Regression Models for Quantitative and Qualitative Predictors

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

8.1 Polynomial regression

- Used when the relationship between Y and the predictor(s) is curvilinear.

8.1 Polynomial regression

- Used when the relationship between Y and the predictor(s) is curvilinear.
- Example: we might add a quadratic term to a simple linear model to get a parabolic mean

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \epsilon_i.$$

8.1 Polynomial regression

- Used when the relationship between Y and the predictor(s) is curvilinear.
- Example: we might add a quadratic term to a simple linear model to get a parabolic mean

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \epsilon_i.$$

- We can no longer interpret β_1 and β_{11} "as usual." Cannot hold x_1 constant and increase x_1^2 by one unit, or vice-versa!

8.1 Polynomial regression

- Used when the relationship between Y and the predictor(s) is curvilinear.
- Example: we might add a quadratic term to a simple linear model to get a parabolic mean

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \epsilon_i.$$

- We can no longer interpret β_1 and β_{11} "as usual." Cannot hold x_1 constant and increase x_1^2 by one unit, or vice-versa!
- Can be done easily in R, e.g. `lm(y ~ x + x*x)`.

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$.

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.
- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ "turning points" or extrema.

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.
- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ "turning points" or extrema.
- A $(k-1)$th-order polynomial can go through $(x_1, Y_1), \cdots, (x_k, Y_k)$ *exactly*!

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.
- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ "turning points" or extrema.
- A $(k-1)$th-order polynomial can go through $(x_1, Y_1), \cdots, (x_k, Y_k)$ *exactly*!
- Polynomials of degree ≥ 4 should rarely be used.

General notes on fitting polynomials

- Predictors can be first centered:

$$x^*_{ij} = x_{ij} - \bar{x}_j,$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.
- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ "turning points" or extrema.
- A $(k-1)$th-order polynomial can go through $(x_1, Y_1), \cdots, (x_k, Y_k)$ *exactly*!
- Polynomials of degree ≥ 4 should rarely be used. High-degree polynomials have wiggly behavior and can provide extremely poor out of sample prediction.

General notes on fitting polynomials

- Predictors can be first centered:

$$x_{ij}^* = x_{ij} - \bar{x}_j,$$

  where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij}$. This may reduce multicollinearity among, for example, $x_{i1}, x_{i1}^2, x_{i1}^3$, etc.
- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ "turning points" or extrema.
- A $(k-1)$th-order polynomial can go through $(x_1, Y_1), \cdots, (x_k, Y_k)$ *exactly*!
- Polynomials of degree ≥ 4 should rarely be used. High-degree polynomials have wiggly behavior and can provide extremely poor out of sample prediction. Extrapolation is particularly dangerous.

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Example: Quadratic regression, two predictors:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{1st order}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}}_{\text{2nd order}} + \epsilon_i .$$

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Example: Quadratic regression, two predictors:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{1st order}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}}_{\text{2nd order}} + \epsilon_i.$$

- This is an example of a **response surface**, or parabolic surface.

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Example: Quadratic regression, two predictors:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{1st order}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}}_{\text{2nd order}} + \epsilon_i.$$

- This is an example of a **response surface**, or parabolic surface.
- "Hierarchical model building," (p. 299) stipulates that a model containing a particular term should also contain all terms of lower order including the cross-product terms.

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Example: Quadratic regression, two predictors:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{1st order}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}}_{\text{2nd order}} + \epsilon_i.$$

- This is an example of a **response surface**, or parabolic surface.
- "Hierarchical model building," (p. 299) stipulates that a model containing a particular term should also contain all terms of lower order including the cross-product terms.
- Degree of cross-product term is obtained by summing power for each predictor. E.g. the degree of $\beta_{1123} x_{i1}^2 x_{i2} x_{i3}$ is $2 + 1 + 1 = 4$.

Hierarchical model building

Hierarchical model building

- When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate.

Hierarchical model building

- When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate.
- With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model.

Hierarchical model building

- When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate.
- With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model.
- One would not drop the quadratic term of a predictor variable but retain the cubic term in the model.

Hierarchical model building

- When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate.
- With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model.
- One would not drop the quadratic term of a predictor variable but retain the cubic term in the model. Since the quadratic term is of lower order, it is viewed as providing more basic information about the shape of the response function;

Hierarchical model building

- When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate.
- With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model.
- One would not drop the quadratic term of a predictor variable but retain the cubic term in the model. Since the quadratic term is of lower order, it is viewed as providing more basic information about the shape of the response function; the cubic term is of higher order and is viewed as providing refinements in the specification of the shape of the response function

When to include polynomial terms

- If the response vs. a predictor is curved in the initial scatterplot, this relationship *may or may not hold* when other predictors are added!

When to include polynomial terms

- If the response vs. a predictor is curved in the initial scatterplot, this relationship *may or may not hold* when other predictors are added! It's better to examine residuals versus each predictor to see if, e.g. adding a quadratic term, might be useful.

When to include polynomial terms

- If the response vs. a predictor is curved in the initial scatterplot, this relationship *may or may not hold* when other predictors are added! It's better to examine residuals versus each predictor to see if, e.g. adding a quadratic term, might be useful.
- Added variable plots are a refined plot to help figure out if the "non-linear" pattern is there when other variables are added (Section 10.1)

When to include polynomial terms

- If the response vs. a predictor is curved in the initial scatterplot, this relationship *may or may not hold* when other predictors are added! It's better to examine residuals versus each predictor to see if, e.g. adding a quadratic term, might be useful.
- Added variable plots are a refined plot to help figure out if the "non-linear" pattern is there when other variables are added (Section 10.1)
- With lots of predictors, say $k \geq 5$, it is easier to reduce to important first-order predictors, look for possible pairwise interactions (if necessary), and then see if any of the residual plots look curved; if so, add quadratic term(s).

When to include polynomial terms

- Sometimes people fit a higher-order model and then start cutting higher order terms with t and F-tests to get a simpler, more interpretable model.

When to include polynomial terms

- Sometimes people fit a higher-order model and then start cutting higher order terms with t and F-tests to get a simpler, more interpretable model. This is called **backward elimination** (Chapter 9).

When to include polynomial terms

- Sometimes people fit a higher-order model and then start cutting higher order terms with t and F-tests to get a simpler, more interpretable model. This is called **backward elimination** (Chapter 9).
- The example on pp. 300–305 starts with a full quadratic function:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x^2 + \epsilon_i$$

When to include polynomial terms

- Sometimes people fit a higher-order model and then start cutting higher order terms with t and F-tests to get a simpler, more interpretable model. This is called **backward elimination** (Chapter 9).
- The example on pp. 300–305 starts with a full quadratic function:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x^2 + \epsilon_i$$

  then throws away higher order terms through the partial F-test that FTR

$$\mathrm{H}_0 : \beta_{11} = \beta_{12} = \beta_{22} = 0,$$

When to include polynomial terms

- Sometimes people fit a higher-order model and then start cutting higher order terms with t and F-tests to get a simpler, more interpretable model. This is called **backward elimination** (Chapter 9).
- The example on pp. 300–305 starts with a full quadratic function:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x^2 + \epsilon_i$$

  then throws away higher order terms through the partial F-test that FTR

$$\mathrm{H}_0 : \beta_{11} = \beta_{12} = \beta_{22} = 0,$$

  finally leaving only the first-order terms as important:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by one unit?

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by one unit?

$$\text{at } x_1 \Rightarrow \mathrm{E}[Y] = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by one unit?

$$\begin{aligned} \text{at } x_1 &\Rightarrow \mathrm{E}[Y] = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \\ \text{at } x_1 + 1 &\Rightarrow \mathrm{E}[Y] = \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_{12} (x_1 + 1) x_2 \end{aligned}$$

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by one unit?

$$\begin{aligned}
\text{at } x_1 &\Rightarrow \mathrm{E}[Y] = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \\
\text{at } x_1 + 1 &\Rightarrow \mathrm{E}[Y] = \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_{12} (x_1 + 1) x_2 \\
\textbf{difference} &= \beta_1 + \beta_{12} x_2
\end{aligned}$$

8.2 Pairwise interactions among predictors

An interaction model includes one or several *cross-product* terms.

Example: two predictors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by one unit?

$$\begin{aligned} \text{at } x_1 &\Rightarrow \mathrm{E}[Y] = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \\ \text{at } x_1 + 1 &\Rightarrow \mathrm{E}[Y] = \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_{12} (x_1 + 1) x_2 \\ \textbf{difference} &= \beta_1 + \beta_{12} x_2 \end{aligned}$$

How the mean changes depends on the other variable.

Interactions

A model with no interactions is like a sheet of paper held "flat." Adding a pairwise interaction is like twisting the two ends of the paper.

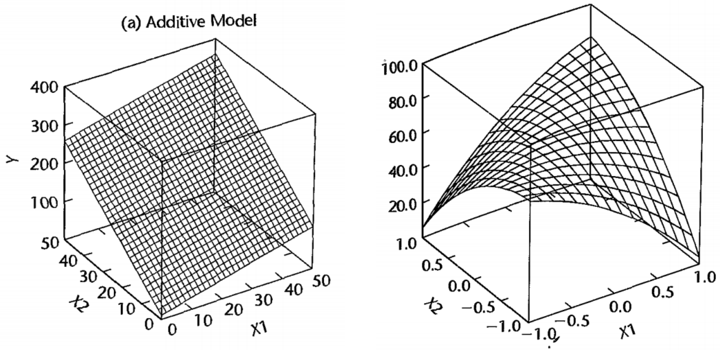


Figure: No interacations (left), With interactions (right)

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.
- The textbookook suggests fitting an additive model, then looking at residuals e_i versus each two-way interaction;

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.
- The textbookook suggests fitting an additive model, then looking at residuals e_i versus each two-way interaction; if theres a pattern you could include that the interaction should be in the model.

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.
- The textbookook suggests fitting an additive model, then looking at residuals e_i versus each two-way interaction; if theres a pattern you could include that the interaction should be in the model.
- A researcher will often have "an idea" of which variables might interact.

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.
- The textbookook suggests fitting an additive model, then looking at residuals e_i versus each two-way interaction; if theres a pattern you could include that the interaction should be in the model.
- A researcher will often have "an idea" of which variables might interact. This can be helpful.

Interactions

- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to filter them out via t-tests and/or F-tests.
- The textbookook suggests fitting an additive model, then looking at residuals e_i versus each two-way interaction; if theres a pattern you could include that the interaction should be in the model.
- A researcher will often have "an idea" of which variables might interact. This can be helpful.
- Can also start with a first-order model, then add interactions one at a time (forward selection!) using R.

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_I\}$.

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_I\}$. We need to allow each level of c to affect $\mathrm{E}[Y]$ differently.

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_I\}$. We need to allow each level of c to affect $\mathrm{E}[Y]$ differently. This is accomplished by the use of dummy variables.

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \dots, \text{category}_I\}$. We need to allow each level of c to affect $\mathrm{E}[Y]$ differently. This is accomplished by the use of dummy variables.

In R, there is a way to define categorical variables;

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_l\}$. We need to allow each level of c to affect $\mathrm{E}[Y]$ differently. This is accomplished by the use of dummy variables.

In R, there is a way to define categorical variables; then calling `lm(y ~ c)` will create and handle all dummy variables automatically.

8.3 Categorical predictors

Let's say we wish to include a categorical variable c that takes on values $c \in \{\text{category}_1, \text{category}_2, \ldots, \text{category}_l\}$. We need to allow each level of c to affect $\mathrm{E}[Y]$ differently. This is accomplished by the use of dummy variables.

In R, there is a way to define categorical variables; then calling `lm(y ~ c)` will create and handle all dummy variables automatically.

Partial F-tests can be used to see whether an entire categorical predictor can be dropped from the model (all of the dummy variables at once).

Creating zero-one dummies

Define $z_1, z_2, \ldots, z_{l-1}$ as follows:

$$z_j = \begin{cases} 1 & c = \text{category}_j \\ 0 & c \neq \text{category}_j \end{cases}$$

Creating zero-one dummies

Define $z_1, z_2, \ldots, z_{I-1}$ as follows:

$$z_j = \begin{cases} 1 & c = \text{category}_j \\ 0 & c \neq \text{category}_j \end{cases}$$

This sets the last category_I as the baseline.

Creating zero-one dummies

Define $z_1, z_2, \ldots, z_{I-1}$ as follows:

$$z_j = \begin{cases} 1 & c = \text{category}_j \\ 0 & c \neq \text{category}_j \end{cases}$$

This sets the last category$_I$ as the baseline. Say $I = 3$, then the model is

$$\mathrm{E}[Y] = \beta_0 + \beta_1 z_1 + \beta_2 z_2$$

Creating zero-one dummies

Define $z_1, z_2, \ldots, z_{l-1}$ as follows:

$$z_j = \begin{cases} 1 & c = \text{category}_j \\ 0 & c \neq \text{category}_j \end{cases}$$

This sets the last category_l as the baseline. Say $l = 3$, then the model is

$$\mathrm{E}[Y] = \beta_0 + \beta_1 z_1 + \beta_2 z_2$$

which gives

$$\begin{aligned} \mathrm{E}[Y] &= \beta_0 + \beta_1 && \text{when} \quad c = \text{category}_1 \\ \mathrm{E}[Y] &= \beta_0 + \beta_2 && \text{when} \quad c = \text{category}_2 \\ \mathrm{E}[Y] &= \beta_0 && \text{when} \quad c = \text{category}_3 \end{aligned}$$

β_1 and β_2 are *offsets to baseline* mean.

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together;

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $x \in \{1, 2, 3\}$ and $z \in \{1, 2, 3, 4\}$. An additive model is

$$\begin{aligned}\mathrm{E}[Y] = \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ + \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3].\end{aligned}$$

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $x \in \{1, 2, 3\}$ and $z \in \{1, 2, 3, 4\}$. An additive model is

$$\begin{aligned} \mathrm{E}[Y] = \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ + \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3]. \end{aligned}$$

The model that includes an interaction between x and z adds $(3 - 1)(4 - 1) = 6$ additional dummy variables accounting for all possible pairwise products.

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $x \in \{1, 2, 3\}$ and $z \in \{1, 2, 3, 4\}$. An additive model is

$$\begin{aligned}\mathrm{E}[Y] = \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ + \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3].\end{aligned}$$

The model that includes an interaction between x and z adds $(3 - 1)(4 - 1) = 6$ additional dummy variables accounting for all possible pairwise products. The new model is rather cumbersome:

$$\begin{aligned}\mathrm{E}[Y] =& \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ &+ \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3]\end{aligned}$$

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $x \in \{1, 2, 3\}$ and $z \in \{1, 2, 3, 4\}$. An additive model is

$$\begin{aligned} \mathrm{E}[Y] = \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ + \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3]. \end{aligned}$$

The model that includes an interaction between x and z adds $(3 - 1)(4 - 1) = 6$ additional dummy variables accounting for all possible pairwise products. The new model is rather cumbersome:

$$\begin{aligned} \mathrm{E}[Y] =& \beta_0 + \beta_1 \mathbb{I}[x = 1] + \beta_2 \mathbb{I}[x = 2] \\ & + \beta_3 \mathbb{I}[z = 1] + \beta_4 \mathbb{I}[z = 2] + \beta_5 \mathbb{I}[z = 3] \\ & + \beta_6 \mathbb{I}[x = 1]\mathbb{I}[z = 1] + \beta_7 \mathbb{I}[x = 1]\mathbb{I}[z = 2] \\ & + \beta_8 \mathbb{I}[x = 1]\mathbb{I}[z = 3] + \beta_9 \mathbb{I}[x = 2]\mathbb{I}[z = 1] \\ & + \beta_{10} \mathbb{I}[x = 2]\mathbb{I}[z = 2] + \beta_{11} \mathbb{I}[x = 2]\mathbb{I}[z = 3]. \end{aligned}$$

Example: Insurance innovation

An economist studied 10 mutual firms and 10 stock firms to relate speed Y (months) with which an insurance innovation is adopted to size (total assets) of insurance firm x_1 and type x_2, stock or mutual.

Example: Insurance innovation

An economist studied 10 mutual firms and 10 stock firms to relate speed Y (months) with which an insurance innovation is adopted to size (total assets) of insurance firm x_1 and type x_2, stock or mutual.

```
> ins_data = read.csv("path/to/insurance.csv", header=FALSE)
> colnames(ins_data) = c('months', 'size', 'type')
> head(ins_data)
  months size   type
1     28  164  stock
2     31   85  stock
3     30  124  stock
4     21  175 mutual
5     12  210 mutual
6     15  272  stock
```

8.5 Categorical and quantitative interaction

Consider the following model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

where x_1 is size of firm and $x_2 = 1$ if stock firm and $x_2 = 0$ otherwise.

8.5 Categorical and quantitative interaction

Consider the following model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

where x_1 is size of firm and $x_2 = 1$ if stock firm and $x_2 = 0$ otherwise.

Then

$$\begin{aligned} \mathrm{E}[Y] &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 && \text{for stock firm} \\ \mathrm{E}[Y] &= \beta_0 + \beta_1 x_1 && \text{otherwise} \end{aligned}$$

8.5 Categorical and quantitative interaction

Consider the following model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

where x_1 is size of firm and $x_2 = 1$ if stock firm and $x_2 = 0$ otherwise.

Then

$$\begin{aligned} \mathrm{E}[Y] &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 && \text{for stock firm} \\ \mathrm{E}[Y] &= \beta_0 + \beta_1 x_1 && \text{otherwise} \end{aligned}$$

Have different intercepts and different slopes.

8.5 Categorical and quantitative interaction

If we instead fit an additive model

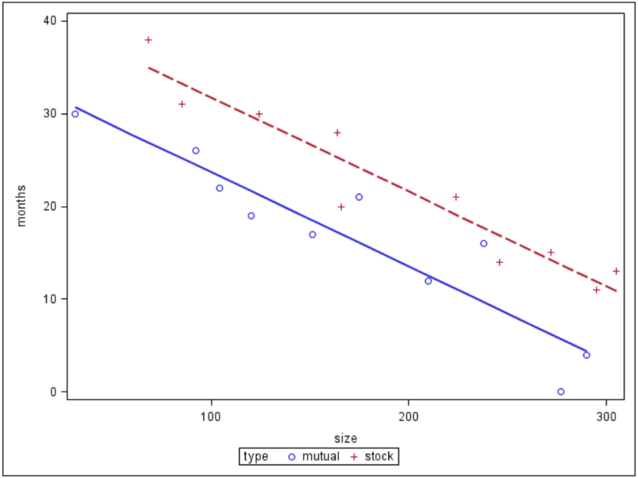
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

then

$$\begin{aligned} \mathrm{E}[Y] &= (\beta_0 + \beta_2) + \beta_1 x_1 && \text{for stock firm} \\ \mathrm{E}[Y] &= \beta_0 + \beta_1 x_1 && \text{otherwise} \end{aligned}$$

These are two parallel lines; the slope is the same. β_2 is how much better (or worse) stock firms do *at any firm size*

Insurance innovation



Do these look parallel?

Insurance innovation

```
> m1 = lm(months ~ size + type + size * type, data=ins_data)
> summary(m1)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    33.8383695  2.4406498  13.864 2.47e-10 ***
size           -0.1015306  0.0130525  -7.779 7.97e-07 ***
typestock       8.1312501  3.6540517   2.225   0.0408 *
size:typestock -0.0004171  0.0183312  -0.023   0.9821

Residual standard error: 3.32 on 16 degrees of freedom
Multiple R-squared:  0.8951,Adjusted R-squared:  0.8754
F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

Insurance innovation

Do the following yourself:

- Look at standard diagnostic plots.
- Test whether a quadratic in firm size is necessary.
- Test whether an interaction between firm size and type is necessary.
- Interpret the model.