# CHAPTER 3

# Multiple Regression

## 3.3 Comparing Two Regression Lines

# 3.3 Comparing Two Regression Lines:

In chapter 1, we consider a simple linear regression model to summarize the relationship between two quantitative variables. Suppose now that we want to investigate whether such a relationship changes between groups determine by some categorical variable.

- Is the relationship between price and mileage different for Porsche offered sale at physical car lots compared to those for sale on the internet.

**Indicator Variable (Dummy Variable, Binary Variable):** An indicator variable uses two values, usually 0 and 1, to indicate whether a data case does (1) or does not (0) belong to a specific category.

Examples:

$$Gender = \begin{cases} 0 & if \quad male \\ 1 & if \quad female \end{cases}$$

$$Surgery = \begin{cases} 0 & if \quad no \\ 1 & if \quad yes \end{cases}$$

# How should indicator variable be interpreted in a regression model?

Assume that, we have two explanatory variables in the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \in$$

where $X_1$ is a quantitative variable and $X_2$ is an indicator variable.

$$X_2 = \begin{cases} 0 \\ 1 \end{cases}$$

- The group $X_2 = 0$ is called the baseline

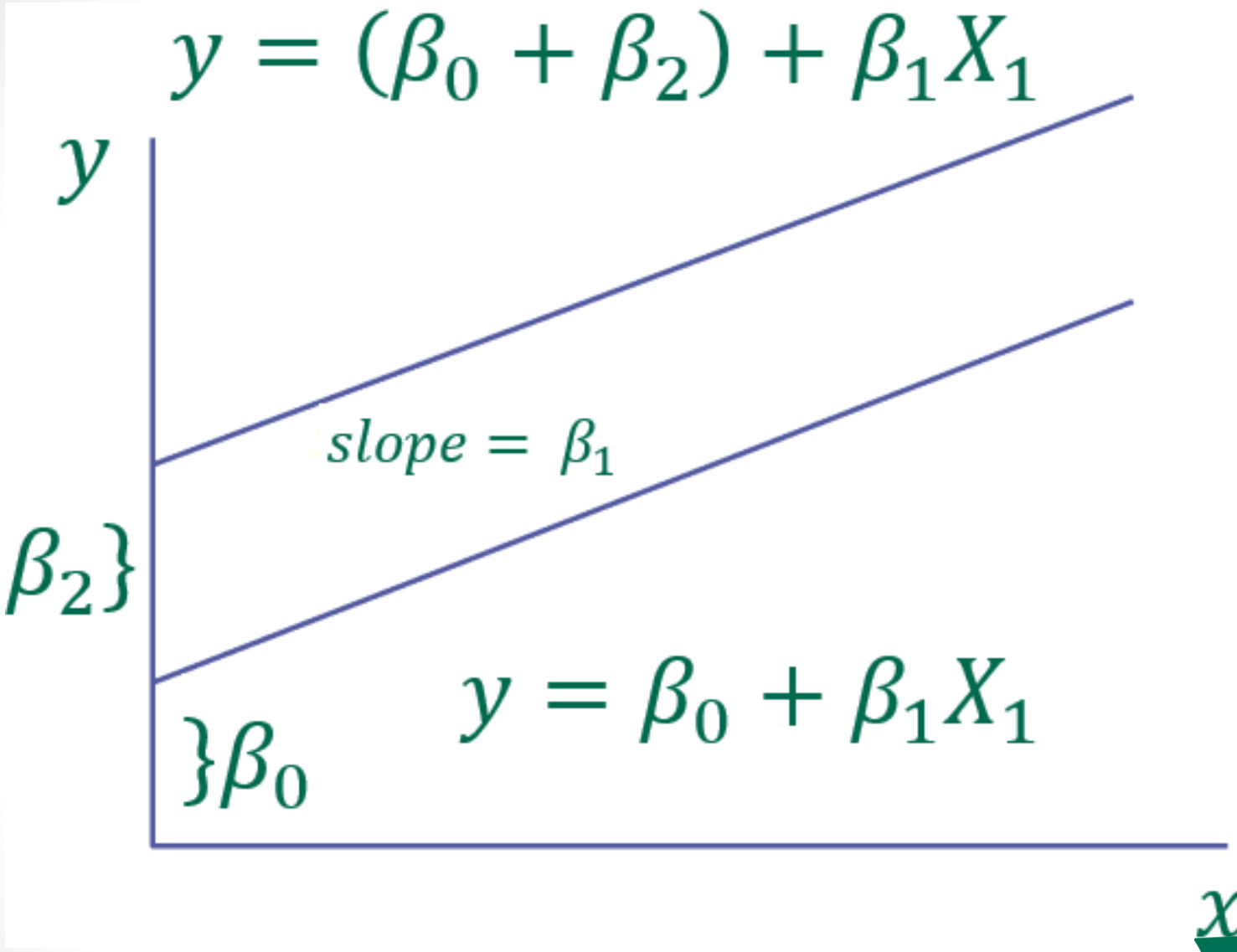
When $X_2 = 0$, the linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(0) + \epsilon$$
$$= \beta_0 + \beta_1 X_1 + \epsilon$$

When $X_2 = 1$, the linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \epsilon$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$
$$= (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$$

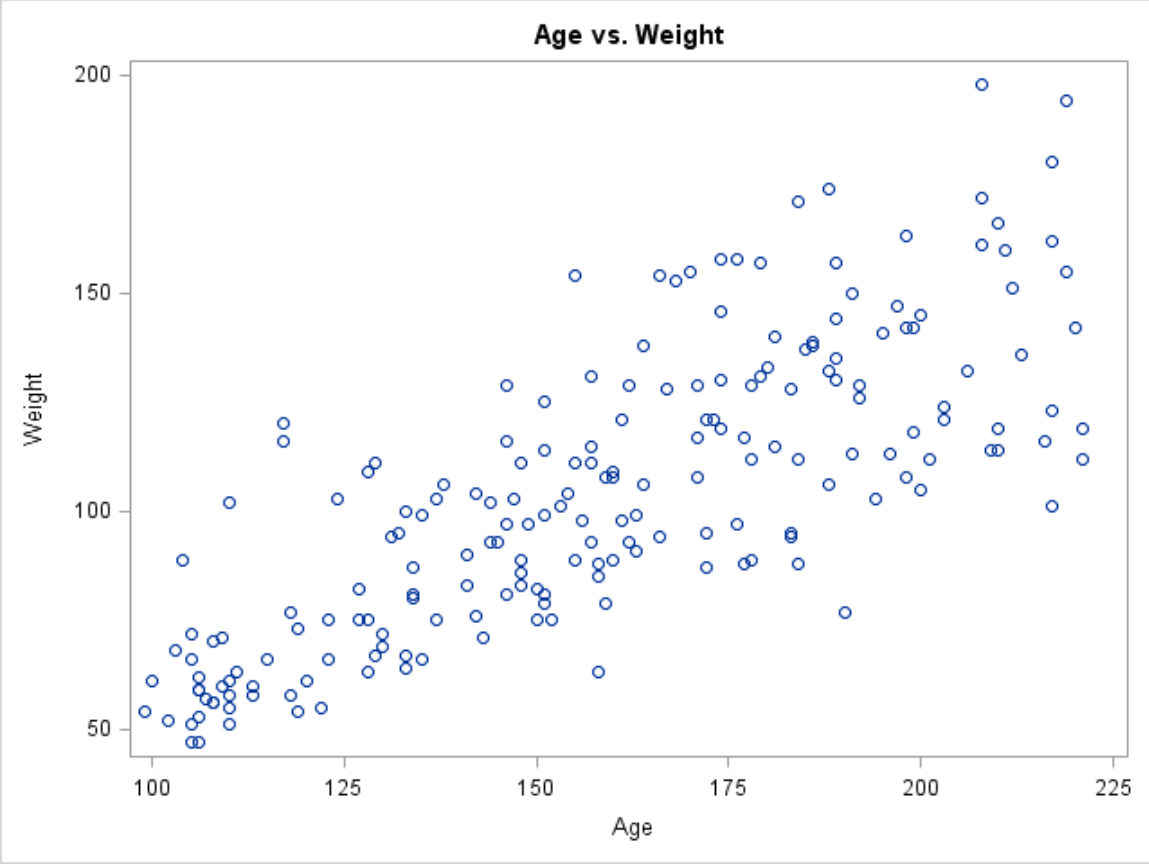
Difference in intercepts


$y = (\beta_0 + \beta_2) + \beta_1 X_1$
$y$
$slope = \beta_1$
$\beta_2\}$
$y = \beta_0 + \beta_1 X_1$
$\}\beta_0$
$x$

# Example 1: *(Kids growth)*

The dataset *Kids198.csv* comes from a 1977 anthropometric study of body measurements for 198 children (8 -18 months), it represents the age (in months), weight (in pounds), height (in inches), sex, and race.

1. Graph the scatterplot of the weight (response) vs. age.



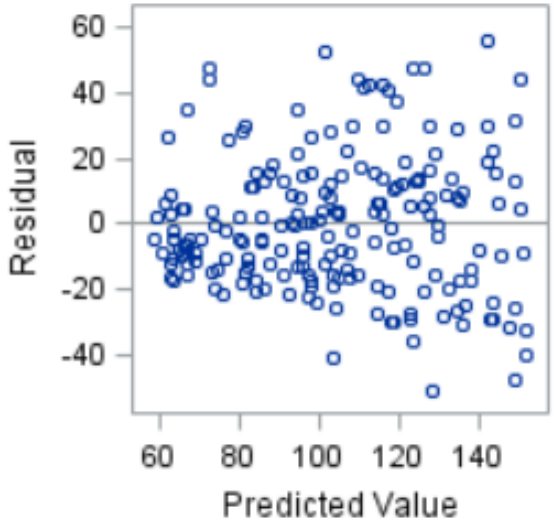
3. Fit the regression model for weight and age.

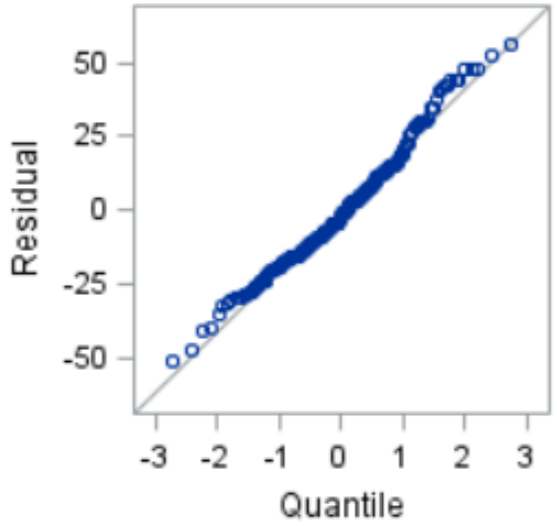
| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -17.01281 | 7.05798 | -2.41 | 0.0169 |
| Age | 1 | 0.76414 | 0.04359 | 17.53 | <.0001 |

| | | | |
|---|---|---|---|
| Root MSE | 20.68077 | R-Square | 0.6106 |
| Dependent Mean | 104.01010 | Adj R-Sq | 0.6086 |
| Coeff Var | 19.88343 | | |

The simple regression model is:

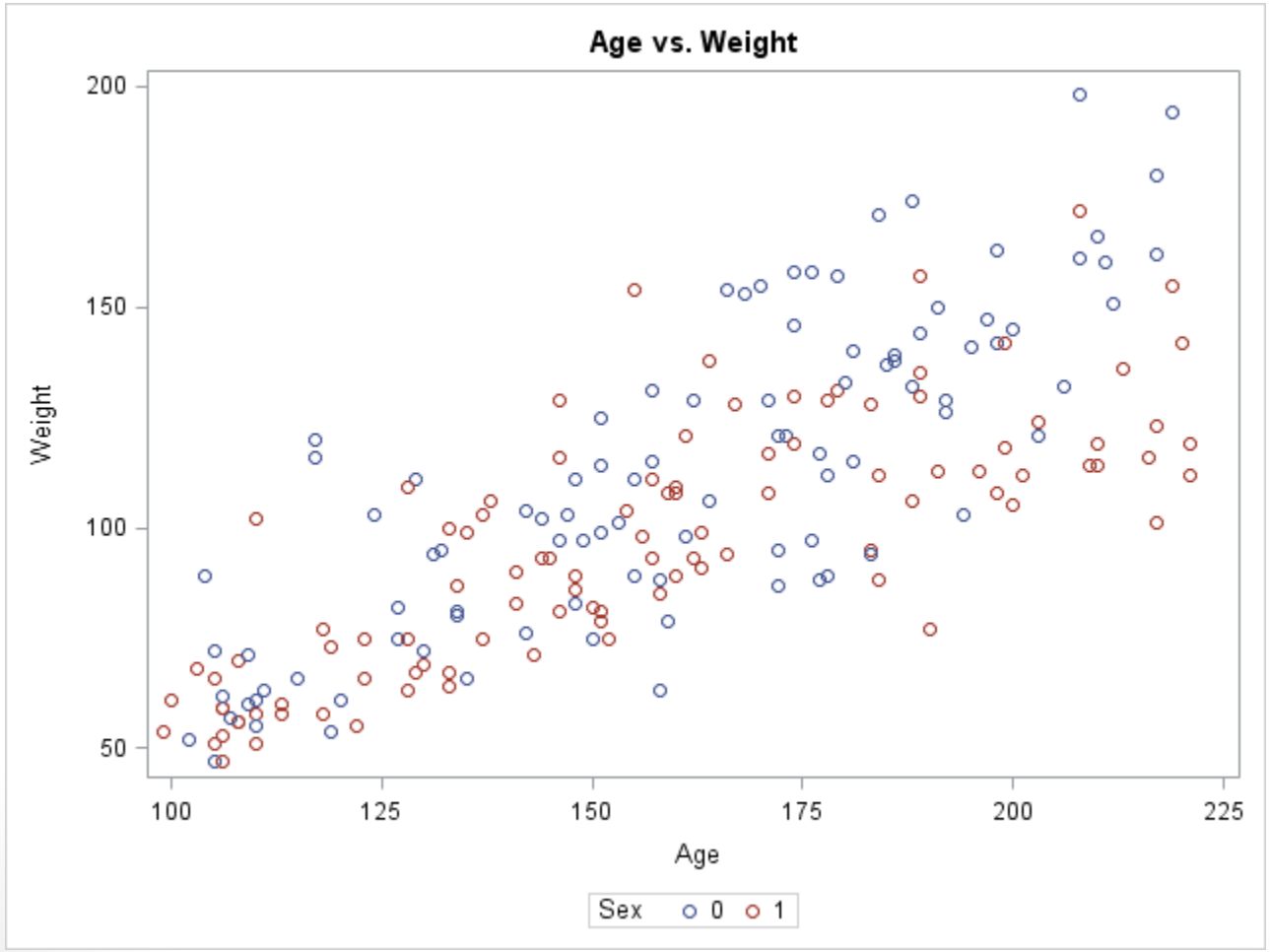
$$\widehat{Weight} = -17.01 + 0.76\,Age$$



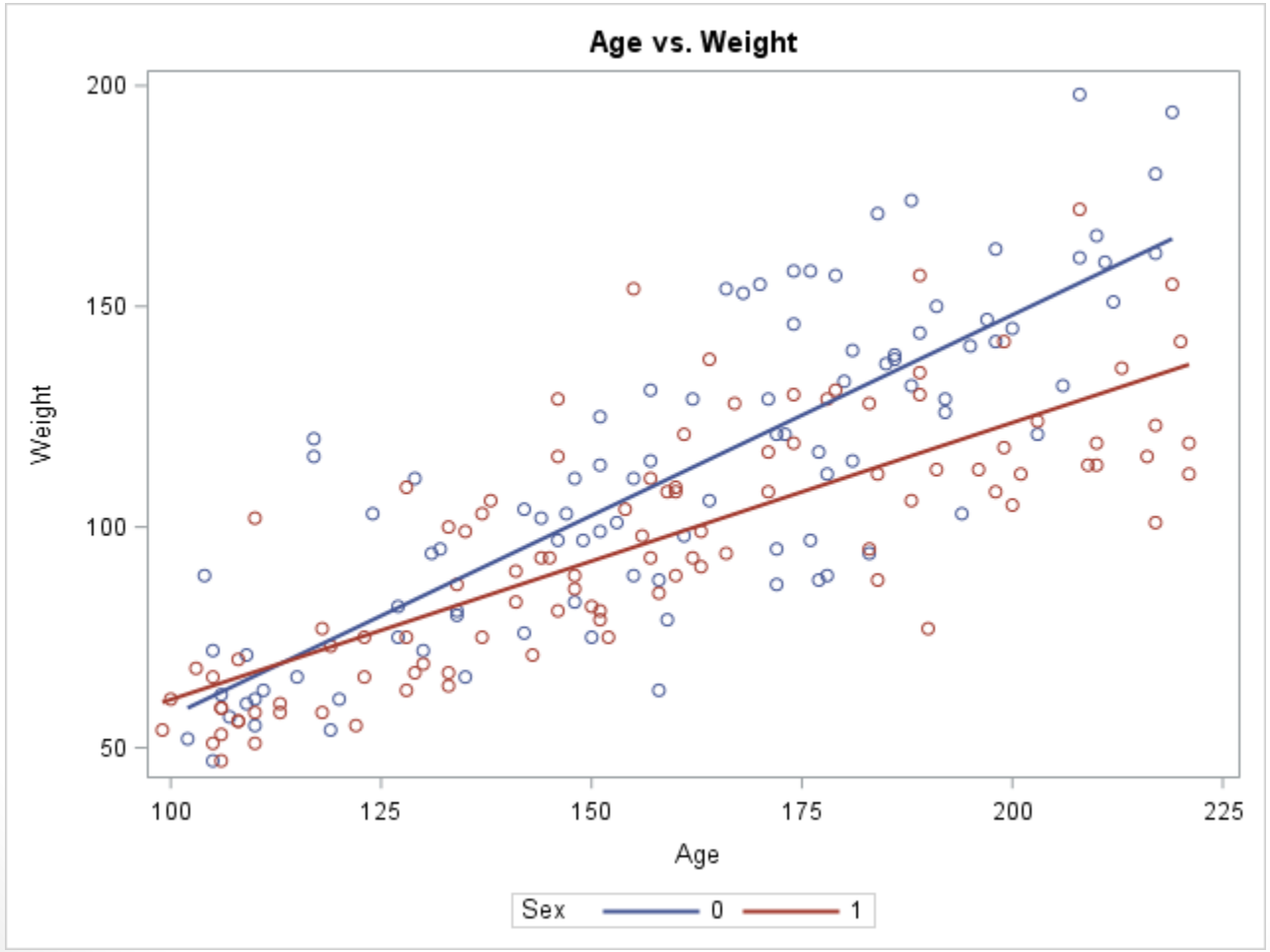
4. Graph the scatterplot of the weight vs. age by gender.

```
proc sgplot data=kids198;
   title 'Age vs. Weight';
   scatter x=age y=weight / group=Sex;
run;
```



5. Graph the scatterplot with the regression line of the weight vs. age for boys (0) and girls (1).

```
proc sgplot data=kids198;
   title 'Age vs. Weight';
   reg x=age y=weight / group=Sex;
run;
```



6. Use “proc sort” to arrange your dataset according to the gender.

**Note:** The new dataset will sort the rows into two groups, males above and females below.

```
proc sort data = kids198 out = kids198_new;
by sex;
run;
```

**Sample of the original dataset:**

| Obs | Height | Weight | Age | Sex | Race |
|---|---|---|---|---|---|
| 1 | 67.8 | 166 | 210 | 0 | 1 |
| 2 | 63 | 93 | 144 | 1 | 0 |
| 3 | 50.1 | 54 | 119 | 0 | 0 |
| 4 | 55.7 | 69 | 130 | 1 | 0 |
| 5 | 63.2 | 115 | 157 | 0 | 0 |

**Sample of the new dataset:**

| Obs | Height | Weight | Age | Sex | Race |
|---|---|---|---|---|---|
| 1 | 67.8 | 166 | 210 | 0 | 1 |
| 2 | 50.1 | 54 | 119 | 0 | 0 |
| 3 | 63.2 | 115 | 157 | 0 | 0 |
| 4 | 48.8 | 52 | 102 | 0 | 0 |
| 5 | 61.3 | 89 | 155 | 0 | 0 |

7. Fit the regression model for weight and age by sex.

i. For boys (0):

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Sex=0

| | |
|---|---|
| Number of Observations Read | 96 |
| Number of Observations Used | 96 |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -33.69254 | 10.87366 | -3.10 | 0.0026 |
| Age | 1 | 0.90871 | 0.06635 | 13.70 | <.0001 |

| | | | |
|---|---|---|---|
| Root MSE | 20.84733 | R-Square | 0.6662 |
| Dependent Mean | 112.35417 | Adj R-Sq | 0.6626 |
| Coeff Var | 18.55501 | | |

The simple linear regression model for boys is:

$$\widehat{Weight} = -33.7 + 0.91\ Age$$

# ii. For girls (1):

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Sex=1

| | |
|---|---|
| Number of Observations Read | 102 |
| Number of Observations Used | 102 |

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -1.84197 | 7.90251 | -0.23 | 0.8162 |
| Age | 1 | 0.62749 | 0.04937 | 12.71 | <.0001 |

| | | | |
|---|---|---|---|
| Root MSE | 17.48150 | R-Square | 0.6176 |
| Dependent Mean | 96.15686 | Adj R-Sq | 0.6138 |
| Coeff Var | 18.18019 | | |

The simple linear regression model for girls is:

$$\widehat{Weight} = -1.84 + 0.62\ Age$$

8. Fit the regression model for weight, age, and sex.

**Note:** First we need to create a new variable which is called the **interaction** between the age and sex.

**Note:** the dataset should be sorted before creating the interaction variable.

```
data kids198_new_1;
set kids198_new;
age_sex = age * sex;
run;
```

| Obs | Height | Weight | Age | Sex | Race | age_sex |
|---|---|---|---|---|---|---|
| 1 | 67.8 | 166 | 210 | 0 | 1 | 0 |
| 2 | 50.1 | 54 | 119 | 0 | 0 | 0 |
| 3 | 63.2 | 115 | 157 | 0 | 0 | 0 |
| 4 | 48.8 | 52 | 102 | 0 | 0 | 0 |
| 5 | 61.3 | 89 | 155 | 0 | 0 | 0 |

8. Fit the regression model for weight, age, and sex.

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | -33.69254 | 10.00727 | -3.37 | 0.0009 |
| Age | 1 | 0.90871 | 0.06106 | 14.88 | <.0001 |
| Sex | 1 | 31.85057 | 13.24269 | 2.41 | 0.0171 |
| age_sex | 1 | -0.28122 | 0.08164 | -3.44 | 0.0007 |

| | | | |
|---|---|---|---|
| Root MSE | 19.18625 | R-Square | 0.6683 |
| Dependent Mean | 104.01010 | Adj R-Sq | 0.6631 |
| Coeff Var | 18.44652 | | |

- All of the predictors variables were significant.

The multiple linear regression (full) model is:

$$\widehat{Weight} = -33.7 + 0.9\,Age + 31.9\,Sex - 0.3\,Age * Sex$$

The multiple linear regression model is:

$$\widehat{Weight} = -33.7 + 0.9\,Age + 31.9\,Sex - 0.3\,Age * Sex$$

Boys model (Sex = 0):

$$\widehat{Weight} = -33.7 + 0.9\,Age + 31.9\,(0) - 0.3\,Age * (0)$$

$$\widehat{Weight} = -33.7 + 0.9\,Age$$

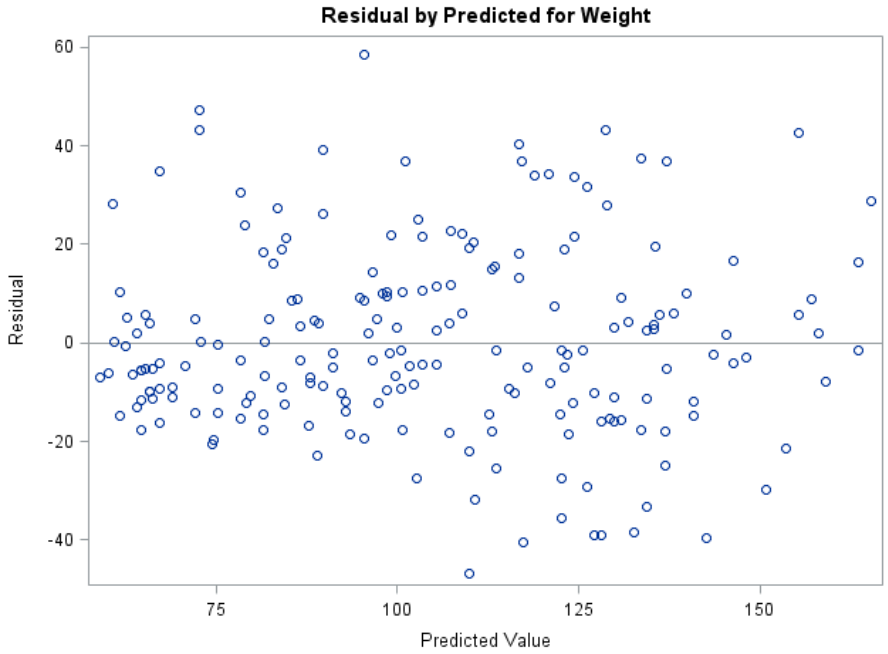
Girls model (Sex = 1):

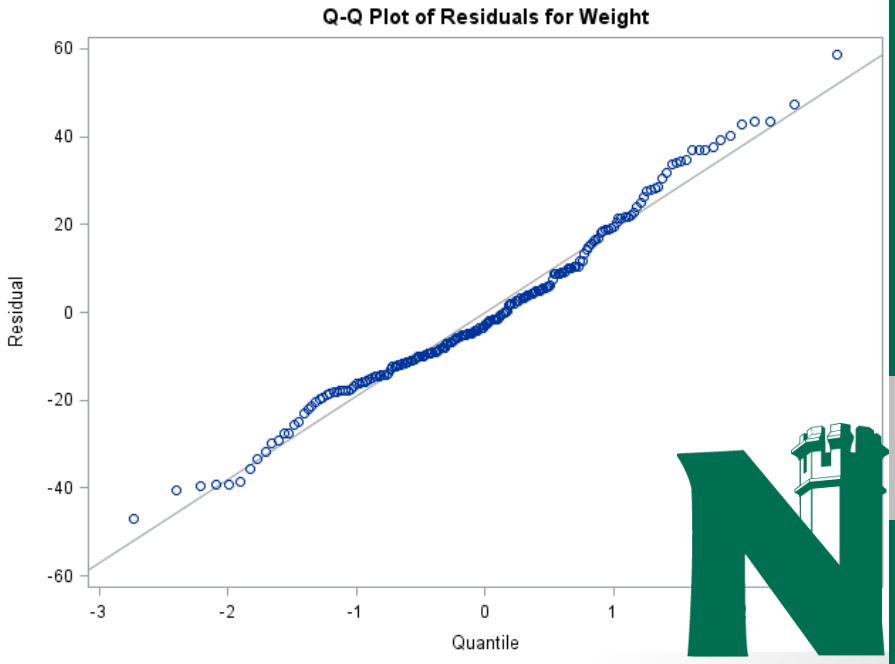
$$\widehat{Weight} = -33.7 + 0.9\,Age + 31.9\,(1) - 0.3\,Age * (1)$$

$$\widehat{Weight} = (-33.7 + 31.9) + (0.9 - 0.3)\,Age$$

$$\widehat{Weight} = -1.8 + 0.6\,Age$$

# 9. Check the regression assumptions.



10. Using the multiple regression model, how much the predicted weight for a boy at age 210 months.

$$\widehat{Weight} = -33.7 + 0.9\ Age + 31.9\ Sex - 0.3\ Age * Sex$$

$$\widehat{Weight}$$

$$= 33.69254 + 0.90871(210) + 31.85057(0)$$

$$- 0.28122(210) * (0)$$

$$\widehat{Weight} = 157.1\ lb$$

| Obs | Height | Weight | Age | Sex | Race | age_sex | Predicted | Residual |
|---|---|---|---|---|---|---|---|---|
| 1 | 67.8 | 166 | 210 | 0 | 1 | 0 | 157.137 | 8.8635 |

11. Using the multiple regression model, how much the predicted weight for a girl at age 144 months.

$$\widehat{Weight} = -33.7 + 0.9\ Age + 31.9\ Sex - 0.3\ Age * Sex$$

$$\widehat{Weight}$$

$$= -33.69254 + 0.90871(144) + 31.85057(1)$$

$$- 0.28122(144) * (1)$$

$$\widehat{Weight} = 88.5\ lb$$

| Obs | Height | Weight | Age | Sex | Race | age_sex | Predicted | Residual |
|---|---|---|---|---|---|---|---|---|
| 97 | 63 | 93 | 144 | 1 | 0 | 144 | 88.516 | 4.4837 |

# Reading Assignment

Read section 3.3