Regression Analysis of Facebook Metrics

1 Introduction

1.1 Objective

With the growing popularity of social media, social media has become a vital platform for promoting an organisation's products, events, and branding. Thus, I would like to investigate the features that have significant influence on post performance and predict the impact of a post before it is published. This analysis could be used to assist in devising data-driven marketing and outreach strategies to maximise reach to current and potential customers as well as boost event engagement rates through posting on social media.

1.2 Data Set

The Facebook metrics data set is obtained from UCI Machine Learning Repository, consisting of information about 500 posts published by a cosmetics brand throughout the year of 2014. (Moro et al. 2016) It contains 7 features known prior to post publication, namely category of post, page total likes, type of content, month the post was published, day the post was published, hour the post was published and whether the company paid for Facebook advertising. The data set also includes 12 features that assess post performance. In this analysis, the target variable will be lifetime post consumers, which is the number of users who clicked anywhere in the post.

2 Methods

2.1 Exploratory Data Analysis

A brief analysis of the relationship between the explanatory variables and the response variable is conducted before the modelling process. According to Figure 1(a), as "Total Page Likes" increased, there was a slight increase in lifetime post consumers, thus we may infer that "Total Page Likes" and "Lifetime Post Consumers" might be linked. From Figure 1(b), we observe that "Status" posts outperformed the other types of posts on average, implying that type of post has great influence on the target variable. Based on Figure 1(c), "Inspiration" posts performed slightly better than the other categories. Figure 1(d) shows that "Paid" posts had greater impact than "Non-Paid" posts on average. In Figure 1(e), we notice that posts that were published in February have the highest lifetime post consumers on average, whereas posts in November have the lowest lifetime post consumers on average. According to Figure 1(f), posts that were published on Wednesdays performed the best, but the posts that were published on the other days performed relatively the same. From Figure 1(g), posts that were published at 4am had the highest average lifetime post consumers, while posts that were published at 7pm had the lowest average lifetime post consumers.

The histogram in Figure 1(h) shows that the distribution of "Lifetime Post Consumers" is highly positively skewed, thus logarithm is applied to "Lifetime Post Consumers" so that it has distribution closer to normal as shown in the histogram in Figure 1(i). Note that 1 is added to "Lifetime Post Consumers" to account for the cases where there are zero lifetime post consumers. The correlation between "Post Month" and "Total Page Likes" is 0.94 and the correlation between "type_photo" and "type_status" is -0.75. Therefore, the predictors "Post Month" and "type_photo" are dropped from the models. Dummy variables are also created for the categorical predictor variables "Type" and "Category".

2.2 Least Squares Regression

Figure 2(a) shows the output for the least squares linear model with 8 predictor variables. We can observe that the predictors "Post Weekday", "Post Hour", "category_action" and "category_product" are not significant at the 1% level. Thus, these predictors are removed from the linear model and the model is re-fit with the remaining 4 predictors, namely "Total Page Likes", "Paid", "type_link" and "type_status". In Figure 2(b), we notice that there is a slight drop in the adjusted R-squared value and increase in the residual standard error for the smaller model. An analysis of variance is conducted to decide whether the bigger or smaller model should be used. From Figure 2(c), the F-statistic p-value is approximately 0.08. Hence, there is insufficient evidence to reject the null hypothesis at the 1% level (corresponding to the smaller model). This means that we can proceed to use the smaller model.

Figure 3 demonstrates the diagnosis plots for the smaller model to check whether the assumptions of the linear regression model are met. In Figure 3(a), the red line is approximately horizontal at zero, indicating there might be a linear relationship between the predictors and outcome. According to Figure 3(b), the points generally fall on the reference line, with the exception of heavy-tails at both ends. The normality assumption can be considered to be met. Figure 3(c) does not show any obvious patterns and the red line stays reasonably constant, implying that the residuals reasonably satisfy the constant variance assumption. From Figure 3(d), the Cook's distance is generally low, but the leverage is considerably high for some of the data points. The diagnosis plots demonstrate that the linear regression assumptions are met at an acceptable level.

A training and test sample is generated, with 80% of the data used for training and remaining 20% for testing. The least squares model is fitted to just the training data and used to predict the response associated with the test set observations. The same train-test split are also used for all the models below.

2.3 Ridge Regression

Cross-validation is run to obtain the optimal λ which minimises the mean square error, which is 0.0335. Figure 4(a) shows that the coefficients of the predictor variables have been slightly shrunk, which can be further affirmed by the table of coefficients in Figure 5.

2.4 Lasso Regression

Similar to Ridge Regression, cross-validation is run to obtain the optimal λ of 0.00166. Figure 4(b) displays that the coefficients of the predictor variables have also been slightly shrunk. In Figure 5, the table of coefficients show that the shrinkage for lasso is not as much as in ridge regression as the optimal penalty parameter for lasso is much smaller.

2.5 Principal Components Regression

According to the validation plot in Figure 4(c), minimum root mean squared error is achieved when all 4 principal components are included in the model, which is the same as the least squares regression model. However, the root mean squared error when 3 principal components are used is only slightly higher than when all 4 principal components are used and using 3 principal components is able to capture approximately 96.9% of the variance. Thus, 3 principal components are used in the model. With reference to the coefficients of the principal components regression model with 3 principal components in Figure 5, we notice that the coefficient of "Total Page Likes" has been shrunk more than ridge and lasso regression. The coefficient of "Paid" has not been shrunk. It is interesting to note that the coefficient of "type_link" has been shrunk significantly in comparison to ridge and lasso regression, whereas the coefficient of "type_status" has been slightly increased instead.

3 Results and Discussion

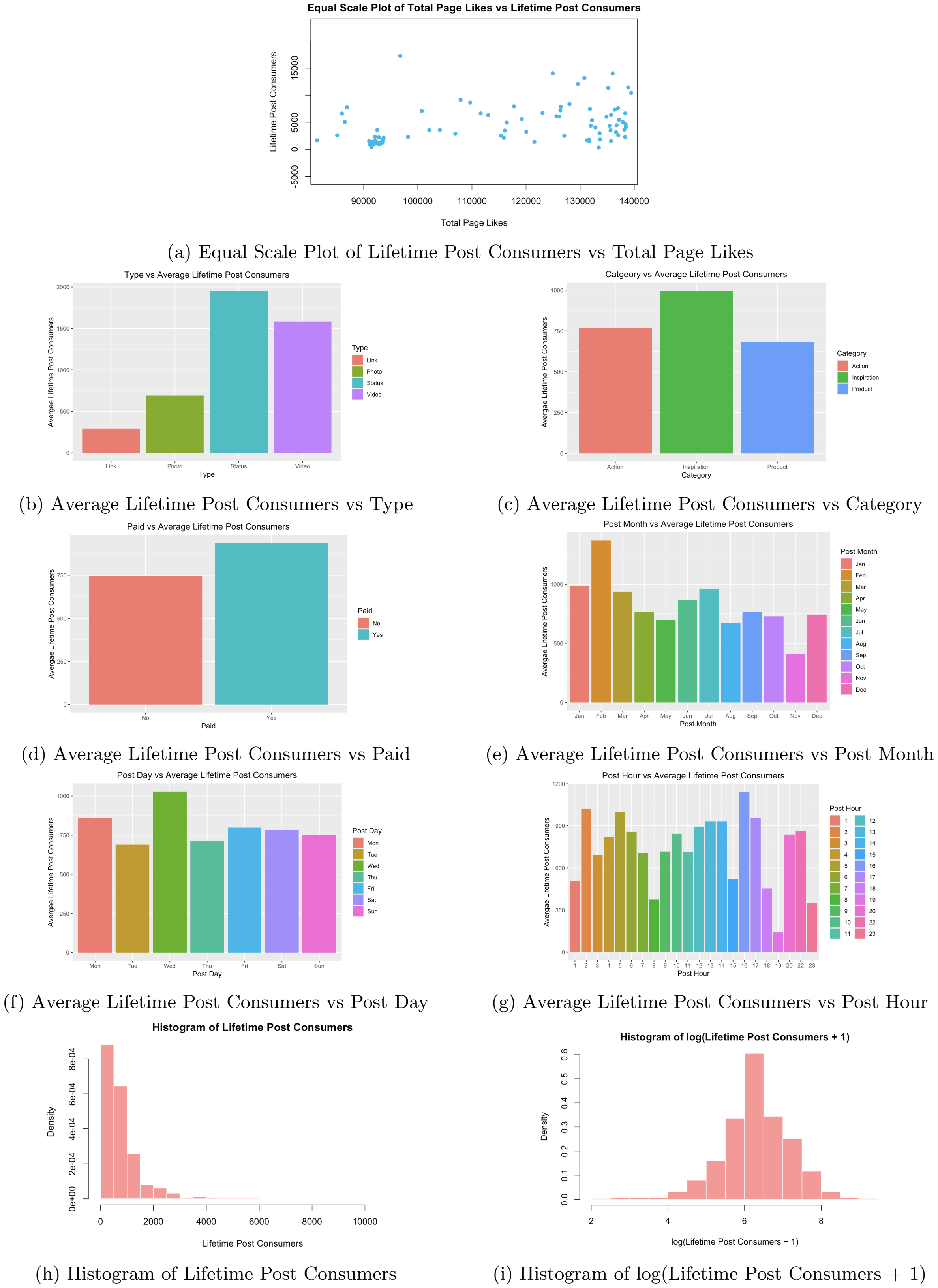
From Figure 5, we can deduce that principal components regression model outperforms the other models as it produced the lowest test mean squared error. It is then followed by ridge regression, lasso regression and least squares regression. In Figure 4(d), the residuals vs true test values plot shows that the models tend to overestimate small values of "Lifetime Post Consumers" and underestimate large values of "Lifetime Post Consumers", but the residuals are fairly random in between. This suggests that there might be outliers, the observations are not independent or the model is unable to fully capture the relationship between the predictor variables and target variable.

4 Conclusion

The features "Total Page Likes", "Paid" and "Type" have significant impact on "Lifetime Post Consumers". Hence, these features should be taken into consideration before publishing posts on social media. The principal components regression model performs the best when predicting lifetime post consumers of a post. However, the R squared value of approximately 0.4 and slight trend shown in the residuals vs true test values plot imply that the current regression model might not be the best way to understand the data. Thus, we should consider improving the model incorporating more features, including interaction terms, using segmentation models, using general linear models.

5 Appendix

5.1 Figures and Plots



(a) Equal Scale Plot of Lifetime Post Consumers vs Total Page Likes

(b) Average Lifetime Post Consumers vs Type

(c) Average Lifetime Post Consumers vs Category

(d) Average Lifetime Post Consumers vs Paid

(e) Average Lifetime Post Consumers vs Post Month

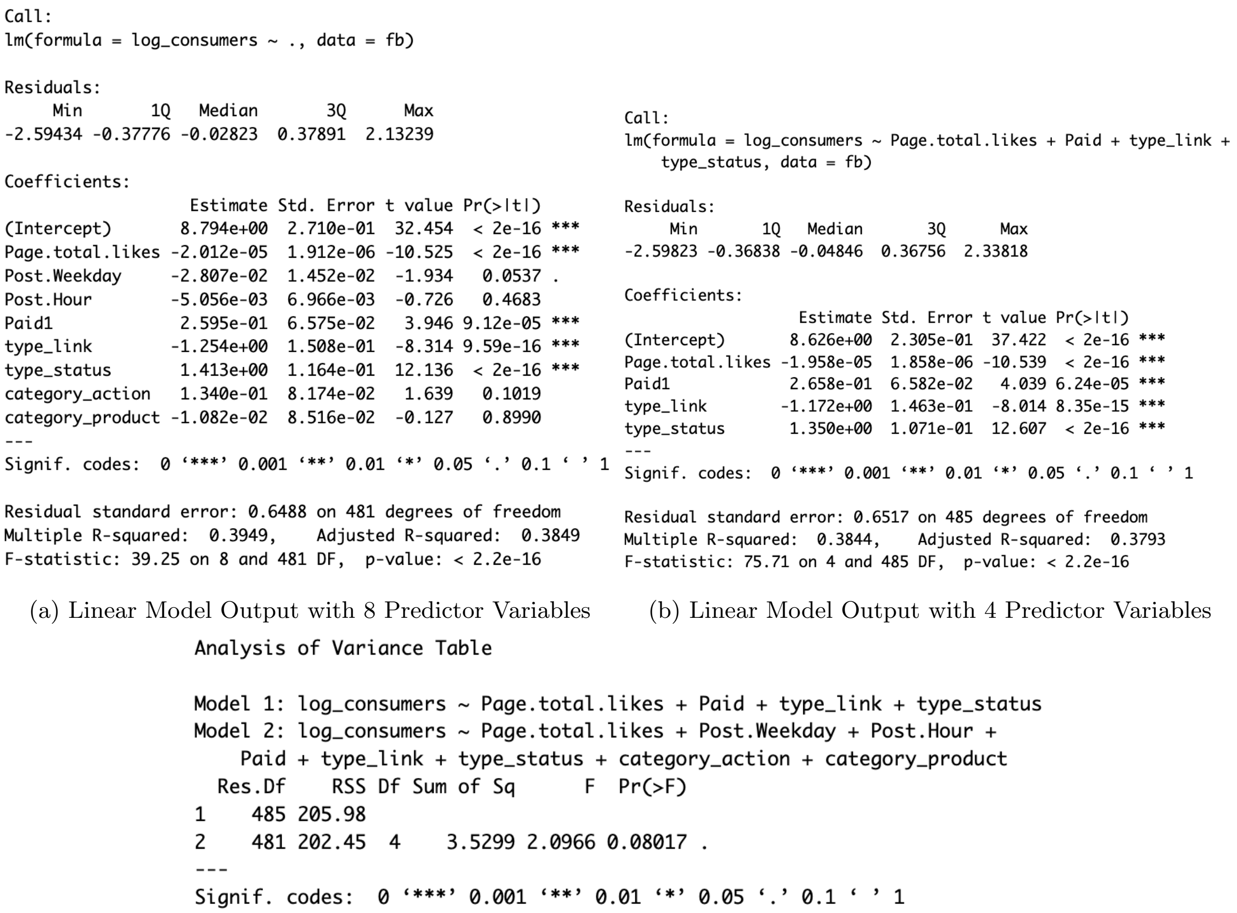
(f) Average Lifetime Post Consumers vs Post Day

(g) Average Lifetime Post Consumers vs Post Hour

(h) Histogram of Lifetime Post Consumers

(i) Histogram of log(Lifetime Post Consumers + 1)

Figure 1: Plots for Explanatory Data Analysis



```
Call:
lm(formula = log_consumers ~ ., data = fb)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.59434 -0.37776 -0.02823  0.37891  2.13239 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      8.794e+00  2.710e-01  32.454  < 2e-16 ***
Page.total.likes -2.012e-05  1.912e-06 -10.525  < 2e-16 ***
Post.Weekday     -2.807e-02  1.452e-02  -1.934   0.0537 .  
Post.Hour        -5.056e-03  6.966e-03  -0.726   0.4683    
Paid1             2.595e-01  6.575e-02   3.946 9.12e-05 ***
type_link        -1.254e+00  1.508e-01  -8.314 9.59e-16 ***
type_status       1.413e+00  1.164e-01  12.136  < 2e-16 ***
category_action   1.340e-01  8.174e-02   1.639   0.1019    
category_product -1.082e-02  8.516e-02  -0.127   0.8990    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6488 on 481 degrees of freedom
Multiple R-squared:  0.3949,	Adjusted R-squared:  0.3849 
F-statistic: 39.25 on 8 and 481 DF,  p-value: < 2.2e-16
```

(a) Linear Model Output with 8 Predictor Variables

```
Call:
lm(formula = log_consumers ~ Page.total.likes + Paid + type_link + 
    type_status, data = fb)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.59823 -0.36838 -0.04846  0.36756  2.33818 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      8.626e+00  2.305e-01  37.422  < 2e-16 ***
Page.total.likes -1.958e-05  1.858e-06 -10.539  < 2e-16 ***
Paid1             2.658e-01  6.582e-02   4.039 6.24e-05 ***
type_link        -1.172e+00  1.463e-01  -8.014 8.35e-15 ***
type_status       1.350e+00  1.071e-01  12.607  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6517 on 485 degrees of freedom
Multiple R-squared:  0.3844,	Adjusted R-squared:  0.3793 
F-statistic: 75.71 on 4 and 485 DF,  p-value: < 2.2e-16
```

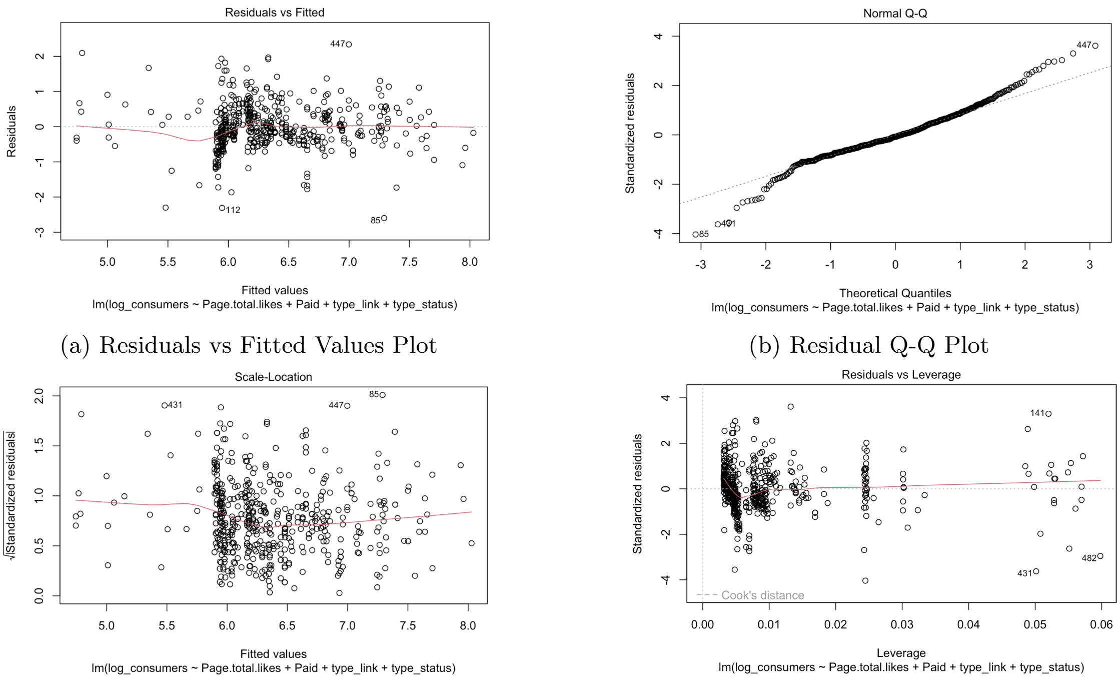
(b) Linear Model Output with 4 Predictor Variables

```
Analysis of Variance Table

Model 1: log_consumers ~ Page.total.likes + Paid + type_link + type_status
Model 2: log_consumers ~ Page.total.likes + Post.Weekday + Post.Hour + 
    Paid + type_link + type_status + category_action + category_product
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1    485 205.98                              
2    481 202.45  4    3.5299 2.0966 0.08017 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c) Analysis of Variance Table

Figure 2: Summary of Linear Models



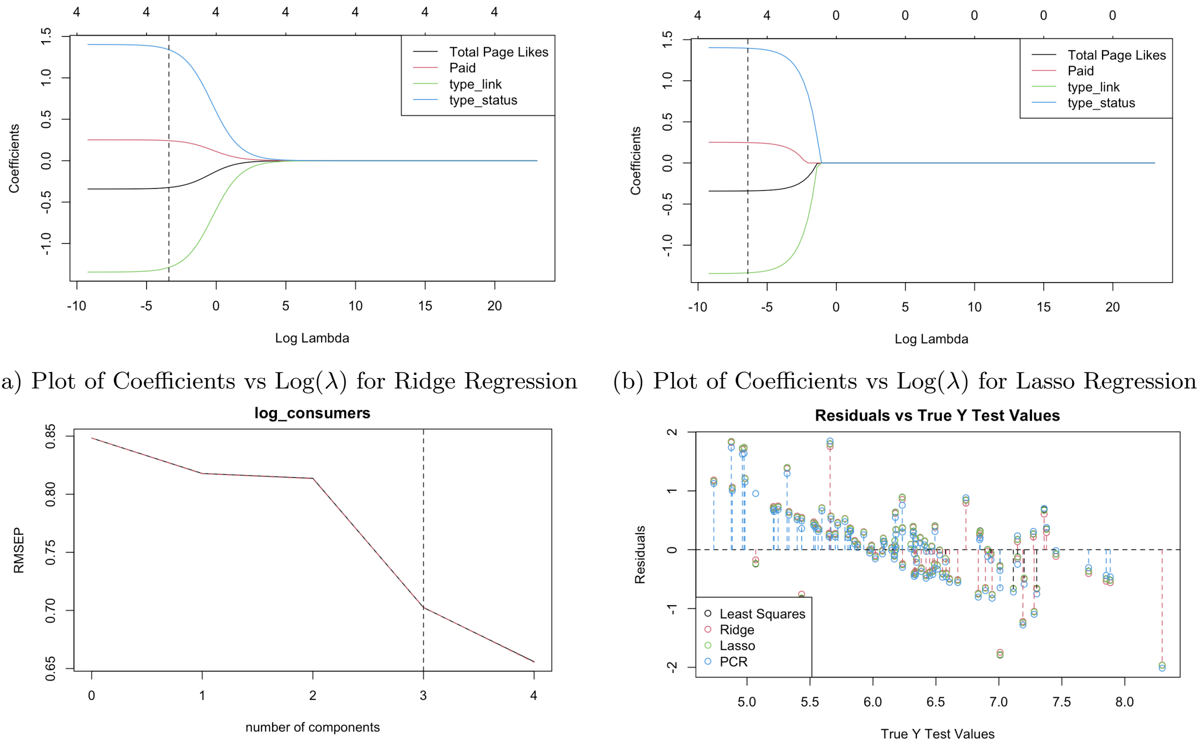
(a) Residuals vs Fitted Values Plot

(b) Residual Q-Q Plot

(c) Standardised Residuals vs Fitted Values Plot

(d) Standardised Residuals vs Leverage Plot

Figure 3: Diagnosis Plots



(a) Plot of Coefficients vs Log(λ) for Ridge Regression

(b) Plot of Coefficients vs Log(λ) for Lasso Regression

(c) Principal Component Regression Validation Plot

(d) Residuals vs True Y Test Values Plot

Figure 4: Regression Models Output

Model	MSE	Total Page Likes	Paid	type_link	type_status
Least Squares Regression	42.985	-0.342	0.252	-1.345	1.403
Ridge Regression	42.169	-0.324	0.242	-1.289	1.340
Lasso Regression	42.874	-0.340	0.248	-1.337	1.396
Principal Components Regression	39.152	-0.314	0.252	-0.112	1.501

Figure 5: Table of MSE and Coefficients for Each Model

References

Moro, S., Rita, P. & Vala, B. (2016), 'Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach', *Journal of Business Research* **69**(9), 3341–3351. **URL:** *https://doi.org/10.1016/j.jbusres.2016.02.010*