SLR

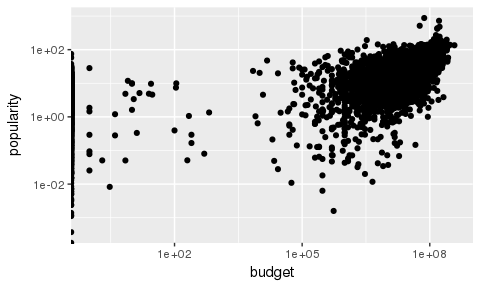
Maryann Zhao and Leyla Akay

## Introduction

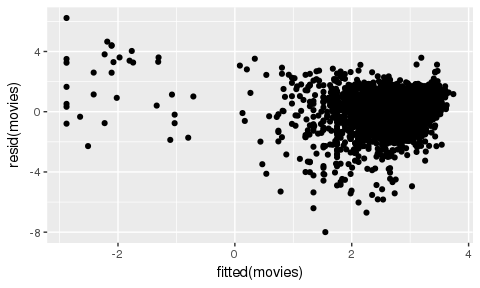
We are studying a dataset of movies, that contains information about their languages, actors, genres, etc. For this project, our variables of interest are budget (in US Dollars) and popularity (based on number of votes per day, and number of views and favorites). We are interested in determining whether there is a linear relationship between the two, and if so, the nature of the relationship. Our null hypothesis is that there is no relationship between budget and popularity; our alternative hypothesis is that there is a positive relationship between budget and popularity. That is, .

To test this, first we must check that the data appears to fit with the technical assumptions necessary to perform a linear regression. The first is linearity. An initial plot of budget values vs. popularity indicates a linear shape. We transformed both explanatory and response variables with a log function, and noticed that the plot became more linear, so we decided to keep the transformation for the remaining tests. The second is independence--we are assuming mathematical independence. That is, spending more money doesn't necessarily translate to increased popularity, or vice versa. The final assumptions are normality and constant errors. To check this, we plotted the residuals and looked at the symmetry. The residuals appear to be symmetrically distributed around 0, with some increased scattering at lower values of budget. The residuals appear more concentrated around 0, and are all in the range of +/- 4.

#plot of log popularity vs. log budget  
ggplot(tmdb2, aes(x=budget, y=popularity)) +  
 geom\_point() +  
 scale\_x\_log10() +  
 scale\_y\_log10()



#residual plot  
bud2 <- tmdb2 %>%   
 select(popularity, budget) %>%  
 filter(budget > 0, popularity > 0)  
pop.log <- log(bud2$popularity)  
bud.log <- log(bud2$budget)  
movies <- lm(pop.log ~ bud.log, data = tmdb2)  
ggplot (movies, aes(x=fitted(movies), y=resid(movies))) + geom\_point()

 ##Testing To determine whether the alternative hypothesis holds, we found a 95% confidence interval for . We found the CI was between 0.317 and 0.353. Because both variables had been logged-transformed, this means that a doubling of budget would be consistent with a multiplicative change in median of popularity between: , which is consistent with 1.25 and 1.28.

We were interested in movies with 1M$ budgets, since these are relatively low compared to many block-buster films, but many appear to achieve comparable levels of popularity. In this situation, we thought of budget as the explanatory variable, since it's determined before popularity, so although budget doesn't necessarily explain popularity, it wouldn't be possible for popularity to inform budget.

#CI and PI for 1 mill budget  
broom::tidy(movies, conf.int = TRUE, conf.level = 0.95)

## term estimate std.error statistic p.value conf.low conf.high  
## 1 (Intercept) -2.881 0.15287 -18.8 8.99e-76 -3.180 -2.581  
## 2 bud.log 0.335 0.00916 36.6 3.51e-251 0.317 0.353

mill <- data.frame(bud.log=c(13.8))  
crit\_val <- qt(.95, glance(movies)$df.resid)  
movie.pred <- broom::augment(movies, newdata=mill, type.predict = "response")  
.se.pred <- sqrt(glance(movies)$sigma^2 + movie.pred$.se.fit)  
  
movie.pred <- movie.pred %>%  
mutate(lower\_CI = .fitted - crit\_val \* .se.fit,  
upper\_CI = .fitted + crit\_val \* .se.fit)  
  
movie.pred <- movie.pred %>%  
 mutate(lower\_PI = .fitted - crit\_val \* .se.pred,  
 upper\_PI = .fitted + crit\_val \* .se.pred,   
 lower\_CI = .fitted - crit\_val \* .se.fit,   
 upper\_CI = .fitted + crit\_val \* .se.fit)  
  
movie.pred

## bud.log .fitted .se.fit lower\_CI upper\_CI lower\_PI upper\_PI  
## 1 13.8 1.74 0.0312 1.69 1.79 -0.106 3.59

## Assessing the fit

The r-squared value is 0.262. The residuals (plotted above) appear to be evenly distributed. Therefore, we believe our model accurately describes the data.

#R squared value  
summary(movies)$r.squared

## [1] 0.262

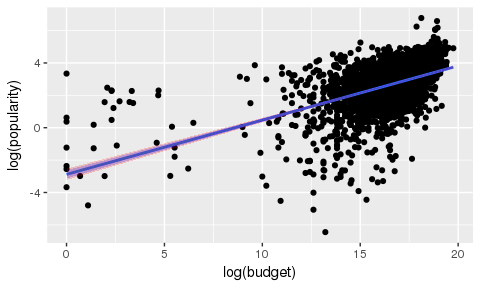
## Conclusion

In general, the positive relationship was expected. The decreased variability at very high budgets was interesting, but upon further reflection it makes sense. Movies that spend tens of millions of dollars probably attract high-profile actors and have lots of advertising campaigns, which probably increase the popularity.

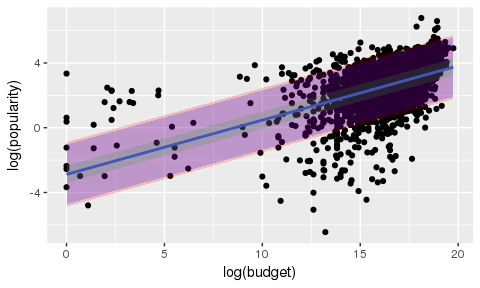
## Pairs: Simultaneous Inference

To find the mean and prediction intervals for all n values, we used the Bonferroni, Working-Hotelling, and Scheffé methods.

#critical values for 3 methods  
num\_int <- 3  
n <- 3766  
crit\_Bonf <- qt((1-.05)/num\_int, glance(movies)$df.resid)  
crit\_WH <- sqrt(2\*qf(.95, num\_int, glance(movies)$df.resid))  
crit\_Sch <- sqrt(num\_int\*qf(.95, n-2, glance(movies)$df.resid))  
  
#mean intervals for all n points  
movie\_gl <- broom::glance(movies)  
movie\_sig <- dplyr::pull(movie\_gl, sigma)  
  
movie\_CI <-broom::augment(movies) %>%  
 mutate(.se.pred = sqrt(movie\_sig^2 + .se.fit^2)) %>%  
 mutate(lower\_CI = .fitted - crit\_val \* .se.fit,  
 upper\_CI = .fitted + crit\_val \* .se.fit,   
 lower\_CI\_B = .fitted - crit\_Bonf\*.se.fit,  
 upper\_CI\_B = .fitted + crit\_Bonf\*.se.fit,   
 lower\_CI\_WH = .fitted - crit\_WH\*.se.fit,  
 upper\_CI\_WH = .fitted + crit\_WH\*.se.fit,  
 lower\_CI\_S = .fitted - crit\_Sch\*.se.fit,  
 upper\_CI\_S = .fitted + crit\_Sch\*.se.fit)  
  
#blue - no adjustment, red - Working-Hotelling, green - Bonferroni  
ggplot(movie\_CI, aes(x = bud.log, y =pop.log)) + geom\_point() +  
 stat\_smooth(method = "lm", se = FALSE) +  
 geom\_ribbon(data=movie\_CI, aes(ymin = lower\_CI\_B, ymax = upper\_CI\_B), alpha = .2, fill = "green") +  
 geom\_ribbon(data = movie\_CI, aes(ymin = lower\_CI\_WH, ymax = upper\_CI\_WH), alpha = .2, fill = "red") +  
 geom\_ribbon(aes(ymin = lower\_CI, ymax = upper\_CI), alpha = .2, fill = "blue") +  
 xlab("log(budget)") + ylab("log(popularity)")



#prediction interval for all n points  
movie\_PI <-broom::augment(movies) %>%  
 mutate(.se.pred = sqrt(movie\_sig^2 + .se.fit^2)) %>%  
 mutate(lower\_PI = .fitted - crit\_val \* .se.pred,  
 upper\_PI = .fitted + crit\_val \* .se.pred,   
 lower\_PI\_B = .fitted - crit\_Bonf\*.se.pred,  
 upper\_PI\_B = .fitted + crit\_Bonf\*.se.pred,   
 lower\_PI\_WH = .fitted - crit\_WH\*.se.pred,  
 upper\_PI\_WH = .fitted + crit\_WH\*.se.pred,  
 lower\_PI\_S = .fitted - crit\_Sch\*.se.pred,  
 upper\_PI\_S = .fitted + crit\_Sch\*.se.pred)  
  
#blue - no adjustment, red - Scheffer, green - Bonferroni  
ggplot(movie\_PI, aes(x = bud.log, y =pop.log)) + geom\_point() +  
 stat\_smooth(method = "lm", se = FALSE) +  
 geom\_ribbon(data=movie\_PI, aes(ymin = lower\_PI\_B, ymax = upper\_PI\_B), alpha = .2, fill = "green") +  
 geom\_ribbon(data = movie\_PI, aes(ymin = lower\_PI\_S, ymax = upper\_PI\_S), alpha = .2, fill = "red") +  
 geom\_ribbon(aes(ymin = lower\_PI, ymax = upper\_PI), alpha = .2, fill = "blue") +  
 xlab("log(budget)") + ylab("log(popularity)")



It's important to control for multiple comparisons, because the more variables you compare, the more likely you are to get a "significant" difference, even though there might not be a real effect. For this reason, it's important to have stricter measures of confidence when testing multiple variables. For mean intervals, it's most useful to use the Working-Hotelling procedure, because it uses the entire range of x's, whereas the Bonferroni only uses the intervals of interest. However, for prediction intervals, Bonferroni is better than Scheffe because it takes into account the multiple comparisons error.