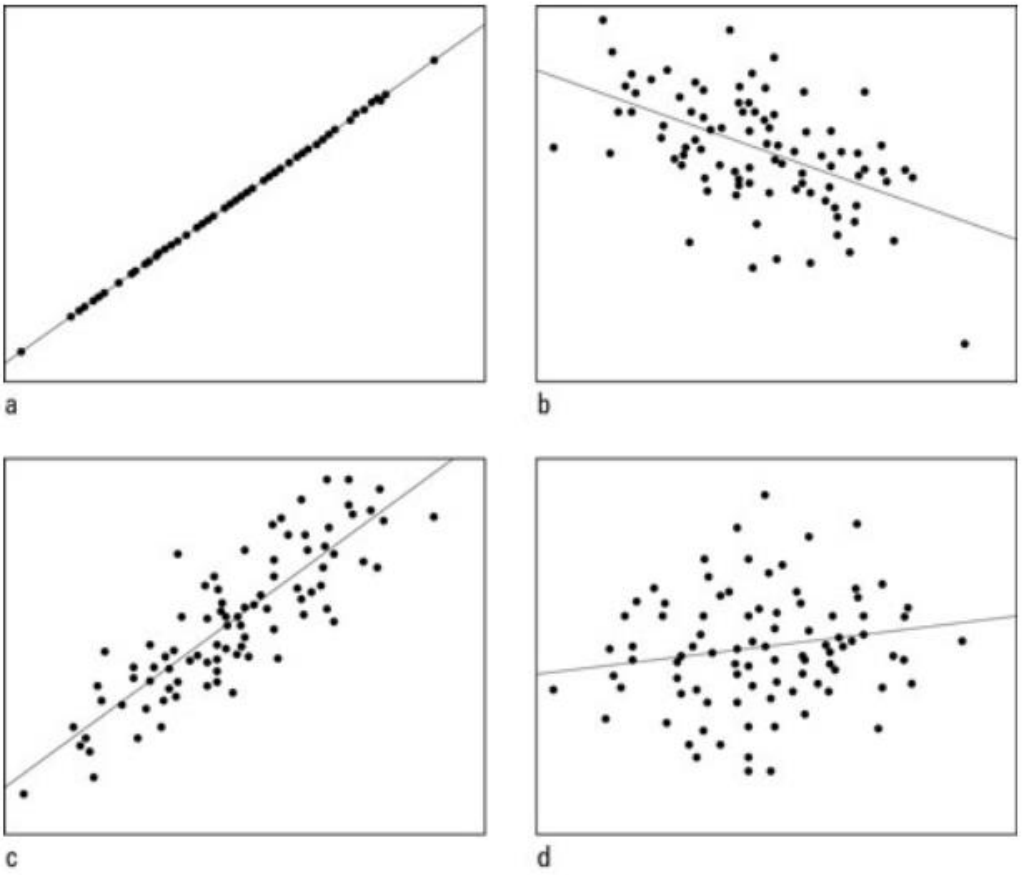
KORELACJA

1. Wykres rozrzutu – ocena związku między zmiennymi X i Y



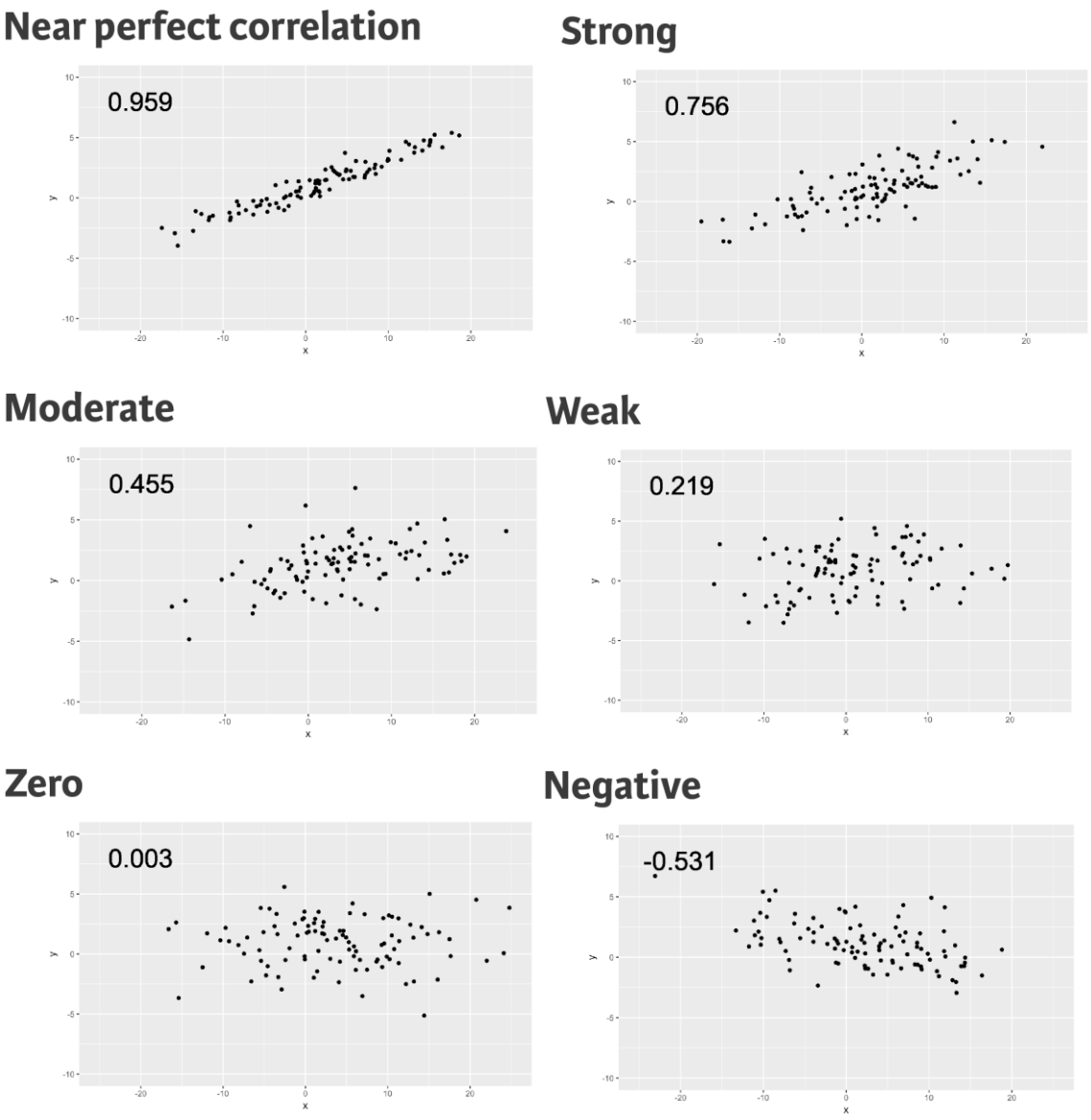
Scatterplots with correlations of a) +1.00; b) –0.50; c) +0.85; and d) +0.15.

2. Współczynnik korelacji Pearsona

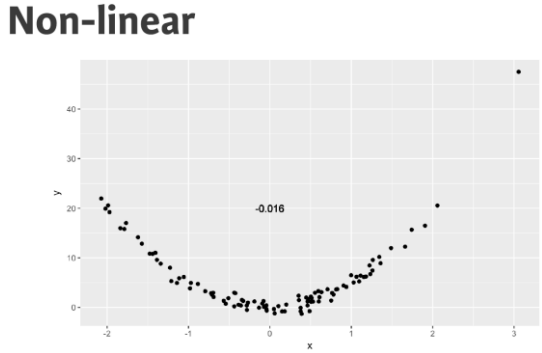
$$r = \frac{1}{n-1}\left(\frac{\sum_x \sum_y (x-\bar{x})(y-\bar{y})}{s_x s_y}\right)$$

$$s_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}.$$

3. Siła i kierunek związku między zmiennymi

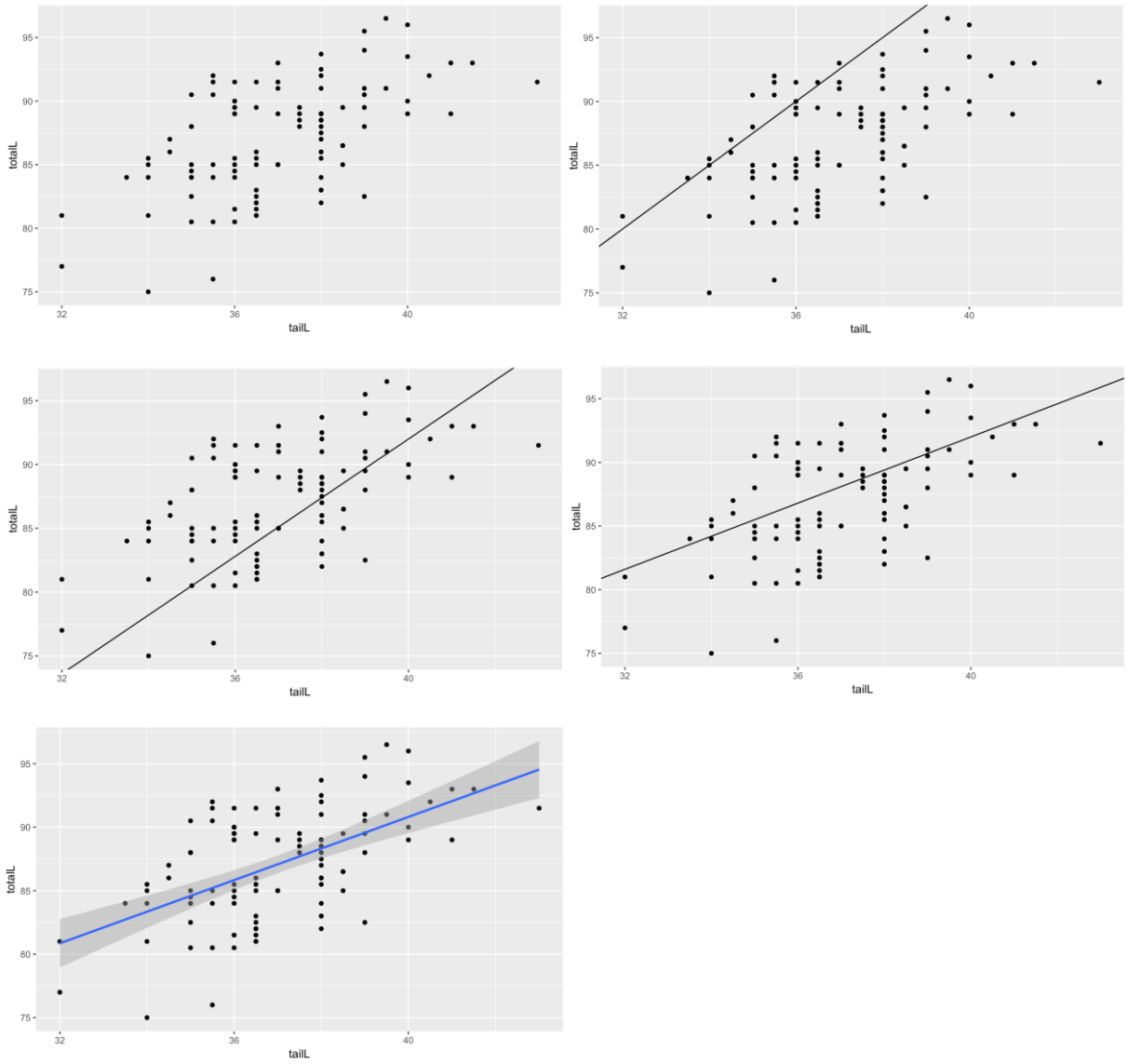


4. Korelacja „ma sens”, tylko wtedy, gdy związek między zmiennymi ma charakter liniowy (zob. tzw. kwartet Anscombe’a).



PROSTA REGRESJA LINIOWA (Simple linear regression)

1. Związek między X i Y można opisać za pomocą równania prostej.



2. Prosty model liniowy

a) postać ogólna – równanie regresji

response = f(explanatory) + noise

*response = intercept + (slope * explanatory) + noise*

b) model regresji z jedną zmienną niezależną

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

c) wartości dopasowane (ang. fitted values)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

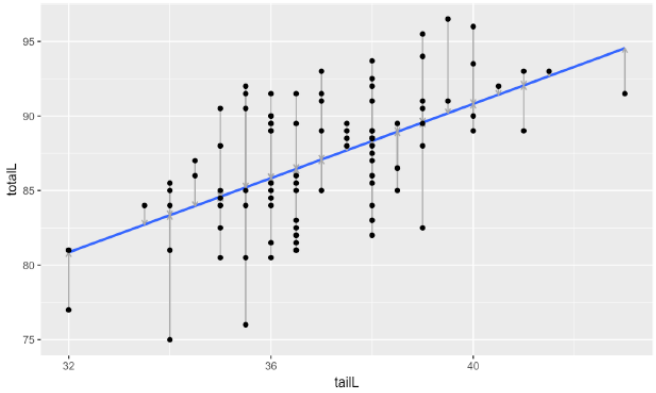
$$\widehat{wgt} = -105.011 + 1.018 \cdot hgt$$

d) reszty

$$e = Y - \hat{Y}$$

e) procedura dopasowania modelu (metoda najmniejszych kwadratów)

- Given n observations of pairs (x_i, y_i)...
- Find $\hat{\beta}_0, \hat{\beta}_1$ that minimize $\sum_{i=1}^{n} e_i^2$



3. Podsumowanie modelu

```
> # Show the coefficients
> coef(mod)
(Intercept)         hgt
-105.011254    1.017617
>
> # Show the full output
> summary(mod)

Call:
lm(formula = wgt ~ hgt, data = bdims)

Residuals:
    Min      1Q  Median      3Q     Max
-18.743  -6.402  -1.231   5.059  41.103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -105.01125    7.53941  -13.93   <2e-16 ***
hgt            1.01762    0.04399   23.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 9.308 on 505 degrees of freedom
Multiple R-squared:  0.5145,  Adjusted R-squared:  0.5136
F-statistic: 535.2 on 1 and 505 DF,  p-value: < 2.2e-16
>
```

a) znaczenie *std. error*, testowanie hipotez (zob. str. 89 SPSS Guidebook)

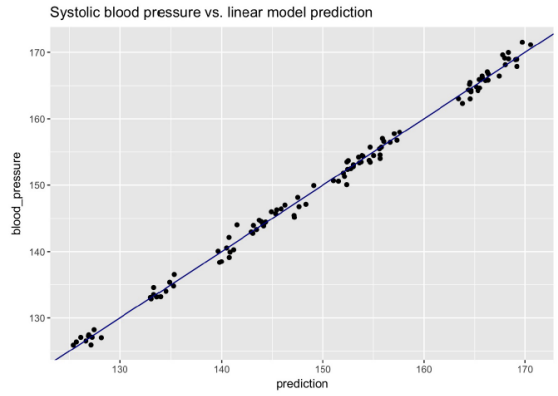
H_0: = 0 (the slope/intercept is zero; there is no linear relationship between the variables).
H_a: $\neq 0$ (the slope/intercept is not zero; there is a linear relationship between the variables).

Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x} - \mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown

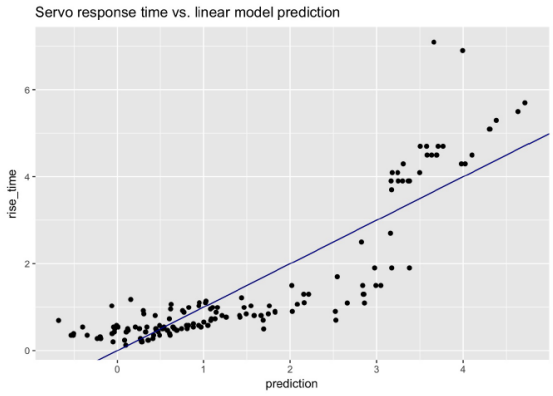
b) ocena modelu

- SSE – sum of squared errors
- RMSE – root mean squared error, RSE – residual standard error
- R squared, adjusted R squared
- null model
- residual plot – wykres reszt

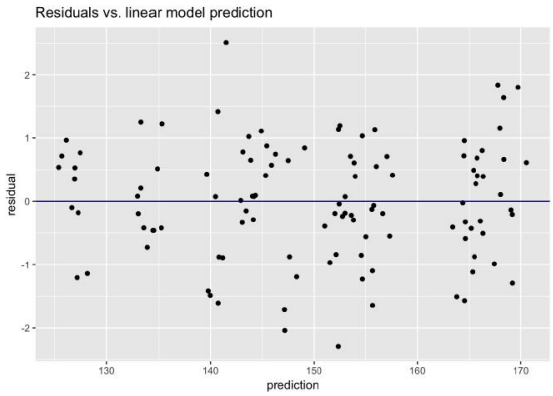
A well fitting model



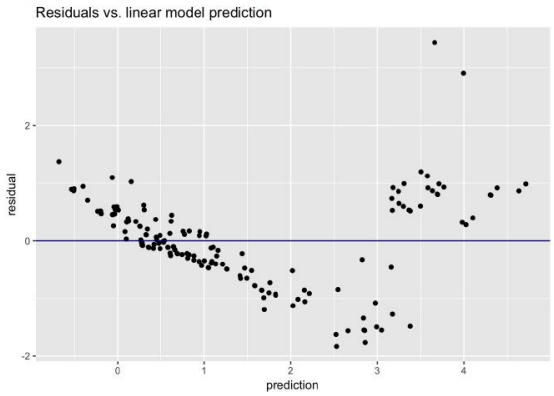
A poorly fitting model



A well fitting model

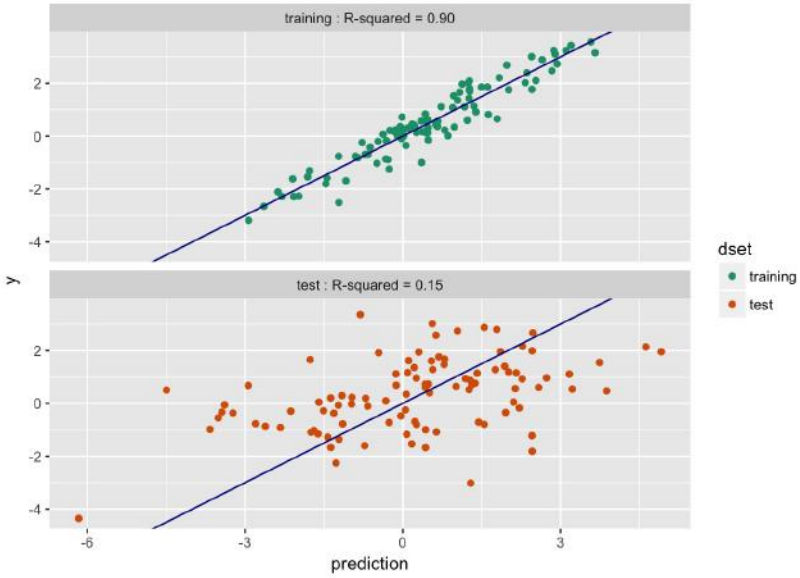


A poorly fitting model

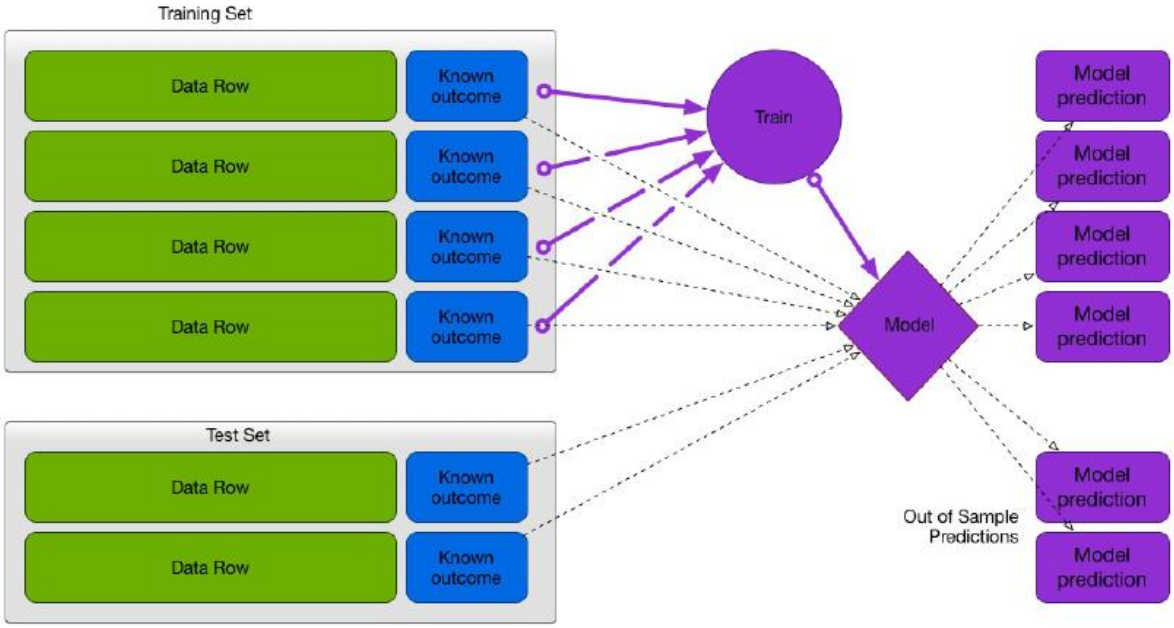


4. Jeżeli model ma zastosowanie wyłącznie w celu zbadania zależności między zmiennymi, to w tym punkcie można zakończyć jego podsumowanie. Jednak jeżeli model ma byc wykorzystywany w celu estymacji wartości zmiennej Y dla nowych wartości zmiennej X, to istotne dla oceny przydatności tego modelu dla tego zastosowania jest sprawdzenie, jak model poradzi sobie w takiej sytuacji.

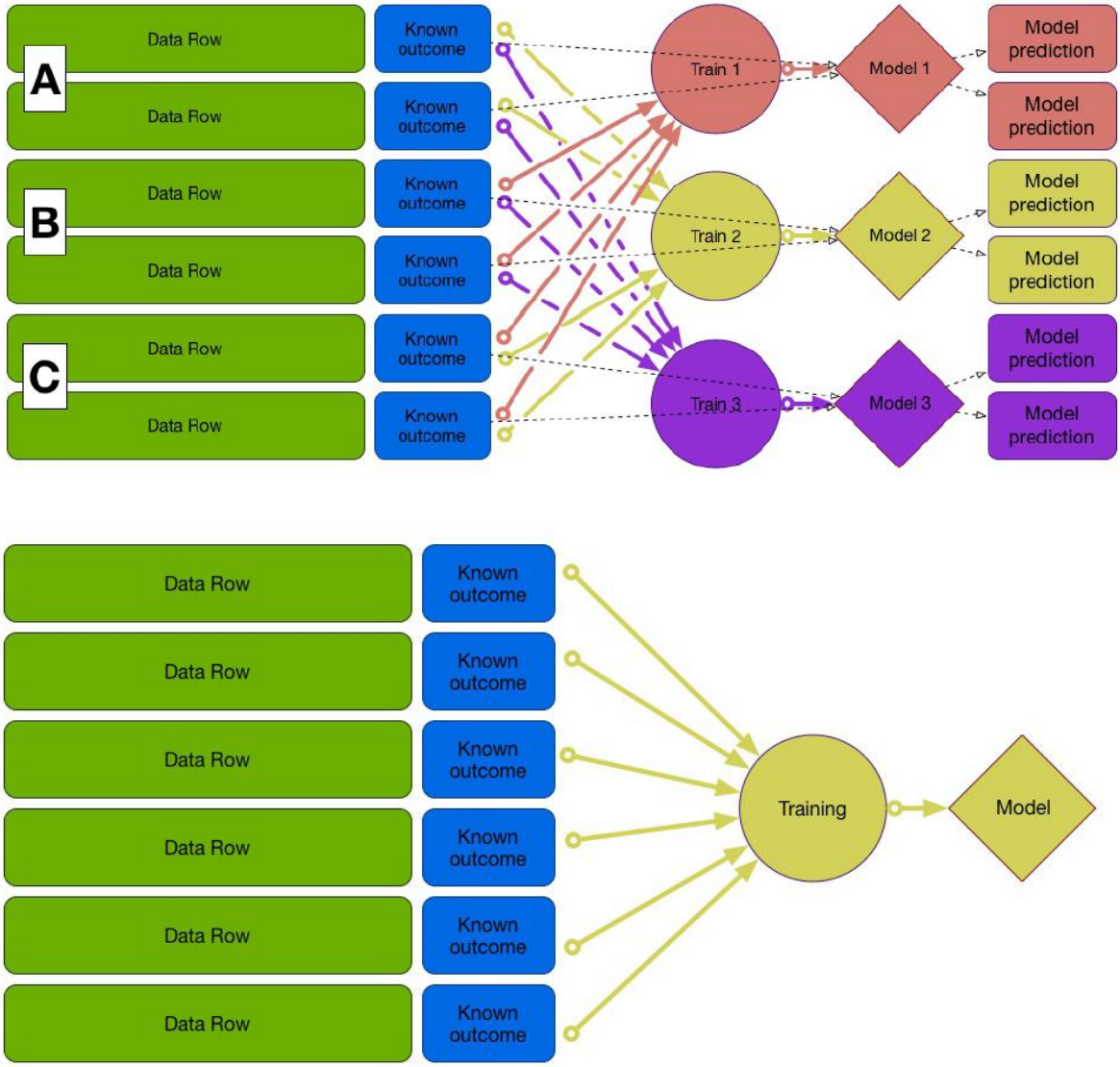
a) model może bardzo dobrze radzić sobie na danych uczących, ale bardzo słabo na danych testowych (-> ang. overfitting)



b) podział na zbiór uczący i testowy



c) walidacja krzyżowa (ang. cross-validation)



Porównanie wyników rzeczywistego modelu na zbiorze uczącym, testowym oraz w walidacji krzyżowej (materiały Data Camp, *Supervised learning in R: Regression*):

Measure type	RMSE	R^2
train	0.7082675	0.8029275
test	0.9349416	0.7451896
cross-validation	0.8175714	0.7635331

5. Założenia (str. 89 SPSS Guidebook):

1. *Normality.* The population of *Y* values for each *X* is normally distributed.

2. *Equal variances.* The populations in Assumption 1 all have the same variance.

3. *Independence.* The dependent variables used in the computation of the regression equation are independent. This typically means that each observed *X*-*Y* pair of observations must be from a separate subject or entity.

REGRESJA WIELORAKA (Multiple linear regression)

1. Postać ogólna – równanie regresji

response = f(explanatory$_1$) + f(explanatory$_2$) + f(explanatory$_n$) + … + noise

*response = intercept + (slope$_1$ * explanatory$_1$) + (slope$_2$ * explanatory$_2$) + (slope$_3$ * explanatory$_3$) + … + noise*

2. Model regresji z wieloma zmiennymi niezależnymi

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \varepsilon$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_i \cdot X$$

H_0: $\beta_i = 0$ (the intercept and slope of the *i* variable is zero; there is no linear relationship between the variables).
H_a: $\beta_i \neq 0$ (the intercept and slope of the *i* variable is not zero; there is a linear relationship between the variables).

Dodatkowo podawana jest wartość statystyki F z testu ANOVA, w którym testowana jest następująca hipoteza zerowa (zob. str. 98 Guidebook SPSS):

H_0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_i = 0$ (there is no linear relationship between the dependent variable and the collection of independent variables).

H_a: At least one of the β_i s is nonzero (there is a linear relationship between the dependent variable and at least one of the independent variables).

3. Podsumowanie przykładowego modelu

```
> fmla.abs
Income2005 ~ Arith + Word + Parag + Math + AFQT
>
> summary(model.abs)

Call:
lm(formula = fmla.abs, data = income_train)

Residuals:
   Min     1Q Median     3Q    Max
-78728 -24137  -6979  11964 648573
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17516.7     6420.1   2.728  0.00642 **
Arith         1552.3      303.4   5.116 3.41e-07 ***
Word          -132.3      265.0  -0.499  0.61754
Parag        -1155.1      618.3  -1.868  0.06189 .
Math           725.5      372.0   1.950  0.05127 .
AFQT           177.8      144.1   1.234  0.21734
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45500 on 2063 degrees of freedom
Multiple R-squared:  0.1165,  Adjusted R-squared:  0.1144
F-statistic:  54.4 on 5 and 2063 DF,  p-value: < 2.2e-16
```

4. Jakie są różnice między regresją prostą a wieloraką?
   a) tzw. *additive terms* (chodzi o to, że y jest modelowany jako suma iksów)
   b) interakcje – wielkość zmiany Y przy zmianie jednego z X o jedną jednostkę zależy od zmiany/poziomu jakiegoś innego X
   c) zmienne kategorialne – konieczne jest ich odpowiednie zakodowanie

[Dwuczynnikowa ANOVA – bez interakcji i z interakcjami]

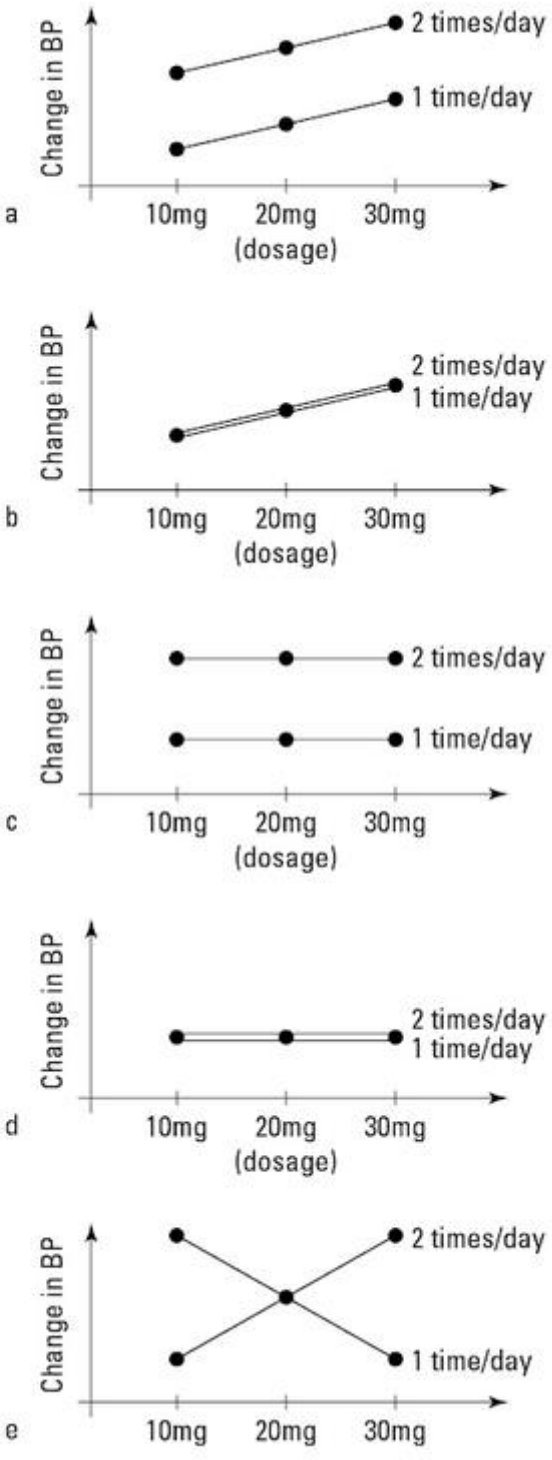


Figure 11-1: Five examples of the results from a two-way ANOVA with interaction.