Econometrics I

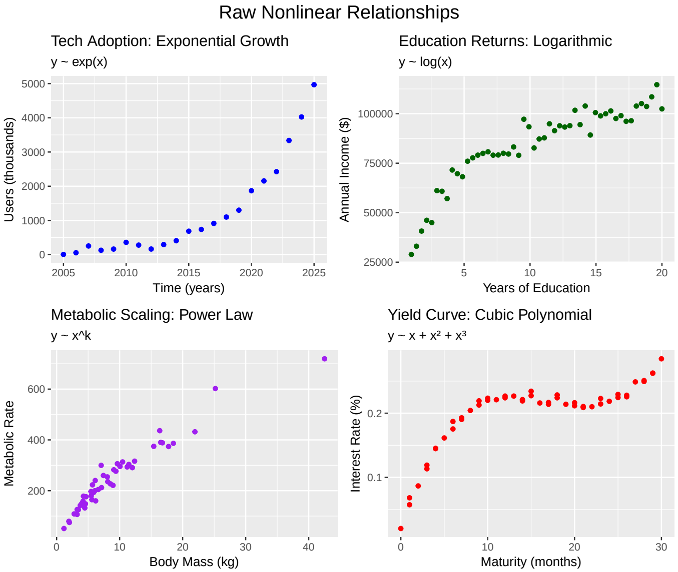
Lecture 3

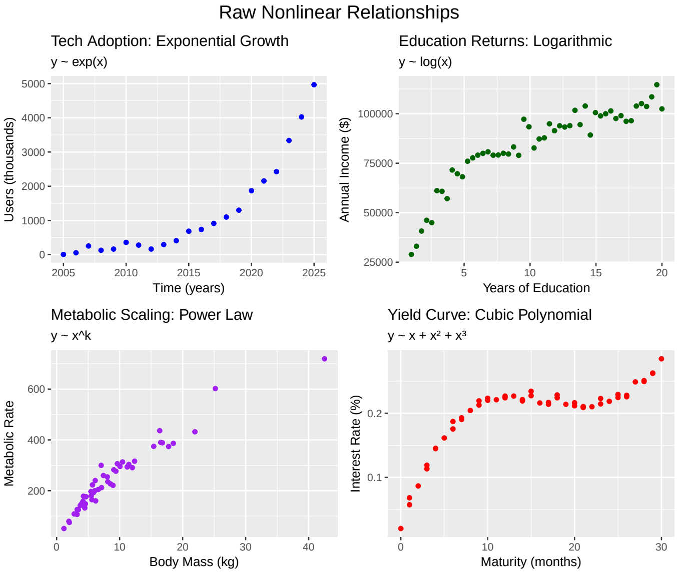
Transformations, outliers and fit

Matías Cabello
matias.cabello@wiwi.uni-halle.de

October 26, 2025

Transformations

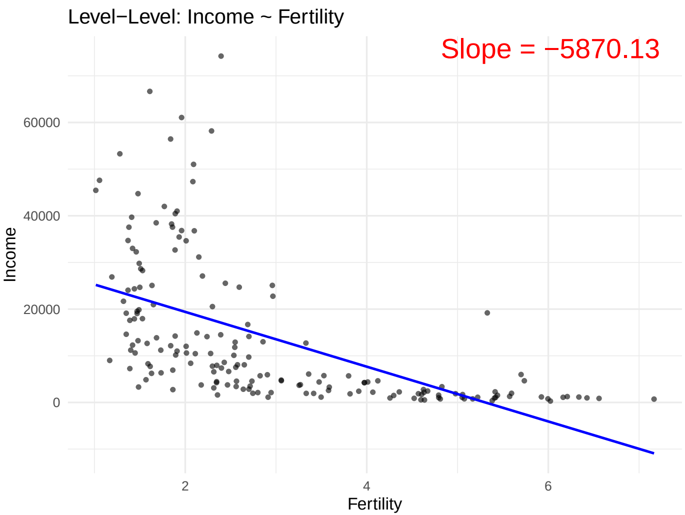

Raw Nonlinear Relationships
Tech Adoption: Exponential Growth
y ~ exp(x)
Users (thousands)
Time (years)
Education Returns: Logarithmic
y ~ log(x)
Annual Income ($)
Years of Education
Metabolic Scaling: Power Law
y ~ x^k
Metabolic Rate
Body Mass (kg)
Yield Curve: Cubic Polynomial
y ~ x + x² + x³
Interest Rate (%)
Maturity (months)


Raw Nonlinear Relationships
Tech Adoption: Exponential Growth
y ~ exp(x)
Users (thousands)
Time (years)
Education Returns: Logarithmic
y ~ log(x)
Annual Income ($)
Years of Education
Metabolic Scaling: Power Law
y ~ x^k
Metabolic Rate
Body Mass (kg)
Yield Curve: Cubic Polynomial
y ~ x + x² + x³
Interest Rate (%)
Maturity (months)

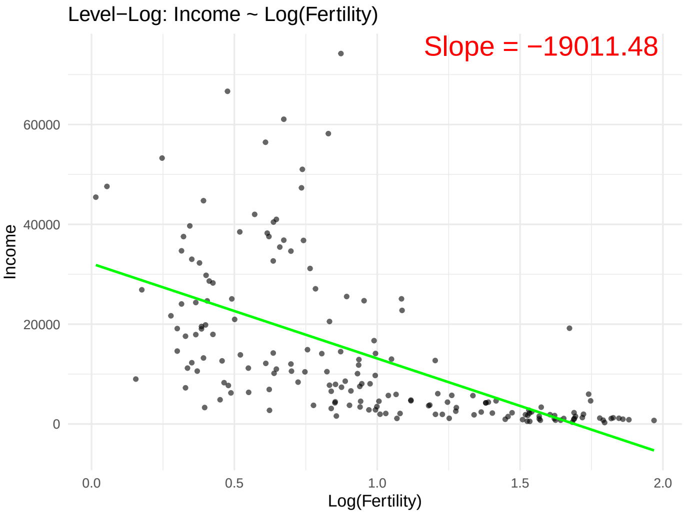
Typical transformations

Name	Specification	Total Differential	Interpretation
Level-Level	$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$	$\Delta\hat{y} = \hat{\beta}_2 \Delta x$	A one-unit increase in x increases $\hat{y}$ by $\hat{\beta}_2$ units.
Log-Log	$\ln \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln x$	$\frac{\Delta\hat{y}}{\hat{y}} \approx \hat{\beta}_2 \frac{\Delta x}{x}$	A 1% increase in x increases $\hat{y}$ by $\hat{\beta}_2$%.
Level-Log	$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln x$	$\Delta\hat{y} \approx \hat{\beta}_2 \frac{\Delta x}{x}$	A 1% increase in x increases $\hat{y}$ by $\hat{\beta}_2/100$ units.
Log-Level	$\ln \hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$	$\frac{\Delta\hat{y}}{\hat{y}} \approx \hat{\beta}_2 \Delta x$	A one-unit increase in x increases $\hat{y}$ by $100 \times \hat{\beta}_2$ %.

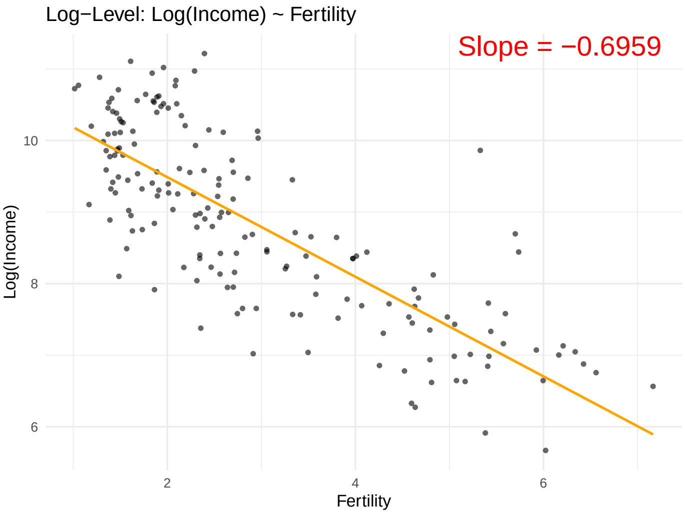
Typical transformations



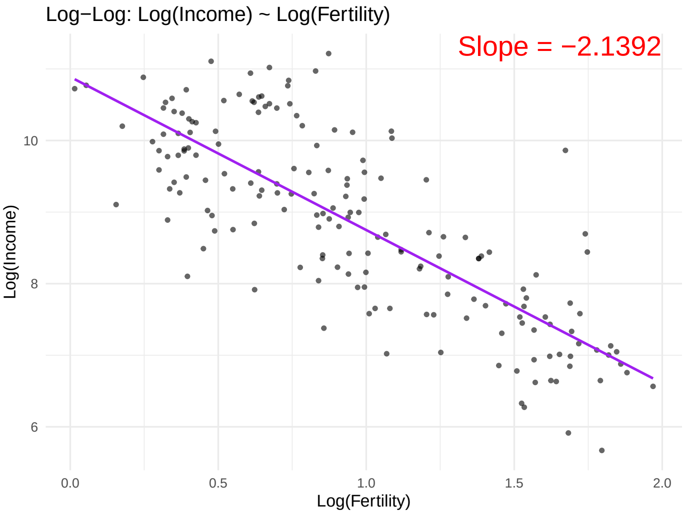
Typical transformations



Typical transformations



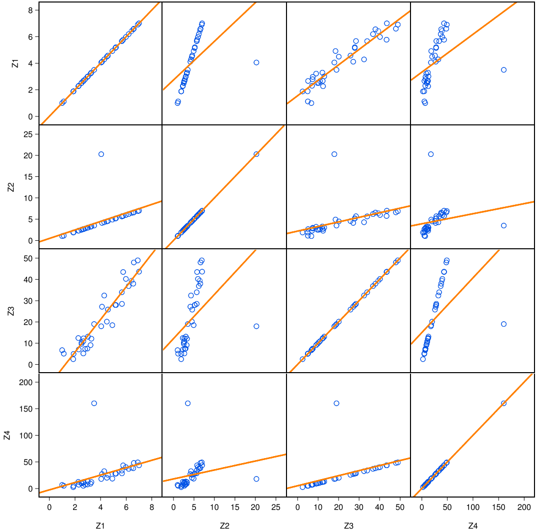
Typical transformations



Outliers

- Outlier = an observation that does not fit the data's overall pattern
- Possible causes:
  - A typo or error in the data
  - A particularly interesting and informative data point
- Should an outlier be removed or included?

When are outliers dangerous?



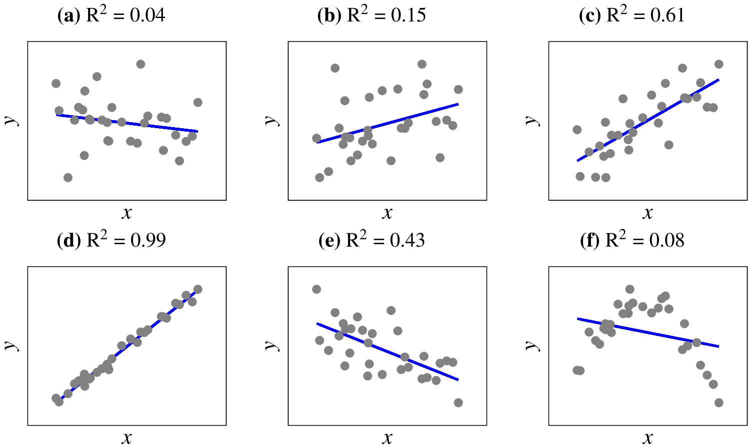
When are outliers dangerous?

- Note that outliers are particularly dangerous when far from the mean in terms of x only.
- y-outliers not dangerous.
- Evaluate if typo/error or needs to be kept in the regression.

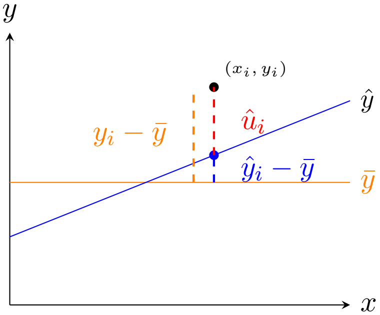
Goodness of fit (R^2)

The coefficient of determination (R^2)

The R^2 tries to capture the goodness of fit in one number between 0 and 1.



The coefficient of determination (R^2)



$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ with $\hat{\beta}_2 = \frac{\text{cov}(x,y)}{\text{var}(x)}$

$$\boxed{R^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (y_i - \bar{y})^2}}$$

Note that the distance $y_i - \bar{y}$ can be decomposed as $\hat{y}_i - \bar{y} + \hat{u}_i$.

- **Perfect fit:** All points fall exactly on $\hat{y}$; hence $\hat{u}_i = 0$ and $y_i - \bar{y} = \hat{y}_i - \bar{y}$ for all $i \implies \boxed{R^2 = 1}$
- **Worst-possible fit:** Zero covariance; hence $\hat{\beta}_2 = 0$ and $y_i - \bar{y} = \hat{u}_i \implies \boxed{R^2 = 0}$.

The coefficient of determination (R^2)

In terms of the **residual** sum of squares (RSS $= \sum_i \hat{u}_i^2$):

$$R^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\frac{1}{n}\sum_i (\hat{u}_i - 0)^2}{\frac{1}{n}\sum_i (y_i - \bar{y})^2} = 1 - \frac{\mathrm{var}(\hat{u})}{\mathrm{var}(y)}$$

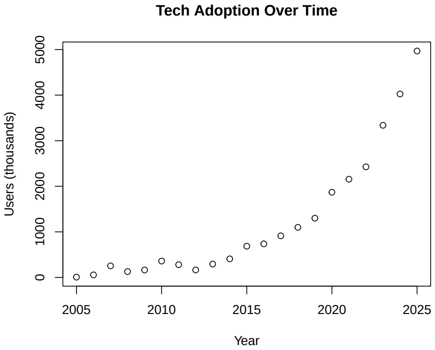
In terms of the **explained** sum of squares (ESS $= \sum_i (\hat{y}_i - \bar{y})^2$):

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_i (y_i - \bar{y})^2} = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)}$$

Interpretation

- $R^2 =$ fraction of explained variance (the ESS) over the total variance (the TSS $= \sum_i (y_i - \bar{y})^2$).
- $R^2 =$ how much of y's variance fitted by the model ($\hat{y}$).
- $R^2 = 1$: Perfect fit; $R^2 = 0$: No explanatory power

Example: R^2 after transformation




```r
# Run linear regression
model_linear <- lm(Users ~ Year, data = tech)

# Run log-linear model
model_log <- lm(log(Users) ~ Year, data = tech)
```

Example: R^2 after transformation

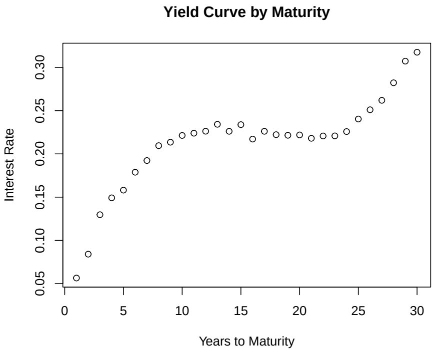
Tech Adoption: Linear vs Log-Linear Models

	Dependent variable:	
	Users	log(Users)
	Linear	Log-Linear
	(1)	(2)
Year	200.094***	0.241***
	(25.205)	(0.022)
Constant	-401,968.400***	-479.229***
	(50,787.630)	(44.104)
Observations	21	21
R2	0.768	0.864
Note:	*p<0.1; **p<0.05; ***p<0.01	

Notice difference in R^2. (Interpretation of slope in (2)? $\rightarrow$ Users increase by 24% each year!)

Example: Polynomial regression

Cannot be transformed with logs, etc.:



Then estimate polynomial regression:

$$\text{yield}_i = \hat{\beta}_1 + \hat{\beta}_2 \text{maturity}_i + \hat{\beta}_3 \text{maturity}_i^2 + \hat{\beta}_4 \text{maturity}_i^3 + \hat{u}_i$$

Example: Polynomial regression

Polynomial regression:

```
# Linear model for yield
model_linear_yield <- lm(Yield ~ Maturity, data = yield)

# Cubic polynomial model
model_poly <- lm(Yield ~ Maturity + I(Maturity^2) + I(
    Maturity^3), data = yield)
```

Comparing linear and polynomial models:

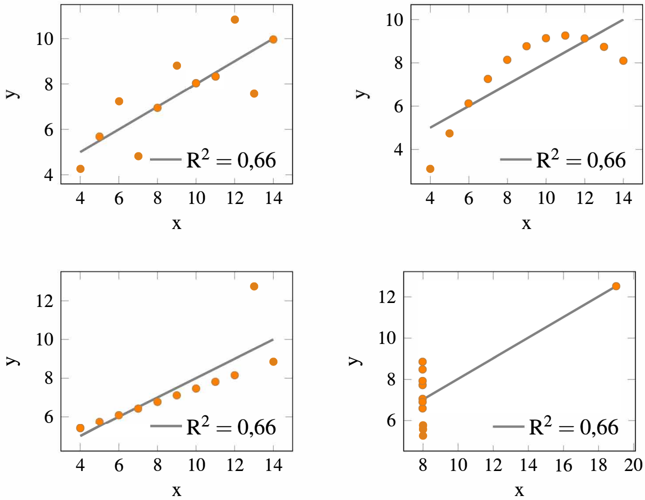
```
stargazer(yield_lm_linear, yield_lm,
          type = "text",
          title = "Yield␣Curve:␣Linear␣vs␣Polynomial␣
              Models",
          column.labels = c("Linear", "Cubic"),
          dep.var.labels = c("Yield", "Yield"))
```

Example: Polynomial regression

	Dependent variable: Yield	
	Linear	Cubic
Maturity	0.005***	0.041***
	(0.001)	(0.001)
I(Maturity2)		-0.003***
		(0.0001)
I(Maturity3)		0.0001***
		(0.00000)
Constant	0.131***	0.018***
	(0.011)	(0.004)
Observations	30	30
R2	0.715	0.993
Note:		*p<0.1; **p<0.05; ***p<0.01

Notice difference in R^2: 71.5% vs. 99.3%.

But R^2 has important limitations



But R^2 has important limitations

Most importantly: **Correlation $\neq$ Causality !!**

1. Direct causality $\longrightarrow$ $\mathbf{x}$ causes $\mathbf{y}$.
2. Reverse causality $\longrightarrow$ $\mathbf{y}$ causes $\mathbf{x}$.
3. Simultaneous causality $\longrightarrow$ $\mathbf{x}$ causes $\mathbf{y}$ and $\mathbf{y}$ causes $\mathbf{x}$.
4. Spurious correlation $\longrightarrow$ Either by pure chance (when samples are small) or when both $\mathbf{x}$ and $\mathbf{y}$ are caused by a common factor $\mathbf{z}$ (called 'confounder').

Takeaways

Key takeaways

You should now know:

- Nonlinear relationships $\rightarrow$ transformations or polynomial regression
- Interpretation of coefficients with logs: $\%$ change
- Outliers: especially dangerous when far from $\bar{x}$.
- R^2: fraction of y's variance fitted
- But correlation $\neq$ causality