

Module 4

Regression Models

Shmueli et al., Chapter 6

Explain vs. predict

Fit & assess

Select

Regularize

Roumen Vesselinov, PhD **Roumen@jhu.edu**

Explore the data · build the model · ask the AI · verify every result

Outcomes and predictors

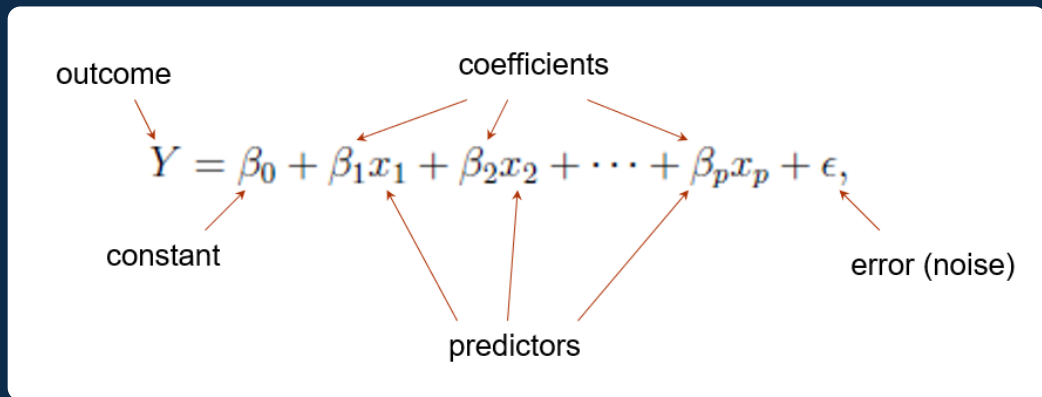
Var	Description	Var	Description
Y1	GDP per capita, constant 2015 US\$	Y2	Armed conflict (yes/no)
X1	Annual GDP growth rate (%)	X11	Total population
X2	Inflation rate (CPI, annual %)	X12	Population aged 65+ (% of total)
X3	Trade (% of GDP)	X13	Urban population (% of total)
X4	FDI net inflows (% of GDP)	X14	Top 10% income share (inequality)
X5	Fraser Economic Freedom Index	X15	Government effectiveness (WGI)
X6	Oil rents (% of GDP): resource channel	X16	Rule of law (WGI)
X7	Human Development Index	X17	Political stability & absence of violence (WGI)
X8	Life expectancy at birth (years)	X18	V-Dem electoral democracy index
X9	Internet users (% of population)	X19	Historical region (10 categories)
X10	Government expenditure on education (% of GDP)	X20	Year (numeric, for time-trend modeling)

X19: historical region

Value	Region	Value	Region
1	Eastern Europe & post-Soviet Union	6	East Asia
2	Latin America	7	South-East Asia
3	North Africa & Middle East	8	South Asia
4	Sub-Saharan Africa	9	The Pacific
5	Western Europe & North America (incl. Australia, NZ)	10	The Caribbean

X19 is categorical: in models it becomes dummy variables.

Regression model



Outcome, constant, coefficients, predictors, error (noise)

- The outcome (dependent variable) is continuous.
- The predictors are continuous/rank, or dummy variables.
- We assume a linear relationship between predictors and outcome.

Topics

1 Explain vs. predict

Explanatory vs. predictive modeling with regression.

2 Fit

Fitting a predictive model.

3 Assess

Assessing predictive accuracy.

4 Select

Selecting a subset of predictors.

5 Multicollinearity

Predictors cannot be highly correlated among themselves.

6 Regularize

Lasso, ridge, and elastic net (new).

Explanatory modeling

- Goal: explain the relationship between predictors (explanatory variables) and the target.
- The familiar use of regression in data analysis.
- Model goal: fit the data well and understand the contribution of explanatory variables to the model.
- "Goodness of fit": R^2 , residual analysis, p-values.

Typical question

Is government effectiveness associated with GDP per capita, holding other factors fixed?

Predictive modeling

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}}$$

RMSE

$$SD = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}}$$

compare to SD

- Goal: predict target values in other data where we have predictor values but not target values.
- Model goal: optimize predictive accuracy.
- Train the model on training data; assess performance on validation (hold-out) data.
- Explaining the role of predictors is not the primary purpose (but useful).
- Measure: root mean squared error (RMSE), compared to the SD.

Explanatory vs. predictive: side by side

	Explanatory	Predictive
Goal	Understand relationships	Predict new records
Data used	Full sample	Training, then validation/test
Judged by	R^2 , p-values, coefficient signs, theory	RMSE (and MAE) on validation data
Variables chosen by	Theory, confounding	Predictive usefulness
Main danger	Confounding, misspecification	Overfitting, leakage

A model can explain well and predict poorly, or predict well without a causal interpretation.

Least squares and interpretation

$$\text{minimize } \Sigma(y_i - \hat{y}_i)^2 \quad \beta = (X^T X)^{-1} X^T y$$

Numeric predictor

β_j = expected change in Y for a one-unit increase in X_j , other predictors held constant.

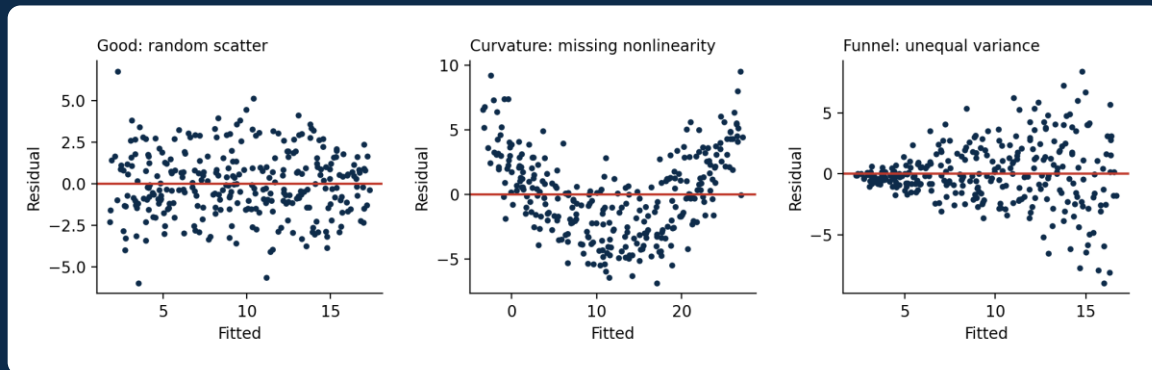
Dummy variable

Difference from the reference category, other predictors equal (e.g., region vs. the reference region).

Caution

With correlated predictors, a coefficient can differ from, even have the opposite sign to, the simple correlation.

Assumptions and residual diagnostics



Residuals vs. fitted values (simulated)

- LINE: Linearity, Independence, Normality, Equal variance.
- Curvature → missing nonlinear term or interaction.
- Funnel → unequal variance (common with skewed GDP per capita).
- Repeated country-years violate independence: cluster by country.

Correlations and multicollinearity

Correlation with Y1

X15	0.82
X16	0.81
X19_5	0.77
X7	0.70
X9	0.67
X8	0.66
X17	0.62
X5	0.62
X13	0.60
X12	0.56
X18	0.48
X14	-0.47
X3	0.37

Correlation with Y1

```

=== Highly Correlated Predictor Pairs (|r| >= 0.9 )
Var1 Var2 Correlation CorY1_Var1 CorY1_Var2 Drop
X8   X7   0.926      0.659      0.705   X8
X16  X15   0.948      0.809      0.817  X16

=== Variables to be EXCLUDED ===
[1] "X8" "X16"

```

Highly correlated predictor pairs ($|r| \geq 0.9$)

- X15 government effectiveness correlates 0.82 with Y1.
- X8 & X7: $r = 0.926$; X16 & X15: $r = 0.948$.
- Variables excluded: X8 and X16.

Why highly correlated predictors are a problem

$$\text{VIF}_j = 1 / (1 - R^2_j)$$

What happens

Coefficients stay unbiased but standard errors inflate; signs can flip when a predictor is added.

Example

$r = 0.95$ between two predictors $\rightarrow$
 $\text{VIF} \approx 10 \rightarrow$ SEs about 3× larger.

What to do

Drop one of a pair, combine into an index, or use regularization. Check with `car::vif()`; $\text{VIF} > 5\text{--}10$ is a warning.

Fitting a regression with all variables

```

Coefficients:
(Intercept)  2.763e+05  1.288e+05  2.144  0.03220 *
X1           9.181e+01  5.535e+01  1.659  0.09746 .
X2          -4.304e+01  3.964e+01 -1.086  0.27775
X3           6.432e+01  5.828e+00  11.037  < 2e-16 ***
X4          -2.769e+01  9.890e+00 -2.800  0.00519 **
X5          -6.774e+02  4.920e+02 -1.377  0.16881
X6           1.991e+02  3.984e+01  4.997  6.64e-07 ***
X7          -2.744e+02  5.386e+03 -0.051  0.95938
X9           1.133e+02  2.039e+01  5.555  3.38e-08 ***
X10          -1.006e+03  1.722e+02 -5.842  6.57e-09 ***
X11           1.298e-07  2.374e-06  0.055  0.95642
X12          -3.026e+02  9.339e+01 -3.240  0.00123 **
X13           3.156e+01  2.115e+01  1.492  0.13600
X14          -7.011e+03  4.337e+03 -1.617  0.10620
X15           7.986e+03  7.300e+02  10.939  < 2e-16 ***
X17          -1.644e+02  4.357e+02 -0.377  0.70610
X18          -2.541e+02  1.792e+03 -0.142  0.88727
X20          -1.317e+02  6.429e+01 -2.048  0.04075 *
X19_1        -1.501e+03  1.579e+03 -0.950  0.34206
X19_2         1.618e+03  1.157e+03  1.398  0.16250
X19_3         3.132e+03  1.367e+03  2.291  0.02214 *
X19_5         2.403e+04  1.697e+03  14.159  < 2e-16 ***
X19_6         8.267e+03  2.285e+03  3.617  0.00031 ***
X19_7        -5.207e+03  1.273e+03 -4.091  4.58e-05 ***
X19_8         1.480e+02  1.443e+03  0.103  0.91833
X19_10        -4.768e+03  4.650e+03 -1.025  0.30544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

lm() output on training data

- Y1 is the outcome; 25 explanatory variables.
- X8 and X16 excluded (multicollinearity).
- Region enters as dummy variables.
- Many significant p-values with thousands of rows: significance $\neq$ practical importance.

Model 1 accuracy on validation data

```
> accuracy(regress1Pred, valid1$Y1)
```

	ME	RMSE	MAE
Test set	366.8278	8474.952	5397.49

RMSE on validation data

- Accuracy based on validation data.
- $RMSE < SD$: \$8,474 < \$17,973.
- Good predictive ability.

```
c(n = length(valid1$Y1),
  mean = mean(valid1$Y1), sd = sd(valid1$Y1),
  min = min(valid1$Y1), max = max(valid1$Y1))
```

n	mean	sd	min	max
1113.0000	13215.3608	17973.9041	321.5478	107591.9659

SD of Y1 on validation data

Accuracy measures on validation data

Measure	Formula	Notes
RMSE	$\sqrt{(\sum e_i^2 / n)}$	Same units as Y; penalizes large errors; main measure
MAE	$\sum e_i / n$	Typical error; robust to large errors
MAPE	$100 \cdot \sum e_i / y_i / n$	Percentage; fails when Y is near 0
RMSE / SD	RMSE $\div$ SD of Y (validation)	< 1: better than predicting the mean
Out-of-sample R^2	$\approx 1 - (\text{RMSE}/\text{SD})^2$	Model 1: $1 - 0.47^2 \approx 0.78$

Model 1: $\text{RMSE}/\text{SD} = 8,474 / 17,973 = 0.47$.

Selecting subsets of predictors

- Goal: find a parsimonious model: the simplest model that performs sufficiently well.
- More robust; higher predictive accuracy.
- We assess predictive accuracy on validation data.
- Stepwise selection algorithms: forward, backward, stepwise.

Why fewer predictors?

Each extra predictor adds variance, cost, and complexity.

Forward, backward, and stepwise selection

F

Forward

Start with no predictors; add the one that improves the criterion most; stop when no addition helps.

B

Backward

Start with all predictors; remove the least useful one at a time.

S

Stepwise

Add and remove at each step (R: `step()`, using AIC).

!

Caveats

Unstable choices; optimistic p-values after selection (Harrell, 2015). Always validate.

Criteria: AIC, BIC, adjusted R^2 ; the final judgment uses validation RMSE.

Stepwise selection regression

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	297615.068	121192.027	2.456	0.014194	*
X1	100.230	54.733	1.831	0.067300	.
X3	61.580	5.406	11.390	< 2e-16	***
X4	-27.891	9.835	-2.836	0.004642	**
X6	210.678	35.411	5.950	3.48e-09	***
X9	118.871	19.387	6.132	1.16e-09	***
X10	-973.977	155.933	-6.246	5.74e-10	***
X12	-289.624	83.146	-3.483	0.000512	***
X13	39.072	18.005	2.170	0.030190	*
X14	-6599.879	3925.624	-1.681	0.092966	.
X15	7227.647	526.070	13.739	< 2e-16	***
X20	-144.952	60.255	-2.406	0.016288	*
X19_1	-2445.320	1115.133	-2.193	0.028501	*
X19_3	2602.629	984.342	2.644	0.008294	**
X19_5	23182.794	1329.100	17.442	< 2e-16	***
X19_6	7046.991	2062.848	3.416	0.000655	***
X19_7	-5255.310	987.189	-5.324	1.20e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Selected model: step()

- Y1 is the outcome; 16 explanatory variables.
- Chosen by stepwise selection using AIC on the training data.

Model 2 accuracy on validation data

```
> accuracy(Step1.pred, valid1$Y1)
```

	ME	RMSE	MAE
Test set	355.4905	8494.745	5402.271

RMSE on validation data

```
c(n = length(valid1$Y1),
  mean = mean(valid1$Y1), sd = sd(valid1$Y1),
  min = min(valid1$Y1), max = max(valid1$Y1))
```

n	mean	sd	min	max
1113.0000	13215.3608	17973.9041	321.5478	107591.9659

SD of Y1 on validation data

- Accuracy based on validation data.
- $RMSE < SD$: $\$8,494 < \$17,973$: good predictive ability.
- Model 1 $RMSE < Model\ 2\ RMSE$ ($\$8,474 < \$8,494$), but Model 2 is more parsimonious.

Same accuracy, fewer predictors

\$8,474

Model 1: 25 predictors

\$8,494

Model 2: 16 predictors

0.2%

Difference in RMSE: negligible

- Prefer Model 2: nearly identical accuracy with nine fewer predictors.
- One validation split is noisy; cross-validation gives a more reliable comparison.

Regularization

- Regularization penalizes large coefficients in a regression model that may lead to overfitting the data.
- Lasso regression shrinks some coefficients completely to zero.
- It minimizes the complexity of the model (number of predictors) by limiting the sum of the absolute values of the coefficients (L1 norm).
- Other regularization models: ridge regression and elastic net regression.

Idea

Accept a little bias in the coefficients to gain a large reduction in variance.

Ridge, lasso, and elastic net

$$\text{minimize } \sum (y_i - \hat{y}_i)^2 + \lambda [\alpha \sum |\beta_j| + (1 - \alpha) \sum \beta_j^2 / 2]$$

Ridge ($\alpha = 0$)

L2 penalty: shrinks all coefficients toward 0; keeps all; good with correlated predictors (Hoerl & Kennard, 1970).

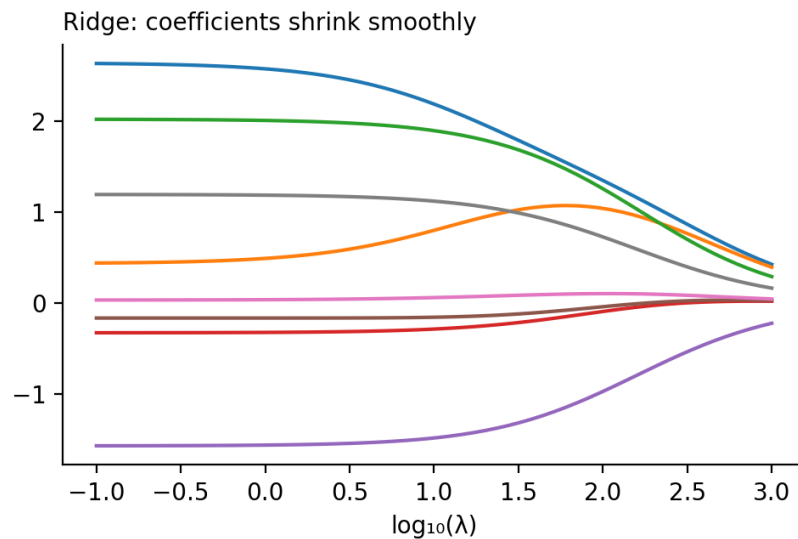
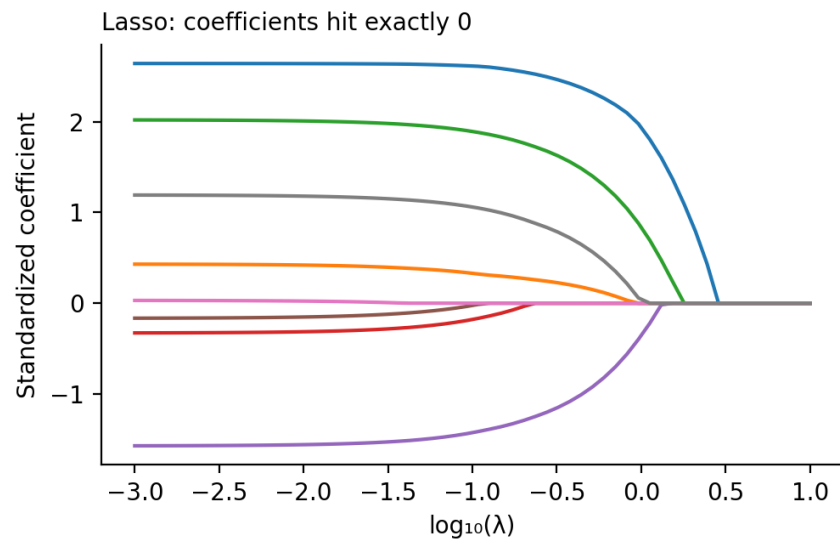
Lasso ($\alpha = 1$)

L1 penalty: sets some coefficients exactly to 0: variable selection (Tibshirani, 1996).

Elastic net

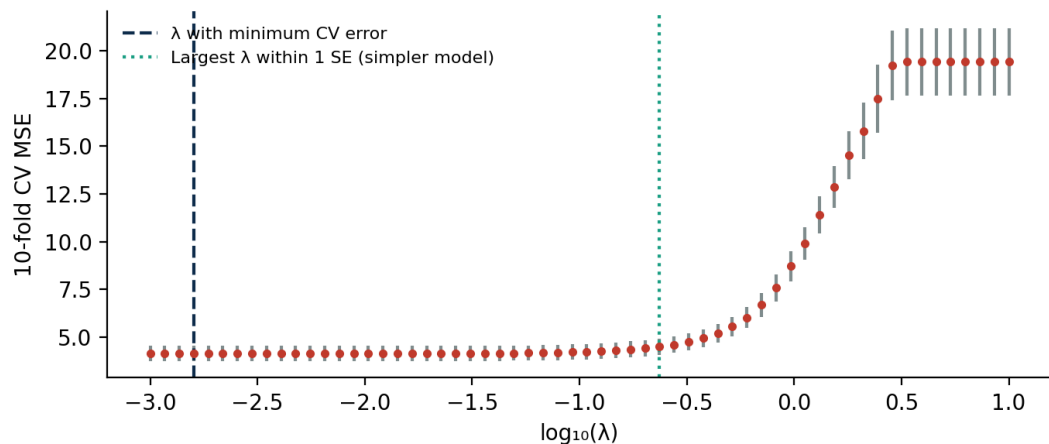
Mix: selects variables while keeping correlated groups together (Zou & Hastie, 2005).

Coefficient paths as the penalty grows



Simulated data, 8 predictors

Choosing λ by cross-validation



10-fold CV error for the lasso (simulated)

- Fit the model for a grid of λ values.
- Choose the λ with minimum CV error, or the largest λ within one SE of it (simpler model).
- Standardize predictors first (glmnet does this by default).

The regression workflow in R

Model 1

```
m1 <- lm(Y1 ~ . - X8 - X16, data = train); summary(m1)
```

Accuracy

```
forecast::accuracy(predict(m1, valid), valid$Y1); sd(valid$Y1)
```

Diagnostics

```
plot(m1); car::vif(m1)
```

Model 2

```
m2 <- step(m1, direction = "both", trace = 0)
```

Lasso

```
x <- model.matrix(Y1 ~ ., train)[, -1]; cv <- glmnet::cv.glmnet(x, train$Y1, alpha = 1)
```

Coefficients

```
coef(cv, s = "lambda.1se"); predict(cv, newx = xvalid, s = "lambda.1se")
```

Intervals and nonlinear terms

Two kinds of intervals

- Confidence interval: uncertainty about the mean Y at given X .
- Prediction interval: for one new record; always wider.
- For predicting countries, use prediction intervals: `predict(m, newdata, interval = "prediction")`.

Beyond straight lines

- Polynomials (X , X^2) and splines for curvature.
- Interactions ($X_1 \times X_2$) when effects depend on each other.
- Trees (Modules 9–10) find these automatically: RMSE $\approx$ \$3,750 for $Y1$.

Summary

- 1 Linear regression models are popular tools, not only for explanatory modeling, but also for prediction.
- 2 A good predictive model has high predictive accuracy (to a useful practical level).
- 3 Predictive models are fit to training data; accuracy is evaluated on a separate validation data set.
- 4 Removing redundant predictors is key to predictive accuracy and robustness.
- 5 Subset selection methods (and regularization) help find good candidate models, which should then be run and assessed.

MODULE 4

Regression Models

Roumen Vesselinov · JHBigData.com