Project EDA

Yash Manne

02/04/2023

Data Source

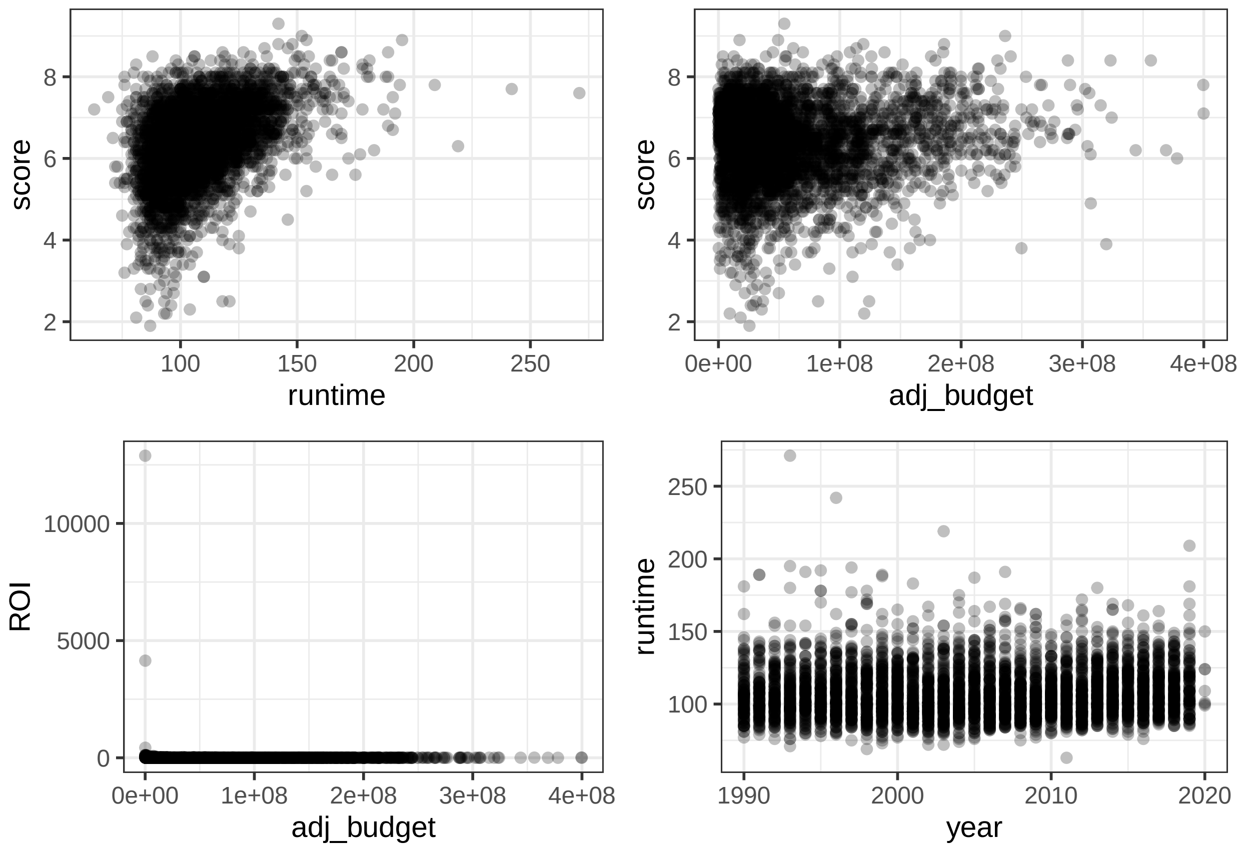
```
raw_df = read.csv('cleaned_data.csv')
```

```
df = raw_df[raw_df$year>=1990,]
df$rating = factor(df$rating, levels=c("G", "PG", "PG-13", "R"))
df$genre = relevel(factor(df$genre), "Other")
df$binned_director = factor(df$binned_director, levels=c("False", "True"))
df$binned_writer = factor(df$binned_writer, levels=c("False", "True"))
df$binned_star = factor(df$binned_star, levels=c("False", "True"))
df$binned_company = factor(df$binned_company, levels=c("False", "True"))
# df$is_sequel = as.factor(df$is_sequel)
# df$is_remake = as.factor(df$is_remake)

df$binROI = df$ROI > 0
```

```
p1 = df %>% ggplot(mapping = aes(x= runtime, y=score)) +
  geom_point(alpha=0.25) + theme_bw()
p2 = df %>% ggplot(mapping = aes(x= adj_budget, y=score)) +
  geom_point(alpha=0.25) + theme_bw()
p3 = df %>% ggplot(mapping = aes(x= adj_budget, y=ROI)) +
  geom_point(alpha=0.25) + theme_bw()
p4 = df %>% ggplot(mapping = aes(x= year, y=runtime)) +
  geom_point(alpha=0.25) + theme_bw()

cowplot::plot_grid(p1,p2,p3,p4,
  ncol = 2, nrow = 2
)
```



Linear Regression with Score

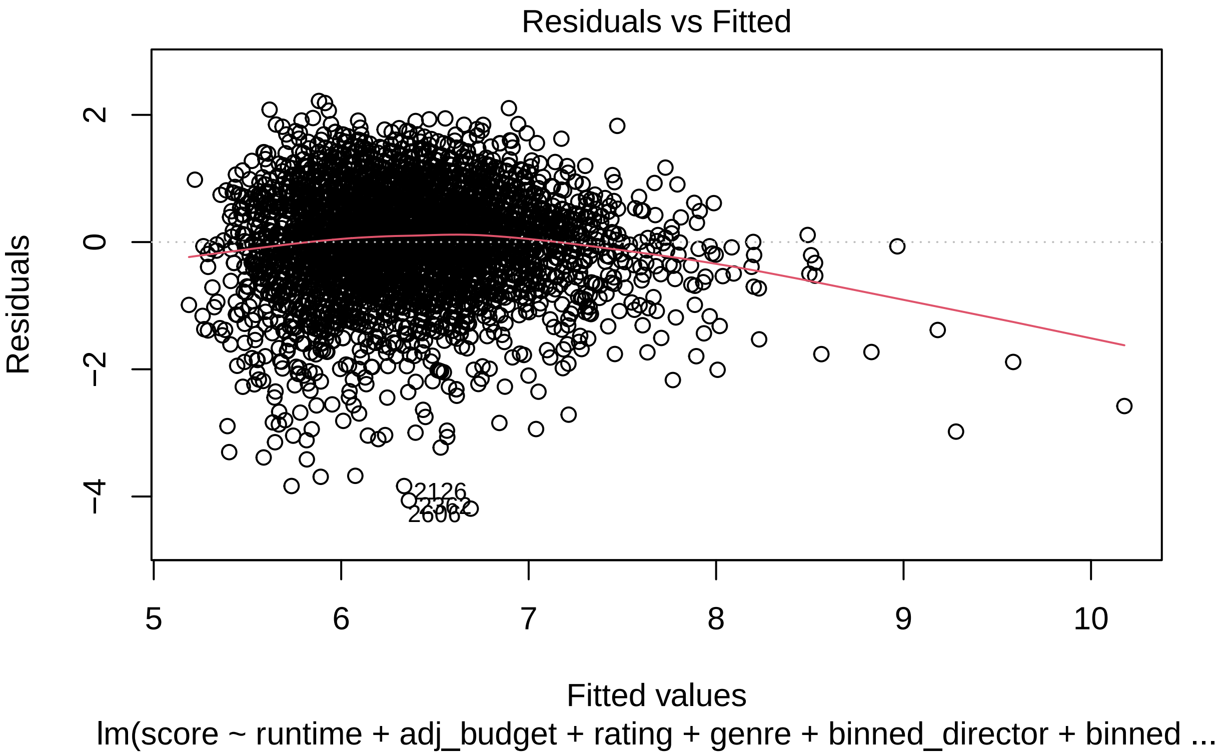
```
summary(df$score)
```

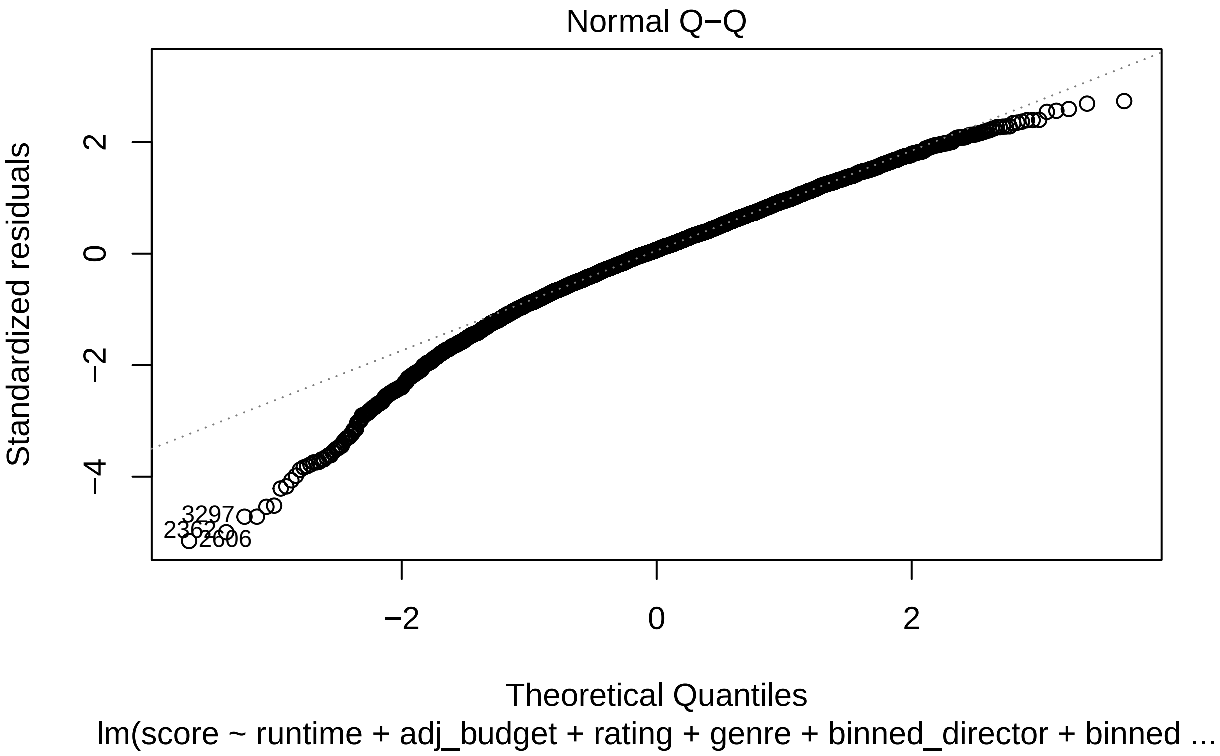
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.90    5.80    6.40    6.36    7.00    9.30
```

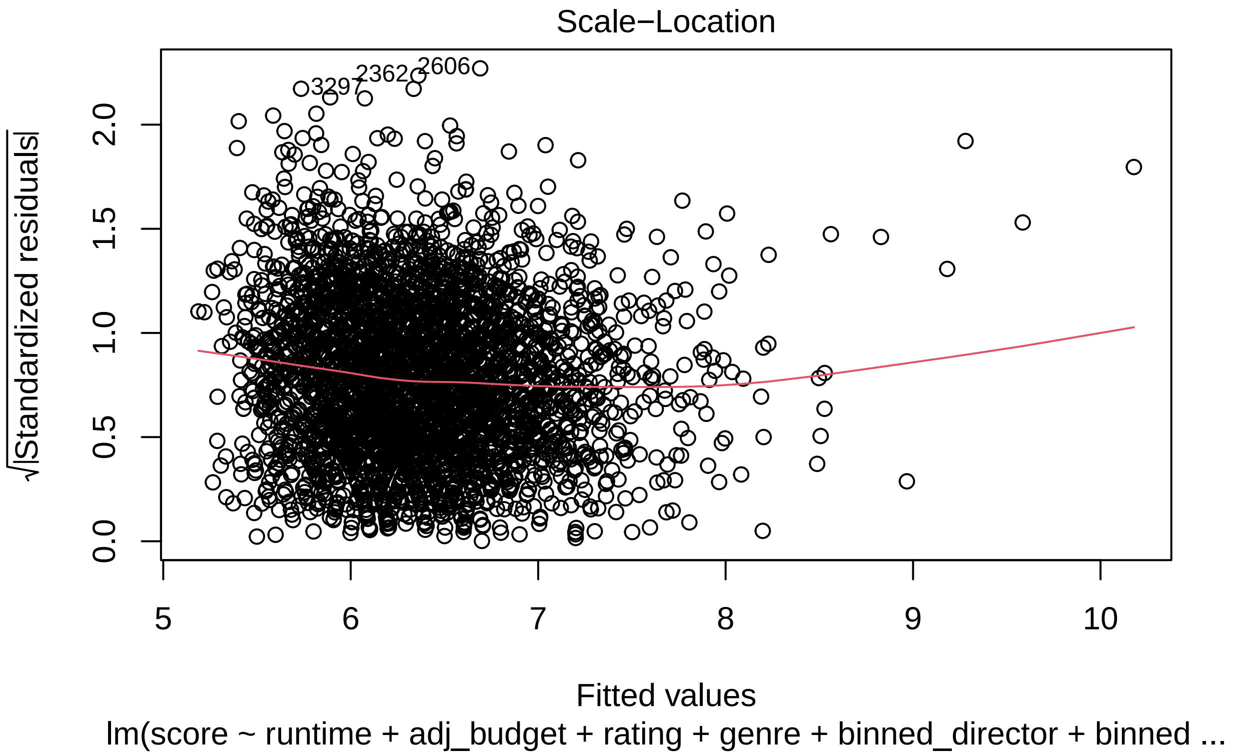
```
linregScore <- lm(score~ runtime + adj_budget + rating + genre + binned_director
             + binned_writer + binned_star + binned_company + is_remake + is_sequel, data = df)
summary(linregScore)
```

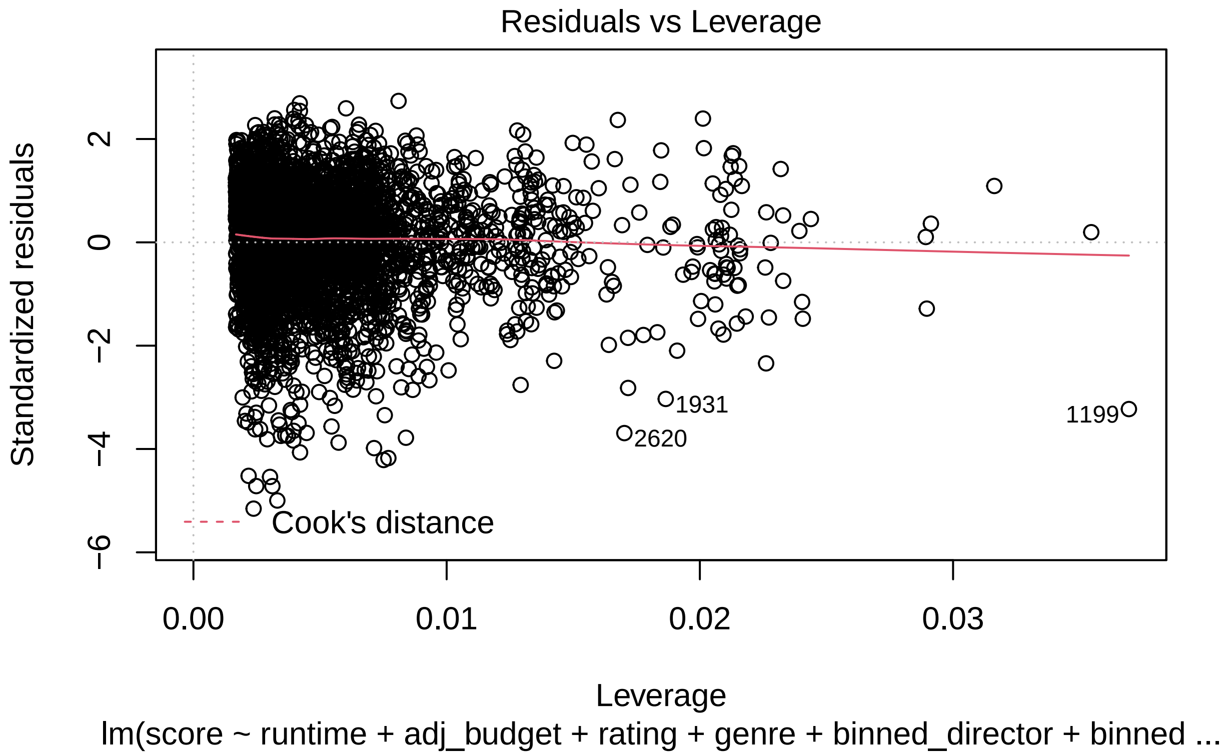
```
##
## Call:
## lm(formula = score ~ runtime + adj_budget + rating + genre +
##     binned_director + binned_writer + binned_star + binned_company +
##     is_remake + is_sequel, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.191 -0.447  0.054  0.535  2.219
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.65e+00   1.70e-01   21.43  < 2e-16 ***
## runtime              2.37e-02   9.07e-04   26.19  < 2e-16 ***
## adj_budget          -7.96e-10   3.15e-10   -2.53   0.0116 *
## ratingPG            -2.05e-01   9.97e-02   -2.06   0.0396 *
## ratingPG-13         -6.98e-02   1.05e-01   -0.67   0.5049
## ratingR              1.49e-01   1.05e-01    1.42   0.1570
## genreAction         -1.54e-02   1.16e-01   -0.13   0.8949
## genreAdventure       2.13e-01   1.27e-01    1.68   0.0931 .
## genreAnimation       1.05e+00   1.34e-01    7.85  5.1e-15 ***
## genreBiography       5.67e-01   1.26e-01    4.52  6.5e-06 ***
## genreComedy          1.07e-01   1.16e-01    0.92   0.3567
## genreCrime           2.50e-01   1.23e-01    2.03   0.0423 *
## genreDrama           3.07e-01   1.18e-01    2.60   0.0093 **
## genreHorror         -2.83e-01   1.29e-01   -2.20   0.0277 *
## binned_directorTrue -5.76e-02   2.82e-02   -2.04   0.0411 *
## binned_writerTrue    2.82e-03   3.10e-02    0.09   0.9274
## binned_starTrue      3.54e-02   2.74e-02    1.29   0.1962
## binned_companyTrue   3.23e-02   2.89e-02    1.12   0.2637
## is_remakeTrue       -9.70e-02   7.67e-02   -1.26   0.2061
## is_sequel           -1.02e-01   4.95e-02   -2.05   0.0402 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.814 on 4057 degrees of freedom
## Multiple R-squared:  0.273,  Adjusted R-squared:  0.27
## F-statistic: 80.2 on 19 and 4057 DF,  p-value: <2e-16
```

```
plot(linregScore)
```


Residuals vs Fitted
Residuals
2
0
−2
−4
5
6
7
8
9
10
Fitted values
2126
2362
2606
lm(score ~ runtime + adj_budget + rating + genre + binned_director + binned ...


Normal Q−Q
Standardized residuals
2
0
−2
−4
3297
2362
2606
−2
0
2
Theoretical Quantiles
lm(score ~ runtime + adj_budget + rating + genre + binned_director + binned ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
3297
2362
2606
5
6
7
8
9
10
Fitted values
lm(score ~ runtime + adj_budget + rating + genre + binned_director + binned ...



```
# std_errors = sqrt(diag(vcov(simpleLinReg)))
# percError <- summary(simpleLinReg)$sigma / mean(pros$lcavol) * 100
```

Robust

```
# coeftest(linregScore,vcov=vcovHC)
summaryLin = coeftest(linregScore, vcov=vcovHC)[,]
summaryLinDf = as.tibble(summaryLin)
```

```
## Warning: ‘as.tibble()‘ was deprecated in tibble 2.0.0.
## Please use ‘as_tibble()‘ instead.
## The signature and semantics have changed, see ‘?as_tibble‘.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was generated.
```

```
# coefci(linregScore,vcov=vcovHC)
summaryLinDf[, "Coefficient"] = c("(Intercept)", "Runtime", "Budget",
  "$\\text{Rating}_{\\text{PG}}$",
  "$\\text{Rating}_{\\text{PG-13}}$",
  "$\\text{Rating}_{\\text{R}}$",
  "$\\text{Genre}_{\\text{Action}}$",
  "$\\text{Genre}_{\\text{Adventure}}$",
  "$\\text{Genre}_{\\text{Animation}}$",
```

```r
  "$\\text{Genre}_{\\text{Biography}}$",
  "$\\text{Genre}_{\\text{Comedy}}$",
  "$\\text{Genre}_{\\text{Crime}}$",
  "$\\text{Genre}_{\\text{Drama}}$",
  "$\\text{Genre}_{\\text{Horror}}$",
  "$I(\\text{Experienced Director})$",
  "$I(\\text{Experienced Writer})$",
  "$I(\\text{Experienced Actor})$",
  "$I(\\text{Big 5 Production Co.})$",
  "$I(\\text{Remake})$",
  "$I(\\text{Sequel})$"
  )
ciL1 = coefci(linregScore, vcov=vcovHC)[,1]
names(ciL1) = summaryLinDf$Coefficient
summaryLinDf[,'2.5'] = ciL1
summaryLinDf[,'97.5'] = unname(coefci(linregScore, vcov=vcovHC)[,2])
```

```r
kable(data.frame("est" = summaryLinDf$Estimate,
  "SE1" = summaryLinDf$`Std. Error`,
  "2.5" = summaryLinDf$`2.5`,
  "975" = summaryLinDf$`97.5`,
  # "z" = summaryLogDf$`z value`,
  'p-val' = summaryLinDf$`Pr(>|t|)`
  ),
  col.names = c("Estimate", "Robust SE", "95% CI", "", "p-value"),
  caption = "Robust Linear Regression for IMDb Score"
)
```

Table 1: Robust Linear Regression for IMDb Score

	Estimate	Robust SE	95% CI		p-value
(Intercept)	3.64998	0.19286	3.27186	4.02810	0.00000
Runtime	0.02375	0.00115	0.02150	0.02599	0.00000
Budget	0.00000	0.00000	0.00000	0.00000	0.01410
$\text{Rating}_{\text{PG}}$	-0.20522	0.11921	-0.43894	0.02850	0.08525
$\text{Rating}_{\text{PG-13}}$	-0.06975	0.12548	-0.31576	0.17626	0.57831
Rating_{R}	0.14852	0.12565	-0.09783	0.39486	0.23729
$\text{Genre}_{\text{Action}}$	-0.01538	0.11838	-0.24747	0.21671	0.89663
$\text{Genre}_{\text{Adventure}}$	0.21271	0.12998	-0.04213	0.46755	0.10183
$\text{Genre}_{\text{Animation}}$	1.05190	0.14035	0.77673	1.32706	0.00000
$\text{Genre}_{\text{Biography}}$	0.56707	0.12364	0.32466	0.80947	0.00000
$\text{Genre}_{\text{Comedy}}$	0.10661	0.11745	-0.12365	0.33688	0.36408
$\text{Genre}_{\text{Crime}}$	0.24991	0.12427	0.00628	0.49354	0.04439
$\text{Genre}_{\text{Drama}}$	0.30739	0.12012	0.07188	0.54289	0.01053
$\text{Genre}_{\text{Horror}}$	-0.28343	0.12973	-0.53778	-0.02908	0.02897
$I(\text{Experienced Director})$	-0.05762	0.02862	-0.11373	-0.00152	0.04413
$I(\text{Experienced Writer})$	0.00282	0.03133	-0.05860	0.06425	0.92817
$I(\text{Experienced Actor})$	0.03542	0.02789	-0.01927	0.09011	0.20425
$I(\text{Big 5 Production Co.})$	0.03235	0.02805	-0.02265	0.08735	0.24893
$I(\text{Remake})$	-0.09700	0.06218	-0.21890	0.02490	0.11881
$I(\text{Sequel})$	-0.10155	0.05518	-0.20974	0.00664	0.06581

```
redModLin_genre = lm(score~ runtime + adj_budget + rating + binned_director
             + binned_writer + binned_star + binned_company + is_remake+ is_sequel, data = df)
waldtest(redModLin_genre, linregScore, vcov=vcovHC)
```

```
## Wald test
##
## Model 1: score ~ runtime + adj_budget + rating + binned_director + binned_writer +
##     binned_star + binned_company + is_remake + is_sequel
## Model 2: score ~ runtime + adj_budget + rating + genre + binned_director +
##     binned_writer + binned_star + binned_company + is_remake +
##     is_sequel
##   Res.Df Df    F Pr(>F)
## 1   4065
## 2   4057  8 42.1 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
redModLin_rating = lm(score~ runtime+ adj_budget +genre + binned_director
             + binned_writer + binned_star+ binned_company, data = df)
waldtest(redModLin_rating, linregScore, vcov=vcovHC)
```

```
## Wald test
##
## Model 1: score ~ runtime + adj_budget + genre + binned_director + binned_writer +
##     binned_star + binned_company
## Model 2: score ~ runtime + adj_budget + rating + genre + binned_director +
##     binned_writer + binned_star + binned_company + is_remake +
##     is_sequel
##   Res.Df Df    F Pr(>F)
## 1   4062
## 2   4057  5 17.1 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Regression with ROI

NOT VALID Due to to OUTLIERS

```
# linregROI <- lm(ROI~ runtime + adj_budget + rating + genre + binned_director
#               + binned_writer + binned_star + binned_company, data = df)
# summary(linregROI)
# plot(linregROI)
```

Log Reg with ROI

```
logregROI <- glm(binROI~ runtime + adj_budget + rating + genre + binned_director
             + binned_writer + binned_star + binned_company + is_remake + is_sequel, data = df, family =
# summary(logregROI)
# plot(logregROI)
```

```r
summaryLog = coeftest(logregROI, vcov=vcovHC)[,]
summaryLogDf = as.tibble(summaryLog)
```

```r
# summaryLog
# coefci(logregROI, vcov=vcovHC)[,]
```

```r
summaryLogDf[, "Coefficient"] = c("(Intercept)", "Runtime", "Budget",
  "$\\text{Rating}_{\\text{PG}}$",
  "$\\text{Rating}_{\\text{PG-13}}$",
  "$\\text{Rating}_{\\text{R}}$",
  "$\\text{Genre}_{\\text{Action}}$",
  "$\\text{Genre}_{\\text{Adventure}}$",
  "$\\text{Genre}_{\\text{Animation}}$",
  "$\\text{Genre}_{\\text{Biography}}$",
  "$\\text{Genre}_{\\text{Comedy}}$",
  "$\\text{Genre}_{\\text{Crime}}$",
  "$\\text{Genre}_{\\text{Drama}}$",
  "$\\text{Genre}_{\\text{Horror}}$",
  "$I(\\text{Experienced Director})$",
  "$I(\\text{Experienced Writer})$",
  "$I(\\text{Experienced Actor})$",
  "$I(\\text{Big 5 Production Co.})$",
  "$I(\\text{Remake})$",
  "$I(\\text{Sequel})$"
  )
ci1 = coefci(logregROI, vcov=vcovHC)[,1]
names(ci1) = summaryLogDf$Coefficient
summaryLogDf[,'2.5'] = ci1
summaryLogDf[,'97.5'] = unname(coefci(logregROI, vcov=vcovHC)[,2])
```

```r
kable(data.frame(
  # "cof"= summaryLogDf$Coefficient,
  "est"= summaryLogDf$Estimate,
  "SE" = summaryLogDf$`Std. Error`,
  "2.5" = summaryLogDf$`2.5`,
  "975" = summaryLogDf$`97.5`,
  # "z" = summaryLogDf$`z value`,
  'p-val' = summaryLogDf$`Pr(>|z|)`
  ),
  col.names = c("Estimate", "Robust SE", "95% CI", "", "p-value"),
  caption = "Robust Logistic Regression for Positive ROI Classification"
)
```

Table 2: Robust Logistic Regression for Positive ROI Classification

	Estimate	Robust SE	95% CI		p-value
(Intercept)	-0.50199	0.53955	-1.55948	0.55551	0.35217
Runtime	0.01332	0.00310	0.00725	0.01938	0.00002
Budget	0.00000	0.00000	0.00000	0.00000	0.94396
$\text{Rating}_{\text{PG}}$	0.25078	0.32501	-0.38623	0.88780	0.44035
$\text{Rating}_{\text{PG-13}}$	0.20457	0.33824	-0.45837	0.86751	0.54531

	Estimate	Robust SE	95% CI		p-value
Rating$_{\text{R}}$	-0.26915	0.33875	-0.93308	0.39479	0.42689
Genre$_{\text{Action}}$	-0.28363	0.34623	-0.96223	0.39498	0.41268
Genre$_{\text{Adventure}}$	-0.40593	0.37645	-1.14375	0.33189	0.28089
Genre$_{\text{Animation}}$	0.84363	0.43245	-0.00396	1.69121	0.05108
Genre$_{\text{Biography}}$	-0.51843	0.36905	-1.24176	0.20489	0.16009
Genre$_{\text{Comedy}}$	-0.36877	0.34393	-1.04286	0.30532	0.28362
Genre$_{\text{Crime}}$	-0.62588	0.35940	-1.33028	0.07853	0.08160
Genre$_{\text{Drama}}$	-0.58113	0.34987	-1.26687	0.10461	0.09672
Genre$_{\text{Horror}}$	0.58490	0.39387	-0.18708	1.35688	0.13755
I(Experienced Director)	-0.07216	0.08120	-0.23131	0.08698	0.37415
I(Experienced Writer)	0.15841	0.08997	-0.01792	0.33474	0.07827
I(Experienced Actor)	-0.01386	0.07613	-0.16308	0.13536	0.85555
I(Big 5 Production Co.)	0.64731	0.08722	0.47635	0.81826	0.00000
I(Remake)	0.32514	0.23413	-0.13374	0.78402	0.16492
I(Sequel)	1.61134	0.23170	1.15721	2.06546	0.00000

```r
redModLog = glm(binROI~ runtime+ adj_budget+genre+ rating+ binned_director
             + binned_writer+binned_star, data = df, family='binomial')
waldtest(redModLog, logregROI, vcov=vcovHC, test = 'Chisq')
```

```
## Wald test
##
## Model 1: binROI ~ runtime + adj_budget + genre + rating + binned_director +
##     binned_writer + binned_star
## Model 2: binROI ~ runtime + adj_budget + rating + genre + binned_director +
##     binned_writer + binned_star + binned_company + is_remake +
##     is_sequel
##   Res.Df Df Chisq Pr(>Chisq)
## 1   4060
## 2   4057  3   103     <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```