R and Basic Linear Regression

Adam J Sullivan, PhD

1/29/2018

Linear Regression

Outline

1. One Categorical Covariate
2. One Continuous Covariate
3. Multiple Covariates

The Data for Class

- We will consider the data behind the story: "Comic Books are Still Made By Men, For Men and About Men".
- This data is part of the `fivethirtyeight` package:
- To explore the variable names run the following code:

```
library(fivethirtyeight)
?comic_characters
```

Appearances

- We will consider appearances on the comic books.
- We will see what predicts the number of appearances.

One Categorical Covariate - Binary

Binary Covariate

- With this type of covariate, we are comparing some outcome against 2 different groups.
- In order to make these comparisons it depends on the outcome we are working with.
- We will perform these tests based on the outcome and then use confidence intervals to assess.

Differences in appearances by publisher

- Let's consider the difference in appearances by publisher

```
library(fivethirtyeight)
library(tidyverse)

cnt <-  comic_characters%>%
  group_by(publisher) %>%
  tally()
mn<- comic_characters%>%
  group_by(publisher) %>%
  summarise(mean_app=mean(appearances, na.rm=T))
full_join(cnt,mn)
```

Differences in appearances by publisher

- Let's consider the difference in appearances by publisher

```
## # A tibble: 2 x 3
##   publisher     n mean_app
##   <chr>     <int>    <dbl>
## 1 DC         6896     23.6
## 2 Marvel    16376     17.0
```

Differences in Appearances by Publisher

- We have learned how to do this previously.
- We first did this comparison with a t-test
- Then we did this with an F-test in ANOVA

Appearance by Publisher: t-test

- Consider this with a t-test

```
t.test(appearances~publisher, comic_characters)
```

```
##
##  Welch Two Sample t-test
##
## data:  appearances by publisher
## t = 4.9476, df = 13552, p-value = 7.605e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.980214 9.203299
## sample estimates:
##     mean in group DC mean in group Marvel
##             23.62513             17.03338
```

Appearances by publisher: ANOVA

- Consider with ANOVA

```
library(broom)
tidy(aov(appearances~publisher, comic_characters))
```

```
##        term    df       sumsq     meansq statistic      p.value
## 1 publisher     1    199019.3 199019.306  22.63549 1.970861e-06
## 2 Residuals 21819 191840415.8   8792.356        NA           NA
```

ANOVA vs t-test

- t-test and ANOVA should give us the same results.
- We can see that in our output this is not true.
- What were the assumptions of ANOVA?

Appearances by publisher: t-test

- Consider this with a t-test

```
t.test(appearances~publisher, comic_characters, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  appearances by publisher
## t = 4.7577, df = 21819, p-value = 1.971e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.876078 9.307436
## sample estimates:
##     mean in group DC mean in group Marvel
##             23.62513             17.03338
```

Linear Regression

```
model <- lm(appearances~publisher, comic_characters)
tidy(model)
glance(model)
```

Linear Regression

```
##              term  estimate std.error statistic      p.value
## 1     (Intercept) 23.625134  1.159393 20.377163 1.893592e-91
## 2 publisherMarvel -6.591757  1.385499 -4.757677 1.970861e-06
##     r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.001036346  0.0009905619 93.76756  22.63549 1.970861e-06  2 -130046.9
##        AIC      BIC  deviance df.residual
## 1 260099.7 260123.7 191840416       21819
```

Interpreting the Coefficients: Categorical

- Intercept is the average for the reference group.
- Each coefficient is the average change between the reference group and the one of interest.

Interpreting the Coefficients: Categorical

- Intercept interpretation: Every DC character has on average 23.6 appearances.
- Marvel Coefficient: Every marvel character has on average 6.59 less appearances than DC.

One Binary Categorical Variable - Continuous Outcome

- We can perform
  - t-test with equal variances
  - ANOVA
  - Linear Regression
- All yield the same exact results

Assumptions of Linear Regression

- Function f is linear.
- Mean of error term is 0.

$$E(\varepsilon) = 0$$

- Error term is independent of covariate.

$$Corr(X, \varepsilon) = 0$$

- Variance of error term is same regardless of value of X.

$$Var(\varepsilon) = \sigma^2$$

- Errors are normally Distributed

What about more categories?

- We can also use linear regression with multiple categories.

```
mod <- lm(appearances~sex, comic_characters)
tidy(mod)
```

What about more categories?

- We can also use linear regression with multiple categories.

```
##                          term    estimate std.error   statistic      p.value
## 1                 (Intercept)  19.6666667  14.75085  1.33325688 0.1824620063
## 2       sexFemale Characters   1.3729391  14.80728  0.09272058 0.9261264351
## 3 sexGenderfluid Characters 262.8333333  69.18760  3.79885030 0.0001457792
## 4  sexGenderless Characters  -6.8245614  26.43048 -0.25820801 0.7962489134
## 5         sexMale Characters  -0.6395696  14.77091 -0.04329926 0.9654633959
## 6 sexTransgender Characters -15.6666667  96.72776 -0.16196660 0.8713337207
```

How do we interpret?

- We need to know the baseline.

```
## # A tibble: 7 x 3
##   sex                          n mean_sex
##   <chr>                    <int>    <dbl>
## 1 Agender Characters          45    19.7
## 2 Female Characters         5804    21.0
## 3 Genderfluid Characters       2   282
## 4 Genderless Characters       20    12.8
## 5 Male Characters          16421    19.0
## 6 Transgender Characters       1     4.00
## 7 <NA>                       979     5.13
```

Working with Factors

- Since we are interested in knowing whether or not male characters appear more often, we need to change how we view the factor.
- We will work on the following:
  - Renaming factors
  - Reordering factor levels.

Working with Factors: Renaming

```
comic_characters <- comic_characters %>%
      mutate(sex = fct_recode(sex,
      "Agender" = "Agender Characters",
      "Female" = "Female Characters",
      "Genderfluid" = "Genderfluid Characters",
      "Genderless" = "Genderless Characters",
      "Male" = "Male Characters",
      "Transgender" = "Transgender Characters"
      ))
```

Working with Factors: Relevel

```
comic_characters <- comic_characters %>%
      mutate(sex = fct_relevel(sex,
                 "Male",
                 "Agender",
                 "Female" ,
                  "Genderfluid" ,
                  "Genderless" ,
                  "Transgender"
      ))
```

Regression again

```
mod <- lm(appearances~sex, comic_characters)
tidy(mod)
```

Regression again

```
##              term     estimate  std.error   statistic       p.value
## 1     (Intercept)   19.0270971  0.7696883 24.72052160 5.155064e-133
## 2      sexAgender    0.6395696 14.7709130  0.04329926  9.654634e-01
## 3       sexFemale    2.0125087  1.5034512  1.33859259  1.807179e-01
## 4 sexGenderfluid  263.4729029 67.6012493  3.89745612  9.751072e-05
## 5   sexGenderless   -6.1849918 21.9448219 -0.28184288  7.780668e-01
## 6 sexTransgender  -15.0270971 95.5995052 -0.15718802  8.750982e-01
```

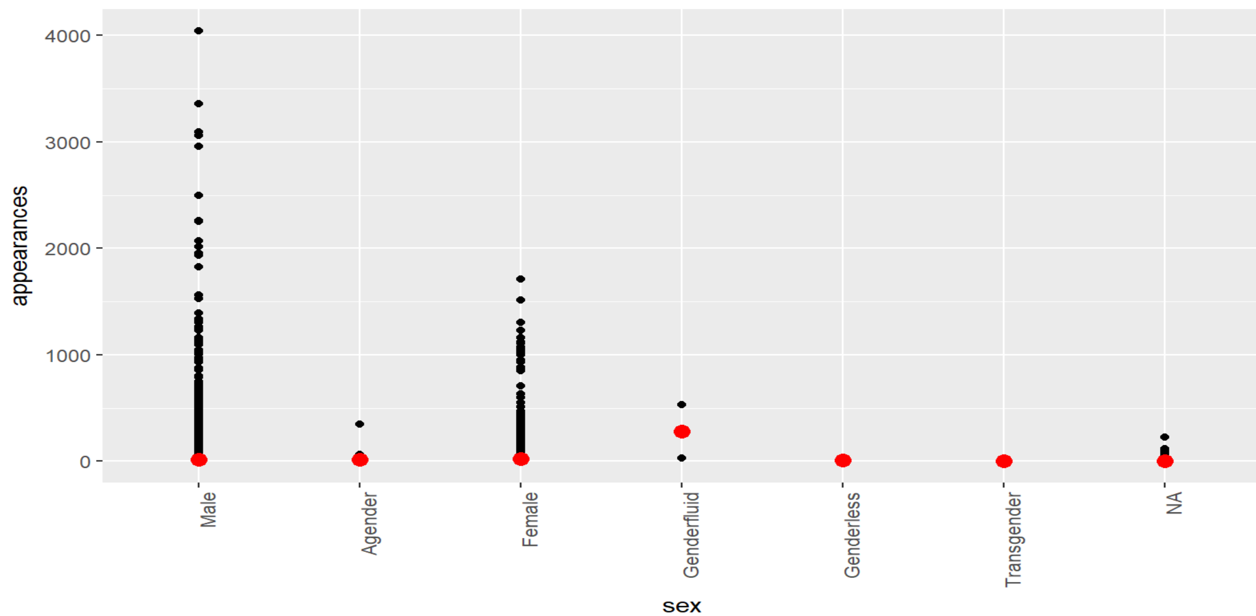
Interpreting the Coefficients: Categorical

- Intercept interpretation: Every Male Character has on average 19 appearances.
- Agender coefficient: Every Agender character has on average 0.64 more appearances than male characters
- …

Whats happening?

```
ggplot(comic_characters, aes(x = sex, y = appearances)) +
  geom_point()  +
  geom_point(stat = "summary", fun.y = "mean", color = "red", size = 3) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Whats happening?



One Continuous

One Continuous Covariate

- We will consider one continuous covariate.
- We will consider year.

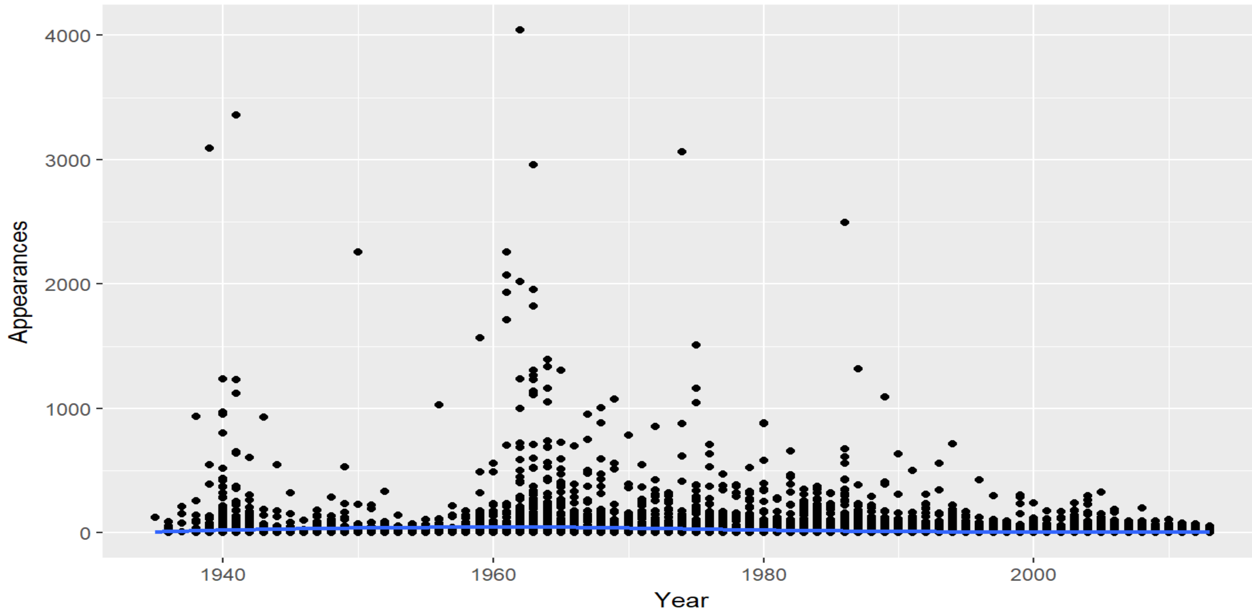
Example: Year and Appearances

- Consider the effect of year on appearances.
- With categorical data we plotted this with box-whisker plots.
- We can now use a scatter plot

Scatter Plot: Year and Appearances

```
ggplot(comic_characters, aes(year, appearances)) +
  geom_point() +
  geom_smooth(method="lm") +
  xlab("Year") +
  ylab("Appearances")
```

Scatter Plot: Year and Appearances



Modeling What We See

- There might not be a connection or there might be a very small one, let's explore further.
- How can we do this?
- How does linear regression work?

How do we Quantify this?

- One way we could quantify this is

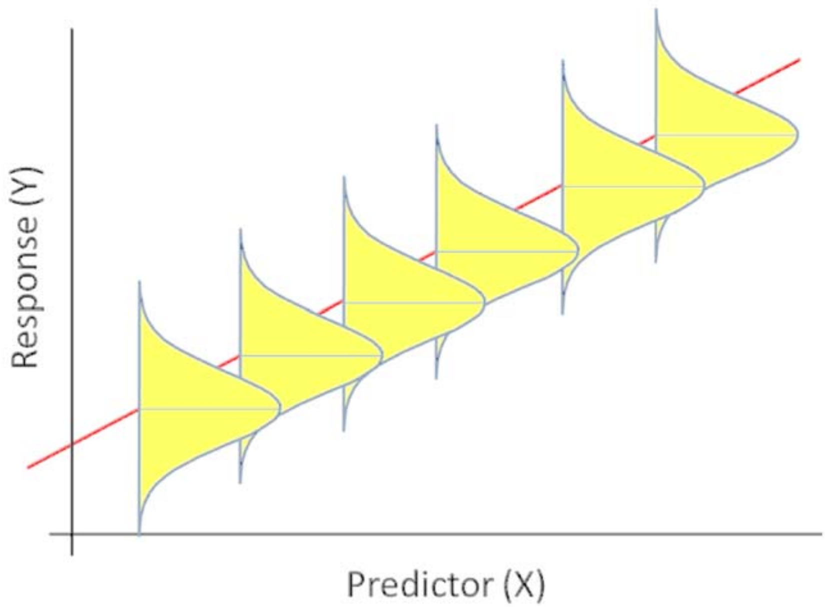
$$\mu_{y|x} = \beta_0 + \beta_1 X$$

- where
  - $\mu_{y|x}$ is the mean time for those whose year is x.
  - β_0 is the y-intercept (mean value of y when $x = 0$, $\mu_y|0$)
  - β_1 is the slope (change in mean value of Y corresponding to 1 unit increase in x).

Population Regression Line

- With the population regression line we have that the distribution of appearances for those at a particular year, x, is approximately normal with mean, $\mu_{y|x}$, and standard deviation, $\sigma_{y|x}$.

Population Regression Line



Distribution of Y and different levels of X.

Population Regression Line

- This shows the scatter about the mean due to natural variation. To accommodate this scatter we fit a regression model with 2 parts:
    - Systematic Part
    - Random Part

The Model

- This leads to the model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Where $\beta_0 + \beta_1 X$ is the systematic part of the model and implies that

$$E(Y|X = x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

- the variation part where we have $\varepsilon \sim N(0, \sigma^2)$ which is independent of X.

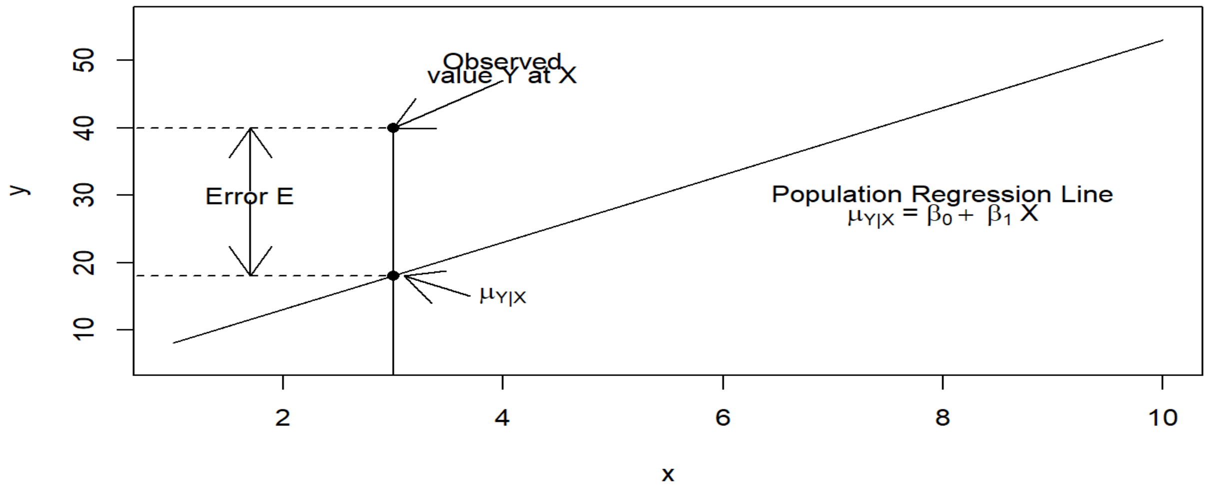
What do We Have?

- Consider the scenario where we have n subjects and for each subject we have the data points (x, y).
- This leads to us having data in the form (X_i, Y_i) for $i = 1, \dots, n$.
- Then we have the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- $Y_i | X_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$
- $E(Y_i | X_i) = \mu_{y|x} = \beta_0 + \beta_1 X_i$
- $Var(Y | X_i) = \sigma^2$

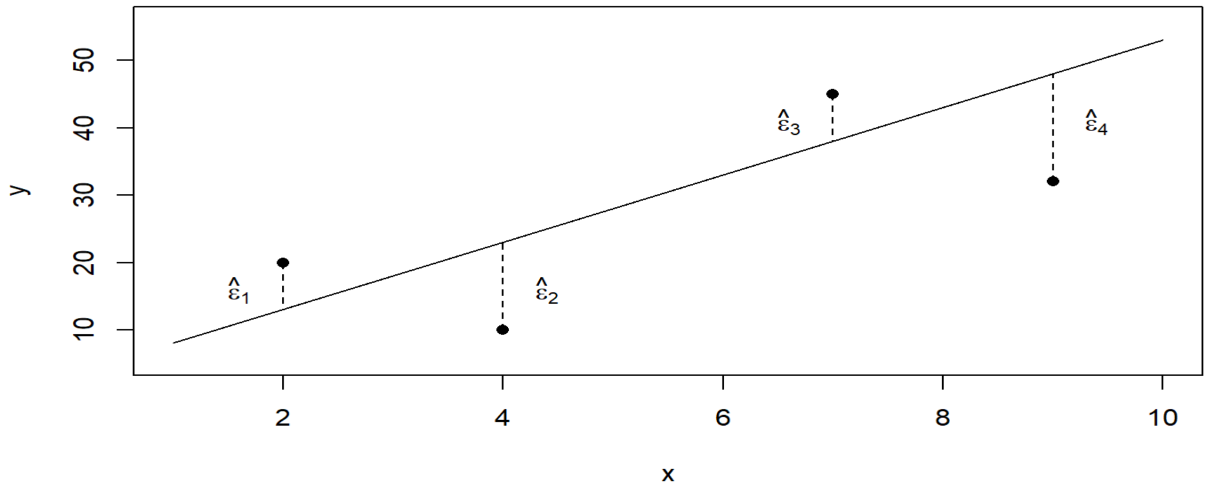
Picture of this



What Does This Tell Us?

- We can refer back to our scatter plot now and discuss what is the "best" line.
- Given the previous image we can see that a good estimator would somehow have smaller residual errors.
- So the "best" line would minimize the errors.

Residual Errors



In Comes Least Squares

- The least squares estimator of regression coefficients in the estimator that minimizes the sum of squared errors.
- We denote these estimators as $\hat{\beta}_0$ and $\hat{\beta}_1$.
- In other words we attempt to minimize

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Inferences on OLS

- Once we have our intercept and slope estimators the next step is to determine if they are significant or not.
- Typically with hypothesis testing we have needed the following:
  - Population/Assumed Value of interest
  - Estimated value
  - Standard error of Estimate

Confidence Interval Creation

- with 95% confidence intervals of

$$\hat{\beta}_1 \pm t_{n-2,0.975} \cdot se\left(\hat{\beta}_1\right)$$

$$\hat{\beta}_0 \pm t_{n-2,0.975} \cdot se\left(\hat{\beta}_0\right)$$

- In general we can find a $100(1-\alpha)\%$ confidence interval as

$$\hat{\beta}_1 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot se\left(\hat{\beta}_1\right)$$

$$\hat{\beta}_0 \pm t_{n-2,1-\frac{\alpha}{2}} \cdot se\left(\hat{\beta}_0\right)$$

Example: Year and Appearances

```
model <- lm(appearances~year, data=comic_characters)
tidy(model, conf.int=TRUE)[,-c(3:4)]
glance(model)
```

Example: Year and Appearances

```
##    r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.01457607    0.01452946 93.75137  312.7551 1.736275e-69  2 -126020.4
##        AIC      BIC  deviance df.residual
## 1 252046.8 252070.6 185841357       21144
```

Example: Year and Appearances

```
##    r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.01457607    0.01452946 93.75137  312.7551 1.736275e-69  2 -126020.4
##        AIC      BIC  deviance df.residual
## 1 252046.8 252070.6 185841357       21144
```

Interpreting the Coefficients: Continuous

- Before we can discuss the regression coefficients we need to understand how to interpret what these coefficients mean.
- β_0 is mean value for Y when $X = 0$.
- What about β_1?

Interpreting the Coefficients: Continuous

- Then we consider β_1 to see the meaning of this we do the following

$$\begin{aligned} E(Y|X=x+1) - E(Y|X=x) &= \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

Interpreting the Coefficients: Continuous

- We consider β_0 first.
- Does this value have meaning with our current data?
  - The estimated value of time level is only applicable to year within the range of our data.
  - Many times the intercept is scientifically meaningless.
  - Even if meaningless on its own, β_0 is necessary to specify the equation of our regression line.
  - **Note:** People do sometimes use mean centered data and the intercept is then interpretable.

Interpreting the Coefficients: Continuous

- This gives us the interpretation that β_1 represents the mean change in outcome Y given a one unit increase in predictor X.
- This is not an actual prescription though, this is considering different subjects or groups of subjects who differ by one unit.
- Below are correct interpretations of β_1 in our example.
  -
  -

Multiple Regression

- We have been discussing simple models so far.
- This works well when you have:
  - Randomized Data to test between specific groups (Treatment vs Control)
- In most situations we need look at more than just one relationship.
- Think of this as needing more information to tell the entire story.

Multiple Linear Regression with appearances

- First start with univariate models
- Then perform the multiple model

Multivariate Models

```
mod3 <- lm(appearances~publisher + year, data=comic_characters)
tidy3 <- tidy(mod3, conf.int=T)[,-c(3:4)]
tidy3
```

```
##               term    estimate      p.value     conf.low    conf.high
## 1     (Intercept) 1265.202320 9.811075e-78 1132.8767591 1397.5278806
## 2 publisherMarvel   -9.539045 1.242355e-11  -12.2971767   -6.7809141
## 3            year   -0.623927 5.927831e-75   -0.6904228   -0.5574312
```

Interpreting Multiple Coefficients

- The intercept is when all coefficients are zero.
- Each other coefficient is interpreted in context to another.

Interpreting Multiple Coefficients: Our Example

- Intercept: DC average appearances at year 0.
- Publisher Coefficient: If we consider 2 characters in the same year, the character from Marvel will have on average 9.54 less appearances than the character from DC.
- Year Coefficient: If we consider 2 characters from the same publisher, an increase in 1 year will lead to on average 0.62 less appearances.