Dain Hall

true

Determine notebook defaults:

Load packages:

Read in the data:

```r
# rm(list=ls())
load("/Users/dain/Programs/R_Projects/MKTG_482_HW3/bbb.Rdata")
```

Assignment answers

Part 1 - Logistic Regression

Question 1

Estimate a logistic regression model using "buyer" as the dependent variable and the following as predictor variables: * gender * last * total * child * youth * cook * do_it * reference * art * geog

```r
lrm <- glm(buyer ~ gender + last + total + child + youth + cook + do_it + reference + art + geog, famil
summary(lrm)
```

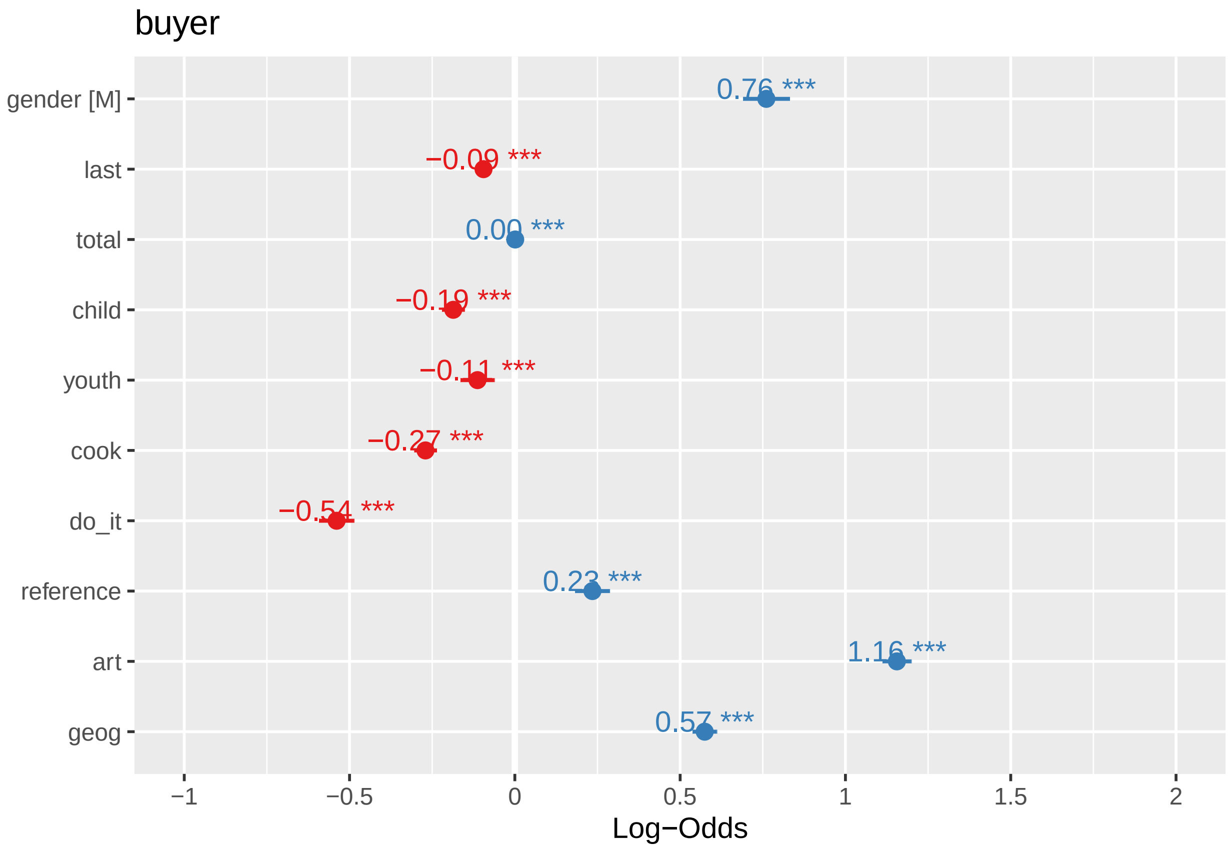
```

Call:
glm(formula = buyer ~ gender + last + total + child + youth +
    cook + do_it + reference + art + geog, family = binomial(logit),
    data = bbb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4031  -0.4129  -0.2807  -0.1839   3.2650

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3608301  0.0492961 -47.891  < 2e-16 ***
genderM      0.7607204  0.0357608  21.272  < 2e-16 ***
last        -0.0947124  0.0027924 -33.918  < 2e-16 ***
total        0.0011160  0.0001982   5.630 1.80e-08 ***
child       -0.1862162  0.0172824 -10.775  < 2e-16 ***
youth       -0.1129745  0.0261087  -4.327 1.51e-05 ***
cook        -0.2703210  0.0171283 -15.782  < 2e-16 ***
do_it       -0.5391648  0.0269657 -19.994  < 2e-16 ***
reference    0.2346876  0.0265583   8.837  < 2e-16 ***
art          1.1555840  0.0221439  52.185  < 2e-16 ***
geog         0.5742763  0.0186311  30.824  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30355  on 49999  degrees of freedom
Residual deviance: 24122  on 49989  degrees of freedom
AIC: 24144

Number of Fisher Scoring iterations: 6
```

Question 2

Use "plot_model(. . . , show.values = TRUE, transform = NULL)" to display the coefficients and confidence intervals. Which variables are statistically significant and which ones seem to be economically 'important'?

```
plot_model(lrm, show.values = TRUE, transform = NULL)
```




```
"
All variables appear to be statistically significant according to their P-values.
However, the variables with the greatest absolute values of their intercepts include art, gender, geog,
These variables, therefore, should have the greatest economic 'importance' as predictors of buyership.
"
```

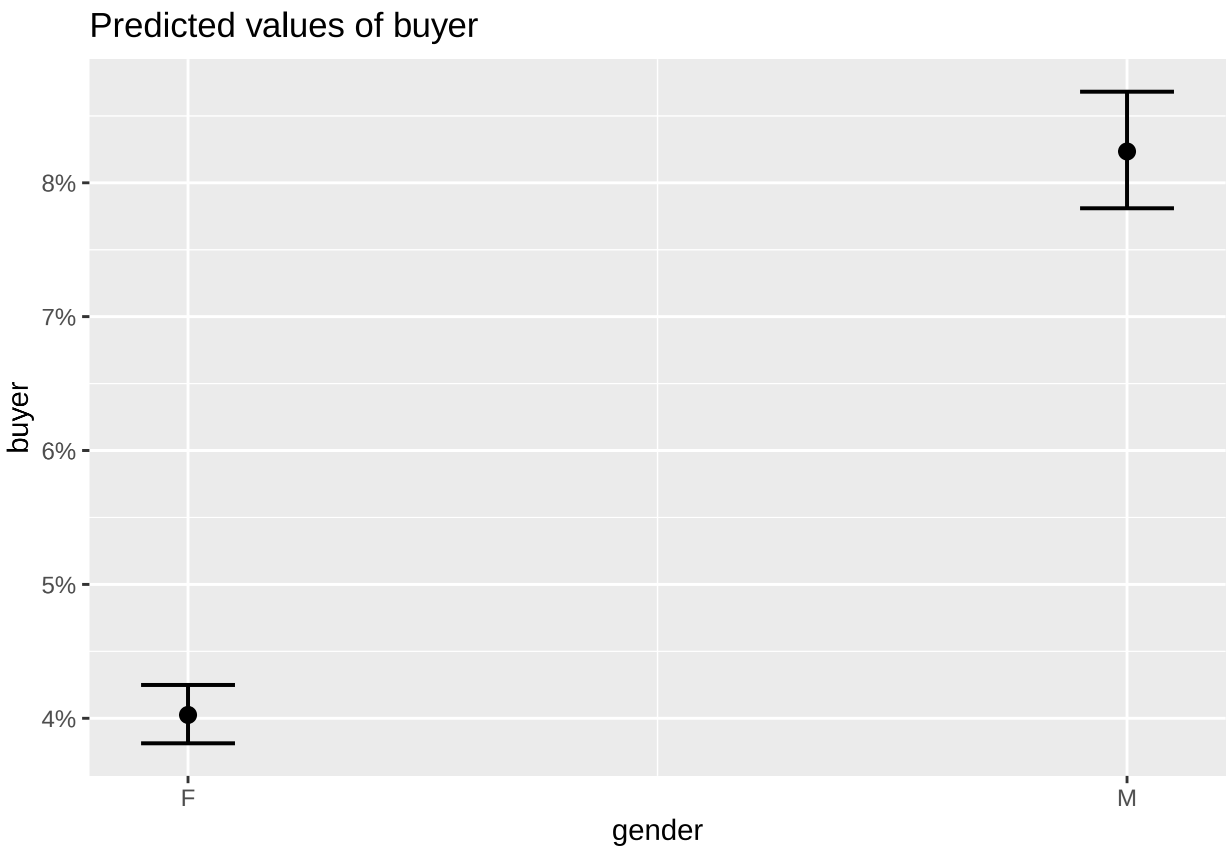
```
[1] "\nAll variables appear to be statistically significant according to their P-values.\nHowever, the
```

Question 3

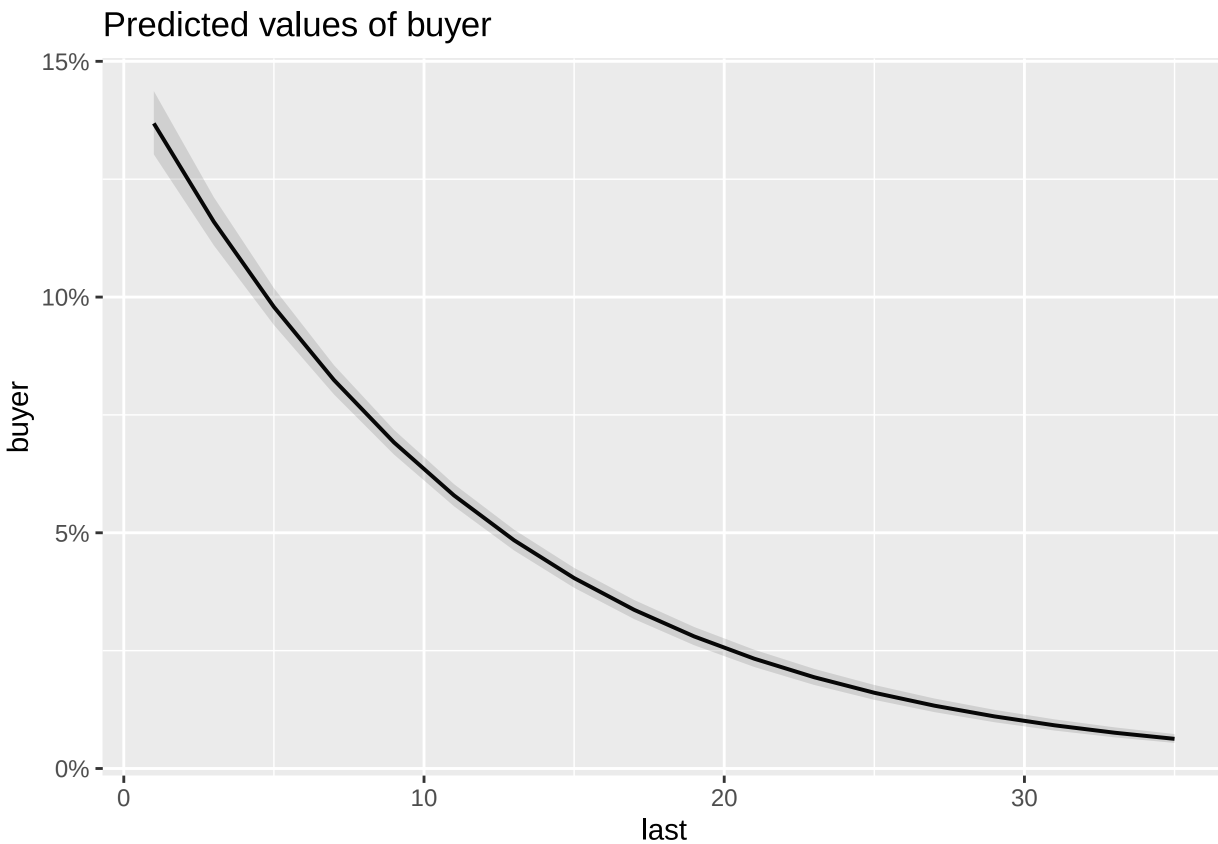
Use the "plot_model(. . . , type="eff")" command to plot marginal effects. For which variables does your assessment of the importance of a variable change and why?

```r
plot_model(lrm, show.values = TRUE, transform = NULL, type = "eff")
```

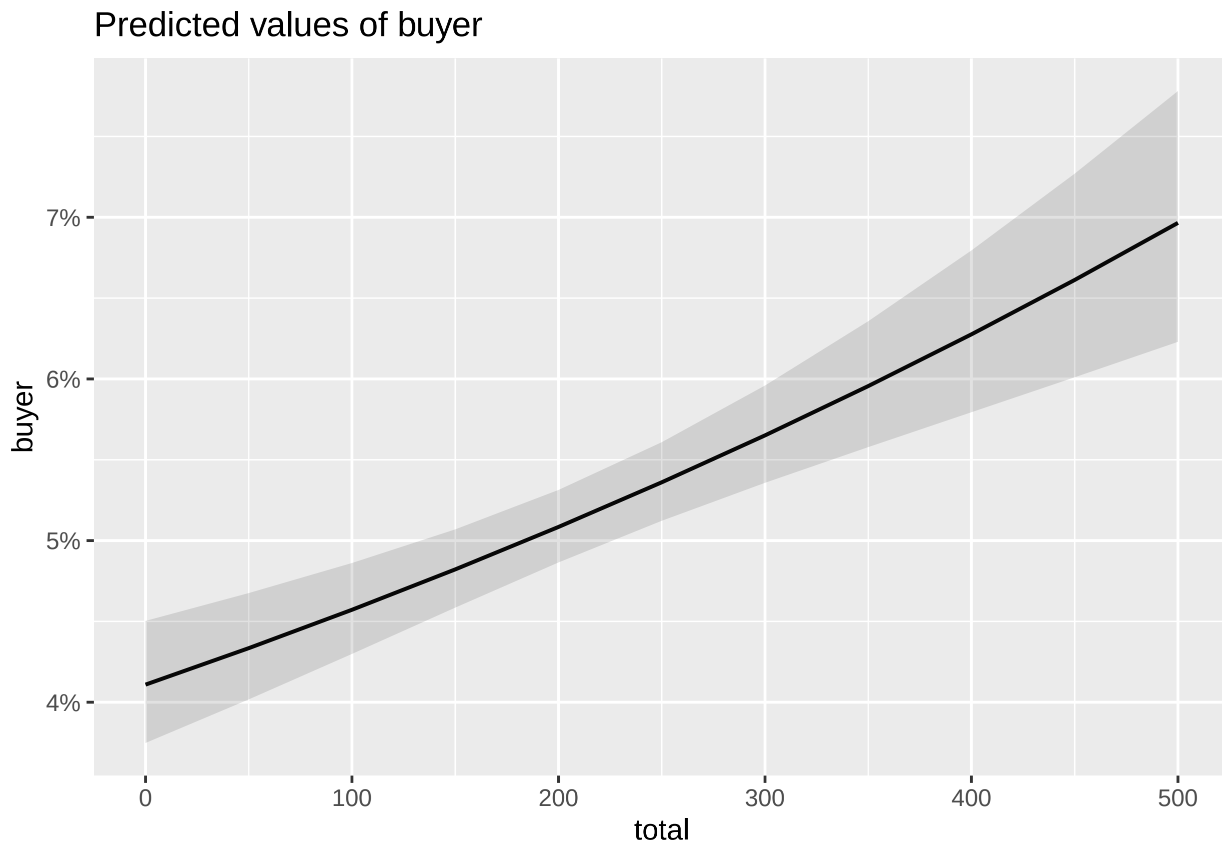
```
$gender
```



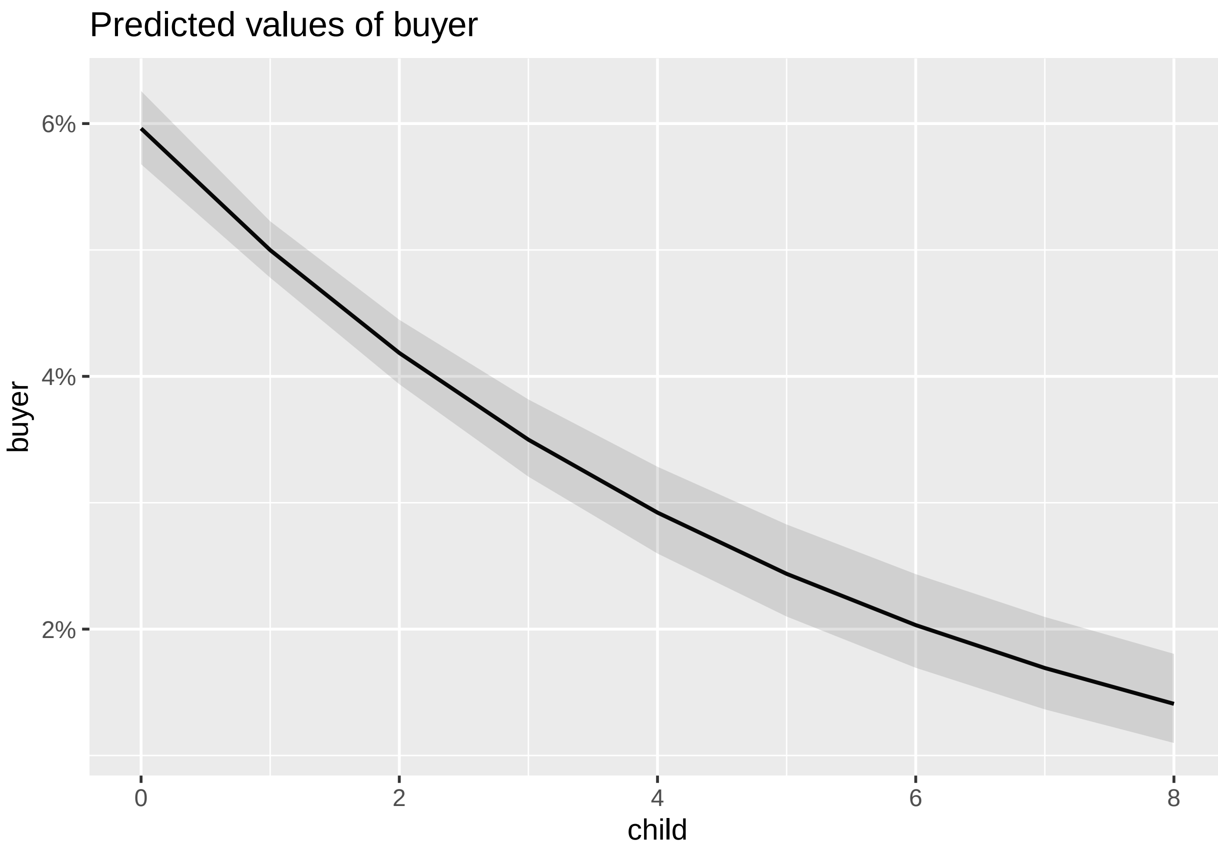

```
$last
```



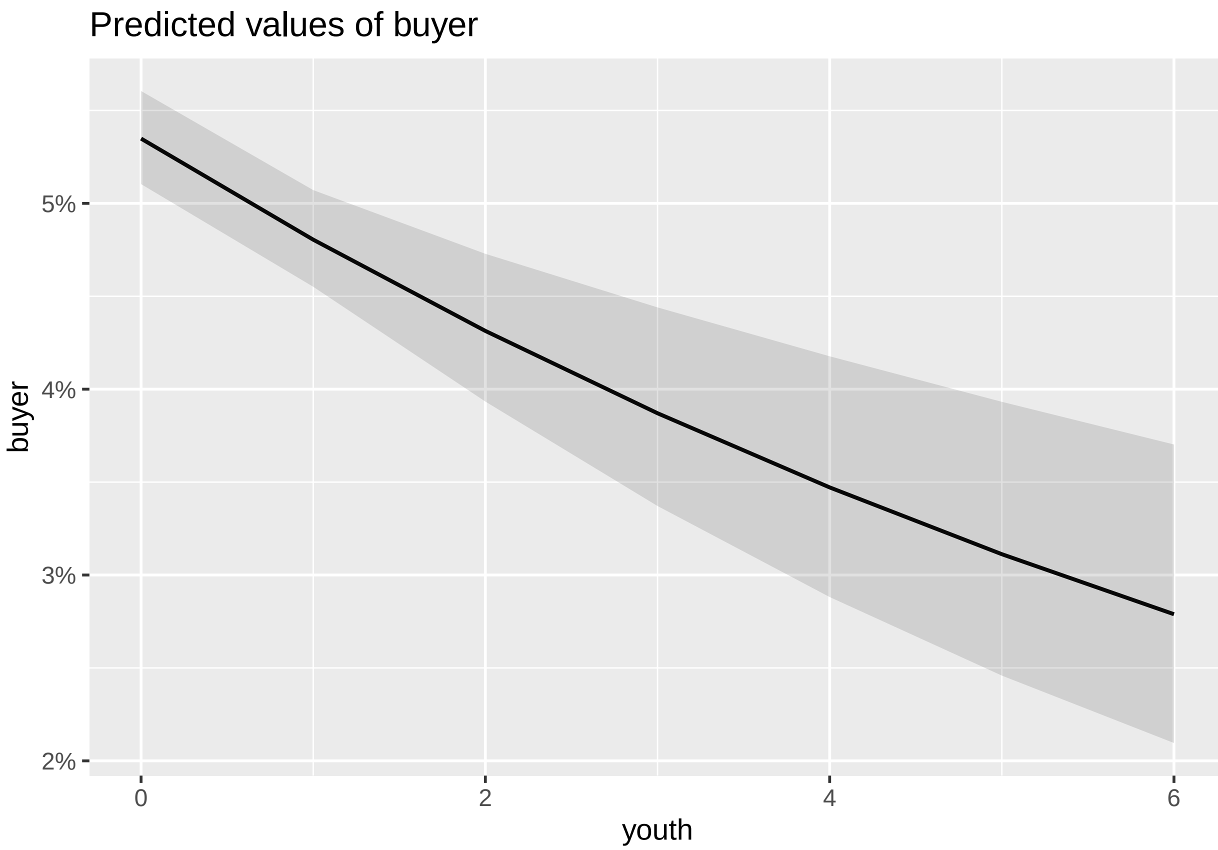
$total


Predicted values of buyer
buyer
7%
6%
5%
4%
0
100
200
300
400
500
total

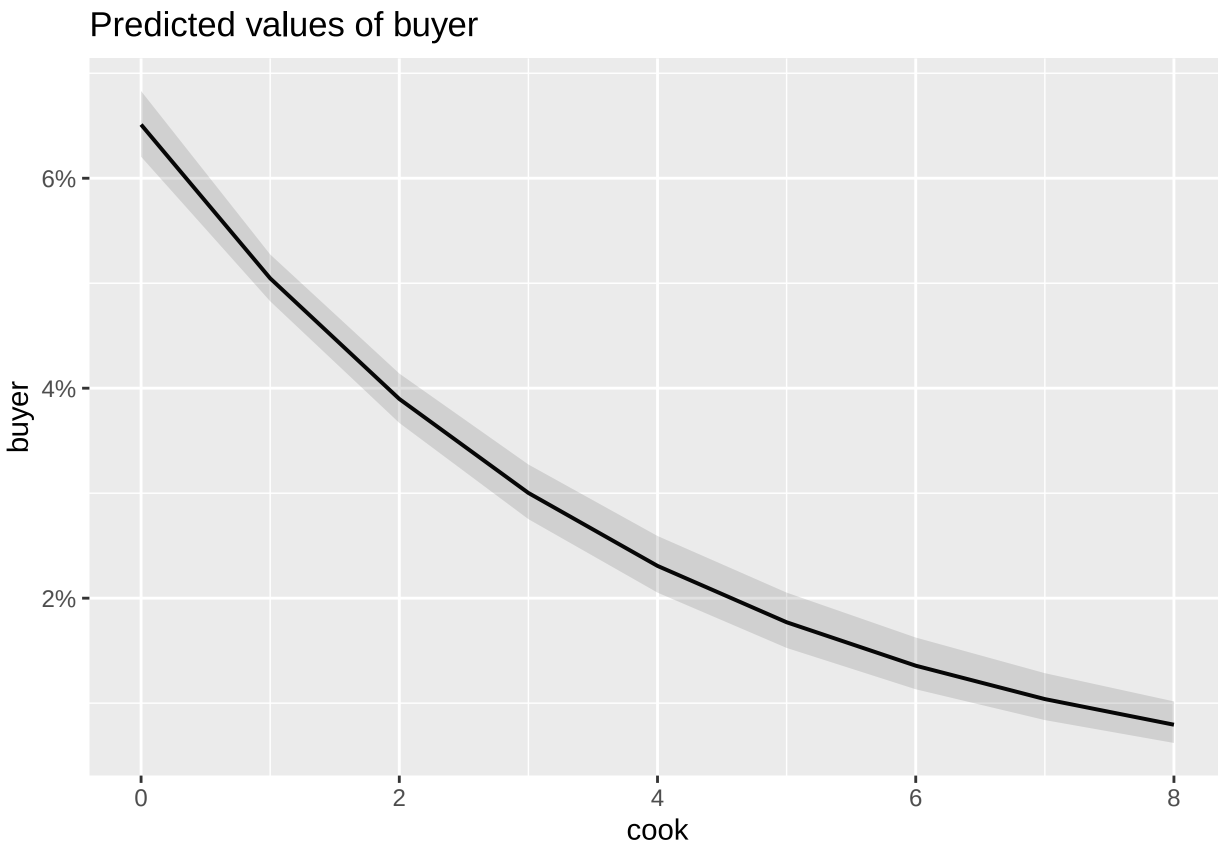
$child



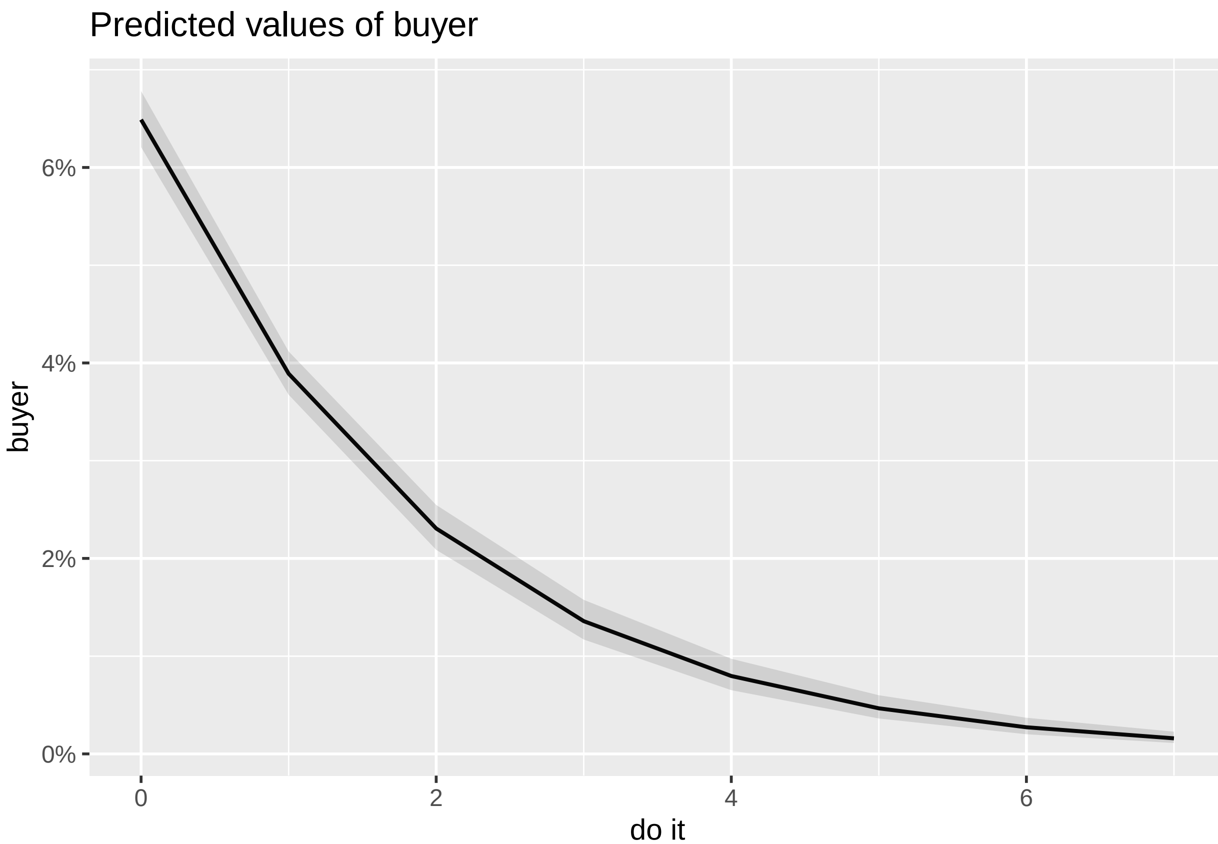
$youth



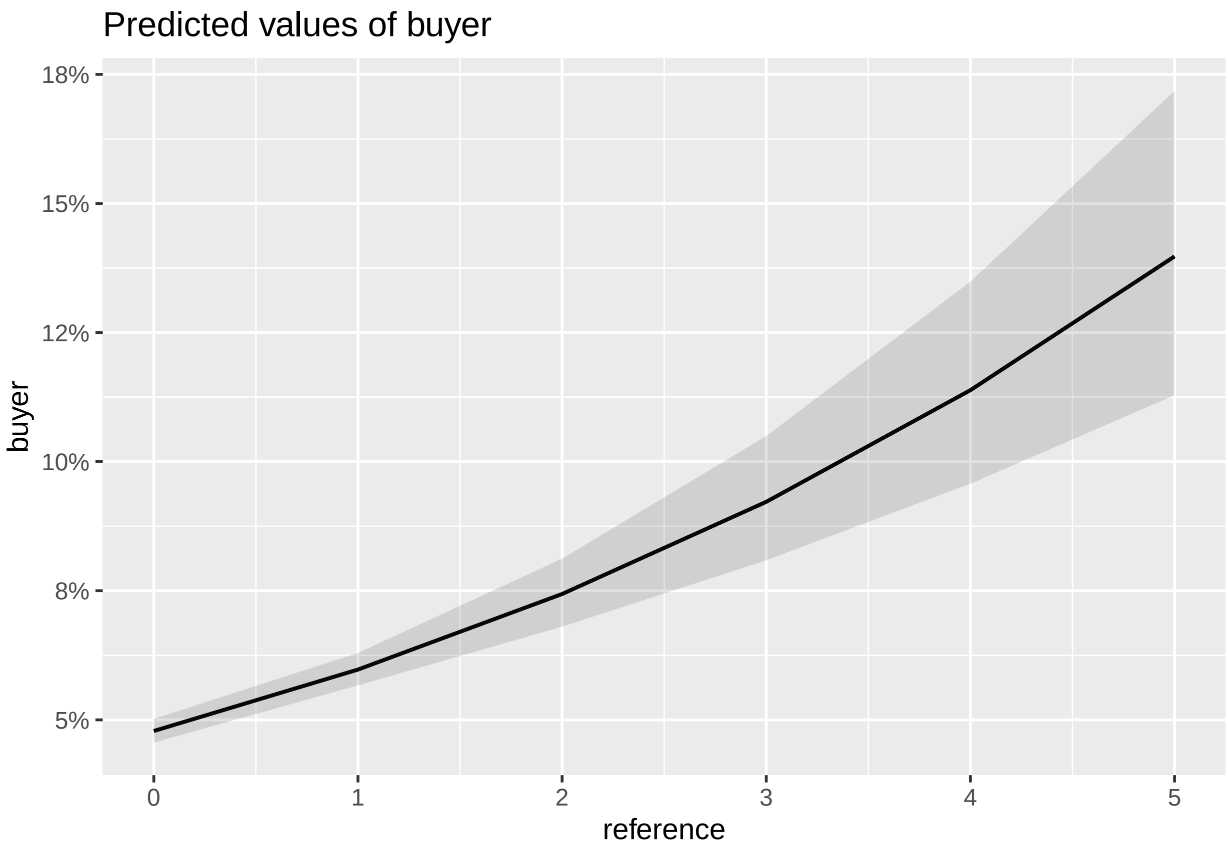
$cook




```
$do_it
```

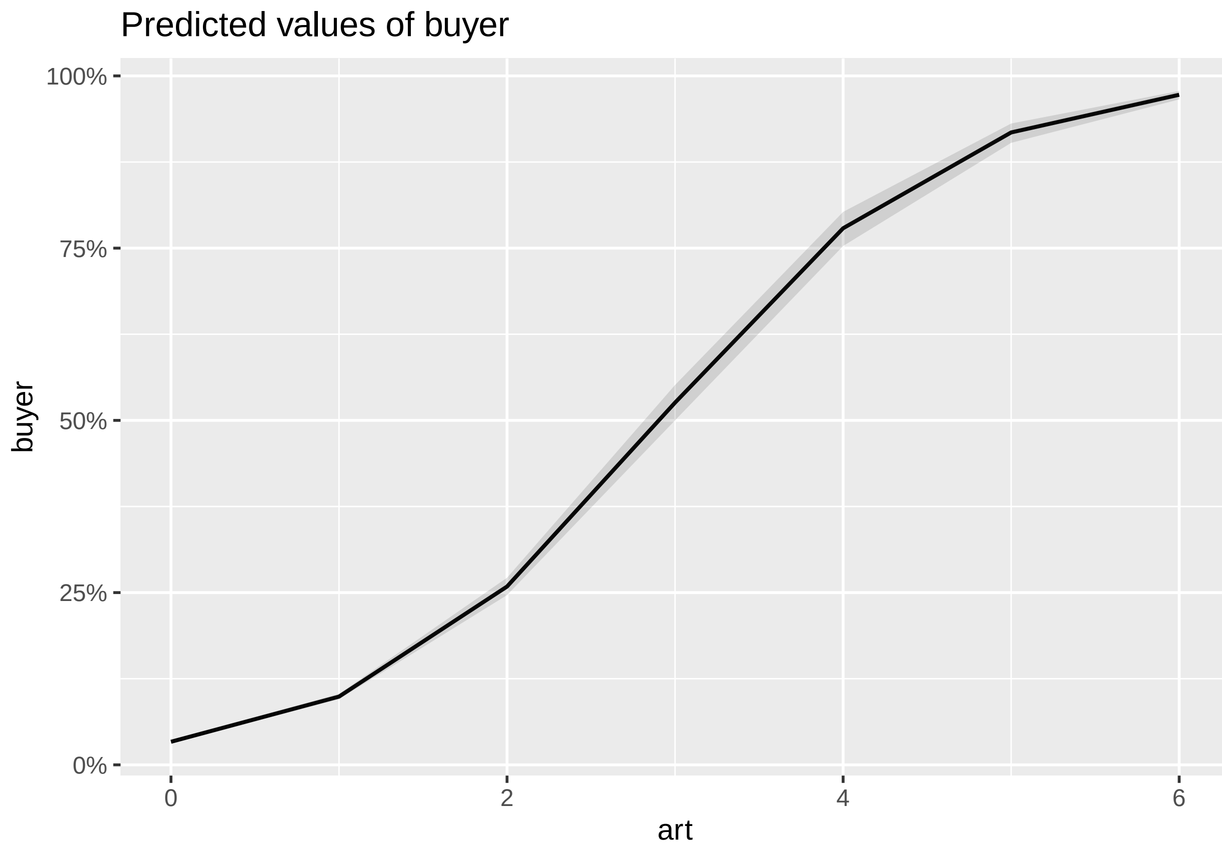


$reference

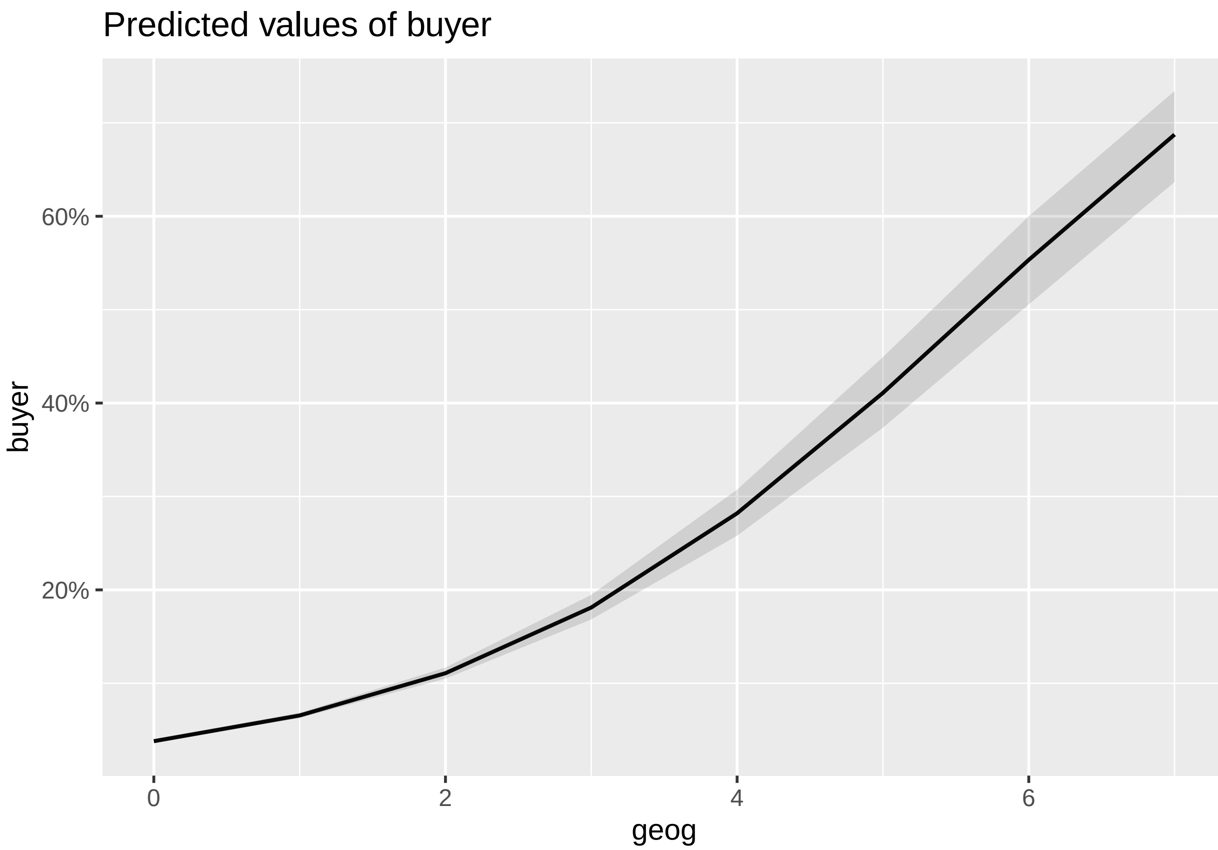


Predicted values of buyer
18%
15%
12%
10%
8%
5%
buyer
0
1
2
3
4
5
reference

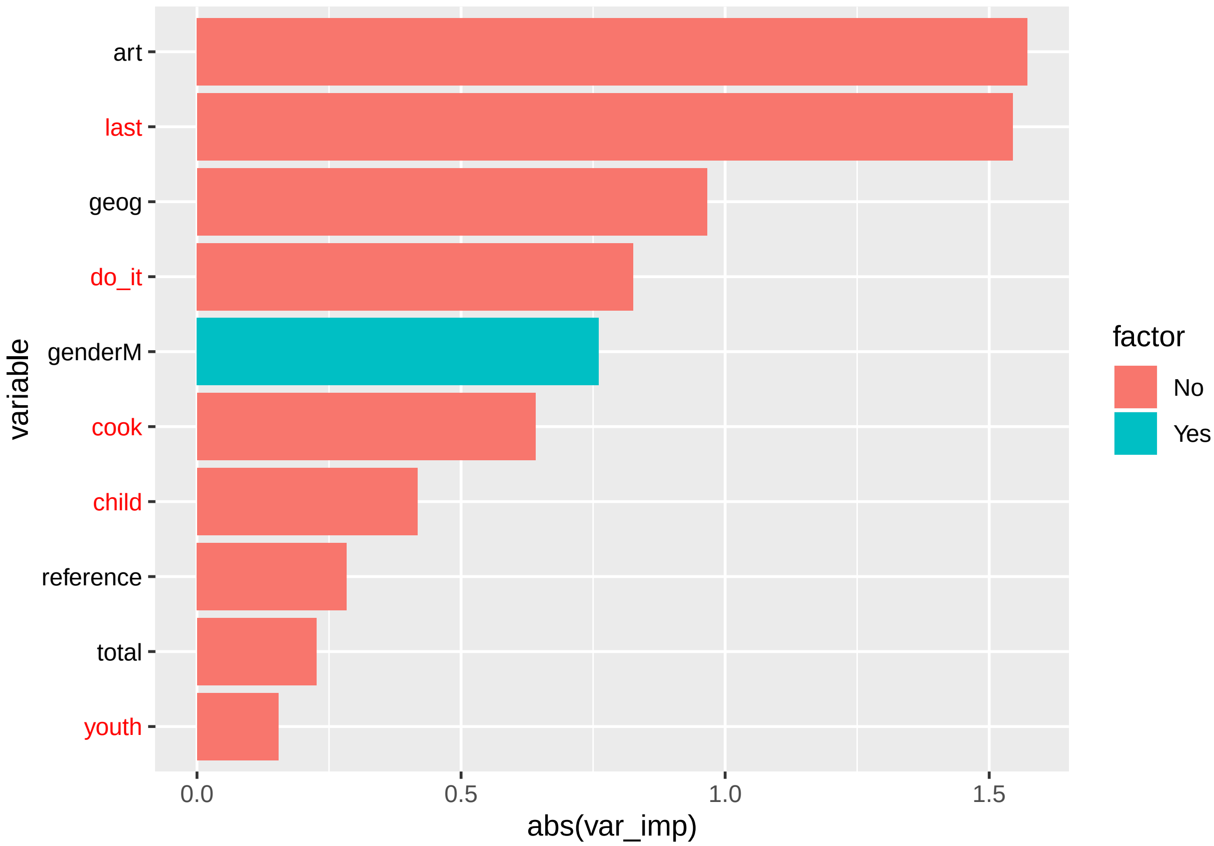

```
$art
```



$geog




```r
varimp.logistic(lrm) %>% plotimp.logistic()
```




```
# A tibble: 10 x 6
   variable  factor var_imp var_imp_lower var_imp_upper p_value
   <chr>     <chr>    <dbl>         <dbl>         <dbl>   <dbl>
 1 art       No       1.57          1.51          1.63        0
 2 last      No      -1.54         -1.63         -1.46        0
 3 geog      No       0.966         0.905         1.03        0
 4 do_it     No      -0.826        -0.907        -0.745       0
 5 genderM   Yes      0.761         0.691         0.831       0
 6 cook      No      -0.641        -0.720        -0.561       0
 7 child     No      -0.417        -0.493        -0.341       0
 8 reference No       0.283         0.221         0.346       0
 9 total     No       0.226         0.147         0.305       0
10 youth     No      -0.154        -0.224        -0.0844      0
```

```
"
↑ art, last, geog, gender - these variables have greater impact than originally expected
↓ do_it, total            - these variables have less impact than originally expected
"
```

```
[1] "\n↑ art, last, geog, gender - these variables have greater impact than originally expected\n↓ do_i
```

Question 4

Add the predicted values of the logistic regression model to the "bbb" data frame. For the first few observations in the data, visually compare the "buyer" variable to the predicted values. Next, for the full dataset, compare the average of the predicted values with the average of the "buyer" variable. What do you notice? Why is that?

```r
bbbPred <- bbb %>%
   mutate(pred_buyer=predict(lrm, type = "response"))

bbbPred %>% select(buyer, pred_buyer) %>% arrange(desc(pred_buyer))
```

```
# A tibble: 50,000 x 2
   buyer pred_buyer
   <int>      <dbl>
 1     1      0.984
 2     1      0.976
 3     1      0.973
 4     1      0.972
 5     1      0.971
 6     1      0.957
 7     1      0.954
 8     1      0.954
 9     1      0.951
10     1      0.948
# ... with 49,990 more rows
```

```r
"
For the entire dataset, the averages of buyer and predicted buyer using the lrm are the same.
It makes sense that this is the case because we used the dataset to generate the lrm model.
"
```

```
[1] "\nFor the entire dataset, the averages of buyer and predicted buyer using the lrm are the same.\nI
```

```r
bbbPred %>% summarise(avg_buyer=mean(buyer), avg_pred_buyer=mean(pred_buyer))
```

```
# A tibble: 1 x 2
  avg_buyer avg_pred_buyer
      <dbl>          <dbl>
1    0.0904         0.0904
```

Part 2 - Decile Analysis of Logistic Regression Results

Question 1

Assign each customer to a decile based on his or her predicted probability of purchase. Assign those with the highest predicted probability of purchase to decile 1 and those with the lowest predicted probability of purchase to decile 10.

```r
decileBBB <- bbbPred %>%
   mutate(pred_buyer_decile = ntile(-pred_buyer, 10)) %>%
   group_by(pred_buyer_decile) %>%
   summarise(num_cust=n(), num_buyers=sum(buyer), resp_rate=sum(buyer)/n(), pred_resp_rate=mean(pred_bu

decileBBB
```

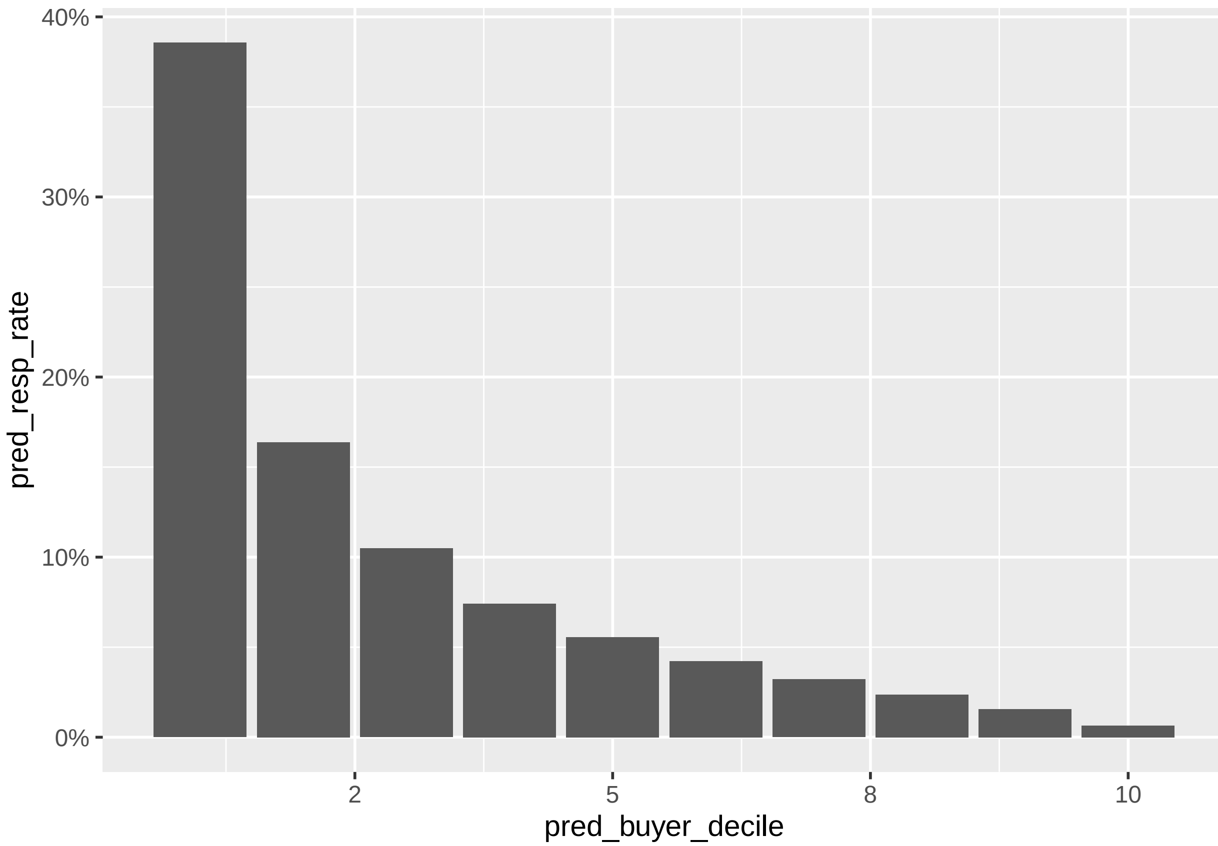
```
# A tibble: 10 x 5
   pred_buyer_decile num_cust num_buyers resp_rate pred_resp_rate
               <int>    <int>      <int>     <dbl>          <dbl>
 1                 1     5000       1935    0.387          0.386
 2                 2     5000        836    0.167          0.164
 3                 3     5000        511    0.102          0.105
 4                 4     5000        368    0.0736         0.0741
 5                 5     5000        284    0.0568         0.0556
```

6	6	5000	196	0.0392	0.0423
7	7	5000	139	0.0278	0.0321
8	8	5000	121	0.0242	0.0237
9	9	5000	90	0.018	0.0157
10	10	5000	42	0.0084	0.00651

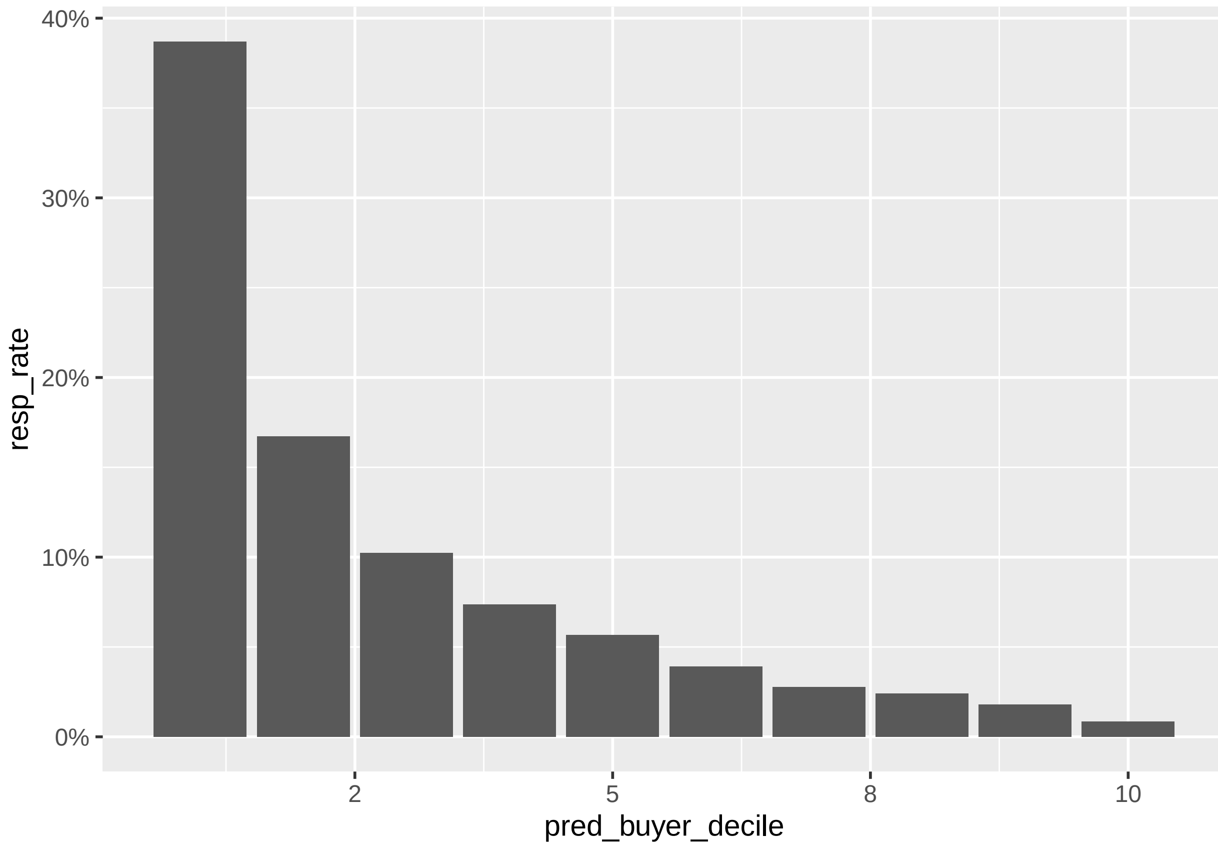
Question 2

Create a bar chart plotting response rate by decile (as just defined above). Hint: The "response rate" is not the same as the "predicted probability of purchase" that the model generated. Instead, it is the actual percentage of customers in a given group (for example a decile) that have bought "The Art History of Florence."

```
ggplot(deciletBBB, aes(x=pred_buyer_decile, y=pred_resp_rate)) +
  geom_col() +
  scale_x_continuous(labels=scales::number_format(accuracy = 1)) +
  scale_y_continuous(labels=scales::percent_format(accuracy = 1))
```




```
ggplot(decileBBB, aes(x=pred_buyer_decile, y=resp_rate)) +
  geom_col() +
  scale_x_continuous(labels=scales::number_format(accuracy = 1)) +
  scale_y_continuous(labels=scales::percent_format(accuracy = 1))
```



Question 3

Generate a report showing number of customers, the number of buyers of "The Art History of Florence' and the response rate to the offer by decile for the random sample (i.e. the 50,000) customers in the dataset.

```
decileBBB %>% select(decile=pred_buyer_decile, num_cust, num_buyers, resp_rate)
```

```
# A tibble: 10 x 4
   decile num_cust num_buyers resp_rate
    <int>    <int>      <int>     <dbl>
 1      1     5000       1935    0.387
 2      2     5000        836    0.167
 3      3     5000        511    0.102
 4      4     5000        368    0.0736
 5      5     5000        284    0.0568
 6      6     5000        196    0.0392
 7      7     5000        139    0.0278
 8      8     5000        121    0.0242
 9      9     5000         90    0.018
10     10     5000         42    0.0084
```

Part 3 - Lifts & Gains

Question 1

Use the information from the report in II.3 above to create a table showing the lift and cumulative lift for each decile. You may want to use Excel for these calculations.

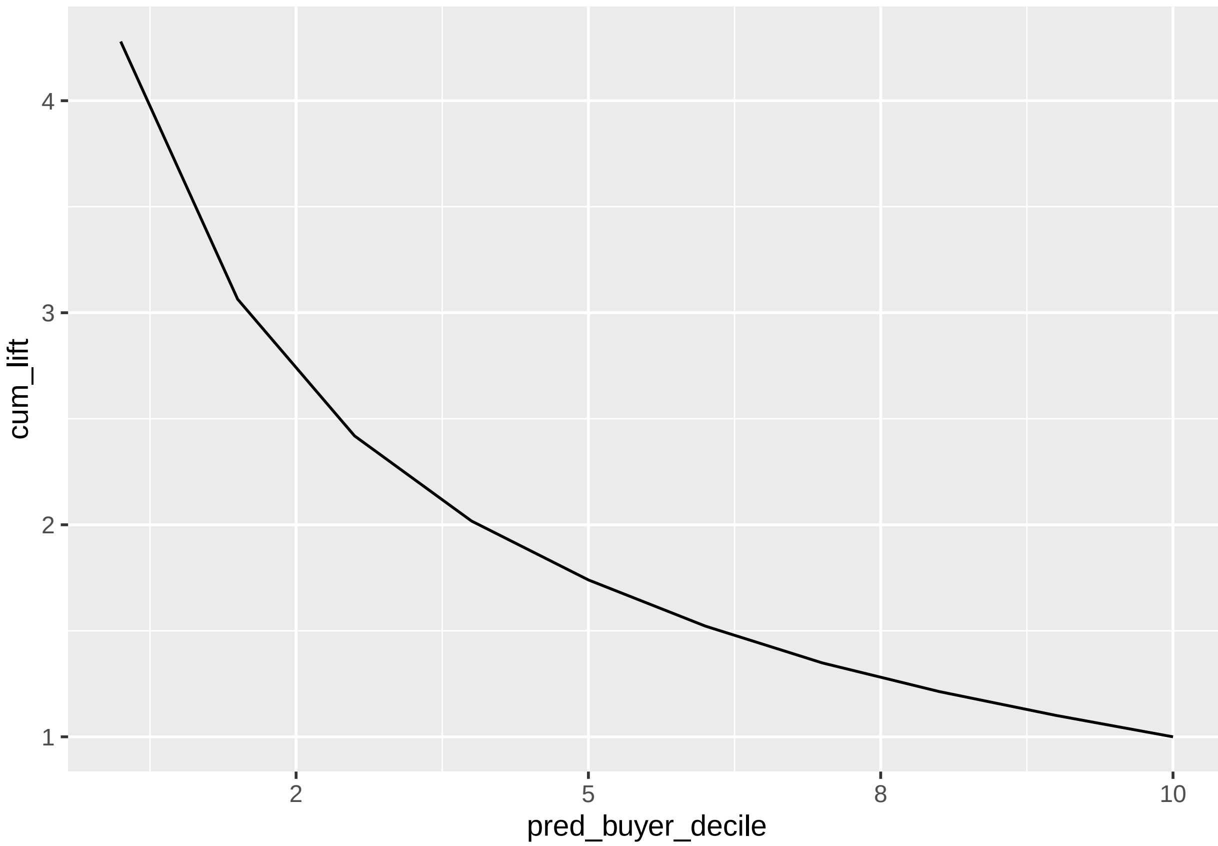
```r
# clipr::write_clip(decileBBB)
total_customers <- sum(decileBBB$num_cust)
total_buyers <- sum(decileBBB$num_buyers)
blended_resp_rate <- total_buyers / total_customers
lift_and_gains <- decileBBB %>%
  mutate(cum_cust=cumsum(num_cust), cum_buyers=cumsum(num_buyers), cum_resp_rate=cumsum(resp_rate), li
  select(pred_buyer_decile, num_cust, cum_cust, num_buyers, cum_buyers, gains, cum_gains, lift, cum_li

lift_and_gains %>% select(everything(), -perc_cum_cust, -gains, -cum_gains)
```

```
# A tibble: 10 x 7
   pred_buyer_decile num_cust cum_cust num_buyers cum_buyers   lift cum_lift
               <int>    <int>    <int>      <int>      <int>  <dbl>    <dbl>
 1                 1     5000     5000       1935       1935 4.28       4.28
 2                 2     5000    10000        836       2771 1.85       3.06
 3                 3     5000    15000        511       3282 1.13       2.42
 4                 4     5000    20000        368       3650 0.814      2.02
 5                 5     5000    25000        284       3934 0.628      1.74
 6                 6     5000    30000        196       4130 0.433      1.52
 7                 7     5000    35000        139       4269 0.307      1.35
 8                 8     5000    40000        121       4390 0.268      1.21
 9                 9     5000    45000         90       4480 0.199      1.10
10                10     5000    50000         42       4522 0.0929     1
```

Question 2

In Excel, create a chart showing the cumulative lift by decile.

```r
ggplot(lift_and_gains, aes(x=pred_buyer_decile, y=cum_lift)) + geom_line() +
  scale_x_continuous(labels=scales::number_format(accuracy = 1))
```



Question 3

Use the information from the report in II.3 above to create a table showing the gains and cumulative gains for each decile. You may want to use Excel for these calculations.

```
lift_and_gains %>% select(everything(), -lift, -cum_lift, -perc_cum_cust)
```

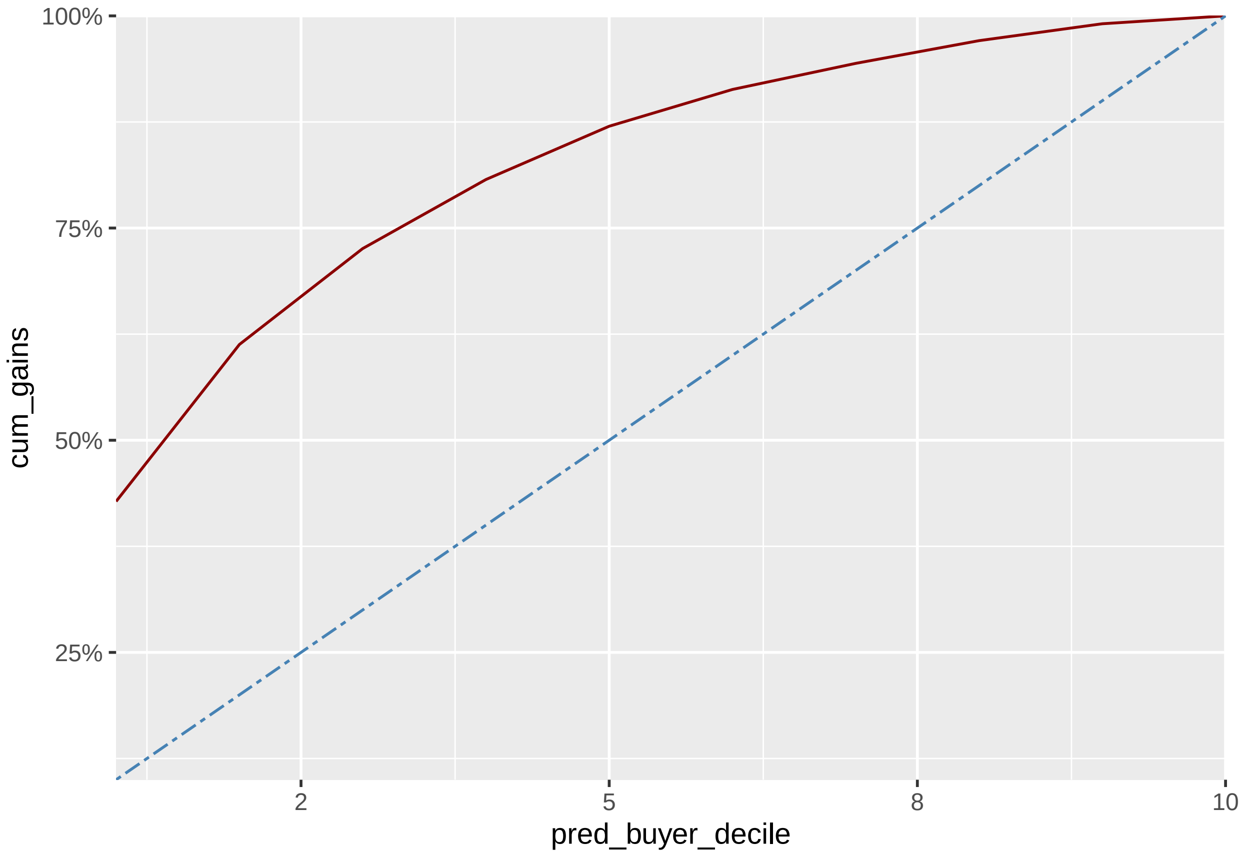
```
# A tibble: 10 x 7
   pred_buyer_decile num_cust cum_cust num_buyers cum_buyers   gains cum_gains
               <int>    <int>    <int>      <int>      <int>   <dbl>     <dbl>
 1                 1     5000     5000       1935       1935 0.428       0.428
 2                 2     5000    10000        836       2771 0.185       0.613
 3                 3     5000    15000        511       3282 0.113       0.726
 4                 4     5000    20000        368       3650 0.0814      0.807
 5                 5     5000    25000        284       3934 0.0628      0.870
 6                 6     5000    30000        196       4130 0.0433      0.913
 7                 7     5000    35000        139       4269 0.0307      0.944
 8                 8     5000    40000        121       4390 0.0268      0.971
 9                 9     5000    45000         90       4480 0.0199      0.991
10                10     5000    50000         42       4522 0.00929     1
```

Question 4

In Excel, create a chart showing the cumulative gains by decile along with a reference line corresponding to 'no model'.

```
ggplot(lift_and_gains, aes(x=pred_buyer_decile)) +
   geom_line(aes(y = cum_gains), color = "darkred") +
```

```
  geom_line(aes(y = perc_cum_cust), color="steelblue", linetype="twodash")  +
  scale_x_continuous(labels=scales::number_format(accuracy = 1), expand = c(0, 0)) +
  scale_y_continuous(labels=scales::percent_format(accuracy = 1), expand = c(0, 0)) +
  theme(
     legend.position = c(0.95, 0.95),
     legend.justification = c("right", "top")
  )
```




```
"
GRADER PLEASE NOTE: there may be a bug in ggplot where with 'expand' - I could not get the x/y axis to 
"
```

```
[1] "\nGRADER PLEASE NOTE: there may be a bug in ggplot where with 'expand' - I could not get the x/y a
```

Hint: Please integrate the Excel-generated charts into the R Notebook you are using for the rest of this assignment. Here is how: • Save the graphs and tables in Excel as pdf files • Place the pdf files into the same directory as your R Notebook for the assignment • Use the "include_graphics()" command to insert each pdf. • For example, suppose your pdf is called "cum_lift.pdf", then insert the code block: • Please note the header of the code block. There you can change the width of the chart (here 70% of page width) and how it is aligned (here centered).

Part 4 - Profitability Analysis

Use the following cost information to assess the profitability of using logistic regression to target customers:

Item	Price/Cost
selling price	$18.00

Item	Price/Cost
cost to mail	$0.50
Wholesale price	$9.00
Shipping costs	$3.00

Question 1

What is the breakeven response rate?

```
price <-  18
cogs <- 9 + 3
marginal_cost <- 0.5

net_rev <- price - cogs
break_even_rate <- marginal_cost / net_rev
percent(break_even_rate, 0.01)
```

```
[1] "8.33%"
```

Question 2

For the customers in the dataset, create a new variable (call it "target") with a value of 1 if the customer's predicted probability is greater than or equal to the breakeven response rate and 0 otherwise. (Hint: in mutate() multiply the TRUE/FALSE expression with "1" to get a 0/1 variable).

```
bbb_final <- bbbPred %>%
  mutate(target=1*(pred_buyer>break_even_rate))
bbb_final %>% select(pred_buyer, target) %>% arrange(desc(target))
```

```
# A tibble: 50,000 x 2
   pred_buyer target
        <dbl>  <dbl>
 1     0.0871      1
 2     0.391       1
 3     0.113       1
 4     0.139       1
 5     0.355       1
 6     0.0867      1
 7     0.254       1
 8     0.587       1
 9     0.170       1
10     0.153       1
# ... with 49,990 more rows
```

Question 3

For the customers in the dataset, if had you used the model to select which customer to target, what percentage of customer would you have targeted? Of those customers you would have targeted, what percentage would have purchased the "Art History of Florence?"

```
targeted_customers <- bbb_final %>%
  filter(target==1) %>%
  summarise(frac_mailed=n()/nrow(bbb_final), resp_rate=mean(buyer))

targeted_customers
```

```
# A tibble: 1 x 2
  frac_mailed resp_rate
        <dbl>     <dbl>
1       0.311     0.214
```

Question 4

For the 500,000 remaining customers, what would the expected profit (in dollars) and the expected return on marketing expenditures have been if BookBinders had mailed the offer to buy "The Art History of Florence" only to customers with a predicted probability of buying that was greater than or equal to the breakeven rate? Make the calculations in R?

```r
ntargeted_customers <- 500000 * targeted_customers$frac_mailed
targeted_resp_rate <- targeted_customers$resp_rate
targeted_costs <- ntargeted_customers * marginal_cost
targeted_revenue <-ntargeted_customers * targeted_resp_rate * net_rev
profit_w_targeting <- targeted_revenue - targeted_costs
paste(
   "Targeted Profit: ", dollar(profit_w_targeting),
   "Targeted ROI: ", percent(profit_w_targeting / (targeted_costs), 0.1)
)
```

```
[1] "Targeted Profit:  $121,580 Targeted ROI:  156.3%"
```

Question 5

For the 500,000 remaining customers, calculate the incremental profit of having used the logistic regression model instead of a mass mailing?

```r
no_targ_revenue <- 500000 * blended_resp_rate * net_rev
no_targ_costs <- 500000 * marginal_cost
profit_wo_targeting <- no_targ_revenue - no_targ_costs
profit_wo_targeting
```

```
[1] 21320
```

```r
paste(
   "Untargeted Profit", dollar(profit_wo_targeting),
   "Untargeted ROI", percent(profit_wo_targeting / (no_targ_costs), 0.1),
   "Incremental Targeted Profit: ", dollar(profit_w_targeting - profit_wo_targeting)
)
```

```
[1] "Untargeted Profit $21,320 Untargeted ROI 8.5% Incremental Targeted Profit:  $100,260"
```