Aalto University
School of Business

Research Methods In Accounting

22E20700

Henry Jarva
Aalto University

Learning outcomes

- The course provides students skills to conduct a **quantitative** / qualitative study.
- The course is designed to facilitate the thesis work.
  - To lower the ”barrier to entry”
  - Improve the quality of the thesis (already very high!)
  - Basics of using SAS 9.4 software

Material - Qualitative research

- Textbook: Boris Blumberg, Donald R. Cooper & Pamela S. Schindler. *Business Research Methods*, Second European Edition, McGraw-Hill, 2008.
- Ahrens, T. and Dent, J. 1998. Accounting and Organizations: Realizing the Richness of Field Research. *Journal of Management Accounting Research*, vol.10, pp. 1–39.
- McKinnon, J. 1988. Reliability and validity in field research: some strategies and tactics. *Accounting, Auditing and Accountability Journal*, vol.1, pp. 34–54.
- Scapens, R. 1990. Researching management accounting practice: the role of case study methods. *British Accounting Review*, 22, pp. 259–281.
- Vaivio, J. 2008. Qualitative management accounting research: rationale, pitfalls and potential. *Qualitative Research in Accounting & Management*, vol.5, no.1, pp. 64–86.
- Videos (part 1–7)

Material - Quantitative research

- Lecture and exercise material
- All material will be available through MyCourses

GRADING

- **GRADING: 1 – 5 (6 cr)**
- **TWO 'OPEN BOOK' EXAMS**
  - You need to pass **both** qualitative and quantitative exams.
  - Both exams are equally weighted.
  - Exams will be available in MyCourses

Why study econometrics?

- **Econometrics** is about how we can use theory and data from economics, business, and the social sciences, along with tools from statistics, to answer ”how much” questions.

PART I
Simple Regression

The simple regression model

- The simple regression model can be used to study the relationship between two variables
- It has limitations as a general tool for empirical analysis
- Learning how to interpret the simple regression model is good practise for studying multiple regression
- Dependent variable: The variable to be explained in a regression model (and a variety of other models)
- Independent variable / explanatory variable: In regression analysis, a variable that is used to explain variation in the dependent variable

Terminology for simple regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor
	Covariate

The simple regression model

- $y = \beta_0 + \beta_1 x + u$
- (y, x, u) are random variables
- y and x are observable while u is not
- Model implies that u captures everything that determines y except for x
- In social sciences, this often includes a lot of stuff!

Definition of the simple regression model

- Much of applied econometric analysis begins with the following premise: *y* and *x* are two variables, representing some population, and we are interested in ”explaining *y* in terms of *x*,” or in ”studying how *y* varies with changes in *x*.”
  - Examples:

1. *y* is soybean crop yield and *x* is amount of fertilizer
2. *y* is hourly wage and *x* is years of education
3. *y* is community crime rate and *x* is number of police officers

Soybean yield and fertilizer

- Suppose that soybean yield is determined by the model:
  - $yield = \beta_0 + \beta_1 fertilizer + u$
  - So that $y = yield$ and $x = fertilizer$.
  - β_0 is the intercept parameter and β_1 is the slope parameter.
  - The effect of fertilizer on yield is given by β_1:
  - $\Delta yield = \beta_1 \Delta fertilizer$
  - The error term (or disturbance) u contains factors such as land quality, rainfall, and so on.

Deriving the ordinary least squares (OLS) estimates

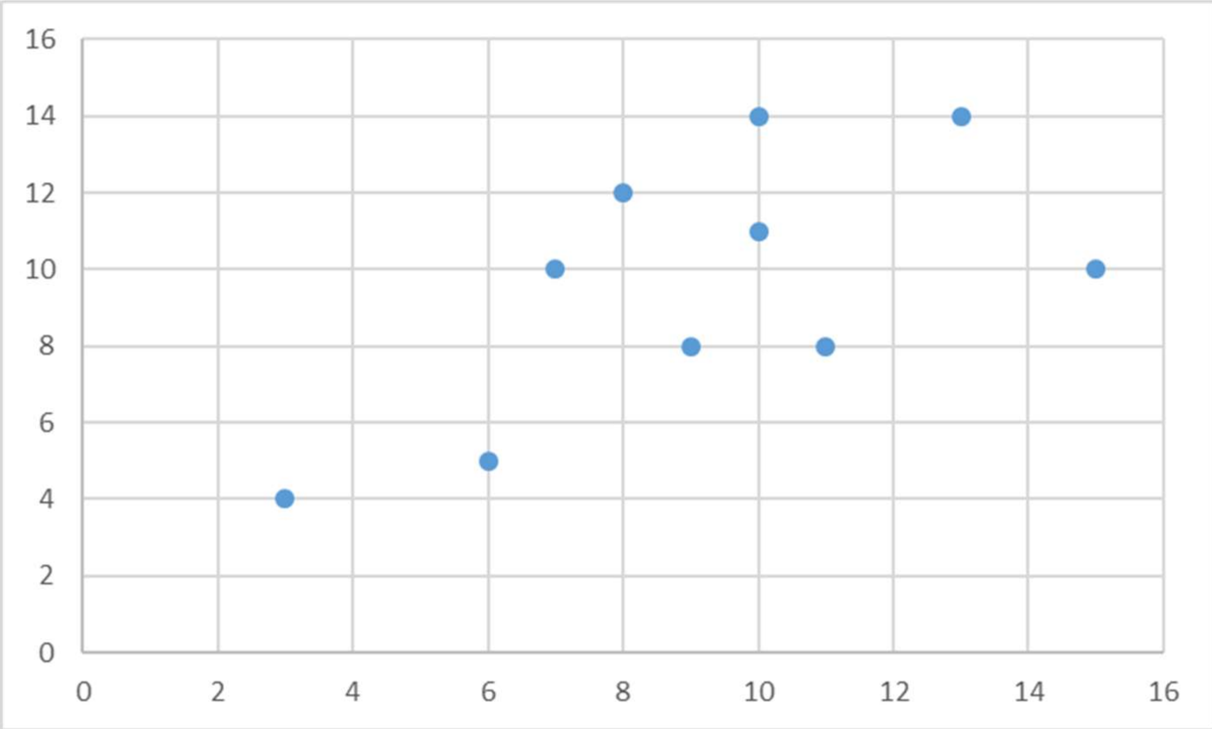
- $y_i = \beta_0 + \beta_1 x_i + u_i$
  - The estimated slope is:
  - $\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - Stated differently:
  - $\beta_1 = \frac{Cov[x,y]}{Var[x]} = \frac{\sigma_x\, \sigma_y \rho_{x,y}}{\sigma_x^2} = \frac{\sigma_y \rho_{x,y}}{\sigma_x}$
  - The intercept estimate $\beta_0 = \bar{y} - \beta_1 \bar{x}$

Deriving the OLS estimates

Data

x	y
3	4
11	8
6	5
7	10
10	11
9	8
15	10
8	12
10	14
13	14

Scatter Plot



Computations

x̄	ȳ
9.2	9.6

	(x-x̄)^2	(x-x̄)(y-ȳ)
	38.44	34.72
	3.24	-2.88
	10.24	14.72
	4.84	-0.88
	0.64	1.12
	0.04	0.32
	33.64	2.32
	1.44	-2.88
	0.64	3.52
	14.44	16.72
∑	107.6	66.8

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$= \frac{66.8}{107.6} = 0.621$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$
$$= 9.6 - 0.621 * 9.2 = 3.887$$

SAS code and output

SAS code:

```
data sample;
input x y;
datalines;
3	4
11	8
6	5
7	10
10	11
9	8
15	10
8	12
10	14
13	14
;
run;
proc reg data=sample;
model y=x;
run;
```

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	41.47063	41.47063	5.27	0.0508
Error	8	62.92937	7.86617		
Corrected Total	9	104.40000			

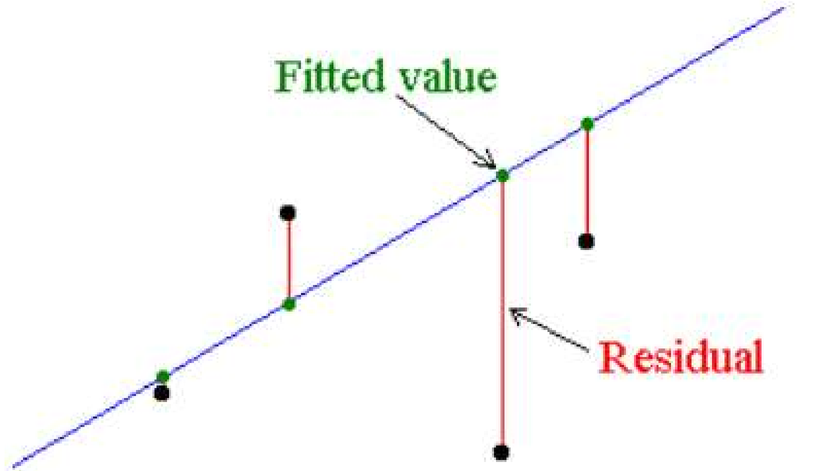
Root MSE	2.80467	R-Square	0.3972
Dependent Mean	9.60000	Adj R-Sq	0.3219
Coeff Var	29.21531		

Parameter Estimates

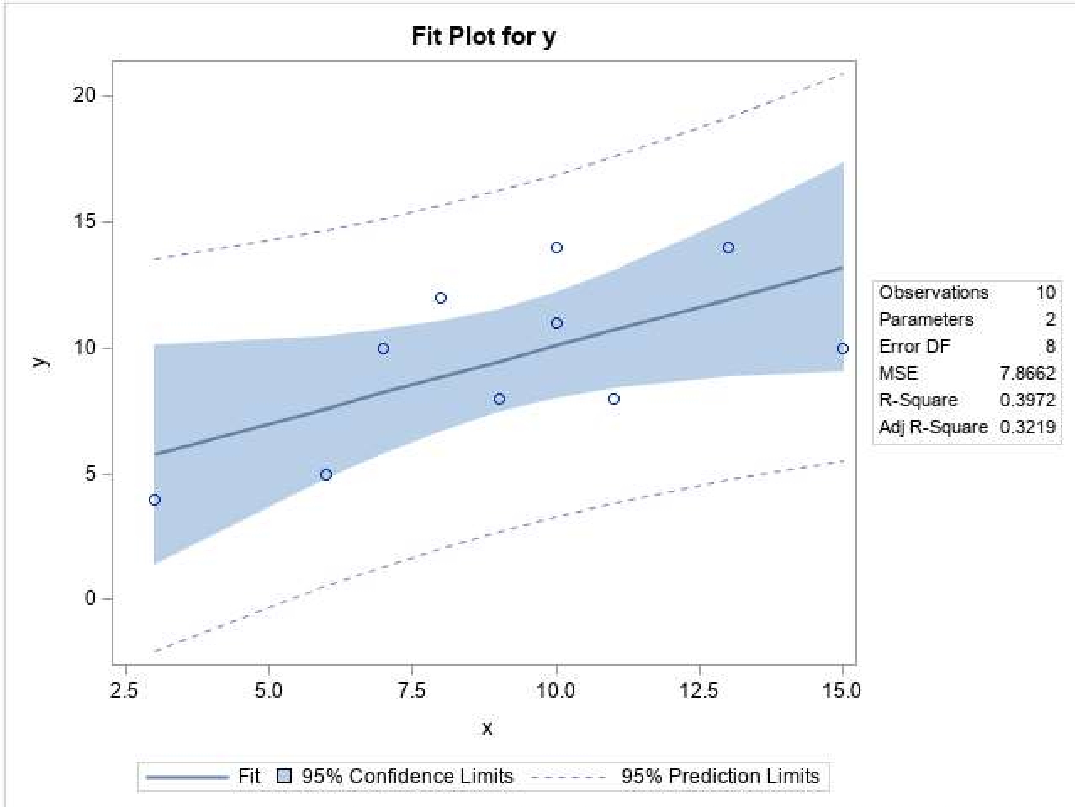
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.88848	2.64089	1.47	0.1791
x	1	0.62082	0.27038	2.30	0.0508

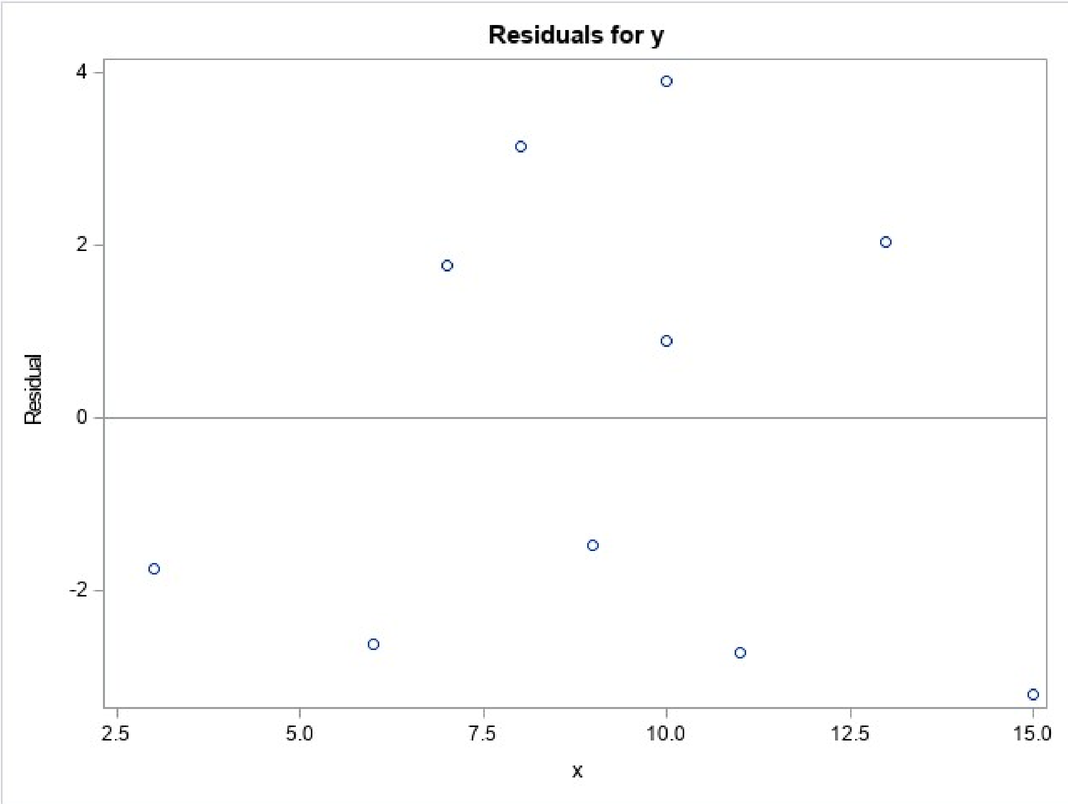
Fitted values and residuals

- β_0 and β_1 define a fitted value for y when $x = x_i$ as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- The residual for observation i is the difference between the actual y_i and its fitted value: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$



SAS output (continued)





A simple wage equation

- A model relating a person's wage to observed education is:
  - $wage = \beta_0 + \beta_1 educ + u$
  - If *wage* is measured in dollars per hour and *educ* is years of education, then β_1 measures the change in hourly wage given another year of education, holding all other factors fixed.
  - Is this the end of the causality issue?
  - How can we hope to learn in general about the ceteris paribus effect of *x* on *y,* holding other factors fixed, when we are ignoring all those other factors?

A simple wage equation

$$wage = \beta_0 + \beta_1 educ + u$$

- This formulation assumes change in wages is constant for all educational levels
- E.g., increasing education from 5 to 6 years leads to the same $ increase in wages as increasing education from 11 to 12, or 15 to 16, etc.
- Maybe a better assumption is that each year of education leads to a constant proportionate (i.e., percentage) increase in wages
- Approximation of this intuition captured by

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Log dependent variable - A simple wage equation

- Percentage change in wage given one unit increase in education is
  $\%\Delta wage \approx (100\beta)\ \Delta\ educ$
- Percent change in wage is constant for each additional year of education

$\Rightarrow$ Change in wage for an extra year of education increases as education increases.

  – I.e., increasing return to education (assuming $\beta > 0$)

- Log wage is linear in education. Wage is nonlinear
  $\log(wage) = \beta_0 + \beta_1 educ + u$
  $\Rightarrow wage = e^{(\beta_0 + \beta_1 educ + u\ )}$

Log dependent variable - A simple wage equation

- Sample of 526 individuals in 1976. Wages measured in $/hour.

Root MSE	0.48008	R-Square	0.1858
Dependent Mean	1.62327	Adj R-Sq	0.1843
Coeff Var	29.57481		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.58377	0.09734	6.00	<.0001
educ	educ	1	0.08274	0.00757	10.94	<.0001

- Interpretation:
  - Each additional year of education leads to an 8.3% increase in wages (NOT log(wages)!!!).
  - For someone with no education, their wage is exp(0.584)...this is meaningless because no one in sample has education=0.

CEO salary and return on equity

- For the population of chief executive officers (CEO), let *y* be annual salary (*salary)* in thousands of dollars.
- Let *x* be the average return on equity (*roe*) for the CEO's firm for the previous three years.
  - $salary = \beta_0 + \beta_1 roe + u$
  - The slope parameter β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point.
  - $\widehat{salary} = 963.191 + 18.501 roe$
  - How do we interpret the equation?

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	963.19133	213.24026	4.52	<.0001
roe	roe	1	18.50119	11.12325	1.66	0.0978

CEO salary and return on equity

- Table below contains a listing of the first 5 observations in the CEO data set, along with the fitted values, called *salaryhat*, and the residuals, called *uhat*.

Obs	roe	salary	salaryhat	uhat
1	14.1	1095	1224.06	-129.058
2	10.9	1001	1164.85	-163.854
3	23.5	1122	1397.97	-275.969
4	5.9	578	1072.35	-494.348
5	13.8	1368	1218.51	149.492

CEO salary and firm sales

- We can estimate a constant elasticity model relating CEO salary to firm sales.
- $\log(salary) = \beta_0 + \beta_1 \log(sales) + u$
  - where β_1 is the elasticity of *salary* with respect to *sales*.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4.82200	0.28834	16.72	<.0001
lsales	lsales	1	0.25667	0.03452	7.44	<.0001

CEO salary and firm sales

- $\log(salary) = \beta_0 + \beta_1 \log(sales) + u$

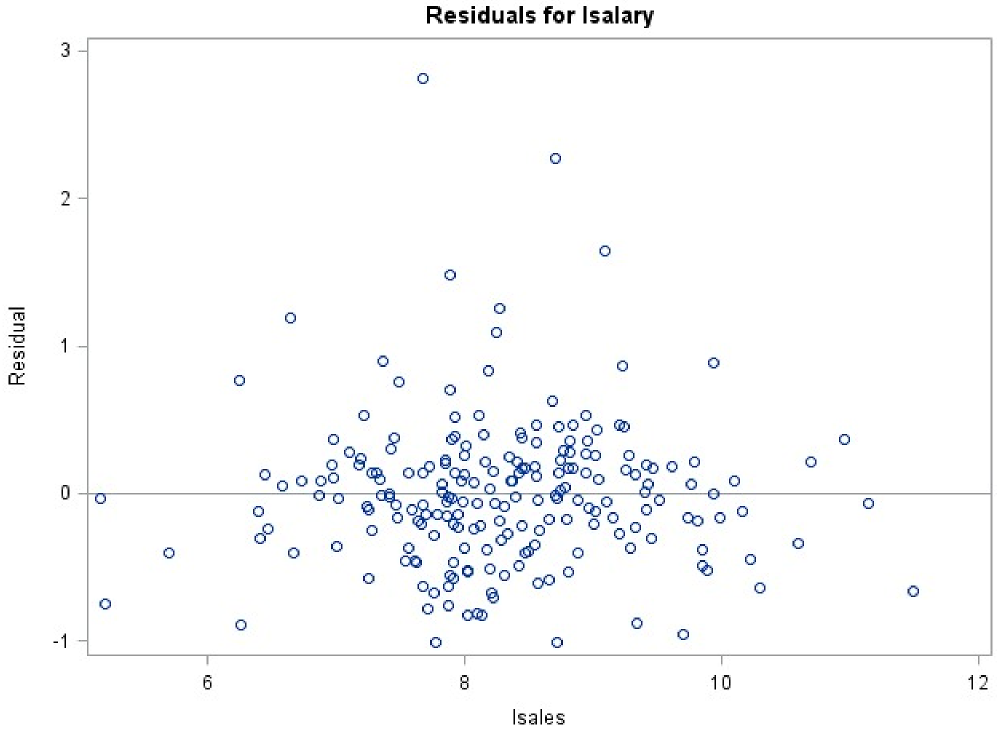
Number of Observations Read	209
Number of Observations Used	209

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	14.06617	14.06617	55.30	<.0001
Error	207	52.65600	0.25438		
Corrected Total	208	66.72217			

Root MSE	0.50436	R-Square	0.2108
Dependent Mean	6.95039	Adj R-Sq	0.2070
Coeff Var	7.25654		

CEO salary and firm sales

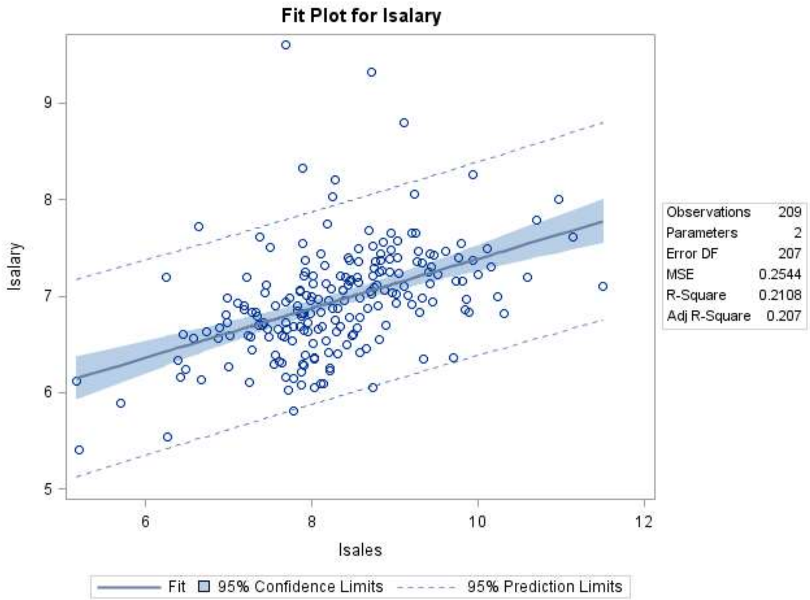
- $\log(salary) = \beta_0 + \beta_1 \log(sales) + u$



Constant elasticity model - CEO salary and firm sales

- $\log(salary) = \beta_0 + \beta_1 \log(sales) + u$

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4.82200	0.28834	16.72	<.0001
lsales	lsales	1	0.25667	0.03452	7.44	<.0001



- Interpretation: For each 1% increase in sales, salary increase by 0.257%

Interpret Regression Coefficient Estimates

Model	Interpretation of β
Level-Level Regression $y = \beta_0 + \beta_1 x + \varepsilon$	Δy=β1Δx "If you change x by one, we'd expect y to change by β1"
Log-Level Regression $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$	%Δy=100·β1·Δx "if we change x by 1 (unit), we'd expect our y variable to change by 100·β1 percent" Technically, the interpretation is the following: $\%\Delta y = 100 \times (e^{\beta_1} - 1)$
Level-Log Regression $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$	Δy=(β1/100)%Δx "If we increase x by one percent, we expect y to increase by (β1/100) units of y."
Log-Log Regression $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$	%Δy=β1%Δx "if we change x by one percent, we'd expect y to change by β1 percent"

PART II
Multiple Regression

Multiple regression analysis

- The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw *ceteris paribus* conclusions about how x affects y.
- Ceteris paribus definition: All other relevant factors are held fixed.
- The assumption that all other factors affecting y are uncorrelated with x is often unrealistic.
- Multiple regression analysis is more amenable to ceteris paribus analysis because it allows us to *explicitly* control for many other factors that simultaneously affect the dependent variable.

The model with two independent variables

- Much of the security price research has focused on the relation between prices and earnings, several empirical studies have adopted a balance sheet approach to relating accounting data to equity valuation.
- Under this approach, the market value of equity (MVE) equals the sum of the market values of assets (MVA) less the sum of the market value of liabilities (MVL).
- By virtue of the accounting identity, the book value of common equity (BVE) equals the book value of assets (BVA) less the book value of liabilities (BVL).
- In a simple setting of perfect and complete markets, market value of equity is a linear function of the market values of the individual assets and liabilities.

The model with two independent variables

- If there were no measurement error in book values, market value of equity would be a linear function of the book values, where the implied intercept (α) is zero, the implied coefficient on each asset and liability component (β) is one, and there is nothing left to be explained (in other words, the residual term, $u = 0$.
- $MVE_{it} = \alpha + \beta_1 BVA_{it} + \beta_2 BVL_{it} + u$
- In the presence of measurement error, the intercept term can be nonzero, the slopes can be different from one, and the residual term is nonzero.

Cross-sectional data set

- Textbook definition: *A data set collected by sampling a population at a given point in time.*

Fiscal Year	Company	Market Value	Total Assets	Total Liabilities
2010	Amazon.com Inc	81180	18797	11933
2010	Apple Inc	259906	75183	27392
2010	Boeing Co	47983	68565	65703
2010	Deere & Co	32423	43267	36963
2010	General Electric Co	194155	751216	627018
2010	HP Inc	92652	124503	83722
2010	International Business Machines Corp	180220	113452	90280
2010	Microsoft Corp	199451	86113	39938

Time series data

- Textbook definition: *Data collected over time on one or more variables.*

Fiscal Year	Company	Market Value	Total Assets	Total Liabilities
2005	3M Co	58477	20513	10102
2006	3M Co	57229	21294	11057
2007	3M Co	59796	24694	12622
2008	3M Co	39906	25547	15244
2009	3M Co	58745	27250	13948
2010	3M Co	61444	30156	14139
2011	3M Co	56800	31616	15754
2012	3M Co	63796	33876	15836
2013	3M Co	93027	33550	15602
2014	3M Co	104365	31269	18127
2015	3M Co	91789	32718	20971

Pooled cross section

- Textbook definition: *A data configuration where independent cross sections, usually collected at different points in time, are combined to produce a single data set.*

Fiscal Year	Company	Market Value	Total Assets	Total Liabilities
2010	International Business Machines Corp	180220	113452	90280
2011	International Business Machines Corp	213886	116433	96197
2012	International Business Machines Corp	214032	119213	100229
2013	International Business Machines Corp	197772	126223	103294
2014	International Business Machines Corp	158920	117532	105518
2015	International Business Machines Corp	132904	110495	96071
2010	Microsoft Corp	199451	86113	39938
2011	Microsoft Corp	217776	108704	51621
2012	Microsoft Corp	256375	121271	54908
2013	Microsoft Corp	287691	142431	63487
2014	Microsoft Corp	343566	172384	82600
2015	Microsoft Corp	354392	176223	96140

- A panel data set consists of a time series for *each* cross-sectional member in the data set.

R-SQUARED

- Textbook definition: *In a multiple regression model, the proportion of the total sample variation in the dependent variable that is explained by the independent variable.*

Goodness-of-fit (r-square)

- How to measure how well the explanatory variables explain the dependent variable?
- The *R*-squared of the regression is defined as

$R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ where

SSE = Sum of squares explained $= \sum_{i=1}^{N} (\hat{y}_i - \bar{y}_i)^2$

SST = Sum of squares total $= \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$

SSR = Sum of squares residual $= \sum_{i=1}^{N} (\hat{u}_i - \bar{u}_i)^2 = \sum_{i=1}^{N} \hat{u}_i^{\ 2}$

Interesting point (if the model includes only intercept and one explanatory variable): $R^2 = [Corr(y, \hat{y})]^2$

Pearson Correlation Coefficients, N = 209
Prob > |r| under H0: Rho=0

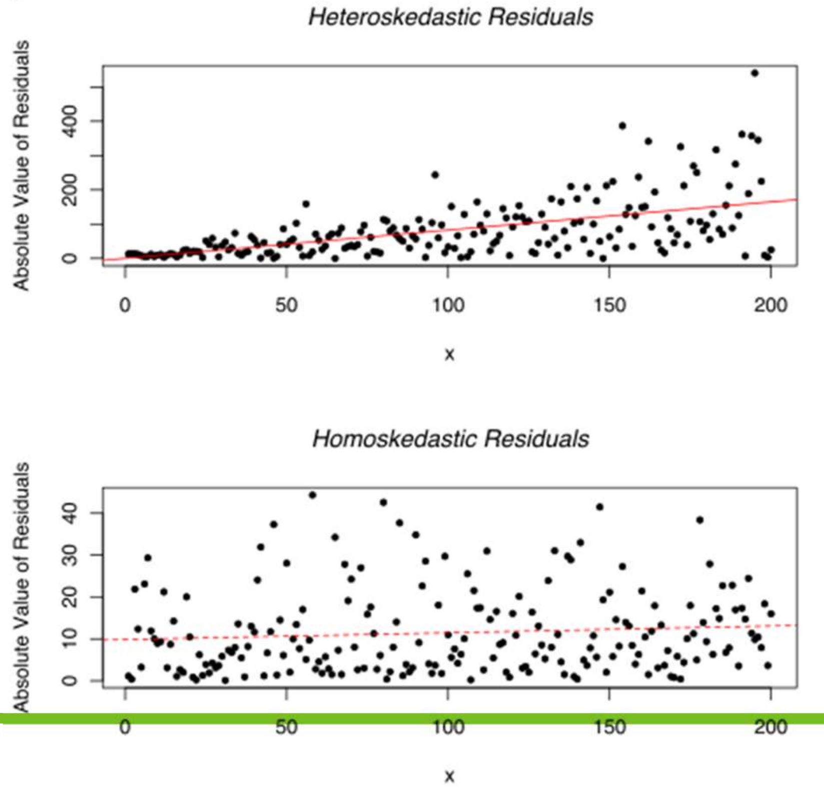
	lsalary	lsales
lsalary	1.00000	0.45915
lsalary		<.0001

Adjusted R-SQUARED

- Textbook definition: *A goodness-of-fit measure in multiple regression analysis that penalizes additional explanatory variables by using a degrees of freedom adjustment in estimating the error variance.*

Heteroscedasticity

- Textbook definition: *The variance of the error term, given the explanatory variables is not constant.*



Multicollinearity

- A term that refers to correlation among independent variables in a multiple regression model
  - For example, suppose we are interested in estimating the effect of various school expenditure categories on student performance. It is likely that expenditures on teacher salaries, instructional materials, athletics, and so on are highly correlated: Wealthier schools tend to spend more on everything, and poorer schools spend less on everything. Nor surprisingly, it can be difficult to estimate the effect of any particular expenditure category on student performance when there is little variation in one category that cannot largely be explained by variations in the other expenditure categories.

Multicollinearity

- The problem of multicollinearity cannot be clearly defined
- We cannot specify how much correlation among explanatory variables is ”too much”
- The most common diagnostic/statistic for individual coefficients is the **variance inflation factor (VIF)**
- We can try dropping other independent variables from the model in an effort to reduce multicollinearity but this can lead to bias
- Note that if our main interest is in the causal effect of x_1 on *y,* then we should ignore entirely the VIFs of other coefficients

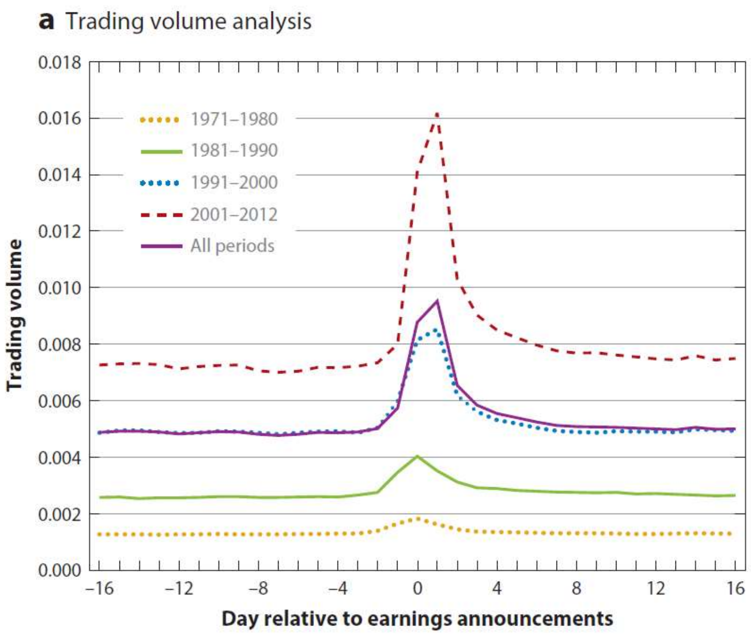
PART III
Accounting Examples

Early studies

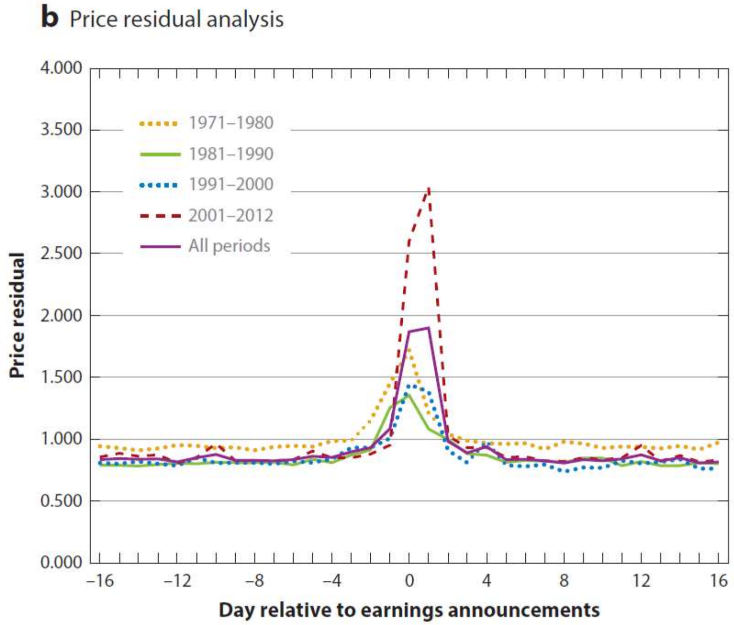
- Beaver (1968) and Ball and Brown (1968) are seminal papers that examine the usefulness of earnings.
  - Beaver (1968) investigates whether earnings announcements lead to significant increases in trading volume and stock price volatility.
  - Ball and Brown (1968) provide important evidence about the link between earnings and stock returns.

Beaver (1968)

TRADING VOLUME



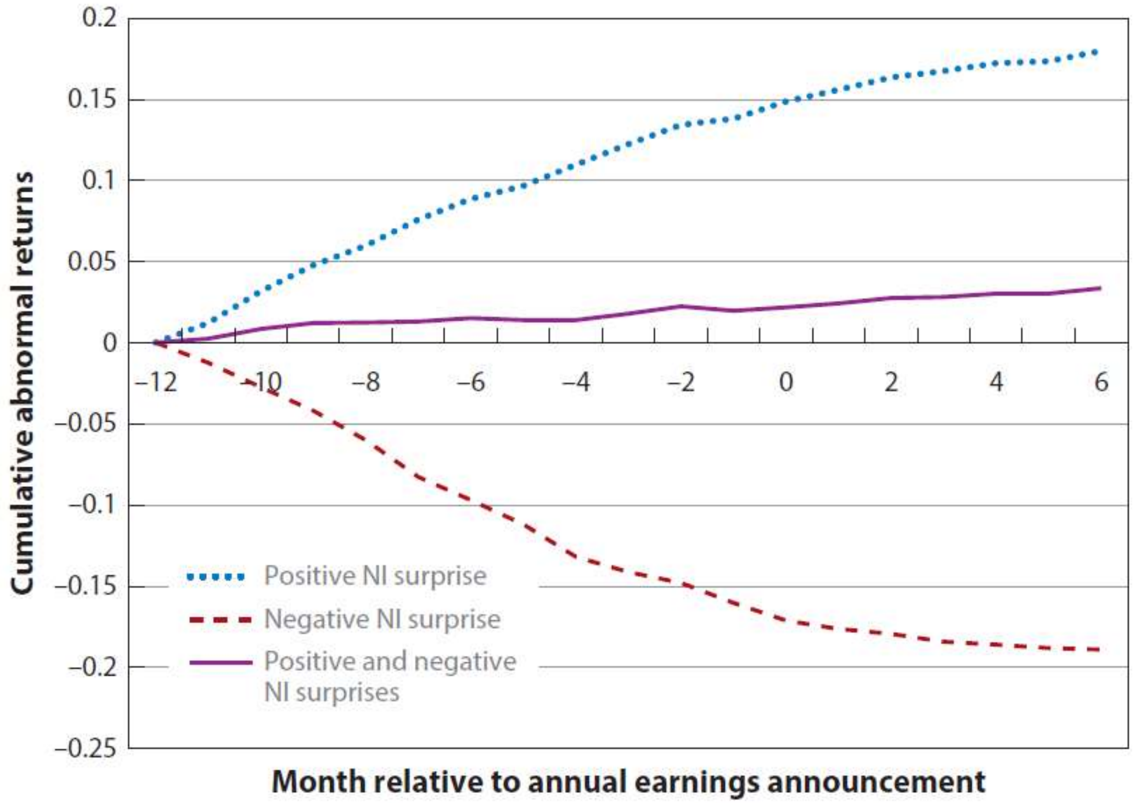
PRICE RESIDUAL



Details

This figure replicates Beaver's (1968) tests of volume and price reactions to earnings announcements over the past 40 years. We plot the (*a*) trading volume and (*b*) price residual during the [−16, 16] window surrounding the earnings announcement. Our sample is comprised of 762,032 firm-quarters from 1971 to 2012. Quarterly earnings announcement dates (RDQ) are available from Compustat starting in 1971. Volume and price information from CRSP must be non-missing from 16 days before the announcement date to 16 days after the announcement date. With less restrictive sample criteria than Beaver (1968), we include all fiscal year-end firms and both quarterly and annual earnings announcements. In panel *a*, trading volume reaction is calculated as the daily volume (VOL) divided by the number of shares outstanding (SHROUT) from CRSP. In panel *b*, price residual is calculated as u^2/s^2, where u^2 is the squared residual of the firm's daily return on the S&P Composite Index return, and s^2 is the variance of all firms' residuals from regressing returns on the S&P Composite Index return that day.

Ball and Brown (1968)



Details

Following Ball & Brown (1968), we plot the cumulative abnormal returns over the [−12, 6] month window for firms with positive (negative) annual earnings surprises and for all firms. Our sample is comprised of 165,224 firm-years from 1971 to 2012 that have non-missing earnings, returns, and earnings announcement dates. Annual earnings are measured as net income (NI) from Compustat. Cumulative abnormal returns are market-adjusted returns using the CRSP equal-weighted return, accumulated from 12 months before the earnings announcement month. Fourth-quarter earnings announcements dates (RDQ) are available from Compustat starting in 1971. Returns from CRSP must be non-missing from 12 months before the announcement month to 6 months after the announcement month. Our sample criteria are less restrictive than those of Ball & Brown (1968), who required CRSP price observations for 100 months and included only December fiscal year-end firms. Earnings surprise is the actual earnings minus the expected earnings ($X_t - E[X_t]$). We use the naïve model for earnings expectations such that a positive (negative) annual change in earnings defines a positive (negative) earnings surprise.

Value relevance

- How well accounting numbers reflect information used by equity investors?
- Many accounting papers investigate the empirical relation between stock market values (or changes in values) and particular accounting numbers for the purpose of assessing (or providing a basis of assessing) those numbers' use (or proposed use) in an accounting standard.
- The group of papers that are at least partially motivated by standard-setting purposes are called the ”value-relevance” literature.

Accounting relations and regression specifications

- The omission of relevant variables produces biased estimates (”the omitted variables bias”).
- A regression specification involving accounting numbers should be determined by the structure that delivers the numbers.
- The point can be illustrated by asking how the cost of goods sold (COGS) number on income statements is priced in the market: is it a reduction of the value of shareholders’ equity as the accounting prescribes?
- One might naively run the following cross-sectional regression using a levels specification:
  - $MVE_{it} = \alpha + \beta_1 COGS_{it} + \varepsilon$

Accounting relations and regression specifications

- Cost of goods sold is an expense (a reduction of shareholder value), yet the estimated slope coefficients from these equations are positive.
  - Using Compustat data from 1963 to 2001, the estimate of coefficient, β, is 1.12 (with a t-statistic of 13.52 calculated from mean estimates from annual cross-sectional regressions).
  - As a matter of statistical correlation, the estimates are appropriate, but they do not inform.
- Coefficients on included variables are affected by correlation with omitted information.

Accounting relations and regression specifications

- Cost of goods sold is part of the calculation of earnings; by accounting principle, it is involved with the sales with which it is matched to determine gross margin, so cost of goods sold cannot be considered without the matching sales
- Specifying regression under this dictate:
  - $MVE_{it} = \alpha + \beta_1 Sales_{it} + \beta_2 COGS_{it} + \varepsilon$
  - Now the estimated coefficient β_2 is reliably negative (–3.94 with a *t*-statistic of –17.74)
  - The estimate of β_1 is reliably positive (3.66)

PART IV
Treatment Effects

Treatment effects

- Consider the question ”Do hospitals make people healthier?”
- The results of a National Health Interview Survey included the questions:
  - ”During the past 12 months, was the respondent a patient in a hospital overnight?”
  - ”Would you say your health in general is excellent, very good, good, fair, or poor?”

Treatment effects (continued)

- Using the number 1 for poor health and 5 for excellent health, those *who had not* gone to the hospital had an average health score of 3.93, and those *who had been* to the hospital had an average score of 3.21.
- That is, individuals who had been to the hospital had poorer health than those who had not.

Treatment effects (discussion)

- Correlation is not the same as causation
  - We observe that those who had been in a hospital are less healthy, but observing this association does not imply that going to the hospital causes a person to be less healthy.
  - Data exhibit a selection bias, because some people chose (or self-selected) to go to the hospital and the others did not.
  - When membership in the treated group is in part determined by choice, then the sample is *not* a random sample.
  - There are systematic factors, in this case health status, contributing to the composition of the sample.

The difference estimator

- In order to understand the measurement of treatment effects, consider a simple regression model in which the explanatory variable is a dummy variable, indicating whether a particular individual is in the treatment or control group.
- Let y be the outcome variable, the measured characteristic the treatment is designed to effect.

The difference estimator

- Medical researchers use white mice to test new drugs, because these mice, surprisingly, are genetically similar to humans.
- Mice that are bred to be identical are randomly assigned to treatment and control groups, making estimation of the treatment effect of a new drug on the mice a relatively straightforward and reproducible process.
- Randomized controlled experiments in the social sciences are equally attractive from a statistician's point of view, but are rare because of the difficulties in organizing and funding them.
  - A notable example of a randomized experiment is Tennessee's Project STAR.

The difference estimator: project star

- The Tennessee class size project is a three-phase study designed to determine the effect of smaller class size in the earliest grades on short-term and long-term pupil performance.
  - A longitudinal experiment was conducted during 1985–1989.
  - A single cohort of students was followed from kindergarten through third grade.

The difference estimator: project star

- In the experiment chldren were randomly assigned within schools into three types of classes:
  1. Small class with 13–17 students
  2. Regular-sized classes with 22–25 students
  3. Regular-sized classes with a full-time teacher aide to assist the teacher
- Student scores on achievement tests were recorded, as was some information about the students, teachers, and schools.
- See the data file *star.xlsx*

The difference estimator: project star

- Descriptive statistics: A statistic used to summarize a set of numbers; the sample average, sample median, and sample standard deviation are the most common
- Let us first compare the performance of students in small classes versus regular classes.
  - The variable *TOTALSCORE* is the combined reading and math achievement scores and *SMALL* = 1 if the student was assigned to a small class, and zero if student is in regular class.
  - The average value of *TOTALSCORE* in the regular classes is 918.0429 and in small classes it is 931.9419, a difference of 13.899 points.
  - The test scores are higher in the smaller classes.

The difference estimator: project star

small=1

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
totalscore	totalscore	1738	931.9419	76.3586	747.0000	1253.0000
tchexper	tchexper	1738	8.9954	5.7316	0.0000	27.0000

regular=1

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
totalscore	totalscore	2005	918.0429	73.1380	635.0000	1229.0000
tchexper	tchexper	2005	9.0683	5.7244	0.0000	24.0000

The difference estimator: project star

- The difference estimator obtain using regression will yield the same estimate, along with significance levels.
  - $TOTALSCORE = \beta_0 + \beta_1 SMALL + \text{e}$

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	918.04289	1.66716	550.66	<.0001
small	small	1	13.89899	2.44659	5.68	<.0001

The difference estimator: project star

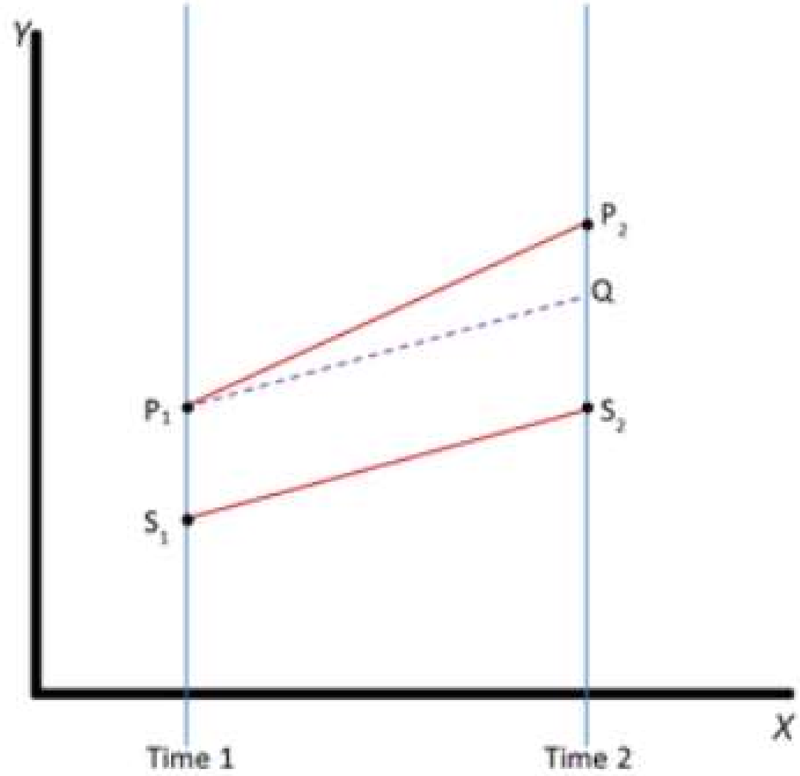
- Let's control for teaching experience.
  - $TOTALSCORE = \beta_0 + \beta_1 SMALL + \beta_1 TCHEXPER + \text{e}$

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	907.56434	2.54241	356.97	<.0001
small	small	1	13.98327	2.43733	5.74	<.0001
tchexper	tchexper	1	1.15551	0.21228	5.44	<.0001

The difference-in-differences estimator

- Suppose that we observe two groups before and after a policy change, with the treatment group being affected by the policy, and the control group being unaffected by the policy.
- Using such data, we will examine any change that occurs to the control group and compare it to the change in the treatment group.

The difference-in-differences



The difference-in-differences estimator

- We can isolate the effect of the treatment by using a control group that is not affected by the policy change.
- The treatment effect is:

$$\delta = (P_2 - S_2) - (P_1 - S_1)$$

- The estimator δ is called a differences-in-differences (or DID) estimator of the treatment effect.
- It can be conveniently calculated using a simple regression:
  - $y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 AFTER_t + \beta_3 (TREAT_i \times AFTER_t) + e$

PART V
Logistic Regression

A binary dependent variable

- So far we have discussed about the simple and multiple linear regression model.
- A binary (or dummy) variable is a variable that takes on the value zero or one (e.g., *Female_CEO* is a dummy variable that equals one if the CEO is a female, and zero otherwise).
- We have also studied how, through the use of binary independent variables, we can incorporate qualitative information as explanatory variables in a multiple regression model.
- What happens if we want to use multiple regression to *explain* a qualitative event?

A binary dependent variable

- What does it mean when y is a binary variable?
- $y = \beta_0 + \beta_1 x + u$
- Because y can take on only two values, β_1 cannot be interpreted as the change in y given a one-unit increase in x; y either changes from zero to one or from one to zero (or does not change).
- Note that when the OLS model includes only the intercept, β_0 is the predicted probability that the dependent variable equals one.

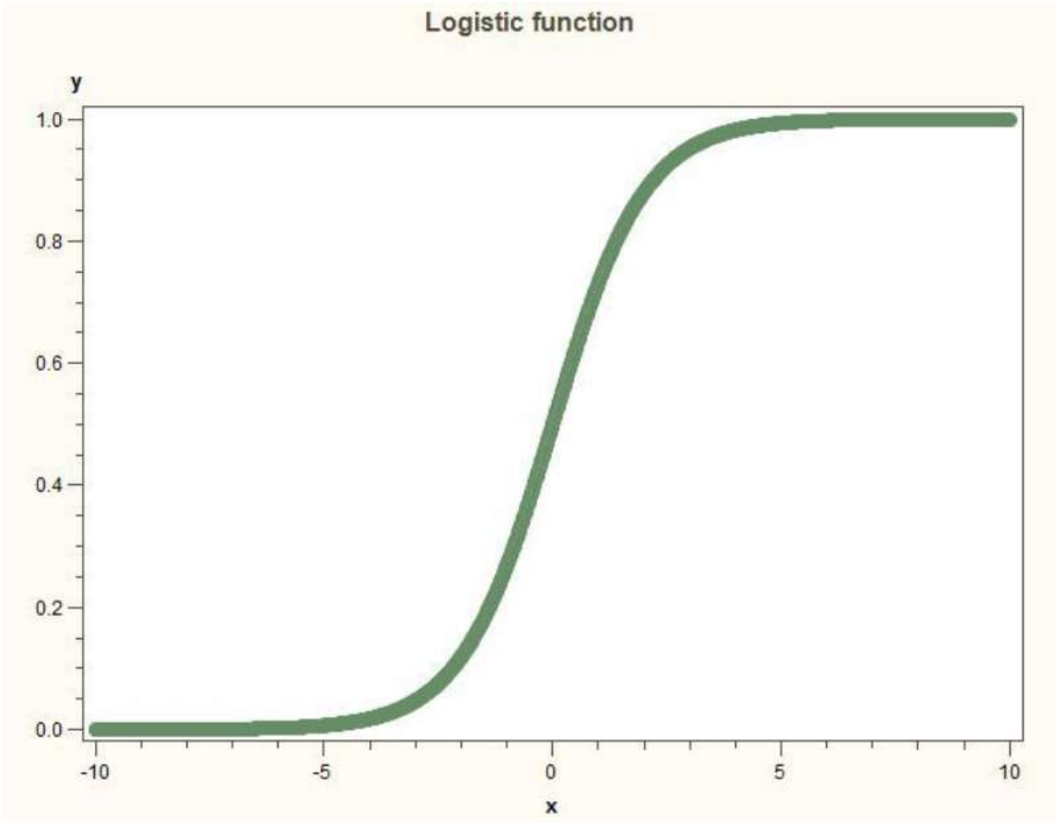
A binary dependent variable

- However, the OLS approach has two main drawbacks:
  1. The fitted probabilities can be less than zero or greater than one
  2. The partial effect of any explanatory variable (appearing in level form) is constant
- **Binary response models** overcome these limitations
- In a binary response models, interest lies primarily in the **response probability**
  – $P(y = 1|\boldsymbol{X}) = P(y = 1|x_1, x_2, \dots, x_k)$ where $\boldsymbol{X}$ denote the full set (vector) of explanatory variables

Logit model

- Binary response models ensure that the estimated response probabilities are strickly between zero and one.
  - $P(y = 1|\boldsymbol{x}) = \mathrm{G}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = G(\boldsymbol{X\beta})$
- We will cover here only the **logit model**
  - $G(X\beta) = \frac{e^{\boldsymbol{X\beta}}}{1+e^{\boldsymbol{X\beta}}} = \frac{1}{1+e^{-\boldsymbol{X\beta}}}$
  - $logit(p) = \ln\left(\frac{p}{1-p}\right) = \boldsymbol{X\beta}$ (equivalent representation)
  - As you can see, logit is a non-linear model (just like Probit)
  - Logit regression models the logit-transformed probability as a *linear* relationship with the predictor variables

Graph of the logistic function



$$G(X\beta) = \frac{e^{\boldsymbol{X\beta}}}{1 + e^{\boldsymbol{X\beta}}}$$

Example: married women's labor force participation

- We now use MROZ.xlsx data to estimate the labor force participation logit model (Mroz 1987 *Econometrica*)
- Let *inlf* (”in the labor force”) be a binary variable indicating labor force participation by a married woman during 1975.
- We assume that labor force participation depends on other sources of income, including husband's earnings, years of education, past years of labor market experience, age, number of children less than six years old, and number of kids between 6 and 18 years of age.

MROZ Variable Description

```
 1. inlf                    =1 if in labor force, 1975
 2. hours                   hours worked, 1975
 3. kidslt6                 # kids < 6 years
 4. kidsge6                 # kids 6-18
 5. age                     woman's age in yrs
 6. educ                    years of schooling
 7. wage                    estimated wage from earns., hours
 8. repwage                 reported wage at interview in 1976
 9. hushrs                  hours worked by husband, 1975
10. husage                  husband's age
11. huseduc                 husband's years of schooling
12. huswage                 husband's hourly wage, 1975
13. faminc                  family income, 1975
14. mtr                     fed. marginal tax rate facing woman
15. motheduc                mother's years of schooling
16. fatheduc                father's years of schooling
17. unem                    unem. rate in county of resid.
18. city                    =1 if live in SMSA
19. exper                   actual labor mkt exper
20. nwifeinc                (faminc - wage*hours)/1000
21. lwage                   log(wage)
22. expersq                 exper^2
```

Descriptive statistics

Variable	N	Mean	Std Dev	Minimum	Median	Maximum
inlf	753	0.568	0.496	0.000	1.000	1.000
nwifeinc	753	20.129	11.635	-0.029	17.700	96.000
educ	753	12.287	2.280	5.000	12.000	17.000
exper	753	10.631	8.069	0.000	9.000	45.000
expersq	753	178.039	249.631	0.000	81.000	2025.000
age	753	42.538	8.073	30.000	43.000	60.000
kidslt6	753	0.238	0.524	0.000	0.000	3.000
kidsge6	753	1.353	1.320	0.000	1.000	8.000

Logit estimates of labor force participation

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4255	0.8604	0.2445	0.6210
nwifeinc	1	-0.0213	0.00842	6.4243	0.0113
educ	1	0.2212	0.0434	25.9228	<.0001
exper	1	0.2059	0.0321	41.2421	<.0001
expersq	1	-0.00315	0.00102	9.6354	0.0019
age	1	-0.0880	0.0146	36.4844	<.0001
kidslt6	1	-1.4434	0.2036	50.2637	<.0001
kidsge6	1	0.0601	0.0748	0.6460	0.4215

labor force participation

- What is the expected probability on being in labor force for a ”median” woman?
  - $P(labor\ force\ participation = 1) = \frac{1}{1+e^{-X\beta}}$ where $X\beta = \hat{\beta}_0 + \hat{\beta}_1$nwifeinc+$\hat{\beta}_2$educ+$\hat{\beta}_3$exper+$\hat{\beta}_4$expersq+$\hat{\beta}_5$age+$\hat{\beta}_6$kidslt6+$\hat{\beta}_7$kidsge6
  - *z=0.4255+(-0.0213)*17.7+0.2212*12+0.2059*9+(-0.00315)*81+(-0.088)*43+(-1.4434)*0+0.0601*1=0.57694*
  - $P(labor\ force\ participation = 1) = \frac{1}{1+e^{-0.57694}} = 0.6404$ (vs. the unconditional probability of 0.568)

Interpretation and Odds ratios

- The odds ratio refers to the ratio of two odds
- The estimated coefficient measures the effect of a unit increase in *x* on the logarithm of the odds ratio of *p,* while holding other independent variables unchanged.
- The interpretation of estimated coefficients for fitted logit function is different from the interpretation of slope coefficients in an OLS model!
- The estimated odds ratio equals:
  - $\frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = e^{[\widehat{\beta}(x_1-x_0)]}$

Interpretation and Odds ratios

- SAS outputs also the odds ratios ($e^{\widehat{\beta}_i}$)

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
nwifeinc	0.979	0.963	0.995
educ	1.248	1.146	1.358
exper	1.229	1.154	1.308
expersq	0.997	0.995	0.999
age	0.916	0.890	0.942
kidslt6	0.236	0.158	0.352
kidsge6	1.062	0.917	1.230

Interpretation

- The figure plots the change in probability $p_1 - p_0$ against the baseline probability p_0 for a selection of positive effect sizes $\beta(x_1 - x_0)$