Met4 Home Exam
Movies: Blockbusters and Flops

Dept. of Business and Management Science

15.11.2017 09:00 - 17.11.2017 14:00

1 Introduction

In this case, we are interested in costs and revenues from movies. How are costs and revenues related? Can we say something about the expected return from movies? And what about risk - do some genres have a higher variability of returns than others?

The Movie Database[1] (TMDB) is a large online database of movies, containing a lot of information about each movie. Movies are classified by genres, and each movie may by tagged in several genres. Further, users on TMBD may rate movies on a scale from 1 to 10, where a higher value indicates a more positive view on the movie.

In this home exam, we will use a subset of the data from TMBD. Firstly, we will only consider movies released in 1990 or later years, and secondly only movies where the first genre is either *"Action", "Adventure", "Comedy", "Drama"* or *"Horror"*. See the appendix for an overview of the variables in the dataset.

Although this is not the primary focus of this home exam, keep in mind the data quality may be low, and therefore that there might be some errors in the dataset. Please see TMDB if you need any additional information on the variables. The dataset is distributed in two files: `data.csv` and `data.RData`. These files are identical except for the formatting. Please use the file most suited for your statistical software.

[1] `https://www.themoviedb.org/`

2 Assignments

1. Present and *briefly* discuss relevant statistics of the data set.

2. Use OLS to regress `log(revenue)` on `log(budget)`, and answer the following questions:

   (a) Briefly discuss your results.

   (b) Are there any exceptional outliers, and if so - which ones? Are some genres more likely to be outliers?[2]

   (c) Are your OLS-estimates sensitive to the presence of the outliers?

3. A risk averse investor is considering buying shares in a movie that has already been made, but not yet been released to the public. The return on investment for the investor is given by:

$$\texttt{r} = \frac{.5 * \texttt{revenue} - \texttt{budget}}{\texttt{budget}} \tag{1}$$

   where `revenue` and `budget` are variables in your dataset. The movie in question has a budget of \$10 000 000, and in a test screening viewers gave the movie on average 7.5 points with a similar measure as the variable `vote_average`. What is the expected return, and what is the variance of the investment?[3] What additional information could potentially improve your estimates?

4. The dataset contains more variables and may provide insights beyond the first three assignments. Use maximum one page to present and discuss one additional finding from the dataset.

[2] *Note: see appendix B for some hints!*

[3] *Note: see appendix C for some hints!*

A Variables in the dataset

The following table lists all the variables in the data set:

Variable name	Definition
budget	Budgeted production costs of movie in million USD
revenue	Box office revenues in million USD
original_language	Original language of movie
original_title	Title of movie in original language
runtime	Runtime of movie in minutes
title	Movie title in English
vote_average	Average vote by users on TMBD
vote_count	Number of user votes for a movie on TMBD
month	Month movie was released
year	Year movie was released
genre	Main genre of movie
prod_comp	Main production company

B Largest entries of a variable

B.1 With R

With R, if `df` is a data frame, the command:
`large.values <- tail(order(abs(df$Variable), na.last = F), n=10)`
creates a collection in `large.values` with the *indices* of the 10 largest, absolute values of `df$Variable`, ignoring missing values. Hence, issuing the following command
`df[large.values,]`
returns the rows of `df` with the ten largest absolute values of `Variable`.

B.2 With Gretl (software used in earlier versions of Met4)

The following commands can be used in Gretl, if we want to sort the dataset by the absolute values of `Variable`:

```
absVariable = abs(Variable)
dataset sortby absVariable
smpl missing(absVariable)==0 --restrict
```

The data set is now sorted by the absolute values of `Variable`, ignoring missing values.

C Variance of predictions with OLS

C.1 With R

If we estimate a model in R with OLS, e.g.
`reg <- lm(Y ~ X, data=data.df)`
and have a data frame `data.new` where we want to make predictions:
`pred <- predict.lm(reg, data.new, se.fit=TRUE)`
We can use `pred` to find the variance of the *expected value of Y*:
`var.conf <- pred$se.fit^2`
If we want the variance of a predicted, single value of Y, we can find this with:
`var.pred <- var.conf+pred$residual_scale^2`

C.2 With Gretl (software used in earlier versions of Met4)

Assuming you have subset the data set to the observations you want to use for running the regression,[4] you run OLS with the command:
`ols Y 0 X`
We first run the following command to activate the whole dataset. This allows us to create predictions not only to the observations used for estimating the model:
`smpl full`
Thereafter, you generate predictions with
`fcast --mean-y`
The third column of the output displays the standard error of the predicted, expected value of Y, and squaring it gives the variance. To obtain the variance of the predicted value of a single observation of Y, we use
`fcast`
where again we square the third column to get the variance.

[4] E.g. the command `smpl 1 10` subsets to the first 10 observations

Regulations for the home exam

The home exam in Met4 must be handed in by groups of sizes 2, 3 or 4 students. See in particular section 9 the REGULATIONS FOR EXAMINATIONS AT NHH, as well as part 2 of SUPPLEMENTARY REGULATIONS TO THE REGULATIONS FOR EXAMINATIONS (FULL-TIME PROGRAMMES. It is not permitted to discuss the exam with students not in your group after the dataset has been released.

Grading and formal requirements

The reports will be graded in accordance with the assessment rubric posted on ItsLearning as well as in table 1. You may write your report in Norwegian or English.

If there is a need for clarifications, you may send an email to Ole-Petter Moe Hansen (s9705@nhh.no) *and* Håkon Otneim (Hakon.Otneim@nhh.no). Any extra information will be announced to all groups on It'sLearning.

The report should be maximum 10 pages. Tables figures and references are included in the ten pages. If the report has a front page without any answers to the assignments, this may come in addition to the 10 pages. Prioritize what you include in the report!

The report must typeset with Times New Roman, with font size 12 and line spacing 1.15. Text in tables and figures may be set as low as font size 9.

The exam is handed in through Wiseflow. We accept a single file of type pdf for submission - MS Word files or r-scripts are not accepted.

Met4 Fall 2017: Rubric for grading home exam

Points:	3	2	1	0	Weight
Presentation of tables and figures	Tables and figures i: are self-explanatory, ii: give a meaningful contribution to the results of the report, and iii: are visually appealing. The report demonstrates independence in the choice of graphics.	Tables and figures i: are self-explanatory, ii: give meaningful contribution to the results of the report, and iii: are visually appealing.	Tables and figures only partially i: are self-explanatory, ii: give meaningful contribution to the results of the report, and iii: are visually appealing.	Tables and figures are difficult to understand, and do not contribute to the results of the report.	0.10
Choice of methods	The methods chosen are appropriate for the problem at hand. The report explains which choices and assumptions are used. The report demonstrates great degree of independence in the choice of methods.	The methods chosen are appropriate for the problem at hand. The report explains which choices and assumptions are used.	The methods may be appropriate, but the justification for choice of methods is lacking.	The methods chosen are not appropriate for the problem at hand. Assumptions and choices are not explained. The report displays lack of independence in the choice of methods.	0.30
Application of methods	The methods are skillfully implemented, without any errors.	The methods are skillfully implemented, with only minor errors .	There are some errors in the implementation, but the the results are overall correct.	There are several errors made in the application of the methods.	0.25
Discussion of findings	Demonstates an outstanding understanding of both the statisticial and economical significance of results. Results are discussed in a relevant context, e.g. policy presciptions, discussion of causality, contrast to theoretical results, as applicable.	Demonstrates an good understanding of both the statisticial and economical significance of results. Results are discussed in a relevant context, e.g. policy presciptions, discussion of causality, contrast to theoretical results, as applicable.	Demonstates an some understanding of both the statisticial and economical significance of results. Results are to a limited extent discussed in a relevant context.	Lists statistical results without demonstrating an understanding of their meaning. Does not highlight the economic importance of the results.	0.30
Ethics	Discusses the ethical implications of the analysis (if applicable).			Ignores the ethical implications of the analysis (if applicable).	0.05
Each divergence from the formal requirements of the report (e.g. page lenght, font..) gives -1 points from the overall score.					

Table 1: We calculate the the overall score for a submission by summing together the points for each row, using the weights in the rightmost column. The cutpoints between grades will be set after the exam.

Solution Proposal

Assignment 1

Table 2 shows summary statistics for the numerical variables in the dataset. Other descriptive statistics, such as histograms or tables with frequencies over categorical variables, could also be relevant.

Table 2: Summary statistics over the numerical variables in the data set.

Statistic	N	Mean	St. Dev.	Min	Max
budget	2,488	39.792	43.517	0.008	380.000
revenue	2,230	118.037	186.196	0.002	2,787.965
runtime	3,216	106.877	20.037	0	276
vote_average	3,217	6.047	1.071	0.000	10.000
vote_count	3,217	691.994	1,265.847	0	13,752
month	3,217			1	12
year	3,217			1990	2017

Assignment 2

Table 3 shows the baseline regression result in column (1), showing that there is almost a one-for-one relationship between `log(budget)` and `log(revenue)`. However, there is considerable heteroskedasticity, as shown in figure C.2.

The ten largest outliers from column (1) in table 3 are shown in table 4. As we see, outliers are present in all categories, but only horror movies have revenues far over budgets.

Finally, we can note that dropping these outliers do not influence the results in the initial regression all that much, as shown in column (2) in table 3.

Table 3: Assignment 2 regressions. Column (1) is on the full dataset, column (2) excludes the ten largest outliers from column (1).

	Dependent variable:	
	`log(revenue)`	
	(1)	(2)
`log(Budget)`	0.932^{***}	0.949^{***}
	(0.024)	(0.022)
Constant	0.786^{***}	0.748^{***}
	(0.081)	(0.077)
Observations	2,125	2,115
R^2	0.425	0.459
Adjusted R^2	0.424	0.458
Residual Std. Error	1.396 (df = 2123)	1.312 (df = 2113)
F Statistic	$1{,}566.366^{***}$ (df = 1; 2123)	$1{,}789.393^{***}$ (df = 1; 2113)
Note:		*p<0.1; **p<0.05; ***p<0.01

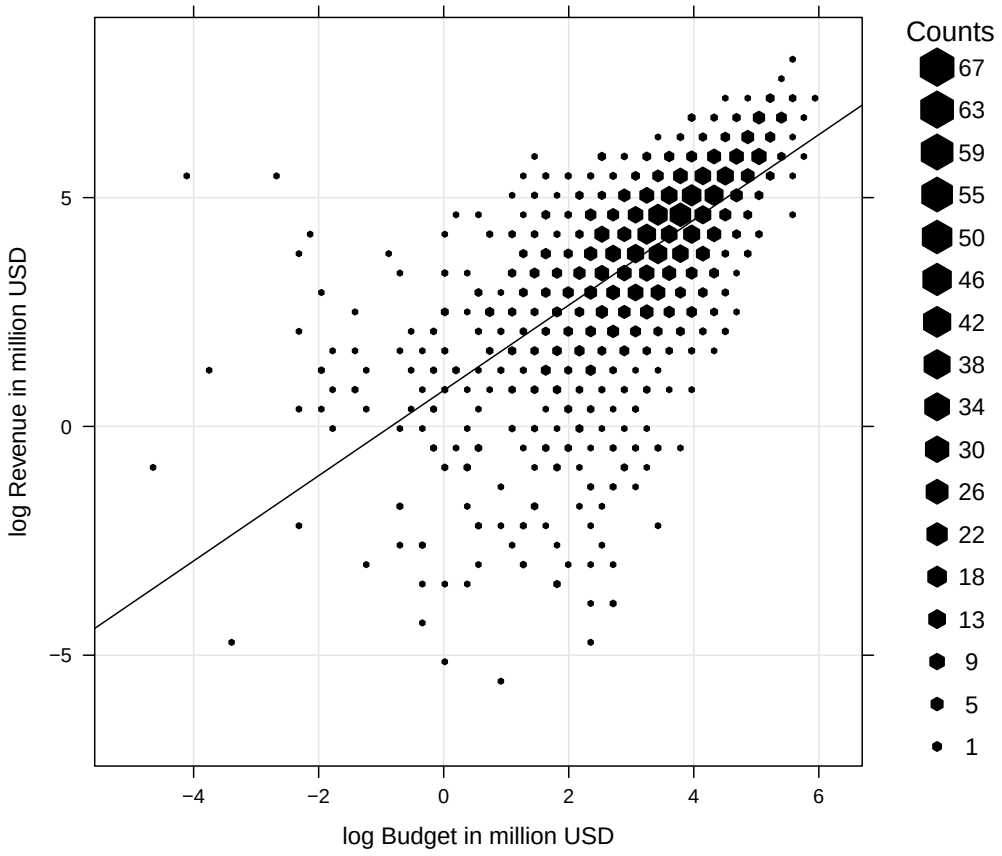

Figure 1: Scatterplot over `log(budget)` and `log(revenue)`, with the data pooled in bins (the "counts" show how many observations there are in each bin). The straight line is the regression line, coming from column(1) in table 3.

Table 4: The ten largest outliers from column (1) in table 3

Budget, mill $	Revenue Mill $	Title	Genre
15	0.023	All The Queen's Men	Action
30	0.108	1911	Adventure
10	0.046	An Alan Smithee Film: Burn, Hollywood, Burn	Comedy
15	0.020	The Good Night	Comedy
10	0.009	Janky Promoters	Comedy
14	0.046	Margaret	Drama
10	0.017	Strangerland	Drama
2.100	0.003	Mi America	Drama
0.060	248	The Blair Witch Project	Horror
0.015	193.356	Paranormal Activity	Horror

Assignment 3

There are several ways to answer this assignment. One way of solving it is to estimate the relationship between `r` and the two variables where we have data for this movie, and use the regression to predict the expected return and variance for the movie in question.

Table 5 shows the results, both in total and by genre. We can note that horror movies are very different than the other categorie, and that R^2 increase by a factor of around 3-8 when we estimate the model by genres separately. Hence, knowing the genre increases the fit of the model considerably.

Finally, table 6 shows expected returns and variance for the movie in question, where we can see that both statistics vary considerable across the categories. We don't know how the investor weighs expected returns against variance, so we cannot make strong claims about which categories it is worthwhile investing in. However, Drama movies have both a lower expected return and higher variance than Comedies, so a Comedy should be preferred over a Drama.

Horror movies have very large variance, so the investor would have to be close to risk neutral to want to invest here. However, a good answer will note that the extreme results of horror movies are driven a few movies with small budgets and very large revenues. Hence, it may be very unlikely that those extreme returns are relevant for a movie with a budget of 10 million. Dropping movies with a budget

smaller than e.g. 1 million makes the results on horror movies look a lot more like the other genres.

Table 5: Regression results on return by genre

	Dependent variable:					
	$\texttt{r} = \frac{.5*\texttt{revenue}-\texttt{budget}}{\texttt{budget}}$					
	All	Action	Adventure	Comedy	Drama	Horror
	(1)	(2)	(3)	(4)	(5)	(6)
vote_average	−1.467	0.470***	0.643***	0.977***	0.468	2.350
	(3.661)	(0.051)	(0.112)	(0.227)	(0.406)	(57.019)
log(budget)	−18.694***	0.022	−0.101	−1.365***	−1.994***	−186.343***
	(2.446)	(0.044)	(0.103)	(0.156)	(0.257)	(28.605)
Constant	73.914***	−2.742***	−3.099***	−0.780	3.662	491.918
	(24.457)	(0.341)	(0.799)	(1.514)	(2.775)	(343.292)
Observations	2,125	521	233	577	644	150
R^2	0.027	0.142	0.125	0.159	0.087	0.225
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 6: Expected return and variance for the movie with 7.5 vote average and budget of 10 mill $, total and by genre.

	All	Action	Adventure	Comedy	Drama	Horror
Expected Return	19.87	0.83	1.49	3.40	2.58	80.48
Variance	21,069.78	0.98	1.91	17.30	78.17	249,930.50

Assignment 4

This is an opportunity for students to show of their skills. However, given that there is much work to be done in the first three assignments, the requirements here are not very high.

Solution proposal - R-Script

```
##  __  __           _    _  _
## |  \/  | ___| |_| || |
## | |\/| |/ _ \ __| || |_
## | |  | |  __/ |_|__   _|
## |_|  |_|\___|\__|  |_|
##  ____            _        _    _
## / ___|  ___ | |_    _| |_(_) ___   _ __
## \___ \ / _ \| | | | | __| |/ _ \ '_ \
##  ___) | (_) | | |_| | |_| | (_) | | | |
## |____/ \___/|_|\__,_|\__|_|\___/|_| |_|
##  ____                                                _
## |  _ \ _ __ ___  _ __    ___  ___  __ _| |
## | |_) | '__/ _ \| '_ \ / _ \/ __|/ _` | |
## |  __/| | | (_) | |_) | (_) \__ \ (_| | |
## |_|   |_|  \___/| .__/ \___/|___/\__,_|_|
##
##
##
## Preamble ------------------

# Start with a clean slate:
rm(list=ls())

# Load in libraries:
library(stargazer)
library(hexbin)

# Read in the dataset:
df <- read.csv2("data.csv")[,-1]


## Assignment 1 - summary statistics ------------------

stargazer(df, type="text")


## Assignment 2 - Regression ------------------

# First, create the log transformed variables:
df$lrev <- log(df$revenue)
df$lbud <- log(df$budget)

# Next, run the regression:
```

```
reg <- lm(lrev~lbud, data=df)

# Next, want to identify the largest outliers. First, calculate
# residuals for each observation:
df$residuals <- df$lrev-predict.lm(reg, df)

# Next, find a list with the ten largest outliers. Use absolute
# deviations from the regression line:
outlier.list <- tail(order(abs(df$residuals), na.last = F),n=10)

# Create a small dataframe with the largest outliers
df.outliers  <- df[outlier.list, c(1,2,6,11)]
df.outliers  <- df.outliers[order(df.outliers$genre),]

# Print list of the largest outliers:
stargazer(df.outliers, summary=F,rownames = F, type="text")

# Re-estimate the regression without the 10 outliers:
reg.no.outliers <- lm(lrev~lbud, data=df,
                      subset=-c(outlier.list))

# Compare with and without outliers:
stargazer(reg, reg.no.outliers, type="text")

# We can also show this with a plot:
pdf("plot.pdf")
hexbinplot(df$lrev~df$lbud,
           inv=function(x) x,
           type=c('g','r'),
           xlab="log Budget in million USD",
           ylab="log Revenue in million USD",
           style="lattice")
dev.off()


## Assignment 3 - Expected return  ----------------

# First, create the variable with returns:
df$r <- (df$revenue*.5 - df$budget)/(df$budget)

# Next, regress returns on vote average and log(budget).
# Do this for the entire dataset, as well as for each
# genre separately:
eq   <- formula(r~vote_average+lbud)
reg  <- lm(eq, data=df                              )
```

```
reg1 <- lm(eq, data=df, subset=genre=="Action"     )
reg2 <- lm(eq, data=df, subset=genre=="Adventure" )
reg3 <- lm(eq, data=df, subset=genre=="Comedy"     )
reg4 <- lm(eq, data=df, subset=genre=="Drama"      )
reg5 <- lm(eq, data=df, subset=genre=="Horror"     )

# Compare the results:
stargazer(reg,
          reg1,
          reg2,
          reg3,
          reg4,
          reg5,
          type="text",
          column.labels = c("All",
                            "Action",
                            "Adventure",
                            "Comedy",
                            "Drama",
                            "Horror")
          )

# Next, want to predict expected utility. Create a data frame
# with a single observation, with values equal to
# the movie the investor is considering:
predframe <- data.frame(vote_average=7.5, lbud=log(10))

# create a prediction list for each of the regressions:
p  <- predict.lm(reg , predframe, se.fit = T)
p1 <- predict.lm(reg1, predframe, se.fit = T)
p2 <- predict.lm(reg2, predframe, se.fit = T)
p3 <- predict.lm(reg3, predframe, se.fit = T)
p4 <- predict.lm(reg4, predframe, se.fit = T)
p5 <- predict.lm(reg5, predframe, se.fit = T)

# Finally, we calculate expected return and variance in
# each of the cases:
exp.ret <- matrix(NA, ncol=6,nrow=2)
colnames(exp.ret) <- c("All",
                       "Action",
                       "Adventure",
                       "Comedy",
                       "Drama",
                       "Horror")
rownames(exp.ret) <- c("Expected Return",
```

```
                        "Variance")
exp.ret[1,1] <- p $fit
exp.ret[1,2] <- p1$fit
exp.ret[1,3] <- p2$fit
exp.ret[1,4] <- p3$fit
exp.ret[1,5] <- p4$fit
exp.ret[1,6] <- p5$fit
exp.ret[2,1] <- p $se.fit^2+p $residual.scale^2
exp.ret[2,2] <- p1$se.fit^2+p1$residual.scale^2
exp.ret[2,3] <- p2$se.fit^2+p2$residual.scale^2
exp.ret[2,4] <- p3$se.fit^2+p3$residual.scale^2
exp.ret[2,5] <- p4$se.fit^2+p4$residual.scale^2
exp.ret[2,6] <- p5$se.fit^2+p5$residual.scale^2

# Print out expected returns. As we see, the investor
# has highest expected returns from comedies. Horrors movies
# have an extreme variance.
stargazer(exp.ret, type="text", digits = 2)


## The End
```

solution_proposal.R