

# R Review Session

Brett Gordon and Joonhyuk Yang

10/07/21

## 1 Setup

This example file uses data from Dominick's, a now defunct grocery store chain in Chicago. First, I read in the data and print out a quick summary of the variables.

```
df <- read_excel("dominicks_oj.xlsx")
str(df)
```

```
tibble [12,512 x 10] (S3: tbl_df/tbl/data.frame)
 $ zone      : chr [1:12512] "CubFighter" "CubFighter" "CubFighter" "CubFighter" ...
 $ week      : num [1:12512] 1 2 3 4 5 6 7 8 9 10 ...
 $ holiday   : num [1:12512] 0 0 0 0 0 0 0 1 0 0 0 ...
 $ brand     : chr [1:12512] "STORE" "STORE" "STORE" "STORE" ...
 $ size      : num [1:12512] 16 16 16 16 16 16 16 16 16 16 ...
 $ brand_size: chr [1:12512] "STORE_16" "STORE_16" "STORE_16" "STORE_16" ...
 $ units     : num [1:12512] 892 1035 1139 690 898 ...
 $ price     : num [1:12512] 1.54 1.54 1.54 1.54 1.54 ...
 $ cost      : num [1:12512] 1.21 1.21 1.21 1.21 1.21 ...
 $ merch     : num [1:12512] 0 0 0 0 0 0 0 0 0 0 ...
```

```
head(df)
```

```
# A tibble: 6 x 10
  zone      week holiday brand  size brand_size units price  cost merch
  <chr>    <dbl>   <dbl> <chr> <dbl> <chr>      <dbl> <dbl> <dbl> <dbl>
1 CubFighter    1       0 STORE    16 STORE_16    892  1.54  1.21     0
2 CubFighter    2       0 STORE    16 STORE_16   1035  1.54  1.21     0
3 CubFighter    3       0 STORE    16 STORE_16   1139  1.54  1.21     0
4 CubFighter    4       0 STORE    16 STORE_16    690  1.54  1.21     0
5 CubFighter    5       0 STORE    16 STORE_16    898  1.54  1.21     0
6 CubFighter    6       0 STORE    16 STORE_16    584  1.54  1.21     0
```

## 2 Manipulating the data

Let's create some new variables.

```
df <- df %>%
  mutate(lnp = log(price),
         lnq = log(units),
         Dmerch = factor(merch),
         Dholiday = factor(holiday),
         Dsize = factor(size),
         Dzone = factor(zone))
head(df)
```

```
# A tibble: 6 x 16
  zone week holiday brand size brand_size units price cost merch lnp lnq
<chr> <dbl> <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 CubF~ 1 0 STORE 16 STORE_16 892 1.54 1.21 0 0.432 6.79
2 CubF~ 2 0 STORE 16 STORE_16 1035 1.54 1.21 0 0.432 6.94
3 CubF~ 3 0 STORE 16 STORE_16 1139 1.54 1.21 0 0.432 7.04
4 CubF~ 4 0 STORE 16 STORE_16 690 1.54 1.21 0 0.432 6.54
5 CubF~ 5 0 STORE 16 STORE_16 898 1.54 1.21 0 0.432 6.80
6 CubF~ 6 0 STORE 16 STORE_16 584 1.54 1.21 0 0.432 6.37
# ... with 4 more variables: Dmerch <fct>, Dholiday <fct>, Dsize <fct>,
# Dzone <fct>
```

Quick tabulations by zone, brand, and brand-size.

```
table(df$zone)
```

CubFighter	High	Low	Medium
3129	3129	3125	3129

```
table(df$brand)
```

CITHI	MMAID	STORE	TROPI
767	4443	4443	2859

```
table(df$brand, df$size)
```

	6	12	16
CITHI	0	767	0
MMAID	1584	1279	1580
STORE	1584	1279	1580
TROPI	0	1279	1580

Summary statistics by brand.

```
# Total sales, mean price, number of observations *for each* brand
df %>%
  group_by(brand) %>%
  summarize(sum_sales = sum(units),
            mean_price = mean(price),
            num_obs = n())
```

```
# A tibble: 4 x 4
  brand sum_sales mean_price num_obs
  <chr>    <dbl>      <dbl>   <int>
1 CITHI  832635      1.51     767
2 MMAID  4831259      1.49    4443
3 STORE  9864500      1.18    4443
4 TROPI  3101784      1.70    2859
```

Summary statistics for a subset of the data.

```
df %>%
  filter(size==12) %>%
  group_by(brand) %>%
  summarize(sum_sales = sum(units),
            mean_price = mean(price),
            num_obs = n())
```

```
# A tibble: 4 x 4
  brand sum_sales mean_price num_obs
  <chr>    <dbl>      <dbl>   <int>
1 CITHI  832635      1.51     767
2 MMAID  3641990      1.46    1279
3 STORE  6301080      1.20    1279
4 TROPI  2615389      1.42    1279
```

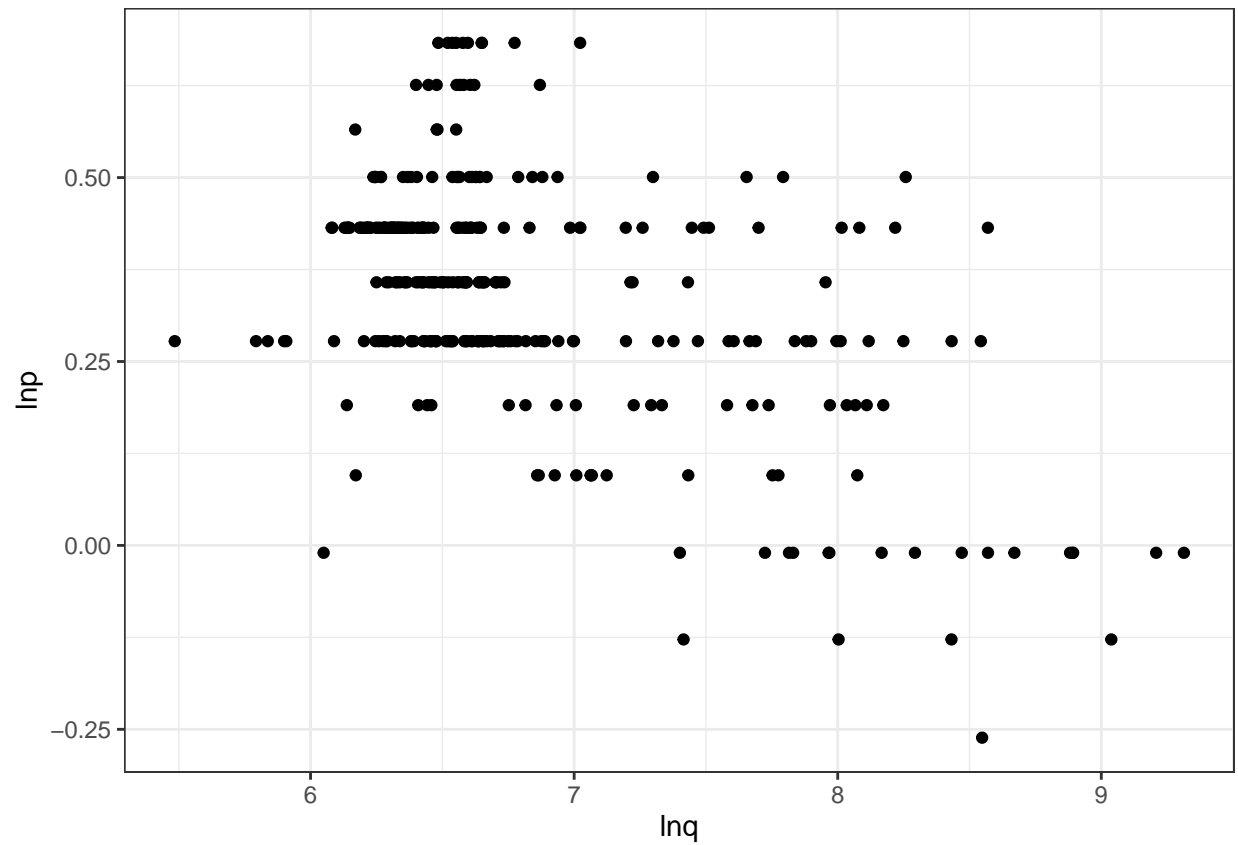
Let me create some handy subsets of the data (and note my variable naming scheme).

```
df.low <- df %>% filter(zone=="Low")
df.mmaid12 <- df %>% filter(brand=="MMAID" & size==12)
df.low.mmaid12 <- df.mmaid12 %>% filter(zone=="Low")
```

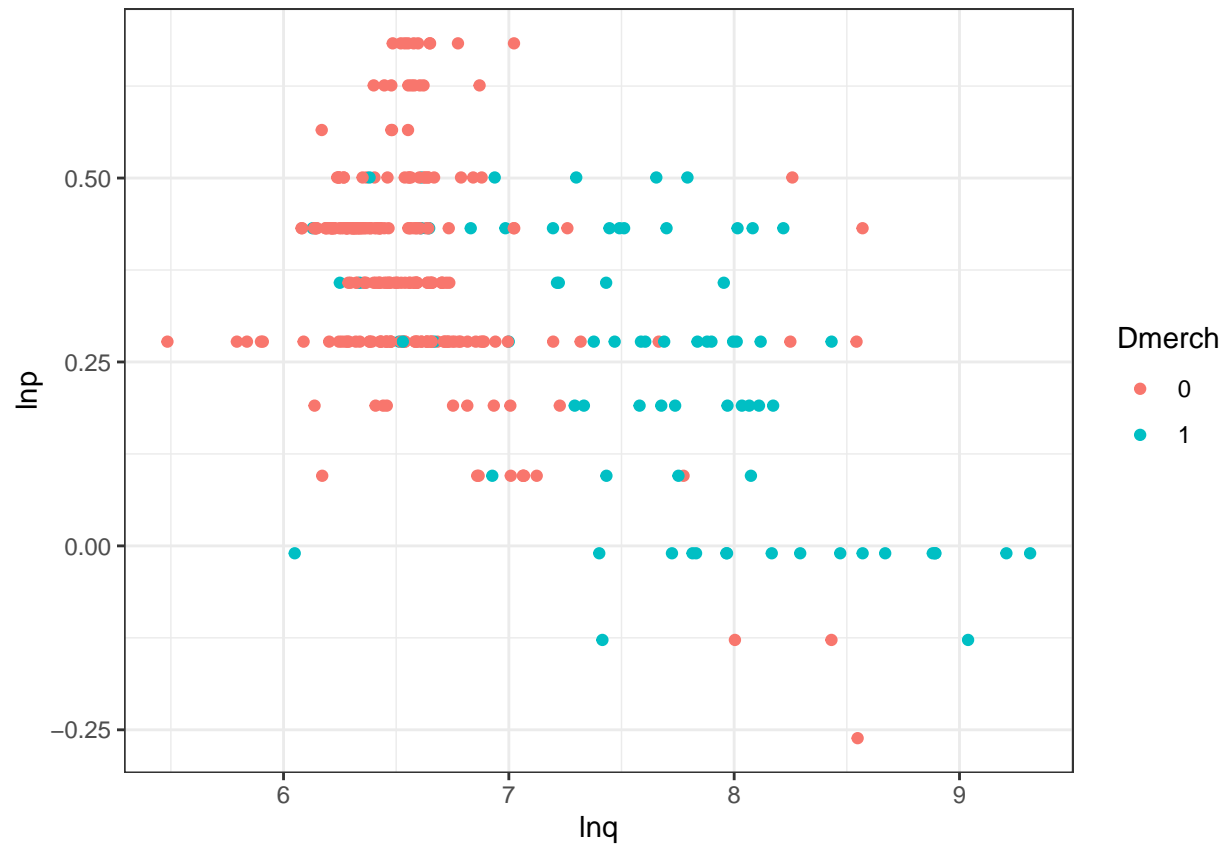
## 3 Plots

### 3.1 Scatter plots

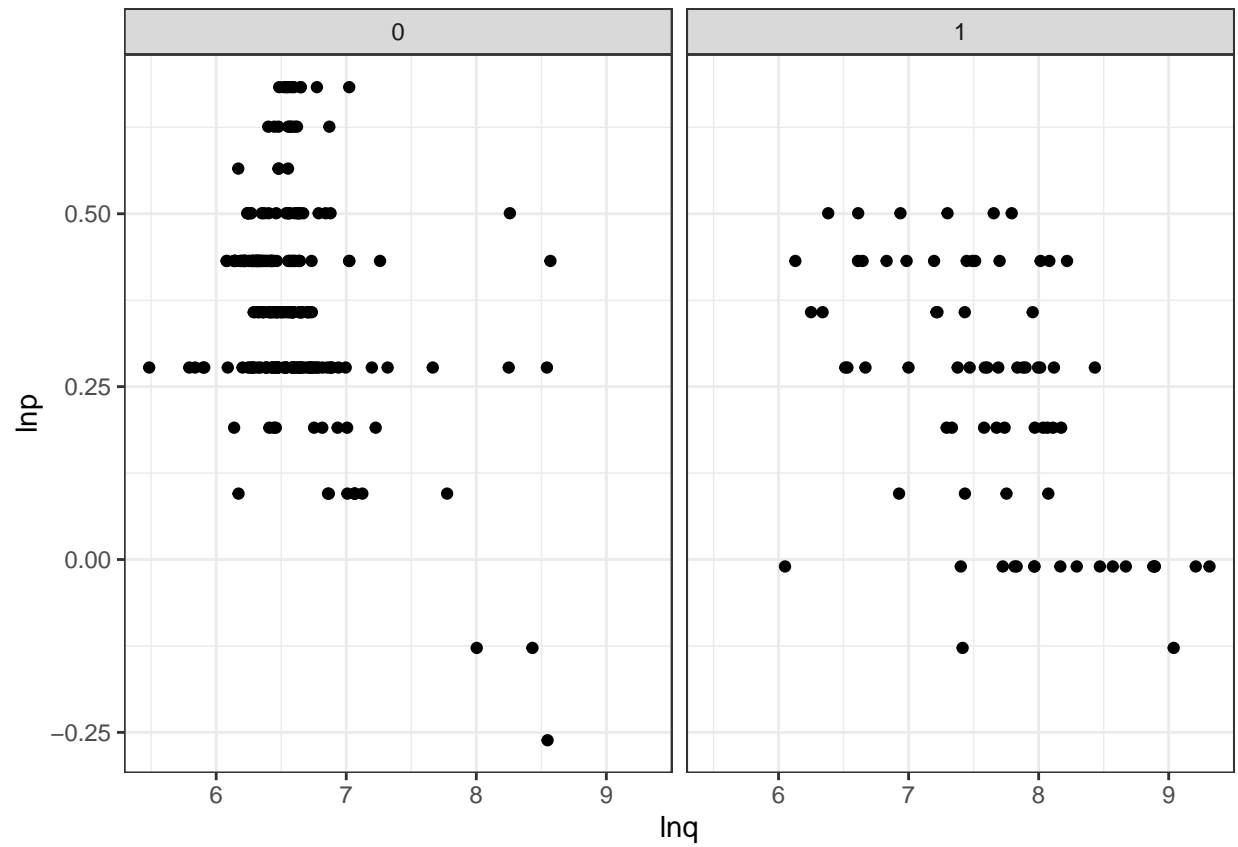
```
ggplot(df.low.mmaid12, aes(x=lnq, y=lnp)) +  
  geom_point()
```



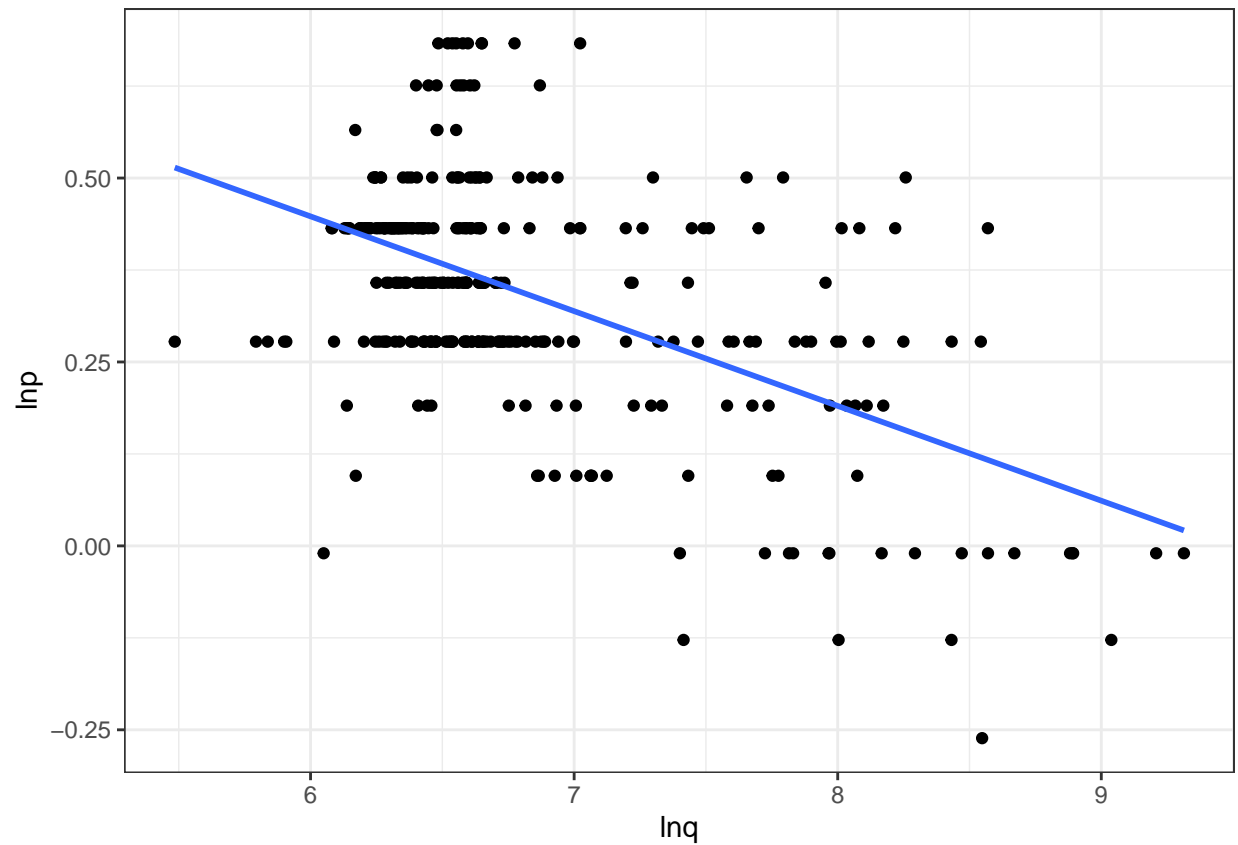
```
ggplot(df.low.mmaid12, aes(x=lnq, y=lnp, color=Dmerch)) +  
  geom_point()
```



```
ggplot(df.low.mmaid12, aes(x=lnq, y=lnp)) +
  geom_point() +
  facet_wrap(~ Dmerch)
```

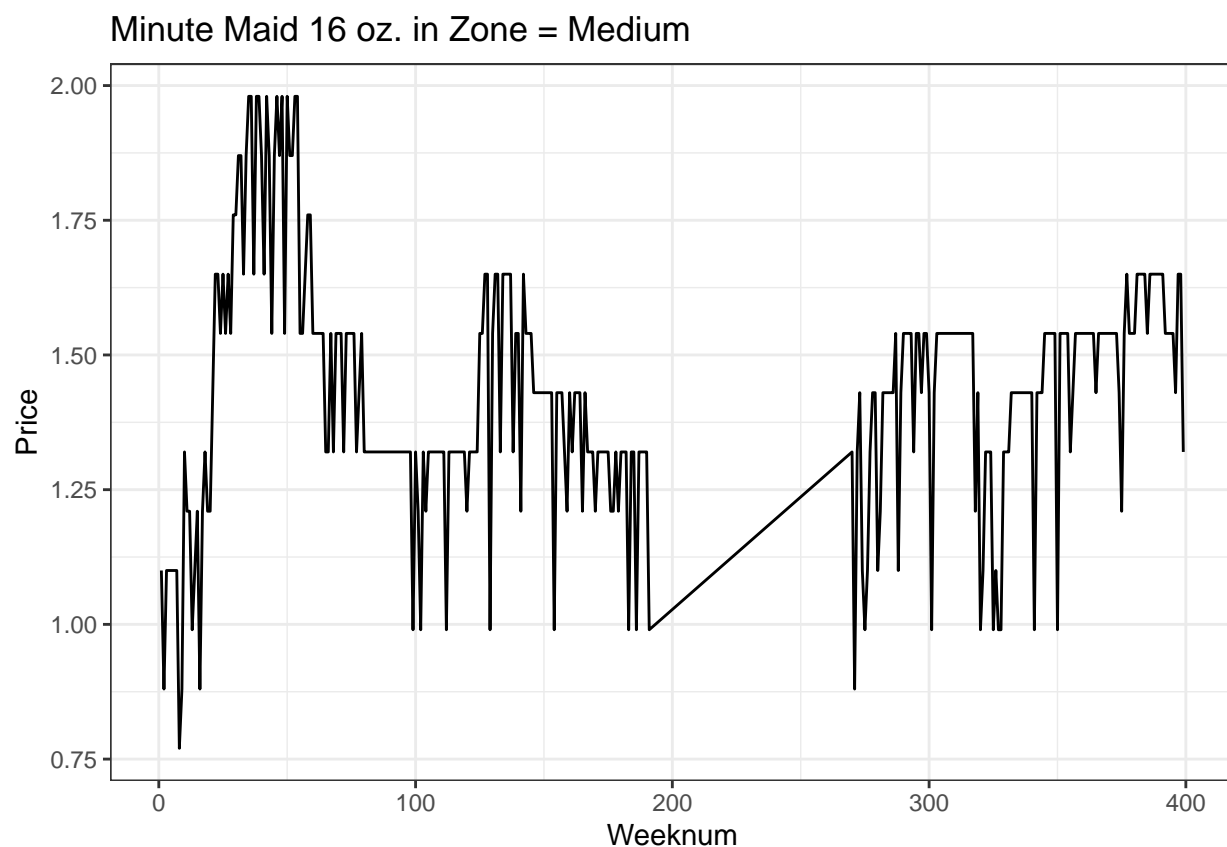


```
ggplot(df.low.mmaid12, aes(x=lnq, y=lnp)) +
  geom_point() +
  geom_smooth(method='lm', se=FALSE)
```



## 3.2 Line plots

```
ggplot(df.low.mmaid12, aes(x=week, y=price)) +
  geom_line() +
  labs(x="Weeknum", y="Price", title="Minute Maid 16 oz. in Zone = Medium")
```



## 4 Regression

```
reg1 <- lm(lnq ~ lnp, data=df.low.mmaid12)
summary(reg1)
```

Call:

```
lm(formula = lnq ~ lnp, data = df.low.mmaid12)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.5579	-0.3608	-0.1922	0.3593	1.9544

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.58509	0.07504	101.08	<2e-16 ***
lnp	-2.24576	0.19768	-11.36	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



Residual standard error: 0.5747 on 317 degrees of freedom  
Multiple R-squared: 0.2893, Adjusted R-squared: 0.2871  
F-statistic: 129.1 on 1 and 317 DF, p-value: < 2.2e-16

The regression indicates that the price elasticity is -2.246. Now I will add zone dummies and an interaction between `lnp` and `Dmerch`.

```
reg2 <- lm(lnq ~ lnp + Dzone + lnp:Dmerch, data=df.mmaid12)
summary(reg2)
```

Call:

```
lm(formula = lnq ~ lnp + Dzone + lnp:Dmerch, data = df.mmaid12)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.8673	-0.2420	-0.0562	0.2439	2.4228

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.68063	0.04651	165.153	< 2e-16 ***
lnp	-2.34245	0.09992	-23.443	< 2e-16 ***
DzoneHigh	1.02904	0.04722	21.794	< 2e-16 ***
DzoneLow	-0.15736	0.04637	-3.394	0.000711 ***
DzoneMedium	1.26856	0.04665	27.193	< 2e-16 ***
lnp:Dmerch1	1.77531	0.12187	14.568	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5857 on 1273 degrees of freedom  
Multiple R-squared: 0.5974, Adjusted R-squared: 0.5958  
F-statistic: 377.8 on 5 and 1273 DF, p-value: < 2.2e-16

Now the price elasticity is -2.342, which is pretty similar to the first regression. Both estimates are within the reasonable range for a price elasticity.