Zadanie 1. Będziemy rozważać pewne dane przy użyciu następującego modelu regresji liniowej wielorakiej

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

dla którego estymatory $\beta_0,\ \beta_1,\ \beta_2,\ \sigma$ wynoszą kolejno $b_0 = 1,\ b_1 = 4,\ b_2 = 3,\ s = 3$.
W pierwszym kroku dokonamy predykcji wartości Y dla $X_1 = 2$ oraz $X_2 = 6$. Korzystając z podanych estymatorów otrzymujemy, że $\hat{Y} = 1 + 4 \cdot 2 + 3 \cdot 6 = 27$.
W dalszej części naszym celem będzie estymacja wartości wariancji błędu predykcji zmiennej Y_h tj. $\sigma^2(pred)$ przy założeniu, że $X_{h,1} = 2,\ X_{h,2} = 6$ oraz $\hat{\mu_h} = 2$. Zauważmy, że

$$s^2(pred) = s^2 + s^2(\hat{\mu_h}) = 4 + 9 = 13,$$

a więc $s(pred)$ jest w przybliżeniu równy 3.61.
Na koniec założymy dodatkowo, że powyższy model zbudowany został na podstawie zbioru 20 obserwacji i estymator odchylenia standardowego b_1, $s(b_1)$, jest równy 1. Naszym celem będzie konstrukcja 95% przedziału ufności dla β_1. Przypomnijmy, że jest on postaci

$$b_1 \pm t_c\ s(b_1),$$

gdzie t_c jest kwantylem rzędu 0.95 z rozkładu studenta studenta z 17 stopniami swobody. Korzystając z `R` otrzymujemy, że t_c wynosi 2.11, a zatem szukany przedział ufności jest równy

$$(1.89,\ 6.11).$$

Zadanie 2. Analizujemy pewne dane przy użyciu poniższego modelu regresji liniowej wielorakiej

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

dla którego sumy I, II typu wynoszą:

	Type I	Type II
X_1	300	30
X_2	40	25
X_3	20	?

Zakładamy dodatkowo, że SST wynosi 760, a n, tj. liczba obserwacji, wynosi 24
W pierwszym kroku wyznaczymy wartość sumy II typu dla X_3. Z definicji jest ona równa sumie I typu dla X_3, a zatem $SSM(X_3|X_1, X_2)$ wynosi 20.
W dalszej części, przy użyciu ogólnego testu F, sprawdzać będziemy istotność regresora X_1, tj. będziemy porównywać modele:

$$H_0:\ \text{dane pochodzą z modelu } Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

$$H_1:\ \text{dane pochodzą z modelu } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

Statystyka testowa F wynosi:

$$F = \frac{SSE(R) - SSE(F)}{(dfE(R) - dfE(F))MSE(F)}.$$

Zauważmy, że

$$SSE(R) - SSE(F) = SSM(F) - SSM(R) = SSM(X_1, X_2, X_3) - SSM(X_2, X_3) = SSM(X_1|X_2, X_3),$$

a zatem, na podstawie tabeli, otrzymujemy, że $SSE(R) - SSE(F) = 30$. Dodatkowo $dfE(F) = 24 - 4 = 20$, a $SSE(F) = 760 - 300 - 40 - 20 = 400$, a zatem ostatecznie $F = 60 \cdot 20^{-1} = 1.5$. Przy użyciu `R` otrzymujemy, że F^* tj. kwantyl rzędu 0.95 z rozkładu Fishera-Snedecora o 1, 20 stopniach

swobody wynosi 4.35, a zatem $F < F^*$ co w szczególności oznacza, że nie możemy odrzucić hipotezy $H_0 : \beta_1 = 0$.
Następnie testować będziemy istotność regresorów X_2, X_3 tj. będziemy porównywać modele:

$$H_0 : \text{ dane pochodzą z modelu } Y = \beta_0 + \beta_1 X_1 + \epsilon$$

$$H_1 : \text{ dane pochodzą z modelu } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

Zauważmy, że

$$SSE(R) - SSE(F) = SSM(F) - SSM(R) = SSM(X_1, X_2, X_3) - SSM(X_1),$$

a zatem, na podstawie tabeli, otrzymujemy, że $SSE(R) - SSE(F) = 60$, zaś $F = 1.5$ Odczytujemy, że F^* tj. kwantyl rzędu 0.95 z rozkładu Fishera-Snedecora o 2, 20 stopniach swobody wynosi 3.49, a zatem $F < F^*$ co w oznacza, że nie możemy odrzucić hipotezy $H_0 : \beta_2 = \beta_3 = 0$.
Kolejny test będzie miał za zadanie zbadanie, czy którakolwiek ze zmiennych objaśniających ma wpływa na zmienną wynikową tzn. testować będziemy

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : (\exists i \in \{1, 2, 3\})\ \beta_i \neq 0.$$

Statystyka testowa F wynosi

$$F = \frac{360}{3 \cdot 20} = 360 \cdot 60^{-1} = 6.$$

F^* tj. kwantyl rzędu 0.95 z rozkładu Fishera-Snedecora o 3, 20 stopniach swobody wynosi 3.1, a zatem $F > F^*$. Otrzymujemy więc wniosek, że przynajmniej jeden z parametrów $\beta_1 = \beta_2 = \beta_3$ jest różny od 0, czyli przynajmniej jeden z regresorów X_1, X_2, X_3 wpływa na Y.
W kolejnym kroku będziemy analizować omawiane dane przy założeniu, że nie uwzględniamy zmiennych X_2, X_3. Testować będziemy czy Y jest w relacji liniowej z X_1 tj.

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

Statystyka testowa F dla powyższego problemu wynosi

$$F = \frac{MSM}{MSE} = \frac{30 \cdot 22}{1 \cdot 400} = 16.5.$$

F^* tj. kwantyl rzędu 0.95 z rozkładu Fishera-Snedecora o 1, 20 stopniach swobody wynosi 4.3, a zatem $F < F^*$. Odrzucamy więc H_0, przyjmując tym samym, że w tym przypadku Y zależy od X_1 w sposób liniowy.
Na koniec wyznaczymy próbkową korelację między Y, a X_1. Otrzymjemy, że

$$R_1^2 = \frac{SSM(X_1|X_2, X_3)}{SSE(F) + SSM(X_1|X_2, X_3)} = 30 \cdot 430^{-1} = 0.096,$$

a więc korelacja próbkowa między Y, a X_1 wynosi w przybliżeniu ± 0.31.

Zadanie 3. Zaczniemy od wygenerowania macierzy X_{100x2}, której wiersze będą iid wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma/100)$, gdzie

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.$$

```
sigma <- matrix(c(1, 0.9, 0.9, 1), 2, 2)
X <- mvrnorm(100, c(0, 0), sigma/100)
```

Następnie tworzymy wektor odpowiedzi $Y = \beta_1 X_1 + \epsilon$, gdzie $\beta_1 = 3$, X_1 jest pierwszą kolumną macierzy Σ, zaś $\epsilon \sim N(0, \mathbb{1})$.

```
Y <- 3*X[ ,1] + rnorm(100)
```

W kolejnym kroku utworzymy 95% przedziały ufności dla współczynników β_0, β_1 dla modelu $Y = \beta_1 X_1 + \epsilon$ oraz współczynników β_0, β_1, β_2 dla modelu $Y' = \beta_1 X_1 + \beta_2 X_2 \epsilon$. Dla modelu $Y = \beta_1 X_1 + \epsilon$ otrzymujemy

```
model_reduced <- lm(Y ~ X[ ,1])
confint(model_reduced, level = .95)

##                   2.5 %    97.5 %
## (Intercept) -0.2425917 0.1441639
## X[, 1]       1.4146269 5.4865062
```

Dla modelu $Y' = \beta_1 X_1 + \beta_2 X_2 + \epsilon$ zaś otrzymujemy

```
model_full <- lm(Y ~ X[ ,1] + X[ ,2])
confint(model_full, level = .95)

##                   2.5 %     97.5 %
## (Intercept) -0.2465579  0.1357853
## X[, 1]       2.6889873 11.1649568
## X[, 2]      -8.7291423  0.3074096
```

Widzimy więc, że przedział ufności dla β_1 w przypadku obu modeli nie zawiera 0, a zatem dla testów

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

odrzucamy hipotezę H_0, co oznacza, że regresor X_1 jest istotny. Ponadto możemy zauważyć, że dla obu modeli przedział ufności dla β_0 zawiera 0, a więc możemy przyjąć hipotezę mówiącą o nieistotności interceptu. Dodatkowo przedział ufności dla β_2 również zawiera zero, a zatem regresor X_2 nie wpływa w sposób istotny na drugi model. Zauważmy, że otrzymane rezultaty są w pełni zgodne ze sposobem, w jaki skonstruowaliśmy wektor Y.

W dalszej części obliczymy odchylenia standardowe estymatorów β_1 dla omawianych powyżej modeli.

```
sd_reduced <- sqrt(sum((model_reduced$residuals^2))
              /(var(X[,1])*(100-2)^2))
s_full <- sd(model_full$residuals)*sqrt(99/97)
m <- (s_full^2*solve(t(model.matrix(model_full))%*%model.matrix(model_full)))
sd_full <- sqrt(m[2,2])
```

Otrzymujemy następujące wartości

	model Y	model Y'
$s(\hat{\beta}_1)$	1.031	2.135

Następnie obliczymy mocy testów dla β_1 dla modeli Y i Y'.

```
delta <- 3/sd_reduced
power_reduced <- 1 - pt(qt(1-0.05/2, 98), 98, delta) +
                 pt(-qt(1-.05/2, 98), 98, delta)

delta <- 3/sd_full
power_full <- 1 - pt(qt(1-0.05/2, 97), 97, delta) +
              pt(-qt(1-0.05/2, 97), 97, delta)
```

Otrzymujemy następujące wartości

	model Y	model Y'
$\pi(0)$	0.821	0.285

Widzimy zatem, że moc testu dla β_1 w przypadku modelu Y jest bliska 1, co jest zgodne ze sposobem, w jaki konstruowaliśmy wektor Y. W przypadku modelu Y' widzimy, że moc testu jest znacznie słabsza, co jest spowodowane uwzględnieniem zmiennej objaśniającej X_2 w modelu Y', które jest niezgodne z konstrukcją wektora Y.
Na koniec wygenerujemy 1000 niezależnych kopii wektora błędów ϵ i 1000 odpowiadających kopii wektora odpowiedzi. Dla każdego z otrzymanych wektorów odpowiedzi estymujemy β_1 oraz wykonujemy testy istotności β_1 dla zdefiniowanych powyżej modeli Y oraz Y'.

```
licznik_r <- 0
licznik_f <- 20
b1_r <- c()
b1_f <- c()

for(i in 1:1000) {
  error <- rnorm(100, 0, 1)
  Y <- 3 * X[ ,1] + error
  model_reduced <- lm(Y ~ X[ ,1])
  model_full <- lm(Y ~ X[ ,1] + X[ ,2])
  iterval1=confint(model_reduced)[2,]
  iterval2=confint(model_full)[2,]

  if(0 >= confint(model_reduced)[2, 1] && 0 <= confint(model_reduced)[2, 2]) {
    licznik_r <- licznik_r + 1
  }

  if(0 >= confint(model_full)[2, 1] && 0 <= confint(model_full)[2, 2]) {
    licznik_f <- licznik_f + 1
  }

  b1_r[i] <- model_reduced$coefficients[2]
  b1_f[i] <- model_full$coefficients[2]
}
```

Na podstawie uzyskanych danych estymujemy wartości odchylenia standardowego β_1 dla obu modeli (ozn. $s(\tilde{\beta_1})$) . Porównując otrzymane wartości z obliczonymi wcześniej teoretycznymi wartościami $s(\beta_1)$ otrzymujemy

	model Y	model Y'
$s(\tilde{\beta_1})$	1.061	2.150
$s(\beta_1)$	1.031	2.135

Widzimy więc, że teoretyczne i empiryczne odchylenia standardowe dla obu modeli są bardzo zgodne. Następnie estymujemy wartości funkcji mocy dla β_1 dla obu modeli (ozn. $\pi(\tilde{0})$). Porównując otrzymane wartości z obliczonymi wcześniej teoretycznymi wartościami $\pi(0)$ otrzymujemy

	model Y	model Y'
$\pi(\tilde{0})$	0.756	0.181
$\pi(0)$	0.821	0.285

Widzimy, że tutaj również empiryczne moce są zbliżone do podanych wcześniej mocy teoretycznych.

Zadanie 4. Zaczniemy od wygenerowania macierzy $X_{1000x95}$, której elementami są iid zmienne losowe z rozkładu $N(0, 0.1)$.

```r
X <- matrix(rnorm(1000 * 950, 0, 0.1), 1000, 950)
```

Następnie generujemy wektor zmiennych objaśnianych odpowiadający modelowi

$$Y = X\beta + \epsilon,$$

gdzie $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$.

```r
beta <- rep(0, 950)
beta[1:5] <- 3
Y <- X%*%beta+rnorm(1000)
```

Następnie, używając kolejno pierwszych 1, 2, 5, 10, 50, 100, 500, 950 kolumn powstałej macierzy będziemy budować modele regresji liniowej. Naszym celem będzie zbadanie, który z powstałych modeli najlepiej dopasowuje się do danych tj. będziemy badać wpływ wymiarów macierzy planu na własności modelu. Dla każdego ze zbudowanych modeli obliczymy SSE, MSE, wartość AIC, p-wartość odpowiadającą dwóm pierwszym zmiennym objaśniającym oraz liczbę zmiennych, które mają poziom istotności poniżej 0.05 ozn. FD.

```r
k <- c(1,2,5,10,50,100,500,950)
SSE <- rep(0,length(k))
MSE <- rep(0,length(k))
AIC_a <- rep(0,length(k))
p.value <- matrix(0,length(k),2)
FD <- rep(0,length(k))

for (i  in 1:length(k)) {
  model <- lm(Y~X[,1:k[i]])
  SSE[i] <- sum(model$residuals^2)
  MSE[i] <- sum((X%*%beta-model$fit)^2)
  AIC_a[i] <- AIC(model)

  if(i==1) p.value[i,1] <- summary(model)$coefficient[2,4]
  else p.value[i,] <- summary(model)$coefficient[2:3,4]

  if(i>3) FD[i]=sum(summary(model)$coefficient[7:(k[i]),4]<0.05)
}
```

Otrzymujemy następujące wartości:

kolumny	1	2	5	10	50	100	500	950
SSE	1333.49	1239.52	984.92	981.33	940.62	895.655	494.32	71.13
MSE	365.58	273.47	2.39	5.99	46.70	91.67	493.00	916.19
AIC	3131.68	3060.60	2836.69	2843.03	2880.66	2931.67	3137.31	2098.63
FD	0	0	0	0	2	5	29	24

Ponadto otrzymujemy, że p-wartości są w każdym przypadku znacznie mniejsze od 0.05.
Możemy zaobserwować, że wraz z doborem kolejnych zmiennych wyraźnie spada wartość SSE tzn. wzrasta dopasowanie predykcji do danych. Z kolei dla MSE minimum osiągane jest dla modelu budowanego na bazie podmacierzy złożonej z 5 kolumn. Następnie MSE wzrasta wraz z doborem kolejnych zmiennych. Oczywiście zależność ta wynika wprost z konstrukcji macierzy X i wektora Y. Dalej możemy zauważyć, że pod względem kryterium AIC najlepszym modelem jest model zbudowany z 950 zmiennych, ponieważ dla niego wartość AIC jest najmniejsza. Należy jednak pamiętać, że ze względu na konstrukcję kryterium AIC od pewnego momentu wartość dodawanego logarytmu moze przewyższać "karę" za dodawanie kolejnej zmiennej.

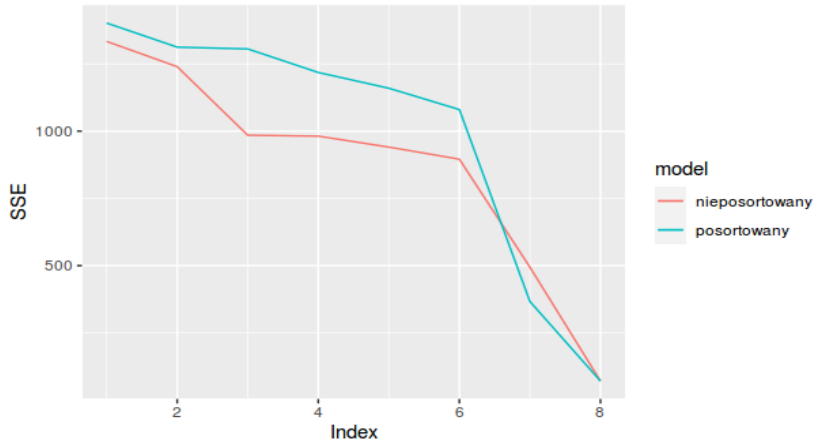
W dalszej części wyznaczymy SSE, MSE, AIC, FD i p-wartości dla pewnej modyfikacji modeli z poprzednich rozważań, która polegać będzie na użyciu zmiennych z największymi estymatorami współczynników regresji.

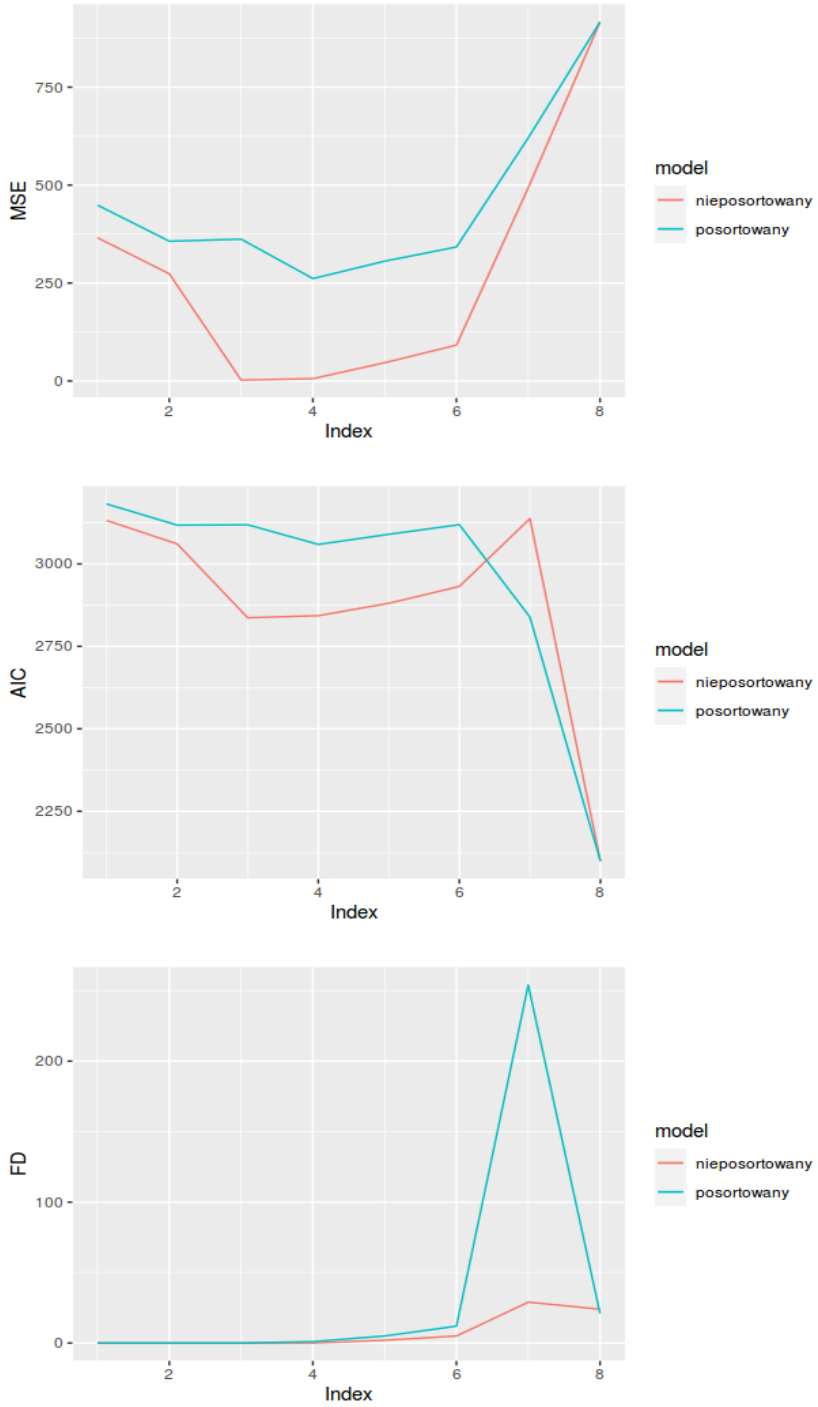
```
k <- c(1,2,5,10,50,100,500,950)
SSE_b <- rep(0,length(k))
MSE_b <- rep(0,length(k))
AIC_b <- rep(0,length(k))
p.value_b <- matrix(0,length(k),2)
FD_b <- rep(0,length(k))
coeff_oreder <- order(abs(summary(model)$coefficient[2:951]),decreasing = TRUE)

for (i  in 1:length(k)) {
  model <- lm(Y~X[,coeff_oreder[1:k[i]]])
  SSE_b[i] <- sum(model$residuals^2)
  MSE_b[i] <- sum((X%*%beta-model$fit)^2)
  AIC_b[i] <- AIC(model)
  if (i==1) p.value_b[i,1] <- summary(model)$coefficient[2,4]
  else p.value_b[i,] <- summary(model)$coefficient[2:3,4]

  if(i>3) FD_b[i] <- sum(summary(model)$coefficient[7:(k[i]),4]<0.05)}
```

Otrzymujemy następujące wartości:

kolumny	1	2	5	10	50	100	500	950
SSE	1401.77	1312.21	1305.59	1217.99	1159.46	1080.12	366.73	71.13
MSE	448.76	356.73	362.02	261.43	305.90	342.08	620.60	916.20
AIC	3181.61	3117.59	3118.53	3059.08	3089.83	3118.95	2838.76	2098.63
FD	0	0	0	1	5	12	254	21

Ponadto otrzymujemy, że p-wartości są w każdym przypadku znacznie mniejsze od 0.05.
Widzimy, że SSE znów jest ściśle malejące. Najmniejsza wartość osiągnięta jest w modelu zawierającym wszystkie zmienne. Dla MSE najmniejsza wartość osiągana jest dla modelu zbudowanego z 50 kolumn macierzy. Kryterium AIC jako najlepszy model znów wskazuje na ten, który zawiera wszystkie zmienne. Najwięcej zmiennych niestotnych pojawia się w przypadku modelu zbudowanego 500 kolumn. W przypadku modeli budowanych z mniejszej liczby kolumn liczba zmiennych nieistotnych jest znacznie mniejsza. W szczególności dla modeli 1, 2, 5-kolumnowych wynosi ona 0.
W dalszej części, przy użyciu odpowiednich wykresów, porównywać będziemy otrzymane modele. Przedstawimy kolejno wykresy SSE, MSE, AIC oraz FD.





Widzimy zatem, że, zgodnie z wynikami podanymi w powyższych tabelach, otrzymane krzywe są dosyć skorelowane. Oznacza to, że zastosowane posortowanie nie miało większego wpływu na wartości SSE, MSE, AIC oraz FD. Z drugiej strony warto zauważyć, że błędy osiągane dla modelu posortowanego są dużo większe niż w przypadku braku sortowania. Również liczba zmiennych nieistotnych jest większa w przypadku modelu posortowanego. Również średni wynik AIC jest lepszy w przypadku modelu nieposortowanego. Podsumowując, na mocy kryterium AIC najlepszym modelem jest model zawierający wszystkie kolumny. Powyższa analiza dostarcza nam dodatkowej informacji o przewadze modelu nieposortowanego nad posortowanym.

Przejdziemy teraz do analizy zbioru danych `CH06PR15.txt`, który zawiera wiek, ciężkość przebiegu choroby, poziom lęku oraz poziom zadowolenia dla pewnej grupy pacjentów.

```
data <- read.table("CH06PR15.txt", header = FALSE)
colnames(data) <- c("age", "severity", "anxiety", "satisfaction")
```

Zadanie 5. W tym zadaniu zajmować będziemy się badaniem zależności pomiędzy poziomem zadowolenia, a wiekiem, ciężkością przebiegu choroby i poziomem lęku. Zaczniemy od zastosowania regresji liniowej wielorakiej zakładając, że regresorem jest poziom zadowolenia, a zmiennymi objaśniającymi są wiek, ciężkość przebiegu choroby i poziom lęku.

```
model <- lm(satisfaction ~ age + severity + anxiety, data)
```

Korzystając z funkcji `summary` otrzymujemy następujące dane dla zbudowanego modelu:

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33589 -0.13333 -0.03347  0.12599  0.52022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.053245   0.613791   1.716  0.09354 .
## age         -0.005861   0.003089  -1.897  0.06468 .
## severity     0.001928   0.005787   0.333  0.74065
## anxiety      0.030148   0.009257   3.257  0.00223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2098 on 42 degrees of freedom
## Multiple R-squared:  0.5415,Adjusted R-squared:  0.5088
## F-statistic: 16.54 on 3 and 42 DF,  p-value: 3.043e-07
```

Widzmy zatem, że dopasowana prosta regresji dana jest wzorem $1.053 - 0.006 * X_1 + 0.002 * X_2 + 0.030 * X_3$, gdzie X_1 to wiek, X_2 to ciężkość przebiegu choroby, a X_3 to poziom lęku. Współczynnik determinacji wynosi jedynie 0.542, a zatem około 54% wariancji w wektorze opowiedzi jest wyjaśniona poprzez regresory. Nie jest to zbyt dobry wynik. Następnie zauważmy, że dla testu

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : (\exists\, i\,)(\beta_i \neq 0)$$

otrzymujemy statystykę testową F równą 16.54 o 3 i 42 stopniach swobody. Obliczając Tc:

```
qf(1 - .05, 3, 42)

## [1] 2.827049
```

widzimy, że wartość F jest znacznie większa od Tc, co świadczy o istnieniu zależności pomiędzy poziomem zadowolenia, a wiekiem, ciężkością przebiegu choroby i poziomem lęku. Oczywiście wynika to również natychmiastowo z faktu, że p-wartość powyższego testu jest bardzo mała - wynosi zaledwie $3.043 \cdot 10^{-7}$.

Zadanie 6. Zaczniemy od podania 95%-przedziałów ufności dla parametrów regresji związanych z wiekiem, ciężkością przebiegu choroby i poziomem lęku w zbudowanym wcześniej modelu.

```
confint(model)[2:4, ]
```

```
##                 2.5 %       97.5 %
## age      -0.01209411 0.0003730895
## severity -0.00974994 0.0136060385
## anxiety   0.01146717 0.0488283055
```

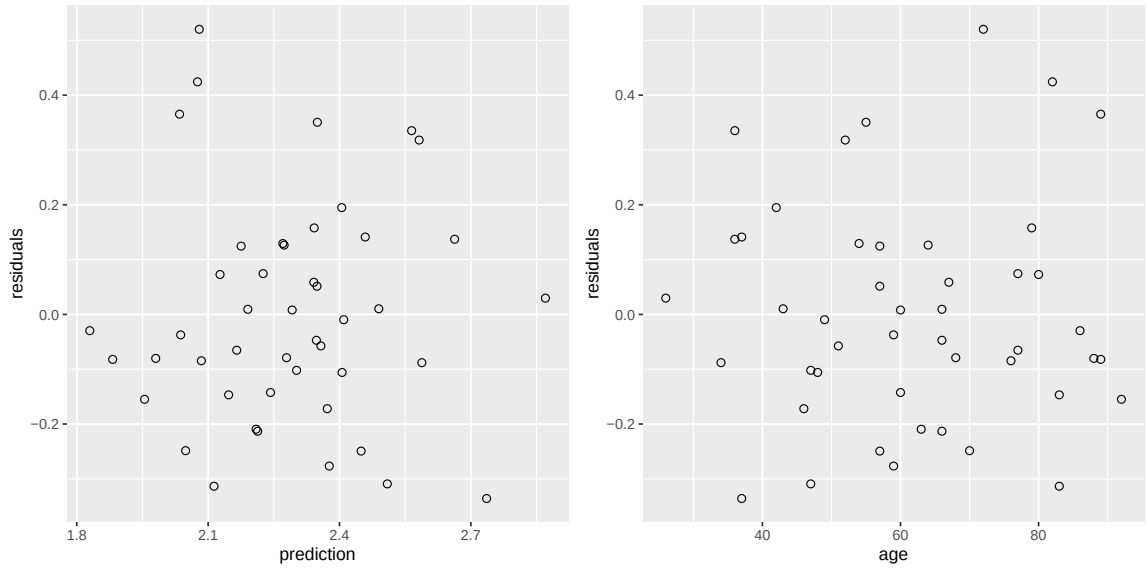
Następnie, korzystając z rezultatów otrzymanych w poprzednim zadaniu za pomocą funkcji `summary` otrzymujemy następujące wyniki:

test	wartość statystyki T	p-wartość
$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$	-1.897	0.065
$H_0: \beta_2 = 0$ vs. $H_1: \beta_2 \neq 0$	0.333	0.741
$H_0: \beta_3 = 0$ vs. $H_1: \beta_3 \neq 0$	3.257	0.002

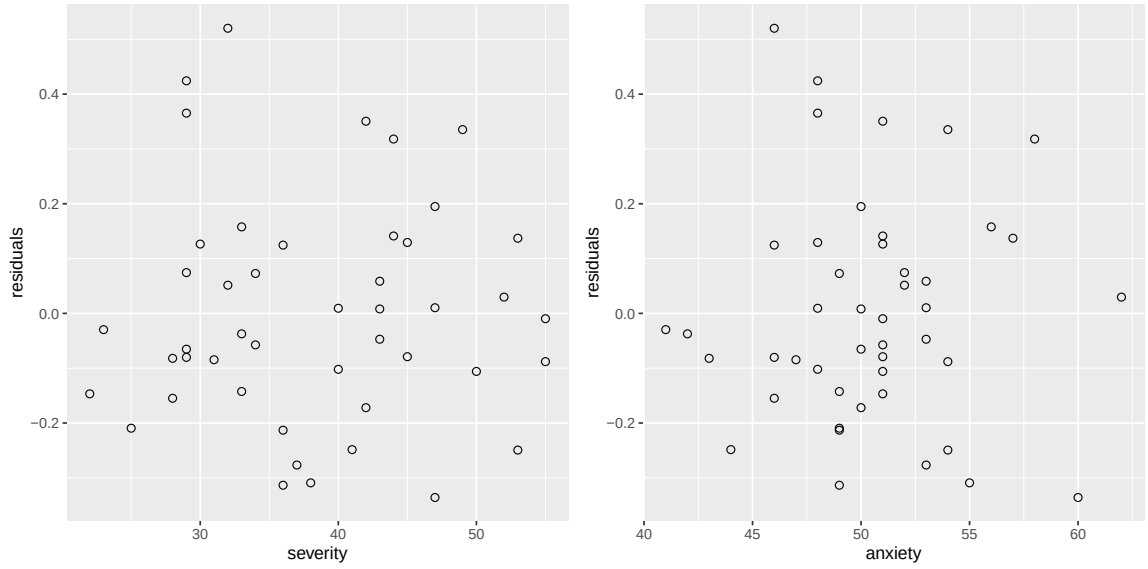
Widzimy zatem na podstawie powyższych testów, że zależność liniowa między poziomem zadowolenia, a rozważanymi zmiennymi objaśniającymi istnieje jedynie w przypadku poziomu lęku. Zauważmy, że jest to w pełni zgodne z otrzymanymi przedziałami ufności. Przedział ufności dla poziomu lęku jako jedyny nie zawiera w sobie 0.

Zadanie 7. W tym zadaniu, na podstawie odpowiednich wykresów rozrzutów podanych niżej, zajmiemy się badaniem zależności między residuami, a kolejno oczekiwanym poziomem satysfakcji i każdą ze zmiennych objaśniających.

```
library(gridExtra)
grid.arrange(r.vs.sat, r.vs.age, ncol = 2)
```




```
grid.arrange(r.vs.severity, r.vs.anxiety, ncol = 2)
```



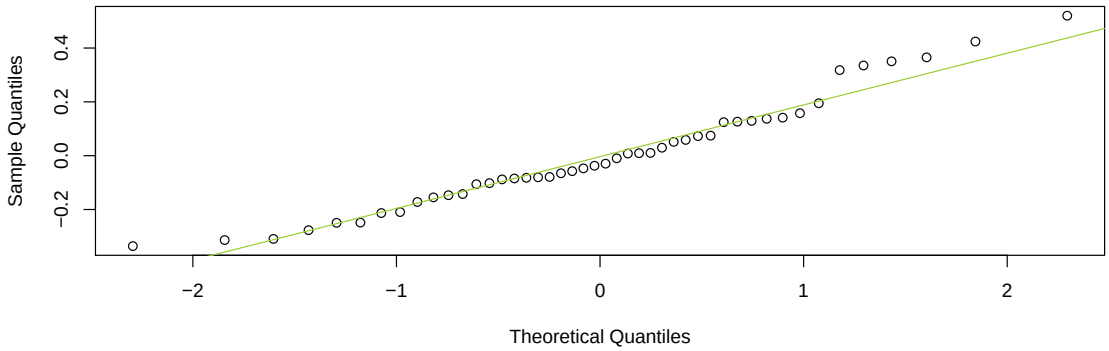
Łatwo zauważyć, że wektor residuów jest niezależny od regresorów, mają strukturę losową. Ponadto wartości residuów są w pewnym stopniu skupione wokół 0. Możemy również zaobserwować występowanie obserwacji odstających.

Zadanie 8. Na koniec zajmiemy się badaniem własności wektora residuów zbudowanego modelu tj. będziemy badać czy pochodzą one z rozkładu normalnego. Zaczniemy od przeprowadzenia testu Shapiro-Wilka, tj. będziemy testować H_0: residua pochodzi z rozkładu normalnego vs. H_1: residua nie pochodzą z rozkładu normalnego.

```
shapiro.test(model$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.96286, p-value = 0.1481
```

Zauważmy, że otrzymana p-wartość jest wysoka, a zatem nie mamy podstaw do odrzucenia hipotezy zerowej. W dalszej części rozważmy wykres kwantylowo-kwantylowy dla residuów.

```
qqnorm(model$residuals, main = "")
qqline(model$residuals, main = "", col = "yellowgreen")
```



Widzimy tutaj, że dopasowanie prostej do reszt jest dość dokładne, a więc rozkład residuów jest bliski rozkładowi normalnemu. Oczywiście jest to w pełni zgodnie z rezultatami uzyskanymi za pomocą testu Shapiro-Wilka.