

# Regression: Interactions and dummy variables

Author: Nicholas G Reich

*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](http://creativecommons.org/licenses/by-sa/3.0/deed.en_US)*

# Outline

- Dummy variables for categorical covariates
- Modeling interactions
- Model selection

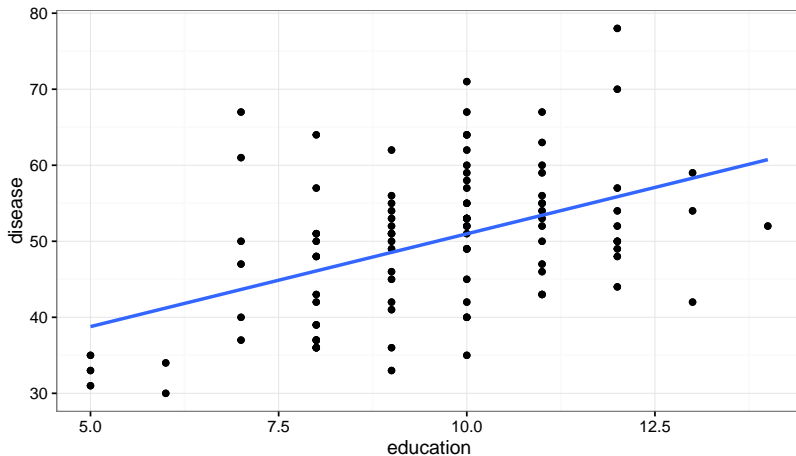
dummy variables

# Categorical predictors

- Assume  $X$  is a categorical / nominal / factor variable with  $k$  levels
- Can't use a single predictor with levels  $1, 2, \dots, K$  – this has the wrong interpretation
- Need to create *indicator* or *dummy* variables

# Categorical predictor example: lung data

```
qplot(education, disease, data=dat) + geom_point() +  
  geom_smooth(method="lm", se=FALSE)
```



# Indicator variables

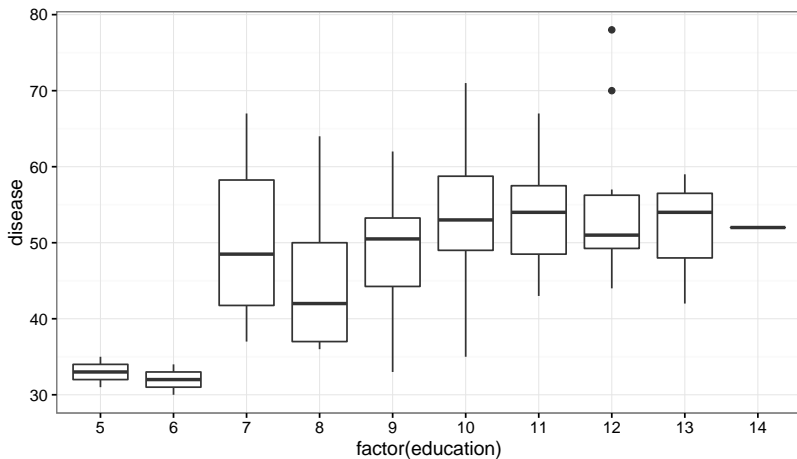
- Let  $x$  be a categorical variable with  $k$  levels (e.g. with  $k = 3$  “red”, “green”, “blue”).
- Choose one group as the baseline (e.g. “red”)
- Create  $(k - 1)$  binary terms to include in the model:

$$x_{1,i} =$$

$$x_{2,i} =$$

## Categorical predictor example: lung data

```
qplot(factor(education), disease, geom="boxplot", data=dat)
```



## Standard model interpretation

Using the model  $y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_{k-1} x_{k-1,i} + \epsilon_i$ , interpret

$$\beta_0 =$$

$$\beta_1 =$$



## Equivalent model

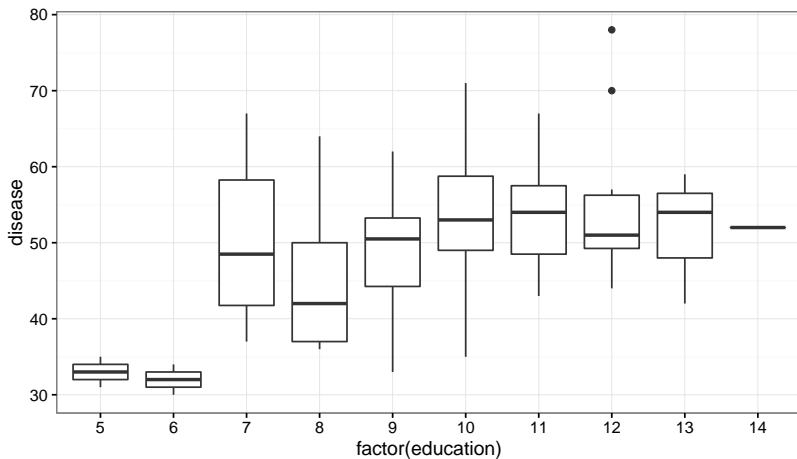
Define the model  $y_i = \beta_1 x_{i1} + \dots + \beta_k x_{i,k} + \epsilon_i$  where there are indicators for each possible group

$$\beta_1 =$$

$$\beta_2 =$$

## Categorical predictor example: lung data

```
qplot(factor(education), disease, geom="boxplot", data=dat)
```



## Categorical predictor example: lung data

$$dis_i = \beta_0 + \beta_1 educ_{6,i} + \beta_2 educ_{7,i} + \cdots + \beta_9 educ_{14,i}$$

```
mlr7 <- lm(disease ~ factor(education), data=dat)
summary(mlr7)$coef
```

##	Estimate	Std. Error	t value
## (Intercept)	33.00000	4.912705	6.7172765
## factor(education)6	-1.00000	7.767669	-0.1287387
## factor(education)7	17.33333	6.016811	2.8808175
## factor(education)8	11.17647	5.328577	2.0974588
## factor(education)9	15.50000	5.353496	2.8953040
## factor(education)10	20.38462	5.188395	3.9288865
## factor(education)11	20.53333	5.381599	3.8154707
## factor(education)12	22.20000	5.601346	3.9633332
## factor(education)13	18.66667	6.947614	2.6867735
## factor(education)14	19.00000	9.825411	1.9337614
##	Pr(> t )		
## (Intercept)	1.689481e-09		
## factor(education)6	8.978549e-01		
## factor(education)7	4.969406e-03		
## factor(education)8	3.878868e-02		

# Categorical predictor releveling

$$dis_i = \beta_0 + \beta_1 educ_{5,i} + \beta_2 educ_{6,i} + \beta_1 educ_{7,i} + \beta_2 educ_{9,i} + \dots + \beta_{14} educ_{14,i}$$

```
dat$educ_new <- relevel(factor(dat$education), ref="8")
mlr8 <- lm(disease ~ educ_new, data=dat)
summary(mlr8)$coef
```

##		Estimate	Std. Error	t value	Pr(> t )
##	(Intercept)	44.176471	2.063749	21.4059318	7.303151e-37
##	educ_new5	-11.176471	5.328577	-2.0974588	3.878868e-02
##	educ_new6	-12.176471	6.360902	-1.9142680	5.879890e-02
##	educ_new7	6.156863	4.040594	1.5237520	1.311162e-01
##	educ_new9	4.323529	2.963834	1.4587624	1.481508e-01
##	educ_new10	9.208145	2.654021	3.4695065	8.059293e-04
##	educ_new11	9.356863	3.014298	3.1041594	2.558604e-03
##	educ_new12	11.023529	3.391086	3.2507375	1.625933e-03
##	educ_new13	7.490196	5.328577	1.4056653	1.633049e-01
##	educ_new14	7.823529	8.755746	0.8935309	3.739828e-01

## Categorical predictor: no baseline group

$$dis_i = \beta_1 educ_{5,i} + \beta_2 educ_{6,i} + \cdots + \beta_{14} educ_{14,i}$$

```
mlr9 <- lm(disease ~ factor(education) - 1, data=dat)
summary(mlr9)$coef
```

##		Estimate	Std. Error	t value
##	factor(education)5	33.00000	4.912705	6.717277
##	factor(education)6	32.00000	6.016811	5.318432
##	factor(education)7	50.33333	3.473807	14.489386
##	factor(education)8	44.17647	2.063749	21.405932
##	factor(education)9	48.50000	2.127264	22.799241
##	factor(education)10	53.38462	1.668763	31.990531
##	factor(education)11	53.53333	2.197029	24.366243
##	factor(education)12	55.20000	2.690800	20.514349
##	factor(education)13	51.66667	4.912705	10.516948
##	factor(education)14	52.00000	8.509055	6.111137
##		Pr(> t )		
##	factor(education)5	1.689481e-09		
##	factor(education)6	7.715960e-07		
##	factor(education)7	3.845787e-25		
##	factor(education)8	7.303151e-37		

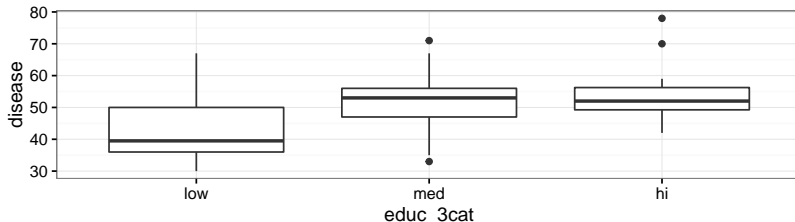
# Creating categories using cut()

$$dis_i = \beta_1 educ_{low,i} + \beta_2 educ_{med,i} + \cdots + \beta_{14} educ_{hi,i}$$

```
dat$educ_3cat <- cut(dat$education, breaks=3,  
                      labels=c("low", "med", "hi"))  
mlr10 <- lm(disease ~ educ_3cat - 1, data=dat)  
coef(mlr10)
```

```
## educ_3catlow educ_3catmed educ_3cathi  
##      43.42857      52.05263      54.21429
```

```
qplot(educ_3cat, disease, geom="boxplot", data=dat)
```

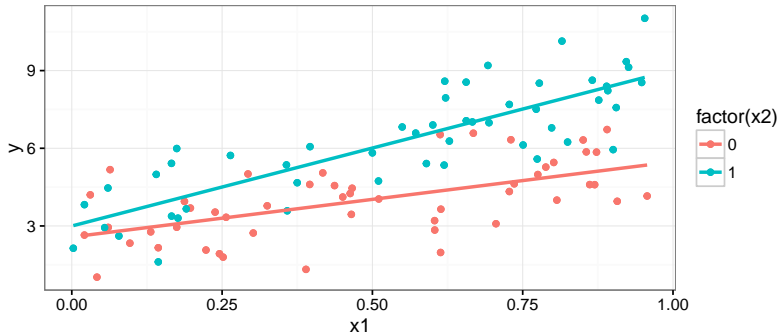


interaction

# What is interaction?

## Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable.





# Interaction vs. confounding

## Definition of interaction

Interaction occurs when the relationship between two variables depends on the value of a third variable. E.g. you could hypothesize that the true relationship between physical activity level and cancer risk may be different for men and women.

## Definition of confounding

Confounding occurs when the measurable association between two variables is distorted by the presence of another variable.

Confounding can lead to biased estimates of a true relationship between variables.

- It is important to include confounding variables (if possible!) when they may be biasing your results.
- Unmodeled interactions do not lead to “biased” estimates in the same way that confounding does, but it can lead to a richer and more detailed description of the data at hand.

Some real world examples?

# How to include interaction in a MLR

Model A:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Model B:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

## Key points

- “easily” conceptualized with 1 continuous, 1 categorical variable
- models possible with other variable combinations, but interpretation/visualization harder
- two variable interactions are considered “first-order” interactions
- still a **linear** model, but no longer a strictly **additive** model

# How to interpret an interaction model

For now, assume  $x_1$  is continuous,  $x_2$  is 0/1 binary.

Model A:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Model B:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

# How to interpret an interaction model

For now, assume  $x_1$  is continuous,  $x_2$  is 0/1 binary.

Model A:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Model B:  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \cdot x_{i2} + \epsilon_i$

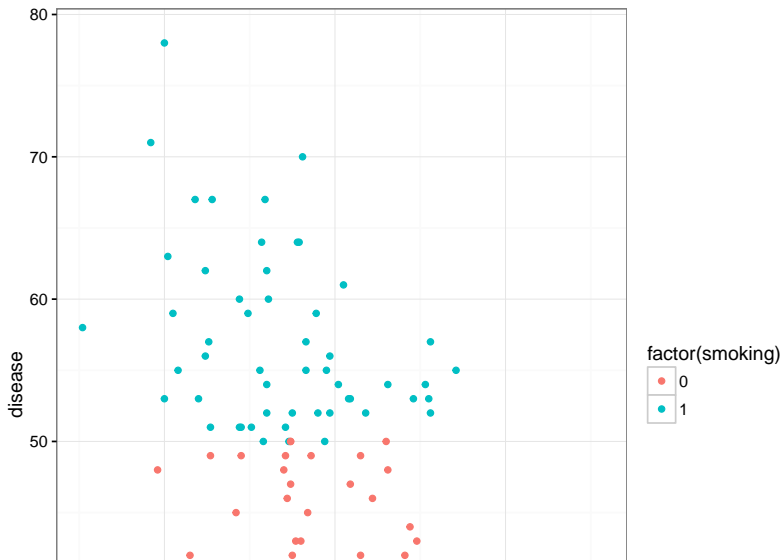
$\beta_3$  is the change in the slope of the line that describes the relationship of  $y \sim x_1$  comparing the groups defined by  $x_2 = 0$  and  $x_2 = 1$ .

$\beta_1 + \beta_3$  is the expected change in  $y$  for a one-unit increase in  $x_1$  in the group  $x_2 = 1$ .

$\beta_0 + \beta_2$  is the expected value of  $y$  in the group  $x_2 = 1$  when  $x_1 = 0$ .

# Example interaction model with lung data

```
ggplot(dat, aes(nutrition, disease, color=factor(smoking))) +  
  geom_point()
```



## Example interaction model with lung data

$$dis_i = \beta_0 + \beta_1 nutrition_i + \beta_2 smoking_i + \beta_3 nutrition \cdot smoking_i + \epsilon_i$$

```
mi1 <- lm(disease ~ nutrition + smoking, data=dat)
mi2 <- lm(disease ~ nutrition*smoking, data=dat)
c(summary(mi1)$adj.r.squared, summary(mi2)$adj.r.squared)
round(summary(mi2)$coef, 2)
```

```
## [1] 0.6190283 0.6483849
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	39.60	1.65	24.05	0.00
## nutrition	0.03	0.02	1.49	0.14
## smoking	20.69	2.15	9.61	0.00
## nutrition:smoking	-0.08	0.03	-3.00	0.00

## Example interaction model with lung data

$$dis_i = \beta_0 + \beta_1 nutrition_i + \beta_2 smoking_i + \beta_3 nutrition \cdot smoking_i + \epsilon_i$$

```
mi1 <- lm(disease ~ nutrition + smoking, data=dat)
mi2 <- lm(disease ~ nutrition*smoking, data=dat)
c(summary(mi1)$adj.r.squared, summary(mi2)$adj.r.squared)
round(summary(mi2)$coef, 2)
```

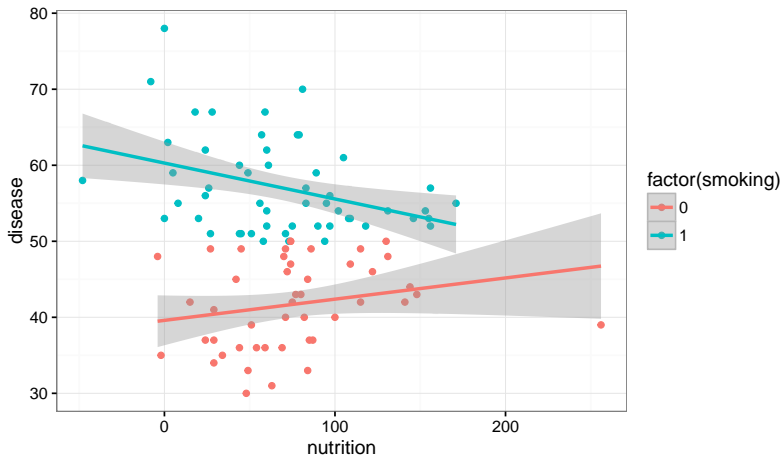
```
## [1] 0.6190283 0.6483849
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      39.60      1.65    24.05   0.00
## nutrition         0.03      0.02     1.49   0.14
## smoking          20.69      2.15     9.61   0.00
## nutrition:smoking -0.08      0.03    -3.00   0.00
```

Among non-smokers there is little evidence to support an association between nutrition and disease status. For every 10 units increase in nutrition score, the expected disease score increases by 0.3 points. The models find evidence that this relationship is significantly different for smokers, estimating that for every 10 unit increase in nutrition, disease score would decrease by 0.5 points.



# Example interaction model with FEV data

```
ggplot(dat, aes(nutrition, disease, color=factor(smoking))) +  
  geom_point() + geom_smooth(method="lm")
```



## Example interaction model with lung data

```
dat$smoking_relevel <- factor(dat$smoking, levels=c(1,0))  
mi3 <- lm(disease ~ nutrition*smoking_relevel, data=dat)  
round(summary(mi3)$coef, 2)
```

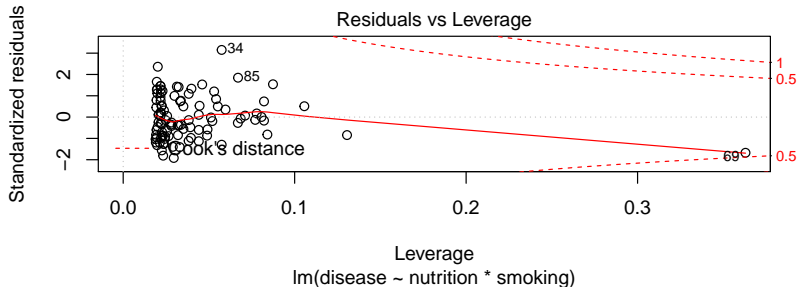
```
##              Estimate Std. Error t value  
## (Intercept)         60.29      1.39   43.46  
## nutrition          -0.05      0.02   -2.84  
## smoking_relevel0   -20.69      2.15  -9.61  
## nutrition:smoking_relevel0    0.08      0.03    3.00  
##              Pr(>|t|)  
## (Intercept)         0.00  
## nutrition           0.01  
## smoking_relevel0    0.00  
## nutrition:smoking_relevel0    0.00
```

Indeed, we see that there is a 'significant' negative slope for smokers.

## Checking influential points

We note that these results are sensitive to the inclusion of an influential outlying observation which had a much higher value of nutrition than any other observation.

```
plot(mi2, which=5)
```



```
dat[69,]
```

```
##      disease education crowding airqual nutrition smoking
## 69      39          8      20      54      256          0
```

## Results sensitivity to outlier

```
round(summary(mi2)$coef, 2)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	39.60	1.65	24.05	0.00
## nutrition	0.03	0.02	1.49	0.14
## smoking	20.69	2.15	9.61	0.00
## nutrition:smoking	-0.08	0.03	-3.00	0.00

```
mi2a <- lm(disease ~ nutrition*smoking, data=dat, subset=-69)
```

```
round(summary(mi2a)$coef, 2)
```

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	38.13	1.85	20.66	0.00
## nutrition	0.05	0.02	2.21	0.03
## smoking	22.15	2.30	9.63	0.00
## nutrition:smoking	-0.10	0.03	-3.47	0.00

# Interaction modeling summary

- Interactions can give you a more detailed story about your data.
- They are 'easier' to interpret/visualize with a binary and continuous variable interaction.
- They are also valid for continuous x continuous variables: as the value of variable  $A$  increases, the association between  $B$  and  $Y$  changes.
- Interaction is sometimes referred to as 'effect modification'.

model selection

# 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

- In practice, issues with sample size, collinearity, and available predictors are real problems.
- There is not a single best algorithm for model selection! It pretty much always requires thoughtful reasoning and knowledge about the data at hand.
- When in doubt (unless you are specifically “data mining”), err on the side creating a process that does not require choices being made (by you or the computer) about which covariates to include.

# 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 (i.e. include non-linear relationships with covariates) 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/parameters you include.
- Rule of thumb for multiple linear regression is to have at least 15 observations for each regression coefficient you include in your model.