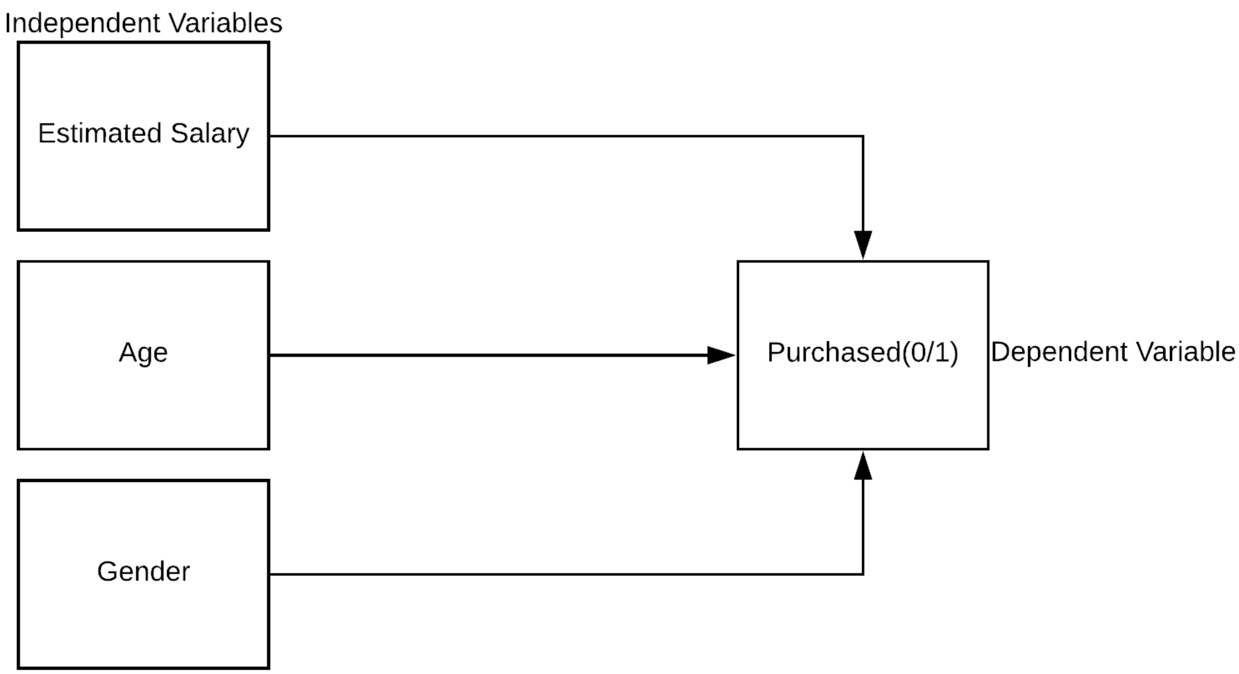
Code

Conceptual model

The underlying conceptual model here is to study if the age,gender and/or estimated salary of a person can explain whether they purchased a product advertised online.



Logistic regression

Hide

```
mydata <- read.csv("D:\\Learning Material\\SDs\\Social_Network_Ads_Cleaned.csv")
mydata$Purchased<-factor(mydata$Purchased, levels = c(0:1), labels = c("No","Yes"))
cat("The levels of gender variable: \n")
```

```
The levels of gender variable:
```

Hide

```
levels(mydata$Gender)
```

```
[1] "Female" "Male"
```

Hide

```
cat("The levels of purchased variable: \n")
```

```
The levels of purchased variable:
```

Hide

```
levels(mydata$Purchased)
```

```
[1] "No"  "Yes"
```

Hide

```
cat("The summary of the data: \n")
```

```
The summary of the data:
```

Hide

```
summary(mydata)
```

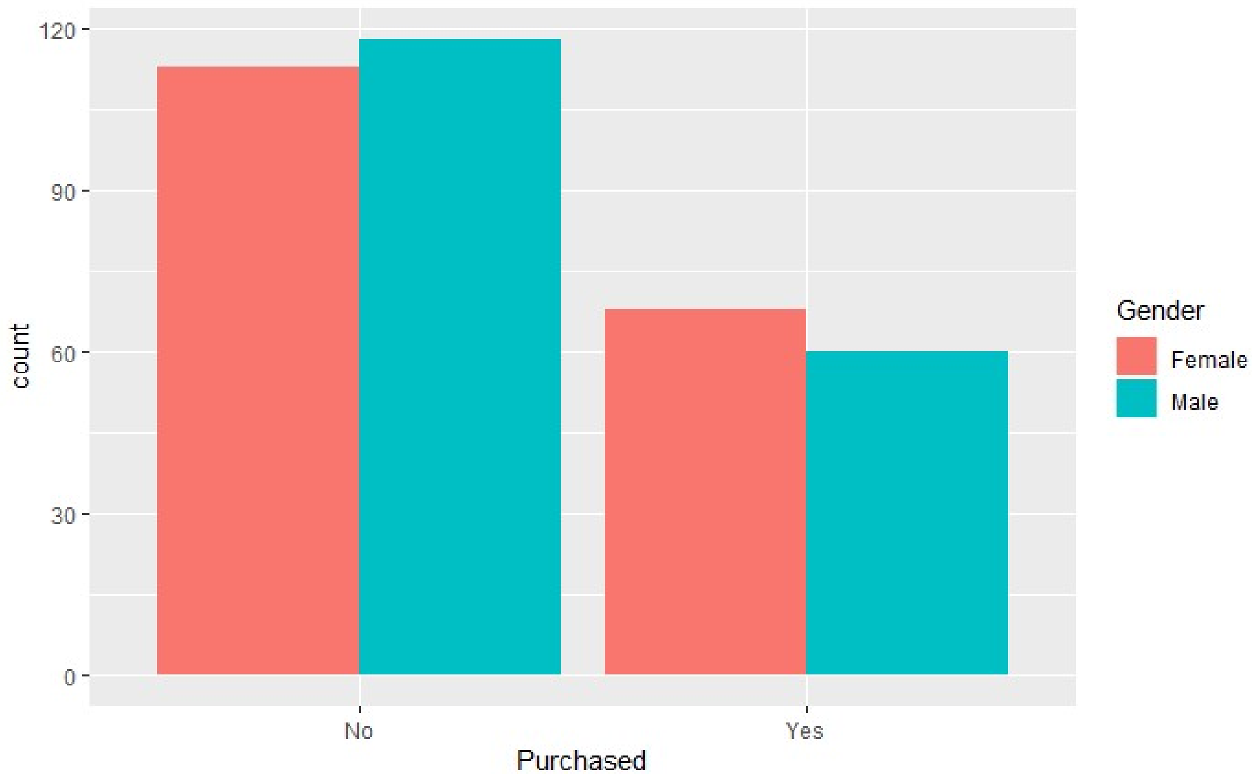
```
    Gender          Age         EstimatedSalary  Purchased
 Female:181   Min.   :18.00   Min.   : 15000   No :231
 Male  :178   1st Qu.:29.50   1st Qu.: 43000   Yes:128
              Median :37.00   Median : 69000
              Mean   :37.64   Mean   : 69462
              3rd Qu.:46.00   3rd Qu.: 88000
              Max.   :60.00   Max.   :150000
```

As taught in class, we made sure to factorize the categorical variable namely gender and purchased as can be seen above. We were also interested in looking at the summary of the data to understand it's contents.

Visualizing the data

Hide

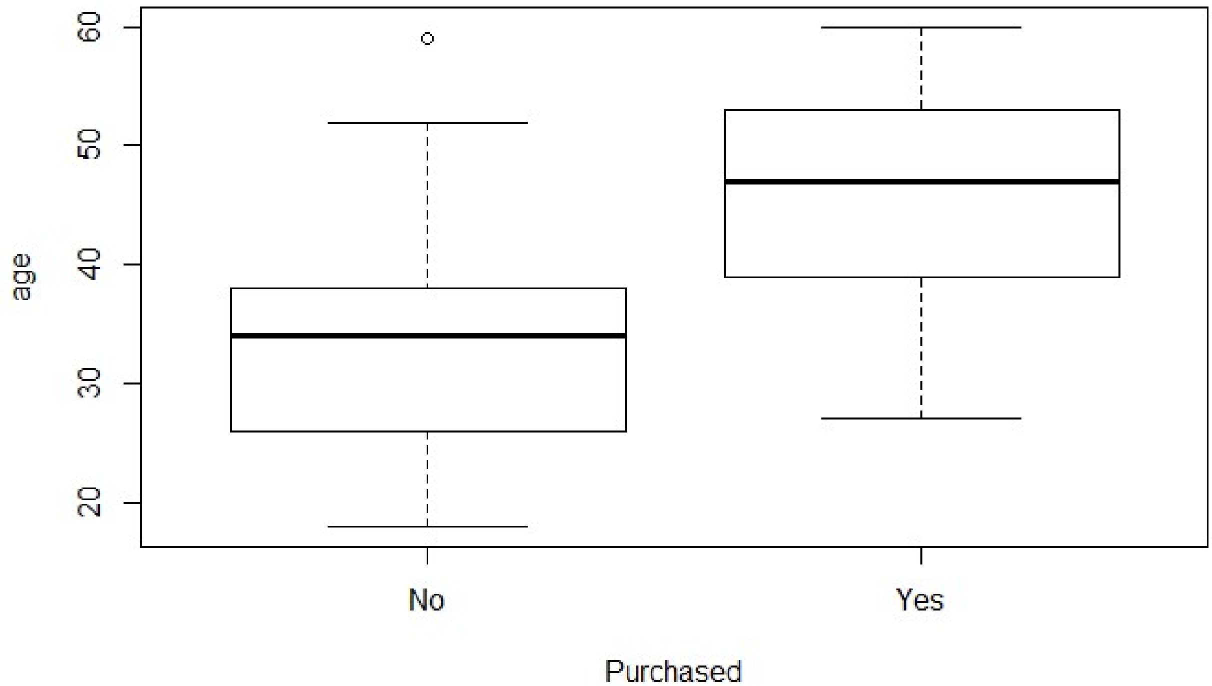
```
ggplot(mydata, aes(Purchased, ..count..)) + geom_bar(aes(fill = Gender), position = "d
odge")
```



Here we see that in general women have made more online purchases than men. And that in general there are less successful purchases online.

Hide

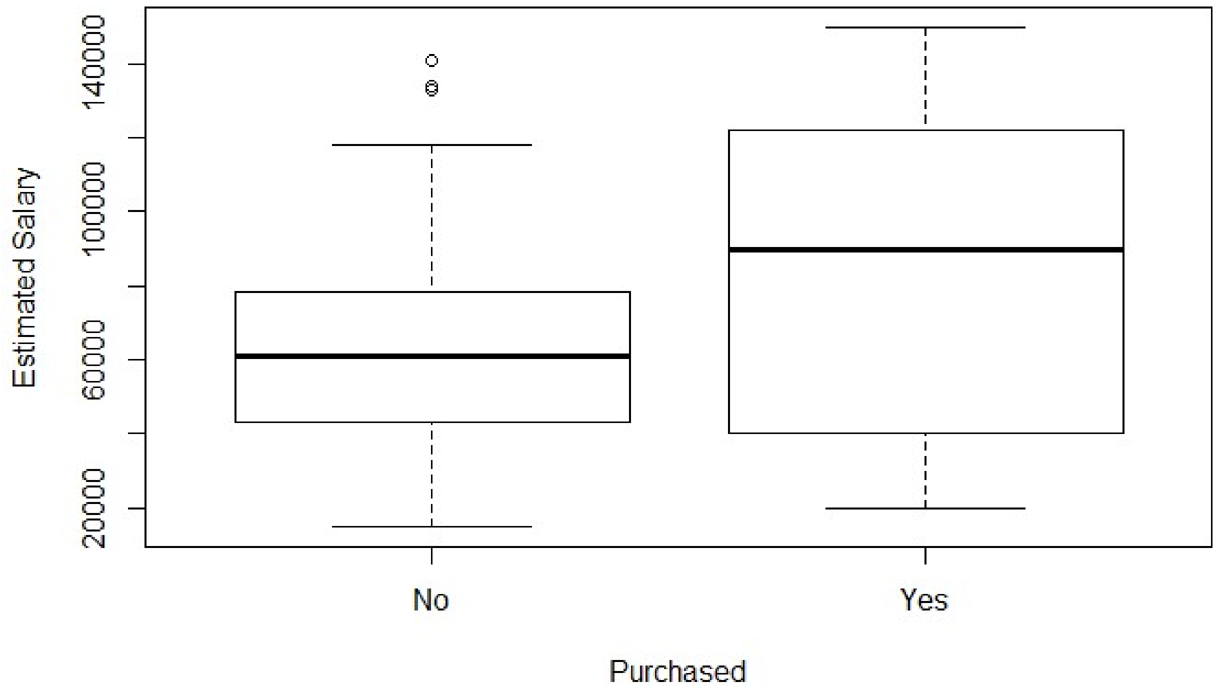
```
plot(mydata$Purchased, mydata$Age, xlab = "Purchased", ylab = "age")
```



Here we clearly see that older people make more purchases.

Hide

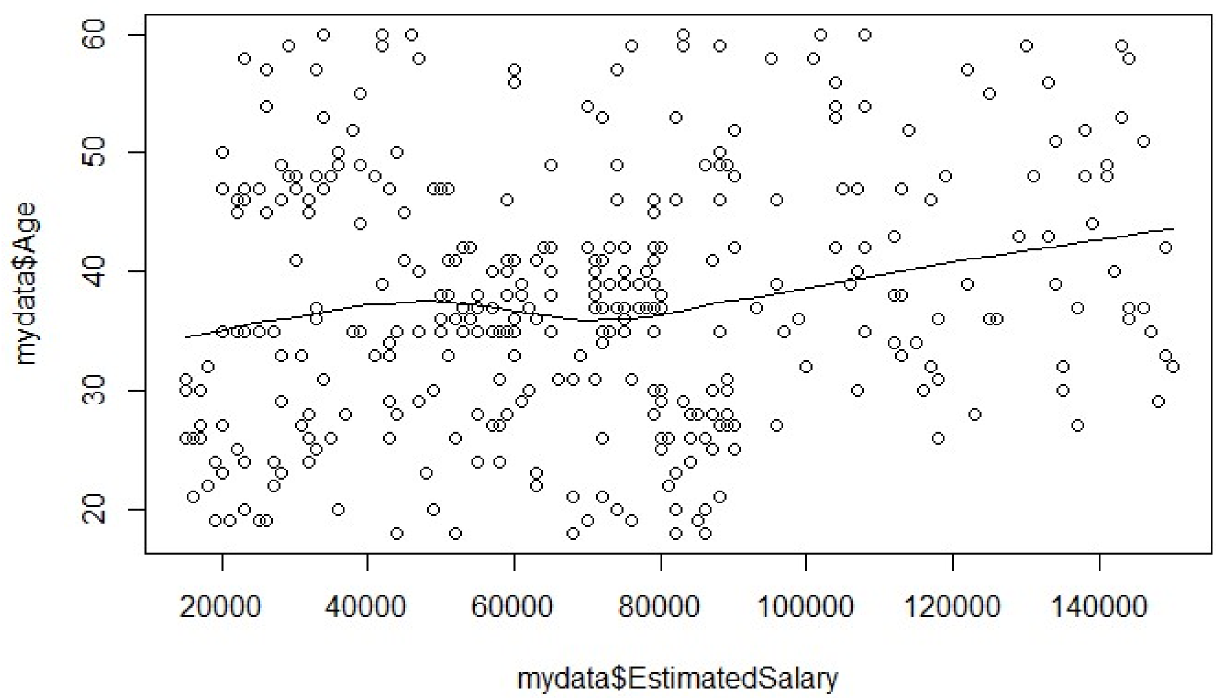
```
plot(mydata$Purchased, mydata$EstimatedSalary, xlab = "Purchased", ylab = "Estimated S
alary")
```



Here we see that usually people with more money make more purchases which seems to make sense.

Hide

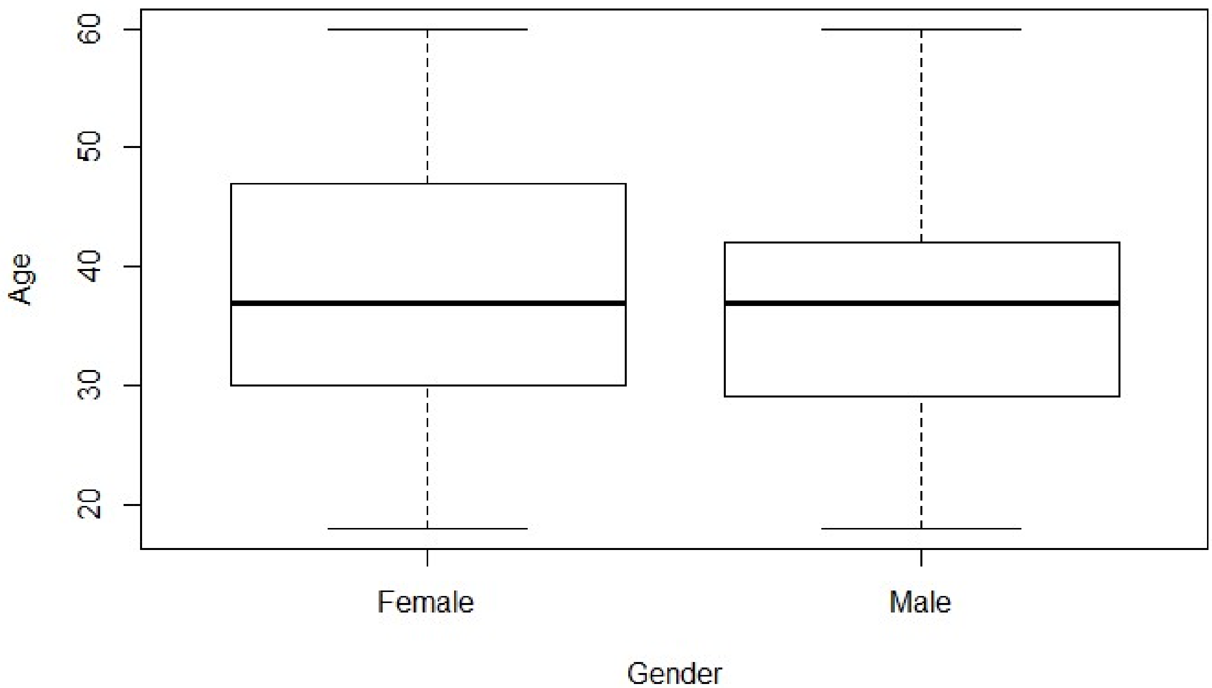
```
scatter.smooth(mydata$EstimatedSalary,mydata$Age)
```



Here as expected we see that there is a slight positive correlation between age and estimated salary. People with more money are usually also older.

Hide

```
plot(mydata$Gender, mydata$Age, xlab = "Gender", ylab = "Age")
```



Here we see that variance of age for both genders is similar in our data. This is good because we can truly consider gender to be independent of age.

Generalised Linear Models

Hide

```
mydata$EstimatedSalary <-mydata$EstimatedSalary - mean(mydata$EstimatedSalary)
mydata$Age <-mydata$Age - mean(mydata$Age)
```

First we re-scale our numeric predictors(age & estimated salary) by centring them around their mean. This allows us to interpret the estimates of our model as deviations from their mean values.

Then we instantiate the base model and incrementally extend it to incorporate all the different combinations of our 3 predictor variables as described in the conceptual model i.e Age, Gender & Estimated Salary. This results in 8 models including the base model as can be seen below.

Hide

```
model0 <- glm(Purchased ~ 1, data = mydata, family = binomial())
model1 <- glm(Purchased ~ Gender , data = mydata, family = binomial())
model2 <- glm(Purchased ~ Age, data = mydata, family = binomial())
model3 <- glm(Purchased ~ EstimatedSalary, data = mydata, family = binomial())
model4 <- glm(Purchased ~ Age+EstimatedSalary, data = mydata, family = binomial())
model5 <- glm(Purchased ~ Gender+Age, data = mydata, family = binomial())
model6 <- glm(Purchased ~ Gender + EstimatedSalary, data = mydata, family = binomial
())
model7 <- glm(Purchased ~ Gender + Age + EstimatedSalary, data = mydata, family = bino
mial())

pander(anova(model0,model1,test = "Chisq" ),
       caption = "The effect of adding gender.")
```

```
---------------------------------------------------
 Resid. Df   Resid. Dev   Df   Deviance   Pr(>Chi)
----------- ------------ ---- ---------- ----------
    358        467.7      NA      NA         NA

    357        467.1      1     0.5835     0.4449
---------------------------------------------------

Table: The effect of adding gender.
```

Hide

```
pander(anova(model0,model2,test = "Chisq" ),
       caption = "The effect of adding Age.")
```

```
----------------------------------------------------
 Resid. Df   Resid. Dev   Df   Deviance   Pr(>Chi)
----------- ------------ ---- ---------- -----------
    358        467.7      NA      NA         NA

    357        305.8      1     161.9     4.245e-37
----------------------------------------------------

Table: The effect of adding Age.
```

Hide

```
pander(anova(model0,model3,test = "Chisq" ),
       caption = "The effect of adding Estimated Salary.")
```

```
--------------------------------------------------------
 Resid. Df   Resid. Dev   Df   Deviance   Pr(>Chi)
----------- ------------ ---- ---------- -----------
    358        467.7      NA      NA         NA

    357        420.7      1     46.97     7.203e-12
--------------------------------------------------------

Table: The effect of adding Estimated Salary.
```

From the above model comparisons using the anova function, we see that including gender does not have a significant impact to the fit of our model whereas including age or estimated salary do have a significant impact with a p value less than 1%.

Now that we know which models are relevant to the analysis, we can look at the summary of the correct model which includes only the significant predictors by using the summary function.

Hide

```
pander(summary(model4),
       caption = "Summary results of model 1")
```

```
---------------------------------------------------------------------
                  Estimate    Std. Error   z value   Pr(>|z|)
--------------------- ----------- ------------ --------- -----------
   **(Intercept)**      -1.139       0.177      -6.432    1.261e-10

       **Age**          0.2246      0.02628      8.545    1.282e-17

 **EstimatedSalary**   3.438e-05    5.47e-06     6.286    3.267e-10
---------------------------------------------------------------------


(Dispersion parameter for  binomial  family taken to be  1 )


-------------------- --------------------------
   Null deviance:     467.7  on 358  degrees of
                               freedom

 Residual deviance:   253.7  on 356  degrees of
                               freedom
-------------------- --------------------------
```

Using the results shown above, we can see the estimates of the significant predictors in our model i.e age and estimated salary. We see based on the estimates that age has a more profound impact to purchases being made with respect to unit changes. Finally we can see that the standard errors are low which means that our estimates can be considered reliable.

Crosstable predicted and observed responses

Hide

```
mydata$Purchasedpred[fitted(model4) <=0.5] <- 0
mydata$Purchasedpred[fitted(model4) > 0.5] <- 1
mydata$Purchasedpred<-factor(mydata$Purchasedpred, levels = c(0:1), labels = c("No","Y
es"))
CrossTable(mydata$Purchasedpred, mydata$Purchased, prop.r=FALSE, prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq=FALSE, format = "SPSS",
           fisher = FALSE, chisq = TRUE,
           expected = FALSE, sresid = FALSE)
```

```
   Cell Contents
|-------------------------|
|                   Count |
|-------------------------|

Total Observations in Table:  359

                     | mydata$Purchased
mydata$Purchasedpred |        No  |       Yes  | Row Total |
---------------------|-----------|-----------|-----------|
                  No |       212 |        39 |       251 |
---------------------|-----------|-----------|-----------|
                 Yes |        19 |        89 |       108 |
---------------------|-----------|-----------|-----------|
        Column Total |       231 |       128 |       359 |
---------------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  147.1726     d.f. =  1     p =  7.194579e-34

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  144.2723     d.f. =  1     p =  3.09783e-33


       Minimum expected frequency: 38.50696
```

From this confusion matrix, we can see we made 19 type 1 errors or false positives and 39 type 2 errors or false negatives. Therefore our model is worse at predicting when an actual purchase has been made.

The overall accuracy of our model is 83.84%.

Odds Ratios

Hide

```
exp(model4$coefficients)
```

```
    (Intercept)             Age EstimatedSalary 
      0.3202232       1.2517688       1.0000344 
```

Based on the results shown above we can say that on average the odds of a person making a purchase is 0.32, and these odds are 1.25 times higher based on a unit increase in age whereas it increases very little based on unit increase in estimated salary by about 1.0000344 times the intercept value.

Hide

```
pander(exp(confint(model4)),
       caption = "95% confidence interval of odds ratio")
```

```
Waiting for profiling to be done...
```

```


--------------------------------------
                  2.5 %    97.5 %
-------------------- -------- --------
   **(Intercept)**     0.2229   0.4472

       **Age**         1.194    1.324

 **EstimatedSalary**     1        1
--------------------------------------

Table: 95% confidence interval of odds ratio
```

From the results shown above, we see that the confidence intervals are small for both our predictor variables indicating that the estimates are reliable with little random error.

Examing Assumptions

First we test the multi-collinearity assumption of our model.

Hide

```
1/vif(model4)
```

```
            Age EstimatedSalary
       0.757185        0.757185
```

We can see that the tolerance values are larger that 0.2 and so there is no indication of collinearity.

Secondly, we test if the errors are indeed independent.

Hide

```
durbinWatsonTest(model4)
```

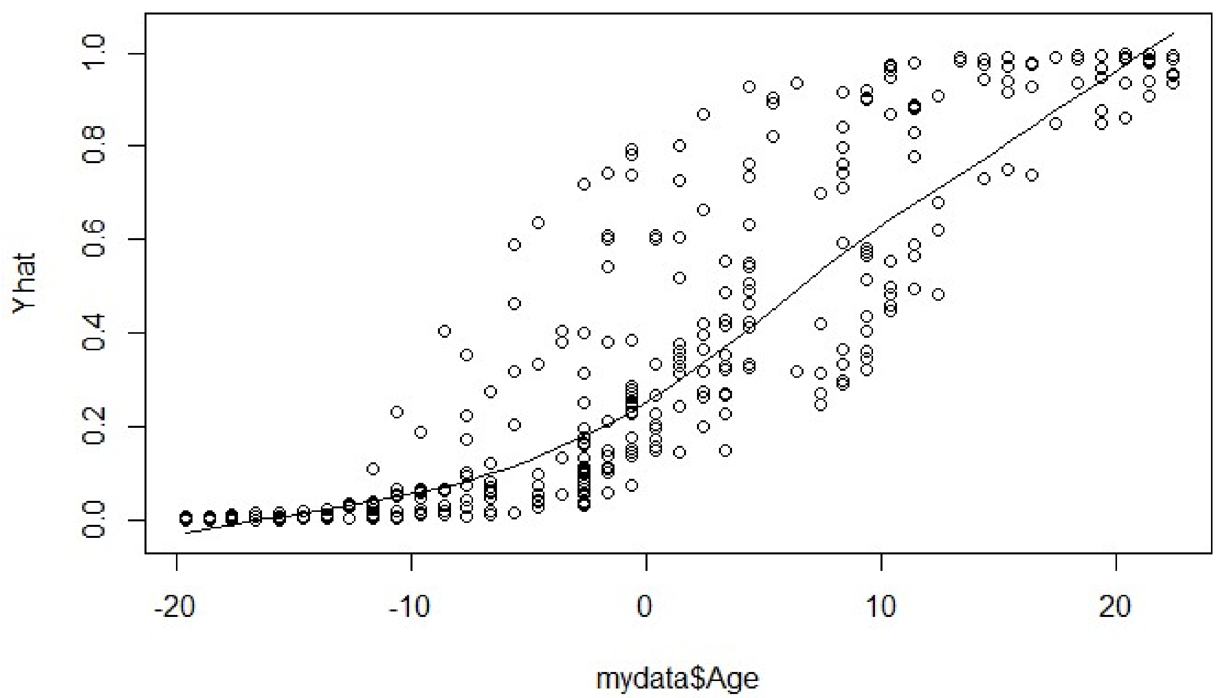
```
lag Autocorrelation D-W Statistic p-value
  1      0.04995433      1.895589   0.272
Alternative hypothesis: rho != 0
```

Based on the findings as shown above, we fail to reject the null hypothesis(lag=1, Autocorrelation=0.05,D-W Statistic 1.9, p-value =0.294). Therefore the errors are indeed independent.

Lastly, we wanted to check if there was some kind of linearity between the log odds and our independent variables.

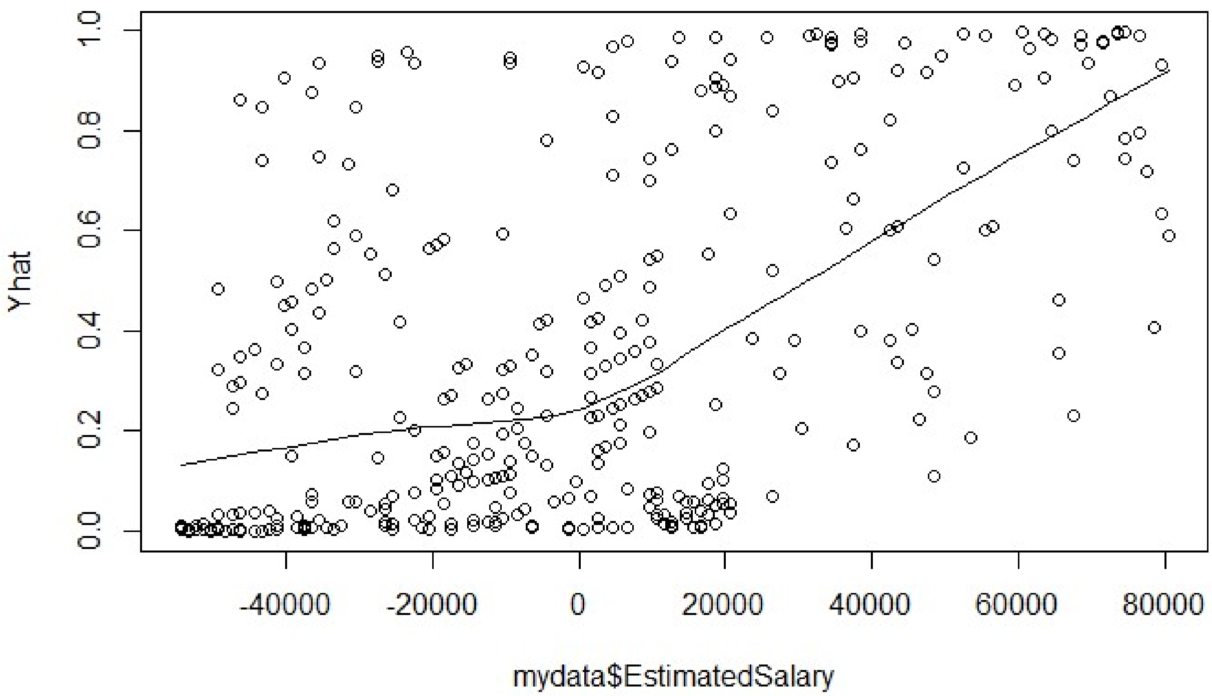
Hide

```
Yhat    <- fitted(model4)
scatter.smooth(mydata$Age,Yhat)
```



Hide

```
scatter.smooth(mydata$EstimatedSalary,Yhat)
```



Therefore, we see that the log odds have a somewhat linear trend to the predictor variables. Thereby verifying our assumptions.

Pseudo R squared Value

Hide

```
logisticPseudoR2s <- function(LogModel) {

        dev <- LogModel$deviance

        nullDev <- LogModel$null.deviance

        modelN <-  length(LogModel$fitted.values)

        R.l <-  1 -  dev / nullDev

        R.cs <- 1- exp ( -(nullDev - dev) / modelN)

        R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))

        cat("Pseudo R^2 for logistic regression\n")

        cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")

        cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")

        cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")

}

logisticPseudoR2s(model4)
```

```
Pseudo R^2 for logistic regression
Hosmer and Lemeshow R^2   0.457
Cox and Snell R^2         0.449
Nagelkerke R^2            0.617
```

Here we present the pseudo R squared values for our model. The pseudo R square value is between 0 and 1 and gives an indication for how well our model can explain the data or how good our model fit is. It’s a useful metric to compare different models.

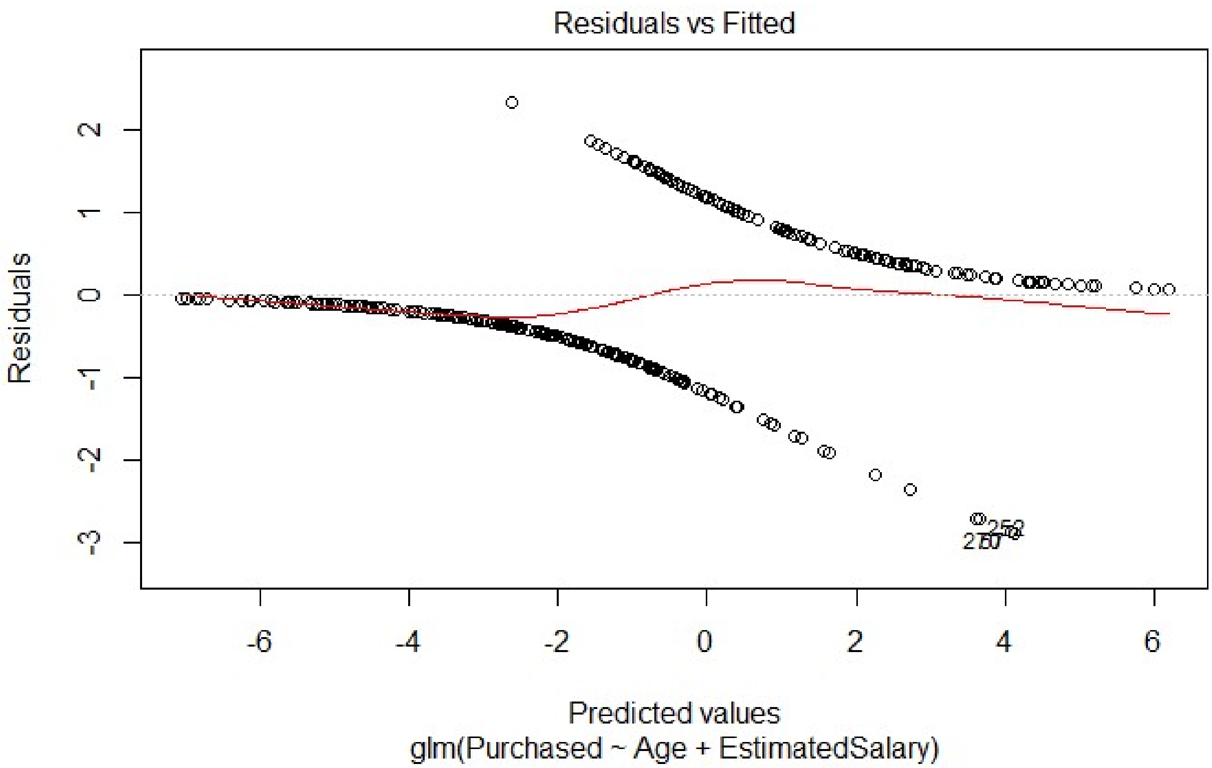
Impact Analysis of Individual Cases.

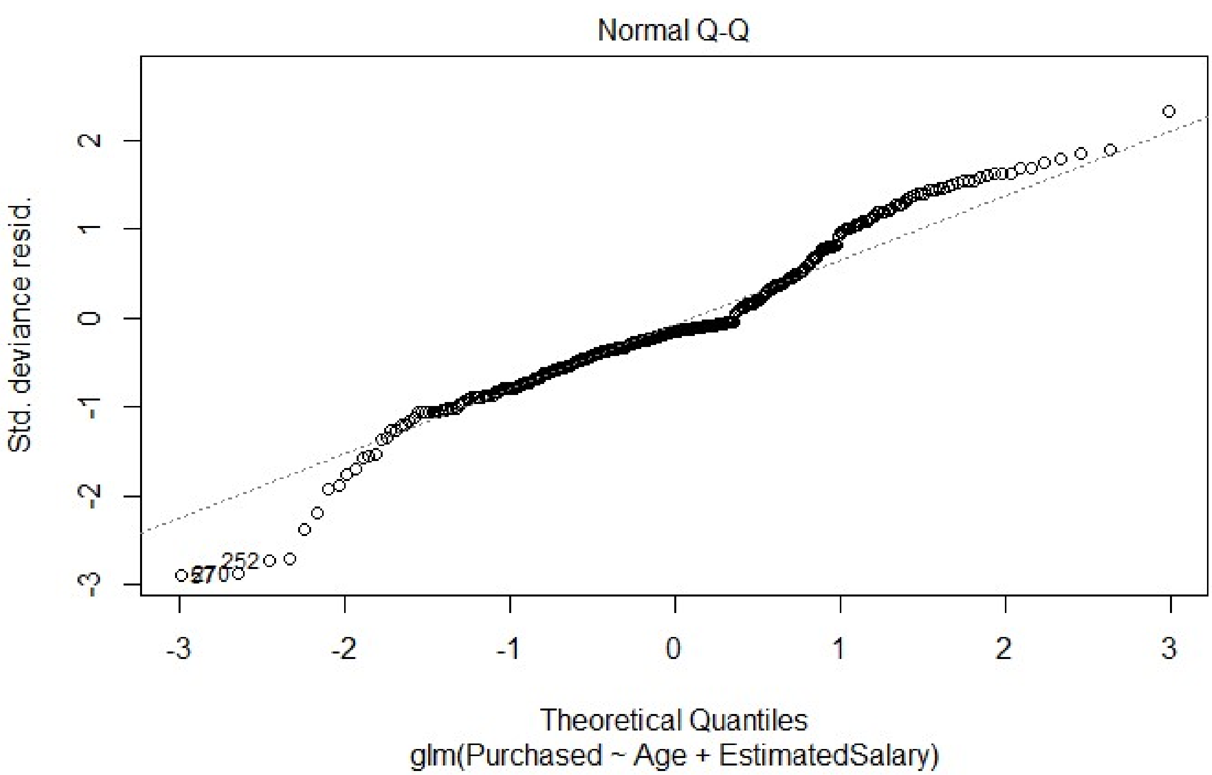
Hide

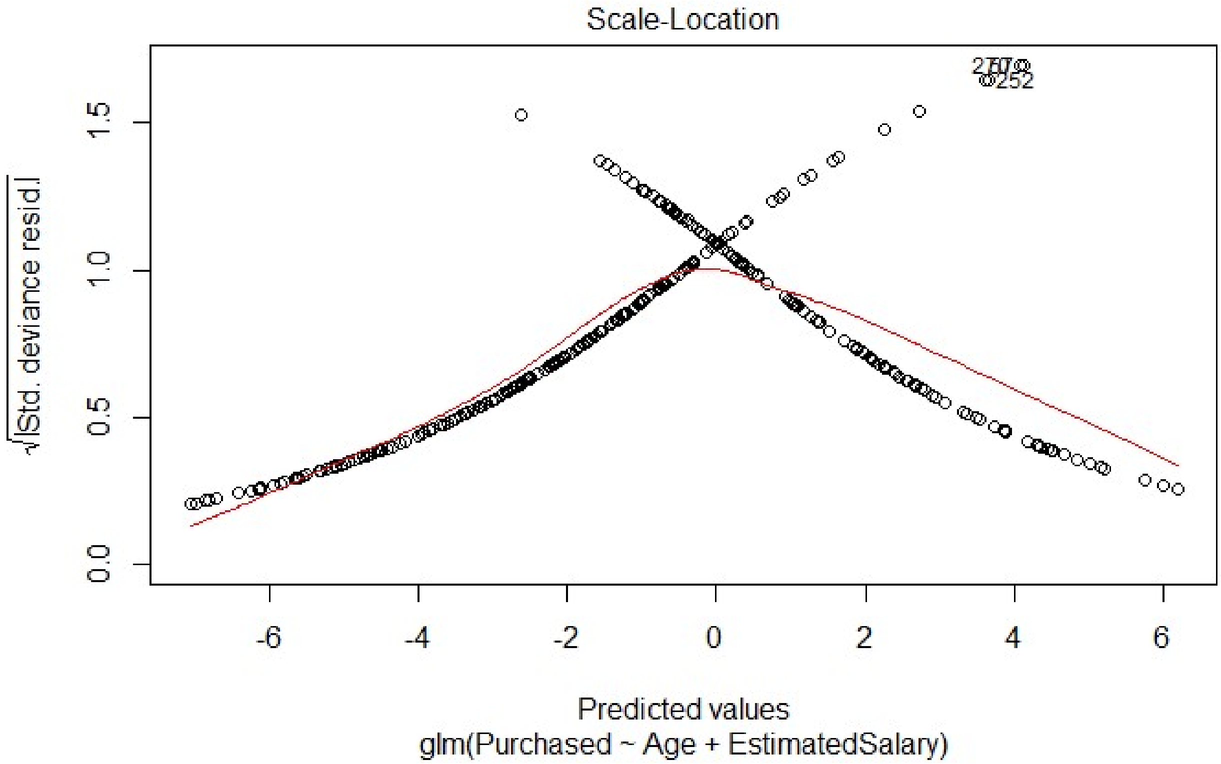
```
mydata$leverage<-hatvalues(model4)
mydata$stud.res<-rstudent(model4)
mydata$dfbeta<-dfbeta(model4)
troublingdata<-subset(mydata, (leverage > 3*.0083) | (abs(stud.res) > 2),
       select = c("leverage", "stud.res", "dfbeta"))
head(troublingdata)
```

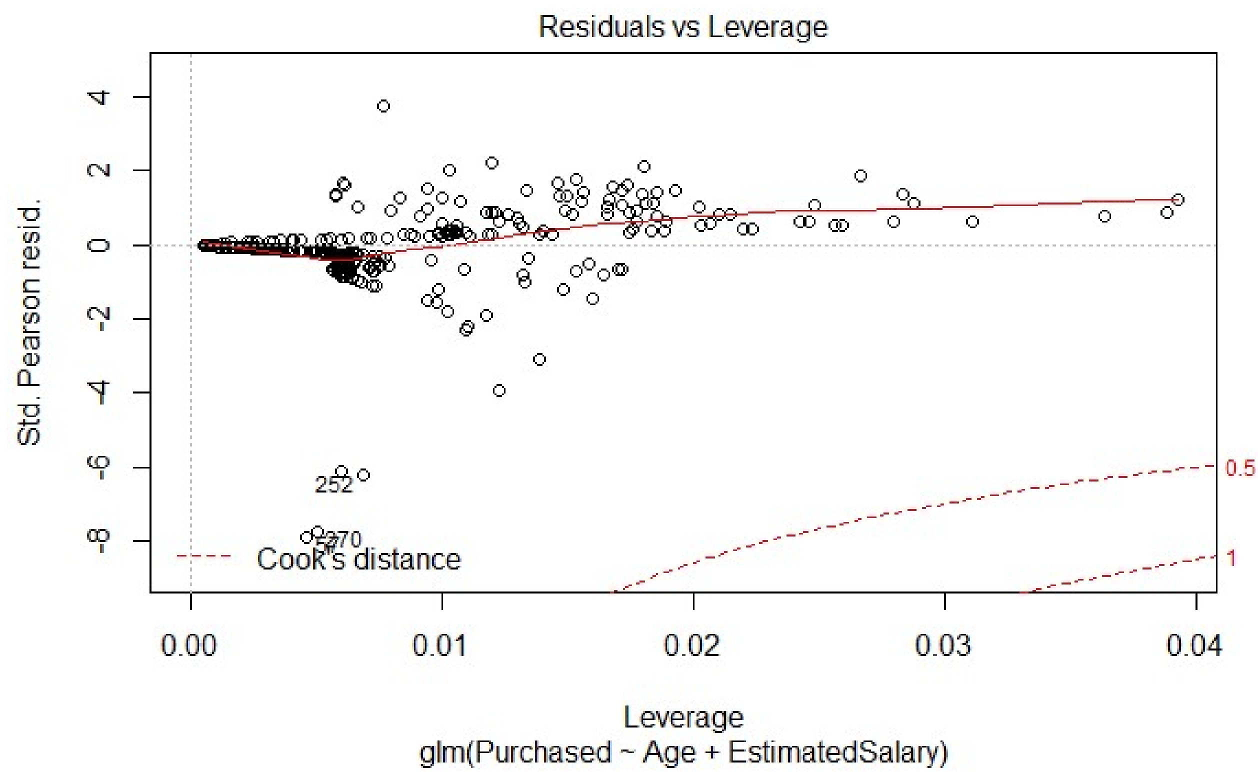
Hide

```
plot(model4)
```









We have also looked into the impact of individual cases as well by looking at the studentized residuals, the leverage values and the dfbeta values. We have calculated the leverage as 1+(# of predictors)/# of observations. We check for values greater than 3 times the average leverage and for values with the absolute studentized residuals greater than 2. In addition we also looked for points outside the dashed red line in the plot of the residuals vs leverage to see if there were points with a large cook's distance as well. We have removed some of these points and noticed an improvement in the pseudo R squared values as well. Infact for this analysis we have used the cleaned dataset after removing values with high leverage based on the previous one as can be seen in the name of the file used too.

Report section for a scientific publication

Our goal in this section was to study the affect of our independent variables, namely age, gender and estimated salary on a binary categorical dependent variable("purchased") which indicated if a purchase was made by a user online or not. Therefore, we compared the fit of our null model with the fit of a model which included either age, gender or estimated salary. Based on our analysis, we found no significant main effect ($\chi2(Df=1,Resid.Df=357)=0.58$, $p.= 0.449$) for gender, whereas there was a significant main effect on the fit of our model for (M:37.6,SD:10.6)age ($\chi2(Df=1, Resid.Df=357)=161.9$, $p.< 0.01$) and (M:69462,SD:34277.3)estimated salary ($\chi2(Df=1,Resid.Df=357)=46.97$, $p.< 0.01$) respectively. Thus, for our significant predictors we report that the odd ratios for the estimates of age and estimated salary are 1.2517688(95% CI, 1.194 to 1.324) and 1.0000344(95% CI, 1 to 1) with age having a stronger impact on average on the odds of a purchase made by a user. We also made sure to verify our assumptions about collinearity, independence of errors and linearity of our significant predictors with the predicted log odds of our model and we believe that they hold as the tolerance values for our predictors were greater that 0.2, the durbin-watson test (D-W Statistic=1.9, p-value =0.294)allowed us to safely fail to reject the null hypothesis about independent errors and based on our plots we can see the linearity between the predictors and the predicted log odds. Finally, we calculated the pseudo R squared values of our model as being (Hosmer and Lemeshow R^2) 0.457,(Cox and Snell R^2) 0.449 and (Nagelkerke R^2) 0.617 which indicate how well our data is explained through the model.