Regression Analysis of A/B test

Motivation

It is challenging to to analyze experiments in a unified and flexible way. For example, click through or sign-up follow binomial distribution, orders per user per week follow Poisson distribution and the amount of money spent on Ritual follow Gaussian distribution (after normalization since it should have negative values). So, it is difficult to choose the correct statistical test for analysis.

Secondly, it is difficult to measure cohort effect in terms different sign-up location (metro_id) and sign-up time(weekly cohort). Running different test within each group may loose statistical significance or may run into Simpson's Paradox in the very end.

In order to solve the problem, I proposed alternative approach, regression models. That is to say, we may try to use logistic regression to solve binomial statistical inference, use Poisson regression and simple linear regression to solve other problems accordingly.

To illustrate my point more, I used the following fake data for illustration purposes.

Logistic Regression

I created some fake click data by binomial distribution and run a logistic regression. Regress city, sign-up channel, tenure and experiment id on click.

```
## -- Attaching packages ---------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Parsed with column specification:
## cols(
##   user_id = col_character(),
##   city = col_character(),
##   tenure = col_double(),
##   channel = col_double(),
##   opu_6_week = col_double()
## )
```

```
## Observations: 16,000
## Variables: 7
## $ user_id    <chr> "31c0fadd6066ab12e90fa931abbb649f7f591cd7927b9bbc76...
## $ city       <chr> "STL", "EDM", "CAL", "MSY", "SAN", "HAL", "STL", "M...
## $ tenure     <dbl> 35, 4, 25, 68, 17, 0, 46, 53, 4, 3, 25, 28, 5, 0, 4...
## $ channel    <chr> "14", "15", "15", "14", "2", "10", "10", "5", "10",...
## $ opu_6_week <dbl> 0.50, 0.50, 0.67, 0.83, 0.33, 0.17, 0.00, 0.00, 0.1...
## $ experiment <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ click      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

by noticing experiment's regression coefficient's p-value is less than 10^{-10}, we may conclude with high statistical confidence that in this case, the experiment drastically changed users' click through behavior across all metro and channel.

Observations	16000
Dependent variable	click
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(46)$	93.43
Pseudo-R² (Cragg-Uhler)	0.02
Pseudo-R² (McFadden)	0.02
AIC	5342.96
BIC	5703.93

Linear Regression

To analyze the average opu for a user in the past 6 weeks, we may use linear regression since the variable should be continuous. Before running regression, centering and scaling have been applied.

```
user$opu_6_week <-  scale(user$opu_6_week)
model2 <- glm(opu_6_week ~ city + tenure + channel + experiment, family = gaussian, data = user)
summ(model2)
```

This time, experiment's coefficient is not significant anymore and we may conclude that this feature haven't been changed(deteriorate) by the experiment.

	Est.	S.E.	z val.	p
(Intercept)	-3.64	0.46	-7.92	0.00
cityBOS	-0.07	0.48	-0.15	0.88
cityCAL	1.05	0.84	1.24	0.22
cityCHI	0.08	0.43	0.19	0.85
cityCLT	0.01	0.72	0.01	0.99
cityCVG	0.21	0.66	0.32	0.75
cityDAL	-0.76	0.82	-0.92	0.36
cityDEN	-0.18	0.65	-0.27	0.79
cityDET	0.62	0.52	1.21	0.23
cityEDM	0.57	0.84	0.69	0.49
cityGNV	-13.06	1199.74	-0.01	0.99
cityHAL	0.31	1.10	0.28	0.78
cityHOU	0.00	0.54	0.01	0.99
cityLAX	0.09	0.46	0.20	0.84
cityLON	0.08	0.43	0.18	0.86
cityMEL	-0.52	0.71	-0.72	0.47
cityMES	0.33	0.55	0.61	0.54
cityMIA	-13.02	619.55	-0.02	0.98
cityMKE	-0.49	0.83	-0.59	0.56
cityMSY	0.47	0.83	0.56	0.58
cityNYC	0.01	0.43	0.01	0.99
cityOTT	-0.15	0.53	-0.28	0.78
cityPDX	-0.01	0.83	-0.02	0.99
cityPHI	0.16	0.52	0.30	0.76
citySAN	-13.07	2399.54	-0.01	1.00
citySEA	0.24	0.53	0.45	0.65
citySFO	0.24	0.50	0.47	0.64
citySTL	-13.05	453.30	-0.03	0.98
citySYD	0.14	0.49	0.28	0.78
cityTOR	0.10	0.49	0.20	0.84
tenure	0.00	0.00	0.04	0.97
channel10	0.10	0.22	0.45	0.65
channel11	-13.31	414.43	-0.03	0.97
channel12	-0.06	0.50	-0.13	0.90
channel13	0.13	0.29	0.46	0.65
channel14	0.08	0.26	0.32	0.75
channel15	0.10	0.21	0.48	0.63
channel16	-13.60	599.14	-0.02	0.98
channel2	0.15	0.23	0.63	0.53
channel3	1.25	0.79	1.58	0.11
channel4	0.04	0.77	0.05	0.96
channel5	0.32	0.38	0.84	0.40
channel6	-13.01	758.80	-0.02	0.99
channel7	-13.63	2399.54	-0.01	1.00
channel8	-13.67	1199.46	-0.01	0.99
channel9	0.70	0.45	1.56	0.12
experiment	0.60	0.24	2.56	0.01

Standard errors: MLE

Observations	16000
Dependent variable	opu_6_week
Type	Linear regression

$\chi^2(46)$	927.80
Pseudo-R^2 (Cragg-Uhler)	0.06
Pseudo-R^2 (McFadden)	0.02
AIC	44545.18
BIC	44913.84

	Est.	S.E.	t val.	p
(Intercept)	-0.23	0.08	-2.99	0.00
cityBOS	-0.04	0.08	-0.47	0.64
cityCAL	0.72	0.20	3.54	0.00
cityCHI	0.13	0.07	1.79	0.07
cityCLT	0.42	0.12	3.66	0.00
cityCVG	0.22	0.11	1.97	0.05
cityDAL	-0.10	0.10	-1.00	0.32
cityDEN	0.10	0.10	1.00	0.32
cityDET	0.27	0.09	2.92	0.00
cityEDM	0.40	0.17	2.40	0.02
cityGNV	0.08	0.49	0.16	0.88
cityHAL	0.16	0.20	0.77	0.44
cityHOU	-0.01	0.09	-0.16	0.87
cityLAX	0.07	0.08	0.90	0.37
cityLON	-0.13	0.07	-1.79	0.07
cityMEL	0.57	0.10	5.73	0.00
cityMES	0.04	0.10	0.47	0.64
cityMIA	0.36	0.26	1.40	0.16
cityMKE	-0.17	0.11	-1.51	0.13
cityMSY	0.17	0.16	1.09	0.28
cityNYC	0.05	0.07	0.67	0.50
cityOTT	0.01	0.09	0.13	0.89
cityPDX	0.43	0.13	3.25	0.00
cityPHI	0.05	0.10	0.49	0.62
citySAN	0.08	0.97	0.08	0.93
citySEA	0.04	0.10	0.45	0.65
citySFO	0.07	0.09	0.83	0.41
citySTL	0.44	0.20	2.24	0.02
citySYD	0.29	0.09	3.39	0.00
cityTOR	0.22	0.09	2.59	0.01
tenure	0.00	0.00	9.39	0.00
channel10	0.12	0.04	2.96	0.00
channel11	0.29	0.17	1.67	0.10
channel12	-0.10	0.09	-1.13	0.26
channel13	0.03	0.05	0.61	0.54
channel14	0.10	0.05	2.06	0.04
channel15	-0.01	0.04	-0.27	0.79
channel16	0.09	0.25	0.37	0.71
channel2	0.20	0.04	4.65	0.00
channel3	4.33	0.25	17.05	0.00
channel4	0.34	0.15	2.23	0.03
channel5	-0.11	0.08	-1.50	0.13
channel6	-0.44	0.31	-1.41	0.16
channel7	-0.67	0.97	-0.69	0.49
channel8	-0.30	0.49	-0.61	0.54
channel9	0.20	0.10	1.92	0.06
experiment	-0.00	0.05	-0.01	0.99

Standard errors: MLE