Specification and Data Issues: Part II

Statistics and Econometrics

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Roadmap

- Regression analysis with cross-sectional data
  - Basics: estimation, inference, analysis with dummy variables
  - More involved: model specification and data issues
- Advanced topics
  - Binary dependent variable models
  - Panel data analysis
  - Time series analysis

Outline (Wooldridge, Chap. 3.3, 3.4, 5.2, 9.5)

- Model diagnostics
- Outliers
- A possible model fitting strategy

Outline

- Model diagnostics
- Outliers
- A possible model fitting strategy

Statistical Properties of OLS Estimators

Theorem (3.1)

With a "good" model, the OLS estimators are unbiased, i.e., $E(\hat{\beta}_j) = \beta_j$, $j = 0, 1, \ldots, k$

Theorem (4.1, Normal Sampling Distribution)

With a "good" model,

$$\hat{\beta}_j \sim Normal\left(\beta_j, Var(\hat{\beta}_j)\right),$$

where the variance is given by

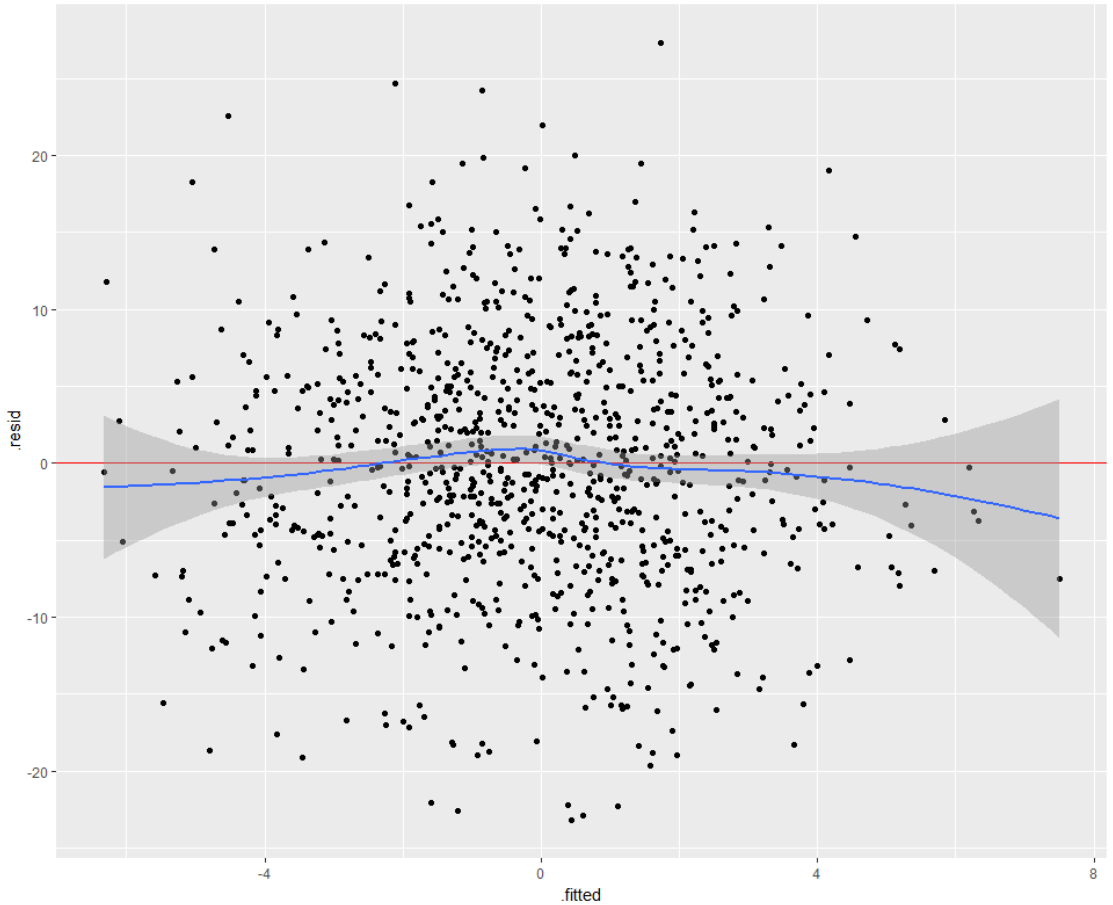
$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \qquad j = 1, \ldots, k.$$

Gauss-Markov Assumptions

- [MLR1] (linear in parameters) In the population model, y is related to x's by $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$, where $(\beta_0, \beta_1, \ldots, \beta_k)$ are population parameters and u is disturbance
  - Common causes lead to violation of this assumption
    - Functional form misspecification: log vs level form, omitting quadratic term
  - Identification of assumption violation
    - RESET, Residual plots

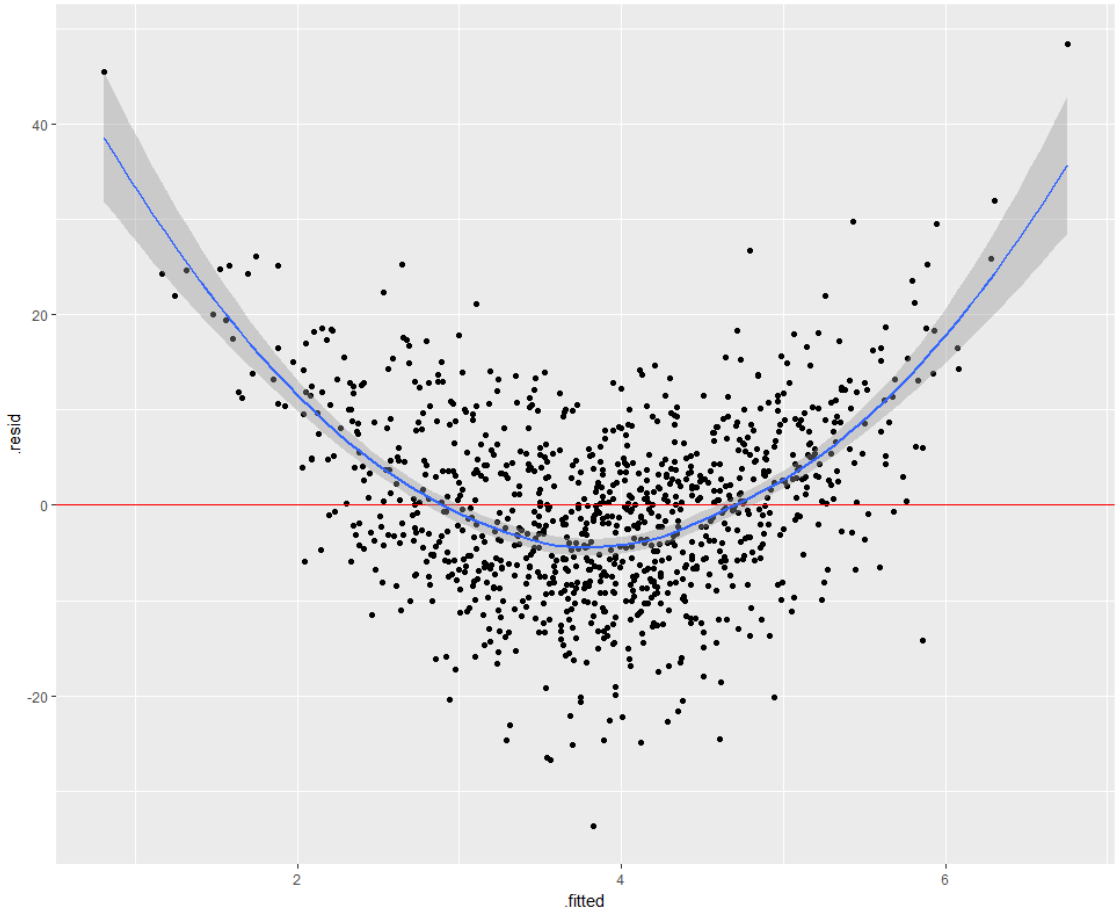
Residual Plots: A Correctly Specified Model

```
> x <- rnorm(1000, mean = 0, sd = 2)
> y <- x + rnorm(1000, mean = 0, sd = 8)
> m1 <- lm(y ~ x)
> ggplot(m1, aes(.fitted, .resid)) + geom_point() + geom_hline(
    yintercept=0, col="red") + stat_smooth(method = "loess")
```



Residual Plots: A Misspecified Model

```
> x <- rnorm(1000, mean = 0, sd = 2)
> y <- x + x^2 + rnorm(1000, mean = 0, sd = 8)
> m2 <- lm(y ~ x)
> ggplot(m2, aes(.fitted, .resid)) + geom_point() + geom_hline(
    yintercept=0, col="red") + stat_smooth(method = "loess")
```



Gauss-Markov Assumptions

- [MLR2] (random sampling) $\{(x_{i1}, \ldots, x_{ik}, y_i), i = 1, 2, \ldots, n\}$ with $n \geq k + 1$ is a random sample drawn from the population model

Missing Data

- If any observation has missing data on one of the variables in the model, it cannot be used.
- Would this practice cause problems?
  - If data is missing at random, then the only consequence is a reduction in the sample size
  - A problem can arise if the data is missing in a systematic way. The sample becomes nonrandom (violation to MLR2)

Nonrandom Samples

- Exogenous sample selection
  - If the sample is chosen on the basis of an independent variable x, the OLS estimators will still be unbiased
  - Eg. Consider the birth weight model (`bwght.RData`)

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

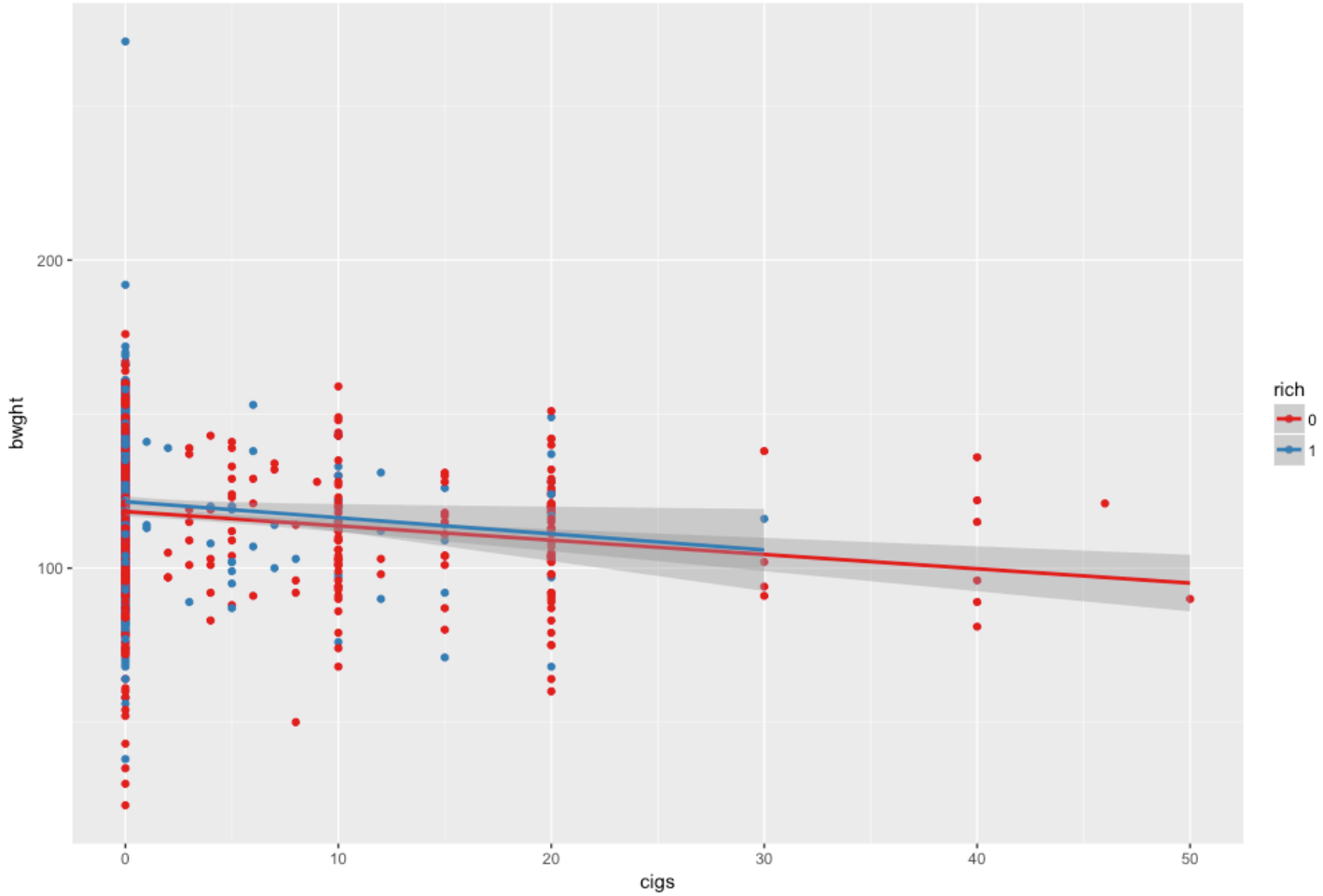
    Suppose the data set is based on a survey of families with annual income of £$30,000$ and over
  - While the sample is nonrandom, zero-conditional mean assumption still holds as

$$E(u|cigs, faminc) = 0,$$

    for any subset of $(cigs, faminc)$

Nonrandom Samples: Exogenous sample selection

```
> data <- data %>% mutate(rich = ifelse(faminc > median(faminc), 1, 0))
> data$rich <- as.factor(data$rich)
> ggplot(data, aes(x = cigs, y = bwght, color = rich)) + geom_point() +
    geom_smooth(method='lm') + scale_color_brewer(palette="Set1")
```



Nonrandom Samples

- Endogenous sample selection
  - If the sample is chosen on the basis of the dependent variable y, the OLS estimators will be biased
  - Eg. Again consider the birth weight model (`bwght.RData`)

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

Suppose the sample only includes infants lighter than 3 kilograms

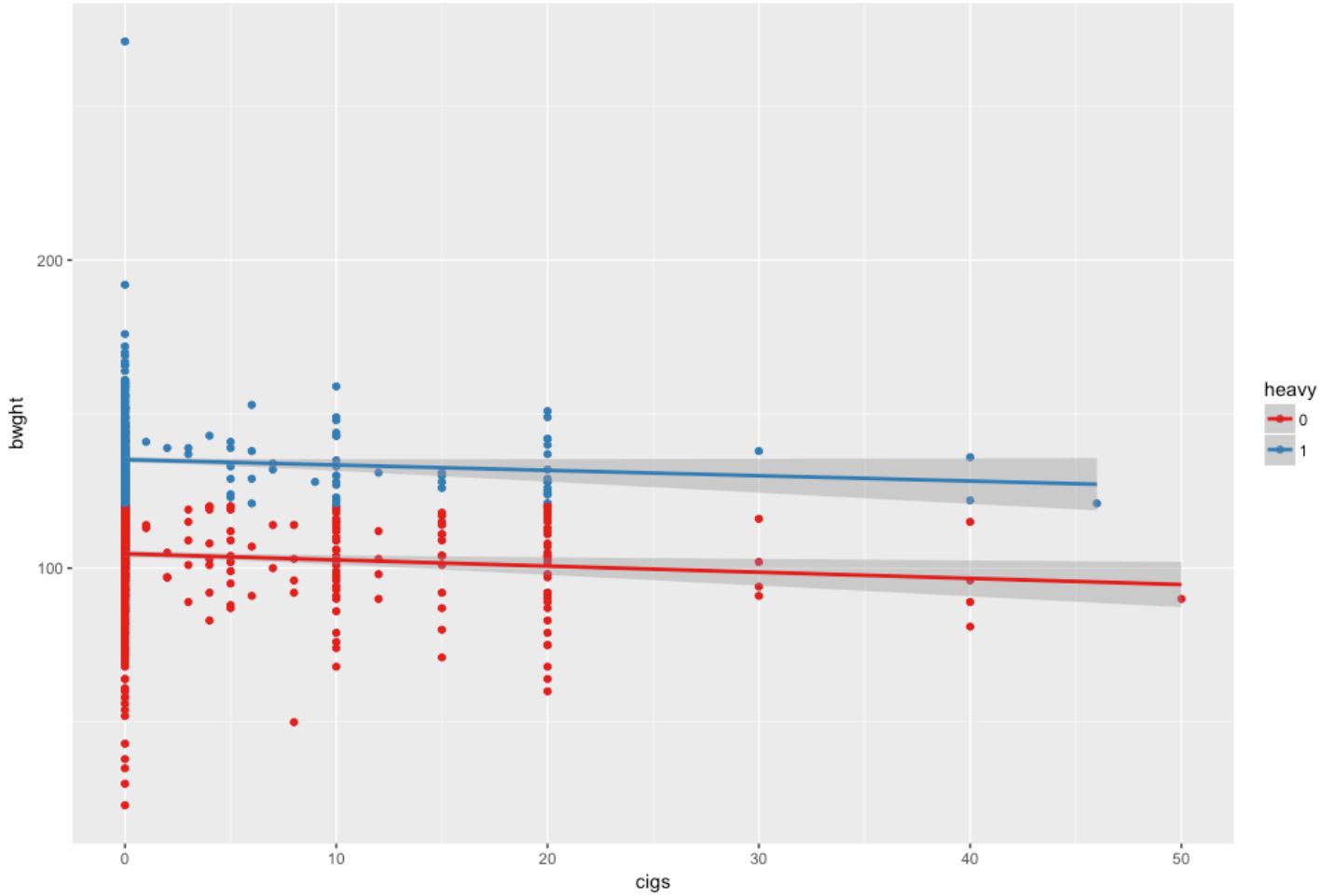
  - The sample is nonrandom and

$$\begin{aligned} & E(bwght|cigs, faminc, bwght < 3) \\ \neq \quad & E(bwght|cigs, faminc) \end{aligned}$$

Zero-conditional mean assumption fails!

Nonrandom Samples: Endogenous sample selection

```
> data <- data %>% mutate(heavy = ifelse(bwght > median(bwght), 1, 0))
> data$heavy <- as.factor(data$heavy)
> ggplot(data, aes(x = cigs, y = bwght, color = heavy)) + geom_point() +
    geom_smooth(method='lm') + scale_color_brewer(palette="Set1")
```



Gauss-Markov Assumptions

- [MLR3] (no perfect collinearity) None of x's is constant and there is no perfect linear relationships among x's
  - Common causes lead to violation of this assumption
    - Multiple variables measure the same thing, dummy variables trap
  - Identification of assumption violation
    - Routinely reported by statistical softwares

Example: A Model with Perfect Collinearity

```
> load("wage1.RData")
> male <- 1 - data$female
> wage.m1 <- lm(lwage ~ educ + exper + male + female, data)
```

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.137239   0.101327   1.354    0.176
educ        0.091290   0.007123  12.816  < 2e-16 ***
exper       0.009414   0.001449   6.496 1.93e-10 ***
male        0.343597   0.037667   9.122  < 2e-16 ***
female            NA         NA      NA       NA
```

Multicollinearity

- High correlation between two or more independent variables is known as multicollinearity
- Multicollinearity does NOT violate MLR3

Question:

Suppose the correlation between two variables is 0.5. Are they highly correlated?

Multicollinearity: An Example

```
> x1 <- rnorm(100, mean = 0, sd = 2)
> x2.low <- x1 + rnorm(100, mean = 0, sd = 2)
> x2.med <- x1 + rnorm(100, mean = 0, sd = 1)
> x2.high <- x1 + rnorm(100, mean = 0, sd = 0.5)
> cor(x1, cbind(x2.low, x2.med, x2.high))
```

```
        x2.low    x2.med   x2.high
[1,] 0.6751081 0.9021813 0.9646901
```

```
> y <- x1 + rnorm(100, mean = 0, sd = 4)
> m1 <- lm(y ~ x1 + x2.low)
> m2 <- lm(y ~ x1 + x2.med)
> m3 <- lm(y ~ x1 + x2.high)
> stargazer(m1, m2, m3, align = TRUE, no.space = TRUE)
```

Multicollinearity: An Example (100 Observations)

	Dependent variable:		
	y		
	(1)	(2)	(3)
x1	0.989***	0.773*	1.041
	(0.249)	(0.425)	(0.698)
x2.low	0.016		
	(0.167)		
x2.med		0.233	
		(0.386)	
x2.high			−0.037
			(0.685)
Constant	0.547	0.563	0.549
	(0.370)	(0.369)	(0.369)
Observations	100	100	100
R^2	0.236	0.238	0.236
Adjusted R^2	0.220	0.223	0.220
Residual Std. Error (df = 97)	3.687	3.680	3.687
F Statistic (df = 2; 97)	14.954***	15.187***	14.950***
Note:			*p<0.1; **p<0.05; ***p<0.01

Variance Inflation Factors

- The variance inflation factor for x_j is

$$VIF_j = \frac{1}{1 - R_j^2},$$

R_j^2 is the R-squared from regressing x_j on all the other independent variables

  - x_j is strongly correlated with other independent variables $\rightarrow$ R_j^2 close to 1 $\rightarrow$ VIF_j is large
  - Rule of thumb: Value of VIF greater than 10 indicates the multicollinearity problem
  - R function: `vif` in multiple packages, such as `HH`, `car`, `fmsb`, `faraway` and `VIF`

Multicollinearity

- Consequence of multicollinearity
  - Important variables can appear to be insignificant and standard errors can be large
  - Makes it hard to separate the roles of independent variables
    - Independent variables are redundant, but we are asking the regression model to separate them
- What do we do about it?
  - Nothing
    - Multicollinearity weakens ability to interpret, but may allow us to draw more reliable causal inference
  - Get rid of one of the offenders

Gauss-Markov Assumptions

- [MLR4] (zero conditional mean) The disturbance u satisfies $E(u|x_1, \ldots, x_k) = 0$ for any given value of $(x_1, \ldots, x_k)$
  - Common causes lead to violation of this assumption
    - Missing important variables in the model
  - Identification of assumption violation
    - Unfortunately, no test for this
  - Mitigation
    - Direction of omitted variable bias
    - If we can find a variable that is closely related to the omitted variable, we can use it as an instrument variable or a proxy variable

Gauss-Markov Assumptions

- [MLR5] (homoskedasticity) $Var(u_i|x_{i1}, \ldots, x_{ik}) = \sigma^2$ for $i = 1, 2, \ldots, n$. (It implies $Var(u_i) = \sigma^2$)
  - Common causes lead to violation of this assumption
    - Data issue
  - Identification of assumption violation
    - Residual plots, Breusch-Pagan test, White test
  - Solutions
    - Robust standard errors

Model Diagnostics

- MLR1-5 are collectively known as the Gauss-Markov Assumptions
  - under which, OLS estimator is the best linear unbiased estimator (BLUE)
  - MLR1-4 are required for OLS estimators to be unbiased
- After fitting a regression model, it is important to determine whether all the necessary model assumptions are valid

Outline

- Model diagnostics
- Outliers
- A possible model fitting strategy

Outliers

- Outliers in the dependent variable
  - Observations lie far from the SRF
  - Rule of thumb: An observation is an outlier if its residual is larger than 3 standard deviations away from the mean
- Outliers in the independent variable
  - Known as high-leverage points, to distinguish them from observations that are outliers in the dependent variable
  - Can be identified using leverage values

Leverage Value

- The fitted values of a multiple regression can be written as

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \cdots + p_{in}y_n$$

- Denote p_{ii} as the leverage value for the ith observation
  - It is the weight (leverage) given to y_i in determining the ith fitted value $\hat{y}_i$
  - It measures the "outlierness" in the independent variables
  - $0 \leq p_{ii} \leq 1$, and average of all leverage values is $(k+1)/n$
  - Rule of thumb: Points with p_{ii} greater than $2(k+1)/n$ are generally regarded as points with high leverage

Influential observations and Cook's Distance

- Influential observations
  - Large absolute residual + high leverage value
    - A point with high leverage may or may not be influential
  - OLS is very sensitive to influential observations
- Cook's distance measures the influence of the jth observation by

$$C_j = \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_{i(j)})^2}{\hat{\sigma}^2(k+1)},$$

  where $\hat{y}_i$ is the fitted value obtained from the full sample, and $\hat{y}_{i(j)}$ is the fitted value obtained by deleting the jth observation
  - If a point is influential, its deletion causes large changes in fitted values, and value of C_j will be large
  - Rule of thumb: Points with C_j values greater than 1 are influential

Outliers: An Example

- Example 9.8. R&D Intensity and Firm Size (rdchem.RData)
  - The regression model is

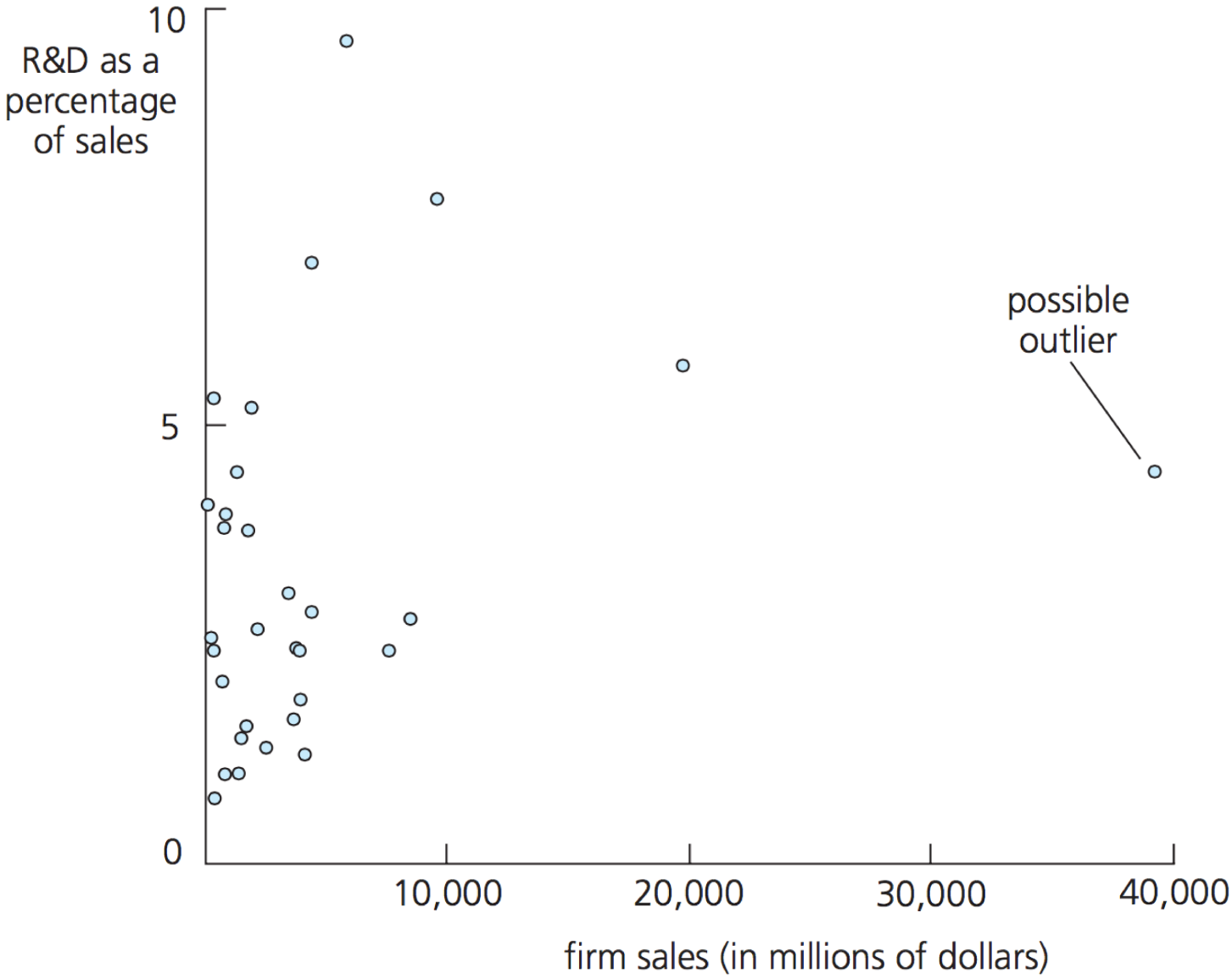
$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u,$$

    where
    - *rdintens*: R&D expenditures as a percentage of sales
    - *sales*: annual sales (in millions)
    - *profmarg*: profits as a percentage of sales
  - The OLS equation using data on 32 chemical companies is

$$\widehat{rdintens} = \underset{(0.586)}{2.625} + \underset{(.000044)}{.000053} sales + \underset{(.0462)}{.0446} profmarg,$$

    $n = 32, R^2 = .0761$
  - Neither *sales* nor *profmarg* is statistically significant at even the 10% level in this regression.

Outliers: An Example



Outliers: An Example

- Of the 32 firms, 31 have annual sales less than \$20 billion, where one firm has annual sales of almost \$40 billion.
- Without the high-leverage observation, the estimated model is given by

$$\widehat{rdintens} = \underset{(0.592)}{2.297} + \underset{(.000084)}{.000186}sales + \underset{(.0445)}{.0478}profmarg,$$

$n = 31, R^2 = .173$

  - Using the sample of smaller firms, there is a statistically significant positive effect between R&D intensity and firm size.
  - The profit margin is still not significant, and its coefficient has not changed by much.

What to do about outliers?

- Don't just remove, understand why a point is unusual!
- Outliers can be simple data entry errors
  - It is always a good idea to check summary statistics (min, max, etc)
  - Not unreasonable to fix observations where it's clear there was just an extra zero entered, etc.
- Outliers can be that the observation is just truly very different from the others
  - Reconsider whether the model is reasonable
  - Report the OLS results with and without suspected outliers

Outline

- Model diagnostics
- Outliers
- A possible model fitting strategy

A Possible Model Fitting Strategy

1. Identify the question of interest, and the goal of the analysis
   - Causal inference vs Prediction
   - Let us revisit the bias-variance tradeoff
     - Suppose $y = f(x) + u$, where $E[u] = 0$ and $Var[u] = \sigma^2$
     - We estimate $f(x)$ with $\hat{f}(x)$, then

$$\begin{aligned} \mathrm{E}[(y - \hat{f}(x))^2] &= \left(\mathrm{E}[\hat{f}(x)] - f(x)\right)^2 + \mathrm{E}\left[\left(\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\right)^2\right] + \sigma^2 \\ &= (Bias[\hat{f}(x)])^2 + Var[\hat{f}(x)] + \sigma^2 \end{aligned}$$

A Possible Model Fitting Strategy

2. Understand the data set (Exploratory Data Analysis - Visualization module)
   - Identify scales, ranges, distributions and etc. for each variable
     - Are data skewed? Outliers? Missing values?
     - Calculate summary statistics, plot histograms and etc.
   - Construct bivariate scatterplots
     - Nonlinear (curvature)? Outliers, leverage points? Correlation between variables?
   - Check for special features in the data
     - Categorical factors

A Possible Model Fitting Strategy

3. Fitting the model - modeling is an iterative process. No one gets it right the first time!
    - Prediction: Select variables based on information criteria
    - Causal inference
        - Check statistical significance
        - Omitted variable bias
    - Check for missed nonlinearity
        - Are variables appropriately transformed?
        - Test for correct functional forms of variables

A Possible Model Fitting Strategy

4. Model validation
   - Causal inference: Residual analysis - ensure satisfactory residual plots and no negative diagnostic messages
     - Check linearity
     - Check for heteroscedasticity, skewness and etc.
     - Look for outliers and influential points
   - Prediction: Avoid overfitting - Machine learning module
     - The model may be fitted by part of the data and validated by the remainder if the sample size is large
     - Otherwise, resampling methods, such as bootstrap and cross-validation, can be used

Last Words

If applied econometrics were easy, theorists would do it. But it's not as hard as the dense pages of *Econometrica* might lead you to believe. Carefully applied to coherent causal questions, regression almost always makes sense. Your standard errors probably won't be quite right, but they rarely are. Avoid embarrassment by being your own best skeptic, and especially, DON'T PANIC!

Source: Mostly harmless econometrics: An empiricist's companion, by Angrist and Pischke