MPO1: Quantitative Research Methods

Session 8: Tests of regression assumptions, continued: Outliers and Influential observations

Thilo Klein

University of Cambridge
Judge Business School

CAMBRIDGE Judge Business School

Wine and Wealth

- Robert Owen CEO of Forestier. A company that produces 4 brands of wine: Almaden, Bianco, Casarosa, and Delacroix.
- Expand the business to neighborhoods not currently covererd.
- Estimate how average income affects sales of wine.
- Data: average monthly income (units of 1,000 GBP) and monthly wine sales (units of 1,000 GBP).

Simple regression using the Almaden data

Sales A on Income A

```
>  data <- read.csv("http://thiloklein.de/R/wineandwealth.csv")
>  attach(data)
>  lmA <- lm(SalesA ~ IncomeA); summary(lmA)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
IncomeA       0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Simple regression using the Bianco data

Sales B on Income B

```
>  lmB <- lm(SalesB ~ IncomeB); summary(lmB)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
IncomeB        0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

Simple regression using the Casarosa data

Sales C on Income C

```
>  lmC <- lm(SalesC ~ IncomeC); summary(lmC)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
IncomeC       0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

Simple regression using the Delacroix data

Sales D on Income D

```
>  lmD <- lm(SalesD ~ IncomeD); summary(lmD)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
IncomeD       0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667, Adjusted R-squared: 0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Forecasting with regressions

To find confidence interval for the predicted value, $\hat{Y}_i|X_i$, we need

- the predicted value $\hat{Y}_i|X_i$
- its standard error $s.e.(\hat{Y}_i|X_i)$
- the appropriate t-distribution quantile for the chosen $(1-\alpha\%)$ confidence interval, so that we can find:

$$\hat{Y}_i|X_i \pm s.e.(\hat{Y}_i|X_i) \cdot t_{n-k,\alpha/2}$$

Variance / Standard error of the predicted value

Variance of the predicted value

for model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$$Var\left(\hat{Y}_i|X_i\right) = Var\left(\hat{\beta}_0\right) + Var\left(\hat{\beta}_1\right) \cdot X_i^2 + 2X_i Cov\left(\hat{\beta}_0, \hat{\beta}_1\right)$$

Aside: (equivalently)

- Residual: $e_i = Y_i - \hat{Y}_i$
- Residual variance estimated by: $[s.e.(e)]^2 = \frac{\sum_i e_i^2}{n-2}$
- Variance of the predicted value:

$$\left[s.e.(\hat{Y}_i|X_i)\right]^2 = [s.e.(e)]^2 \left[\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{(n-1)Var(X)}\right]$$

CAMBRIDGE Judge Business School

Confidence interval for the predicted value

So the **confidence** interval depends on:

- Residual variance
- Distance of the data point from the mean
- Sample size
- Variance of independent variable

```
# R-code; Conf interval at IncomeA = 15
>  x <- data.frame(IncomeA = 15)
>  conf_a <- predict(lmA, x, se.fit=T, level=.95,
   interval = "confidence"); conf_a
$fit
       fit      lwr      upr
1 10.50145 8.692466 12.31044
$se.fit
[1] 0.799674
```

Variance / Standard error of $Y_i|X_i$

Prediction interval:

$$Y_i|X_i \pm s.e.(Y_i|X_i) \cdot t_{n-k,\alpha/2}$$

Note: not $\hat{Y}_i|X_i$ but $Y_i|X_i$

$$\begin{aligned} Var\left(Y_i|X_i\right) &= Var\left(\hat{\beta}_0\right) + Var\left(\hat{\beta}_1\right) \cdot X_i^2 + 2X_i Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ &\quad + [s.e.(e)]^2 \end{aligned}$$

```
# R-code; Prediction interval at IncomeA = 15
>  pred_a <- predict(lmA, x, se.fit=T, level=.95,
   interval="prediction"); pred_a
$fit
       fit      lwr     upr
1 10.50145 7.170113 13.8328
$se.fit
[1] 0.799674
```

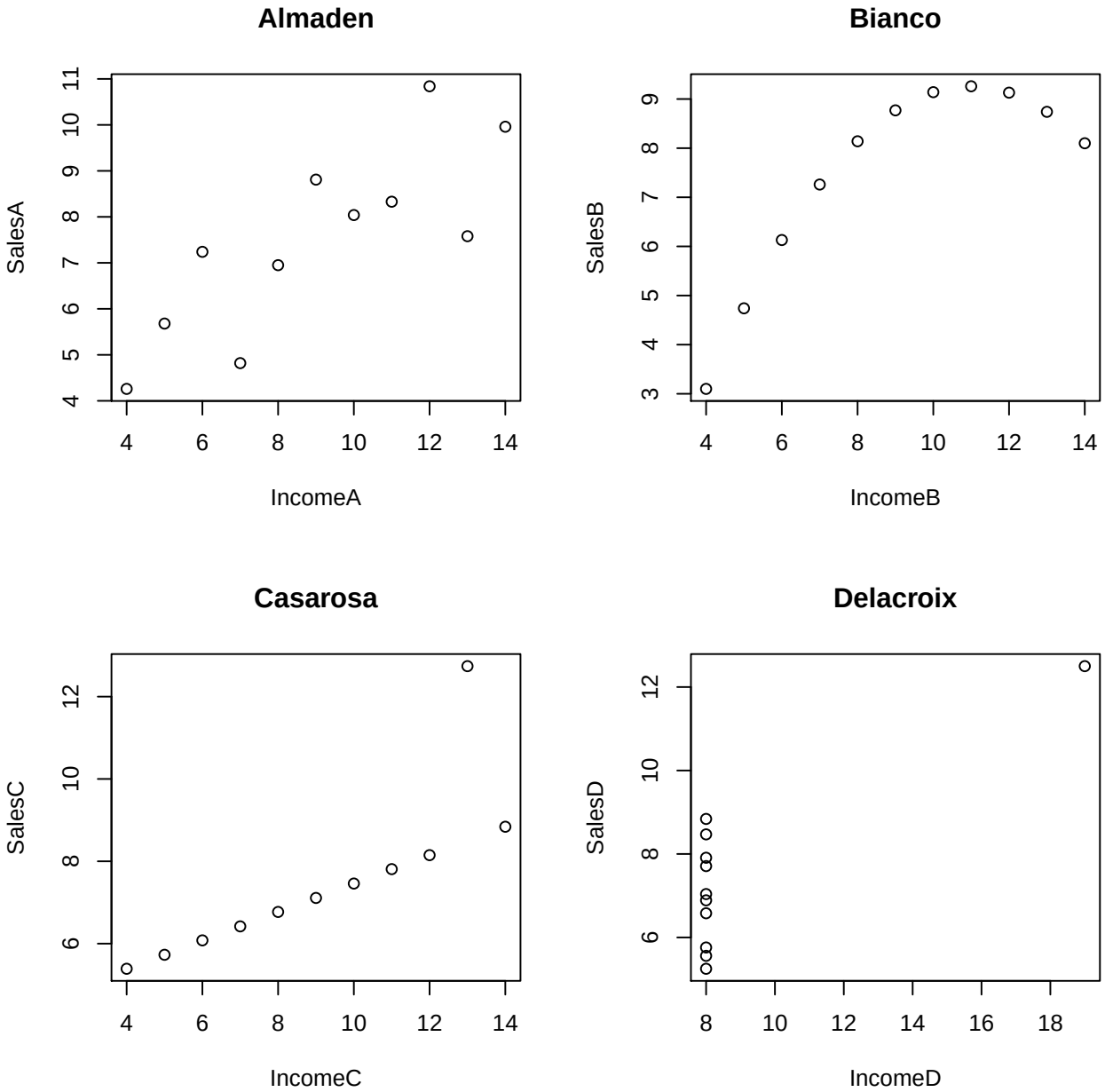
Wine and Wealth: Confidence interval and Prediction Interval

For all four regressions:

- 95% Confidence Intervals for income of 15: $(8.69, 12.31)$
- 95% Prediction Intervals for income of 15: $(7.17, 13.83)$

Conclusion: All four wines have a similar income effects.

Scatterplots



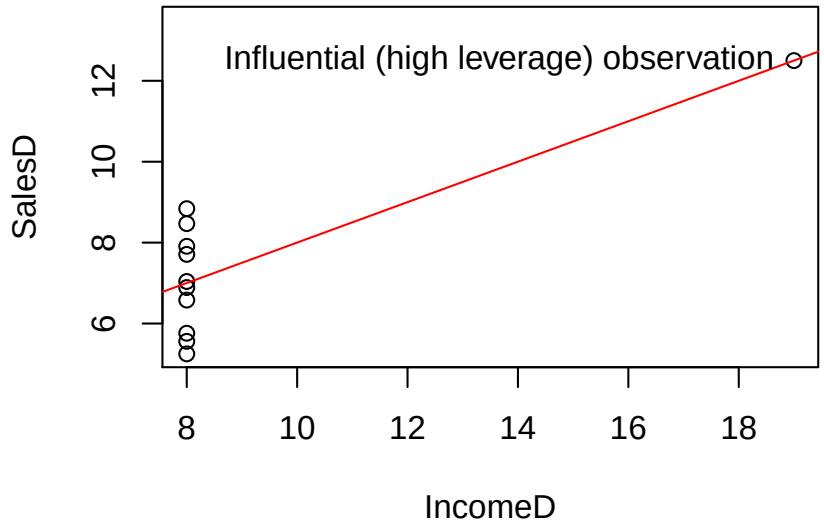
Delacroix

The data set is totally different from the other three

- The regression is being driven by one data point.
- This data point is called an influential observation due to its impact on the regression equation.
- In particular, this point has a high "leverage" as its X value is "far away" from the bulk of the other X values.
- Leverage of the observation is a measure of the potential for the observation to have undue influence on the model: ranging from 0 to 1.

Simple regression Delacroix

Influential observation




```
# R-code:
> plot(SalesD~IncomeD, ylim=c(min(SalesD), max(SalesD)+1))
> abline(lm(SalesD~IncomeD), col="red")
> at <- which(SalesD==max(SalesD)); at
[1] 8
> text(x=IncomeD[at], y=SalesD[at], "Influential obs", pos=2)
```

CAMBRIDGE Judge Business School

Influential Observations

Leverage

- An influential observation is a data point which has a large effect on the regression results.
- An influential observation can be an outlier.
- In this example, it is not an outlier. In fact, the regression line goes through this point.

Aside: Formula for Leverage, $h_i = \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)Var(X)}\right]$

```
# R-code; Leverage
> hD <- hat(model.matrix(lmD))
> plot(hD)
# Criterion: Is leverage > 2K/n ?
> 2*2/11
[1] 0.3636364
> sort(hD)
[1] 1.0 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
```

Leverage

Leverage

- The slope of the regression line is a weighted sum of slopes of lines passing between each point and the "mean point" $(\bar{X}, \bar{Y})$.
- If a point has large leverage, then the slope of the regression line follows more closely the slope of the line between that point and the mean point: points that have large leverage are important. Neither good nor bad.
- Points that have small leverage "do not count" in the regression – we could move them or remove them from the data and the regression line does not change very much.
- **Farther the observation is from the mean, the larger the leverage**. Leverage is reduced by sample size, and by large variance of independent variable.

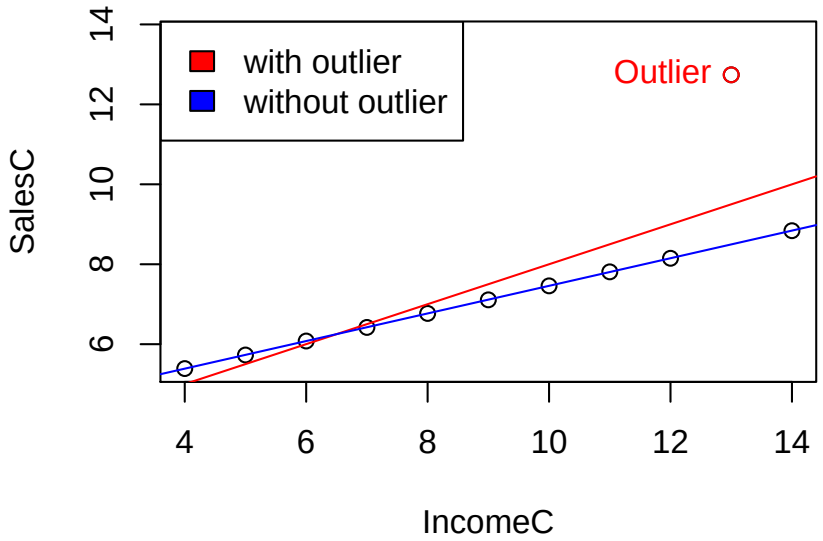
Casarosa

Outlier

- The relationship is linear, with an outlier
- The pattern is very clearly linear, BUT the regression line is being pulled from the 'line' by ONE outlier
- Outlier: sample observations whose Y values are a long way away from the the model's predicted Y value $\hat{Y}_i|X_i$: so, observations with a large residual.
- Question: Is the observation lying too far from the main trend to have happened by chance? In other words, is the corresponding residual too large to have happened by chance?

Simple regression Casarosa

Outlier




```
> plot(SalesC~IncomeC, ylim=c(min(SalesC), max(SalesC)+1))
> abline(lm(SalesC~IncomeC), col="red")
> at <- which(SalesC==max(SalesC)); at
[1] 3
> points(SalesC[at]~IncomeC[at], col="red")
> text(x=IncomeC[at], y=SalesC[at], "Outlier", pos=2, col="red")
> abline(lm(SalesC[-at]~IncomeC[-at]), col="blue")
> legend("topleft", legend=c("with","~out"), fill=c("red","blue"))
```

Outliers

Outliers

- An outlier is an observation with an unusually large *studentised* residual (calculated by dividing the residual by *its* standard error).
- Measurement scale of residuals is the same as measurement scale of Y variable, hence the standardisation.
- Check if it is not due to a mistake. If not, worthwhile to find out more about it.
- For example, in financial data an outlier might be linked to a stock market crash.
- In this example the outlier could be related to a single buyer who is particularly fond of Casarosa wine.

Aside: Studentised residuals

Studentised residuals

$$\text{Rstudent}_i = \frac{e_i}{s_{(i)}\sqrt{1-h_i}},$$

where

- $s_{(i)}$ is the error term standard deviation after deleting the ith observation
- h_i is the leverage of the ith observation

```
# R code:
> sort(rstudent(lmC), decreasing=TRUE)[1:4]
           3            8           11            7
1.203539e+03 3.618514e-01 2.002634e-01 6.640743e-02
> qt(0.025,9,lower.tail=F)
[1] 2.262157
```

CAMBRIDGE Judge Business School

Outliers

Outliers

- Studentised residual (follows the t distribution) gives the number of standard deviations a residual is away from zero.
- If studentised residual is x, this observation is x standard deviations away from zero.
- Under t-distribution of residual with, mean 0, we could expect to see 5% of "extreme residuals" in a typical regression.
- Do NOT remove outliers from your data set unless you are certain they are due to a mistake.

Cook's distance

Cook's distance, D_i

$$D_i = \frac{e_i^2}{K \cdot s.e.(e)} \cdot \frac{h_i}{(1 - h_i^2)} = \frac{\sum_{j-1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{K \cdot s.e.(e)}$$

- Cook's distance measures the effect of deleting a given observation. A measure of the global influence of this observation on **all** the predicted values
- Cook's distance considers both Outliers and observations with unusual predictors together in one statistic.
- Points with a Cook's distance of 1 or more are considered to merit closer examination in the analysis.

Example: Casarosa case

CAMBRIDGE Judge Business School

Influential Observations

Influential Observations

- As with outliers, one should check that the influential observation is not due to a data error of some sort. If it is not due to error then it should not be dropped.
- Should not rely on the results from the Delacroix regression. The results are driven by one observation.

```
# R code, Cook's distance
> cookD <- cooks.distance(lmD)
> round(cookD, 2)
   1    2    3    4    5    6    7    8    9   10   11
0.01 0.06 0.02 0.14 0.09 0.00 0.12  NaN 0.08 0.03 0.00
```

Watch out for results reported as 'NaN' (Not a Number)!

CAMBRIDGE Judge Business School

Summary for this part

Summary

- Regression must not be interpreted mechanically.
- Assumptions must be checked.
- Careful specification
- Outliers and influential observations should not be modified or deleted unless there is measurement error or data entry error.

Quick: Review – Least Squares

Assumptions

1. Model linear in parameters
   $Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i$
2. Zero conditional mean: $E(u|X = x) = 0$
3. Random sampling, i.e., errors are uncorrelated
   $(X_i, Y_i), i = 1, ..., n$ are i.i.d., i.e.,
   $E(u_i, u_j|X_1, X_2, ..., X_k) = 0$
4. No perfect multicollinearity
5. Large outliers are rare (finite moments)
6. u is homoskedastic: $V(u|X_1, X_2, ..., X_k) = \sigma^2$
7. u is distributed $N(0, \sigma^2)$

Efficiency of OLS

The Gauss-Markov Theorem

- Under LS assumptions 1-6 (including homoskedasticity), $\hat{\beta}_i$ has the smallest variance among *all linear estimators* (estimators that are linear functions of the variables).
- Under LS assumptions 1-7 (including normally distributed errors) $\hat{\beta}_i$ has the smallest variance of all consistent estimators (linear *or* nonlinear functions of $Y_1, ..., Y_n$), as $n \mapsto \infty$.

A common problem

Problem

- Suppose that you have a model in which Y is determined by X: $Y = \beta_0 + \beta_1 X_1 + u$.
- but you have reason to believe that u is not distributed independently of X (because of omitted variables, or any other reason) – a Gauss-Markov condition is violated.
- An OLS regression would then yield biased and inconsistent estimates.

Why?

If $Cov(X, u) \neq 0$, OLS estimator of β_1 is biased (and inconsistent)

OLS estimator: $\hat{\beta}_1^{OLS} = \frac{Cov(X,Y)}{V(X)}$

$$\begin{aligned}
\hat{\beta}_1^{OLS} &= \frac{Cov(X, \beta_0 + \beta_1 X + u)}{V(X)} \\
&= \beta_1 \frac{Cov(X, X)}{V(X)} + \frac{Cov(X, u)}{V(X)} \\
&= \beta_1 + \frac{Cov(X, u)}{V(X)}
\end{aligned}$$

$\Rightarrow$ If $\frac{Cov(X,u)}{V(X)} \neq 0$, then OLS estimator is biased! (can be shown that it is inconsistent as well)

Errors-in-variables bias

So far we have assumed that X is measured without error..

In reality, economic data often have measurement error.

- Data entry errors in administrative data
- Recollection errors in surveys (when did you start your current job?)
- Ambiguous questions problem (what was your income last year?)
- Intentionally false response problems with surveys(What is the current value of your financial assets? How often do you drink and drive?)

In general, measurement error in a regressor results in "errors-in-variables" bias.

Illustration:

- Suppose
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
is "correct" in the sense that the three least squares assumptions hold (in particular $E(u_i|X_i) = 0$).
- Let
X_i = unmeasured true value of X
$\tilde{X}_i$ = imprecisely measured version of X
- Then
$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1((X_i - \tilde{X}_i) + u_i], \end{aligned}$$
or $Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i$, where $\tilde{u}_i = \beta_1(X_i - \tilde{X}_i) + u_i$.

CAMBRIDGE Judge Business School

Errors-in-variables bias (cont'd)

Illustration (cont'd)

But $\tilde{X}_i$ typically is correlated with $\tilde{u}_i$ so $\hat{\beta}_1$ is biased:

$$\begin{aligned} cov(\tilde{X}_i, \tilde{u}_i) &= cov(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 cov(\tilde{X}_i, X_i - \tilde{X}_i) + cov(\tilde{X}_i, u_i) \\ &= \beta_1[cov(\tilde{X}_i, X_i) - var(\tilde{X}_i)] + 0 \neq 0, \end{aligned}$$

because in general $cov(\tilde{X}_i, X_i) \neq var(\tilde{X}_i)$

CAMBRIDGE Judge Business School

Errors-in-variables bias (cont'd)

Illustration (cont'd)

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i,$$

where $\tilde{u}_i = \beta_1(X_i - \tilde{X}_i) + u_i$.

- If X_i is measured with error, $\tilde{X}_i$ is in general correlated with $\tilde{u}_i$, so $\hat{\beta}_1$ is biased and inconsistent.
- It is possible to derive formulas for this bias, but they require making specific mathematical assumptions about the measurement error precess (for example, the $\tilde{u}_i$ and X_i are uncorrelated). Those formulas are special and particular, but the observation that measurement error in X results in bias is general.

Potential solutions to errors-in-variables bias

Solutions

1. Obtain better data.
2. Develop a specific model of the measurement error process.
3. This is only possible if a lot is known about the nature of the measurement error - for example a subsample of the data are cross-checked using administrative records and the discrepancies are analysed and modeled. (We won't pursure this)
4. Instrumental variables regression (next term)

Sample selection bias

Sample selection

So far we have assumed simple random sampling of the population. In some cases, simple random sampling is thwarted because the sample, in effect, "selects itself".

Sample selection bias arises when a selection process:

1. influences the availability of data and
2. that process is related to the dependent variable.

Example

Firm performance

- Do old firms outperform young firms?
- Empirical strategy:
  - Sampling scheme: simple random sampling of firms on which data is available on a given data.
  - Is there sample selection bias?

Sample selection bias induces correlation between a regressor and the error term.

$$\text{performance}_i = \beta_0 + \beta_1 \text{firm-age}_i + u_i$$

Being an old firm in the sample means that your return was better than failed firms who by definition are not in the sample "old" – so $corr(\text{firm-age}_i, u_i) \neq 0$.

Potential solutions to sample selection bias

Solutions

- Collect the sample in a way that avoids sample selection.
- Randomised controlled experiment.
- Construct a model of the sample selection problem and estimate that model (will explore this next term)

Simultaneous causality bias in equations

Simultaneous causality

So far we have assumed that X causes Y. What if Y causes X, too?

1. Causal effect on Y of X $\quad Y_i = \beta_0 + \beta_1 X_i + u_i$
2. Causal effect on X of Y $\quad X_i = \gamma_0 + \gamma_1 Y_i + v_i$

- Large u_i means large Y_i, which implies large X_i(if $\gamma_1 > 0$)
- Thus $corr(X_i, u_i) \neq 0$
- Thus $\hat{\beta}_1$ is biased and inconsistent

Potential solutions to simultaneous causality bias

Solutions

1. Randomized controlled experiment. Because X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i (assuming perfect compliance).
2. Develop and estimate a complete model of both directions of causality. *This is extremely difficult in practice.*
3. Use instrumental variables regression to estimate the causal effect of interest (effect of X on Y, ignoring effect of Y on X). To be covered next term.

In Conclusion

Conclusion

- Your arguments must be grounded with some "theory", which in turn must be grounded in your understanding of the "world".
- Going from anecdote to generalisation and inference is an essential part of the research process
- Accurate measurements of the relative size of various responses is very useful

This module

This module has:

- Enabled you to estimate reasonably rich regression models,
- Alerted you to some problems associated with regression modelling, and strategies available to solve them
- Enabled you to appreciate critically the empirical literature in your field

Interpretation

Interpretation

Statistical and econometric estimates are simply a sophisticated way of describing the data.

- Really a way of obtaining a set of conditional correlations
- Can tell us what relationships are stronger than would arise if the variables were random

Theory (and common sense) tell us

- What relationships are interesting
- How to interpret them

Interpretation (cont'd)

Interpretation (cont'd)

Statistical and econometric estimates are simply a sophisticated way of describing the data.

- In general what we observe in our data do not correspond one to one with the (counterfactual) experiment we would like to conduct
  - Some variables are unobserv(ed/able)
  - We do not know what is causing the variation in our explanatory variables (endogeniety/selection)
- There are useful and imaginative solutions to all these problems, and that will be the focus of the module next term.