Wielokrotna (wieloraka) regresja liniowa

Na dzisiejszych zajęciach opowiemy sobię trochę o wielokrotnej (wielorakiej) regresji liniowej (and. *multiple linear regression*). Pracować będziemy na przykładzie z rozdziału 15 podręcznika Howella, który znajduje się na stronie w materiałach dostępnych do pobrania po zalogowaniu. Proszę Państwa o przeczytanie tego rozdziału z podręcznika. Kod, który wklejam poniżej, ma stanowić bazę do ćwiczeń i ilustracje, jak różne rzeczy omówione w książce możemy zrobić w R. Kod będzie objaśniony, ale nie zastąpi (!) to lektury podręcznika. Dodatkowo pomyślałem, że możnaby przemycić trochę ciekawych i przydatnych R-owych bibliotek oraz technik, więc będę czasami korzystał z nieznanych Państwu funkcji, ale proszę się nie martwić – wszystkie będę objaśniał i tłumaczył co robią.

Na początek wczytajmy dwie biblioteki, których użyjemy do stworzenia tego pliku oraz dane znajdujące się w pliku `Tab15-1.dat`.

Biblioteka `pander` służy do ładnego wyświetlania wyników różnych testów statystycznych. Bez niej *output* różnych operacji, które wykonujemy, wygląda jak wydruk kodu źródłowego programu (szarawe tło oraz czcionka o stałej szerokości). Pander sprawia, że zamiast wydruku testy statystyczne zwracają ładnie sformatowane wyniki w formie np. tabelki. Z pakietu `knitr` będziemy używać funkcji `kable`, która w zasadzie służy do tego samego, czyli do ładnego wyświetlania danych w dokumentach HTML (takich jak ten). Od razu mogą Państwo zobaczyć działanie tego pakietu – dzięki funkcji `kable` nasza ramka danych jest estetycznie sformatowana.

Uwaga! Wczytaliśmy dane za pomocą funkcji `read.table`. Funkcja ta pozwala wczytać pliki, w których wartości oddzielone są białymi znakami (np. tabulatorami). Domyślnie jednak wczytuje pliki bez nagłówka (zakłada, że wszystkie wiersze to dane). My jednak ustawiliśmy argument `header` na `TRUE`, co sprawiło, że R potraktował pierwszy wiersz jako wiersz z nazwami kolumn.

```
library(pander)
library(knitr)
data <- read.table("Tab15-1.dat", header = T)
if (knitr::is_latex_output()) {
  pander(data, split.table = 70)
```

```
} else {
  kable(data)
}
```

Tabela 1: Table continues below

id	State	Expend	PTratio	Salary	PctSAT	Verbal	Math
1	Alabama	4.405	17.2	31.14	8	491	538
2	Alaska	8.963	17.6	47.95	47	445	489
3	Arizona	4.778	19.3	32.17	27	448	496
4	Ark	4.459	17.1	28.93	6	482	523
5	Calif	4.992	24	41.08	45	417	485
6	Col	5.443	18.4	34.57	29	462	518
7	Conn	8.817	14.4	50.05	81	431	477
8	Del	7.03	16.6	39.08	68	429	468
9	Florida	5.718	19.1	32.59	48	420	469
10	Georgia	5.193	16.3	32.29	65	406	448
11	Hawaii	6.078	17.9	38.52	57	407	482
12	Idaho	4.21	19.1	29.78	15	468	511
13	Illino	6.136	17.3	39.43	13	488	560
14	Indiana	5.826	17.5	36.78	58	415	467
15	Iowa	5.483	15.8	31.51	5	516	583
16	Kansas	5.817	15.1	34.65	9	503	557
17	Ken	5.217	17	32.26	11	477	522
18	Louisia	4.761	16.8	26.46	9	486	535
19	Maine	6.428	13.8	31.97	68	427	469
20	Marylan	7.245	17	40.66	64	430	479
21	Mass	7.287	14.8	40.8	80	430	477
22	Mich	6.994	20.1	41.9	11	484	549
23	Minn	6	17.5	35.95	9	506	579
24	Mississ	4.08	17.5	26.82	4	496	540
25	Missour	5.383	15.5	31.19	9	495	550
26	Montana	5.692	16.3	28.79	21	473	536
27	Neb	5.935	14.5	30.92	9	494	556
28	Nev	5.16	18.7	34.84	30	434	483
29	NH	5.859	15.6	34.72	70	444	491
30	NJ	9.774	13.8	46.09	70	420	478
31	NM	4.586	17.2	28.49	11	485	530
32	NY	9.623	15.2	47.61	74	419	473
33	NC	5.077	16.2	30.79	60	411	454
34	ND	4.775	15.3	26.33	5	515	592

id	State	Expend	PTratio	Salary	PctSAT	Verbal	Math
35	Ohio	6.162	16.6	36.8	23	460	515
36	Ok	4.845	15.5	28.17	9	491	536
37	Oregon	6.436	19.9	38.55	51	448	499
38	Penn	7.109	17.1	44.51	70	419	461
39	RI	7.469	14.7	40.73	70	425	463
40	SC	4.797	16.4	30.28	58	401	443
41	SD	4.775	14.4	25.99	5	505	563
42	Tenn	4.388	18.6	32.48	12	497	543
43	Texas	5.222	15.7	31.22	47	419	474
44	Utah	3.656	24.3	29.08	4	513	563
45	Vermont	6.75	13.8	35.41	68	429	472
46	Va	5.327	14.6	33.99	65	428	468
47	Wash	5.906	20.2	36.15	48	443	494
48	WV	6.107	14.8	31.94	17	448	484
49	Wisc	6.93	15.9	37.75	9	501	572
50	Wyoming	6.16	14.9	31.29	10	476	525

SATcombined	PctACT	ACTcombined	LogPctSAT
1029	61	20.2	2.079
934	32	21	3.85
944	27	21.1	3.296
1005	66	20.3	1.792
902	11	21	3.807
980	62	21.5	3.367
908	3	21.7	4.394
897	3	21	4.22
889	36	20.7	3.871
854	16	20.2	4.174
889	17	21.6	4.043
979	62	21.4	2.708
1048	69	21.2	2.565
882	19	21.2	4.06
1099	64	22.1	1.609
1060	74	21.7	2.197
999	65	20.1	2.398
1021	80	19.4	2.197
896	2	21.5	4.22
909	11	20.7	4.159
907	6	21.6	4.382

SATcombined	PctACT	ACTcombined	LogPctSAT
1033	68	21.3	2.398
1085	60	22.1	2.197
1036	79	18.7	1.386
1045	64	21.5	2.197
1009	55	21.9	3.045
1050	73	21.7	2.197
917	39	21.3	3.401
935	4	22.3	4.248
898	3	20.8	4.248
1015	59	20.3	2.398
892	16	21.9	4.304
865	11	19.3	4.094
1107	78	21.4	1.609
975	60	21.3	3.135
1027	66	20.6	2.197
947	12	22.3	3.932
880	8	21	4.248
888	2	21.4	4.248
844	13	18.9	4.06
1068	68	21.3	1.609
1040	83	19.7	2.485
893	30	20.2	3.85
1076	69	21.5	1.386
901	7	21.9	4.22
896	6	20.7	4.174
937	16	22.4	3.871
932	57	20	2.833
1073	64	22.3	2.197
1001	70	21.4	2.303

Wczytane przez nas dane dotyczą edukacji w Stanach Zjednoczonych i zostały zebrane przez Gubner (1999). Jak widzimy, w każdym wierszu mamy informacje dotyczące jednego stanu.

Nas dzisiaj będą interesowały dane z kolumn `SATcombined`, `PTratio`, `PctSAT`, oraz `LogPctSAT`.

W kolumnie `SATcombined` znajduje się średni wynik testu SAT w danym stanie. Test ten umożliwia zarekrutowanie się na studia wyższe. Nie jest to jednak jedyny taki test. Uczniowie mogą również zdawać test ACT. Generalnie rzecz biorąc test SAT jest popularniejszy w jednych stanach, w innych zaś większą popularnością cieszy się ACT. Różnice w popularności możemy ocenić przyglądając się wartościom z kolumny `PctSAT`. Znajduje się tam odsetek uczniów,

którzy przystąpili do egzaminu SAT. W kolumnie `LogPctSAT` znajdziemy logarytm naturalny z tej wartości. Dodatkowo warto pamiętać, że SAT wymagany jest przez wiele prestiżowych uczelni. W kolumnie `PTratio` znajdziemy informacje o tym ilu uczniów przypada na jednego nauczyciela. W kolumnie `Expend` znajdują się wydatki na edukacje w danym stanie.

Zanim przystąpimy do analizy danych, obejrzyjmy rozkład naszych zmiennych. Dla każdej z nich narysujemy histogram, wykres kwantyl-kwantyl (który posłuży nam do oceny normalności rozkładu) oraz wykres rozrzutu, gdzie zmienną na osi Y będzie `SATcombined` (oprócz samego `SATcombined`).

Żeby to zrobić, posłużymy się odpowiednimi funkcjami z pakietu `ggplot2`. Jeżeli nigdy nie korzystali Państwo z tego pakietu, to zachęcam do przerobienia odpowiedniego notebooka ze strony kursu (nagrałem też do nich filmiki, które dostępne są na YT). Dla tych z Państwa, którzy korzystali z tego pakietu, kilka ciekawych twistów:

- Użyliśmy pakietu `gridExtra` i zawartej w nim funkcji `grid.arrange`, żeby narysować kilka wykresów stworzonych za pomocą `ggplot2` na jednym panelu. Jesli rysowalibyśmy te wykresy używając standardowych funkcji R takich jak `hist`, to moglibyśmy po prostu użyć `par(mfrow=c(4,3))`. W przypadku `ggplot2` technika ta jednak nie działa.
- Kazdy z wykresów jest obiektem, który można przypisać do zmiennej, co zrobiliśmy.
- Wszystkie te obiekty przekazaliśmy do funkcji `grid.arrange`, która narysowała te wykresy.

To, w jaki sposób rysuje się zawarte poniżej typy wykresów w `ggplot2`, omówione jest w notatniku poświęconym temu pakietowi, więc odsyłam do niego (oraz do dokumentacji `ggplot2`!), jeżeli ktoś chciałby dowiedzieć się więcej.

```
library(ggplot2)
library(gridExtra)
expend_qq <- ggplot(data, aes(sample=Expend))+stat_qq() + stat_qq_line()
expend_hist <- ggplot(data, aes(x=Expend))+ geom_histogram(bins = 8)
expend_scatter <- ggplot(data, aes(x=Expend, y=SATcombined)) +geom_point() +
 ↪  geom_smooth(method=lm)

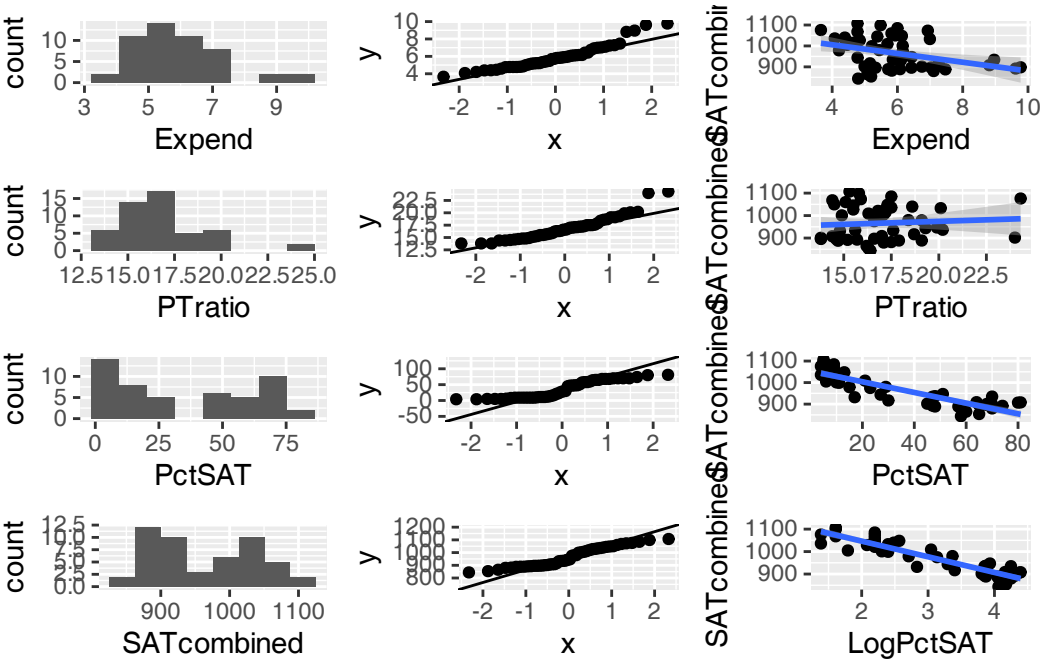

ptratio_qq <- ggplot(data, aes(sample=PTratio))+stat_qq() + stat_qq_line()
ptratio_hist <- ggplot(data, aes(x=PTratio))+ geom_histogram(bins = 8)
ptratio_scatter <- ggplot(data, aes(x=PTratio, y=SATcombined)) +geom_point()
 ↪  + geom_smooth(method=lm)

pctsat_qq <- ggplot(data, aes(sample=PctSAT))+stat_qq() + stat_qq_line()
pctsat_hist <- ggplot(data, aes(x=PctSAT))+ geom_histogram(bins = 8)
pctsat_scatter <- ggplot(data, aes(x=PctSAT, y=SATcombined)) +geom_point() +
 ↪  geom_smooth(method=lm)
```

```
satcombined_qq <- ggplot(data, aes(sample=SATcombined))+stat_qq() +
 ↪  stat_qq_line()
satcombined_hist <- ggplot(data, aes(x=SATcombined))+ geom_histogram(bins =
 ↪  8)
logpctsat_scatter <- ggplot(data, aes(x=LogPctSAT, y=SATcombined))
 ↪  +geom_point() + geom_smooth(method=lm)

grid.arrange(expend_hist, expend_qq, expend_scatter, # dajemy tutaj wszystkie
 ↪  wykresy
             ptratio_hist, ptratio_qq, ptratio_scatter,
             pctsat_hist, pctsat_qq, pctsat_scatter,
             satcombined_hist, satcombined_qq, logpctsat_scatter,
             ncol = 3 # liczba kolumn, w które chcemy je ułożyć
             )
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```



Rzeczą, która powinna zwrócić naszą uwagę jest to, że relacja między `Expend` i `SATCombined` jest negatywna, to znaczy im więcej pieniędzy wydajemy na edukacje, tym gorsze wyniki

uzyskują uczniowie (sic!). Jest to dość kontrintuicyjne, ponieważ spodziewalibysmy się zupełnie odwrotnego efektu – wydatki na edukacje powinny przecież polepszać jej jakość.

Spróbumy teraz obliczyć współczynnik korelacji, między zmiennymi, które nas interesują. Jak Państwo wiedzą, do obliczenia współczynnika korelacji oraz przeprowadzania odpowiedniego testu statystycznego ($H_0 : \rho = 0$) używamy funkcji `cor.test`. Dla przykładu możemy obliczyć korelację między zmienną `SATcombined` oraz `Expend`:

```
pander(cor.test(~ SATcombined + Expend, data = data))
```

Tabela 2: Pearson's product-moment correlation: `SATcombined` and `Expend`

Test statistic	df	P value	Alternative hypothesis	cor
-2.851	48	0.006408 * *	two.sided	-0.3805

Widzimy, że uzyskaliśmy negatywny współczynnik korelacji — niezbyt wysoki, ale statystycznie istotny! Jest to dość zaskakujące.

Efekt ten, przynajmniej według podręcznika, z którego pochodzi ten przykład, związany jest z faktem, że w tych stanach, w których mały odsetek uczniów przystępuje do testu SAT, przystępują do niego zazwyczaj najlepsi i to zaburza wynik. W dalszej części notatnika pokażemy w jaki sposób możemy za pomocą wielokrotnej regresji to pokazać!

Krótka uwaga na marginesie. Gdybyśmy mieli policzyć współczynnik korelacji dla wszystkich kombinacji zmiennych, musielibyśmy napisać bardzo dużo kodu! Możemy sobie z tym jednak poradzić. W pakiecie `rstatix` znajduje się sporo ciekawych funkcji do „masowego'' przeprowadzania testów statystycznych. My skorzystamy z funkcji `cor_test` (która jest odpowiednikiem `cor.test` z biblioteki standardowej R).

```
library(rstatix)
# Do funkcji cor_test przekazujemy ramkę danych oraz nazwy kolumn.
results <- cor_test(data = data, Expend, PTratio, Salary, PctSAT,
 ↪  SATcombined, LogPctSAT)
# Dane wyjściowe `cor_test` to ramka danych więc możemy ją sformatować za
 ↪  pomocą `kable`
kable(results[, c("var1", "var2", "cor", "p")])
```

var1	var2	cor	p
Expend	Expend	1.0000	0.0000000
Expend	PTratio	-0.3700	0.0079873

var1	var2	cor	p
Expend	Salary	0.8700	0.0000000
Expend	PctSAT	0.5900	0.0000058
Expend	SATcombined	-0.3800	0.0064080
Expend	LogPctSAT	0.5600	0.0000227
PTratio	Expend	-0.3700	0.0079873
PTratio	PTratio	1.0000	0.0000000
PTratio	Salary	-0.0011	0.9936975
PTratio	PctSAT	-0.2100	0.1374041
PTratio	SATcombined	0.0810	0.5748329
PTratio	LogPctSAT	-0.1300	0.3610337
Salary	Expend	0.8700	0.0000000
Salary	PTratio	-0.0011	0.9936975
Salary	Salary	1.0000	0.0000000
Salary	PctSAT	0.6200	0.0000018
Salary	SATcombined	-0.4400	0.0013913
Salary	LogPctSAT	0.6100	0.0000022
PctSAT	Expend	0.5900	0.0000058
PctSAT	PTratio	-0.2100	0.1374041
PctSAT	Salary	0.6200	0.0000018
PctSAT	PctSAT	1.0000	0.0000000
PctSAT	SATcombined	-0.8900	0.0000000
PctSAT	LogPctSAT	0.9600	0.0000000
SATcombined	Expend	-0.3800	0.0064080
SATcombined	PTratio	0.0810	0.5748329
SATcombined	Salary	-0.4400	0.0013913
SATcombined	PctSAT	-0.8900	0.0000000
SATcombined	SATcombined	1.0000	0.0000000
SATcombined	LogPctSAT	-0.9300	0.0000000
LogPctSAT	Expend	0.5600	0.0000227
LogPctSAT	PTratio	-0.1300	0.3610337
LogPctSAT	Salary	0.6100	0.0000022
LogPctSAT	PctSAT	0.9600	0.0000000
LogPctSAT	SATcombined	-0.9300	0.0000000
LogPctSAT	LogPctSAT	1.0000	0.0000000

Przejdźmy teraz do głównego tematu dzisiejszych zajęć, to znaczy do wielorakiej regresji liniowej. Jak już wspomnieliśmy, chcielibyśmy zobaczyć związek między wydatkami na szkolnictwo i testem SAT kontrolując przy tym odsetek przystępujących do SAT uczniów, to znaczy zmienną `LogPctSAT` (logarytm z odsetka przystępujących osób, wybraliśmy go bo ma troszkę lepszy rozkład i zależność jest bardziej liniowa).

Stworzymy dwa modele w których zmienną objaśnianą jest `SATcombined`. W pierwszym z nich (`fit1`) uwzględniać będziemy tylko `Expend`.

```
fit1 <- lm(SATcombined ~ Expend, data = data)
pander(summary(fit1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1089	44.39	24.54	8.168e-29
Expend	-20.89	7.328	-2.851	0.006408

Tabela 3: Fitting linear model: SATcombined ~ Expend

Observations	Residual Std. Error	R^2	Adjusted R^2
50	69.91	0.1448	0.127

Teraz stworzymy drugi model, w którym to uwzględnimy jednak dodatkową zmienną - `LogPctSAT`.

```
fit2 <- lm(SATcombined ~ Expend + LogPctSAT, data = data)
pander(summary(fit2))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1147	16.7	68.68	8.447e-49
Expend	11.13	3.264	3.409	0.001346
LogPctSAT	-78.2	4.471	-17.49	3.292e-22

Tabela 4: Fitting linear model: SATcombined ~ Expend + LogPctSAT

Observations	Residual Std. Error	R^2	Adjusted R^2
50	25.78	0.8861	0.8813

Widzimy, że zmiana jest gigantyczna! W poprzednim wypadku współczynnik regresji był dla `Expand` ujemny (-20.89) ale teraz, kiedy kontrolujemy zmienną `LogPctSAT` jest delikatnie dodatni (11.13). To jest dokładnie to, czego oczekiwaliśm! Relacja między `LogPctSAT` a `SATcombined` jest negatywna - im mniej osób przystępuje w danym stanie do egzaminu, tym wyższy wynik średnio uzyskują. To również jest zgodne z naszymi oczekiwaniami. Warto zwrócić uwagę, że dramatycznie zwiększyło się również dopasowanie naszego modelu do danych - współczynnik R^2 wzrósł z 0.145 do 0.886.

Interpretacja wielokrotnej regresji

Spróbujmy prześledzi, co tak naprawdę robi wielokrotna regresja liniowa po to, aby zrozumieć jak interpretować jej wyniki.

Stwórzmy model, w którym zmienną objaśnianą będzie `SATcombined` a predyktorem `LogPctSAT`.

```
fit_sat <- lm(SATcombined ~LogPctSAT, data = data)
```

Obliczmy jakie wartości dla `SATcombined` przewiduje nasz model. Posłużymy się tutaj funkcją `predict`. Działa ona tak, jakbyśmy po prostu wzięli wyraz wolny regresji oraz współczynniki kierunkowe a następnie dla każdej obserwacji obliczali przewidywaną wartość.

```
predictsat <- predict(fit_sat)
```

Następnie obliczmy różnicę między prawdziwymi (rzeczywiście zaobserwowanymi) wartościami a wartościami przewidywanymi przez regresje (czyli residua/reszty!). Przypiszmy je do zmiennej `residsat`.

```
residsat <- data$SATcombined - predictsat
```

W tym momencie musimy dokonać waznej obserwacji. Wartości naszej nowej zmiennej (`residsat`) są nieskorelowane z `LogPctSAT` w naszej próbie — usunęliśmy z nich liniowy związek z tą zmienną. Dokładnie to samo możemy zrobić z naszą drugą zmienną, czyli `Expend`.

```
fit_expend <- lm(Expend ~ LogPctSAT, data = data)
predictexpend <- predict(fit_expend)
residexpend <- data$Expend - predictexpend
```

Mamy teraz dwie zmienne, które są nieskorelowane z `LogPctSAT` w naszej próbie. W takim razie, możemy stworzyć model, w którym przewidujemy wartości `residsat` na podstawie wartości `residexpend`!

```
fit_resid <- lm(residsat ~ residexpend)
pander(summary(fit_resid))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.526e-13	3.608	-7.001e-14	1
residexpend	11.13	3.23	3.445	0.001194

Tabela 5: Fitting linear model: residsat ~ residexpend

Observations	Residual Std. Error	R^2	Adjusted R^2
50	25.51	0.1983	0.1816

Jak widać współczynnik regresji który uzyskaliśmy dla tego modelu z jedną zmienną jest dokładnie taki sam jak wóœczas, gdy mieliśmy dwie zmienne. Dlaczego? Dlatego, że „manualnie'' zrobiliśmy to, co regresja robi za nas – obliczyliśmy współczynnik regresji kontrolując efekty, które można przypisać trzeciej zmiennej (bo „usunęliśmy'' te efekty zarówno z predyktora jak i ze zmiennej objaśnianej).

O wielokrotnej regresji liniowej możemy też myśleć w inny sposób. Wiemy, że w przypadku naszego modelu z dwoma predyktorami równanie regresji ma postać

$$\hat{Y} = b_1 X_1 + b_2 X_2 + b_0$$

Czyli umując to w kategoriach naszych zmiennych:

$$\widehat{SAT} = b_1 Expend + b_2 LogPctSAT + b_0$$

W naszym wypadku znamy b_1 oraz b_2 (bo są to nasze współczynniki regresji) więc równanie to możemy przedstawić również tak:

$$\widehat{SAT} = 11.129 \times Expend - 78.203 \times LogPctSAT + 1147.100$$

Możemy więc korzystając z tego wzoru obliczyć $\widehat{SAT}$.

```
predictions <- predict(fit2)
```

Obliczmy więc korelację między $\widehat{SAT}$ (naszymi predykcjami w zmiennej `predictions`) oraz SAT (czyli wartościami zaobserwowanymi). Tak obliczony współczynnik korelacji podnieśmy do kwadratu.

```r
cor.test(data$SATcombined, predictions)$estimate ** 2
```

```
      cor
0.8861068
```

Przypomnijmy sobie jak wyglądał nasz model dla dwóch zmiennych, który stworzyliśmy wcześniej.

```r
pander(summary(lm(SATcombined ~ Expend + LogPctSAT, data = data)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1147	16.7	68.68	8.447e-49
Expend	11.13	3.264	3.409	0.001346
LogPctSAT	-78.2	4.471	-17.49	3.292e-22

Tabela 6: Fitting linear model: SATcombined ~ Expend + LogPctSAT

Observations	Residual Std. Error	R^2	Adjusted R^2
50	25.78	0.8861	0.8813

Jak widzimy uzyskaliśmy dokładnie tę samą wartość co w przypadku R^2 wcześniej! Nakierowuje nas to na pewien sposób myślenia o regresji, zgodnie z którym zasadniczo sprowadza się ona do utworzenia nowej zmiennej ($\widehat{SAT}$), będącej najlepszą liniową kombinacją predyktorów w naszym modelu (`Expend` oraz `LogPctSAT`). Przez „najlepszą liniową kombinację'' rozumiemy to, że nie ma żadnych innych wag, które moglibyśmy przypisać naszym predyktorom (współczynników regresji), które lepiej przewidywałyby naszą zmienną objaśnianą.

O współczynniku R dla wielokrotnej regresji liniowej możemy myśleć po prostu jako o współczynniku korelacji r Pearsona tyle że dla zmiennej stanowiącej najlepszą liniową kombinację predyktorów.