Introduction to Statistical Modeling Categorical variables

Joris Vankerschaver

Example: BIRNH study

- Epidemiological follow-up study in the mid 80s where nutritional and health data in Belgium were measured (n=5,815)
- Goal: effect of smoking on cholesterol
- Since people from different provinces might have a different smoking and dietary behaviour, we want to correct for province
- Possible values of this variable:
 - 1: West Flanders
 - 2: East Flanders
 - 3: Flemish Brabant
 - 7: Antwerp
 - 8: Limburg

A first analysis ...

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 420.48851244 7.684314e+01 5.472037 4.646481e-08
SMOKING 1.60405509 1.384697e+00 1.158416 2.467450e-01
AGE -14.85347652 4.442769e+00 -3.343292 8.334607e-04
I(AGE^2) 0.34173838 8.359639e-02 4.087956 4.414182e-05
I(AGE^3) -0.00234162 5.126759e-04 -4.567447 5.045194e-06
SEX 9.77881436 1.290010e+00 7.580415 4.021198e-14
PROVINCE -1.68519064 2.307958e-01 -7.301653 3.249025e-13
```

Model implicitly assumes that mean difference in cholesterol

- between Limburg and West-Flanders is 7 times as large as
- the one between East- and West-Flanders

Dummy variables

Create 4 dummy variables

$$P_2 = \begin{cases} 1 & \text{East Flanders} \\ 0 & \text{other} \end{cases}$$

$$P_3 = \begin{cases} 1 & \text{Flemish Brabant} \\ 0 & \text{other} \end{cases}$$

$$P_7 = \begin{cases} 1 & \text{Antwerp} \\ 0 & \text{other} \end{cases}$$

$$P_8 = \begin{cases} 1 & \text{Limburg} \\ 0 & \text{other} \end{cases}$$

Dummy variables

- These 4 dummy variables carry same information as variable PROVINCE:
 - In West Flanders: $(P_2, P_3, P_7, P_8) = (0, 0, 0, 0)$
 - In East Flanders: $(P_2, P_3, P_7, P_8) = (1, 0, 0, 0)$
 - In Flemish Brabant: $(P_2, P_3, P_7, P_8) = (0, 1, 0, 0)$
 - $\bullet \ \ \text{In Antwerp:} \ (P_2,P_3,P_7,P_8) = (0,0,1,0)$
 - In Limburg: $(P_2, P_3, P_7, P_8) = (0, 0, 0, 1)$
- Each categorical variable with k levels can be transformed into k-1 dummy variables by choosing 1 level as **reference**:
- In R:

```
m <- lm(TCHOL ~ SMOKING + AGE + I(AGE<sup>2</sup>) + I(AGE<sup>3</sup>)
+ SEX + factor(PROVINCE))
```

Analysis with dummy variables

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	427.509211438	7.681468e+01	5.565462	2.738817e-08
SMOKING	1.691720424	1.383183e+00	1.223063	2.213583e-01
AGE	-15.186319875	4.440163e+00	-3.420217	6.302833e-04
I(AGE^2)	0.347689341	8.354742e-02	4.161581	3.208966e-05
I(AGE^3)	-0.002376407	5.123749e-04	-4.638023	3.599409e-06
SEX	9.777916700	1.288500e+00	7.588602	3.777949e-14
factor(PROVINCE)2	-8.004574314	2.043346e+00	-3.917385	9.060444e-05
factor(PROVINCE)3	-8.332945248	1.487402e+00	-5.602350	2.217790e-08
factor(PROVINCE)7	-11.567195456	1.858041e+00	-6.225478	5.157986e-10
factor(PROVINCE)8	-13.559736351	1.999453e+00	-6.781721	1.312645e-11

How to test for effect of province?

```
Std. Error t value
                                                          Pr(>|t|)
                       Estimate
(Intercept)
                427.509211438 7.681468e+01
                                             5.565462 2.738817e-08
SMOKING
                    1.691720424 1.383183e+00 1.223063 2.213583e-01
AGE.
                 -15.186319875 4.440163e+00 -3.420217 6.302833e-04
I(AGE^2)
                   0.347689341 8.354742e-02 4.161581 3.208966e-05
I(AGE^3)
                  -0.002376407 5.123749e-04 -4.638023 3.599409e-06
SEX
                   9.777916700 1.288500e+00 7.588602 3.777949e-14
factor(PROVINCE)2 -8.004574314 2.043346e+00 -3.917385 9.060444e-05
factor(PROVINCE)3 -8.332945248 1.487402e+00 -5.602350 2.217790e-08
factor(PROVINCE)7 -11.567195456 1.858041e+00 -6.225478 5.157986e-10
factor(PROVINCE)8 -13.559736351 1.999453e+00 -6.781721 1.312645e-11
```

Necessary to test if multiple coefficients are zero

Partial F-test

Assume we want to compare 2 nested models:

• Complete model (C) with p_C parameters; e.g., $p_C=10$ and

$$E(Y|X,P) = \beta_0 + \beta_1 X + \beta_2 P$$

 \bullet Reduced model (R) with p_R parameters; e.g., $p_R=6$ and

$$E(Y|X,P) = \beta_0^* + \beta_1^* X$$

• Testing $H_0: \beta_2 = 0$ is equivalent to testing if complete and reduced model are equal.

Partial F-test

- Under null hypothesis, residual sums of squares of both models will be approximately same.
- Test statistic:

$$SSE(R) - SSE(C)$$
.

• What is distribution under null hypothesis?

$$\frac{SSE(R) - SSE(C)}{p_C - p_R} \div \frac{SSE(C)}{n - p_C} \sim F_{p_C - p_R, n - p_C}.$$

under null hypothesis

Partial F-test in R

Analysis of Variance Table

```
Model 1: TCHOL ~ SMOKING + AGE + I(AGE^2) + I(AGE^3) + SEX

Model 2: TCHOL ~ SMOKING + AGE + I(AGE^2) + I(AGE^3) + SEX + factor(PRO

Res.Df RSS Df Sum of Sq F Pr(>F)

1 5480 10460957

2 5476 10329117 4 131840 17.474 2.943e-14 ***
---
```

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