Interactions and Nonlinear Trends

November 18, 2019

Consider

Consider an experiment that examines the impact of Vitamin C from two sources on the growth of teeth in Guinea pigs.

- Each Guinea pig was randomly assigned to one of two possible levels of each variable.
  - `supp` indicates a supplement type for Vitamin C, with levels `VC` for ascorbic acid and `OJ` for orange juice.
  - `dose` indicates the amount of Vitamin C, which takes values of either 1 or 2 mg.

The outcome is the tooth length of the Guinea pigs.

Example

- We want to build a regression model for these data.
- We start with the following model:

$$Y = \beta_0 + \beta_1 x_{supp} + \beta_2 x_{dose} + \epsilon$$

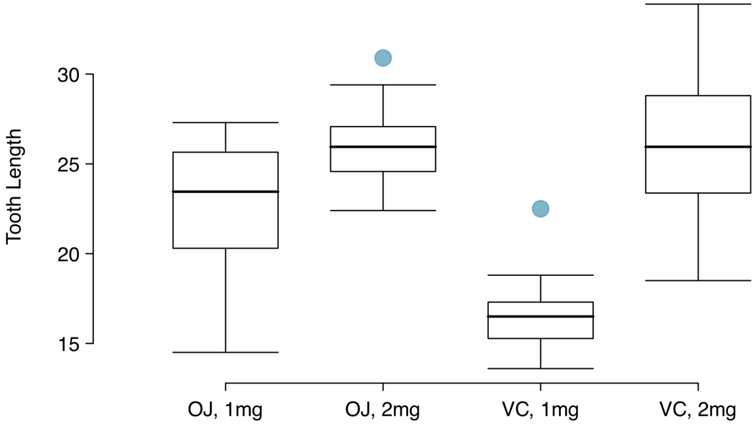
Example

The model output is

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	14.8325	2.0319	7.30	0.0000
suppVC	-2.9250	1.2253	-2.39	0.0222
dose	6.3650	1.2253	5.19	0.0000

Write the regression model for these data.

Example



Predict an outcome for each possible combination of variables. How do these means compare to the boxplots?

Example

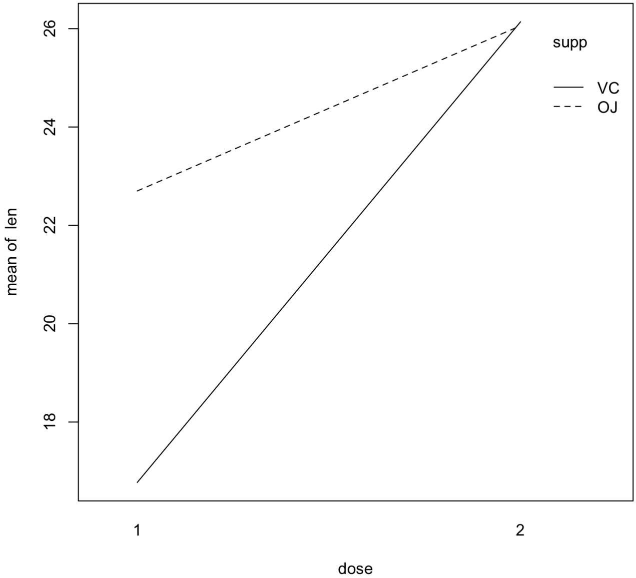
For this model,

- $R^2 = 0.5553$
- $R^2_{adj} = 0.5183$

When we built our model, we assumed that the effects of the supplement and dose are independent.

- What if this isn't the case?
- Can we improve the model?

Interactions



Interactions

- There appears to be some interaction between `dose` and `supp`
- Recall: an **interaction** means that the effect of one may partially depend on the value of the other.
- We can model this effect by including an additional term:

$$Y = \beta_0 + \beta_1 x_{supp} + \beta_2 x_{dose} + \beta_3 x_{dose} x_{supp} + \epsilon$$

Interactions

Consider a few rows of the tooth growth data:

len	supp	dose
16.5	VC	1
16.5	VC	1
25.5	VC	2
19.7	OJ	1
23.3	OJ	1
26.4	OJ	2

What does this look like with the interaction term?

Example

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	19.3400	2.5419	7.61	0.0000
suppVC	-11.9400	3.5948	-3.32	0.0021
dose	3.3600	1.6076	2.09	0.0437
suppVC:dose	6.0100	2.2735	2.64	0.0121

R^2: 0.7296, R^2_{adj}: 0.7151

Write out the regression model. Calculate the predicted value for each group.

Nonlinear Trends

- One of our regression assumptions is of linear relationships.
- What happens when this is violated?
- We may be able to use a **transformation** to ”fix” things.

Transforming the Response Variable

These techniques may be useful for violations of

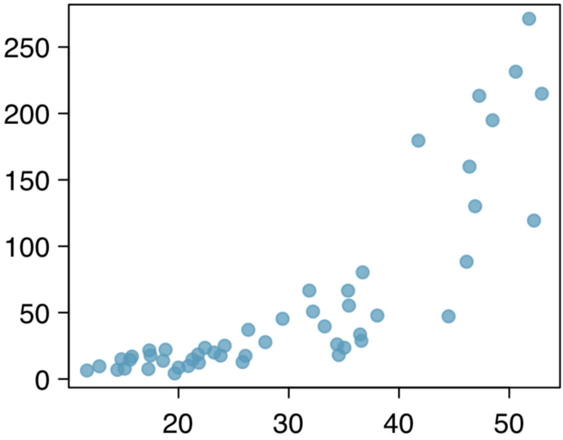
- linearity.
- normality of residuals.
- constant variability of residuals.

Transformations for Nonlinear Trends

We will consider two types of transformation:

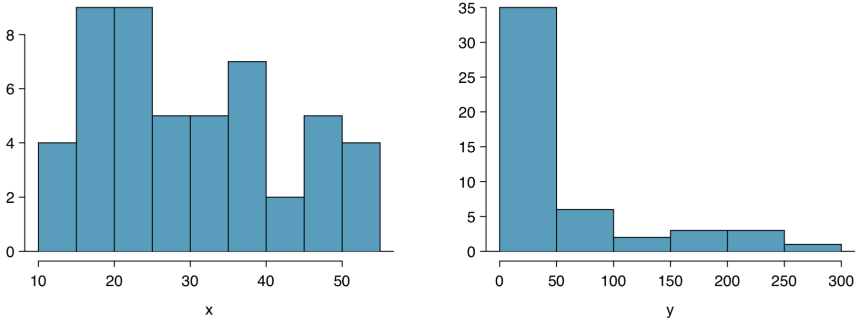
1. transforming the response variable.
2. using polynomial terms in multiple regression.

Transformations



- The trend may be nonlinear.
- The variability increases with x.

Transformations



- x looks pretty reasonable.
- y is extremely right-skewed.
- So we probably want to transform y.

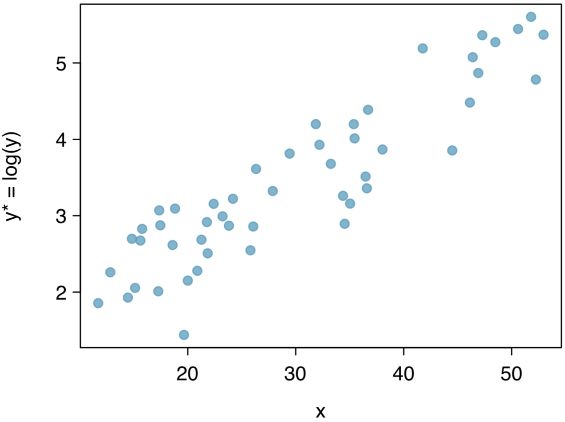
Transformations on the Response

One of the more common transformations is the **natural log** (ln).

$$y^* = \log y$$

Note: In statistics, "log" almost always implies the natural log.

Transformations on the Response



Now, the relationship looks linear and the variability looks pretty constant.

Back-Transforming

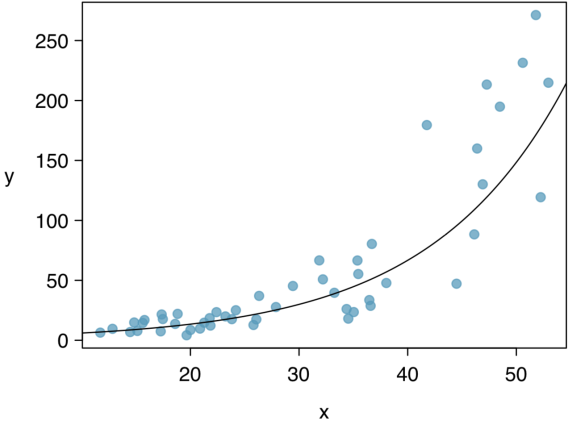
The model for these transformed data is

$$\hat{y}^* = 1.03 + 0.08x$$

To get predictions for y, we need to **back-transform** the data by solving for y.

Back transform the data and predict the value of y when $x = 31$.

Back-Transforming



The back-transformed equation overlaid on the original data.

Interpreting Coefficients

For a model that used $\log y$ and fits the data well, we say

- y tends to grow (or decay) **exponentially** relative to x.

A Word of Caution

- In theory, there are infinite possible transformations.
- If we keep trying different transformations until one ”works”, we haven’t effectively modeled our data.
- In actuality, this is a form of data fishing!
- This is one reason to think carefully and stick to mostly standard transformations.