

Regression with a Categorical Predictor

Ryan Miller

Introduction

From our previous lab, consider the following model for the Happy Planet data:

$$\text{Happiness} = \beta_0 + \beta_1 \cdot \text{Life Expectancy} + \epsilon$$

In the regression framework, this model:

- ▶ Uses a country's life expectancy to predict its happiness score
- ▶ Assumes the distribution of happiness scores at each life expectancy value is Normal with a common standard deviation
- ▶ Has a slope and intercept that can be estimated by minimizing the sum of squared residuals

Another Model

- ▶ Instead of life expectancy, suppose we'd like to use Region to predict a country's happiness
 - ▶ It wouldn't make sense to Happiness = $\beta_0 + \beta_1 \cdot \text{Region} + \epsilon$ because Region is categorical
 - ▶ Multiplying a category label doesn't make sense, and coding Region using the numbers 1-7 would force an arbitrary linear relationship
- ▶ To come up with a sensible model formula, we need to represent the many categories of Region using a set of many binary variables

One-hot Encoding

In this example, we map the Type of a college to the **dummy variable** named Type = "Public" using **one-hot encoding**:

College	Type
Grinnell College	"Private"
University of Iowa	"Public"
University of Minnesota	"Public"
Middlebury College	"Private"
Carlton College	"Private"



College	Type = "Public"
Grinnell College	0
University of Iowa	1
University of Minnesota	1
Middlebury College	0
Carlton College	0

Note that private colleges are the **reference group** in this example, they do not have a dummy variable representation.

One-hot Encoding

One-hot encoding is frequently used for nominal categorical variables with many categories:

College	State
Grinnell College	IA
University of Iowa	IA
University of Minnesota	MN
Middlebury College	VT
Carlton College	MN



College	State = "MN"	State = "VT"
Grinnell College	0	0
University of Iowa	0	0
University of Minnesota	1	0
Middlebury College	0	1
Carlton College	1	0

In this example, "IA" is the reference category, and dummy variables are used to represent the other two categories appearing in the data.

Practice

Shown below are the first few entries for the Region variable in the Happy Planet data:

Country	Region	Happiness
Albania	7	5.5
Algeria	3	5.6
Angola	4	4.3
Argentina	1	7.1
Armenia	7	5.0
Australia	2	7.9

How would one-hot encoding represent the data shown above if Region 1 is used as the reference category?

Practice

Country	Region2	Region3	Region4	Region5	Region6	Region7	Happiness
Albania	0	0	0	0	0	1	5.5
Algeria	0	1	0	0	0	0	5.6
Angola	0	0	1	0	0	0	4.3
Argentina	0	0	0	0	0	0	7.1
Armenia	0	0	0	0	0	1	5.0
Australia	1	0	0	0	0	0	7.9

Regression with One-Hot Predictors

To form a regression model using these new one-hot variables, we will need many different “slope” coefficients:

$$\text{Happiness} = \beta_0 + \beta_1 \cdot \text{Region2} + \beta_2 \cdot \text{Region3} + \dots + \beta_6 \cdot \text{Region7} + \epsilon$$

Like before, the components of the population model (each β_j) are estimated using the sample data by minimizing the model's sum of squared residuals.

Regression with One-Hot Predictors

Shown below are the estimates of these coefficients using R:

```
region_model = lm(Happiness ~ Region, data = happy_planet)
summary(region_model)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	6.9125000	0.1364981	50.641736	3.863078e-90
## Region2	0.6416667	0.1930374	3.324053	1.140295e-03
## Region3	-0.9250000	0.2158224	-4.285931	3.422258e-05
## Region4	-2.8640152	0.1793936	-15.964977	5.587494e-33
## Region5	-1.3267857	0.2872489	-4.618941	8.844326e-06
## Region6	-0.5958333	0.2364216	-2.520215	1.288422e-02
## Region7	-1.1754630	0.1875987	-6.265838	4.560674e-09

- ▶ The model's intercept, 6.91, is the predicted happiness score for the reference category (countries in Region 1)
- ▶ Each "slope" is the estimated difference between that region and the reference category
 - ▶ The expected happiness for Region 2 is $6.91 + 0.64 = 7.55$

Hypothesis Tests

Like we saw with simple linear regression, R will test $H_0 : \beta_j = 0$ for every coefficient:

```
region_model = lm(Happiness ~ Region, data = happy_planet)
summary(region_model)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	6.9125000	0.1364981	50.641736	3.863078e-90
## Region2	0.6416667	0.1930374	3.324053	1.140295e-03
## Region3	-0.9250000	0.2158224	-4.285931	3.422258e-05
## Region4	-2.8640152	0.1793936	-15.964977	5.587494e-33
## Region5	-1.3267857	0.2872489	-4.618941	8.844326e-06
## Region6	-0.5958333	0.2364216	-2.520215	1.288422e-02
## Region7	-1.1754630	0.1875987	-6.265838	4.560674e-09

- ▶ These T -tests compare the *difference in means* between each one-hot encoded category and the reference category
- ▶ For example, conclude there is strong evidence ($p = 0.0011$) of a difference in means for Regions 1 and 2.

Practice

In this model I've changed the reference category to Region 7:

```
region7_mod = lm(Happiness ~ Region, data = happy_planet)
summary(region7_mod)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	5.7370370	0.1286916	44.5797217	4.916846e-83
## Region1	1.1754630	0.1875987	6.2658383	4.560674e-09
## Region2	1.8171296	0.1875987	9.6862604	3.390549e-17
## Region3	0.2504630	0.2109719	1.1871866	2.372235e-01
## Region4	-1.6885522	0.1735278	-9.7307325	2.620489e-17
## Region5	-0.1513228	0.2836225	-0.5335357	5.945337e-01
## Region6	0.5796296	0.2320021	2.4983806	1.366751e-02

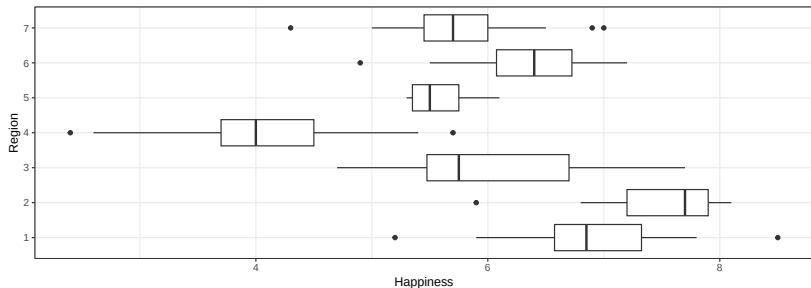
- ▶ What is the estimated mean happiness score in Region 7?
- ▶ What is the estimated mean happiness score in Region 5?
- ▶ Is there statistical evidence of a difference in mean happiness between Regions 7 and 5?

Model Assumptions

- ▶ The main assumption of regression is that the model's errors follow identical Normal distributions
- ▶ For a single categorical predictor, the model's prediction is always the mean outcome in the category of the observation being predicted
 - ▶ Thus, residuals are simply the observed outcomes minus their group means.
 - ▶ For example, we saw 7.55 was the prediction for Region 2, and Australia (a country in Region 2) has a happiness score of 7.9, so a residual of $7.9 - 7.55 = 0.35$
- ▶ So, the modeling assumption we must check is that the data in each group (category) are approximately Normal with a similar spread (standard deviation)

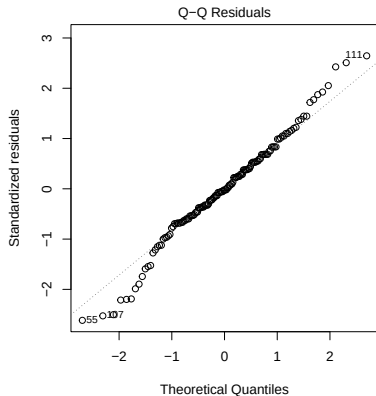
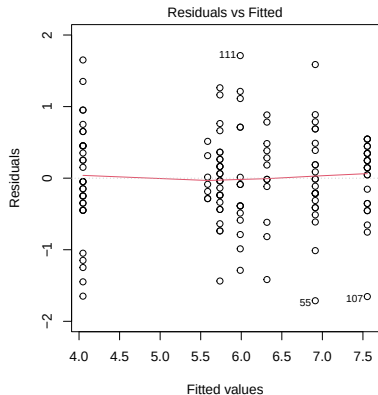
Model Assumptions

For a regression with a single categorical predictor, we now have two ways to assess the model assumptions:



Here we see that happiness scores within a region are generally symmetric, and aside from Region 3 (larger spread), there is similar variability in happiness scores across regions.

Model Assumptions



Key Takeaways

- ▶ Categorical predictors are represented using a set of one-hot variables (dummy variables)
 - ▶ A variable with k categories requires $k - 1$ dummy variables to represent it
- ▶ The intercept in this model is the mean value in the reference category
 - ▶ Each “slope” coefficient measures the difference in means from the corresponding category and the reference category
- ▶ Changing the reference category changes the “slope” coefficients and their interpretations, but not the model predictions