Advertisement Placement in Mobile Applications

Chase Baggett

Introduction

In 2015, Avazu, an online digital marketing company released 10 days of their data representing marketing compaigns within mobile platforms. It amounts to over 40 million ad views, and associated to that, a click through result of true or false. This data does not come from a conducted experiment, but from the real world, therefore is observational in nature. In addition to that, like the real world, the data is dirty.

It is made more complicated by data anonymization where category descriptions or even names have been hidden from the researchers for reasons of confidentiality. For instance, we are given a cagegorical banner position, but not where that position is, which means we cannot hypothesize from a position of knowledge about the similarity of any positions. We are also given categorical variables they believe are important, with no context. Despite this, it is a valuable insight into the world of online marketing, and the size and reach of the data provides interesting research opportunities.

To put it in context, companies pay to advertise within the platform, and have many options for advertising. The goal of advertising for many companies is to maximize click per view, or click-through-rate. Specifically for this exercise, I intend to try to understand the effect of advertisement position on the click-through rate, after adjustment for other variables.

Data

To create a dataset from the raw level clicks and views, I am recording the view and click count of advertiements over the 10 days. The data I am working with is raw web traffic over the 240 hours. Given I have assembled the data from raw clicks we could consider this data simulated, though it is drawn from the real world, I cannot be certain of many aspects of the design and am making some assumptions.

It is a crossed factoral observational data Every adertisement has an associated position that it was placed within the app. All factor levels are represented within each of the other factor levels randomly, though not all of the factor levels combinations were ever clicked. There is no nesting. It is an entirely random design. I am treating all variables as fixed becsause I have no need to generalize beyond the levels of the data. However, we a continuous variavble of importance, and therefore must test the slope to determine in an ANCOVA is needed.

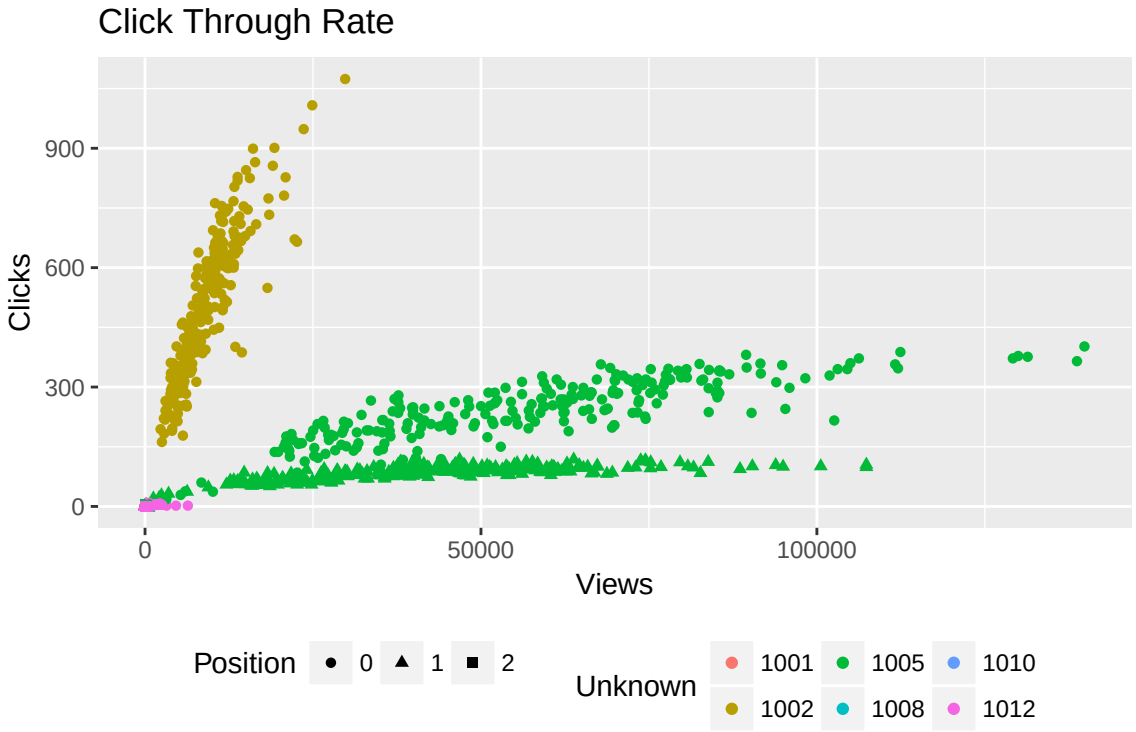
There are a total of four variables in the data.

Variable	Structure	Type	Lvls	Description
Clicks	Continuous	Response	NA	The nunber of times the advertisement was clicked.
Views	Continuous	Fixed	NA	The number of times a given advertisement was viewed.
Position	Factor	Fixed	3	The position on the site of the advertisement.
Unknown	Factor	Fixed	7	An unkown anonymized categorical variable.

Exploring the Data

The purpose of an advertiser is to maximize the numbers of clicks per view, or stated differently, they are trying to increase the slope between Views and Clicks. This lends itself naturally to an ANCOVA model with Views as a continuous predictor. Our primarily point of concern is finding factors that increase the slope between clicks and views, or industry terms, the click-through-rate.

As we can see below, there are some very clear and identifiable linear trends in the data, where slope and intercept both change, suggesting an interaction term will be necessary. For Unknown, Avazu would not tell the researchers what it meant, yet chose to include it in the data anyway. It has been been made entirely anonymous, and we therefore cannot attribute any specific meaning to it. Position represents one of three locations to place an advertisement within the mobile app.



Picking a Model

Slopes Equal to Zero

In order to know which kind of model to fit, we must first test the slope of Click to Views. If we establish the slope is zero, we can use an ANOVA model, but if the slope is non-zero, we must use an ANCOVA model.

For a function of $y = \beta_0 + \beta_1 Views$:

$H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$

As we can see below, with a near zero p-value, we reject the null hypothesis in favor the alternative that the slope is non-zero. This means we must use an ANCOVA model to account for the covariate.

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Views	1	9894694	9894694.31	363.2439	0
Residuals	2066	56277446	27239.81	NA	NA

All Slopes Equal by Position

Now, we have established the slope is non-zero, but we have not established that the slopes are different. We must use this to decide to use an equal or different slopes model. For a function of $y = \beta_0 + \beta_1 Views + \beta_2 Position + \beta_3 Views * Position$:

$H_0 : \beta_3 = 0$ vs $H_a : \beta_3 \neq 0$

As we can see below, with a near zero p-value, we reject the null hypothesis in favor the alternative that the slope are not equal, and therefore must use a different slopes model, because the interaction term is significant.

	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Views	1	9894694.3	9894694.31	444.03969	0.0000000
Position	2	9821046.5	4910523.24	220.36731	0.0000000
Views:Position	2	508120.2	254060.08	11.40134	0.0000119
Residuals	2062	45948279.4	22283.36	NA	NA

Unequal Slopes ANCOVA

I am using a generalized linear model of the following form to conduct the ANCOVA. Technically one of the factor levels will be "baked in" to the model, but I am including a beta for it here for statements on the hypothesis.

$y = \beta_0 + \beta_1 Views + \beta_2 Position_1 + \beta_3 Position_2 + \beta_4 Position_3 +$
$\beta_5 Unknown_1 + \beta_6 Unknown_2 + \beta_7 Unknown_3 + \beta_8 Unknown_4 +$
$\beta_9 Unknown_5 + \beta_10 Unknown_6 + \beta_{11} Unknown_7 +$
$Views \times ($
$\beta_{12} Position_1 + \beta_{13} Position_2 + \beta_{14} Position_3 + \beta_{15} Unknown_1 + \beta_{16} Unknown_2 +$
$\beta_{17} Unknown_3 + \beta_{18} Unknown_4 + \beta_{19} Unknown_5 + \beta_{20} Unknown_6 + \beta_{21} Unknown_7$
$)$

The Type 3 Test of Fixed Effects is Below. The model has added significant terms for the Unknown category, as well as the interraction term.

	SS	DF	F	P
(Intercept)	0.2412948	1	0.0002007	0.9886990
Views	6.3304147	1	0.0052647	0.9421649
Position	51800.4729386	2	21.5398731	0.0000000
Unknown	1205651.2732322	5	200.5354313	0.0000000
Views:Position	982152.5066556	2	408.4024562	0.0000000
Views:Unknown	4993376.0235274	5	830.5459770	0.0000000
Residuals	2467390.8211141	2052	NA	NA

Hypothesis Tests

This model has Five Hypothesis Tests.

Test for Intercept

$H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$

Outcome: Reject the Null in Favor of the Alternative

	SS	DF	F	P
(Intercept)	0.2412948	1	0.0002007	0.988699

Test For Slope Across Factors

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Outcome: Fail to Reject the Null. We could think of this as no slope between click and views in common across the factors. The interaction terms are significant so it stays in the model.

	SS	DF	F	P
Views	6.330415	1	0.0052647	0.9421649

Test for Position Mean

At Least One Inequality Satisfies rejection of the null

$H_0 : \beta_2 = \beta_3 = \beta_4$ vs $H_0 : \beta_2 \neq \beta_3 \neq \beta_4$

Outcome: Reject the null in favor of the alternative.

	SS	DF	F	P
Position	51800.47	2	21.53987	0

Test for Unknown Category Mean

At Least One Inequality Satisfies rejection of the null

$H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_10 = \beta_{11}$ vs $H_a : \beta_5 \neq \beta_6 \neq \beta_7 \neq \beta_8 \neq \beta_9 \neq \beta_10 \neq \beta_{11}$

Outcome: Reject the null in favor of the alternative.

	SS	DF	F	P
Unknown	1205651	5	200.5354	0

Test for Difference in Slope by Position

At Least One Inequality Satisfies rejection of the null

$H_0 : \beta_{12} = \beta_{13} = \beta_{14}$ vs $H_0 : \beta_{12} \neq \beta_{13} \neq \beta_{14}$

Outcome: Reject the null in favor of the alternative.

	SS	DF	F	P
Views:Position	982152.5	2	408.4025	0

Test for Difference in Slope by Unknown Category

At Least One Inequality Satisfies rejection of the null

$H_0 : \beta_{15} = \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = \beta_{20} = \beta_{21}$ vs $H_a : \beta_{15} \neq \beta_{16} \neq \beta_{17} \neq \beta_{18} \neq \beta_{19} \neq \beta_{20} \neq \beta_{21}$

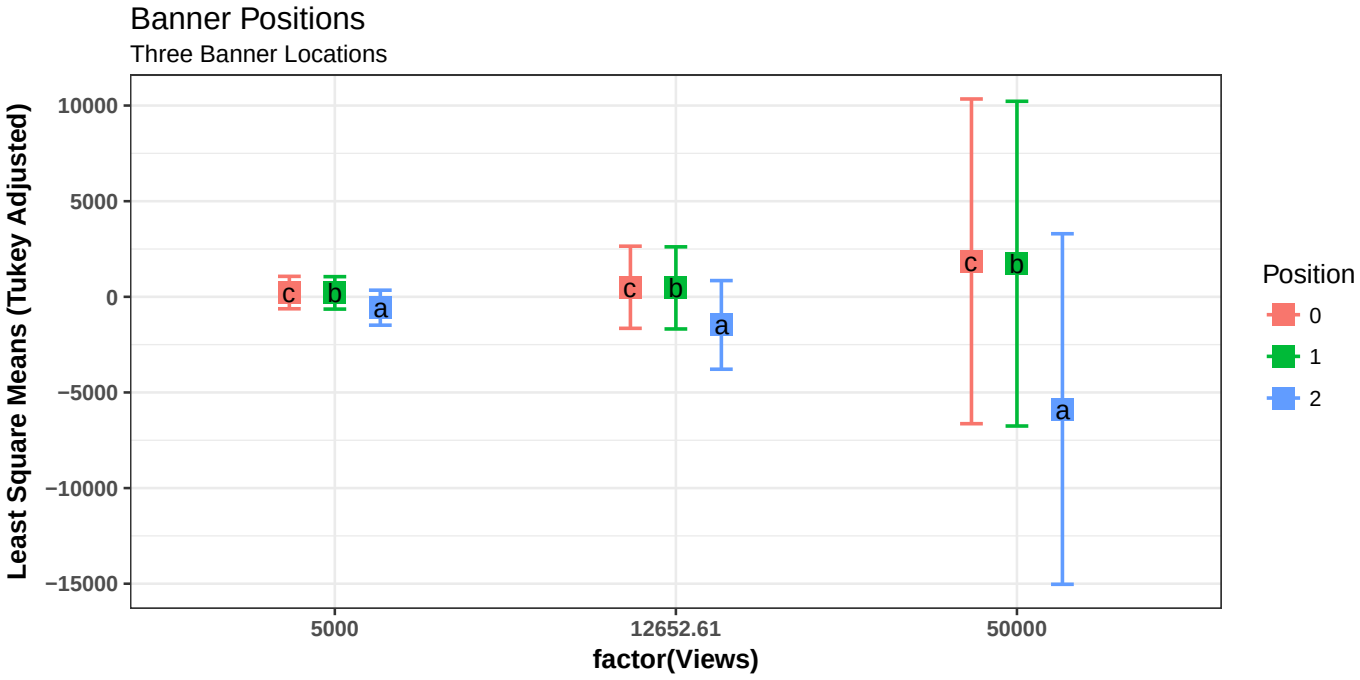
Outcome: Reject the null in favor of the alternative.

	SS	DF	F	P
Views:Unknown	4993376	5	830.546	0

Mean Comparisons

Position

Because I have a covariate and my primary area of interest in the interaction between views and clicks, I am doing a Tukey means comparison at 3 points along the values. As Views increases they each start to separate into their own groups. At high numbers of views we can establish a signficiant difference between all 3 positions, with banner position 0 having the highest mean. The middle View amount represents the mean for the data.



Conclusion

We've established that clicks do increase with views, and that position as well as our unknown category both effect the mean as well as the slope between click and views. Considering the unknown category as a fixed and unchangeable variable, we can provide guidance to advertisers that banner position 0 has a significantly higher click-through-rate as compared to banner position 1 and 2. Knowing the most beneficial advertisement within a mobile application could provide considerible value to advertisers.

Appendix

Location of the Raw Data: https://www.kaggle.com/c/avazu-ctr-prediction Full Code and Aggregated Data on Github: https://github.com/cbagg/avazu_click_through

The entire project was performed in R.

The entire document is reproducible via the R code and included data on github.

Notes

I conducted my study only on app category 07d7df22 and banner_position 0,1,2, taking the others out of the study because they were so rare it created missing data in the comnbinations. Over 35 milllion clicks remained. I aggredated those clicks into a dataset with 2,068 records, once for each application. Because it is click log data I could have reobserved the data at the app, site, or other levels, and chose to observe it at the application level because that seemed most logical to me. I consider this data more akin to a video feed– there are many things I could have observed from it because it is extremely granular.

I could not test the interaction term between Views,Unknown and Position because some combinations had only a click of zero.

Table for Tukey Means

```
knitr::kable(data.frame(pos_means_letters), row.names = F)
```

Position	Views	lsmean	SE	df	lower.CL	upper.CL	.group
2	5000.00	-568.1410	330.2460	2052	-1482.4092	346.1273	a
1	5000.00	206.1607	306.1976	2052	-641.5309	1053.8524	b
0	5000.00	221.9675	306.1820	2052	-625.6809	1069.6160	c
2	12652.61	-1469.2586	837.0145	2052	-3786.4883	847.9710	a
1	12652.61	465.9629	775.5229	2052	-1681.0306	2612.9565	b
0	12652.61	499.6191	775.5167	2052	-1647.3572	2646.5954	c
2	50000.00	-5867.0250	3310.2433	2052	-15031.2557	3297.2056	a
1	50000.00	1733.8878	3066.0297	2052	-6754.2506	10222.0262	b
0	50000.00	1854.6550	3066.0451	2052	-6633.5260	10342.8361	c

GLM Model Coefficients

```
knitr::kable(tidy(ancova_model_full))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0523957	3.6987225	0.0141659	0.9886990
Views	0.0217246	0.2994097	0.0725581	0.9421649
Position1	-4.1445343	3.1944955	-1.2973987	0.1946399
Position2	-19.9339072	3.1642592	-6.2997075	0.0000000
Unknown1002	168.1734825	6.3674544	26.4114151	0.0000000
Unknown1005	43.8113967	4.7765092	9.1722627	0.0000000
Unknown1008	20.6083672	5.5077826	3.7416813	0.0001878
Unknown1010	0.2469975	5.0802858	0.0486188	0.9612278
Unknown1012	10.1922750	4.3928107	2.3202172	0.0204268
Views:Position1	-0.0023325	0.0000821	-28.4138484	0.0000000
Views:Position2	-0.1540349	0.0306640	-5.0233122	0.0000006
Views:Unknown1002	0.0143112	0.2994101	0.0477981	0.9618818
Views:Unknown1005	-0.0184191	0.2994097	-0.0615180	0.9509527
Views:Unknown1008	0.1334821	0.3029541	0.4406017	0.6595478
Views:Unknown1010	-0.0167563	0.3650331	-0.0459035	0.9633916
Views:Unknown1012	-0.0252738	0.2994252	-0.0844078	0.9327404