HW1\_Template

Lina Cao , Thursdays 3:30PM

SDGB 7844; Prof. Nagaraja; Fall 2017

## Question 1

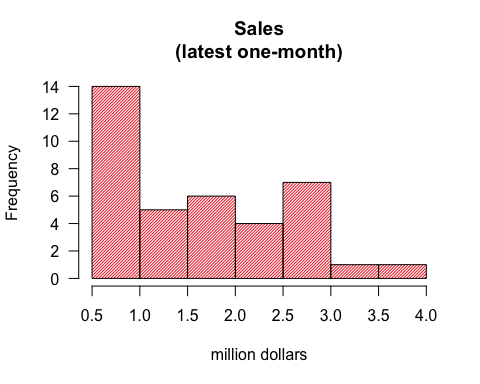
Identify the population and sample. What is the sample size?

1. population: All the stores in the large chain of nationwide sporting goods stores.
2. sample: The randomly chosen 38 sporting goods stores.
3. sample size: 38

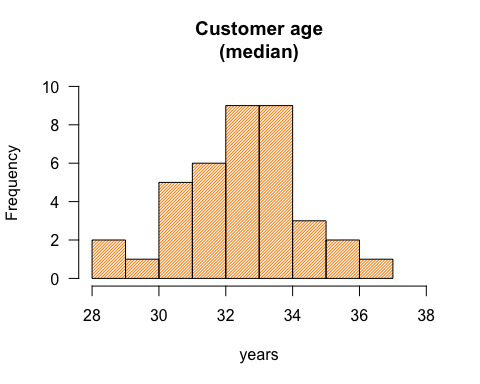
## Question 2

Identify and describe the variables in the dataset and the type of variable (continuous, discrete, nominal, or ordinal). Include histograms for each variable and discuss skewness/outliers/etc. Also, provide the following summary statistics for each variable in a table: mean, median, standard deviation, minimum and maximum. Remember to include units of measurement.

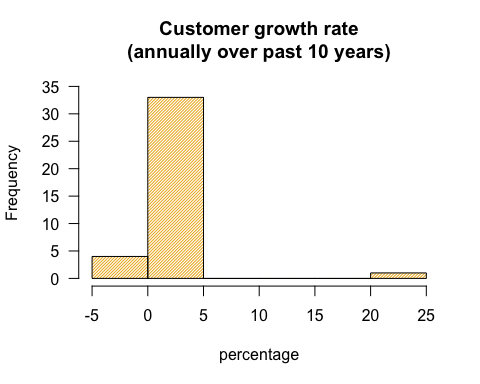
# import data  
s <- read.table("Sporting.csv",header = TRUE, sep = ",",stringsAsFactors = FALSE)  
# histograms for each variable  
  
# sales  
hist(s$Sales/1000000, las = TRUE,  
 main = "Sales\n(latest one-month)", xlab = "million dollars",  
 col = "firebrick2",density = 45, angle = 45, border = "black")



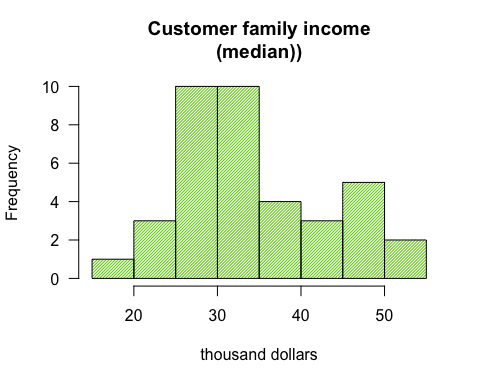
# age  
hist(s$Age, las = TRUE,  
 main = "Customer age\n(median)", xlab = "years",   
 xlim = c(28,38), ylim = c(0,10),  
 col = "darkorange",density = 45, angle = 45, border = "black")



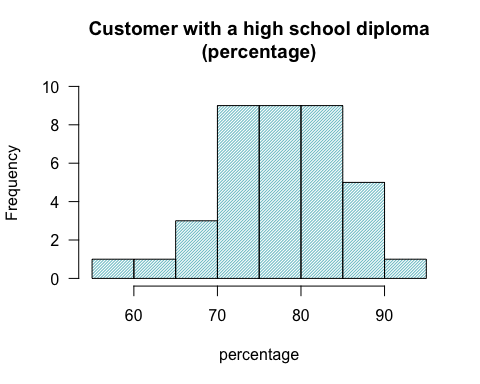
# growth  
hist(s$Growth, las = TRUE,  
 main = "Customer growth rate\n(annually over past 10 years)",   
 xlab = "percentage", ylim = c(0,35),  
 col = "goldenrod2",density = 45, angle = 45, border = "black")



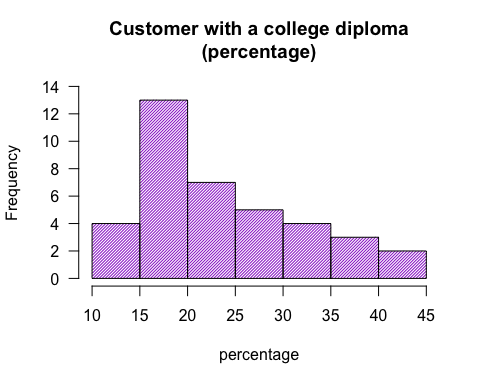
# income  
hist(s$Income/1000, las = TRUE,  
 main = "Customer family income\n(median))",   
 xlab = "thousand dollars", xlim = c(15,55),  
 col = "chartreuse3",density = 45, angle = 45, border = "black")



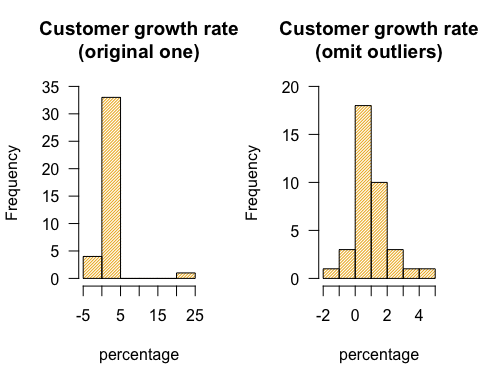
# HS  
hist(s$HS, las = TRUE,  
 main = "Customer with a high school diploma\n(percentage)",   
 xlab = "percentage", xlim = c(55,95), ylim = c(0,10),  
 col = "cadetblue3",density = 45, angle = 45, border = "black")



# college  
hist(s$College, las = TRUE,  
 main = "Customer with a college diploma\n(percentage)",   
 xlab = "percentage", ylim = c(0,14),  
 col = "darkorchid3",density = 45, angle = 45, border = "black")



# omit the outliers of Growth  
par(mfrow=c(1,2))  
mar=c(5,5,4,6)+1  
  
hist(s$Growth, las = TRUE,  
 main = "Customer growth rate\n(original one)", xlab = "percentage",  
 ylim = c(0,35),  
 col = "goldenrod2",density = 45, angle = 45, border = "black")  
hist(s$Growth[s$Growth<10], las = TRUE,  
 main = "Customer growth rate\n(omit outliers)", xlab = "percentage",  
 ylim = c(0,20),  
 col = "goldenrod2",density = 45, angle = 45, border = "black")



# since standard deviation is not included in summary() function  
# I choose to calculate the required explantory statistics first  
# and merge them into a new data.frame to apply kable()  
mean <- apply(s[,c("Sales","Age","Growth","Income","HS","College")],2,mean,na.rm=TRUE)  
median <- apply(s[,c("Sales","Age","Growth","Income","HS","College")],2,median,na.rm=TRUE)  
sd <- apply(s[,c("Sales","Age","Growth","Income","HS","College")],2,sd,na.rm=TRUE)  
min <- apply(s[,c("Sales","Age","Growth","Income","HS","College")],2,min,na.rm=TRUE)  
max <- apply(s[,c("Sales","Age","Growth","Income","HS","College")],2,max,na.rm=TRUE)  
  
table <- data.frame(mean,median,sd,max,min,stringsAsFactors = FALSE)  
rownames(table) <- c("Sales(dollars)","Age","Growth(%)","Income(dollars)","HS(%)","College(%)")  
  
# create table of summary statistics when knit  
# http://rmarkdown.rstudio.com/lesson-7.html  
library(knitr)  
kable(table)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | median | sd | max | min |
| Sales(dollars) | 1.638487e+06 | 1.507047e+06 | 9.077581e+05 | 3860007.3160 | 529215.4590 |
| Age | 3.245048e+01 | 3.267380e+01 | 1.722670e+00 | 36.1056 | 28.5166 |
| Growth(%) | 1.599332e+00 | 7.628500e-01 | 3.796262e+00 | 23.4630 | -1.0400 |
| Income(dollars) | 3.417460e+04 | 3.187626e+04 | 8.890971e+03 | 53063.7900 | 18730.8800 |
| HS(%) | 7.724139e+01 | 7.833635e+01 | 7.462192e+00 | 93.4996 | 58.3610 |
| College(%) | 2.366922e+01 | 2.082375e+01 | 8.147172e+00 | 41.7057 | 10.6729 |

rm(mean, median, sd, min, max)

All the variables belong to continuous variable. Only “Growth” has outliers, the rest five variables do not have outliers. From the histograms above,

1. “Sales” | right skewed
2. “Age” | slightly left skewed
3. “Growth” | a little right skewed after ignoring outliers(between 20~25)
4. “Income” | right skewed
5. “HS” | left skewed
6. “College”| right skewed

## Question 3

Compute the sample correlation coefficient r and scatter polts for all variables with sales (sales vs age; sales vs HS, etc). Is correlation an appropriate measure to describe the relationship between these variables? Justify your answer.

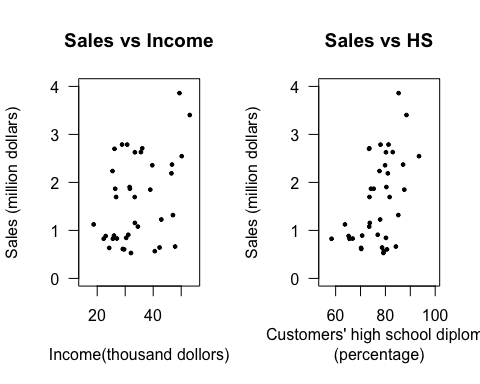
# correlation  
cor(s,use = "complete.obs")

## Sales Age Growth Income HS College  
## Sales 1.00000000 0.04133758 0.11221839 0.38364527 0.4904336 0.3217927  
## Age 0.04133758 1.00000000 -0.20294528 0.38740587 0.1926220 0.2648512  
## Growth 0.11221839 -0.20294528 1.00000000 0.03206261 0.2165781 0.3811376  
## Income 0.38364527 0.38740587 0.03206261 1.00000000 0.8035409 0.6815267  
## HS 0.49043361 0.19262195 0.21657808 0.80354092 1.0000000 0.7117698  
## College 0.32179270 0.26485125 0.38113756 0.68152665 0.7117698 1.0000000

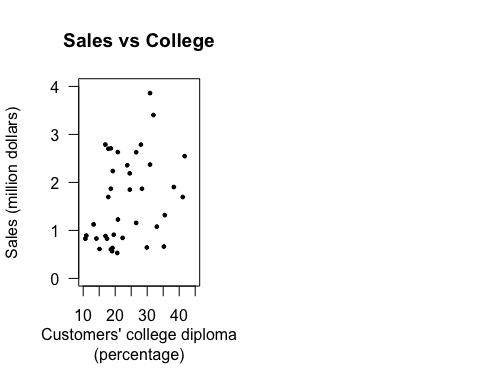
# scatter plots  
mar=c(10,10,10,10)+0.5  
par(mfrow=c(1,2))  
plot(s$Age, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.5, main = "Sales vs Age" ,   
 ylab = "Sales (million dollars)", xlab = "Age",  
 ylim = c(0,4), xlim = c(28,38))  
plot(s$Growth, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.5, main = "Sales vs Growth" ,   
 ylab = "Sales (million dollars)", xlab = "Growth(percentage)",  
 ylim = c(0,4), xlim = c(0,24))



par(mfrow=c(1,2))  
plot(s$Income/1000, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.5, main = "Sales vs Income" ,   
 ylab = "Sales (million dollars)", xlab = "Income(thousand dollors)",  
 ylim = c(0,4), xlim = c(15,55))  
plot(s$HS, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.5, main = "Sales vs HS" ,   
 ylab = "Sales (million dollars)", xlab = "Customers' high school diploma\n(percentage)",  
 ylim = c(0,4), xlim = c(55,100))



plot(s$College, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.5, main = "Sales vs College" ,   
 ylab = "Sales (million dollars)", xlab = "Customers' college diploma\n(percentage)",  
 ylim = c(0,4), xlim = c(10,45))



I don’t think correlation is an appropriate enough measurement used to describe the relationship between variables.  
Take “Sales vs. Growth” as an example.  
The correlation between these two variables is about 0.1122, indicating a positive relationship. However, the dots are literally cluster within the “band” where x-axis is between 0 and 5(%). Additionally, correlation cannot show information of outliers, as the point of which “Growth=23.46%” in the plot. When it comes to the cases between “Sales vs Age” and “Sales vs Income”, they have large difference between their correlations although they share a similar irregular pattern of scattered dotted plot.  
Therefore, only correlation itself will lose some other data information, it is not so appropriate when it comes alone.

## Question 4

Based on your analysis so far, choose the best explanatory variable to use in your model. Justify your choice.

I chose HS (Percentage of customer base with a high school diploma). Justify: It has the highest correlation, 0.4904, among all five variables. Also, from the scatter plot above, the linear relationship is more obvious than “Sales” with other variables.

## Question 5

Fit a liear regression model between sales and your explanatory variable. In your solution, include the output of the summary() function when applied to your linear model object. Write out the estimated regression line and remember to include the units of measurement.

lm.s <- lm(Sales/1000000~HS, data = s)  
summary(lm.s)

##   
## Call:  
## lm(formula = Sales/1e+06 ~ HS, data = s)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.38899 -0.55131 -0.05001 0.52498 1.74050   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.96974 1.37096 -2.166 0.03700 \*   
## HS 0.05966 0.01767 3.377 0.00177 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.802 on 36 degrees of freedom  
## Multiple R-squared: 0.2405, Adjusted R-squared: 0.2194   
## F-statistic: 11.4 on 1 and 36 DF, p-value: 0.001773

lm.s$coefficients

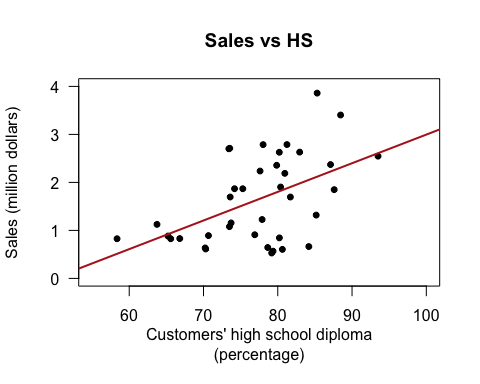
## (Intercept) HS   
## -2.96974123 0.05966009

Estimated Regression Line:  
Sales = -2.9697(million dollars) + 0.0597(million dollars/percent)\*HS(percent)

## Question 6

Include a scatter plot of sales versus your explanatory variable, add the estimated regression line to your scatter plot, and describe the fit of your model.

plot(s$HS, s$Sales/1000000, las = TRUE,  
 pch = 19, cex = 0.8, main = "Sales vs HS" ,   
 ylab = "Sales (million dollars)", xlab = "Customers' high school diploma\n(percentage)",  
 ylim = c(0,4), xlim = c(55,100))  
abline(a=lm.s$coefficients[1], b=lm.s$coefficients[2], lwd=2, col = "firebrick")



I don’t think the dots fit the model very well.  
Although the linear regression model is based on the least squared estimate, there are actually only two points that are very close to the line.  
Additionally, from the regression result in Question 5, the R-squared is 0.2405, which means only a small amount (about a quarter) of the data can be explained by this model.  
However, this line is already the best fit as a linear model using the dataset os “Sales” and “HS”。

## Question 7

Interpret the slope of the regression line. Over what range of x is the interpretation meaningful?

# slope   
lm.s$coefficients[2]

## HS   
## 0.05966009

# x range  
range(s$HS)

## [1] 58.3610 93.4996

When other variables remain the same, one percent higher percentage of customer base with a high school diploma can on average increase 0.0597 million dollars of sales.  
Since the regression is calculated based on the HS ranging from 58.3610 to 93.4996, the x in this range can make interpretation meaningful.

## Question 8

Interpret the y-intercept of the regression line. Does it have a practical meaning for this application? Justify your answer.

lm.s$coefficients[1]

## (Intercept)   
## -2.969741

The y-intercept comes from where x (HS) equals 0, which means the sales in the situation that percentage of customer base with a high school diploma is zero (i.e. all the customers has a lower than high school diploma).  
In this regression line, the y-intercept means when the customers all have lower than high school diploma, the sales can be -2.9697 million dollars. This indicates that the sporting stores will have negative number of sales.  
Thus in practical this means nothing, since sales number will always be positive, even a store has deficit in total revenue. A negative sales number means when the store sales something, it will give the customer money instead of receiving.

## Question 9

What is the coefficient of determination (r-squared)? Interpret it.

From Question 5, the R-squared is 0.2405.  
R-squared = 1 - SSE/SST, where SSE/SST is the fraction of variation in y that is not explained(SSE stands for total variation of the residuals) by x using the regression model. Therefore R-squared is the fraction of y that can be explained by the x using the model. Here, only around a quarter of the data can be explained by the model.

## Question 10

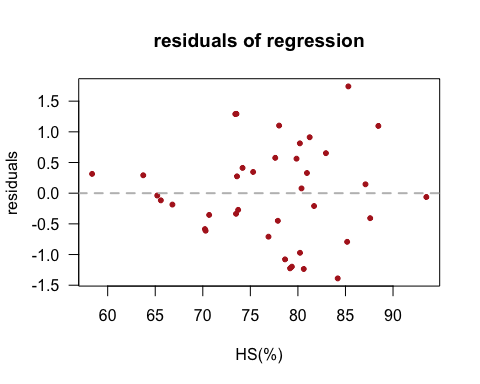
Check the regression assumptions and determine the adequacy of the fit of your regression model. Include your residual plot and the normal quantile plot of the residuals.

This regression model fits the assumptions:  
(1) The parameters are linearly related but data in one variable are independent;  
(2) The observations are randomly sampled with accuracy;  
The following assumptions need further computation to determine whether it is true in this regression:  
(3) The conditional mean of error should be zero;  
(4) There is no multi-collinearity;  
(5) There is no autocorrelation (should be homoscedasticity);  
(6) additionally, error(residuals) should be normally distributed.

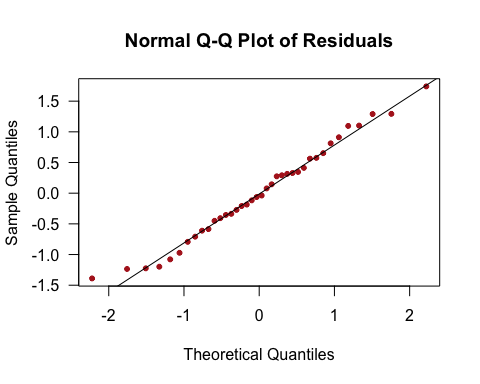
# mean of residuals  
mean(lm.s$residuals)

## [1] -3.30325e-18

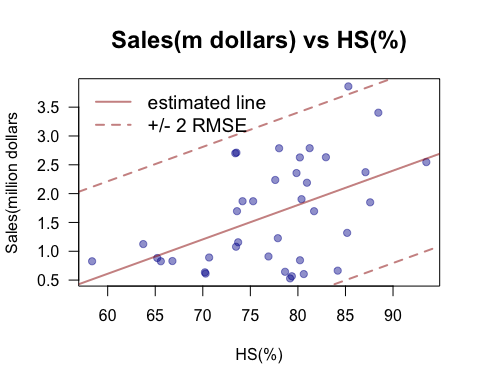
#residual plot  
plot(s$HS, lm.s$residuals, pch = 20, cex = 1,  
 col = "firebrick", las = TRUE,  
 main = "residuals of regression",   
 xlab = "HS(%)", ylab = "residuals")  
abline(h=0, lty=2, col="gray", lwd=2)



# quntile plot  
qqnorm(lm.s$residuals, las=TRUE, pch=20, col="firebrick",   
 main="Normal Q-Q Plot of Residuals")  
qqline(lm.s$residuals)



# check homoscedasticity  
plot(s$HS, s$Sales/1000000, las=TRUE, pch=19, col="#00009970",   
 xlab="HS(%)", ylab="Sales(million dollars",  
 main="Sales(m dollars) vs HS(%)", cex.lab=1, cex.axis=1, cex.main=1.5)  
# the regression line  
abline(coef(lm.s), col="#99000080", lwd=2)  
# moving parallel line for two sd  
abline(a=coef(lm.s)[1]+2\*summary(lm.s)$sigma, b=coef(lm.s)[2], col="#99000080", lwd=2, lty=2)  
abline(a=coef(lm.s)[1]-2\*summary(lm.s)$sigma, b=coef(lm.s)[2], col="#99000080", lwd=2, lty=2)  
legend("topleft", legend=c("estimated line", "+/- 2 RMSE"), bty="n", col="#99000080",   
 lwd=2, lty=c(1,2), cex=1.2)

 From the first calculation, the mean of the residuals is -3.30325e-18, which is small enough to regard it as zero. Assumption (3) is satisfied.  
The residual plot shows no pattern of residuals, which is a good sign. From the second plot above, the distributioin of residuals in this model is close to normal distribution, especially in central part of the data. However, there’s still some points at two sides show deviation from the normal quantile line, which implies the distribution may have a slightly heavier tail than normal distribution.  
For simple linear regression, there is just one variable, no multi-collinearity will occur between explanatory variables, assumption (4) is satisfied.  
The last plot illustrates that for almost all x-value, the corresponding value varies bwtween two standard deviation of error, there’s only one dot slightly outside the area. Assumption (5) is satisfied.

## Question 11

What is the RMSE for your regression model? (Remember to include units of measurement.) Explain what it means in this context. Is it high or low? Justify your answer.

# extract RMSE from regression  
summary(lm.s)$sigma

## [1] 0.8020038

# approximation for RMSE  
#sd(s$Sales/1000000)\*sqrt(1-cor(s$Sales, s$HS)^2)  
# the result is 0.7910917  
range(s$Sales/1000000)

## [1] 0.5292155 3.8600073

# the proportion of RMSE in the range of Sales  
summary(lm.s)$sigma/((max(s$Sales)-min(s$Sales))/1000000)

## [1] 0.2407847

Here RMSE= 0.8020,  
RMSE measures the standard deviation around the estimated line given a HS value. Therefore, here, given an HS value, the estimated sales will have a standard deviation of 0.8020 million dollars around the line.  
As a number lower than 1, this RMSE seems to be a small one. However, compared to the magnitude of “Sales”(in million dollars) and its range, (0.5292, 3.8600), RMSE takes up almost one fourth of the range. Therefore I think RMSE is high in this case.

## Question 12

At the alpha=0.05 level of significance, is there evidence of a linear relationship between the response and explanatory variable? (Answer this quesiton regardless of your response to quesiton 10). Make shure to specify the hypotheses, test statistics, degree of freedom, p-value and your conclusion.

# test statistic = (hat(beta)\_0 - 0)/se for hat(beta)\_0  
test.stat <- (lm.s$coef[2] - 0)/summary(lm.s)$coefficients[2,2]  
test.stat

## HS   
## 3.376562

# calculate two-sided p-value  
2\*pt(test.stat, df=nrow(s)-2, lower.tail=FALSE)

## HS   
## 0.001772702

# short way:  
confint(lm.s, parm="HS", level=0.95)

## 2.5 % 97.5 %  
## HS 0.02382593 0.09549425

We want to test the following hypotheses:

# H\_0: beta\_0 = 0  
# H\_1: beta\_0 != 0

The calculation tells that the test statistic is 3.3766. With the degree of freedom to be 38-2 = 36, the calculated p-value is 0.0018.  
Two-sided p-value is less than the significance level alpha, which means we reject the null hypothesis, i.e. the linear relationship between Sales and HS is significant under 0.05 significance level.  
The quick way to compute the confidence interval of the slope can be interpreted as: there is a interval (0.0238, 0.0955) that has 95% chance to include the true population slope, beta\_0. Hence, the slope has a large propability not likely to be zero, indicating evidence of a linear relationship between HS and Sales.

## Question 13

Based on your model, predict the expected sales revenues for counties A, B, C. For each county, is your prediction an interpolation or an extrapolation? Justify.

# from the table in paper, HS is 50,75,70 for County A,B,C respectively  
new.HS <- data.frame("HS"=c(50, 75, 70))  
# use predict function  
predict(lm.s, newdata=new.HS)

## 1 2 3   
## 0.0132634 1.5047657 1.2064652

# range of HS  
range(s$HS)

## [1] 58.3610 93.4996

The expected sales revenue for each County A,B,C is 0.0133, 1.5048, 1.2065 million dollars respectively.  
The predictions of B and C are both interpolation predictions, because the x-value HS ranges from 58.3610 to 93.4996, which contains their x-value 75, 70. However, the prediciton of A is extrapolation, since the x-value of A is smaller than the lower bound of (58.3610, 93.4996), and thus outside the range.

## Question 14

To learn some information about the quality of the predictions for the three counties, is it better to compute a 95% confidence interval or a 95% prediction interval? Justify your answer.

It is better to compute a 95% prediction interval. Because 95% confidence interval means there is a 95% chance that the confidence interval for estimated sales covers the true value. But here we are given specific HS values, in which case we may expect more uncertainty. Thus a prediction interval is better.

Compute the intervals for all three counties for the type of interval you have chosen.

# the predicted data for conties are as above, 50,75,70  
predict(lm.s, newdata=new.HS, interval="prediction", level=0.95)

## fit lwr upr  
## 1 0.0132634 -1.9019817 1.928508  
## 2 1.5047657 -0.1449926 3.154524  
## 3 1.2064652 -0.4616432 2.874574

The prediction intervals for each county are:  
A. (-1.9020, 1.9285)  
B. (-0.1450, 3.1545)  
C. (-0.4616, 2.8746)

## Question 15

What county do you recommend the company to open a store in? Justify your answer.

I will recommend County B.  
(1) From Q13, B will have the highest Sales estimation (or the same number in “fit” in Q14);  
(2) From Q14, the prediction interval of B has larger chance to get a positive and higher sales value; therefore, if the new store is open in County B, the sporting store can expect a higher profit.