Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Dummy Variable

Qualitative Predictors

Qualitative variables, a.k.a. *categorical variables*, represent certain characteristics of a subject:

- a fixed set of possible values/levels/classes
- Examples:
  - Blood type: A, B, AB or O
  - Smoke status: smoke or not smoke $\Longrightarrow$ *binary variable*
  - Income level: high, medium or low
  - Education level: high school, college, or advanced degree

Dummy Variables

- To include qualitative predictors in regression models, we need to quantitatively identify their levels/classes
- A popular approach is to use *dummy variables (a.k.a. indicator variables)* – variables only take on the values 0 or 1

Quantify a Binary Variable with a Dummy Variable

- Suppose the two classes are labelled as C_1, C_2. Then the dummy variable can be defined as

$$X = \begin{cases} 1 & \textit{if} \quad C_1 \\ 0 & \textit{if} \quad C_2 \end{cases}$$

- For example, to code smoking status:

$$X = \begin{cases} 1 & \textit{if} \quad \textit{smoker} \\ 0 & \textit{if} \quad \textit{nonsmoker} \end{cases}$$

- The coding is **not unique**, since the *reference class* – the class coded as 0 – is arbitrarily chosen.

Qualitative Variables with More than Two Classes

A qualitative variable with r classes, labeled as $C_1, \cdots, C_r$, need to be represented by $r - 1$ dummy variables:

$$X_1 = \begin{cases} 1 & \textit{if} \quad C_1 \\ 0 & \textit{if} \quad \textit{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \textit{if} \quad C_2 \\ 0 & \textit{if} \quad \textit{otherwise} \end{cases}$$

$$\vdots$$

$$X_{r-1} = \begin{cases} 1 & \textit{if} \quad C_{r-1} \\ 0 & \textit{if} \quad \textit{otherwise} \end{cases}$$

For C_r (the *reference class*), $X_1 = \cdots = X_{r-1} = 0$.

First Order Model with Qualitative and Quantitative Predictors

Insurance

An economist wanted to relate the speed with which a particular insurance innovation is adopted by an insurance firm to the size of the firm and the type of the firm. He collected data on 20 insurance firms, 10 stock firms and 10 mutual firms.

Firm	Number_of_month_elapsed	Firm_size	Firm_Type
1	17	151	mutual
2	26	92	mutual
3	21	175	mutual
...	...	...	...
18	13	305	stock
19	30	124	stock
20	14	246	stock

- Y – – number of months elapsed before the firm adopted the innovation and X_1 – – the amount of total assets of the firm are quantitative variables.
- Type of the firm is a binary variable taking on two values: “stock” or “mutual”. If we choose “mutual” as the reference class, then it can be quantified by a dummy variable:

$$X_2 = \begin{cases} 1 & \textit{if} \quad \textit{stock} \\ 0 & \textit{if} \quad \textit{mutual} \end{cases}$$

Figure: Side-by-Side Boxplots

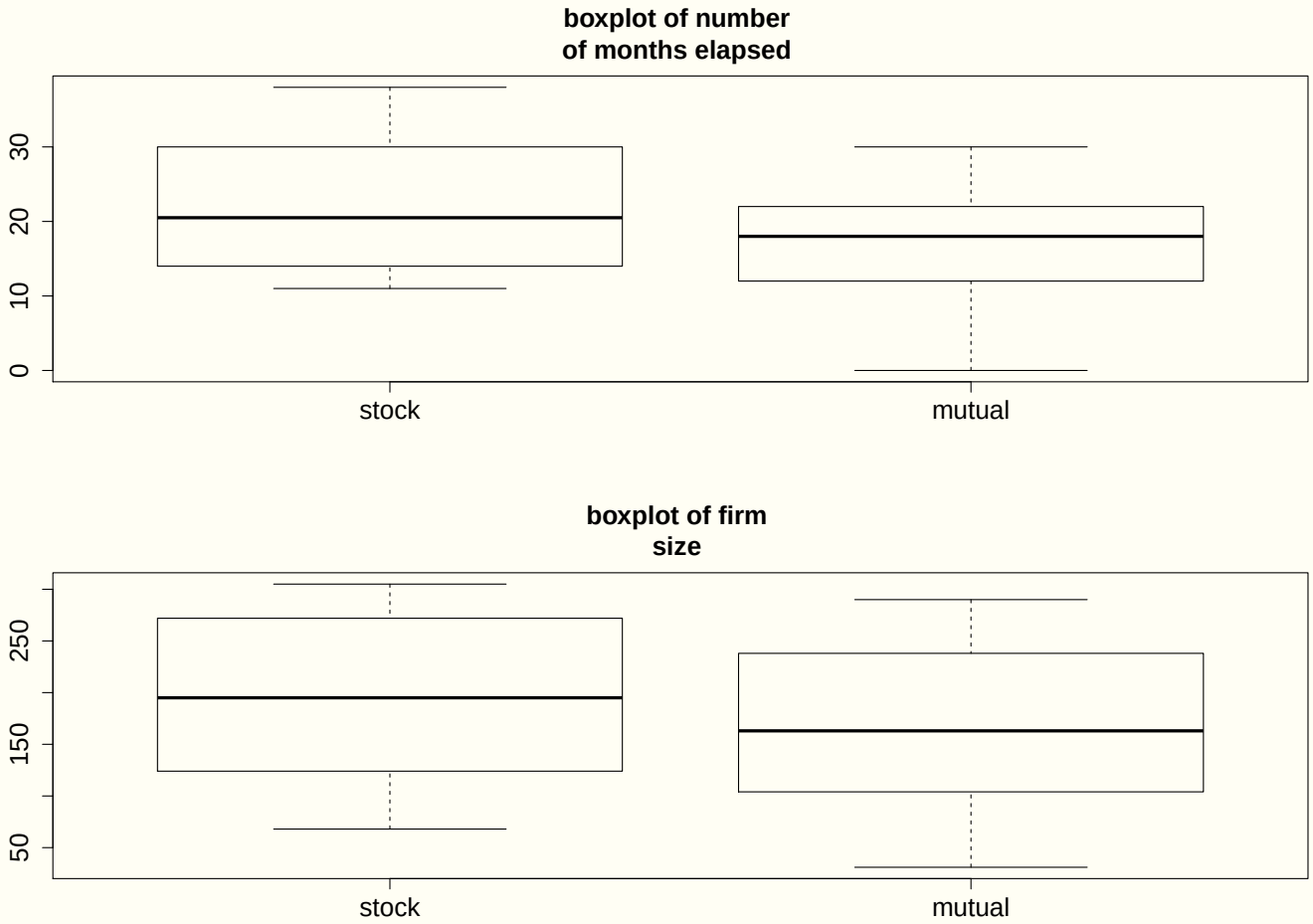
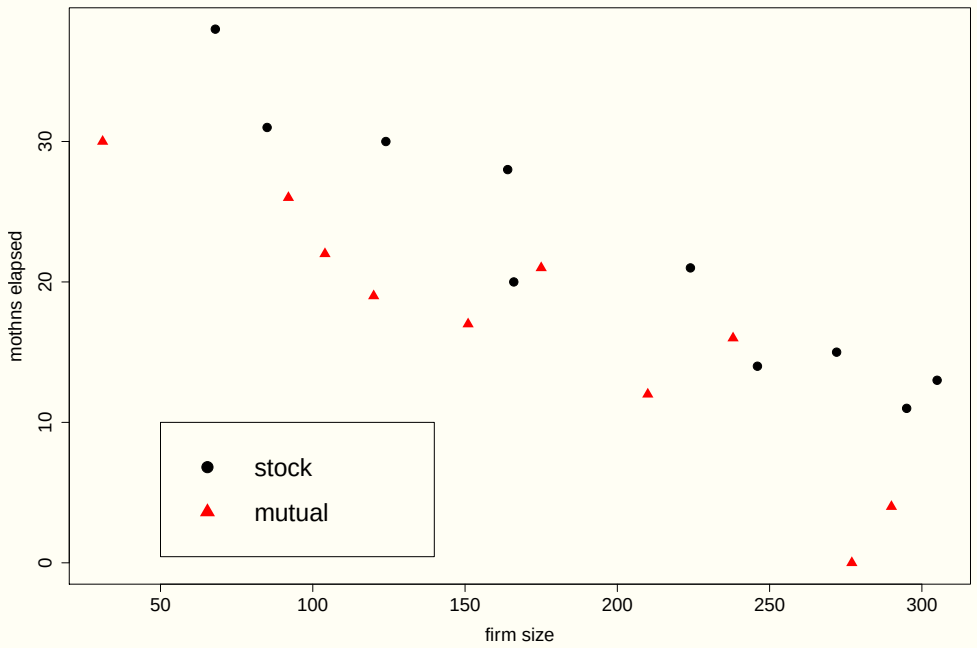


Figure: Scatter plot of months elapsed versus firm size



The slope appears to be similar for the two types of firms, whereas the intercept appears to be different. This means that a first-order model is probably sufficient.

First Order Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20$$

Response function: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

- when $x_2 = 0$ (i.e., mutual firms), the response function:

$$y = \beta_0 + \beta_1 x_1$$

- when $x_2 = 1$ (i.e., stock firms), the response function:

$$y = (\beta_0 + \beta_2) + \beta_1 x_1$$

- Both are straight lines with the same slope β_1 but with intercepts differing by β_2

- β_1 is the common slope of the mean response line under both classes.
- β_0 is the baseline intercept under class 0 (i.e., the reference class).
- β_2 shows how much higher (if positive) or lower (if negative) the mean response line is for class 1 for any given value of X_1.
- The effect of one variable is the same no matter the value of the other variable.

First Order Model: R Output

```
Call:
lm(formula = Y ~ X1 + factor(X2), data = data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     33.874069   1.813858  18.675 9.15e-13 ***
X1              -0.101742   0.008891 -11.443 2.07e-09 ***
factor(X2)stock  8.055469   1.459106   5.521 3.74e-05 ***
---
Residual standard error: 3.221 on 17 degrees of freedom
Multiple R-squared: 0.8951,     Adjusted R-squared: 0.8827
F-statistic:  72.5 on 2 and 17 DF,  p-value: 4.765e-09
```

Stock firms response line: $y = (33.874 + 8.055) - 0.1017x_1$

Mutual firms response line: $y = 33.874 - 0.1017x_1$

Figure: Response lines for stock firms (black) and mutual firms (red)

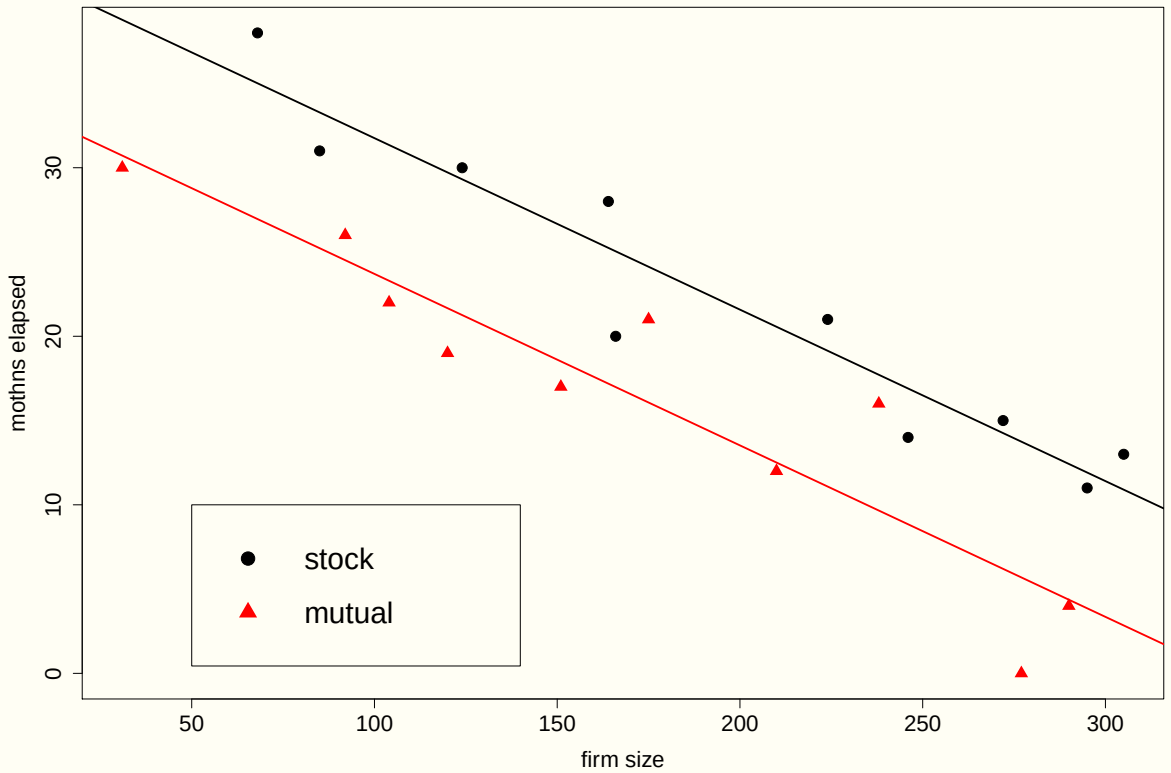
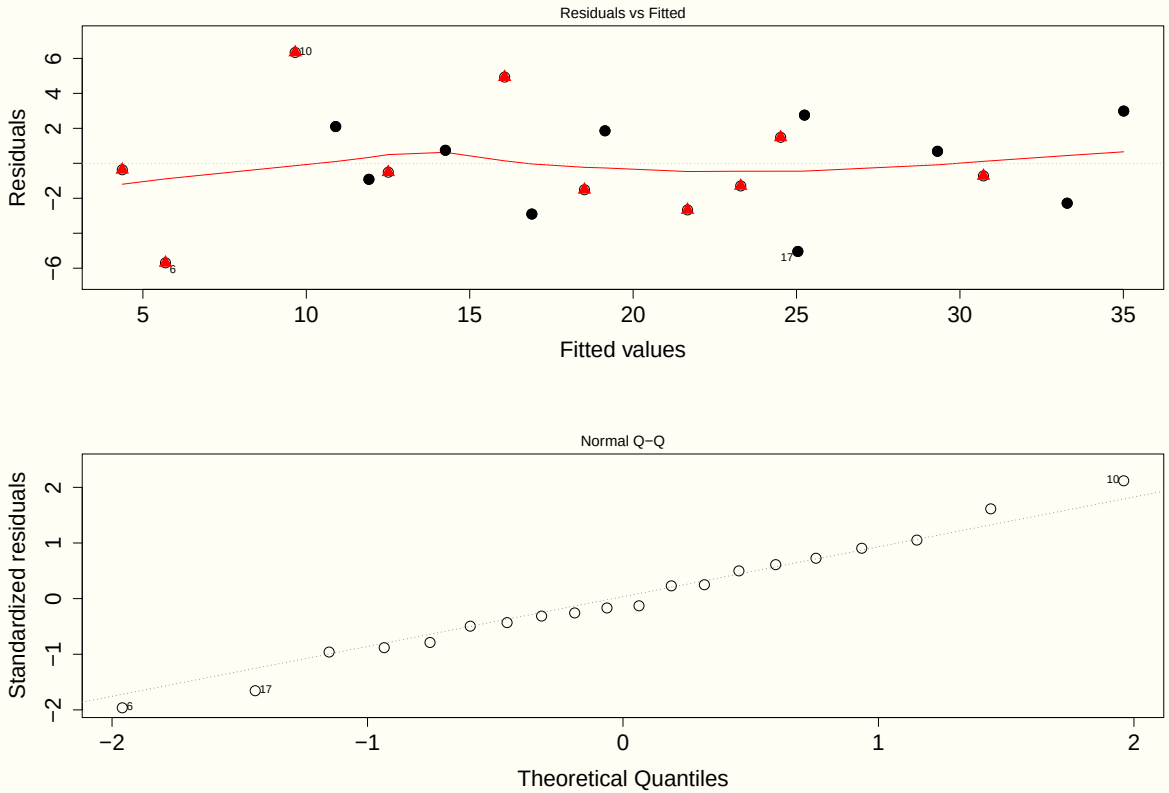


Figure: Residual plots



No obvious violation of the linearity, constant error variance and normal error assumptions

The economist was most interested in the effect of firm type on the speed to adopt an innovation.

- $\hat{\beta}_2 = 8.055$ means that for any given firm size, on average, it takes stock firms 8 more months to adopt an innovation than mutual firms of the same size.
- A 95% confidence interval for β_2:

$$8.055 \pm 2.11 \times 1.459 = [4.98, 11.13].$$

With 95% confidence, we conclude that on average stock firms takes between 5 to 11 more months to adopt an innovation than mutual firms.

Interaction Model with Qualitative and Quantitative Predictors

Interaction Model

Interaction between qualitative and quantitative predictors can be introduced into the model through cross-product term.

- Insurance company:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20$$

Response function: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

- when $x_2 = 0$ (i.e., mutual firms), the response function:

$$y = \beta_0 + \beta_1 x_1$$

  which is a straight line with slope β_1 and intercept β_0.

- when $x_2 = 1$ (i.e., stock firms), the response function:

$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$$

  which is a straight line with slope $\beta_1 + \beta_3$ and intercept $\beta_0 + \beta_2$.

- β_0 and β_1 are baseline intercept and slope, respectively, of the response function for class 0 (i.e., the reference class).
- β_2 indicates how much greater (if positive) or smaller (if negative) is the intercept of the response function for class 1.
- β_3 indicates how much greater (if positive) or smaller (if negative) is the slope of the response function for class 1.
- The effect of one variable depends on the value of the other variable.

Insurance: Interaction Model R Ouput

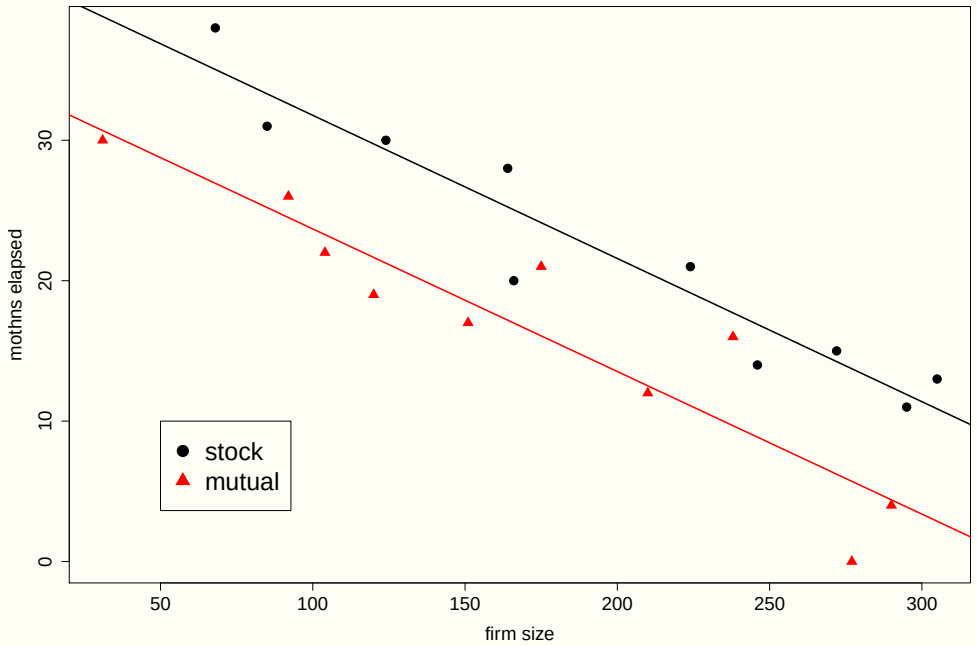
```
Call:
lm(formula = Y ~ X1 + factor(X2) + X1:factor(X2), data = data)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        33.8383695  2.4406498  13.864 2.47e-10 ***
X1                 -0.1015306  0.0130525  -7.779 7.97e-07 ***
factor(X2)stock     8.1312501  3.6540517   2.225   0.0408 *
X1:factor(X2)stock -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
```

Stock firms: $y = (33.838 + 8.131) - (0.1015 + 0.00042)x_1$

Mutual firms: $y = 33.838 - 0.1015x_1$.

Figure: Response lines for stock firms (black) and mutual firms (red)



The two lines are nearly parallel because $\hat{\beta}_3 = -0.00042$ is very small compared to $\hat{\beta}_1 = -0.1015$. Moreover, β_3 is not statistically significant. So the first-order model suffices.

Joint Model vs. Separate Models

Why not simply fit two separate regression models for stock firms and mutual firms?

- Joint Model:
  - stronger assumptions: e.g., constant error variance for both types of firms; same regression slope for firm size under first-order model
  - more efficient through pooling data together: less sampling variability
- Separate Models:
  - less assumptions, more flexible
  - smaller sample size for each model, overall more parameters