Final Project STATS-101A

Mary McSweeney, Lior Levy Meruk, Henry Johnson, Ainsley Strang, and Evan Sang

2024-02-01

Change Data

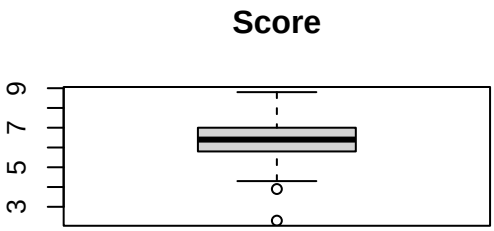
```
movies <- read.csv("movies.csv")
attach(movies)
movies <- movies[!is.na(score) & !is.na(budget) & !is.na(gross) & !is.na(runtime), ]

## Transform budget and gross to millions of dollars
movies$budget <- movies$budget / 1e9
movies$gross <- movies$gross / 1e9
```

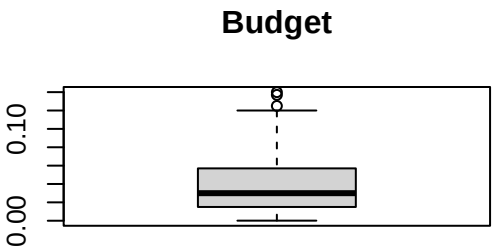
Summaries

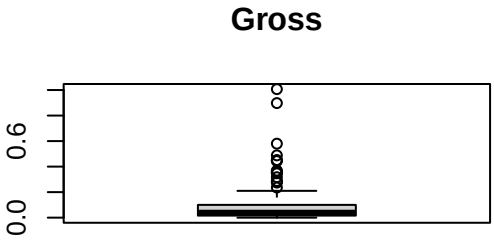
```
attach(movies)
```

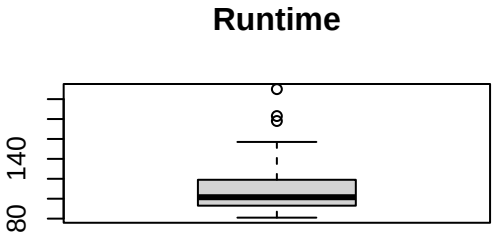
```
## The following objects are masked from movies (pos = 3):
##
##     budget, gross, runtime, score
```

```
par(mfrow = c(2, 2))
boxplot(score, main = "Score")
boxplot(budget, main = "Budget")
boxplot(gross, main = "Gross")
boxplot(runtime, main = "Runtime")
```










```
sapply(movies, mean)
```

```
##        score       budget        gross      runtime
##   6.34397590   0.03883313   0.09289490 107.25301205
```

```
sapply(movies, sd)
```

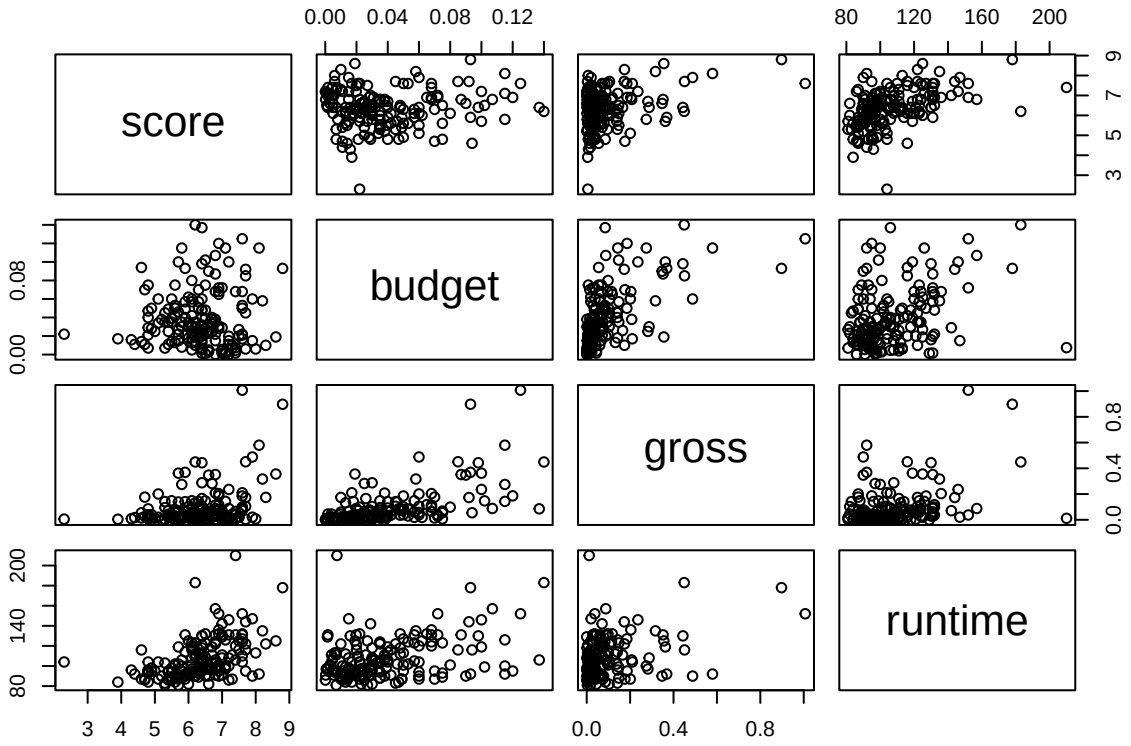
```
##       score      budget       gross     runtime
##  0.96655896  0.03143674  0.14427589 20.33587269
```

```
sapply(movies, summary)
```

```
##            score     budget       gross runtime
## Min.    2.300000 0.00010000 0.000080631  81.000
## 1st Qu. 5.800000 0.01500000 0.016158239  93.000
## Median  6.400000 0.03000000 0.038577035 101.500
## Mean    6.343976 0.03883313 0.092894899 107.253
## 3rd Qu. 6.975000 0.05600000 0.099609142 119.000
## Max.    8.800000 0.14000000 1.006968171 210.000
```

First Model

```
pairs(movies)
```




```
movies_model <- lm(score ~ budget + gross + runtime)
summary(movies_model)
```

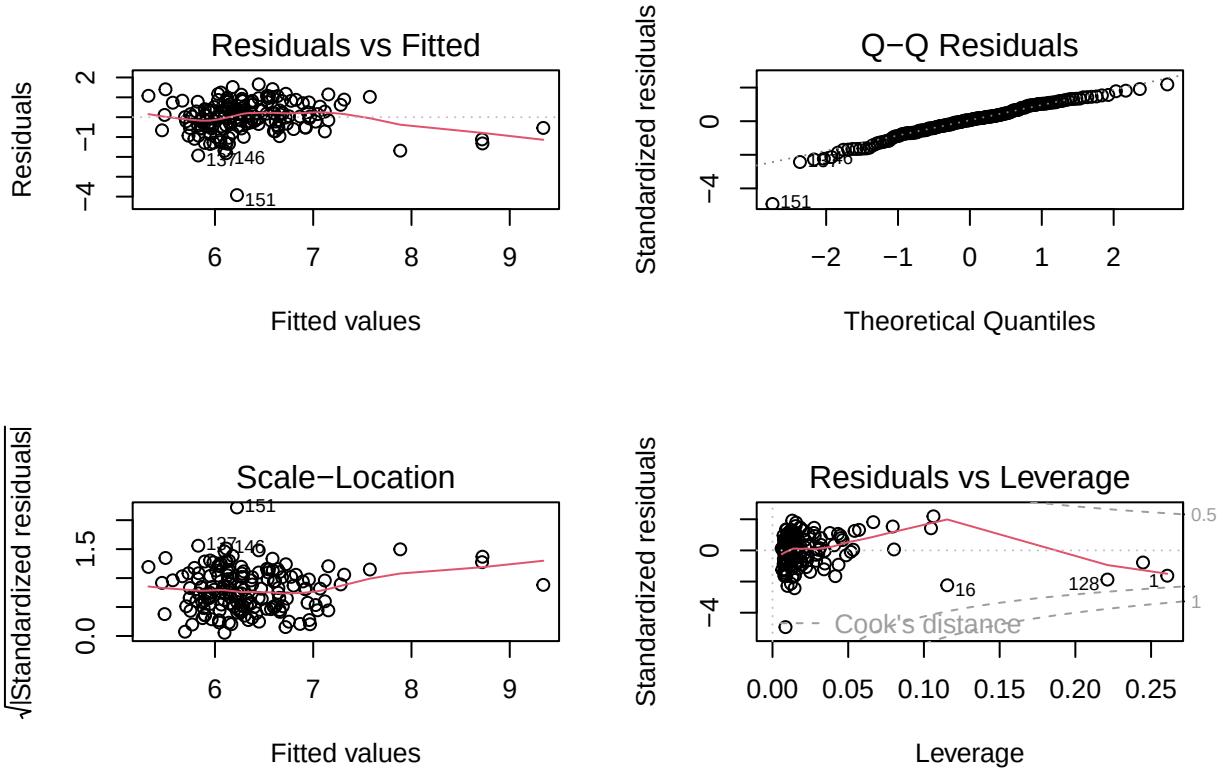
```
##
## Call:
## lm(formula = score ~ budget + gross + runtime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9243 -0.4411  0.0526  0.5306  1.6546
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.134388   0.344458  12.003  < 2e-16 ***
## budget      -9.920431   2.528044  -3.924 0.000128 ***
## gross        2.451998   0.552143   4.441 1.65e-05 ***
## runtime      0.022070   0.003335   6.617 5.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8001 on 162 degrees of freedom
## Multiple R-squared:  0.3272, Adjusted R-squared:  0.3147
## F-statistic: 26.26 on 3 and 162 DF,  p-value: 6.771e-14
```

```
anova(movies_model)
```

```
## Analysis of Variance Table
##
## Response: score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## budget      1   0.584  0.5838  0.9118    0.3411
```

```
## gross       1  21.823 21.8233 34.0882 2.798e-08 ***
## runtime     1  28.029 28.0291 43.7817 5.105e-10 ***
## Residuals 162 103.713  0.6402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



Transformed Model

```
library(car)
```

```
## Loading required package: carData
```

```
summary(powerTransform(cbind(score, budget, gross, runtime) ~ 1))
```

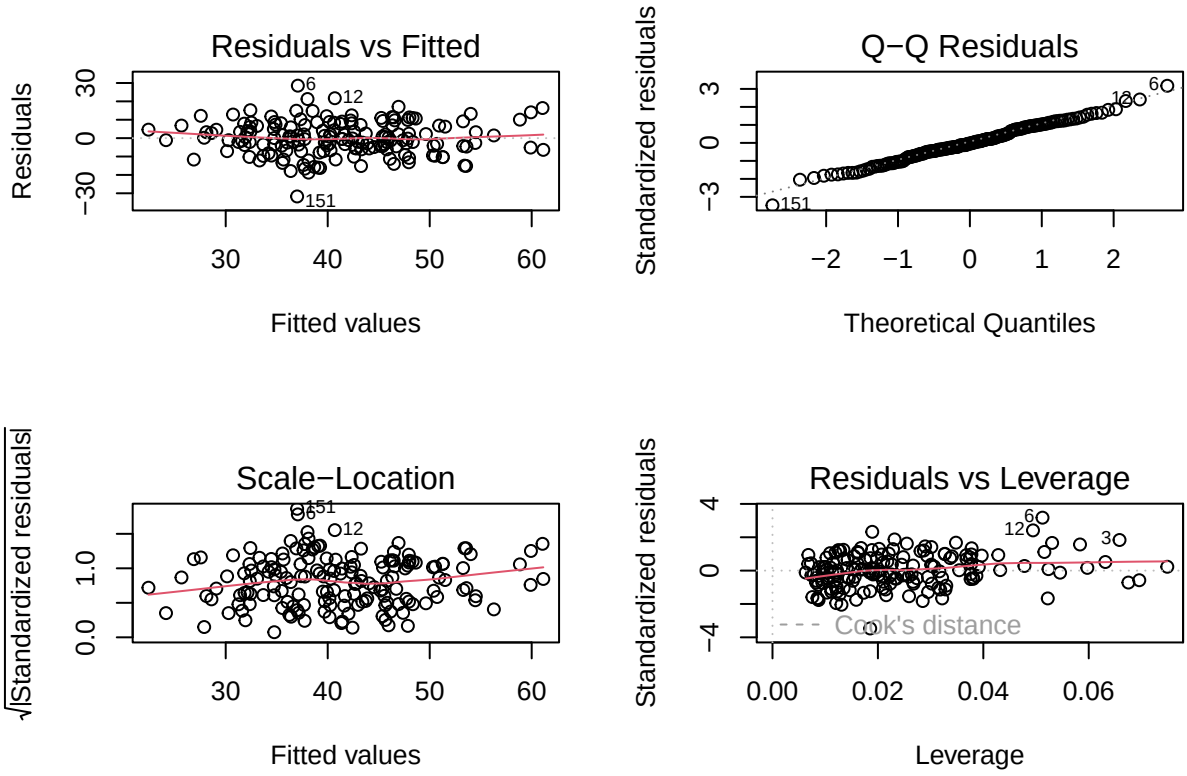
```
## bcPower Transformations to Multinormality
##          Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## score       2.0112        2.00       1.3745       2.6478
## budget      0.3495        0.33       0.2459       0.4530
## gross       0.1690        0.17       0.0997       0.2382
## runtime    -1.8979       -2.00      -2.6836      -1.1123
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 151.8116  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
```

```
##                                  LRT df       pval
## LR test, lambda = (1 1 1 1) 507.9174  4 < 2.22e-16
```

```
t_score <- score^2
t_budget <- budget^.33
t_gross <- gross^0.17
t_runtime <- runtime^(-2)
t_movies_model <- lm(t_score ~ t_budget + t_gross + t_runtime)
summary(t_movies_model)
```

```
##
## Call:
## lm(formula = t_score ~ t_budget + t_gross + t_runtime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.713  -5.575  -0.319   6.920  28.538
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.999e+01  4.552e+00  13.180  < 2e-16 ***
## t_budget    -6.105e+01  1.011e+01  -6.037 1.03e-08 ***
## t_gross      3.669e+01  6.746e+00   5.438 1.95e-07 ***
## t_runtime   -2.239e+05  2.660e+04  -8.419 1.92e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.228 on 162 degrees of freedom
## Multiple R-squared:  0.4152, Adjusted R-squared:  0.4043
## F-statistic: 38.33 on 3 and 162 DF,  p-value: < 2.2e-16
```

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




```
anova(t_movies_model)
```

```
## Analysis of Variance Table
##
## Response: t_score
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## t_budget    1    30.1    30.1  0.3534     0.553
## t_gross     1  3728.1  3728.1 43.7754 5.118e-10 ***
## t_runtime   1  6035.8  6035.8 70.8735 1.919e-14 ***
## Residuals 162 13796.4    85.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable Selection

```
library(leaps)

vif(t_movies_model)
```

```
##  t_budget   t_gross t_runtime
##  1.938831  1.940866  1.112464
```

```
X <- cbind(t_budget, t_gross, t_runtime)
b <- regsubsets(as.matrix(X), t_score)
summary(b)
```

```
## Subset selection object
## 3 Variables  (and intercept)
##           Forced in Forced out
## t_budget      FALSE      FALSE
```

```
## t_gross       FALSE      FALSE
## t_runtime     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          t_budget t_gross t_runtime
## 1  ( 1 ) " "      " "     "*"
## 2  ( 1 ) "*"      " "     "*"
## 3  ( 1 ) "*"      "*"     "*"
```

Optimal Models: 1 predictor: score^(2) ~ runtime^(-2) 2 predictors: score^(2) ~ budget^(0.33) + runtime^(-2) 3 predictors: score^(2) ~ budget^(0.33) + gross^(0.17) + runtime^(-2)

```
om1 <- lm(t_score ~ t_runtime)
om2 <- lm(t_score ~ t_budget + t_runtime)
om3 <- lm(t_score ~ t_budget + t_gross + t_runtime)
Radj2_vect <- summary(b)$adjr2
om_list <- list(om1, om2, om3)
AIC_vect <- numeric(3)
AICc_vect <- numeric(3)
BIC_vect <- numeric(3)

for (i in seq_len(3)) {
  AIC_vect[i] <- extractAIC(om_list[[i]])[2]
}

for (i in seq_len(3)) {
  AICc_vect[i] <- extractAIC(om_list[[i]])[2] + 2 * (i + 2) * (i + 3) / (nrow(movies) - i - 1)
}

for (i in seq_len(3)) {
  BIC_vect[i] <- extractAIC(om_list[[i]], k = log(nrow(movies)))[2]
}

data.frame("Size" = 1:3, "Radj2" = Radj2_vect, "AIC" = AIC_vect, "AICc" = AICc_vect, "BIC" = BIC_vect)
```

```
##   Size     Radj2      AIC     AICc      BIC
## 1    1 0.2656625 774.5293 774.6756 780.7533
## 2    2 0.2999124 767.5853 767.8307 776.9213
## 3    3 0.4043372 741.7497 742.1200 754.1976
```

R^2, AIC, AICc, and BIC all suggest the model with 3 predictors.

```
backAIC <- step(t_movies_model, direction = "backward", data = movies)
```

```
## Start:  AIC=741.75
## t_score ~ t_budget + t_gross + t_runtime
##
##             Df Sum of Sq   RSS    AIC
## <none>                   13796 741.75
## - t_gross    1    2518.7 16315 767.59
## - t_budget   1    3104.1 16901 773.44
## - t_runtime  1    6035.8 19832 799.99
```

Back AIC suggests the model with 3 predictors.

```
backBIC <- step(t_movies_model, direction = "backward", data = movies,
                k = log(nrow(movies)))
```

```
## Start:  AIC=754.2
## t_score ~ t_budget + t_gross + t_runtime
##
##             Df Sum of Sq   RSS    AIC
## <none>                   13796 754.20
## - t_gross    1    2518.7 16315 776.92
## - t_budget   1    3104.1 16901 782.77
## - t_runtime  1    6035.8 19832 809.33
```

Back BIC suggests the model with 3 predictors.

```
mint <- lm(score ~ 1, data = movies)
forwardAIC <- step(mint, scope = list(lower = ~ 1, upper = ~ t_budget + t_gross + t_runtime),
                   direction = "forward", data = movies)
```

```
## Start:  AIC=-10.3
## score ~ 1
##
##             Df Sum of Sq    RSS     AIC
## + t_runtime  1    39.422 114.73 -57.325
## + t_gross    1     9.911 144.24 -19.326
## <none>                   154.15 -10.295
## + t_budget   1     0.174 153.97  -8.483
##
## Step:  AIC=-57.33
## score ~ t_runtime
##
##            Df Sum of Sq    RSS     AIC
## + t_budget  1    5.5191 109.21 -63.509
## + t_gross   1    1.8720 112.86 -58.056
## <none>                  114.73 -57.325
##
## Step:  AIC=-63.51
## score ~ t_runtime + t_budget
##
##           Df Sum of Sq    RSS     AIC
## + t_gross  1    15.128  94.08 -86.262
## <none>                 109.21 -63.509
##
## Step:  AIC=-86.26
## score ~ t_runtime + t_budget + t_gross
```

Forward AIC suggests the model with 3 predictors

```
forwardBIC <- step(mint, scope = list(lower = ~ 1, upper = ~ t_budget + t_gross + t_runtime),
                   direction = "forward", data = movies, k = log(nrow(movies)))
```

```
## Start:  AIC=-7.18
## score ~ 1
##
##             Df Sum of Sq    RSS     AIC
## + t_runtime  1    39.422 114.73 -51.101
## + t_gross    1     9.911 144.24 -13.102
## <none>                   154.15  -7.183
## + t_budget   1     0.174 153.97  -2.259
##
```

```
## Step:  AIC=-51.1
## score ~ t_runtime
##
##            Df Sum of Sq    RSS     AIC
## + t_budget  1    5.5191 109.21 -54.173
## <none>                  114.73 -51.101
## + t_gross   1    1.8720 112.86 -48.720
##
## Step:  AIC=-54.17
## score ~ t_runtime + t_budget
##
##           Df Sum of Sq    RSS     AIC
## + t_gross  1    15.128  94.08 -73.814
## <none>                 109.21 -54.173
##
## Step:  AIC=-73.81
## score ~ t_runtime + t_budget + t_gross
```

Forward BIC suggests the model with 3 predictors.

Final Model

```
t_movies_model
```

```
##
## Call:
## lm(formula = t_score ~ t_budget + t_gross + t_runtime)
##
## Coefficients:
## (Intercept)     t_budget      t_gross    t_runtime
##       59.99       -61.05        36.69   -223906.15
```

```
summary(t_movies_model)
```

```
##
## Call:
## lm(formula = t_score ~ t_budget + t_gross + t_runtime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.713  -5.575  -0.319   6.920  28.538
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.999e+01  4.552e+00  13.180  < 2e-16 ***
## t_budget    -6.105e+01  1.011e+01  -6.037 1.03e-08 ***
## t_gross      3.669e+01  6.746e+00   5.438 1.95e-07 ***
## t_runtime   -2.239e+05  2.660e+04  -8.419 1.92e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.228 on 162 degrees of freedom
## Multiple R-squared:  0.4152, Adjusted R-squared:  0.4043
## F-statistic: 38.33 on 3 and 162 DF,  p-value: < 2.2e-16
```