Introducing Multiple Regression

November 8, 2019

Example

The regression line for the `faithful` data using `waiting` to predict `eruption` was

$$\hat{y} = -1.874 + 0.076x$$

Example

The ANOVA table for this regression is

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
faithful$waiting	1	286.478	286.478	1162.1	< 2.2e-16
Residuals	270	66.562	0.247		

1. Find the appropriate interval for the average eruption time when the wait time is 70 minutes.
2. Find the appropriate interval for the specific eruption time when the wait time is 70 minutes.

Multiple Regression

- An outcome may be simultaneously influenced by many variables.
- Think back to our randomized block and factorial designs.
- Multiple regression extends this idea into the regression framework.
- We will extend the simple linear regression to include many predictor variables.

Consider

We will work with a data set that contains patient data for a clinical trial with 1000 participants.

- Each patient entered the study with high LDL cholesterol.
- Patients were randomly assigned to either a medication to manage their cholesterol or to a placebo.
- We can always examine treatment and change in LDL cholesterol.
- But what about other variables?

The Data

	ldl.post	trt	sex	age	weight	sys.bp	dia.bp	income	ldl.pre
1	176	0	M	47	186	119	64	low	178
2	191	0	M	37	185	121	72	med	191
3	155	1	M	48	208	106	71	med	203
4	123	1	F	46	159	106	57	med	168
5	120	1	M	37	117	100	74	low	168
6	134	1	F	38	228	128	71	low	190
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

How might weight impact the medication's effectiveness? What about blood pressure?

We start by fitting a linear regression model using `trt` to predict `ldl.post - ldl.pre`.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.1588	0.1576	1.007	0.314
trt	-48.1471	0.2196	-219.216	<2e-16

Write the regression line. Then, interpret the slope and intercept.

Indicator and Categorical Variables as Predictors

We can also fit models using categorical variables with more than 2 levels.

- The `income` variable has 3 levels: high, medium, and low.
- Suppose we want to know whether socioeconomic status is relevant to treatment outcomes.

Indicator and Categorical Variables as Predictors

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	-23.6000	1.8949	-12.454	<2e-16
incomelow	-0.9113	2.1937	-0.415	0.678
incomemed	-1.7000	2.2986	-0.740	0.460

- Each row represents the relative difference for each level.
- Notice we are missing income:high. This is the **reference level** that the other variables are measured against.
- I let `R` choose the reference level, but we could pick any one of the income levels to act as the "default".

Example

None of the levels of income appear to be good predictors for our treatment outcomes, but let's think about how to use the regression equation.

Predictors with Several Categories

- When fitting a regression model with a categorical variable that has k levels, standard software will provide a coefficient for $k - 1$ of them.
- For the level that does not receive a coefficient, this is the reference level.
- The coefficients listed for the other levels are all considered relative to this reference level.

Including and Assessing Many Variables in a Model

- The world is complex! More information is typically better information.
- If we have the ability to collect and use many variables, we should use them!
- This is the idea behind **multiple linear regression**.

Including and Assessing Many Variables in a Model

We want to construct a model for our cholesterol data using all of the variables simultaneously:

$$\begin{aligned}\widehat{\texttt{ldl.post}} = \beta_0 &+ \beta_1 \times \texttt{trt} + \beta_2 \times \texttt{sex} + \beta_3 \times \texttt{age} \\ &+ \beta_4 \times \texttt{weight} + \beta_5 \times \texttt{sys.bp} + \beta_6 \times \texttt{dia.bp} \\ &+ \beta_7 \times \texttt{income}_{med} + \beta_8 \times \texttt{income}_{low} + \beta_9 \times \texttt{ldl.pre}\end{aligned}$$

Estimating Parameters

We estimate $\beta_0, \beta_1, \ldots, \beta_9$ the same way we did for our linear regression with only two parameters, by minimzing the sum of squared residuals

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

But... this time we'll definitely use a computer.