Co to jest (mieszany) model liniowy i co można z nim zrobić

26 października 2011

Spis treści

1 Regresja liniowa z jedną zmienną niezależną 1

2 ANOVA jednoczynnikowa 6

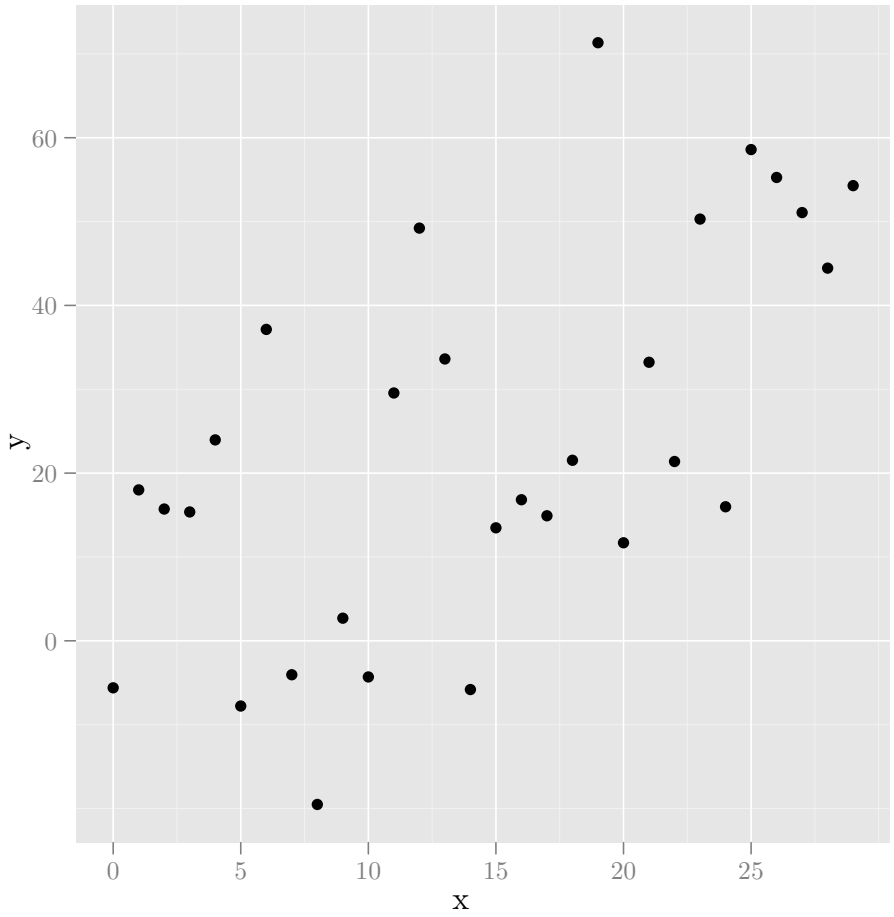
3 Model liniowy z interakcją zmiennej ilościowej i czynnika 14

4 Model liniowy ze zmienną ilościową i czynnikiem bez interakcji 17

5 Liniowy model mieszany z jedną ilościową zmienną niezależną 20

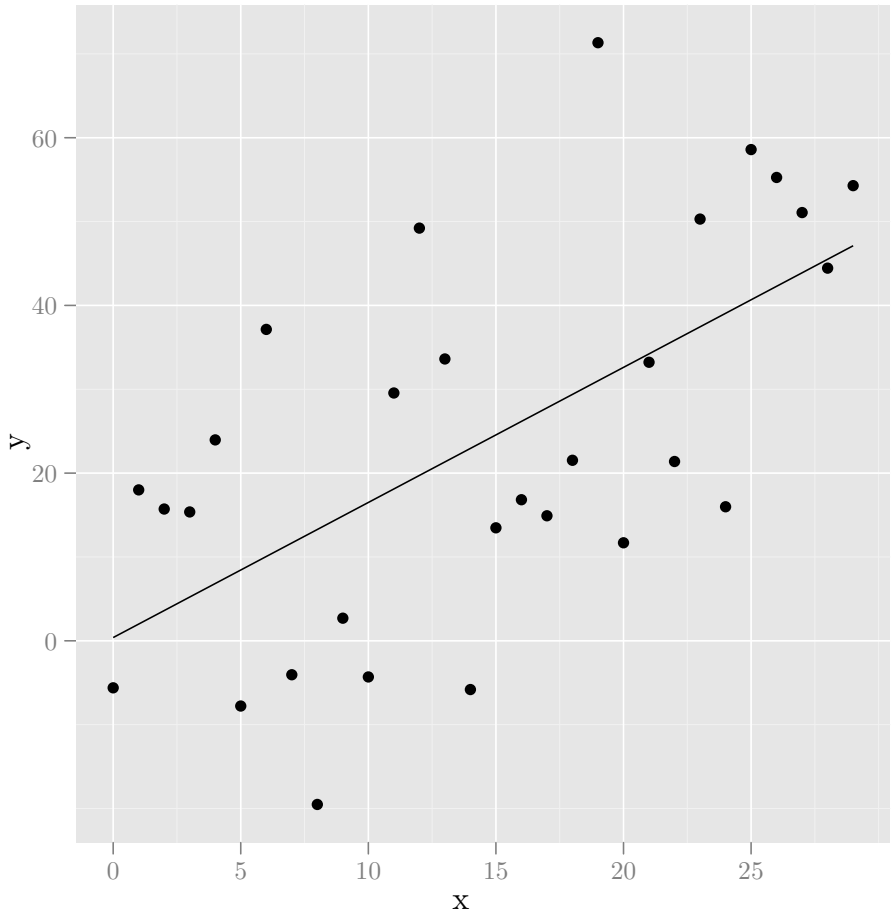
1 Regresja liniowa z jedną zmienną niezależną

Na początek popatrzymy sobie na dane pochodzące z symulacji:



Wygląda na zależność w przybliżeniu prostoliniową. Nie napisałem liniową, bo liniowość w kontekście ogólnego modelu liniowego oznacza coś innego, niż zależność wyglądająca na wykresie jak linia prosta, ale o tym później.

Popatrzymy sobie teraz na dopasowanie modelu liniowego:



Kreska wygląda sensownie. Teraz popatrzymy sobie na dopasowanie tego samego modelu, ale w postaci tabelki współczynników regresji:

```
y ~ x

            Wspolcz. Blad std.     p sig
(Intercept)    0.386     6.636 0.954
x              1.611     0.393 0.000 ***

k = 2 n =  30 blad std. reszt = 18.629
```

Mamy dwa współczynniki, `(Intercept)` równy około 0.39 i `x` równy około 1.61 (odtąd będziemy często zaokrąglać do dwóch miejsc po przecinku). Chociaż doskonale wiemy, co to są za współczynniki, powtórzymy to sobie teraz, bo wkrótce będzie nieco trudniej.

Formalnie rzecz ujmując dopasowaliśmy model:

$$y = \alpha_0 + \alpha_1 x + \epsilon$$

gdzie ϵ to coś, co czasem bywa nazywane błędem pomiaru, ale co lepiej jest nazywać *resztami*, ponieważ rozproszenie danych wokół wartości oczekiwanej wynika zwykle z wielu przyczyn, nie tylko z błędu pomiaru zmiennych. Właściwie powinniśmy przedstawić równanie dla tego modelu z odpowiednimi indeksami:

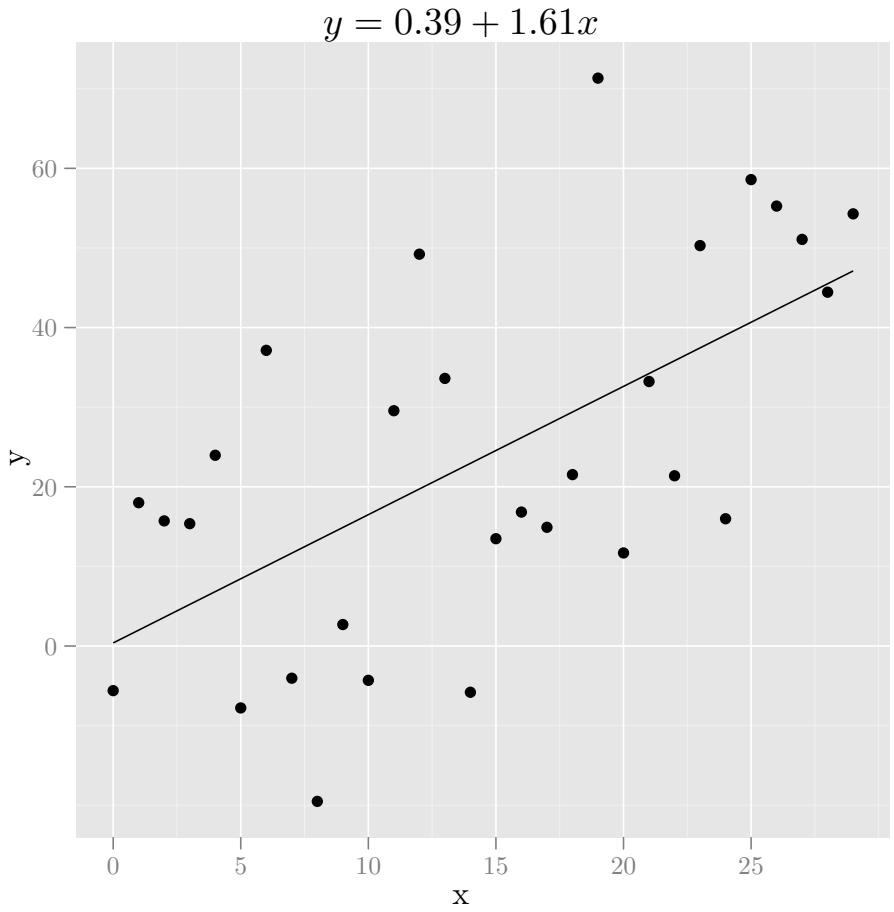
$$y_i = \alpha_0 + \alpha_1 x_i + \epsilon_i$$

Tutaj indeks i identyfikuje nam poszczególne obserwacje. Każda konkretna wartość y_i odpowiada pewnej wartości x_i i pewnej reszcie ϵ_i. Parametry α_0 i α_1 z założenia mają przyjmować te same wartości dla każdej obserwacji, dlatego tych parametrów nie indeksujemy, albo raczej, indeksujemy je inaczej.

W modelu liniowym zakładamy, że reszty (ϵ) mają rozkład normalny, ze średnią zero i nieznaną wariancją. Model ma tak naprawdę *trzy* wolne parametry, to jest α_0, α_1 i σ. W tabelce regresji współczynnik o nazwie `(Intercept)` to nasz α_0, inaczej nazywany *wyrazem wolnym*. Ten współczynnik szacuje punkt przecięcia z osią Y. Współczynnik o nazwie `x` to nasz α_1, czyli *nachylenie* linii regresji, a parametr σ to nieznane przed dopasowaniem modelu odchylenie standardowe reszt. W ogólnym przypadku dopasowanie modelu polega na znalezieniu „optymalnych" wartości wszystkich (tutaj trzech) wolnych parametrów. Współczynniki, które widzieliśmy w tabelce, to właśnie owe w pewien sposób optymalne oszacowania. Dopasowany model, pomijając reszty, które nas zwykle nie interesują, wygląda zatem tak:

$$y_i = 0.39 + 1.61 x_i$$

Odtąd często będziemy pomijać reszty. Popatrzmy jak się to wszystko ma do wykresu:



Te wartości współczynników można i warto potraktować dosłownie. Pomijając „szum" (reszty), gdy zmienna x jest równa 0, to zmienna y przyjmuje wartość:

$$y = 0.39 + 1.61 \times 0 = 0.39$$

Patrzymy na wykres i widzimy, że wszystko się zgadza. Gdy zmienna x przyjmuje na przykład wartość 20, to:

$$y = 0.39 + 1.61 \times 20 = 32.61$$

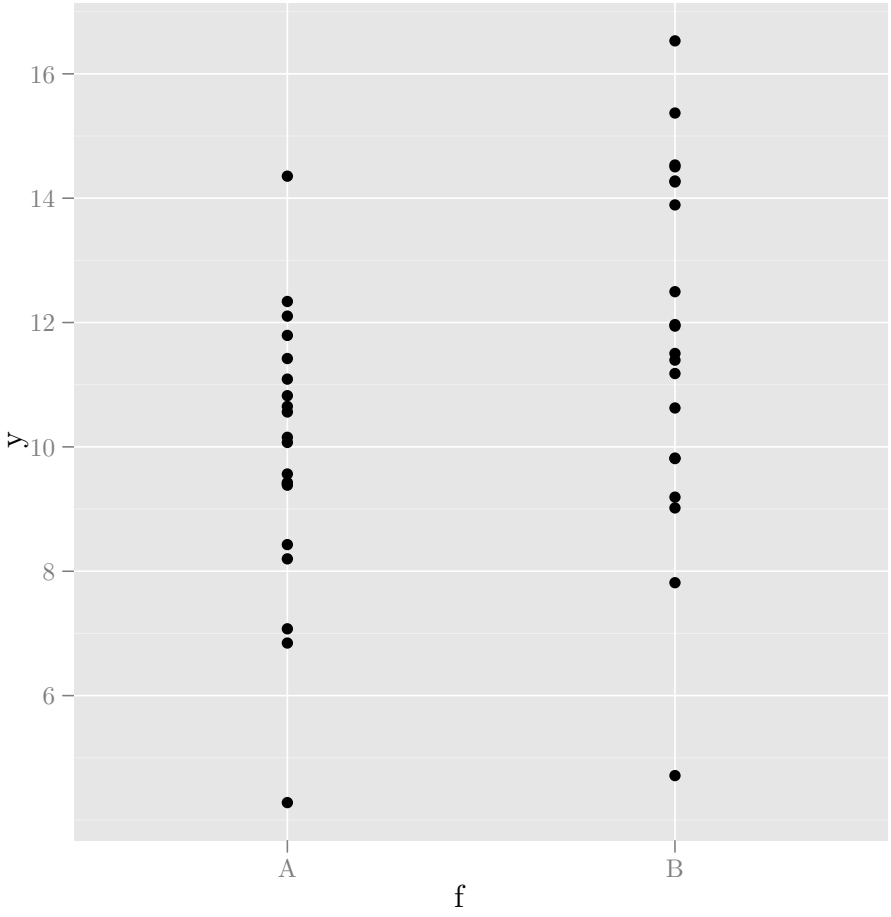
Patrzymy na wykres i znowu wszystko się zgadza. Mówiąc prościej, współczynnik `x` z tabelki, którą widzieliśmy wcześniej, mówi nam, o ile przeciętnie wzrasta (lub maleje) wartość zmiennej y gdy zmienna x wzrasta o 1. Współczynnik `(Intercept)` mówi nam, jaka jest oczekiwana (średnia) wartość y gdy *wszystkie pozostałe predyktory* są równe 0. Wyjaśnimy sens terminu predyktor nieco później. Na razie zaznaczymy tylko, że predyktor to niekoniecznie to samo co zmienna niezależna.

Myślenie w kategoriach wyrazu wolnego - inaczej punktu przecięcia z osią Y - i nachylenia jest naturalne, gdy mamy do czynienia ze zmienną niezależną

ilościową. Ten sam sposób myślenia można zastosować do sytuacji, gdy zmienna niezależna to *czynnik*.

2 ANOVA jednoczynnikowa

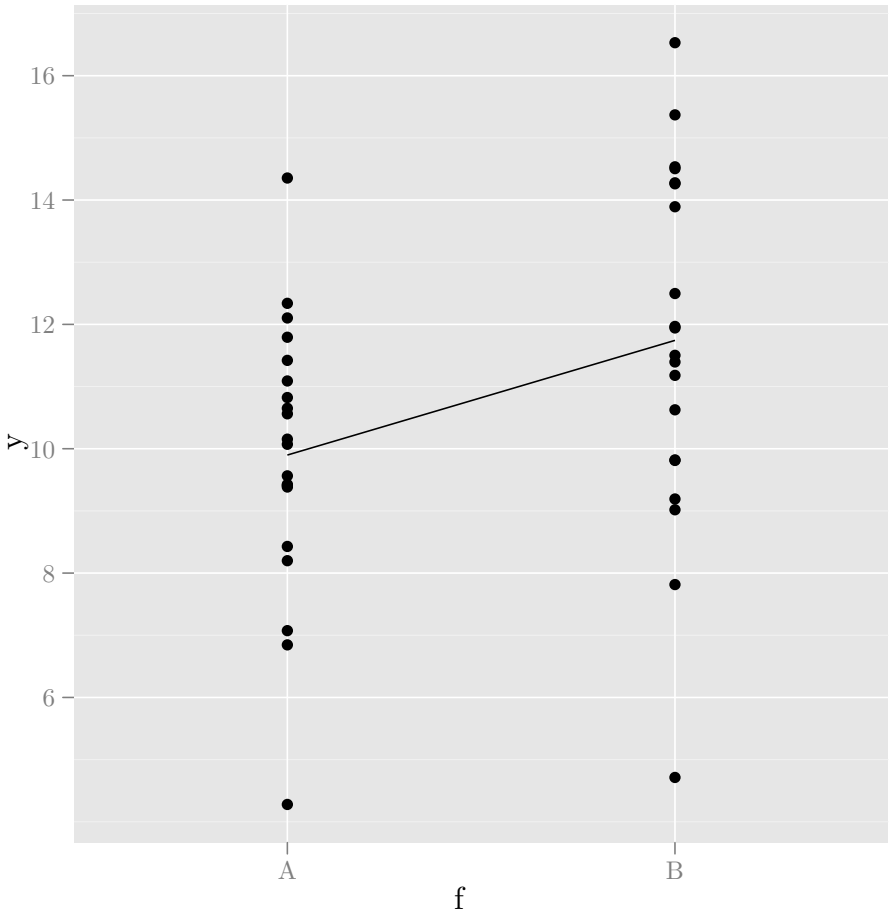
Zastosowaliśmy dwa poziomy (A i B) czynnika f i otrzymaliśmy takie oto dane:



Ponieważ czujemy się komfortowo w towarzystwie ANOVy, zrobiliśmy ANOVę i otrzymaliśmy taką oto tabelkę:

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
f          1  34.00  34.000   5.123 0.0294 *
Residuals 38 252.19   6.637
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Popatrzymy sobie teraz na dopasowanie modelu:



Zależność reprezentowaną przez zaznaczony na wykresie odcinek również możemy opisać dosłownie i wyczerpująco za pomocą kilku liczb. Gdy $f = A$, to y powinno być (według naszego dopasownego modelu przedstawionego za pomocą kreski) równe około 9.9. Gdy $f = B$, to y powinno wynosić około 11.74. Tak to wygląda na wykresie. Oczywiście tabelka ANOVy niczego takiego nam nie mówi. Tabelka ANOVy w ogóle nie mówi nam zbyt wiele. Warto podkreślić, że to nie jest *nasza* wina.

ANOVA to nic innego, jak ogólny model liniowy. Model regresji liniowej to nic innego jak ogólny model liniowy.

„ANOVA" tym się różni od „regresji liniowej", że inaczej przedstawiamy wyniki dopasowania tego samego modelu.

Można niemal powiedzieć, że ANOVA to tylko *specjalny rodzaj tabelki*, za pomocą którego psychologowie lubią przedstawiać wyniki dopasowania modelu liniowego. Tabelka ANOVy, którą widzieliśmy przed chwilą, powstała na podstawie dopasowania modelu liniowego, którego współczynniki wraz z istotnościami wyglądają tak:

```
y ~ f
```

```
            Wspolcz. Blad std.     p sig
(Intercept)    9.898     0.576 0.000 ***
fB             1.844     0.815 0.029   *

k = 2 n =  40 blad std. reszt = 2.576
```

W wartości wyrazu wolnego natychmiast rozpoznajemy oczekiwaną wartość y dla $f = A$, czyli lewy koniec odcinka na wykresie obrazującym dopasowanie modelu. W wartości współczynnika o dziwnej nazwie `fB` na razie niekoniecznie cokolwiek rozpoznajemy. Chcielibyśmy wiedzieć, skąd wzięły się te współczynniki. Skoro to są współczynniki modelu liniowego, to czym w tym modelu był x?

Otóż jednoczynnikowa ANOVA została wyrażona za pomocą tak zwanej *macierzy modelu*. Zaraz okaże się, że pojęcie macierz modelu jest Twoim przyjacielem. Pierwsze 10 rzędów macierzy modelu dla naszej ANOVy wygląda tak:

```
   (Intercept) fB
1            1  0
2            1  1
3            1  0
4            1  1
5            1  0
6            1  1
7            1  0
8            1  1
9            1  0
10           1  1
```

Mamy dwie kolumny, `(Intercept)` i `fB`. Żeby od razu dostrzec sens kolumny `fB` popatrzymy na macierz modelu połączoną z kolumną kodującą warunki:

```
   (Intercept) fB f
1            1  0 A
2            1  1 B
3            1  0 A
4            1  1 B
5            1  0 A
6            1  1 B
7            1  0 A
8            1  1 B
9            1  0 A
10           1  1 B
```

Należy zaznaczyć, że kolumna `f` *nie* jest częścią macierzy modelu, dokleiliśmy ją tylko po to, żeby coś zademonstrować. Od razu widzimy, że kolumna `fB` przyjmuje wartość 0 gdy $f = A$ i wartość 1 gdy $f = B$. To po prostu pewien *kontrast*, który reprezentuje czynnik f.

Każdy model liniowy jest reprezentowany dla potrzeb dopasowania przez macierz modelu. Jeżeli nie zastosujemy szczególnej parametryzacji (a zwykle

tego nie robimy) macierz modelu dla modelu liniowego zawsze jako pierwszą będzie miała kolumnę samych jedynek. W modelu regresji z jednym predyktorem też tak naprawdę była ukryta taka macierz modelu z kolumną samych jedynek o nazwie `(Intercept)`. Wrócimy teraz na chwilę do macierzy modelu dla modelu regresji omawianego na początku. Macierz modelu regresji liniowej, który omawialiśmy wcześniej, wygląda tak (pierwsze 10 rzędów):

```
   (Intercept) x
1            1 0
2            1 1
3            1 2
4            1 3
5            1 4
6            1 5
7            1 6
8            1 7
9            1 8
10           1 9
```

Widzimy tutaj jakie wartości przyjmowała zmienna x dla kolejnych obserwacji. Możemy sobie połączyć tą macierz modelu z kolumną wartości y odpowiadających poszczególnym obserwacjom:

```
        y (Intercept) x
1   -5.61           1 0
2   18.00           1 1
3   15.72           1 2
4   15.37           1 3
5   23.96           1 4
6   -7.78           1 5
7   37.14           1 6
8   -4.05           1 7
9  -19.52           1 8
10   2.70           1 9
```

Teraz wyjaśnimy o co chodzi z tą kolumną jedynek. Jak pamiętamy, model liniowy z jedną zmienną niezależną wygląda tak:

$$y_i = \alpha_0 + \alpha_1 x_i + \epsilon_i$$

Ale to jest to samo co:

$$y_i = \alpha_0 1_i + \alpha_1 x_i + \epsilon_i$$

(uzupełniliśmy wzór o 1_i) W nieszkodliwym uproszczeniu, taki właśnie jest sens kolumny samych jedynek w macierzy modelu. Po prostu każdy współczynnik modelu regresji jest związany z jakimś predyktorem, a każdy *predyktor to pewna kolumna w macierzy modelu*. Dla wyrazu wolnego (punkt przecięcia z osią Y) taki predyktor został specjalnie utworzony. Teraz wracamy do macierzy modelu dla naszej jednoczynnikowej ANOVy:

```
   (Intercept) fB
1            1  0
2            1  1
3            1  0
4            1  1
5            1  0
6            1  1
7            1  0
8            1  1
9            1  0
10           1  1
```

Numery obserwacji w zbiorze danych (wartości indeksu i) znajdują się tutaj po lewej stronie macierzy. Tak się składa, że w naszym symulowanym zbiorze danych pozycje nieparzyste to te, dla których czynnik f przyjmował wartość A, natomiast pozycje parzyste to te, dla których f był równy B. Czynnik f, będący *pojedynczą* zmienną niezależną, został zamieniony na *dwa* predyktory. Teraz rozumiemy, dlaczego „predyktor" w naszej terminologii znaczy niekoniecznie to samo, co „zmienna niezależna". Czynnik f jest pojedynczą zmienną zależną, którą reprezentują w modelu dwa predyktory. Teraz to jest już tylko szkolna arytmetyka.

Współczynniki regresji dla naszej ANOVy wyglądają tak:

```
y ~ f

            Wspolcz. Blad std.     p sig
(Intercept)    9.898     0.576 0.000 ***
fB             1.844     0.815 0.029   *

k = 2 n =  40 blad std. reszt = 2.576
```

a są to współczynniki, przez które należy pomnożyć predyktory, czyli kolumny macierzy modelu, która to macierz wygląda tak:

```
   (Intercept) fB
1            1  0
2            1  1
3            1  0
4            1  1
5            1  0
6            1  1
7            1  0
8            1  1
9            1  0
10           1  1
```

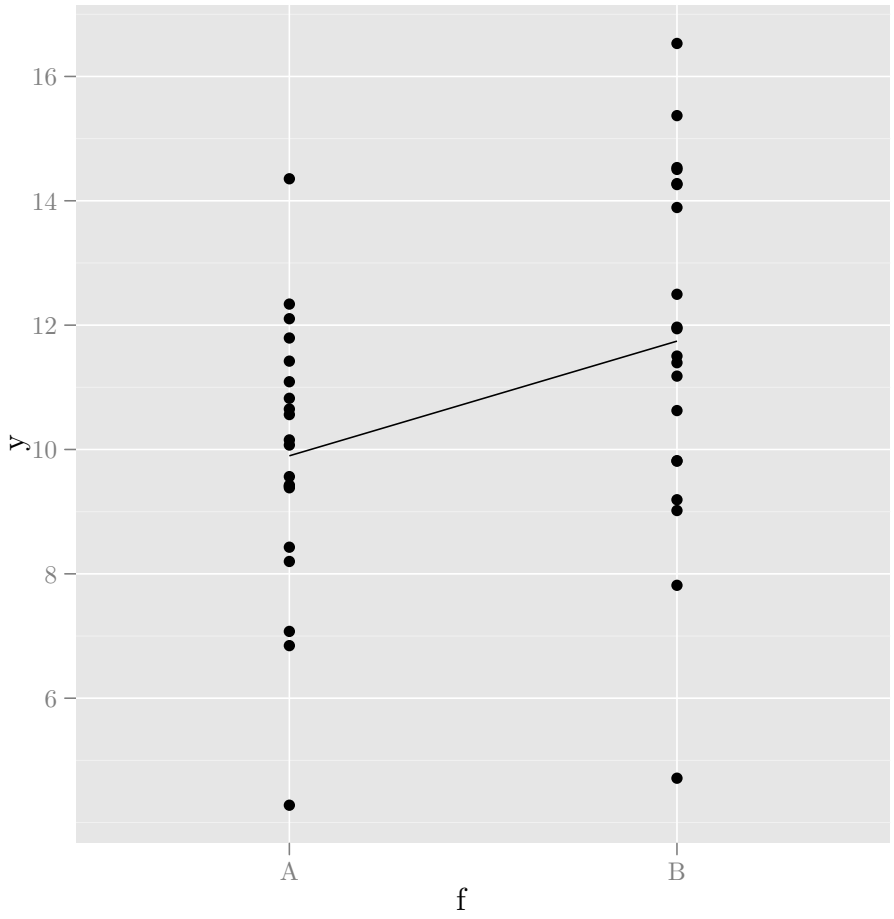
Wiemu już, że predyktor `fB` przyjmuje wartość 0 dla tych obserwacji, dla których f był równy A i wartość 1 dla tych obserwacji, dla których f był równy B — stąd nazwa `fB`. Skoro tak, to gdy f jest równe A, to zgodnie z dopasowanym modelem:

$$y = 9.9 + 1.84 \times 0 = 9.9$$

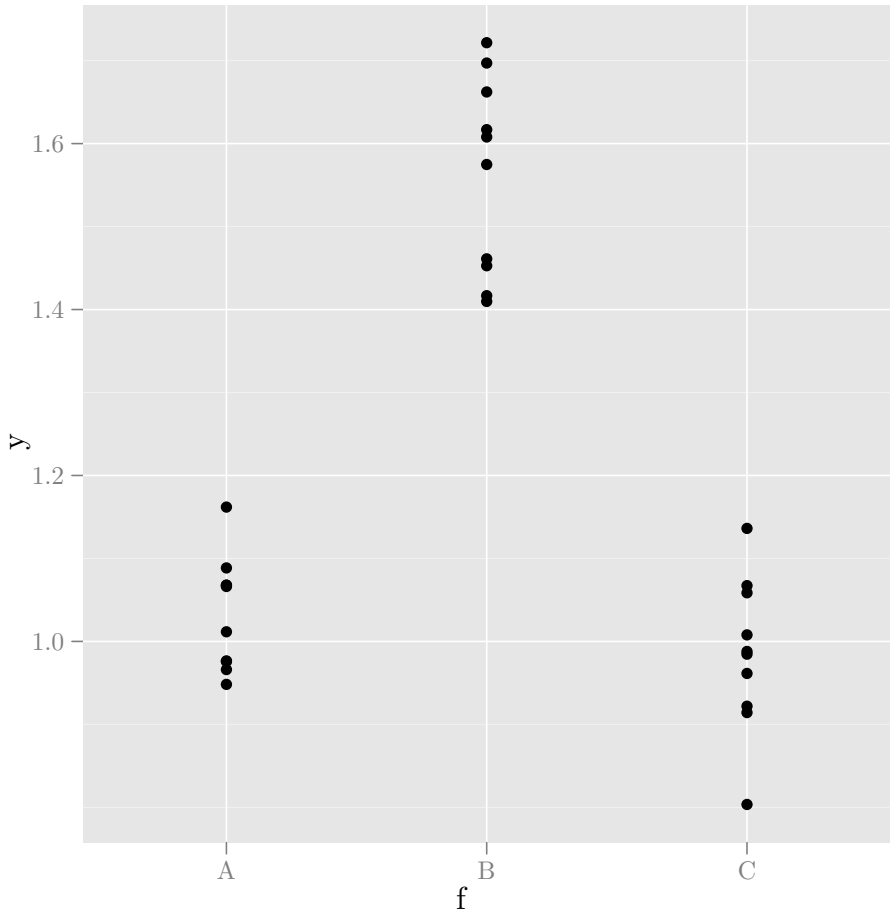
co zgadza się z wykresem. Natomiast gdy f jest równe B, to:

$$y = 9.9 + 1.84 \times 1 = 11.74$$

co też zgadza się z wykresem, na który popatrzymy jeszcze raz:



Jeżeli woleliśmy wczesniej tabelkę ANOVy, to w tym momencie przestajemy rozumieć, dlaczego woleliśmy tabelkę ANOVy. Rozważymy teraz przypadek, gdy czynnik ma trzy poziomy, a dane wyglądają tak:



Jeszcze raz te same dane, tym razem w postaci tabelki (pierwsze 10 rzędów):

```
   f    y
1  A 1.09
2  B 1.62
3  C 0.91
4  A 1.16
5  B 1.61
6  C 0.92
7  A 0.97
8  B 1.72
9  C 0.80
10 A 1.01
```

Macierz modelu wygląda wtedy tak:

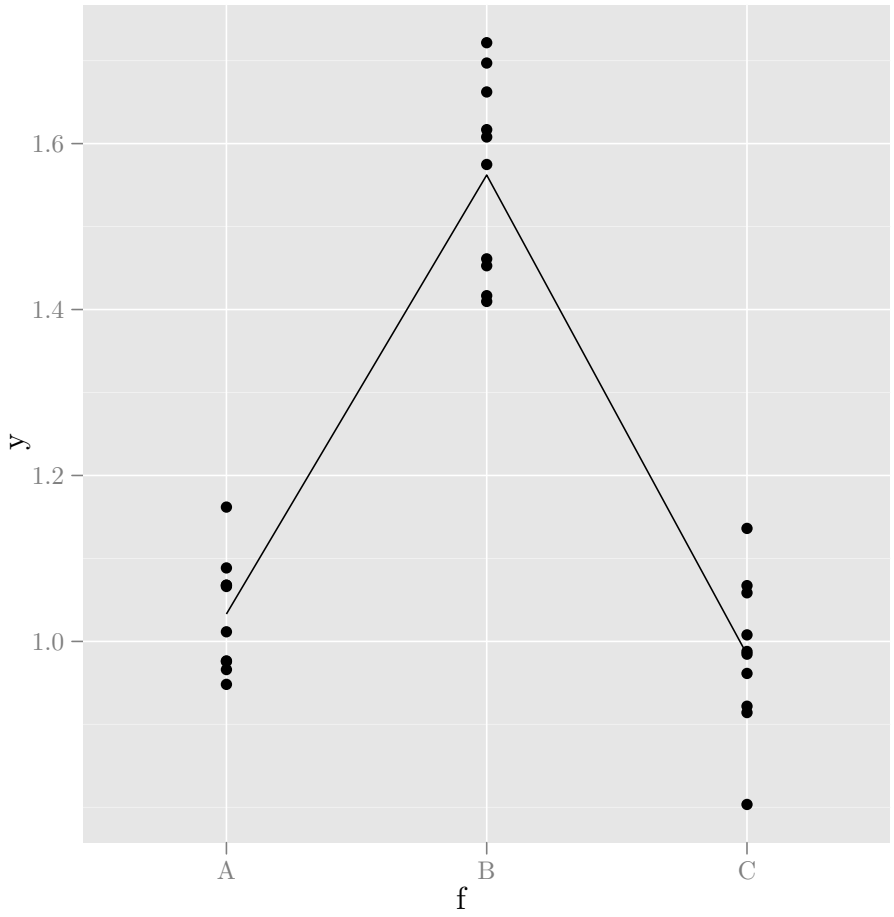
```
   (Intercept) fB fC
1            1  0  0
2            1  1  0
```

```
3            1  0  1
4            1  0  0
5            1  1  0
6            1  0  1
7            1  0  0
8            1  1  0
9            1  0  1
10           1  0  0
```

Domyślamy się już, że każdy czynnik będzie zamieniony na $k-1$ predyktorów, gdzie k to liczba poziomów czynnika. Predyktor `fB` przyjmuje wartość 1 dokładnie wtedy, gdy $f = B$, a predyktor `fC` przyjmuje wartość 1 dokładnie wtedy, gdy $f = C$. Tabelka regresji wygląda tak:

```
y ~ f

            Wspolcz. Blad std.     p sig
(Intercept)    1.033     0.030 0.000 ***
fB             0.529     0.043 0.000 ***
fC            -0.049     0.043 0.263

k = 3 n =  30 blad std. reszt = 0.095
```

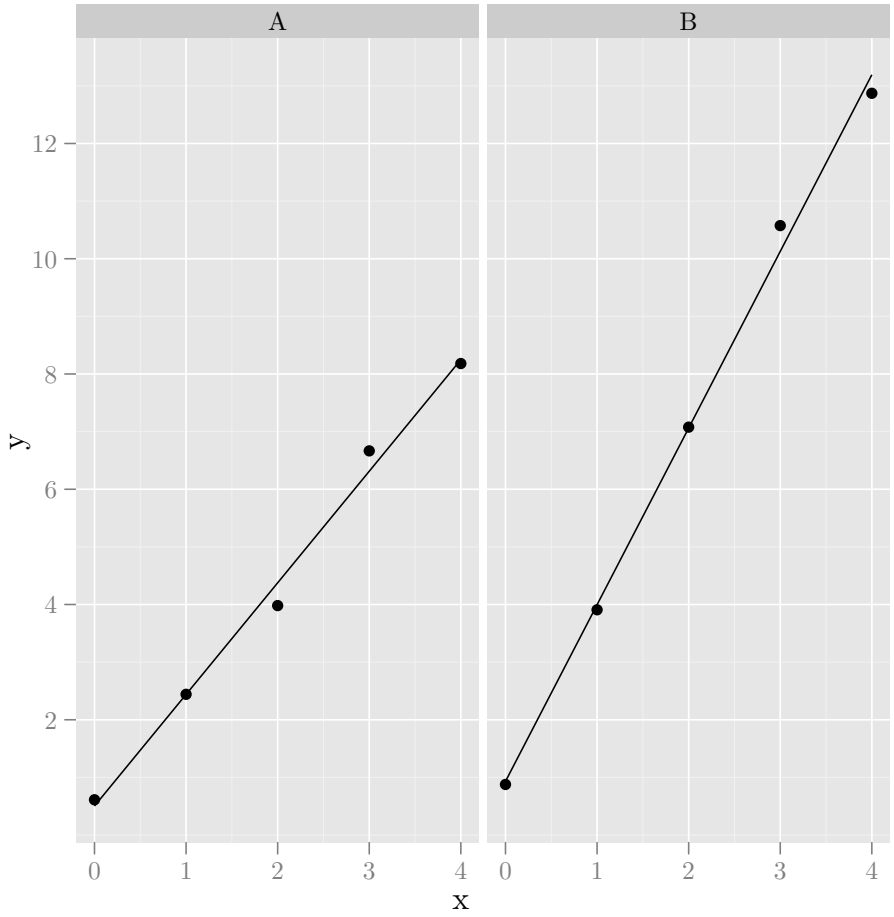
a więc zgodnie z dopasowanym modelem liniowym oczekiwana wartość y dla poziomu A wynosi 1.03, dla poziomu B to będzie $1.03 + 0.53 = 1.56$, a dla poziomu C to będzie $1.03 + -0.05 = 0.98$. Wszystko to zgadza się z tym, co widzimy na wykresie obrazującym dopasowanie modelu:



Podsumowując, w przypadku gdy jedyną zmienną niezależną jest czynnik, współczynniki regresji mówią nam, jaka jest wartość oczekiwana zmiennej zależnej gdy ten czynnik przyjmuje poziom bazowy (tutaj akurat A, ale możemy przyjąć dowolny inny) i o ile większa (lub mniejsza) jest oczekiwana wartość zmiennej zależnej, gdy czynnik przyjmuje pozostałe poziomy. W tle ANOVA jednoczynnikowa jest wyrażona za pomocą macierzy modelu, składającej się z kolumny samych jedynek i kolumn kodujących poziomy inne niż bazowy.

3 Model liniowy z interakcją zmiennej ilościowej i czynnika

Niech dane i dopasowany model liniowy wyglądają tak:



Żeby nieco ułatwić sobie życie dodaliśmy bardzo mało szumu. Mamy dwa poziomy czynnika f i dla każdego poziomu mamy 5 wartości *ilościowej* zmiennej niezależnej x. Dane zestawione razem z macierzą modelu wyglądają tak:

```
   f x     y (Intercept) fB x fB:x
1  A 0  0.61           1  0 0    0
2  A 1  2.44           1  0 1    0
3  A 2  3.98           1  0 2    0
4  A 3  6.67           1  0 3    0
5  A 4  8.18           1  0 4    0
6  B 0  0.88           1  1 0    0
7  B 1  3.91           1  1 1    1
8  B 2  7.08           1  1 2    2
9  B 3 10.58           1  1 3    3
10 B 4 12.87           1  1 4    4
```

A oto współczynniki regresji:

```
y ~ f * x
```

```
            Wspolcz. Blad std.     p sig
(Intercept)    0.504     0.247 0.088
fB             0.427     0.350 0.268
x              1.936     0.101 0.000 ***
fB:x           1.129     0.143 0.000 ***

k = 4 n =  10 blad std. reszt = 0.319
```

`(Intercept)` wynosi 0.5, a więc nasz dopasowany model mówi, że gdy wszystkie pozostałe *predyktory* są równe 0, oczekiwana wartość y wynosi 0.5. Wszystkie pozostałe predyktory to kolumny macierzy modelu o nazwach `fB`, `x` i `fB:x`. Te predyktory są równe 0 dokładnie wtedy, gdy $x = 0$ i $f = A$ (poziom bazowy czynnika). Wobec tego współczynnik dla wyrazu wolnego szacuje nam punkt przecięcia linii regresji z osią Y ($x = 0$ to właśnie ten punkt przecięcia) w warunku bazowym $f = A$. Patrzymy na wykres i wszystko się zgadza. Ponieważ `fB` jest równy 0.43, to gdy `x` jest równe 0 i `fB:x` jest równe 0 oczekiwana wartość y wynosi $0.5 + 0.43 = 0.93$. To jest właśnie punkt przecięcia z osią Y w warunku $f = B$.

Współczynnik `x` jest równy 1.94 i szacuje nachylenie linii regresji dla zmiennej niezależnej x w warunku $f = A$. Żeby zrozumieć, że jest tak właśnie, trzeba zastanowić się przez chwilę nad macierzą modelu. Żeby policzyć, jakie jest nachylenie linii regresji w warunku $f = B$, trzeba dodać współczynniki `x` i `fB:x`. Żeby zrozumieć, że jest tak właśnie, trzeba ponownie zastanowić się przez chwilę nad macierzą modelu. Wzrost oczekiwanej wartości y związany z jednostkowym wzrostem wartości x dany jest przez:

$$\begin{aligned}
&\alpha_0 + \alpha_1 fB + \alpha_2(x+1) + \alpha_3(fBx+1) - (\alpha_0 + \alpha_1 fB + \alpha_2 x + \alpha_3 fBx) \\
&= \alpha_2(x+1) + \alpha_3(fBx+1) - (\alpha_2 x + \alpha_3 fBx) \\
&= \alpha_2 + \alpha_3
\end{aligned}$$

jeżeli `fB:x` *w ogóle może wzrastać o* 1. Jak widać w macierzy modelu predyktor `fB:x` zmienia się tylko w warunku $f = B$, a w warunku $f = A$ jest stały i równy zawsze 0. Wobec tego nachylenie jest równe sumie współczynników `x` i `fB:x` (nasze odpowiedniki α_2 i α_3) tylko w warunku $f = B$. W warunku $f = A$ nachylenie jest dane przez wartość współczynnika `x`.

Mówiąc krótko, wyraz wolny mówi nam tutaj, jaki jest punkt przecięcia z osią Y w warunku bazowym, współczynnik `fB` mówi nam, o ile punkt przecięcia w warunku $f = B$ jest większy (lub mniejszy) niż w warunku bazowym, współczynnik `x` mówi nam, jakie jest nachylenie linii regresji dla zmiennej niezależnej x w warunku bazowym, a współczynnik `fB:x` mówi nam, o ile jest większe (lub mniejsze) nachylenie tej linii w warunku $f = B$. Kolumna `fB:x` powstała przez pomnożenie kolumn `x` i `fB`. Tak właśnie powstają kolumny kodujące efekty interakcyjne. Tak samo będziemy interpretować współczynniki w sytuacji, gdy czynnik ma więcej niż dwa poziomy i mamy w modelu interakcję tego czynnika ze zmienną ilościową. Uzyskamy punkt przecięcia dla poziomu bazowego i różnicę między punktami przecięcia dla pozostałych poziomów a punktem przecięcia dla poziomu bazowego i tak samo z nachyleniami. To wszystko można przedstawić za pomocą tabelki ANOVy:

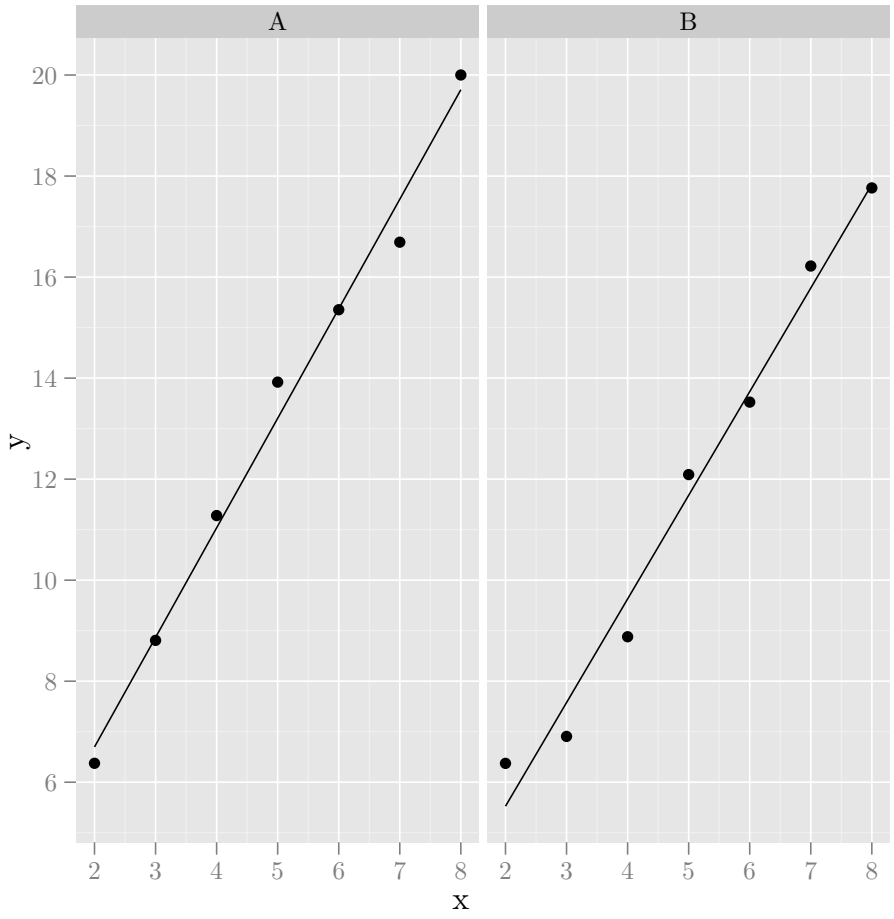
```
Analysis of Variance Table
```

```
Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
f          1  18.030  18.030  176.775 1.119e-05 ***
x          1 125.085 125.085 1226.415 3.613e-08 ***
f:x        1   6.374   6.374   62.499 0.0002173 ***
Residuals  6   0.612   0.102
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

tylko nie wiadomo po co.

4 Model liniowy ze zmienną ilościową i czynnikiem bez interakcji

Oto dane i dopasowany model w którym *uwzględniamy* możliwą interakcję:



A oto tabelka:

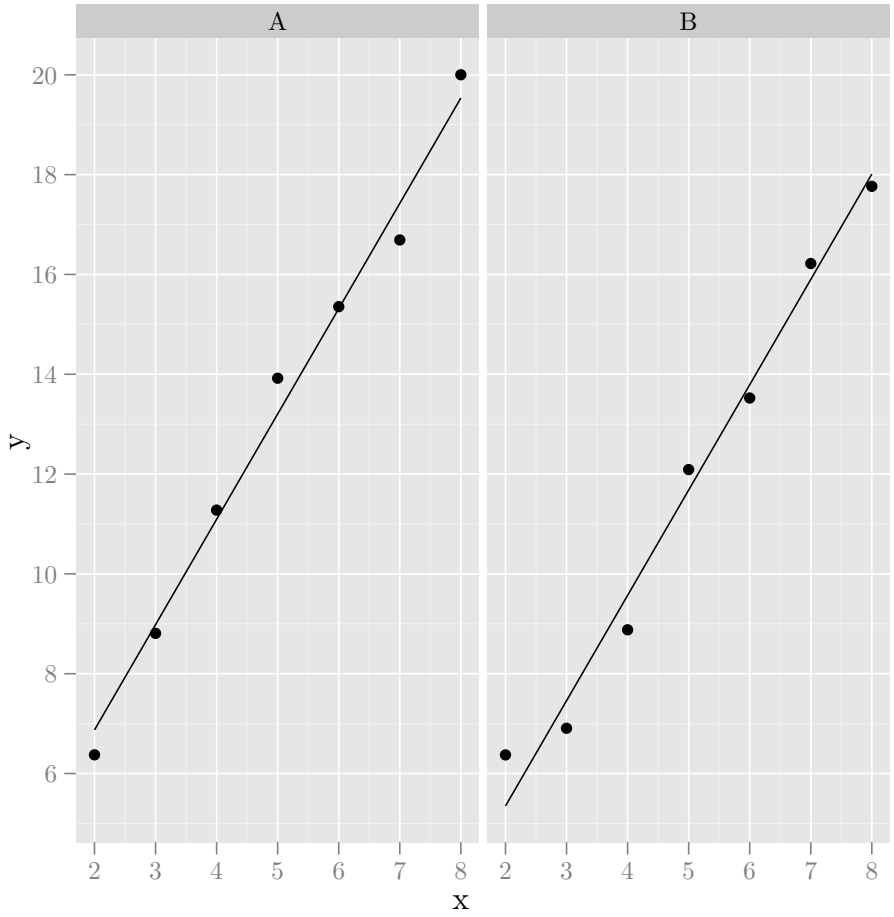
```
y ~ f * x
```

```
            Wspolcz. Blad std.     p sig
(Intercept)    2.361     0.613 0.003  **
fB            -0.939     0.866 0.304
x              2.169     0.114 0.000 ***
fB:x          -0.117     0.161 0.484

k = 4 n =  14 blad std. reszt = 0.602
```

Efekt interakcyjny, to jest zmiana w nachyleniu linii regresji dla x związana z przejściem z poziomu A do B jest daleki od statystycznej istotności, więc sprawdzamy, czy prostszy model, bez termu interakcyjnego (czyli kolumny `fB:x` w macierzy modelu) nie będzie pasował równie dobrze:

```
Analysis of Variance Table

Model 1: y ~ f + x
Model 2: y ~ f * x
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 3.8150
2     10 3.6234  1   0.19157 0.5287 0.4838
```

Wynik jest nieistotny, a więc nie ma powodów do odrzucania hipotezy zerowej. Tutaj hipotezą zerową jest model prostszy, pozbawiony termu interakcyjnego. W tym prostszym modelu nie dopuszczamy, aby nachylenie zmieniało się między warunkami. Dopasowanie modelu bez interakcji wygląda tak:



czyli jak widać równie znakomicie. Współczynniki modelu wyglądają tak:

```
y ~ f + x
```

```
            Wspolcz. Blad std.     p sig
(Intercept)    2.653     0.452 0.000 ***
fB            -1.524     0.315 0.001 ***
x              2.110     0.079 0.000 ***
```

```
k = 3 n =  14 blad std. reszt = 0.589
```

A więc gdy $x = 0$ i $f = A$ to y powinno być średnio równe 2.65. Patrzymy na wykres i nie widzimy nigdzie x równego 0. W tym przypadku szczególnie nas to nie martwi, gdyby jednak zmienna x reprezentowała na przykład wzrost, wyraz wolny nie miałby sensu. Odejmiemy sobie od wartości zmiennej x średnią z x, czyli wycentrujemy tą zmienną na średniej. Współczynniki wyglądają wtedy tak:

```
y ~ I(x - mean(x)) + f
```

```
               Wspolcz. Blad std.     p sig
(Intercept)      13.204     0.223 0.000 ***
I(x - mean(x))    2.110     0.079 0.000 ***
fB               -1.524     0.315 0.001 ***

k = 3 n =  14 blad std. reszt = 0.589
```

Średnia wartość niecentrowanej zmiennej x wynosi 5. Wyraz wolny zgadza się z oczekiwaną wartością y dla $x = 5$, co widać na wykresie. Dodanie lub odjęcie ustalonej wartości od zmiennej ilościowej lub pomnożenie tej zmiennej przez ustaloną (niezerową) wartość to tylko zmiana skali. Takie przekształcenie wpływa na *oszacowania* współczynników, ale nie wpływa na *istotności* współczynników i służy jedynie zwiększeniu ich czytelności. Macierz modelu (pierwsze 10 rzędów) wygląda tak:

```
   (Intercept) x fB
1            1 2  0
2            1 3  0
3            1 4  0
4            1 5  0
5            1 6  0
6            1 7  0
7            1 8  0
8            1 2  1
9            1 3  1
10           1 4  1
```

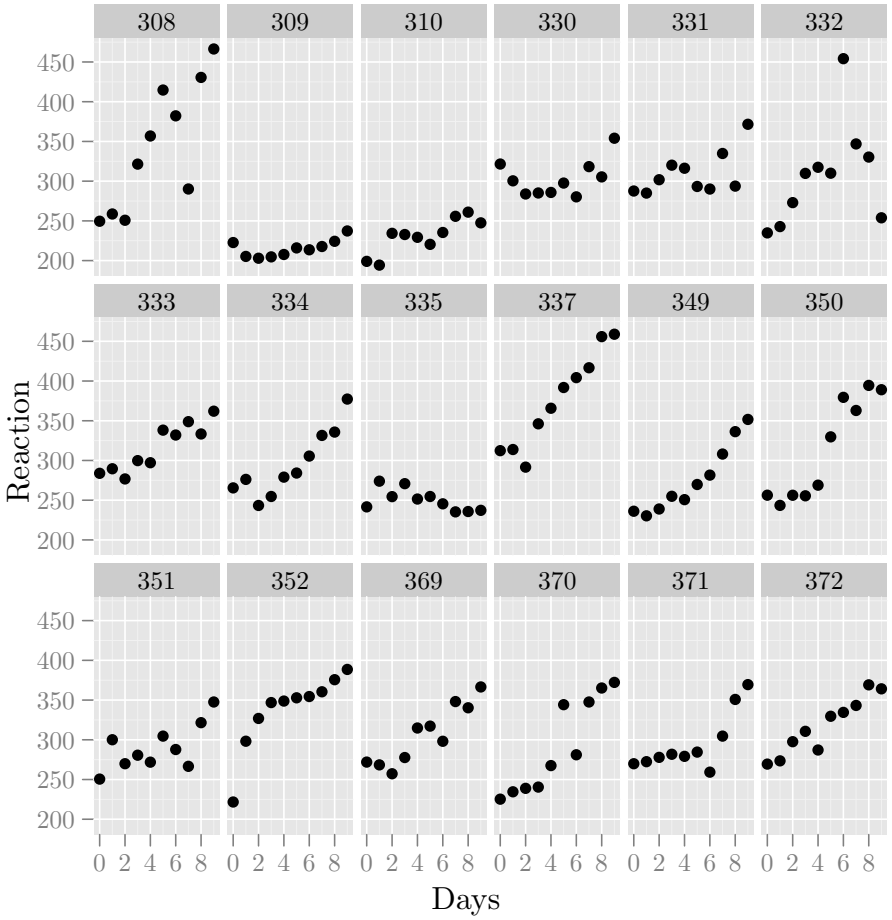
Być może wcześniej nie było jasne, czemu nie dodajemy wyrazu wolnego do nachyleń gdy chcemy poznać nachylenia. Nachylenia mówią nam, jak *zmienia się* oczekiwana wartość zmiennej zależnej gdy wartość zmiennej niezależnej wzrasta o 1. W tym przypadku wzrost oczekiwanej wartości y związany z jednostkowym przyrostem x dany jest przez:

$$\alpha_0 + \alpha_1(x+1) + \alpha_2 fB - (\alpha_0 + \alpha_1 x + \alpha_2 fB) = \alpha_1(x+1) - \alpha_1 x = \alpha_1$$

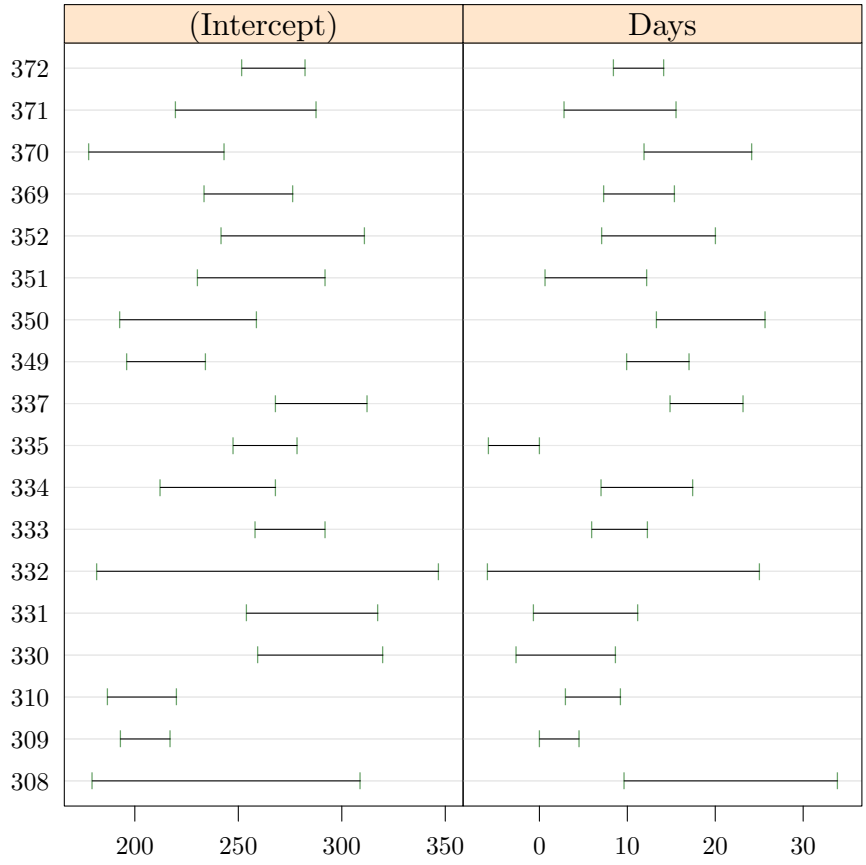
Dodajemy coś do współczynników dla nachyleń tylko wtedy, gdy odpowiednie zmienne niezależne ilościowe wchodzą w interakcje z innymi zmiennymi niezależnymi, ilościowymi lub nie. Nadszedł czas na model mieszany.

5 Liniowy model mieszany z jedną ilościową zmienną niezależną

Dane pochodzą z badania dotyczącego efektów deprywacji snu na średni czas reakcji obliczony dla zestawu testów. W dniu 0 wszyscy byli wyspani jak należy, ale już od następnej nocy mogli spać tylko po 3 godziny na dobę. Punkty danych dla wszystkich uczestników eksperymentu wyglądają tak:



Numery od 308 do 372 to identyfikatory osób badanych. Widzimy, że średni czas reakcji geleralnie rośnie w trakcie kolejnych dni, ale widzimy też, że zarówno początkowe średnie czasy reakcji (punkty przecięcia) jak i tempa wzrostu czasów reakcji (nachylenia) różnią się między osobami. Możemy dopasować sobie osobne modele liniowe z punktami przecięcia i nachyleniami dla każdej osoby. Uzyskamy wtedy po dwa współczynniki na każdy model. Dla każdego punktu przecięcia i nachylenia możemy policzyć przedziały ufności i popatrzeć jak się zmieniają:



Ewidentnie zarówno punkty przecięcia jak i nachylenia różnią się między osobami. Możemy też dopasować zwykły model liniowy z interakcją zmiennych *Days* i *Subject*:

```
Reaction ~ Days * Subject
```

```
                Wspolcz. Blad std.     p sig
(Intercept)      244.193    15.042 0.000 ***
Days              21.765     2.818 0.000 ***
Subject309       -39.138    21.272 0.068
Subject310       -40.708    21.272 0.058
Subject330        45.492    21.272 0.034   *
Subject331        41.546    21.272 0.053
Subject332        20.059    21.272 0.347
Subject333        30.826    21.272 0.149
Subject334        -4.030    21.272 0.850
Subject335        18.842    21.272 0.377
Subject337        45.911    21.272 0.033   *
Subject349       -29.081    21.272 0.174
```

```
Subject350       -18.358    21.272 0.390
Subject351        16.954    21.272 0.427
Subject352        32.179    21.272 0.133
Subject369        10.775    21.272 0.613
Subject370       -33.744    21.272 0.115
Subject371         9.443    21.272 0.658
Subject372        22.852    21.272 0.284
Days:Subject309  -19.503     3.985 0.000 ***
Days:Subject310  -15.650     3.985 0.000 ***
Days:Subject330  -18.757     3.985 0.000 ***
Days:Subject331  -16.499     3.985 0.000 ***
Days:Subject332  -12.198     3.985 0.003  **
Days:Subject333  -12.623     3.985 0.002  **
Days:Subject334   -9.512     3.985 0.018   *
Days:Subject335  -24.646     3.985 0.000 ***
Days:Subject337   -2.739     3.985 0.493
Days:Subject349   -8.271     3.985 0.040   *
Days:Subject350   -2.261     3.985 0.571
Days:Subject351  -15.331     3.985 0.000 ***
Days:Subject352   -8.198     3.985 0.041   *
Days:Subject369  -10.417     3.985 0.010  **
Days:Subject370   -3.709     3.985 0.354
Days:Subject371  -12.576     3.985 0.002  **
Days:Subject372  -10.467     3.985 0.010  **

k = 36 n =  180 blad std. reszt = 25.592
```

Dostajemy w ten sposób współczynniki szacujące punkt przecięcia dla poziomu bazowego czynnika *Subject*, czyli osoby, która została potraktowana jako bazowa (współczynnik `(Intercept)`), nachylenie linii regresji dla zmiennej *Days* u tej osoby, różnice między punktami przecięcia dla wszystkich pozostałych osób w porównaniu do osoby bazowej i różnice w nachyleniach dla wszystkich pozostałych osób w porównaniu do bazowej. Tabelka ANOVy dla takiego modelu wygląda tak:

```
Analysis of Variance Table

Response: Reaction
             Df Sum Sq Mean Sq  F value    Pr(>F)
Days          1 162703  162703 248.4234 < 2.2e-16 ***
Subject      17 250618   14742  22.5093 < 2.2e-16 ***
Days:Subject  17  60322    3548   5.4178 3.272e-09 ***
Residuals    144  94312     655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ale ta tabelka *nie* odpowiada tabelce ANOVy dla pomiarów powtarzanych, bo też zastosowany przez nas model z interakcją dni i osób nie uwzględnia faktu, że osoby można i należy potraktować jako losową próbkę z pewnej populacji. W ogóle nie będziemy się zagłębiać w zawiłości analizy wariancji dla pomiarów powtarzanych, ponieważ:

Model mieszany radzi sobie znakomicie ze wszystkim, z czym radzi sobie analiza wariancji dla pomiarów powtarzanych, ale nie odwrotnie.

Na początek warto zwrócić uwagę, że interesuje nas tutaj efekt czasu deprywacji, a różnice między osobami interesują nas znacznie mniej. Widać jednak wyraźnie, że punkty przecięcia i nachylenia różnią się znacznie między osobami. Gdy traktujemy zmienną *Subject* tak jak każdy inny czynnik, dopasowanie modelu liniowego z interakcją polega na ustaleniu optymalnych oszacowań punktów przecięcia i nachyleń *dla każdej osoby niezależnie*. W takim modelu zakładamy, że to, jakie jest nachylenie lub punkt przecięcia na przykład dla osoby numer 308 nie mówi nam *absolutnie nic* o tym, jakie jest nachylenie lub punkt przecięcia na przykład dla osoby numer 309.

Wyobraźmy sobie jednak, że w pewnym badaniu dla wszystkich spośród 1000 osób poza jedną punkty przecięcia były bliskie zeru, a dla jednej osoby punkt przecięcia był znacznie większy, w dodatku dane dla tej odstającej osoby były znacznie mocniej zaszumione niż dla pozostałych. Załóżmy także, że osoby trafiły do badania dosyć przypadkowo (stanowiły losową próbkę z populacji). Jasne jest, że odstający punkt przecięcia powinien budzić poważne wątpliwości. Co więcej, na podstawie tych wszystkich informacji możemy słusznie stwierdzić, że odstający punkt przecięcia był przypuszczalnie przeszacowany i prawdziwa wartość dla tej osoby znajduje się bliżej typowego punktu przecięcia dla pozostałych osób. Zwykły model liniowy z interakcją osób i dni deprywacji jest wobec takiego problemu bezradny, ponieważ w zwykłym modelu liniowym wszystkie współczynniki traktowane są jako niezależne wielkości.

Wartości współczynników w modelach liniowych i nie tylko nazywa się czasem *efektami*. Można więc mówić o efekcie poziomu B czynnika f (różnicy między poziomami B i A), albo o efekcie dni deprywacji (nachyleniu). Współczynniki, dla których nie zakładamy, że pochodzą z jakiegoś rozkładu, nazywa się *efektami ustalonymi* (fixed effects). Wszystkie współczynniki w rozważanych do tej pory modelach liniowych były efektami ustalonymi. Zwykle jesteśmy zainteresowani konkretnymi wartościami efektów ustalonych, na przykład jesteśmy zainteresowani tym, jak bardzo zwiększa się czas reakcji w miarę postępującej deprywacji snu.

Współczynniki, o których zakładamy, że *pochodzą z jakiegoś rozkładu*, w przypadku modeli mieszanych prawie zawsze rozkładu normalnego, nazywa się *efektami losowymi*. Często można się spotkać z definicją efektów losowych zgodnie z którą są to efekty czynników, których poziomy *nie zostały wybrane z premedytacją*. Na przykład, w badaniu deprywacji snu osoby (poziomy czynnika) zostały wybrane do badania na zasadzie mniej lub bardziej przypadkowej. W kolejnym badaniu raczej nie użylibyśmy powtórnie tych samych poziomów (czyli osób). Taka definicja efektu losowego ma sens, ale nie oddaje dobrze akurat tego, co z punktu widzenia wnioskowania statystycznego jest najważniejsze. Niebawem zademonstrujemy konsekwencje założenia, że efekty losowe są próbą z rozkładu normalnego o nieznanej wariancji.

Mieszany model liniowy tym różni się od zwykłego modelu liniowego, że zawiera nie tylko efekty ustalone, ale również efekty losowe.

Sensownie jest założyć, że punkty przecięcia dla poszczególnych osób pochodzą z rozkładu w przybliżeniu normalnego o nieznanej średniej i wariancji. Średnią tego rozkładu punktów przecięcia będzie globalny (typowy) punkt przecięcia. Jeżeli punkty przecięcia dla poszczególnych osób wyrazimy sobie jako *odchylenia* od globalnego (typowego) punktu przecięcia, to będą one pochodziły z rozkładu

normalnego o nieznanej wariancji, ale średniej równej 0. Wtedy punkt przecięcia dla konkretnej osoby będzie sumą globalnego punktu przecięcia i właściwego dla tej osoby odchylenia od wartości globalnej. Mamy więc jakiś typowy punkt przecięcia, ale każda osoba może mieć trochę inny. *To samo możemy zrobić z każdym innym efektem ustalonym, który daje się oszacować dla poszczególnych osób*. Gdyby na przykład osoby wykonywały wiele prób zadania z dwoma warunkami, mielibyśmy efekt ustalony warunku, który dawałby się oszacować dla każdej z osób. Moglibyśmy wtedy założyć, że istnieje jakiś typowy efekt tego warunku, ale u każdej osoby może być nieco inny. Tabelka dla modelu mieszanego, w którym mamy ustalony efekt dni deprywacji i zakładamy normalny rozkład odchyleń punktów przecięcia dla poszczególnych osób wygląda tak:

```
Random effects

 Groups    Name        Variance  SD
  Subject  Intercept   1378.18     37.12
  Residual               960.46     30.99

Fixed effects

            Coef. Coef. SE     t     p sig
Intercept 251.41     9.75 25.80 0.000 ***
Days       10.47     0.80 13.02 0.000 ***


t df:  161
```

W górnej tabelce mamy oszacowanie wariancji odchyleń punktów przecięcia, czyli naszych efektów losowych i oszacowanie wariancji reszt. Jak widać, zmienność punktu przecięcia między osobami jest znacznie większa niż zmienność wyników resztowych. Mamy też dwa efekty ustalone, to jest globalny punkt przecięcia i globalne nachylenie linii regreji dla zmiennej *Days*. Wygląda na to, że na początku eksperymentu średni czas reakcji (po wszystkich osobach) wynosił 251.41 i z każdym kolejnym dniem deprywacji rósł przeciętnie o 10.47 milisekundy.

Tabelka nie informuje nas w żaden sposób o wartościach punktów przecięcia dla poszczególnych osób, otrzymujemy tylko oszacowanie ich wariancji. Dzieje się tak, ponieważ konkretne wartości efektów losowych zwykle nas nie interesują. Efekty losowe to zwykle coś, co chcemy i powinniśmy kontrolować, a nie coś, co chcemy dokładniej poznać, chociaż zdarzają się wyjątki od tej reguły.

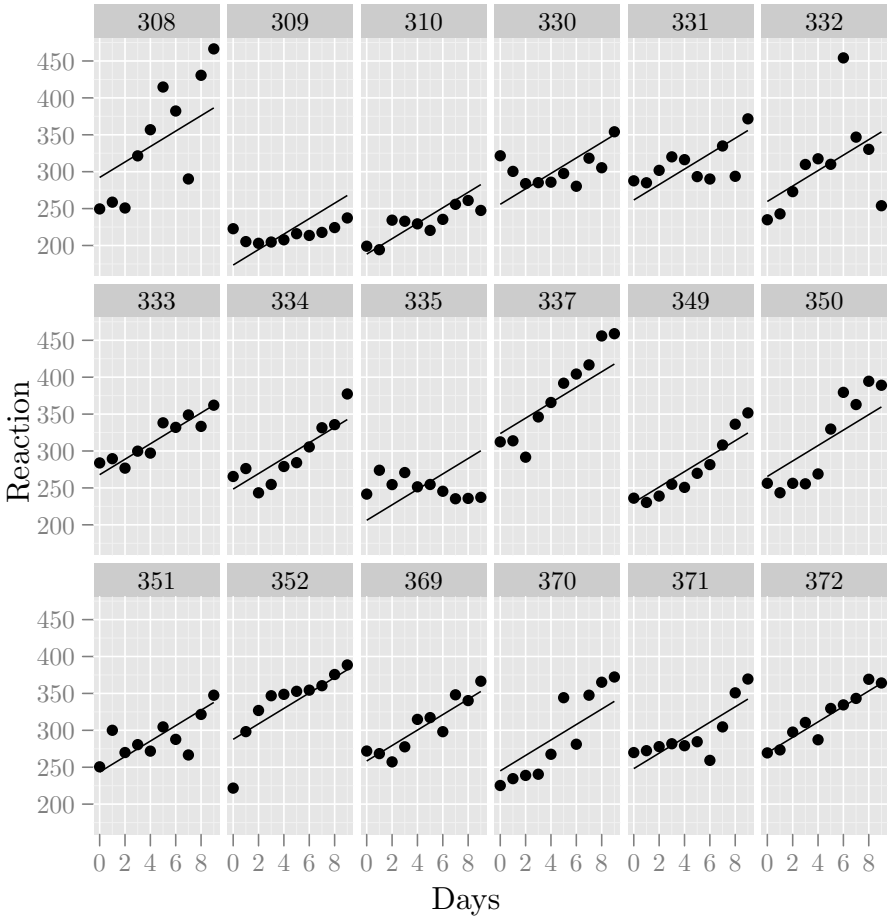
W przypadku tych danych musimy wyrazić w modelu *co najmniej* zmienność punktu przecięcia między osobami, inaczej model byłby *zupełnie błędny*. Byłby błędny, ponieważ jeśli uwzględnimy tylko jedną zmienną niezależną, to jest liczbę dni deprywacji, wyniki resztowe nie będą z pewnością statystycznie niezależne, a przecież w modelu liniowym zakładamy, że są statystycznie niezależne. Reszty nie będą statystycznie niezależne, ponieważ nawet jeśli poznamy ich wariancję (średnia reszt jest z konieczności równa 0), to i tak wartości poszczególnych reszt będą nam mówiły coś o wartościach innych reszt. Na przykład, dowolna reszta dla osoby numer 309 mówi nam coś, nawet jeśli niewiele, o pozostałych resztach dla tej osoby, ponieważ wszystkie reszty dla tej osoby będą relatywnie

duże lub małe zależnie od tego, czy ta osoba reagowała przeciętnie stosunkowo szybko czy wolno. Jeżeli w modelu zakładamy, że reszty są niezależne, a w istocie nie są, to błędnie szacujemy zmienność reszt. Wynika to stąd, że niezależne próbki z jakiegokolwiek rozkładu mówią nam więcej o tym rozkładzie niż próbki skorelowane. Próbki skorelowane mówią trochę o rozkładzie, a trochę o sobie nawzajem. Jeżeli źle szacujemy zmienność reszt, to źle szacujemy błędy standardowe współczynników i źle obliczamy poziomy istotności związane ze współczynnikami.

Aby wyprowadzić poprawne wnioski w modelu mieszanym musimy uwzględnić wszystkie ważne źródła zmienności reprezentowane w danych.

Pojawia się pytanie, które źródła zmienności koniecznie trzeba uwzględnić, a których nie trzeba. Zwykle źródła zmienności dzielimy na ważne i nieważne na podstawie porównania dopasowania modeli mieszanych, które te źródła uwzględniają, z modelami, które ich nie uwzględniają.

Możemy teraz ocenić dopasowanie naszego wstępnego modelu:

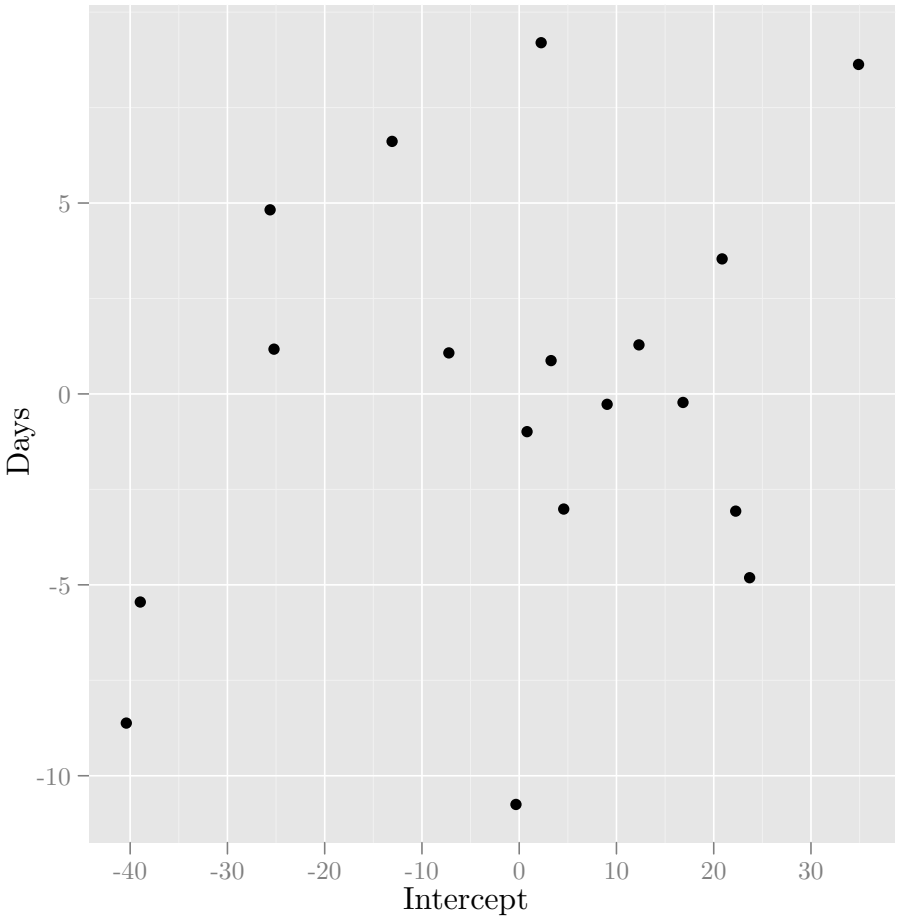


Od razu widać, co tu nie gra. Model radzi sobie tylko z tymi osobami, dla których nachylenie jest zbliżone do nachylenia globalnego (uśrednionego po wszystkich osobach). Dodanie zmienności punktów przecięcia z pewnością tutaj nie

wystarcza. Jeżeli uwzględnimy dodatkowo zmienność nachyleń uzyskamy model:

```
Random effects

 Groups    Name        Variance  SD        Correlations
  Subject  Intercept    612.09     24.74
           Days          35.07      5.92      0.07
  Residual              654.94     25.59

Fixed effects

            Coef. Coef. SE     t     p sig
Intercept 251.41     6.82 36.84 0.000 ***
Days       10.47     1.55  6.77 0.000 ***


t df:  144
```

Warto zwrócić uwagę, że efekty ustalone nie uległy zmianie na skutek wprowadzenia zmienności nachyleń. Może się jednak zdarzyć i faktycznie często się zdarza, że oszacowania efektów ustalonych ulegają zmienia na skutek wprowadzenia do modelu dodatkowych efektów losowych. Wariancja nachyleń jest mniejsza niż wariancja punktów przecięcia (górna tabelka o nazwie Random effects), ale model z losowymi punktami przecięcia i nachyleniami pasuje istotnie lepiej niż model bez losowych nachyleń. Ponieważ punkty przecięcia i nachylenia dla poszczególnych osób traktujemy tutaj jako zmienne losowe i każda z tych zmiennych ma swój (normalny) rozkład prawdopodobieństwa, możemy dopuścić (a więc wyrazić w modelu i oszacować) ich korelację. Ta korelacja została oszacowana na 0.07. Patrzymy na efekty losowe dla puntków przecięcia i nachyleń:



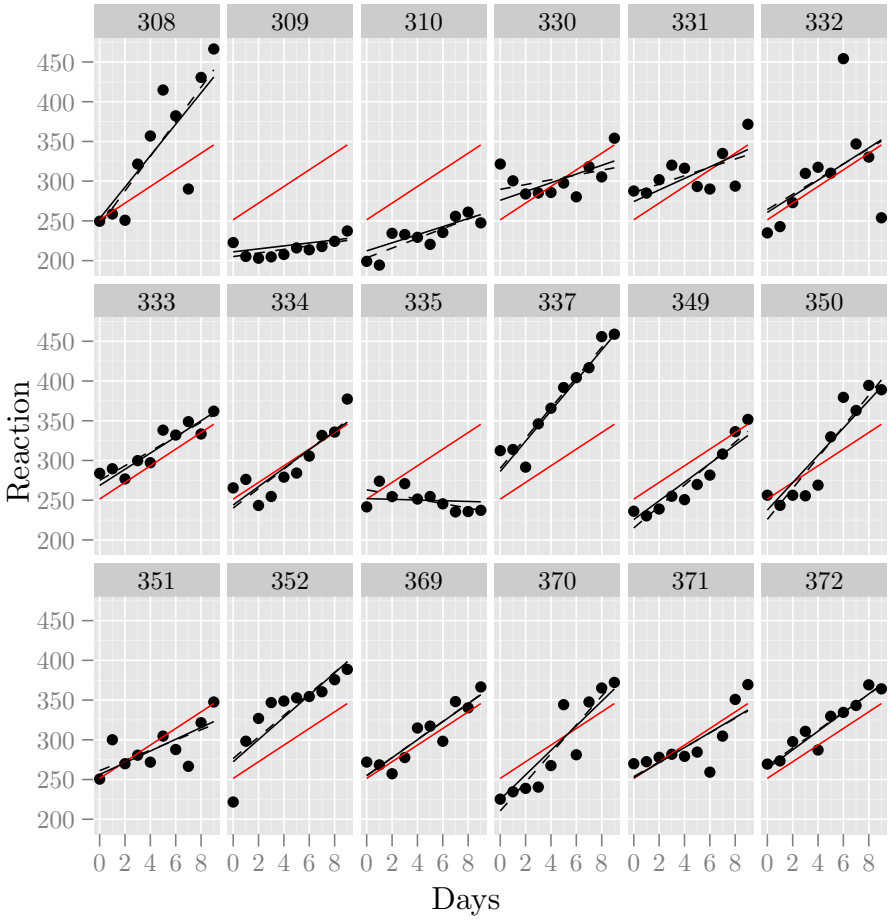
i widzimy, że nie są jakoś szczególnie skorelowane. Oczywiście, mogłyby być. Mogłoby być na przykład tak, że im dłuższy średni czas reakcji w warunkach bez deprywacji (dzień 0, czyli punkt przecięcia), tym silniejszy efekt deprywacji (nachylenie dla zmiennej *Days*). Najwyraźniej tak jednak nie jest, albo nie widać tego w naszej próbie. Widzimy, że na wykresie średnia losowych punktów przecięcia i nachyleń jest równa zero. Dzieje się tak, ponieważ w modelu mieszanym efekty losowe to odchylenia od efektu ustalonego (tutaj globalnego punktu przecięcia lub nachylenia) związane z poziomami czynnika grupującego (tutaj tym czynnikiem jest tożsamość osoby badanej). Inny przykład: moglibyśmy badać efekt metody nauczania (czynnik o dwóch poziomach) w różnych szkołach i różnych klasach (dwa czynniki grupujące). Szkoły i klasy można spokojnie potraktować jako próbki z pewnej populacji. W ewentualnym następnym badaniu nie powtórzylibyśmy tych samych szkół i klas z premedytacją. Nie można wykluczyć, że efekt metody zmienia się nieco między klasami i między szkołami. Każda klasa i każda szkoła miałyby więc swoje *odchylenie* (stąd średnia zero) od ogólnego-typowego-średniego-globalnego-ustalonego efektu metody. Wracamy do deprywacji snu i porównujemy model, w którym dopuszczalna jest korelacja efektów losowych z modelem, w którym nie dopuszczamy takiej korelacji:

```
Data: sleepstudy
Models:
lmer0: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
lmer2: Reaction ~ Days + (Days | Subject)
      Df  AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmer0  5 1762 1778.0 -876.02
lmer2  6 1764 1783.1 -875.99 0.0609      1     0.8051
```

i widzimy, że korelacja nic nie wnosi. Porównujemy jeszcze model, który właśnie wygrał z modelem, w którym nachylenia nie mogą się zmieniać między osobami:

```
Data: sleepstudy
Models:
lmer00: Reaction ~ Days + (1 | Subject)
lmer0: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
       Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
lmer00  4 1802.1 1814.9 -897.05
lmer0   5 1762.0 1778.0 -876.02 42.053      1  8.885e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i stanowczo odrzucamy prostszy model, w którym zmienia się tylko punkt przecięcia. Patrzymy na predykcje zwycięskiego modelu mieszanego bez korelacji (linia ciągła), globalny punkt przecięcia i nachylenie (linia czerwona) i predykcje zwykłego modelu liniowego z interakcją osób i dni (linia przerywana) dla wszystkich osób:

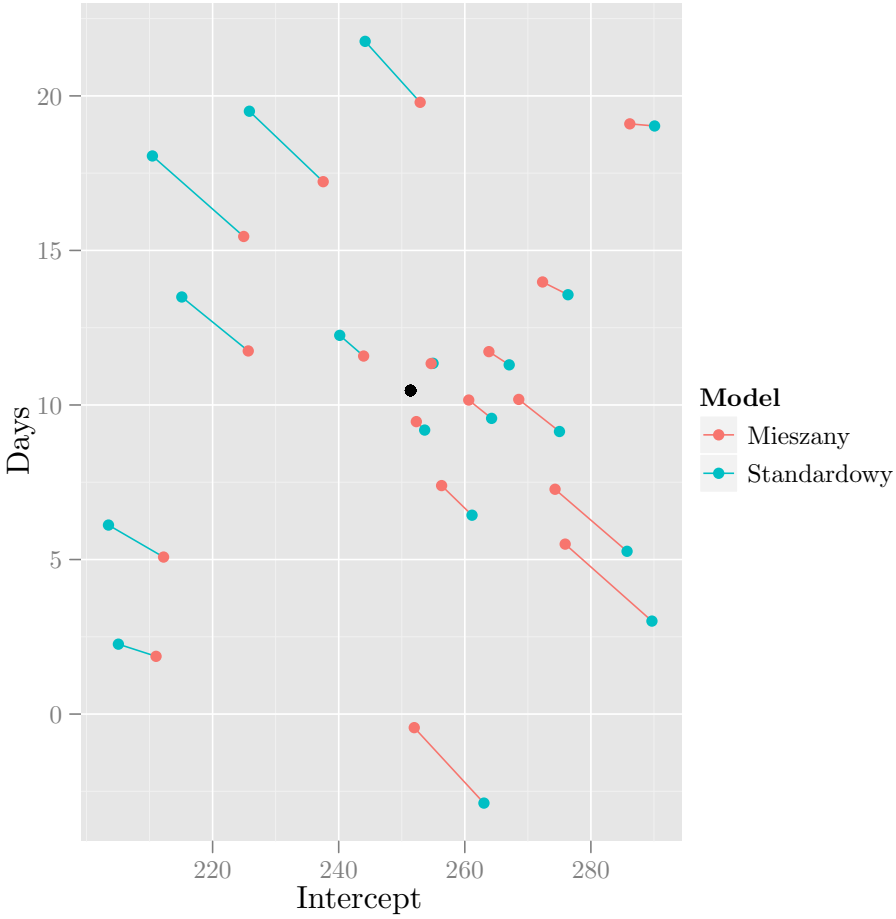


Widzimy, że predykcje obydwu modeli są podobne, ale nie identyczne. Ponieważ w modelu mieszanym zakładamy, że efekty losowe punktów przecięcia i nachyleń pochodzą z dwóch, odrębnych (nieskorelowanych) rozkładów normalnych, to wartości losowych przecięć mówią coś o sobie nawzajem (bo każde z nich dostarcza informacji na temat wariancji przecięć) i wartości losowych nachyleń mówią coś o sobie nawzajem, ale wartości losowych przecięć nie mówią nam nic o wartościach losowych nachyleń (brak korelacji między przecięciami i nachyleniami).

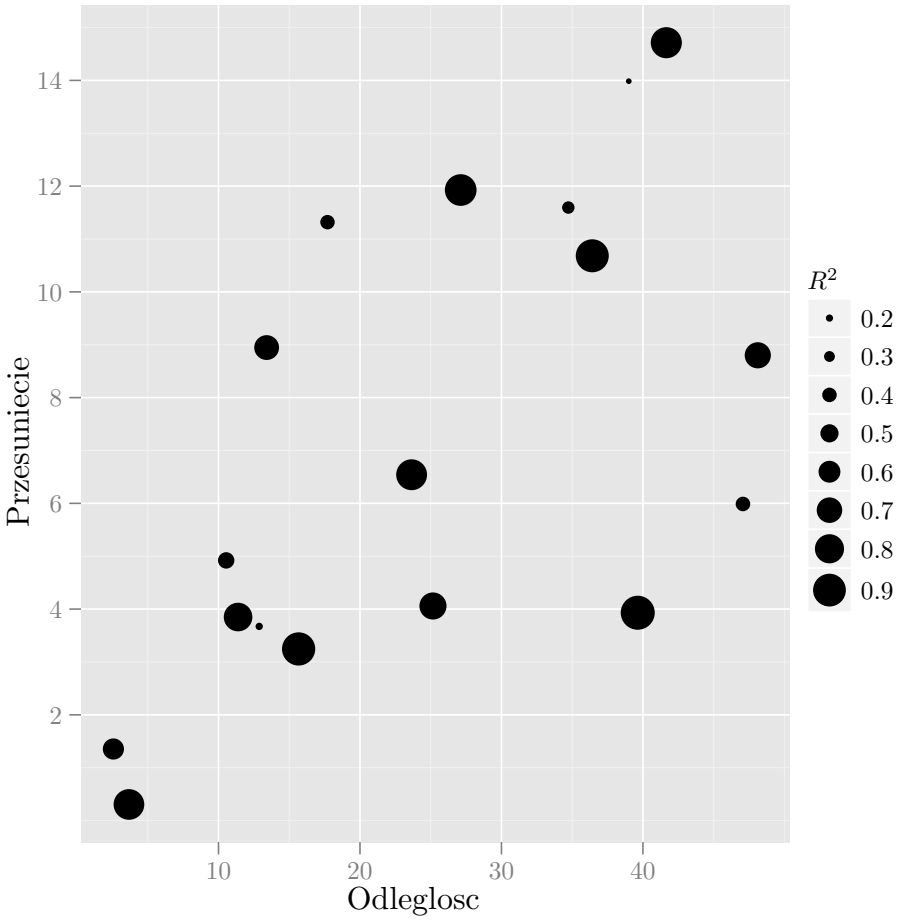
Każdy co najmniej dwuelementowy podzbiór odchyleń punktów przecięcia mówi coś o tym, jaka jest wariancja punktów przecięcia. Wszystko, co możemy wywnioskować na temat średniej i wariancji tych efektów informuje nas jednocześnie o tym, jak powinny wyglądać pozostałe efekty. Im bardziej efekt losowy odstaje od rozkładu pozostałych lub im większa jest niepewność związana z oszacowaniem danego efektu losowego (tutaj efektu losowego dla konkretnej osoby), tym bardziej jego oszacowanie będzie *przesunięte* w stronę wartości globalnej. Ponieważ efekty losowe są wyrażone jako odchylenia od wartości globalnej, to przesunięcie w kierunku wartości globalnej będzie polegało na ich *skurczeniu* (zbliżeniu do zera). Istotą zjawiska kurczenia się efektów losowych jest korygo-

wanie oszacowań efektów losowych ze względu na to, jak bardzo wydają się być nietypowe i jak bardzo wydają się być niepewne, zjawisko to jest więc teoretycznie uzasadnioną i porządaną własnością modeli mieszanych.

Spróbujemy zobrazować zjawisko kurczenia się efektów losowych. Porównamy w tym celu punkty przecięcia i nachylenia z modelu mieszanego (uzyskane przez dodanie odpowiednich odchyleń do odpowiednich wartości globalnych) z punktami przecięcia i nachyleniami ze zwykłego modelu liniowego zawierającego interakcję osób i dni.



Czarny punkt to globalny punkt przecięcia i globalne nachylenie. Efekty w modelu mieszanym są wszystkie mniej lub bardziej przesunięte w stronę wartości globalnej w porównaniu do efektów w modelu liniowym ze zwykłą interakcją osób i dni. Każde takie przesunięcie to wektor. Możemy policzyć sobie długość tego wektora i zobaczyć, jak wielkość tego przesunięcia ma się do odległości efektu ze zwykłego modelu liniowego od wartości globalnej.



Gdyby procent wariancji wyjaśnionej dla każdej z osób miał znaczący wpływ na skurczenie się efektów losowych, na wykresie kółka znajdujące się niżej byłyby większe niż kółka znajdujące się wyżej. Najwyraźniej decydujący wpływ ma tutaj odległość od wartości globalnej, ale nie specyficzna dla osób wariancja resztowa. Nasz model nie może jednak konsekwentnie uwzględniać międzyosobowych różnic w wariancji reszt, ponieważ nasz model zakłada, że *wszystkie reszty pochodzą z tego samego rozkładu*, a więc wszystkie reszty mają z założenia tą samą wariancję. Ważnym czynnikiem odpowiedzialnym za stopień skurczenia się efektów losowych, który w tym wypadku nie występuje, jest również zróżnicowana liczba obserwacji dla każdego poziomu czynnika grupującego. Gdyby na przykład niektóre osoby miały znacznie mniej punktów danych, efekty losowe tych osób byłyby silniej zbliżone do zera.

Został nam jeszcze jeden kłopot. Na wykresie obrazującym dopasowanie modelu dla wszystkich osób widać coś niepokojącego. Czasy reakcji nie są przypadkowo rozrzucone wokół linii regresji, tylko często tworzą mniej lub bardziej regularne krzywe, na oko sinusoidalne. Tego rodzaju zjawiska mogą się pojawić gdy dane mają charakter szeregów czasowych. Przypuszczalnie każdy ma lepsze okresy, które przechodzą stopniowo w okresy gorsze, które znowu prze-

chodzą stopniowo w lepsze, i tak dalej. To oznacza, że reszty w naszym modelu mieszanym nadal nie są statystycznie niezależne, ponieważ mówią coś o sobie nawzajem. Niestety, żeby uwzględnić tą korelację reszt w modelu mieszanym musielibyśmy dodać do modelu proces autoregresyjny, a tego nie da się zrobić za pomocą aktualnej wersji flagowego pakietu R służącego do dopasowywania modeli mieszanych (lme4), chociaż można to zrobić za pomocą starszego pakietu (nlme).

Spis tabel

Spis rysunków