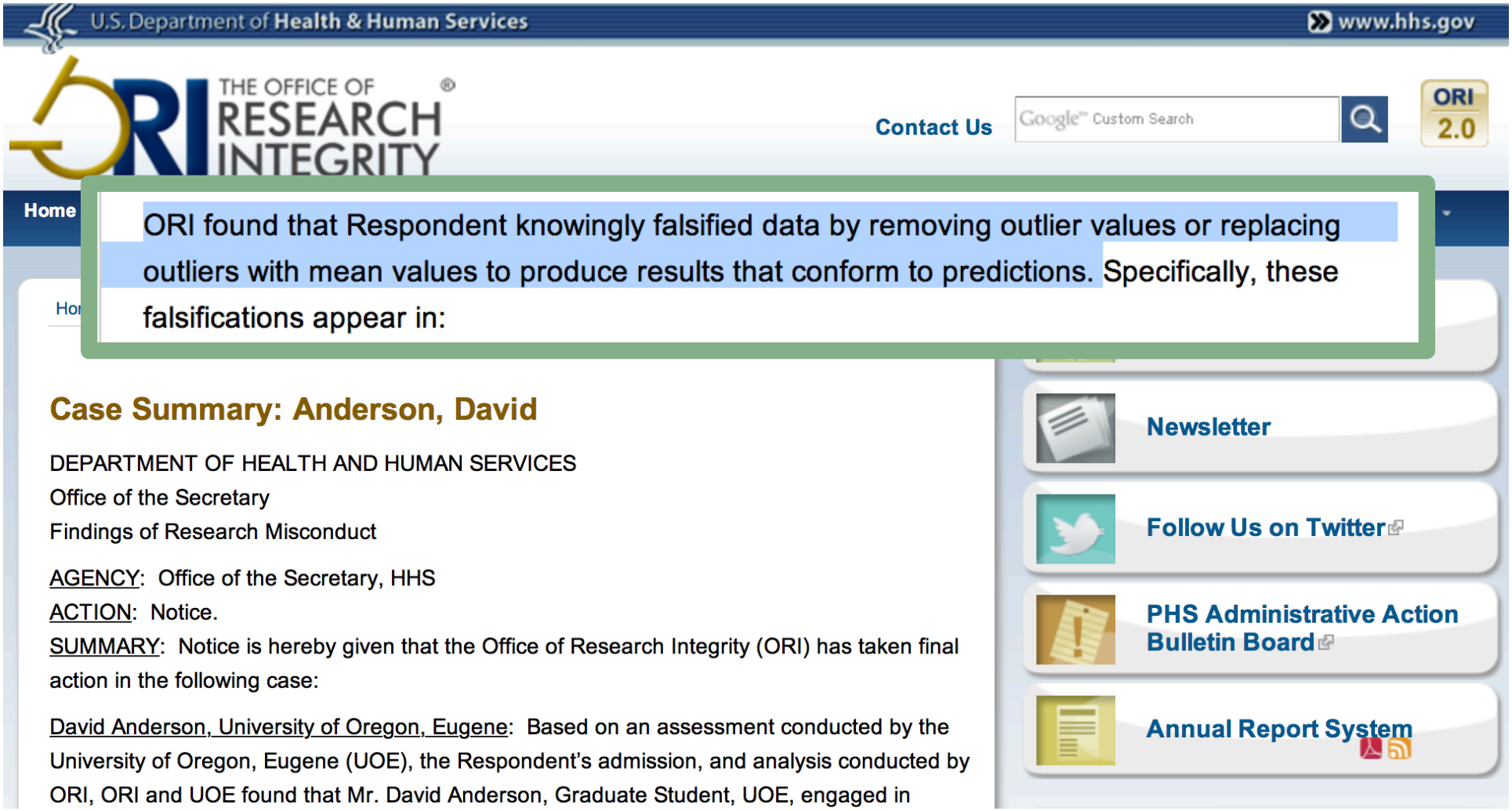
Class 12:
Multiple linear regression

Alison Presmanes Hill

Outliers


U.S. Department of Health & Human Services
www.hhs.gov
THE OFFICE OF RESEARCH INTEGRITY
Contact Us
Google™ Custom Search
ORI 2.0
Home
ORI found that Respondent knowingly falsified data by removing outlier values or replacing outliers with mean values to produce results that conform to predictions. Specifically, these falsifications appear in:
Case Summary: Anderson, David
DEPARTMENT OF HEALTH AND HUMAN SERVICES
Office of the Secretary
Findings of Research Misconduct
AGENCY: Office of the Secretary, HHS
ACTION: Notice.
SUMMARY: Notice is hereby given that the Office of Research Integrity (ORI) has taken final action in the following case:
David Anderson, University of Oregon, Eugene: Based on an assessment conducted by the University of Oregon, Eugene (UOE), the Respondent's admission, and analysis conducted by ORI, ORI and UOE found that Mr. David Anderson, Graduate Student, UOE, engaged in
Newsletter
Follow Us on Twitter
PHS Administrative Action Bulletin Board
Annual Report System

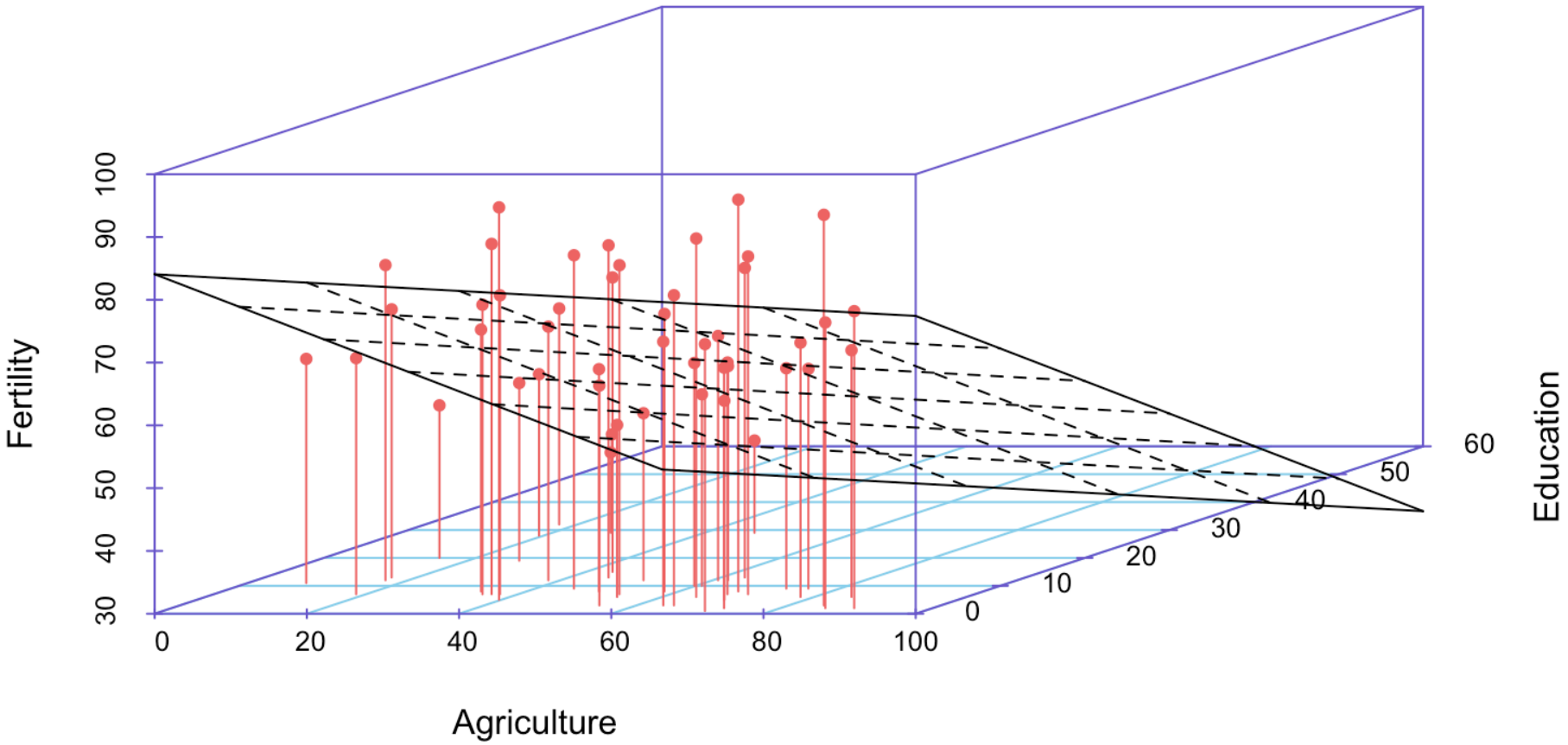
MULTIPLE* LINEAR REGRESSION

* Refers to the number of predictors, not response variables

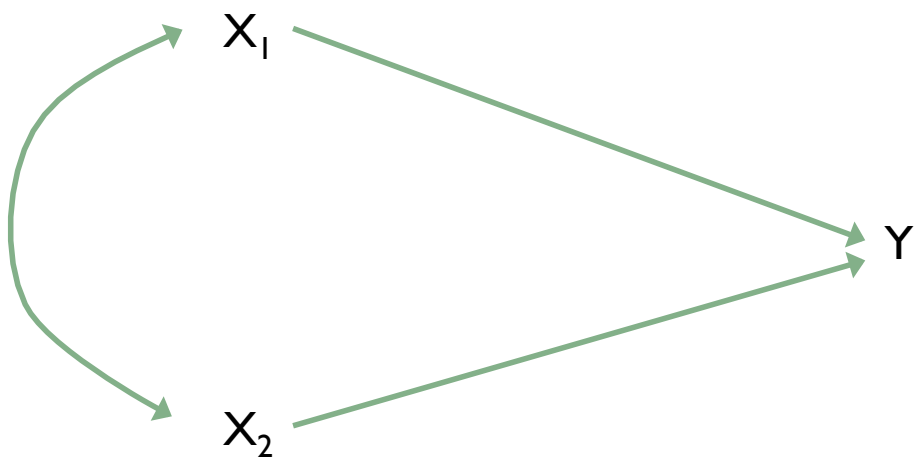
Regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

2D line → 3D plane



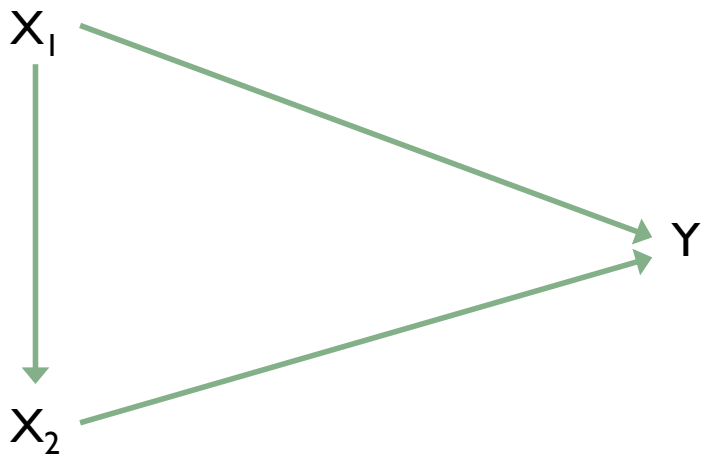
Partial redundancy



Both Xs exert direct effects on Y

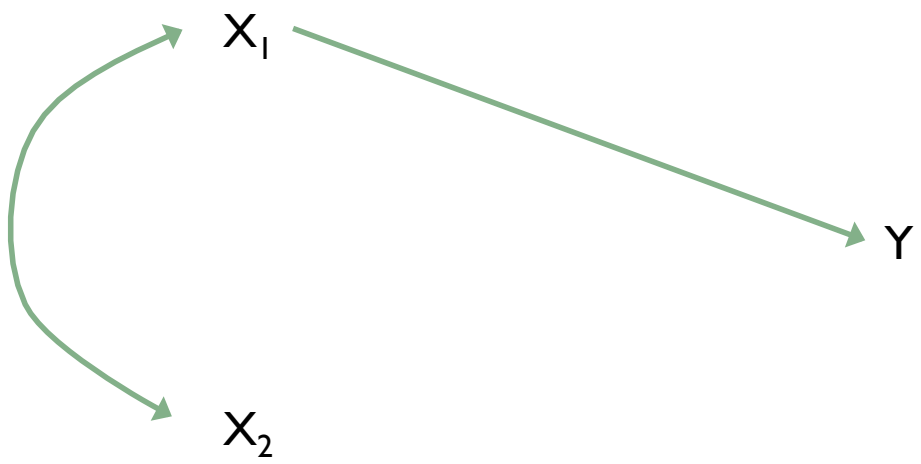
Example: predict school achievement (Y) from parental income (X_1) and education (X_2)

Partial redundancy



Both Xs exert direct effects on Y
X_1 also has an indirect effect on Y via X_2
Example: predict salary (Y) from years since PhD (X_1) and publications (X_2)

Full redundancy: spurious relationship



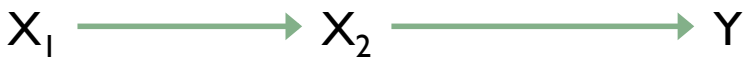
X_2 is completely redundant with X_1 in terms of predicting Y
X_1 is a confounder of the relationship between X_2 and Y
Example: predict infant behavior (Y) from maternal nutrition (X_1)
and maternal substance abuse (X_2)
Does maternal nutrition impact infant behavior independent of substance abuse?

Full redundancy: spurious relationship

Mediator variables - "In general, a given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion. Mediators explain how external physical events take on internal psychological significance. Whereas moderator variables specify when certain effects will hold, mediators speak to how or why such effects occur." *p. 1176, Baron & Kenny, 1986*

The general test for mediation is to examine the relation between the predictor and the criterion variables, the relation between the predictor and the mediator variables, and the relation between the mediator and criterion variables. All of these correlations should be significant. The relation between predictor and criterion should be reduced (to zero in the case of total mediation) after controlling the relation between the mediator and criterion variables.

Full redundancy: indirect effect

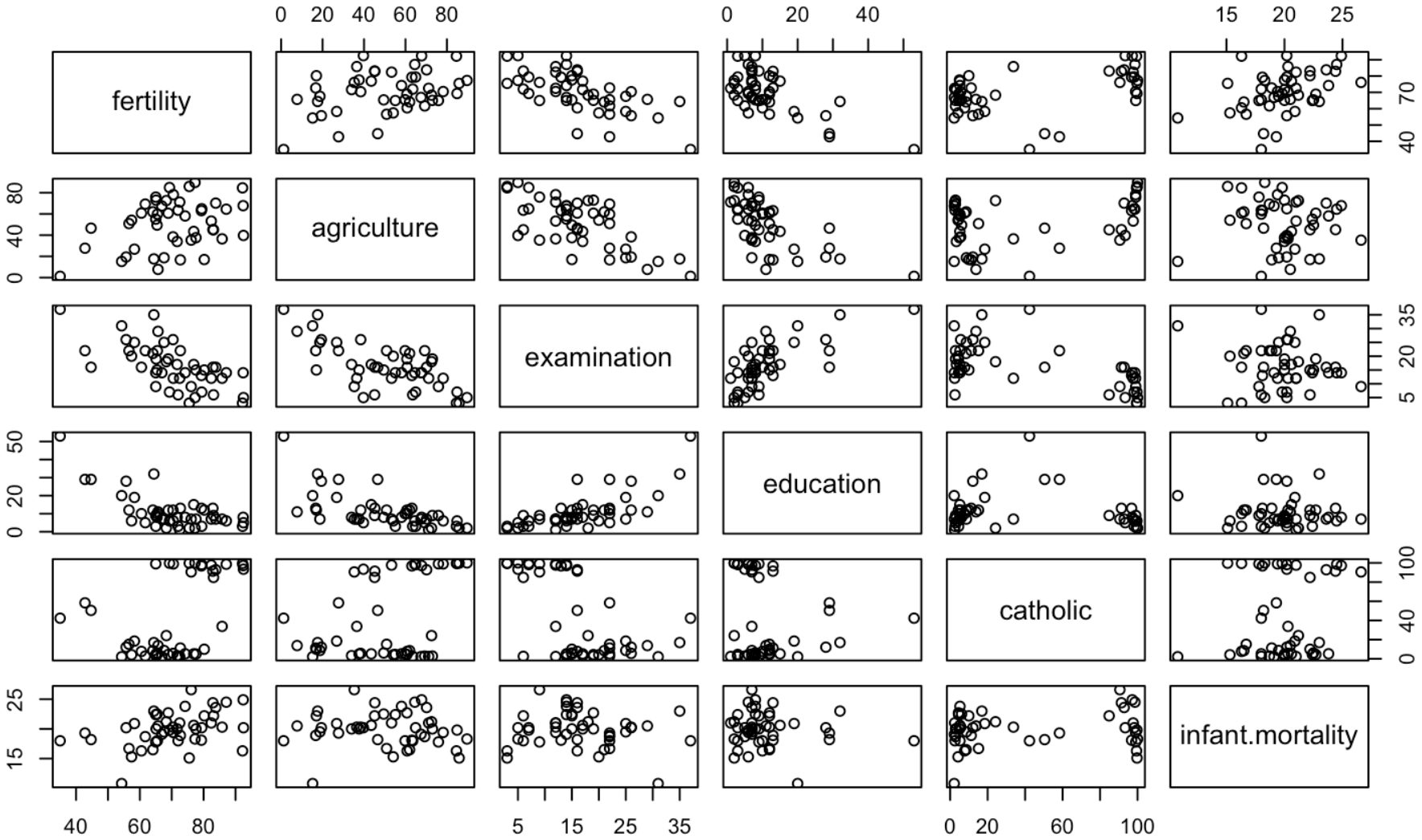


Total mediation: no direct effect of X_1 on Y- only direct effect is X_2 on Y
Example: Predict academic achievement (Y) from ethnicity (X_1) and economic opportunity (X_2)

Modeling

- In most cases, the statistics we calculate would be the same, but our interpretations of the effects may be quite different based on current knowledge and evidence-based theories
- We can only demonstrate consistency of sample data with a particular model
- Can never prove one specific model's accuracy

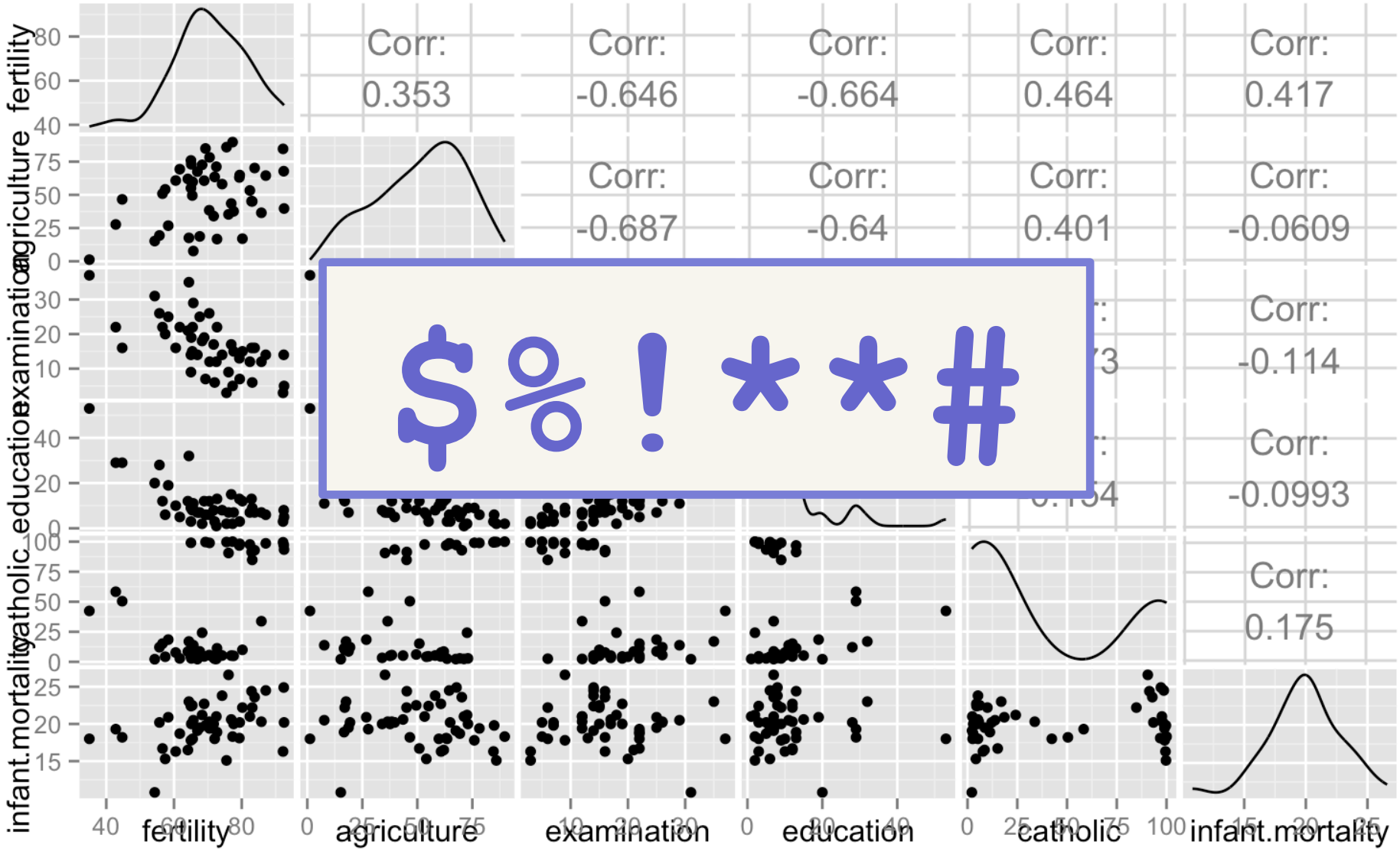
pairs(swiss)


fertility
agriculture
examination
education
catholic
infant.mortality

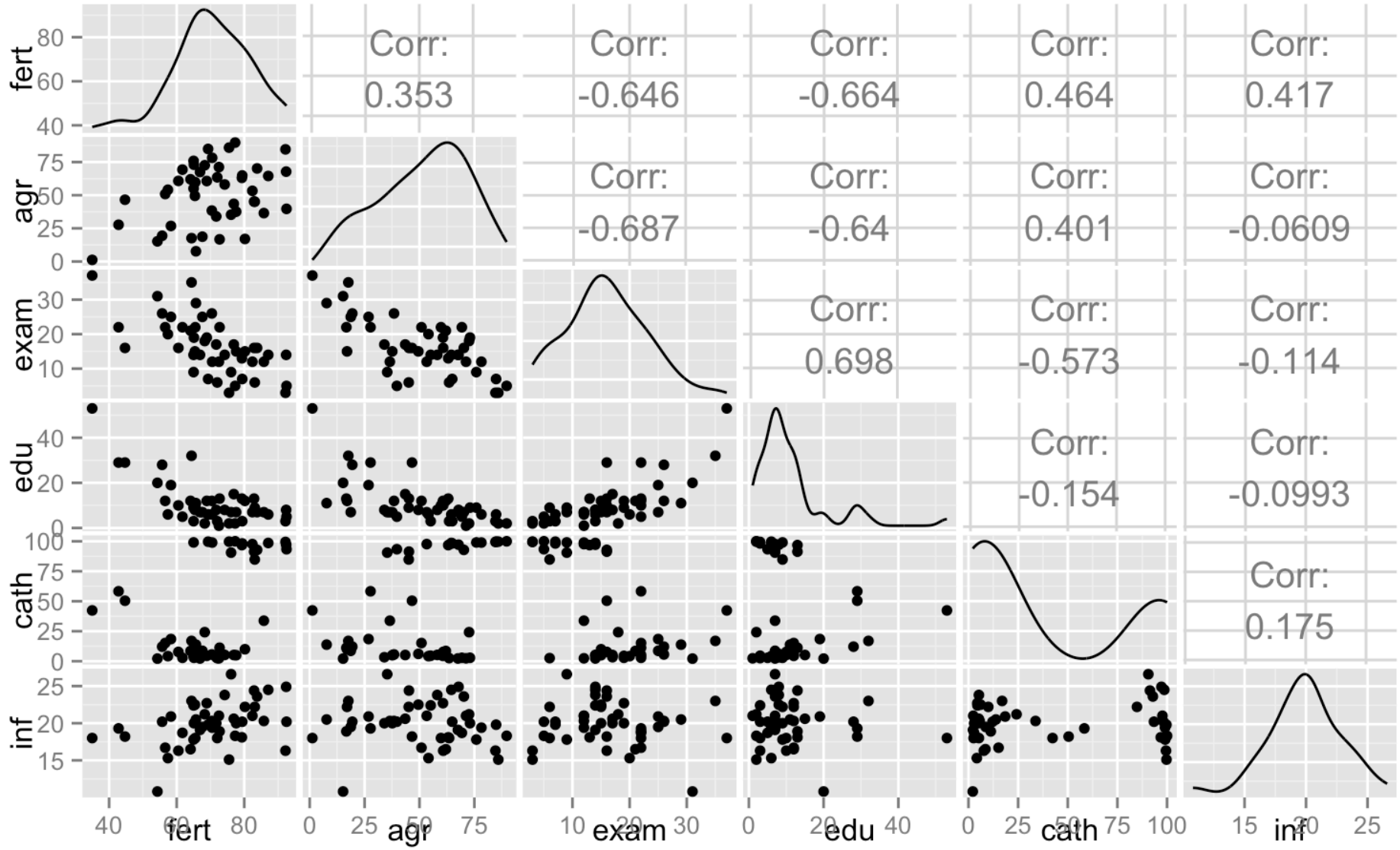
```
> corr.test(swiss) # library(psych)
Call:corr.test(x = swiss)
Correlation matrix
                 fertility agriculture examination education catholic infant.mortality
fertility             1.00        0.35       -0.65     -0.66     0.46             0.42
agriculture           0.35        1.00       -0.69     -0.64     0.40            -0.06
examination          -0.65       -0.69        1.00      0.70    -0.57            -0.11
education            -0.66       -0.64        0.70      1.00    -0.15            -0.10
catholic              0.46        0.40       -0.57     -0.15     1.00             0.18
Infant.mortality      0.42       -0.06       -0.11     -0.10     0.18             1.00
Sample Size
[1] 47
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                 fertility agriculture examination education catholic infant.mortality
fertility             0.00        0.09        0.00      0.00     0.01             0.03
agriculture           0.01        0.00        0.00      0.00     0.04             1.00
examination           0.00        0.00        0.00      0.00     0.00             1.00
education             0.00        0.00        0.00      0.00     1.00             1.00
catholic              0.00        0.01        0.00      0.30     0.00             1.00
Infant.mortality      0.00        0.68        0.45      0.51     0.24             0.00

 To see confidence intervals of the correlations, print with the short=FALSE option
> # print(corr.test(swiss), short = FALSE)
```

```
library(GGally)
ggpairs(swiss)
```



```
library(GGally)
ggpairs(swiss, columnLabels = c("fert", "agr", "exam", "edu", "cath", "inf"))
```



Kitchen sink linear regression

```
m_all <- lm(fertility ~ ., data = swiss)
```

```
fertility ~ agriculture + examination +
education + catholic + infant.mortality
```

New design matrix

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p} \end{pmatrix}$$

Where n is the number of observations;
p is the number of predictors in our regression formula

```
> model.matrix(m_all)
```

Kitchen sink linear regression

```
m_all <- lm(fertility ~ ., data = swiss)
tidy(m_all)
```

```
              term   estimate   std.error statistic      p.value
1      (Intercept) 66.9151817 10.70603759  6.250229 1.906051e-07
2      agriculture -0.1721140  0.07030392 -2.448142 1.872715e-02
3      examination -0.2580082  0.25387820 -1.016268 3.154617e-01
4        education -0.8709401  0.18302860 -4.758492 2.430605e-05
5         catholic  0.1041153  0.03525785  2.952969 5.190079e-03
6 infant.mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

```
glance(m_all)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.706735      0.670971 7.165369  19.76106 5.593799e-10  6 -156.0358
       AIC      BIC deviance df.residual
1 326.0716 339.0226 2105.043          41
```

```
fertility ~ agriculture + examination +
education + catholic + infant.mortality
```

Ordinary least squares

- Just as in simple linear regression, model is fit by minimizing the residual sums of squares…

$$RSS(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j \ X_{ij}))^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Where n is the number of observations;
p is the number of predictors in our regression formula

Estimating residual variance

- Also as in simple linear regression…

$$\widehat{\sigma}^2 = \frac{RSS}{n-p-1} \sim \sigma^2 \cdot \frac{\chi^2_{n-p-1}}{n-p-1}$$

Where n is the number of observations;
p is the number of predictors in our regression formula

Estimating residual variance

```
m_all <- lm(fertility ~ ., data = swiss)
glance(m_all) # broom
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.706735      0.670971 7.165369  19.76106 5.593799e-10  6 -156.0358
       AIC      BIC deviance df.residual
1 326.0716 339.0226 2105.043          41
sigma.hat <- sqrt(sum(resid(m_all)^2) / m_all$df.resid)
sigma.hat
[1] 7.165369
```

Interpreting regression coefficients

- Simple linear regression: coefficient is…

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\widehat{Cov}(X,Y)}{\widehat{Var}(X)}$$

- Multiple linear regression: coefficients are partial correlations, keeping all other predictors constant

$$\hat{\beta}_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \times \frac{sd_Y}{sd_1}$$

Here, 1 = X_1
2 = X_2

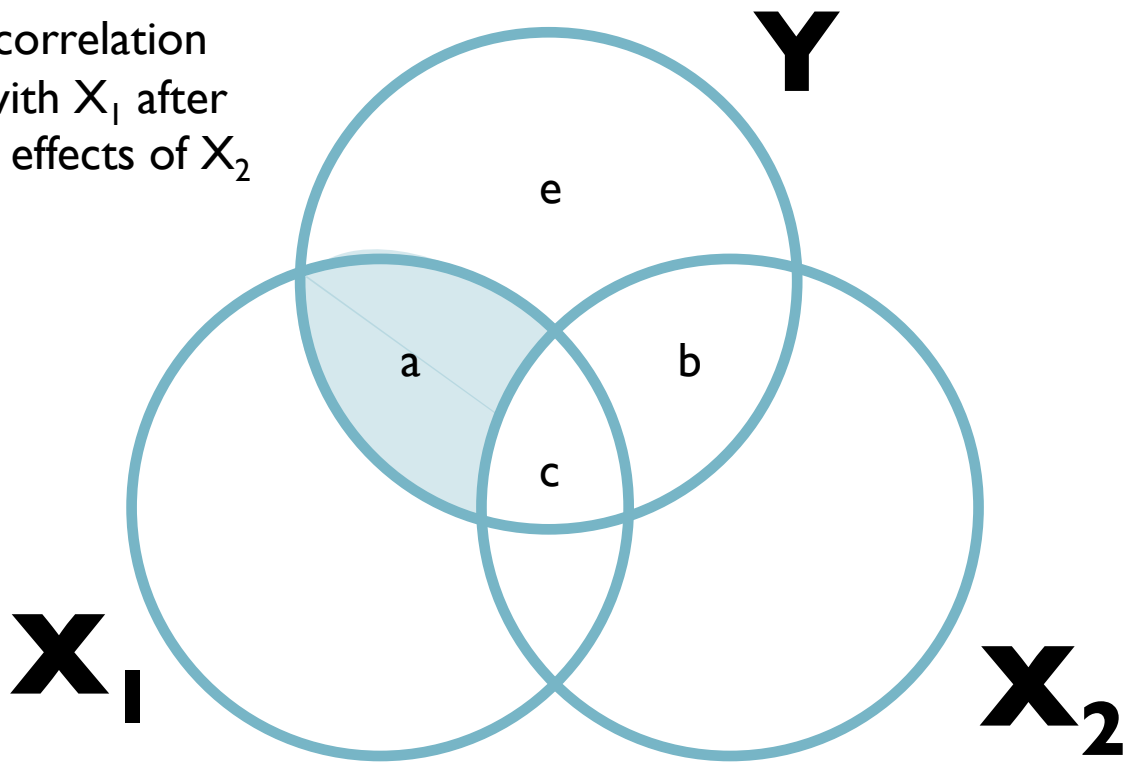
$$\hat{\beta}_{Y2.1} = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r_{12}^2} \times \frac{sd_Y}{sd_2}$$

NEW NOTATION!

- I'm adding dots to things! Even things with hats!
- The variables to the left of the dots are what we were dealing with in simple linear regression, as in the estimated regression coefficient predicting Y from X_1 (with no other predictors)
- Because the predictors now impact each other, you can read the first regression coefficient as the partial regression coefficient for Y on X_1 when X_2 is also in the equation
- Likewise, R^2 with .12 on the right means the multiple R squared like in simple linear regression with both X_1 and X_2 in the equation

$$R^2_{Y.12} = a + b + c$$

In words:
"a is the squared correlation between all of Y with X_1 after partialling out the effects of X_2 from X_1"



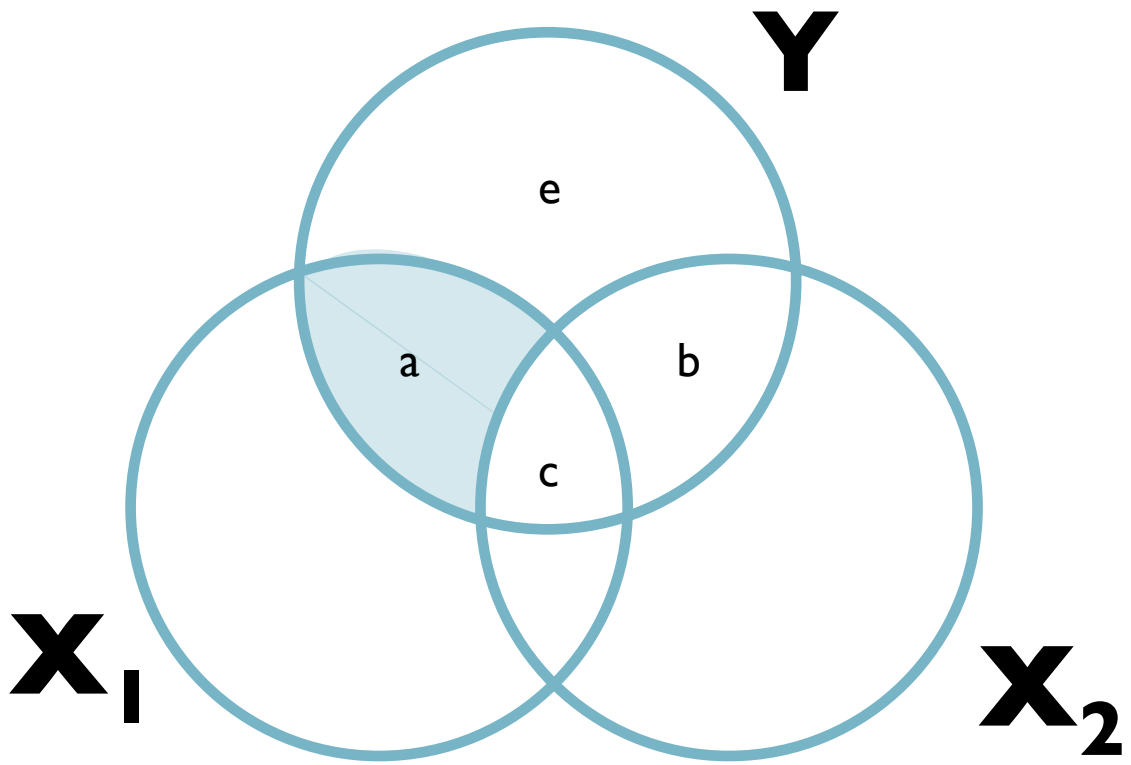
$$a = R^2_{Y.12} - r^2_{Y2}$$

$$b = R^2_{Y.12} - r^2_{Y1}$$

$$r^2_{Y1} = a + c$$

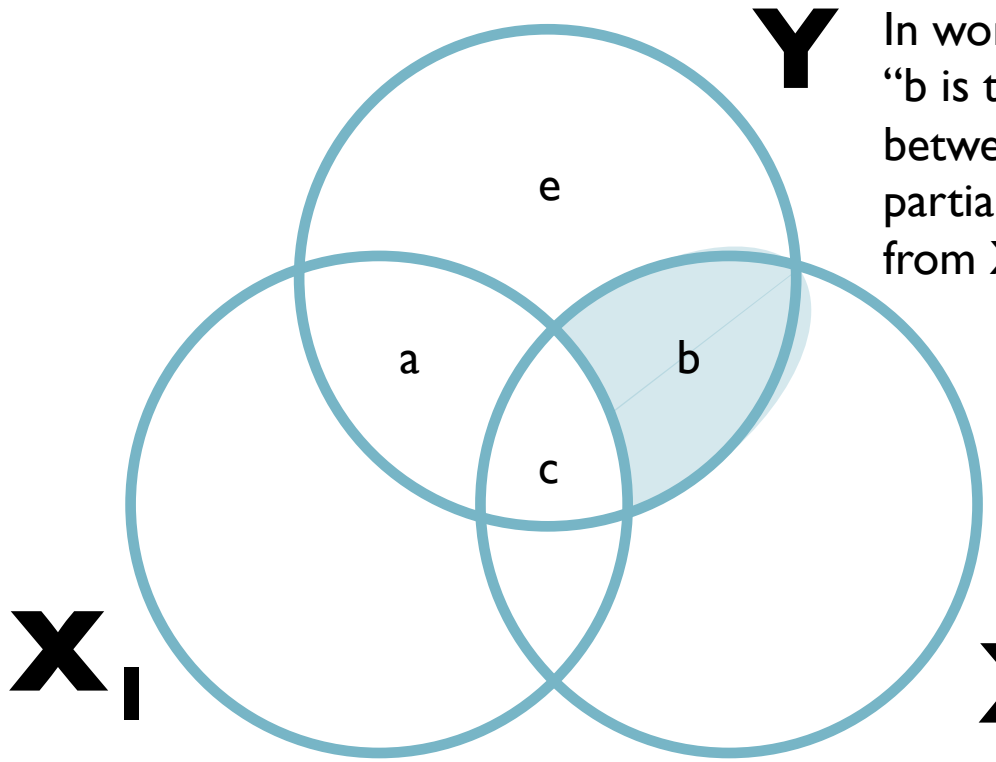
$$r^2_{Y2} = b + c$$

$$R^2_{Y.12} = a + b + c$$



$$\hat{\beta}_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r^2_{12}} \times \frac{sd_Y}{sd_1}$$

$$R^2_{Y.12} = a + b + c$$



In words:
"b is the squared correlation between all of Y with X_2 after partialling out the effects of X_1 from X_2"

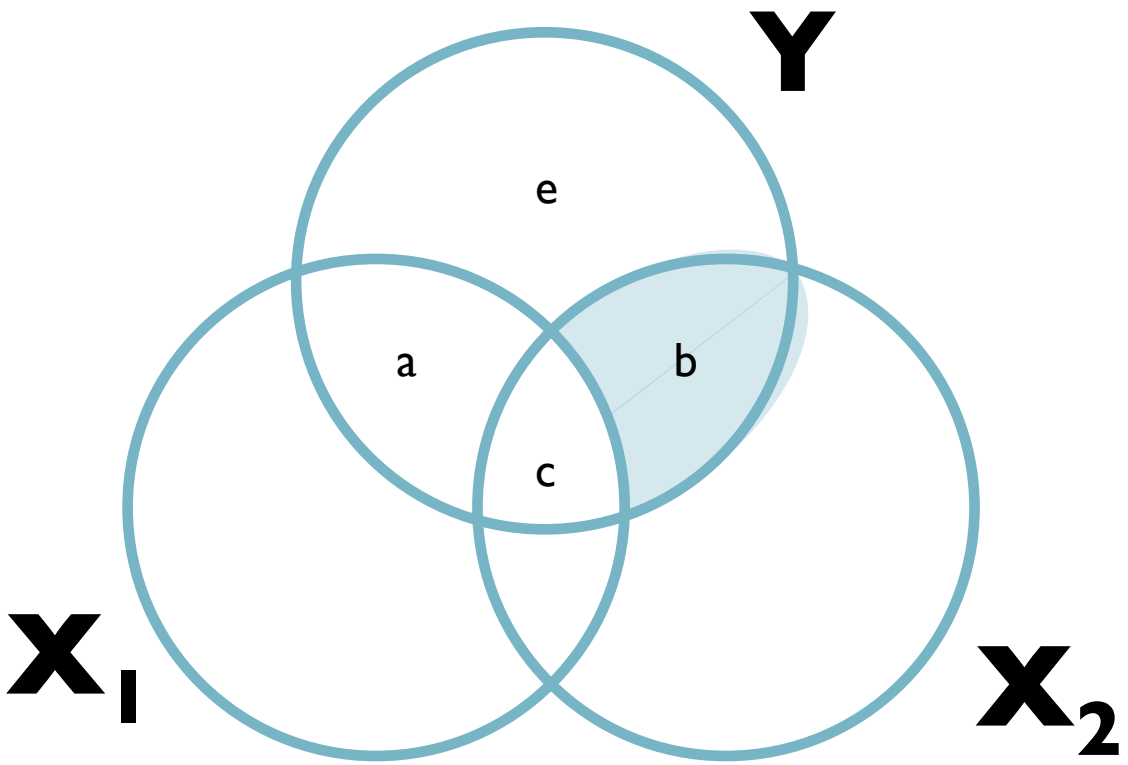
$$a = R^2_{Y.12} - r^2_{Y2}$$

$$b = R^2_{Y.12} - r^2_{Y1}$$

$$r^2_{Y1} = a + c$$

$$r^2_{Y2} = b + c$$

$$R^2_{Y.12} = a + b + c$$



$$\hat{\beta}_{Y2.1} = \frac{r_{Y2} - r_{Y1}r_{12}}{1 - r^2_{12}} \times \frac{sd_Y}{sd_2}$$

```
> partial_resid_agr <- resid(lm(agriculture ~  . - fertility, data=swiss))
> partial_resid_fert <- resid(lm(fertility ~  . - agriculture, data=swiss))
> summary(lm(partial_resid_fert ~ partial_resid_agr))
```

Kitchen sink linear regression

```
m_all <- lm(fertility ~ ., data = swiss)
tidy(m_all)
```

```
              term   estimate   std.error statistic      p.value
1      (Intercept) 66.9151817 10.70603759  6.250229 1.906051e-07
2      agriculture -0.1721140  0.07030392 -2.448142 1.872715e-02
3      examination -0.2580082  0.25387820 -1.016268 3.154617e-01
4        education -0.8709401  0.18302860 -4.758492 2.430605e-05
5         catholic  0.1041153  0.03525785  2.952969 5.190079e-03
6 infant.mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

```
glance(m_all)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.706735      0.670971 7.165369  19.76106 5.593799e-10  6 -156.0358
       AIC      BIC deviance df.residual
1 326.0716 339.0226 2105.043          41
```

```
fertility ~ agriculture + examination +
education + catholic + infant.mortality
```

```
> partial_resid_agr <- resid(lm(agriculture ~  . - fertility, data=swiss))
> partial_resid_fert <- resid(lm(fertility ~  . - agriculture, data=swiss))
> summary(lm(partial_resid_fert ~ partial_resid_agr))

Call:
lm(formula = partial_resid_fert ~ partial_resid_agr)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                              Estimate              Std. Error t value Pr(>|t|)
(Intercept)       -0.00000000000000007773  0.99764330584362781 45   0.000   1.0000
partial_resid_agr -0.17211397094145544 64  0.06710659947454739 04  -2.565   0.0137 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.839 on 45 degrees of freedom
Multiple R-squared:  0.1275,      Adjusted R-squared:  0.1081
F-statistic: 6.578 on 1 and 45 DF,  p-value: 0.01373
```

```
> partial_resid_agr <- resid(lm(agriculture ~  . - fertility, data=swiss))
> partial_resid_fert <- resid(lm(fertility ~  . - agriculture, data=swiss))
> summary(lm(partial_resid_fert ~ partial_resid_agr))

Call:
lm(formula = partial_resid_fert ~ partial_resid_agr)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                              Estimate              Std. Error t value Pr(>|t|)
(Intercept)       -0.00000000000000007773  0.99764330584362781 45   0.000   1.0000
partial_resid_agr -0.17211397094145544 64  0.06710659947454739 04  -2.565   0.0137 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.839 on 45 degrees of freedom
Multiple R-squared:  0.1275,      Adjusted R-squared:  0.1081
F-statistic: 6.578 on 1 and 45 DF,  p-value: 0.01373
```

We need new plots

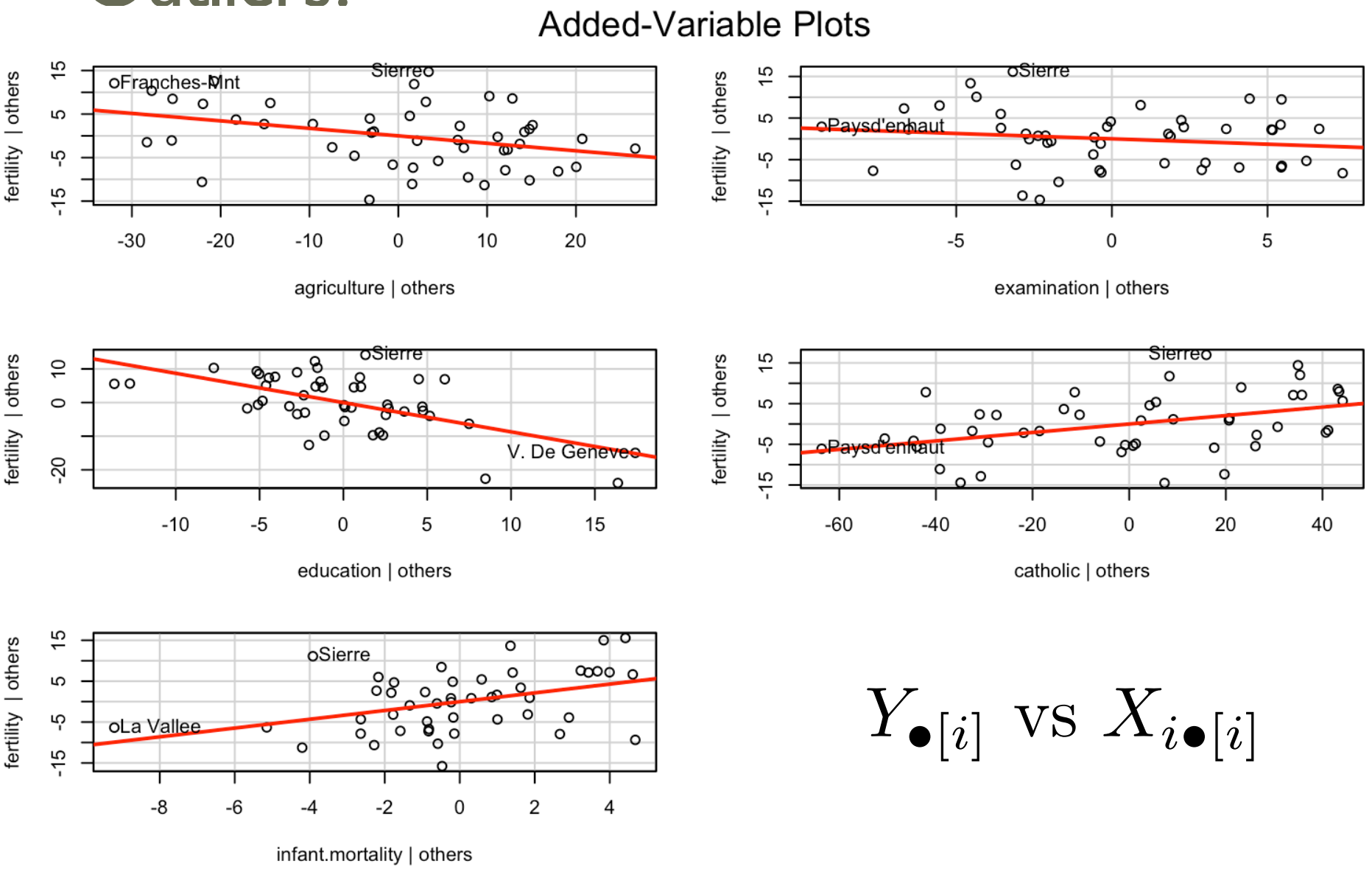
- Although it is useful in multiple regression to plot Y against each X, these plots can be misleading
- Our interest centers on the *partial* relationship between Y and each X, controlling for the other X's
- **Not** interested in the *marginal* relationship between Y and an individual X, ignoring the other X's (this is what we see in the scatterplot matrix)
- *Note:* scatterplot matrix **is** necessary for evaluating correlations *among* predictors

Added variable or partial regression plots

- The partial regression plot shows the partial relationship between the response and an explanatory variable, after the effect of the other explanatory variables has been removed
- X-axis is **not** X_i
- Plots show the correct strength of the linear relationship between Y and X_i
- Use to identify observations that exert joint influence
- Can reveal nonlinearity and suggest whether a relationship is monotone

$$Y_{\bullet[i]} \text{ vs } X_{i\bullet[i]}$$

Outliers?



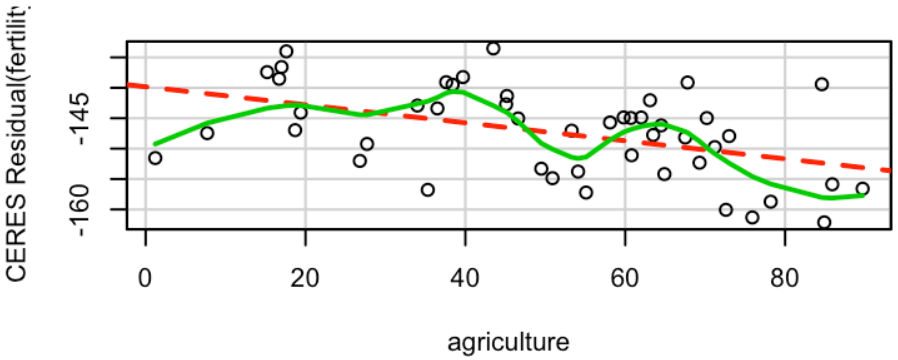
$$Y_{\bullet[i]} \text{ vs } X_{i\bullet[i]}$$

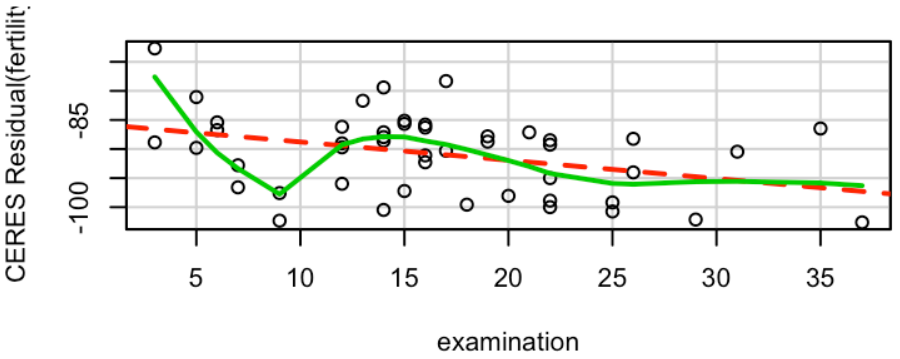
Component-residual or partial residual plots

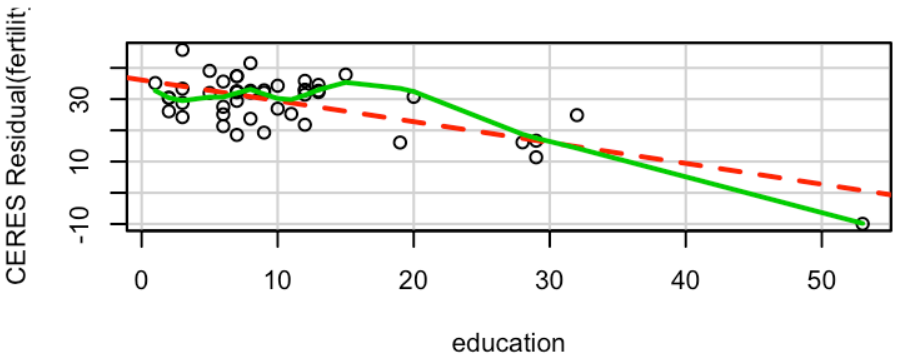
- Plotting residuals or studentized residuals against each X is frequently helpful for detecting departures from linearity.
- Partial residual plots are better for the detection of linearity; added variable plots are better for the detection of outliers and influential data points.
- X-axis is X_i, but plots do **not** show the correct strength of the linear relationship between Y and X_i

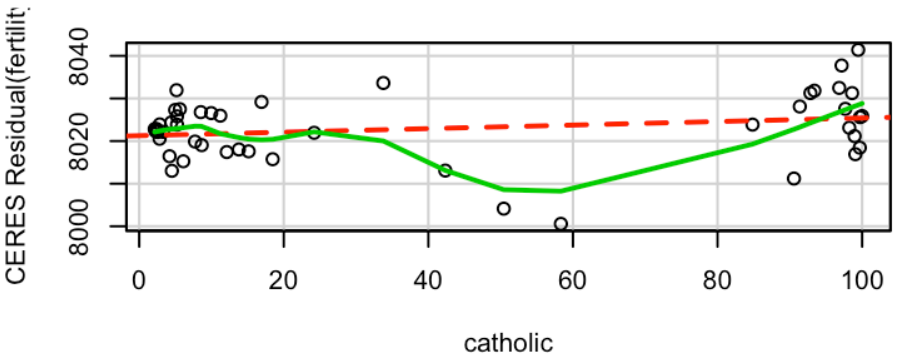
$$(e_i + \hat{\beta}_i X_i) \text{ vs. } X_i$$

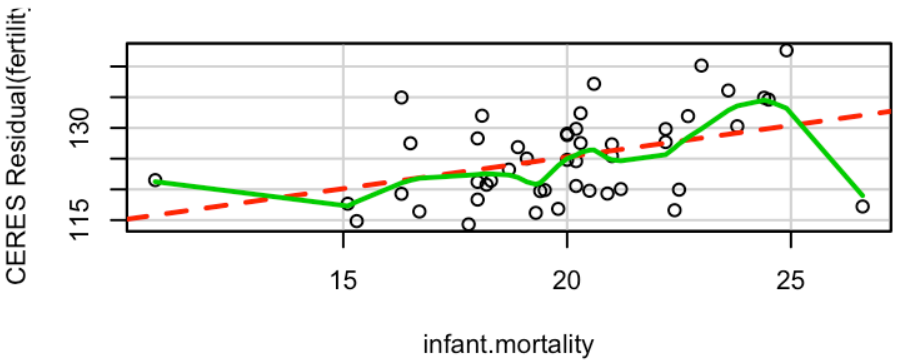
Linear?











$$(e_i + \hat{\beta}_i X_i) \text{ vs. } X_i$$

Partitioning variation in Y

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

$$MSS = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$R^2 = \frac{MSS}{TSS}$$

Squared multiple correlation, R^2

Total sums of squares (SS_{tot})	**Model sums of squares (SS_{mod})**	**Residual sums of squares (SS_{res})**
$\sum (y_i - \overline{y})^2$	$\sum (\hat{y}_i - \overline{y})^2$	$\sum (y_i - \hat{y}_i)^2$

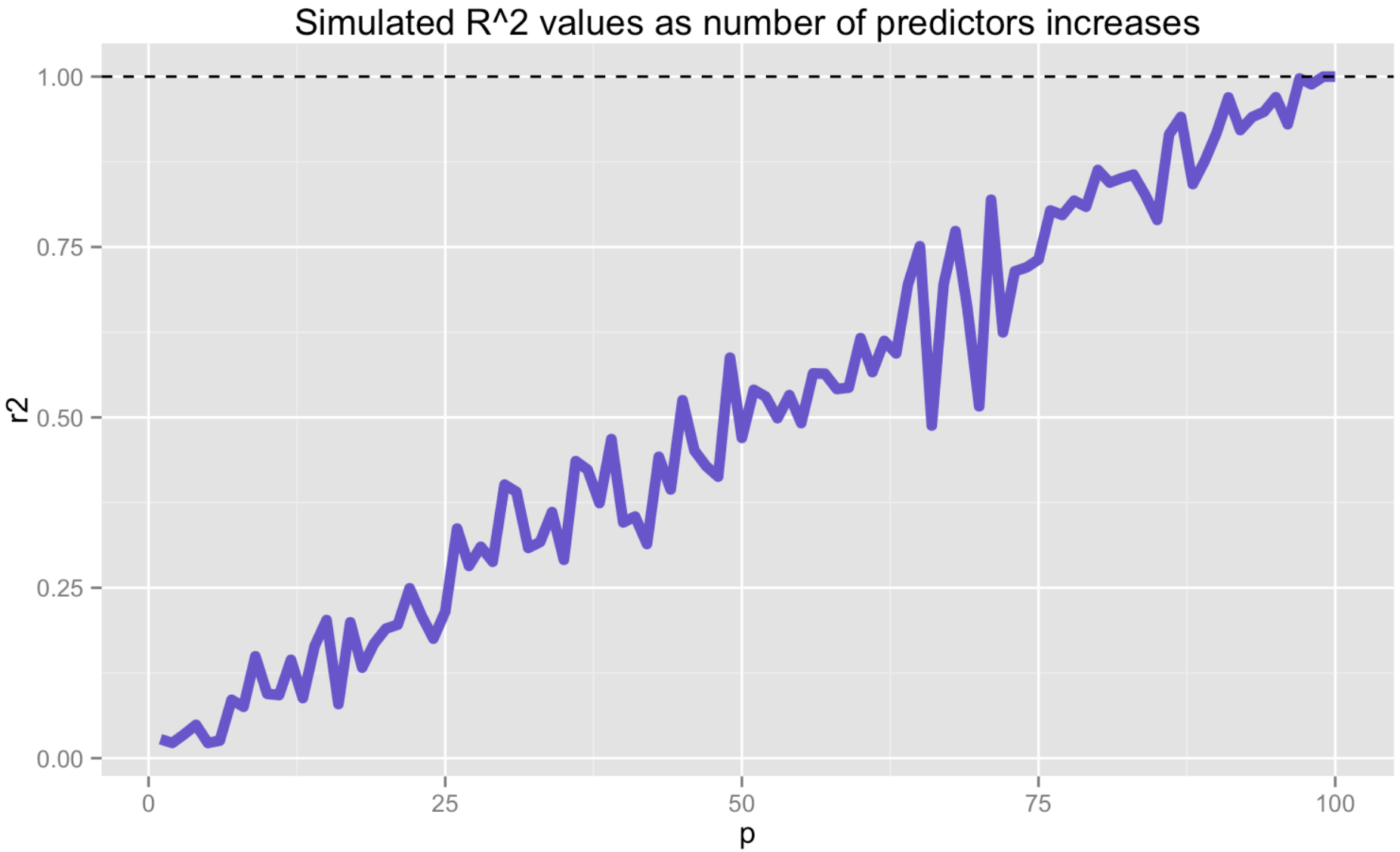
total variation = "explained" variation + residual variation

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Interpreting R^2 with > 1 predictor

- Weakness: denominator is fixed
- Numerator can ONLY INCREASE
- Capitalizes on *chance*
- Therefore, each additional variable added will at least not decrease the numerator, and will likely increase at least slightly
- Even if new predictor causes equation to become less efficient (even with higher R^2)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$


Simulated R^2 values as number of predictors increases
1.00
0.75
0.50
0.25
0.00
r2
0
25
50
75
100
p

Adjusted multiple R^2

Tries to account for bias in R^2 by replacing sums of squares by mean sums of squares, dividing by degrees of freedom

$$R_a^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{RMS}{TMS}$$

Kitchen sink linear regression

```
m_all <- lm(fertility ~ ., data = swiss)
tidy(m_all)
```

```
              term   estimate   std.error statistic      p.value
1      (Intercept) 66.9151817 10.70603759  6.250229 1.906051e-07
2      agriculture -0.1721140  0.07030392 -2.448142 1.872715e-02
3      examination -0.2580082  0.25387820 -1.016268 3.154617e-01
4        education -0.8709401  0.18302860 -4.758492 2.430605e-05
5         catholic  0.1041153  0.03525785  2.952969 5.190079e-03
6 infant.mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

```
glance(m_all)
```

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.706735      0.670971 7.165369  19.76106 5.593799e-10  6 -156.0358
       AIC      BIC deviance df.residual
1 326.0716 339.0226 2105.043          41
```

```
fertility ~ agriculture + examination +
education + catholic + infant.mortality
```

Overall fit of model: F-statistic

Total sums of squares (SS_{tot})	Model sums of squares (SS_{mod})	Residual sums of squares (SS_{res})
$\sum(y_i - \overline{y})^2$	$\sum(\hat{y}_i - \overline{y})^2$	$\sum(y_i - \hat{y}_i)^2$

```
> Y <- swiss$fertility
> n <- length(Y)
> TSS <- sum((Y - mean(Y))^2)
> TMS <- TSS / (n - 1)
> RSS <- sum(resid(m_all)^2)
> RMS <- RSS / m_all$df.residual
> MSS <- TSS - RSS
> MMS <- MSS / (n - 1 - m_all$df.residual)
> print(c(TMS,RMS,MMS))
[1]  156.04250   51.34251 1014.58239
> MMS/RMS
[1] 19.76106
```

$$F = \frac{MMS}{RMS}$$

Sequences of nested models (revisited)

- Full(er) model: $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots \beta_p X_{ip} + \varepsilon_i$

- (Default) reduced model: $Y_i = \beta_0 + \varepsilon_i$

Nested models

$$F = \frac{(RSS(R) - RSS(F))/(df_R - df_F)}{RSS(F)/df_F}.$$

Nested models

```
m_mean <- lm(fertility ~ 1, data = swiss) # intercept ONLY
m_all <- lm(fertility ~ ., data = swiss)
anova(m_mean, m_all)
Analysis of Variance Table

Model 1: fertility ~ 1
Model 2: fertility ~ agriculture + examination + education + catholic +
    infant.mortality
  Res.Df  RSS Df Sum of Sq      F    Pr(>F)
1     46 7178
2     41 2105  5    5072.9 19.761 5.594e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Kitchen sink linear regression

```
m_all <- lm(fertility ~ ., data = swiss)
tidy(m_all)
```

```
              term   estimate   std.error statistic      p.value
1      (Intercept) 66.9151817 10.70603759  6.250229 1.906051e-07
2      agriculture -0.1721140  0.07030392 -2.448142 1.872715e-02
3      examination -0.2580082  0.25387820 -1.016268 3.154617e-01
4        education -0.8709401  0.18302860 -4.758492 2.430605e-05
5         catholic  0.1041153  0.03525785  2.952969 5.190079e-03
6 infant.mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

```
glance(m_all)
```

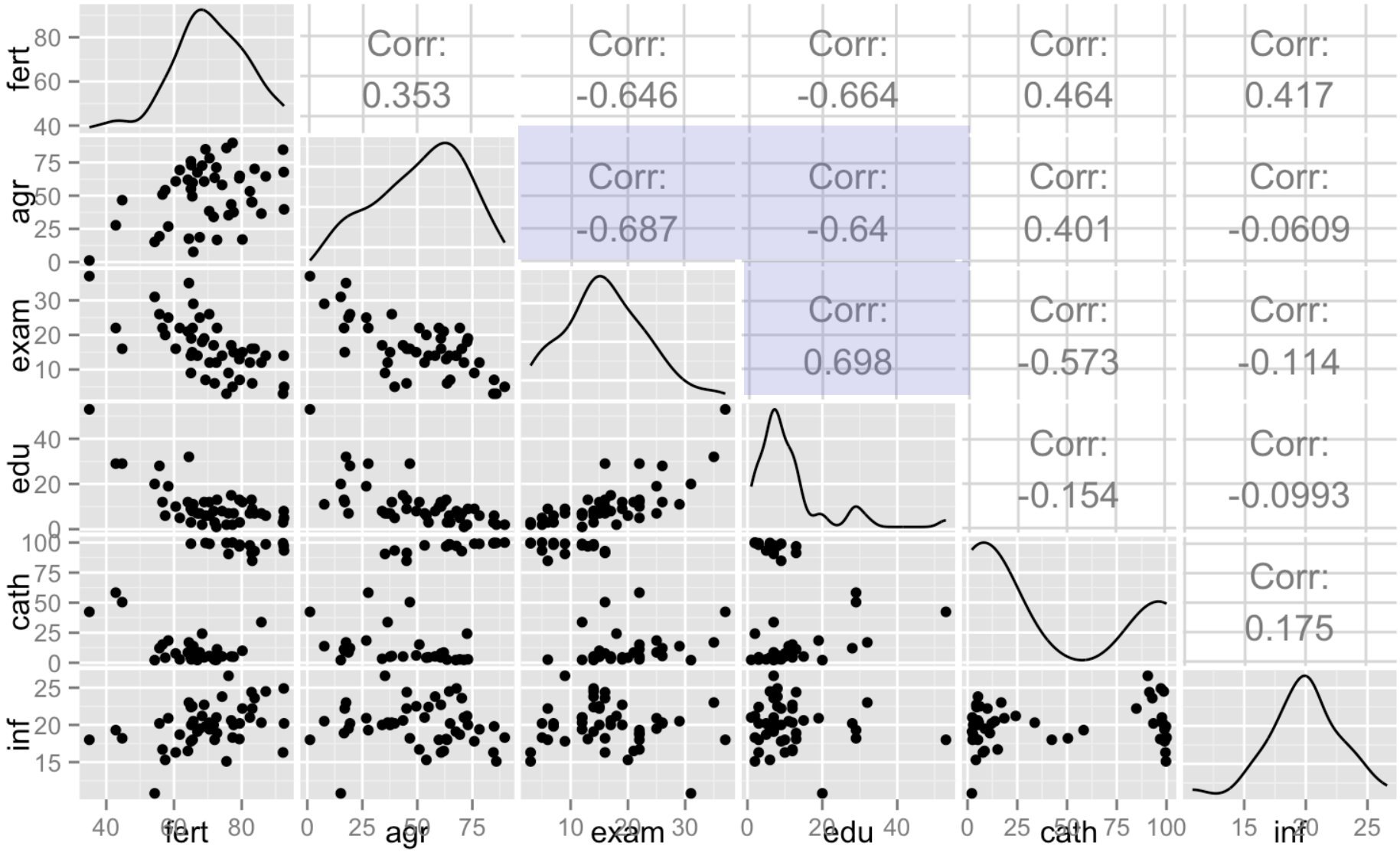
```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik
1  0.706735      0.670971 7.165369  19.76106 5.593799e-10  6 -156.0358
       AIC      BIC deviance df.residual
1 326.0716 339.0226 2105.043          41
```

```
fertility ~ agriculture + examination +
education + catholic + infant.mortality
```

What may have gone wrong here?

- A lot!
- One predictor looks to be bi-modal, and should probably be converted to a categorical predictor as its mean does not appear to be very meaningful (catholic)
- 3 predictors are highly positively correlated (agriculture, education, and examination)
- You may not have noticed, but agriculture in our simple linear regression model had a positive coefficient. Now it is negative! (Simpson's paradox)

```
library(GGally)
ggpairs(swiss, columnLabels = c("fert", "agr", "exam", "edu", "cath", "inf"))
```



Multicollinearity

- Predictor variables are linearly dependent
- 2 types:
  - Extreme
  - Near extreme
- Only problematic for interpreting either of the regression coefficients for the collinear predictors
- Consequence: increases standard errors for collinear coefficient estimates

Extreme multicollinearity

- At least one predictor is a perfect linear function of one or more of the other predictors
- Let's make a trivial one up in our swiss dataset: we'll add 5 to every value of agriculture

```
swiss2 <- swiss %>%
  mutate(agr_5 = agriculture + 5)
```

```
> m_col <- lm(fertility ~ agriculture + agr_5, data = swiss2)
> summary(m_col)

Call:
lm(formula = fertility ~ agriculture + agr_5, data = swiss2)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5374  -7.8685  -0.6362   9.0464  24.4858

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.30438    4.25126  14.185   <2e-16 ***
agriculture  0.19420    0.07671   2.532   0.0149 *
agr_5             NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.82 on 45 degrees of freedom
Multiple R-squared:  0.1247,     Adjusted R-squared:  0.1052
F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
```

Matrix rank

- The *rank* of a matrix is the number of columns that are independent of all the others.
- If the rank is smaller than the number of columns, then the LSE are not unique.
- In R, we can obtain the rank of matrix with the function qr:

```
> m_col_X <- model.matrix(m_col)
> ncol(m_col_X) # number of columns in my model matrix
[1] 3
> qr(m_col_X)$rank
[1] 2
```

Near-extreme multicollinearity

- Much more common!
- No standard cut-off, but I get concerned if any zero order correlations are > .6
- Consequence: standard errors of collinear coefficients gets large
- As multicollinearity increases, gets harder and harder to separate and estimate effects of predictors, making it harder to detect differences from 0

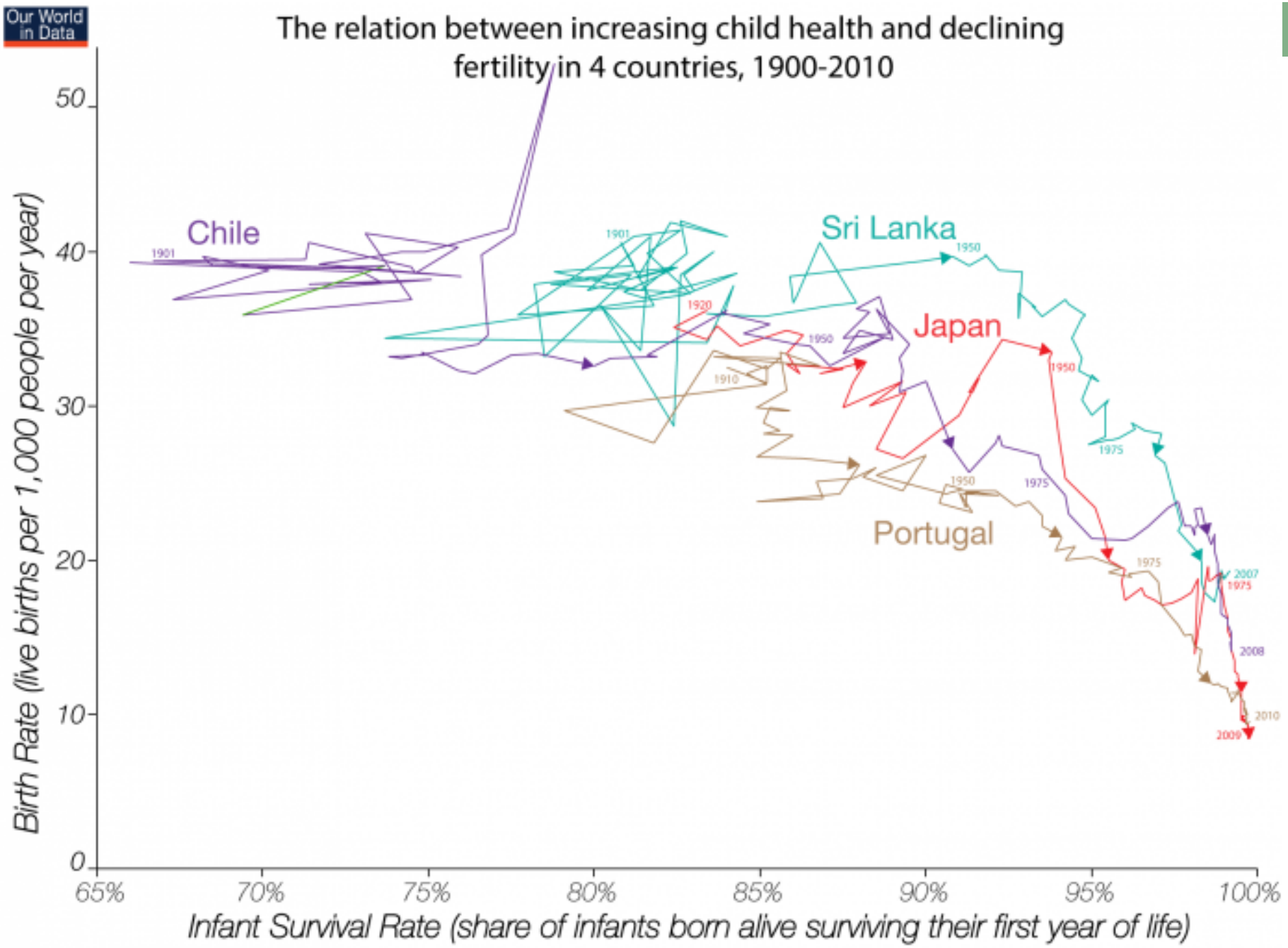
Variance inflation factor (vif)

```
> car::vif(m_all) # library(car)
     agriculture      examination        education         catholic infant.mortality
        2.284129         3.675420         2.774943         1.937160         1.107542
> sqrt(car::vif(m_all)) # library(car)
     agriculture      examination        education         catholic infant.mortality
        1.511334         1.917138         1.665816         1.391819         1.052398
```

index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.
Proposed cut-offs? 2? 3? 4? 10?

How should I have gone about this?

- Formulate educated hypotheses
  - Let's pretend I collected this data, and picked these 6 variables based on theory
- Inspect scatterplot matrix
  - Infant.mortality is in
  - Agriculture, examination and education are too highly correlated. I should pick one or two and assess variance inflation
  - Split catholicism somewhere that makes sense
- Do some nested models



Births per woman by income level (latest data) – By Max Roser

Our World in Data

Pink bubbles ● show country averages for income (GDP per capita, PPP adjusted) and for the total fertility rate.
For all other countries the fertility rate is shown for each wealth quintile within the country. It is plotted against the average income per quintile in the same country.

Data sources: World Bank for all income measures. Fertility national averages from WDI. Within country inequality of fertility from WHO (based on DHS) – except China for which data was added from various research papers. Most data are from 2013 – none of the data refer to a year earlier than 2005.