# SAT

František Kareš

2025-04-22

```
library(Sleuth2)
library(ggplot2)
library(patchwork)
library(dplyr)
library(olsrr)
library(performance)
library(car)
```

## Úkol č.1

```
head(case1201)
```

```
##         State  SAT Takers Income Years Public Expend Rank
## 1        Iowa 1088      3    326 16.79   87.8  25.60 89.7
## 2 SouthDakota 1075      2    264 16.07   86.2  19.95 90.6
## 3 NorthDakota 1068      3    317 16.57   88.3  20.62 89.8
## 4      Kansas 1045      5    338 16.30   83.9  27.14 86.3
## 5    Nebraska 1045      5    293 17.25   83.6  21.05 88.5
## 6     Montana 1033      8    263 15.91   93.7  29.48 86.4
```

### Popis dat

Data obsahují informace o průměrných výsledcích SAT testů ve všech amerických státech v roce 1982 a o faktorech, které mohou být s těmito výsledky spojeny.

- Datový rámec obsahuje **50 pozorování** a následujících **8 proměnných**:
  - **State**: Americký stát
  - **SAT**: Průměrné celkové skóre SAT testu
  - **Takers**: Procento všech způsobilých studentů (studenti posledních ročníků střední školy), kteří test absolvovali
  - **Income**: Median příjem rodin studentů, kteří test absolvovali (ve stovkách dolarů)
  - **Years**: Průměrný počet let formálního studia v oblasti společenských věd, přírodních věd a humanitních věd
  - **Public**: Procento studentů, kteří navštěvovali veřejné střední školy
  - **Expend**: Celkové výdaje státu na střední školy (ve stovkách dolarů na studenta)
  - **Rank**: Mediánové percentilové pořadí studentů v rámci jejich středních škol

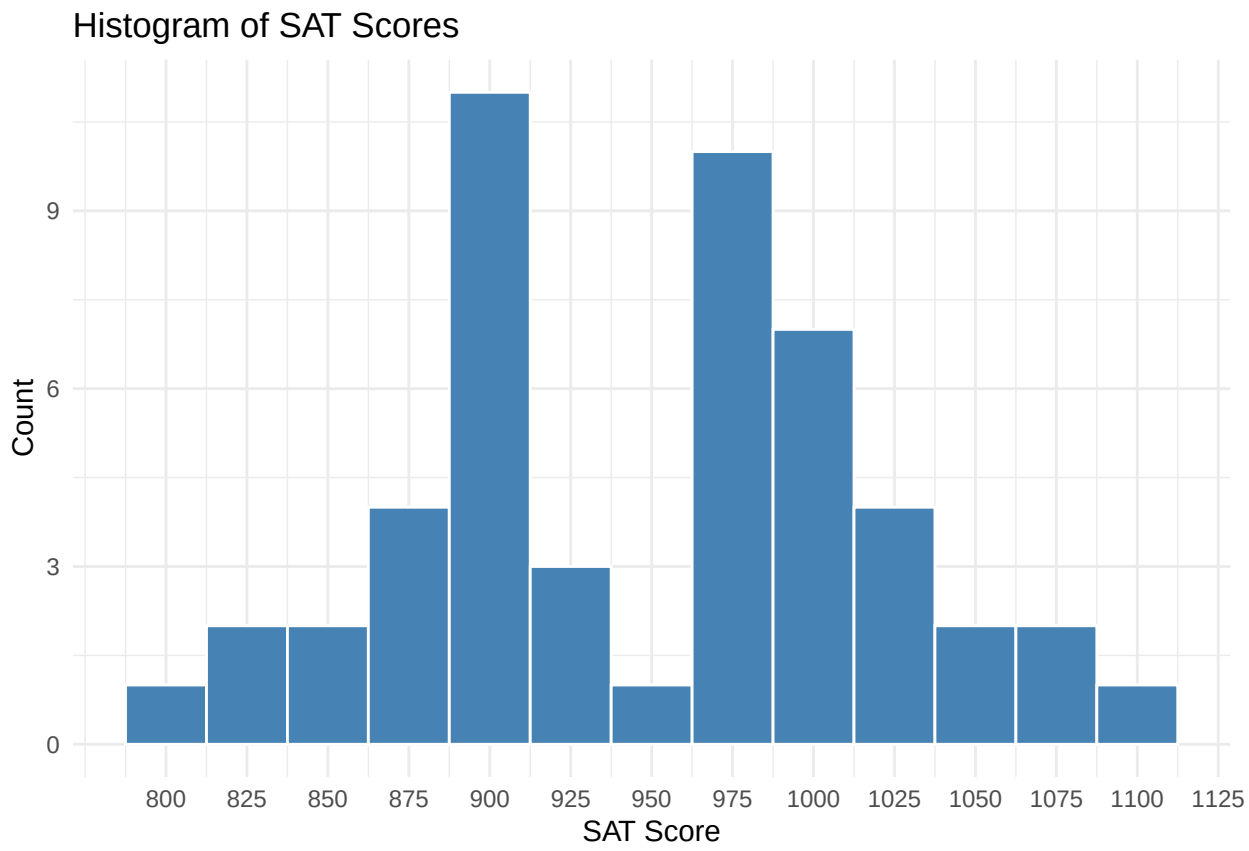
```r
summary(case1201)
```

```
##     State                SAT             Takers          Income
##  Length:50          Min.   : 790.0   Min.   : 2.00   Min.   :208.0
##  Class :character   1st Qu.: 889.2   1st Qu.: 6.25   1st Qu.:261.5
##  Mode  :character   Median : 966.0   Median :16.00   Median :295.0
##                     Mean   : 947.9   Mean   :26.22   Mean   :294.0
##                     3rd Qu.: 998.5   3rd Qu.:47.75   3rd Qu.:325.0
##                     Max.   :1088.0   Max.   :69.00   Max.   :401.0
##      Years            Public          Expend           Rank
##  Min.   :14.39   Min.   :44.80   Min.   :13.84   Min.   :69.80
##  1st Qu.:15.91   1st Qu.:76.92   1st Qu.:19.59   1st Qu.:74.03
##  Median :16.36   Median :80.80   Median :21.61   Median :80.85
##  Mean   :16.21   Mean   :81.20   Mean   :22.97   Mean   :79.99
##  3rd Qu.:16.76   3rd Qu.:88.25   3rd Qu.:26.39   3rd Qu.:85.83
##  Max.   :17.41   Max.   :97.00   Max.   :50.10   Max.   :90.60
```

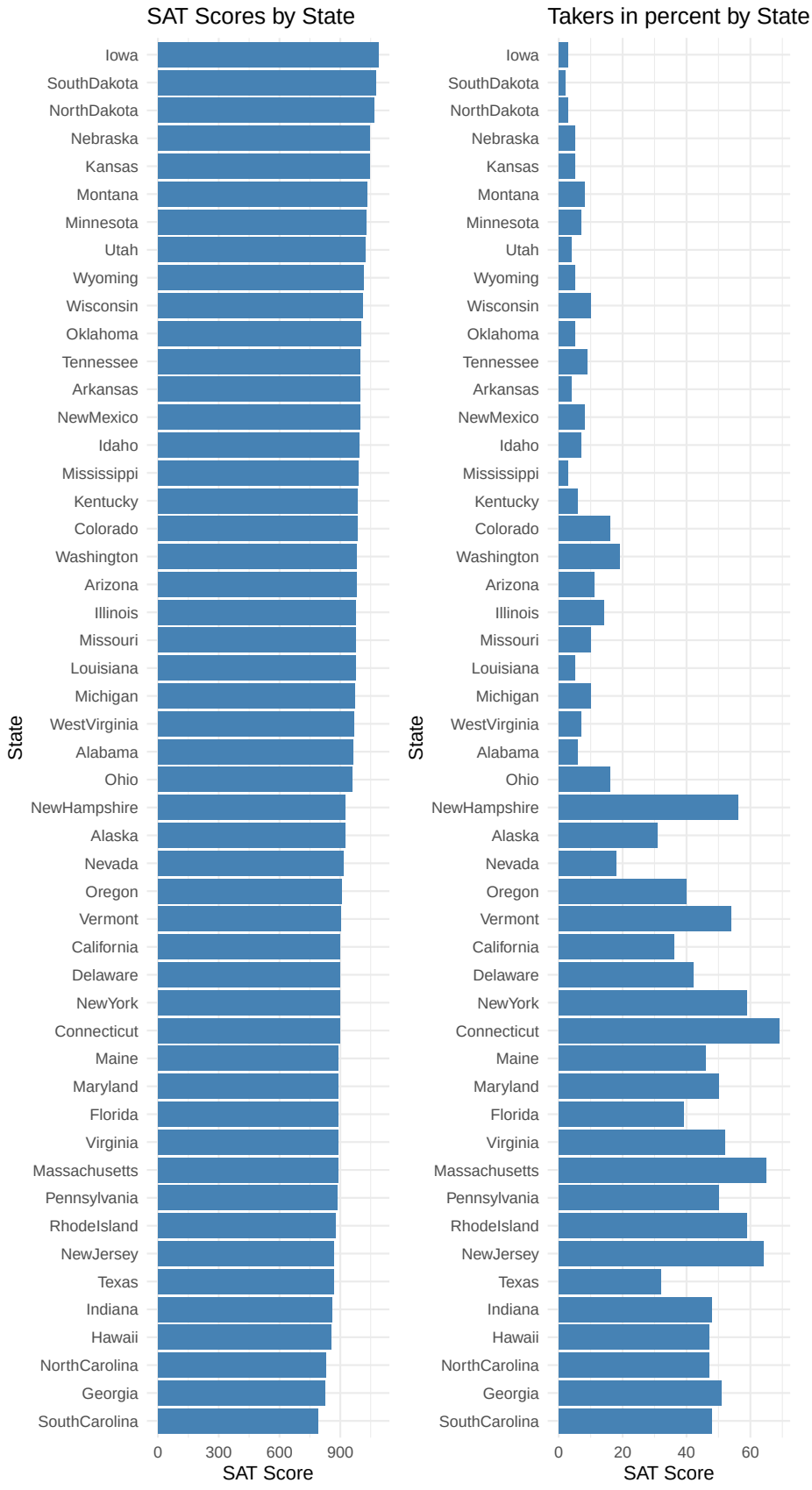
## Základní pozorování

- **State:** 50 států
- **SAT:** skóre se pohybuje v rozmezí 790-1088 s průměrem 947.9
- **Takers:** vysoké rozpětí absolvovaných studentů v procentech 2%-69% s průměrem 26.22% a mediánem 16%, což naznačuje asymetričnost směrem k vyšším hodnotám.
- **Income:** rozpětí 208$-401$ s průměrem 294$ (ve stovkách dolarů). Jedná se tedy přibližně o dvojnásobný rozdíl. Očištěno o inflaci, tak se jedná o rozpětí 689$-1328$.
- **Years:**
- **Public:** podíl studentů ve veřejných školách je mezi 44.8 % a 97 %, s průměrem 81.2 %. To značí silnou převahu veřejného školství v USA.
- **Expend:** výdaje na studenta středních škol kolísají od 13.84 do 50.10, průměrně 22.97. Toto vysoké rozpětí naznačuje výrazné rozdíly ve financování mezi státy.
- **Rank:** rozmezí 69.8-90.6 s průměrem 79.99 naznačuje relativně vyrovnanou úroveň studentů

```r
ggplot(case1201, aes(x = SAT)) +
  geom_histogram(binwidth = 25, fill = "steelblue", color = "white") +
  scale_x_continuous(breaks = seq(800, 1200, by = 25)) +
  labs(
    title = "Histogram of SAT Scores",
    x = "SAT Score",
    y = "Count"
  ) +
  theme_minimal()
```

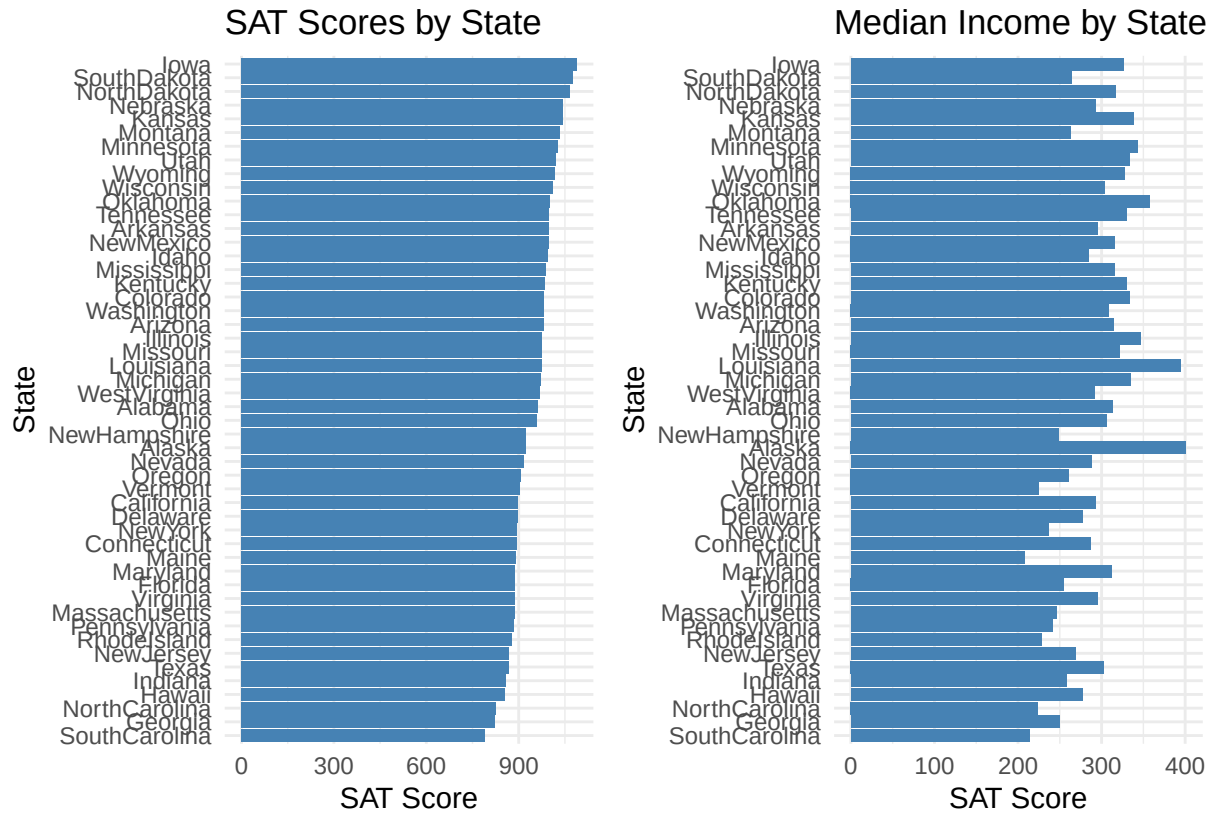



```r
p1 <- ggplot(case1201, aes(x = reorder(State, SAT), y = SAT)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +  # Flip axes for better readability
  labs(
    title = "SAT Scores by State",
    x = "State",
    y = "SAT Score"
  ) +
  theme_minimal()

p2 <- ggplot(case1201, aes(x = reorder(State, SAT), y = Takers)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +  # Flip axes for better readability
  labs(
    title = "Takers in percent by State",
    x = "State",
    y = "SAT Score"
  ) +
  theme_minimal()

p1 + p2
```


SAT Scores by State
Takers in percent by State
State
Iowa
SouthDakota
NorthDakota
Nebraska
Kansas
Montana
Minnesota
Utah
Wyoming
Wisconsin
Oklahoma
Tennessee
Arkansas
NewMexico
Idaho
Mississippi
Kentucky
Colorado
Washington
Arizona
Illinois
Missouri
Louisiana
Michigan
WestVirginia
Alabama
Ohio
NewHampshire
Alaska
Nevada
Oregon
Vermont
California
Delaware
NewYork
Connecticut
Maine
Maryland
Florida
Virginia
Massachusetts
Pennsylvania
RhodeIsland
NewJersey
Texas
Indiana
Hawaii
NorthCarolina
Georgia
SouthCarolina
0
300
600
900
SAT Score
0
20
40
60
SAT Score

můžeme si všimnout, že státy, kde je vyšší SAT skóre, tak mají malé procento lidí, kteří vůbec test podstoupili. Což naznačuje bias vůči tomu, že test jdou psát pouze lidé, kteří mají nějaké větší ambice, či celkově lepší školní výsledky.

```
p3 <- ggplot(case1201, aes(x = reorder(State, SAT), y = Income)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Median Income by State",
    x = "State",
    y = "SAT Score"
  ) +
  theme_minimal()

p1 + p3
```




z grafu se může zdát, že income nemá velký vliv na SAT skóre, kvůli žadnému patrnému trendu. Ale na tuto otázku bude schopni kvalifikovaněji odpovědět v následujícím úkolu.

# Úkol č.2

```
model <- lm(SAT ~ Income, data = case1201)
summary(model)
```

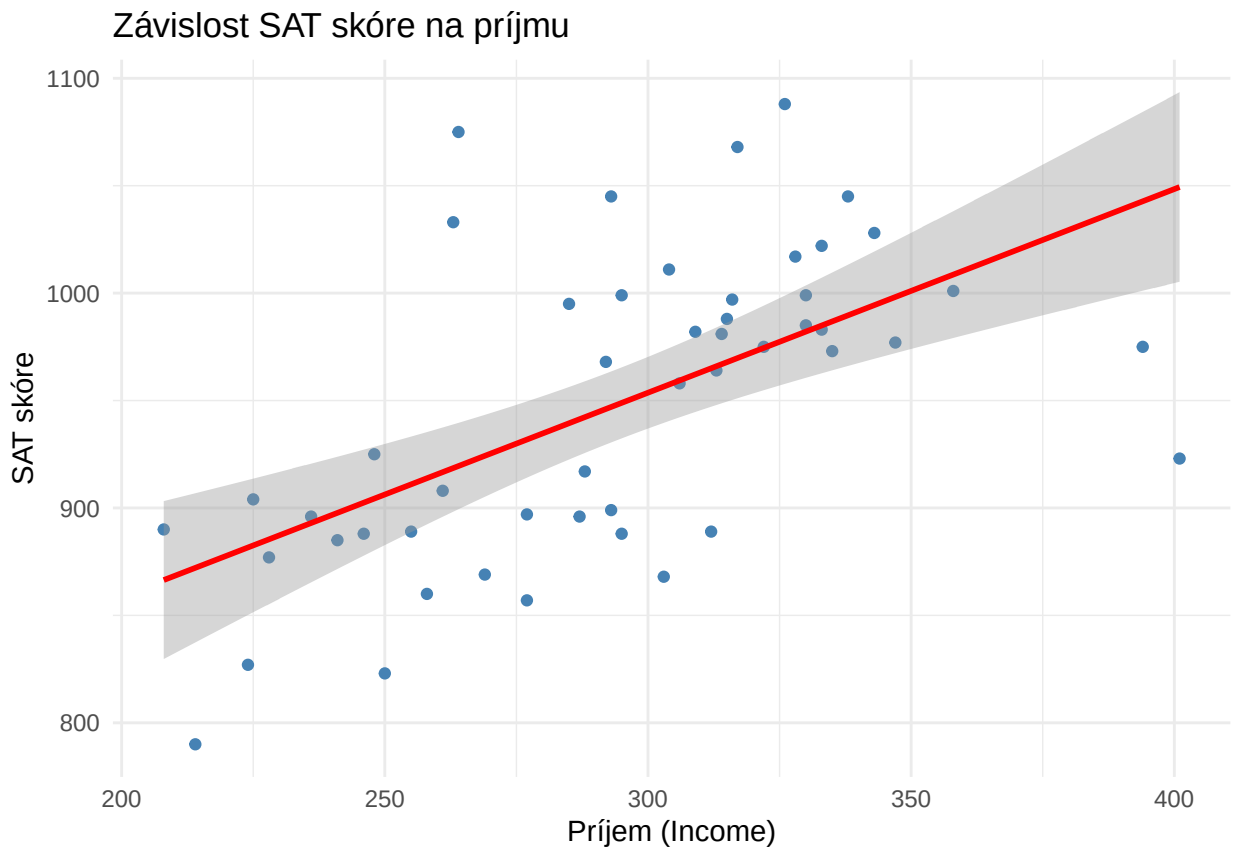
```
##
## Call:
## lm(formula = SAT ~ Income, data = case1201)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.376  -42.705   -1.628   27.030  155.476
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 669.2994    56.4364   11.86 7.15e-16 ***
## Income        0.9478     0.1899    4.99 8.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.09 on 48 degrees of freedom
## Multiple R-squared:  0.3416, Adjusted R-squared:  0.3279
## F-statistic:  24.9 on 1 and 48 DF,  p-value: 8.329e-06
```

**Regresní rovnice:**

$$\text{SAT} = 669.30 + 0.95 \times \text{Income}$$

Jelikož je regresní koeficient kladný, nejspíše existuje pozitivní korelace, která v našem modelu může být interpretována následovně: s každými dalšími sto dolary příjmu na rodinu by se SAT skóre mělo zvýšit přibližně o 0,95 bodu. Hodnota R-squared (0,3416) ukazuje, že příjem vysvětluje přibližně 34 % variability v SAT skórech. P-hodnota u F-statistiky je menší než 0,05, což znamená, že zamítáme nulovou hypotézu (že model pouze s interceptem je lepší). To potvrzuje, že zařazení proměnné Income do modelu významně zlepšuje vysvětlení výsledků SAT testů oproti modelu, který by pracoval jen s interceptem.

```r
ggplot(case1201, aes(x = Income, y = SAT)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Závislost SAT skóre na příjmu",
    x = "Příjem (Income)",
    y = "SAT skóre"
  ) +
  theme_minimal()
```




## Úkol č.3

```
case1201 <- case1201 %>%
  mutate(PublicCat = ifelse(Public > median(Public), "High", "Low")) %>%
  mutate(PublicCat = factor(PublicCat))


model_cat <- lm(SAT ~ PublicCat, data = case1201)
```

Vytvořil jsem kategorickou proměnnou, kde Low, je stát, kde procento lidí, kteří chodili do veřejných škol je nižší než median.

**Median: 80.8**

```
# Výpis výsledků
summary(model_cat)
```

```
##
## Call:
## lm(formula = SAT ~ PublicCat, data = case1201)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -161.04  -55.84   18.06   51.11  136.96
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    951.04      14.30  66.487   <2e-16 ***
## PublicCatLow    -6.20      20.23  -0.306    0.761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.52 on 48 degrees of freedom
## Multiple R-squared:  0.001953,   Adjusted R-squared:  -0.01884
## F-statistic: 0.09394 on 1 and 48 DF,  p-value: 0.7606
```

**Regresní rovnice:**

$$\text{SAT} = 951.04 - 6.20 \times \text{PublicCatLow}$$

**Intercept (951.04):** Tento odhad znamená, že pro státy, kde procento lidí, kteří chodili do veřejných škol, je vyšší než medián (tzn. kategorie High), je průměrné SAT skóre 951.04.

**Koeficient pro PublicCatLow (-6.20):** Tento koeficient ukazuje, že pro státy, kde je procento lidí, kteří chodili do veřejných škol, nižší než medián (kategorie Low), se průměrné SAT skóre snižuje o 6.20 bodů ve srovnání s kategorií High. Nicméně tento rozdíl není statisticky významný, protože p-hodnota (0.761) je výrazně vyšší než 0.05, což znamená, že rozdíl mezi těmito dvěma skupinami není statisticky relevantní.

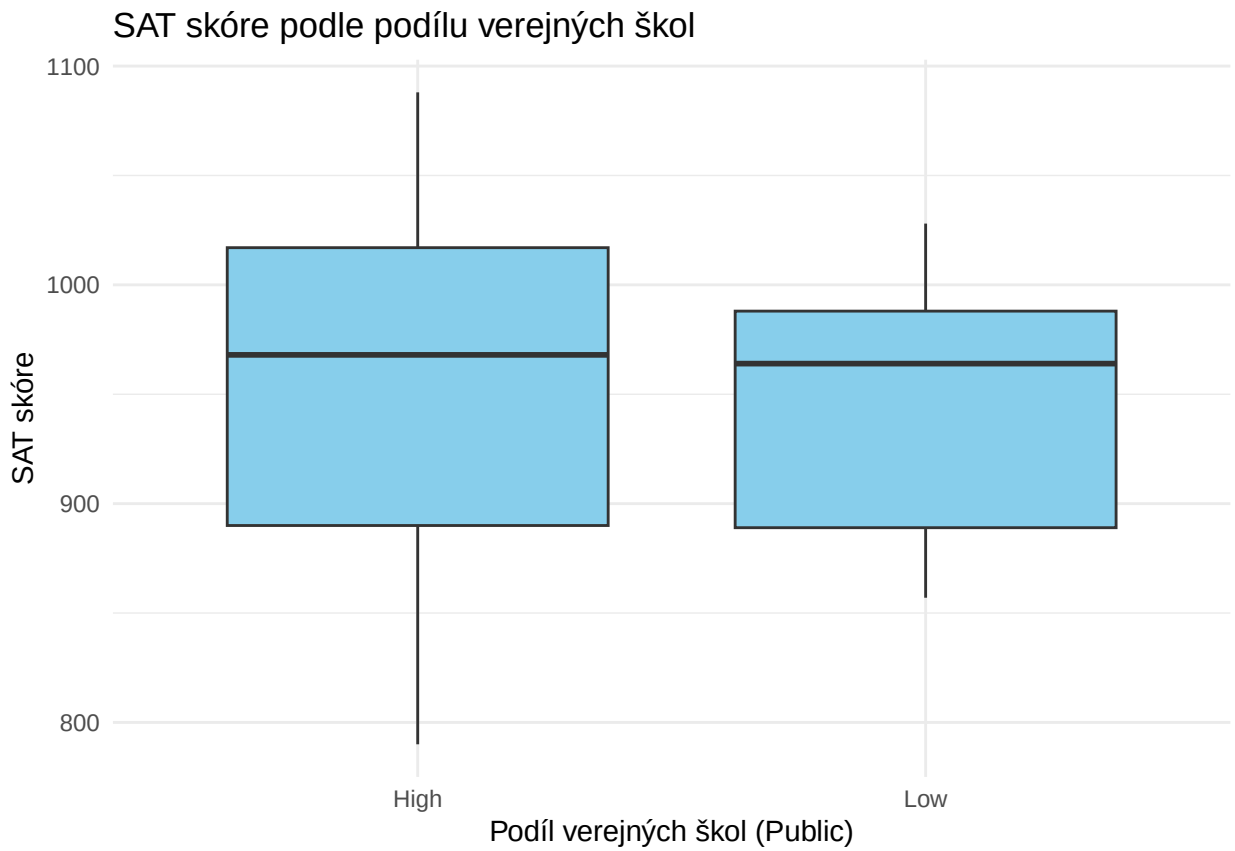
**R-squared (0.001953):** Tato hodnota naznačuje, že model vysvětluje pouze velmi malou část variability SAT skóre. V podstatě to znamená, že proměnná PublicCat (rozdělení států podle veřejných škol) téměř neovlivňuje výsledky SAT.

**F-statistika (0.09394) a p-hodnota (0.7606):** Ukazují, že model neprokázal významnou závislost mezi SAT a kategorií "veřejné školství" podle stanoveného mediánu.

```
levels(case1201$PublicCat)
```

```
## [1] "High" "Low"
```

```
ggplot(case1201, aes(x = PublicCat, y = SAT)) +
  geom_boxplot(fill = "skyblue") +
  labs(
    title = "SAT skóre podle podílu veřejných škol",
    x = "Podíl veřejných škol (Public)",
    y = "SAT skóre"
  ) +
  theme_minimal()
```




I z vizualizace je zřejmé, že se nejedná o velký rozdíl.

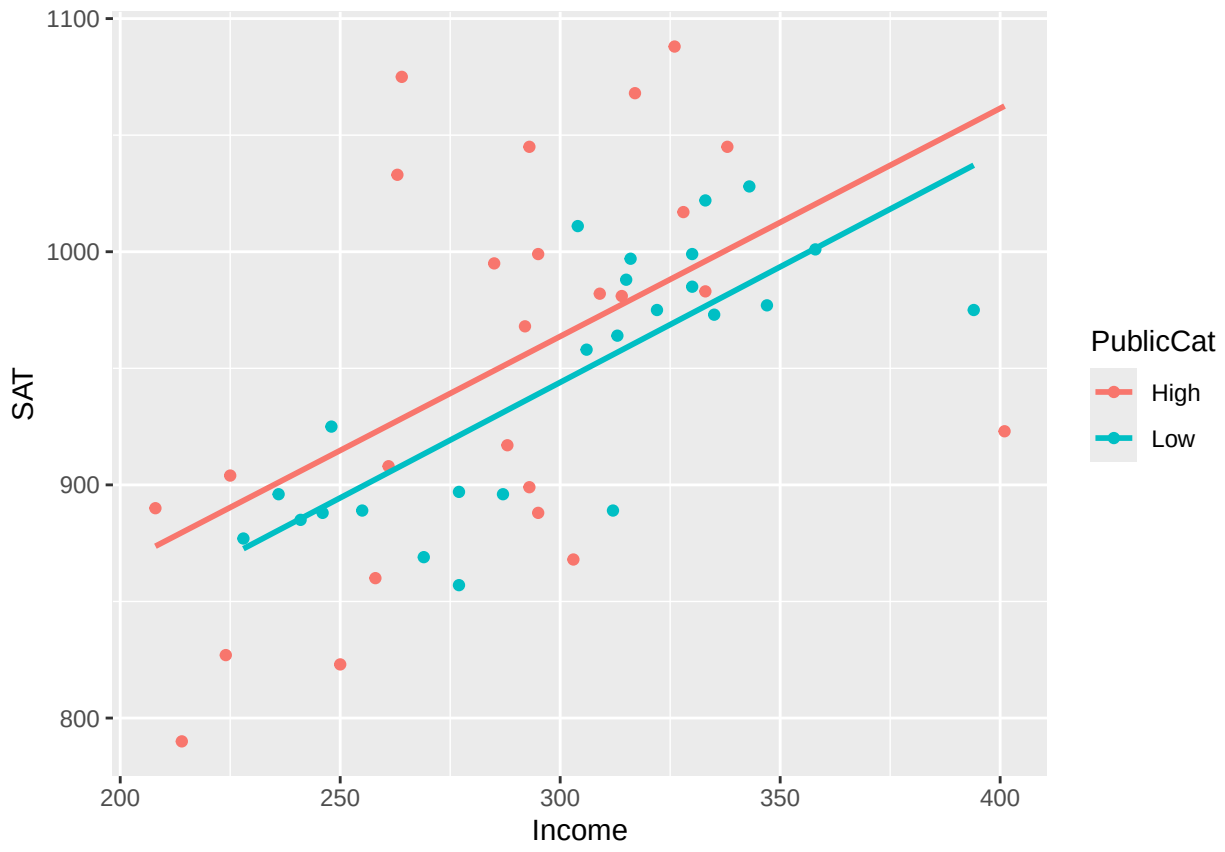
## Úkol č.4

```
model_expend <- lm(SAT ~ PublicCat*Income, data = case1201)
summary(model_expend)
```

```
##
## Call:
## lm(formula = SAT ~ PublicCat * Income, data = case1201)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -139.470  -35.050    5.761   31.709  146.535
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         670.23527   78.38065   8.551 4.62e-11 ***
## PublicCatLow        -23.51550  115.18490  -0.204 0.839133
## Income                0.97814    0.26997   3.623 0.000724 ***
## PublicCatLow:Income   0.01269    0.38739   0.033 0.974016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 58.46 on 46 degrees of freedom
## Multiple R-squared:  0.361,  Adjusted R-squared:  0.3193
## F-statistic: 8.662 on 3 and 46 DF,  p-value: 0.000115
```

```
ggplot(case1201, aes(x = Income, y = SAT, color = PublicCat)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```




### Statistická významnost:

- **Příjem (Income)** je **statisticky významným** prediktorem SAT skóre ($p = 0.0007$).
- **Kategorie veřejné školy (PublicCatLow)** sama o sobě **není významná** ($p = 0.839$).
- **Interakce (PublicCatLow:Income)** také **není významná** ($p = 0.974$).

**To znamená:**

- Efekt příjmu na SAT je podobný pro oba typy států (High i Low).
- Rozdíl mezi skupinami High a Low v závislosti na příjmu není podstatný.
- **Multiple R-squared = 0.361** → Model vysvětluje 36.1 % variability v SAT skóre.
- **Adjusted R-squared = 0.319** → Po zohlednění počtu prediktorů vysvětluje model 31.9 % variability.
- **F-statistika = 8.662, p-hodnota = 0.000115** → Model jako celek je statisticky významný.

# Úkol č.5

```r
fullmodel <- lm(SAT ~ PublicCat + Income + Takers*Rank, data = case1201)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = SAT ~ PublicCat + Income + Takers * Rank, data = case1201)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.046 -16.457   3.786  19.962  42.782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   49.13708  226.87732   0.217   0.8295
## PublicCatLow  13.46420   11.63072   1.158   0.2533
## Income         0.19434    0.14019   1.386   0.1727
## Takers        10.66149    4.14095   2.575   0.0135 *
## Rank          10.51688    2.37760   4.423  6.3e-05 ***
## Takers:Rank   -0.14549    0.05896  -2.468   0.0176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.58 on 44 degrees of freedom
## Multiple R-squared:  0.8435, Adjusted R-squared:  0.8257
## F-statistic: 47.43 on 5 and 44 DF,  p-value: < 2.2e-16
```

### Statistická významnost:

- **Income: není** statisticky významným prediktorem SAT skóre (p = 0.1727).
  (V modelu s příznakem Takers)
- **Kategorie PublicCatLow**: **není** statisticky významným prediktorem (p = 0.2533).
- **Takers je** statisticky významným prediktorem (p = 0.0135).
- **Rank je** statisticky významným prediktorem (p = 6.3e-05)
- Kombinace **Takers** a **Rank je** statisticky významným prediktorem ( p = 0.0176)

**To znamená:**

- Že **takers** a především **rank** má velký vliv na na SAT skóre
- **Multiple R-squared = 0.8435** → Model vysvětluje 84.4 % variability v SAT skóre.
- **Adjusted R-squared = 0.8257** → Po zohlednění počtu prediktorů vysvětluje model 82.6 % variability.
- **F-statistika = 47.43, p-hodnota = 2.2e-16** → Model jako celek je statisticky významný.

```r
reducedModel_without <- lm(SAT ~ 0 + Takers*Rank, data = case1201)
summary(reducedModel_without)
```

```
##
## Call:
## lm(formula = SAT ~ 0 + Takers * Rank, data = case1201)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.829 -13.027   6.878  21.451  51.634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## Takers      14.56934    3.71489   3.922 0.000285 ***
## Rank        11.94198    0.10479 113.963  < 2e-16 ***
## Takers:Rank -0.19826    0.05234  -3.788 0.000431 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31 on 47 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.9989
## F-statistic: 1.566e+04 on 3 and 47 DF,  p-value: < 2.2e-16
```

```r
anova(reducedModel_without,fullmodel)
```

```
## Analysis of Variance Table
##
## Model 1: SAT ~ 0 + Takers * Rank
## Model 2: SAT ~ PublicCat + Income + Takers * Rank
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     47 45161
## 2     44 38499  3    6662.7 2.5383 0.06874 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
reducedModel_with <- lm(SAT ~ Takers*Rank, data = case1201)
summary(reducedModel_with)
```

```
##
## Call:
## lm(formula = SAT ~ Takers * Rank, data = case1201)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.036 -15.781   8.505  17.908  54.225
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 336.45371  174.30985   1.930 0.059760 .
## Takers       12.63777    3.74771   3.372 0.001520 **
## Rank          8.07834    2.00427   4.031 0.000207 ***
## Takers:Rank  -0.18633    0.05126  -3.635 0.000699 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.14 on 46 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8191
## F-statistic: 74.96 on 3 and 46 DF,  p-value: < 2.2e-16
```

```
AIC(reducedModel_with)
```

```
## [1] 488.2986
```

```
AIC(reducedModel_without)
```

```
## [1] 490.1926
```

```
AIC(fullmodel)
```

```
## [1] 488.2116
```

Zde vidíme, že menší AIC skóre má **fullmodel** a **reducedModel_with**, což ukazuje, že při použití těchto modelů, bude pravděpodobně menší information loss (lepší fit na naše data).

```
reducedModel_with$coeff
```

```
## (Intercept)      Takers        Rank Takers:Rank
## 336.4537125  12.6377717   8.0783375  -0.1863255
```

**Regresní model:**

$$\text{SAT} = 336.54 + 12.64 \times \text{Takers} + 8.08 \times \text{Rank} - 0.19 \times (\text{Takers} \times \text{Ranks})$$

**Vysvětlení koeficientů:**

- **Intercept (336.54):** Při nulových příznacích, je očekávané SAT skóre 336.54, což ale v realitě nenastane, jelikož by Takers bylo 0% a Rank taktéž.
- **Koeficient u `Takers` (12,63):** S každým dalším zvýšením procenta Takers, by se podle našeho modelu mělo zvýšit o 14.57 bodů, za předpokladu, že ostatní proměnné zůstanou neměnné a nulové interakce mezi **Takers** a **Rank**
- **Koeficient u `Rank` (8,07):** S každým zvýšením procenta Rank, by se podle našeho modelu mělo zvýšit SAT skóre o 11.94 bodů, za předpokladu, že ostatní proměnné zůstanou neměnné a nulové interakce mezi **Takers** a **Rank**
- **Koeficient u interakční proměnné `Takers × Rank` (-0,19):** Tento záporný koeficient značí, že vliv `Takers` na SAT skóre **klesá**, pokud `Rank` roste a naopak. Čím více studentů skládá test ve státech s horším hodnocením (`Rank`), tím **menší přínos** to má pro výsledné SAT skóre. Každá jednotková změna v součinu `Takers × Rank` sníží SAT skóre o **0,19 bodu**.

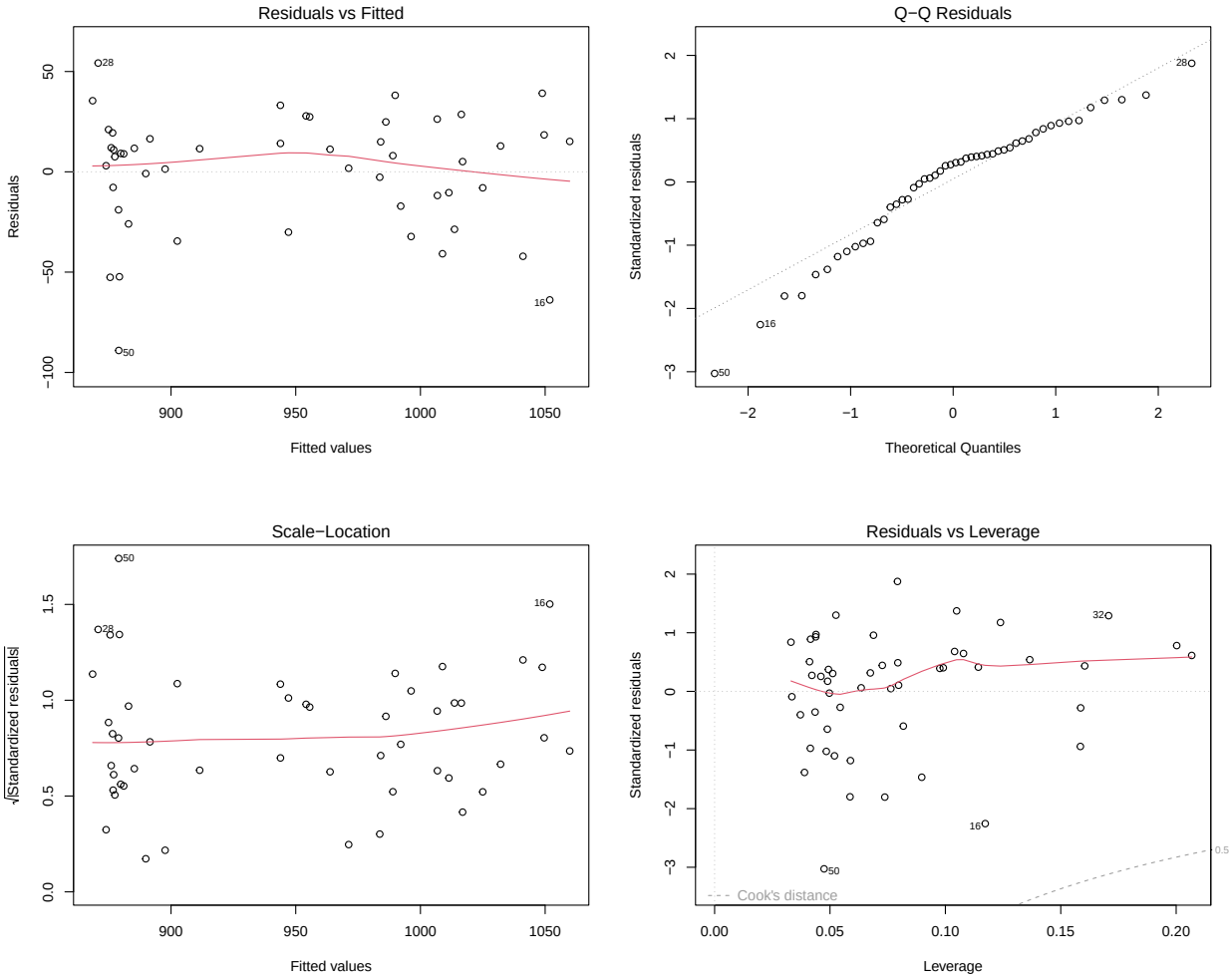
### Výběr modelu `SAT ~ Rank * Takers`

Model `SAT ~ Rank * Takers` jsem vybral na základě následujících důvodů:

- Jednodušší než **finalmodel** (méně regresorů), které jsou zároveň více důležité.
- Rozdíl v různých hodnotách mezi **fullmodel** a **reducemodel_with** nebyl veliký, ale rule of thumb je lepší jít s jednodušším modelem, kvůli jeho podobnému výkonu a menšímu vlivu šumu.
- Zvolil jsem si model s interceptem, jelikož měl nižší AIC a model bez interceptu se hůře interpretuje.
- **Vysoká predikční schopnost:** Adjusted $R^2$ je 0.82, což znamená, že model vysvětluje přibližně 82% variability v SAT skóre. To ukazuje na velmi dobrou shodu modelu s daty.
- **Významné prediktory:** Všechny koeficienty (`Rank`, `Takers`, i interakce `Rank:Takers`) jsou statisticky významné na standardní hladině významnosti (p-hodnoty $< 0.05$), což potvrzuje, že mají skutečný vliv na SAT skóre.
- **Interakční efekt:** Model zahrnuje interakční člen `Rank:Takers`, což znamená, že vliv počtu studentů (`Takers`) na SAT skóre závisí na hodnocení školy (`Rank`). Tento vztah je realistický a odráží komplexnější strukturu v datech.
- **Nízká reziduální chyba:** Reziduální standardní chyba je pouze 30.14 bodu SAT skóre, což je relativně nízké vzhledem k rozsahu SAT výsledků.
- **Logická interpretace:** Očekáváme, že jak hodnocení školy (Rank), tak počet studentů přistupujících ke zkoušce (Takers) ovlivní celkové SAT skóre, a jejich vzájemná interakce dává smysl. Například ve školách s vyšším Rankem může mít počet Takers jiný vliv než ve školách s nižším Rankem.

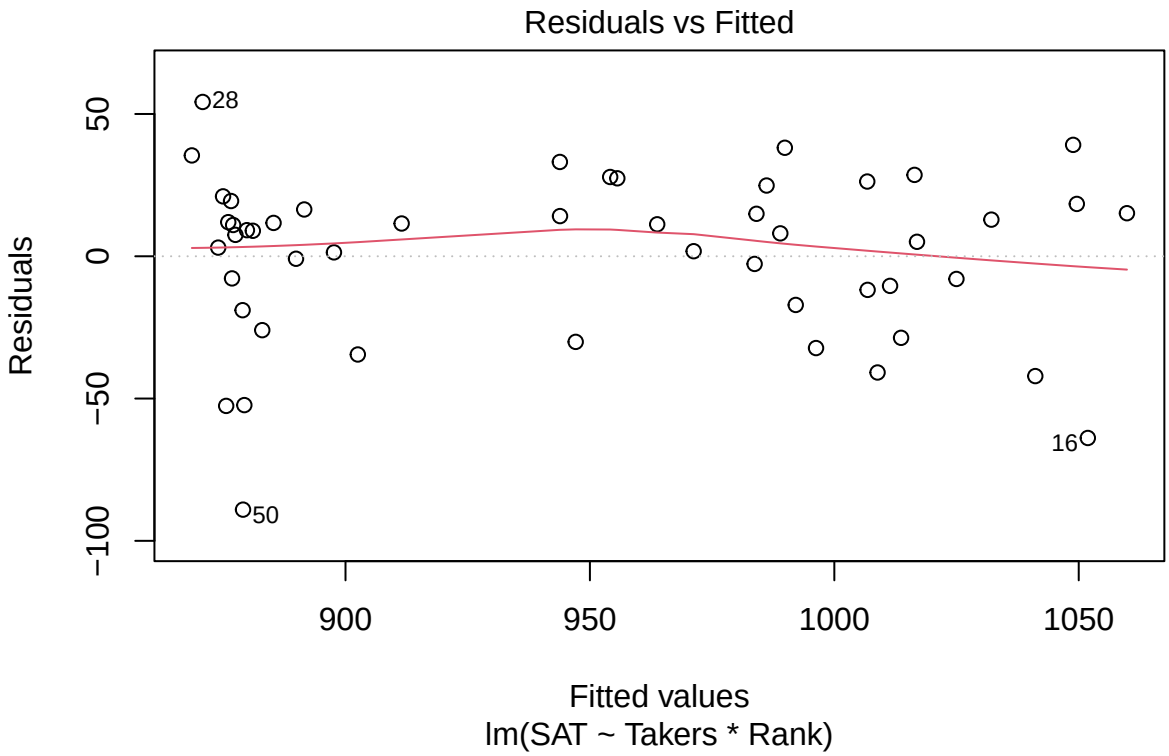
## Úkol č.6

```
par(mfrow = c(2, 2))
plot(reducedModel_with)
```




## Linearita

```
plot(reducedModel_with,1)
```



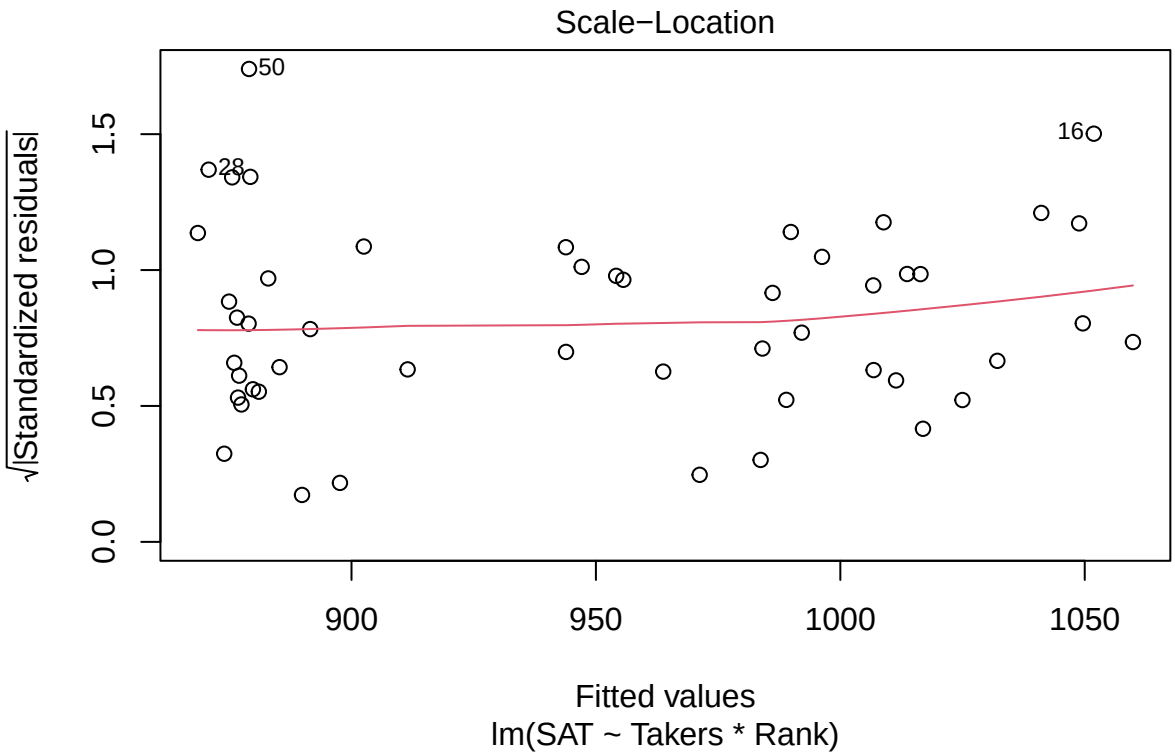

Residual plot nenaznačuje žádný zjevný pattern až na hodnoty mezi 925 až 975, kde rezidua jsou spíše kladná(menší zakřívení). Ideálnější by bylo, pokud by LOESS křivka byla horizontální na úrovni nuly.

### Homoskedasticita

```
ols_test_breusch_pagan(reducedModel_with)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  -------------------------------
##  Response : SAT
##  Variables: fitted values of SAT
##
##         Test Summary
##  ---------------------------
##  DF            =    1
##  Chi2          =    0.225447
##  Prob > Chi2   =    0.6349206
```

```
plot(reducedModel_with,3)
```



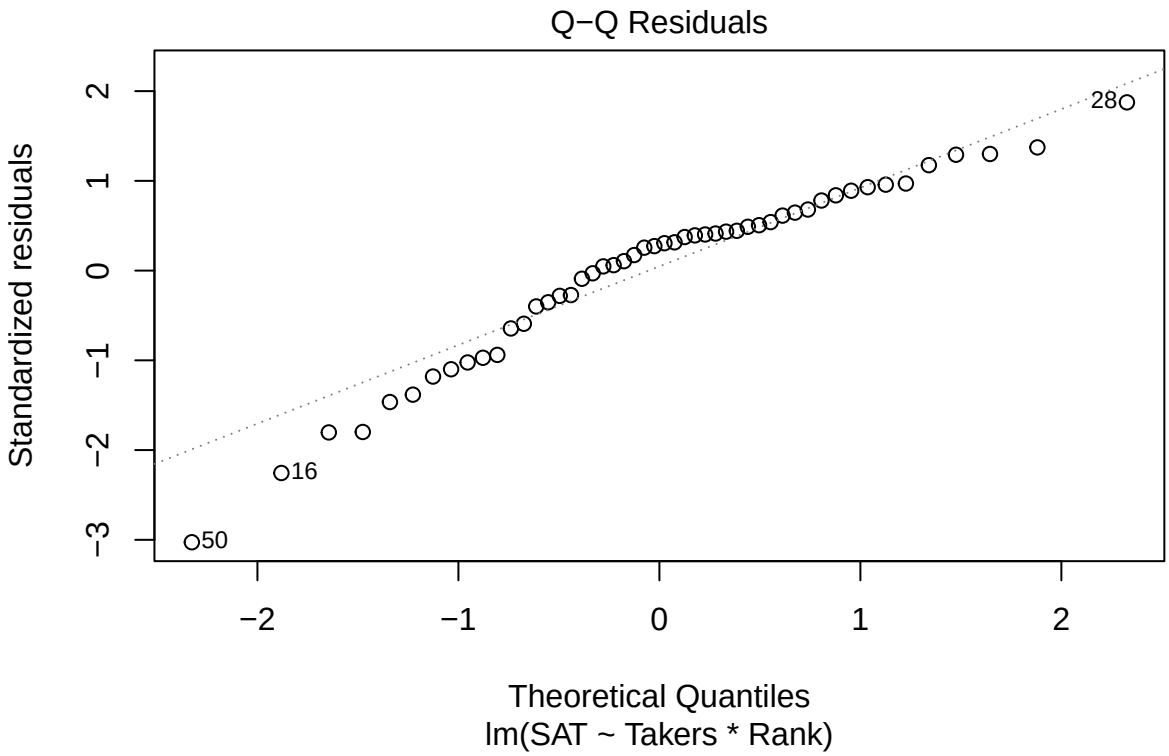

Breusch-Pagan test s **p-hodnotou**(>0.05) **nezamítáme** nulovou hypotézu, tudíž můžeme předpokládat homoskedasticitu reziduí. I z grafu je patrný stejný rozptyl reziduí napříč hodnotami.

### Normalita reziduí

```
ols_test_normality(reducedModel_with)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9434         0.0184
## Kolmogorov-Smirnov        0.1416         0.2442
## Cramer-von Mises          4.8867         0.0000
## Anderson-Darling          1.0638         0.0078
## -----------------------------------------------
```

```
plot(reducedModel_with,2)
```




Na základě Shapiro-Wilk normality testu s **p-hodnotou**(<0.05), zamítáme nulovou hypotézu, tudíž rezidua **nejsou** normálně rozdělena. (3/4 z testů na normalitu vyšly s p-hodnotou menší než $< 0.05$)

```
wilcox.test(reducedModel_with$residuals)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  reducedModel_with$residuals
## V = 707, p-value = 0.5054
## alternative hypothesis: true location is not equal to 0
```

nulovou hypotézu **nezamítáme** (p-hodnota $> 0.05$), tudíž můžeme předpokládat, že střední hodnota reziduí je rovna 0.

## Korelace

tento test lze provést pouze za podmínky normality, kterou jsme v předešlých testech zamítli.

```
durbinWatsonTest(reducedModel_with)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.421767      0.930035       0
##  Alternative hypothesis: rho != 0
```

```
ols_test_correlation(reducedModel_with)
```

```
## [1] 0.9699608
```

```
check_model(reducedModel_with)
```

```
ols_vif_tol(reducedModel_with)
```

```
##     Variables   Tolerance        VIF
## 1      Takers 0.002720975 367.515324
## 2        Rank 0.108436328   9.222002
## 3 Takers:Rank 0.002915051 343.047181
```

Je jasné že v modelu je vysoká korelace, jelikož jsme si ji tam sami zavedli s příznakem **Takers:Rank.** VIF skóre u každého regresoru je vyšší než 5, což naznačuje vysoko multikolinearitu.

# Poznámka na závěr

při evaluaci dvou modelů, kde jediný rozdíl byl v přítomnosti interceptu, jsem zjistil, že nějaké hodnoty mají jinou, či obtížnou interpretaci, a že intercept hraje důležitou roli.

Například R-squared:

bez interceptu

$$R_0^2 = \frac{\sum_i \hat{y}_i}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2}$$

s interceptem:

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

tudíž je to hůře porovnatelnémezi sebou.