Data Analytics for Social Science
Linear regression

Johan A. Elkink

School of Politics & International Relations
University College Dublin

28 February 2017

Outline

Linear model

The regression equation here is

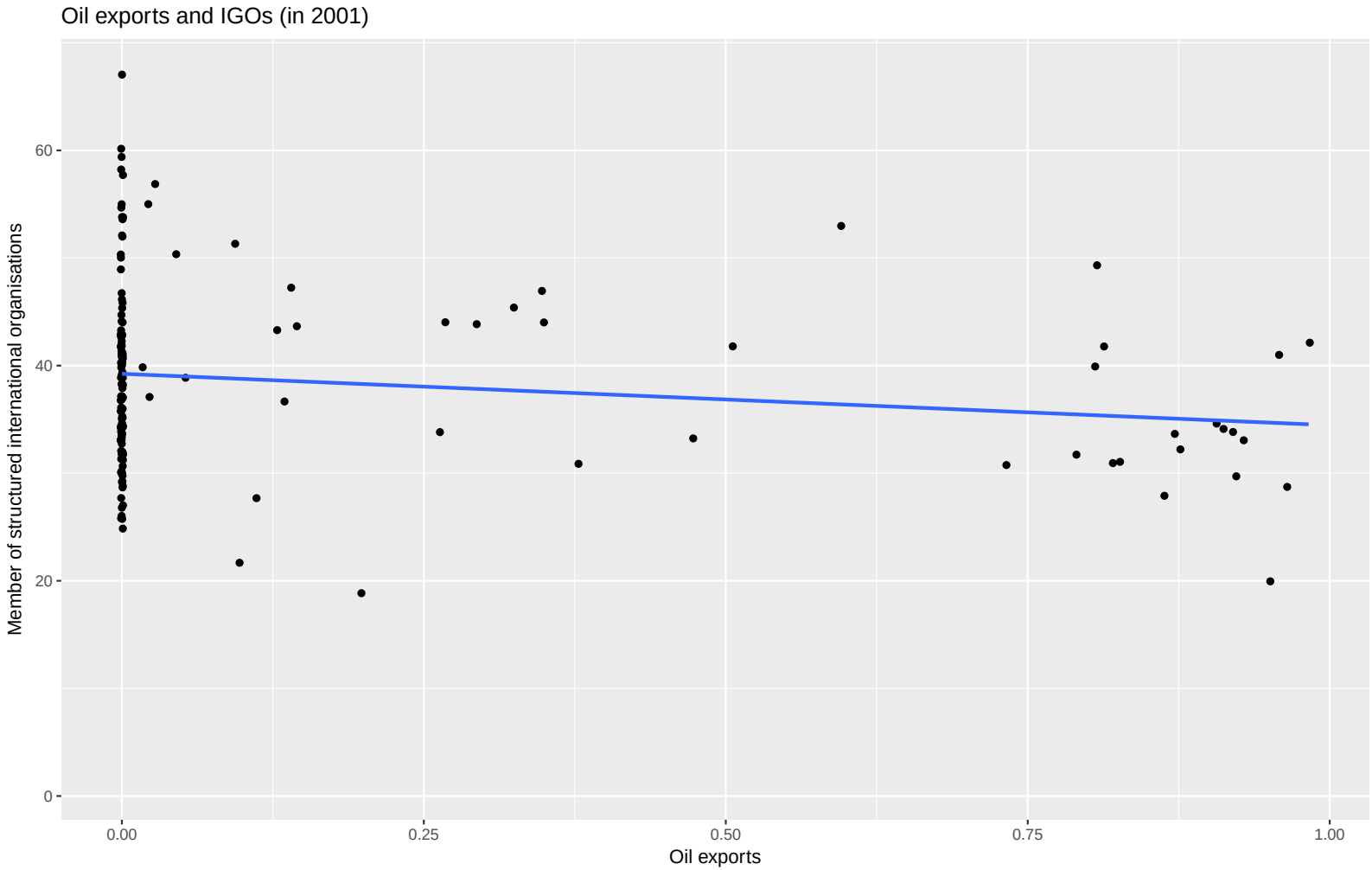
$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

whereby **y** is the dependent variable, **x** the independent variable, i an indicator of the case (e.g. country), β_1 and β_2 the model parameters, and ε the error term.

The model thus uses a linear combination of the independent variables to predict or explain the dependent variable, whereby both are assumed to be interval or ratio variables.

The high level of **parsimony** of the model reduces the risk of **overfitting** and thus helps us to generalize (cf. the lines we saw with `geom_smooth()`).

Linear model



Residuals

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

The linear prediction given the parameters would be $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$.

The extend to which the real value differs from the predicted value is:

$$y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i = e_i.$$

By this formulation, the **residuals** ($\boldsymbol{e}$) are the vertical distances between a point and the regression line (i.e. not the shortest distance between the point and the line).

Terminology

- **y** is the **dependent** variable
  - also known as **regressand**
- **X** are the **independent** variables
  - also known as **explanatory** variables
  - also known as **regressors** or **predictors** (or **factors**, **carriers**)
  - **X** is sometimes called the **design matrix** or **factor space**
- **y** is **regressed on X**
- β is the population **parameter**, $\hat{\beta}$ the estimated **coefficient**.
- The **error** term or **disturbance** $\varepsilon = \mathbf{y} - \mathbf{X}\beta$.
- The difference between the observed and predicted dependent variable is the **residual** $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

Linear model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$$

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$$

Components

Two components of the model:

$$\begin{array}{l|l} \mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2) & \text{Stochastic} \\ \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} & \text{Systematic} \end{array}$$

Generalised version (not necessarily linear):

$$\begin{array}{l|l} \mathbf{y} \sim f(\boldsymbol{\mu}, \boldsymbol{\alpha}) & \text{Stochastic} \\ \boldsymbol{\mu} = g(\mathbf{X}, \boldsymbol{\beta}) & \text{Systematic} \end{array}$$

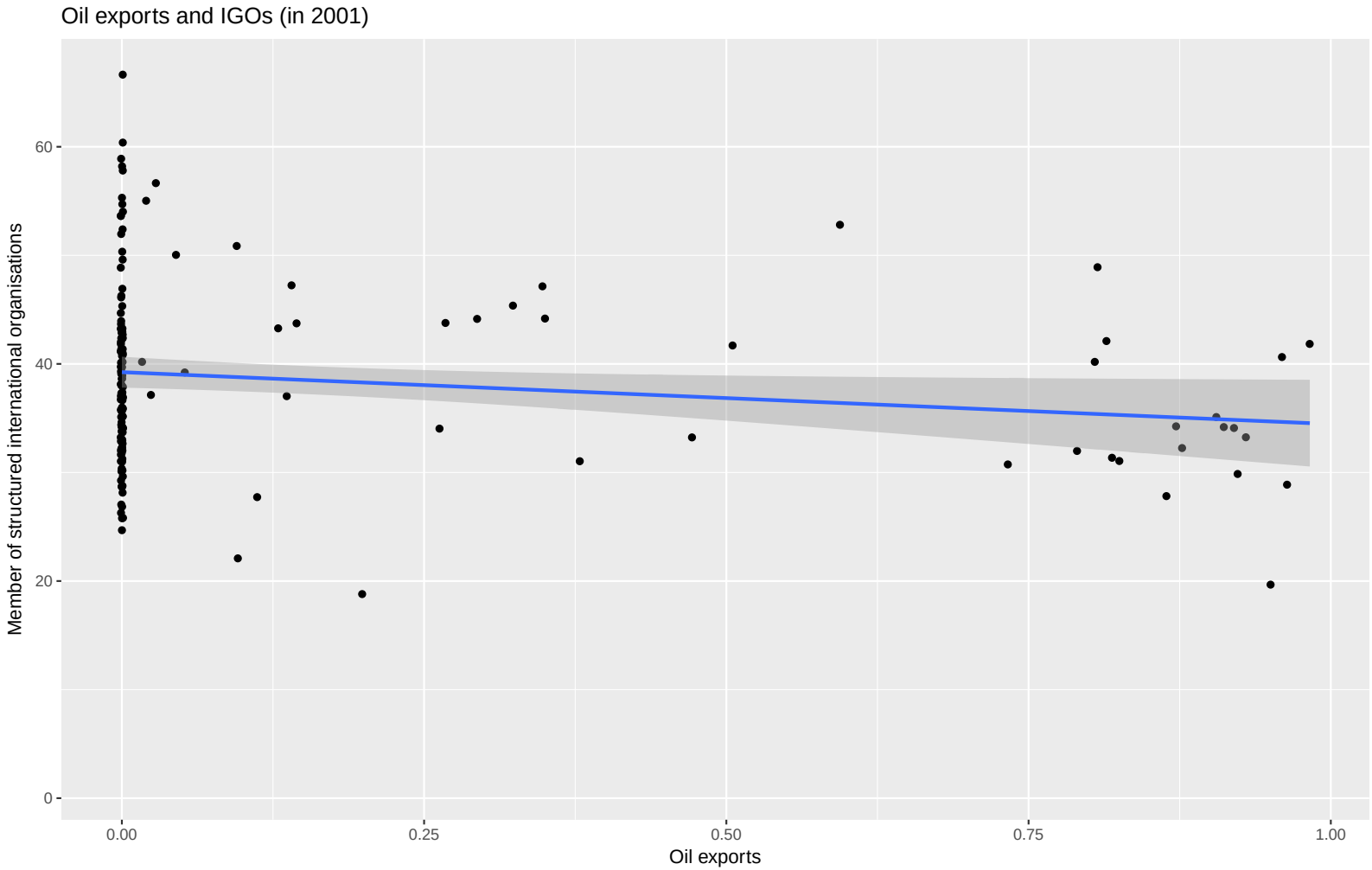
Two types of uncertainty:

Estimation uncertainty: lack of knowledge about $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$; can be reduced by increasing n.

Fundamental uncertainty: represented by stochastic component and exists independent of researcher.

(King, 1998, 8)

Linear model



Outline

Ordinary Least Squares

To estimate the regression line, we need to estimate the parameters β_1 and β_2.

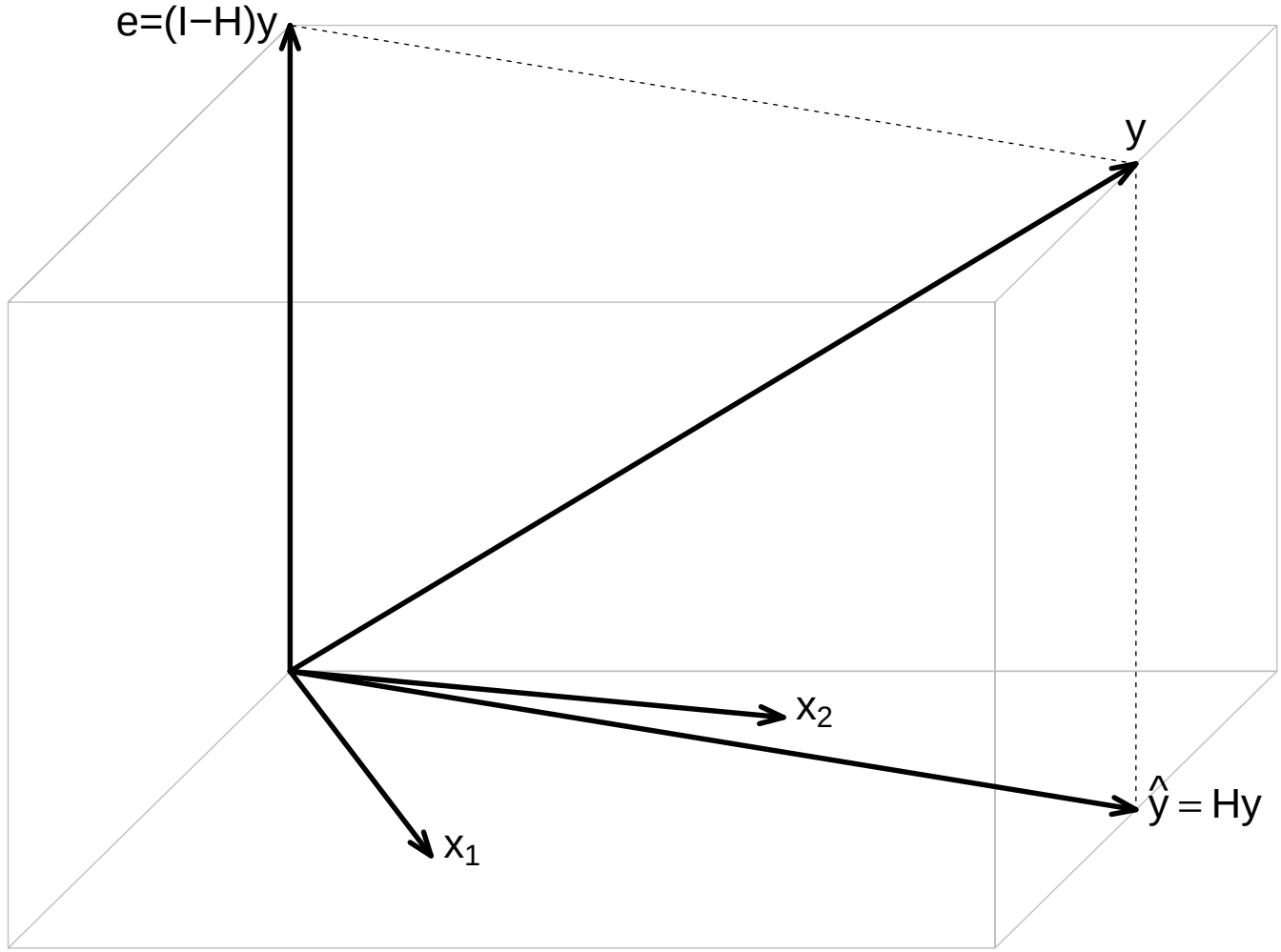
For the **linear** model, the most popular method of estimation is **ordinary least squares** (OLS).

$$\hat{\boldsymbol{\beta}}^{OLS} = \arg\min_{\hat{\boldsymbol{\beta}}} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

With OLS, we estimate the parameters such that the **sum of squared residuals** are minimized. This is the same as minimizing the variance of the residuals.

OLS is the **best linear unbiased estimator** (BLUE).

OLS as projection



$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

OLS assumptions

specification

Linear in parameters (i.e. $f(\mathbf{X}\beta) = \mathbf{X}\beta$ and $E(\mathbf{y}) = \mathbf{X}\beta$)
No extraneous variables in $\mathbf{X}$
No omitted independent variables
Parameters to be estimated are constant
Number of parameters is less than the number of cases, $k < n$

disturbances

Errors have an expected value of zero, $E(\varepsilon|\mathbf{X}) = 0$
Errors are normally distributed, $\varepsilon \sim N(0, \sigma^2)$
Errors have a constant variance, $var(\varepsilon|\mathbf{X}) = \sigma^2 < \infty$
Errors are not autocorrelated, $cov(\varepsilon_i, \varepsilon_j|\mathbf{X}) = 0 \quad \forall \quad i \neq j$
Errors and $\mathbf{X}$ are uncorrelated, $cov(\mathbf{X}, \varepsilon) = 0$

regressors

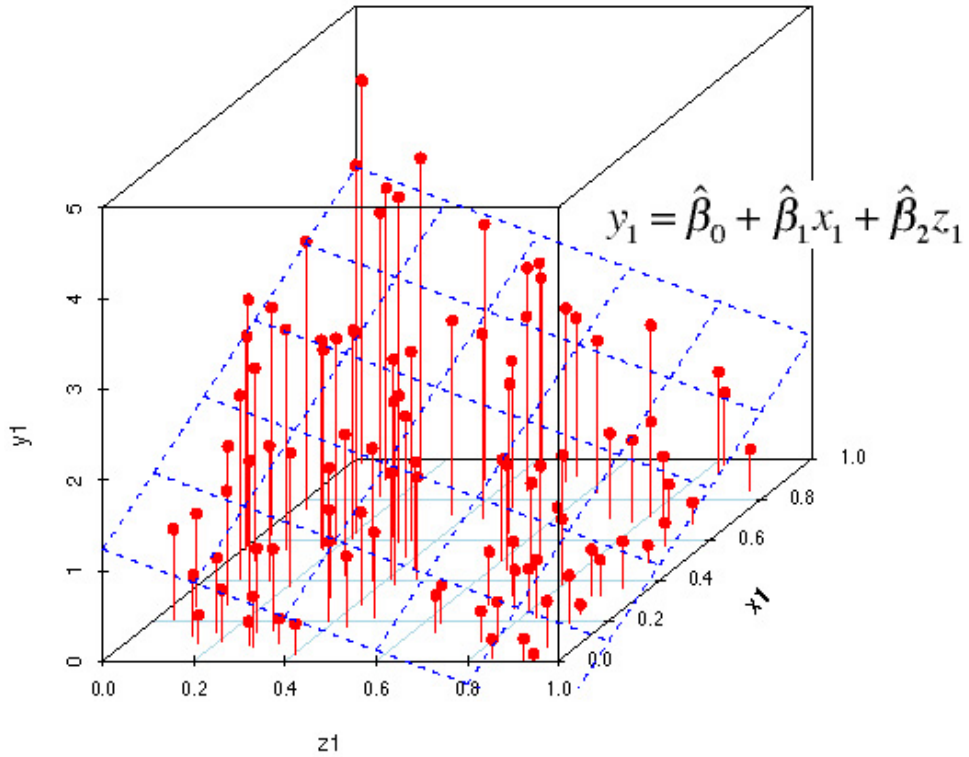
$\mathbf{X}$ varies and is of full column rank (note: requires $k < n$)
No measurement error in $\mathbf{X}$
No endogenous variables in $\mathbf{X}$

Outline

Multiple regression

For causal inference, we can add **control variables** to capture confounding factors, such as here a dummy variable where the country is a democracy, the log of GDP, and the log of population size.

For predictive inference, we can add variables to get a more accurate prediction.



Multiple regression

	Dependent variable:
	strucint
oilexp	-4.322^{**}
	(1.901)
democracy	3.726^{***}
	(1.188)
lngdp	2.306^{***}
	(0.346)
lnpop	1.920^{***}
	(0.266)
Constant	-11.670^{**}
	(5.296)

$$strucint_i = \beta_1 + \beta_2 oilexp_i + \beta_3 democracy_i + \beta_4 lngdp_i + \beta_5 lnpop_i + \varepsilon_i$$

t- and *F*-tests

The two most important tests in regression are, for each coefficient, the ***t*-test**:

$H_0 : \beta = 0$ and $H_1 : \beta \neq 0$

If a β is zero, it means there is no relationship between the respective independent variable and the dependent variable.

For all coefficients together we have the ***F*-test**:

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = 0$ and the alternative that not all of them are zero.

If the *F*-test is not significant, the model explains very little of the dependent variable.

(Insignificant *t*-tests in combination with a significant *F*-test is an indication of multicollinearity.)

Dummy variables

Linear regression is only suitable for continuous variables.

However, when a categorical variable has only two categories, one coded as "1" and the other as "0", it is reasonable to insert into a regression as explanatory variable, take x_i a continuous and d_i a **dummy** variable:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i$$

We can distinguish two scenarios:

$$\begin{array}{l|l} d_i{=}0 & y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i = \beta_1 + \beta_2 x_i \\ d_i{=}1 & y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i = (\beta_1 + \beta_3) + \beta_2 x_i \end{array}$$

So the coefficient β_3 is the *difference in intercept* for those where $d_i = 0$ and those where $d_i = 1$.

Nominal variables

Dealing with a nominal variable with multiple categories as explanatory variable is then straightforward: create multiple dummy variables, one for each category.

party	ff	fg	lab	other
ff	1	0	0	0
ff	1	0	0	0
fg	0	1	0	0
labour	0	0	1	0
fg	0	1	0	0
other	0	0	0	1

Remember to always leave one of the dummy variables out of the regression, which then becomes the **reference category**.

$$leftRight_i = \beta_1 + \beta_2 ff_i + \beta_3 lab_i + \beta_4 other_i$$

This means that β_2 represents the difference in intercept between FF and FG; β_3 between Labour and FG; etc.

Outline

Sums of squares

SST Total sum of squares $\sum(y_i - \bar{y})^2$

SSE Explained sum of squares $\sum(\hat{y}_i - \bar{y})^2$

SSR Residual sum of squares
$\sum e_i^2 = \sum(\hat{y}_i - y_i)^2 = \mathbf{e}'\mathbf{e}$

The key to remember is that **SST = SSE + SSR**

Sometimes instead of "explained" and "residual", "regression" and "error" are used, respectively, so that the abbreviations are swapped (!).

R^2

Once we have estimated a line, we might ask how well this line summarizes the relationship between those two variables.

A common measure is R^2:

$$R^2 = 1 - \frac{\text{residual sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

This can be interpreted as the proportion of the variation in **y** explained by this model.

Note the relation with correlation coefficient Peason's r: $r = \sqrt{R^2}$.

R^2 rises with the addition of more explanatory variables. For this reason we often report the **adjusted** R^2:

$$1 - (1 - R^2)\frac{n-1}{n-k}.$$

Akaike Information Criterion (AIC)

Another approach to making a similar balance between parsimony and explained variance is **Akaike Information Criterion**:

$$AIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{2k}{n}$$

Thus the smaller AIC, the better.

$$BIC = \log\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) + \frac{(\log n)k}{n}$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

and there are many more similar variations.

Multicollinearity

When two or more variables in **X** are highly correlated:

- It will be harder to estimate **b**.
- Each individual **x** adds less to the prediction of **y**.

We can identify high multicollinearity by looking at **variance inflation factors**, or VIF scores—a VIF score of 4 or higher, approximately, raises concerns.

```
library(faraway)
m <- lm(y ~ x1 + x2 + x3, data)
vif(m)
```

Stepwise regression

Stepwise regression is an algorithm where the computer adds or removes variables and evaluates the improvements in fit (e.g. AIC).

Step-Forward: adding variables until candidates add too little in terms of prediction.

Step-Backward: removing variables until candidates remove too much in terms of prediction.

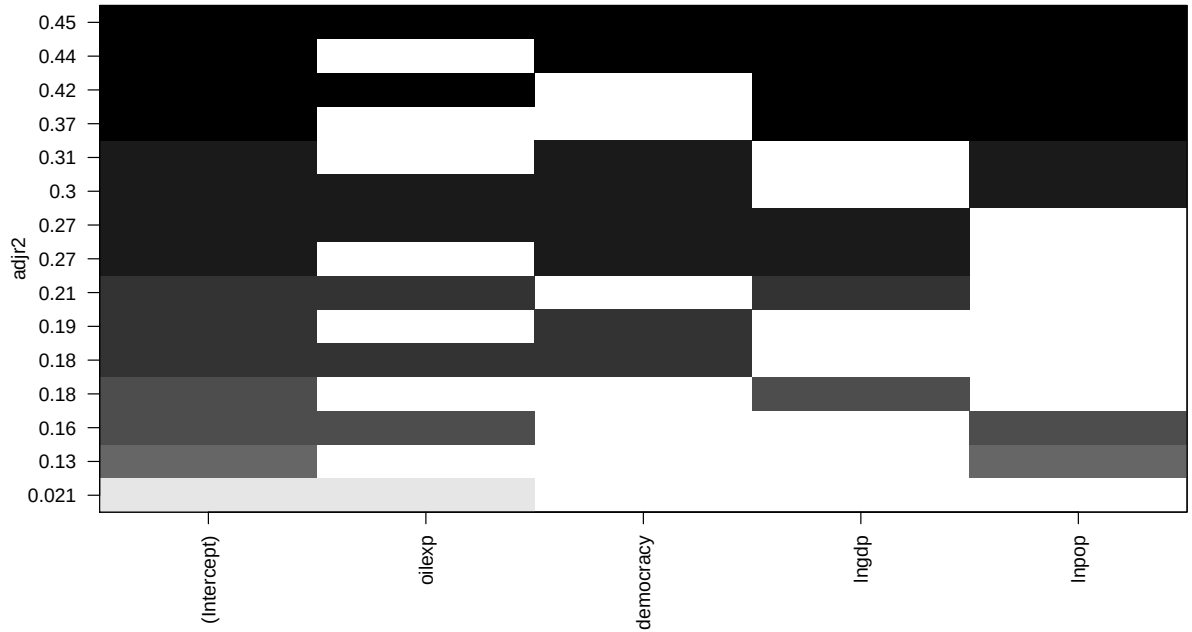
Mixed: allow both direction to find the best fitting model.

Note that this is only relevant for prediction, not causal inference! For the latter, we select variables on the basis of potential confounders.

All subsets regression

If the model is relatively quick to estimate—not too many variables and not a very large data set—then we can also just estimate models for all reasonable subsets, and select based on some fit criterion (R^2, AIC, etc.).

The *leaps* package in R provides a useful plot to select a model:



Ridge regression

Instead of adding or dropping variables, we can also simply put less weight on those variables that contribute less. In **ridge regression** we shrink β-estimates for variables which contribute less.

Ridge regression is similar to regular linear regression, but instead of only minimizing the sum of squared residuals, it also adds a penalty for high values of β.

$$\hat{\beta}^{OLS} = \arg\min_{\hat{\beta}} (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

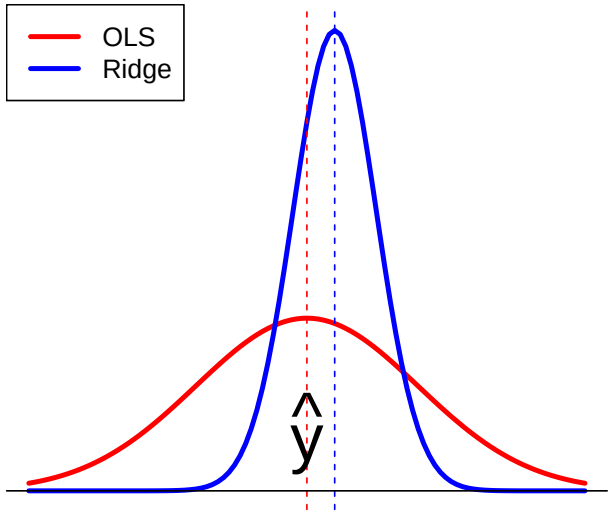
$$\begin{aligned}\hat{\beta}^{ridge} &= \arg\min_{\hat{\beta}} \left[(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\beta'\beta\right] \\ &= \left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

The penalty is set by λ. If $\lambda = 0$, $\hat{\beta}^{OLS} = \hat{\beta}^{ridge}$.

Ridge regression: benefits

Ridge regression is particularly useful with **high multicollinearity**, as highly multicollinear variables will be downweighted.

$\hat{\beta}^{ridge}$ a biased estimate of β, but typically has a lower prediction variance for $\hat{\mathbf{y}}$, such that the **mean squared error** might in fact be smaller.

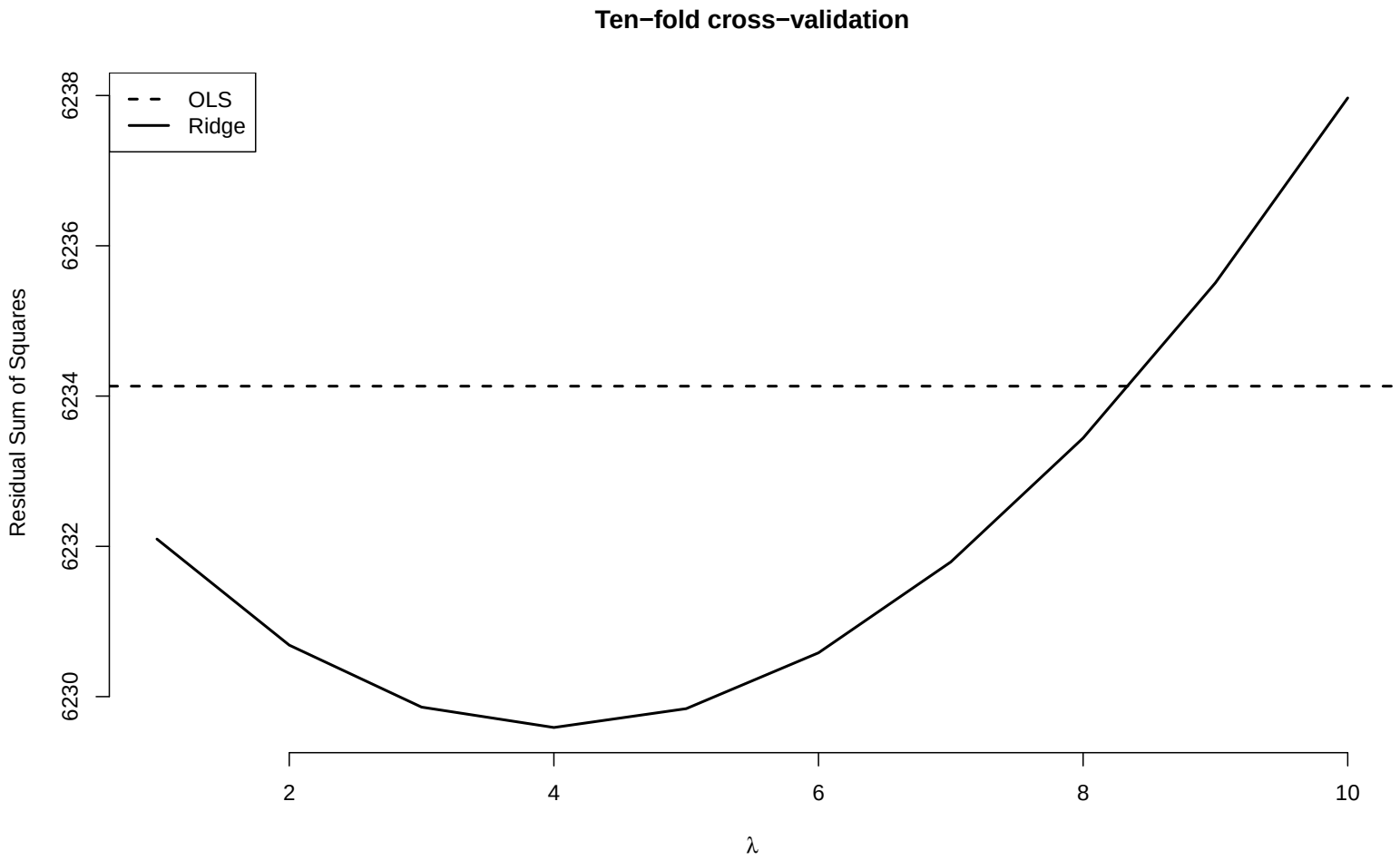


Ridge regression also protects against **overfitting**.

Other variations, such as the **lasso**, are also common in machine learning.

(Hastie, Tibshirani and Friedman, 2009, 61–68)

Ridge regression: selecting λ



Outline

Plotting the regression line

```
ggplot(ross[ross$Year == 2001,],
            aes(x = oilexp, y = strucint)) +
   geom_jitter() +
   geom_smooth(method = "lm", se = FALSE) +
   labs(x = "Oil exports", y = "IGO memberships",
      title = "Oil exports and IGOs (in 2001)")
```

Regression lines are added with the `geom_smooth` command where you need to add that `method` is equal to "lm", i.e., the linear model, instead of the standard smoothened curve that is plotted.

You can set `se` equal to `TRUE` to also plot the confidence intervals (standard errors).

Getting the regression coefficients

```
library(stargazer)

stargazer(lm(oilexp ~ strucint, ross,
   subset = Year == 2001), type = "html")
```

If you combine this with the

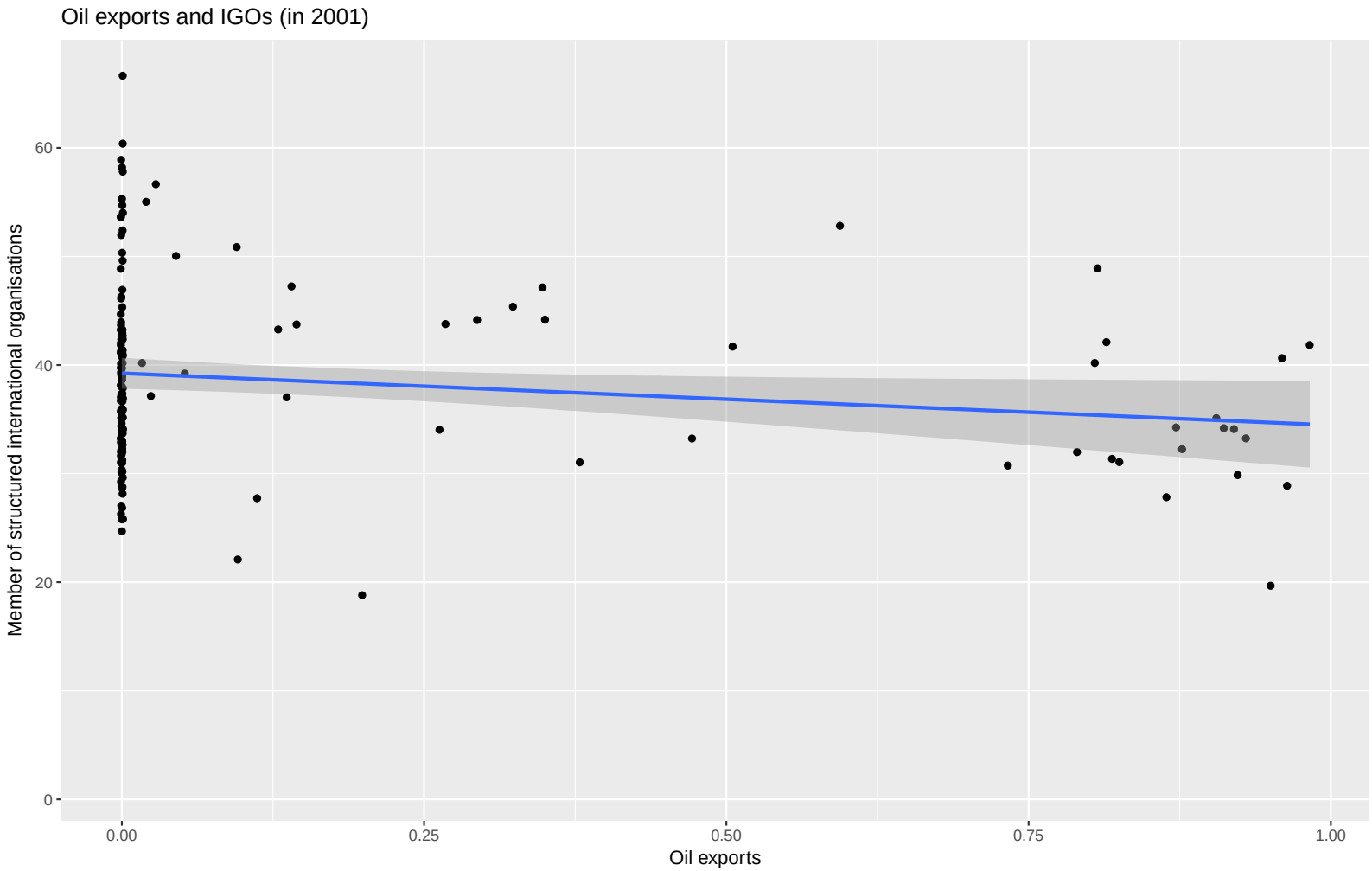
```
```{r results = "asis"}
```

option in the RMarkdown chunk, you get a nice looking regression table in your output file.
```

	Dependent variable:
	strucint
oilexp	-4.780^{**}
	(2.257)
Constant	39.240^{***}
	(0.720)
Observations	161
R^2	0.027
Adjusted R^2	0.021
Residual Std. Error	8.282 (df = 159)
F Statistic	4.485^{**} (df = 1; 159)
Note:	*p<0.1; **p<0.05; ***p<0.01

$$strucint_i = \hat{\beta}_1 + \hat{\beta}_2 oilexp_i = 39.240 - 4.780 \cdot oilexp_i$$

Linear model



$$strucint_i = \hat{\beta}_1 + \hat{\beta}_2 oilexp_i = 39.240 - 4.780 \cdot oilexp_i$$

Multiple regression

```
stargazer(lm(strucint ~ oilexp + democracy
    + lngdp + lnpop, ross,
    subset = Year == 2001), type = "html")
```

$$strucint_i = \beta_1 + \beta_2 oilexp_i + \beta_3 democracy_i + \beta_4 lngdp_i + \beta_5 lnpop_i + \varepsilon_i$$

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition ed. New York: Springer.

King, Gary. 1998. *Unifying political methodology. The likelihood theory of statistical inference*. University of Michigan Press.