

Regresja liniowa - ciąg dalszy

W tej części kursu zajmiemy się kilkoma bardziej zaawansowanymi aspektami prostej regresji liniowej z jednym predyktorem.

Błąd standardowy estymacji - odchylenie standardowe reszt. Wielkość ta informuje o przeciętnej wielkości odchylen empirycznych wartości zmiennej zależnej od wartości wyliczonych z modelu (teoretycznych). Parametr ten jest ważny w analizie regresji, gdyż stanowi miarę rozproszenia elementów populacji wokół linii regresji. Na jego podstawie możemy ocenić stopień “dopasowania” modelu do danych empirycznych.

Dokładność oszacowania regresji można ocenić za pomocą współczynnika ϕ^2 . Punktem wyjścia jest wariancja Y . Możemy pokazać, że:

$$S^2(Y) = S^2(\hat{Y}) + S^2(U)$$

Równość ta (*równość wariancyjna*) głosi, że całkowity obszar zmienności zmiennej zależnej jest sumą zmienności wyjaśnianej regresją i zmienności resztowej (niewyjaśnianej przez regresję).

Współczynnikiem zbieżności ϕ^2 lub **indeterminacji** nazywamy stosunek tej części zmienności badanego zjawiska, która nie jest wyjaśniana przez zmiany zmiennych objaśniających w funkcji regresji do całkowitej zmienności zmiennej objaśnianej.

$$\phi^2 = \frac{S^2(U)}{S^2(Y)}$$

Współczynnik determinacji R^2 mierzy, jaka część ogólnej zmienności zmiennej zależnej jest wyjaśniana przez regresję liniową. Obliczamy go według wzoru:

$$R^2 = \frac{S^2(\hat{Y})}{S^2(Y)} = 1 - \phi^2$$

Interpretacja współczynnika R^2

- Im większe R^2 tym lepiej. Musimy jednak pamiętać, że dołączenie nowej zmiennej do modelu zawsze spowoduje zwiększenie jego wartości. Dlatego w przypadku kilku predyktorów lepiej używać poprawionego R^2 , który uwzględnia, że R^2 obliczony jest z próby i jest “zbyt dobrze” dopasowany.

$$R_{popr}^2 = R^2 - \frac{1}{n-2}(1 - R^2)$$

Założenia regresji

- **homogeniczność wariancji w grupach** - założenie że wariancja Y dla każdej wartości X jest stała w populacji
- **normalność w grupach** - założenie, że wartości Y dla każdej wartości X mają rozkład normalny
- (takie meta-założenie) **relacja między X a Y jest relacją, którą oddaje funkcja liniowa**

Narzędzia diagnostyczne:

- **wykres przedstawiający na jednej osi wartości dopasowane przez model, a na drugiej residua (reszty) lub standaryzowane residua**
 - dla modelu adekwatnego średnia wartość resuduum nie powinna zależeć od wartości dopasowania (powinniśmy w wyniku dostać pas punktów losowo rozmieszczonych wokół prostej $y = 0$)
 - jeżeli można dostrzec jakiś wzór w ułożeniu punktów, może to oznaczać, że relacja między zmiennymi nie jest liniowa
 - w szczególności niepokoją nas regularne krzywe
- **wykresy kwantylowe dla standaryzowanych residuów**
 - powinny wskazywać na normalność
 - nie przejmujemy się za bardzo niewielkimi odchyleniami przy wartościach skrajnych
 - przejmujemy się, jeżeli wykres kwantyl-quantyl reszt wskazuje na wyraźną skośność tzn. nienormalność
- **wykres, na którym dla każdej wartości zmiennej objaśniającej wyznaczono pierwiastek z wartości bezwzględnej jego residuum standaryzowanego**
 - nie powinniśmy zaobserwować żadnego trendu, jeśli takowy występuje oznacza to, że wariancja błędu nie jest stała
 - powinniśmy zaobserwować równomiernie rozłożone punkty i poziomą linię przebiegającą przez nie
- **wykres dźwigni**
 - nie ma znaczenia wzór - interesują nas obserwacje odstające

- przerywane linie reprezentują dystans Cooka - nie chcemy, żeby nasze obserwacje odstające się tam znalazły, bo to by znaczyło, że miały duży wpływ na naszą regresję
- wyniki regresji mogą znacznie się zmienić, jeśli wykluczymy te obserwacje - czasami warto to zrobić

- **wykres pudełkowy/skrzypcowy Y warunkowanej na X**

- można użyć, jeśli X ma tylko kilka poziomów
- pudełka powinny wyglądać w miarę podobnie

Więcej o wykresach diagnostycznych można przeczytać tutaj w języku angielskim i w książkach też:

<http://data.library.virginia.edu/diagnostic-plots/>

Przykład

Większość z nas potrafi całkiem nieźle sobie poradzić z egzaminem, jeżeli wcześniej uczyliśmy się materiału. Jak dobrze jednak poradziłibyśmy sobie, gdybyśmy przystąpili do egzaminu nawet nie patrząc wcześniej na materiał?

Katz, Lautenschlager, Blackburn i Harris (1990) postanowili odpowiedzieć na to pytanie i poprosili studentów o zapoznanie się z krótkim fragmentem tekstu a następnie o odpowiedzenie na dotyczące go pytania wielokrotnego wyboru. W drugiej grupie studentów autorzy również poprosili o odpowiedź na te pytanie, uczestnikom nie dali jednak wcześniej przeczytania rzeczonoego fragmentu. Pytania na teście były bardzo podobne do tych, z którymi stykają się uczniowie w USA na teście SAT. Wyniki uzyskane przez badaczy doprowadziły ich do wniosku, że uczniowie, którzy dobrze poradzi sobie na SAT poradzą sobie równie dobrze na tym teście, ponieważ oba angażują podstawowe umiejętności rozwiązywania testów, takie jak eliminacja nieprawdopodobnych odpowiedzi itp.

Nas będą interesowały odpowiedzi tylko tych studentów, którzy nie czytali fragmentu tekstu. Zmienną objaśnianą będzie wynik na SAT, zmienną objaśniającą zaś - wynik na teście, który badacze dali studentom.

Na początek przyjrzyjmy się naszym danym - wczytajmy je do R, zobaczmy pierwsze 6 wierszy oraz obliczymy podstawowe statystyki deskryptywne.

```
df <- read.table('Tab9-6.dat', header = TRUE)
head(df)
```

	ID	Score	SATV
1	1	58	590
2	2	48	580
3	3	34	550

4	4	38	550
5	5	41	560
6	6	55	800

```
summary(df)
```

ID	Score	SATV
Min. : 1.00	Min. :33.00	Min. :490.0
1st Qu.: 7.75	1st Qu.:41.00	1st Qu.:560.0
Median :14.50	Median :46.50	Median :590.0
Mean :14.50	Mean :46.21	Mean :598.6
3rd Qu.:21.25	3rd Qu.:50.75	3rd Qu.:620.0
Max. :28.00	Max. :60.00	Max. :800.0

Następnie dopasujemy model liniowy do naszych danych:

```
fit <- lm(SATV ~ Score, data= df)
summary(fit)
```

Call:

```
lm(formula = SATV ~ Score, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-87.529	-29.285	-3.204	15.443	158.686

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	373.736	70.938	5.269	1.66e-05 ***
Score	4.865	1.520	3.202	0.00359 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

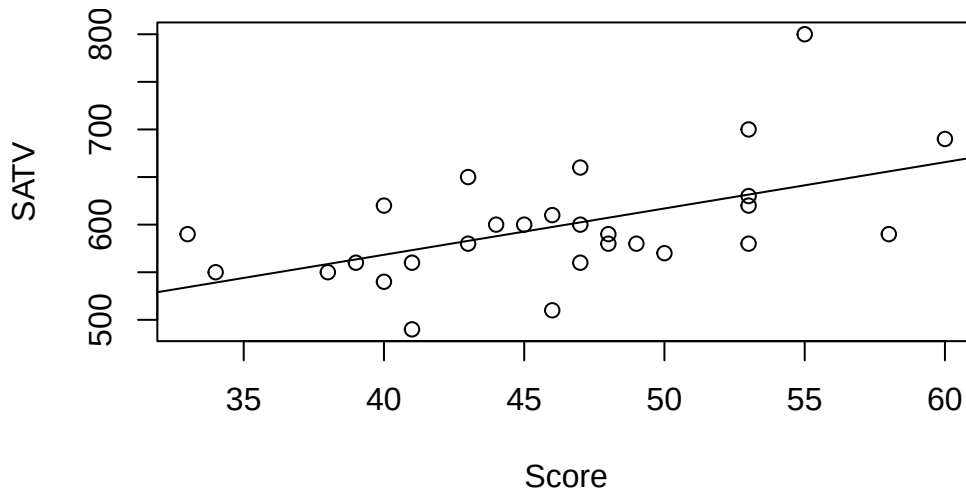
Residual standard error: 53.13 on 26 degrees of freedom

Multiple R-squared: 0.2828, Adjusted R-squared: 0.2552

F-statistic: 10.25 on 1 and 26 DF, p-value: 0.003588

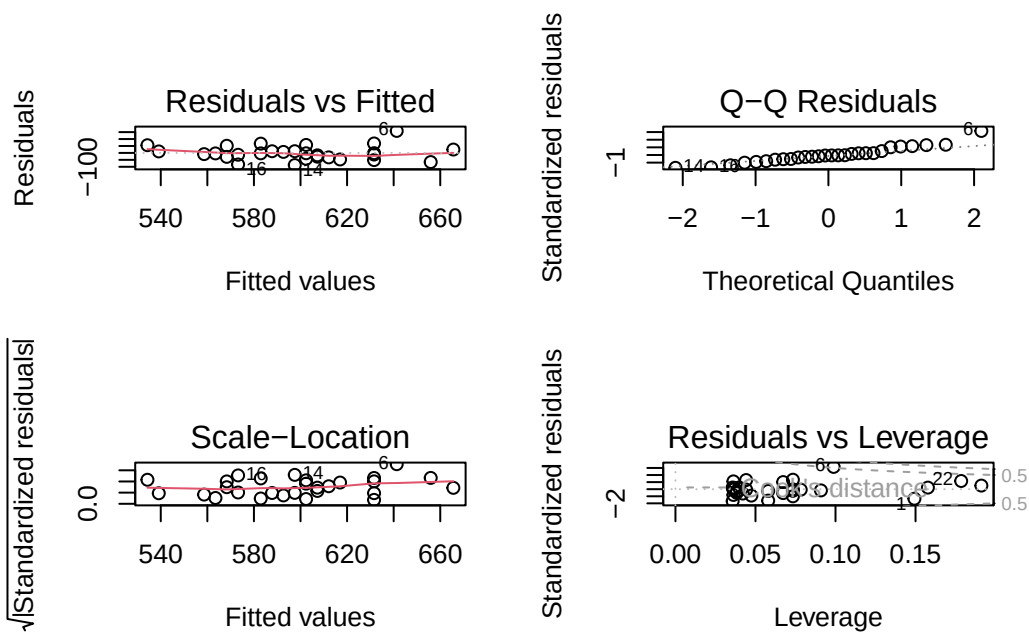
Zobaczmy, jak wygląda wykres punktowy (wykres rozrzutu) dla zmiennej objaśniającej i objaśnianej oraz dorysujmy linię regresji.

```
plot(SATV ~ Score, data = df)
abline(fit)
```



Teraz użyć możemy naszych narzędzi diagnostycznych, które omówiliśmy wcześniej i zobaczyć, czy założenia regresji liniowej nie zostały poważnie naruszone.

```
par(mfrow = c(2,2))
plot(fit)
```



Co możemy stwierdzić na podstawie stworzonych przez nas wykresów diagnostycznych?

Predykcja

Modelu liniowego możemy użyć do wyprowadzania z danych przewidywań. Na początek spróbujmy obliczyć przewidywane wartości zmiennej objaśnianej ($\hat{Y}$) dla każdej naszych wartości zmiennej objaśniającej (X). R słusznie przypomni nam, że wartości te dotyczą przyszłych wartości.

```
predict(fit, interval = "predict")
```

```
Warning in predict.lm(fit, interval = "predict"): predictions on current data  
  ↪ refer to _future_ responses
```

	fit	lwr	upr
1	655.9096	538.8214	772.9977
2	607.2590	495.9682	718.5498
3	539.1483	421.6324	656.6641
4	558.6085	444.5348	672.6821
5	573.2036	460.8659	685.5414
6	641.3144	526.8261	755.8027
7	582.9338	471.3303	694.5372
8	602.3940	491.2159	713.5720
9	602.3940	491.2159	713.5720
10	597.5289	486.3760	708.6819
11	568.3386	455.5057	681.1715
12	563.4735	450.0616	676.8855
13	616.9891	505.2110	728.7672
14	597.5289	486.3760	708.6819
15	607.2590	495.9682	718.5498
16	573.2036	460.8659	685.5414
17	582.9338	471.3303	694.5372
18	631.5843	518.4307	744.7379
19	665.6397	546.4400	784.8394
20	587.7988	476.4329	699.1647
21	612.1241	500.6331	723.6151
22	534.2832	415.7165	652.8499
23	568.3386	455.5057	681.1715
24	631.5843	518.4307	744.7379
25	592.6639	481.4482	703.8795
26	602.3940	491.2159	713.5720
27	631.5843	518.4307	744.7379
28	631.5843	518.4307	744.7379

Za pomocą tej samej funkcji możemy również dowiedzieć się, jakie przewidywane wartości daje nasz model dowolnej innej wartości predyktora.

```
predict(fit, data.frame(Score = 60), interval = "predict")
```

```
      fit      lwr      upr
1 665.6397 546.44 784.8394
```

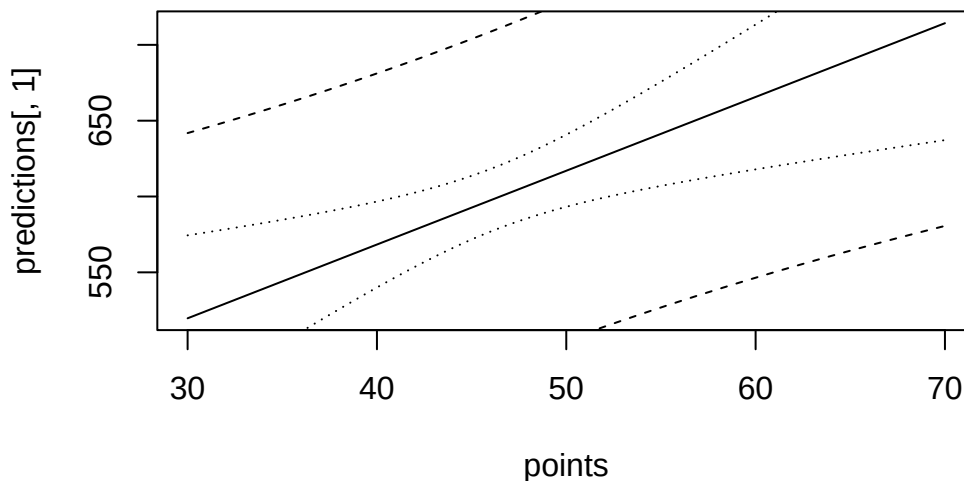
W regresji liniowej warto rozróżnić dwa rodzaje przedziałów. Przedział ufności dla średniej (`interval = "confidence"`) mówi o niepewności oszacowania średniej wartości Y dla danego X. Przedział predykcji (`interval = "prediction"`) dotyczy wyniku pojedynczej nowej obserwacji przy tym samym X. Jest szerszy, bo uwzględnia dodatkowo rozrzut pojedynczych wyników wokół średniej.

O różnicy między tymi dwoma przedziałami więcej można przeczytać np. tu:

<http://www.ma.utexas.edu/users/mks/statmistakes/CIsvsPI.html>.

Poniżej znajduje się wykres ilustrujący różnicę, między tymi dwoma pojęciami. Szerszy przedział to przedział dla predykcji, węższy - dla średniej.

```
points <- seq(30, 70, length.out = 1000)
predictions <- predict(fit, data.frame(Score = points), interval = "predict")
plot(points, predictions[,1], type = 'l')
points(points, predictions[,2], type = 'l', lty = 2)
points(points, predictions[,3], type = 'l', lty = 2)
predictions = predict(fit, data.frame(Score = points), interval =
  ↪ "confidence")
points(points, predictions[,2], type = 'l', lty = 3)
points(points, predictions[,3], type = 'l', lty = 3)
```



To samo możemy zrobić za pomocą funkcji z pakietu `rms`. Bardzo polecam zapoznanie się w wolnym czasie z możliwościami tego pakietu - dużo różnych rodzajów regresji i przydatne funkcje.

```
library(rms)
dd <- datadist(df)
options(datadist = "dd")

model_rms <- ols(SATV ~ Score , data = df)
model_rms
```

Linear Regression Model

```
ols(formula = SATV ~ Score, data = df)
```

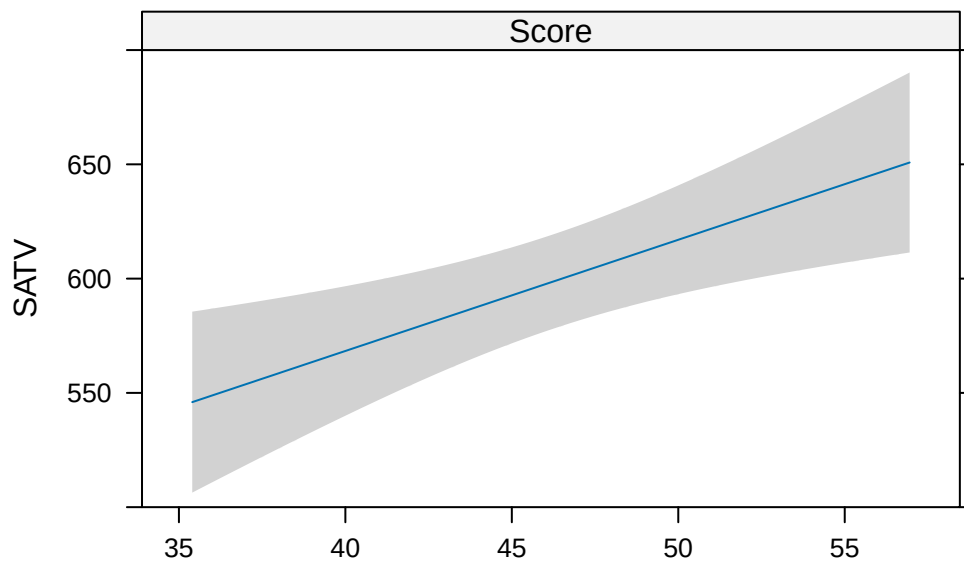
		Model Likelihood	Discrimination
		Ratio Test	Indexes
Obs	28	LR chi2	9.31
		R2	0.283
sigma	53.1336	d.f.	1
		R2 adj	0.255
d.f.	26	Pr(> chi2)	0.0023
		g	37.736

Residuals

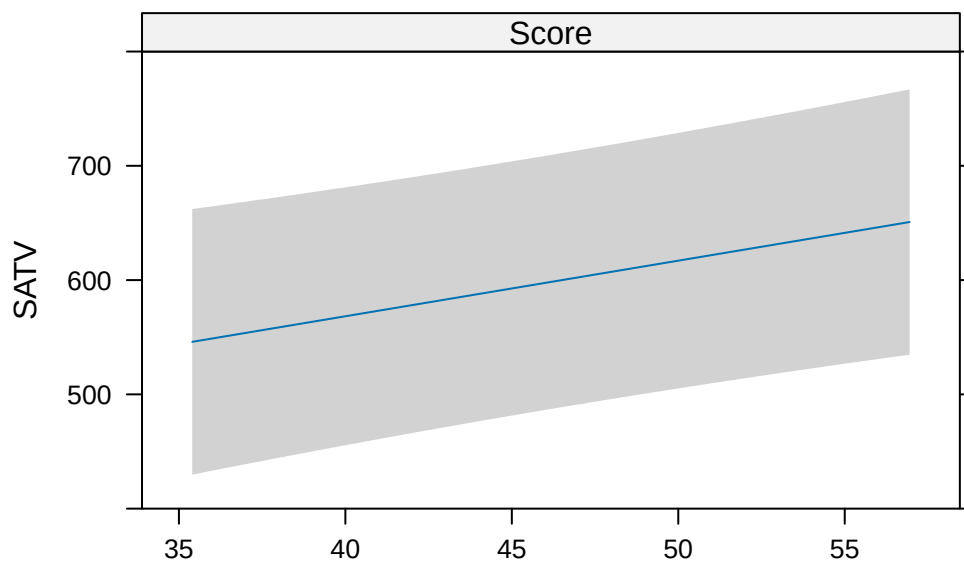
Min	1Q	Median	3Q	Max
-87.529	-29.285	-3.204	15.443	158.686

	Coef	S.E.	t	Pr(> t)
Intercept	373.7364	70.9379	5.27	<0.0001
Score	4.8651	1.5195	3.20	0.0036

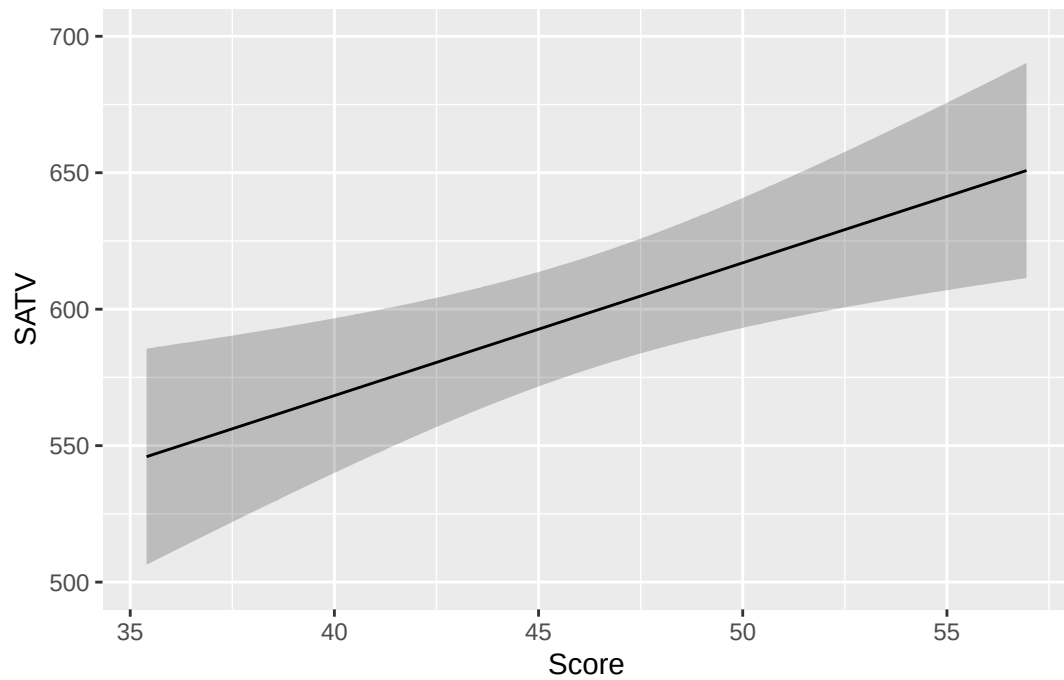
```
plot(Predict(model_rms))
```



```
plot(Predict(model_rms, conf.type = 'individual'))
```



```
ggplot(Predict(model_rms))
```



```
ggplot(Predict(model_rms, conf.type = 'individual'))
```

