Topic 1:

Sales data and regression models

Takeaki Sunada[1]

[1]Simon Business School

University of Rochester

Utilizing sales data to estimate demand

- Now that our objective is to estimate demand (entire P-Q relationship) with various kinds of data.
- Let's start from one most commonly used data type - "sales data". Virtually every firm records the history of product sales and prices for accounting purposes.
- To simplify, suppose your firm offers only one product at a single price.

A typical structure of sales data

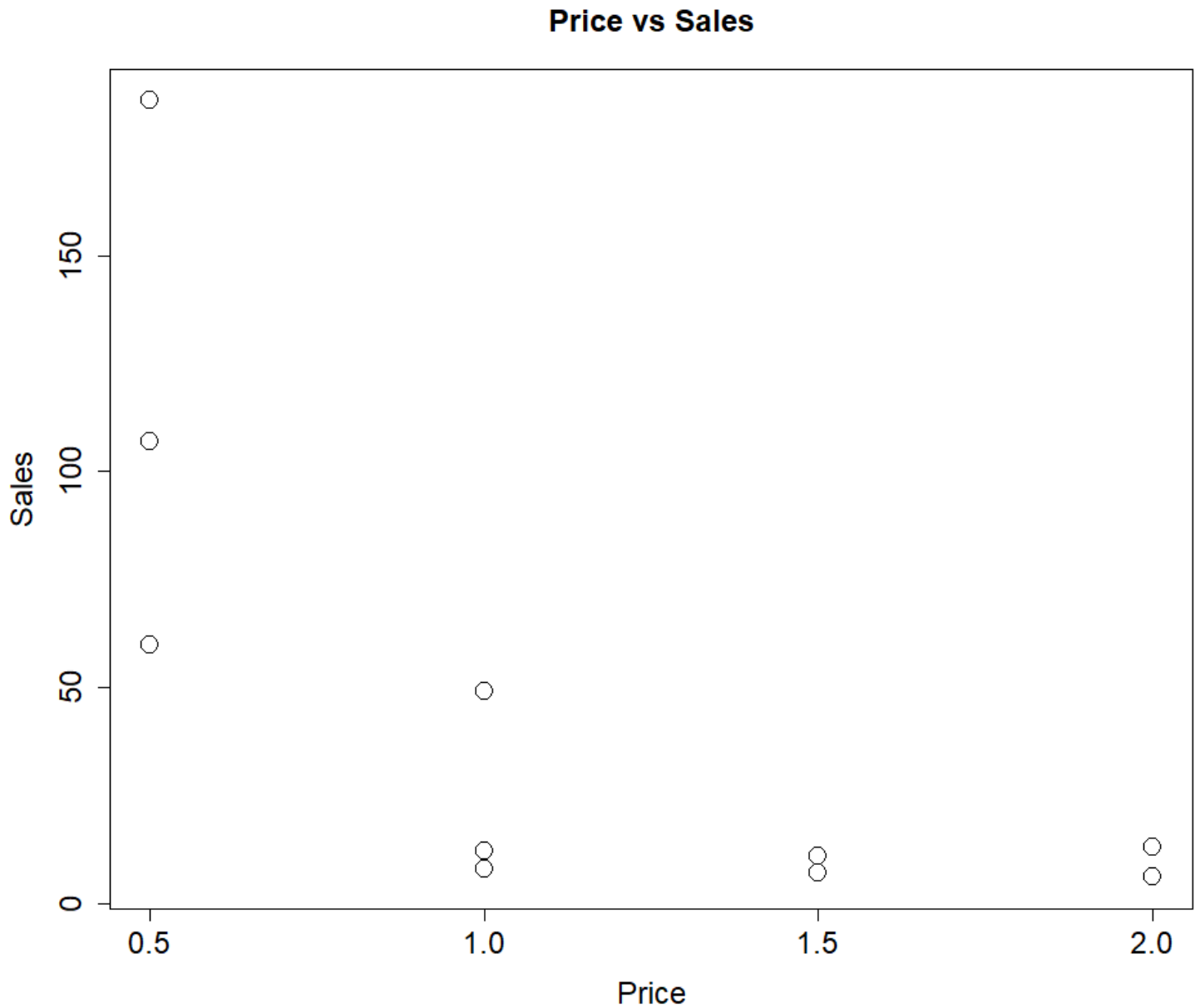
	A	B	C
1	month	sales	price
2	1	11	1.5
3	2	8	1
4	3	107	0.5
5	4	7	1.5
6	5	13	2
7	6	186	0.5
8	7	12	1
9	8	60	0.5
10	9	6	2
11	10	49	1

simulate_example_data

Cross-sectional, time-series and panel data

- If a data comes from one market (e.g. entire U.S) across multiple time periods (e.g. monthly sales), we call it "time-series data".
- If a data comes from multiple markets (e.g. state by state) but just one time period (e.g. 2016 yearly sales only), we call it "cross-sectional data".
- If a data comes from multiple markets across multiple time periods, we call it "panel data".
- For now no need to distinguish between them. However some tools we study later require a panel data.

Descriptive analysis



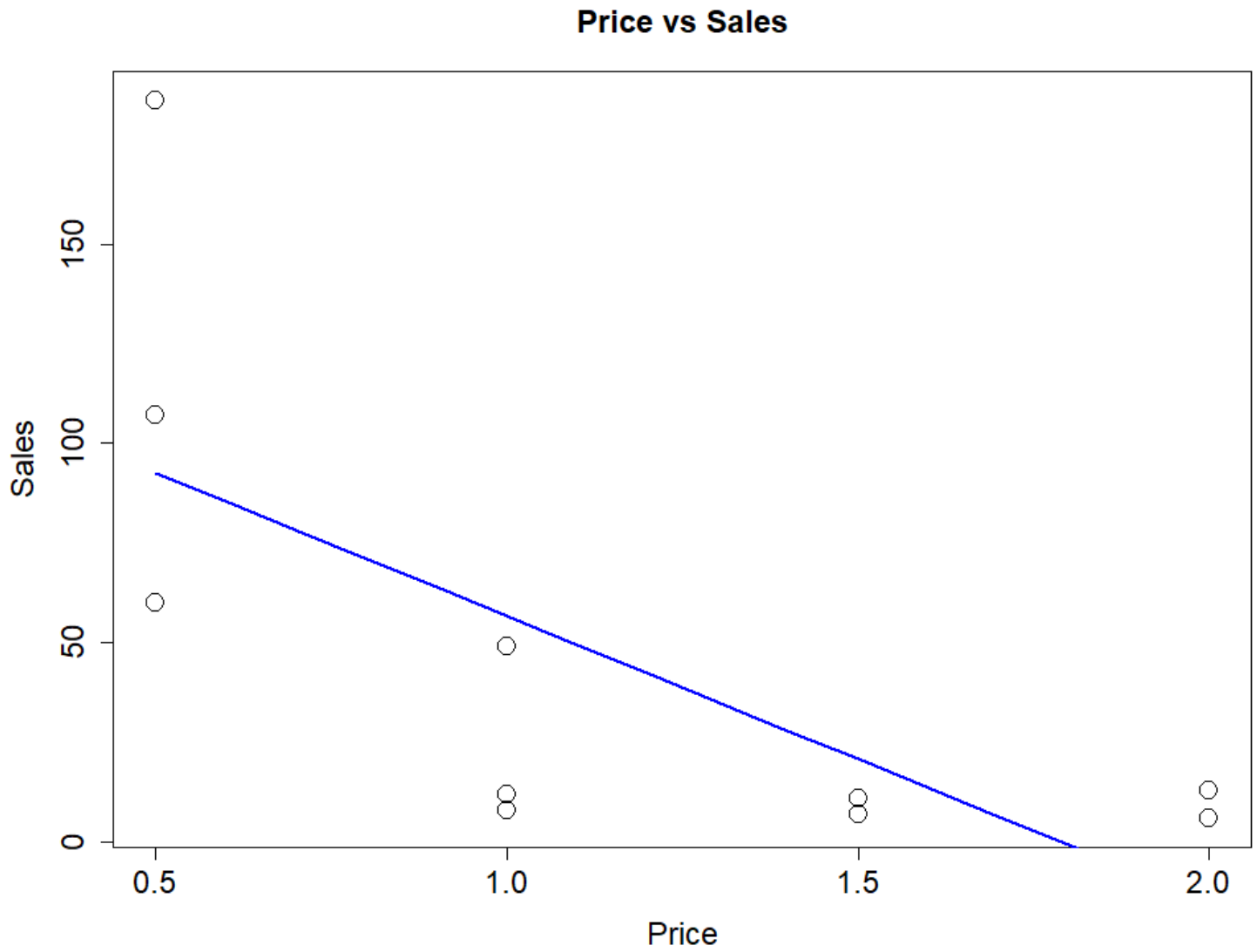
Regression models

- In this course, we use regression models to study sales data.
- Why regression? Recall our model selection criteria: the data richness.
  - Estimating a complicated (flexible) model requires rich data - data needs to be informative enough.
  - If data is crappy, the model needs to be simple (Note that "simple model" = "more restrictive" = "suitable for less informative data"!).
- Regression models are particularly suitable here - results easily interpretable, mechanism very straightforward, and easy to code.

"lm" for regression in R

```
#Example 1
#Read data and run a regression
#Use "fread" function from "data.table" package
simdata=fread("simulate_example_data.csv",stringsAsFactors = F)

reg=lm(sales~price,data=simdata)
```

Regression line



A quick review of regression models

- I assume we already know regression models. But let's run a quick refresher:
- Suppose we run the following OLS.

$$Q = \beta_0 + \beta_1 P + \epsilon,$$

  - What is the interpretation of β_0?
  - What is the interpretation of β_1?
  - What does ϵ represent?

A quick review of regression models

- Formally, the regression line corresponds to "the best linear predictor of Q given P".
- In other words, the regression line represents our best guess (more formally, expectation) of Q given P, assuming that Q is linear in P.

$$E(Q \mid P) = \beta_0 + \beta_1 P.$$

- $E(Q \mid P)$ denotes "expectation of Q given P".

A quick review of regression models

- Then β_0 corresponds to "the expected value (our best guess) of Q when $P = 0$".
- β_1 represents how Q differs *in expectation* at different values of P.

A quick review of regression models

- The regression line represents our expectation of Q, given the P we set.
- Of course at the P values that we see in the data, the predicted $E(Q \mid P)$ does not necessarily match with the actual Q.

$$\begin{aligned} Q &= \beta_0 + \beta_1 P + \epsilon, \\ &= E(Q \mid P) + \epsilon. \end{aligned}$$

- ϵ is the error, or the difference between our expectation and actual realization.

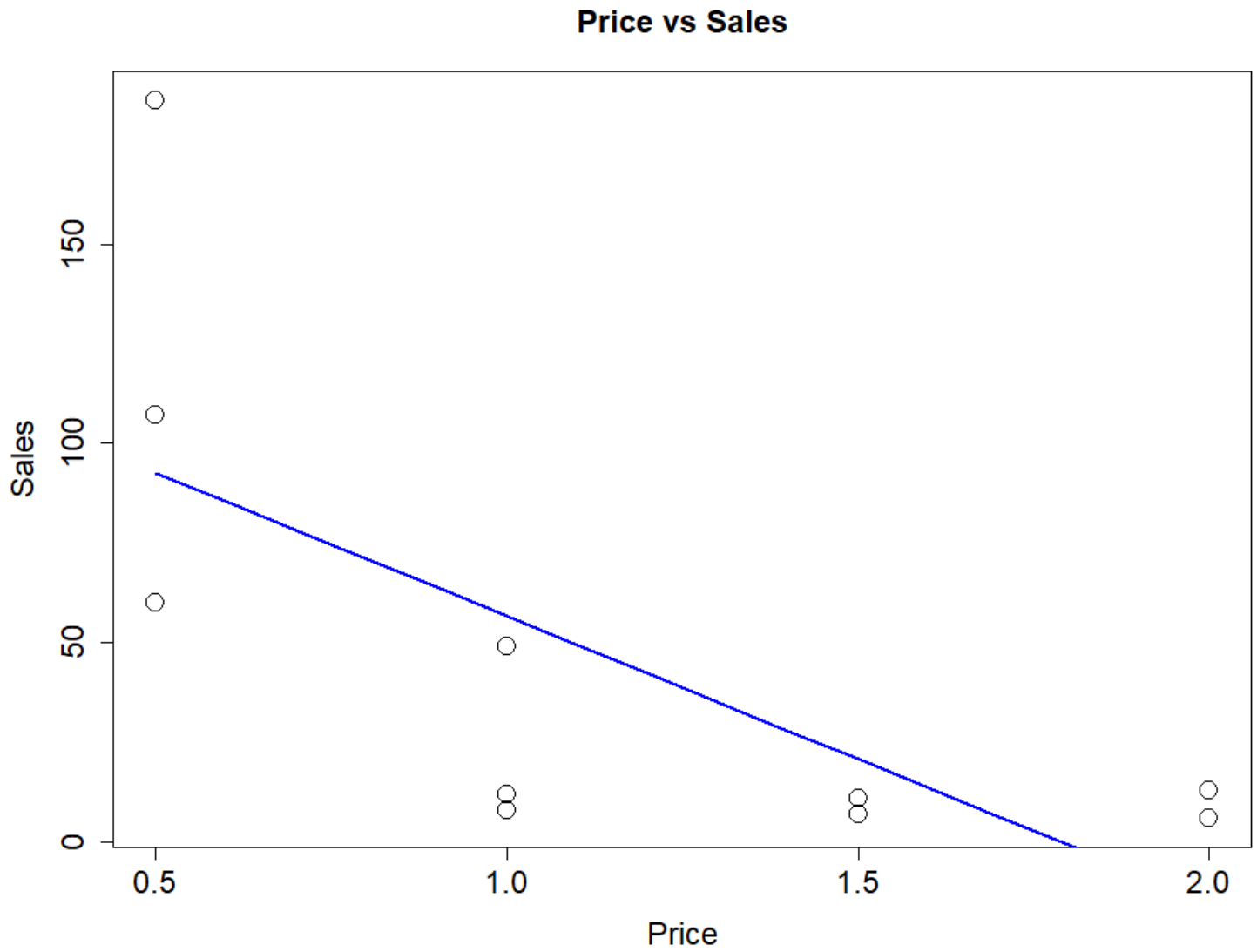
What we study about regression models

- "Regression in log"
- Causal analysis with regression models
- Estimating competition

What we study about regression models

- "Regression in log"
- Causal analysis with regression models
- Estimating competition

Regression line



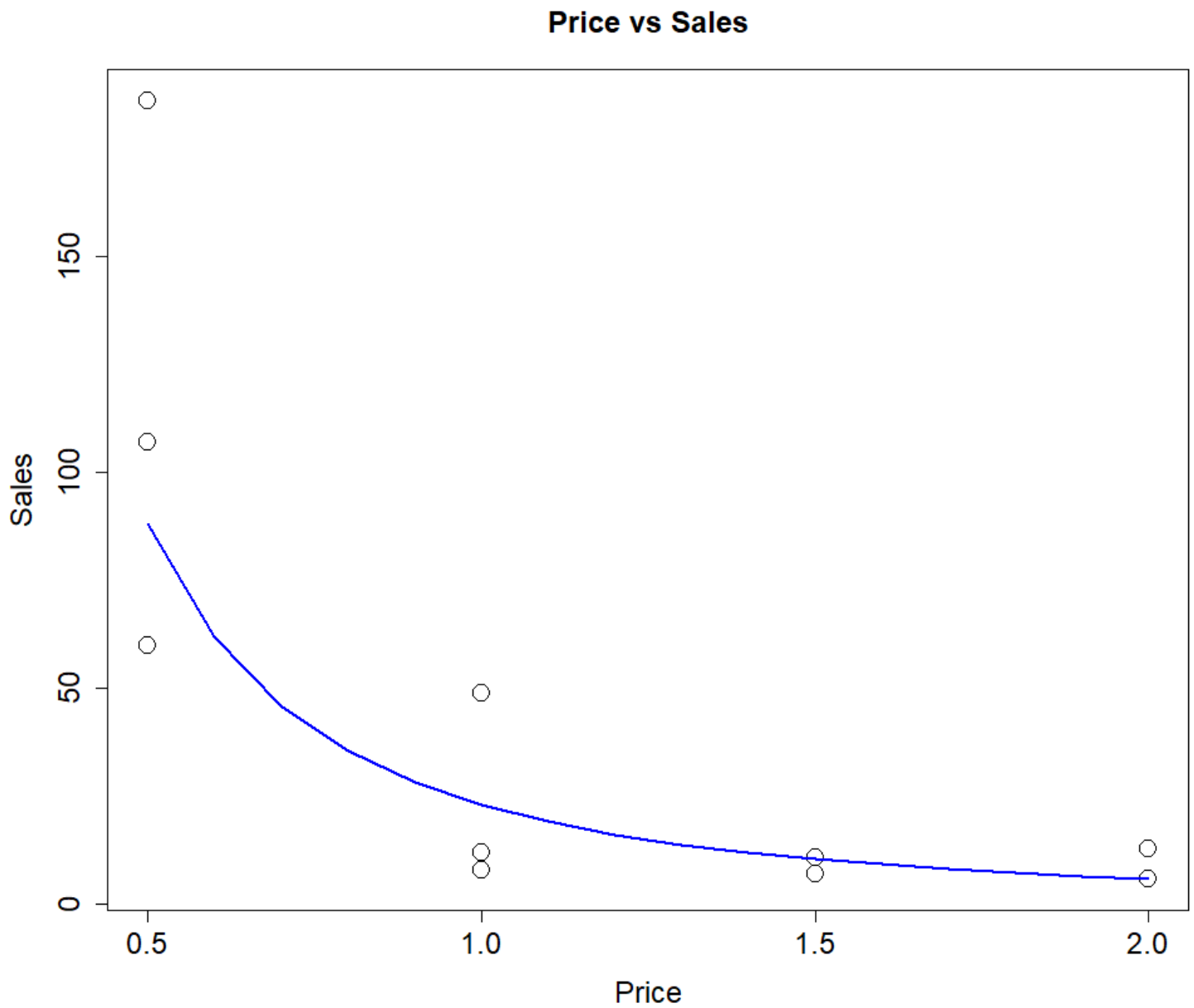
Regressing log-log

$$Q = \beta_0 + \beta_1 P + \epsilon.$$

- The fit doesn't look good at all.
- The scatter plot clearly indicates a nonlinear relationship. In practice, such nonlinear patterns are present in many industries.
- As a matter of fact, when we run a regression model with sales data, we often run "log-log" regression of the following form instead.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon.$$

Log-log regression line



Regressing log-log

- As log-log regression is linear in log, it is nonlinear in P-Q space.
- It fits significantly better (R^2 of linear model $= 0.4317$, that of log-log model $= 0.7062$). Again this tends to be the case in many environments.
- And there's more...

Regressing log-log

$$Q = \beta_0 + \beta_1 P + \epsilon.$$

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon.$$

- With a usual linear regression, β_1 represents "one unit (dollar) change in P would change Q by β_1 unit".
- If we do log-log, then β_1 represents "one percent change in P would change Q by β_1 percent": β_1 is the price elasticity of demand.
- This is because a small change in $log(P)$ is roughly equal to a small percentage change in P (and the same for Q).

Regressing log-log

- Note that by doing regression in log, we are imposing a completely different assumption.
- If we regress Q on P, the assumption is "the relationship between Q and P is linear".
- If we regress $\log(Q)$ on $\log(P)$, the assumption is "the relationship between $\log(Q)$ and $\log(P)$ is linear".
- Note log-log regression is not any more flexible than the original OLS!

What we study about regression models

- "Regression in log"
- Causal analysis with regression models
- Estimating competition

Predictive vs causal

- Causal analysis is *the most important element* when we use sales data to estimate demand for pricing purposes.
- *Causal* analysis contrasts with *predictive* analysis and the difference is as follows:
  - Predictive analysis measures *any* relationship *between* P and Q.
  - Causal analysis measures a *directional* relationship *from* P *to* Q.

Predictive vs causal

- Causal analysis is *the most important element* when we use sales data to estimate demand for pricing purposes.
- *Causal* analysis contrasts with *predictive* analysis and the difference is as follows:
  - Predictive analysis measures *any* relationship *between* P and Q.
  - Causal analysis measures a *directional* relationship *from* P *to* Q.

Predictive vs causal

- Suppose you find that a low P often comes with a high Q in the data.
- It could mean one of the three happened during the data period:
  1. Low P caused high Q
  2. High Q caused low P
  3. No direct relationship between P and Q - there exists some outside factor X, which affects both P and Q at the same time, so if you measure the relationship between Q and P ignoring X, it *seems* that they are correlated.

  ... or some combination of them.
- Predictive analysis does not require distinguishing between them. Regardless of (1) - (3), if we *see* low P, we can still *predict* high Q associated with it.

Predictive vs causal

- In causal analysis, we focus exclusively on the first effect.
- Consider adjusting prices. To see how profit changes, we need to know how Q will change *as a result of this*. The direction of the effect must be "from P to Q".
- If, what the regression captures is the reverse ("high Q causes low P"), then changing a price will NOT affect Q, because what the regression line tells you is "if you change Q, how it affects P".
- Clearly, *we need a causal analysis for the purpose of pricing*. Any price recommendations based on a predictive estimate will be off.

Predictive vs causal in a regression model

- Suppose that we run a log-log regression.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon.$$

- β_1 from this regression is predictive. To estimate β_1, we take all data variations in P and associate them with the variations of Q. But not all of them are "managers set P, and so the sales Q realizes" causal direction - for example:
  1. When a manager sees higher Q, they may set higher P.
  2. Some observations may come from rural regions (low Q low P), while others from NY city (high Q high P)
- The estimated β_1 = "true causal relationship + everything else".

Predictive vs causal in a regression model

- To get the right causal effect in β_1, we need to eliminate "everything else" part from our β_1.
- Because not-causal β_1 is due to not-causal part of the data, intuitively, we should be able to estimate a causal β_1 if we *only use causal variations in the data*.

How do we eliminate non-causal variations?

	A	B	C
1	month	sales	price
2	1	11	1.5
3	2	8	1
4	3	107	0.5
5	4	7	1.5
6	5	13	2
7	6	186	0.5
8	7	12	1
9	8	60	0.5
10	9	6	2
11	10	49	1

simulate_example_data

Causal β_1 = independence between P and ϵ

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon$$

- β_1 represents a causal relationship when P (or $\log(P)$ in this case) is independent of ϵ.
-
-

"Omitted factors" in ϵ causes causality issue

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon$$

- β_1 represents a causal relationship when P (or $\log(P)$ in this case) is independent of ϵ.
- In other words, non-causal effects contaminate β_1 when "ϵ and P move together (i.e., correlated)".
- Remember ϵ is "everything that is omitted in the regression" here - very likely, there *are* factors in the ϵ that are correlated with P.

Why "Omitted factors" in ϵ causes causality issue

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon$$

- Imagine demand seasonality - it affects Q (like observations from winter show low Q whereas summer shows high Q). Because it is not included as a variable on the right hand side, it is an omitted factor - hence in the ϵ.
- Likely firm adjusts pricing month to month if demand shows seasonality. i.e. P and ϵ move together.

Why "Omitted factors" in ϵ causes causality issue

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon$$

- Imagine estimating demand with such data. When high Q is observed, likely ϵ is high (because it's summer), and also P is high (because manager knows it's summer).
- We need to estimate β_1, which measures the effect of P on Q exclusively. In other words, of the variations in Q, we need to distinguish: how much comes from P vs how much comes from ϵ.
- This is *impossible*, because ϵ is unobservable to us.

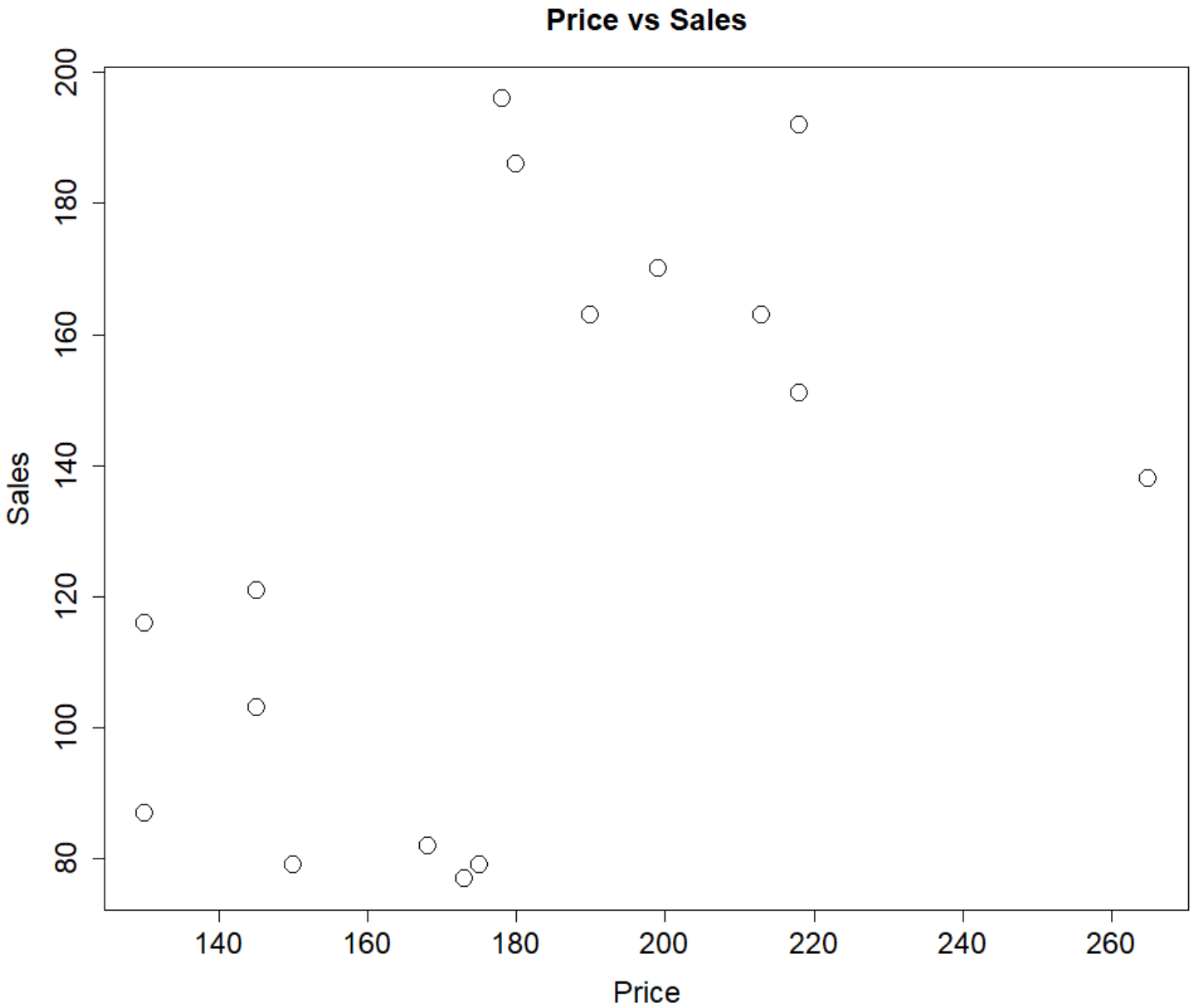
Why "Omitted factors" in ϵ causes causality issue

- However, if we do "lm" in R, we should get a number for β_1 off of it nevertheless - how come?
- Indeed, if both ϵ and P affects Q at the same time, we implicitly assume that *it all comes from* P and estimate β_1 - hence β_1 reflects both the effect of P and ϵ.
- Note the analogy to the statement before - the effect driven by ϵ is equivalent to the non-causal effect.
- Hence, estimating a causal β_1
  $=$ eliminating non-causal variation from P
  $=$ eliminating any omitted factor in ϵ that moves together with P.

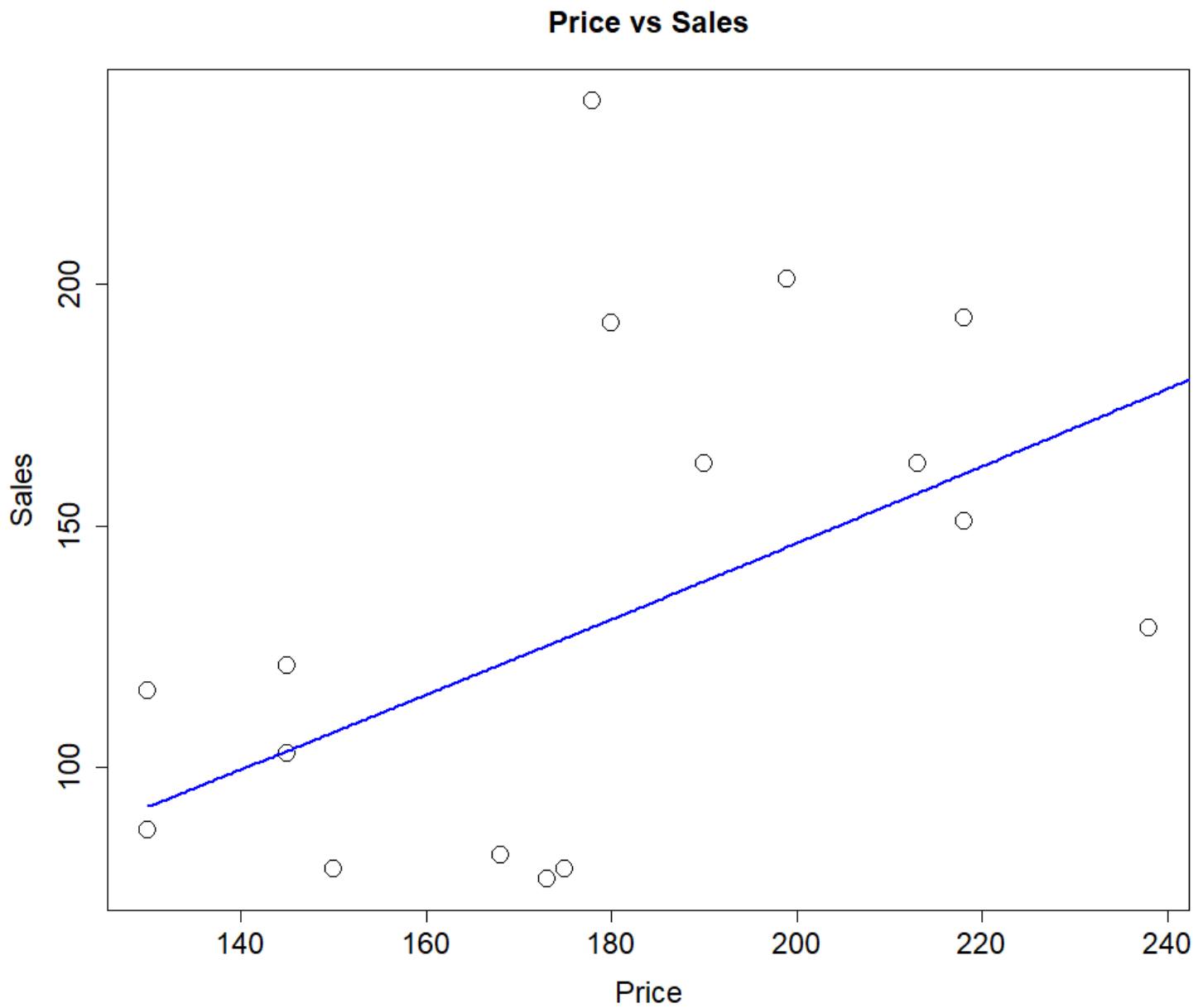
Example 1

- Suppose that you want to set prices for airline tickets. You have access to historical data of Q and P.
- Likely, during the summer period the travel demand spikes, and Q, the observed sales, also spikes. Airline managers know this, and they raise prices. So you tend to see high Q and high P.
- During winter months, due to low travel demand airlines have to cut prices, and not many people take the flight. You see low Q and low P.
- If you regress $\log(Q)$ on $\log(P)$, then...

Descriptive analysis



Regression result



Example 1 issues

- Clearly, the relationship does give you prediction.
  - If you *see* that the price is 220 dollars in the data, then the sales should be around 160 seats.
- However, the relationship cannot be causal. Higher price cannot cause the demand to increase.
  - Even if you *choose* the price at 220 dollars, the sales may not be around 160 seats.
- Rather, this is a spurious relationship - there exists a latent variable (seasonality) that causes both higher sales and higher demand, which is omitted from the regression.

Example 1 issues

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \epsilon.$$

- The omitted "seasonality" is then part of ϵ. Historical P was set accounting for demand seasonality. Hence, P and ϵ move together.
- β_1 then captures two opposing effects:
  1. P decreases Q (this is the causal effect we want to measure).
  2. Demand seasonality increases Q, and P increases in response to seasonality.

  If seasonality is omitted from the regression, the second effect dominates the first and we *predict* higher Q at higher P.

Example 1 solution

- In this example, we know what in ϵ causes contamination - seasonality (note in reality, we usually don't even know it!).
- One solution is include a variable that represents seasonality in the regression, thereby taking the seasonality out of ϵ.
- Suppose that the data are collected during February and August.
- We now run a regression including a dummy that is one if it is August.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_2 I\{August\} + \epsilon.$$

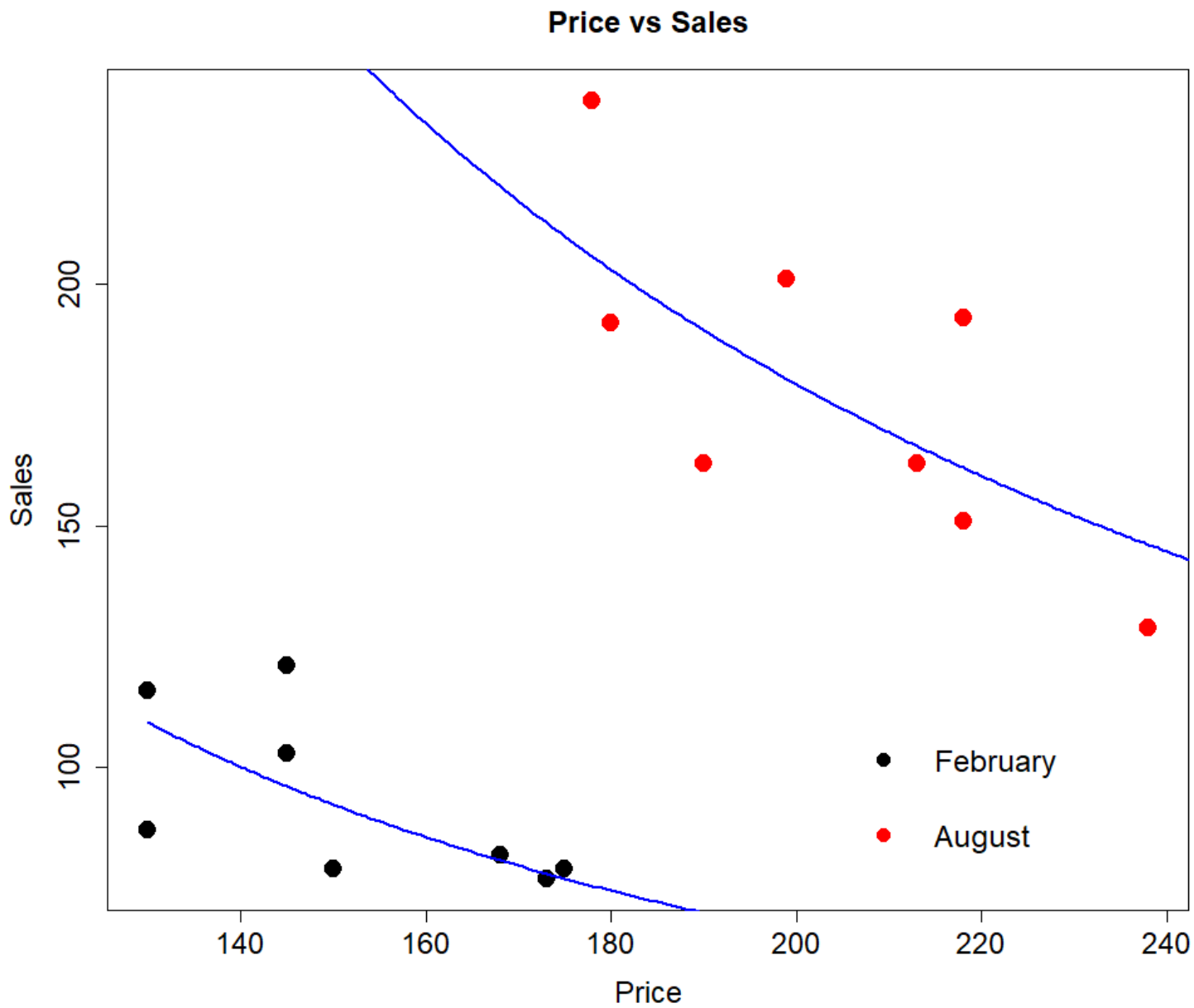
$I\{August\}$ is 1 if the observation is collected in August and zero if February.

Example 1 solution

- This takes "seasonality" out of ϵ. i.e. seasonality is no longer "omitted". Hence according to the definition, it eliminates the contamination of the estimate. But what is the mechanism?
- With $I\{August\}$ in the regression, β_1 is estimated by the variation of $\log(P)$ and $\log(Q)$ *within* February and August only. In other words, we don't use variations of P and Q *across* months to estimate β_1.
- In other words, we estimate one $\log(P)$ - $\log(Q)$ relationship for February, and the other relationship for August. The β_1 reported on R is (sort of) the average of the two.

Including month dummy

Regression result



Example 1 solution

- A general lesson: including a variable X in a regression $=$ to estimate β_1, we don't compare P and Q *across* different values of X anymore. We only use variation of P and Q *within* a given value of X.
- Another way to put it: by including X, we are not using certain (non-causal) variation of P to estimate β_1. See the analogy to the statement we made earlier?

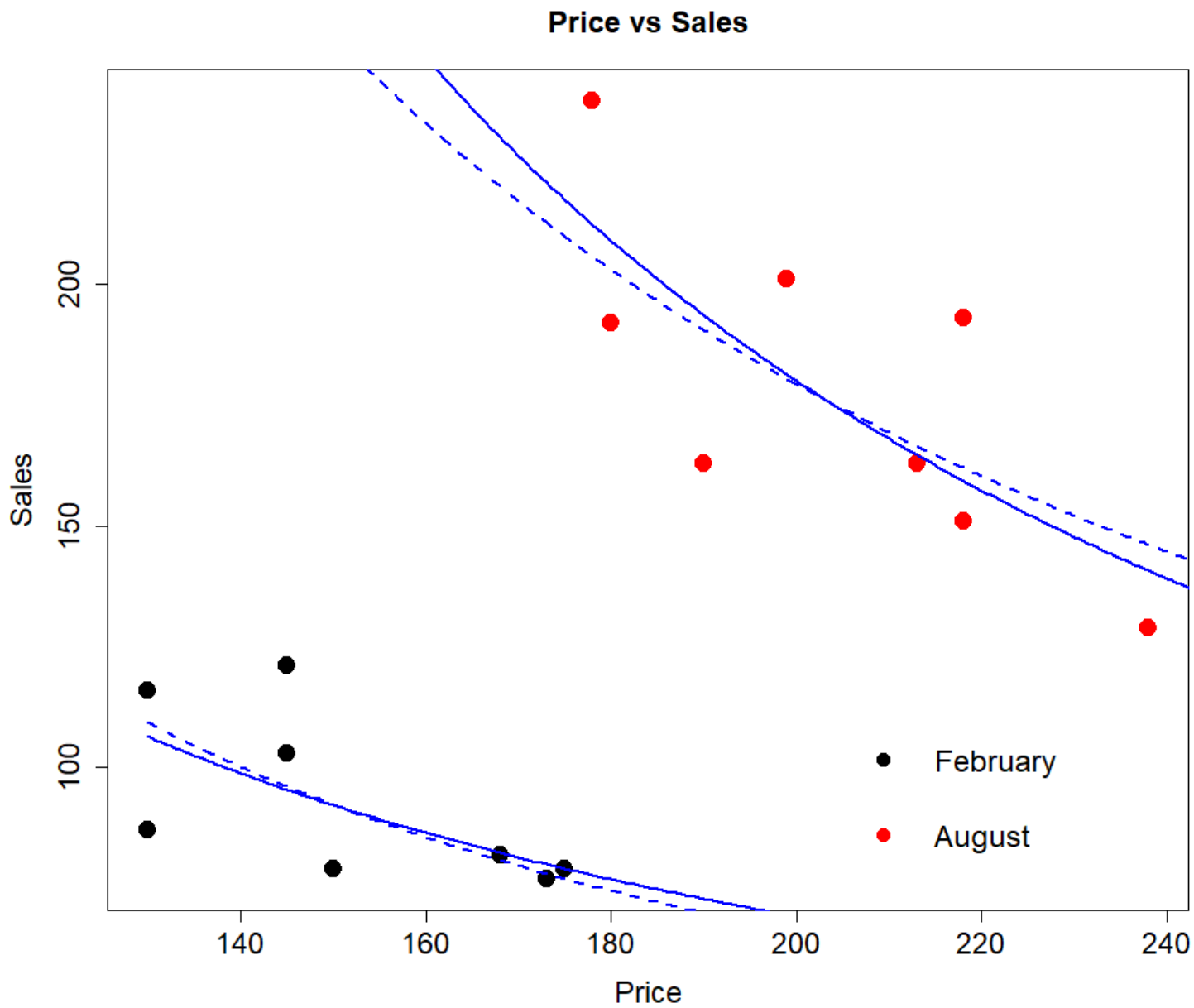
Optional - Example 1 extension

- The previous expression assumes that seasonality only impacts demand intercept, and elasticity remains the same between February and August. We may also want to relax that assumption.
- Let's interact the August dummy with log price.

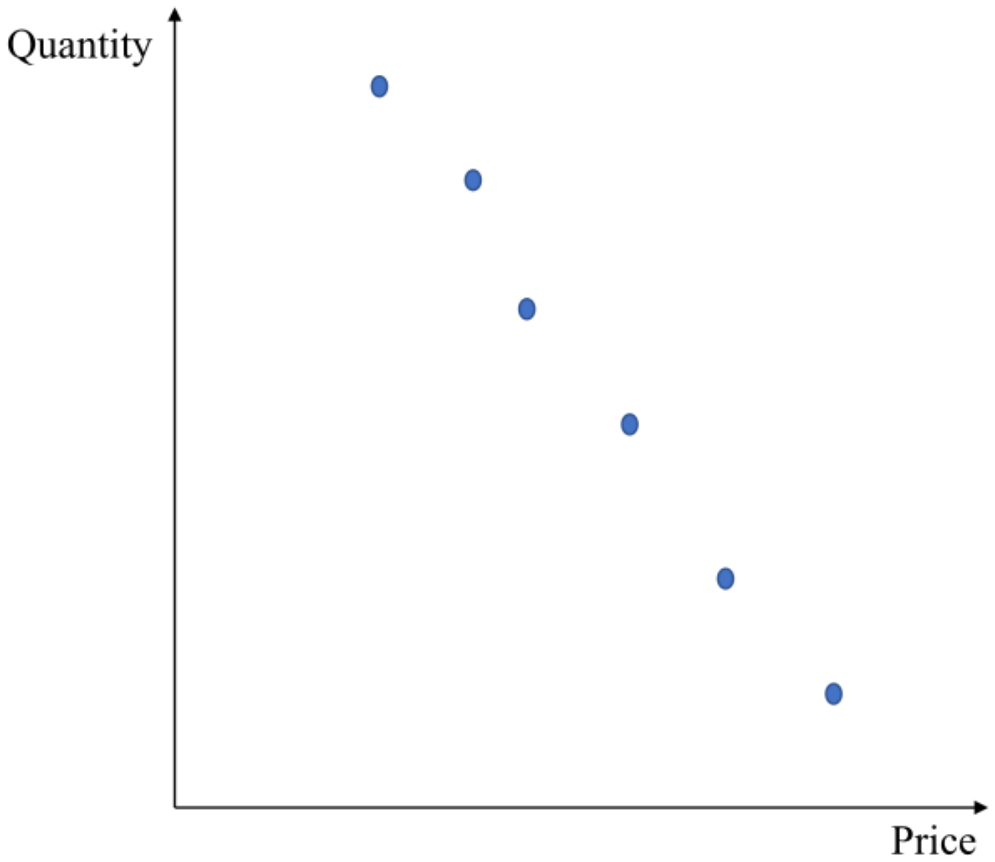
$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_2 I\{August\} + \beta_3 \log(P) I\{August\} + \epsilon.$$

- Then β_1 is the elasticity in February, and $\beta_1 + \beta_3$ is the elasticity in August.

Optional - Regression result

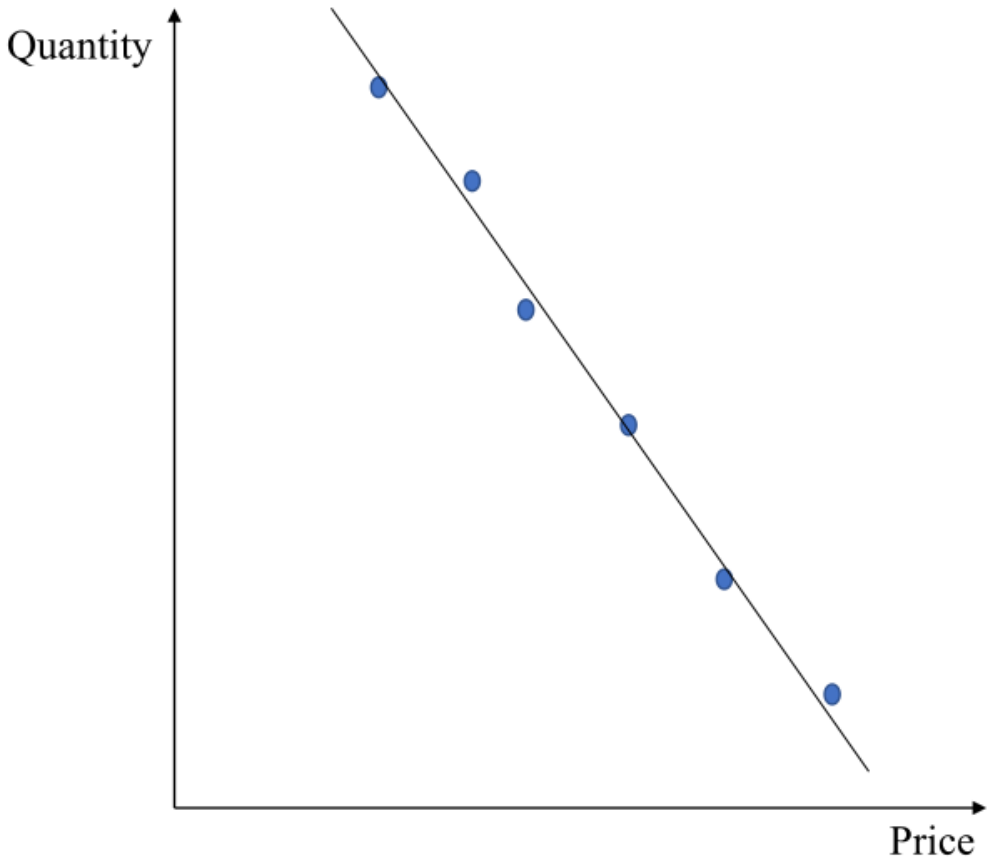


Example 2



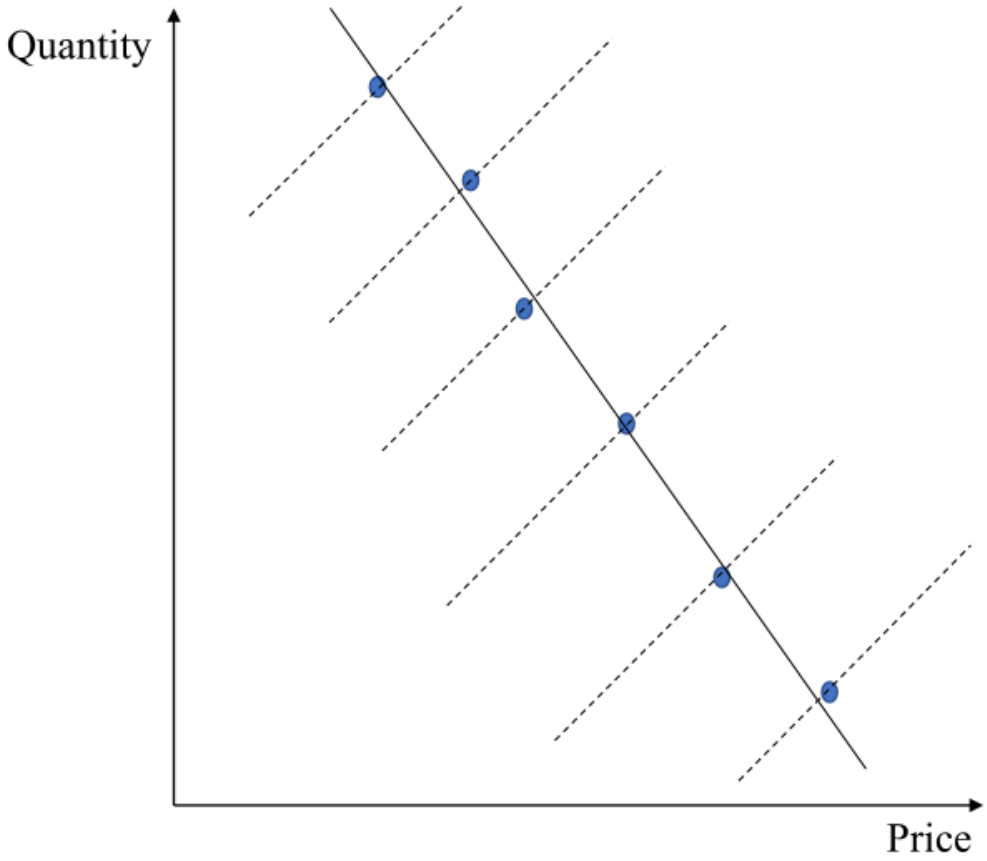
- Suppose that you sell some retail-packaged product, and observe these price-quantity combinations in the data.

Example 2



- Running regression gives you great fit - happy with that?
- Think about *why* we see different prices in the data.

Example 2

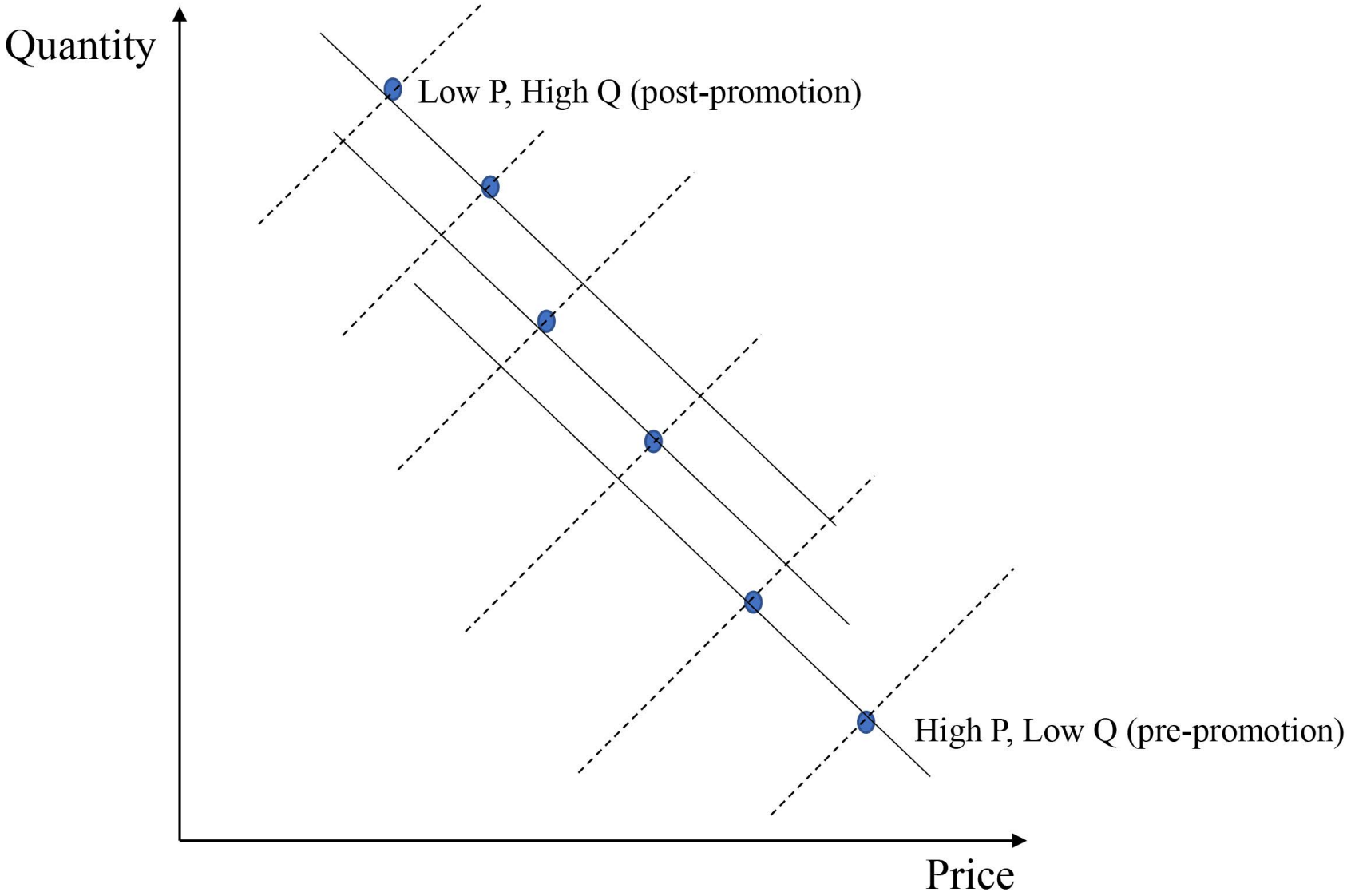


- By running a regression, we are assuming that the firm somehow changes prices, despite facing a fixed demand (Q-P relationship) and the sales realize according to that fixed demand.

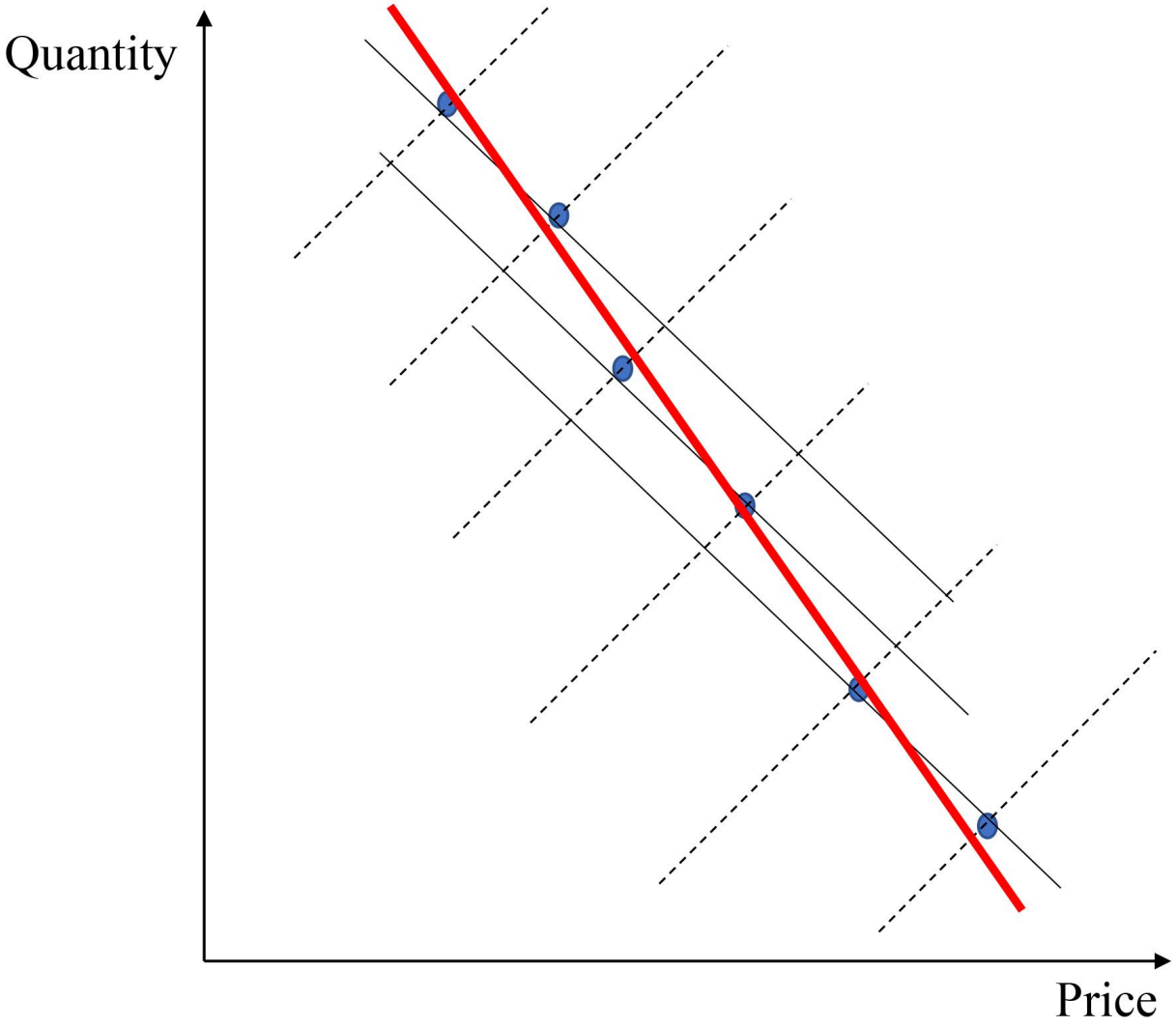
Example 2

- What if, in reality, the firm was running an advertising promotion during the data period, and lower prices were part of that promotion?
- Then:
  1. the ads boost the demand (i.e. increases Q for a given P - changes the Q-P relationship),
  2. and lower P would boost Q even further (i.e. increases Q because of a lower P, for a given Q-P relationship).

Example 2


Quantity
Low P, High Q (post-promotion)
High P, Low Q (pre-promotion)
Price

Example 2



Example 2

- In this case, β_1 from the simple regression is not causal - the effect of ad is omitted from the regression, and hence is reflected in ϵ. β_1 reflects both the causal effect of price (change in P) and also the effect of ads (change in ϵ).
- To resolve this, we may include some measure of ad effect (e.g. ad expenditure) in the regression.
- Then we compare different values of P *within a fixed value of ad*. We don't compare P across different values of ad. We eliminate contamination.

A general lesson from the two examples

- Suppose any demand-side factor (seasonality, advertisement, etc.) that affects the firm's pricing decision during the data period is omitted from the regression. This causes correlation between P and ϵ and thus non-causal β_1.
- Put differently, if you suspect that there's any demand-side factor that affects the pricing decisions, you may collect data about it and include it as a control variable (X) in the regression. It is then out of ϵ and does not cause contamination.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_X X + \epsilon.$$

Note: Causal β_1 vs improving model fit

- Seemingly, adding X variables in the regression is something you've done before.
- Note that our objective is to establish the causal β_1. Our current objective is NOT to improve model fit (i.e. R^2).
- We are concerned about correlation between P and ϵ, whereas higher model fit means smaller ϵ, relative to the explained part of the model ($\beta_0 + \beta_1 \log(P)$). These two are different things.
- Imagine a model that fits perfect with the data, but β_1 is not causal. It's a fantastic model for predictive purposes, but is completely useless for pricing analytics.

A general lesson from the two examples

- Given that our goal is to establish causal β_1, what demand-side factors to include as an X? A short answer: we are never sure.
  - ϵ is not observable - unlike the examples before, we generally don't know what's in there.
  - Unlike the case of model fit (R^2), there's no metric that you can rely on.
- Hence, the decision of what to include in X is based on our guesswork. If you are more familiar with the industry, you can include "more right" X's and can get a better causal β_1.

Side: Project 1

- Of course you are not an expert of a car market. What I expect from you is to choose a set of X with your guesswork, and try to justify your specification as much as possible.
- Your reasoning can arise from anything: your pure guesswork, our commonsense, etc. In fact, there may be things in the data we can use as a guide in choosing which variable to include too (e.g., how does the estimated coefficient change as we add variables?).

Establishing causation

- Often, obtaining a proxy (X) for all the latent factors is just not possible. You may just need to control for lots of things: firm's marketing activities, product attributes, measure of competitiveness, measure of consumer tastes, etc.
- If you have panel data (data from multiple markets and across time), there's a shortcut...

Fixed effects

- Suppose that you have access to panel data of sales from the U.S. (e.g. sales in each state at each month).
- Different geographical locations likely have different underlying demand and hence the firm has been offering different prices. You may want to control for latent demand-side factors.
- However, what exactly varies across states? Different consumption habits? Different underlying demographics (hence different consumer preference)? Obtaining data for all those seems just impossible.

Location fixed effects

- In such cases, including dummies for each state may help you out.
- Suppose that you have data of Q and P from NY, TX, and CA.
- Consider running a following regression.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{TX} I\{TX\} + \beta_{CA} I\{CA\} + \epsilon,$$

where $I\{State\}$ is a dummy, which is one if the observation is from that state, and otherwise zero.

Location fixed effects

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{TX} I\{TX\} + \beta_{CA} I\{CA\} + \epsilon,$$

- Then β_1 is estimated based on *within-state* variation of P and Q. We don't compare NY observations with TX observations. Rather, we calculate one relationship, each from NY, TX and CA. Estimated β_1 is (again, sort of) the average of the three.
- Those dummy variables hence let us not compare Q and P across states - difference of unobserved factors across states would not contaminate our β_1. Note that we don't even need to specify what causes the differences.

Location fixed effects

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{TX} I\{TX\} + \beta_{CA} I\{CA\} + \epsilon,$$

- This approach works with panel data, because there we have multiple dimensions of data. Even if we throw away across-state variations of data, we still have within-state across-time variations to estimate β_1.
- Such a series of dummy variables we insert to kill a dimension of data variation is called "fixed effects".

Time fixed effects

- Sometimes you may believe that contamination arises from over-time variations (seasonality) rather than geographical differences.
- We can include "time fixed effect" (e.g. dummies for each month) that varies across time but constant across locations.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{Feb} I\{Feb\} + \beta_{Mar} I\{Mar\} + ... + \beta_{Dec} I\{Dec\} + \epsilon.$$

- Then across-time variations are eliminated from estimating β_1. We are only using P-Q variations across locations.

Location and time fixed effects

- In fact, we could include both location and time fixed effects if we suspect that both geographical and seasonal demand variations cause causality issues.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{TX} I\{TX\} + \beta_{CA} I\{CA\} + \beta_{Feb} I\{Feb\} + \beta_{Mar} I\{Mar\} + ... + \beta_{Dec} I\{Dec\} + \epsilon.$$

- When we include multiple FEs, keep in mind that each of them takes some variations away from estimating β_1.

Interactive fixed effects

- Sometimes, we may include interactive fixed effects - if we suspect that there's some specific demand shock in "NY in July", we include a dummy that is one for observations from NY in July.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_{NYJ} I\{NYinJul\} + \beta_{TX} I\{TX\} + \beta_{CA} I\{CA\} + \beta_{Feb} I\{Feb\} + \beta_{Mar} I\{Mar\} + ... + \beta_{Dec} I\{Dec\} + \epsilon.$$

- Then we don't compare "NY in July" observations with others to estimate β_1.

Too many fixed effects

- Of course, too many fixed effects aren't good either.
- Recall that inserting fixed effects is equivalent to taking away some of the price variations from estimating β_1 - if we have too many FEs, there's simply no variation of P left to estimate β_1.
- For now, a good rule of thumb is when R gives you errors as you add FEs, you have probably added too many.
- Punchline - if we have panel data and don't know what to include as a control variable, fixed effects are a good starting point.

Instrumental variables

- The approach we pursued so far is to include all the demand factors explicitly as X and control for them. Hence we have less things left in ϵ and hence less issues. This is "do something with ϵ" approach.
- Alternatively, in some cases we may have variations in P that are absolutely independent of any possible contaminants in ϵ - i.e. changes in prices that have nothing to do with any demand shift. e.g. cost shocks, institutional changes (e.g. consumption tax),
- If we can run a regression only using such price variations independent of ϵ, the estimated β_1 is the true causal effect of price on the sales. This is "do something with P" approach.

Instrumental variables

	A	B	C
1	month	sales	price
2	1	11	1.5
3	2	8	1
4	3	107	0.5
5	4	7	1.5
6	5	13	2
7	6	186	0.5
8	7	12	1
9	8	60	0.5
10	9	6	2
11	10	49	1

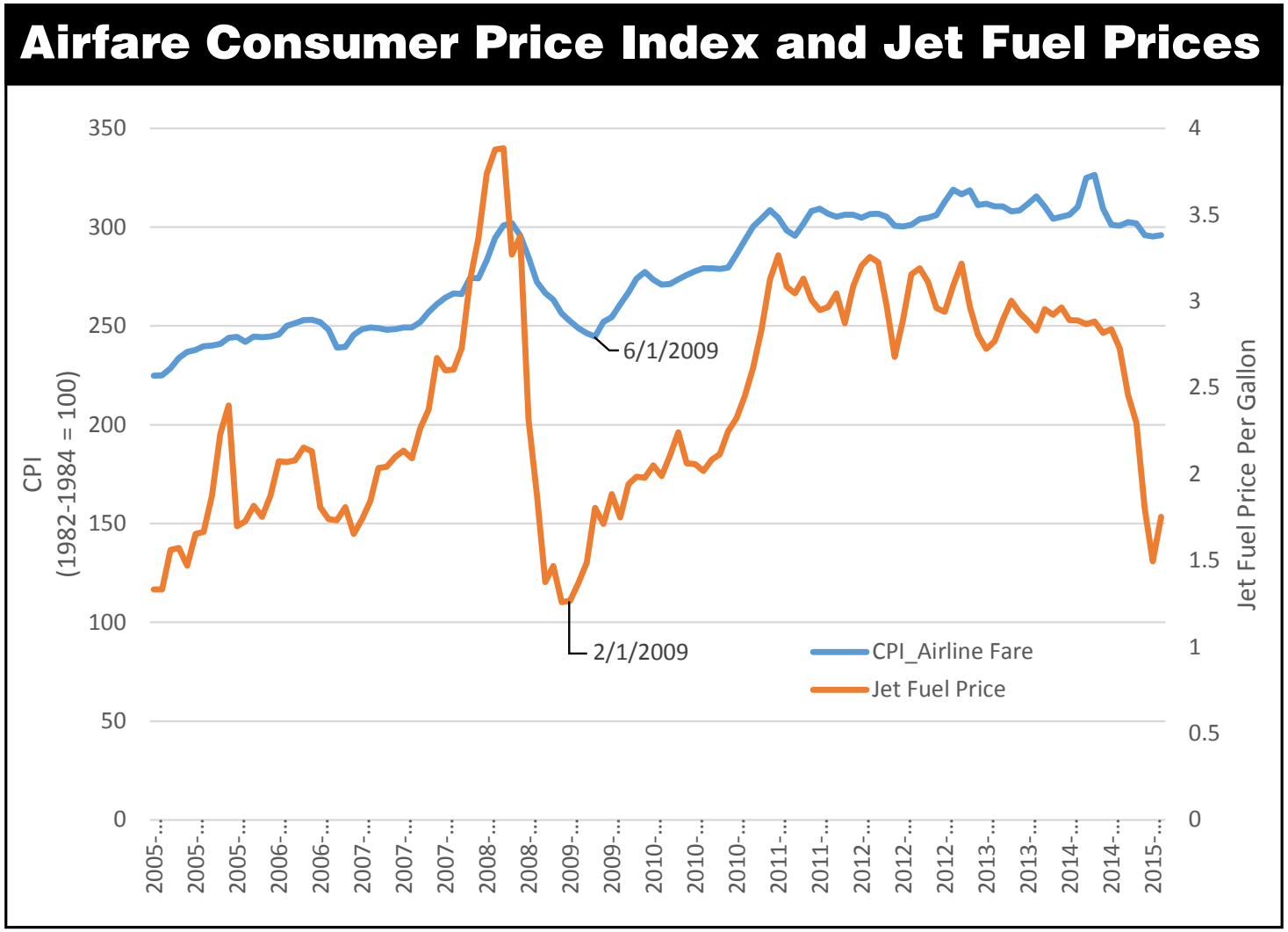
simulate_example_data

- In the data, you would probably see a lot of past price changes. Is there any way to find if each of them is due to demand shift (hence invalid), or due to other exogenous factors (hence valid)?

Instrumental variables

- If there's any price changes independent of ϵ, there must be a cause for it - suppose that you observe some proxy for that "cause".
  - Some P change caused by change in input prices - you may observe input prices in the data.
  - P may vary across observations because of institutional regulatory differences (e.g. different consumption tax across states) - you of course observe consumption tax at each state.
- Then we can apply "Instrumental variable regression".
- Consider airline industry. It is well known that airfare jumps around due to the change in fuel prices (cost-side factors).

Airfare and cost



[1]

[1] Source: NDSU Agriculture communication

Instrumental variables

- Some of the price variations are demand-driven (hence invalid), but others are cost-driven (valid).
- Instrumental variable regression (IV regression) is based on the following idea: "use price variation that is highly correlated with the cost to estimate β_1 - those are the ones likely due to cost change and not to demand change".
- If cost changes are not related to demand, cost-driven price changes can be used to establish the causal relationship.
- The cost variable in this context is called "instrumental variable (IV)".

Instrumental variables

- If price jumps likely take place at the same time as the cost jumps, then we conclude that price is affected more by costs. If price jumps and cost jumps occur at different time, then price change may not be attributable to cost changes.
- IV regression re-weights the price, such that in estimating β_1, we put more weights on changes in P that are highly correlated with cost changes, and less weights on the ones less correlated with cost changes.
- Hence as a whole, β_1 is estimated with "more clean" variation of prices, less contaminated by demand-side factors, while still using all the data.

Data we use

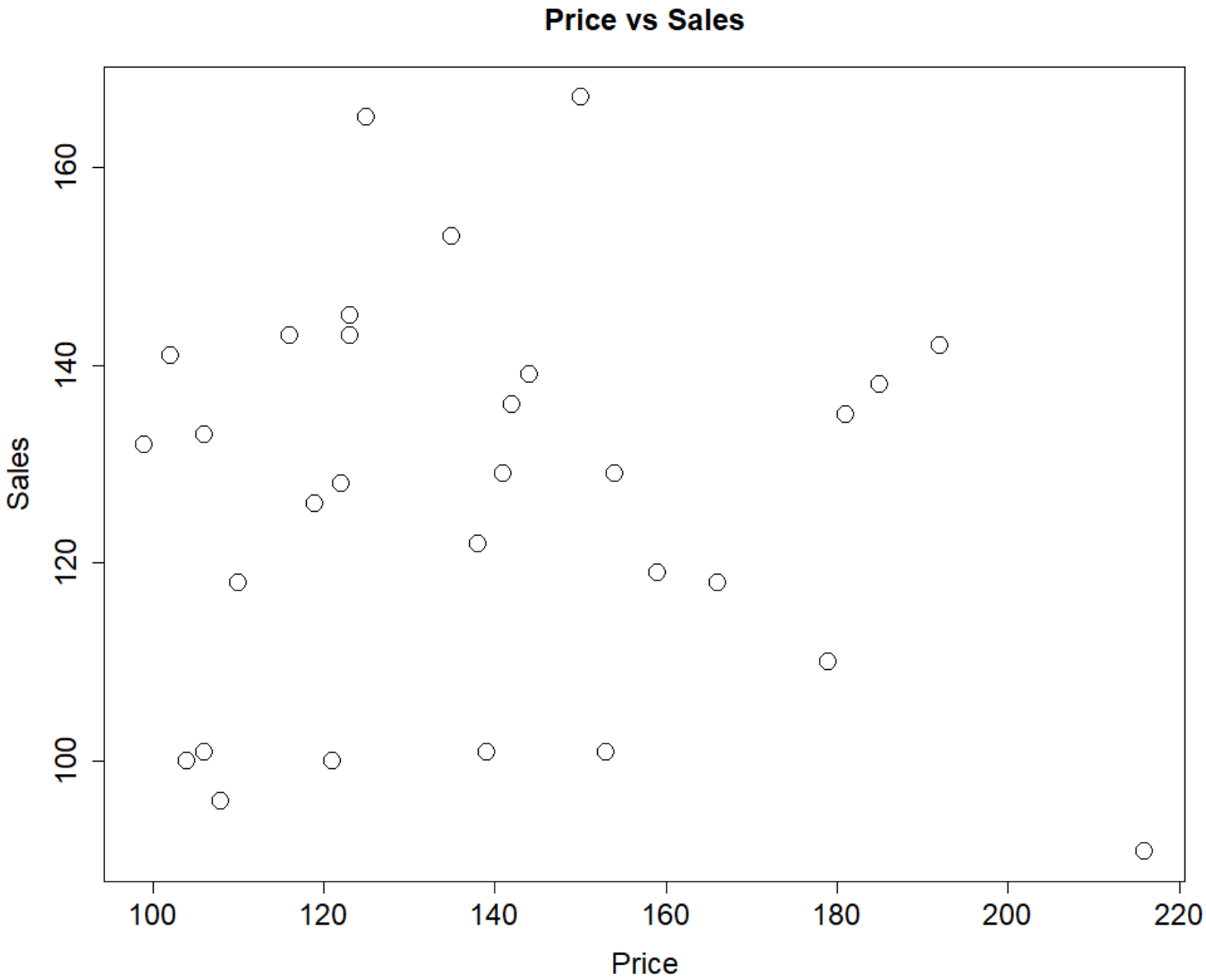
	A	B	C	D
1	period	sales	price	fuel price
2	1	101	153	63
3	2	135	181	69
4	3	110	179	71
5	4	91	216	86
6	5	138	185	59
7	6	142	192	58
8	7	118	166	58
9	8	129	154	52
10	9	119	159	44
11	10	136	142	38
12	11	101	139	38

"felm" function

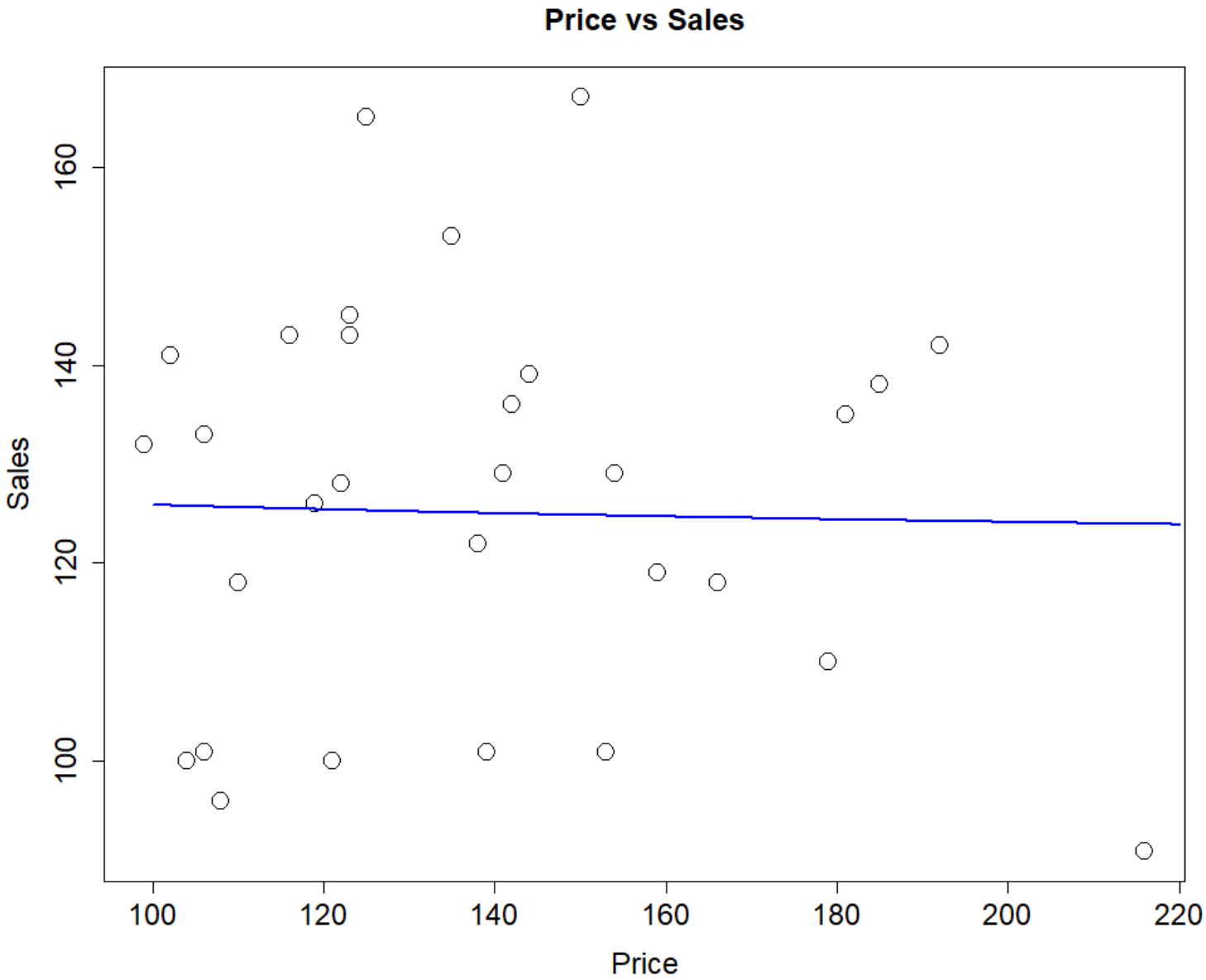
```
#IV regression example
library("lfe")
library("data.table")
#Read data
simdata=fread("simulate_example_data3.csv",stringsAsFactors = F)

#Standard OLS
reg=lm(log(sales)~log(price),data=simdata)

#IV regression using "fuel_price" as an instrument for log(price)
regiv=felm(log(sales)~ 1| 0|(log(price)~fuel_price),data=simdata)
```

I recommend using "felm" function available in "lfe" package. Example syntax is provided in project 1.

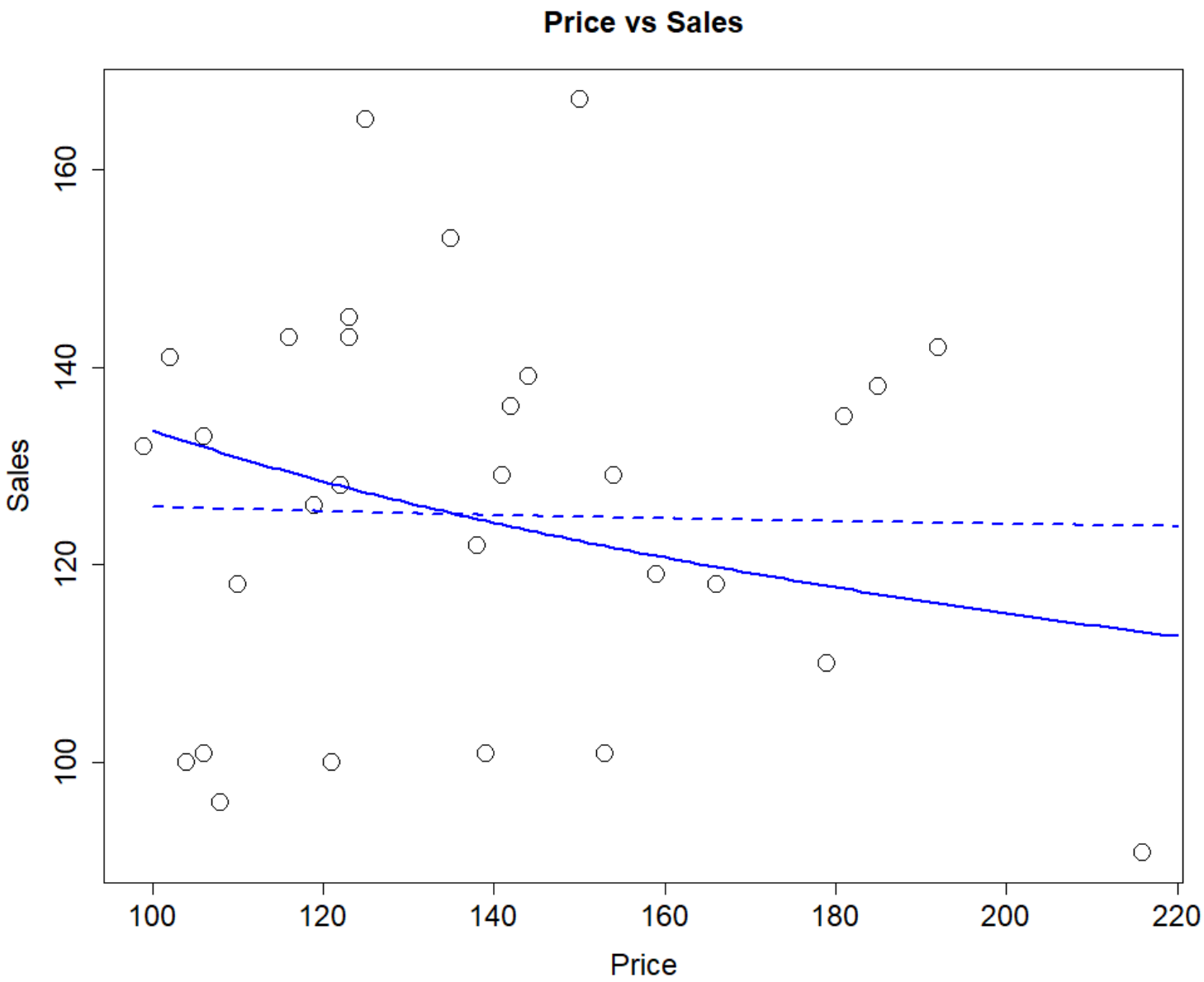
Scatter plot



Standard OLS



IV regression



IV regression

- Once you find a good candidate for a instrumental variable, implementation isn't hard.
- The challenge is to find a right instrumental variable in a given environment.

IV regression

- Formally, a variable Z can be used as an instrumental variable if it satisfies the following two conditions.
  1. It is correlated with price: $Cov(P, Z) \neq 0$.
  2. It is independent from unobservable demand-side factors: $E(\epsilon \mid Z) = 0$.
- In the airline case, both of them are likely satisfied - airfare is subject to fuel price, and fuel price may not affect travel demand directly.
- Cost is often a good candidate for an IV. But in some industries there may exist more suitable IVs. This is where your expertise on the industry shines!

Multiple IVs

- In some cases, you have multiple candidates for an IV.
- IV doesn't have to be a single variable - include all of them!

```
#IV regression using fuel, labor and airport fees as a instruments for log(price)
regiv=felm(log(sales)~ 1| 0|(log(price)~fuel_price+labor+airport_fee),data=simdata)
```

- In this case, price variations are used as long as they are correlated with any of the IVs included - you can use more information in the data.

Optional - IV notes

- Cost shifters are usually a good candidate of IV, but not always.
- For example, consider your firm not changing your product price even though your input price surges (i.e. managerial decisions to eat up the cost). Then, cost shifters and prices are not correlated - cost shifters are not a valid IV.

Optional - IV for P is not a proxy for P

- It is tempting to believe that an IV Z can instead be used as a proxy for the price.

$$\log(Q) = \beta_0 + \beta_1 \log(Z) + \beta_X X + \epsilon.$$

- This is *incorrect*. β_1 in the regression above measures the causal effect of *fuel price* on sales. Given our objectives, this regression just makes no sense.

Summary of establishing causality

- Pricing requires a causal estimate of demand.
- Find any demand-side factors that may contaminate the estimate (i.e. ones that create correlations between P and ϵ) and include them in the regression. If you have panel data, fixed effects will give you a shortcut.
- The other approach is to find variation in P that is independent of demand-side factors ϵ, and only use those variations. To do so, you need a measure of likely cause of supply-side-driven price change. Use it as an instrumental variable.
- Of course you can (and often do) combine both approaches.

What we study about regression models

- "Regression in log"
- Causal analysis with regression models
- Estimating competition

Rival prices as an X variable

- If we own another product and/or there exists a rival product, we may well include its price (say $\log(P_2)$) in the regression as an X.

$$\log(Q) = \beta_0 + \beta_1 \log(P) + \beta_2 \log(P_2) + \text{other } X \text{ variables} + \epsilon.$$

- Recall that β_1 is the price elasticity of demand - using the same logic, β_2 now represents *cross price elasticity of demand*: "if P_2 changes by one percent, Q of my product changes by what percent?".
- β_2 lets us study product substitution and competition.
- Note that we also need a causal relationship from P_2 to Q.

Presence of other products

month	location	quantity	price	price2	price3
1	101	9391	10433.618	4834.9863	5663.1409
2	101	6346	10180.527	4902.9757	5598.9584
3	101	6462	10520.427	3483.454	5819.4764
4	101	4670	10697.118	3751.8279	5962.2228
5	101	2750	7709.6709	3974.7472	4434.3434
6	101	6447	7975.0215	4200.6266	5048.086
7	101	5590	9928.5	3779.5991	6108.4691
8	101	4349	11145.028	3588.9502	6897.3372
9	101	4984	11401.249	7907.269	7063.4566
10	101	4274	11746.366	6411.6709	7232.1602
11	101	4838	15304.676	6854.9102	8775.0753
12	101	3130	14874.94	7290.2031	8613.6638
1	102	1269	15339.186	7021.8208	8579.9897

Presence of other products

- Suppose that we run a regression without including rival prices, and we get this:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
log(price)   -1.984      0.237  -8.369 3.33e-11 ***
```

- Then we include prices of two rival products, and we get this:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
log(price)    -2.7328     0.3683  -7.420 1.33e-09 ***
log(price2)    0.5529     0.2774   1.993   0.0517 .
log(price3)    1.5712     0.5999   2.619   0.0116 *
```

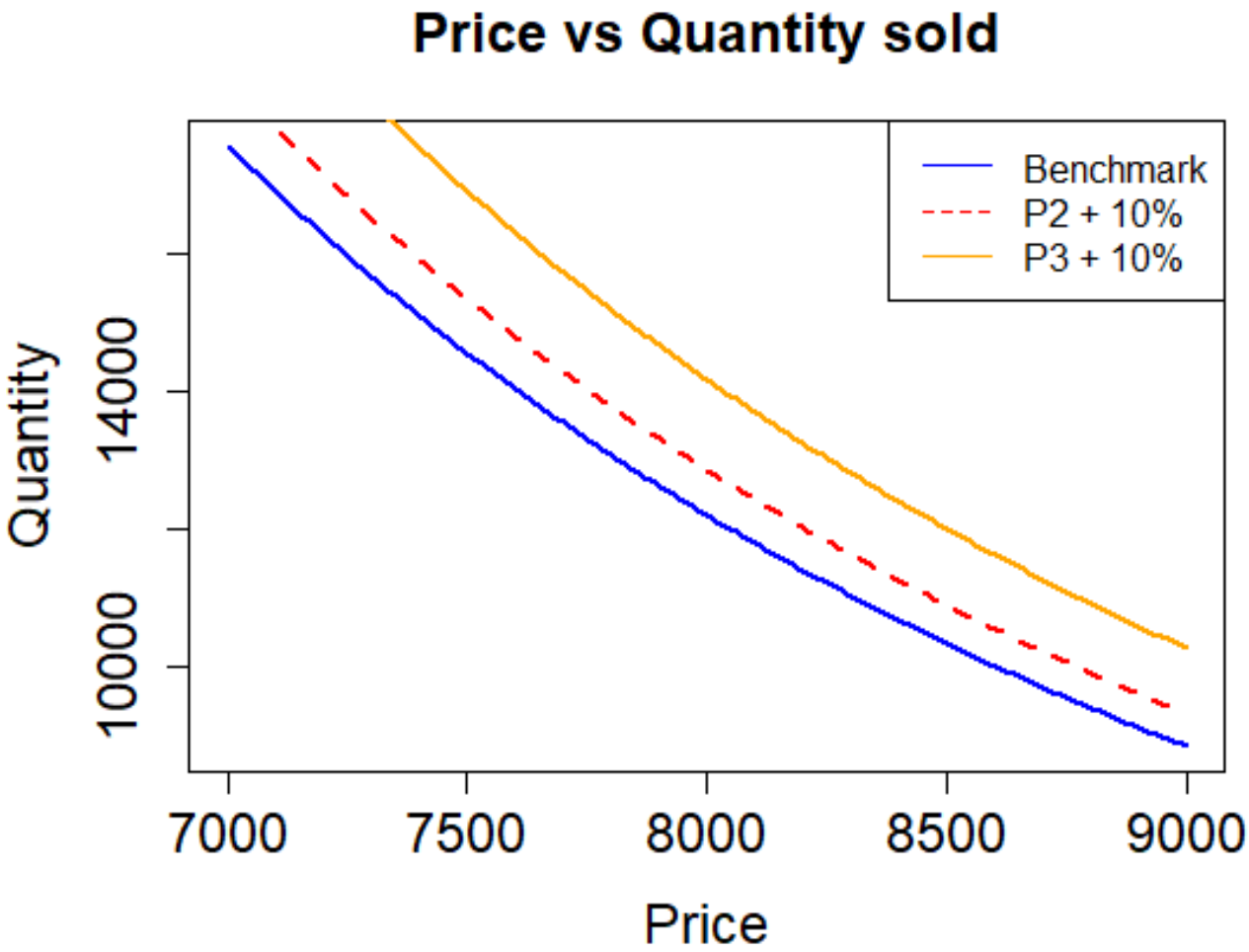
- What can we infer from those results?

Presence of other products

- Observation 1: product 3's price change impacts our demand more - product 3 is a closer substitute.

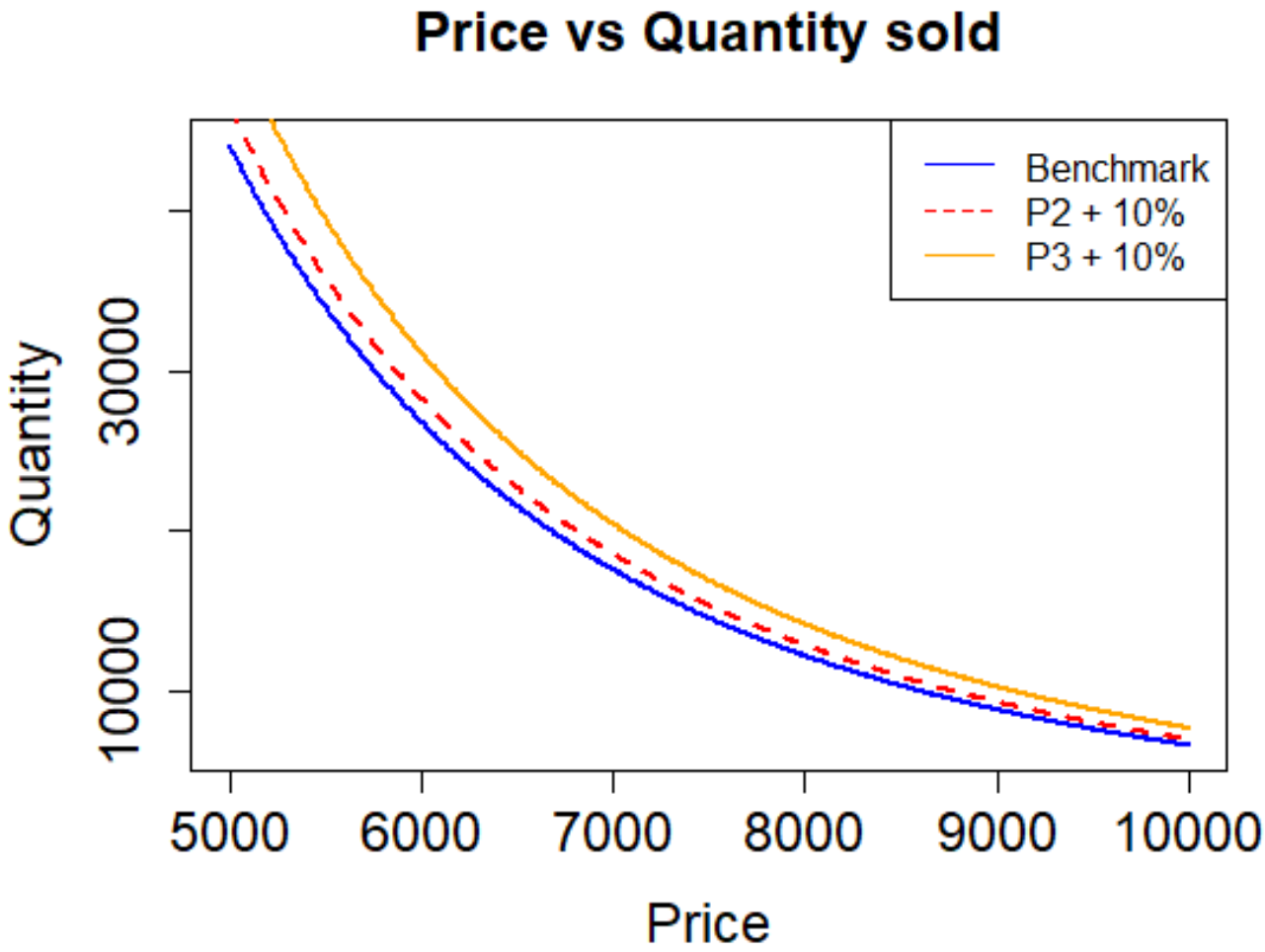
Presence of other products

- Observation 1: product 3's price change impacts our demand more - product 3 is a closer substitute.



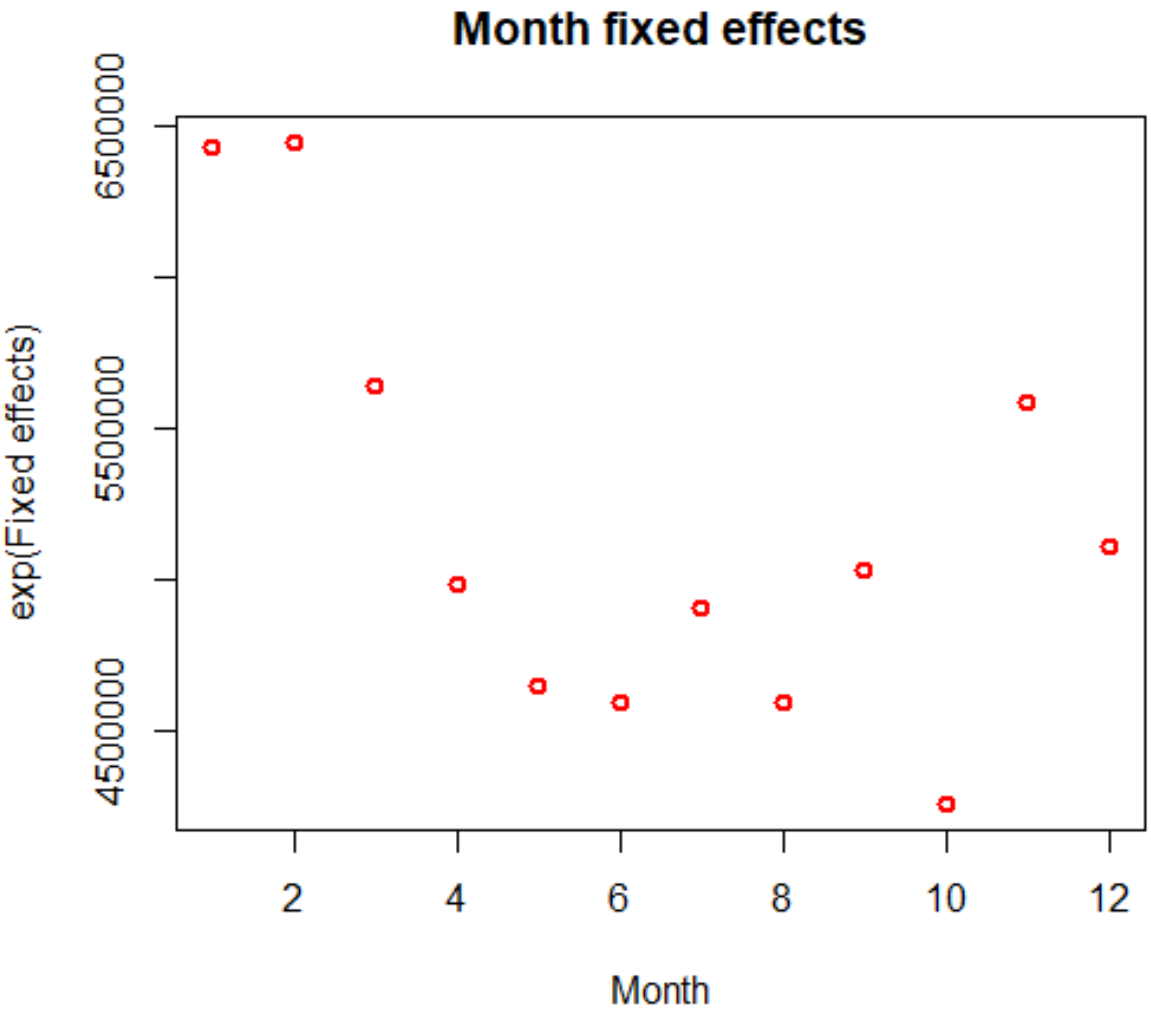
Presence of other products

- Observation 1.1: The cross effect is larger when prices are low.



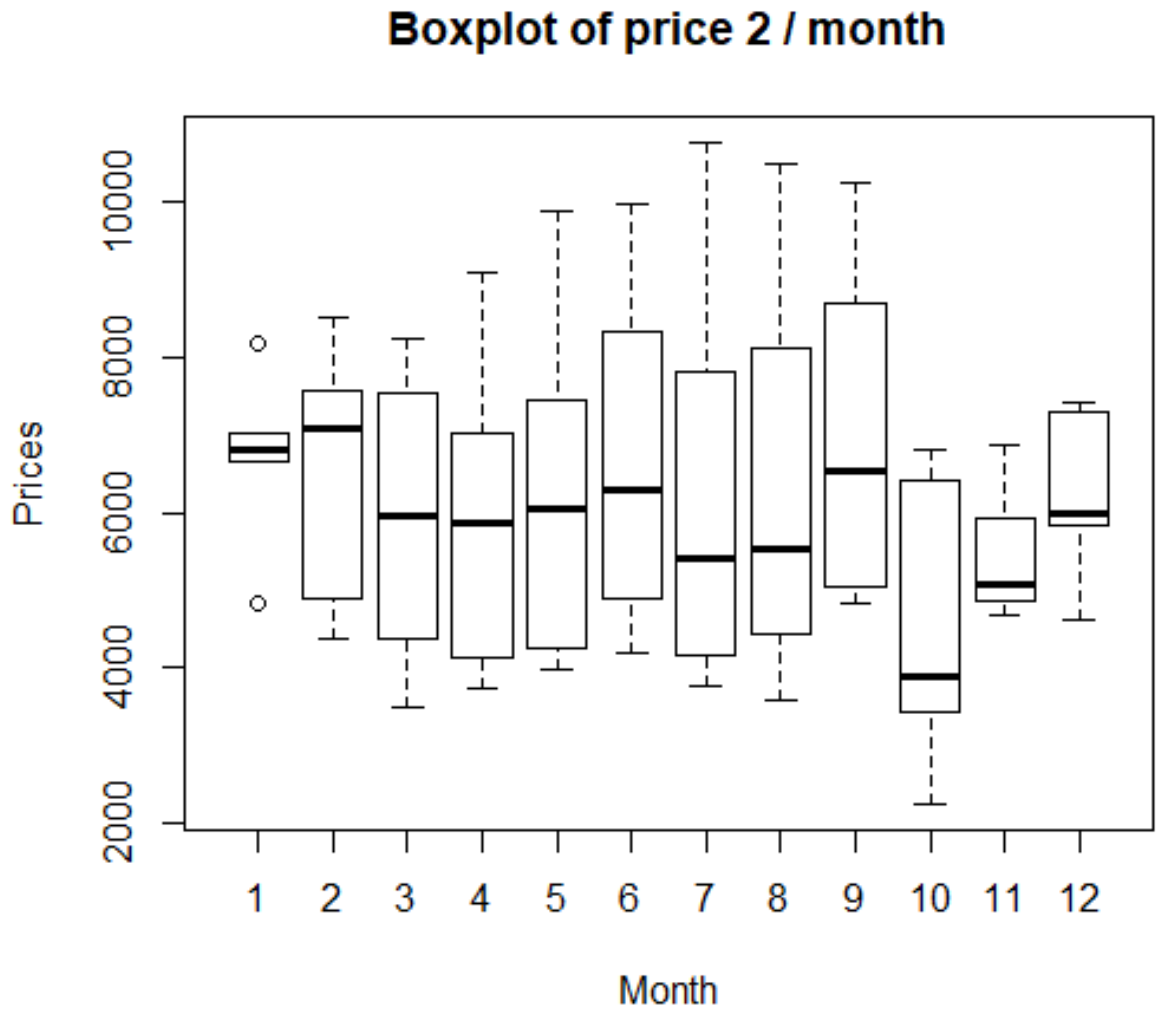
Presence of other products

- Observation 2: The demand systematically varies across months.



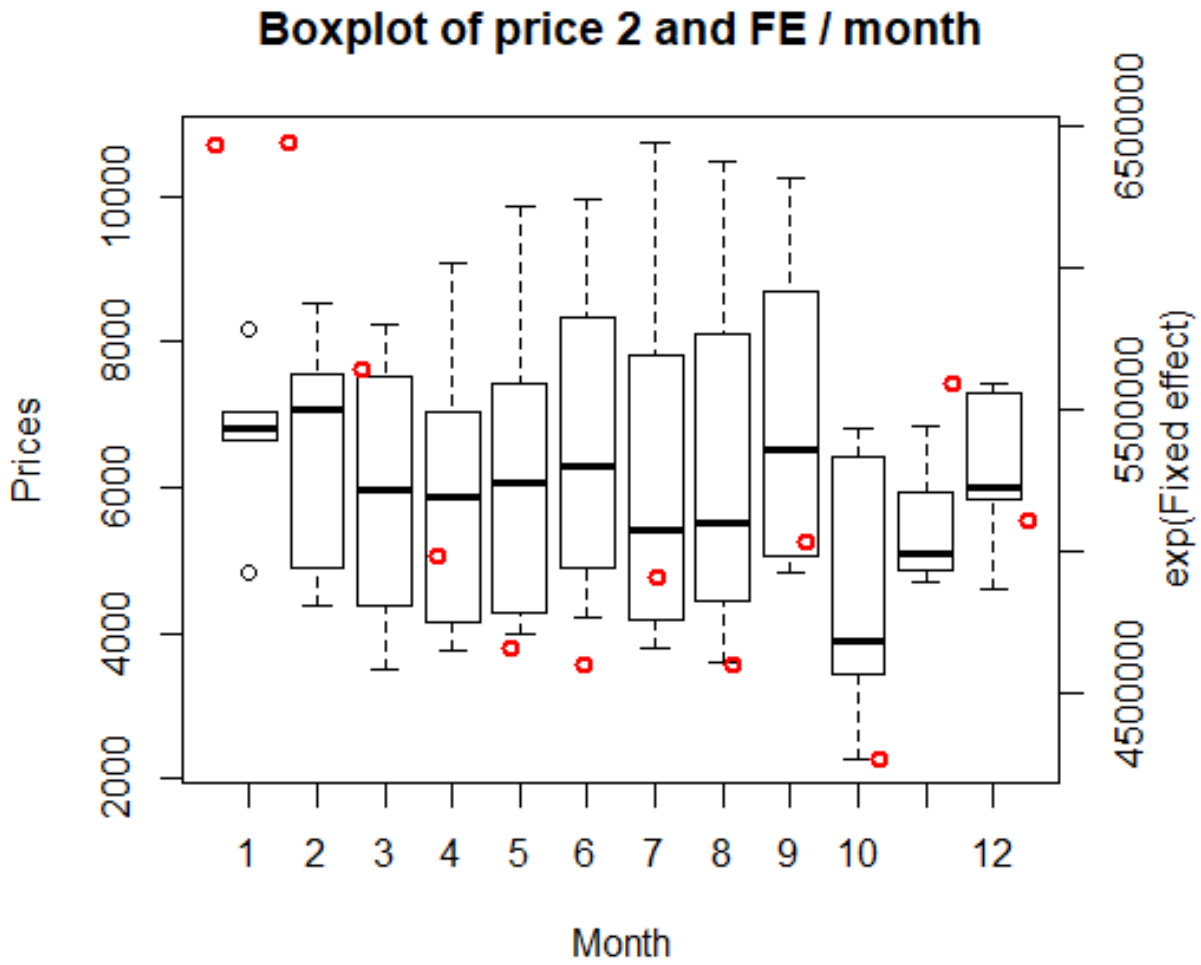
Presence of other products

- Observation 2: The demand systematically varies across months.



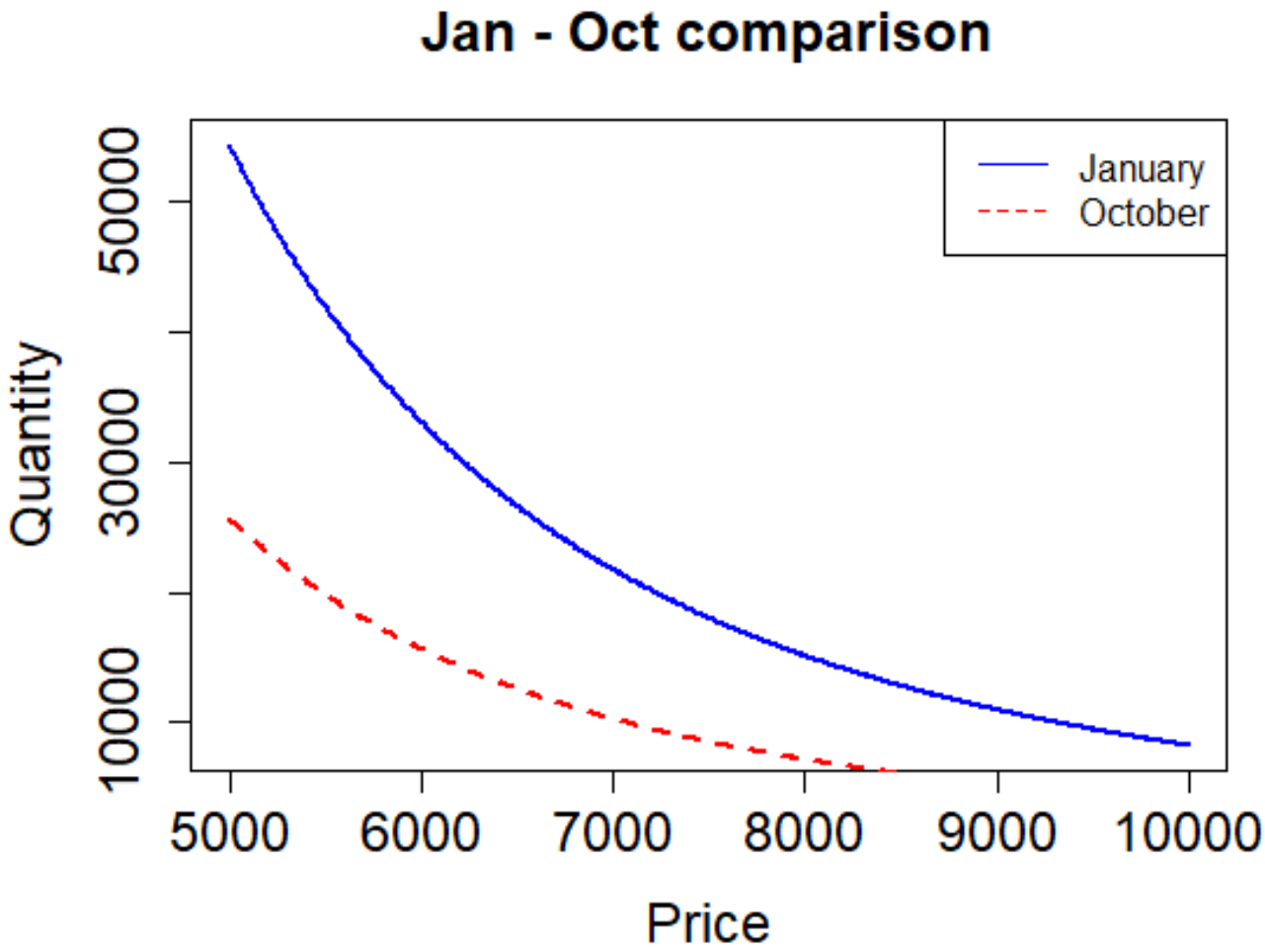
Presence of other products

- Observation 2: The demand systematically varies across months.



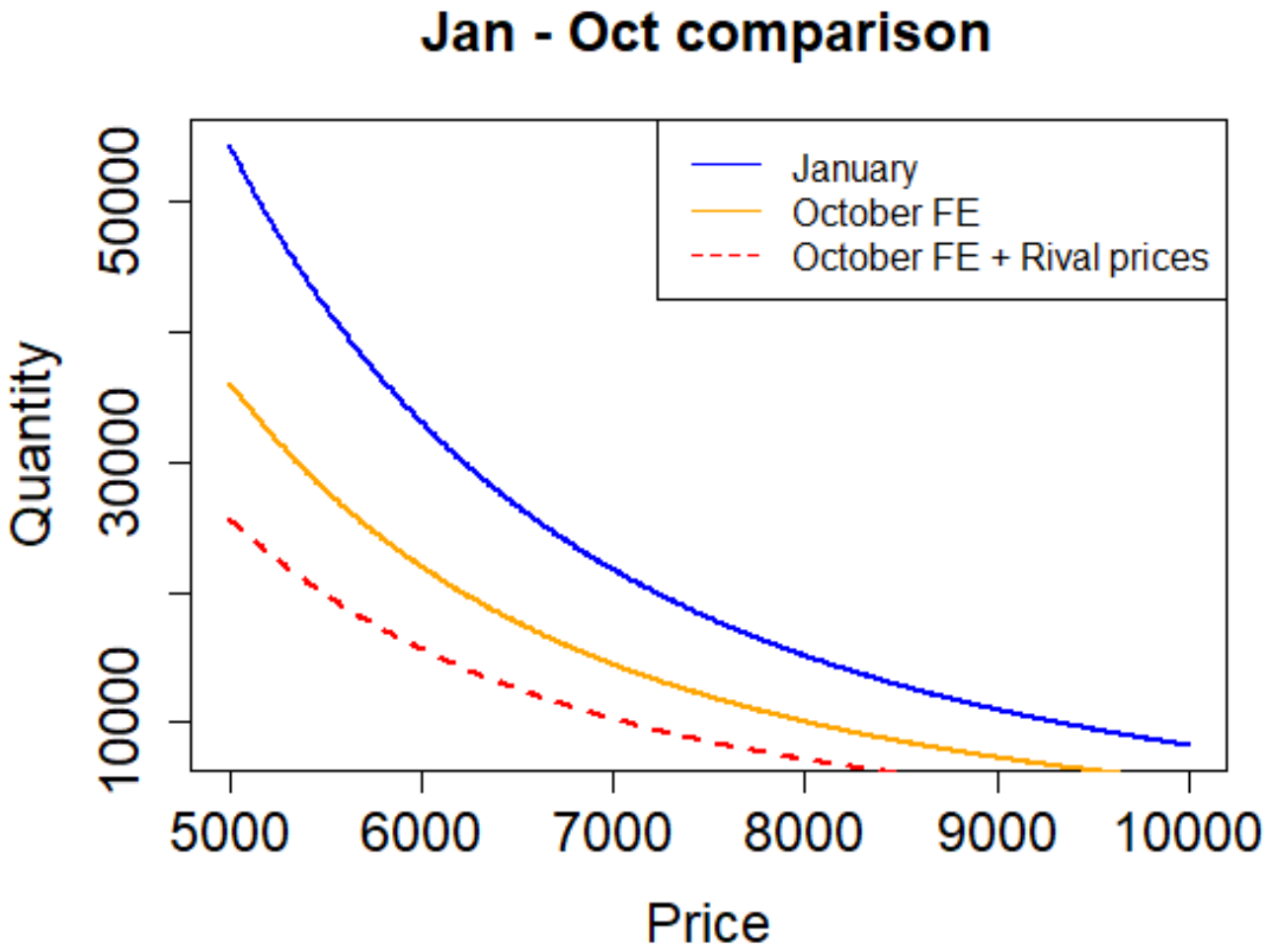
Presence of other products

- Observation 2: The demand systematically varies across months.



Presence of other products

- Observation 2: The demand systematically varies across months.



Repost: Presence of other products

- Suppose that we run a regression without including rival prices, and we get this:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
log(price)   -1.984      0.237  -8.369 3.33e-11 ***
```

- Then we include prices of two rival products, and we get this:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
log(price)   -2.7328     0.3683  -7.420 1.33e-09 ***
log(price2)   0.5529     0.2774   1.993   0.0517 .
log(price3)   1.5712     0.5999   2.619   0.0116 *
```

- What can we infer from those results?

Presence of other products

- Observation 3: Including cross effect increases own elasticity. Why?

Presence of other products

- No cross effect:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
log(price)   -1.984      0.237  -8.369 3.33e-11 ***
```

- Only product 2:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
log(price)   -1.9503     0.2274  -8.577 1.84e-11 ***
log(price2)   0.6881     0.2878   2.391   0.0205 *
```

- Both product 2 and 3:

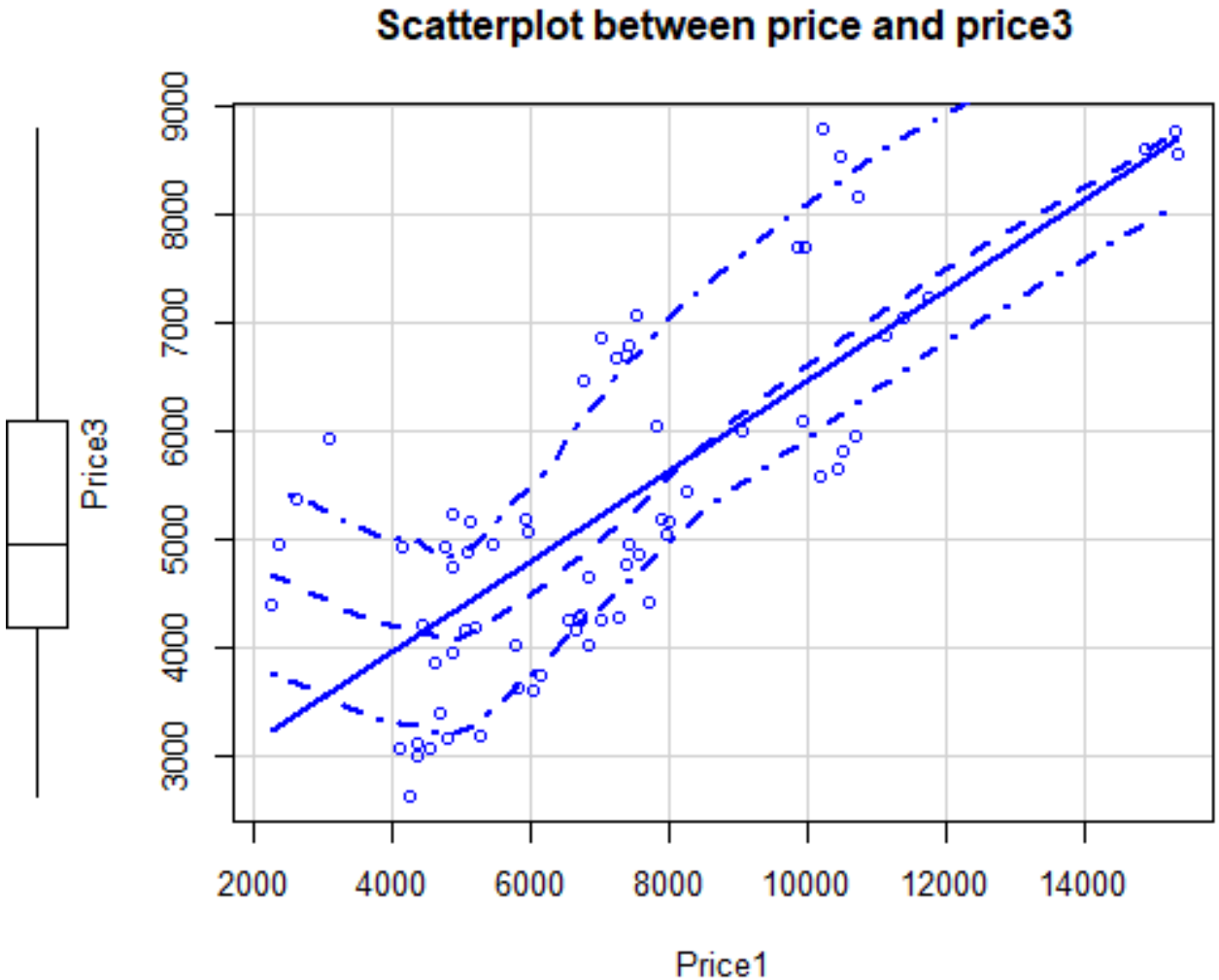
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
log(price)   -2.7328     0.3683  -7.420 1.33e-09 ***
log(price2)   0.5529     0.2774   1.993   0.0517 .
log(price3)   1.5712     0.5999   2.619   0.0116 *
```

Presence of other products

- Presence of a new variable changes β_1 if that new variable affects causal interpretation of β_1.
- Likely, the original β_1 was not causal due to some contamination. By including $\log(P_3)$, we have a better causal estimate. Why?
- Suppose that there's some time-varying demand-side factors in ϵ that fixed effects cannot eliminate. Suppose also that both managers from firm 1 and 3 see the same ϵ and set prices in response to it.
- Then $\log(P_3)$ serves as a proxy to that time-varying factor. It is observable, and is correlated with the contaminating ϵ. Including it in the regression improves β_1.

Presence of other products

- High correlation between price and price 3 - both firms are likely setting prices in response to the same demand shock.



Summary of sales data

- Sales data is a type of data that you tend to face a lot of issues with and cannot get much insights from. Nevertheless, it's, so, prevalent.
- When you need to work on a sales data, do make sure you have a causal estimate. That is the first-order priority (not the model fit!). Care about R^2 only after you are all done with the causality issue.
- With the right causal analysis, regression models tell you own- and cross-elasticities in the market. These measures are good input for simple pricing objectives.

Limitation of sales data

- With sales data, we never see "who buys what". Hence segmentation is just way too difficult. Segmentation is a key to price targeting and product positioning - the mainstream of what modern pricing analytics.
- For the rest of the course, we switch our focus to *choice data*, to which increasing number of firms have access.
- Choice data provide much richer information about each consumer's historical purchase patterns. We can hence study more realistic demand systems, including the presence of multiple segments of consumers.