第二次作业

刘琪

2024-11-25

目录

1 Question #1: BigBangTheory. (Attached Data: BigBangTheory)

a. 观众人数的最小值为 13.3, 观众人数的最大值为 16.5。

b. 平均数为 15.0428571; 中位数为 15; 众数为 13.6。

c. 第一四分位数为 14.1; 第三四分位数为 16。

d. 从下图可中，我们无法发现 2011-2012 季度观众人数变化的任何趋势。

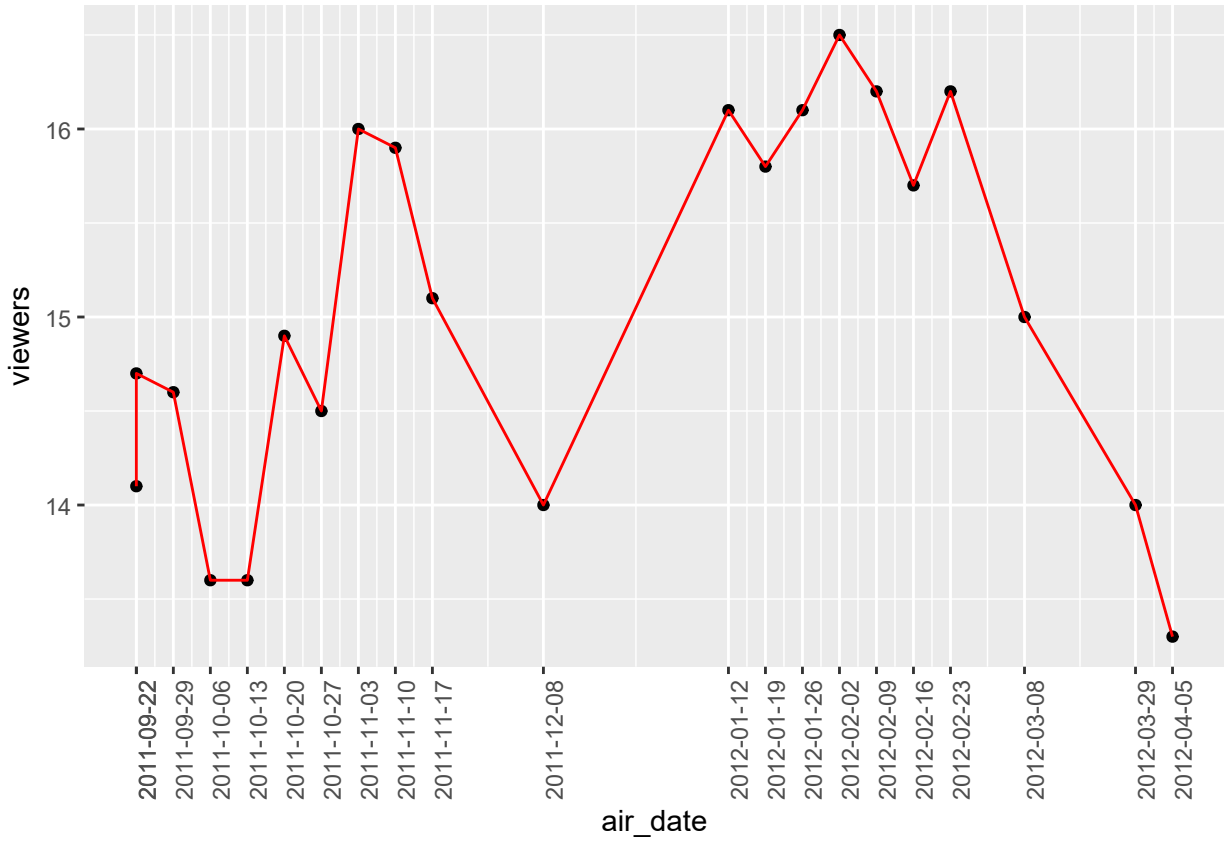


图 1: Plot between date and viewers

2 Question #2: NBAPlayerPts. (Attached Data: NBAPlayerPts)

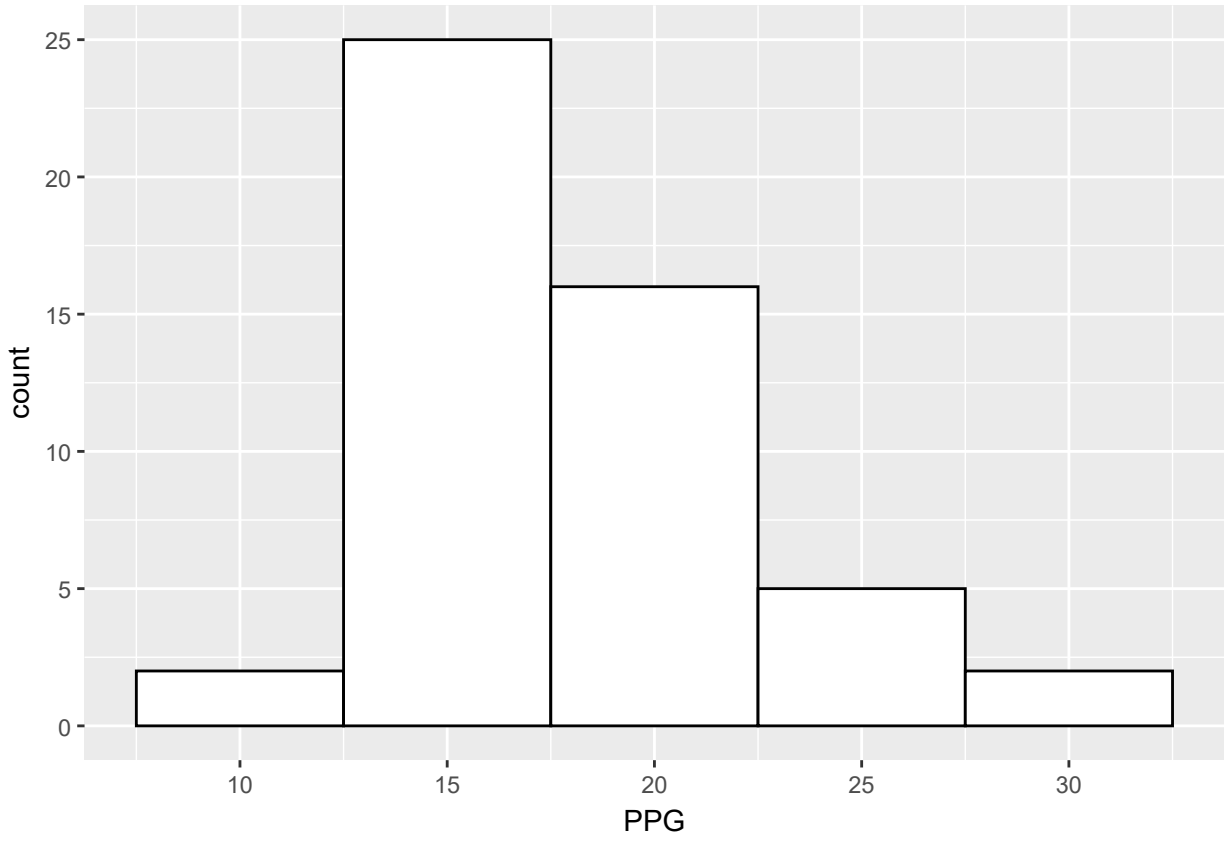
```
#> [10,12] (12,14] (14,16] (16,18] (18,20] (20,22] (22,24] (24,26] (26,28] (28,30]
#>    0.02    0.10    0.22    0.62    0.78    0.86    0.90    0.90    0.96    1.00
```

a. 频率分布: 1, 4, 6, 20, 8, 4, 2, 0, 3, 2

b. 相对频率分布: 0.02, 0.08, 0.12, 0.4, 0.16, 0.08, 0.04, 0, 0.06, 0.04

c. 累积百分比频率分布: \n 0.02, 0.1, 0.22, 0.62, 0.78, 0.86, 0.9, 0.9, 0.96, 1

d. 平均得分直方图:



e. 数据看起来向右偏斜, 因为分布的尾部在右侧延伸。

f. 1-78% = 22%.

3 Question #3

a. 这次调查中使用的样本大小为 625。

b. 样本量很大，所以样本均值的分布是正态分布。点估计在总体均值 ±25 以内的概率是 0.7887005.。

4 Question #4:Young Professional Magazine (Attached Data: Professional)

管理报告：a. 描述性统计数据如下。

skim_type	skim_variable	n_missing	complete_rate	factor.ordered	factor.n_unique	factor.top_counts
factor	gender	0	1	FALSE	2	Mal: 229, Fem: 181

factor	real_estate	0	1	FALSE	2	No: 229, Yes: 181
factor	has_broadband	0	1	FALSE	2	Yes: 256, No: 154
factor	have_children	0	1	FALSE	2	Yes: 219, No: 191
numeric	age	0	1	NA	NA	NA
numeric	investments	0	1	NA	NA	NA
numeric	num_trans	0	1	NA	NA	NA
numeric	income	0	1	NA	NA	NA

b. 为订阅者的平均年龄 95% 置信区间为 30 岁至 31 岁；家庭收入 95% 置信区间为 71079 美元至 77840 美元。

```
#> 95% CI for average age: 29.72153 30.50286
```

```
#> 95% CI for average income: 71079.26 77839.77
```

c. 家中有宽带接入的订阅者比例 95% 置信区间为 58% 至 67%；有孩子的订阅者比例 95% 置信区间为 48% 至 58%。

```
#> # A tibble: 1 x 6
#>   statistic chisq_df     p_value alternative lower_ci upper_ci
#>       <dbl>    <int>       <dbl> <chr>          <dbl>    <dbl>
#> 1      24.9        1 0.000000610 two.sided      0.575    0.671
```

```
#> # A tibble: 1 x 6
#>   statistic chisq_df p_value alternative lower_ci upper_ci
#>       <dbl>    <int>   <dbl> <chr>          <dbl>    <dbl>
#> 1      1.78        1   0.182 two.sided      0.485    0.583
```

d. 我认为这本杂志是在线经纪人的良好广告渠道，有数据可以发现几乎全部的订阅者除了他们的房产外还有金融投资，平均金额达到 2.8538293×10^4 美元，最高者达到了 133400 美元。其次是股票、债券和共同基金的交易数量，几乎全部的订阅者都有过交易次数，平均每年大约是 5.9731707 次，而有些订阅者的交易数量远超这个数字。

e. 我认为这本杂志是为销售教育软件和儿童电脑游戏的公司做广告的好地方，调查结果使我们发现估计订阅者的平均年龄在 30 岁和 31 岁直接，并且 53.41% 的订阅者有孩子。由订阅者的年龄普遍偏年轻，其中有小孩的又占多数，我们可以推断这些订阅者的小孩都很小，他们对于教育软件和儿童电脑游戏会有一定需求。可以得出结论，《Young Professional Magazine》的订阅者是销售儿童教育软件和电脑游戏公司的一个很好的目标市场。

f. 从调查结果来看，我认为订阅者最会感兴趣的文章类型应该是金融类、投资类的文章，因为这些订阅者或订阅者的家庭中几乎全部都有金融投资。其次感兴趣会是育儿方面的文章，因为订阅者平均年龄很年轻，且超过半数都有小孩。再者会是关于房地产的文章，因为接近半数的订阅者在未来两年内有买房计划。

5 Question #5: Quality Associate, Inc. (Attached Data: Quality)

a. 每个测试的 p 值如下:

```
#>          s1          s2          s3          s4
#> 0.281008276 0.454650325 0.003790318 0.033893355
```

你可以使用区间估计来检验假设

```
#> [1] 11.91081 12.08919
```

```
#> $s1
#> [1] 11.95867
#>
#> $s2
#> [1] 12.02867
#>
#> $s3
#> [1] 11.889
#>
#> $s4
#> [1] 12.08133
```

不需要采取措施。

b. 四个样本的标准差如下：

```
#> $s1
#> [1] 0.220356
#>
#> $s2
#> [1] 0.220356
#>
#> $s3
#> [1] 0.2071706
#>
#> $s4
#> [1] 0.206109
```

可以合理地假设标准差是 0.21。

c. 样本均值为 12.

```
#> [1] 11.91081 12.08919
```

d. 将显著性水平由 0.01 提高至 0.05：

```
#> [1] 11.93694 12.06306
```

随着显著性水平的提高，第一类错误会增加。

6 Question #6

a. 估计 2007 年 3 月第一周出租的比例为 35%，2008 年 3 月第一周出租比例为 47%。

```
#> [1] 0.35
```

```
#> [1] 0.4666667
```

b. 比例差异的 95% 置信区间

比例差异的 95% 置信区间为：-0.2203182, -0.0130152.

c. 根据你的发现，2008 年 3 月的租赁率似乎比一年前有所上升。

7 Question #7: ir Force Training Program (data file: Training)

a.

skim_type	skim_variable	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25
numeric	Current	0	1	75.06557	3.944907	65	72
numeric	Proposed	0	1	75.42623	2.506385	69	74

由上表可知，两种方法的训练时间的平均值相差不大，均为 75 小时。但是标准差相差较大，说明通过文本进行学习的学习时间的分散程度较大。

b.

```
#>
#>  Welch Two Sample t-test
#>
#> data:  data_q7$Current and data_q7$Proposed
#> t = -0.60268, df = 101.65, p-value = 0.5481
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -1.5476613  0.8263498
#> sample estimates:
#> mean of x mean of y
#>  75.06557  75.42623
```

两种方法的学习时间，均值相差不大。在 95% 的置信区间内未发现差异，均值相同。

c.

```
#> $Current
#> [1] 3.944907
#>
#> $Proposed
#> [1] 2.506385

#> $Current
#> [1] 15.5623
#>
#> $Proposed
#> [1] 6.281967

#>
#>  F test to compare two variances
#>
#> data:  data_q7$Current and data_q7$Proposed
#> F = 2.4773, num df = 60, denom df = 60, p-value = 0.000578
#> alternative hypothesis: true ratio of variances is not equal to 1
#> 95 percent confidence interval:
#>  1.486267 4.129135
#> sample estimates:
#> ratio of variances
#>           2.477296
```

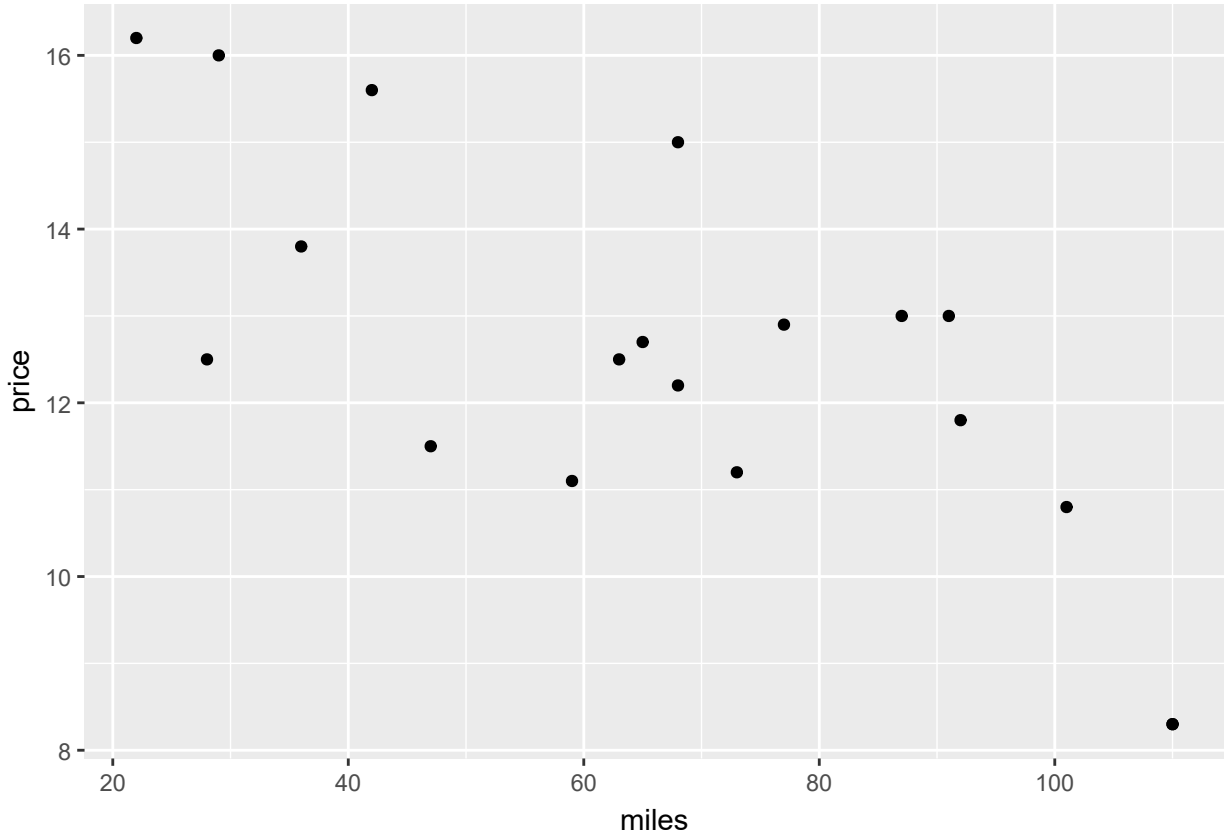
使用指导文本进行学习，标准差为 3.9，方差为 16；使用提议中的计算机辅助教学方法，标准差为 2.5，方差为 6.3. 两种标准差或方差是不同的。使用指导文本进行学习，学习时间的分散程度较大。

d. 两种方法学习完成时间的均值非常接近。但是提议的方法具有显著较低的方差。在提议的方法下，多数学生可能在更接近的时间内完成培训。方便把握整体学习进度。

e. 当前仅根据学习时间，作出结论过于草率。建议收集两种方法下学习量、学习效果和学生学习满意度的数据。或者进行考试，统计两种不同方法的考试分数。通过以上数据，综合分析两种学习方法的好坏再作结论。时间数据支持转向提议的方法。然而，提议方法的培训质量是否与当前方法相同或更好？两组都可以在培训计划结束时进行考试。对考试成绩的分析将确定这些计划在提供的学习方法上是否相似或不同。在最终决定采用提议的方法之前，应该进行这项分析。

8 Question #8

a. 散点图如下（x 轴里程数，y 轴价格）：



b. 根据图中的点分布显示，随着里程数的增加，价格呈现下降趋势。这表明里程数与价格之间存在负相关关系；数据点在低里程数（20-40 千英里）时价格较高，随着里程数增加，价格逐渐降低。在高里程数（80-100

千英里）区域，价格下降得更为明显。虽然数据点的分布并不完全沿着一条直线，但整体趋势可以近似地用一条直线来描述，这表明里程数与价格之间可能存在线性关系。在低里程数区域，有一些点的价格明显高于其他点，这可能是由于车辆的其他因素（如车况、额外配置等）导致的。

c.

```
#>
#> Call:
#> lm(formula = price ~ miles, data = data_q8)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.32408 -1.34194  0.05055  1.12898  2.52687
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 16.46976    0.94876  17.359 2.99e-12 ***
#> miles       -0.05877    0.01319  -4.455 0.000348 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.541 on 17 degrees of freedom
#> Multiple R-squared:  0.5387, Adjusted R-squared:  0.5115
#> F-statistic: 19.85 on 1 and 17 DF,  p-value: 0.0003475
```

回归方程为;

$$Price = 16.470 - 0.059 * miles$$

d. 显著关系 0.05 的显著性水平: $p - value = 0.000348 < \alpha = .05$

e. 我觉得并没有提供良好的拟合度，因为并没有考虑车况、配置等其他因素。

f. 估计回归方程的斜率是 -0.059。因此，x 值每增加一个单位，y 值就会相应减少 0.059。由于数据是以千为单位记录的，汽车里程表上每增加 1000 英里，预计价格将下降 59.0 美元。

g. 行驶了 60,000 英里的二手 2007 年款凯美瑞，根据拟合的回归方程，价格为 12942 美元，这会是一个参考价格，通过这个价格为起点，再考虑车况、配置等其他因素进一步报价。

9 Question #9

a. 流失等于 0 和流失等于 1 的情况之间的比较进行可视化探索。

```
#> Rows: 6,347
#> Columns: 13
#> $ id           <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
#> $ churn        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
#> $ happy_index  <dbl> 0, 62, 0, 231, 43, 138, 180, 116, 78, 78, 91, 40, 215, 0,~
#> $ chg_hi       <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, -37, -1, 14, 15, 0, 63,~
#> $ support      <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, ~
#> $ chg_supprt   <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
#> $ priority     <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
#> $ chg_priority <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
#> $ log_in_fre   <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7, 14, 0, 71, 0, 5, 0, 4~
#> $ chg_blog_fre <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3, 0, 9, 0, 1, 0, 0, 0, 6~
#> $ chg_vis      <dbl> 0, -16, 0, 21996, 9, -33, 907, 38, 0, 30, 0, 15, 8658, 0,~
#> $ y_age        <dbl> 72, 72, 60, 68, 62, 63, 62, 51, 61, 61, 58, 61, 62, 62, 6~
#> $ chg_interval <dbl> 33, 33, 33, 2, 33, 2, 2, 8, 9, 16, 2, 33, 2, 33, 2, 33, 3~
```

churn	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg_blog_fr
0	88.60591	5.530213	0.7242696	-0.0092961	0.8295759	0.0326818	16.13894	0.171148
1	63.27245	-3.736842	0.3715170	0.0371517	0.4995577	-0.0166962	8.06192	-0.102167

在所有 11 个指标中，流失与未流失的客户之间存在差异。未流失的客户的 “当月客户幸福指数” 较高，为 89，而流失的客户的 “当月客户幸福指数” 则较低，为 63。未流失的客户在 “客户幸福指数相比上月变化” 上为 5.5，而流失的客户为-3.7，这可能意味着流失的客户经历了负面的变化。在 “当月客户支持” 上，未流失的客户为 0.72，而流失的客户为 0.37，表明流失的客户可能获得的支持较少。“登录频率” 也显示出差异，未流失的客户登录频率为 16.1，而流失的客户为 8.1。

b. 使用 t.test 来检查这些差异是否具有统计学意义

variable	estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	co
chg_blog_fre	0.2733159	0.1711487	-0.1021672	2.5315145	0.0115761	695.9510	0.0613390	0.4
chg_hi	9.2670546	5.5302125	-3.7368421	5.7835224	0.0000000	365.7132	6.1161371	12.4
chg_interval	-4.9746139	3.5114542	8.4860681	-4.0971030	0.0000522	346.0344	-7.3627124	-2.5
chg_priority	0.0493780	0.0326818	-0.0166962	0.6411575	0.5218233	364.4864	-0.1020692	0.2
chg_supprt	-0.0464479	-0.0092961	0.0371517	-0.6319825	0.5277532	406.9016	-0.1909261	0.0
chg_vis	202.3773636	106.6095618	-95.7678019	1.9136102	0.0563070	448.0016	-5.4637294	410.2
happy_index	25.3334639	88.6059097	63.2724458	7.6242176	0.0000000	369.3571	18.7995591	31.8
log_in_fre	8.0770247	16.1389442	8.0619195	3.5708588	0.0004037	362.6743	3.6288837	12.5
priority	0.3300182	0.8295759	0.4995577	5.1427709	0.0000004	373.1266	0.2038355	0.4

support	0.3527526	0.7242696	0.3715170	5.5098545	0.0000001	419.2152	0.2269082	0.4
y_age	-1.5342161	18.8187251	20.3529412	-2.9811315	0.0030568	379.8984	-2.5461200	-0.5

根据表格，我们可以得出以下结论：除了 “客户支持相比上月的变化” 和 “服务优先级相比上月的变化”，其他所有指标的差异都是显著的。这意味着在这些指标上，流失客户与未流失客户之间存在统计学上的显著差异。对于 “客户支持相比上月的变化” 和 “服务优先级相比上月的变化”，我们没有足够的证据表明流失与未流失客户之间存在显著差异。这可能意味着这些指标的变化对于客户是否流失的影响不大，或者这种影响在统计上不够显著。

c. 以流失为因变量，以当月客户幸福指数、客户幸福指数相比上月变化、当月客户支持、当月服务优先级、当月登录次数、博客数相比上月的变化、访问次数相比上月的增加、客户使用期限、访问间隔变化为自变量建立回归方程

```
#>
#> Call:
#> glm(formula = churn ~ happy_index + chg_hi + support + priority +
#>     log_in_fre + chg_blog_fre + chg_vis + y_age + chg_interval,
#>     family = binomial(link = "logit"), data = we_data)
#>
#> Coefficients:
#>                Estimate Std. Error z value Pr(>|z|)
#> (Intercept)  -2.874e+00  1.215e-01 -23.661  < 2e-16 ***
#> happy_index  -5.225e-03  1.161e-03  -4.500 6.78e-06 ***
#> chg_hi       -9.501e-03  2.424e-03  -3.920 8.87e-05 ***
#> support      -3.522e-02  7.438e-02  -0.474  0.63581
#> priority     -3.727e-02  7.514e-02  -0.496  0.61985
#> log_in_fre    9.104e-04  1.952e-03   0.466  0.64098
#> chg_blog_fre -2.357e-05  2.080e-02  -0.001  0.99910
#> chg_vis      -1.170e-04  4.069e-05  -2.877  0.00401 **
#> y_age         1.418e-02  5.260e-03   2.696  0.00701 **
#> chg_interval  1.700e-02  4.277e-03   3.975 7.03e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 2553.1  on 6346  degrees of freedom
#> Residual deviance: 2445.9  on 6337  degrees of freedom
```

```
#> AIC: 2465.9
#>
#> Number of Fisher Scoring iterations: 6

#>  happy_index       chg_hi      support     priority   log_in_fre chg_blog_fre
#>     1.513596     1.240227     2.166698     2.128518     1.293839     1.068660
#>      chg_vis        y_age chg_interval
#>     1.034792     1.247978     1.197948
```

d. 根据以上回归方程，流失可能性最大的前 100 名用户 ID 列表如下：

id	churn	happy_index	chg_hi	support	chg_supprt	priority	chg_priority	log_in_fre	chg_blog_fr
2287	0	227	7	5	5	2.8	2.8	11	-
109	0	0	-125	0	0	0.0	0.0	-8	
1971	0	0	-113	0	0	0.0	0.0	-23	
2025	0	18	-15	0	0	0.0	0.0	0	
1	0	0	0	0	0	0.0	0.0	0	
929	0	123	35	0	0	0.0	0.0	7	
2076	0	29	-69	0	0	0.0	0.0	0	
76	0	1	-70	0	0	0.0	0.0	-7	-
14	0	0	0	0	0	0.0	0.0	0	
18	0	0	0	0	0	0.0	0.0	0	
3	0	0	0	0	0	0.0	0.0	0	
2244	0	16	-38	0	0	0.0	0.0	0	
21	0	0	0	0	0	0.0	0.0	0	
1287	0	24	-72	0	-1	0.0	-3.0	-6	-
1929	0	7	-40	0	0	0.0	0.0	0	
1459	0	0	-22	0	0	0.0	0.0	0	
51	0	1	0	0	0	0.0	0.0	0	
128	0	31	-26	0	0	0.0	0.0	0	
183	0	0	-17	0	0	0.0	0.0	-1	
59	0	0	0	0	0	0.0	0.0	0	
55	0	3	0	0	0	0.0	0.0	0	
121	0	0	0	0	0	0.0	0.0	0	
2240	0	0	-15	0	0	0.0	0.0	0	
1520	0	0	-67	0	0	0.0	0.0	-4	
2599	0	7	-30	0	0	0.0	0.0	0	

1236	0	0	-35	0	0	0.0	0.0	0
137	0	0	0	0	0	0.0	0.0	0
1862	0	0	-27	0	0	0.0	0.0	0
2080	0	4	-25	0	0	0.0	0.0	0
1143	0	0	-17	0	0	0.0	0.0	0
154	0	0	0	0	0	0.0	0.0	0
1286	0	20	-77	0	0	0.0	0.0	-2
2546	0	8	-86	0	-1	0.0	-3.0	-9
146	0	0	-2	0	0	0.0	0.0	0
119	0	0	0	0	0	0.0	0.0	0
171	0	0	0	0	0	0.0	0.0	0
190	0	0	0	0	0	0.0	0.0	0
42	0	25	0	0	0	0.0	0.0	0
5	0	43	-1	0	0	0.0	0.0	0
2	0	62	4	0	0	0.0	0.0	0
123	0	10	-25	0	-1	0.0	-2.0	-4
101	0	0	0	0	0	0.0	0.0	0
1616	0	2	-42	0	0	0.0	0.0	-1
95	0	6	6	0	0	0.0	0.0	10
2680	0	2	-72	0	0	0.0	0.0	-10
2838	0	0	-28	0	0	0.0	0.0	-1
61	0	7	7	0	0	0.0	0.0	4
2289	0	30	-10	0	0	0.0	0.0	0
1438	0	0	0	0	0	0.0	0.0	0
1392	0	0	0	0	0	0.0	0.0	0
2481	0	21	-94	0	-1	0.0	-3.0	-9
2924	0	0	-19	0	0	0.0	0.0	0
106	0	0	0	0	0	0.0	0.0	0
3671	0	0	-40	0	0	0.0	0.0	0
203	0	15	-4	0	0	0.0	0.0	0
1393	0	3	3	0	0	0.0	0.0	0
69	0	10	10	0	0	0.0	0.0	10
1574	0	3	-87	0	0	0.0	0.0	-10
1204	0	25	-40	0	0	0.0	0.0	0
68	0	0	0	1	1	2.0	2.0	0
2255	0	0	0	0	0	0.0	0.0	0

1395	0	0	0	0	0	0.0	0.0	0
1478	0	0	0	0	0	0.0	0.0	0
2235	0	0	0	0	0	0.0	0.0	0
89	0	4	4	1	1	2.0	2.0	10
798	0	0	-30	0	0	0.0	0.0	0
1141	0	3	-73	0	-1	0.0	-3.0	-14
2739	0	31	-25	0	0	0.0	0.0	0
62	0	14	14	0	0	0.0	0.0	4
4245	0	20	-72	0	0	0.0	0.0	-2
1151	0	3	-26	0	0	0.0	0.0	-1
2830	0	0	-14	0	-1	0.0	-3.0	0
1693	0	0	-23	0	0	0.0	0.0	0
3042	0	22	-25	0	0	0.0	0.0	0
12	0	40	14	0	0	0.0	0.0	0
142	0	15	-27	0	0	0.0	0.0	-2
1908	0	0	0	0	0	0.0	0.0	0
10	0	78	-37	0	0	0.0	0.0	-7
868	0	43	-59	0	-1	0.0	-3.0	-1
2286	0	10	0	0	0	0.0	0.0	0
3076	0	5	-15	0	0	0.0	0.0	0
57	0	0	-21	0	0	0.0	0.0	-8
2242	0	61	-71	0	0	0.0	0.0	-12
1951	0	0	0	0	0	0.0	0.0	0
3124	0	19	-71	0	0	0.0	0.0	-4
1019	0	33	-77	0	-1	0.0	-3.0	3
1110	0	8	-3	0	0	0.0	0.0	0
2062	0	0	0	0	0	0.0	0.0	0
2903	0	72	-93	0	-2	0.0	-3.0	-16
2913	0	58	-43	0	0	0.0	0.0	-1
2047	0	0	0	0	0	0.0	0.0	0
104	0	0	0	0	0	0.0	0.0	0
1953	0	0	0	0	0	0.0	0.0	0
2656	0	22	-42	0	0	0.0	0.0	-3
1155	0	41	-33	0	0	0.0	0.0	0
2744	0	0	0	0	0	0.0	0.0	0
1446	0	32	-1	0	0	0.0	0.0	0

2306	0	20	0	0	0	0.0	0.0	0
163	0	36	-4	0	0	0.0	0.0	-1
240	0	0	-22	0	0	0.0	0.0	0