QRM II Graded Assignment (2), Period 1 2025

Material by Sjoerd van Alten and Klervie Toczé

Group 21 - Carlijn Calori, Julia Koeleman, Marit Springer & Collin Veenstra

03-10-2025

# Week 4

1. There is another argument going on in the movie studio. *Bob* claims that production budgets are getting out of hand, and that the studio should focus on making cheaper movies. *Chantal* disagrees. She tells Bob that ``Every dollar we spend on movie production is more than offset by the increase in movie profits’’.
2. Set up a regression model to test Chantal’s claim, and estimate it. That is, estimate:

* Print a summary of your estimated model.

1. What is the estimated value of and how do you interpet it?
2. Test for the null hypothesis that . Report the p-value and state your conclusion.
3. Next, estimate the model

* When creating the variables Log Profits and Log Budget, make sure that movies with a Revenue or Budget of zero are assigned the value “NA”. Print a summary of your estimated model

1. What is the estimated value of and how do you interpet it?
2. Which model has better fit? The level-level model or the log-log model? Explain.
3. Who do you think is correct? Bob or Chantal? What would you advise the movie studio to do?

*step a*

#WRITE YOUR CODE HERE  
movies <- read.table("movies1.tsv", header=TRUE)  
  
model <- lm(revenue ~ budget, data = movies)  
summary(model)

##   
## Call:  
## lm(formula = revenue ~ budget, data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -420678243 -37658344 14833690 17600279 1964947903   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.760e+07 7.165e+06 -2.456 0.0144 \*   
## budget 3.547e+00 1.302e-01 27.248 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 131700000 on 503 degrees of freedom  
## Multiple R-squared: 0.5961, Adjusted R-squared: 0.5953   
## F-statistic: 742.5 on 1 and 503 DF, p-value: < 2.2e-16

**Your Answer:**

Profit = -1,76x10^7 + 3,547 x Budget. If a movie had a budget of 0, expected profits would be about -17,6 million euros. For every 1 euro increase in budget, expected profits rise by 3,55 euro on average.

*step b*

#WRITE YOUR CODE HERE  
summary(model)

##   
## Call:  
## lm(formula = revenue ~ budget, data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -420678243 -37658344 14833690 17600279 1964947903   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.760e+07 7.165e+06 -2.456 0.0144 \*   
## budget 3.547e+00 1.302e-01 27.248 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 131700000 on 503 degrees of freedom  
## Multiple R-squared: 0.5961, Adjusted R-squared: 0.5953   
## F-statistic: 742.5 on 1 and 503 DF, p-value: < 2.2e-16

**Your Answer:**

The estimated value of Beta 1 is 3.547. This means that each additional dollar spent on production, is an increase of 3.547 dollar in profits.

*step c*

#WRITE YOUR CODE HERE  
coefs <- summary(model)$coefficients  
t\_val <- coefs["budget", "t value"]  
  
p\_value <- pt(t\_val, df = model$df.residual, lower.tail = FALSE)  
print(p\_value)

## [1] 2.135988e-101

**Your Answer:**

H0: B1 <= 0 H1: B1 > 0

Since the p-value is essentially 0 (2,1x10^-101), we reject the null hypothesis and conclude there is very strong evidence that B1 > 0. This means higher budgets are associated with significantly higher profits

*step d*

#WRITE YOUR CODE HERE  
movies$profit <- movies$revenue - movies$budget  
  
movies$LogProfit <- ifelse(movies$profit > 0, log(movies$profit), NA)

## Warning in log(movies$profit): NaNs produced

movies$LogBudget <- ifelse(movies$budget > 0, log(movies$budget), NA)  
  
model\_log <- lm(LogProfit ~ LogBudget, data = movies)  
summary(model\_log)

##   
## Call:  
## lm(formula = LogProfit ~ LogBudget, data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3205 -0.6734 0.2365 0.8935 3.6926   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.12020 1.06215 5.762 2.41e-08 \*\*\*  
## LogBudget 0.67348 0.06171 10.914 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.429 on 253 degrees of freedom  
## (250 observations deleted due to missingness)  
## Multiple R-squared: 0.3201, Adjusted R-squared: 0.3174   
## F-statistic: 119.1 on 1 and 253 DF, p-value: < 2.2e-16

**Your Answer:**

LogProfit = 6,12 + 0.67 x LogBudget.

*step e*

#WRITE YOUR CODE HERE

**Your Answer:**

The estimated value of Beta 1 is 0,67348. Since this is a log-log model, a 1% increase in budget is associated with an estimated 0,673% increase in profits, on average.

*step f*

#WRITE YOUR CODE HERE

**Your Answer:**

The level–level model fits the data better since it has a higher R^2: Level R^2 = 0.5961 Log R^2 = 0.3201

Het level–level model heeft duidelijk de beste statistische fit. Dit model is geschikt als je geïnteresseerd bent in absolute effecten. Het log–log model daarentegen heeft een lagere fit, maar is economisch vaak zinvoller omdat het in elasticiteiten interpreteerbaar is: een 1% stijging in budget gaat samen met een 0,67% hogere winst.

*step g*

#WRITE YOUR CODE HERE

**Your Answer:**

Chantal is correct in the sense that, at the margin, the studio currently gets back more than 1 dollar in profits for every extra dollar spent, because B1 is 3,547. The level–level model supports her argument directly.

Bob has a point too: the log–log model shows diminishing returns, because B1 is 0.673. So Bob’s concern is not entirely unfounded. I would advise the studio to keep investing in production budgets but to be cautious with spending too much on very large projects, focusing instead on the range of budgets where returns are strongest.

2 a. Make a plot with a 95% confidence interval with the mean log of budget on the y-axis, and whether the first actor of the movie is male or female on the x-axis. What do you conclude? b. Estimate the following simple OLS model: $log(budget)\_i=\beta\_0+\beta\_1 \text(FirstActorMale)\_i + \varepsilon\_i.$ Is the estimated coefficient for significantly different from zero? How do you interpret its estimate, and how does this relate to your conclusion in 2a? c. Now, have a close look at your data frame. Can you find any instances of male first actors who are wrongly labeled as being female, or vice versa? What would such mislabelling mean for the coefficient you estimated under 2b?

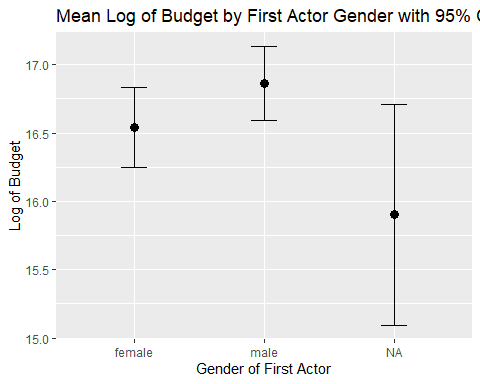
*step a*

#WRITE YOUR CODE HERE  
install.packages("Hmisc", repos = "https://cloud.r-project.org")

library (tidyverse)

library (Hmisc)

movies <- subset(movies, budget >0)  
movies$log\_budget <- log(movies$budget)  
  
ggplot(movies, aes(x = first\_actor\_gender, y = log\_budget)) +  
 stat\_summary(fun = mean, geom = "point", size = 3) +  
 stat\_summary(fun.data = mean\_cl\_normal, geom = "errorbar", width = 0.2) +  
 labs(x = "Gender of First Actor", y = "Log of Budget",  
 title = "Mean Log of Budget by First Actor Gender with 95% Confidence Interval")



**Your Answer:**

This plot shows that the average log-budget of films with a male first actor is slightly higher than that of films with a female first actor, but there is a large overlap of confidence intervals. This means there is no clear statistical difference between the groups. For the NA group, the uncertainty is so large that no conclusion can be drawn.

*step b*

#WRITE YOUR CODE HERE  
model <- lm(log\_budget ~ movies$first\_actor\_gender, data = movies,)  
summary(model)

##   
## Call:  
## lm(formula = log\_budget ~ movies$first\_actor\_gender, data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.1708 -0.4835 0.3527 1.0459 2.6171   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.5407 0.2144 77.16 <2e-16 \*\*\*  
## movies$first\_actor\_gendermale 0.3232 0.2486 1.30 0.194   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.045 on 353 degrees of freedom  
## (16 observations deleted due to missingness)  
## Multiple R-squared: 0.004767, Adjusted R-squared: 0.001948   
## F-statistic: 1.691 on 1 and 353 DF, p-value: 0.1943

percent\_effect <- (exp(0.3232)-1)\*100  
print(percent\_effect)

## [1] 38.15416

**Your Answer:**

The estimated coefficient for β1 is 0.3232 with a p-value of 0.194, so it is not significantly different from zero. This is because 0.194 > 0.05. This means we cannot conclude that movies with a male first actor have different budgets than those with a female first actor. The positive estimate would suggest about a 38% higher budget for male leads, but since it’s not significant, this difference is not reliable. This matches your conclusion in 2a: There is no significant difference in mean log-budgets between male and female first actors.

*step c*

#WRITE YOUR CODE HERE

**Your Answer:**

We have detected some wrongly labeled actors. Some examples are: Cameron Diaz, Jackie Chan and Blake Lively. Because some actors are wrongly labeled as male or female, the data contains errors. This makes the estimated coefficient for β1 smaller than it should be, pushing it closer to zero. As a result, any real difference in budgets between male and female leads becomes harder to see.

# Week 5

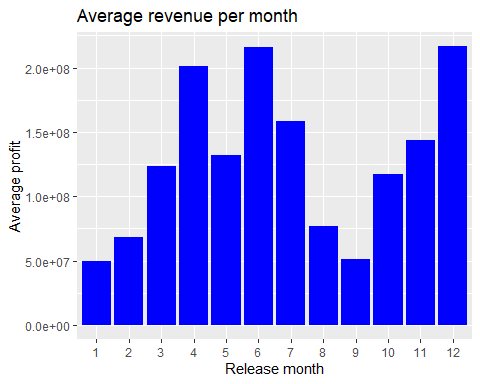
1. Create a plot of the mean profits by month of release. Do you see any indication that month of release matters to the profits of the movie?
2. Estimate an OLS model which has as dependent variable the log of profits of a movie, and as independent variable the log of budget, a dummy for whether the movie was released in english or not, and a linear term for the month of release. Show a summary of the resulting model and interpret each coefficient.
3. Test for the hypothesis that the coefficient that belongs to month of release is zero.
4. Based on your plot in a.) do you consider the choice that month of release enters the model linearly under b.) reasonable? Estimate a specification that allows for a more flexible curve. In this new specification, test for the null hypothesis that month of release does not impact profits. This might require testing multiple terms at once.
5. One executive at the studio wants to time the release of the movie to a specific month of the year such that they can maximize revenue. Based on your model under d.), What would you advise the movie studio regarding the timing of the release of the movie?

The movie studio that you work at is releasing a new movie in 2026. It will be an English-spoken Thriller movie with a budget of 40,000,0000.

1. Estimate a model that is able to predict the revenue of this movie. Give its predicted revenue and include a 99% prediction interval.

*step a*

#WRITE YOUR CODE HERE  
library(tidyverse)  
  
profits\_by\_month<- movies %>%   
 group\_by(release\_month) %>%   
 summarise(meanrevenue=mean(revenue,na.rm=T))  
ggplot(profits\_by\_month, aes(x=factor(release\_month), y=meanrevenue))+  
 geom\_col(fill="blue")+  
 labs(  
 title="Average revenue per month",   
 x="Release month",  
 y="Average profit"  
 )



**Your Answer:**

The month of release matters. It is better to release the movie in months June or December, because those generated the most profit. January and September generated the least profit, so the movie should not release in those months.

*step b*

#WRITE YOUR CODE HERE  
movies$english <- ifelse(movies$original\_language == "en", 1, 0)  
movies$english<- as.factor(movies$english)  
summary(movies)

## index budget keywords original\_language   
## Min. : 0.0 Min. : 2 Length:371 Length:371   
## 1st Qu.: 820.5 1st Qu.: 11500000 Class :character Class :character   
## Median :1689.0 Median : 27000000 Mode :character Mode :character   
## Mean :1917.0 Mean : 43141437   
## 3rd Qu.:2911.0 3rd Qu.: 55500000   
## Max. :4796.0 Max. :250000000   
##   
## title popularity release\_date revenue   
## Length:371 Min. : 0.05063 Length:371 Min. :0.000e+00   
## Class :character 1st Qu.: 9.81746 Class :character 1st Qu.:9.102e+06   
## Mode :character Median : 19.74746 Mode :character Median :5.243e+07   
## Mean : 29.53585 Mean :1.286e+08   
## 3rd Qu.: 38.90758 3rd Qu.:1.456e+08   
## Max. :203.73459 Max. :2.788e+09   
##   
## runtime vote\_average vote\_count genre   
## Min. : 47.0 Min. :0.000 Min. : 0.0 Length:371   
## 1st Qu.: 97.0 1st Qu.:5.700 1st Qu.: 189.5 Class :character   
## Median :105.0 Median :6.200 Median : 505.0 Mode :character   
## Mean :109.5 Mean :6.182 Mean : 1066.3   
## 3rd Qu.:122.8 3rd Qu.:6.800 3rd Qu.: 1266.5   
## Max. :178.0 Max. :8.100 Max. :11800.0   
## NA's :1   
## release\_year release\_month release\_day first\_actor   
## Min. :1990 Min. : 1.000 Min. : 1.00 Length:371   
## 1st Qu.:2002 1st Qu.: 4.000 1st Qu.: 9.00 Class :character   
## Median :2008 Median : 8.000 Median :15.00 Mode :character   
## Mean :2006 Mean : 7.213 Mean :15.43   
## 3rd Qu.:2012 3rd Qu.:10.000 3rd Qu.:22.00   
## Max. :2016 Max. :12.000 Max. :31.00   
##   
## first\_actor\_gender director\_first\_name director\_gender profit   
## Length:371 Length:371 Length:371 Min. : -90000000   
## Class :character Class :character Class :character 1st Qu.: -2050292   
## Mode :character Mode :character Mode :character Median : 24281723   
## Mean : 85471108   
## 3rd Qu.: 94024068   
## Max. :2550965087   
##   
## LogProfit LogBudget log\_budget english  
## Min. :10.86 Min. : 0.6931 Min. : 0.6931 0: 16   
## 1st Qu.:16.89 1st Qu.:16.2579 1st Qu.:16.2579 1:355   
## Median :17.91 Median :17.1113 Median :17.1113   
## Mean :17.67 Mean :16.7431 Mean :16.7431   
## 3rd Qu.:18.79 3rd Qu.:17.8319 3rd Qu.:17.8319   
## Max. :21.66 Max. :19.3370 Max. :19.3370   
## NA's :116

movies$LogRevenue <- ifelse(movies$revenue > 0, log(movies$revenue), NA)  
  
movies<- movies[movies$profit>0&movies$budget >0,]  
0

## [1] 0

model<- lm(LogRevenue ~ LogBudget + english + release\_month, data=movies)  
summary(model)

##   
## Call:  
## lm(formula = LogRevenue ~ LogBudget + english + release\_month,   
## data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6896 -0.5364 -0.0271 0.4738 3.4615   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.75592 0.61383 9.377 <2e-16 \*\*\*  
## LogBudget 0.74907 0.03251 23.040 <2e-16 \*\*\*  
## english1 -0.34898 0.34003 -1.026 0.306   
## release\_month 0.01092 0.01381 0.791 0.430   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7451 on 251 degrees of freedom  
## Multiple R-squared: 0.6821, Adjusted R-squared: 0.6783   
## F-statistic: 179.5 on 3 and 251 DF, p-value: < 2.2e-16

**Your Answer:**

LogRevenue = 5,75592 + 0,749 x LogBudget - 0,349 x English + 0.01 x Release Month

An increase of 1% in budget, increases the revenue with 0.749%. English movies have an estimated 34,9% lower revenue than non-English movies. For each month later in the year, the revenue is 1.1% higher if the rest is constant.

*step c*

#WRITE YOUR CODE HERE  
model<- lm(LogRevenue~LogBudget + english+release\_month, data=movies)  
summary(model)

##   
## Call:  
## lm(formula = LogRevenue ~ LogBudget + english + release\_month,   
## data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.6896 -0.5364 -0.0271 0.4738 3.4615   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.75592 0.61383 9.377 <2e-16 \*\*\*  
## LogBudget 0.74907 0.03251 23.040 <2e-16 \*\*\*  
## english1 -0.34898 0.34003 -1.026 0.306   
## release\_month 0.01092 0.01381 0.791 0.430   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7451 on 251 degrees of freedom  
## Multiple R-squared: 0.6821, Adjusted R-squared: 0.6783   
## F-statistic: 179.5 on 3 and 251 DF, p-value: < 2.2e-16

**Your Answer:** H0: B3 = 0 H1: B3 ≠ 0

If you look at the release month, the t value is 0.791 and the p value is 0.430. The p value is much larger than 0.05, so we fail to reject the null hypothesis at the 5% significance level. Thus, there is no statistically significant evidence that the release month affects movie revenues, keeping other variables constant.

*step d*

#WRITE YOUR CODE HERE  
  
model\_flex <- lm(LogRevenue ~ LogBudget + english + as.factor(release\_month), data = movies)  
summary(model\_flex)

##   
## Call:  
## lm(formula = LogRevenue ~ LogBudget + english + as.factor(release\_month),   
## data = movies)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.48880 -0.48711 -0.06171 0.49089 3.13405   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.39047 0.63612 10.046 <2e-16 \*\*\*  
## LogBudget 0.72103 0.03357 21.476 <2e-16 \*\*\*  
## english1 -0.32491 0.34373 -0.945 0.3455   
## as.factor(release\_month)2 -0.33618 0.26883 -1.251 0.2123   
## as.factor(release\_month)3 -0.22362 0.26601 -0.841 0.4014   
## as.factor(release\_month)4 -0.01065 0.27963 -0.038 0.9696   
## as.factor(release\_month)5 -0.11659 0.27723 -0.421 0.6745   
## as.factor(release\_month)6 0.16234 0.25773 0.630 0.5294   
## as.factor(release\_month)7 0.06813 0.25087 0.272 0.7862   
## as.factor(release\_month)8 -0.43928 0.26708 -1.645 0.1013   
## as.factor(release\_month)9 -0.41257 0.24374 -1.693 0.0918 .   
## as.factor(release\_month)10 -0.08024 0.25673 -0.313 0.7549   
## as.factor(release\_month)11 0.06066 0.25740 0.236 0.8139   
## as.factor(release\_month)12 0.08875 0.24669 0.360 0.7193   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.733 on 241 degrees of freedom  
## Multiple R-squared: 0.7046, Adjusted R-squared: 0.6887   
## F-statistic: 44.22 on 13 and 241 DF, p-value: < 2.2e-16

anova(model, model\_flex)

## Analysis of Variance Table  
##   
## Model 1: LogRevenue ~ LogBudget + english + release\_month  
## Model 2: LogRevenue ~ LogBudget + english + as.factor(release\_month)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 251 139.37   
## 2 241 129.48 10 9.8828 1.8394 0.05467 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Your Answer:**

There are strong seasonal peaks, which causes a non-linear relationship. Therefore a simple linear trend of: later month leads to higher profits, is incorrect. So the assumption in part b that month enters linearly is not reasonable.

In the new model the reference category is January (month 1). All other coefficients are differences in log revenue relative to January, holding budget and language constant.

The hypothesis test gives: F = 1,84 p-value = 0.055

Since the p-value is just above 0.05, we fail to reject the null at the 5% level. This means we cannot conclude that release month has a statistically significant effect on revenues. However, at the 10% level, we would reject the null, suggesting weak evidence that release month might matter. Overall, the evidence for a month effect is marginal, but not strong.

*step e*

#WRITE YOUR CODE HERE

**Your Answer:**

The joint test (p = 0.055) provides only weak evidence that release month matters. Most months show coefficients close to zero with large p-values, suggesting no effect. Only August and September display relatively large negative coefficients and marginal significance (around 10% interval), hinting at lower revenues in late summer. The studio should not rely on month-timing to boost revenues, although they may want to be cautious about late-summer releases (August/September).

*step f*

#WRITE YOUR CODE HERE  
  
new\_movie <- data.frame(  
 LogBudget = log(40000000),  
 english = factor(1, levels = levels(movies$english)),  
 release\_month = factor(1, levels = 1:12)  
)  
  
prediction <- predict(model\_flex, newdata = new\_movie,  
 interval = "prediction", level = 0.99)  
  
exp(prediction)

## fit lwr upr  
## 1 130475964 18055222 942883832

**Your Answer:**

The predicted revenue is about 130 million dollar. The 99% prediction interval is very wide reaching from about 18 million dollar up to 943 million dollar.