Regresja liniowa w R

Piotr J. Sobczyk

Uwaga Poniższe notatki mają charakter roboczy. Mogą zawierać błędy. Za przesłanie mi informacji zwrotnej o zauważonych usterkach serdecznie dziękuję.

Weźmy dane dotyczące wzrostu małżeństw

```
library(ggplot2)
library(PBImisc)
data(heights)
```

Od czego zaczynamy analizę danych? Nie ma jednej odpowiedzi. Warto posłuchać 11 odcinka podcastu Not so standard deviations, żeby zapoznać się z różnymi strategiami.

Pierwszą rzeczą może być podglądnięcie danych. Zobaczymy jakie typy zmiennych mamy, ile jest obserwacji itp.

```
str(heights) #od structure
```

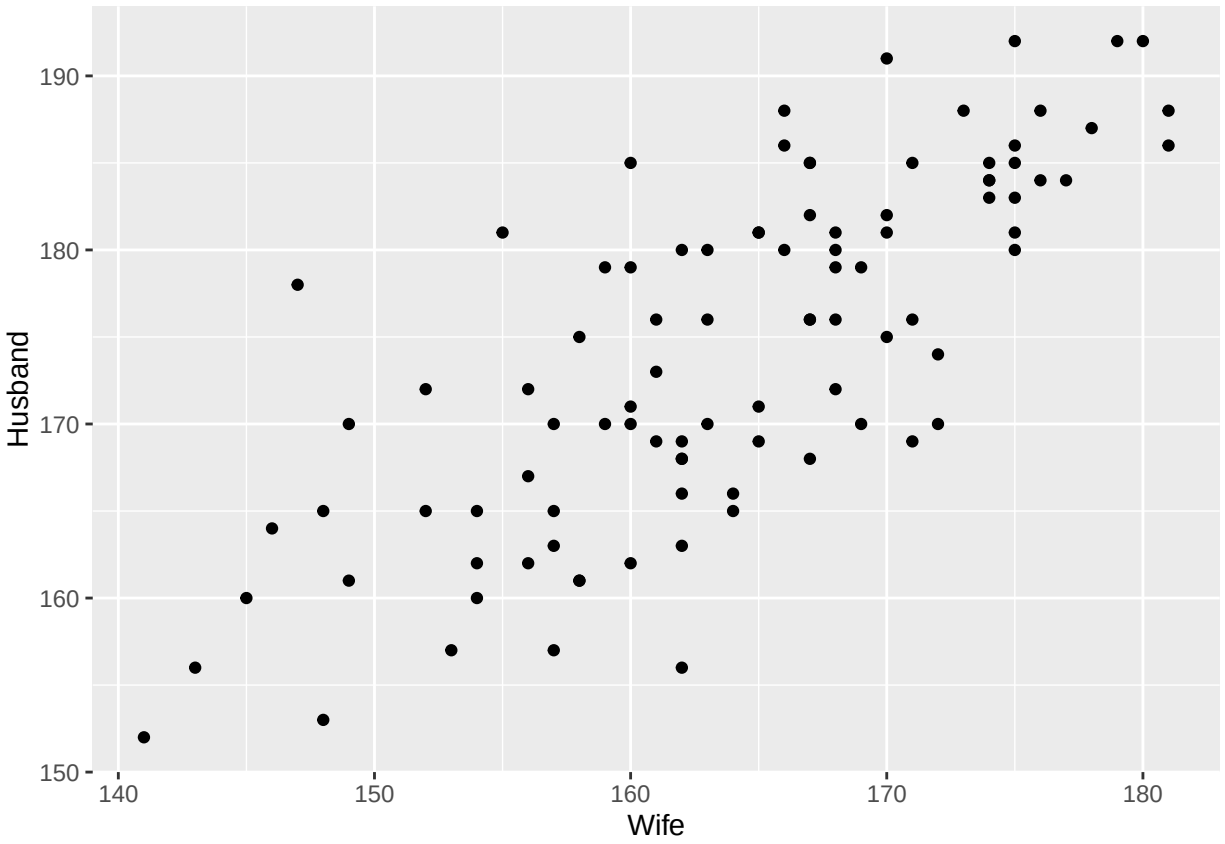
```
## 'data.frame':    96 obs. of  2 variables:
##  $ Husband: int  186 180 160 186 163 172 192 170 174 191 ...
##  $ Wife   : int  175 168 154 166 162 152 179 163 172 170 ...
```

```
summary(heights)
```

```
##     Husband           Wife      
##  Min.   :152.0   Min.   :141.0  
##  1st Qu.:166.8   1st Qu.:158.0  
##  Median :175.5   Median :164.5  
##  Mean   :174.3   Mean   :163.9  
##  3rd Qu.:182.2   3rd Qu.:170.2  
##  Max.   :192.0   Max.   :181.0
```

Kiedy już wiemy co, pod względem czysto technicznym, zawierają dane, możemy przejść do analizy eksploracyjnej. Można liczyć więcej statystyk opisowych, ale dobrą praktyką jest wizualizacja danych, wykres, dzięki któremu zobaczmy co w nich siedzi

```
ggplot(heights, mapping = aes(x=Wife, y=Husband)) +
  geom_point()
```



Ha! Coś już wiemy! Widzimy, że wzrost męża i wzrost żony są ze sobą powiązane. To znaczy ludzie wysocy łączą się w pary z wysokimi a niscy z niższymi. Jak zapiszemy to w języku modelu liniowego?

$$\text{wzrost męża} = \beta_0 + \beta_1 * \text{wzrost żony} + b\text{łąd}$$

Dlaczego β_0?

Przypomnijmy sobie pytanie, jakie nas interesują w związku z modelem. Chcielibyśmy poznać β, σ i potrafić oceniać na podstawie wzrostu żony, jak wysoki jest jej mąż.

Ogólny model liniowy ma postać

$$y = X\beta + \epsilon$$

gdzie $X \in M_{n\times p}$ - deterministyczna macierz plany. $\epsilon \sim N(0, \sigma^2 I)$.

Jak wyglądał nasz estymator ML?

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

Obliczmy to w R

```
X=cbind(1, heights$Wife)
y=heights$Husband
beta=solve(t(X)%*%X)%*%t(X)%*%y
beta
```

```
##            [,1]
## [1,] 37.8100491
## [2,]  0.8329246
```

Czy dostaniemy to samo używają standardowej funkcji eRowej?

```
lm.fit(x = X, y = y) -> lm.husband
lm.husband$coefficients
```

```
##         x1         x2
## 37.8100491  0.8329246
```

Alternatywny, i wygodniejszy sposób konstrukcji modelu liniowego w R

```
lm.husband=lm(Husband~Wife, data=heights)
summary(lm.husband)
```

```
##
## Call:
## lm(formula = Husband ~ Wife, data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7438  -4.2838  -0.1615   4.2562  17.7500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.81005   11.93231   3.169  0.00207 **
## Wife         0.83292    0.07269  11.458  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.468 on 94 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5783
## F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

Na dzisiejszych zajęciach zrozumiemy całość zwracanego podsumowania.

Residuals

To już było. Co to są residua (nie mylić z analizą zespoloną!)

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

Na marginesie, zauważmy, że $\hat{y}$ jest liniową funkcją y:

$$\hat{y} = \underbrace{X(X^T X)^{-1} X^T}_{H} y = Hy$$

H to hat matrix, „daszkuje" y.

Sprawdźmy jakie czy nasze residua ze wzoru zgadzają się z R

```
husband_residuals=y-X%*%solve(t(X)%*%X)%*%t(X)%*%y
summary(husband_residuals)
```

```
##        V1
##  Min.   :-16.7438
##  1st Qu.: -4.2838
##  Median : -0.1615
##  Mean   :  0.0000
```

```
##  3rd Qu.:  4.2562
##  Max.   : 17.7500
```

Kolejna uwaga na marginesie. Przypomnienie z algebry liniowej. Co to jest macierz $X(X^TX)^{-1}X^T$? Jak w związku z tym można interpretować nasze oszacowania? Co w związku z tym wiemy o własnościach macierzy H?

t-value

Zauważmy, że nasz estymator $\hat{\beta}$ jest zmienną losową

$$\hat{\beta} = (X^TX)^{-1}X^Ty \sim (X^TX)^{-1}X^T \cdot \mathcal{N}(X\beta, \sigma^2 I)$$

Co dalej?

$$\begin{aligned}
\hat{\beta} &\sim \mathcal{N}((X^TX)^{-1}X^TX\beta, (X^TX)^{-1}X^T\sigma^2 I((X^TX)^{-1}X^T)^T) && (1)\\
&\sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}) && (2)\\
&\sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}) && (3)
\end{aligned}$$

Jakie płyną stąd wnioski?

1. Estymatory poszczególnych elementów β są skorelowane. Kiedy są nieskorelowane?
2. Jaki jest rozkład pojedynczego β_i?
3. Czego potrzebujemy, aby policzyć wariancję estymatora?

Estymator wariancji w modelu liniowym

$$\hat{\sigma^2} = \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{n - p}$$

Z tw. Fishera, jest on niezależny od $\hat{beta}$. Przejdźmy przez szybki rachunek.

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2(X^TX)^{-1}_{ii})$$

a zatem

$$\frac{\hat{\beta}_i - \beta_i}{\sigma^2(X^TX)^{-1}_{ii}} \sim \mathcal{N}(0,1)$$

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma^2}(X^TX)^{-1}_{ii}} \sim t(n-2)$$

```
beta_hat=solve(t(X)%*%X)%*%t(X)%*%y
RSS=sum( (y-X%*%beta_hat)^2 )
sigma_hat=sqrt(RSS/(nrow(heights)-2))
sigma_hat
```

```
## [1] 6.468001
```

```
sqrt(diag(solve(t(X)%*%X)))*sigma_hat
```

```
## [1] 11.93230880  0.07269272
```

Możemy też policzyć statystykę testową i p-wartości

```
test_stat=beta_hat/(sqrt(diag(solve(t(X)%*%X)))*sigma_hat)
test_stat
```

```
##           [,1]
## [1,]  3.168712
## [2,] 11.458156
```

```
2*(1-pt(test_stat, df = nrow(heights)-2))
```

```
##              [,1]
## [1,] 0.002066323
## [2,] 0.000000000
```

F-statistic i modele zagnieżdżone

Kiedy budujemy model liniowy, powinniśmy zadać sobie pytanie czy ma to jakikolwiek sens. To znaczy, czy zmienne objasniające dają jakąś istotną informację o zmiennej objaśnianej. Robimy to porównując nasz model z tzw modelem zerowym, czyli takim, który ma jedynie Intercept.

Porównanie jest robione w oparciu i RSS z obu modeli.

```
lm(Husband~1, data=heights) -> lm.husband0
anova(lm.husband0, lm.husband)
```

```
## Analysis of Variance Table
##
## Model 1: Husband ~ 1
## Model 2: Husband ~ Wife
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     95 9425.0
## 2     94 3932.5  1    5492.5 131.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2

Współczynnik R^2 mówi o tym, jak procent zmienności danych (czyli wariancji zmiennej objaśnianej) jest wyjaśniany w modelu. Wiemy, że nigdy nie będzie to 100% procent, ponieważ nasz model zawiera element losowy ϵ.

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - RSS(\hat{\beta})}{\sum_i (y_i - \bar{y})^2}$$

```
1-RSS/var(heights$Husband)/(nrow(heights)-1)
```

```
## [1] 0.5827588
```

Korzystając z R^2, można ze sobą porównać dwa konkurecyjne modele. Zależy nam, aby R^2 było jak największe.

Co zatem robi adjusted R^2. Jak zachowuje się RSS gdy zwiększamy liczbę zmiennych w modelu. Czy to jest pożądane zachowanie?

```
1-RSS/(nrow(heights)-2)/var(heights$Husband)
```

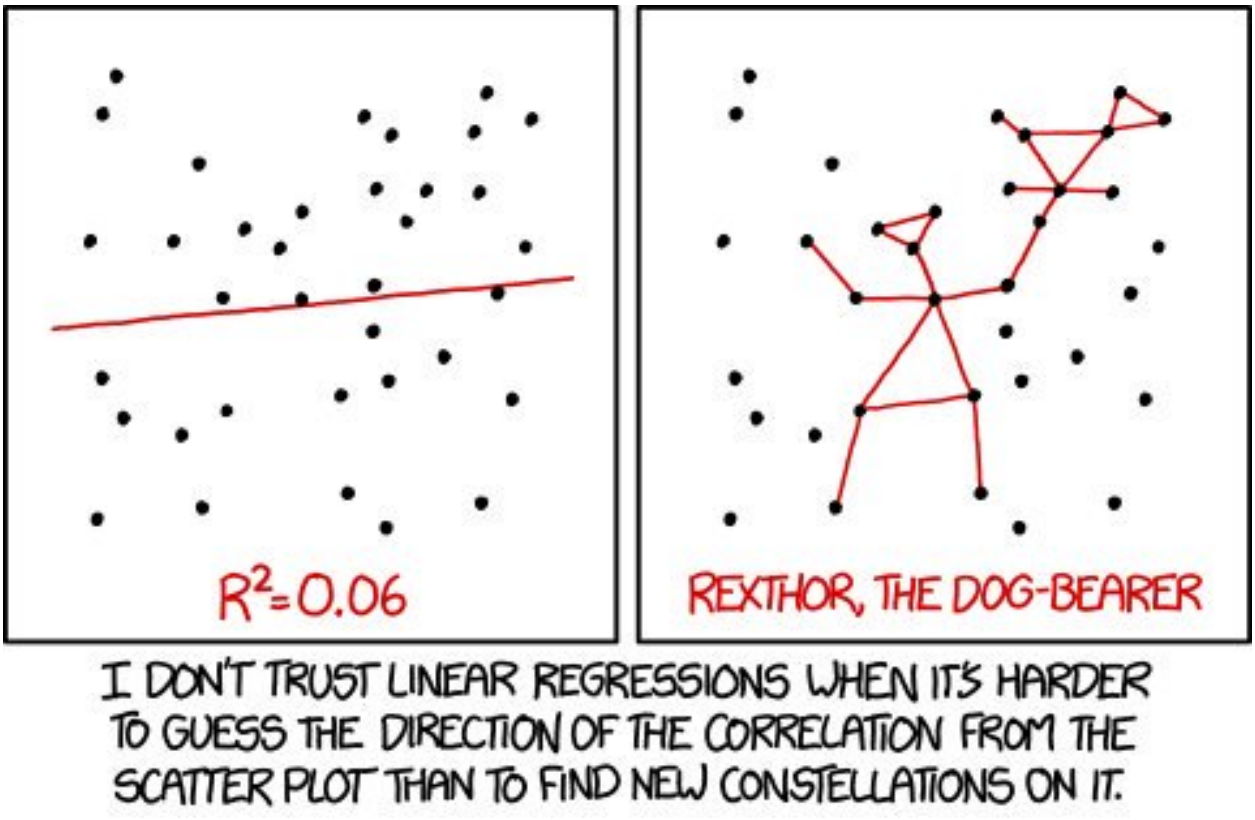


Figure 1: Dlaczego warto robić wykresy? źródło: https://pbs.twimg.com/media/Cqyf1RcXgAIlsSP.jpg

```
## [1] 0.5783201
```

$$\text{adjusted } R^2 = \frac{var(y) - RSS(\hat{\beta})/(n-p)}{var(y)}$$

Zerknijmy jeszcze raz na podsumowanie regresji liniowej

```
summary(lm.husband)
```

```
##
## Call:
## lm(formula = Husband ~ Wife, data = heights)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7438  -4.2838  -0.1615   4.2562  17.7500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.81005   11.93231   3.169  0.00207 **
## Wife         0.83292    0.07269  11.458  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.468 on 94 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5783
## F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

Rozumiemy już wszystkie zawarte w nim informacje i może przejść dalej, tzn. do diagnozowania i oceny stworzonego modelu liniowego.