Regresja liniowa wieloraka

Contents


```r
library(tidyverse)
library(ISLR)
library(caret)
library(ggcorrplot)
library(car)
library(broom)
carseats <- tibble::as_tibble(ISLR::Carseats)
head(carseats)
```

```
## # A tibble: 6 x 11
##   Sales CompPrice Income Advertising Population Price ShelveLoc   Age Education
##   <dbl>     <dbl>  <dbl>       <dbl>      <dbl> <dbl> <fct>     <dbl>     <dbl>
## 1  9.5        138     73          11        276   120 Bad          42        17
## 2 11.2        111     48          16        260    83 Good         65        10
## 3 10.1        113     35          10        269    80 Medium       59        12
## 4  7.4        117    100           4        466    97 Medium       55        14
## 5  4.15       141     64           3        340   128 Bad          38        13
## 6 10.8        124    113          13        501    72 Bad          78        16
## # i 2 more variables: Urban <fct>, US <fct>
```

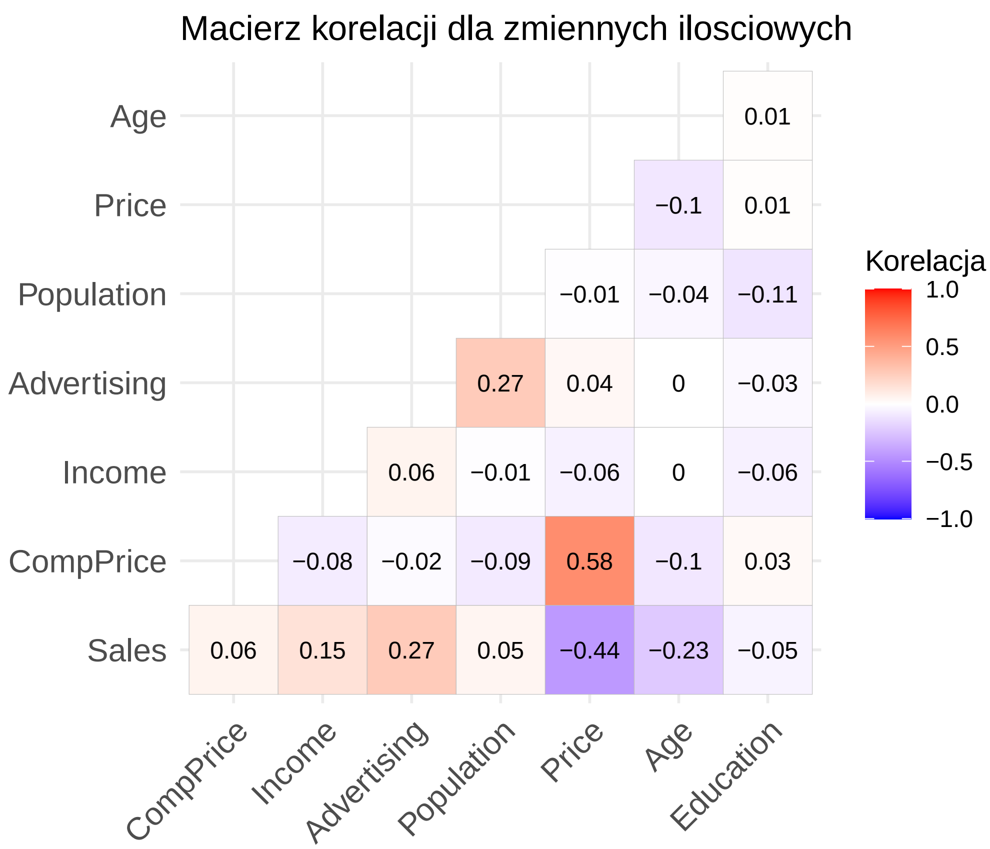
```r
# Podział zbioru na zbiór treningowy i testowy
set.seed(44)
partition <- caret::createDataPartition(carseats$Sales, list=FALSE, p=0.75)
carseats_train <- carseats[partition,]
carseats_test <- carseats[-partition,]
```

```r
model_summary <- function(model, test_data, test_y) {
  model_glance <- broom::glance(model)
  model_augment <- broom::augment(model)
  train_mae <- mean(abs(model_augment$.resid))
  train_mape <- mean(abs(model_augment$.resid/dplyr::pull(model_augment, var=1)))*100
  predicted_y <- predict(model, test_data)
  test_rmse <- sqrt(mean((test_y - predicted_y)^2))
  test_mae <- mean(abs(test_y - predicted_y))
   test_mape <- mean(abs((test_y - predicted_y)/test_y))*100

  cat("\n=========================================\n")
```

```r
  cat("           Podsumowanie modelu    \n")
  cat("=========================================\n\n")
  cat("Metryki treningowe:\n")
  cat("-----------------------------------------\n")
  cat(sprintf("  R-squared (R²):         %.4f\n", model_glance$r.squared))
  cat(sprintf("  Adjusted R-squared:    %.4f\n", model_glance$adj.r.squared))
  cat(sprintf("  Kryterium informacyjne Akaikego (AIC): %.2f\n", model_glance$AIC))
  cat("-----------------------------------------\n\n")
  cat("Charakterystyki \"out-of-sample\":\n")
  cat("==================================================================\n")
  cat(sprintf("  RMSE (trening):         %.4f    |    RMSE (test): %.4f\n", model_glance$sigma, test_rms
  cat(sprintf("  MAE (trening):          %.4f    |    MAE (test):  %.4f\n", train_mae, test_mae))
  cat(sprintf("  MAPE (trening):         %.2f%%    |    MAPE (test): %.2f%%\n", train_mape, test_mape))
  cat("==================================================================\n\n")
}
```

Korelacje między zmiennymi ilościowymi

```r
numeric_vars <- carseats[, sapply(carseats, is.numeric)]
cor_matrix <- cor(numeric_vars)

ggcorrplot(cor_matrix,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           title = "Macierz korelacji dla zmiennych ilościowych",
           legend.title = "Korelacja")
```



Na podstawie macierzy korelacji wybieramy zmienne do modelu, mając na uwadze problemy związane z współliniowością.

Pierwszy model

```
model <- lm(Sales ~ Advertising + ShelveLoc + Price, data = carseats_train)
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + ShelveLoc + Price, data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6799 -1.1957 -0.0414  1.1489  4.0132
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.129964   0.568114  21.351  < 2e-16 ***
## Advertising      0.093584   0.016195   5.778 1.91e-08 ***
## ShelveLocGood    4.648604   0.302192  15.383  < 2e-16 ***
## ShelveLocMedium  1.800412   0.249777   7.208 4.75e-12 ***
## Price           -0.062172   0.004587 -13.555  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.784 on 296 degrees of freedom
```

```
## Multiple R-squared:  0.5908, Adjusted R-squared:  0.5853
## F-statistic: 106.9 on 4 and 296 DF,  p-value: < 2.2e-16
```

```r
# Liniowa niezależność zmiennych objaśniających
vif(model)
```

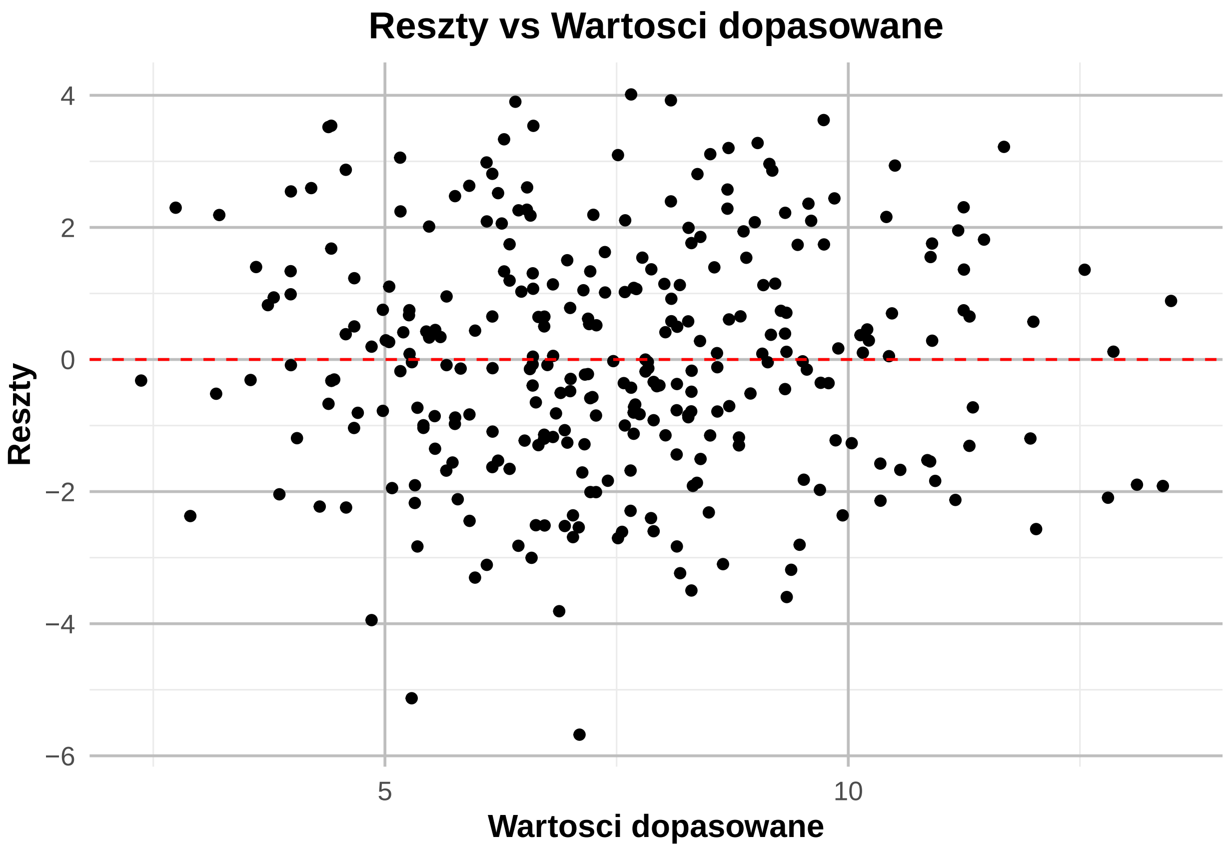
```
##                 GVIF Df GVIF^(1/(2*Df))
## Advertising 1.001132  1        1.000566
## ShelveLoc   1.007814  2        1.001948
## Price       1.007338  1        1.003662
```

Wszystkie zamienne w modelu wykazują bardzo niskie wartości, co wskazuje na bardzo niski poziom kolinearności.

Założenia modelu regresji wielorakiej

```r
# liniowa zależność między zmienną objaśnianą, a objaśniającą postaci
ggplot(augment(model), aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    x = "Wartości dopasowane",
    y = "Reszty",
    title = "Reszty vs Wartości dopasowane"
  ) + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```

```
## Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
## i Please use the `linewidth` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Na wykresie nie widać wyraźnych wzorców ani zakrzywień, co wskazuje na poprawność założenia o liniowej zależności. Punkty są równomiernie rozmieszczone wokół poziomej linii, co dodatkowo sugeruje, że zmienne objaśniające oddziałują na zmienną zależną w sposób liniowy.

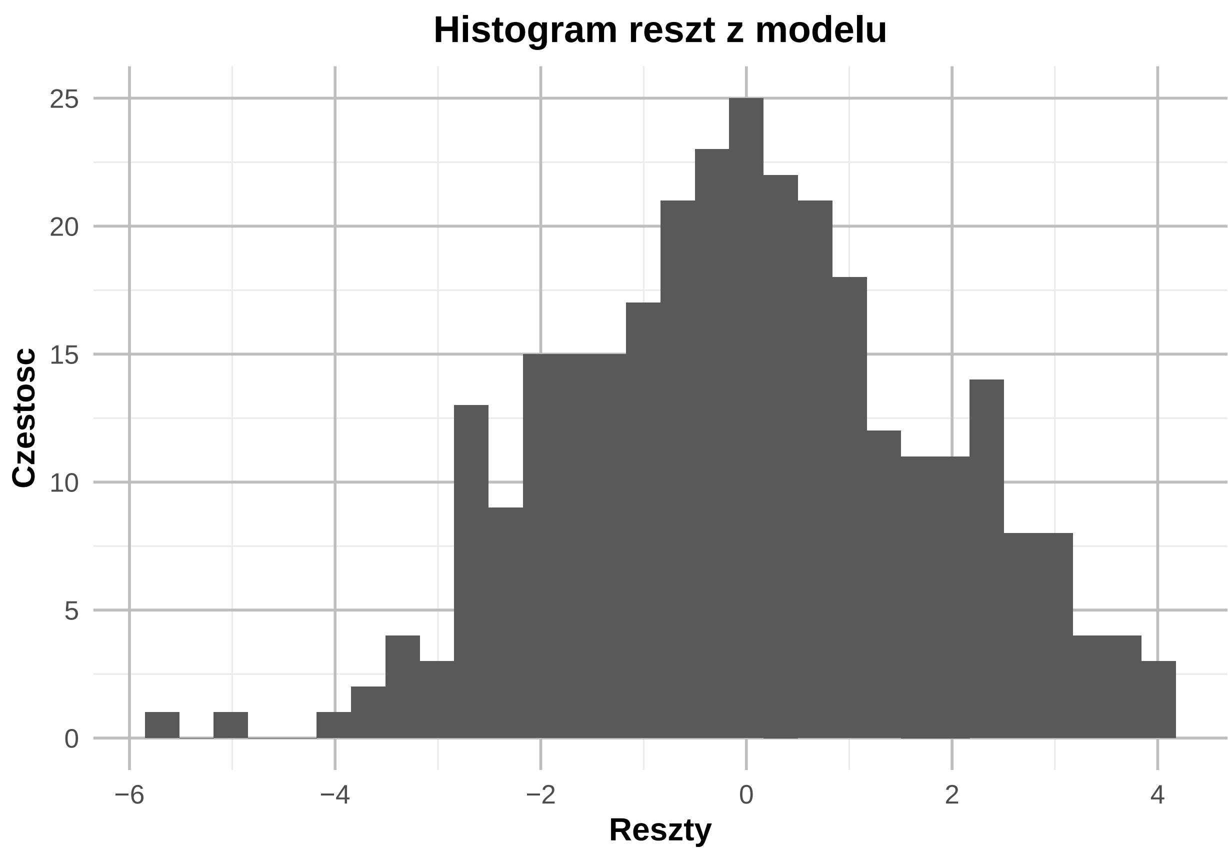
```
# średnia wektora losowego równa 0
t.test(model$residuals)
```

```
##
##  One Sample t-test
##
## data:  model$residuals
## t = 7.3487e-17, df = 300, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.2010052  0.2010052
## sample estimates:
##    mean of x
## 7.506119e-18
```

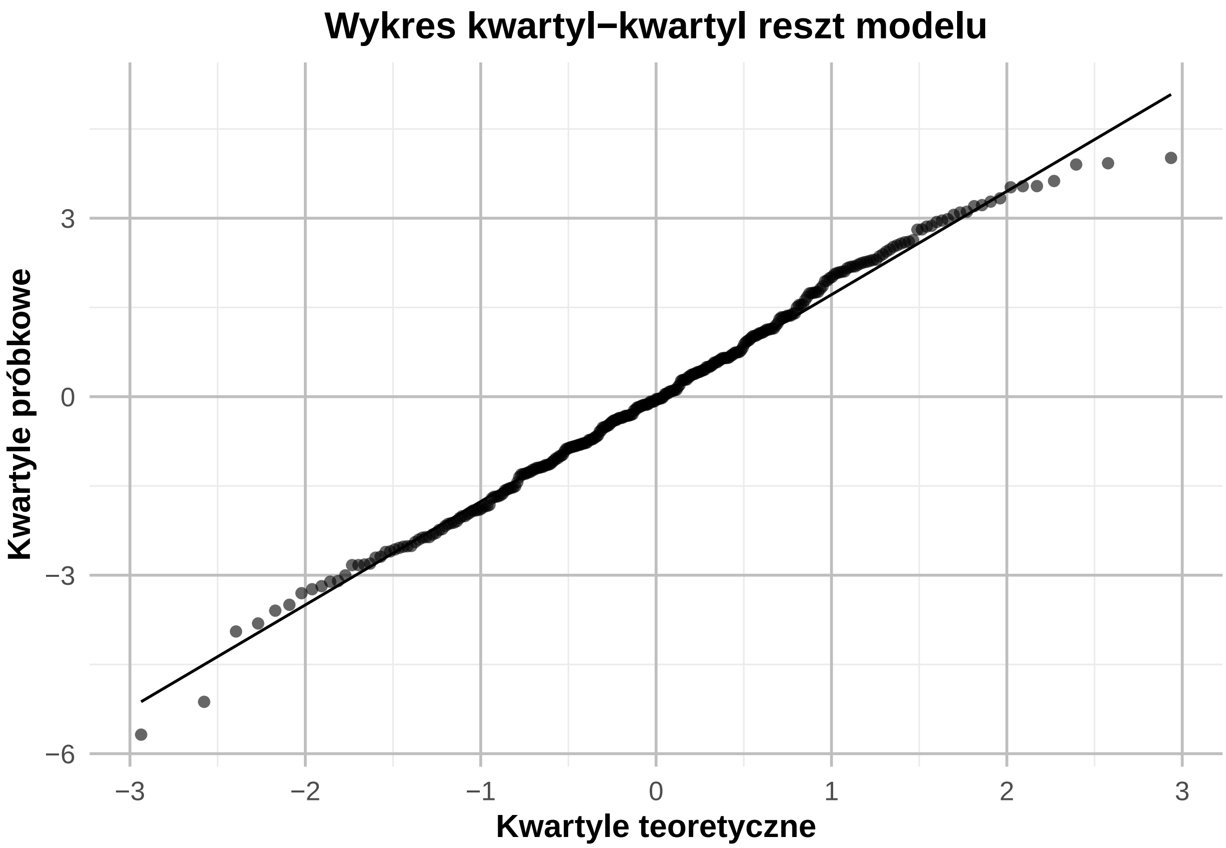
Test wykazał, że należy odrzucić hipotezę alternatywną oraz możemy przyjąć, że prawdziwa jest hipoteza zerowa mówiąca, że średnia reszt jest równa zero.

```
# Sprawdzenie rozkładu reszt
ggplot(model, aes(x=.resid)) + geom_histogram(bins=30) +
  labs(title='Histogram reszt z modelu', x='Reszty', y='Częstość') + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
```

```
)
```




```
ggplot(model, aes(sample = .resid)) +
  geom_qq(size = 1.5, color = "black", alpha = 0.6) +
  geom_qq_line() +
  labs(title = 'Wykres kwartyl-kwartyl reszt modelu',
       x = 'Kwartyle teoretyczne',
       y = 'Kwartyle próbkowe') +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```r
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.99407, p-value = 0.289
```

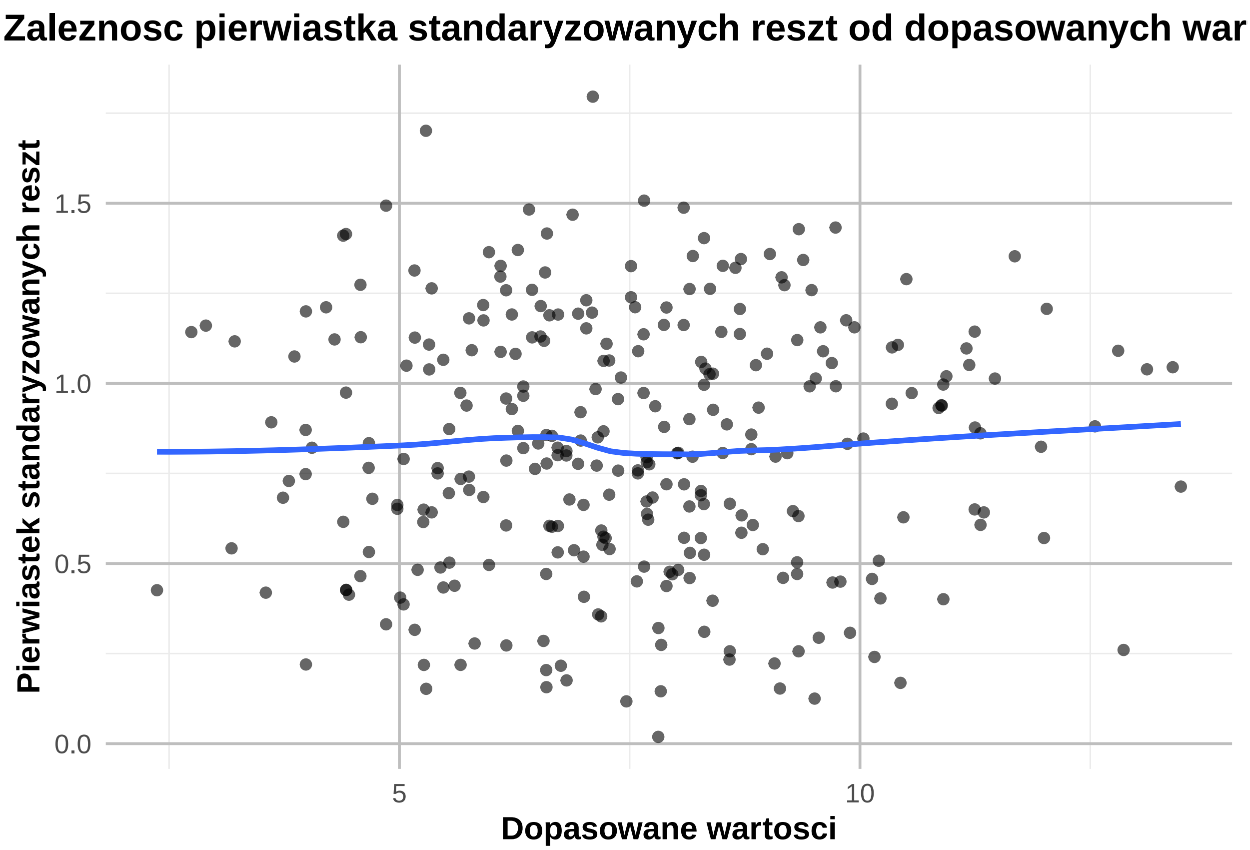
Na histogramie możemy zauważyć lekkie odchylenie reszt modelu, wykres Q-Q pokazuje jednak, że większość punktów skupia się na prostej. Test Shapiro-wilka wskazuje na brak podstaw do odrzucenia hipotezy zerowej mówiącej, że reszty modelu pochodzą z rodkładu normalnego. Więc możemy stwierdzić, że reszty są normalne.

```r
# Sprawdzenie niezależności reszt
lmtest::dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.1637, p-value = 0.9221
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# Homoskedastyczność
ggplot(model, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(size = 1.5, color = "black", alpha = 0.6
  labs(title='Zależność pierwiastka standaryzowanych reszt od dopasowanych wartości', x='Dopasowane war
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
```

```r
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



Punkty na wykresie są równomiernie rozproszone wokół linii, co sugeruje, że wariancja reszt nie zmienia się znacząco w miarę wzrostu wartości dopasowanych.

```r
lmtest::bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.0106, df = 4, p-value = 0.9082
```

Wartość p-value wyniosłą znacznie więcej niż $\alpha = 0{,}05$, oznacza to, że nie mamy istotnych dowodów heteroskedastyczności. Dlatego też możemy wnioskować,że założenie o homoskedastyczności jest prawdziwe dla naszego modelu.

```r
model_summary(model, carseats_test, carseats_test$Sales)
```

```
##
## ========================================
##             Podsumowanie modelu
## ========================================
##
## Metryki treningowe:
## ----------------------------------------
##   R-squared (R²):         0.5908
##   Adjusted R-squared:    0.5853
```

```
##   Kryterium informacyjne Akaikego (AIC): 1209.64
## -----------------------------------------
##
## Charakterystyki "out-of-sample":
## ======================================================================
##   RMSE (trening):         1.7840    |    RMSE (test): 1.7811
##   MAE (trening):          1.4262    |    MAE (test):  1.3664
##   MAPE (trening):         36.21%    |    MAPE (test): Inf%
## ======================================================================
```

- **Interpretacja**:
  - Bardzo niskie p-value oznacza, że co najmniej jedna zmienna w modelu mocno wpływa na zmienną **Sales**.
  - Ujemna korelacja między **Price** a **Sales** wskazuje na spadek sprzedaży wraz ze wzrostem ceny.
  - Największy wpływ na sprzedaż ma dobra lokalizacja półki w sklepie (**ShelveLoc**).
  - Wartości **RMSE** dla treningu i testu są bardzo zbliżone, co sugeruje, że model dobrze generalizuje na danych testowych i nie ma nadmiernego dopasowania.
  - Wartości **MAE** na danych treningowych i testowych są bardzo zbliżone, co jest pozytywnym sygnałem, wskazującym na to, że model dobrze przewiduje zarówno w zbiorze treningowym, jak i testowym.
  - Wartość **MAPE** na danych treningowych wynosi 36.21%, co sugeruje, że model jest dość niedokładny. Wartość **MAPE** na danych testowych wynosi Inf% co wskazuje na to, że model prawdopodobnie ma problemy z wystepującymi zerami.

Pełny model

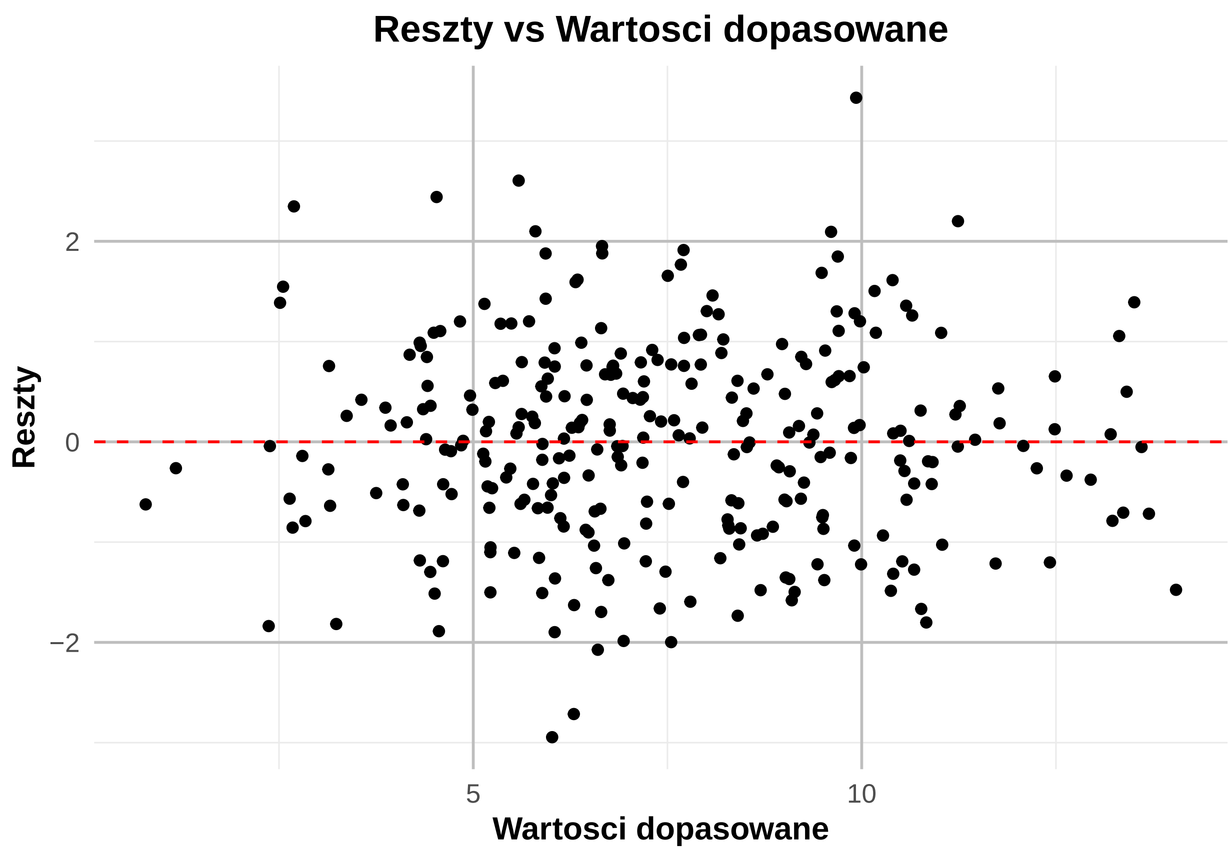
```r
full_model <- lm(Sales ~ ., data = carseats_train)
summary(full_model)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9454 -0.6672  0.0102  0.6733  3.4317
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.261e+00  7.138e-01   8.771  < 2e-16 ***
## CompPrice        9.154e-02  4.845e-03  18.896  < 2e-16 ***
## Income           1.561e-02  2.160e-03   7.228 4.38e-12 ***
## Advertising      1.186e-01  1.345e-02   8.822  < 2e-16 ***
## Population      -3.946e-05  4.212e-04  -0.094    0.925
## Price           -9.592e-02  3.115e-03 -30.791  < 2e-16 ***
## ShelveLocGood    4.767e+00  1.755e-01  27.158  < 2e-16 ***
## ShelveLocMedium  1.873e+00  1.453e-01  12.891  < 2e-16 ***
## Age             -4.564e-02  3.666e-03 -12.450  < 2e-16 ***
## Education       -3.701e-02  2.298e-02  -1.610    0.108
## UrbanYes         3.131e-02  1.380e-01   0.227    0.821
## USYes           -1.373e-01  1.721e-01  -0.798    0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.019 on 289 degrees of freedom
## Multiple R-squared:  0.8697, Adjusted R-squared:  0.8648
## F-statistic: 175.4 on 11 and 289 DF,  p-value: < 2.2e-16
```

Założenia modelu regresji wielorakiej

```r
# liniowa zależność między zmienną objaśnianą, a objaśniającą postaci
ggplot(augment(full_model), aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    x = "Wartości dopasowane",
    y = "Reszty",
    title = "Reszty vs Wartości dopasowane"
  ) + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



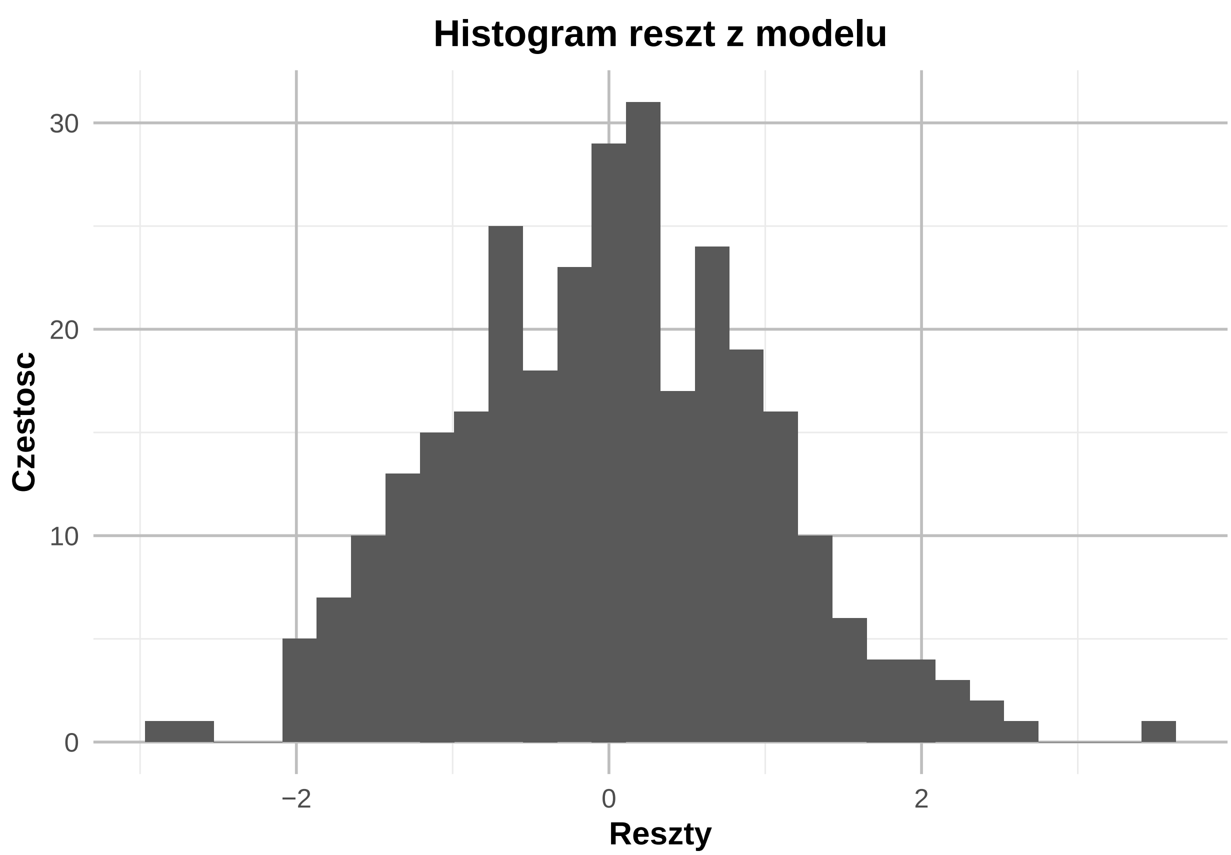
Na wykresie nie widać wyraźnych wzorców ani zakrzywień, co wskazuje na poprawność założenia o liniowej zależności. Punkty są równomiernie rozmieszczone wokół poziomej linii, co dodatkowo sugeruje, że zmienne objaśniające oddziałują na zmienną zależną w sposób liniowy.

```r
# średnia wektora losowego równa 0
t.test(full_model$residuals)
```

```
##
##  One Sample t-test
##
## data:  full_model$residuals
## t = 6.9103e-16, df = 300, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1134107  0.1134107
## sample estimates:
##    mean of x
## 3.982444e-17
```

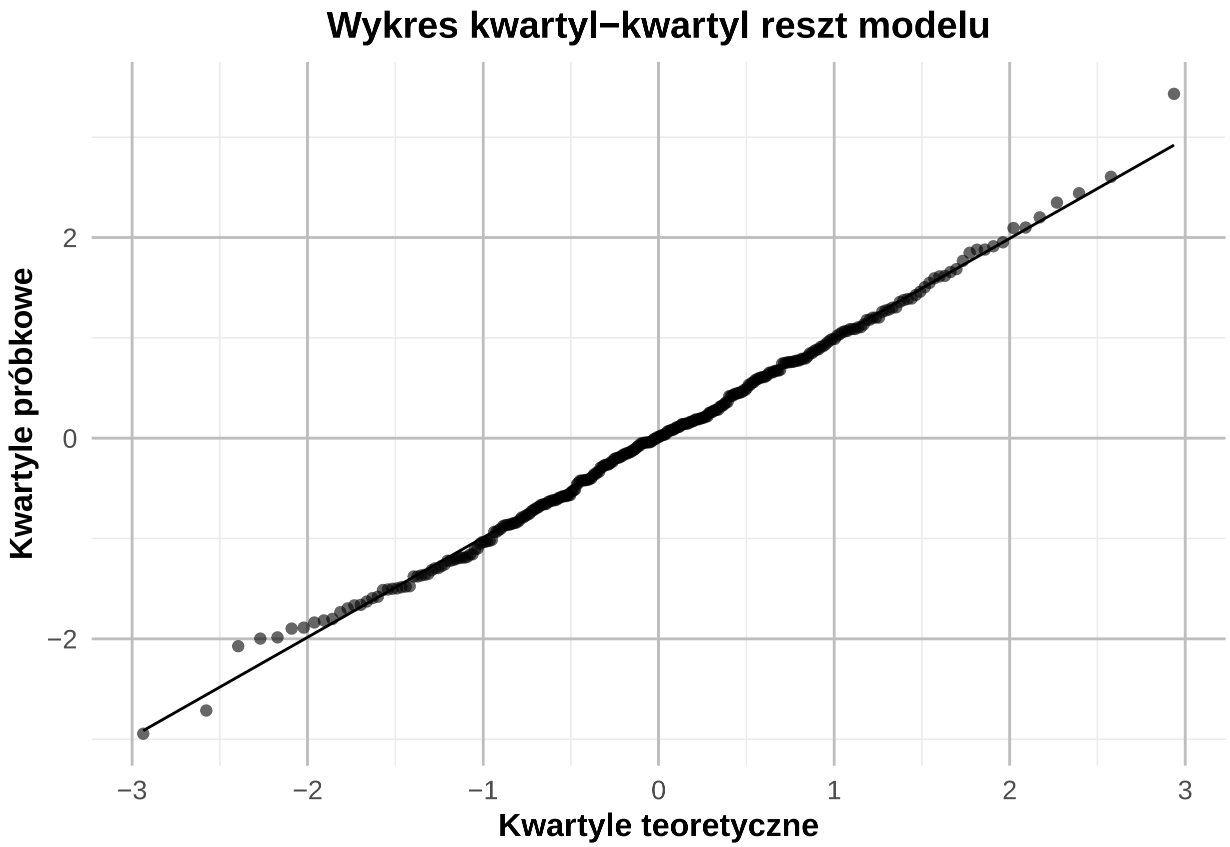
Test wykazał, że należy odrzucić hipotezę alternatywną oraz możemy przyjąć, że prawdziwa jest hipoteza zerowa mówiąca, że średnia reszt jest równa zero.

```r
# Sprawdzenie rozkładu reszt
ggplot(full_model, aes(x=.resid)) + geom_histogram(bins=30) +
  labs(title='Histogram reszt z modelu', x='Reszty', y='Częstość') + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```r
ggplot(full_model, aes(sample = .resid)) +
  geom_qq(size = 1.5, color = "black", alpha = 0.6) +
  geom_qq_line() +
  labs(title = 'Wykres kwartyl-kwartyl reszt modelu',
```

```r
      x = 'Kwartyle teoretyczne',
      y = 'Kwartyle próbkowe') +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```r
shapiro.test(full_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  full_model$residuals
## W = 0.99762, p-value = 0.9441
```

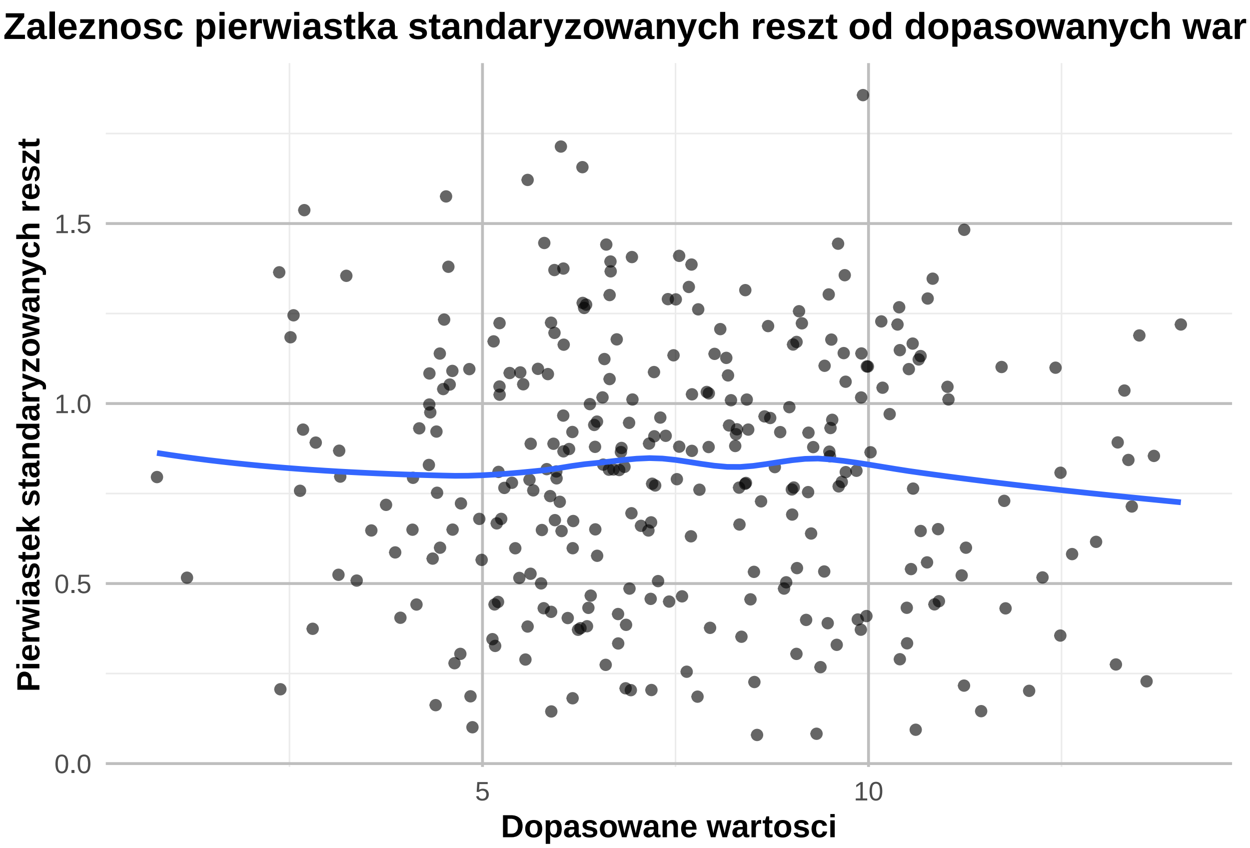
Na histogramie możemy zauważyć lekkie odchylenie reszt modelu, wykres Q-Q pokazuje jednak, że większość punktów skupia się na prostej. Test Shapiro-wilka wskazuje na brak podstaw do odrzucenia hipotezy zerowej mówiącej, że reszty modelu pochodzą z rodkładu normalnego. Więc możemy stwierdzić, że reszty są normalne.

```r
# Sprawdzenie niezależności reszt
lmtest::dwtest(full_model)
```

```
##
##  Durbin-Watson test
##
## data:  full_model
```

```
## DW = 2.1608, p-value = 0.9201
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# Homoskedastyczność
ggplot(full_model, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(size = 1.5, color = "black", alpha =
  labs(title='Zależność pierwiastka standaryzowanych reszt od dopasowanych wartości', x='Dopasowane wart
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



Punkty na wykresie są równomiernie rozproszone wokół linii, co sugeruje, że wariancja reszt nie zmienia się znacząco w miarę wzrostu wartości dopasowanych.

```
lmtest::bptest(full_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  full_model
## BP = 7.7949, df = 11, p-value = 0.7316
```

Wartość p-value wyniosłą znacznie więcej niż $\alpha = 0{,}05$, oznacza to, że nie mamy istotnych dowodów heteroskedastyczności. Dlatego też możemy wnioskować,że założenie o homoskedastyczności jest prawdziwe dla naszego modelu.

```
model_summary(full_model, carseats_test, carseats_test$Sales)
```

```
##
## ==========================================
##             Podsumowanie modelu
## ==========================================
##
## Metryki treningowe:
## ------------------------------------------
##   R-squared (R²):         0.8697
##   Adjusted R-squared:    0.8648
##   Kryterium informacyjne Akaikego (AIC): 879.11
## ------------------------------------------
##
## Charakterystyki "out-of-sample":
## ============================================================================
##   RMSE (trening):         1.0187    |    RMSE (test): 1.0355
##   MAE (trening):          0.7930    |    MAE (test):  0.8425
##   MAPE (trening):         15.02%    |    MAPE (test): Inf%
## ============================================================================
```

```r
vif(full_model)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## CompPrice   1.457139  1        1.207120
## Income      1.030029  1        1.014903
## Advertising 2.116967  1        1.454980
## Population  1.122434  1        1.059450
## Price       1.425198  1        1.193817
## ShelveLoc   1.065004  2        1.015869
## Age         1.037762  1        1.018706
## Education   1.031747  1        1.015749
## Urban       1.069185  1        1.034014
## US          2.007858  1        1.416989
```

Pełny model wyjaśnia 85% zmienności w danych. Kryterium AIC sugeruje, że jest najlepiej dopasowany, ale jego złożoność może być problematyczna.

Model uproszczony

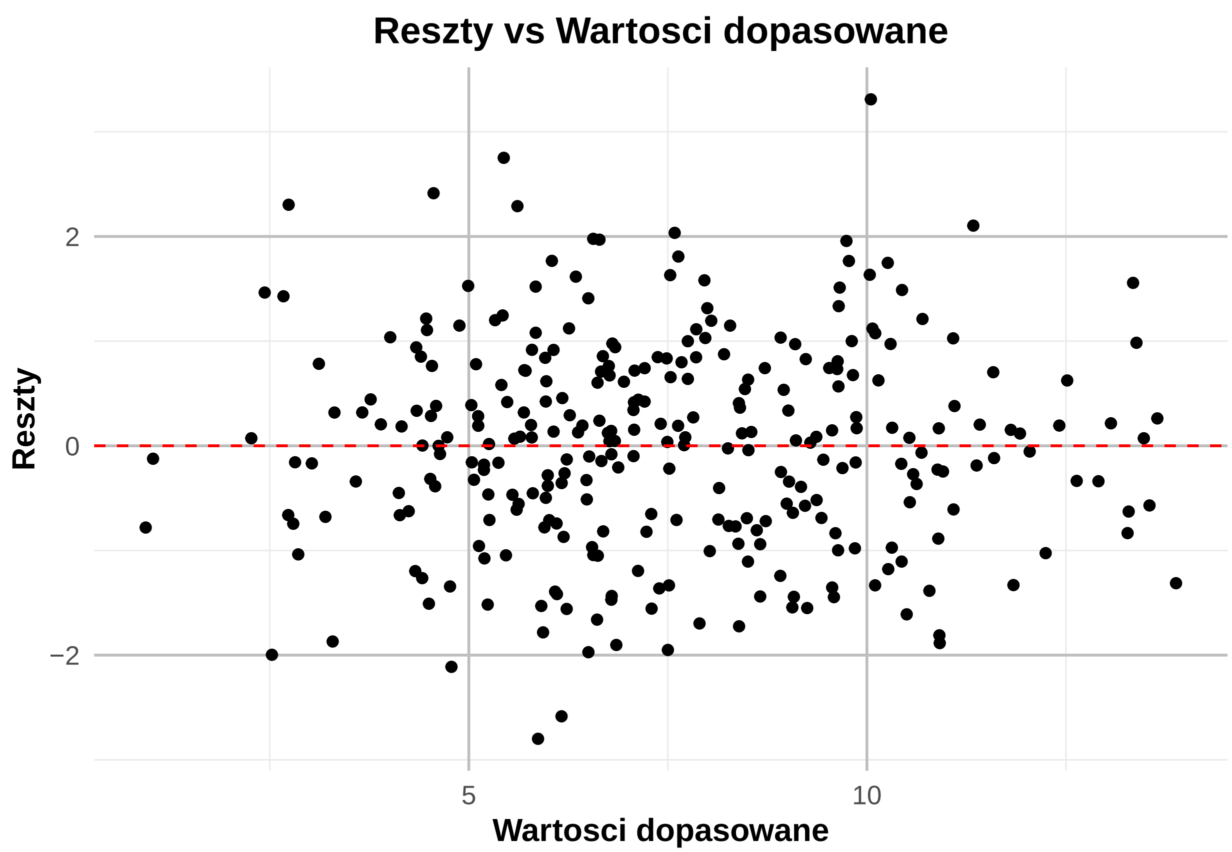
```r
model4 <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age, data = carseats_train)
summary(model4)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = carseats_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7992 -0.7044  0.0462  0.7169  3.3101
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.782599   0.625425   9.246  < 2e-16 ***
## CompPrice       0.090669   0.004781  18.964  < 2e-16 ***
## Income          0.015768   0.002150   7.334  2.2e-12 ***
```

```
## Advertising      0.111663   0.009288  12.022  < 2e-16 ***
## Price           -0.095647   0.003090 -30.951  < 2e-16 ***
## ShelveLocGood    4.762565   0.172639  27.587  < 2e-16 ***
## ShelveLocMedium  1.881786   0.143089  13.151  < 2e-16 ***
## Age             -0.045672   0.003642 -12.540  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.017 on 293 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8651
## F-statistic: 275.9 on 7 and 293 DF,  p-value: < 2.2e-16
```

Założenia modelu regresji wielorakiej

```r
# liniowa zależność między zmienną objaśnianą, a objaśniającą postaci - trzeba dopisac
ggplot(augment(model4), aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    x = "Wartości dopasowane",
    y = "Reszty",
    title = "Reszty vs Wartości dopasowane"
  ) + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



Na wykresie nie widać wyraźnych wzorców ani zakrzywień, co wskazuje na poprawność założenia o liniowej zależności. Punkty są równomiernie rozmieszczone wokół poziomej linii, co dodatkowo sugeruje, że zmienne objaśniające oddziałują na zmienną zależną w sposób liniowy.

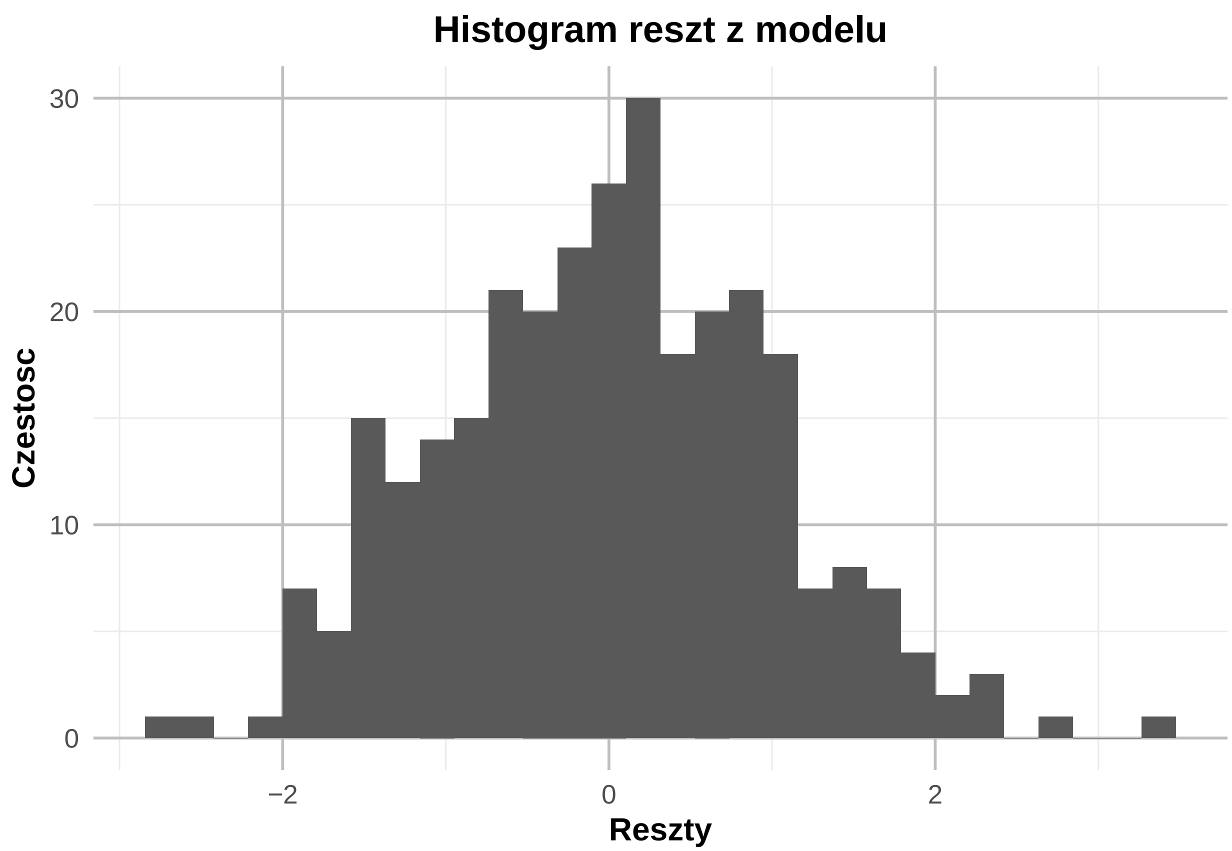
```
# średnia wektora losowego równa 0
t.test(model4$residuals)
```

```
##
##  One Sample t-test
##
## data:  model4$residuals
## t = 8.9109e-16, df = 300, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.1140412  0.1140412
## sample estimates:
##    mean of x
## 5.163936e-17
```

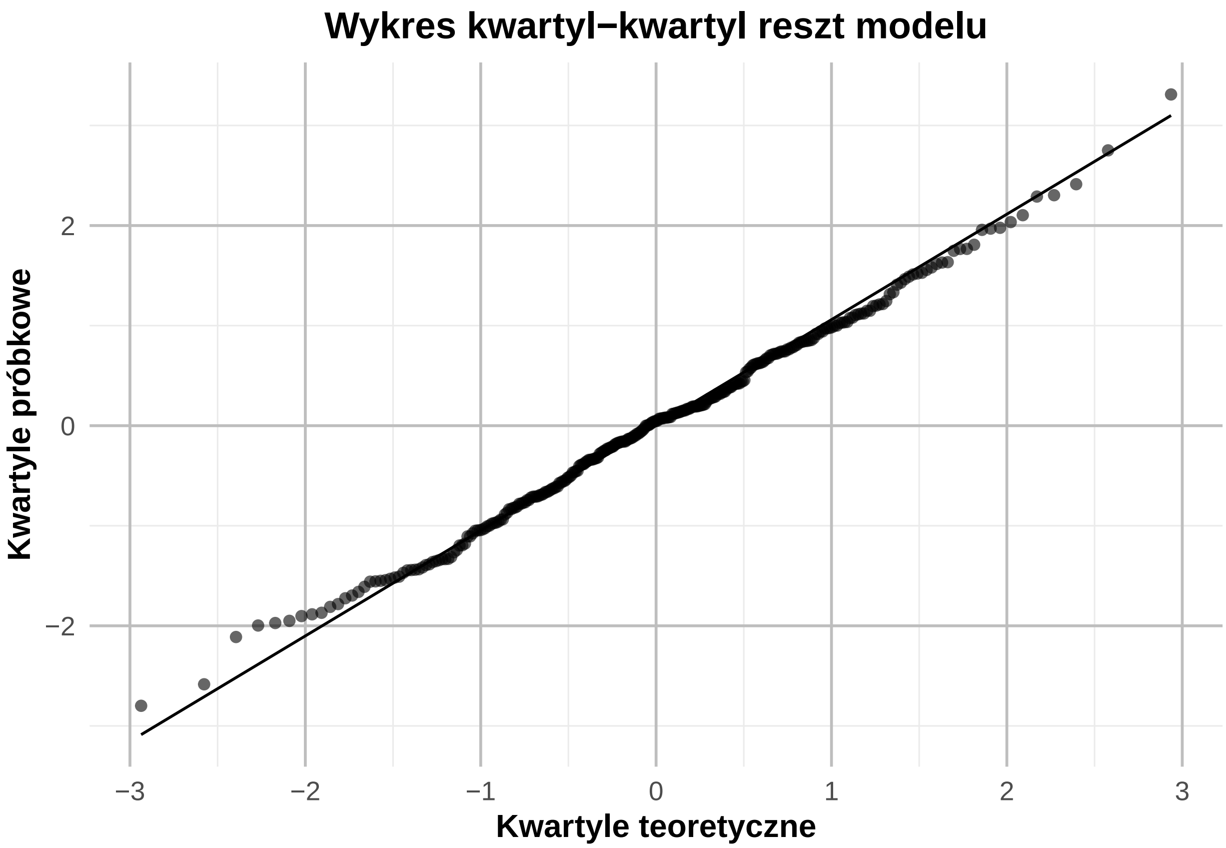
Test wykazał, że należy odrzucić hipotezę alternatywną oraz możemy przyjąć, że prawdziwa jest hipoteza zerowa mówiąca, że średnia reszt jest równa zero.

```
# Sprawdzenie rozkładu reszt
ggplot(model4, aes(x=.resid)) + geom_histogram(bins=30) +
  labs(title='Histogram reszt z modelu', x='Reszty', y='Częstość') + theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
```

```
)
```




```
ggplot(model4, aes(sample = .resid)) +
  geom_qq(size = 1.5, color = "black", alpha = 0.6) +
  geom_qq_line() +
  labs(title = 'Wykres kwartyl-kwartyl reszt modelu',
       x = 'Kwartyle teoretyczne',
       y = 'Kwartyle próbkowe') +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```




```
shapiro.test(model4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model4$residuals
## W = 0.99698, p-value = 0.8472
```

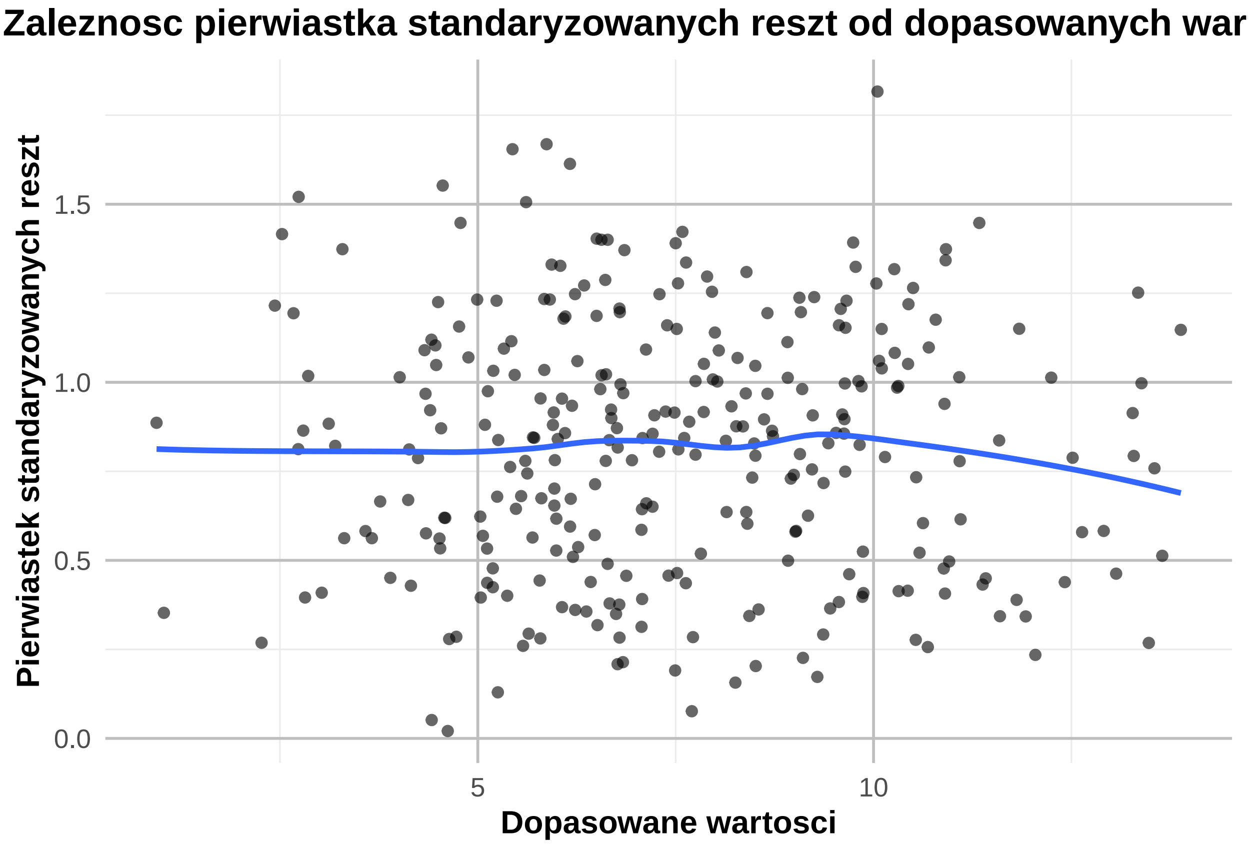
Na histogramie możemy zauważyć lekkie odchylenie reszt modelu, wykres Q-Q pokazuje jednak, że większość punktów skupia się na prostej. Test Shapiro-wilka wskazuje na brak podstaw do odrzucenia hipotezy zerowej mówiącej, że reszty modelu pochodzą z rodkładu normalnego. Więc możemy stwierdzić, że reszty są normalne.

```
# Sprawdzenie niezależności reszt
lmtest::dwtest(model4)
```

```
##
##  Durbin-Watson test
##
## data:  model4
## DW = 2.1322, p-value = 0.876
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# Homoskedastyczność
ggplot(model4, aes(.fitted, sqrt(abs(.stdresid)))) + geom_point(size = 1.5, color = "black", alpha = 0.
  labs(title='Zależność pierwiastka standaryzowanych reszt od dopasowanych wartości', x='Dopasowane war
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
```

```r
    panel.grid.major = element_line(color = "gray", size = 0.5)
  )
```



Punkty na wykresie są równomiernie rozproszone wokół linii, co sugeruje, że wariancja reszt nie zmienia się znacząco w miarę wzrostu wartości dopasowanych.

```r
lmtest::bptest(model4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model4
## BP = 4.4141, df = 7, p-value = 0.731
```

Wartość p-value wyniosłą znacznie więcej niż $\alpha = 0{,}05$, oznacza to, że nie mamy istotnych dowodów heteroskedastyczności. Dlatego też możemy wnioskować,że założenie o homoskedastyczności jest prawdziwe dla naszego modelu.

```r
model_summary(model4, carseats_test, carseats_test$Sales)
```

```
##
## =========================================
##             Podsumowanie modelu
## =========================================
##
## Metryki treningowe:
## -----------------------------------------
##   R-squared (R²):         0.8683
```

```
##   Adjusted R-squared:    0.8651
##   Kryterium informacyjne Akaikego (AIC): 874.44
## -----------------------------------------
##
## Charakterystyki "out-of-sample":
## ====================================================================
##   RMSE (trening):         1.0173    |    RMSE (test): 1.0332
##   MAE (trening):          0.7981    |    MAE (test):  0.8484
##   MAPE (trening):         15.50%    |    MAPE (test): Inf%
## ====================================================================
```

```
vif(model4)
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## CompPrice   1.423016  1        1.192902
## Income      1.023365  1        1.011615
## Advertising 1.012627  1        1.006294
## Price       1.406142  1        1.185809
## ShelveLoc   1.018134  2        1.004503
## Age         1.027325  1        1.013571
```

Model uproszczony wyjaśnia tyle samo zmienności co pełny model (R^2=0.85), ale ma niższą wartość AIC (874.44 vs. 879.11). RMSE wskazuje na lepszą predykcję.

Porównanie z regresją prostą

Przypominając wyniki z części pierwszej, model oparty na zmiennej **Price** wyglądał następująco:

```
price_model <- lm(Sales ~ Price, data = carseats_train)
```

```
model_summary(price_model, carseats_test, carseats_test$Sales)
```

```
##
## =========================================
##             Podsumowanie modelu
## =========================================
##
## Metryki treningowe:
## -----------------------------------------
##   R-squared (R²):         0.2061
##   Adjusted R-squared:    0.2034
##   Kryterium informacyjne Akaikego (AIC): 1403.16
## -----------------------------------------
##
## Charakterystyki "out-of-sample":
## ====================================================================
##   RMSE (trening):         2.4726    |    RMSE (test): 2.7079
##   MAE (trening):          2.0151    |    MAE (test):  2.1401
##   MAPE (trening):         50.52%    |    MAPE (test): Inf%
## ====================================================================
```

Model regresji wielorakiej znacząco przewyższa model prosty w zakresie dopasowania (R^2 dla modeli wielorakich wynosi od 0.60 do 0.85, podczas gdy dla modelu prostego tylko 0.35).