# Problem 1

## i

We will use BIC to select the best model.

```
# best subset
s = leaps::regsubsets(medv ~ ., data = Boston, method = "exhaustive", nvmax = ncol(Boston) -
    1)
s  #forward
f = Rcmdr::stepwise(lm(medv ~ ., data = Boston), direction = "forward", criterion = "BIC")
# backward
b = Rcmdr::stepwise(lm(medv ~ ., data = Boston), direction = "backward", criterion = "BIC")
# best model for backward selection and best subset selection
sb = glm(medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black +
    lstat, data = Boston)
# best moel for forward selection
f = glm(medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn, data = Boston)
```

We can see the selected models in the above code. Best subset selection and backward selection choose medv crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat while forward selection chooses medv  lstat + rm + ptratio + dis + nox + chas + black + zn.

## ii

```
# 5 fold cv MSE for best subset/backwards regression model
set.seed(12345)
boot::cv.glm(Boston, sb, K = 5)$delta[1]

## [1] 23.58

# 5 fold cv MSE for forwards selection model
set.seed(12345)
boot::cv.glm(Boston, f, K = 5)$delta[1]

## [1] 24.38
```

Using this seed the best subset/backwards selection model medv crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat performs better.

# Problem 3

## i

```
enzyme = read.table("Enzyme.txt", head = T)
enzyme.prime = 1/enzyme
alpha = coef(lm(Y ~ x, data = enzyme.prime))
beta = as.numeric(c(1/alpha[1], alpha[2]/alpha[1]))
beta

## [1] 29.62 13.45
```

The initival values of $\beta$ are 29.622, 13.4488

## ii

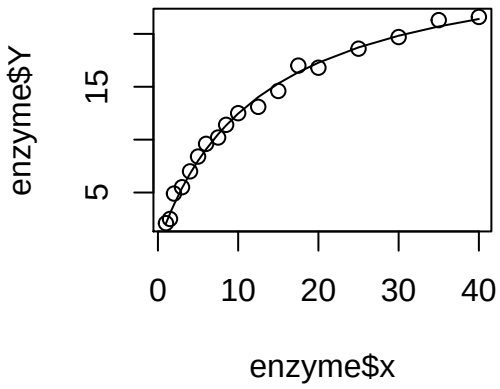
```
m0 = nls(formula = Y ~ b0 * x/(b1 + x), data = enzyme, start = c(b0 = beta[1],
    b1 = beta[2]))
coef(m0)

##    b0    b1
## 28.14 12.57
```

The OLS estimtates of $\beta$ are 28.137, 12.5745

## iii

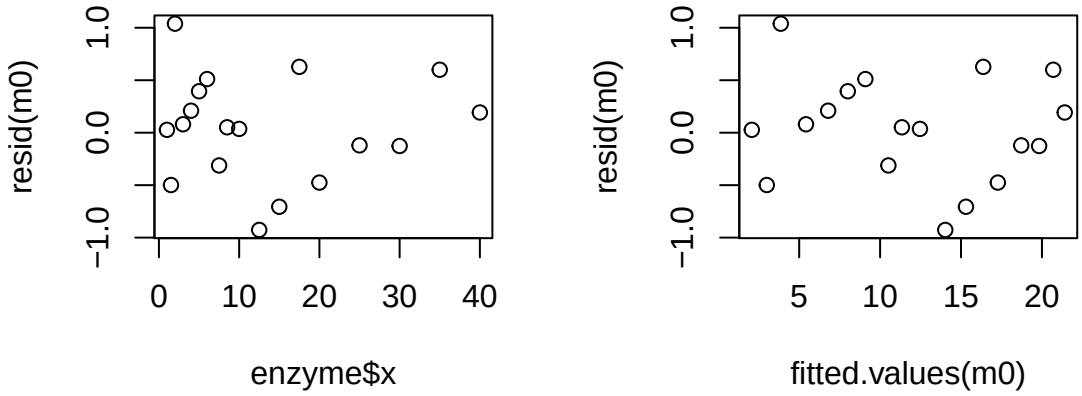
```
plot(enzyme$x, enzyme$Y)
lines(enzyme$x, fitted.values(m0))
```



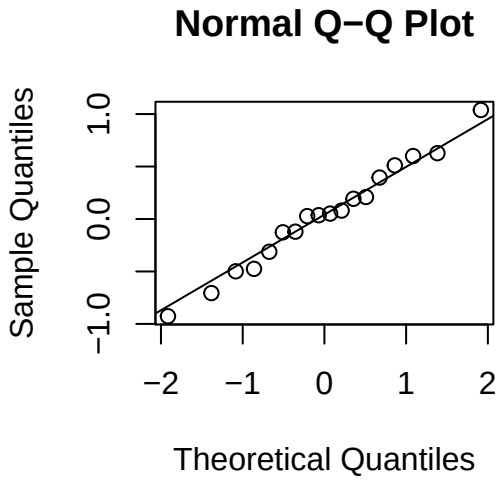

The fit is adequate.

**iv**

```
plot(enzyme$x, resid(m0))
plot(fitted.values(m0), resid(m0))
```




```
qqnorm(resid(m0))
qqline(resid(m0))
```




These plots seem to show that our model is a good fit. The first two plots show no obvious patterns and small residuals. The QQ plot seems to show that our residuals are normally distributed.

### v

Although there is no hypothesis test with null 'this is a bad model' or 'this is not a bad model', there are many ways to test if the resdiauls are normally distributed which is a more formal way to check if that particular assumption holds than a qq plot. According to (NM Razali, YB Wah, 2012) the ShapiroWilk test is the highest power test of normality.

```
shapiro.test(resid(m0))$p

## [1] 0.9943
```

We see that the p-value is 0.9943 and thus fail to reject that the residuals are normally distributed.

### vi

If we collect more datapoints that are closer together (or indeed replications) then we will have lower standard error and thus lower confidence and prediction intervals locally around the points we have sampled. If we collect data points that are farther apart (and no replications) then we will not have as small of a local standard error, confidence intervals, or prediction intervals, but at each concentration value, but we know more about a larger range of concentration values. There will always be a tradeoff between local precision and trying to get a more global view of the phenomenon.

### vii

#### 1

```
est = summary(m0)$coefficients[1, 1]
sd = summary(m0)$coefficients[1, 2]
ci = c(est - 1.96 * sd, est + 1.96 * sd)
ci

## [1] 26.71 29.56
```

The 95% confidence interval is 26.7102, 29.5639.

#### 2

```
est = summary(m0)$coefficients[1, 1]
sd = summary(m0)$coefficients[1, 2]
pval = 1 - pchisq(((est - 20)/sd)^2, 1)
pval

## [1] 0
```

The p-value is 0

## Problem 4

### a

Based on the analysists reasons we can not conclude that the model is highly effective for producing prediction ingervals for suggested retail price. We can only say that dealer cost as significantly correlated

with suggested retail price. There are many reasons that this is in fact a bad model which I will discuss in the following sections.

### b

In the plot of the studentized residuals vs the predictor we see that there is a heteroskedacity problem and our residuals vairance increases as the dealer cost increases. We also can see a clear increasing trend in the square root of the residuals vs the dealer cost. showing that we will make larger errors as the price of the cars increase. The qqplot is also very poor in the two extremes showing that there are many outliers that are vasly underpriced or overpriced in the model.

### c

The new model seems to perform better than the old one. The studentized residuals show less outliers and the plot vs the predictor shows less of a pattern. The square root of the residuals vs the predictor now shows no trend. There are still a few outliers and these can be seen in the normal QQ plot, indicating that this model may still not be as good as we may like.

### d

If the dealer cost increases by 1% then the suggested retail price will increase by $\beta_1\%$

### e

We still see that there are a few studentized residuals with large negative values when the dealer cost is low. This means that our model may underprice some low cost cars. The interpretation of the model is also much less clear than the first model. If this model will not be used by a computer to set the price in advance, but rather is going to be used to give the sales people quick and simple insights into reasonable prices of the cars then it may not be very good.

## Problem 5

### a

This model contains two insignificant variables (hybrid and wheelbase) which should be removed (remove one at a time, maybe removing one will make the other significant) and is thus not a 'valid' model. We should always aim for more parsimonious modes if possible. We also see from the diagnoistic plots that the model does not fit very well. In particular we see that there are a few very large residuals for high priced cars indicating that we are going to underprice expensive cars.

### b

From the curve in the plot we see that we will tend to underprice very cheap and very expensive cars. Cars with average prices will tend to be priced better.

### c

According to wikipedia there are three main heuristics to use Cook's distance to classify if points have high influence or not, $D_i > 1$, $D_i > 4/n$ and $D_i > F_{p,n-p,1-\alpha}$. We have $4/n$ =0.0171 and $F_{p,n-p,1-\alpha}$ = 1.9795. Using a cutoff of 1, or $D_i > F_{p,n-p,1-\alpha}$ we find no highly influential points. Using a cutoff of $4/n$ we find theat most points are highly influental. We will conclude that there are no highly influential points.

## d

This model contains two non significant predictors (thighwaympg and twheelbase) which should be removed (remove one at a time, maybe removing one will make the other significant) and is thus not a 'valid' model. We should always aim for more parsimonious modes if possible. THe diagnoistic plots are better than before but still not quite satisfying.

## e

## f

Add j-1 new predictors where j is the number of unique manufacturers, each variable will be an indicator if the car is one of j-1 specific manufacturer, and if all the variables are 0 then the car will be the reaming manufacturer. We may wish to do further analysis to see if we can collapse several brands into one category (for example luxury car vs non luxury car).