Wstęp do uczenia maszynowego

Regresja liniowa 1

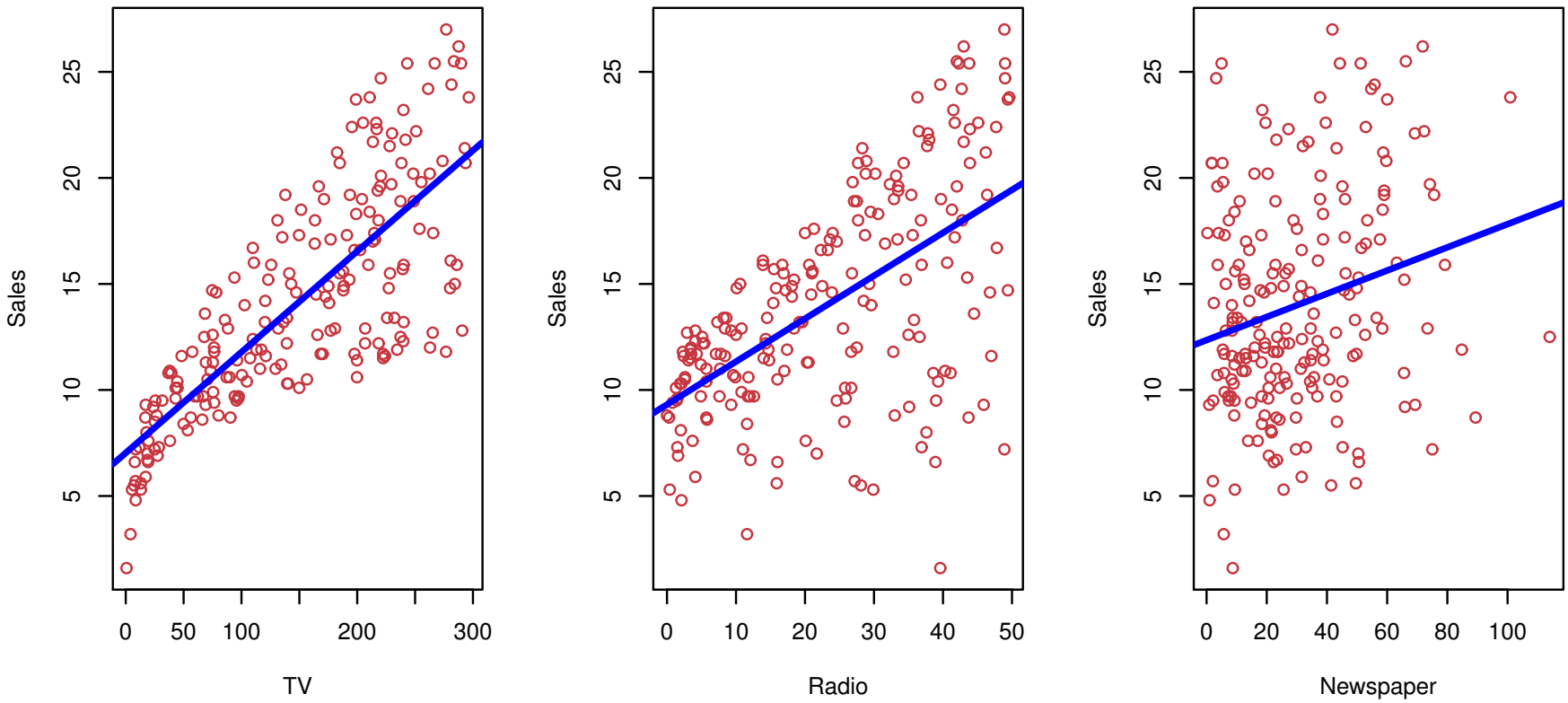
Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

25 marca 2024

UNIWERSYTET
WARSZAWSKI

Dane: wpływ reklam produktu na sprzedaż w 200 sklepach



- **Oś X:** nakłady na reklamę w danym medium w tysiącach $
- **Oś Y:** liczba sprzedanych jednostek produktu (w tysiącach)

Przykładowe pytania

- Czy jest związek pomiędzy wielkością reklamowego budżetu a wielkością sprzedaży?
- Jak silny jest ten związek, np. czy umiemy przewidzieć wielkość sprzedaży w zależności od wysokości budżetu?
- Które media mają największy wpływ na wysokość sprzedaży?
- Jak dokładnie umiemy przewidzieć, dla każdego z mediów, wysokość sprzedaży w zależności od wysokości budżetu reklamowego?
- Jak dokładnie umiemy przewidzieć wysokość sprzedaży w przyszłości?
- Czy zależność jest liniowa?
- Czy istnieje synergia pomiędzy mediami?

Prosta regresja liniowa z jedną zmienną objaśniającą

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

gdzie X to zmienna objaśniająca, Y to zmienna objaśniana, ε to błąd, $\mathbb{E}(\varepsilon) = 0$.

Na przykład:

$$sales \approx \beta_0 \ + \ \beta_1 \times (TV),$$

gdzie β_0, β_1 to (nieznane) parametry.

Prosta regresja liniowa z jedną zmienną objaśniającą

Mamy n obserwacji (dane treningowe):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

Nasz model zakłada, że

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

oraz, że

- Wartości wejściowe x_i deterministyczne (nie losowe) obserwowane
- Zmienne ε_i niezależne o identycznym rozkładzie
- $\mathbb{E}[\varepsilon_i] = 0$
- $\mathrm{Var}[\varepsilon_i] = \sigma^2 < \infty$

dla każdego $i = 1, \ldots, n$.

Predykcja w regresji liniowej z jedną zmienną objaśniającą

Mając estymacje $\hat{\beta}_0$ oraz $\hat{\beta}_1$ obliczamy predykcję wartości dla $\hat{Y}$ na podstawie wartości $X = x$ jako

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Estymacja parametrów (jedna zmienna objaśniająca)

Przy zadanych estymacjach parametrów $\hat{\beta}_0$, $\hat{\beta}_1$, błąd RSS (suma kwadratów reszt, ang. *residual sum of squres*):

$$RSS = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

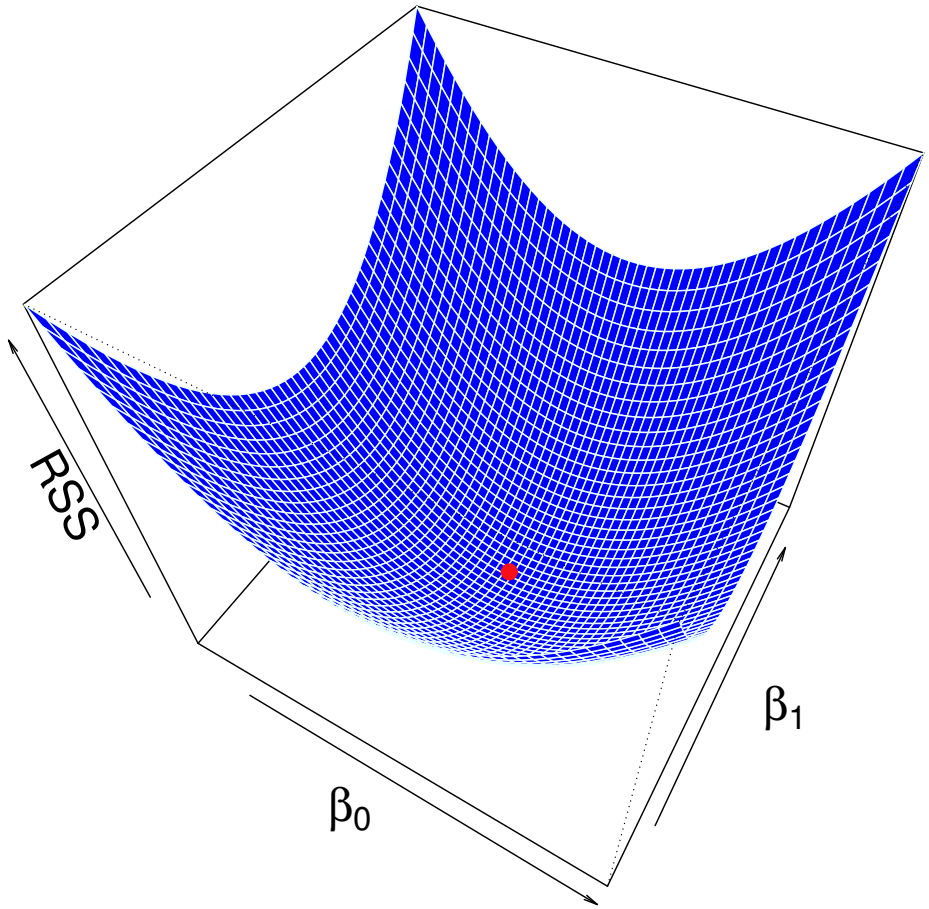
Minimalizacja tego błędu prowadzi do *estymatorów najmniejszych kwadratów dla współczynników*:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

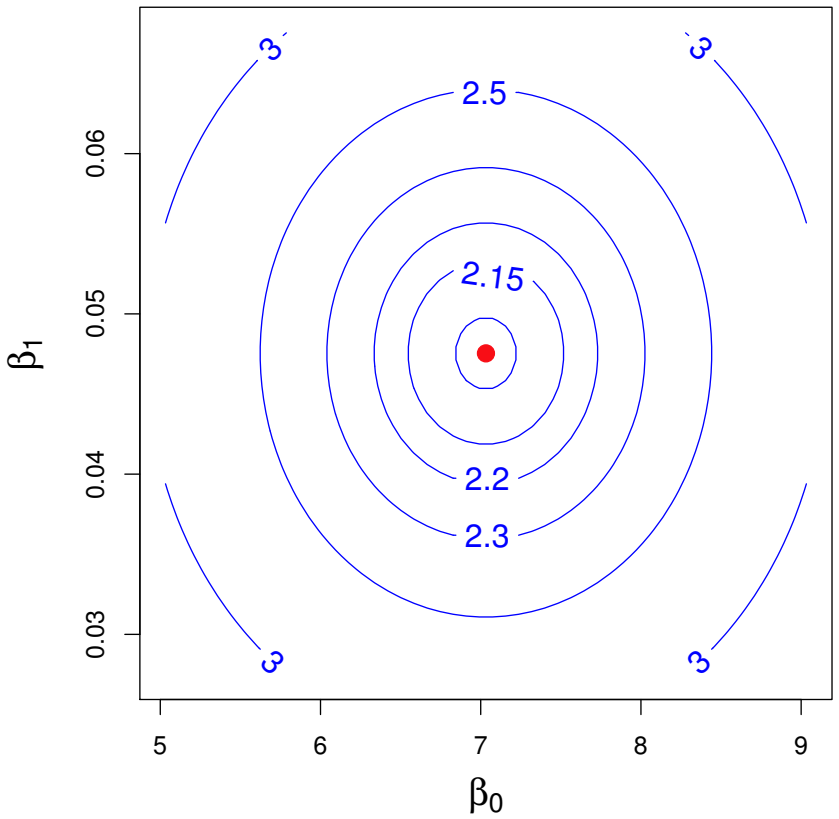
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

gdzie $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$ oraz $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$.

Błąd RSS w zależności od β_0, β_1

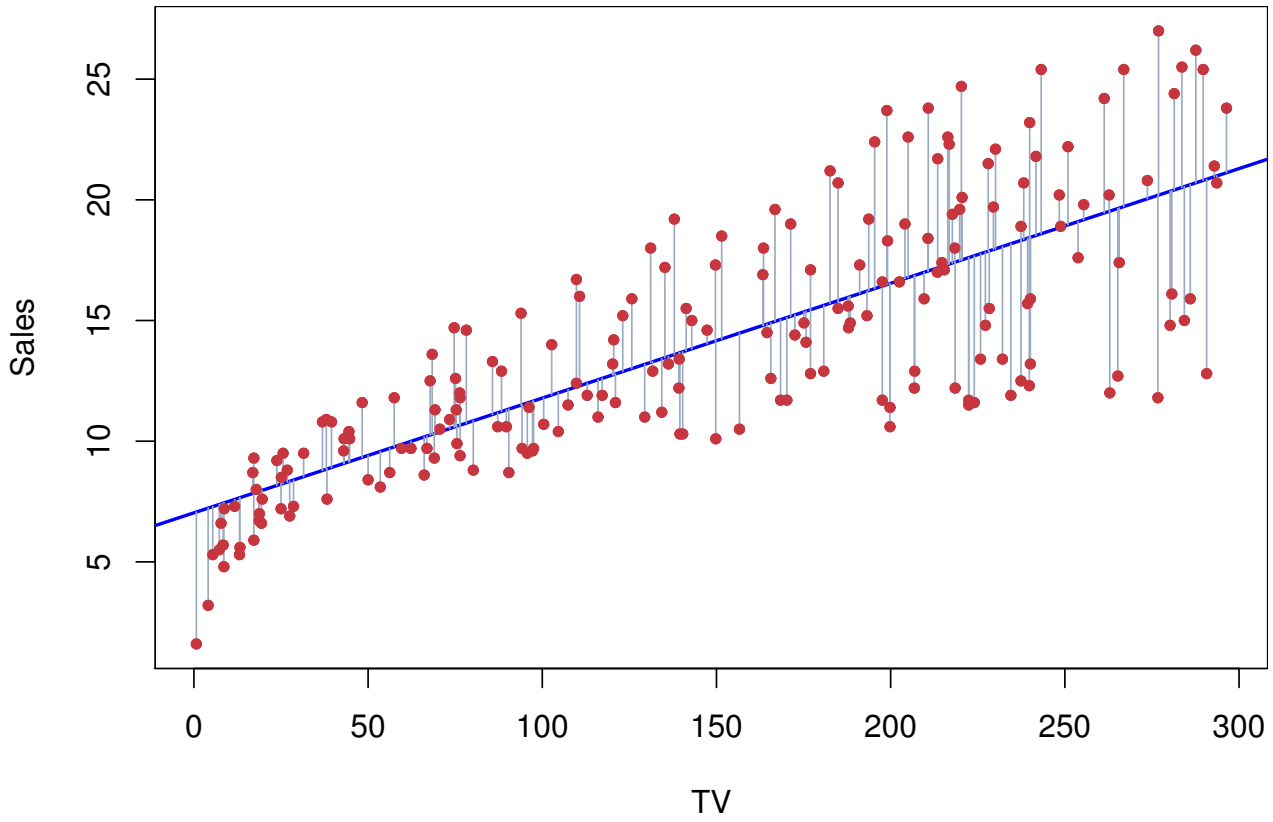


Optymalne parametry: $\hat{\beta}_0 = 7.03$, $\hat{\beta}_1 = 0.0475$



Dołożenie \$1000 na reklamę w TV powoduje zwiększenie sprzedaży średnio o 47.5 jednostek produktu.

Dopasowanie dla danych TV przy pomocy liniowej regresji



Optymalne dopasowanie: szare odcinki reprezentują błąd dopasowania dla poszczególnych obserwacji.

Regresja liniowa z wieloma zmiennymi objaśniającymi

Model regresji:

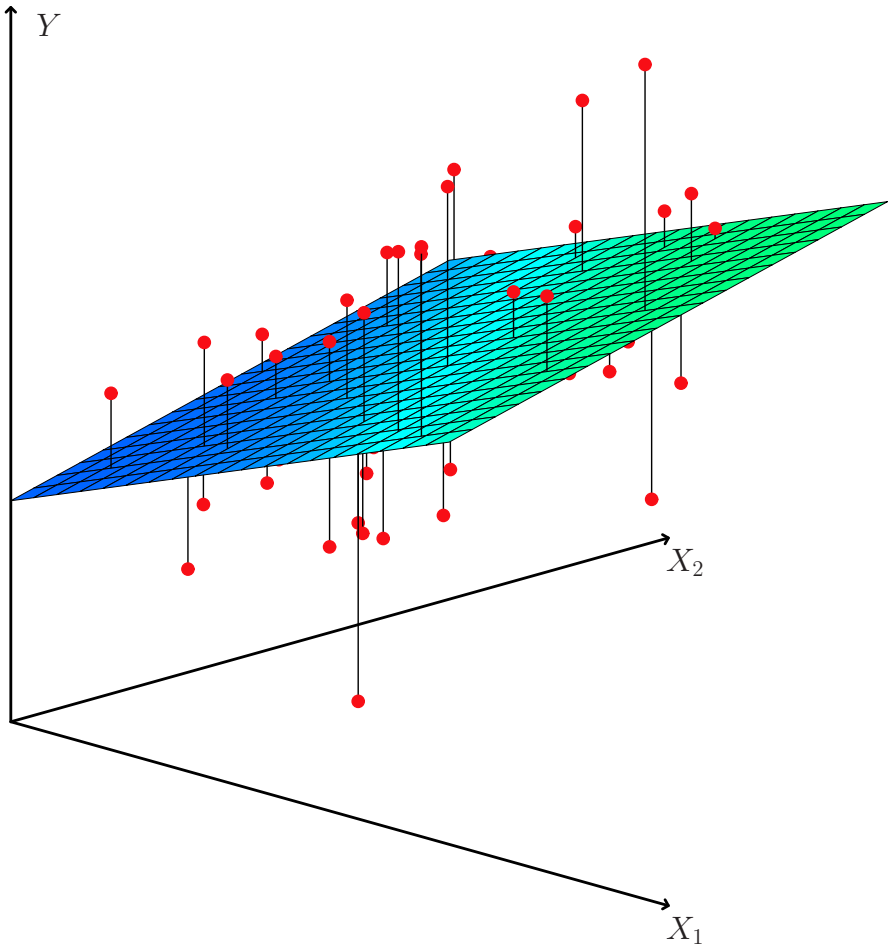
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon,$$

gdzie $X_1, \ldots, X_p$ są zmiennymi objaśniającymi.

Dla modelu sprzedaży mamy:

$$sales = \beta_0 \ + \ \beta_1 \times (TV) \ + \ \beta_2 \times (radio) \ + \ \beta_3 \times (newspaper) \ + \ \varepsilon$$

Regresja liniowa dla danych o dwóch predyktorach daje płaszczyznę regresji (zamiast prostej)



Regresja liniowa z wieloma zmiennymi objaśniającymi

- Dla danych $(x_1, y_1), \ldots, (x_n, y_n)$, gdzie każdy $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})^T$ jest wektorem wejściowych wartości p zmiennych objaśniających
- model regresji ma postać

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \varepsilon_i,$$

- Weźmy $n \times p$ macierz o wejściach $x_{i,j}$ w wierszu i i kolumnie j (każdy wiersz i odpowiada x_i^T)
- Niech $\mathbf{X}$ będzie $n \times (p+1)$ macierzą, otrzymaną z tej macierzy przez dołożenie jako pierwszej kolumny z wartościami 1
- Niech $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ będzie wektorem wartości zmiennej objaśnianej, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ wektorem błędów
- Model w postaci macierzowej:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Liniowa regresja z wieloma zmiennymi objaśniającymi

Dla modelu w postaci macierzowej

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

założenia są postaci

- Wartości macierzy $\mathbf{X}$ deterministyczne (nie losowe) obserwowane
- ϵ wektor losowy niezależnych zmiennych losowych o identycznym rozkładzie
- $\mathbb{E}[\epsilon] = 0$
- $\mathrm{Var}[\epsilon] = \sigma^2 I_n$, $\sigma^2 < \infty$

Dodatkowo założymy, że $\mathbf{X}$ ma pełny rząd, $rank(\mathbf{X}) = p$

Estymacja parametrów dla regresji z wieloma zmiennymi objaśniającymi

Zauważmy, że

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{X}\beta + \varepsilon] = \mathbf{X}\beta + \mathbb{E}[\epsilon] = \mathbf{X}\beta.$$

Szukamy (*metodą najmniejszych kwadratów*) wartości $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, które minimalizują wyrażenie

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \ldots - \hat{\beta}_p x_{i,p})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{i,j})^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \\ &= ||\mathbf{y} - \mathbf{X}\beta||^2 \end{aligned}$$

RSS jest kwadratową funkcją parametrów β.

Estymacja parametrów dla regresji z wieloma zmiennymi objaśniającymi

- Różniczkując po β mamy

$$\begin{aligned} \frac{\delta RSS}{\delta \beta} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\delta^2 RSS}{\delta \beta \delta \beta^T} &= 2\mathbf{X}^T\mathbf{X} \end{aligned}$$

- Dla macierzy $\mathbf{X}$ pełnego rzędu p mamy $\mathbf{X}^T\mathbf{X}$ dodatnio określoną i istnieje dokładnie jedno minimum RSS, które wyznaczamy przez

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

otrzymując

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Predykcja dla regresji liniowej z wieloma zmiennymi objaśniającymi

- Dla danych treningowych

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- Macierz $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ nazywamy **macierzą daszkową** (bo zakłada daszek na $\mathbf{y}$).
- Dla danej nowej obserwacji testowej x_0

$$\hat{y}_0 = (1 : x_0)^T \hat{\beta}$$

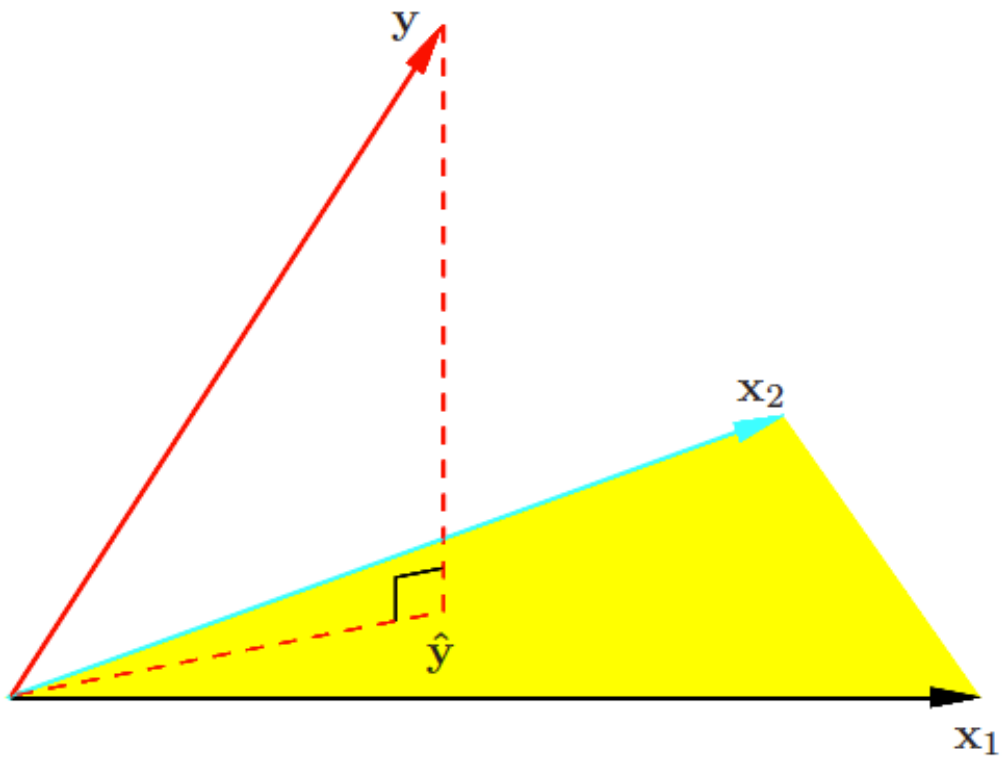
Interpretacja geometryczna

- Niech $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p$ wektory odpowiadające kolumnom macierzy $\mathbf{X}$, gdzie $\mathbf{x}_0 = 1$
- Wektory $\mathbf{x}_i$ rozpinają podprzestrzeń $\mathbb{R}^n$ ($lin(\mathbf{X})$, przestrzeń kolumn $\mathbf{X}$)
- Minimalizując $RSS = ||\mathbf{y} - \mathbf{X}\beta||^2$, wybieramy taki $\hat{\beta}$, dla którego wektor $\mathbf{y} - \hat{\mathbf{y}}$ jest ortogonalny do $lin(\mathbf{X})$.
- Widać to z tego równania

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

- $\hat{\mathbf{y}}$ jest rzutem ortogonalnym $\mathbf{y}$ na $lin(\mathbf{X})$.

Regresja liniowa w n wymiarach



Własności estymatorów $\hat{\beta}$

- Wartość oczekiwana

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbb{E}[\mathbf{y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$$

Stąd, $\hat{\beta}$ jest nieobciążony.

- Macierz (ko-)wariancji

$$\text{Var}[\hat{\beta}] = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

- Nieobciążony estymator σ^2

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = RSS/(n-p-1).$$

$\hat{\beta}$ jest BLUE

BLUE-Best Linear Unbiased Estimator

Twierdzenie Gaussa-Markowa

Dla błędów ϵ_i nieskorelowanych ($\mathbb{E}[\epsilon_i \epsilon_j] = 0$] dla $i \neq j$) i homoskedastycznych ($\mathbb{E}[\epsilon_i^2] = \sigma^2 < \infty$), estymator $\hat{\beta}$ jest liniowym, nieobciążonym estymatorem o najmniejszej wariancji parametru β.

Testowanie istotności danego predyktora

- Przy dodatkowym założeniu, że $\epsilon \sim N(0, \sigma^2 I_n)$ mamy

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

- Aby przetestować $H_0 : \beta_i = 0$ dla i-tego współczynnika korzystamy ze statystyki

$$z_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{v_i}},$$

gdzie v_i to i-ty element na diagonali macierzy $(\mathbf{X}^T\mathbf{X})^{-1}$

- Przy założeniu hipotezy zerowej
  - z_i ma rozkład $t(n-p-1)$
  - Dla hipotezy $H_1 : \beta_i \neq 0$ obszar krytyczny na poziomie istotności α: $W = (-\infty, -t(1-\alpha/2, n-p-1)] \cup [t(1-\alpha/2, n-p-1), \infty)$
- Przy zastąpieniu $\hat{\sigma}$ znanym σ, rozkład $N(0,1)$
- Dla dużych prób n mamy dobre przybliżenie rozkładu t rozkładem standardowym normalnym i korzysta się z kwantyli tego rozkładu

Przedział ufności dla estymatora $\hat{\beta}_i$

- Na poziomie ufności $1-\alpha$ przedział ufności dla $\hat{\beta}_i$

$$(\hat{\beta}_i - z(1-\alpha/2)\sqrt{v_i}\hat{\sigma}, \hat{\beta}_i + z(1-\alpha/2)\sqrt{v_i}\hat{\sigma}),$$

gdzie $z(1-\alpha/2)$ kwantyl rzędu $1-\alpha/2$ rozkładu standardowego normalnego.

- Ponieważ $z(1-0.025) = 1.96$ dla przedziału ufności na poziomie 95% korzysta się często z przybliżenia

$$(\hat{\beta}_i - 2SE(\hat{\beta}_i), \hat{\beta}_i + 2SE(\hat{\beta}_i)),$$

gdzie $SE(\hat{\beta}_i) = \sqrt{v_i}\hat{\sigma}$ nazywane jest błędem standardowym.

Przykład cd: model *sales* w zależności od *TV*

	Współczynnik	Std. błąd	t-statystyka	p-wartość
Intercept	$\hat{\beta}_0 = 7.0325$	0.4578	15.36	< 0.0001
TV	$\hat{\beta}_1 = 0.0475$	0.0027	17.67	< 0.0001

- Według testu t, zmienna *TV* wygląda na istotny predyktor *sales*.

Przedziały ufości na poziomie 95%:

- dla $\hat{\beta}_0$: [6.130, 7.935],
- dla $\hat{\beta}_1$: [0.042, 0.053].

Oznacza to, że z prawdopodobieństwem 95%

- bez żadnych nakładów na reklamę sprzedaż wyniesie pomiędzy 6130 a 7935 jednostek,
- z każdym wydanym \$1000 na reklamę, wzrost sprzedaży wyniesie pomiędzy 42 a 53 jednostek.

Testowanie istotności kilku predyktorów

- Chcemy przetestować istotność kilku predyktorów naraz
- Korzystamy ze statystyki

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(n - p_1 - 1)},$$

gdzie

  - RSS_1 to resztowa suma kwadratów modelu większego, o $p_1 + 1$ parametrach,
  - RSS_0 to resztowa suma kwadratów dla mniejszego, zagnieżdżonego modelu o $p_0 + 1$ parametrach, w kótrym $p_1 - p_0$ parametrów wyzerowano.

- Przy założeniu $\varepsilon \sim N(0, \sigma^2)$ i hipotezy zerowej, że mniejszy model jest prawdziwy i $p_1 - p_0$ parametrów równa się 0, statystyka F ma rozkład Snedecora-Fishera $F(p_1 - p_0, n - p_1 - 1)$

F-test dla wszystkich predyktorów (ang. overall F-test)

- Chcemy przetestować istotność zależności od conajmniej jednego z predyktorów
- Korzystamy ze statystyki

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

gdzie

- $TSS = \sum(y_i - \bar{y})^2$ (ang. *total sum of squares*). RSS dla modelu, w którym współczynniki wszystkich predyktorów są wyzerowane.
- $RSS = \sum(y_i - \hat{y}_i)^2$ dla modelu ze wszystkimi predyktorami
- Zatem jeśli F-statystyka daje wartość bliską 1, to nie możemy odrzucić hipotezy zerowej. Natomiast, gdy F-statystyka daje wartość istotnie większą od 1, to możemy H_0 odrzucić.

Ocena dokładności modelu – odchylenie standardowe składnika resztowego

$$RSE = \hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}}$$

RSE mierzy poziom niedopasowania modelem regresji do danych. Małe RSE oznacza, że model liniowy dobrze pasuje do danych.

Ocena dokładności modelu – statystyka R^2

Współczynnik determinacji:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

gdzie

- $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, całkowita suma kwadratów, proporcjonalna do **całkowitej wariancji** wartości zmiennej objaśnianej **y**
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ resztowa suma kwadratów, proporcjonalna do wariancji, która **pozostała niewyjaśniona** po wykonaniu operacji regresji.

Zatem statystyka R^2 mierzy proporcję zmienności y, którą można wyjaśnić używając modelu $\hat{\mathbf{y}}$.

Ocena dokładności modelu – statystyka R^2

Dla modelu liniowego

- z wyrazem wolnym (β_0)
- z parametrami estymowanymi metodą najmniejszych kwadratów

Zachodzi

- Wartości otrzymane ze statystyki R^2 mieszczą się w przedziale [0,1].
- Wartość statystyki R^2 bliska 1 sugeruje, że duża część zmienności zmiennej Y zostaje wyjaśniona operacją regresji.
- R^2 zadane jest korelacją Pearsona

$$R^2 = r^2 = Cor(\mathbf{y}, \hat{\mathbf{y}})^2$$

a dla prostej regresji z jednym predyktorem $\mathbf{x}$

$$R^2 = r^2 = Cor(\mathbf{y}, \mathbf{x})^2.$$

Przykład cd: model *sales* w zależności od *TV*

Miara	Wartość
RSE	3.26
R^2	0.612
F	312.1

Oznacza to, że

- Ponieważ $RSE = 3.26$, to prawdziwe dane o sprzedaży w każdym sklepie mogą się różnić od prawdziwych średnio o 3260 jednostek.
- Ponieważ średnia wysokość sprzedaży ze wszystkich 200 sklepów wynosi 14000, to błąd procentowy wynosi $3260/14000 = 23\%$.
- 0.612 wariancji w *sales* wytłumaczone jest przez regresję na *TV*.

Zależność *sales* od pozostałych predyktorów

	Współczynnik	Std. błąd	t-statystyka	p-wartość
Intercept	$\hat{\beta}_0 = 9.312$	0.563	16.54	< 0.0001
radio	$\hat{\beta}_1 = 0.203$	0.020	9.92	< 0.0001

	Współczynnik	Std. błąd	t-statystyka	p-wartość
Intercept	$\hat{\beta}_0 = 12.351$	0.621	19.88	< 0.0001
newspaper	$\hat{\beta}_1 = 0.055$	0.017	3.30	< 0.0001

Estymacja parametrów dla regresji wielokrotnej

	Współczynnik	Std. błąd	t-statystyka	p-wartość
Intercept	$\hat{\beta}_0 = 2.939$	0.3119	9.42	< 0.0001
TV	$\hat{\beta}_1 = 0.046$	0.0014	32.81	< 0.0001
radio	$\hat{\beta}_2 = 0.189$	0.0086	21.89	< 0.0001
newspaper	$\hat{\beta}_3 = -0.001$	0.0059	-0.18	0.8599

- wpływ *newspaper* na sprzedaż jest nieistotny, gdy rozważane są wszystkie trzy predyktory
- regresja liniowa dla każdego predyktora z osobna pokazuje, że wpływ tego predyktora na sprzedaż jest istotna
- istnieje duża korelacja (0.3541) pomiędzy zmiennymi *radio* i *newspaper*
- przeprowadzając regresję liniową dla tylko jednego predyktora *newspaper* wykrywamy wpływ na wysokość sprzedaży pochodzący od niewidocznego predyktora jakim jest *radio*

Przykład cd: dyskusja dopasowania modelu

- Dla pełnego modelu F-statystyka=570.
- p-wartość jest praktycznie równa 0, co oznacza, że co najmniej jeden predyktor ma wpływ na wysokość sprzedaży.

Predyktory w modelu	R^2
TV, *radio*, *newspaper*	0.8972
TV, *radio*	0.89719
TV	0.61

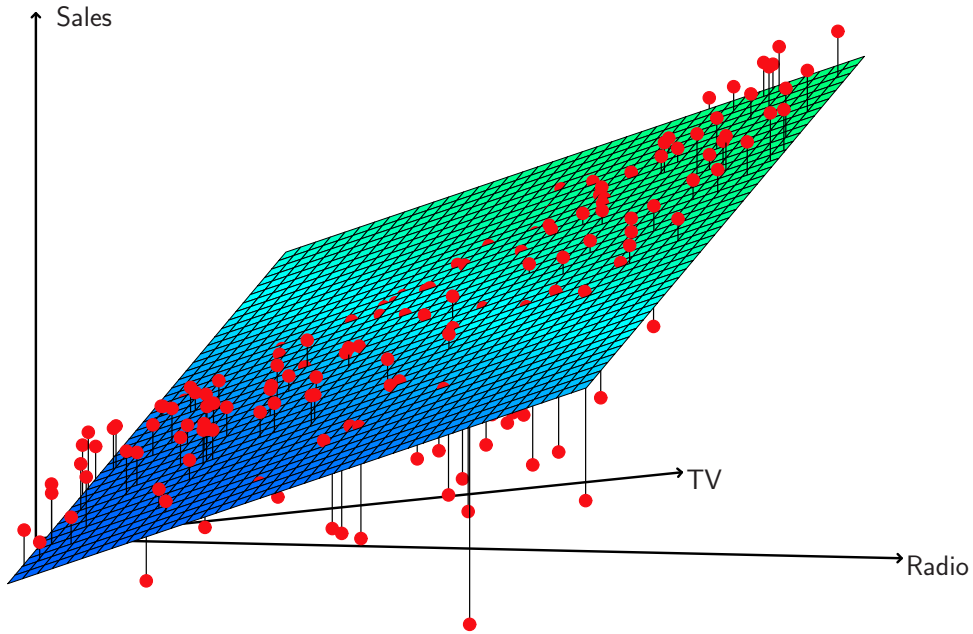
- R^2 na danych treningowych wzrasta wraz z liczbą predyktorów
- Mały wzrost i p-wartość w teście t sugerują, że *newspaper* może zostać usunięty z modelu.
- Zatrzymanie go może powodować problemy z przeuczeniem.
- Wg wzrostu R^2 i p-wartości dodanie predyktora *radio* istotnie poprawia dopasowanie.

Przykład cd: dyskusja dopasowania modelu

Predyktory w modelu	*RSE*
TV, *radio*, *newspaper*	1.686
TV, *radio*	1.681
TV	3.26

- Model oparty na wszystkich predyktorach ma większy błąd RSE niż model oparty tylko na *TV* i *radio* $\rightarrow$ Nie warto mieć *newspaper* w modelu
- Duży błąd modelu opartego tylko na *TV* $\rightarrow$ Warto dodać *radio* do modelu
- Dodanie predyktora do modelu powoduje zmniejszenie RSS.
- Ponieważ $RSE = \sqrt{RSS/(n-p-1)}$, dodanie jednego predyktora powoduje, że maleją zarówno licznik, jak i mianownik. Zwiększenie wartości *RSE* jest możliwe, jeśli spadek wartości *RSS* jest dostatecznie mały.

Przykład cd: Wizualizacja dopasowania



- Model przeszacowuje sprzedaż w sytuacjach gdy wiekszość nakładów jest głównie skierowana na reklamę w TV lub głównie w radiu. **Nieliniowość w danych.**
- Model niedoszacowuje sprzedaż w sytuacjach, gdy reklama jest dzielona pomiędzy TV i radio. **Efekt synergii**.

Które predyktory mają istotny wpływ na zmienną objaśnianą?

Problem wyboru zmiennych.

- Przeszukiwanie wyczerpujące (ang. exhaustive search) dla wielu zmiennych objaśniających niewykonalne.

Podstawowe techniki

- Wyszukiwanie zachłanne (ang. forward search)
- Wyszukiwanie wsteczne (ang. backward search)
- Wyszukiwanie mieszane

Uwaga: to wszystko są heurystyki i mają swoje problemy.

Wyszukiwanie zachłanne

- Zaczynamy z pustym zbiorem predyktorów - model zawiera tylko wyraz wolny i przewiduje średnią wartość **y**.
- Dla każdego z p predyktorów wykonujemy liniową regresję w oparciu o ten jeden predyktor. Wybieramy ten predyktor, dla którego resztowa suma kwadratów (RSS) jest najmniejsza.
- Następnie do niego dobieramy drugi predyktor tak, aby RSS dla modelu z dwoma predyktorami było najmniejsze.
- Powtarzamy tę procedurę aż do momentu spełnienia jakiegoś warunku stopu. (Np. RSE przestaje maleć)

Wyszukiwanie wsteczne

- Zaczynamy budowę modelu z wszystkimi predyktorami i usuwamy tę zmienną, dla której p-wartość jest największa.
- Budujemy model dla tak zmniejszonego zbioru zmiennych i usuwamy tę zmienną, dla której p-wartość jest największa.
- Powtarzamy tę procedurę aż do momentu spełnienia jakiegoś warunku stopu, np. wszystkie p wartości pozostających w grze predyktorów są poniżej pewnego progu.
- Nie można zastosować gdy liczba predyktorów p większa od liczby obserwacji n (Dlaczego?)

Wyszukiwanie mieszane

Kombinacja zachłannego wyszukiwania z wstecznym.

- Startujemy z pustym zbiorem predyktorów.
- Dodajemy zmienne jak w podejściu zachłannym.
- W procesie dodawania zmiennych p-wartości dla pewnych zmiennych mogą wzrosnąć.
- Na każdym kroku sprawdzamy te p-wartości. Jeśli dla pewnego predyktora, jego p-wartość wzrosła powyżej pewnego progu, to usuwamy ten predyktor z modelu.
- Kontynuujemy tę procedurę do momentu aż wszystkie predyktory w modelu mają p-wartości poniżej pewnego progu, a wszystkie predyktory poza modelem otrzymałyby p-wartość powyżej progu, w przypadku dodania tego predyktora do modelu.

Literatura

- G. James, D. Witten, T. Hastie and R. Tibshirani *An Introduction to Statistical Learning, with applications in R*, **Springer Verlag**, 2015. `https://www.statlearning.com/`
- T. Hastie, R. Tibshirani and J. Friedman *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, **Springer Verlag**, Second Edition, 2013. `https://hastie.su.domains/ElemStatLearn/`

Podsumowanie

- Model regresji liniowej
- Założenia modelu regresji liniowej
- Metoda najmnijeszych kwadratów
- Twierdzenie Gaussa-Markowa
- Własności estymatora wektora parametrów β modelu regresji liniowej
- Testowanie istotnosci danego predyktora
- Przedział ufnosci dla estymatora $\hat{\beta}_i$
- Testowanie istotnosci kilku predyktorów
- Overall F test
- odchylenie standardowe składnika resztowego
- Statystyka R^2
- Algorytmy wyboru zmiennych dla regresji linowej