# STAT 302 - Chapter 3 : Multiple Regression - Part 2

Harsha Perera

# Key Topics

- Comparing Two Regression Lines
- Indicator Variables
- Comparing Intercepts
- Comparing Slopes

# Comparing Two Regression Lines

- In Chapter 1, we considered a simple linear regression model to summarize a linear relationship between two quantitative variables.
- Suppose we want to investigate whether such a relationship changes between groups determined by some categorical variable.
- Then we can fit separate linear regression models by considering each categorical group as a different dataset.
- After that we would like to test whether the linear relationship (such as the slope, the intercept, or possibly both) is significantly different between the two groups.
- To make these judgements, we examine multiple regression models that allow us to fit and compare linear relationships for different groups determined by a categorical variable.

# Comparing Two Regression Lines

- Simple Linear Regression: We looked at the relationship between
  - a continuous response
  - a continuous predictor

- Multiple Linear Regression: We looked at the relationship between
  - a continuous response
  - continuous predictors

- Multiple Linear Regression: What if we have
  - a continuous response
  - a combination of continuous and categorical predictors

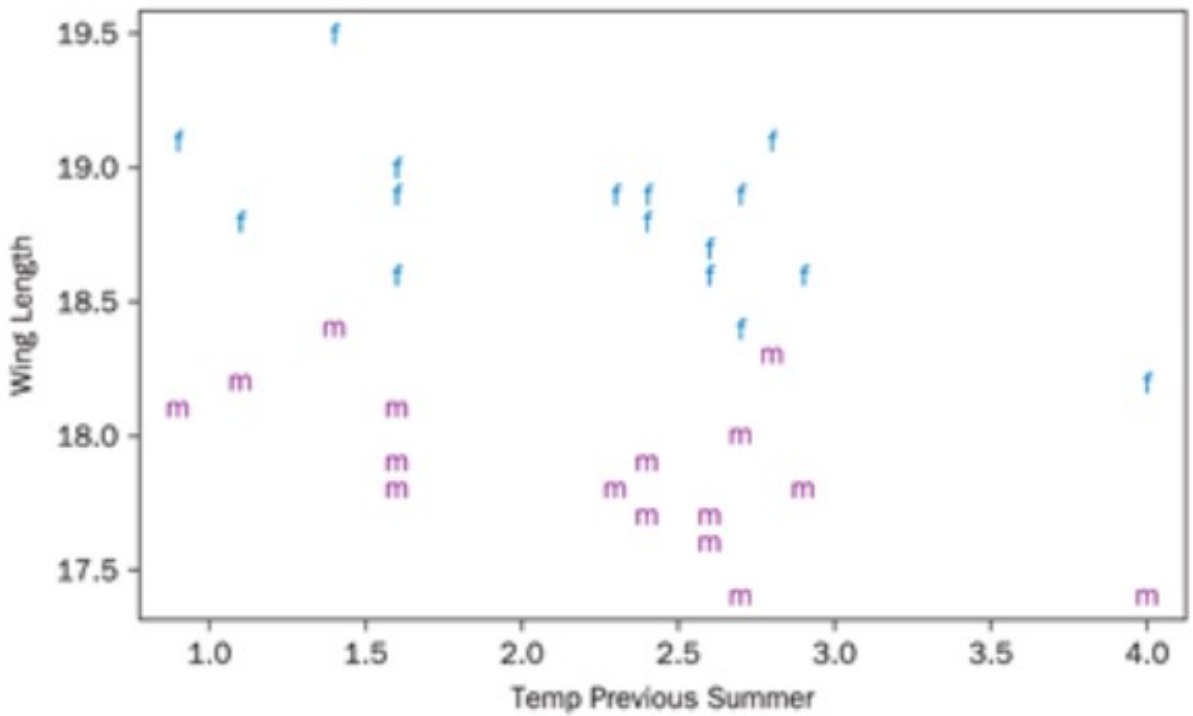
- For the simplicity lets look at a Multiple Linear Regression model with
  - a continuous response
  - a continuous predictor
  - a categorical predictor

# Example 3.9 - Butterfly Size: Females and Males

- In section 2.5 we only looked at the wing length of female butterflies against the summer temperatures.
- In this example we also account for the wing lengths of male butterflies of the same species.
- Question : Is the relationship between temperature and wing length the same for males as for females ?

# Example 3.9 - Butterfly Size: Females and Males

- The scatterplot of wing length versus temperature show the relationship by using different symbols for males and females.
- For each sex there is a reasonably linear trend.



# Example 3.9 - Butterfly Size: Females and Males

- Since the females appear to be quite larger than males, we should not use the same linear model to describe the decreasing trend with temperature for each sex.

- If we consider each sex as its own dataset, we can fit separate linear regression models for each sex :

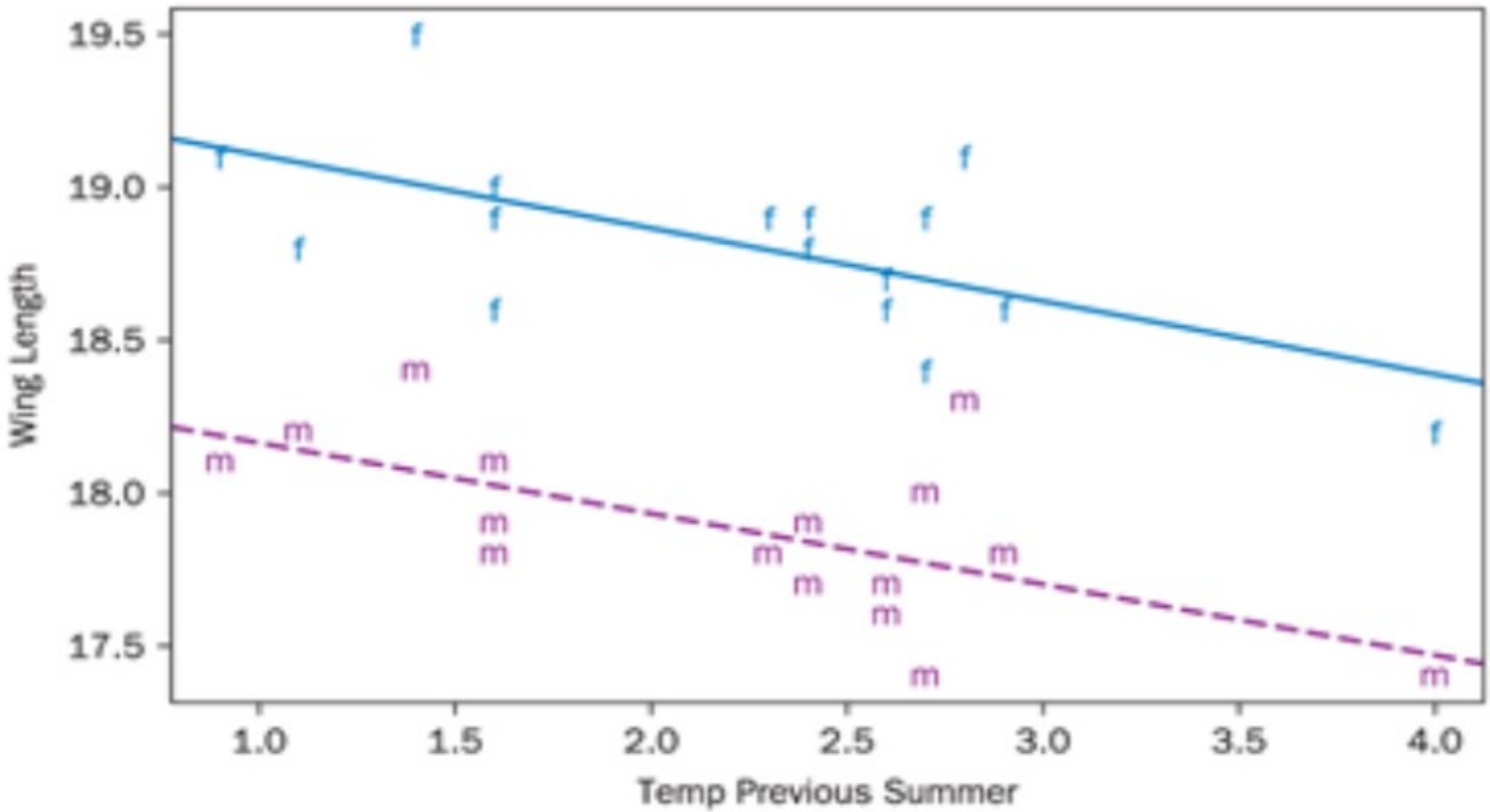
  For Males: $\widehat{Wing} = 18.396 - 0.231 \cdot Temp$
  For Females: $\widehat{Wing} = 19.344 - 0.239 \cdot Temp$

- Separate regression lines for the two sexes have almost similar slopes ( roughly 0.23 mm drop in wing length for every 1 degree increase in temperature) but have different intercepts ( 18.40 for males versus 19.34 for females).

- Therefore the two lines are roughly parallel, the difference in the intercepts gives an indication of how much larger females are than males after taking into account the effect of temperature.

# Example 3.9 - Butterfly Size: Females and Males

Separate regression lines for Male and Female butterflies.



# Indicator Variables

- Using multiple regression, we can examine the Wing vs Temp relationship for both sexes in a single model.
- The key idea here is to use an **indicator variable** that distinguishes between the two groups, in this case male and female.
- We generally use the values 0 and 1 for an indicator variable so that 1 indicates that a data case does belong to a particular group and 0 signifies that the case is not in that group.
- So an indicator for males would be defined as IMale.
- IMale is 1 if Sex = Male and IMale is 0 if Sex = Female

# Indicator Variables

- Therefore in the dataset the new indicator variable IMale takes values :

| Sex | IMale |
|---|---|
| Female | 0 |
| Male | 1 |
| Female | 0 |
| Female | 0 |
| Male | 1 |
| Male | 1 |

- Using the Sex indicator variable IMale, we consider the following multiple regression model :

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot IMale + \in$$

# Indicator Variables

- For data cases from Females (where IMale $= 0$), this model becomes:

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot (0) + \in$$

$$Wing = \beta_0 + \beta_1 \cdot Temp + \in$$

- For data cases from Males (where IMale $= 1$), this model becomes:

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot (1) + \in$$

$$Wing = (\beta_0 + \beta_2) + \beta_1 \cdot Temp + \in$$

# Indicator Variables

- The intercept is adjusted by the amount of $\beta_2$.
- Therefore the coefficient of the IMale indicator variable measures the difference in the intercepts between regression lines for Males and Females that have the same slope.
- If we fit the multiple regression model :

$$\widehat{Wing} = 19.34 - 0.235 \cdot Temp - 0.931 \cdot IMale$$

# Comparing Intercepts

- By substituting the two values (0 and 1) of the indicator variable into the equation, we can obtain a least squares line for each sex :

For Males :

$$\widehat{Wing} = 19.3355 - 0.2350 \cdot Temp - 0.9312 \cdot (1)$$

$$\widehat{Wing} = (19.3355 - 0.9312) - 0.2350 \cdot Temp$$

$$\widehat{Wing} = 18.4043 - 0.2350 \cdot Temp$$

For Females :

$$\widehat{Wing} = 19.3355 - 0.2350 \cdot Temp - 0.9312 \cdot (0)$$

$$\widehat{Wing} = 19.3355 - 0.2350 \cdot Temp$$

# Comparing Intercepts

- In this example we are especially interested in the coefficient ($\beta_2$) of the indicator variable since that reflects the magnitude of the difference in wing length between female and male butterflies.
- From the multiple linear regression equation, we can estimate the average female wing length is 0.93 mm longer than the average male wing length.
- From the R output, the very small p-value gives strong evidence that the observed difference is not due to chance.
- In addition to the estimate and the hypothesis test for the difference in intercepts, we may be interested in producing a confidence interval for the size of the female - male difference.

# Comparing Slopes

Butterfly size data suggests,

- Intercept for males and females are different
- Slopes are similar

Is there a significant difference in slopes? In order to test this claim we need to introduce an additional predictor variable to the model

- Interaction term
- "Temp x IMale"
- Interpretation: change in slope when we move from one category to other

# Comparing Slopes - Butterfly Size Example

The indicator variable: "IMale" (Female=0 / Male=1)

Model:

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot IMale + \beta_3 \cdot Temp \cdot IMale + \in$$

For Male :

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot (1) + \beta_3 \cdot Temp \cdot (1) + \in$$

$$Wing = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot Temp + \in$$

For Female :

$$Wing = \beta_0 + \beta_1 \cdot Temp + \beta_2 \cdot (0) + \beta_3 \cdot Temp \cdot (0) + \in$$

$$Wing = \beta_0 + \beta_1 \cdot Temp + \in$$

# Example 3.10 - Growth Rates of Kids

Children tend to get bigger as they get older, but we might be interested in how growth rates compare. Do boys and girls gain weight at the same rates?

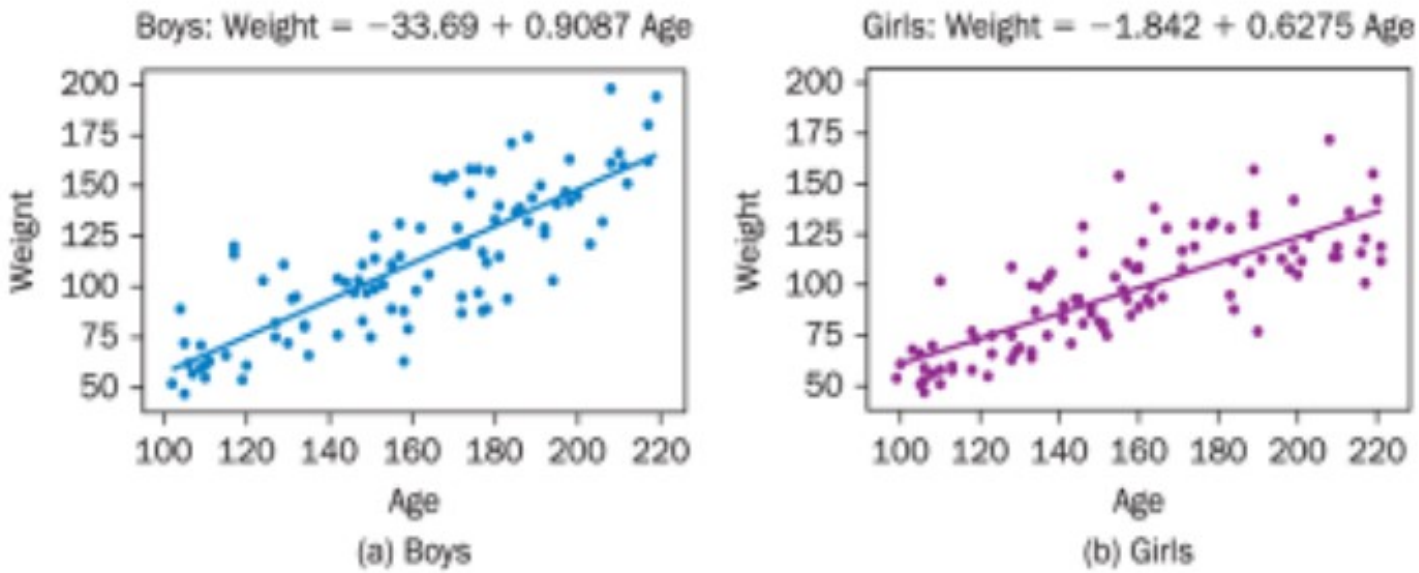



Figure shows separate plots for boys and girls with the regression line. The line is slightly steeper for the boys ( 0.91 pound per month) than it is for girls(0.63 pound per month).

# Example 3.10 - Growth Rates of Kids

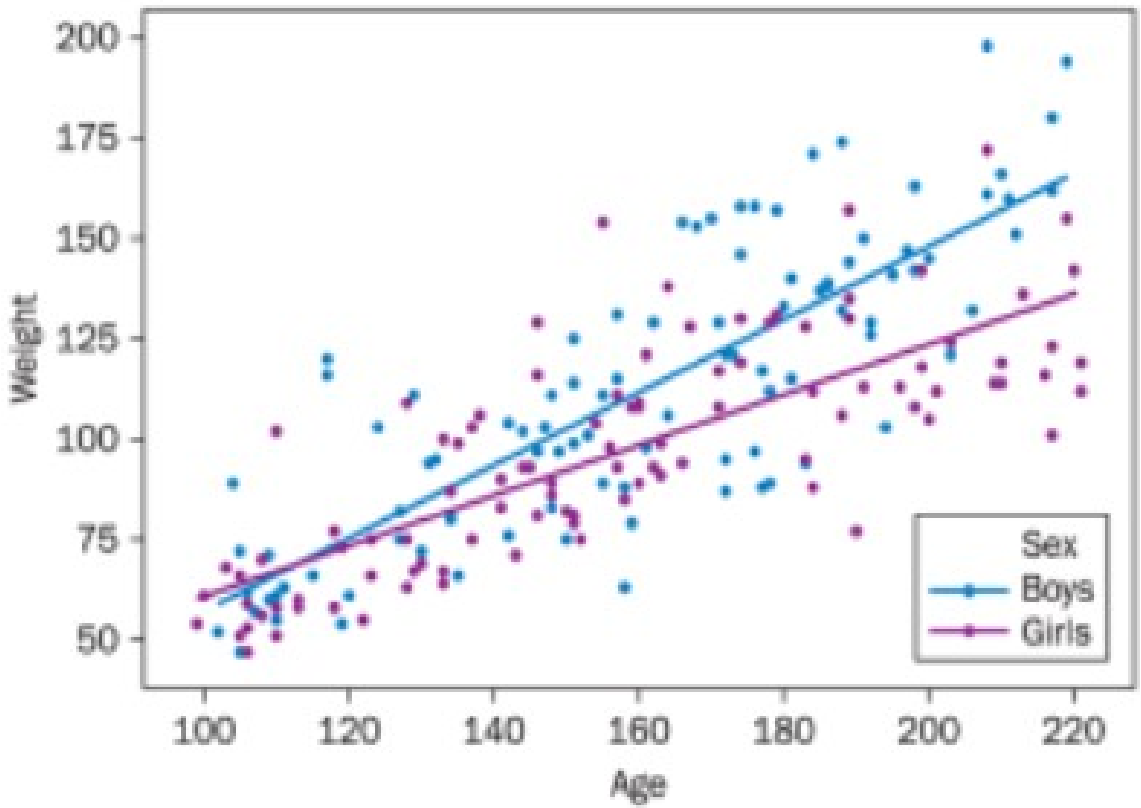



Figure above shows both regression lines on the same plot. Does the difference in slopes for these two samples indicate that the typical growth rate is really larger for boys than it is for girls or could a difference this large occur due to random chance ?

# Example 3.10 - Growth Rates of Kids

- Define the indicator variable IGirl.
- IGirl to be 0 for the boys and 1 for the girls.
- Adding IGirl indicator variable allows for different intercepts for the two groups.
- The new predictor we add now is the product of IGirl and the Age ( Age*IGirl).
- This gives the model :

$$Weight = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot IGirl + \beta_3 \cdot Age \cdot IGirl + \in$$

# Example 3.10 - Growth Rates of Kids

- For the boys in the study (IGirl $= 0$), the model becomes:

$$Weight = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot (0) + \beta_3 \cdot Age \cdot (0) + \in$$

$$Weight = \beta_0 + \beta_1 \cdot Age + \in$$

- For girls (IGirl $= 1$) :

$$Weight = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot (1) + \beta_3 \cdot Age \cdot (1) + \in$$

$$Weight = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot Age + \in$$

The new coefficient $\beta_3$, shows how much the slopes change as we move from the regression line for boys to the line for girls.

# Example 3.10 - Growth Rates of Kids

If we fit the multiple linear regression model :

$$\widehat{Weight} = -33.6925 + 0.9087 \cdot Age + 31.8506 \cdot IGirl - 0.2812 \cdot Age \cdot IGirl$$

Boys :

$$\widehat{Weight} = -33.6925 + 0.9087 \cdot Age$$

Girls :

$$\widehat{Weight} = (31.8506 - 33.6925) + (0.9087 - 0.2812) \cdot Age$$

$$\widehat{Weight} = -1.8419 + 0.6275 \cdot Age$$

# Example 3.10 - Growth Rates of Kids

- The multiple regression model allows us to fit both regressions in the same model and provides a parameter $\beta_3$, that specifically measures the difference in the slopes between the two groups.
- Hypotheses: $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$
- From the R output, the small p-value provides very strong evidence that the coefficient is different from zero and thus that the growth rates are different for boys and girls.
- Moreover we can construct a confidence interval for the $\beta_3$ as well.