzie

$$(X_i - \psi)_+ = \begin{cases} X_i - \psi, & \text{dla } X_i - \psi > 0 \\ 0, & \text{w p.p.} \end{cases}$$

oraz ψ jest punktem zmiany, β_1 współczynnikiem kierunkowym i β_2 różnicą we współczynnikach kierunkowych.

Muggeo, V. (2003), Estimating regression models with unknown break-points. *Statistics in Medicine* 22:3055–3071.