HW3_ISYE6414_ashish_dhiman

Ashish Dhiman | adhiman9@gatech.edu

2022-09-12

```
library(ggplot2)
setwd("~/data_projects/fall22_hw/isye6414_hw/hw3")
```

Read Data and Summary

```
head -5 ./6414_HW3_Clean.csv
```

```
## Demand,PriceDif
## 7.38,-0.05
## 8.51,0.25
## 9.52,0.60
## 7.50,0.00
```

```
df_demand_price = read.table(file ="./6414_HW3_Clean.csv", sep=",",header=TRUE)
head(df_demand_price)
```

```
##   Demand PriceDif
## 1   7.38    -0.05
## 2   8.51     0.25
## 3   9.52     0.60
## 4   7.50     0.00
## 5   9.33     0.25
## 6   8.28     0.20
```

```
dim(df_demand_price)
```

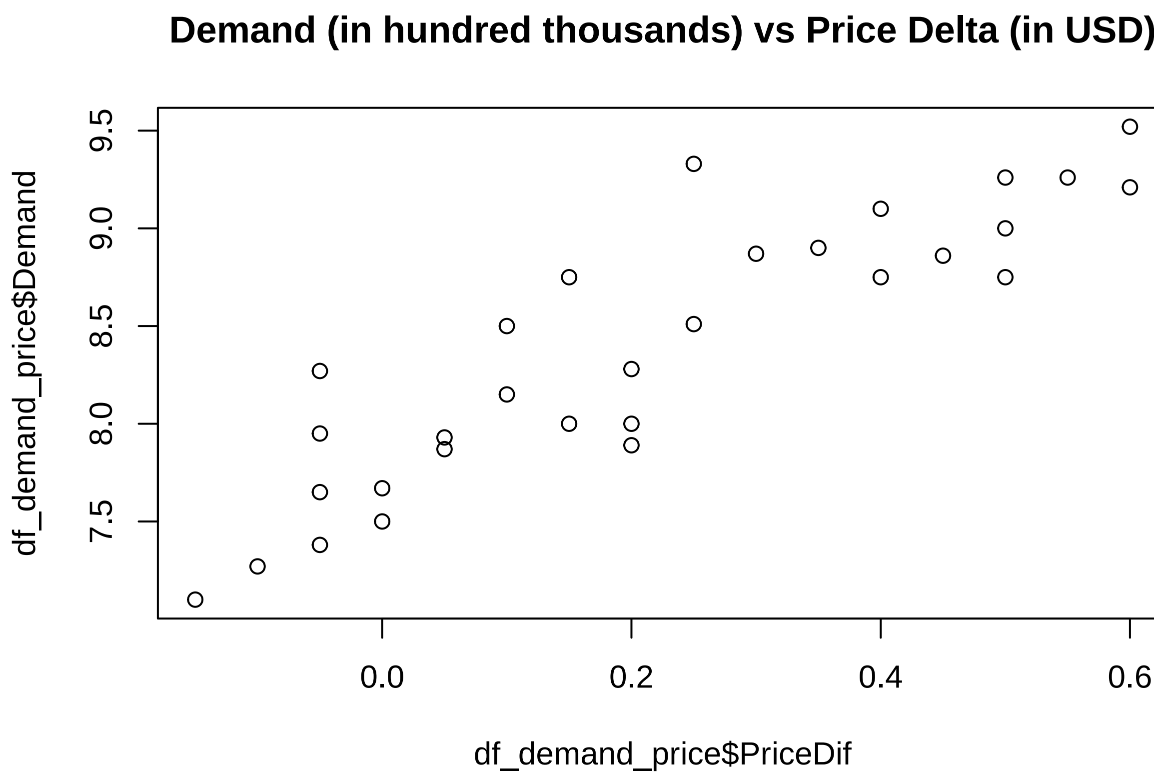
```
## [1] 30  2
```

```
summary(df_demand_price)
```

```
##      Demand         PriceDif
##  Min.   :7.100   Min.   :-0.1500
##  1st Qu.:7.900   1st Qu.: 0.0125
##  Median :8.390   Median : 0.2000
##  Mean   :8.383   Mean   : 0.2133
##  3rd Qu.:8.893   3rd Qu.: 0.4000
##  Max.   :9.520   Max.   : 0.6000
```

Question 1: Scatter Plot

```
title_i = "Demand (in hundred thousands) vs Price Delta (in USD)"
plot(x=df_demand_price$PriceDif, y=df_demand_price$Demand, type ="p",main = title_i)
```



From the above plot, a linear realationship between Demand and Price Difference is apparent

The strength of the linear realtionship can also be tested with corealtion between x and y

```
cor_xy = cor(df_demand_price$Demand,df_demand_price$PriceDif)
print (paste("Corealtion on full data:",round(cor_xy,2)))
```

```
## [1] "Corealtion on full data: 0.89"
```

A corelation of 0.89 is pretty significant and further supports strong linear realtionship between x and y

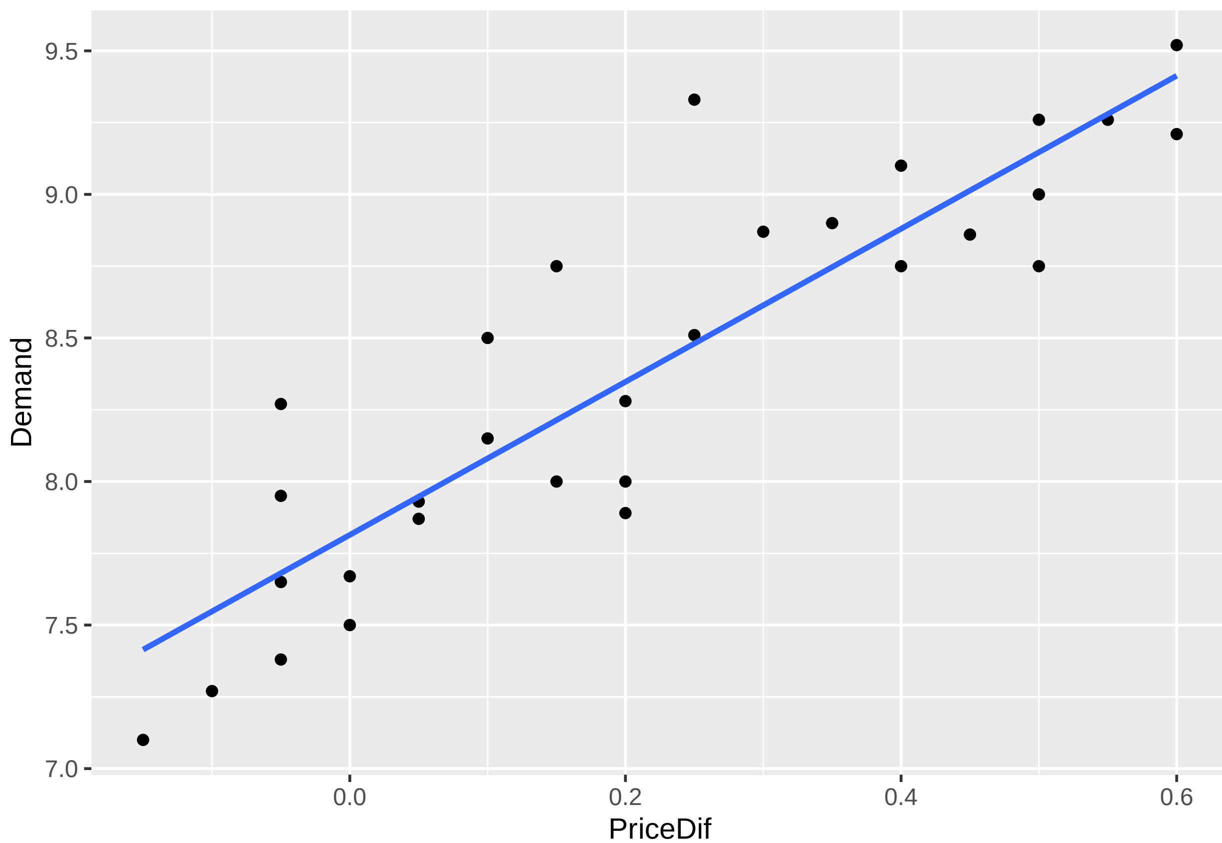
Question 2: Simple Linear Regression and Intercepts

```
#Fit SLR
slr_model <- lm(Demand ~ PriceDif, data = df_demand_price)
slr_model
```

```
##
## Call:
## lm(formula = Demand ~ PriceDif, data = df_demand_price)
##
## Coefficients:
## (Intercept)     PriceDif
##       7.814        2.665
```

```r
#Superpositioning regression line on
ggplot(df_demand_price, aes(PriceDif, Demand)) + #aes(x,y)
  geom_point() +
  stat_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
summary(slr_model)
```

```
##
## Call:
## lm(formula = Demand ~ PriceDif, data = df_demand_price)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45713 -0.21121 -0.04898  0.14314  0.84961
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.81409    0.07988   97.82  < 2e-16 ***
## PriceDif     2.66521    0.25850   10.31 4.88e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3166 on 28 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7841
## F-statistic: 106.3 on 1 and 28 DF,  p-value: 4.881e-11
```

From above summary we have:

$$\hat{\beta}_0 = 7.81409$$
$$\hat{\beta}_1 = 2.66521$$
$$\hat{\sigma} = 0.3166$$
$$se(\hat{\beta}_0) = 0.07988$$
$$se(\hat{\beta}_1) = 0.25850$$

Question 3, 95% CI for $\hat{\beta}_1$

$$CI\ for\ \hat{\beta}_1\ at\ (1-\alpha)\% = \hat{\beta}_1 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{S_{xx}}}$$

```
#In R this is given as:
confint(slr_model,level = 0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 7.650452 7.977723
## PriceDif    2.135702 3.194727
```

```
print (paste("95% CI for beta1 is (2.135702,3.194727)"))
```

```
## [1] "95% CI for beta1 is (2.135702,3.194727)"
```

```
print (paste("Length of CI in terms of sd:",round((3.194727-2.135702)/0.25850,2)))
```

```
## [1] "Length of CI in terms of sd: 4.1"
```

From 95% CI we can ascertain that beta1 lies within (2.135702,3.194727) range with 95% probability. B'cos the above CI is taken from t-distribution, which is fatter at tails (relative to normal), we get 4.1sd compared to 4 for normal.

Question 4: Hypothesis Test on if x is statistically significant

For predictor x to be statistically significant, β_1 should not be 0. Let us conclude a Hypothesis Test for it:

Null Hypothesis $= H_0 : \hat{\beta_1} = 0$

Alternate Hypothesis $= H_1 : \hat{\beta}_1 \neq 0$

Then from the model we have, Test Statistic $= \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{MSE}{S_{xx}}}} = 10.31$

```
print ("Critical t value, for alpha 5%:")
```

```
## [1] "Critical t value, for alpha 5%:"
```

```
qt(p=0.975,df=28)
```

```
## [1] 2.048407
```

```
qt(p=0.025,df=28)
```

```
## [1] -2.048407
```

Here we are performing a two tailed test for our Null Hypothesis using $\alpha = 5\%$. To reject H_0 we want, the test statistic (i.e. the t value) to be:

$$t \in (-\inf, -2.048407) \cup (2.048407, inf)$$

In this case our t value of 10.31 lies in the rejection region. Therefore we can conclude that at alpha = 5%, there is statistical evidence s.t. beta1 is not 0. In other words, given the current data there is <5% chance that β_1 is 0. This further implies that there is support for linear relationship between x & y, else there would not be enough statistical evidence to refute $H_0 : \beta_0 = 0$

```
summary(slr_model)$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 7.814088 0.07988432 97.81754 4.851255e-37
## PriceDif    2.665214 0.25849959 10.31032 4.881335e-11
```

Question 5: p value for β_0

From model summary the p value here is $< 4.851255 * 10^{-37}$. This is very strong(or low) p-value and for almost any typical alpha level of 1%, 5%, 10%, we have statistical evidence to reject the null hypothesis $H_0 : beta_0 = 0$

Question 6: p value for β_1

From model summary the p value here is $(4.88 * 10^{-11})$

For Null Hypothesis $= H_0 : \hat{\beta}_1$

Because p-value is less than 10%, 5% and 0.5%, we can reject null hypothesis at all these alpha levels.

Now we have such a low value, this implies given this data it is highly highly improbable (1 in 10^{11}) times that we fail to reject H0 when H0 is correct. In other words, we have support for very very strong linear relationship between x & y.

Question 7: point estimate & 95% CI for mean demand value for $x = 0.1$

We have to find $E[\hat{y}|x = 0.1]$

```
test = data.frame(PriceDif = 0.1)
predict.lm(slr_model, test, interval = "confidence", level = 0.95)
```

```
##        fit      lwr     upr
## 1 8.080609 7.947878 8.21334
```

From above the point estimate is 8.080609, and 95% CI is (7.947878,8.21334)

Question 8: point estimate & 95% prediction interval for actual demand value for $x = 0.1$

```
predict.lm(slr_model, test, interval = "predict", level = 0.95)
```

```
##        fit      lwr    upr
## 1 8.080609 7.418719 8.7425
```

From above the point estimate in this case is 8.080609, and 95% CI is (7.418719,8.7425)

```
half_length_ci = (8.21334-7.947878)/2
half_length_pi = (8.7425-7.418719)/2

print (paste("half Length CI",half_length_ci))
```

```
## [1] "half Length CI 0.132731"
```

```
print (paste("half Length PI",half_length_pi))
```

```
## [1] "half Length PI 0.6618905"
```

```
print (half_length_pi/half_length_ci)
```

```
## [1] 4.986706
```

Because prediction variance has extra 1 in the variance term, prediction interval is 5 times larger than CI.

Question 9: point estimate & 95% CI for mean demand value for $x = 0.25$

```
test2 = data.frame(PriceDif = 0.25)
predict.lm(slr_model, test2, interval = "confidence", level = 0.95)
```

```
##        fit     lwr      upr
## 1 8.480391 8.36042 8.600362
```

Point Estimate is 8.480391 with 95% CI is (8.36042,8.600362)

```r
half_length_ci2 = (8.600362-8.36042)/2
print(paste("mean x and Hal Length CI for x=0.25 = ",mean(df_demand_price$PriceDif),half_length_ci2))
```

```
## [1] "mean x and Hal Length CI for x=0.25 =  0.213333333333333 0.119971"
```

Because 0.25 is closer to mean x vs 0.1, the half length of ci in this is case smaller compared to ci for x=0.1. This is because the CI term has a factor of $x_i - \bar{x}$

Question 10: Derivation

For $\hat{\beta}_0 = 0$, then we have:

$$\hat{y}_i = \hat{\beta}_1 x_i; \quad then\ SSE = \sum (y_i - \hat{y}_i)^2 = (y_i - \hat{\beta}_1 . x_i)^2$$

We want to find beta1 which minimises SSE. So we take derivative wrt beta1 and equate it to 0.

$$\frac{\partial SSE}{\partial \beta_1} = \sum_i^n [2(y_i - \hat{\beta}_1 x_i).(-x_i)] = 0$$

$$\implies \sum_i^n x_i.y_i = \hat{\beta}_1 . \sum_i^n x_i^2 \quad or\ \hat{\beta}_1 = \frac{\sum_i^n x_i.y_i}{\sum_i^n x_i^2}$$

Hence Proved