The Unknowns behind Movie Data

Group 5

Data Collection

- The Movie Database (TMDb)
- 5000 Movies
- 8 Variables

Hypothesis 1

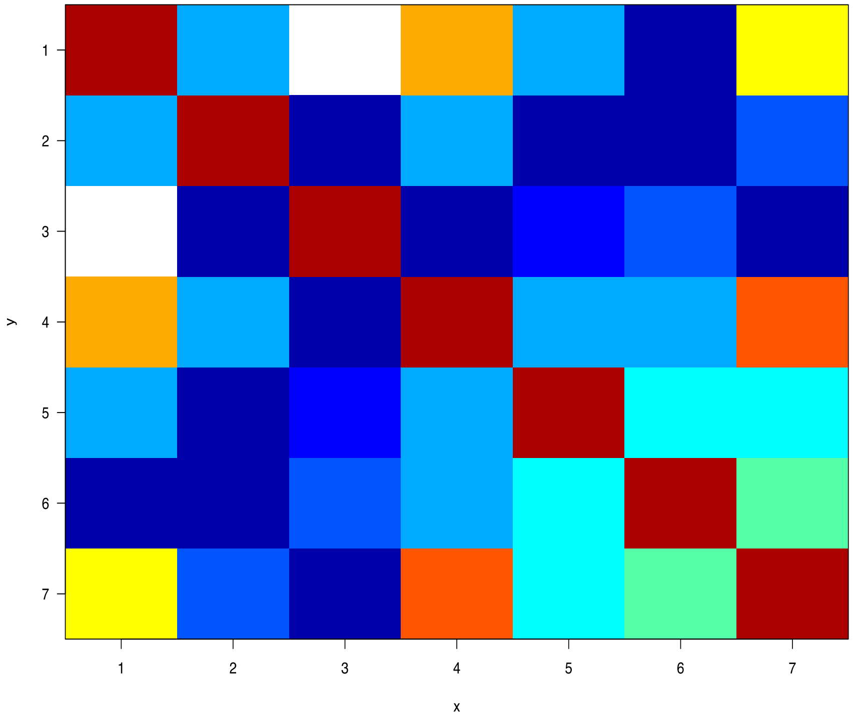
Different factors affect movies' revenue according to a linear relationship.

Hypothesis 1: Preprocessing

- Select 7 variables: revenue, budget, production company, release quarter, runtime, vote average, vote count
- Eliminate missing data
- Choose data from year 2012 to year 2015
- Remain 458 movie data

Hypothesis 1: Relationship among variables

```
> cor(x)
                      revenue     budget production_company release_quarter    runtime vote_average vote_count
revenue           1.000000000 0.72919591        0.268809408     0.001578318 0.26158648  0.258322178 0.77613431
budget            0.729195909 1.00000000        0.303253177    -0.055162857 0.31963125  0.041490950 0.58952558
production_company 0.268809408 0.30325318       1.000000000     0.012813545 0.06413295  0.002363084 0.21063118
release_quarter   0.001578318 -0.05516286       0.012813545     1.000000000 0.13597366  0.235057612 0.01974299
runtime           0.261586476 0.31963125        0.064132955     0.135973658 1.00000000  0.356350478 0.37028139
vote_average      0.258322178 0.04149095        0.002363084     0.235057612 0.35635048  1.000000000 0.42499480
vote count        0.776134310 0.58952558        0.210631185     0.019742992 0.37028139  0.424994796 1.00000000
```

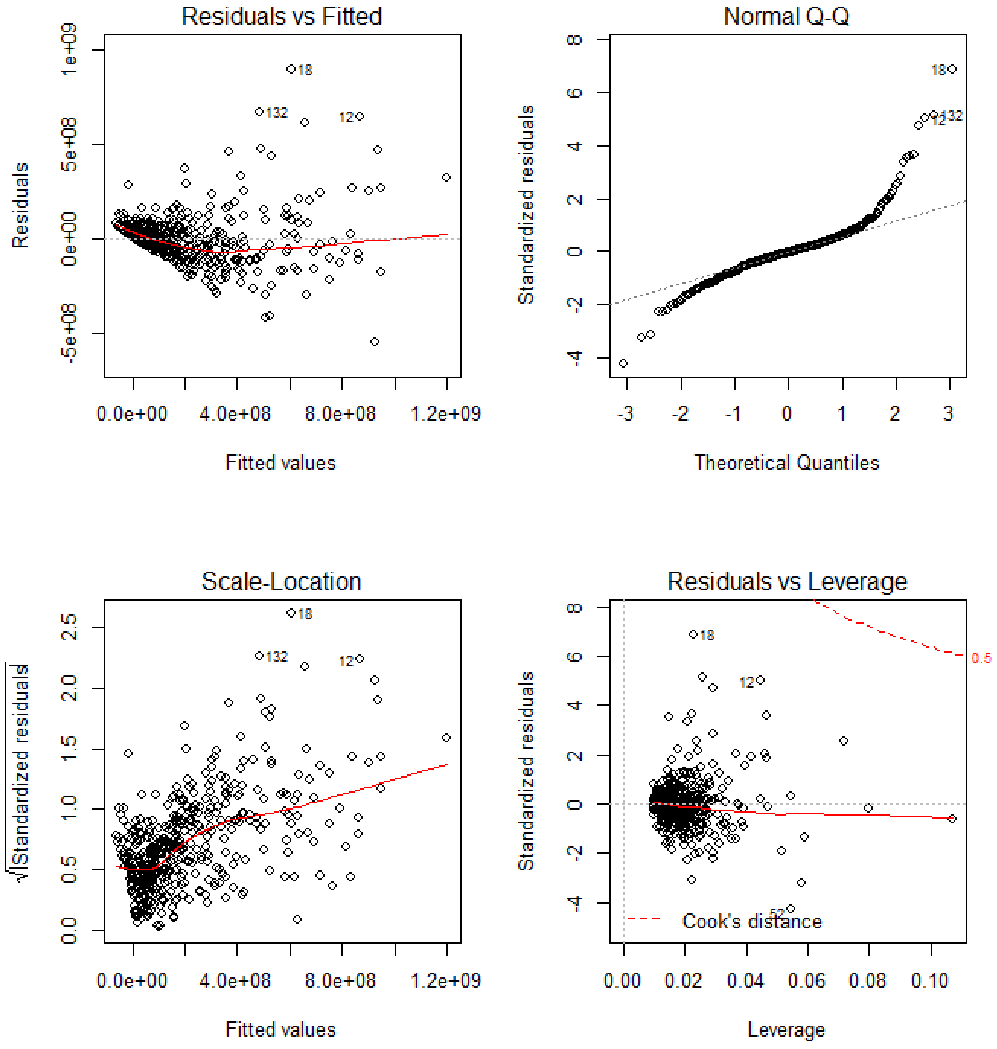


Hypothesis 1: Multiple Linear Regression

```
> fit<-lm(revenue~budget+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
> summary(fit)

Call:
lm(formula = revenue ~ budget + production_company + factor(release_quarter) +
    runtime + vote_average + vote_count, data = x)

Residuals:
       Min         1Q     Median         3Q        Max
-546781577  -58857399   -4629406   46116471  901052683

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.093e+07  6.920e+07   0.447 0.655091
budget                    1.827e+00  1.454e-01  12.569  < 2e-16 ***
production_company        1.184e+07  1.331e+07   0.889 0.374287
factor(release_quarter)2  4.316e+07  1.838e+07   2.348 0.019293 *
factor(release_quarter)3 -1.127e+06  1.730e+07  -0.065 0.948105
factor(release_quarter)4  2.855e+07  1.857e+07   1.538 0.124770
runtime                  -1.689e+06  4.885e+05  -3.457 0.000598 ***
vote_average              1.408e+07  1.023e+07   1.376 0.169455
vote_count                7.200e+04  4.851e+03  14.844  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132800000 on 448 degrees of freedom
Multiple R-squared:  0.7286,    Adjusted R-squared:  0.7238
F-statistic: 150.4 on 8 and 448 DF,  p-value: < 2.2e-16
```

Hypothesis 1: Homoscedasticity & Normality

- ✓ Equal variances assumption does NOT hold.
- ✓ The residuals are NOT normally distributed.

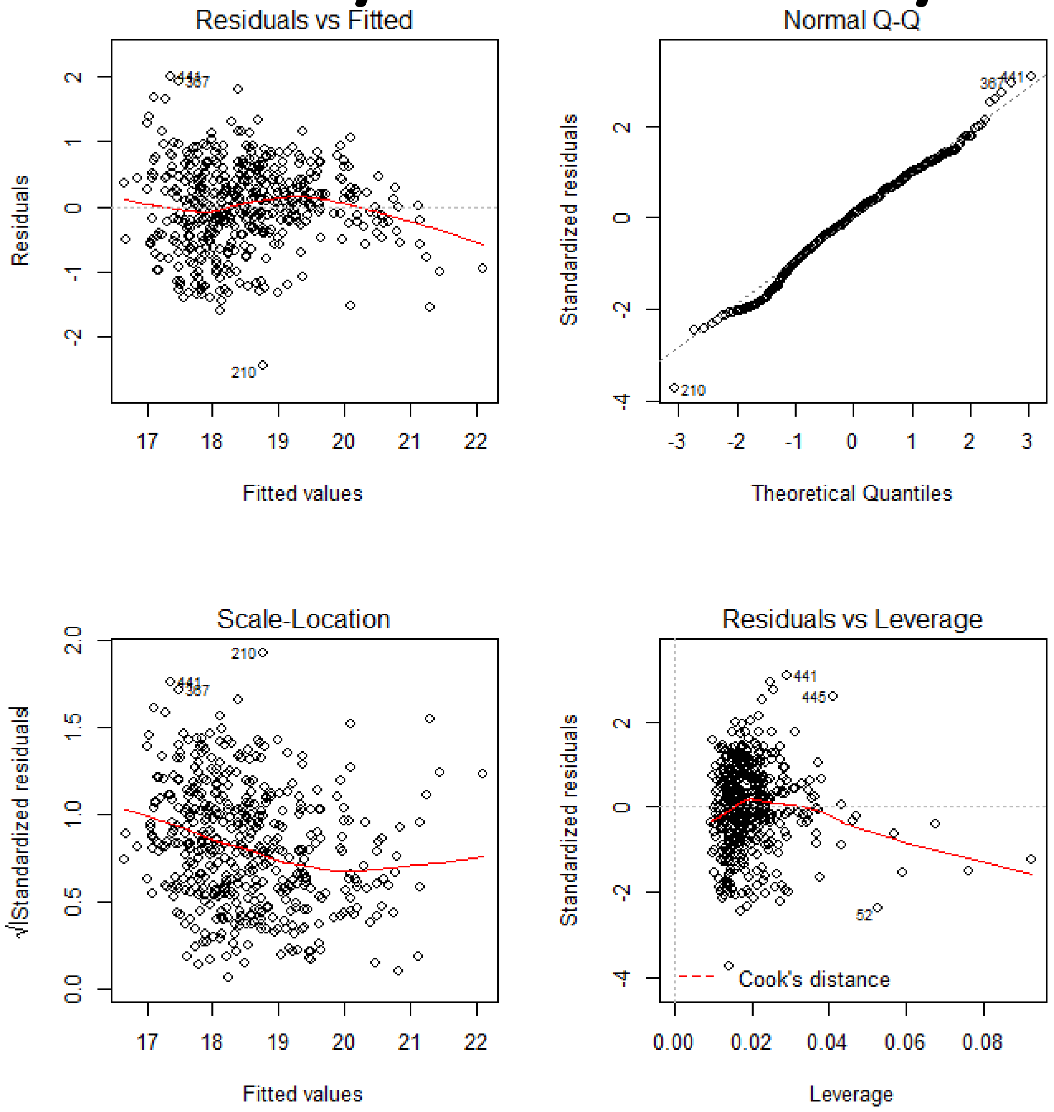


Hypothesis 1: Multiple Linear Regression With Log Transformation On Revenue And Budget

```
> fit<-lm(log(revenue)~log(budget)+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
> summary(fit)

Call:
lm(formula = log(revenue) ~ log(budget) + production_company +
    factor(release_quarter) + runtime + vote_average + vote_count,
    data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4366 -0.4002  0.0449  0.4314  2.0046

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               7.9698360  0.8150435   9.778  < 2e-16 ***
log(budget)               0.5691405  0.0414548  13.729  < 2e-16 ***
production_company        0.1664532  0.0669745   2.485 0.013307 *
factor(release_quarter)2  0.2128106  0.0903399   2.356 0.018920 *
factor(release_quarter)3  0.1364737  0.0856947   1.593 0.111965
factor(release_quarter)4  0.1583921  0.0918846   1.724 0.085431 .
runtime                  -0.0092760  0.0024168  -3.838 0.000142 ***
vote_average              0.1466406  0.0508486   2.884 0.004118 **
vote_count                0.0002596  0.0000229  11.340  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6572 on 448 degrees of freedom
Multiple R-squared:  0.6715,    Adjusted R-squared:  0.6656
F-statistic: 114.5 on 8 and 448 DF,  p-value: < 2.2e-16
```

Hypothesis 1: Homoscedasticity& Normality

- ✓ Equal variances assumption valid.
- ✓ The residuals are NORMALLY distributed.

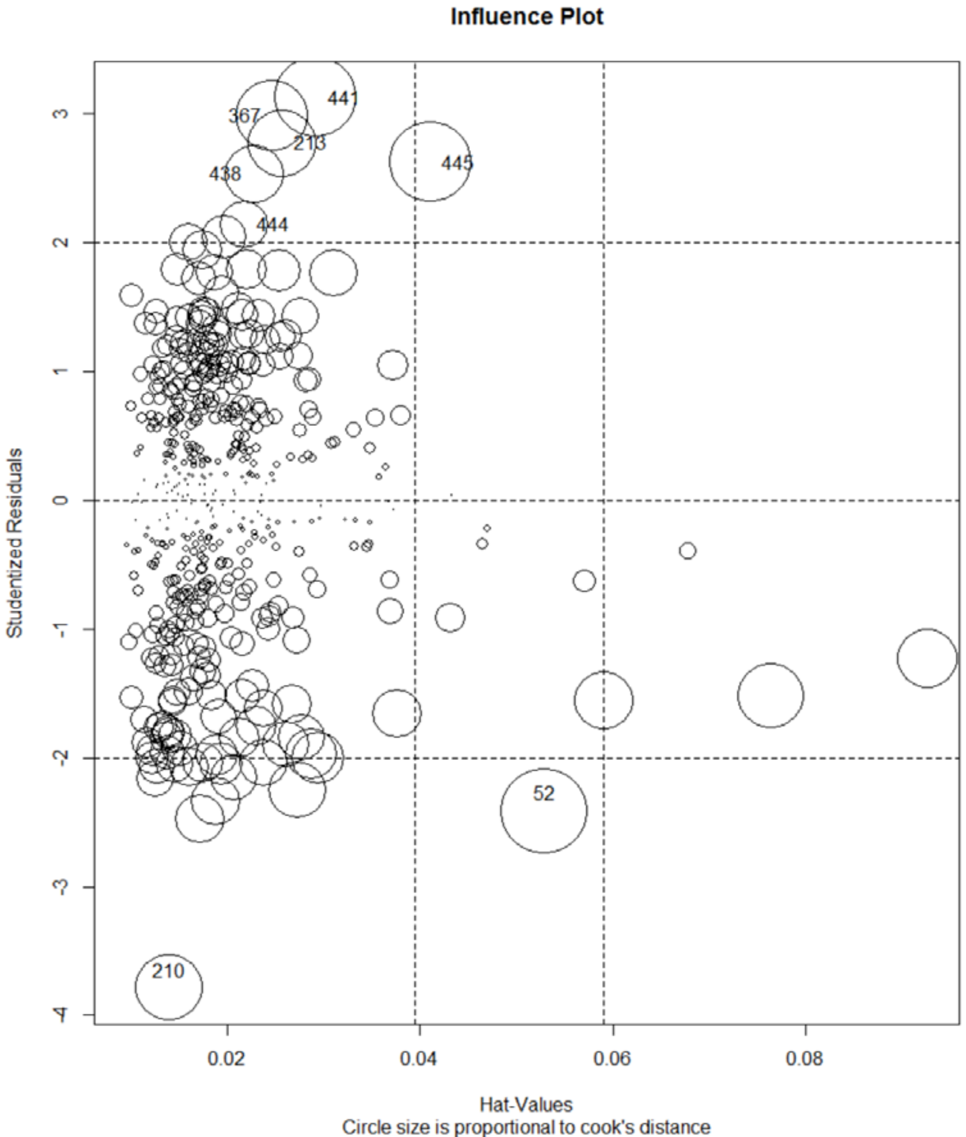


Hypothesis 1: Evaluate Multicollinearity

```
> vif(fit)
                              GVIF Df GVIF^(1/(2*Df))
log(budget)               1.730199  1        1.315370
production_company        1.157884  1        1.076050
factor(release_quarter)   1.179594  3        1.027911
runtime                   1.317103  1        1.147651
vote_average              1.540106  1        1.241010
vote_count                1.876520  1        1.369861
```

✓ VIF values are acceptable.

Hypothesis 1: Detect Outliers, High-leverage Points and Influential Observations

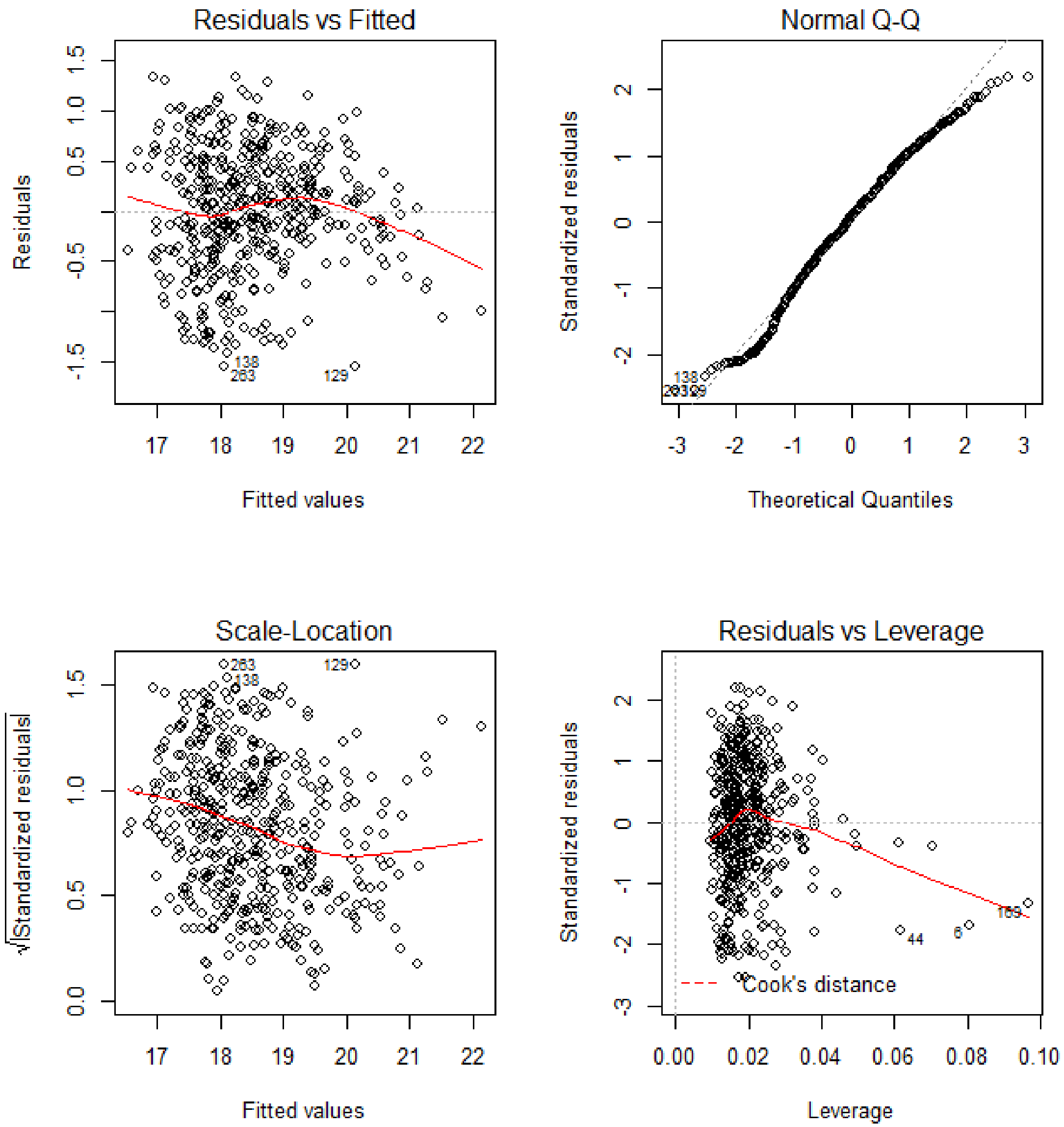


Hypothesis 1: New Multiple Linear Regression

```
> x<-read.table("C:/Users/admin/Desktop/variables.csv", header=TRUE, sep=",")
> fit<-lm(log(revenue)~log(budget)+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
> summary(fit)

Call:
lm(formula = log(revenue) ~ log(budget) + production_company +
    factor(release_quarter) + runtime + vote_average + vote_count,
    data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55176 -0.38863  0.05274  0.43582  1.33437

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.863e+00  7.879e-01   8.710  < 2e-16 ***
log(budget)               6.189e-01  3.971e-02  15.587  < 2e-16 ***
production_company        1.713e-01  6.317e-02   2.712  0.00695 **
factor(release_quarter)2  1.956e-01  8.500e-02   2.302  0.02182 *
factor(release_quarter)3  1.170e-01  8.049e-02   1.454  0.14670
factor(release_quarter)4  1.380e-01  8.702e-02   1.586  0.11340
runtime                  -9.916e-03  2.266e-03  -4.376 1.51e-05 ***
vote_average              1.966e-01  4.879e-02   4.029 6.61e-05 ***
vote_count                2.529e-04  2.203e-05  11.481  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6138 on 440 degrees of freedom
Multiple R-squared:  0.7133,    Adjusted R-squared:  0.7081
F-statistic: 136.9 on 8 and 440 DF,  p-value: < 2.2e-16
```

Hypothesis 1: Homoscedasticity & Normality

- ✓ Equal variances assumption valid.
- ✓ The residuals are NORMALLY distributed.

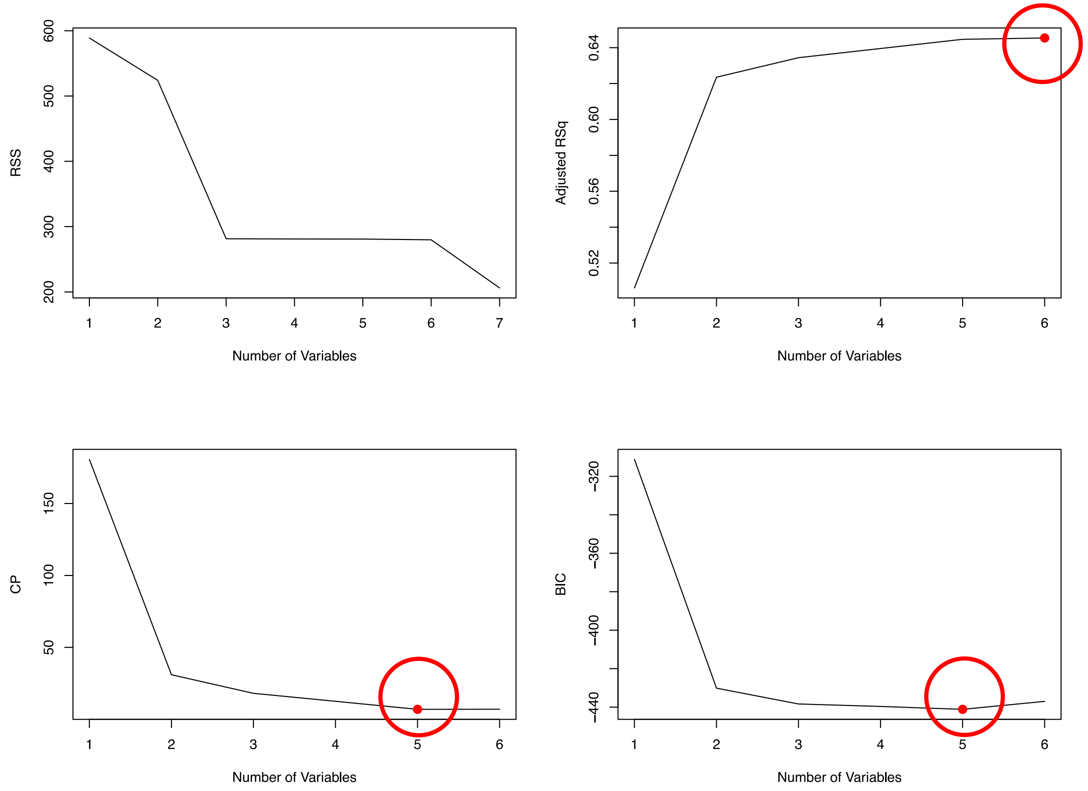


Hypothesis 1: Model Optimization

- Best Subset Model Selection

	budget	production_companies	release_quarter	runtime	vote_average	vote_count
1 (1)	"*"	" "	" "	" "	" "	" "
2 (1)	"*"	" "	" "	" "	" "	"*"
3 (1)	"*"	"*"	" "	" "	" "	"*"
4 (1)	"*"	"*"	" "	"*"	" "	"*"
5 (1)	"*"	"*"	" "	"*"	"*"	"*"
6 (1)	"*"	"*"	"*"	"		

- Optimality Criteria

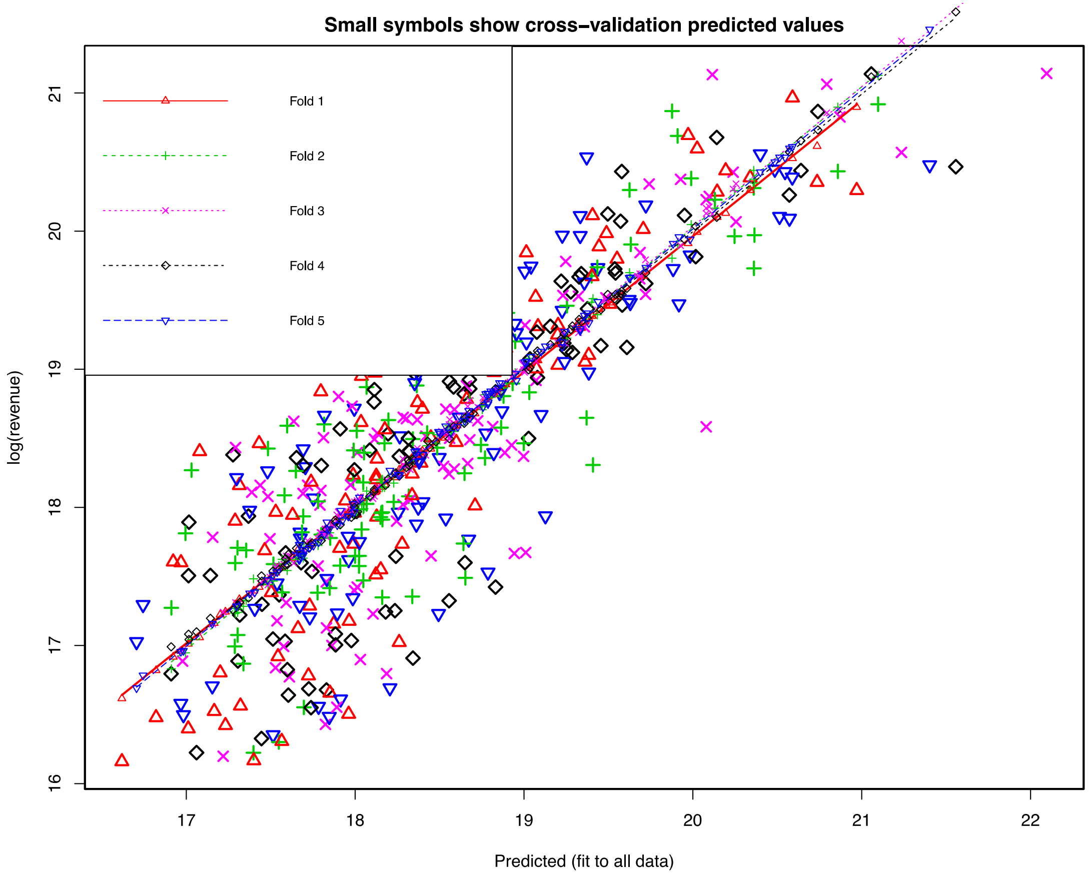


  1. Regression Subset Selection
  2. Adjusted r_p^2-Criterion
  3. C_p-Criterion
  4. Bayesian information criterion

Hypothesis 1: Model Optimization

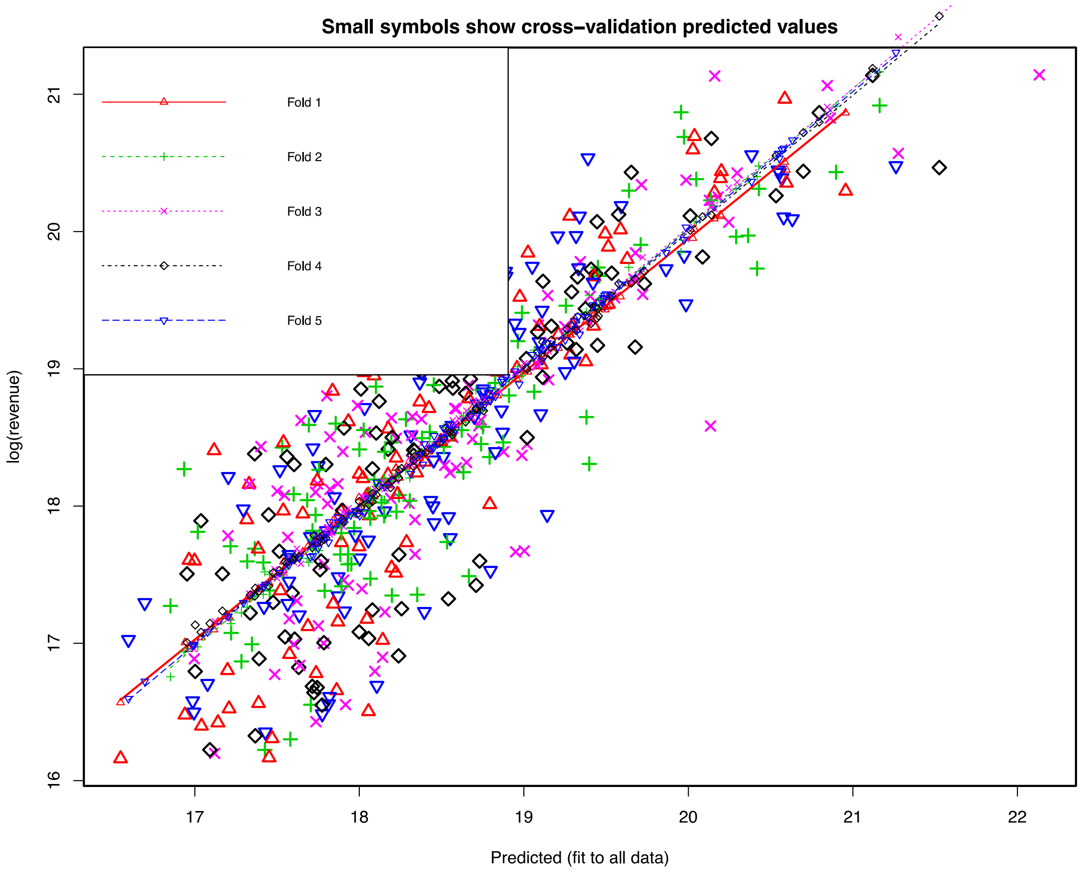
- Cross Validation

5-variable model



Overall MS=0.385

6-variable model



Overall MS=0.385

Hypothesis 1: Conclusion

✓ $\log(\mathrm{y}) = 0.6189 \log(x_1) + 0.1713x_2 + 0.1956\, x_3 - 0.009916\, x_4 + 0.1966\, x_5 + 0.0002529\, x_6 + 6.863$

$y: revenue$	$x_1: budget$	$x_2: production\ company$
$x_3: factor(release\ quarter2)$		$x_4: runtime$
$x_5: vote\ average$		$x_6: vote\ count$

✓ Since our model has demonstrated over-fitting, this implies that a more sufficient non-linear model is needed to appropriately to model the data relationship.

Hypothesis 2

Revenue Depends On Genres

Hypothesis 2: Summary Statistics

Summary statistics by groups- count, mean & standard deviation

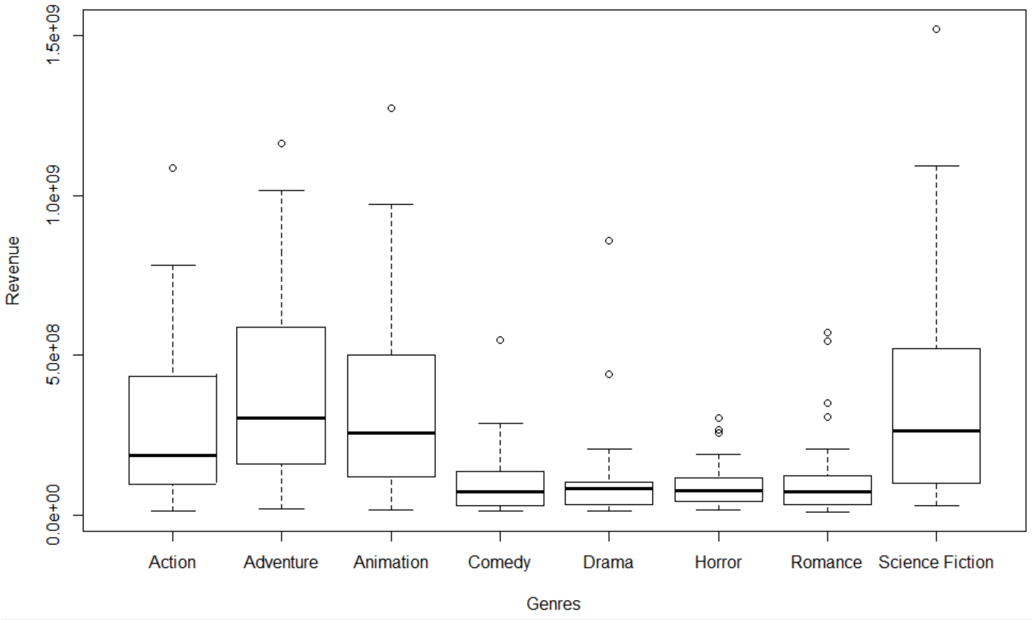
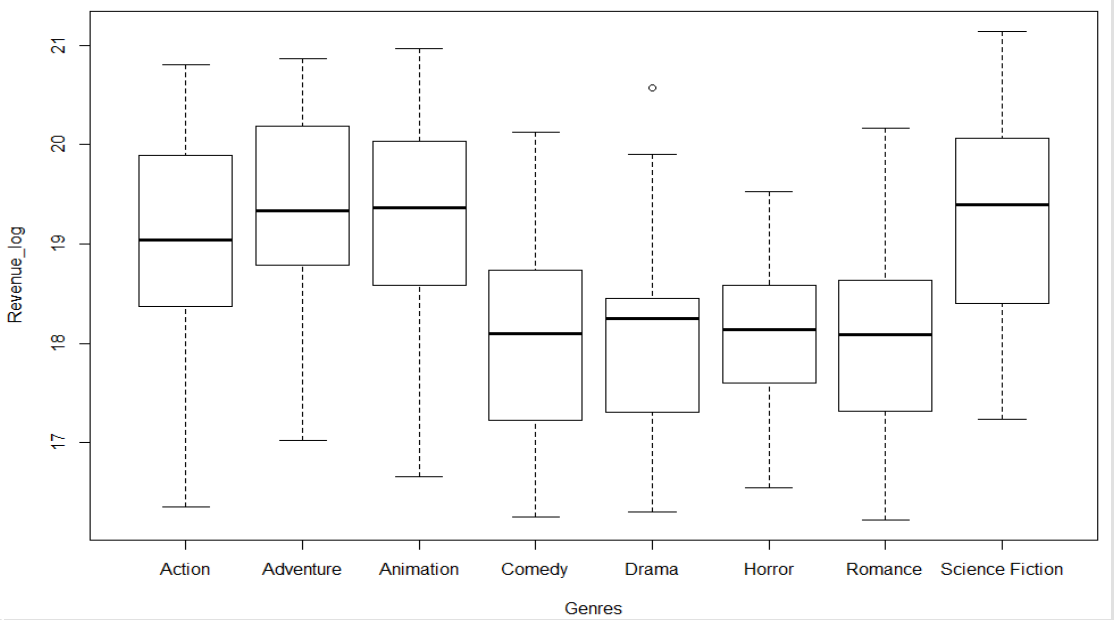
```
# A tibble: 8 x 4
          genres count      mean        sd
           <chr> <int>     <dbl>     <dbl>
1         Action    30 292772290 283023146
2      Adventure    30 365221122 301640784
3      Animation    30 358753585 326682246
4         Comedy    30 106464838 110403360
5          Drama    30 113976380 163743480
6         Horror    30  98024225  75466320
7        Romance    30 120341542 144614826
8 Science Fiction   30 361934480 351419344
```

Hypothesis 2: Homogeneity of Variance

Check Equality of Variances	
Without Log Transformation	With Log Transformation
Equal variance assumption does NOT hold	Equal variance assumption valid

`> bartlett.test(category$revenue ~ category$genres)`
 	
`Bartlett test of homogeneity of variances`
 	
`data:  category$revenue by category$genres`
`Bartlett's K-squared = 102.6, df = 7, p-value < 2.2e-16`
	`> bartlett.test(category$revenue_log ~ category$genres)`
	
	`Bartlett test of homogeneity of variances`
	
	`data:  category$revenue_log by category$genres`
	`Bartlett's K-squared = 5.681, df = 7, p-value = 0.5775`

Hypothesis 2: Boxplot

Boxplot without Log Transformation	Boxplot with Log Transformation
	

Hypothesis 2: One-way ANOVA

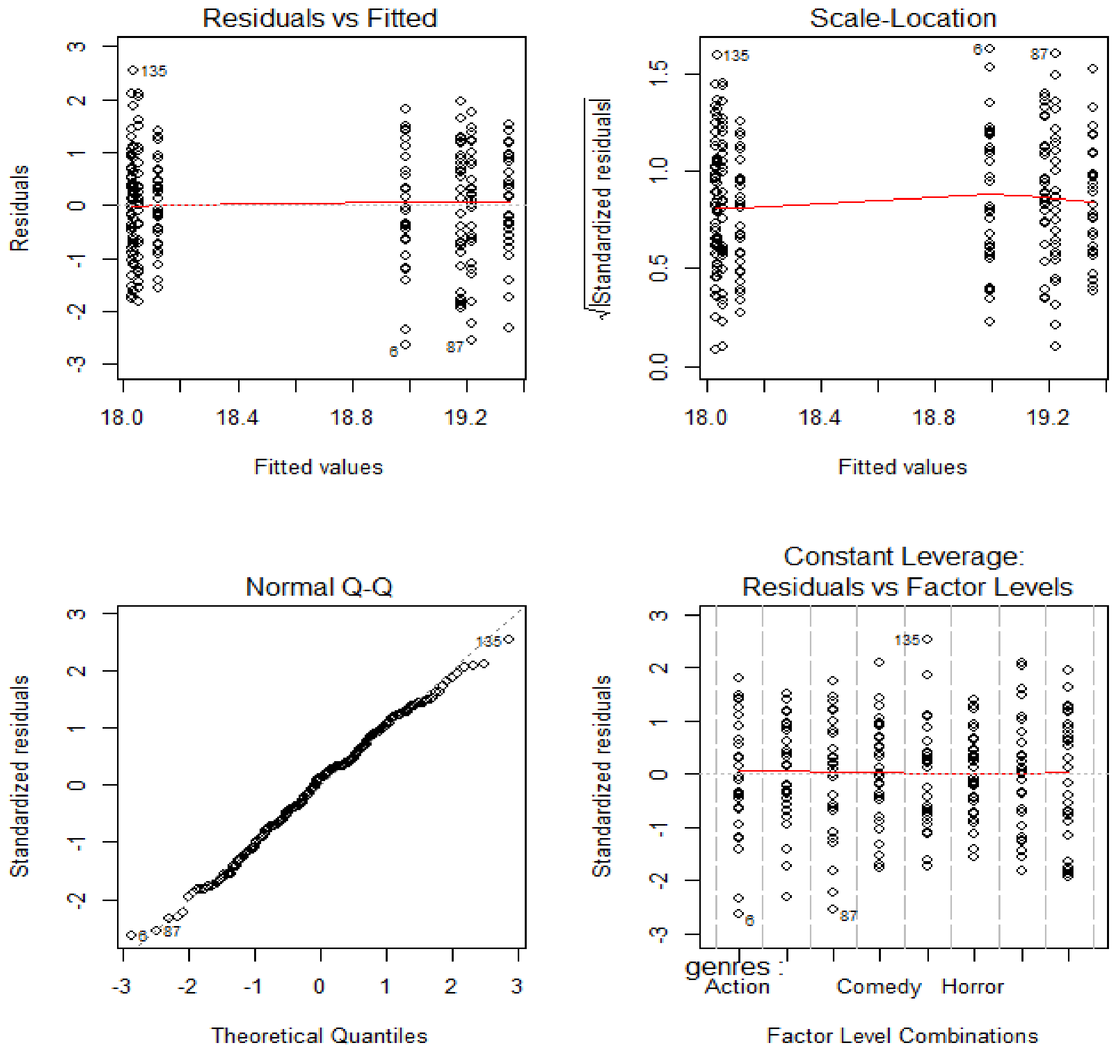
- F test

```
> fit <- aov(revenue_log ~ genres, data = category)
> summary(fit)
              Df Sum Sq Mean Sq F value    Pr(>F)    
genres         7  78.27  11.181   10.73 1.06e-11 ***
Residuals    232 241.70   1.042                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There do appear to be SIGNIFICANT DIFFERENCES among the genres in terms of revenue.

- Normality

✓ The residuals are NORMALLY distributed.



Hypothesis 2: Scheffe Test vs. Tukey Test

	Scheffe Test	Tukey Test
Group 1	Adventure, Animation, Science Fiction	Adventure, Animation, Science Fiction , Action
Group 2	Horror, Romance, Drama, Comedy	Horror, Romance, Drama, Comedy
Group 3	Action	

```
> library(agricolae)
Warning message:
package 'agricolae' was built under R version 3.4.2
> comparison <- scheffe.test(fit, "genres", alpha = 0.05)
> comparison
$statistics
   MSerror  Df        F     Mean       CV  Scheffe CriticalDifference
  1.041811 232 2.049195 18.62398 5.480522 3.787396          0.9981355

$parameters
     test name.t ntr alpha
  Scheffe genres   8  0.05

$means
                revenue_log       std  r      Min      Max      Q25      Q50      Q75
Action             18.99130 1.1156226 30 16.35232 20.80479 18.41113 19.04077 19.82519
Adventure          19.35168 0.9354952 30 17.02309 20.86590 18.80719 19.33094 20.13219
Animation          19.22045 1.1091492 30 16.65677 20.96560 18.60582 19.36494 20.01664
Comedy             18.03067 1.0101446 30 16.25065 20.12428 17.31041 18.09983 18.72666
Drama              18.03502 0.9682944 30 16.30685 20.56966 17.32811 18.25358 18.44772
Horror             18.12134 0.7824412 30 16.54988 19.52933 17.62235 18.13867 18.58211
Romance            18.05627 1.0591694 30 16.22344 20.16291 17.32909 18.08796 18.57875
Science Fiction    19.18511 1.1368686 30 17.23400 21.14169 18.41559 19.39201 20.04687

$comparison
NULL

$groups
                revenue_log groups
Adventure          19.35168      a
Animation          19.22045      a
Science Fiction    19.18511      a
Action             18.99130     ab
Horror             18.12134      b
Romance            18.05627      b
Drama              18.03502      b
Comedy             18.03067      b

attr(,"class")
[1] "group"
```

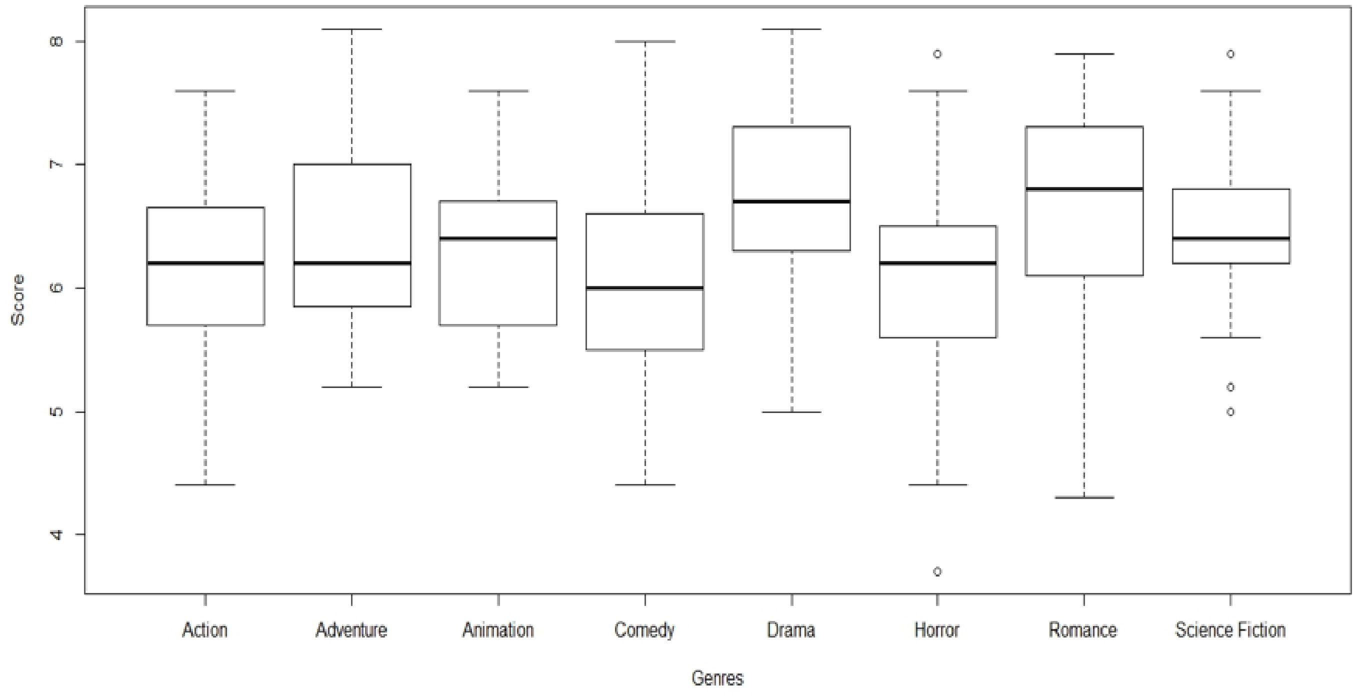
```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = revenue_log ~ genres, data = category)

$genres
                                  diff        lwr         upr     p adj
Adventure-Action            0.360384801 -0.4457391  1.16650875 0.8710404
Animation-Action            0.229156019 -0.5769679  1.03527997 0.9884399
Comedy-Action              -0.960624555 -1.7667485 -0.15450061 0.0078267
Drama-Action               -0.956274921 -1.7623989 -0.15015097 0.0082875
Horror-Action              -0.869952493 -1.6760764 -0.06382854 0.0243247
Romance-Action             -0.935021707 -1.7411457 -0.12889776 0.0109166
Science Fiction-Action      0.193813059 -0.6123109  0.99993701 0.9958256
Animation-Adventure        -0.131228782 -0.9373527  0.67489517 0.9996628
Comedy-Adventure           -1.321009356 -2.1271333 -0.51488541 0.0000290
Drama-Adventure            -1.316659722 -2.1227837 -0.51053577 0.0000314
Horror-Adventure           -1.230337294 -2.0364612 -0.42421334 0.0001376
Romance-Adventure          -1.295406508 -2.1015305 -0.48928256 0.0000455
Science Fiction-Adventure  -0.166571742 -0.9726957  0.63955221 0.9983981
Comedy-Animation           -1.189780574 -1.9959045 -0.38365662 0.0002677
Drama-Animation            -1.185430940 -1.9915549 -0.37930699 0.0002872
Horror-Animation           -1.099108512 -1.9052325 -0.29298456 0.0011039
Romance-Animation          -1.164177726 -1.9703017 -0.35805378 0.0004034
Science Fiction-Animation  -0.035342959 -0.8414669  0.77078099 1.0000000
Drama-Comedy                0.004349634 -0.8017743  0.81047358 1.0000000
Horror-Comedy               0.090672062 -0.7154519  0.89679601 0.9999721
Romance-Comedy              0.025602848 -0.7805211  0.83172680 1.0000000
Science Fiction-Comedy      1.154437614  0.3483137  1.96056156 0.0004706
Horror-Drama                0.086322429 -0.7198015  0.89244638 0.9999801
Romance-Drama               0.021253214 -0.7848707  0.82737716 1.0000000
Science Fiction-Drama       1.150087981  0.3439640  1.95621193 0.0005039
Romance-Horror             -0.065069214 -0.8711932  0.74105473 0.9999971
Science Fiction-Horror      1.063765552  0.2576416  1.86988950 0.0018649
Science Fiction-Romance     1.128834767  0.3227108  1.93495872 0.0007015
```

Hypothesis 3

- Vote Score Depends On Genres
- Drama Has Higher Vote Score Than The Others

Hypothesis 3: Summary Statistics

Summary Statistics of Vote Scores								
Genres	Action	Adventure	Animation	Comedy	Drama	Horror	Romance	Science Fiction
Sample mean	6.20	6.45	6.29	6.06	6.73	6.08	6.68	6.48
Sample SD	0.71	0.76	0.61	0.75	0.71	0.80	0.85	0.66



Hypothesis 3: Homogeneity of Variance

```
> bartlett.test(category$score ~ category$genres)

        Bartlett test of homogeneity of variances

data:  category$score by category$genres
Bartlett's K-squared = 6.4562, df = 7, p-value = 0.4876
```

✓Equal variance assumption valid

Hypothesis 3: One-way ANOVA

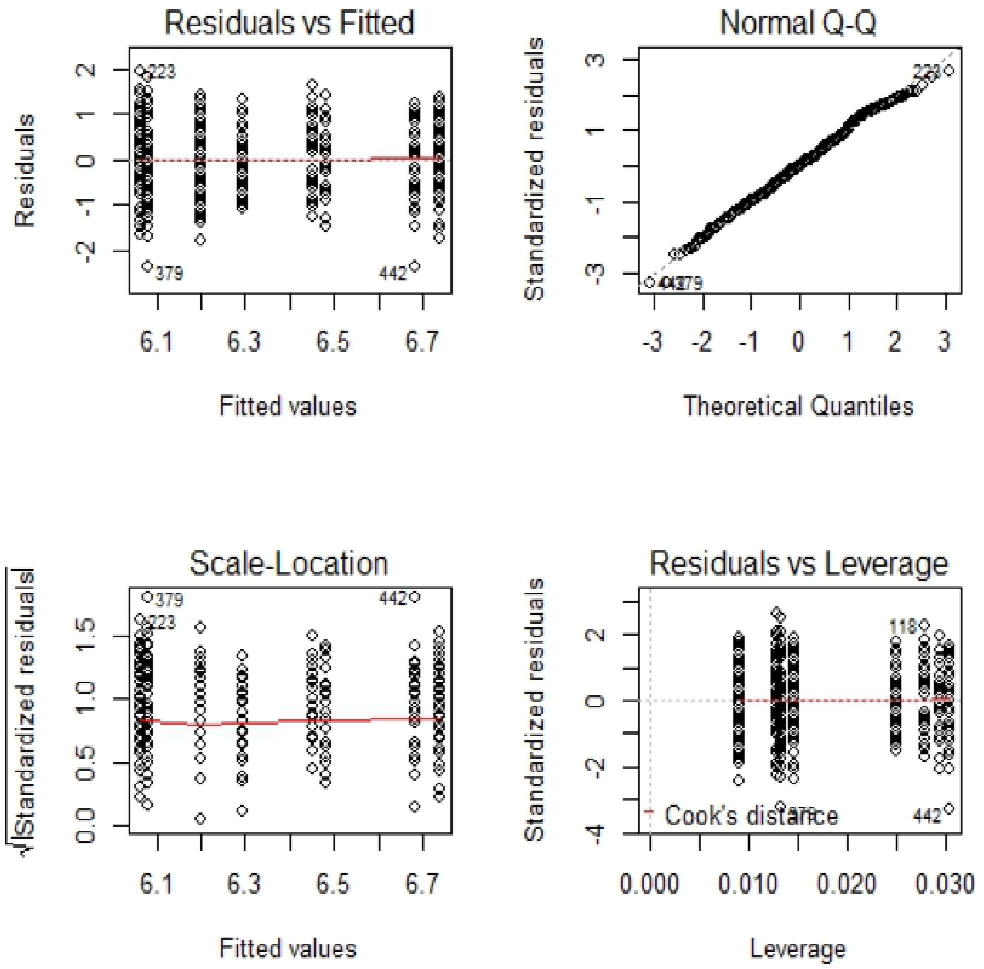
- F test

```
> fit <- aov(score ~ genres, data = category)
> summary(fit)
             Df Sum Sq Mean Sq F value   Pr(>F)
genres        7  28.92   4.131   7.601 1.07e-08 ***
Residuals   468 254.38   0.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

✓ There do appear to be SIGNIFICANT DIFFERENCES among the genres in terms of vote average scores.

- Normality

✓ The residuals are NORMALLY distributed.



Hypothesis 3: Inference On Two Population Variances (σ^2_{drama} & $\sigma^2_{the\ others}$)

```
> var.test(drama$score,others$score)

        F test to compare two variances

data:  drama$score and others$score
F = 0.88278, num df = 68, denom df = 406, p-value = 0.5367
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6273313 1.3034151
sample estimates:
ratio of variances
         0.8827772
```

✓At α=0.05, we DO NOT have sufficient evidence to reject null hypothesis($H_0: \sigma_1^2 = \sigma_2^2$) and use EQUAL variance t test.

Hypothesis 3: Inference On Two Population Means(μ_{drama} & $\mu_{the\ others}$)

```
> t.test(drama$score,others$score,alternative='greater',var.equal=TRUE)

        Two Sample t-test

data:  drama$score and others$score
t = 5.0028, df = 474, p-value = 3.986e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3290849       Inf
sample estimates:
mean of x mean of y
 6.737681  6.246929
```

✓ We conclude that the average score of drama is HIGHER than the average score of others.

R Code

```
Hypothesis 1:
x<-read.table("C:/Users/admin/Desktop/variables.csv", header=TRUE, sep=",")
library(car)

#####Evaluate Correlation#####
cor(x)

#####Construct Multiple Linear Regression#####
fit<-lm(revenue~budget+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
summary(fit)
par(mfrow=c(2,2))
plot(fit)

#####Do Log Transformation#####
fit<-lm(log(revenue)~log(budget)+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
summary(fit)
par(mfrow=c(2,2))
plot(fit)

#####Evaluate Multicollinearity#####
vif(fit)

#####Detect Outliers, High-leverage Points and Influential Observations#####
influencePlot(fit, id.method = "identify", main="Influence Plot", sub="Circle size is proportional to cook's distance")

#####Construct New Multiple Linear Regression After Deleting Outliers, High-leverage Points and Influential Observations#####
x<-read.table("C:/Users/admin/Desktop/variables_new.csv", header=TRUE, sep=",")
fit<-lm(log(revenue)~log(budget)+production_company+factor(release_quarter)+runtime+vote_average+vote_count,data=x)
summary(fit)
par(mfrow=c(2,2))
plot(fit)
```

```
#####Optimize the Model#####
library(leaps)

best.subset <- regsubsets(log(revenue)~., x , nvmax=6)
best.subset.summary <- summary(best.subset)
best.subset.summary$outmat

best.subset.by.adjr2 <- which.max(best.subset.summary$adjr2)
best.subset.by.adjr2

best.subset.by.cp <- which.min(best.subset.summary$cp)
best.subset.by.cp

best.subset.by.bic <- which.min(best.subset.summary$bic)
best.subset.by.bic

par(mfrow=c(2,2))
plot(best.subset$rss, xlab="Number of Variables", ylab="RSS", type="l")
plot(best.subset.summary$adjr2, xlab="Number of Variables", ylab="Adjusted RSq", type="l")
points(best.subset.by.adjr2, best.subset.summary$adjr2[best.subset.by.adjr2], col="red", cex =2, pch =20)
plot(best.subset.summary$cp, xlab="Number of Variables", ylab="CP", type="l")
points(best.subset.by.cp, best.subset.summary$cp[best.subset.by.cp], col="red", cex =2, pch =20)
plot(best.subset.summary$bic, xlab="Number of Variables", ylab="BIC", type="l")
points(best.subset.by.bic, best.subset.summary$bic[best.subset.by.bic], col="red", cex =2, pch =20)

#####Cross-validation for 5-variable model#####
CVlm(data = x, form.lm=formula(log(revenue)~log(budget)+production_company+runtime+vote_average+vote_count), m = 5)

#####Cross-validation for 6-variable model#####
CVlm(data = x, form.lm=formula(log(revenue)~log(budget)+production_company+factor(release_quarter)+runtime+vote_average+vote_count), m = 5)
```

Hypothesis 2:

```
category <- read.table("C:/Users/admin/Desktop/genre.csv", header=TRUE, sep=",")
boxplot(revenue ~ genres, category, xlab="Genres", ylab="Revenue")

#####Test of Homogeneity of Variance#####
bartlett.test(category$revenue ~ category$genres)

category$revenue_log <- log(category$revenue)
boxplot(revenue_log ~ genres, category, xlab="Genres", ylab="Revenue_log")
bartlett.test(category$revenue_log ~ category$genres)

#####One-way ANOVA: F test#####
fit <- aov(revenue_log ~ genres, data = category)
summary(fit)
par(mfcol=c(2,2))
plot(fit)

library(agricolae)
comparison <- scheffe.test(fit, "genres", alpha = 0.05)
comparison
TukeyHSD(fit)
```

Hypothesis 3:

```
category <- read.table("C:/Users/admin/Desktop/genre1.csv", header=TRUE, sep=","
boxplot(score ~ genres, category, xlab="Genres", ylab="Score")

#####Summary Statistics#####
action=category[2:111,]
adventure=category[112:147,]
animation=category[148:187,]
comedy=category[188:265,]
drama=category[266:334,]
horror=category[335:409,]
romance=category[410:442,]
sciencefiction=category[443:476,]
mean(genres$score)
sd(genres$score)

#####Test of Homogeneity of Variance#####
bartlett.test(category$score ~ category$genres)

#####One-way ANOVA: F test#####
fit <- aov(score ~ genres, data = category)
summary(fit)
par(mfrow=c(2,2))
plot(fit)

#####Inference for Two Sample Variances#####
drama=category[266:334,]
others=category[-c(266:334),]
var.test(drama$score,others$score)

#####Inference for Two Sample Means#####
t.test(drama$score,others$score,alternative='greater',var.equal=TRUE)
```