Analysis

Tom Yedwab

August 18, 2016

Read in data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
processed <- read.csv("~/source/berkeleymidsreview/export_data/processed.csv")
```

Data cleanup

```
# Parse some binaries so R recognizes them as such
processed$MobileBin = processed$Mobile. == "True"
processed$ScrolledBin = processed$Scrolled. == "True"
processed$DismissedBin = processed$Dismissed. == "True"
processed$DisclosureBin = processed$Show.Disclosure == "True"
processed$DialogBin = processed$Show.Dialog == "True"

# Pages viewed have to be at least 1, otherwise we wouldn't be recording data for the subject
processed[processed$Pages.Viewed == 0, "Pages.Viewed"] = 1

# Define a "bounce" as only viewing one page, no scrolling
processed$Bounce = processed$Pages.Viewed < 2 & !processed$ScrolledBin
```

Randomization check: These should be close to 0.5

```
mean(processed$DisclosureBin)
```

```
## [1] 0.4950593
```

```
mean(processed$DialogBin)
```

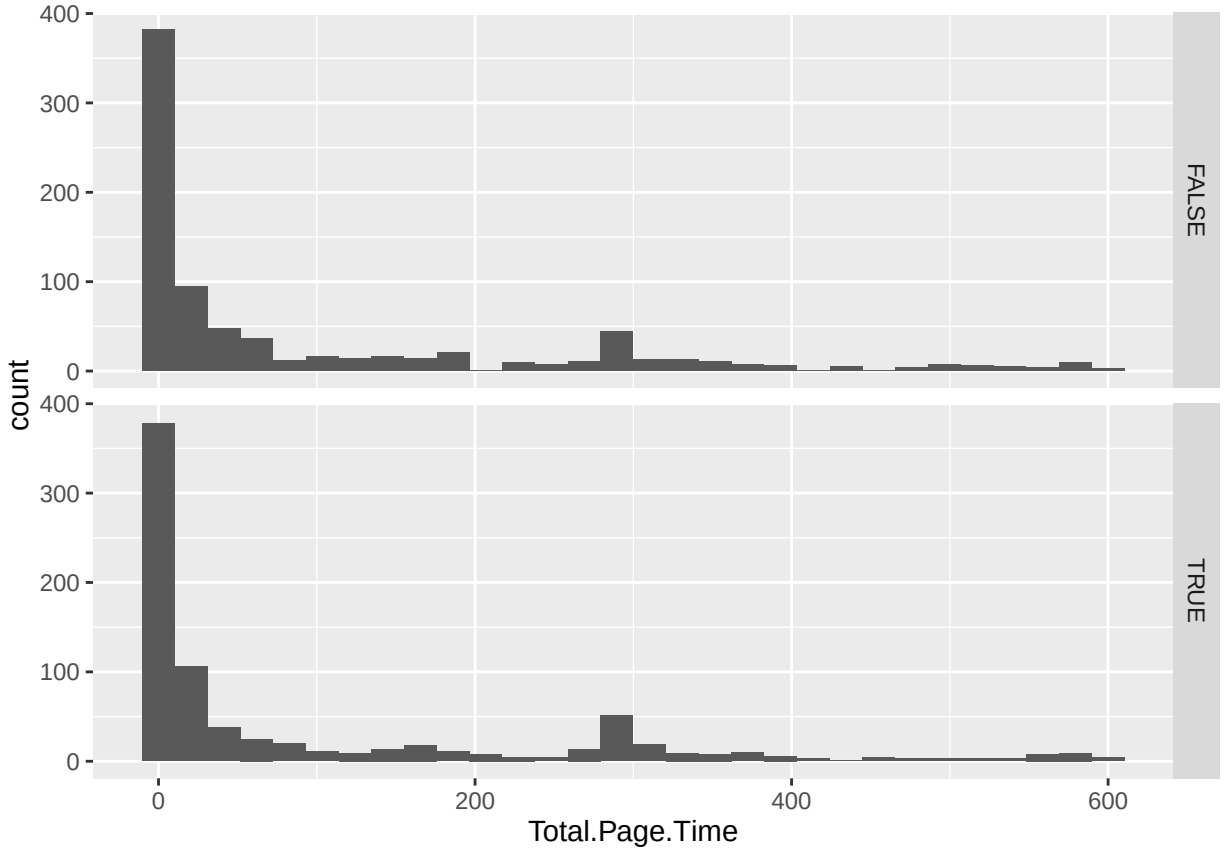
```
## [1] 0.4703557
```

Look at the distributions of outcome variables in treatment/control

Total Page Time:

```
processedSubset = processed[processed$Total.Page.Time <= 600,]
ggplot(processedSubset, aes(Total.Page.Time)) + geom_histogram() + facet_grid(DisclosureBin ~ .)
```

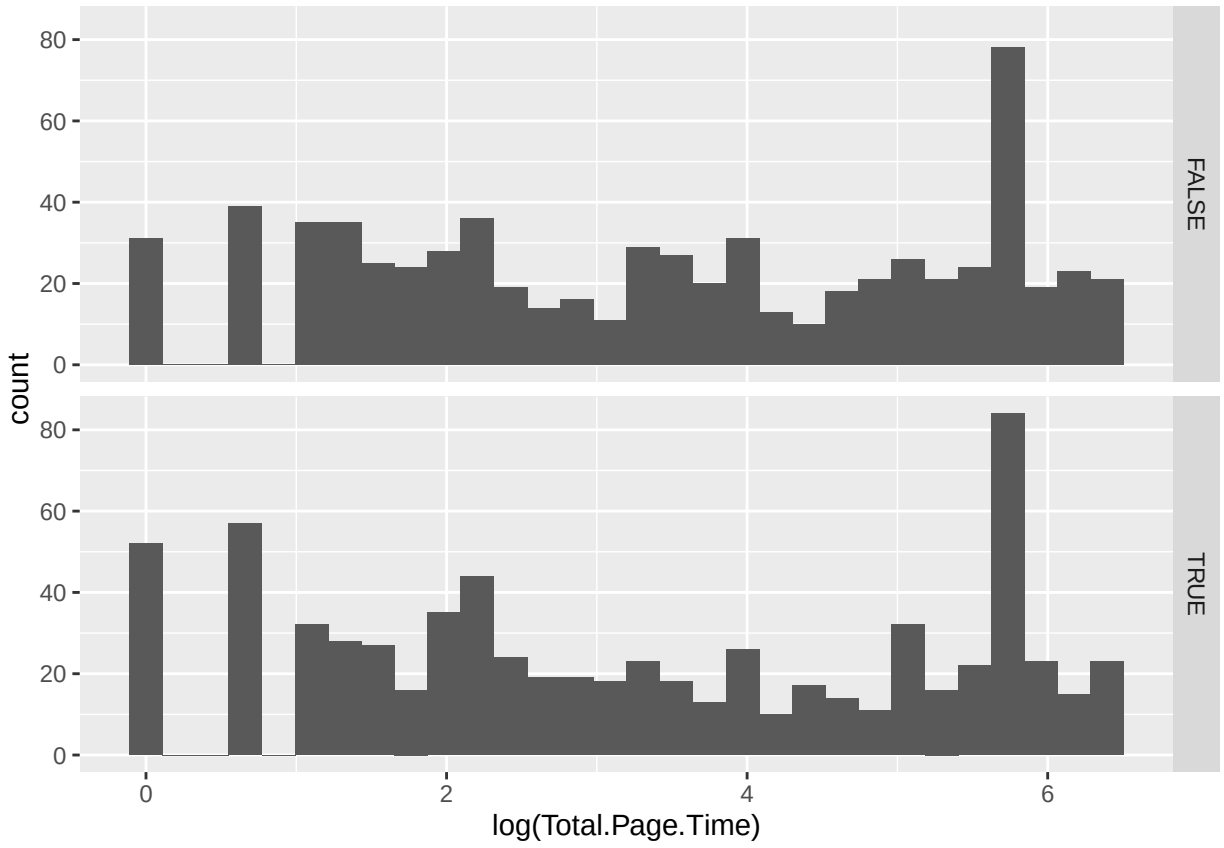
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



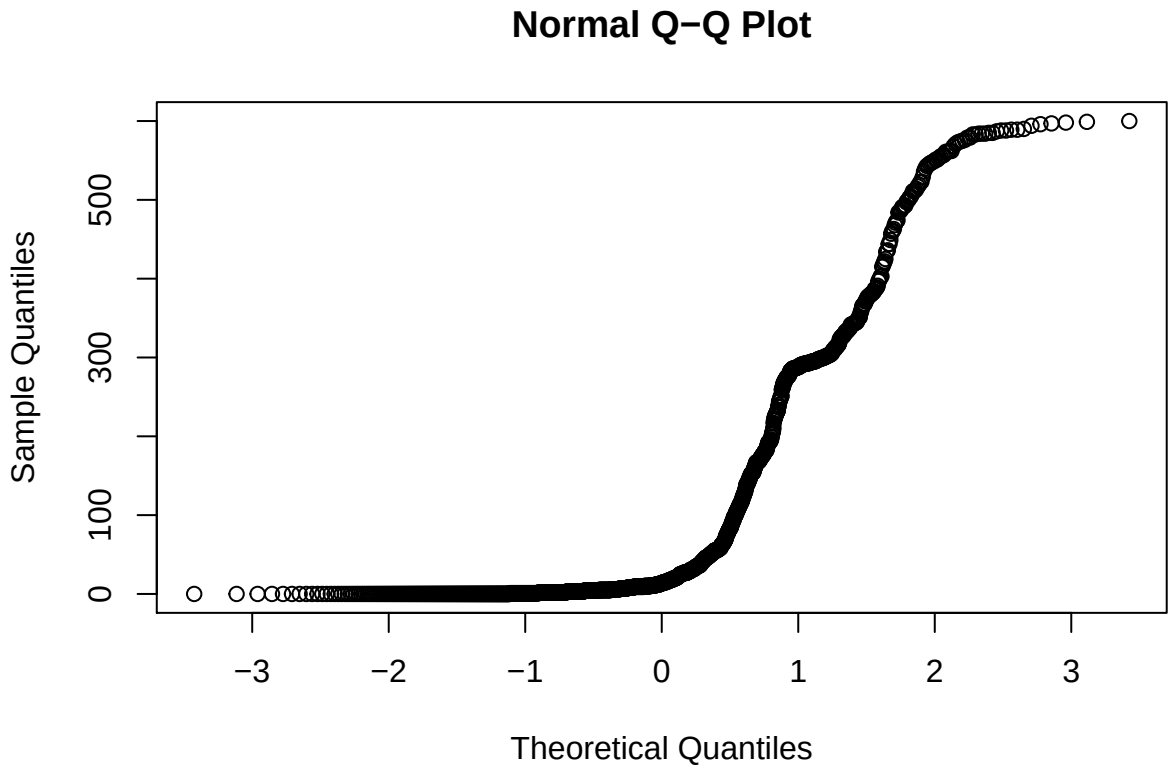

```
ggplot(processedSubset, aes(log(Total.Page.Time))) + geom_histogram() + facet_grid(DisclosureBin ~ .)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 216 rows containing non-finite values (stat_bin).
```

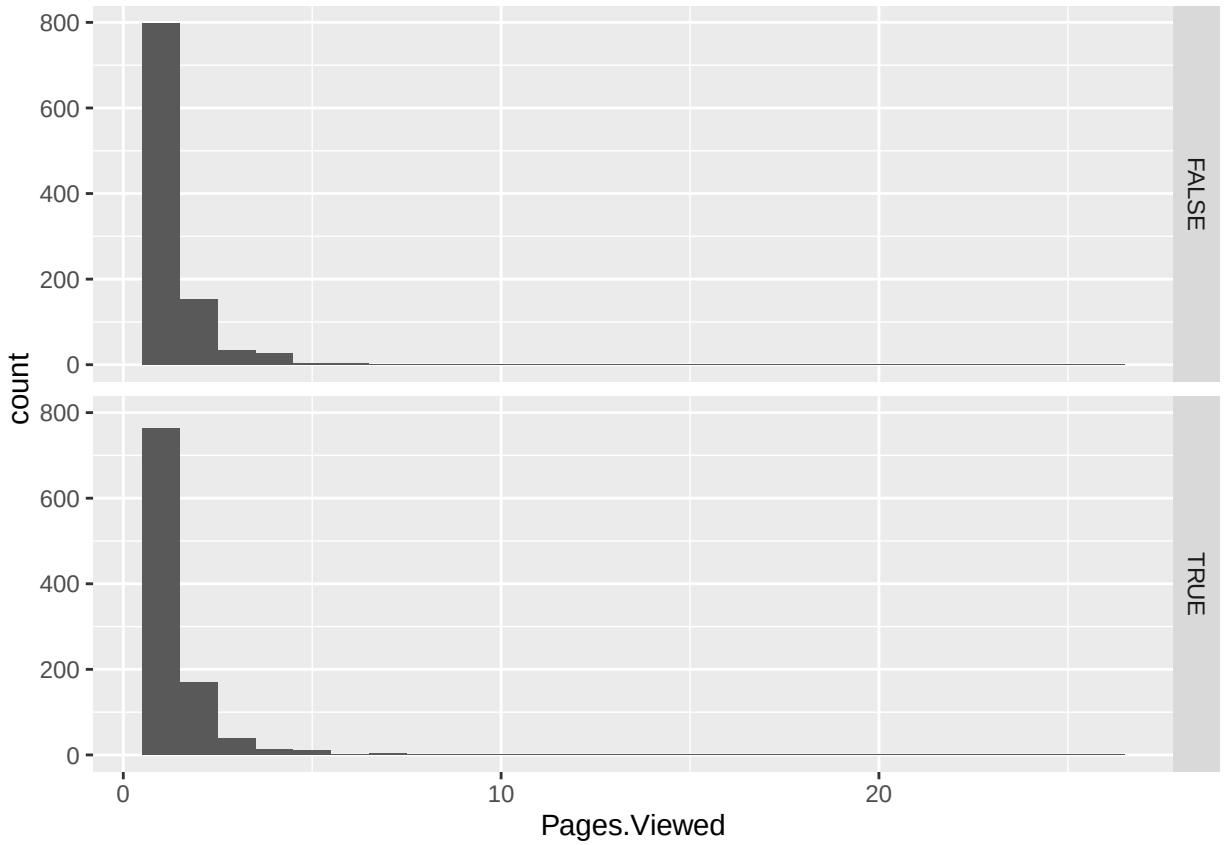



```
qqnorm(processedSubset$Total.Page.Time)
```

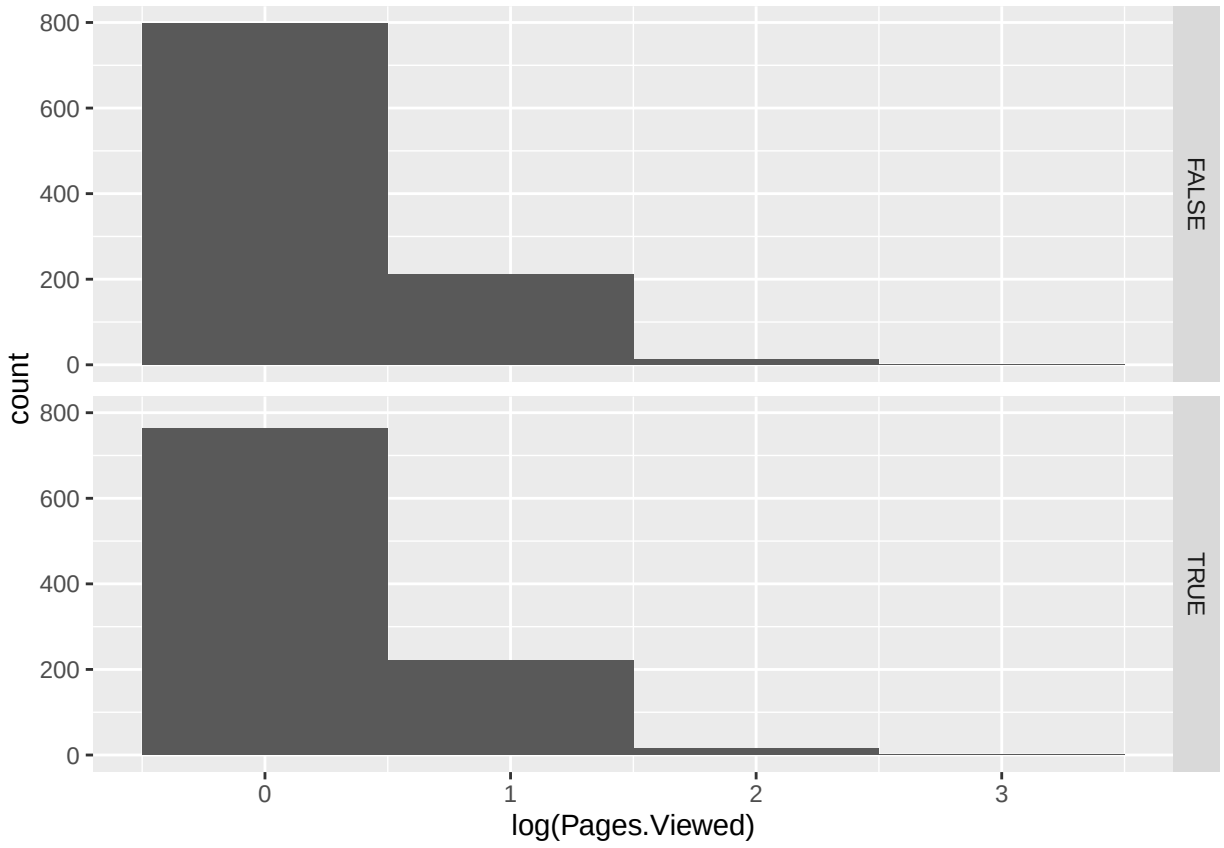


Pages Viewed:

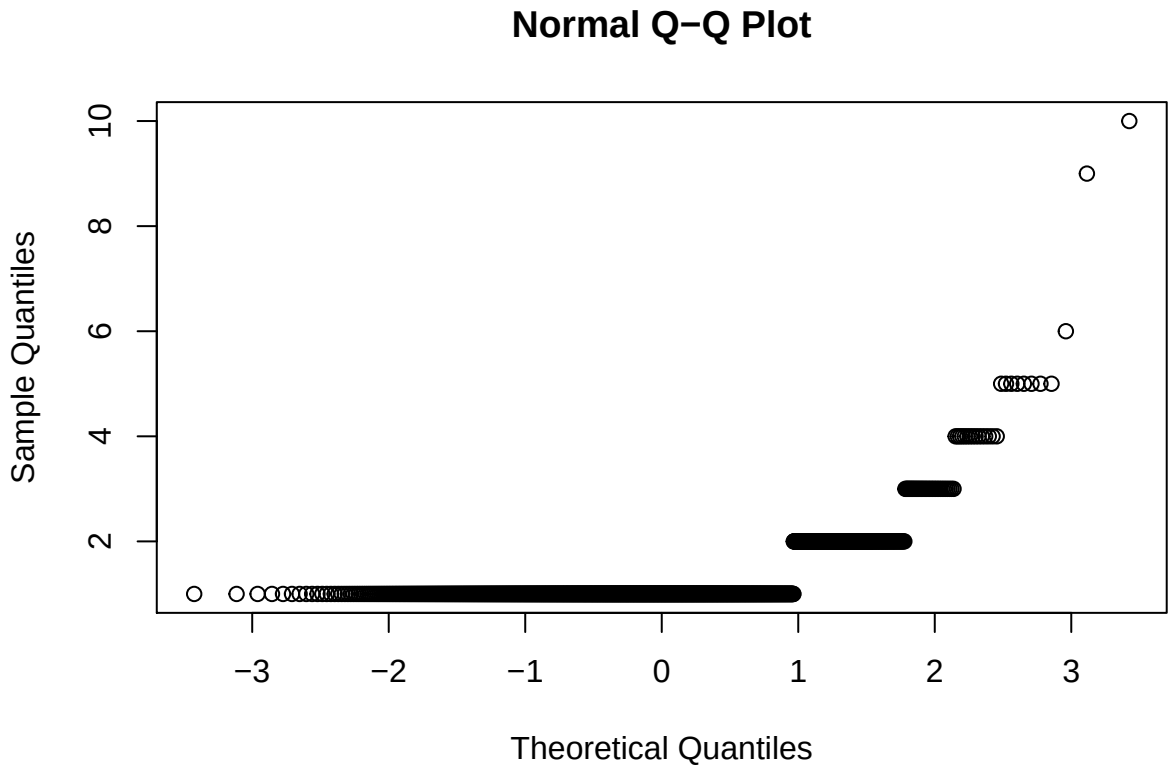
```
ggplot(processed, aes(Pages.Viewed)) + geom_histogram(binwidth=1) + facet_grid(DisclosureBin ~ .)
```




```
ggplot(processed, aes(log(Pages.Viewed))) + geom_histogram(binwidth=1) + facet_grid(DisclosureBin ~ .)
```




```
qqnorm(processedSubset$Pages.Viewed)
```



Estimate some effects

```
time.m = lm(Total.Page.Time ~ DisclosureBin + DialogBin +
              DisclosureBin * DialogBin + Source + MobileBin +
              Age + Gender, data=processed)
pages.m = lm(log(Pages.Viewed) ~ DisclosureBin + DialogBin +
               DisclosureBin * DialogBin + Source + MobileBin +
               Age + Gender, data=processed)
scrolled.m = glm(ScrolledBin ~ DisclosureBin + DialogBin +
                   DisclosureBin * DialogBin + Source + MobileBin +
                   Age + Gender, data=processed, family=binomial(link='logit'))
dismissed.m =  glm(DismissedBin ~ DisclosureBin + DialogBin +
                     DisclosureBin * DialogBin + Source + MobileBin +
                     Age + Gender, data=processed, family=binomial(link='logit'))
bounce.m =  glm(Bounce ~ DisclosureBin + DialogBin +
                     DisclosureBin * DialogBin + Source + MobileBin +
                     Age + Gender, data=processed, family=binomial(link='logit'))
```

```
summary(time.m)
```

```
##
## Call:
## lm(formula = Total.Page.Time ~ DisclosureBin + DialogBin + DisclosureBin *
##     DialogBin + Source + MobileBin + Age + Gender, data = processed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7559  -1103   -894   -491  77590
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       268.44     517.57   0.519  0.60405
## DisclosureBinTRUE                  90.82     273.16   0.332  0.73957
## DialogBinTRUE                     -77.18     280.34  -0.275  0.78311
## SourceFacebook                   -738.26     617.67  -1.195  0.23214
## SourceGoogle                     6512.00    2027.52   3.212  0.00134 **
## MobileBinTRUE                     702.34     473.78   1.482  0.13839
## Age18-24                         1206.11     692.46   1.742  0.08170 .
## Age25-54                         1249.38     691.49   1.807  0.07094 .
## GenderFemale                     -323.98     593.28  -0.546  0.58507
## GenderMale                       -543.31     594.61  -0.914  0.36098
## DisclosureBinTRUE:DialogBinTRUE   199.69     396.01   0.504  0.61415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4425 on 2013 degrees of freedom
## Multiple R-squared:  0.008863,   Adjusted R-squared:  0.003939
## F-statistic:   1.8 on 10 and 2013 DF,  p-value: 0.0557
```

```
summary(pages.m)
```

```
##
```

```
## Call:
## lm(formula = log(Pages.Viewed) ~ DisclosureBin + DialogBin +
##     DisclosureBin * DialogBin + Source + MobileBin + Age + Gender,
##     data = processed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40788 -0.21662 -0.19655 -0.09541  3.03438
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.095406   0.047635   2.003   0.0453 *
## DisclosureBinTRUE               0.006009   0.025140   0.239   0.8111
## DialogBinTRUE                  -0.029958   0.025802  -1.161   0.2457
## SourceFacebook                 -0.028000   0.056849  -0.493   0.6224
## SourceGoogle                    0.235910   0.186607   1.264   0.2063
## MobileBinTRUE                   0.093358   0.043606   2.141   0.0324 *
## Age18-24                        0.091379   0.063732   1.434   0.1518
## Age25-54                        0.054517   0.063643   0.857   0.3918
## GenderFemale                   -0.011622   0.054604  -0.213   0.8315
## GenderMale                     -0.018728   0.054727  -0.342   0.7322
## DisclosureBinTRUE:DialogBinTRUE  0.007150   0.036448   0.196   0.8445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4073 on 2013 degrees of freedom
## Multiple R-squared:  0.007528,   Adjusted R-squared:  0.002598
## F-statistic: 1.527 on 10 and 2013 DF,  p-value: 0.1235
```

```
summary(scrolled.m)
```

```
##
## Call:
## glm(formula = ScrolledBin ~ DisclosureBin + DialogBin + DisclosureBin *
##     DialogBin + Source + MobileBin + Age + Gender, family = binomial(link = "logit"),
##     data = processed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5982  -0.3930  -0.3603  -0.3159   2.5604
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.768409   0.496240  -5.579 2.42e-08 ***
## DisclosureBinTRUE               0.092769   0.225918   0.411  0.68134
## DialogBinTRUE                  -0.444777   0.264412  -1.682  0.09254 .
## SourceFacebook                  0.075180   0.558898   0.135  0.89300
## SourceGoogle                    3.218069   1.012295   3.179  0.00148 **
## MobileBinTRUE                   0.407912   0.467239   0.873  0.38265
## Age18-24                       -0.624339   0.596020  -1.048  0.29486
## Age25-54                       -0.190655   0.589098  -0.324  0.74621
## GenderFemale                    0.114942   0.547591   0.210  0.83374
## GenderMale                      0.294853   0.546176   0.540  0.58930
## DisclosureBinTRUE:DialogBinTRUE -0.006073   0.364294  -0.017  0.98670
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1023.21  on 2023  degrees of freedom
## Residual deviance:  999.89  on 2013  degrees of freedom
## AIC: 1021.9
##
## Number of Fisher Scoring iterations: 5
```

```
summary(dismissed.m)
```

```
##
## Call:
## glm(formula = DismissedBin ~ DisclosureBin + DialogBin + DisclosureBin *
##     DialogBin + Source + MobileBin + Age + Gender, family = binomial(link = "logit"),
##     data = processed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3042  -0.4727  -0.1627  -0.1336   3.1721
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -4.70422    0.72495  -6.489 8.64e-11 ***
## DisclosureBinTRUE              -0.09643    0.61245  -0.157   0.8749
## DialogBinTRUE                   2.48226    0.43775   5.671 1.42e-08 ***
## SourceFacebook                 -0.63770    0.57143  -1.116   0.2644
## SourceGoogle                    2.51513    1.16710   2.155   0.0312 *
## MobileBinTRUE                   0.38547    0.53469   0.721   0.4710
## Age18-24                        0.73414    0.76088   0.965   0.3346
## Age25-54                        1.04481    0.76018   1.374   0.1693
## GenderFemale                   -0.32775    0.55636  -0.589   0.5558
## GenderMale                     -0.70596    0.56443  -1.251   0.2110
## DisclosureBinTRUE:DialogBinTRUE  0.02981    0.64608   0.046   0.9632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 894.16  on 2023  degrees of freedom
## Residual deviance: 774.55  on 2013  degrees of freedom
## AIC: 796.55
##
## Number of Fisher Scoring iterations: 7
```

```
summary(bounce.m)
```

```
##
## Call:
## glm(formula = Bounce ~ DisclosureBin + DialogBin + DisclosureBin *
##     DialogBin + Source + MobileBin + Age + Gender, family = binomial(link = "logit"),
```

```
##     data = processed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2088  -1.4962   0.7730   0.8346   1.6013
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     2.168771   0.350537   6.187 6.13e-10 ***
## DisclosureBinTRUE              -0.106784   0.135663  -0.787  0.43121
## DialogBinTRUE                   0.179328   0.144342   1.242  0.21410
## SourceFacebook                  0.095474   0.323212   0.295  0.76770
## SourceGoogle                   -3.305146   1.177859  -2.806  0.00502 **
## MobileBinTRUE                  -1.006676   0.322944  -3.117  0.00183 **
## Age18-24                       -0.421475   0.379675  -1.110  0.26696
## Age25-54                       -0.374782   0.379222  -0.988  0.32301
## GenderFemale                    0.025161   0.311064   0.081  0.93553
## GenderMale                     -0.007184   0.311022  -0.023  0.98157
## DisclosureBinTRUE:DialogBinTRUE  0.115664   0.202315   0.572  0.56752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2377.8  on 2023  degrees of freedom
## Residual deviance: 2347.1  on 2013  degrees of freedom
## AIC: 2369.1
##
## Number of Fisher Scoring iterations: 4
```

Statistical power checks

```
simulate.regression <- function(mean.control, sd, treatment.effect, n) {
  tbl = data.frame(
    group=c(rep(0, n/2), rep(1, n/2)),
    val=c(rnorm(n/2, mean=mean.control, sd=sd),
          rnorm(n/2, mean=mean.control+treatment.effect, sd=sd)))
  model = lm(val ~ group, data=tbl)
  p.value = summary(model)$coefficients[2,4]
  return(p.value)
}

calculate.power <- function(mean.control, sd, treatment.effect, n) {
  p.value.distribution = replicate(
    10000,
    simulate.regression(mean.control, sd, treatment.effect, n))
  return(mean(p.value.distribution < 0.05))
}
```

What is our statistical power to detect a 10% increase/decrease in total page time?

```
cov.model = lm(Total.Page.Time ~ Source + MobileBin + Age + Gender, data=processed)
processed$Total.Page.Time.Prediction = predict(cov.model, processed)
processed$Total.Page.Time.Resid = processed$Total.Page.Time.Prediction - processed$Total.Page.Time

mean.value = mean(processed[processed$DisclosureBin==F,"Total.Page.Time"])
mean.control = mean(processed[processed$DisclosureBin==F,"Total.Page.Time.Resid"])
sd.control = sd(processed[processed$DisclosureBin==F,"Total.Page.Time.Resid"])
treatment.effect = mean.value * -0.1
time.power = calculate.power(mean.control, sd.control, treatment.effect,
                             nrow(processed))
print(time.power)
```

```
## [1] 0.0896
```

What is our statistical power to detect a 10% increase/decrease in pages viewed?

```
cov.model = lm(log(Pages.Viewed) ~ Source + MobileBin + Age + Gender, data=processed)
processed$Pages.Viewed.Prediction = predict(cov.model, processed)
processed$Pages.Viewed.Resid = processed$Pages.Viewed.Prediction - processed$Pages.Viewed

mean.value = mean(log(processed[processed$DisclosureBin==F,"Pages.Viewed"]))
mean.control = mean(processed[processed$DisclosureBin==F,"Pages.Viewed.Resid"])
sd.control = sd(processed[processed$DisclosureBin==F,"Pages.Viewed.Resid"])
# Increase of log(1.1) = 0.04
treatment.effect = mean.value * 0.04
pages.power = calculate.power(mean.control, sd.control, treatment.effect,
                              nrow(processed))
print(pages.power)
```

```
## [1] 0.0578
```

What is our statistical power to detect a 10% increase/decrease in bounce rate?

```
simulate.logistic.regression <- function(control.p, treatment.p, n) {
  tbl = data.frame(
    group=c(rep(0, n/2), rep(1, n/2)),
    val=c(runif(n/2, 0, 1) <= control.p,
          runif(n/2, 0, 1) <= treatment.p))
  model = glm(val ~ group, data=tbl, family=binomial(link='logit'))
  p.value = summary(model)$coefficients[2,4]
```

```
  return(p.value)
}

calculate.logistic.power <- function(control.p, treatment.p, n) {
  p.value.distribution = replicate(
    10000,
    simulate.logistic.regression(control.p, treatment.p, n))
  return(mean(p.value.distribution < 0.05))
}

control.p = mean(processed[processed$DisclosureBin==F,"Bounce"])
treatment.p = control.p * 1.10
bounce.power = calculate.logistic.power(control.p, treatment.p, nrow(processed))
print(bounce.power)
```

```
## [1] 0.9746
```