Multiple Regression

More than one explanatory (independent) variable, one response (dependent) variable

\[ \hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k \quad k=\#\text{independent variables} \]

\( b_i = \text{change in } Y \text{ when } x_i \text{ changes by one unit and all the other } x \text{ variables are held constant} \)

\( R^2 = \% \text{ of variability in } Y \text{ that is explained by the model} \)

as you add variables, \( R^2 \) can only go up - don’t want to add noise

Adjusted \( R^2 \) - penalty for each extra variable in the model

Could choose a model that maximizes adjusted \( R^2 \)-square

Significance Tests

A) To see if the model as a whole is significant

1) Let \( \beta_i \) be the population slope for variable \( X_i \)

\[ H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \]

\( H_1: \text{at least one } \beta_i \neq 0 \)

\[ Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k + \epsilon \quad \text{EN N}(0, \sigma^2) \]

2) \( \alpha = .05 \)

3) F-test compute in JMP

4) get p-value from JMP

5) reject \( H_0 \) if p-value < \( \alpha \)

6) draw conclusions in context

"The model is helpful"

Ex: There is evidence of a linear relationship between latitude/longitude coordinates and temperature.

Ex: Not enough evidence of a linear relationship between either latitude or longitude and temperature

"The model is not helpful"
B) To see if a particular variable is significant, when all the others are included in the model.

1) \( H_0: \beta_j = 0 \) when all other variables are included in the model.
   \( H_1: \beta_j \neq 0 \) \( \beta_j \) is the population slope for variable \( X_j \)

2) \( \alpha = .05 \)
3) t-test \( df = n - k - 1 \)
4) get p-value from JMP
5) reject \( H_0 \) if \( p < \alpha \)
6) draw conclusions

Example: Conclude that there is a linear relationship between temperature and latitude when longitude is in the model.

**Model Building**

The basic idea of model building is that you want to get the most accurate predictions with the smallest possible prediction variance.

To increase predictive accuracy:
- add relevant variables
- transform variables

To decrease predictive variance:
- remove statistically insignificant variables
- remove highly correlated "independent" variables
- use as few terms as possible

Need to balance these:
- Adjusted R\(^2\)
- Stepwise t-tests

**Logistic Regression**

What if we have a binary (nominal, 2 category) response? Can't model directly with linear regression.

Consider the probability \( P \)

Model the logistic transformation \( \log \frac{P}{1-P} \) spreads the probability to the whole real line.

Logistic regression: \( \log \frac{P}{1-P} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \)