HW2

Jingshi

2/1/2017

Question 3

a)

According to the question, the equation is :

$$Predicted\ Starting\ Salary = 50 + 20 \times GPA + 0.07 \times IQ + 35 \times Gender + 0.01 \times GPA \times IQ - 10 \times GPA \times Gender$$

(i) and (ii) are incorrect because there is an negative interaction between Gender and GPA. So the predicted value of starting salary also depends on the value of GPA. Given a fixed value of IQ and GPA, if GPA is high enough (higher than 3.5), males earn more on average; otherwise, if GPA is low (lower than 3.5), females earn more on average. Therefore, (iii) is correct and (iv) is incorrect.

b)

By plugging in to the above function:

$$Predicted\ Starting\ Salary = 50 + 20 \times 4.0 + 0.07 \times 110 + 35 \times 1 + 0.01 \times 4.0 \times 110 - 10 \times 4.0 \times 1 = 137.1$$

So the predicted starting salary is 137.1 (in thousands of dolars) or $137,100.

c)

It is false because we can not comment on the evidence of an interaction effect only based on the magnitude of the coefficient for the interaction term. We need to compute its p-value in order to make a decition.

Question 8

a)

```
library(ISLR)
#help(Auto)
auto.lm=lm(mpg~horsepower, data =Auto)
summary(auto.lm)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i/ii/iii)

There is a strong negative linear relationship between horsepower (the predictor) and mpg (the response).

iv)

```
# compute the predicted mpg and the associated 95 % confidence interval
predict(auto.lm,data.frame(horsepower =98), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

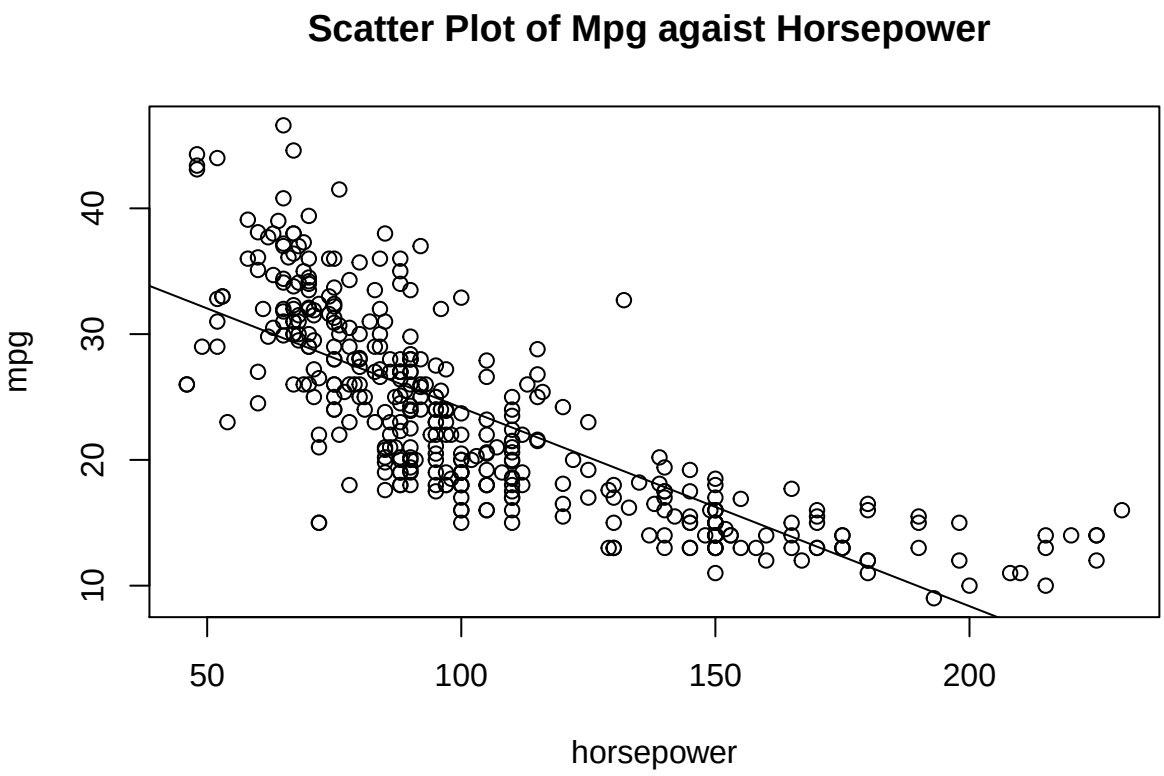
```
# compute the predicted mpg and the associated 95 % prediction interval
predict(auto.lm,data.frame(horsepower =98), interval = "prediction")
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

According to the output, the predicted mpg is 24.46708. The confidence interval is (23.97308,24.96108). The prediction interval is (14.8094, 34.12476) which is wider than the confidence interval.

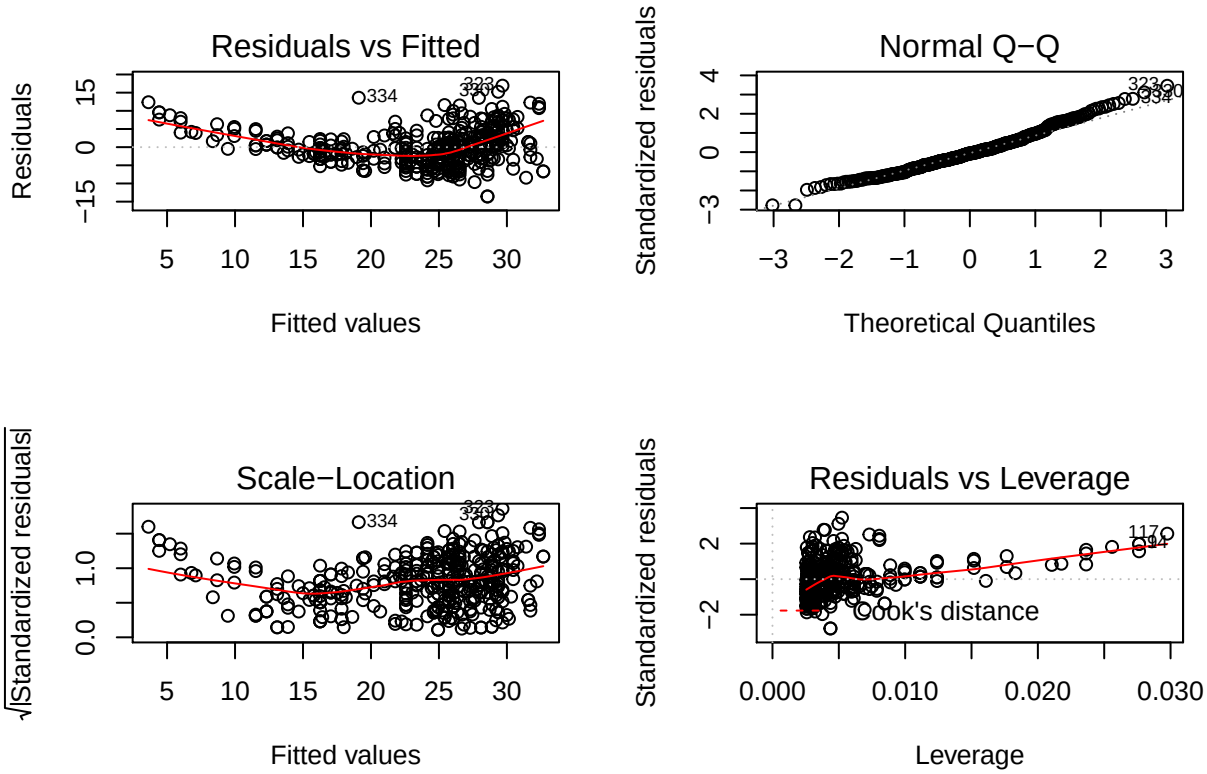
b)

```
plot(Auto$horsepower,Auto$mpg, main = "Scatter Plot of Mpg agaist Horsepower ",xlab="horsepower",
     ylab="mpg")
abline(auto.lm)
```



c)

```
par(mfrow=c(2,2))
plot(auto.lm)
```



There appears to be a curved trend. In the residuals vs fitted plots, the residuals are not homoscedastic and do not spread evenly. The standarized residuals are large at lower and higher fitted values and small at middle fitted values. The normal Q-Q plot also shows a curved trend. The scale-location plot also shows a curved trend. Additionally, the plots in residuals vs leverage plot are not normally distributed. These evidences suggest that the relationship between mpg and horsepower is not linear.

As you can see on the residuals vs leverage plot, both the points with high leverage and low leverage appear to have higher standardized residuals. They could possibly be outliers. There are some outliers with high leverage that are influencial such as observation 117 and 94.

Question 10

a)

```
seats.lm<-lm(Sales ~ Price + Urban + US, data=Carseats)
help(Carseats)
summary(seats.lm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b)

According to the output:

Intercept: 13.043469 Interpretation: When price is 0, the region of store is non-urban and non-US, the expected unit sales is about 13.043 (in thousands).

Coefficient for price: -0.054459 Interpretation: When price increases by 1 unit, sales is expected to decrease by 0.054 (in thousands), holding "urban" and "us" constant.

Coefficient for Urban: -0.021916 Interpretation: Unit sales for urban stores is expected to be lower than it is for rural by about 0.022 (in thousands), holding "price" and "us" constant.

Coefficient for US: 1.200573 Interpretation: Unit sales for US stores is expected to be higher than it is for non-US stores by about 1.201 (in thousands), holding "price" and "urban" constant.

c)

$$Predicted\ Sales = 13.043 - 0.054 \times Price - 0.022 \times Urban + 1.201 \times US$$

Urban = 1 if the store is in an urban location; otherwise, urban = 0. US = 1 if the store is in the US; otherwise, US = 0.

d)

The null hypothesis can be rejected for "Price" and "US" because their p-values (< 2e-16 and 4.86e-06 respectively)are much smaller than alpha = 0.05. The null hypothesis cannot be rejected for "Urban" because the p-value for "Urban" (0.936) is much higher than alpha = 0.05. Thus, "Price" and "US" are significant predictors in this model.

e)

According to the previous question, "Price" and "US" are significant predictors of sales.

```
# fit a new model with "Price" and "US" as predictors
seats2.lm<-lm(Sales ~ Price + US, data=Carseats)
summary(seats2.lm)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f)

The model in (e) fits slightly better than the modle in (a) because it has a slightly higher adjusted R-squared (0.2354) and all the predictors are significant (because their p-values ($<$ 2e-16 for "Price" and 4.71e-06 for "US") are much lower than alpha = 0.05). Both models are significant because their p-values (both are $<$ 2.2e-16) from F test are much lower than alpha = 0.05.

g)

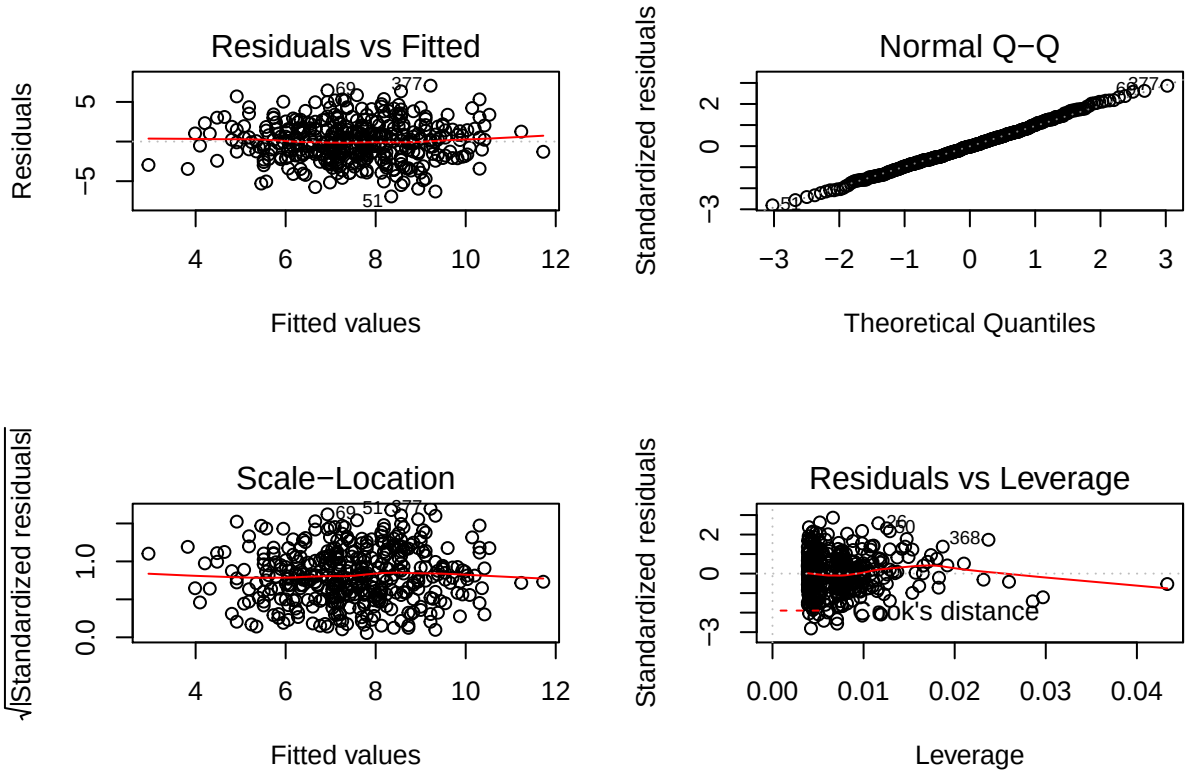
```r
# obtain 95% confidence intervals for the coefficients
confint(seats2.lm)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

According to the output, the 95% confidence interval for "Price" is (-0.06475984, -0.04419543). The 95% confidence interval for "US" is (0.69151957, 1.70776632).

h)

```r
# produce diagnostic plots of the least squares regression fit
# of model from question e.
par(mfrow=c(2,2))
plot(seats2.lm)
```



As the output shows, the residuals spread evenly in the residuals vs fitted plot. There are some outliers such as obervation 51, 69 and 377. The normal Q-Q plot shows a straight line. There appears to be a few points with higher leverage. Only a few points have higher leverage and higher standardized residuals such as observation 368, 50 and 26. These points are influential.

Question 15

a)

```
library(MASS)
help("Boston")
simple.zn=lm(crim~zn,data=Boston)
simple.indus=lm(crim~indus,data=Boston)
simple.chas=lm(crim~chas,data=Boston)
simple.nox=lm(crim~nox,data=Boston)
simple.rm=lm(crim~rm,data=Boston)
simple.age=lm(crim~age,data=Boston)
simple.dis=lm(crim~dis,data=Boston)
simple.rad=lm(crim~rad,data=Boston)
simple.tax=lm(crim~tax,data=Boston)
simple.ptratio=lm(crim~ptratio,data=Boston)
simple.black=lm(crim~black,data=Boston)
simple.lstat=lm(crim~lstat,data=Boston)
simple.medv=lm(crim~medv,data=Boston)

summary(simple.zn)
summary(simple.indus)
summary(simple.chas)
```

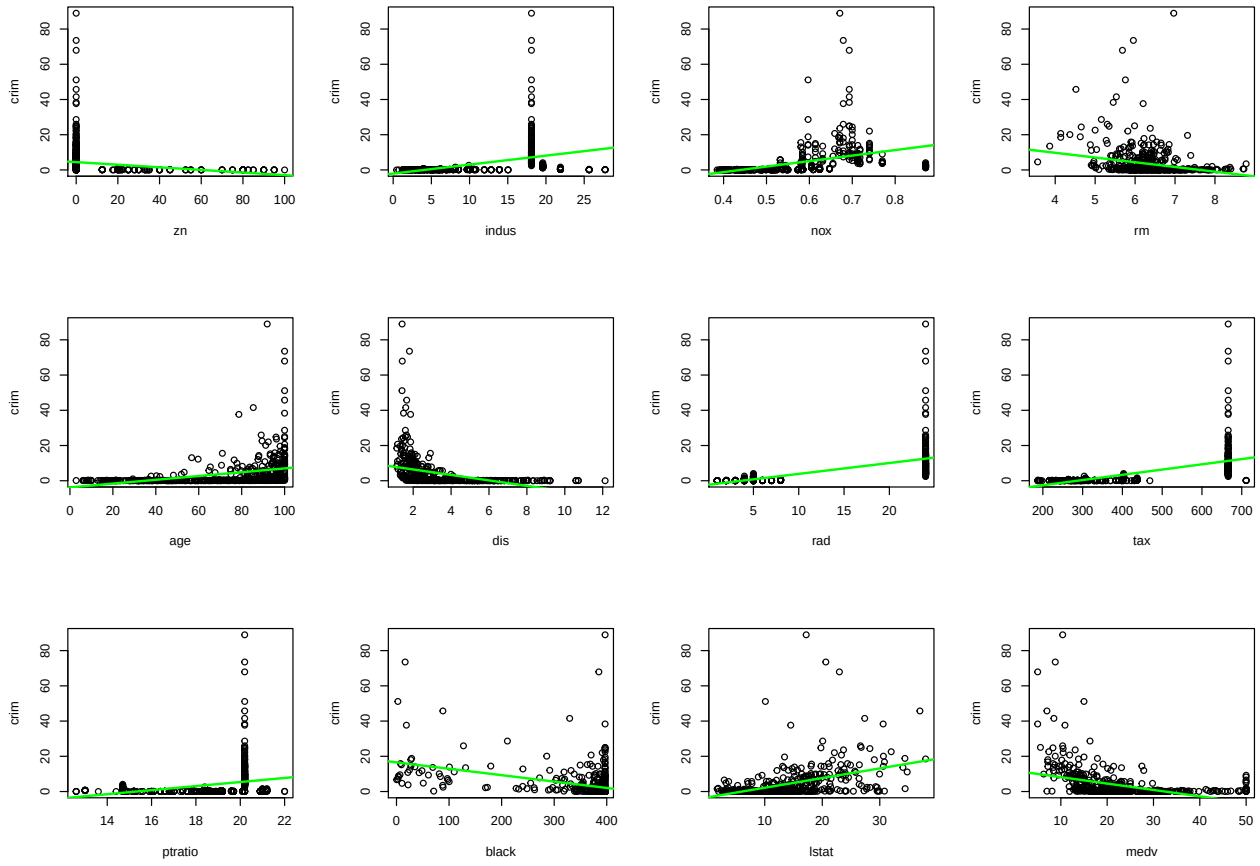
```
summary(simple.nox)
summary(simple.rm)
summary(simple.age)
summary(simple.dis)
summary(simple.rad)
summary(simple.tax)
summary(simple.ptratio)
summary(simple.black)
summary(simple.lstat)
summary(simple.medv)
```

Predictor	P-value	R-squared
zn	5.51×10^{-6}	0.03828
indus	$< 2 \times 10^{-16}$	0.1637
chas	0.209	0.001146
nox	$< 2 \times 10^{-16}$	0.1756
rm	6.35×10^{-7}	0.04618
age	2.85×10^{-16}	0.1227
dis	$< 2 \times 10^{-16}$	0.1425
rad	$< 2 \times 10^{-16}$	0.39
tax	$< 2 \times 10^{-16}$	0.3383
ptratio	2.94×10^{-11}	0.08225
black	$< 2 \times 10^{-16}$	0.1466
lstat	$< 2 \times 10^{-16}$	0.206
medv	$< 2 \times 10^{-16}$	0.1491

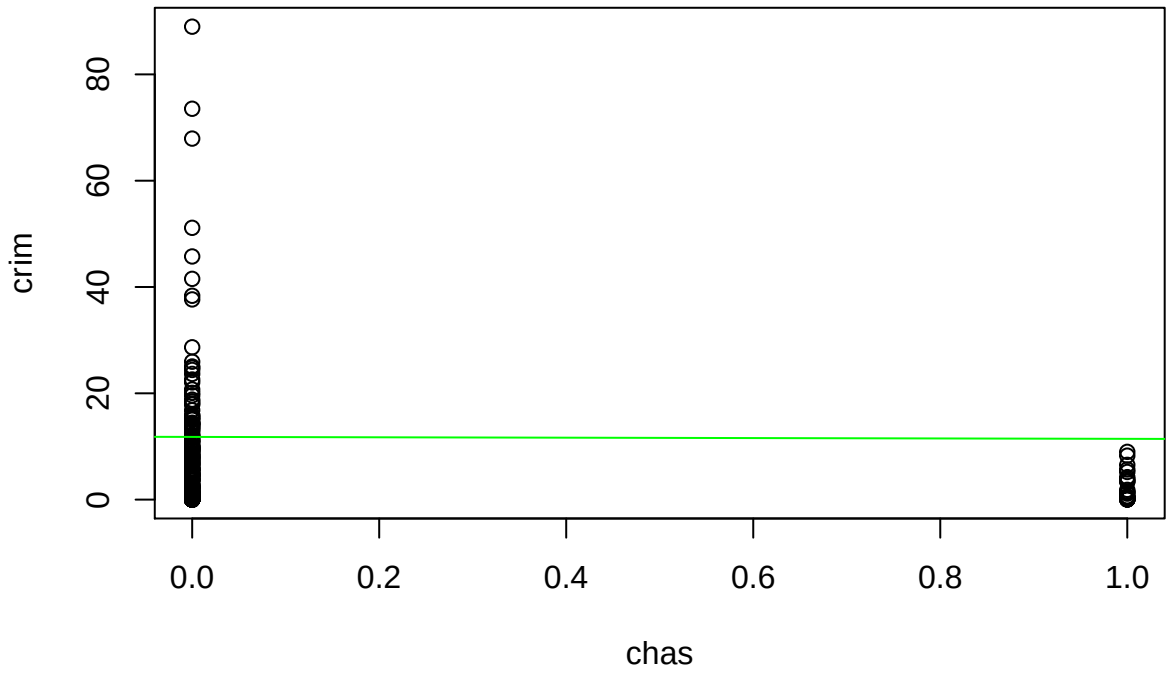
According to the results, most of the models, except for the model with chas as predictor, have statistically significant association between the predictor and the response.

```
# Plots of the models that have statistically significant
# association between the predictor and the response.
par(mfrow=c(3,4))
plot(crim~zn, data = Boston)
abline(simple.zn, col="green",lwd=2)

plot(crim~indus, data = Boston)
abline(simple.indus, col="green",lwd=2)

plot(crim~nox, data = Boston)
abline(simple.nox, col="green",lwd=2)

plot(crim~rm, data = Boston)
abline(simple.rm, col="green",lwd=2)

plot(crim~age, data = Boston)
abline(simple.age, col="green",lwd=2)

plot(crim~dis, data = Boston)
abline(simple.dis, col="green",lwd=2)

plot(crim~rad, data = Boston)
abline(simple.rad, col="green",lwd=2)
```

```r
plot(crim~tax, data = Boston)
abline(simple.tax, col="green",lwd=2)

plot(crim~ptratio, data = Boston)
abline(simple.ptratio, col="green",lwd=2)

plot(crim~black, data = Boston)
abline(simple.black, col="green",lwd=2)

plot(crim~lstat, data = Boston)
abline(simple.lstat, col="green",lwd=2)

plot(crim~medv, data = Boston)
abline(simple.medv, col="green",lwd=2)
```




```r
# Plots of the model (with chas as predictor)that
# dosn't have statistically significant association
# between the predictor and the response.
plot(crim~chas, data = Boston)
abline(simple.medv, col="green")
```



b)

```
# fit a multiple regression model to predict
# crim using all of the predictors
model.all=lm(crim~.,data=Boston)
summary(model.all)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

The result shows less predictors that are statistically significant for predicting crim than the previous question. This may due to that some predictor variables are correlated. The predictors that the null hypothesis can be rejected are zn, dis, rad, black and medv.