Report on Capstone 2020

Chao Wang 19000151

6/4/2020

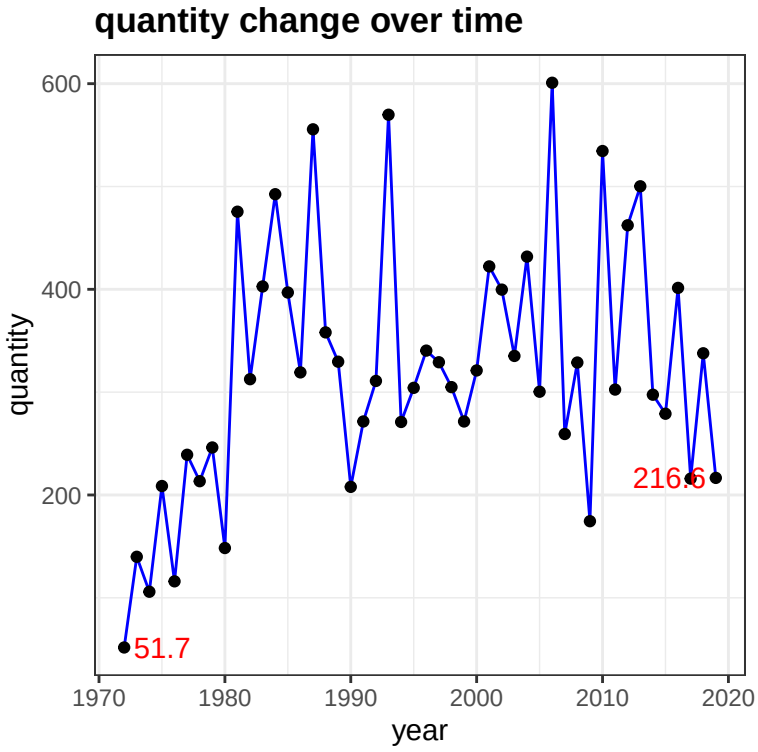
GBXQ3

In order to write a report on avocado data, I started an R project which is also stored in Github. 'https://github.com/Kennethws/capstone-stat-2020.git'

Question 1

First, I studied how the variables `quantity` and `price` vary over time, respectively.

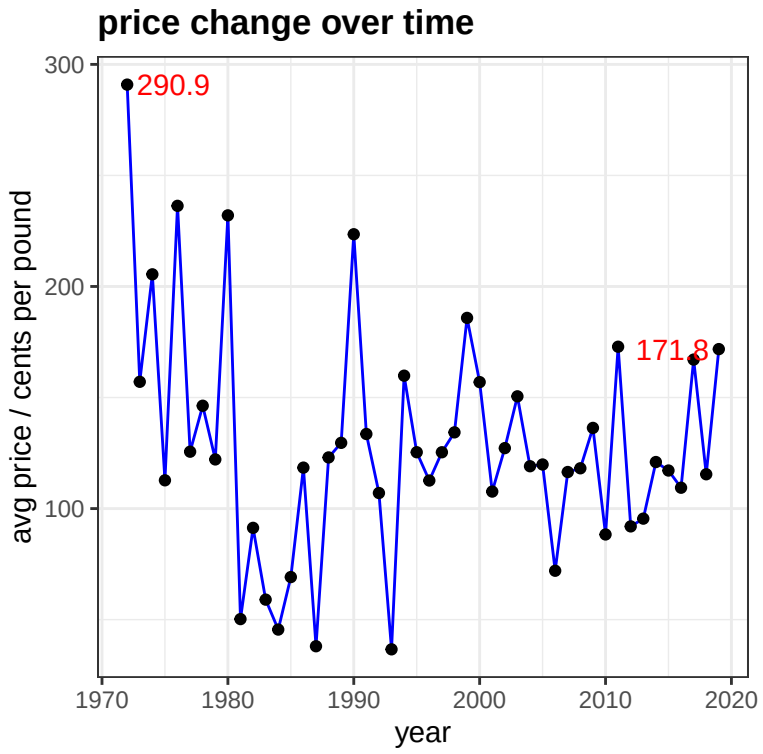
On one hand, the following is a simple time series plot for `quantity` over time:



As shown in the plot above, the quantity increased from 51.7 million pounds in 1972 to 216.6 million pounds in 2019 with dramatic fluctuation. This implies the development of

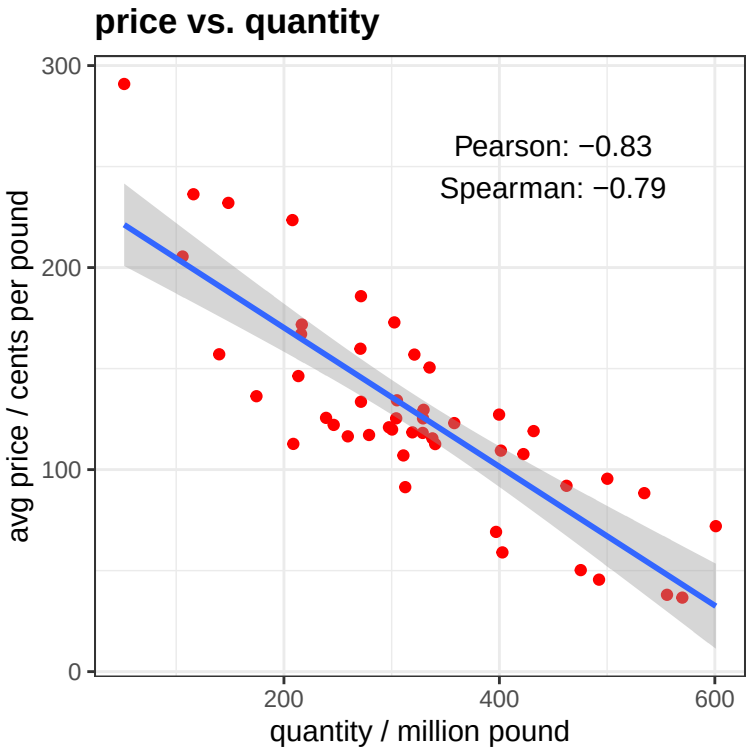
productivity as well as the surging demand worldwide.

On the other hand, below is the time series plot for `price` over time:



We can therefore conclude from the graph that the average price suffered a great loss from 290.9 to 171.8 cents per pound between 1970 and 2019, which was also accompanied by huge fluctuation.

Finally, I looked closer into the relationship between `quantity` and `price` by plotting a scatterplot supplemented by Pearson and Spearman correlation coefficients.



It seems safe to state that `quantity` and `price` are negatively correlated to a great extent.

Question 2

(a)

Mathematical equation:

$$E(Y|X = x) = \beta x + \alpha \tag{1}$$

After fitting a linear model:

term	estimate	std.error	statistic	p.value
(Intercept)	238.99	11.77	20.31	0
quantity	-0.34	0.03	-10.05	0

Given the table, the estimate of α means that the expected average price will be 238.99 cents per pound when the quantity is 0 pound, assuming the linear relationship still holds towards the end.

The estimate of β signifies that the expected average price will decrease by 0.34 million pound for every unit increase of the quantity.

(b)

To calculate a 95% confidence interval for beta:

```
n <- nrow(dat)
x.bar <- mean(dat$quantity)
```

```r
y.bar <- mean(dat$price)
Cxy <- cov(dat$quantity, dat$price) * (n - 1)
Cxx <- crossprod(dat$quantity - x.bar)[1]
beta.hat <- Cxy / Cxx
alpha.hat <- y.bar - beta.hat*x.bar
RSS <- crossprod(dat$price - alpha.hat - beta.hat*dat$quantity)[1]
sigma.hat <- sqrt(RSS / (n-2))
# 95% CI for beta
t <- qt(p = .975, df = n-2)
round(beta.hat + c(-1, 1) * t * sigma.hat / sqrt(Cxx), 3)
```

```
## [1] -0.413 -0.275
```

The confidence interval simply means that there is 95% chance where the estimate of beta computed would be within the interval shown above.

Also, using this confidence interval, we could reject the null hypothesis $H_0 : \beta = 0$, as 0 is not included within the range.

Question 3

(a)

L denotes the event that this farmer makes a loss in 2020.
Let $T_1 = \frac{112.1-Y}{29.53} \sim t_{46}$, so
$P(L) = P(Y \leq 80) = P(T_1 \geq \frac{121.8-80}{29.53}) = 0.085$

(b)

A denotes that `quantity` is 369 million pounds.
B denotes that `quantity` is 300 million pounds.
Let $T_2 = \frac{135.8-Y}{29.48} \sim t_{46}$, then
$P(L) = P(A, L) + P(B, L) \ = P(A)P(L|A) + P(B)P(L|B)$
$= \frac{1}{2}P(T_1 \geq \frac{121.8-80}{29.53}) + \frac{1}{2}P(T_2 \geq \frac{135.8-Y}{29.48})$
$= 0.059$

(c)

By Bayes Theorem,
$P(A|L) = \frac{P(A)P(L|A)}{P(L)} = 0.723$
$P(B|L) = \frac{P(B)P(L|B)}{P(L)} = 0.274$
Since $P(A|L) > P(B|L)$, CAC's forecast has the greater probability of being correct.