Understanding the Reach of Facebook Posts Through Linear Regression

Lab class C4 Team 1
Delaney Swann, Mengze Li (Vanessa), Suan Jung, Ting Hsu, Xianjue Huang

Abstract

Facebook's power to reach a broad audience makes it a powerful tool for advertising. However, given the complex nature of marketing, the underlying cause of the popularity level of a post can be challenging to understand and predict. One of the ways to tackle this problem mathematically is to build a linear regression model with variable selection, and evaluate the model performance using a separate test set of data. In this analysis, two linear regression models for "paid" (posts that got paid for promotion) and "unpaid" Facebook posts were built separately in R on a set of Facebook posts' attributes. Results showed that, although paid posts receive better popularity overall, when it comes to each individual post, it highly depends on multiple factors. For both paid and unpaid posts, the number of Lifetime Consumers (number of people who view the post during its posting life), which is an indicator of the popularity and effectiveness of a post, are affected by Type and Category of the post, and Total Likes a post receives. Nevertheless, both the amount and the direction of the effects of these three input variables can be different in paid and unpaid posts. In addition, the number of consumers in paid posts appears to be more difficult to predict compared to unpaid posts.

Introduction

Facebook is a 434 billion dollar company with 2.5 billion monthly users[1]. It dominates the market share in minutes spent on a social media platform, with the average user spending 50 minutes per day scrolling[2]. Such a broad captive audience's marketing potential provides great opportunities for platform creators as well as other companies. Given the potential reach afforded to companies through social media, there is a vested interest in understanding how to optimize reach through different marketing strategies. However, because of the sophistication of the marketing tools now afforded by social media, it can be challenging to determine the value of a post given the range of metrics and factors involved. In a study conducted by University of Lisbon business researchers, metrics including reach, or the number of people who saw a post, impressions, or the number of times a post was displayed and engaged users - the number of unique people who clicked on a post, were sampled and serves as a reliable source of marketing analysis. Some of these metrics have overlap, while some may be irrelevant to determining how successful a post actually is. Therefore, choosing the metrics that are more indicative of reach is vital in determining an ad's success. Understanding the relationships of these decisions to metrics of success is important to create a marketing strategy. Using statistical analysis to determine the relationships between post attributes and metrics could unlock economically valuable insights on the best usage of Facebook's enormous audience to achieve a better advertising effect.

Member Contribution

Delaney Swann: Coding.
Mengze Li (Vanessa): Report writing (discussion part). Report proofreading. Presentation.
Suan Jung: Report writing (introduction and abstract). Report proofreading.
Ting Hsu: Coding. Report writing (model part).
Xianjue Huang: Coding. Report writing (model part).

Background

This analysis aims to determine how the decisions a company makes to post an advertisement — such as timing, type, and whether they pay Facebook to distribute the post — when posting on their Facebook profile can contribute to the post's overall reach of consumers. The data used in this analysis was procured by a "worldwide renown cosmetics brand" for the original researchers to analyze. Facebook allows a business account to extract most of the metrics included in this study from their posts, and so most of the data come directly from Facebook itself. The cosmetic brand in question added the input variable 'Category' to determine how their social media strategies impacted their posts' performance. This analysis will focus on how a set of input features — Category, Page Total Likes, Type, Post Hour, Post Weekday, Post Month, and Paid — impacts a single output: Lifetime Post Consumers.

Modeling and Analysis

Our plan is to predict lifetime post consumers using category, page total likes, type, month, hour, weekday, and paid, through linear regression. The very first step is to calculate the correlation matrix. The regression model was assumed to be:

$$Lifetime\ Post\ Consumers\ =\ Page\ Total\ Likes\ +\ Type\ +\ Month\ +\ Hour\ +\ Weekday\ + Paid\ +\ Category$$

where Page Total Likes, Month, Hour, and Weekday were all treated as quantitative variables, and Category, Type and Paid were categorical variables. The goal of the analysis was to determine the scale of the impact of each input variable and possible interactions between each of the variables. Given that no particularly strong relationship between the input variables were observed, (details in Appendix Figure A1), we proceed to evaluate the correlation between the output and input variables.

Table 1 below shows the multivariate linear regression result. The statistical estimates of Paid, Type and Category have the strongest significance among the input variables based on the results of a t-test run on each of the coefficients. The F-test of this linear regression shows that this model is more accurate than a model with no input variables, with a p-value close to zero.

Table 1. Statistic estimates of the full model: Lifetime Post Consumers = Page Total Likes + Type + Month + Hour + Weekday +Paid + Category. Type, Paid, and Category appear to be significant in affecting the number of Lifetime Consumers.

	Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	1.69E+03	5.88E+02	2.881	0.00414	**
Page.total.likes	-9.65E-03	6.44E-03	-1.497	0.13492	
Type: Photo	5.59E+02	1.74E+02	3.209	0.00142	**
Type: Status	2.02E+03	2.17E+02	9.306	< 2.00E-16	***
Type: Video	1.48E+03	3.35E+02	4.413	1.26E-05	***
Post.Month	-2.68E+01	3.17E+01	-0.844	0.39903	
Post.Hour	-6.29E+00	8.19E+00	-0.768	0.4427	
Post.Weekday	-2.43E+01	1.70E+01	-1.431	0.15295	
Paid	1.95E+02	7.69E+01	2.534	0.01161	*
Category: 2	-1.26E+02	9.67E+01	-1.304	0.19278	
Category: 3	-2.07E+02	8.45E+01	-2.445	0.01486	*

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 760.6 on 484 degrees of freedom
Multiple R-squared:  0.2766,	Adjusted R-squared:  0.2616
F-statistic:  18.5 on 10 and 484 DF,  p-value: < 2.2e-16
```

Given that a high residual standard error was observed, we suspected different subgroups of the data follow a different pattern.

Data separation

The following plots are lifetime post consumers in unpaid and paid posts, colored by Type and Category. Although the general trend is similar, the pattern of paid and unpaid posts show different underlying patterns under certain circumstances. For instance, in Figure 1, although in “Status” type paid posts receives more consumers than unpaid, the same trend was not observed in “Video” type. Similarly, paid posts gain more consumers in “category” 1 and 2 but not in 3.As a result, we decided to separate the data according to the paid and unpaid posts and build two models separately.

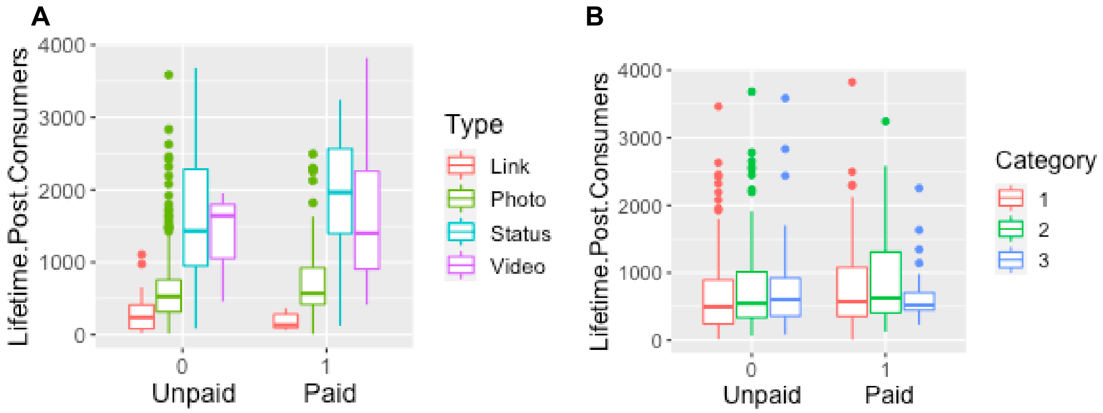


Figure 1 (A) Comparison of consumer numbers between paid and unpaid posts in different types.
(B) Comparison of consumer numbers between paid and unpaid posts in different categories of the posts.

Modeling and Analysis

For Paid and Unpaid posts, data was separated into training and testing sets. Two separate models were built from paid and unpaid posts individually. Variable selection was performed using the forward algorithm.

Modeling and analysis of Paid

Type, Likes, and Category were found to be significant for the model building (see Table 2 below).

The final multiple linear regression model for Paid was:

Lifetime Post Consumers = Page Total Likes + Type + Category

The standardized residuals were generally distributed around zero with a few outliers, and most of the standard residual data were normally distributed (see Figure 2 below). The model was in good shape.

Table 2. Final selection result of Type, Likes, and Category for Paid. All three had a decent AIC.

```
                              Selection Summary
------------------------------------------------------------------------------
        Variable                       Adj.
Step    Entered        R-Square      R-Square     C(p)          AIC          RMSE
------------------------------------------------------------------------------
   1    Type             0.2640        0.2305    10.3950    1116.3592    673.8768
   2    Likes            0.3395        0.2988     4.5626    1110.7873    643.2904
   3    Category         0.4172        0.3617    -1.5028    1106.0268    613.7867
------------------------------------------------------------------------------
```

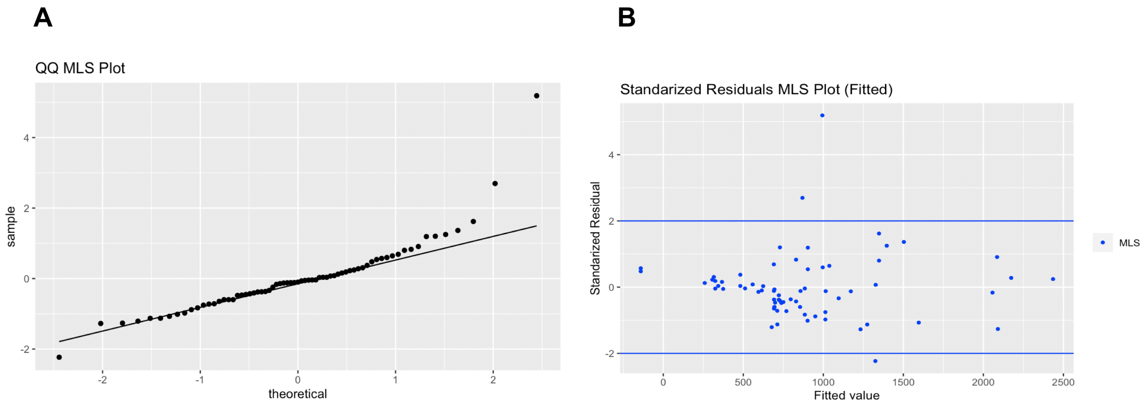


Figure 2. (A) QQ plot to test the normality for standardized residuals of multiple linear regression for Paid posts.
(B) Fitted standardized residuals of multiple linear regression for Paid posts.

Modeling and analysis for Unpaid

Similar to the model for paid posts, only Type, Likes, and Category were significant enough to affect the response variable (see Table 3 below).

The final multiple linear regression model for Unpaid posts was:

Lifetime Post Consumers = Page Total Likes + Type + Category

The residuals were generally distributed around zero with a few outliers, and most of the standardized residuals were normally distributed, but some extreme value deviates from the distribution (see Figure 3 below). Overall, the model was valid enough to provide valuable information.

Table 3. Selection Summary of Type, Likes, and Category for Unpaid

```
                              Selection Summary
------------------------------------------------------------------------------
        Variable                       Adj.
Step    Entered        R-Square      R-Square     C(p)          AIC          RMSE
------------------------------------------------------------------------------
   1    Type             0.2557        0.2472    20.6931    2761.2732    557.5063
   2    Likes            0.3424        0.3310     0.0225    2741.2369    525.5469
   3    Category         0.3530        0.3342    -0.7670    2742.3268    524.2903
------------------------------------------------------------------------------
```

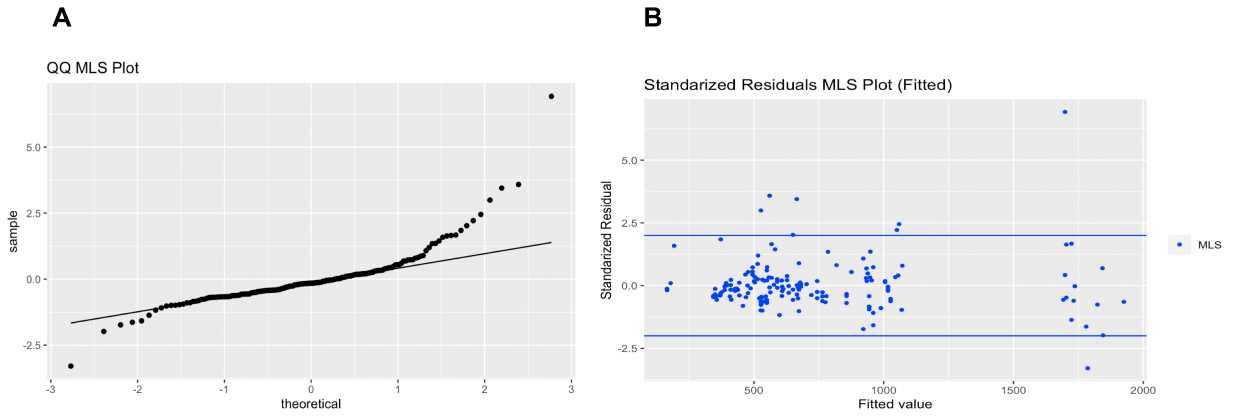


Figure 3. (A) QQ plot to test the normality for standardized residuals of multiple linear regression for Unpaid posts.

(B) Fitted standardized residuals of multiple linear regression for Unpaid posts.

Prediction

For both Paid and Unpaid data, prediction was performed using the model generated by the training set and the data in the test set.

Prediction for Paid

The model captured some of the data variation (see Figure 4B below) . However, the model failed to predict the outliers. The relative Mean Square Error for validation data was 0.7214199, which is acceptable.

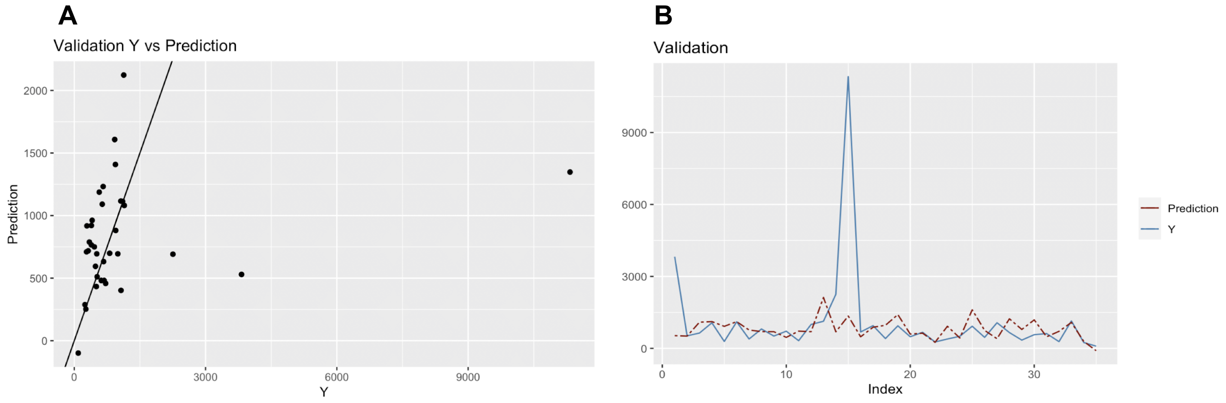


Figure 4 (A) Validation of Lifetime Post Consumers (real value) vs. Prediction in paid posts.

(B) Validation comparison between Lifetime Post Consumers (real value) and prediction in paid posts.

Prediction for Unpaid

The model captured most of the variation of the data. The real value (blue line) shows some extreme peaks that the model failed to capture (see Figure 5B), however, extreme values are difficult to predict by nature. The relative Mean Square Error for validation data is 0.3825531, which is much better than the paid model.

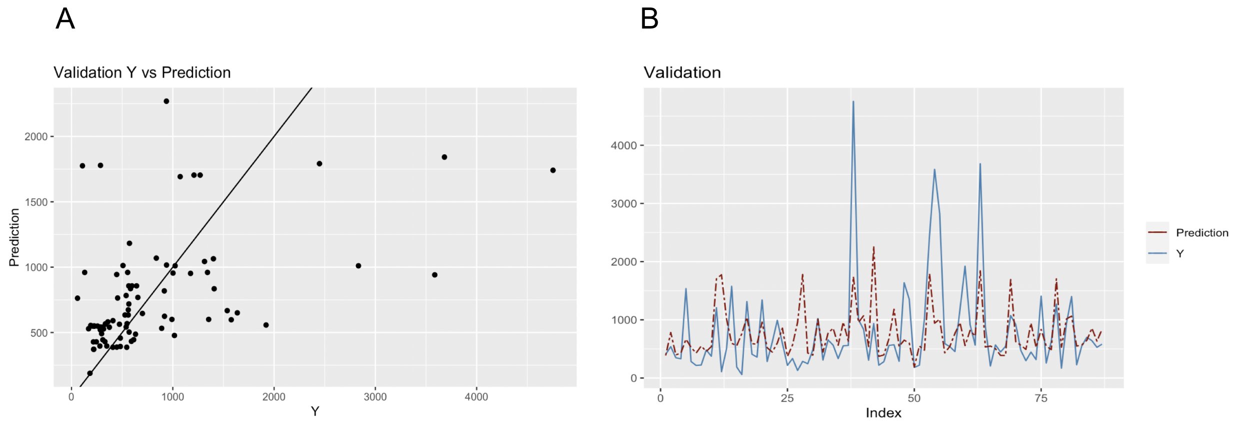


Figure 5 (A) Validation of Lifetime Post Consumers (real value) vs. Prediction in unpaid posts.
(B) Validation comparison between Lifetime Post Consumers (real value) and prediction in unpaid posts.

Discussion

The goal of this analysis was to determine how a company could improve the number of people who see the post over the posts lifetime (lifetime consumers) by posting under certain circumstances. The analysis was done individually in paid and unpaid posts given their different underlying mechanisms to increase the number of consumers. After variable selection, although both of the models yield the same variables, which are category, types, and likes, the parameter estimates in the model are different in values and direction.

Variables could affect lifetime consumers including:

1. Whether it is a "paid" post. Generally, posts that get paid to be promoted tend to gain more consumers, which corresponds with the common sense in the market. However, an inspiring finding was that the effects of paid posts may depend on if the post is precise and clear. Our analysis revealed that a paid post has more potential to be promoted efficiently than unpaid posts when it is directly related to a specific product or brand, and this increase of potential was not observed when posts are not related to specific advertising products.
2. Category and Type of a post. In general, "Status" and "Video" types tend to receive better outcomes than "Link" and "Photo" types. On the other hand, as mentioned above, posts in "action" and "promotion" categories (those with a clear advertising content) have a higher number of lifetime consumers while "inspiration" posts are the opposite.
3. Total "likes" a post receives. The relationship between total likes and lifetime consumers was significant. Nevertheless, the underlying cause-effect relationship between them is still unclear. It is possible that the two variables affect each other recursively as a circle: more consumers would result in more likes, and posts received more likes would get promoted and be viewed by more people. This assumption can explain the relatively frequent existence of extreme high lifetime consumer values in the dataset.

Moreover, the paid posts appear to be more fluctuated and difficult to predict. The model for unpaid posts did a better decent job in prediction compared to the model for paid posts. It's possible that the behavior of paid posts are difficult to predict, or there are other hidden factors that affect the post popularity that was not collected in the dataset.

Nevertheless, further investigation is still in demand given the nature of the limitation of a single analysis. In addition, both the models for paid and unpaid posts performed well when the response value (lifetime post consumers) was around the middle range. Extreme values are more difficult to predict since they could be affected by hidden factors unidentified in the dataset. Examples include the creativity of an advertisement, the quality and market demand of a product itself, and the amount of the discount the sealer is offering.

Acknowledgement

Thanks for Prof. Julio Castrillon throughout the semester for guiding us through the learning process and providing feedback.

Appendix

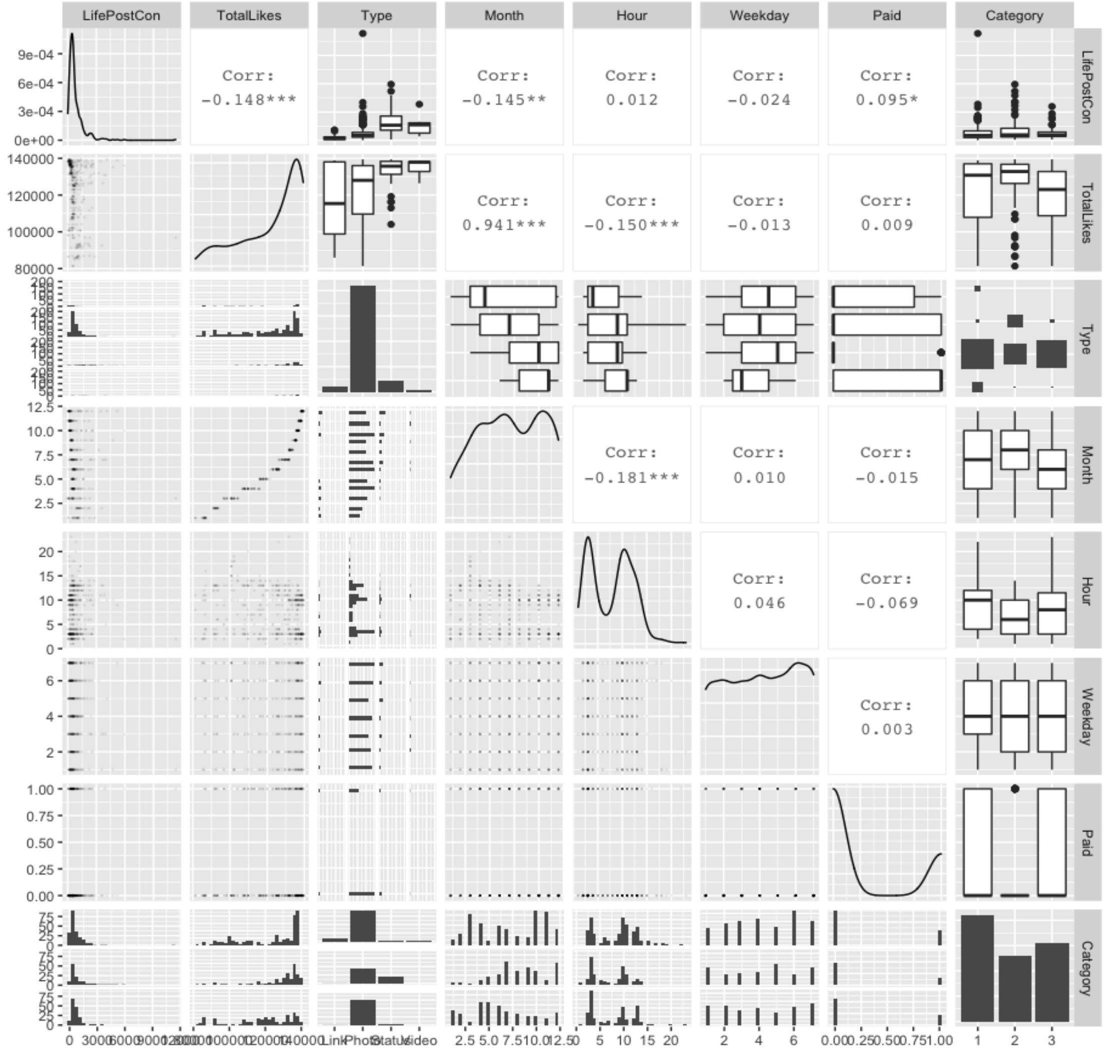
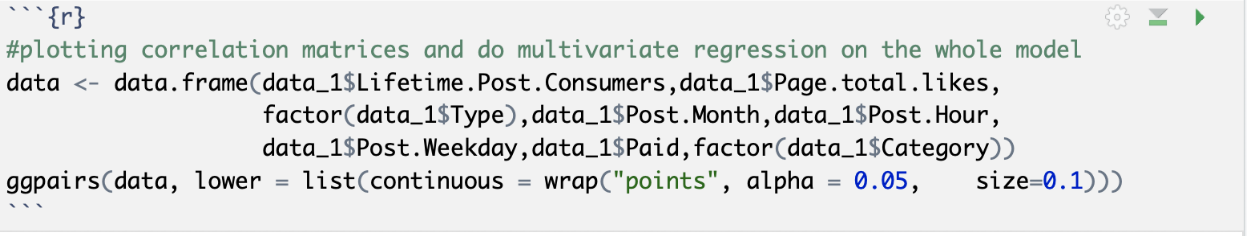
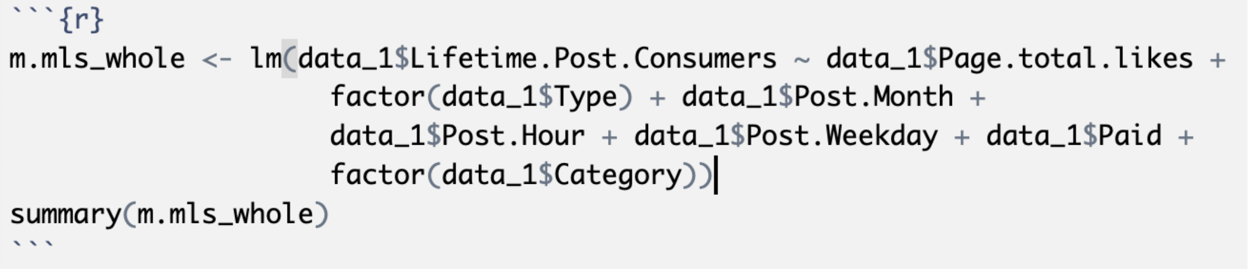


Figure A1. Scatter matrix of output (Lifetime Post Consumer) and input features (Page Total Likes, Type, Month, Hour, Weekday, Paid, Category)



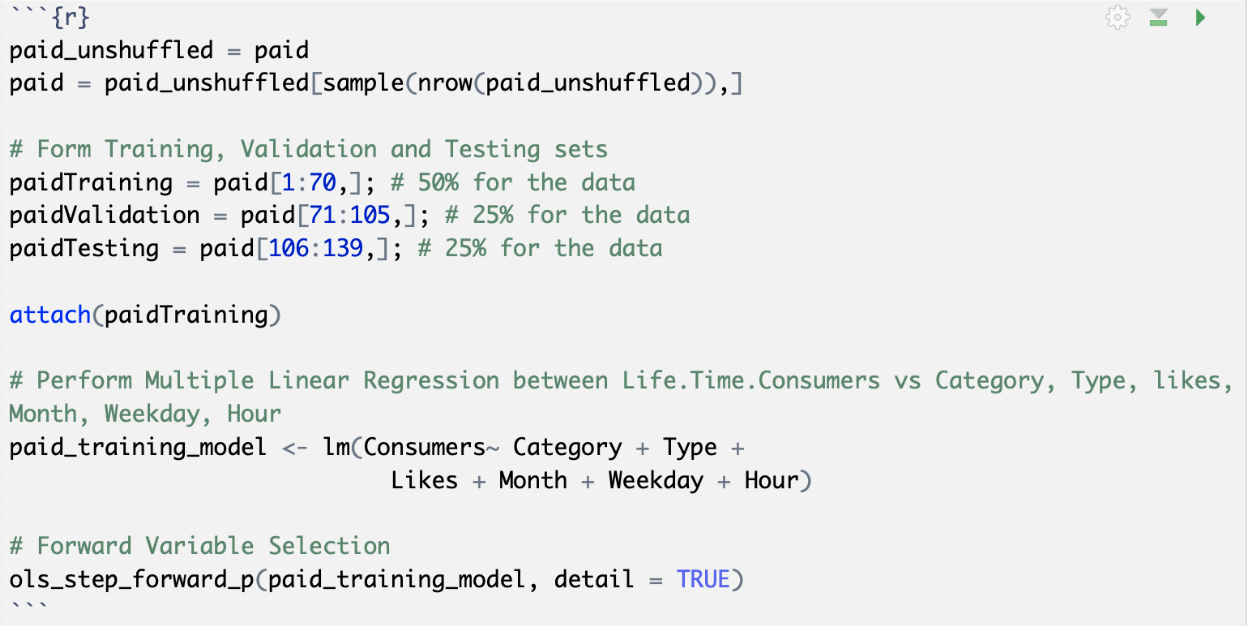
~~~
```{r}
#plotting correlation matrices and do multivariate regression on the whole model
data <- data.frame(data_1$Lifetime.Post.Consumers,data_1$Page.total.likes,
                   factor(data_1$Type),data_1$Post.Month,data_1$Post.Hour,
                   data_1$Post.Weekday,data_1$Paid,factor(data_1$Category))
ggpairs(data, lower = list(continuous = wrap("points", alpha = 0.05,    size=0.1)))
```
~~~

Code for Figure A1



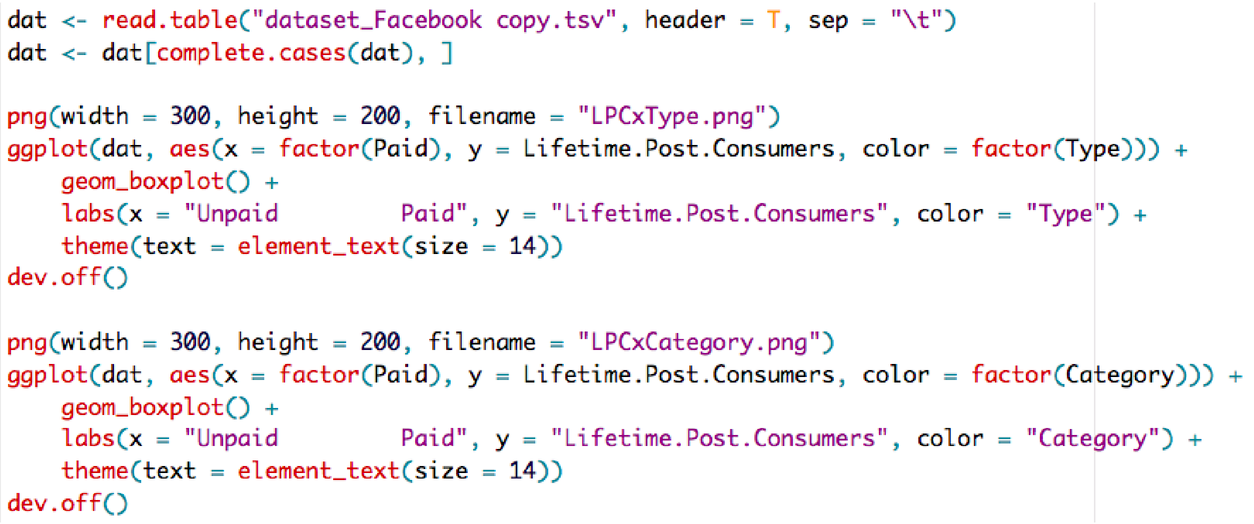
~~~
```{r}
m.mls_whole <- lm(data_1$Lifetime.Post.Consumers ~ data_1$Page.total.likes +
                        factor(data_1$Type) + data_1$Post.Month +
                        data_1$Post.Hour + data_1$Post.Weekday + data_1$Paid +
                        factor(data_1$Category))
summary(m.mls_whole)
```
~~~

Code for Table 1



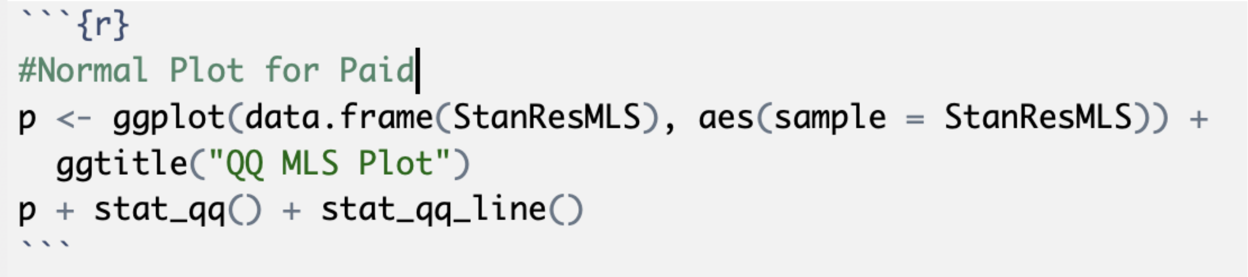
~~~
```{r}
paid_unshuffled = paid
paid = paid_unshuffled[sample(nrow(paid_unshuffled)),]

# Form Training, Validation and Testing sets
paidTraining = paid[1:70,]; # 50% for the data
paidValidation = paid[71:105,]; # 25% for the data
paidTesting = paid[106:139,]; # 25% for the data

attach(paidTraining)

# Perform Multiple Linear Regression between Life.Time.Consumers vs Category, Type, likes,
Month, Weekday, Hour
paid_training_model <- lm(Consumers~ Category + Type +
                            Likes + Month + Weekday + Hour)

# Forward Variable Selection
ols_step_forward_p(paid_training_model, detail = TRUE)
```
~~~

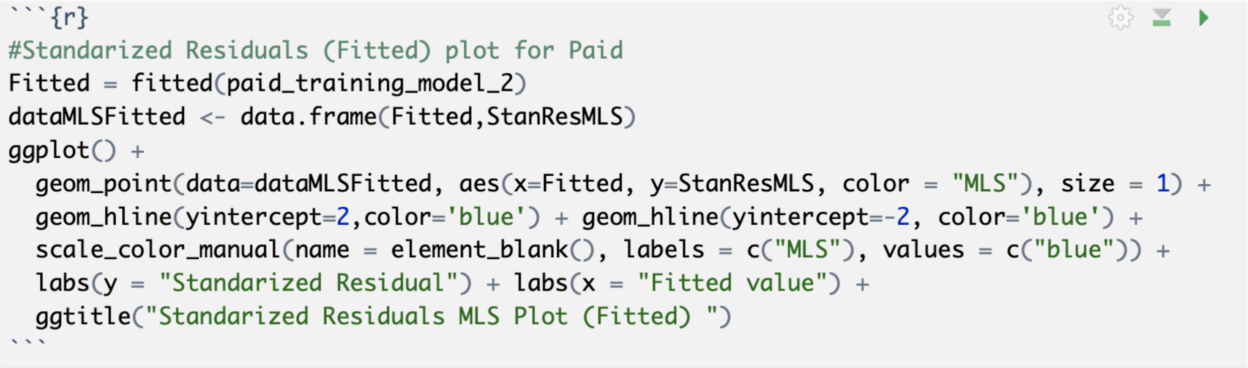
Code for Table 2

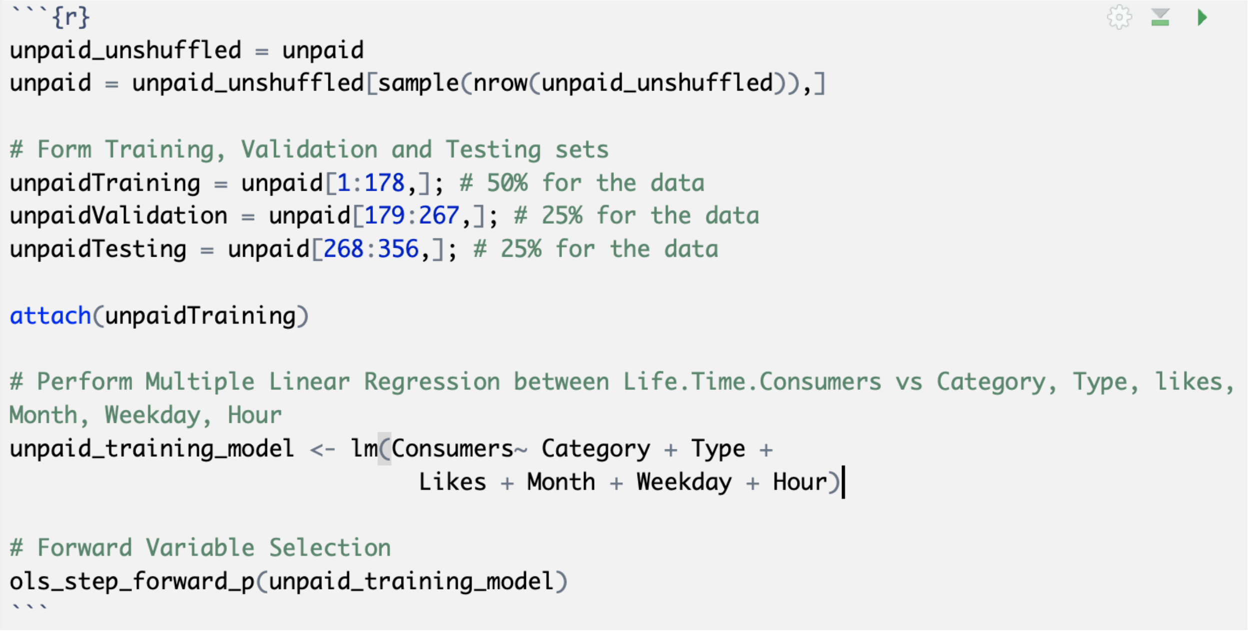


```r
dat <- read.table("dataset_Facebook copy.tsv", header = T, sep = "\t")
dat <- dat[complete.cases(dat), ]

png(width = 300, height = 200, filename = "LPCxType.png")
ggplot(dat, aes(x = factor(Paid), y = Lifetime.Post.Consumers, color = factor(Type))) +
    geom_boxplot() +
    labs(x = "Unpaid         Paid", y = "Lifetime.Post.Consumers", color = "Type") +
    theme(text = element_text(size = 14))
dev.off()

png(width = 300, height = 200, filename = "LPCxCategory.png")
ggplot(dat, aes(x = factor(Paid), y = Lifetime.Post.Consumers, color = factor(Category))) +
    geom_boxplot() +
    labs(x = "Unpaid         Paid", y = "Lifetime.Post.Consumers", color = "Category") +
    theme(text = element_text(size = 14))
dev.off()
```

Code for Figure 1

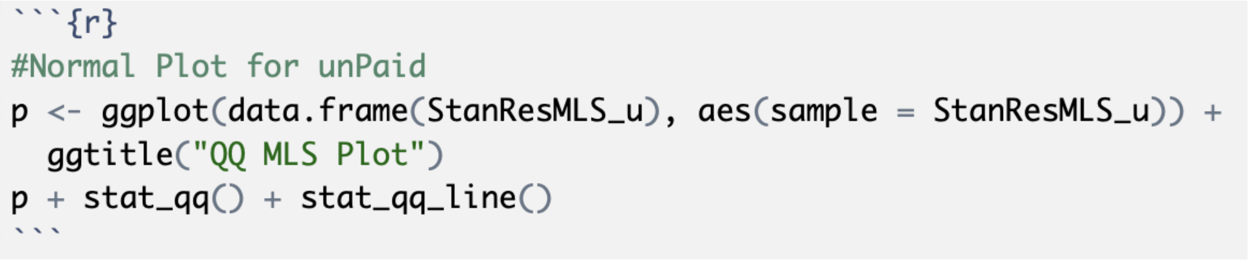


~~~r
```{r}
#Normal Plot for Paid
p <- ggplot(data.frame(StanResMLS), aes(sample = StanResMLS)) +
  ggtitle("QQ MLS Plot")
p + stat_qq() + stat_qq_line()
~~~
~~~

Code for Figure 2A



```r
```{r}
#Standarized Residuals (Fitted) plot for Paid
Fitted = fitted(paid_training_model_2)
dataMLSFitted <- data.frame(Fitted,StanResMLS)
ggplot() +
  geom_point(data=dataMLSFitted, aes(x=Fitted, y=StanResMLS, color = "MLS"), size = 1) +
  geom_hline(yintercept=2,color='blue') + geom_hline(yintercept=-2, color='blue') +
  scale_color_manual(name = element_blank(), labels = c("MLS"), values = c("blue")) +
  labs(y = "Standarized Residual") + labs(x = "Fitted value") +
  ggtitle("Standarized Residuals MLS Plot (Fitted) ")
~~~
```

Code for Figure 2B
```



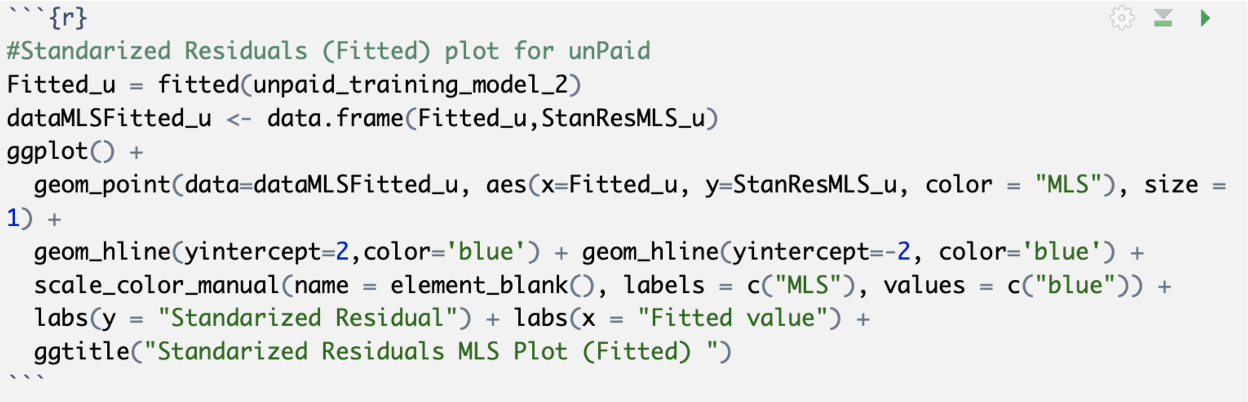
~~~
```{r}
unpaid_unshuffled = unpaid
unpaid = unpaid_unshuffled[sample(nrow(unpaid_unshuffled)),]

# Form Training, Validation and Testing sets
unpaidTraining = unpaid[1:178,]; # 50% for the data
unpaidValidation = unpaid[179:267,]; # 25% for the data
unpaidTesting = unpaid[268:356,]; # 25% for the data

attach(unpaidTraining)

# Perform Multiple Linear Regression between Life.Time.Consumers vs Category, Type, likes,
Month, Weekday, Hour
unpaid_training_model <- lm(Consumers~ Category + Type +
                              Likes + Month + Weekday + Hour)

# Forward Variable Selection
ols_step_forward_p(unpaid_training_model)
```
~~~

Code for Table 3



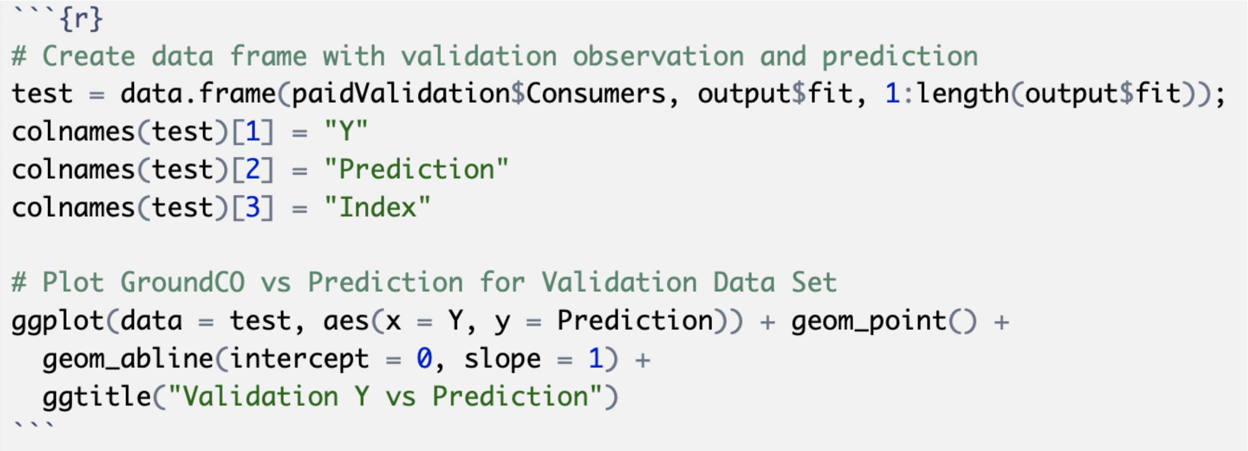
~~~
```{r}
#Normal Plot for unPaid
p <- ggplot(data.frame(StanResMLS_u), aes(sample = StanResMLS_u)) +
  ggtitle("QQ MLS Plot")
p + stat_qq() + stat_qq_line()
```
~~~

Code for Figure 3A



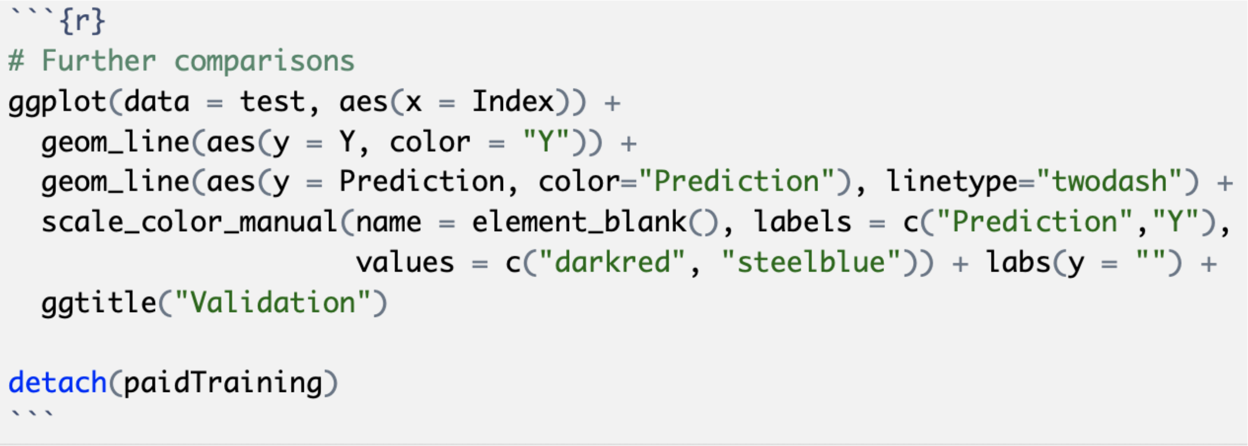
~~~
```{r}
#Standarized Residuals (Fitted) plot for unPaid
Fitted_u = fitted(unpaid_training_model_2)
dataMLSFitted_u <- data.frame(Fitted_u,StanResMLS_u)
ggplot() +
  geom_point(data=dataMLSFitted_u, aes(x=Fitted_u, y=StanResMLS_u, color = "MLS"), size =
1) +
  geom_hline(yintercept=2,color='blue') + geom_hline(yintercept=-2, color='blue') +
  scale_color_manual(name = element_blank(), labels = c("MLS"), values = c("blue")) +
  labs(y = "Standarized Residual") + labs(x = "Fitted value") +
  ggtitle("Standarized Residuals MLS Plot (Fitted) ")
```
~~~

Code for Figure 3B



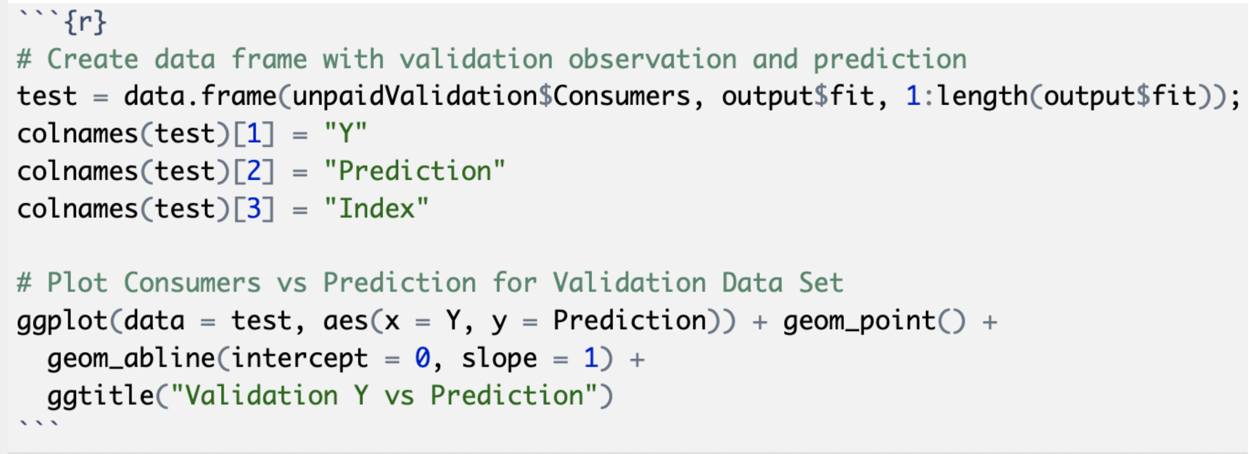
~~~
```{r}
# Create data frame with validation observation and prediction
test = data.frame(paidValidation$Consumers, output$fit, 1:length(output$fit));
colnames(test)[1] = "Y"
colnames(test)[2] = "Prediction"
colnames(test)[3] = "Index"

# Plot GroundCO vs Prediction for Validation Data Set
ggplot(data = test, aes(x = Y, y = Prediction)) + geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  ggtitle("Validation Y vs Prediction")
```
~~~

Code for Figure 4A



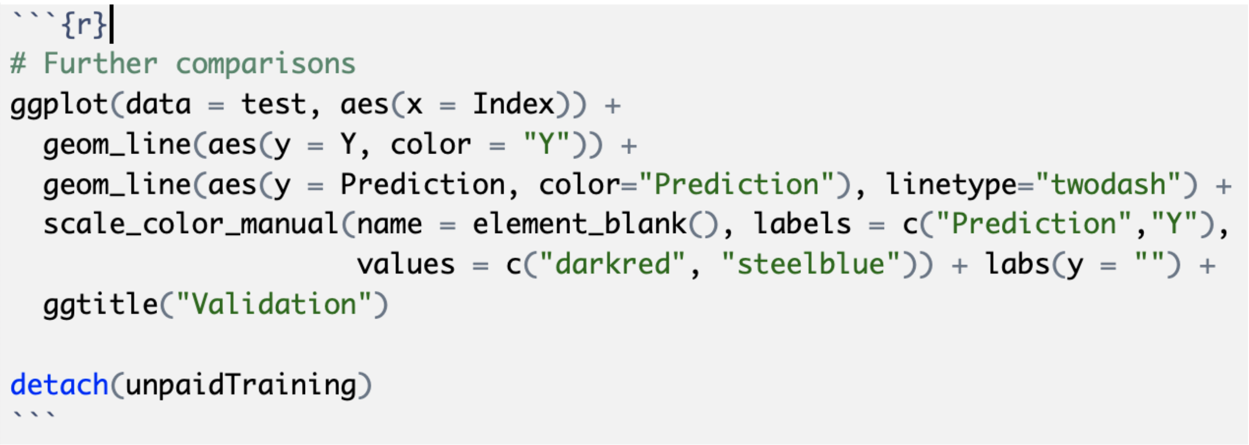
~~~
```{r}
# Further comparisons
ggplot(data = test, aes(x = Index)) +
  geom_line(aes(y = Y, color = "Y")) +
  geom_line(aes(y = Prediction, color="Prediction"), linetype="twodash") +
  scale_color_manual(name = element_blank(), labels = c("Prediction","Y"),
                     values = c("darkred", "steelblue")) + labs(y = "") +
  ggtitle("Validation")

detach(paidTraining)
```
~~~

Code for Figure 4B



~~~
```{r}
# Create data frame with validation observation and prediction
test = data.frame(unpaidValidation$Consumers, output$fit, 1:length(output$fit));
colnames(test)[1] = "Y"
colnames(test)[2] = "Prediction"
colnames(test)[3] = "Index"

# Plot Consumers vs Prediction for Validation Data Set
ggplot(data = test, aes(x = Y, y = Prediction)) + geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  ggtitle("Validation Y vs Prediction")
```
~~~

Code for Figure 5A



~~~
```{r}
# Further comparisons
ggplot(data = test, aes(x = Index)) +
  geom_line(aes(y = Y, color = "Y")) +
  geom_line(aes(y = Prediction, color="Prediction"), linetype="twodash") +
  scale_color_manual(name = element_blank(), labels = c("Prediction","Y"),
                     values = c("darkred", "steelblue")) + labs(y = "") +
  ggtitle("Validation")

detach(unpaidTraining)
```
~~~

Code for Figure 5B