Biostatistics & Epidemiological Data Analysis using R

9

Linear regression II

Stefan Konigorski

Health Intervention Analytics Group, HPI

December 14, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

Overview

Learning objectives of today

- Review of simple and multiple linear regression.
- Comparison of factor & numeric variables as predictors in linear regression.
- Assumptions of linear regression.
- Overview on approaches for evaluating and selecting linear regression models, and how to do this in R.
- (Example how to do linear regression with multiple imputation).

Simple linear regression

Review: simple linear regression

- Aim: Predict a variable Y by a variable X under the assumption of a linear relationship.
- Model equation: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$.
- β_1 tells you about the association of X and Y.

Factors/dummy variables as predictors in `lm`

- Dummy variable = binary 0-1 variable.
- In R, the `lm` function automatically creates $k - 1$ dummy variables for a factor variable with k groups. They capture the differences between each group in comparison to the reference group.

Factors/dummy variables as predictors in `lm`

Example:

- Factor variable age with 3 categories 1, 2, 3.
- R sets the first level as reference, creates variables age2, age3:

age	age2	age3
1	0	0
1	0	0
2	1	0
2	1	0
3	0	1
3	0	1

- Regression with age as ordinal variable (`as.numeric(age)`):

$$Y_i = \beta_0 + \beta_1 age_i + \varepsilon_i$$

- Regression with age as factor variable:

$$Y_i = \beta_0 + \beta_1 age2_i + \beta_2 age3_i + \varepsilon_i$$

Multiple linear regression

Review of multiple linear regression

- Aim: Predict a variable Y by multiple variables $X_1, \ldots, X_k$ under the assumption of a linear relationship.
- Model equation:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- In matrix formulation: $Y = X^T \beta + \varepsilon$
- Least squares estimate of the regression coefficient vector β:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- Standard error of $\widehat{\beta}$:

$$SE(\widehat{\beta}) = \sqrt{\sigma^2 (X^T X)^{-1}}$$

Interpretation of the regression coefficients

- The regression coefficient β_j describes the effect of X_j on Y (increase of Y by β_j when X_j is increased by 1 unit), while holding all other X variables in the model constant.
- I.e. the effect of the other X variables in the model has been removed (=adjusted, corrected) from the estimated effect of X_j on Y i.e. is "considered" in $\hat{\beta}_j$.

Hypothesis tests

- Hypothesis tests and confidence intervals of the regression coefficients, and tests of the full model ($H_0 : \beta_1 = \cdots = \beta_k = 0$ vs. H_1 : at least β_j is $\neq 0$) are analogous to simple linear regression.

Linear regression - Model assumptions

Overview

- So far we have neglected the assumptions that are made in linear regression

- What are they?
- How can we check them?
- How bad is it if they are violated?

Assumptions of linear regression

- Y is continuous
- The relationships between Y and all X_j are linear
- All relevant variables (covariates, confounders) are in the model
- All observations are independent
- There is no multicollinearity ($\approx$ not a "super strong" correlation between the X_j)
- Homoscedasticity (equal variance) of the residuals
- Normal distribution of the residuals

How do you check and deal with these assumptions?

Assumption: Y is continuous

Check

- yes/no (or approx. yes/no?)

What if assumption is not satisfied?

- Other regression (multinomial, ordinal?)

Assumption: Linearity

Check

- Visually in scatter plot
- in R: `plot(X, Y)`

What if assumption is not satisfied?

- Add quadratic/polynomial terms or splines
- Adding a quadratic term in R: `lm(Y ~ X + X^2)`
- Transform predictor to factor?

Assumption: All relevant variables are in the model

Check

- Think, draw directed acyclic graphs, look at literature
- Compare the estimates of the regression coefficients in different models

What if assumption is not satisfied?

- Results can be strongly biased, in any direction
- Throw results into garbage and use other statistical model, or consider in interpretation

Assumption: Independent observations

Check

- Theoretical: Is there a structure in the data (time, hierarchy/cluster)?
- Compute ICC (intraclass correlation coefficient), Durbin-Watson statistic of autocorrelation

What if assumption is not satisfied?

- Use other statistical model (linear mixed models, time series)

Assumption: No multicollinearity

Definition

- 2 or more variables are collinear, if one variable can be written as a linear combination of the other variables.
- Multicollinearity here: $\approx$ no "super strong" correlation between the predictors X_j

Check

- Compute the correlation (and maybe VIF, variance inflation factor) between predictors
- in R: `cor(Xmatrix)`
- Check the predictive power of the model (R^2) eg using cross-validation

Assumption: No multicollinearity

What if assumption is not satisfied?

- If predictors are highly correlated (eg $r = 0.99$), then the estimates of the regression coefficients are still unbiased, and the standard error estimates of the regression coefficient estimates as well as the respective hypothesis tests are still valid. But: the standard error of the regression coefficient estimates is larger ($\longrightarrow$ smaller power) and R^2 is larger (SS_{Resid} and MSE are not affected, $Var(Y)$ and SS_{Regr} larger).
- See `R_8b_assumpt_multicoll.R`
- Consider in interpretation (with respect to validity of estimates/tests, effect of X_1 or X_2?), remove effect of one variable, factor analysis/principal component analysis.

Assumption: Homoscedasticity of residuals

Check

- Graphically: scatter plot of residuals vs. predicted Y values, e.g. with `plot(lm(...))` function
- Extract residuals and predicted Y values in R: `residuals(lm(...))`, `predict(lm(...))`

What if assumption is not satisfied?

- With heteroscedastic residuals, the estimates of the regression coefficients are still unbiased, but the standard errors are underestimated, i.e. the p-values are too small (inflation of type I error)
- See `R_8b_assumpt_homosced.R`
- Solutions: Consider in regression (eg weighted least-squares regression), use robust standard error estimates

Assumption: Normal distribution of residuals

Check

- Graphically: histograms and Q-Q-plots of the residuals
- For example, using the `plot(lm(...))` function

What if assumption is not satisfied?

- Approximatively normally-distributed residuals are sufficient so that estimates, standard error estimates, and hypothesis tests of the regression coefficients are still valid - but the power decreases.
- Only for extremely non-normal residuals, estimates are biased.
- See `R_8b_assumpt_normality.R`
- Solutions: transform Y (log, Box-Cox), in R with `log()` and eg `MASS::boxcox()`

Further regression diagnostics

- The `plot(lm(...))` function yields further diagnostic plots, e.g. of outliers and leverage points.

Overview

Possible practical questions in linear regression:

- Is variable X (or are variables X_j) associated with Y?
- Which variables should you select to predict Y well?
- How good can you predict Y?

$\longrightarrow$ model selection and model evaluation

Linear regression - Model selection

Model selection - Overview

- Question: Which predictors should be included in the regression model to predict Y?
- Alternative question: which variables are important/ relevant/ associated?
- Which model is best suited for predicting Y?

Approach?

(First: think whether linear regression is appropriate)

Model selection - Approaches

Question: Which predictors should be included in the regression model to predict Y?

Approaches

1. Determine all relevant variables X theoretically (expert knowledge, literature review), include them in model, compute model, done.
2. Determine all relevant variables X theoretically and using DAGs (directed acyclic graphs), only include "relevant" variables into the model, compute model, done.
3. Use stepwise statistical approaches (data-driven) to determine all relevant variables from all available variables and include them in final model (stepwise forward/ backward).
4. Use regularized regression model with all predictors, let the model filter out all non-relevant variables (data-driven).

Model selection - Stepwise variable selection

Forward selection

- Start with null model (only intercept), add stepwise those variables which increase the model fit maximally (i.e. with respect to AIC).

Backward selection

- Start with full model (i.e. all predictors), delete variables stepwise so that the model fit is "maximally not decreased".

In R

- E.g. with `step(lm(...), direction = "forward")` or with `MASS:stepAIC()`
- E.g. with `step(lm(...), direction = "backward")`
- See `R_9a_example_model_selection_stepwise.R`

Model selection - lasso

lasso (least absolute shrinkage and selection, Tibshirani 1996)

- Estimate regression coefficients in linear regression model by minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji})^2$$

  with the constraint $\sum_{j=1}^{k}|\beta_j| \leq t$ for a constant t.

- This is equivalent to minimizing

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ji})^2 + \lambda\sum_{j=1}^{k}|\beta_j|$$

Model selection - lasso

lasso (least absolute shrinkage and selection, Tibshirani 1996)

- Implementation in R: eg in package `glmnet`.
- Documentation: `https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet_beta.pdf`
- Extension: elastic net (Zou and Hastie, 2005).
- See `R_9a_example_model_selection_lasso.Rmd`

Hypothesis tests in selected models

NOTE

- If you want to do statistical association tests in a (regression) model (e.g. to test if X is associated with Y), which has been selected from different possible models, this model selection has to be considered in the testing!

Linear regression - Model evaluation

Model evaluation of one model

Question: How good can you predict Y based on the X_j?

Measures of model fit

- F value or p-value of F test of entire regression model
- Explained variance R^2, adjusted R^2: `summary(lm(...))$r.squared`, `...$adj.r.squared`
- Deviance of a model $= -2$ log-likelihood of a model (χ^2 distribution): `deviance(lm(...))`
- MSE (Mean Squared Error) $= E((Y - \widehat{Y})^2)$
- AIC (Akaike Information Criterion) $= 2k - 2 \cdot$log-likelihood, here: $k =$ number of predictors in the model.
- BIC (Bayesian Information Criterion) $=$ $log(n) \cdot k - 2 \cdot$log-likelihood, where $n =$ sample size
- `AIC(lm(...))` and `BIC(lm(...))` in R

Model evaluation of two models

Question: Is model 1 better than model 2 to predict Y?

Compare the model fit of 2 models

- Deviance between 2 models = difference of the deviance of the models (χ^2 distribution) $\longrightarrow$ Likelihood ratio test
- In R: `anova(lm(model1_nested_in_model2), lm(model2))`
- Also possible: comparison of AIC or BIC (smaller = better model fit!)

Model evaluation of one model - revisited

Question: How well can you predict Y based on the X_j?

Problem

All model fit measures on slide 32 don't consider the model selection[1] i.e. generally report a model fit that is overestimating the actual model fit when it is applied to a new dataset.

[1] and other potential challenges such as that the sample is not random and representative

Model evaluation - cross-validation

One approach to obtain a better estimate of model fit:

Cross-validation

- Idea: Building and evaluating your model should be done using different datasets.
- Approach of k-fold cross-validation:
  - Split sample into k parts.
  - Use all parts except of one for building your model (training set), and use the remaining part for model evaluation (validation set).
  - Do this for every subset, and then aggregate.
  - Optimal: After that apply on a test dataset to evaluate there.
- In R: eg `DAAG:cv.lm(data = dataset, lm(...), m = 5)`

Exercise

In the KiGGS dataset:

1. Get together in groups, and apply any of the four approaches on slide 26 to extract variables relevant for predicting BMI.
2. Note: `glmnet` cannot deal with missing values!
3. Primary focus: variable selection.
4. Possible secondary focus: model fit/prediction accuracy (but for this, you need cross-validation etc.)

Multiple imputation

See `R_9b_linear_regression_MI.Rmd` for an example.

Questions?

References

- Knight K (1999). Mathematical statistics. CRC Press.
- Friedman, Tibshirani & Hastie (2001). Elements of statistical learning. Springer. (https://web.stanford.edu/~hastie/Papers/ESLII.pdf)
- Harrell (2015). Regression modeling strategies. Springer.
- Tibshirani (1996). Regression shrinkage and selection via the lasso. J R Stat Soc Series B. (https://www.jstor.org/stable/2346178)
- Zou & Hastie (2005). Regularization and variable selection via the elastic net. J R Statist Soc B. (https://web.stanford.edu/~hastie/Papers/elasticnet.pdf)

Additional homework

Additional homework

See file `R_9_additional_homework.Rmd`.