Using R for Basic Statistical Analysis at Bookbinders.

Name Section XX

Preliminaries

IMPORTANT: Before starting this assignment you must go through logistic regression handout we covered in class to correctly install the `kelloggmktg482` package.

```
# install.packages("remotes")
# library(remotes)
# install_version("vip", version="0.3.2", upgrade="never")
# devtools::install_github("blakemcshane/kelloggmktg482", upgrade = "never", force = TRUE)
```

Load packages:

```
library(tidyverse)
library(kelloggmktg482)       # Always load last
```

Read in the data:

```
# use load("filename.Rdata") for .Rdata files

load("bbb.Rdata")
```

Assignment questions and answers

1. Previously, you reported the correlation between customers' total spending on non-book products and on books (see the `nonbook` and `book` variables). Now, report a 95% confidence interval for this correlation (check the `cor.test` function).

```
cor.test(bbb$nonbook,bbb$book)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bbb$nonbook and bbb$book
## t = 35.648, df = 49998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1488761 0.1659721
## sample estimates:
##       cor
## 0.1574359
```

2. Report the output of a linear regression predicting customers' total spending on non-book products from their spending on book products as well as 95% confidence intervals for the intercept and slope.

```r
colnames(bbb)
```

```
##  [1] "acctnum"   "gender"    "state"     "zip"       "zip3"      "first"
##  [7] "last"      "book"      "nonbook"   "total"     "purch"     "child"
## [13] "youth"     "cook"      "do_it"     "reference" "art"       "geog"
## [19] "buyer"     "training"
```

```r
regress <- lm(nonbook ~ book, data = bbb)
summary(regress)
```

```
##
## Call:
## lm(formula = nonbook ~ book, data = bbb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -168.985  -75.382   -0.253   75.190  164.199
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 144.47388    0.61572  234.64   <2e-16 ***
## book          0.36331    0.01019   35.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.98 on 49998 degrees of freedom
## Multiple R-squared:  0.02479,    Adjusted R-squared:  0.02477
## F-statistic:  1271 on 1 and 49998 DF,  p-value: < 2.2e-16
```

3. Previously, you reported the proportion of customers by gender who bought "The Art History of Florence" (see the `buyer` variable). Now, report a 95% confidence interval for the difference in the proportion of buyers between genders (check the `prop.test` function).

```r
prop.test(table(bbb$gender, bbb$buyer))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  table(bbb$gender, bbb$buyer)
## X-squared = 423.34, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06181926 -0.05018557
## sample estimates:
##    prop 1    prop 2
## 0.8722602 0.9282626
```

4. Report the results of a chi-square test of the association between the state in which the customer lives and whether or not the customer bought "The Art History of Florence" (check the `chisq.test` function).

```
chisq_result <- chisq.test(table(bbb$state, bbb$buyer))
```

```
## Warning in chisq.test(table(bbb$state, bbb$buyer)): Chi-squared approximation
## may be incorrect
```

```
print(chisq_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(bbb$state, bbb$buyer)
## X-squared = 23.549, df = 14, p-value = 0.0519
```

5. Previously, you reported the total number of purchases and the average number of purchases by gender (see the `purch` variable). Now, report the result of a t-test comparing the average number of purchases by gender as well as a 95% confidence interval for the difference in the average number of purchases between genders (check the `t.test` function).

```
 t.test(purch ~ gender, data = bbb)
```

```
##
##  Welch Two Sample t-test
##
## data:  purch by gender
## t = 46.919, df = 29665, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.515036 1.647136
## sample estimates:
## mean in group M mean in group F
##        4.943287        3.362200
```

6. Repeat the same analysis but using a linear regression rather than a t-test.

```
regress <- lm(purch ~ gender, data = bbb)
summary(regress)
```

```
##
## Call:
## lm(formula = purch ~ gender, data = bbb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.943 -2.362 -1.362  2.057  8.638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.94329    0.02628  188.13   <2e-16 ***
## genderF     -1.58109    0.03220  -49.11   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.395 on 49998 degrees of freedom
## Multiple R-squared:  0.04601,    Adjusted R-squared:  0.04599
## F-statistic:  2412 on 1 and 49998 DF,  p-value: < 2.2e-16
```

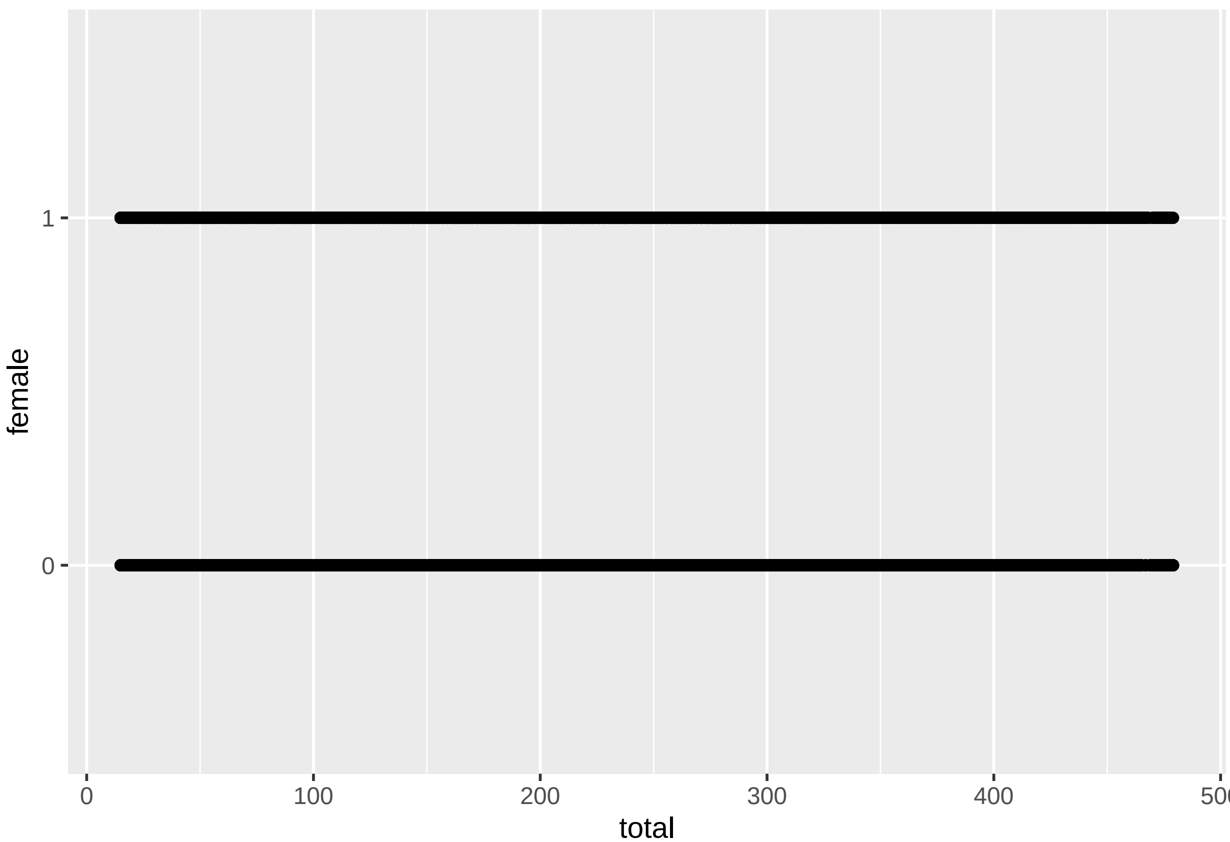
7. In class, we examined a logistic regression predicting the `buyer` variable. In this assigment, we want to predict whether the customer is female versus male. First, we must create `female` binary variable using the following:

```r
bbb <- bbb %>% mutate(female = factor(1 * (gender=="F")))
```

Before proceeding to fit a logistic regression, first provide graphical evidence that the variables `do_it`, `total`, and `last` are or are not predictive of whether the customer is female versus male.

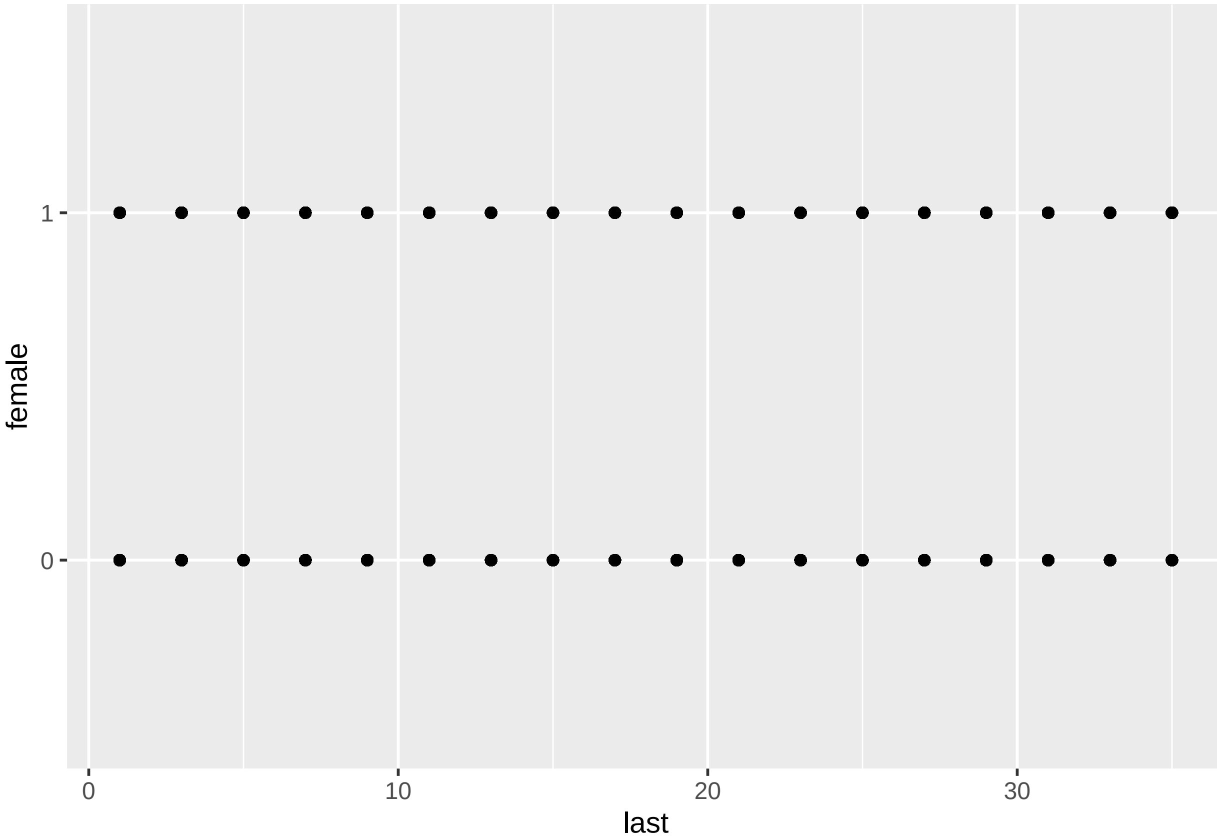
```r
library(dplyr)

library(ggplot2)
gendF <- bbb %>%
  ggplot(aes(x = total, y = female)) +  geom_point()
print(gendF)
```

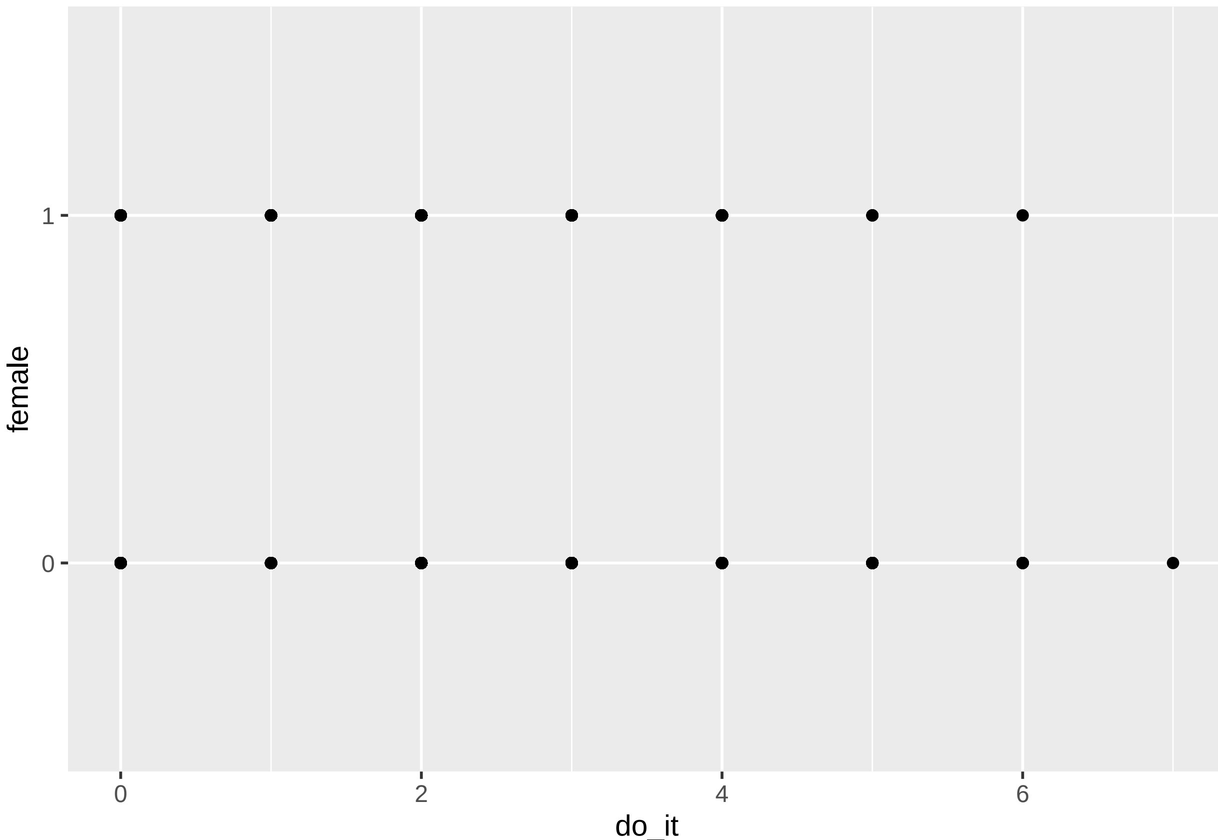



```r
gendF <- bbb %>%
  ggplot(aes(x = last, y = female)) +  geom_point()
```

```r
print(gendF)
```




```r
gendF <- bbb %>%
  ggplot(aes(x = do_it, y = female)) +  geom_point()
print(gendF)
```




```
# df <- bbb %>%
 #  summarise(sum_total = sum(total))
#  ggplot(df, aes(x = female, y = sum_total)) + geom_bar(stat="identity")
#   print (df)
```

8. Report the output of a logistic regression predicting `female` using the variables `last`, `total`, `child`, `youth`, `cook`, `do_it`, `reference`, `art`, and `geog` as well as 95% confidence intervals for the intercept and coefficients.

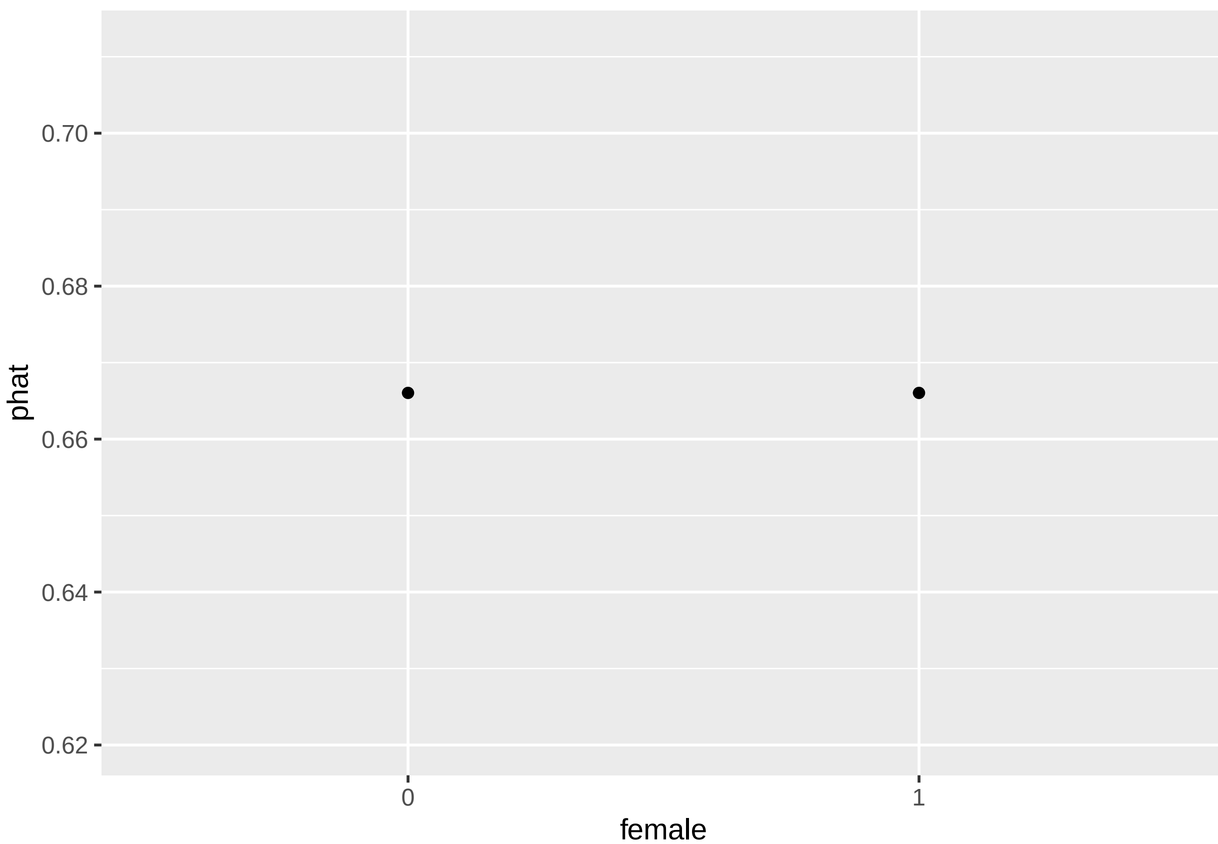
```
logistic_model <- glm(female ~ last + total + child + youth + cook + do_it + reference + art + geog,  fa

summary(logistic_model)
```

```
##
## Call:
## glm(formula = female ~ last + total + child + youth + cook +
##     do_it + reference + art + geog, family = binomial, data = bbb)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7307  -1.2323   0.7287   0.7774   2.6434
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.228e+00  2.750e-02  44.660  < 2e-16 ***
```

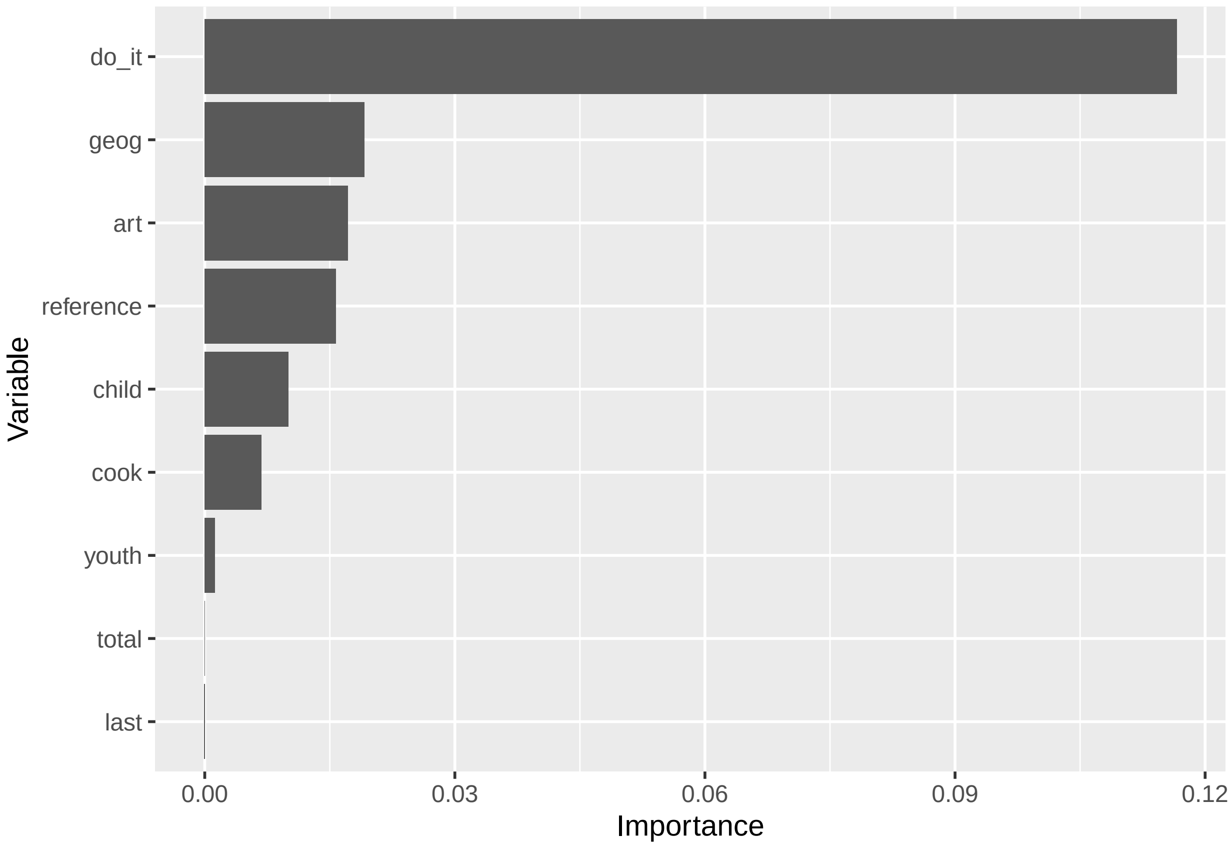
```
## last        -4.903e-04  1.220e-03  -0.402 0.687666
## total        9.244e-05  1.147e-04   0.806 0.420139
## child       -3.582e-02  9.824e-03  -3.646 0.000266 ***
## youth       -1.137e-02  1.525e-02  -0.745 0.456040
## cook        -2.723e-02  9.381e-03  -2.902 0.003704 **
## do_it       -7.634e-01  1.449e-02 -52.705  < 2e-16 ***
## reference   -9.022e-02  1.693e-02  -5.330 9.84e-08 ***
## art         -8.414e-02  1.518e-02  -5.542 3.00e-08 ***
## geog        -7.399e-02  1.257e-02  -5.885 3.97e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 63695  on 49999  degrees of freedom
## Residual deviance: 59172  on 49990  degrees of freedom
## AIC: 59192
##
## Number of Fisher Scoring iterations: 4
```

```r
pardepplot(logistic_model, pred.var = "female", data=bbb)
```



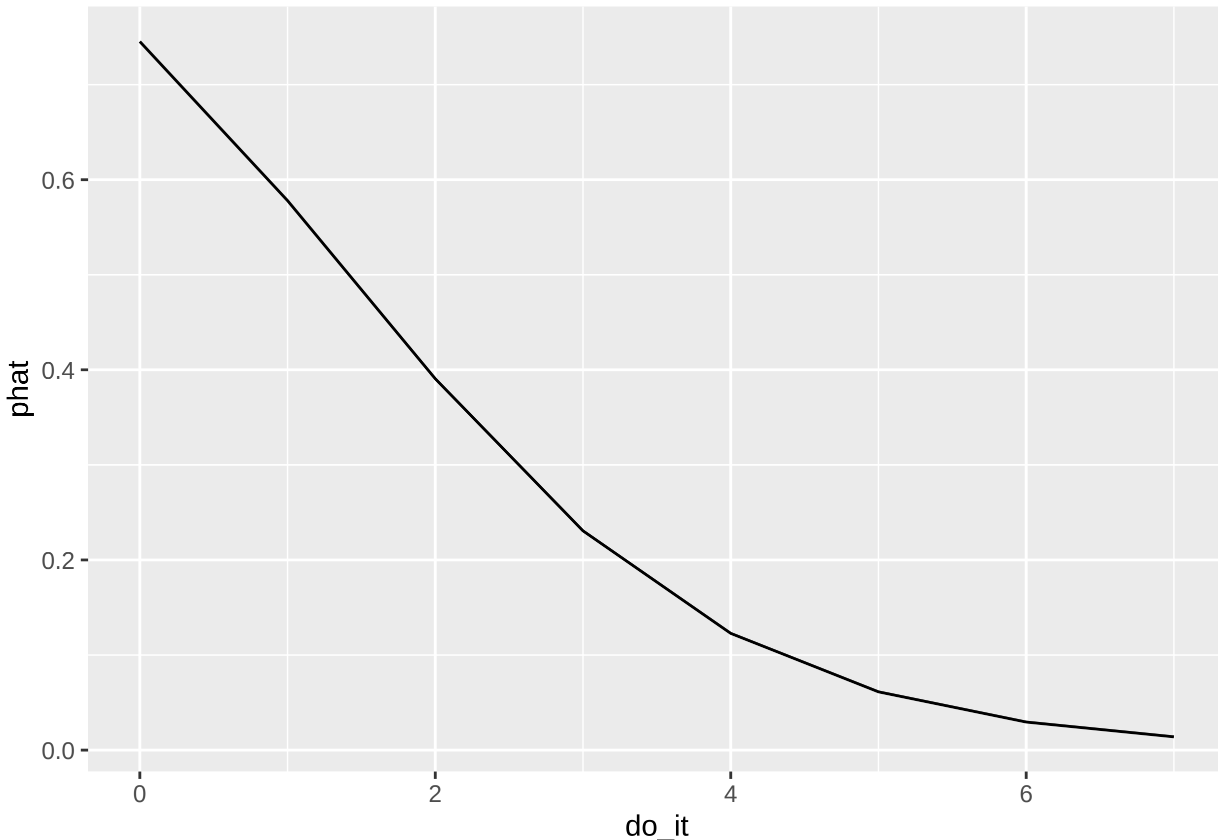
9. Report the variable importance for all variables included in the logistic regression.

```r
varimpplot(logistic_model, target = "female")
```

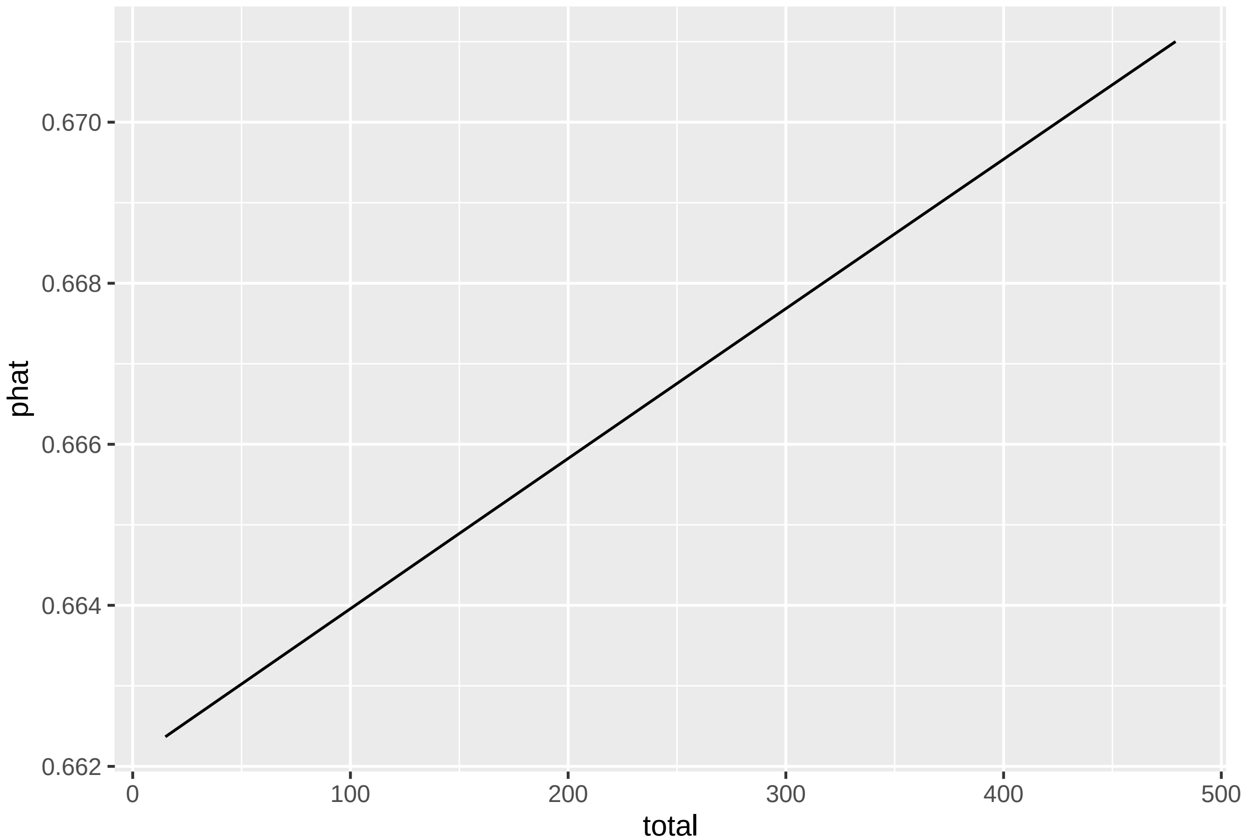


10. Report partial dependence plots for all variables included in the logistic regression. Comment on whether the partial dependence plots for `do_it`, `total`, and `last` are consistent or inconsistent with those you found in Question 7 and explain any consistency or inconsistency.

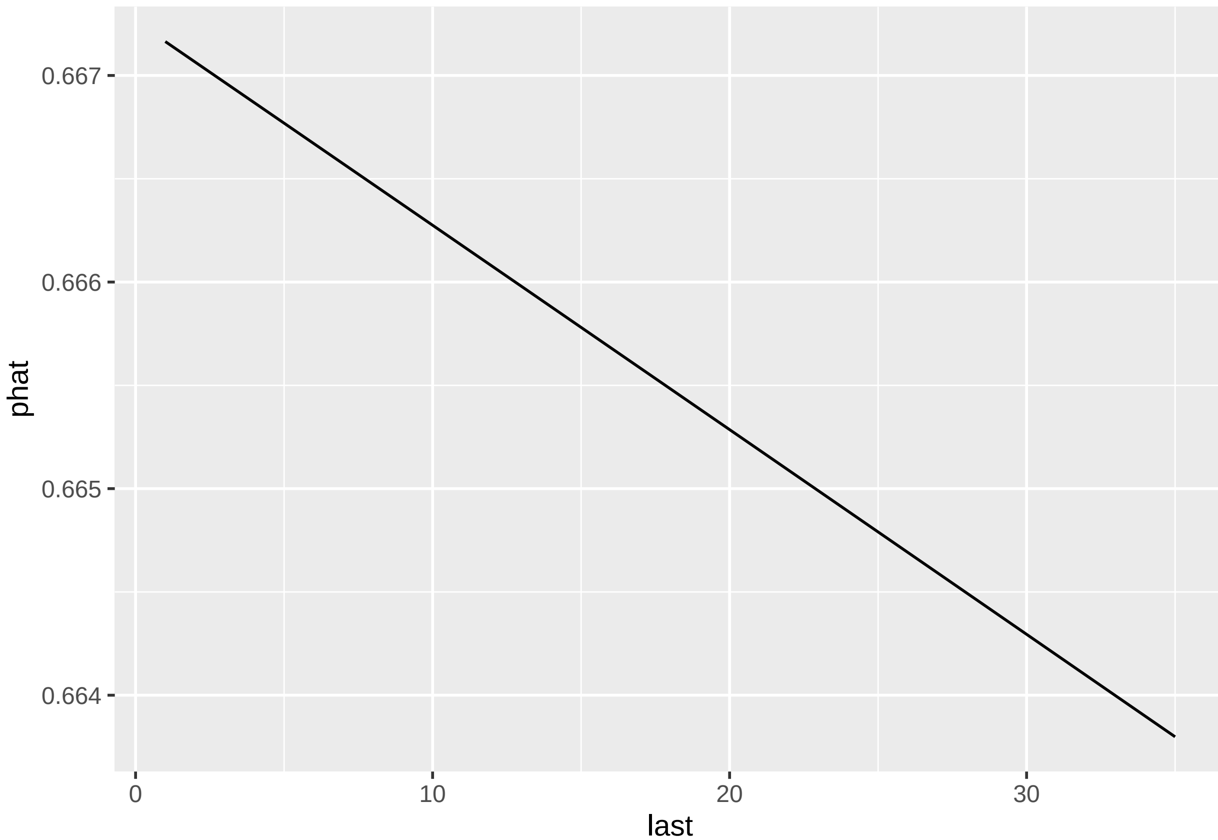
```r
pardepplot(logistic_model, pred.var="do_it", data=bbb)
```




```r
pardepplot(logistic_model, pred.var="total", data=bbb)
```




```
pardepplot(logistic_model, pred.var="last", data=bbb)
```

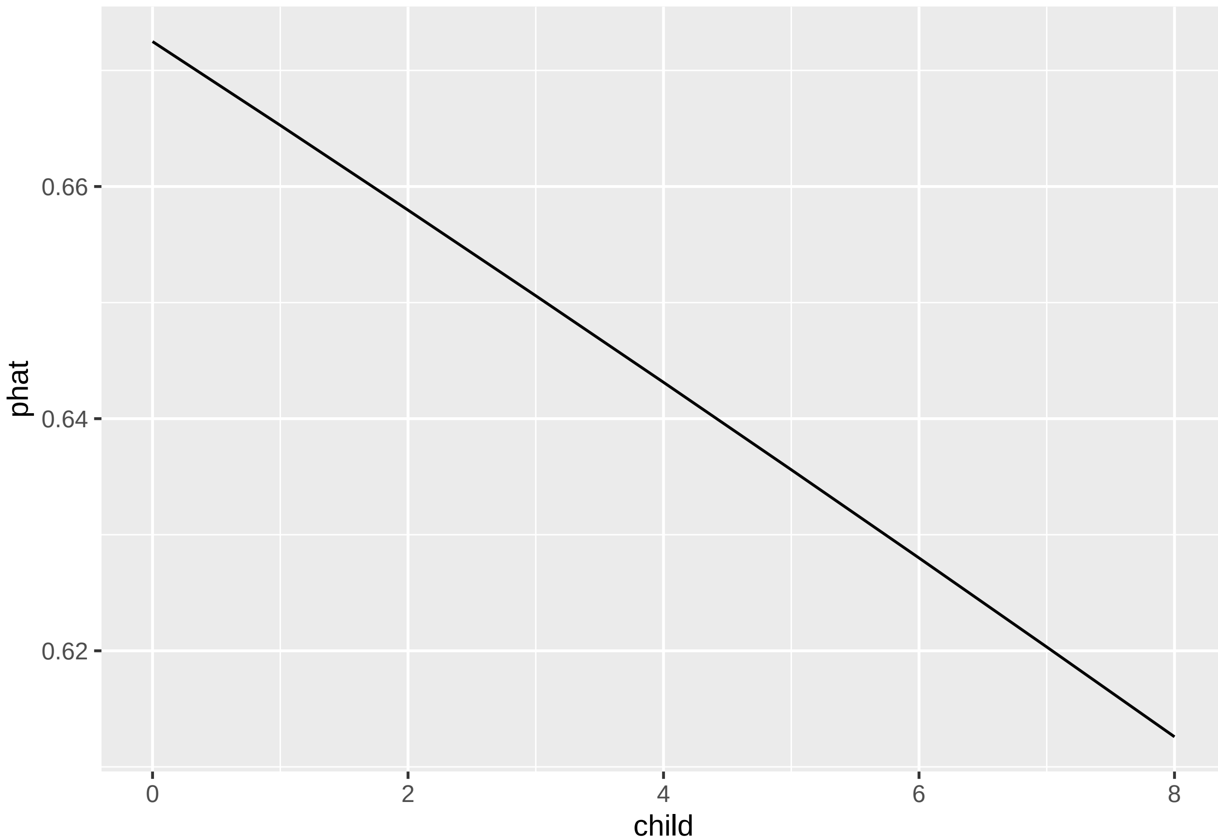


The partial dependency plots of do_it and last are decreasing while that of total is increasing. The slope of do_it is smaller than that of last.

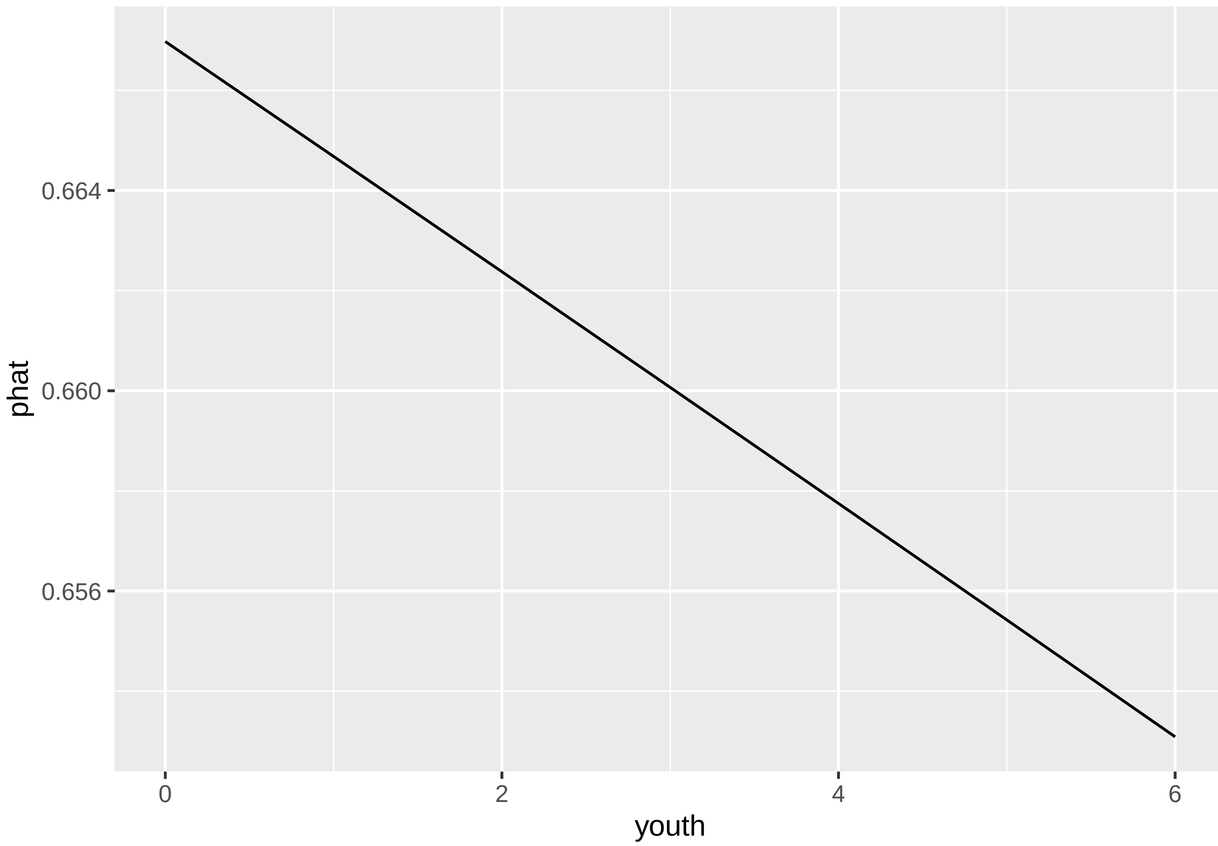
When compared to that of Q7 plots, these plots are much clearer. The Q7 plots are not useful in predicting the gender while the logistic regression provides a much better way to predict.

The do_it plot do not have a linear relationship while other variables have.

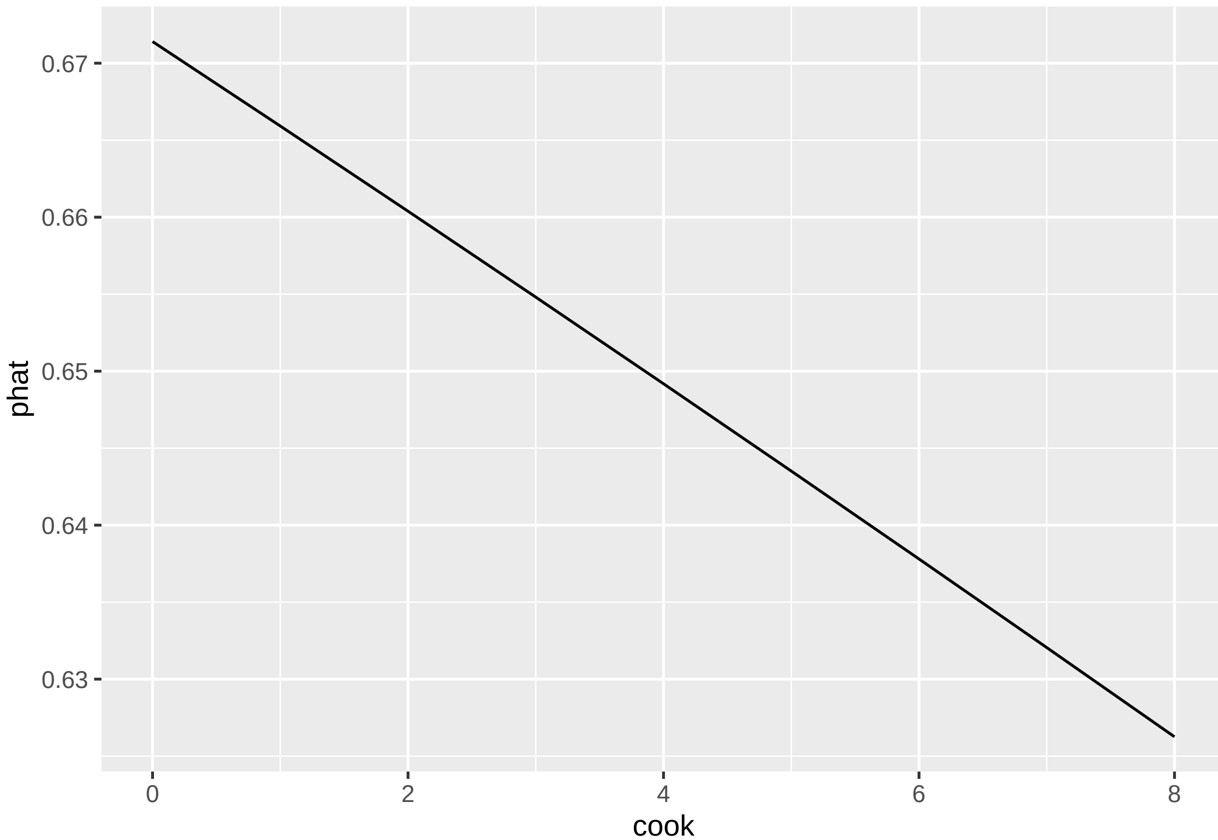
```
pardepplot(logistic_model, pred.var="child", data=bbb)
```



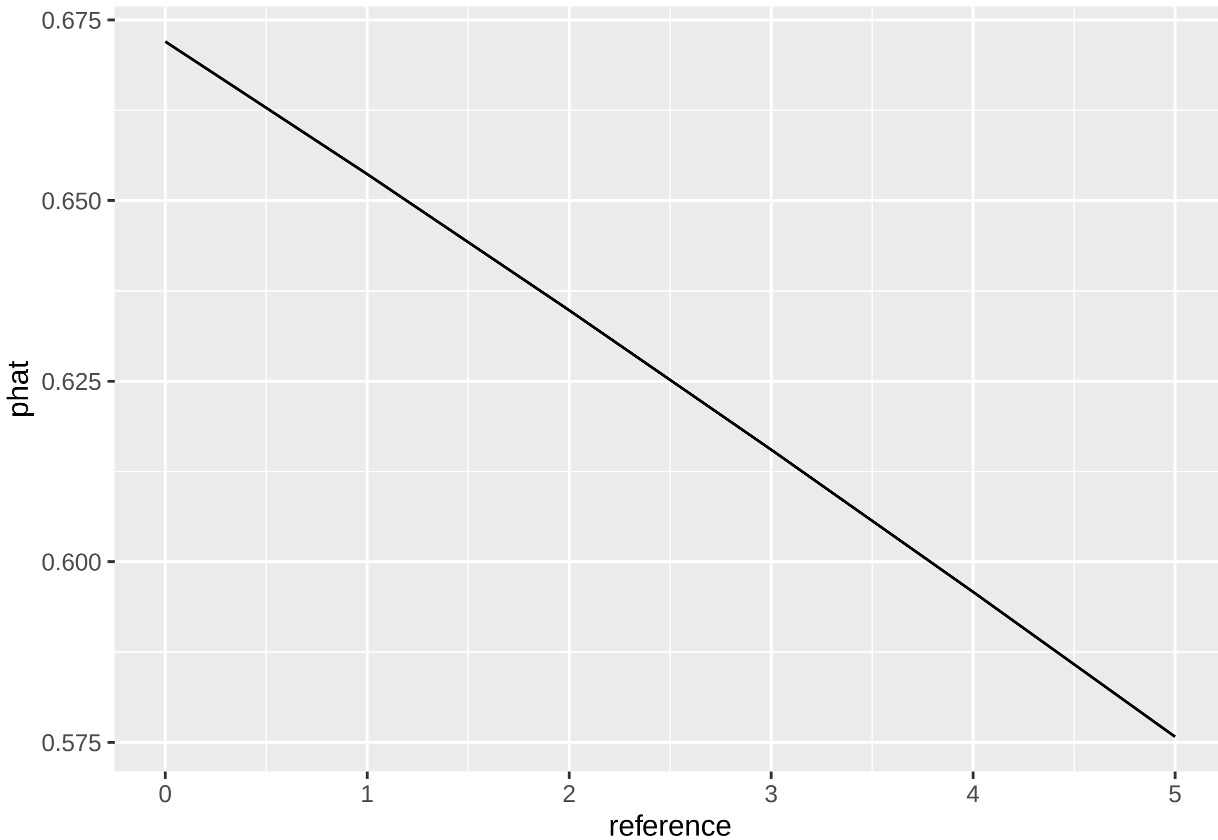

```
pardepplot(logistic_model, pred.var="youth", data=bbb)
```



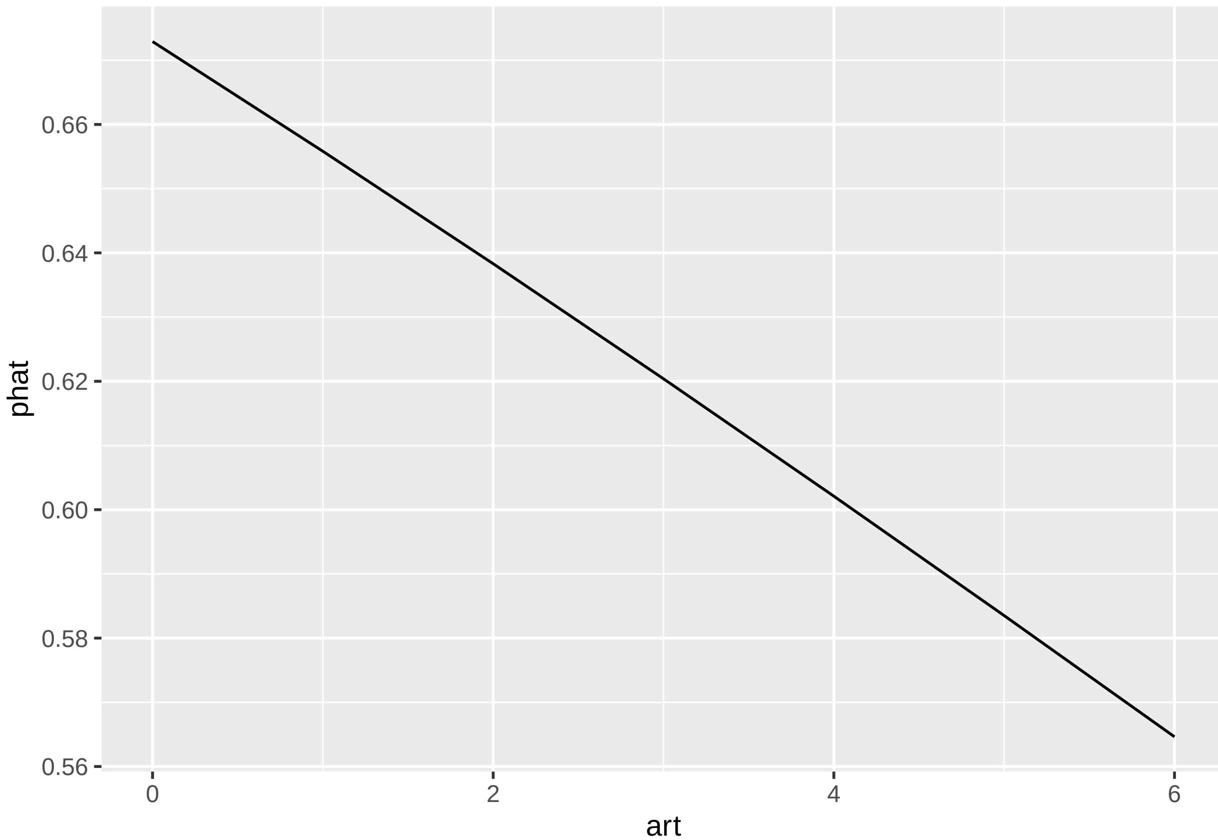

```r
pardepplot(logistic_model, pred.var="cook", data=bbb)
```



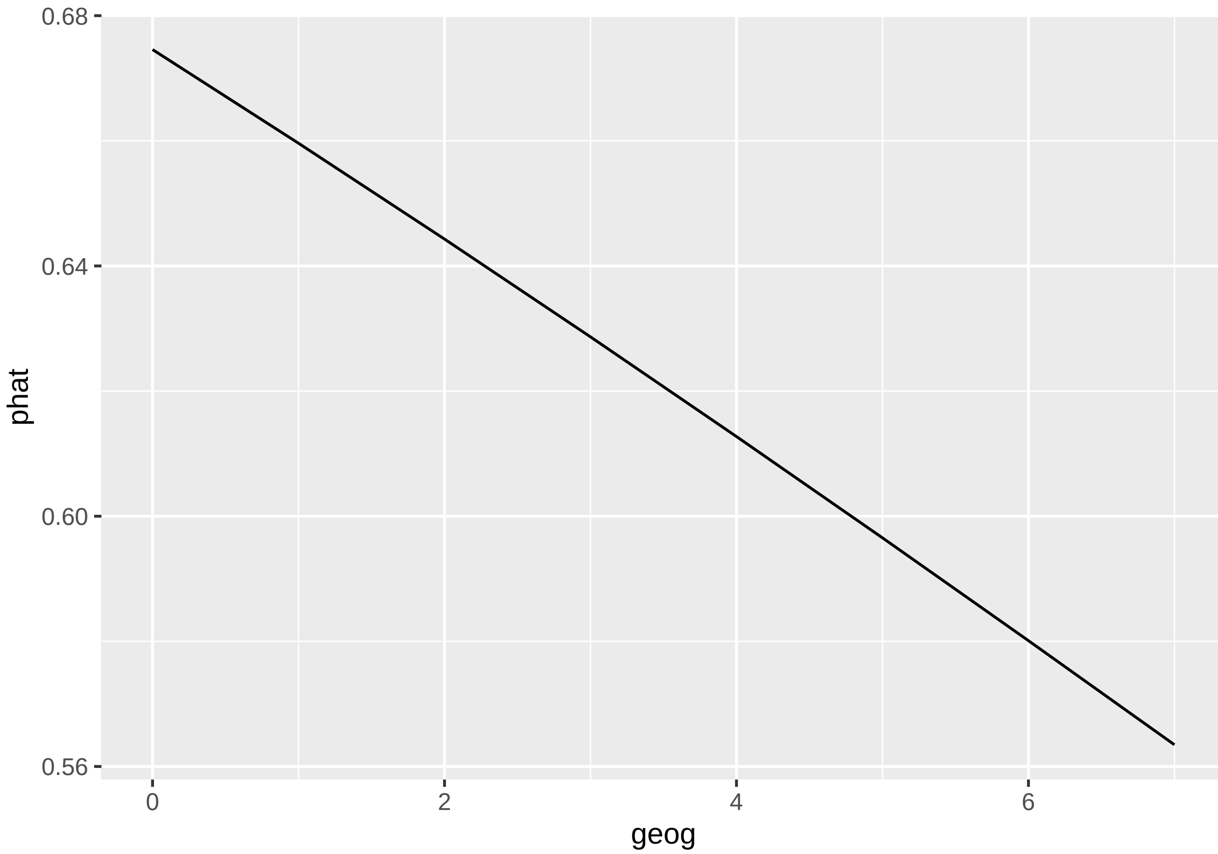

```r
pardepplot(logistic_model, pred.var="reference", data=bbb)
```




```
pardepplot(logistic_model, pred.var="art", data=bbb)
```




```
pardepplot(logistic_model, pred.var="geog", data=bbb)
```



11. Add the predicted probabilities from the logistic regression to the bbb dataframe. Then report histograms of the predicted probabilities separately for each value of `do_it`. Comment on whether the results are consistent with the results of the partial dependence plots from Question 10.

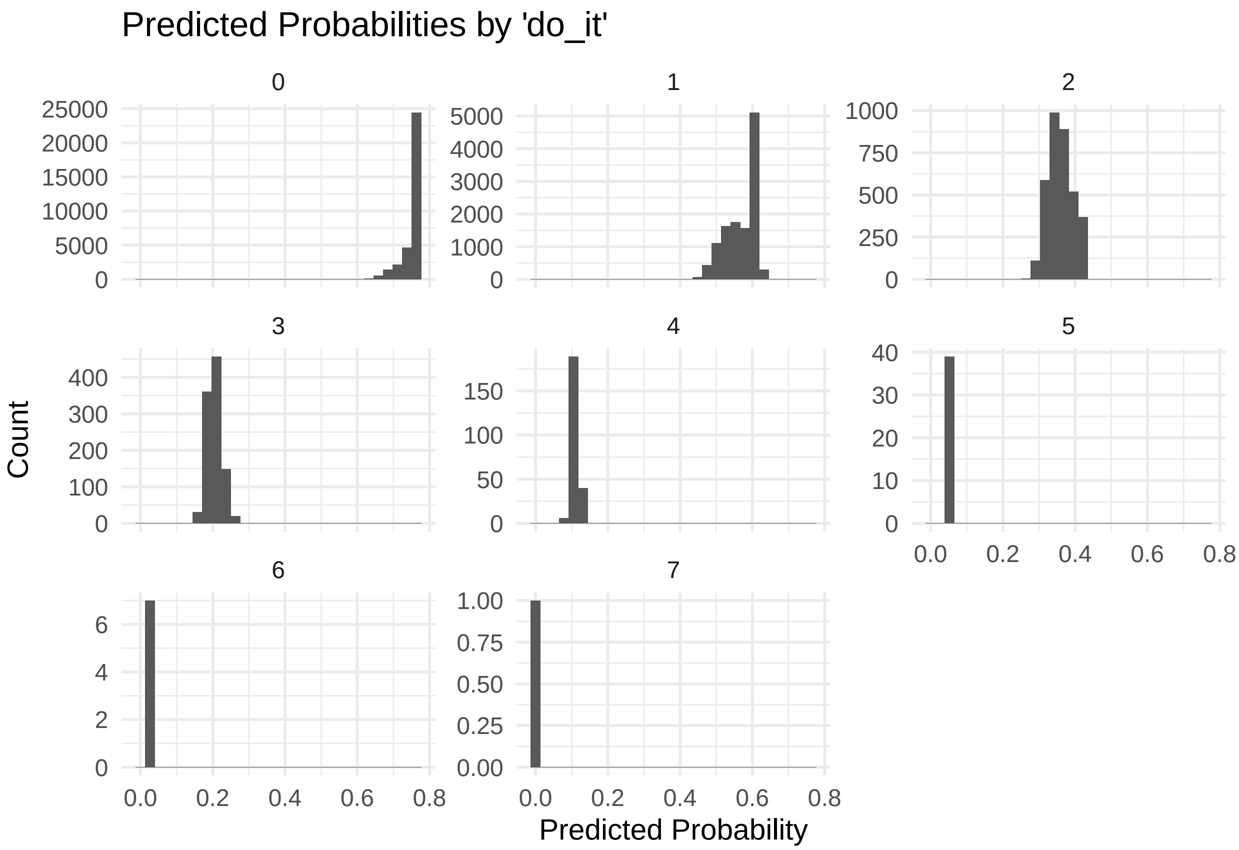
Hint: To accomplish this, recall facet_wrap() from the R tutorial "R_Kellogg_Tutorial_mktg482.pdf." You may want to use "scales='free_y' " within facet_wrap() because relatively few people bought a larger number of do-it-yourself books (an alternative way to achieve the same purpose is to use aes(y=..density..) within geom_histogram()).

```
bbb$predicted_prob <- predict(logistic_model, type = "response")

library(ggplot2)

ggplot(bbb, aes(x = predicted_prob)) +
  geom_histogram() +
  facet_wrap(~ do_it, scales = "free_y") +
  labs(title = "Predicted Probabilities by 'do_it'",
       x = "Predicted Probability",
       y = "Count") +
  theme_minimal()
```

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



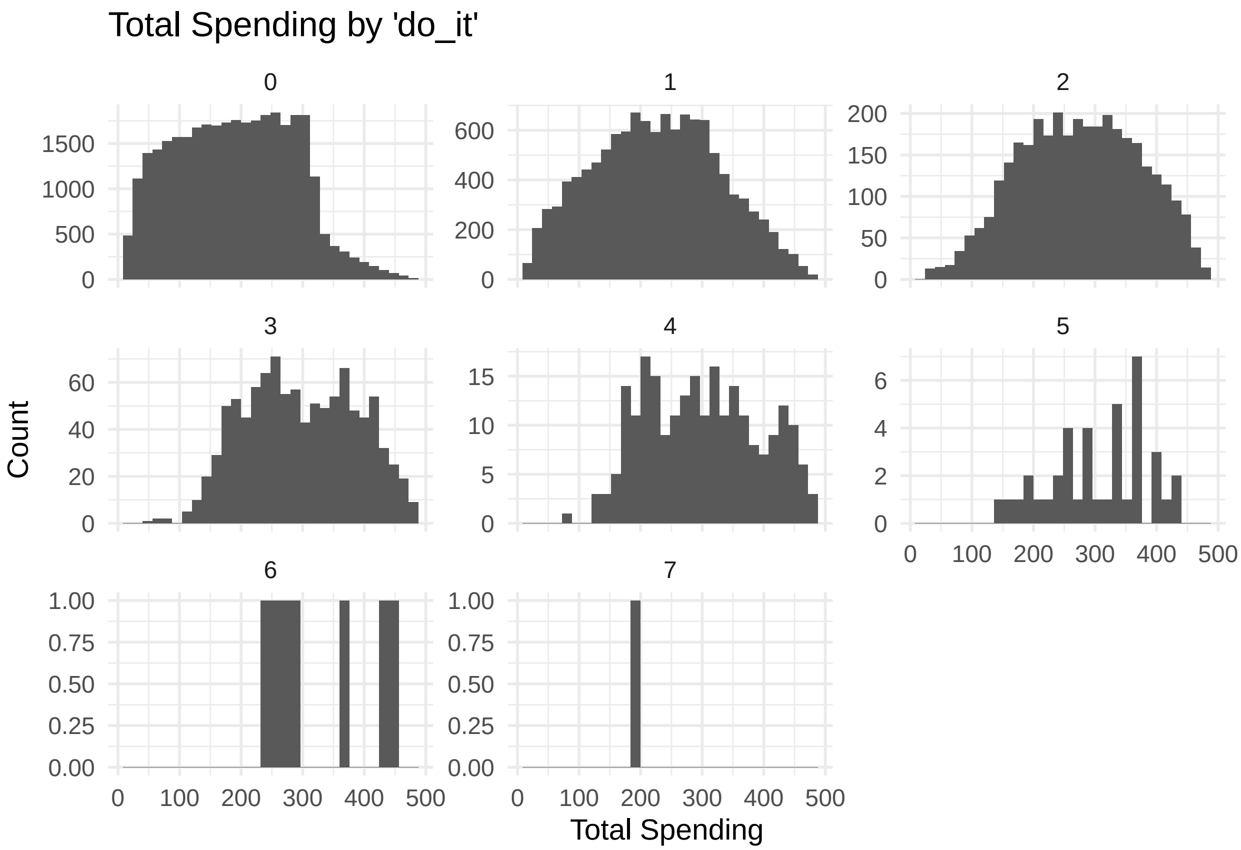
#The do_it plot do not have a linear relationship while other variables have. This non linearity is a bit clear from the predicted probabilities separated by do_it values

12. Report histograms of `total` separately for each value of `do_it`. Does this plot help explain consistency or inconsistency discussed in Question 10?

```
library(ggplot2)

# Create histograms of 'total' faceted by 'do_it'
ggplot(bbb, aes(x = total)) +
  geom_histogram() +
  facet_wrap(~ do_it, scales = "free_y") +
  labs(title = "Total Spending by 'do_it'",
       x = "Total Spending",
       y = "Count") +
  theme_minimal()
```

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



#this explains the non linearity of do_it better