

Multiple Linear Regression: confidence and colinearity

Author: Nicholas G Reich, Jeff Goldsmith

*This material is part of the **statsTeachR** project*

*Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported
License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US*

Today's topics

- estimation and our confidence in our model
- collinearity and non-identifiability
- categorical predictors

Example: predicting respiratory disease severity (“lung” dataset)

Multiple linear regression model

- Observe data (y, x_1, \dots, x_p) . Want to estimate $\beta_0, \beta_1, \dots, \beta_p$ in the model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

Assumptions

- Residuals have mean zero, constant variance, are independent.
- Model is true.

Least squares

As in simple linear regression, we want to find the β that minimizes the residual sum of squares.

$$RSS(\beta) = \sum_i \epsilon_i^2 = \sum_i (\hat{y}_i - y)^2$$

Lung Data Example

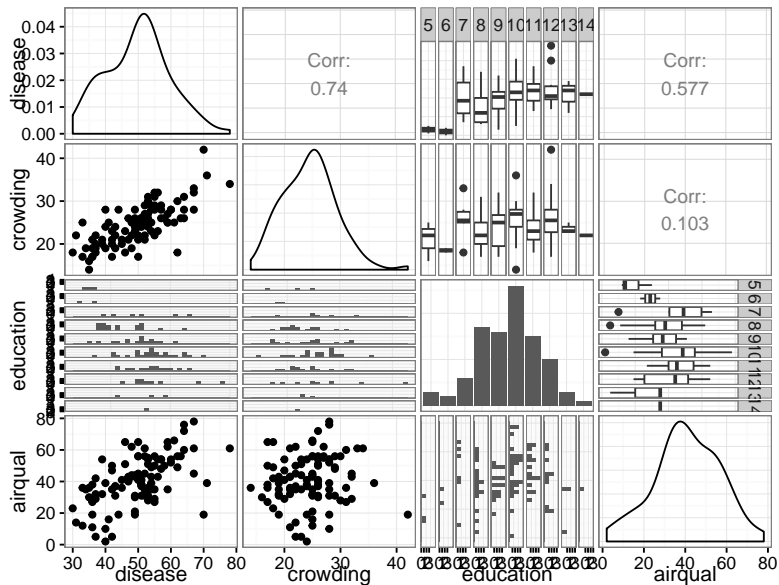
99 observations on patients who have sought treatment for the relief of respiratory disease symptoms.

```
dat <- read.table("lungc.txt", header=TRUE)
dat$education <- factor(dat$education)
```

The variables are:

- disease measure of disease severity (larger values indicates more serious condition).
- education highest grade completed
- crowding measure of crowding of living quarters (larger values indicate more crowding)
- airqual measure of air quality at place of residence (larger number indicates poorer quality)
- nutrition nutritional status (larger number indicates better nutrition)
- smoking smoking status (1 if smoker, 0 if non-smoker)

```
library(GGally)
ggpairs(dat[c("disease", "crowding", "education", "airqual")])
```



Lung Data Example

```
mlr1 <- lm(disease ~ crowding + education + airqual, data=dat)
summary(mlr1)$coef
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	-0.3586956	2.76552352	-0.1297026	8.971011e-01
##	crowding	1.3262116	0.08234084	16.1063641	7.740071e-28
##	education6	-0.7336823	3.35015465	-0.2189995	8.271634e-01
##	education7	2.9068360	2.70190821	1.0758456	2.849708e-01
##	education8	3.1398986	2.34214900	1.3406058	1.835386e-01
##	education9	5.6692646	2.36109920	2.4011124	1.847505e-02
##	education10	5.7785688	2.36728284	2.4410133	1.667257e-02
##	education11	8.2823722	2.43116863	3.4067453	9.972669e-04
##	education12	8.1534760	2.51262476	3.2450034	1.667638e-03
##	education13	13.2612311	2.99374060	4.4296527	2.733931e-05
##	education14	12.8540674	4.23758361	3.0333484	3.187806e-03
##	airqual	0.2950850	0.02549204	11.5755737	2.660791e-19

Least squares estimates: identifiability issues

If two of your variables are identical, or simple transformations of one another, least squares won't work

- This means that there will be an infinite number of mathematically equivalent least squares solutions.
- In practice, true **non-identifiability** (there really are infinite solutions) is rare.
- Can happen if \mathbf{X} is not of full rank, i.e. the columns of \mathbf{X} are linearly dependent (for example, including weight in Kg and lb as predictors)
- Can happen if there are fewer data points than terms in \mathbf{X} : $n < p$ (having 100 predictors and only 50 observations)
- More common, and perhaps more dangerous, is **collinearity**.

Infinite solutions

Suppose I fit a model $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$.

- I have estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2$
- I put in a new variable $x_2 = x_1$
- My new model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
- Possible least squares estimates that are equivalent to my first model:
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \hat{\beta}_2 = 0$
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0, \hat{\beta}_2 = 2$
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1002, \hat{\beta}_2 = -1000$
 - ▶ ...

Non-identifiability example: lung data

```
mlr3 <- lm(disease ~ airqual, data=dat)
summary(mlr3)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	35.4444812	2.23127089	15.885333	9.706236e-29
## airqual	0.3537389	0.05085138	6.956329	4.105421e-10

```
dat$x2 <- dat$airqual/100
mlr4 <- lm(disease ~ airqual + x2, data=dat)
summary(mlr4)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	35.4444812	2.23127089	15.885333	9.706236e-29
## airqual	0.3537389	0.05085138	6.956329	4.105421e-10

Non-identifiability: causes and solutions

- Often due to data coding errors (variable duplication, scale changes)
- Pretty easy to detect and resolve
- Can be addressed using *penalties* (might come up much later)
- A bigger problem is near-unidentifiability (collinearity)

Diagnosing collinearity

- Arises when variables are highly correlated, but not exact duplicates
- Commonly arises in data (perfect correlation is usually there by mistake)
- Might exist between several variables, i.e. a linear combination of several variables exists in the data
- A variety of tools exist (correlation analyses, multiple R^2 , eigen decompositions)

Effects of collinearity

Suppose I fit a model $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$.

- I have estimates $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2$
- I put in a new variable $x_2 = x_1 + \text{error}$, where *error* is pretty small
- My new model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
- Possible least squares estimates that are nearly equivalent to my first model:
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 2, \hat{\beta}_2 = 0$
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0, \hat{\beta}_2 = 2$
 - ▶ $\hat{\beta}_0 = 1, \hat{\beta}_1 = 1002, \hat{\beta}_2 = -1000$
 - ▶ ...
- A unique solution exists, but it is hard to find

Effects of collinearity

- Collinearity results in a “flat” RSS
- Makes identifying a unique solution difficult
- Dramatically inflates the variance of LSEs

Collinearity example: lung data

```
dat$crowd2 <- dat$crowding + rnorm(nrow(dat), sd=.1)
mlr5 <- lm(disease ~ crowding + airqual, data=dat)
summary(mlr5)$coef
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	2.8841197	2.49069149	1.157959	2.497533e-01
## crowding	1.4027587	0.09341356	15.016650	6.154176e-27
## airqual	0.3104388	0.02808020	11.055436	8.202723e-19

```
mlr6 <- lm(disease ~ crowding + crowd2 + airqual, data=dat)
summary(mlr6)$coef
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	2.3834406	2.49147457	0.9566385	3.411779e-01
## crowding	-5.0029890	4.04507807	-1.2368090	2.192068e-01
## crowd2	6.4133265	4.04880062	1.5840065	1.165157e-01
## airqual	0.3176746	0.02823405	11.2514732	3.613440e-19

Using VIFs: lung data

Variance Inflation Factors can help diagnose variables that are highly related.

The VIF for the k^{th} predictor in your model is

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2 from the model with X_k as the response and all other X variables as the predictors.

```
car::vif(mlr5)

## crowding  airqual
## 1.010657  1.010657

car::vif(mlr6)

##      crowding      crowd2      airqual
## 1924.915129 1923.446881  1.037824
```


Some take away messages

- Collinearity can (and does) happen, so be careful
- Often contributes to the problem of variable selection, which we'll touch on later

Model selection

Why are you building a model in the first place?

Model selection: considerations

Things to keep in mind...

- **Why am I building a model?** Some common answers
 - ▶ Estimate an association
 - ▶ Test a particular hypothesis
 - ▶ Predict new values
- What predictors will I allow?
- What predictors are needed?

Different answers to these questions will yield different final models.

Model selection: realities

All models are wrong. Some are more useful than others.

- George Box

- If we are asking which is the “true” model, we will have a bad time
- In practice, issues with sample size, collinearity, and available predictors are real problems
- It is often possible to differentiate between better models and less-good models, though
- The key decisions in model selection almost always involve balancing model complexity with the potential for overfitting.

Basic ideas for model selection

For association studies, when your sample size is large

- Include key covariates of interest.
- Include covariates needed because they might be confounders.
- Include covariates that your colleagues/reviewers/collaborators will demand be included for face validity.
- Do NOT go on a fishing expedition for significant results!
- Do NOT use “stepwise selection” methods!
- Subject the selected model to model checking/diagnostics, possibly adjust model structure as needed.

Basic ideas for model selection

For association studies, when your sample size is small

- Same as above, but may need to be more frugal with how many predictors you include.
- Rule of thumb for multiple linear regression is to have at least 15 observations for each covariate you include in your model.

Today's big ideas

- least squares estimation
- dangers of collinearity and non-identifiability
- model selection

Lab

Analyze the NHANES dataset. Create a model with the outcome variable of cholesterol (chol) that estimates relationships with other variables in the dataset.

```
library(NHANES)  
data(NHANES)  
?NHANES
```