# STAT387 HOMEWORK#2

INTRODUCTION TO STATISTICAL LEARNING

Saeah Go

Due January 29, 12:00 PM

## Conceptual

### 1. 2

Carefully explain the differences between the KNN classifier and KNN regression methods.

The KNN classifier and KNN regression methods are quite similar. They both are given a value for $K$ and a prediction point $x_0$. But the difference is that, KNN classifier is usually used to solve **classification** problems with a qualitative response, by identifying the neighborhood of the study variable $x_0$ and identifying the estimation of the conditional probability, namely $P(Y = j|X = x_0)$ is made, for the class $j$, as a fraction of points in the neighborhood whose response values equals $j$. But the KNN regression method, is usually used to solve **regression** problems which are generally associated with a quantitative response by again identifying the neighborhood of $x_0$, and the the $f(x_0)$ is estimated as an average of all the training points of the neighborhood.

### 2. 3(a, b)

Suppose we have a data set with five predictors, $X_1$ = GPA, $X_2$ = IQ, $X_3$ = Gender (1 for Female and 0 for Male), $X_4$ = Interaction between GPA and IQ, and $X_5$ = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.
(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

**iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.**

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Reason:
The least square line is:

$$\begin{aligned}\hat{y} &= 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.04X_4 - 10X_5 \\ &= 50 + 20GPA + 0.07IQ + 35Gender \\ &+0.01GPA \times IQ - 10GPA \times Gender\end{aligned}$$

If we only observed for males, then, (Note that 0 is male)

$$\begin{aligned}\hat{y} &= 50 + 20GPA + 0.07IQ + 35 \times 0 + 0.04GPA \times IQ - 10GPA \times 0 \\ &= 50 + 20GPA + 0.07IQ + 0.01GPA \times IQ\end{aligned}$$

Similarly, for female, (female is 1)

$$\begin{aligned}\hat{y} &= 50 + 20GPA + 0.07IQ + 35 \times 1 + 0.04GPA \times IQ - 10GPA \times 1 \\ &= 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ\end{aligned}$$

Remember the the response is starting salary after graduation (in thousands of dollars). We can check that the starting salary for males is much higher than for females if $GPA \geq 3.5$. ($\because 50 + 20GPA \geq 85 + 10GPA$)

$$\begin{aligned}50 + 20GPA &\geq 85 + 10GPA, \\ 10GPA &\geq 35 \\ \therefore GPA &\geq 3.5\end{aligned}$$

Thus the option **iii** is correct.
(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.
For female, the fitted line we get is:

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01GPA \times IQ$$

And we are given that IQ is 110 and GPA is 4.0. We can just put the values.

$$\hat{y} = 85 + 10 \times 4.0 + 0.07 \times 110 + 0.01 \times 4.0 \times 110 = 137.1$$

Thus we predict the starting salary for the female is: \$137,100 (Since the response is starting salary in thousand of dollars)

# Applied

## 3. 8(a, b, c)

This question involves the use of simple linear regression on the Auto data set.
(a) Use the $lm()$ function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the $summary()$ function to print the results. Comment on the output.

I used R to answer this question.

```
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
```

For example:

i. Is there a relationship between the predictor and the response?

Yes. As we can see in the result above, the p-values for the regression coefficients are nearly zero. ($Pr(> |t|)$ is $< 2e-16$) This implies statistical significance, thus we can say that there is a relationship between the predictor and the response.

ii. How strong is the relationship between the predictor and the response?

The R-squared value is 0.605948. The $R^2$ value indicates that about 61% of the variation in the response variable ($mpg$) is due to the predictor variable ($horsepower$).

iii. Is the relationship between the predictor and the response positive or negative?

Negative. Since the coefficient of the variable 'horsepower' is negative ($-0.157845$), the relationship is negative, or specifically we can say inverse relationship exists.

iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

$$\begin{aligned} mpg &= \beta_0 + \beta_1 horsepower \\ mpg &= 39.94 - 0.16 \times 98 \\ &= 39.94 - 15.68 = 24.26 \end{aligned}$$

Thus, the predicted mpg associated with a horsepower of 98 is 24.26 miles per gallon.

Confidence Interval & Prediction Interval

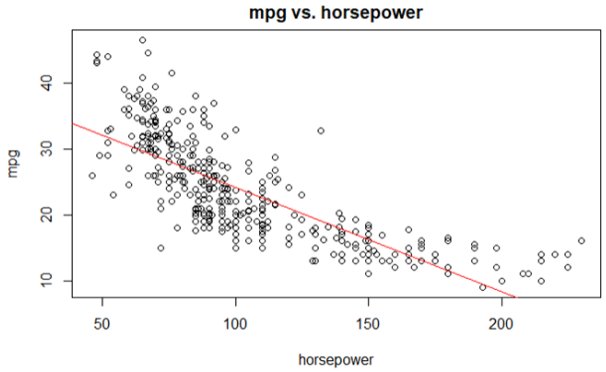
```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
       fit      lwr      upr
1 24.46708 14.8094 34.12476
```

Confidence interval is: [23.97, 24.96]
Prediction interval is: [14.81, 34.12]

(b) Plot the response and the predictor. Use the *abline*() function to display

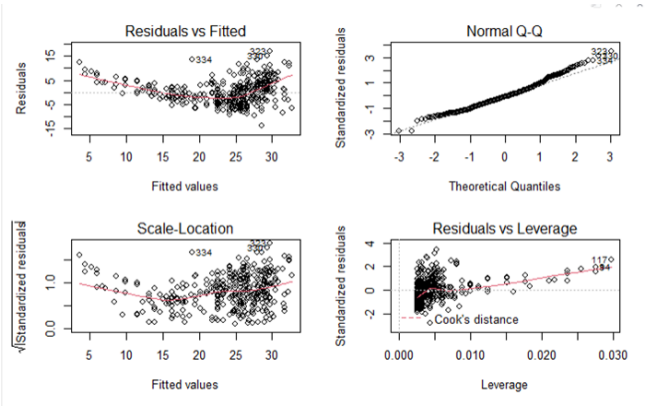
the least squares regression line.

I used R to answer this question.




(c) Use the *plot*() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

I used R to answer this question.




- The first plot (Residuals vs. Fitted) shows a pattern (U-shaped) between the residuals and the fitted values. This indicates a non-linear relationship between the predictor and response variables.
- The Normal Q-Q plot (second plot) shows that the residuals are normally distributed.
- The third plot (Scale-Location) shows that the variance of the errors is constant.
- The fourth plot (Residuals vs. Leverage) indicates that there are no leverage points in the data.

## 4. 10(a, b, c, d, e, f)

This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict *Sales* using *Price*, *Urban*, and

$US$.

I used R to answer this question.

```
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) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative!

The coefficient of Price is $-0.054459$. Note that Sales is unit sales (in thousands) at each location. So, when price increases by \$1000 and other predictors are held constant, sales decrease by 54.459 unit sales. In other words, when price increases by \$1000, the number of car-seats sold decrease by 54.459.

Note that the predictors Urban and US are qualitative. A store's sale is not much affected much by whether the store is in Urban area or not. We could notice this since the p-value of Urban is really high ($0.936 > 0.05$ which means the predictor Urban is not significant) and the coefficient is close to zero ($-0.021916$). But we can still say that a store's sale is approximately 22 less car-seats than a store in rural area.
Since the coefficient of the predictor US is 1.200573, store in the US sales 1201 more car-seats (in average) than a store that is abroad.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

I rounded to the nearest thousandth.

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

Or we can rewrite this for each situation.

$Sales = 13.0435–0.054Price–0.022 + 1.201$ (if store is in urban US area)
$Sales = 13.0435–0.054Price + 1.201$ (if store is in the rural US area)
$Sales = 13.0435–0.054Price–0.0219$ (if store is non-US urban area)
$Sales = 13.0435–0.054Price$ (if store is not in the US and not in urban area)

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

We reject the null hypothesis when p-value is less than 0.05. In this case, we can check that the p-value of the predictor Price is $< 2e-16$ and the p-value of the predictor US is $4.86e-06$. Thus the predictors Price and US, we can

reject the null hypothesis $H_0 : \beta_j = 0$.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

I used R to answer this question.

```
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) How well do the models in $(a)$ and $(e)$ fit the data?

The smaller (reduced) model that we made in $(e)$ is a bit better, since the adjusted $R^2$ value is slightly higher. The adjusted R-squared for $(a)$ is 0.2335 and the adjusted R-squared value for $(e)$ is 0.2354.

## 5. 13(a, b, c, d, e)

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

(a) Using the $rnorm()$ function, create a vector, $x$, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, $X$.

I used R to answer this question.

(b) Using the $rnorm()$ function, create a vector, $eps$, containing 100 observations drawn from a $N(0, 0.25)$ distribution *i.e.* a normal distribution with mean zero and variance 0.25.

I used R to answer this question.

(c) Using $x$ and $eps$, generate a vector $y$ according to the model

$$Y = -1 + 0.5X + \epsilon$$

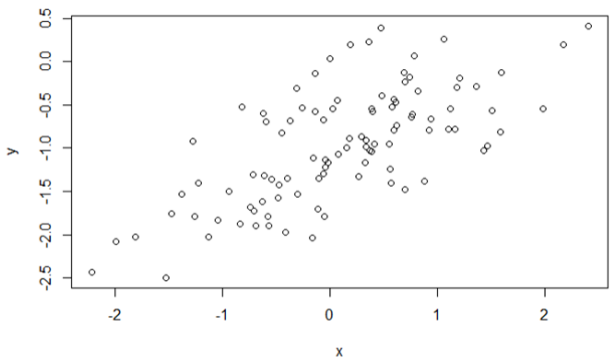
What is the length of the vector $y$? What are the values of $\beta_0$ and $\beta_1$ in this linear model?

I used R to answer this question.

The length of the vector $y$ is 100. The values of $\beta_0$ is $-1$ and $\beta_1$ is 0.5.

(d) Create a scatter plot displaying the relationship between $x$ and $y$. Comment on what you observe.

I used R to answer this question.




The data points are moderately spread out, but the values of $y$ increases as we move right along the $x$ axis. Thus we can conclude that between the variables $x$ and $y$, the relationship is quite positive.

(e) Fit a least squares linear model to predict $y$ using $x$. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to $\beta_0$ and $\beta_1$?

I used R to answer this question.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93842 -0.30688 -0.06975  0.26970  1.17309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
x            0.49947    0.05386   9.273 4.58e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4814 on 98 degrees of freedom
Multiple R-squared:  0.4674,    Adjusted R-squared:  0.4619
F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

Based on the result above, the estimated regression line is:

$$\hat{y} = -1.01885 + 0.49947x$$

The $R^2$ value is 0.4674, and this indicates that 46.94% variation in $y$ is explained by the model. The predictor $x$ is significant because the p-value is $4.583e-15$. The positive coefficient 0.4995 (slope) indicates that $x$ and $y$ have a positive relationship. In other words, it means that for one unit increase in $x$, $y$ increases by 0.4995.

We can check that the estimated values of the coefficients $(\hat{\beta}_0, \hat{\beta}_1)$ are pretty close to the true values $(\beta_0, \beta_1)$. $\hat{\beta}_0 = -1.0188, \hat{\beta}_1 = 0.4995$
$\beta_0 = -1, \beta_1 = 0.5$

# R Code

```r
# Applied

#################################
## 3.8 (a, b, c)
#################################
### Problem (a)
data(Auto) # load the data
fix(Auto)
fit1 <- lm(mpg ~ horsepower, data = Auto)
summary(fit1)

# get the associated 95 % confidence interval
predict(fit1, data.frame(horsepower = 98), interval = "confidence")

# get the associated 95 % prediction interval
predict(fit1, data.frame(horsepower = 98), interval = "prediction")


### Problem (b)
plot(Auto$horsepower, Auto$mpg,
     main = "mpg vs. horsepower",
     xlab = "horsepower",
     ylab = "mpg")
abline(fit1,
       col = "red") # make the line color to red to easily see


### Problem (c)
par(mfrow = c(2, 2)) # two columns two rows
plot(fit1)









#################################
## 4. 10 (a, b, c, d, e, f)
#################################
### Problem (a)
data(Carseats)
fix(Carseats)

fit2 <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(fit2)


### Problem (e)
fit3 <- lm(Sales ~ Price + US, data = Carseats)
```

```r
summary(fit3)
```

```r
###################################
## 5. 13(a, b, c, d, e)
###################################
### Problem (a)
set.seed(1)
x = rnorm(100)


### Problem (b)
eps = rnorm(100, sd = sqrt(0.25)) # variance = sd^2 # eps is
    epsilon (error term)


### Problem (c)
y = -1 + 0.5 * X + eps
length(y)


### Problem (d)
plot(x, y)


### Problem (e)
fit4 <- lm(y ~ x)
summary(fit4)
```