**Kurs: Rachunek prawdopodobieństwa i statystyka**

**Instrukcja do ćwiczeń 7:**

- Analiza regresji i korelacji liniowej

**Potrzebne funkcje**

**Funkcje do tworzenia wykresów**

- **plot(**x=X, y=Y**)**

- **ggplot(**data=My\_Data, aes(x = X, y = Y)**) + geompoint() + stat\_smooth(**method = lm**)**

gdzie X i Y to są odpowiednie wektory danych.

**- scatterplot3d(**x,y,z, *inne opcje*) – do tworzenia wykresów punktów w przestrzeni trójwymiarowej. Pierwszy parametr (x) określa „frontową” oś poziomą (ale może też być macierzą, określającą wszystkie trzy osie), drugi parametr (y) określa drugą oś poziomą (idącą „w głąb”), a trzeci parametr (z) określa oś pionową (zwykle to jest zmienna zależna). Dalsze opcje określają rodzaj punktów, kolory, opisy osi itd. W wersji domyślne dają wykres wystarczający na potrzeby

- **sactterplot3d$plane3d**(*Output\_z\_funkcji\_lm*) – dodaje do wykresu obraz płaszczyzny, będącej wynikiem dopasowania płaszczyzny regresji z modelu z dwoma predyktorami.

**Funkcje do wykonywania analiz i testów statystycznych**

**Statystyki opisowe:**

- **summary()**, **summarize()** – do przedstawienia statystyk opisowych

**Analiza regresji - oszacowanie parametrów modelu liniowego oraz statystyki dotyczące tych oszacowań:**

**- lm()** - standardowa funkcja R do wykonania analizy regresji liniowej. Wywołanie:

**Model\_output <- lm(*Y* ~ *X1 + X3 + …*, data = *My\_data*)**

gdzie *Y* to ilościowa zmienna zależna, a *X1, X2,* … to predyktory ilościowe (zmienne niezależne). Wynikowy obiekt Model\_output przedstawia na wydruku tylko oszacowania współczynników regresji, ale obiekt ten zawiera dane, na podstawie których generowane są pełne wyniki:

- **summary**(Model\_output) - tworzenie tabeli z oszacowaniami współczynników regresji, ich błędami standardowymi i testami istotności, oraz dodatkowymi wynikami: błędem standardowym wartości resztowych, testem F istotności całego modelu, współczynnikiem determinacji R2 (kwadrat współczynnika korelacji);

- **confint**(Model\_output) - raportuje przedziały ufności dla współczynników regresji;

- wykresy i analizy diagnostyczne związane z wynikami modelu liniowego:

**- hist**(residuals(Model\_output)) – histogram rozkładu wartości resztowych;

- **plot**(fitted(Model\_output), residuals(Model\_output)), albo lepiej

- **plot**(Model\_output, *Liczba*) – tworzy różne wykresy, przy czym *Liczba* (1-5) wskazuje na rodzaj generowanego wykresu (1: residual vs. fitted, 2: quantile plot, 3: Cook;s distance).

- **shapiro.test()** – do badania normalności rozkładu reszt.

- **lm()** z opcją “**offset()**” pozwala na testowanie hipotezy zerowej, że współczynnik regresji jest równy jakiejś konkretnej wartości, innej niż zero (co jest wartością domyślną). Na przykład, żeby testować hipotezę zerową, że nachylenie prostej regresji jest równe 1, należy oszacować model:

**lm(*Y* ~ *X* + offset(*1\*X*), data = *My\_data*)**

Raportowany współczynnik nachylenia dla zmiennej X oznaczać będzie wartość, o jaką nachylenie różni się od 1. Test *t* Studenta istotności tego współczynnika też będzie informował, czy różni się od 1. Więcej o możliwości testowania specyficznych hipotez zerowych:

<https://stackoverflow.com/questions/33060601/test-if-the-slope-in-simple-linear-regression-equals-to-a-given-constant-in-r>

**Analiza korelacji**

- **cor.test**() – standardowa funkcja R do obliczania współczynnika Pearsona:

- **cor.test**(werktor\_1, wektor\_2)

- albo: **cor.test**(~ X1 + X2; data = My\_data)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**ZADANIA**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Zadanie 1. Taniny: wpływ zawartości tanin na tempo wzrostu larw motyla**

**Problem badawczy:** rośliny produkują rożnego rodzaju substancje, utrudniające zjadanie ich tkanek przez roślinożerców oraz ataki grzybów i bakterii. Taniny („garbniki”) to substancje które, między innymi, utrudniają trawienie pokarmu roślinnego. Postawiono pytanie, jak zawartość tanin wpływa na tempo wzrostu gąsienic motyli żerujących na tych roślinach.

**Dane:** w pliku **Tannins.txt** są wyniki prostego eksperymentu, w którym zbadano wpływ zawartości tanin w sztucznie przygotowanej pożywce, w której stężenia tanin wynosiło od 0 do 8% (co 1%), na tempo wzrostu gąsienic:

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| Nr | numer rekordu danych |
| TC | Zawartość tanin (*Tannins Concentration*) (%) |
| LM | Masa gąsienic (*Larvae Mass*) (mg) |

**1) Decyzje podejmowane PRZED rozpoczęciem analiz danych** (a najlepiej nawet przed rozpoczęciem wykonywania pomiarów/obserwacji).

***Uwaga:*** *odpowiedzi wpisz bezpośrednio pod kolejnymi pytaniami. Dla ułatwienia sprawdzania, odpowiedzi wpisz niebieską czcionką.*

* 1. Sformułuj adekwatną do problem badawczego i dostępnych danych hipotezę roboczą.

Im wyższa zawartość tanin w roślinie tym wolniejsze tempo wzrostu gąsienic.

* 1. Która z blisko powiązanych analiz będzie miała dla tego problemu badawczego zastosowanie: analiza regresji czy analiz korelacji?

Analiza regresji (model 1)-

Regresja ponieważ badamy w jaki sposób zmienna niezależna jest powiązana ze zmienną zależną. Model 1 gdyż, masa gąsienicy nie wpływa na zawartość tanin. Zawartość tanin (predyktor) została ustalona przez eksperymentatora. (Kierunek zależności jest oczywisty.)

1.3) Sformułuj adekwatną hipotezę zerową i alternatywną.

Zerowa: Zawartość tanin nie wpływa na tempo wzrostu gąsienic motyli żerujących na roślinach.

\*Współczynnik regresji b =0

Alternatywna: Zawartość tanin wpływa na tempo wzrostu gąsienic motyli żerujących na roślinach.

\*Współczynnik regresji b różny od 0

* 1. Jakie założenia teoretycznie są wymagane dla wykonania tej analizy zależności między zmiennymi i testu istotności tej zależności?
* rozkład reszt zbliżony do rozkładu normalnego
* liczba obserwacji większa bądź równa liczbie parametrów wyprowadzonych z analizy regresji
* zależność liniowa
* równomierna wariancja reszt
* nie występuje autokorelacja reszt

**2) Analizy statystyczne**

***Uwaga:*** *w odpowiedziach umieść zarówno kod R jak i wyniki jego wykonania.*

2.1) Przedstaw adekwatne dla tego projektu statystyki opisowe i wykresy „eksploracyjne”, które będą pomocne w decyzjach dotyczących wykonywanych dalej analiz.

tannin <- read.delim("Tannins.txt", header = TRUE)

> summary(tannin)

Index TC LM

Min. :1 Min. :0 Min. : 2.000

1st Qu.:3 1st Qu.:2 1st Qu.: 3.000

Median :5 Median :4 Median : 7.000

Mean :5 Mean :4 Mean : 6.889

3rd Qu.:7 3rd Qu.:6 3rd Qu.:10.000

Max. :9 Max. :8 Max. :12.000

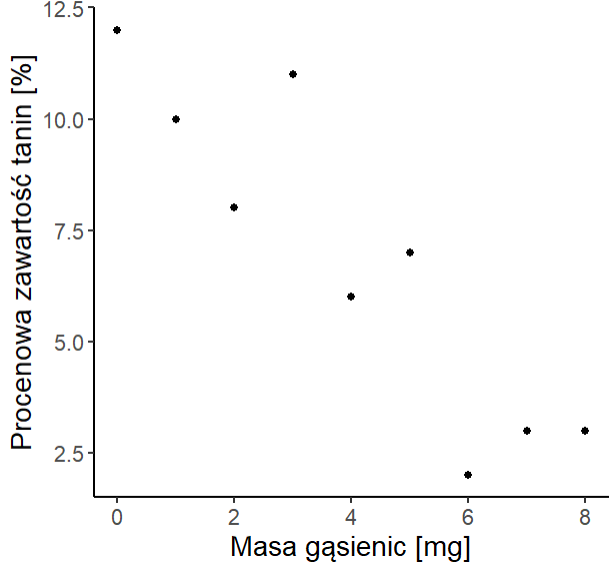
ggplot(data = tannin, mapping = aes(x = TC,

y = LM)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Masa gąsienic [mg]", y = "Procenowa zawartość tanin [%]")



2.2) Użyj odpowiedniej dla danego problemu badawczego funkcji R do analiz zależności między badanymi zmiennymi.

> tannin\_lm <- lm(formula = LM~TC, data = tannin)

> tannin\_lm

Call:

lm(formula = LM ~ TC, data = tannin)

Coefficients:

(Intercept) TC

11.756 -1.217

#funkcja lm() tworzy model regresji (Y~X)

#“offset()” - pozwala zadeklarować konkretną wartość nachylenia jako hipotezę zerową

#Gdy nie użyjemy będzie ustawiony domyślnie na zero

# 11.756 – a (punkt przecięcia osią Y przy X = 0)

#-1.217 - b (nachylenie)

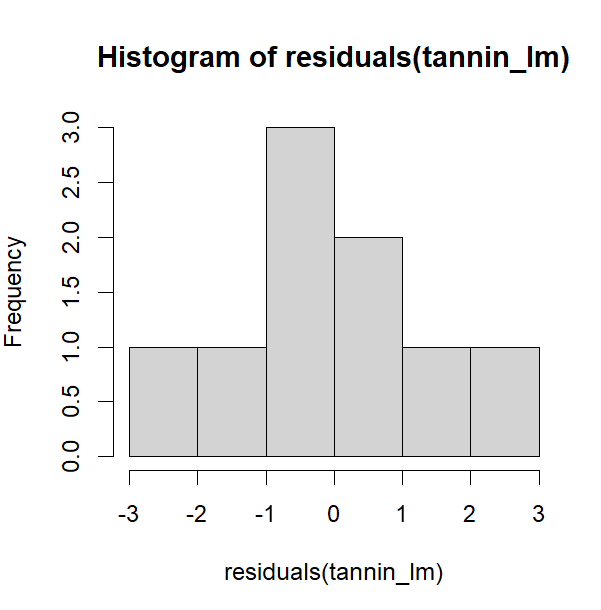
> shapiro.test(tannin\_lm$residuals)

Shapiro-Wilk normality test

data: tannin\_lm$residuals

W = 0.98794, p-value = 0.9926

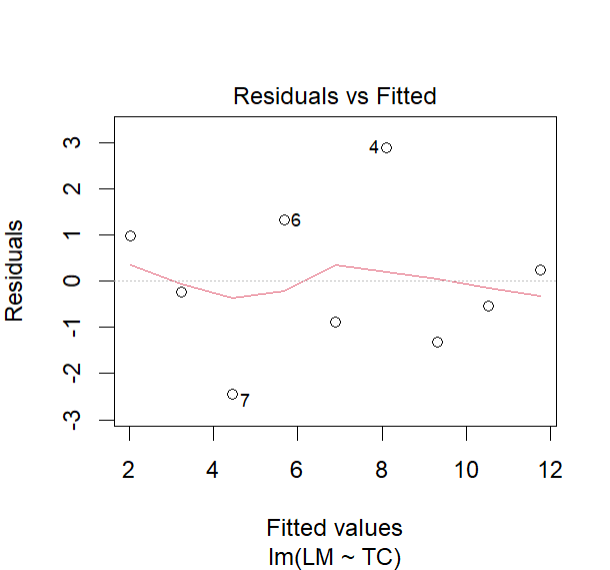
hist(residuals(tannin\_lm))



2.3) Przedstaw wykresy diagnostyczne mające na celu sprawdzenie założeń analizy i zdecyduj, czy te założenia można przyjąć. Jeśli nie, zaproponuj i wykonaj odpowiednie działania zaradcze.

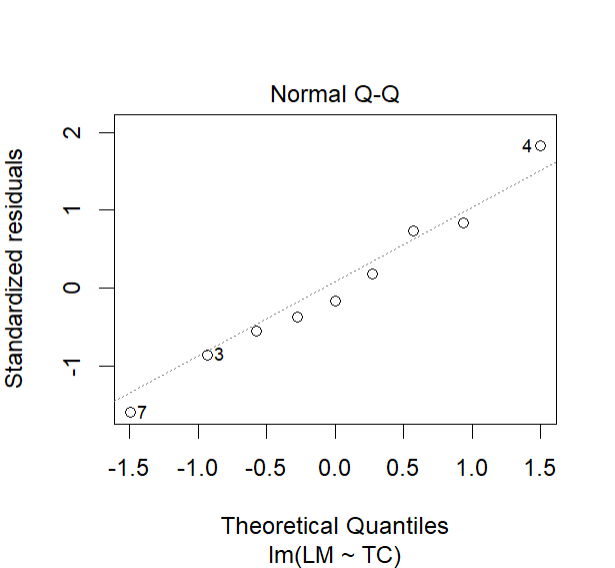
plot(tannin\_lm, 1)

#Punkty są symetrycznie rozmieszczone wokół 0. Rozrzut wartości (wariancja) jest równomierny

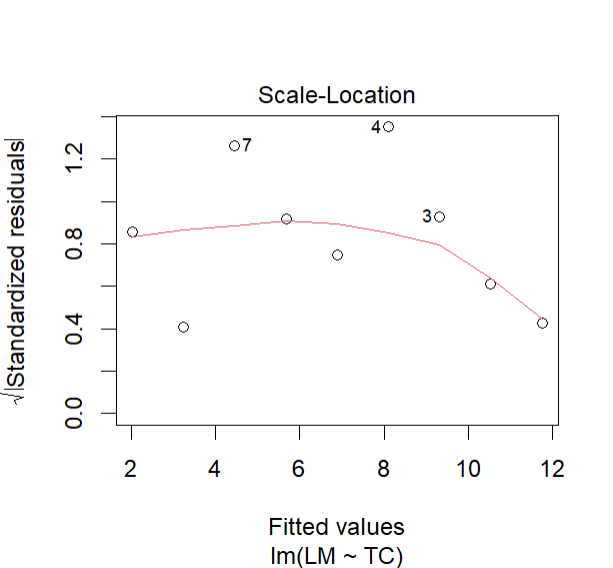


plot(tannin\_lm, 2)

#Punkty są rozrzucone prawie idealnie dla normalnego rozkładu

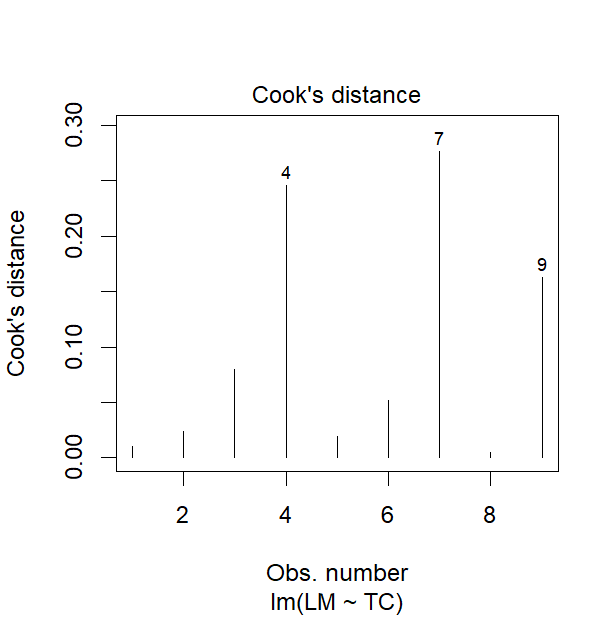


plot(tannin\_lm, 3)



plot(tannin\_lm, 4)

#Ocena jak bardzo konkretne wartości wpływaja na wynik ostateczny



2.4) Przedstaw tabele z kompletnymi wynikami analizy, zawierającymi oszacowania parametrów badanej zależności z ich błędami standardowymi i przedziałami ufności, testy istotności oraz informację o tym, jak duża część wariancji jest wyjaśniana przez model. Objaśnij sens wartości pojawiających się na wydrukach.

> summary(tannin\_lm)

Call:

lm(formula = LM ~ TC, data = tannin)

Residuals:

Min 1Q Median 3Q Max

-2.4556 -0.8889 -0.2389 0.9778 2.8944

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.7556 1.0408 11.295 9.54e-06 \*\*\*

TC -1.2167 0.2186 -5.565 0.000846 \*\*\*

---

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

Residual standard error: 1.693 on 7 degrees of freedom

Multiple R-squared: 0.8157, Adjusted R-squared: 0.7893

F-statistic: 30.97 on 1 and 7 DF, p-value: 0.0008461

#Estimate- Wartość oszacowana

#TC = b (nachylenie)

#(Intercept) = a (wyraz wolny, przecięcie z osią Y dla X=0)

#Std. Error - błąd standardowy odpowiednio do a i b

2.5) Przedstaw adekwatny wykres ilustrujący najważniejsze wyniki.

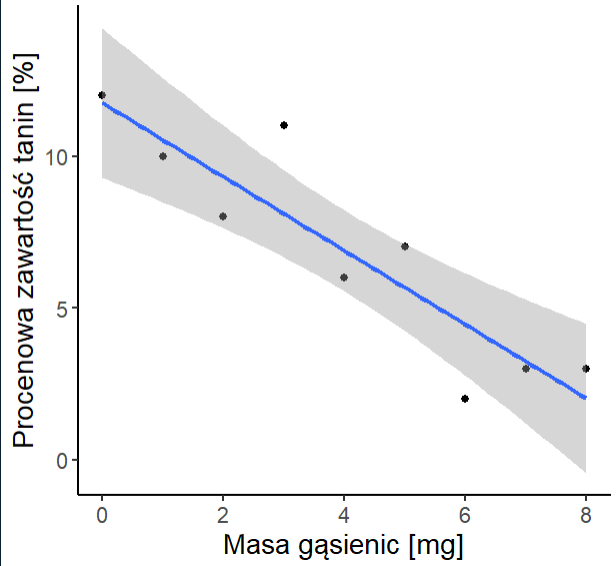
ggplot(data = tannin, mapping = aes(x = TC,

y = LM)) +

geom\_point() + theme\_classic() + stat\_smooth(method = lm, se=TRUE) +

theme(text = element\_text(size = 14)) +

labs(x = "Masa gąsienic [mg]", y = "Procenowa zawartość tanin [%]")



Zaciemniony obszar obrazuje 95% przedział ufności dla położenia tej prostej. Obszar jest skonstruowany na podstawie błędów standardowych dla oszacowanego punktu przecięcia i nachylenia.

2.6) Przedstaw i uzasadnij wniosek z wykonanej analizy, odnoszący się do stawianej hipotezy roboczej. Przedstaw ten wniosek w możliwie najpełniejszej formie, a więc nie poprzestający tylko stwierdzeniu obecności albo braku zależności, ale też na ilościowej charakterystyce tej zależności.

Otrzymaliśmy wartość p-value mniejszą od 0.05, zatem należy odrzucić hipotezę zerową i przyjąć alternatywną iż, zawartość tanin wpływa na wzrost gąsienic. Powyższy wykres również może prowadzić do głębszego wniosku. Możemy przyjąć, że im większa zawartość tanin w roślinie tym tempo wzrostu masy gąsienic jest mniejsze.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Zadanie 2. Allometria: zależność spoczynkowego tempa metabolizmu (RMR) od masy ciała**

**Problem badawczy:** Już od XIX weku wiadomo, że spoczynkowe tempo metabolizmu zwierząt (RMR) rośnie z masą ciała (M) nie liniowo, a „allometrycznie”: w przybliżeniu zgodnie z funkcją potęgową: ***RMR = a M b***. Początkowo sądzono, że zależność ta jest proporcjonalna do powierzchni ciała, w więc że potęga b powinna wynosić około 2/3. Później jednak okazało się, że nachylenie to jest bliższe 3/4 (0.75) i rozwinięto szereg teorii uzasadniających dlaczego właśnie taki powinien być ten współczynnik. Kwestia ta w dalszym ciągu jest przedmiotem badań i sporów. **Czy spoczynkowe tempo metabolizmu u gryzoni rośnie z masą ciała do potęgi 0.75? A może raczej zgodnie z „prawem powierzchni” (nachylenie 2/3)?**

**Dane:** w pliku **RMR\_gryzonie.txt** są wyniki pomiarów RMR u 114 gatunków gryzoni:

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| Species\_No | Numer gatunku (nazwy gatunków zostały dla uproszczenia pominięte) |
| M | Masa ciała (g), średnia dla danego gatunku |
| RMR | Spoczynkowe tempo metabolizmu, mierzone jako tempo zużywania tlenu (mlO2/h), średnia dla danego gatunku |

**1) Decyzje podejmowane PRZED rozpoczęciem analiz danych** (a najlepiej nawet przed rozpoczęciem wykonywania pomiarów/obserwacji).

***Uwaga:*** *odpowiedzi wpisz bezpośrednio pod kolejnymi pytaniami. Dla ułatwienia sprawdzania, odpowiedzi wpisz niebieską czcionką.*

* 1. Zdecyduj, jakich transformacji danych należy użyć, żeby można było wykorzystać model liniowy do analizy zależności opisywanej funkcją potęgową (odpowiedź jest w prezentacji wykładu).

Logarytmowanie

1.2) Sformułuj adekwatne do problem badawczego i dostępnych danych hipotezy robocze dotyczące trzech pytań:

a) czy tempo metabolizmu rośnie z masą ciała?

Tempo metabolizmu gryzoni rośnie wraz z masą ciała.

b) czy nachylenie tej zależności (wykładnik w równaniu allometrycznym) jest zgodne z „prawem” potęgi 3/4 (wynosi 0.75)?

Tempo metabolizmu gryzoni rośnie z masą ciała do potęgi 0.75.

c) czy nachylenie jest zgodne z „prawem powierzchni” (2/3)?

Tempo metabolizmu gryzoni rośnie z masą ciała zgodnie z “prawem powierzchni” (2/3).

1.3) Która z blisko powiązanych analiz będzie miała dla tego problemu badawczego zastosowanie: analiza regresji czy analiz korelacji?

Analiza regresji

1.4) Sformułuj hipotezy zerowe i alternatywne, adekwatne do stawianych pytań (i hipotez roboczych).

Zerowe:

Tempo metabolizmu gryzoni nie rośnie wraz z masą ciała.

Nachylenie linii zależności spoczynkowego tempa metabolizmu gryzoni jest zgodne z prawem potęgi 0.75. (Nachylenie linii = 0)

Nachylenie linii zależności spoczynkowego tempa metabolizmu gryzoni jest zgodne z “prawem powierzchni” (2/3). (Nachylenie linii = 0)

Alternatywne:

Tempo metabolizmu gryzoni rośnie wraz z masą ciała.

Nachylenie linii zależności spoczynkowego tempa metabolizmu gryzoni nie jest zgodne z prawem potęgi 0.75. (Nachylenie linii nie równa się 0)

Nachylenie linii zależności spoczynkowego tempa metabolizmu gryzoni nie jest zgodne z “prawem powierzchni” (2/3). (Nachylenie linii nie równa się 0)

1.5) Jakie założenia dotyczące charakteru rozkładu wartości resztowych są wymagane dla wykonania tej analizy i testu?

-rozkład reszt zbliżony do rozkładu normalnego

-liczba obserwacji większa bądź równa liczbie parametrów wyprowadzonych z analizy regresji

-zależność liniowa

-nie występuje autokorelacja reszt

-równa wariancja reszt

1.6) Oprócz założeń dotyczących charakteru rozkładu, prosty model liniowy wymaga jeszcze jednego ważnego założenia, które dla tego rodzaju danych nie jest spełnione, a które w takich porównawczych analizach międzygatunkowych było aż do lat 80. XX wieku ignorowane. My też wykonamy analizy w sposób uproszczony, bo poprawne rozwiązanie wymaga narzędzi statystycznych na znacznie wyższym poziomie (wymaga użycia uogólnionego modelu liniowego). Ważne jest jednak, byśmy mieli świadomość tego uproszczenia. O jakie założenie tu chodzi?

-niezależność obserwacji (niezależność wartości resztowych)

**2) Analizy statystyczne**

***Uwaga:*** *w odpowiedziach umieść zarówno kod R jak i wyniki jego wykonania.*

2.1) Przedstaw adekwatne dla tego projektu statystyki opisowe i wykresy „eksploracyjne”, które są pomocne w decyzjach dotyczących wykonywanych dalej analiz, zarówno dla danych surowych jak i odpowiednio transformowanych.

> RMR\_Rodents <- read.delim(file = "RMR\_Rodents.txt", header = TRUE)

DANE SUROWE

> summary(RMR\_Rodents)

No M RMR

Min. : 1.00 Min. : 5.20 Min. : 13.50

1st Qu.: 29.25 1st Qu.: 24.60 1st Qu.: 41.90

Median : 57.50 Median : 42.00 Median : 65.10

Mean : 57.50 Mean : 89.69 Mean : 93.96

3rd Qu.: 85.75 3rd Qu.: 82.95 3rd Qu.: 109.22

Max. :114.00 Max. :1870.90 Max. :1029.00

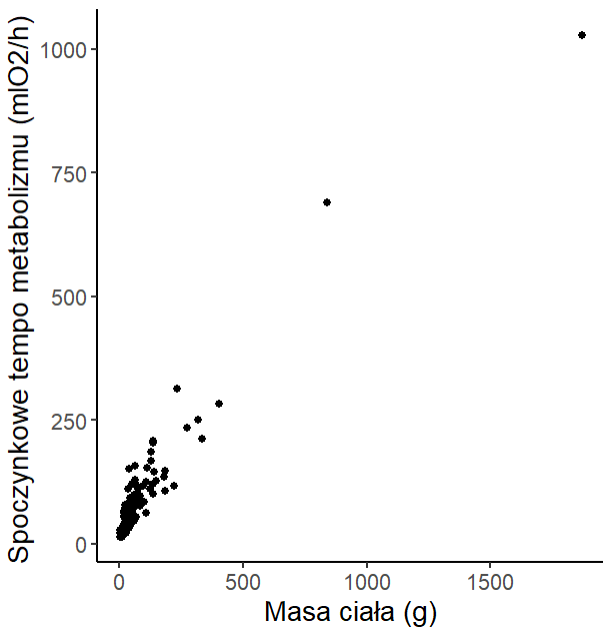
ggplot(data = RMR\_Rodents, mapping = aes(x = M,

y = RMR)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Masa ciała (g)", y = "Spoczynkowe tempo metabolizmu (mlO2/h) ")



DANE TRANSFORMOWANE

> RMR\_Rodents\_log <- RMR\_Rodents

> RMR\_Rodents\_log$RMR <- log(RMR\_Rodents\_log$RMR, 10)

> RMR\_Rodents\_log$M <- log(RMR\_Rodents\_log$M, 10)

> summary(RMR\_Rodents\_log)

No M RMR

Min. : 1.00 Min. :0.716 Min. :1.130

1st Qu.: 29.25 1st Qu.:1.391 1st Qu.:1.622

Median : 57.50 Median :1.623 Median :1.814

Mean : 57.50 Mean :1.677 Mean :1.829

3rd Qu.: 85.75 3rd Qu.:1.919 3rd Qu.:2.038

Max. :114.00 Max. :3.272 Max. :3.012

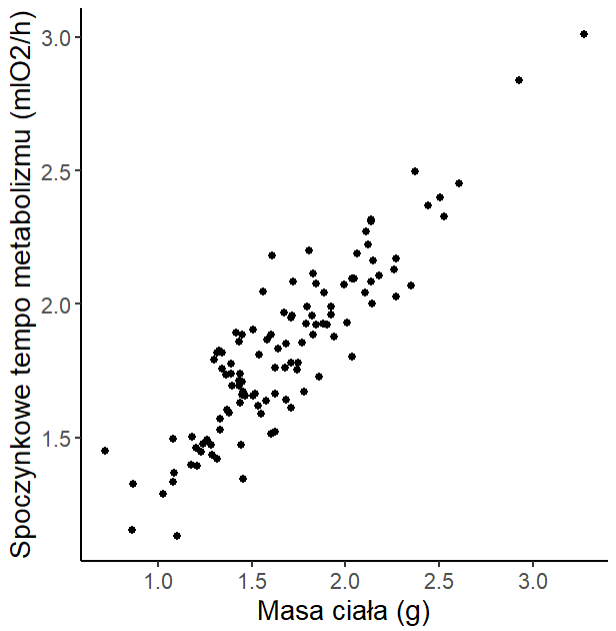
ggplot(data = RMR\_Rodents\_log, mapping = aes(x = M,

y = RMR)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Masa ciała (g)", y = "Spoczynkowe tempo metabolizmu (mlO2/h)")



2.2) Użyj odpowiedniej dla danego problemu badawczego funkcji R do analiz zależności między badanymi zmiennymi.

> RMR\_Rodents\_log\_lm <- lm(formula = RMR~M, data = RMR\_Rodents\_log)

> RMR\_Rodents\_log\_lm

Call:

lm(formula = RMR ~ M, data = RMR\_Rodents\_log)

Coefficients:

(Intercept) M

0.6918 0.6782

(Intercept) -wyraz wolny (a)

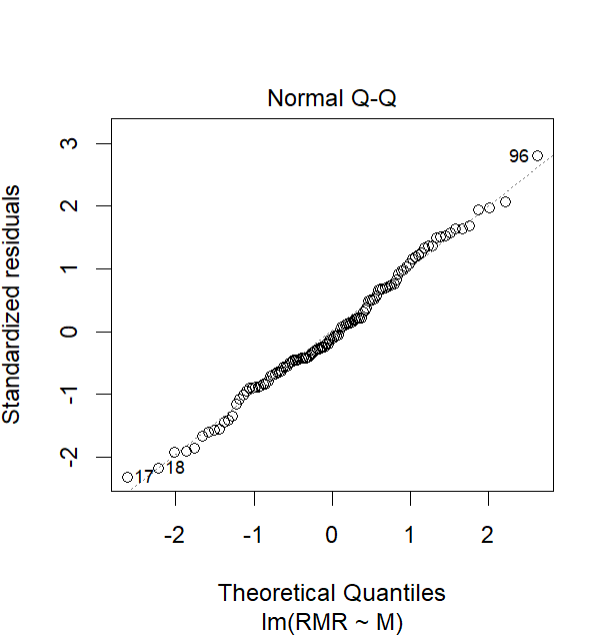
M – nachylenie (b)

2.3) Przedstaw wykresy diagnostyczne mające na celu sprawdzenie założeń analizy i zdecyduj, czy te założenia można przyjąć. Jeśli nie, zaproponuj i wykonaj odpowiednie działania zaradcze.

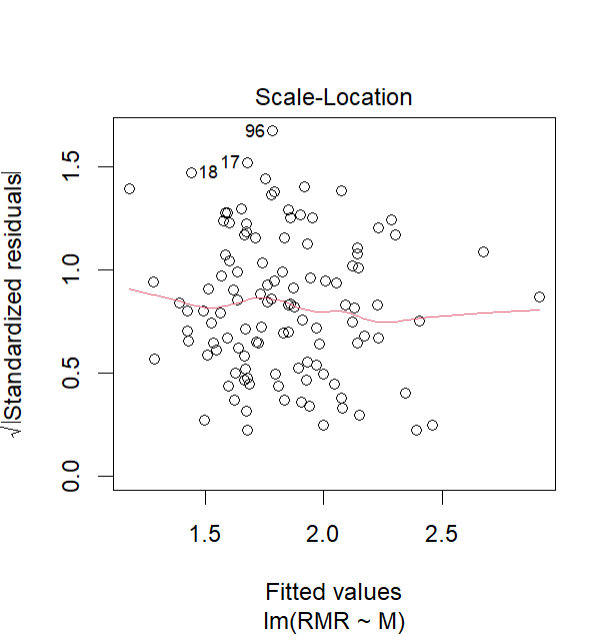
plot(RMR\_Rodents\_log\_lm, 1)



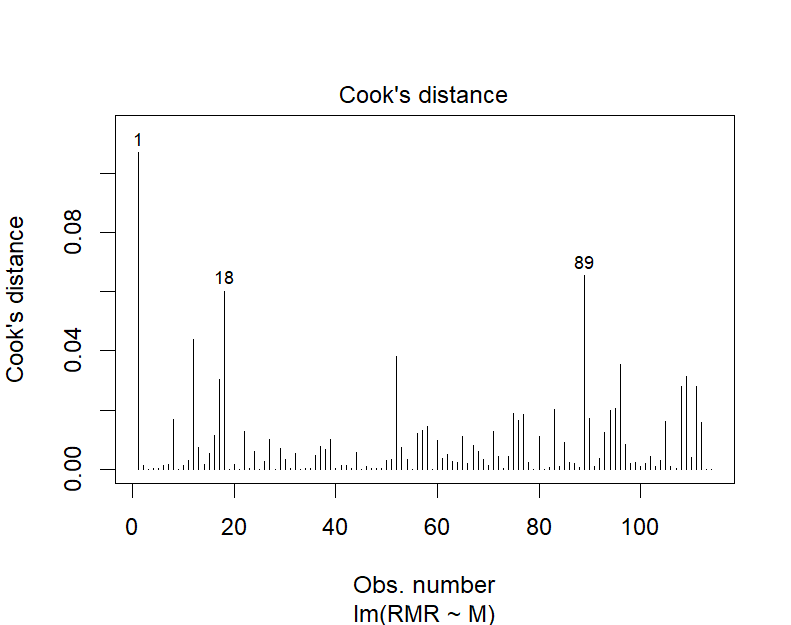
plot(RMR\_Rodents\_log\_lm, 2)



plot(RMR\_Rodents\_log\_lm, 3)



plot(RMR\_Rodents\_log\_lm, 4)



2.4) Przedstaw tabele z kompletnymi wynikami analizy, zawierającymi oszacowania parametrów badanej zależności z ich błędami standardowymi i przedziałami ufności, testy istotności oraz informację o tym, jak duża część wariancji jest wyjaśniana przez model. Objaśnij sens wartości pojawiających się na wydruku.

> summary(RMR\_Rodents\_log\_lm)

Call:

lm(formula = RMR ~ M, data = RMR\_Rodents\_log)

Residuals:

Min 1Q Median 3Q Max

-0.33113 -0.09080 -0.01536 0.09915 0.40068

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.69180 0.05503 12.57 <2e-16 \*\*\*

M 0.67823 0.03182 21.31 <2e-16 \*\*\*

---

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

Residual standard error: 0.1439 on 112 degrees of freedom

Multiple R-squared: 0.8022, Adjusted R-squared: 0.8005

F-statistic: 454.3 on 1 and 112 DF, p-value: < 2.2e-16

2.5) Testy istotności najważniejszego parametru opisującego badaną zależność. Zwróć uwagę, że należy nie tylko odpowiedzieć na pytanie, czy jest zależność między RMR a masą ciała, ale też czy nachylenie jest zgodne z dwoma alternatywnymi przewidywaniami.

> RMR\_of2na3 <- lm(formula = RMR~M+offset((2/3)\*M), data = RMR\_Rodents\_log)

> summary(RMR\_of2na3)

Call:

lm(formula = RMR ~ M + offset((2/3) \* M), data = RMR\_Rodents\_log)

Residuals:

Min 1Q Median 3Q Max

-0.33113 -0.09080 -0.01536 0.09915 0.40068

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.69180 0.05503 12.571 <2e-16 \*\*\*

M 0.01156 0.03182 0.363 0.717

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1439 on 112 degrees of freedom

Multiple R-squared: 0.8022, Adjusted R-squared: 0.8005

F-statistic: 454.3 on 1 and 112 DF, p-value: < 2.2e-16

Dla M wartość p-value jest wysoka zatem spoczynkowe tempo metabolizmu u gryzoni rośnie zgodnie z „prawem powierzchni”

#Górny i dolny przedział ufności

> confint(RMR\_of2na3)

2.5 % 97.5 %

(Intercept) 0.58276294 0.80084067

M -0.05149035 0.07460745

Dla M wartość p-value jest wysoka zatem spoczynkowe tempo metabolizmu u gryzoni rośnie zgodnie z „prawem powierzchni”

> RMR\_of3na4 <- lm(formula = RMR~M+offset((3/4)\*M), data = RMR\_Rodents\_log)

> summary(RMR\_of3na4)

Call:

lm(formula = RMR ~ M + offset((3/4) \* M), data = RMR\_Rodents\_log)

Residuals:

Min 1Q Median 3Q Max

-0.33113 -0.09080 -0.01536 0.09915 0.40068

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.69180 0.05503 12.571 <2e-16 \*\*\*

M -0.07177 0.03182 -2.256 0.026 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1439 on 112 degrees of freedom

Multiple R-squared: 0.8022, Adjusted R-squared: 0.8005

F-statistic: 454.3 on 1 and 112 DF, p-value: < 2.2e-16

>

> confint(RMR\_of3na4)

2.5 % 97.5 %

(Intercept) 0.5827629 0.800840673

M -0.1348237 -0.008725879

Powyższe wyniki dodatkowo potwierdzają że spoczynkowe tempo metabolizmu u gryzoni nie rośnie z masą ciała do potęgi 0.75 gdyż wartość p-value jest dla mniejsza od 0.05

2.6) Przedstaw adekwatny wykres ilustrujący najważniejsze wyniki.

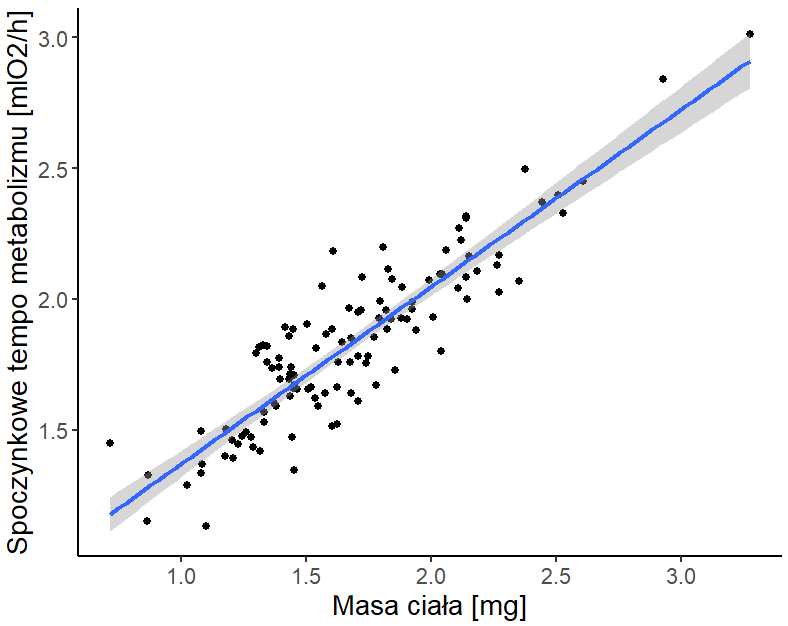
ggplot(data = RMR\_Rodents\_log\_lm, mapping = aes(x = M,

y = RMR)) +

geom\_point() + theme\_classic() + stat\_smooth(method = lm, se=TRUE) +

theme(text = element\_text(size = 14)) +

labs(x = "Masa ciała [mg]", y = "Spoczynkowe tempo metabolizmu [mlO2/h]")



2.7) Przedstaw i uzasadnij wniosek z wykonanej analizy, odnoszący się do stawianych hipotez roboczych. Przedstaw ten wniosek w możliwie najpełniejszej formie, a więc nie poprzestający tylko stwierdzeniu obecności albo braku zależności, ale też na ilościowej charakterystyce tej zależności.

Wykonana analiza wykazała że należy odrzucić hipotezę zerową i przyjąć alternatywną która twierdzi, że spoczynkowe tempo metabolizmu rośnie wraz z masą ciała.

Dodatkowo możemy stwierdzić że wzrost ten jest zgodny z „prawem powierzchni”.

**Zadanie 3. Małpy: dominacja a stopień infekcji pasożytami**

**Problem badawczy:** Czy struktura dominacji w relacjach socjalnych jest powiązana z „jakością” osobnika – np. z jego wydolnością fizjologiczną, odpornością na choroby i pasożyty, itp.? A może osobniki poświęcające dużo energii na walkę o wysoką pozycję w hierarchii „płacą” za to obniżoną odpornością? Wykonano badania, w których określono strukturę dominacji w grupie 6 samców gerezy abisyńskiej (*Colobus guereza*) oraz liczbę jaj pasożytniczego nicienia (*Trichuris sp.*) w odchodach tych osobników. Czy pozycja w rankingu dominacji jest związana ze stopniem zapasożycenia?

**Dane:** w pliku **Monkeys.txt** są wyniki tych badań:

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| Nr | numer rekordu danych |
| Name | „Imię” osobnika |
| DomRank | Pozycja w ranking dominacji – od 1 (najwyższa) do 6 (najniższa) |
| Eggs | Liczba jaj nicienia na 1g odchodów danego osobnika |

**1) Decyzje podejmowane PRZED rozpoczęciem analiz danych** (a najlepiej nawet przed rozpoczęciem wykonywania pomiarów/obserwacji).

***Uwaga:*** *odpowiedzi wpisz bezpośrednio pod kolejnymi pytaniami. Dla ułatwienia sprawdzania, odpowiedzi wpisz niebieską czcionką.*

* 1. Sformułuj adekwatną do problem badawczego i dostępnych danych hipotezę roboczą.

Pozycja w rankingu dominacji jest związana ze stopniem zapasożycenia.

* 1. Która z blisko powiązanych analiz będzie miała dla tego problemu badawczego zastosowanie: analiza regresji czy analiz korelacji? Jakiego testu/analizy użyjesz „z wyboru”?

Analiza korelacji

1.3) Sformułuj adekwatną hipotezę zerową i alternatywną.

Zerowa: Pozycja w rankingu dominacji nie jest związana ze stopniem zapasożycenia.

Alternatywna: pozycja w rankingu dominacji jest związana ze stopniem zapasożycenia.

* 1. Jakie założenia teoretycznie są wymagane dla wykonania tej analizy zależności między zmiennymi i testu istotności tej zależności?

Zależność liniowa

**2) Analizy statystyczne**

***Uwaga:*** *w odpowiedziach umieść zarówno kod R jak i wyniki jego wykonania.*

2.1) Przedstaw adekwatne dla tego projektu statystyki opisowe i wykresy „eksploracyjne”, które będą pomocne w decyzjach dotyczących wykonywanych dalej analiz.

> Monkeys <- read.delim(file = "Monkeys.txt", header = TRUE)

> summary(Monkeys)

Nr Name DomRank Eggs

Min. :1.00 Length:6 Min. :1.00 Min. : 749.0

1st Qu.:2.25 Class :character 1st Qu.:2.25 1st Qu.: 889.8

Median :3.50 Mode :character Median :3.50 Median :1061.5

Mean :3.50 Mean :3.50 Mean :1857.3

3rd Qu.:4.75 3rd Qu.:4.75 3rd Qu.:1512.2

Max. :6.00 Max. :6.00 Max. :5777.0

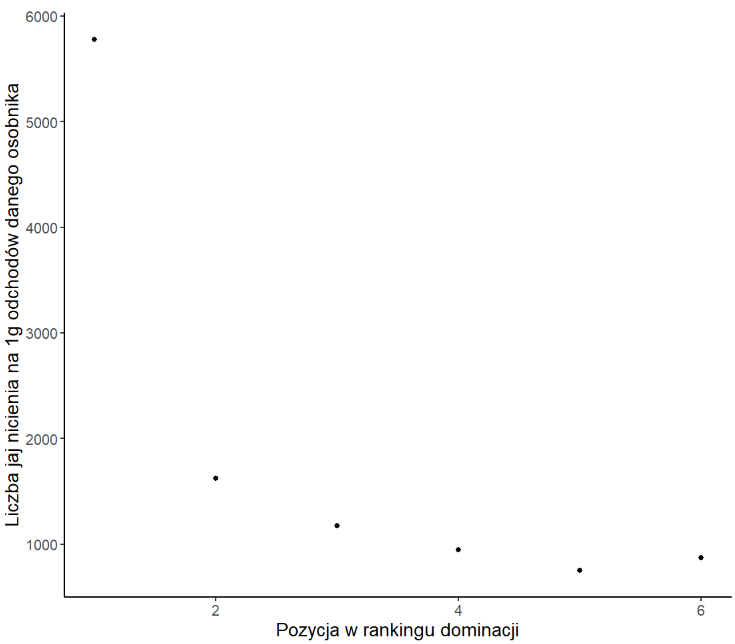
ggplot(data = Monkeys, mapping = aes(x = DomRank,

y = Eggs)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Pozycja w rankingu dominacji", y = "Liczba jaj nicienia na 1g odchodów danego osobnika")



2.2) Zdecyduj, czy dane spełniają założenia dla testu parametrycznego. Jeśli nie, wykonaj odpowiednie działania zaradcze.

Nie powinniśmy zakładać że nasze dane mają zależność liniową ponieważ mamy rang czyli skalę porządkową.

Zatem zastosujemy transformację logarytmiczną

> MonkeysLog <- Monkeys

> MonkeysLog$Eggs <- log(MonkeysLog$Eggs, 10)

> MonkeysLog$DomRank <- log(MonkeysLog$DomRank,10)

> summary(MonkeysLog)

Nr Name DomRank Eggs

Min. :1.00 Length:6 Min. :0.0000 Min. :2.874

1st Qu.:2.25 Class :character 1st Qu.:0.3451 1st Qu.:2.949

Median :3.50 Mode :character Median :0.5396 Median :3.023

Mean :3.50 Mean :0.4762 Mean :3.139

3rd Qu.:4.75 3rd Qu.:0.6747 3rd Qu.:3.176

Max. :6.00 Max. :0.7782 Max. :3.762

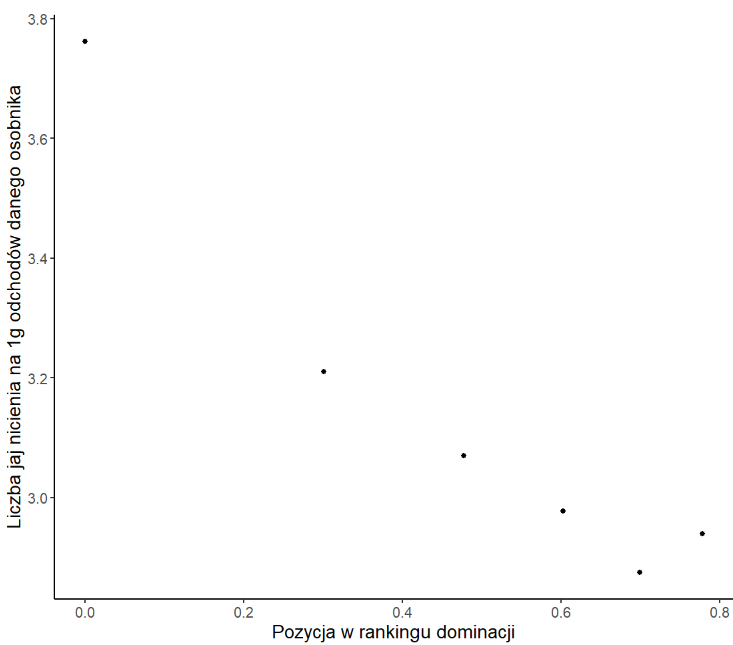
ggplot(data = MonkeysLog, mapping = aes(x = DomRank,

y = Eggs)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Pozycja w rankingu dominacji", y = "Liczba jaj nicienia na 1g odchodów danego osobnika")



2.3) Użyj odpowiedniej dla danego problemu badawczego funkcji R do analiz zależności między badanymi zmiennymi.

Logarytmowanie w pełni nie pomogło, założeni dotyczę liniowości rozkładu nie są spełnione. Zatem stosujemy test nieparametryczny.

> Monkeys\_output <- cor.test(MonkeysLog$Eggs, MonkeysLog$DomRank, method = "spearman")

2.4) Przedstaw tabele z kompletnymi wynikami analizy. Objaśnij sens wartości pojawiających się na wydrukach.

> Monkeys\_output

Spearman's rank correlation rho

data: MonkeysLog$Eggs and MonkeysLog$DomRank

S = 68, p-value = 0.01667

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

-0.9428571

2.5) Przedstaw adekwatny wykres ilustrujący najważniejsze wyniki.

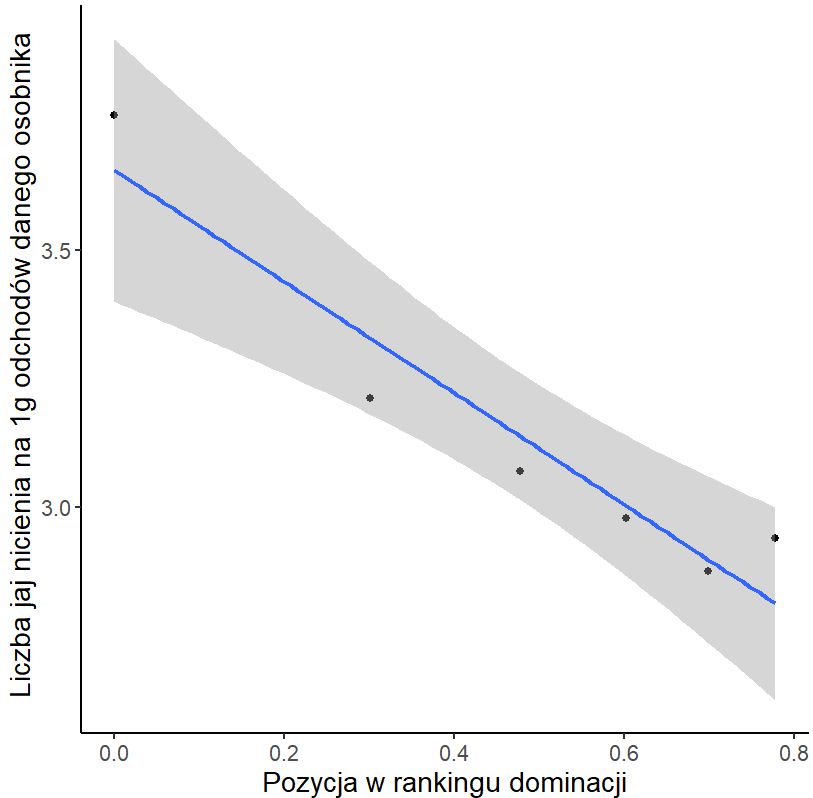
ggplot(data = MonkeysLog, mapping = aes(x = DomRank,

y = Eggs)) +

geom\_point() + theme\_classic() + stat\_smooth(method = lm, se=TRUE) +

theme(text = element\_text(size = 14)) +

labs(x = "Pozycja w rankingu dominacji", y = " Liczba jaj nicienia na 1g odchodów danego osobnika")



2.6) Przedstaw i uzasadnij wniosek z wykonanej analizy, odnoszący się do stawianej hipotezy roboczej. Przedstaw ten wniosek w możliwie najpełniejszej formie, a więc nie poprzestający tylko stwierdzeniu obecności albo braku zależności, ale też na ilościowej charakterystyce tej zależności.

Po wykonanej analizę otrzymujemy p-value mniejsze od 0.05 co sugeruje odrzucenie hipotezy zerowej i przyjęcie alternatywnej. Pozycja w rankingu dominacji jest związana ze stopniem zapasożycenia. Z powyższych wykresów widzimy również zależność że im wyżej w rankingu dominacji znajduje się osobnik tym wyższy będzie jego stopień zapasożycenia.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Zadanie 4. Poroża: wpływ masy ciała i wieku jeleni na wielkość poroży**

**Problem badawczy:** Wielkość poroży u samców jeleniowatych jest ważną cechą w doborze płciowym. Postawiono hipotezę, że u jeleni wielkość poroży rośnie zarówno z masą ciała jak i wiekiem byków, a więc na podstawie tych cech można przewidzieć jak duże poroże będzie miał dany osobnik. Należy sprawdzić tę hipotezę i opracować odpowiedni model predykcyjny.

**Dane:** w pliku **Deer.txt** są potrzebne dane:

|  |  |
| --- | --- |
| **Zmienna** | **Opis** |
| Nr | numer rekordu danych |
| ID | numer osobnika |
| AM | Masa poroża (*antler mass*) [kg] |
| BM | Masa ciała (*body mass*) [kg] |
| Age | Wiek [lata] |

**1) Decyzje podejmowane PRZED rozpoczęciem analiz danych** (a najlepiej nawet przed rozpoczęciem wykonywania pomiarów/obserwacji).

***Uwaga:*** *odpowiedzi wpisz bezpośrednio pod kolejnymi pytaniami. Dla ułatwienia sprawdzania, odpowiedzi wpisz niebieską czcionką.*

* 1. Która z blisko powiązanych analiz będzie miała dla tego problemu badawczego zastosowanie: analiza regresji czy analiz korelacji?

Analiza regresji

* 1. Sformułuj adekwatne hipotezy zerowe i alternatywne. Uwaga: zwróć uwagę, że w tym przypadku pytamy o wpływ więcej niż jednej zmiennej na raz!

Zerowa: Wielkość poroży jeleni nie zależy od ich masy i wieku.

Alternatywna: Wielkość poroży jeleni zależy od ich masy i wieku.

Robocza: Im cięższe i starsze jelenie, tym wielkość ich poroża wzrasta.

* 1. Jakie założenia teoretycznie są wymagane dla wykonania tej analizy zależności między zmiennymi i wybranego testu?
* rozkład reszt zbliżony do rozkładu normalnego
* liczba obserwacji większa bądź równa liczbie parametrów wyprowadzonych z analizy regresji
* zależność liniowa
* równa wariancja reszt
* nie występuje autokorelacja reszt

**2) Analizy statystyczne**

***Uwaga:*** *w odpowiedziach umieść zarówno kod R jak i wyniki jego wykonania.*

2.1) Przedstaw adekwatne dla tego projektu statystyki opisowe i wykresy „eksploracyjne”, które są pomocne w decyzjach dotyczących wykonywanych dalej analiz. Zwróć uwagę, że z punktu widzenia interpretacji wyników ważne jest nie tylko to, jak masa poroża zależy od masy ciała i wieku byka, ale też to, czy i jak dwa predyktory – masa i wielkość byka – są ze sobą związane. Warto wykonać zarówno proste wykresy dwuwymiarowe (XY) jak i trójwymiarowy wykres (XYZ), który pokazuje zależność masy poroża równocześnie od masy ciała i wieku.

> Deer <- read.delim(file = "Deer.txt", header = TRUE)

> summary(Deer)

Nr ID AM BM

Min. : 1.0 Length:35 Min. :1.800 Min. :103.0

1st Qu.: 9.5 Class :character 1st Qu.:4.180 1st Qu.:141.5

Median :18.0 Mode :character Median :5.360 Median :158.0

Mean :18.0 Mean :5.482 Mean :156.8

3rd Qu.:26.5 3rd Qu.:6.950 3rd Qu.:174.0

Max. :35.0 Max. :9.090 Max. :202.0

Age

Min. : 3.000

1st Qu.: 5.000

Median : 8.000

Mean : 8.143

3rd Qu.:11.000

Max. :13.000

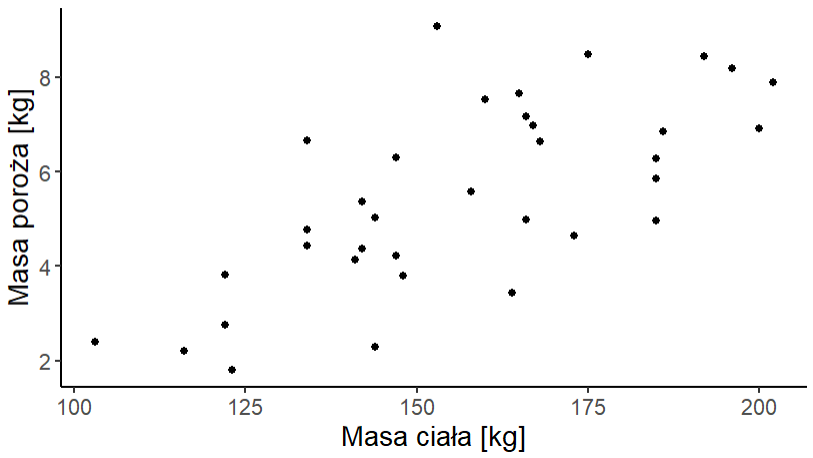
ggplot(data = Deer, mapping = aes(x = BM,

y = AM)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Masa ciała [kg]", y = "Masa poroża [kg]")



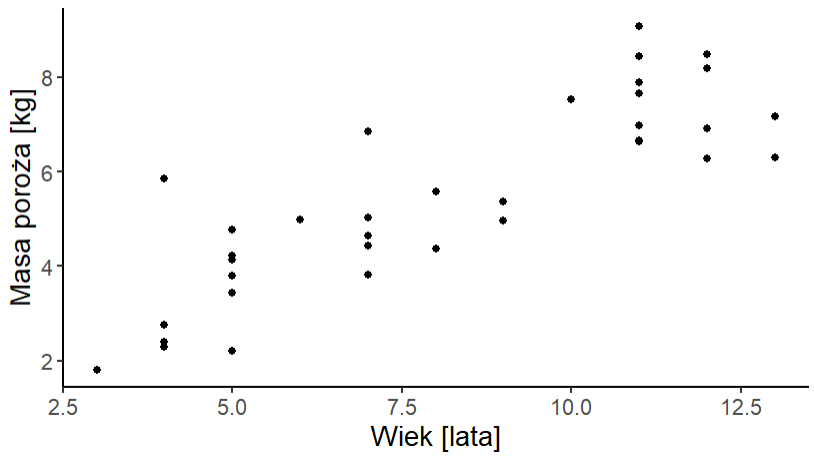
ggplot(data = Deer, mapping = aes(x = Age,

y = AM)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Wiek [lata]", y = "Masa poroża [kg]")



Powiązanie predykatorów

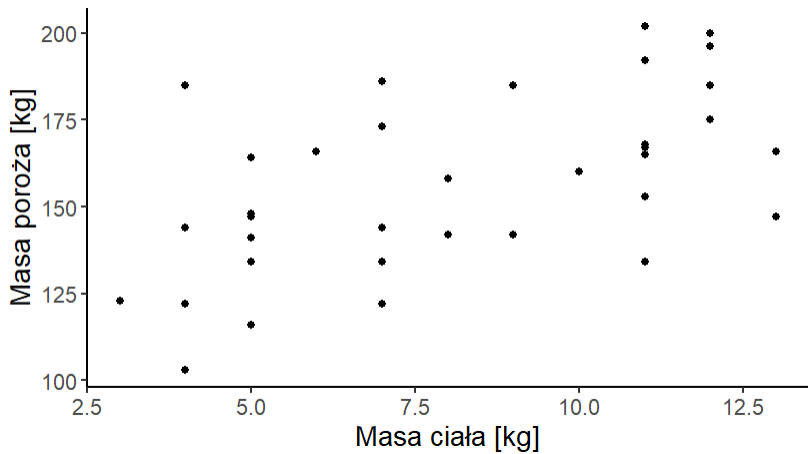
ggplot(data = Deer, mapping = aes(x = Age,

y = BM)) +

geom\_point() + theme\_classic() +

theme(text = element\_text(size = 14)) +

labs(x = "Masa ciała [kg]", y = "Masa poroża [kg]")



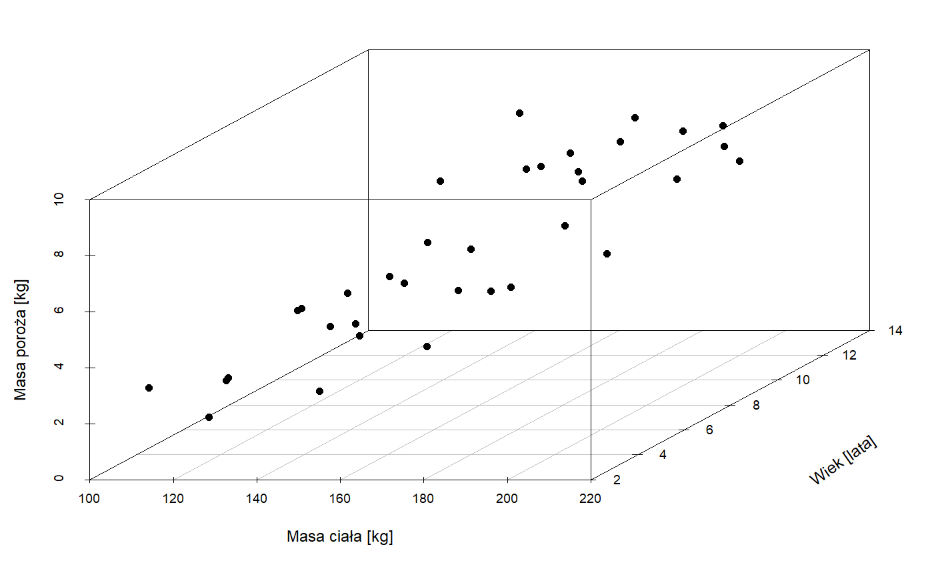
scatterplot3d(x = Deer$BM, y = Deer$Age, z = Deer$AM, , xlab = "Masa ciała [kg]", ylab = "", zlab = "Masa poroża [kg]", pch=16)

dims <- par("usr")

x <- dims[1]+ 0.9\*diff(dims[1:2])

y <- dims[3]+ 0.08\*diff(dims[3:4])

text(x,y,expression("Wiek [lata]"),srt=35)



2.2) Użyj odpowiedniej dla danego problemu badawczego funkcji R do analiz zależności między badanymi zmiennymi. Na potrzeby tych ćwiczeń wykonaj osobo analizy zależności wielkości poroża od a) masy ciała, b) wieku byka, i c) obu tych zmiennych równocześnie.

> Deer\_AM\_BM <- lm(formula = AM~BM, data = Deer)

> Deer\_AM\_BM

Call:

lm(formula = AM ~ BM, data = Deer)

Coefficients:

(Intercept) BM

-3.14749 0.05503

> Deer\_AM\_Age <- lm(formula = AM~Age, data = Deer)

> Deer\_AM\_Age

Call:

lm(formula = AM ~ Age, data = Deer)

Coefficients:

(Intercept) Age

1.0882 0.5396

> Deer\_AM\_Age\_BM <- lm(formula = AM~BM+Age, data=Deer)

> Deer\_AM\_Age\_BM

Call:

lm(formula = AM ~ BM + Age, data = Deer)

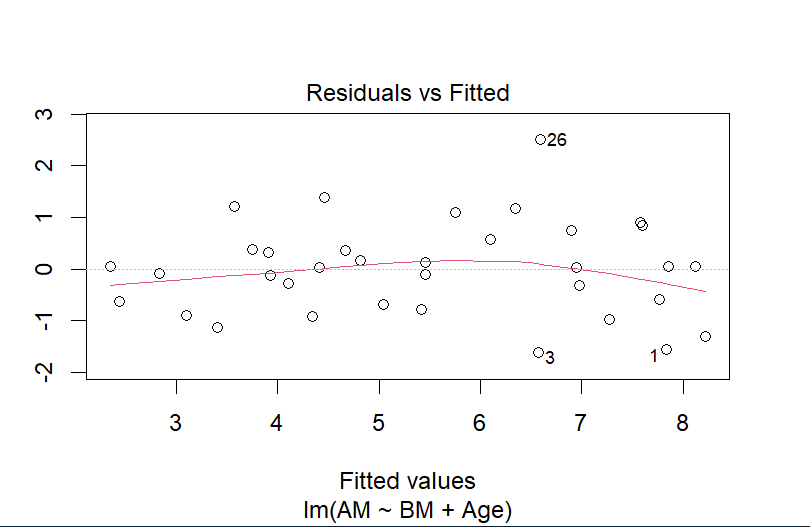
Coefficients:

(Intercept) BM Age

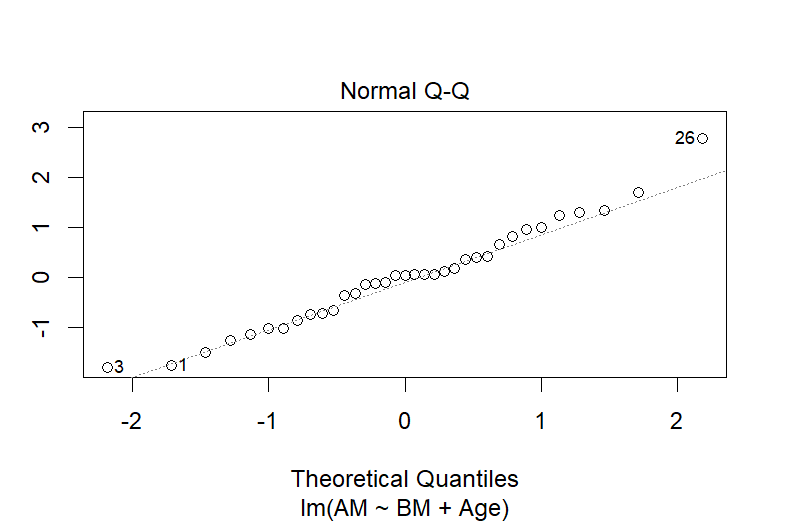
-2.00927 0.02587 0.42177

2.3) Przedstaw wykresy diagnostyczne mające na celu sprawdzenie założeń analizy i zdecyduj, czy te założenia można przyjąć. Jeśli nie, zaproponuj i wykonaj odpowiednie działania zaradcze.

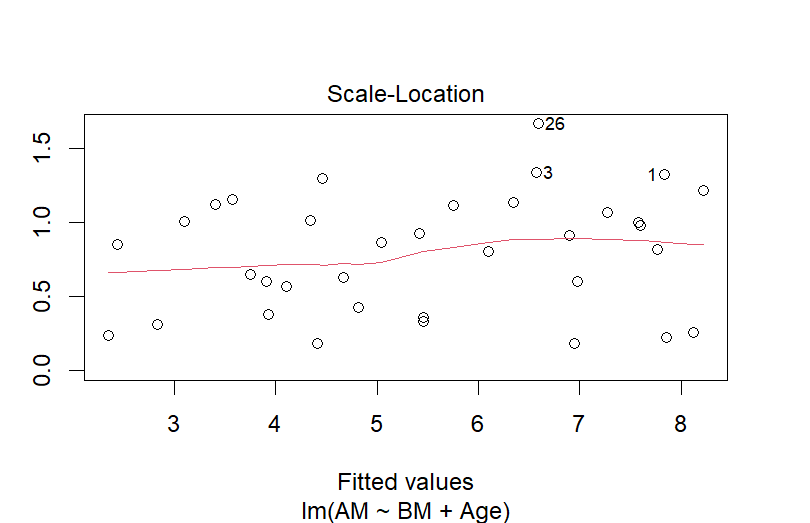
plot(Deer\_AM\_Age\_BM, 1)



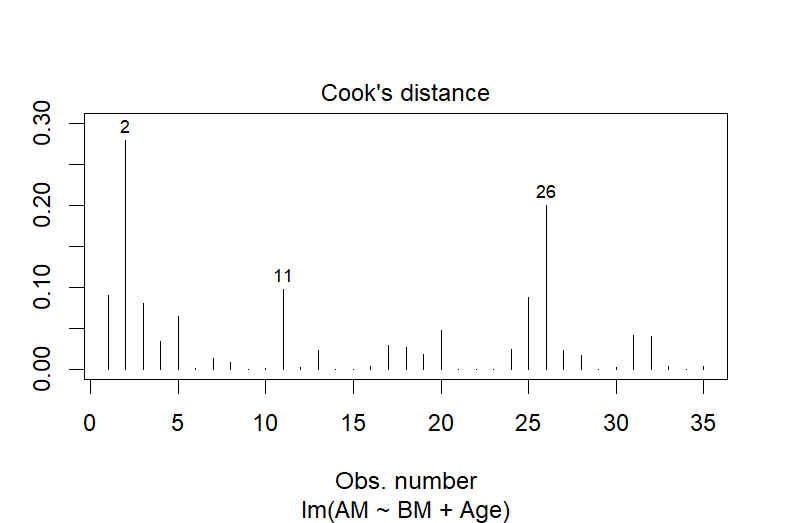
plot(Deer\_AM\_Age\_BM, 2)



plot(Deer\_AM\_Age\_BM, 3)



plot(Deer\_AM\_Age\_BM, 4)



2.4) Przedstaw tabele z kompletnymi wynikami analizy, zawierającymi oszacowania parametrów badanej zależności z ich błędami standardowymi i przedziałami ufności, testy istotności oraz informację o tym, jak duża część wariancji jest wyjaśniana przez model. Objaśnij sens wartości pojawiających się na wydruku.

> summary(Deer\_AM\_BM)

Call:

lm(formula = AM ~ BM, data = Deer)

Residuals:

Min 1Q Median 3Q Max

-2.4964 -0.9724 -0.0679 0.8158 3.8184

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.147488 1.541834 -2.041 0.0493 \*

BM 0.055027 0.009708 5.668 2.56e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.441 on 33 degrees of freedom

Multiple R-squared: 0.4933, Adjusted R-squared: 0.478

F-statistic: 32.13 on 1 and 33 DF, p-value: 2.565e-06

> summary(Deer\_AM\_Age)

Call:

lm(formula = AM ~ Age, data = Deer)

Residuals:

Min 1Q Median 3Q Max

-1.8033 -0.8819 -0.2256 0.6400 2.6033

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.08821 0.51031 2.132 0.0405 \*

Age 0.53962 0.05859 9.210 1.22e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.071 on 33 degrees of freedom

Multiple R-squared: 0.7199, Adjusted R-squared: 0.7114

F-statistic: 84.82 on 1 and 33 DF, p-value: 1.22e-10

> summary(Deer\_AM\_Age\_BM)

Call:

lm(formula = AM ~ BM + Age, data = Deer)

Residuals:

Min 1Q Median 3Q Max

-1.61259 -0.65823 0.03031 0.47299 2.50169

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.009266 1.010621 -1.988 0.05541 .

BM 0.025870 0.007583 3.411 0.00177 \*\*

Age 0.421773 0.061560 6.851 9.46e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9316 on 32 degrees of freedom

Multiple R-squared: 0.7946, Adjusted R-squared: 0.7818

F-statistic: 61.9 on 2 and 32 DF, p-value: 1.003e-11

2.5) Dla modelu z dwoma zmiennymi uwzględnionymi na raz wykonaj osobno test istotności całego modelu.

> confint(Deer\_AM\_Age\_BM)

2.5 % 97.5 %

(Intercept) -4.06783360 0.04930207

BM 0.01042306 0.04131643

Age 0.29637979 0.54716594

2.6) Przedstaw adekwatny wykres ilustrujący najważniejsze wyniki.

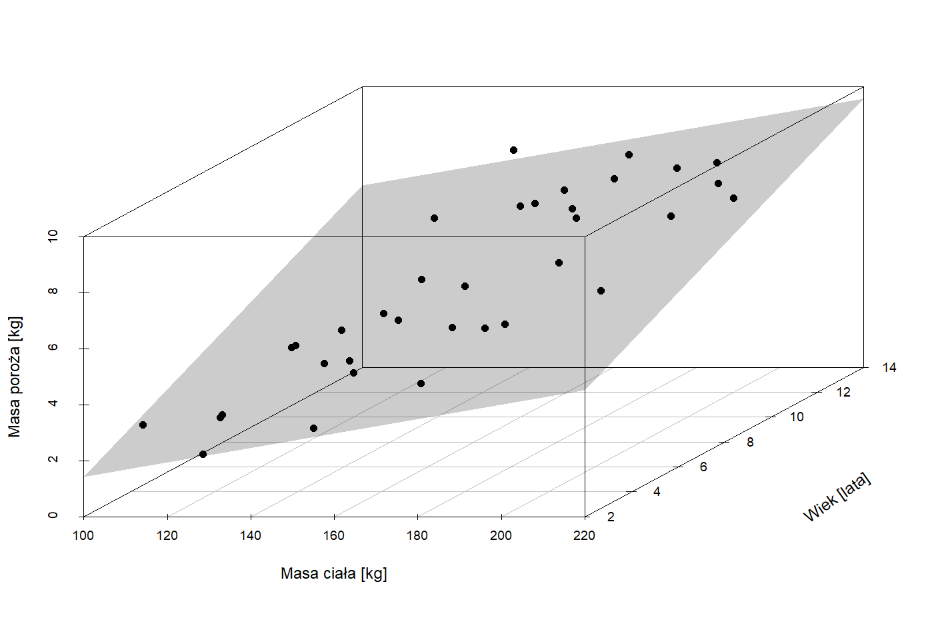
scatterplot3d(x = Deer$BM, y = Deer$Age, z = Deer$AM, , xlab = "Masa ciała [kg]", ylab="", zlab = "Masa poroża [kg]", pch=16)$plane3d(Deer\_AM\_Age\_BM, draw\_polygon=TRUE, draw\_lines = FALSE)

dims <- par("usr")

x <- dims[1]+ 0.9\*diff(dims[1:2])

y <- dims[3]+ 0.08\*diff(dims[3:4])

text(x,y,expression("Wiek [lata]"),srt=35)



2.7) Przedstaw i uzasadnij wniosek z wykonanej analizy, odnoszący się do stawianej hipotezy roboczej. Przedstaw ten wniosek w możliwie najpełniejszej formie, a więc nie poprzestający tylko stwierdzeniu obecności albo braku zależności, ale też na ilościowej charakterystyce tej zależności.

Po wykonanej analizie odrzucamy hipotezę zerową i przyjmujemy alternatywną gdyż p-value jest mniejsze od 0.05. Zatem masa ciała jak i wiek jelenia wpływa na masę jego poroża. Powyższe wykresy również pokazują że im cięższe i starsze są jelenie, tym ich poroża mają większą masę.

2.8) Oblicz, jaka jest przewidywana masa poroża dla byka o masie 180kg i wieku 8 lat.

Y = −2.009266 + 0.025870 ∗ 180kg + 0.421773 ∗ 8 = 6,021518kg

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*