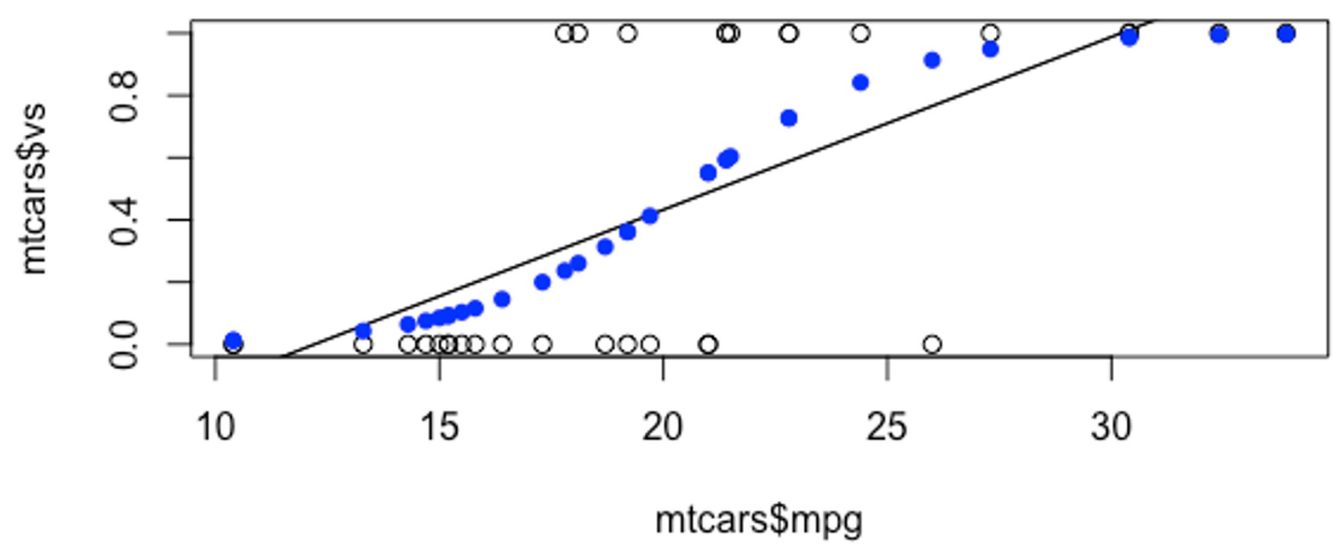
logistic regression for binary classification

```
head(cars)
               mpg  cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4     21.0    6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag 21.0    6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710    22.8    4  108  93 3.85 2.320 18.61  1  1    4    1

carglm=glm(vs ~ mpg, family=binomial, data=mtcars)

coef(carlm)
(Intercept)         mpg
-0.67816541  0.05553164


coef(carglm)
(Intercept)         mpg
 -8.8330726   0.4304135
```




```
points(mtcars$mpg,logistic(
mtcars$mpg * coef(carglm)[2]+ coef(carglm)[1]),col="blue",pch=16)
```

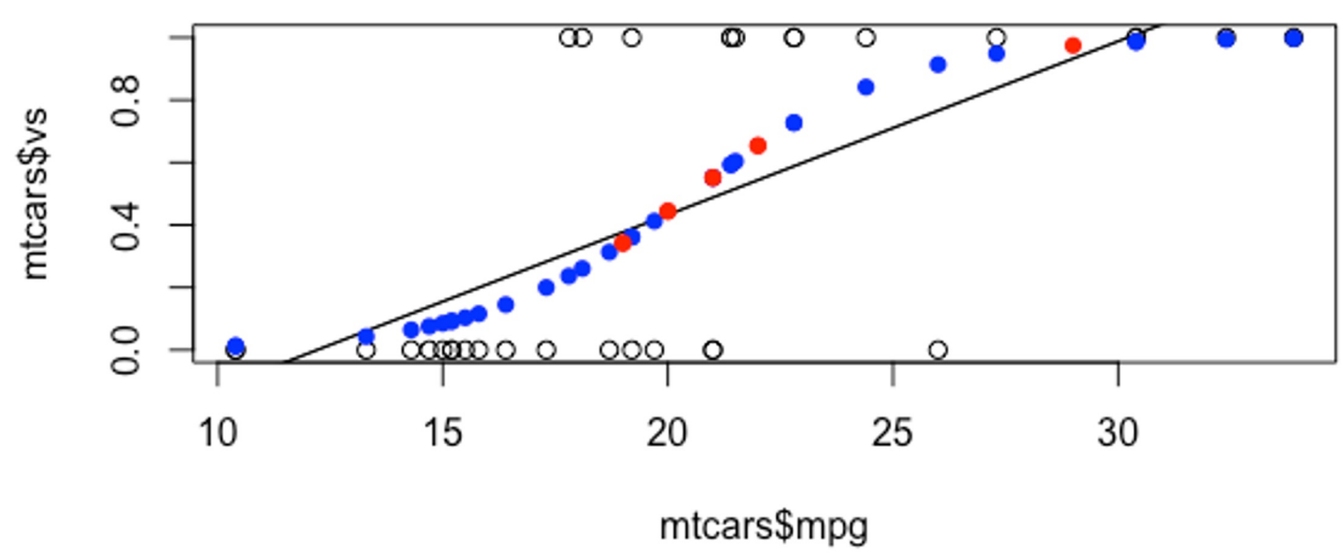
logistic regression for binary classification

Predicting using `carglm` model:

```
data2pred <- data.frame(mpg = c(19:22,29))

data2pred=mutate(data2pred,probvs=
  predict(carglm,type="response",newdata=data2pred))
```

```
data2pred
 mpg     probvs
1   19 0.3418152
2   20 0.4440349
3   21 0.5512225
4   22 0.6538543
5   29 0.9746406
```



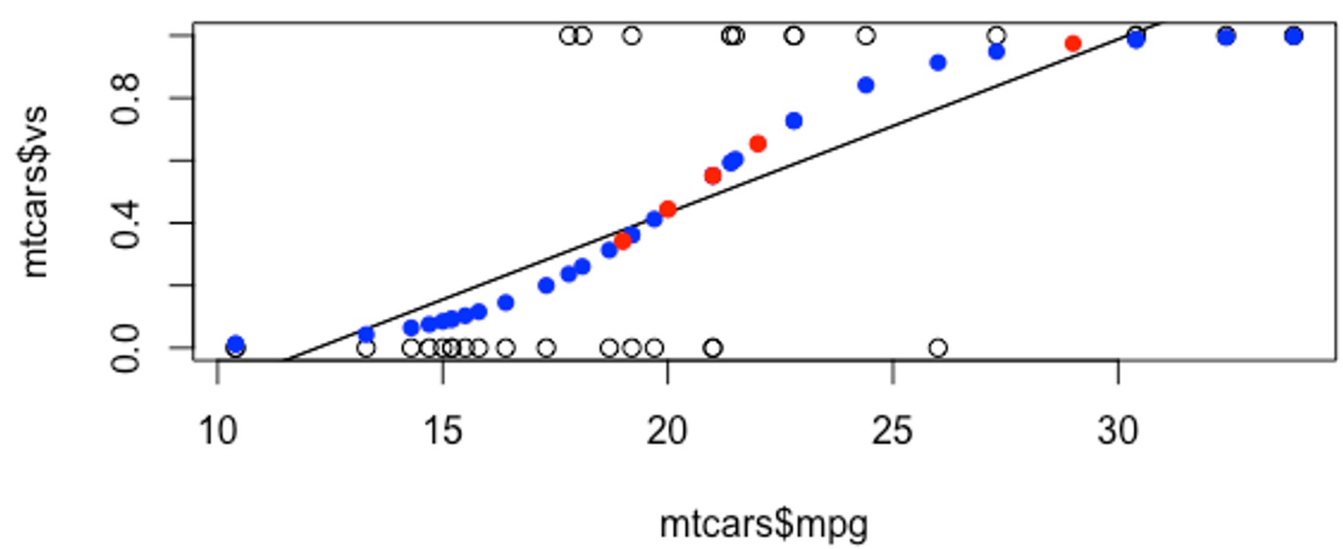

```
points(data2pred$mpg,data2pred$probvs,col="red",pch=16)
```

logistic regression for binary classification

Predicting using `carglm` model:

What is probability that a car has a v-shaped engine (vs==1) given that it gets 29 mpg?

```
data2pred
 mpg     probvs
1   19 0.3418152
2   20 0.4440349
3   21 0.5512225
4   22 0.6538543
5   29 0.9746406
```




```
points(data2pred$mpg,data2pred$probvs,col="red",pch=16)
```

Exercise (background)

Pima Indian diabetes data set (UCIML; Kaggle)

A data frame with patient data from NIDDK.

```
pregnant:    Number of (prior) pregnancies
glucose:     Plasma glucose concentration at 2 hours
pressure:    Diastolic blood pressure (mm Hg)
triceps:     Triceps skin fold thickness (mm)
insulin:     2 hour serum insulin (muU/ml)
mass:        Body mass index, or BMI
pedigree:    Indication of family history of diabetes
age:         Patient age in years
diabetes:    Logical variable indicating a diabetes diagnosis
```

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

Logistic regression exercise

Load data set `diabetes.rds` from the Schedule page.

First, plot the variable glucose (fasting glucose level) vs. diabetes. Then use `glm` to create a model, `dbglm`, of the probability of diabetes as a function of glucose.

create a new data frame to predict using the following code:

```
newvals <- data.frame(glucose = seq(100,160,5))
```

Use the `predict` function with `type="response"` to determine the probabilities (to 2 significant digits) that someone with these glucose levels has diabetes. Send the probability of having diabetes with a fasting glucose level of 110 to me in Piazza.