BUS41100 Applied Regression Analysis

Week 4: MLR Issues and (Some) Fixes

R^2, multicollinearity, F-test
nonconstant variance, clustering, panels

Max H. Farrell
The University of Chicago Booth School of Business

Quick Recap

Everything is in the context of our **linear model**

$$Y|X_1, \ldots, X_d \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d, \sigma^2)$$

So far:

Week 1: What is a line and how should I draw it?

Week 2: Is this *really* the line?

Week 3: I need more than a line, but I don't want to learn anything new.

Today: Some MLR issues, fixes, & extensions

A (bad) goodness of fit measure: R^2

How well does the least squares fit explain variation in Y?

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{Total sum of squares (SST)}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{Regression sum of squares (SSR)}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{Error sum of squares (SSE)}}$$

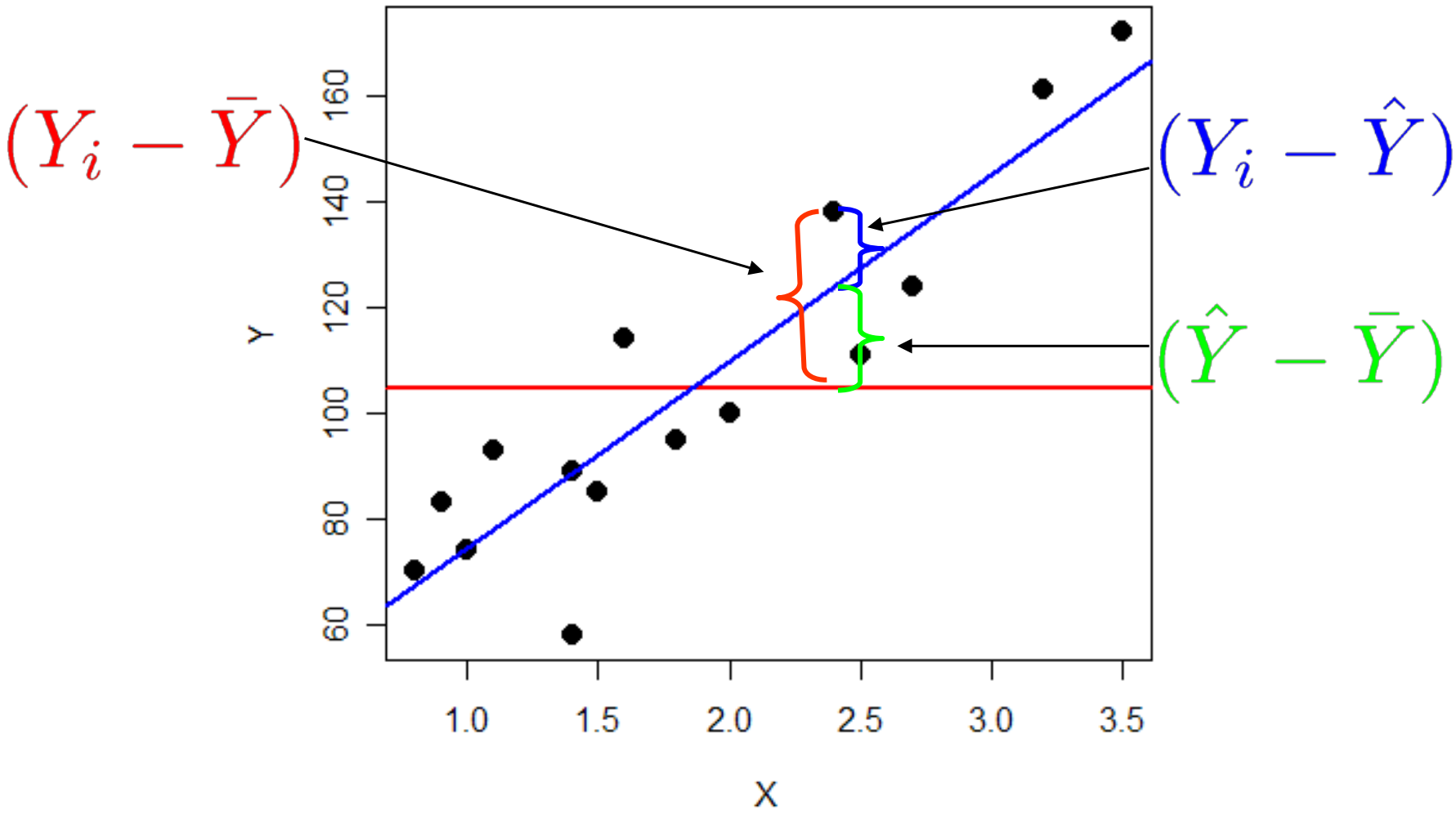
SSR: Variation in Y explained by the regression.

SSE: Variation in Y that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

Be careful of similar acronyms; e.g. SSR for "residual" SS.

How does that breakdown look on a scatterplot?



A (bad) goodness of fit measure: R^2

The coefficient of determination, denoted by R^2, measures goodness-of-fit:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

- SLR or MLR: same formula.
- $R^2 = \text{corr}^2(\hat{Y}, Y) = r^2_{\hat{y}y}$ $(= r^2_{xy}$ in SLR$)$
- $0 < R^2 < 1$.
- R^2 closer to 1 $\rightarrow$ better fit . . . *for these data points*
  - No surprise: the higher the sample correlation between X and Y, the better you are doing in your regression.
  - So what? What's a "good" R^2? For prediction? For understanding?

Adjusted R^2

This is the reason some people like to look at adjusted R^2

$$R_a^2 = 1 - s^2/s_y^2$$

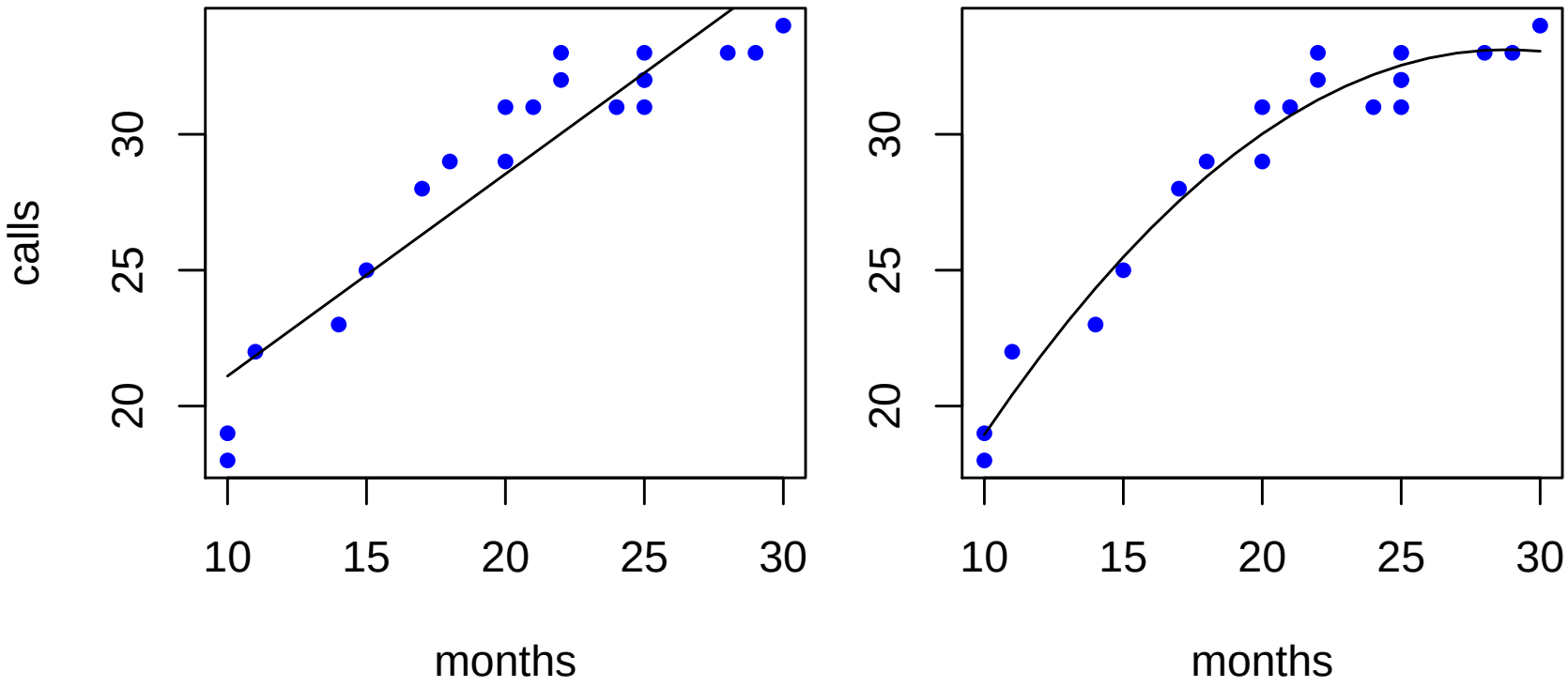
Since s^2/s_y^2 is a ratio of variance estimates, R_a^2 will not necessarily increase when new variables are added.

Unfortunately, R_a^2 is useless!

- The problem is that there is no theory for inference about R_a^2, so we will not be able to tell "how big is big".

For a silly example, back to the call center data.

- The quadratic model fit better than linear.



- But how far can we go?

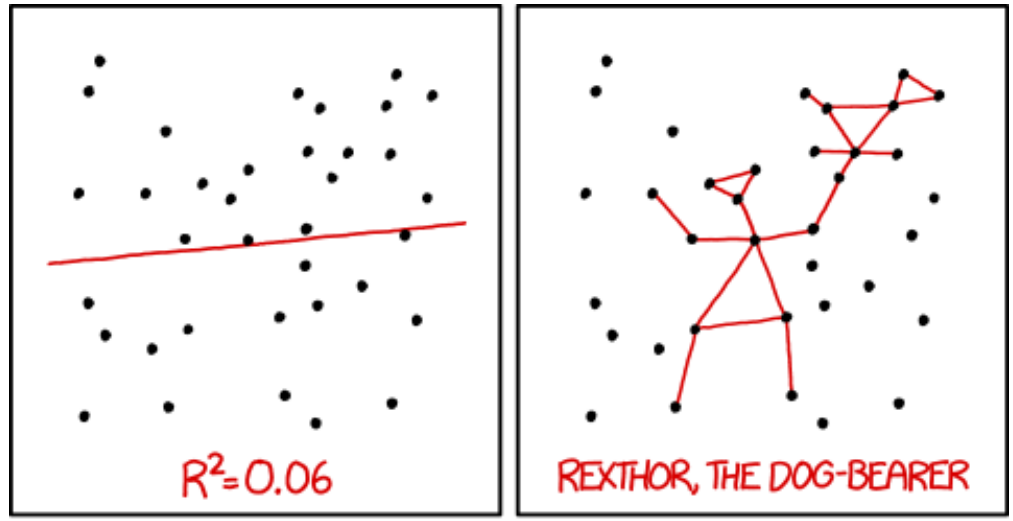
bad R^2?

bad model?

bad data?

bad question?

. . . or **just reality**?




```
> summary(trucklm1)$r.square   ## make
[1] 0.021
> summary(trucklm2)$r.square   ## make + miles
[1] 0.446
> summary(trucklm3)$r.square   ## make * miles
[1] 0.511
> summary(trucklm6)$r.square   ## make * (miles + miles^2)
[1] 0.693
```

- Is `make` useless? Is 45% *significantly* better?
- Is adding `miles^2` worth it?

We will do a good job at building models later in the course.

Remember, **interpretation** matters

- Choose the model that answers the question

Machine learning tools blindly build **prediction** models

- . . . which is great if that's what you want

Multicollinearity

Our next issue is Multicollinearity: strong linear dependence between some of the covariates in a multiple regression.

The usual marginal effect interpretation is lost:

- change in one X variable leads to change in others.

Coefficient standard errors will be large (since you don't know which X_j to regress onto)

- leads to large uncertainty about the b_j's
- therefore you may fail to reject $\beta_j = 0$ for all of the X_j's even if they do have a strong effect on Y.

Suppose that you regress Y onto X_1 and $X_2 = 10 \times X_1$.

Then

$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + \beta_1 X_1 + \beta_2 (10 X_1)$$

and the marginal effect of X_1 on Y is

$$\frac{\partial \mathbb{E}[Y|X_1, X_2]}{\partial X_1} = \beta_1 + 10\beta_2$$

- X_1 and X_2 do not act independently!

We saw this once already, on homework 1.

```
> teach <- read.csv("teach.csv", stringsAsFactors=TRUE)
> summary(reg.sex <- lm(salary ~ sex, data=teach))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1598.76      66.89  23.903  < 2e-16
sexM          283.81      99.10   2.864  0.00523

> summary(reg.marry <- lm(salary ~ marry, data=teach))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1834.84      61.38  29.894  < 2e-16
marryTRUE    -300.38     102.93  -2.918  0.00447

> summary(reg.both <- lm(salary ~ sex + marry, data=teach))

            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1719.8      113.1  15.209   <2e-16
sexM           162.8      134.5   1.210    0.229
marryTRUE     -185.3      139.9  -1.324    0.189
```

How can sex and marry each be significant, but not together?

Because they do not act independently!

```
> cor(as.numeric(teach$sex),as.numeric(teach$marry))
[1] -0.6794459
> table(teach$sex,teach$marry)
    FALSE TRUE
  F    17   32
  M    41    0
```

Remember our MLR interpretation. Can't separate if women or married people are paid less. But we can see significance!

```
> summary(reg.both)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1719.8      113.1  15.209   <2e-16 ***
sexM           162.8      134.5   1.210    0.229
marryTRUE     -185.3      139.9  -1.324    0.189

Residual standard error: 466.2 on 87 degrees of freedom
Multiple R-squared:  0.1033,        Adjusted R-squared:  0.08272
F-statistic: 5.013 on 2 and 87 DF,  p-value: 0.008699
```

The F-test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0$$
$$H_1 : \text{at least one } \beta_j \neq 0.$$

The F-test asks if there is any "information" in a regression. Tries to formalize what's a **"big"** R^2, instead of testing one coefficient.

- The test statistic is not a t-test, not even based on a Normal distribution. We won't worry about the details, just compare p-value to pre-set level α.

The Partial F-test

Same idea, but test if additional regressors have information.

Example: Adding interactions to the pickup data

```
> trucklm2 <- lm(price ~ make + miles, data=pickup)
```

$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 \mathbb{1}_F + \beta_2 \mathbb{1}_G + \beta_3 M$$

```
> trucklm3 <- lm(price ~ make * miles, data=pickup)
```

$$\mathbb{E}[Y|X_1, X_2] = \beta_0 + \beta_1 \mathbb{1}_F + \beta_2 \mathbb{1}_G + \beta_3 M + \beta_4 \mathbb{1}_F M + \beta_5 \mathbb{1}_G M$$

We want to test $H_0 : \beta_4 = \beta_5 = 0$ versus $H_1 : \beta_4$ or $\beta_5 \neq 0$.

```
> anova(trucklm2,trucklm3)
Analysis of Variance Table

Model 1: price ~ make + miles
Model 2: price ~ make * miles
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1     42 777981726
2     40 686422452  2  91559273 2.6677 0.08174
```

The **F-test** is common

but it is not a useful model selection method.

Hypothesis testing only gives a yes/no answer.

- Which $\beta_j \neq 0$?
- How many?
- Is there a lot of information, or just enough?
- What X's should we add? Which combos?
- Where do we start? What do we text "next"?

In a couple weeks, we will see modern variable selection methods, for now just be aware of testing and its limitations.

Multicollinearity is not a big problem in and of itself, you just need to know that it is there.

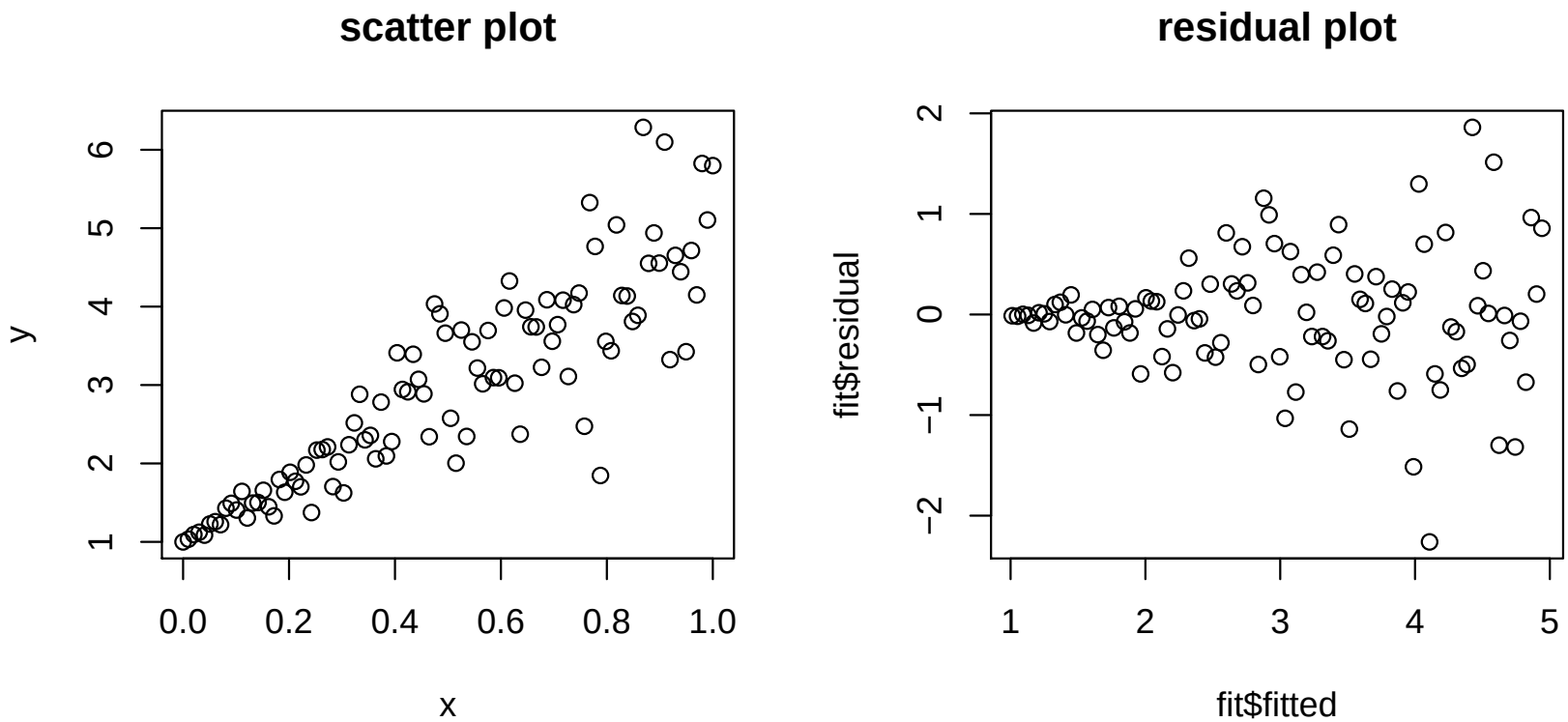
If you recognize multicollinearity:

- Understand that the β_j are not true marginal effects.
- Consider dropping variables to get a more simple model.
- Expect to see big standard errors on your coefficients (i.e., your coefficient estimates are unstable).

Nonconstant variance

One of the most common violations (problems?) in real data

- E.g. A trumpet shape in the scatterplot



We can try to stabilize the variance . . . or do robust inference

Plotting e vs $\hat{Y}$ is your #1 tool for finding fit problems.

Why?

- Because it gives a quick visual indicator of whether or not the model assumptions are true.

What should we expect to see if they are true?

1. No pattern: X has linear information ($\hat{Y}$ is made from X)
2. Each ε_i has the same variance (σ^2).
3. Each ε_i has the same mean (0).
4. The ε_i collectively have a Normal distribution.

Remember: $\hat{Y}$ is made from all the X's, so one plot summarizes across the X even in MLR.

Variance stabilizing transformations

This is one of the most common model violations; luckily, it is usually fixable by transforming the response (Y) variable.

$\log(Y)$ is the most common variance stabilizing transform.

- If Y has only positive values (e.g. sales) or is a count (e.g. # of customers), take $\log(Y)$ (always natural log).

Also, consider looking at Y/X or dividing by another factor.

In general, think about in what scale you expect linearity.

For example, suppose $Y = \beta_0 + \beta_1 X + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, (X\sigma)^2)$.

- This is not cool!
- $\text{sd}(\varepsilon_i) = |X_i|\sigma \;\Rightarrow\;$ nonconstant variance.

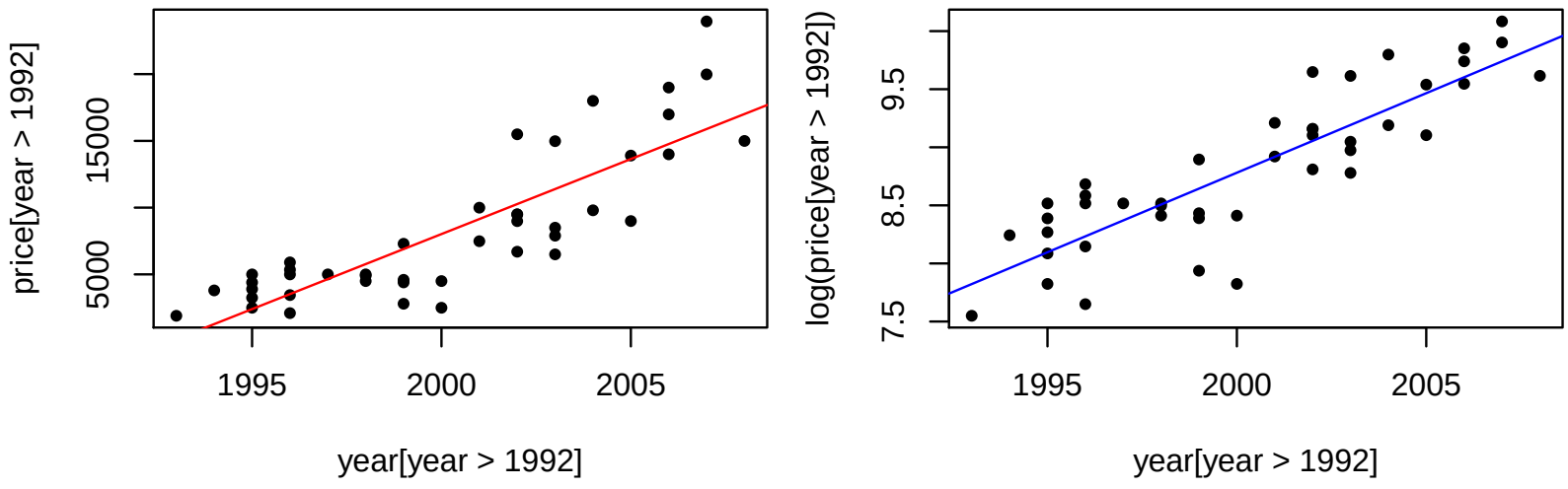
But we could look instead at

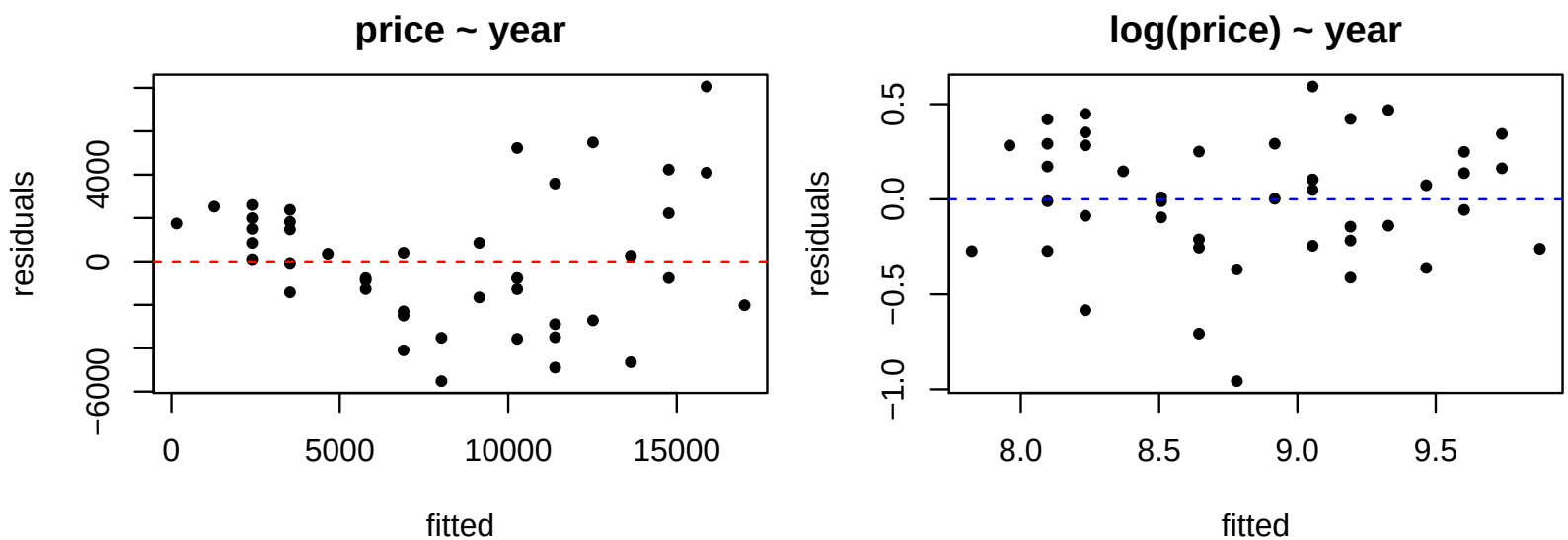
$$\frac{Y}{X} = \frac{\beta_0}{X} + \beta_1 + \frac{\varepsilon}{X} = \beta_0^\star + \frac{1}{X}\beta_1^\star + \varepsilon^\star$$

where $\text{var}(\varepsilon^\star) = X^{-2}\text{var}(\varepsilon) = \sigma^2$ is now constant.

Hence, the proper linear scale is to look at $Y/X \sim 1/X$.

Reconsider the regression of truck `price` onto `year`, after removing trucks older than 1993 (`truck[year>1992,]`).





Warning: be careful when interpreting the transformed model.

If $\mathbb{E}[\log(Y)|X] = b_0 + b_1 X$, then $\mathbb{E}[Y|X] \approx e^{b_0} e^{b_1 X}$.

We have a multiplicative model now!

Also, you cannot compare R^2 values for regressions corresponding to different transformations of the response.

- Y and $f(Y)$ may not be on the same scale,
- therefore var(Y) and var$(f(Y))$ may not be either.

Look at residuals to see which model is better.

Heteroskedasticity Robust Inference

What if σ^2 is not constant?

✓ Predictions, point estimates $\hat{Y}_f = b_0 + b_1 X_f$

- Everything from week 1 still applies

X Inference: CI: $\sigma_{b_1} \neq \sigma / \sqrt{(n-1)s_x^2}$

- But week 2 is all wrong!
- Luckily, we can find different (more complicated) variance formulas.

$\Rightarrow$ Keep the original model

- Same scale, same interpretation
- New standard errors (bigger $\rightarrow$ less precision)
- Impacts confidence intervals, tests
- What about prediction intervals?

Example: back to the full pickup regression of `price` on `years`, all trucks.

Ignoring the violation:

```
> truckreg <- lm(price ~ year)
> coef(summary(truckreg))
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  -1468663.94   202492.62   -7.2529   4.8767e-09
year             738.54      101.28    7.2920   4.2764e-09
```

Accounting for nonconstant variance:

```
> library(lmtest)
> library(sandwich)
> coeftest(truckreg, vcov = vcovHC)
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  -1468663.94   574787.49 -2.5551  0.01415
year             738.54      287.37  2.5700  0.01363
```

Clustering

We assumed:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_d X_{d,i} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

which in particular means

$$\mathbb{COV}(\varepsilon_i, \varepsilon_j) = 0 \qquad \text{for all } i \neq j.$$

Clustering is a very common violation of constant variance and independence. Each observation is allowed to have

- unknown correlation with a **small** number others
- . . . in a known pattern.
- E.g., (i) children in classrooms in schools, (ii) firms in industries, (iii) products made by companies
- How much independent information?

The MLR model with clustering

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_d X_{d,i} + \varepsilon_i, \qquad \varepsilon_i \overset{\cancel{iid}}{\sim} \mathcal{N}(0, \cancel{\sigma^2}),$$

Instead

$$\mathbb{COV}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2 & \text{if } i = j, \qquad \text{just } \mathbb{V}[\varepsilon_i] \\ \sigma_{ij} & \text{if } i \neq j, \text{ but in the same cluster} \\ 0 & \text{otherwise.} \end{cases}$$

So **only** standard errors change!

- Same slope β_1 for everyone

Cluster methods aim for **robustness**:

- No assumptions about σ_i^2 and σ_{ij}
- Assume we have **many** clusters G, each with a **small** number of observations n_g: $\qquad n = \sum_{g=1}^{G} n_g$

Example: Patents and R&D in 1991, by `firm.id`

```
> head(D91)
     year sector    rdexp firm.id patents
1449 1991      4 6.287435       1      55
1450 1991      5 5.150736       2      67
1451 1991      2 4.172710       3      55
1452 1991      2 6.127538       4      83
1453 1991     11 4.866621       5       0
1454 1991      5 7.696947       6       4
```

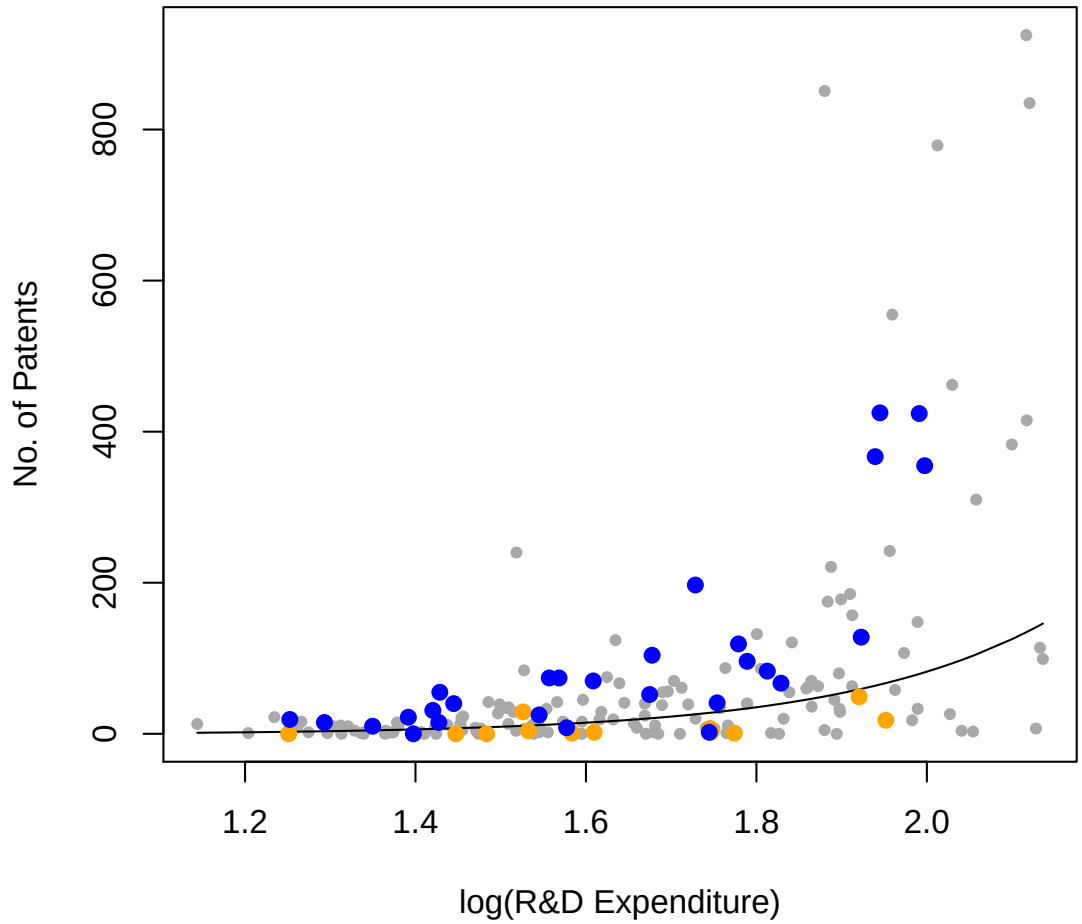
Are these rows independent? If they were . . .

```
> D91$newY <- log(D91$patents + 1)
> summary(slr <- lm(newY ~ log(rdexp), data=D91))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.9226     0.7551  -5.195 5.54e-07
log(rdexp)    4.1723     0.4531   9.208  < 2e-16

Residual standard error: 1.451 on 179 degrees of freedom
```

What happens when errors are correlated?

- If $\varepsilon_i > 0$ we expect $\varepsilon_j > 0$. (if $\sigma_{ij} > 0$)

$\Rightarrow$ Both observation i and j are above the line.



We want our inference to be **robust** to this problem.

```
> library(multiwayvcov); library(lmtest)
> vcov.slr <- cluster.vcov(slr, D91$sector)
> coeftest(slr, vcov.slr)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -3.92263    0.90933 -4.3138 2.649e-05
log(rdexp)   4.17226    0.56036  7.4457 3.920e-12

> summary(slr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.9226     0.7551  -5.195 5.54e-07
log(rdexp)    4.1723     0.4531   9.208  < 2e-16
```

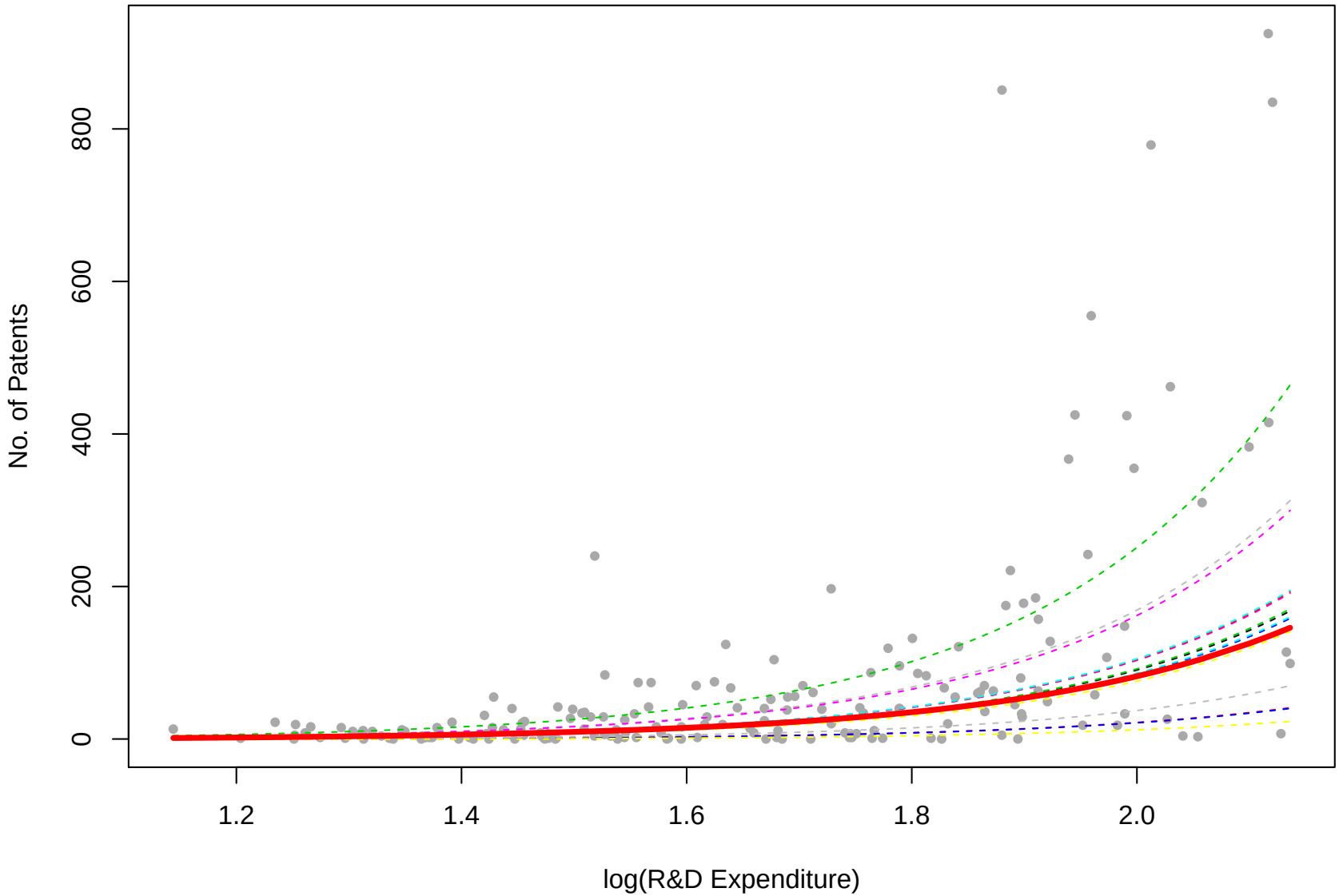
Can we just control for clusters? No!

- Not different slopes (and intercepts?) for each cluster . . .
  we want one slope with the right standard error!

```
> coeftest(slr, vcov.slr)

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -3.92263    0.90933 -4.3138 2.649e-05
log(rdexp)   4.17226    0.56036  7.4457 3.920e-12

> slr.dummies <- lm(newY ~ log(rdexp) + as.factor(sector) - 1)
> summary(slr.dummies)

                    Estimate Std. Error t value Pr(>|t|)
log(rdexp)            4.5007     0.5145   8.747 2.43e-15
as.factor(sector)1   -5.8800     0.9235  -6.367 1.83e-09
as.factor(sector)2   -3.4714     0.8794  -3.947 0.000117
...                     ...
```

Can we just control for clusters? No!

- Not different slopes (and intercepts?) for each cluster . . .
  we want one slope with the right standard error!



Panel Data

So far we have seen i.i.d. data and clustered data.

Panel data adds time:

- units $i = 1, \ldots, n$
- followed over time periods $t = 1, \ldots, T$

$\Rightarrow$ dependent over time, possibly clustered

More and more datasets are **panels**, also called **longitudinal**

- Tracking consumer decisions
- Firm financials over time
- Macro data across countries
- Students in classrooms over several grades

Distinct from a *repeated cross-section*:

- New units sampled each time $\Rightarrow$ independent over time

The linear regression model for panel data:

$$Y_{i,t} = \beta_1 X_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t}$$

Familiar pieces, just like SLR:

- β_1 – the **general** trend, same as always. *(Where's β_0?)*
- $Y_{i,t}$, $X_{i,t}$, $\varepsilon_{i,t}$ – Outcome, predictor, mean zero idiosyncratic shock (clustered?)

What's new:

- α_i – unit-specific effects. Different people are different!
  - Cars: Camry/Tundra/Sienna. S&P500: Hershey/UPS/Wynn
- γ_t – time-specific effects. Different years are different!
- For now, $\gamma_t = 0$. Same concepts/methods.

Just the familiar **same slope**, different intercepts model!

Well, almost ...

Estimation strategy depends on how we think about α_i

1. $\alpha_i = 0 \quad \Longrightarrow \quad Y_{i,t} = \boldsymbol{\beta}_1 X_{i,t} + \varepsilon_{i,t}$

- `lm` on $N = nT$ observations. Cluster if needed.

2. `random` effects: $\text{cor}(\alpha_i, X_{i,t}) = 0$

- Still possible to use `lm` on $N = nT$ (and cluster on unit) . . .

$$Y_{i,t} = \boldsymbol{\beta}_1 X_{i,t} + \tilde{\varepsilon}_{i,t}, \qquad \tilde{\varepsilon}_{i,t} = \alpha_i + \varepsilon_{i,t}$$

- . . . but lots of variance!

3. `fixed` effects: $\text{cor}(\alpha_i, X_{i,t}) \neq 0$

- **same slope**, but $\boldsymbol{n}$ different intercepts!

$$Y_{i,t} = \boldsymbol{\beta}_1 X_{i,t} + \alpha_i + \varepsilon_{i,t}$$

- Too many parameters to estimate. `patent` data has $n = 181$.
- No time-invariant $X_{i,t} = X_i$.

The real patent data is a panel with clustering:

- unit is a `firm`: $i = 1, \ldots, 181$
- time is `year` $= 1983, \ldots, 1991$
- clustered by `sector`?

```
> table(D$year)

1983 1984 1985 1986 1987 1988 1989 1990 1991
 181  181  181  181  181  181  181  181  181

> table(D$firm.id, D$year)

     1983 1984 1985 1986 1987 1988 1989 1990 1991
  1     1    1    1    1    1    1    1    1    1
  2     1    1    1    1    1    1    1    1    1
  3     1    1    1    1    1    1    1    1    1
  4     1    1    1    1    1    1    1    1    1
  5     1    1    1    1    1    1    1    1    1
  ...
```

Estimation in R: using `lm` or the `plm` package.

1. $\alpha_i = 0$

```
> slr <- lm(newY ~ log(rdexp), data=D)
> plm.pooled <- plm(newY ~ log(rdexp), data=D,
+       index=c("firm.id", "year"), model="pooling")
```

2. `random` effects: $\text{cor}(\alpha_i, X_{i,t}) = 0$

```
> vcov.model <- cluster.vcov(slr, D$firm.id)
> coeftest(slr, vcov.model)
> plm.random <- plm(newY ~ log(rdexp), data=D,
+       index=c("firm.id", "year"), model="random")
```

3. `fixed` effects: $\text{cor}(\alpha_i, X_{i,t}) \neq 0$

```
> many.dummies <- lm(newY ~ log(rdexp) + as.factor(firm.id) - 1,
> plm.fixed <- plm(newY ~ log(rdexp), data=D,
+      index=c("firm.id", "year"), model="within")
```

Choosing between `fixed` or `random` effects.

- Fixed effects are more general, more realistic: isolate changes due to X vs due to specific person.
- If α_i don't matter, then $b_{\mathsf{RE}} \approx b_{\mathsf{FE}}$

```
> phtest(plm.random, plm.fixed)

        Hausman Test

data:  newY ~ log(rdexp)
chisq = 22.162, df = 1, p-value = 2.506e-06
alternative hypothesis: one model is inconsistent
```

Using `year` fixed effects (γ_t).

```
> lm(newY ~ log(rdexp) + as.factor(year) - 1, data=D)
> plm(newY ~ log(rdexp), data=D,
+   index=c("firm.id", "year"), model="within", effect="time")
```

Both `firm` and `year` fixed effects $\rightarrow$ `effect="twoways"`

Clustered Panels

A panel is not exempt from the concern of clustered data.

$$Y_{i,t} = \beta_1 X_{i,t} + \alpha_i + \gamma_t + \varepsilon_{i,t} \qquad \operatorname{cor}(\varepsilon_{i_1,t_1}, \varepsilon_{i_2,t_2}) \stackrel{?}{=} 0$$

```
> summary(plm.fixed)

  Estimate Std. Error t-value  Pr(>|t|)
log(rdexp)  2.22611    0.22642   9.832 < 2.2e-16
```

```
> vcov <- cluster.vcov(many.dummies, D$sector)
> coeftest(plm.fixed, vcov)

  Estimate Std. Error t value Pr(>|t|)
log(rdexp)  2.22611    0.80872  2.7527 0.005985
```

$\hookrightarrow$ **Four times** less information!

Prediction in Panels

Just use the usual prediction?

$$\hat{Y}_{f,i,t} = b_1 X_{f,i,t} + \hat{\alpha}_i + \hat{\gamma}_t$$

Predicting for who? when?

Only works if $\hat{\alpha}_i \approx \alpha_i$ and $\hat{\gamma}_t \approx \gamma_t$

- Long panels (large T) and no γ_t
- Many units (large n) and no α_i
- How big is big enough?

Uncertainty, same idea as before.

- Prediction intervals: same logic, similar formula, but **more** uncertainty.
- Intervals can be wide!

Further Issues in Panel Data

More general models

- Dynamic models – adding $X_{i,t} = Y_{i,t-1}$?
- Nonlinear model – binary Y?
- . . . **lots** more.

Specification Tests

- Breusch-Pagan – time effects
- Wooldridge – serial correlation
- Dickey-Fuller – non-stationarity over time
- . . . **lots** more.

Coming Up Next

Next Week:

- MIDTERM! Hurray!
- Plus correlation implies causation

After the exam: More advanced topics in regression

- Different outcomes
- Different data structures
- Some machine learning