College of
EDUCATION

REVIEW OF DESCRIPTIVE STATISTICS AND CONCEPTUALIZATIONS OF VARIANCE

EDF 9780: Multivariate Educational Research

Lecture #1

Learning Objectives

Univariate descriptive statistics

- Central tendency: Mean, median, mode
- Variation/spread: Standard deviation, variance, range

Bivariate descriptive statistics

- Correlation
- Covariance

Types of variable distributions:

- Marginal
- Joint
- Conditional

Bias in estimators

Data for Today's Lecture

To help demonstrate the concepts of today's lecture, we will be using a data set with three variables:

Female (Sex): Male (=0) or Female (=1)

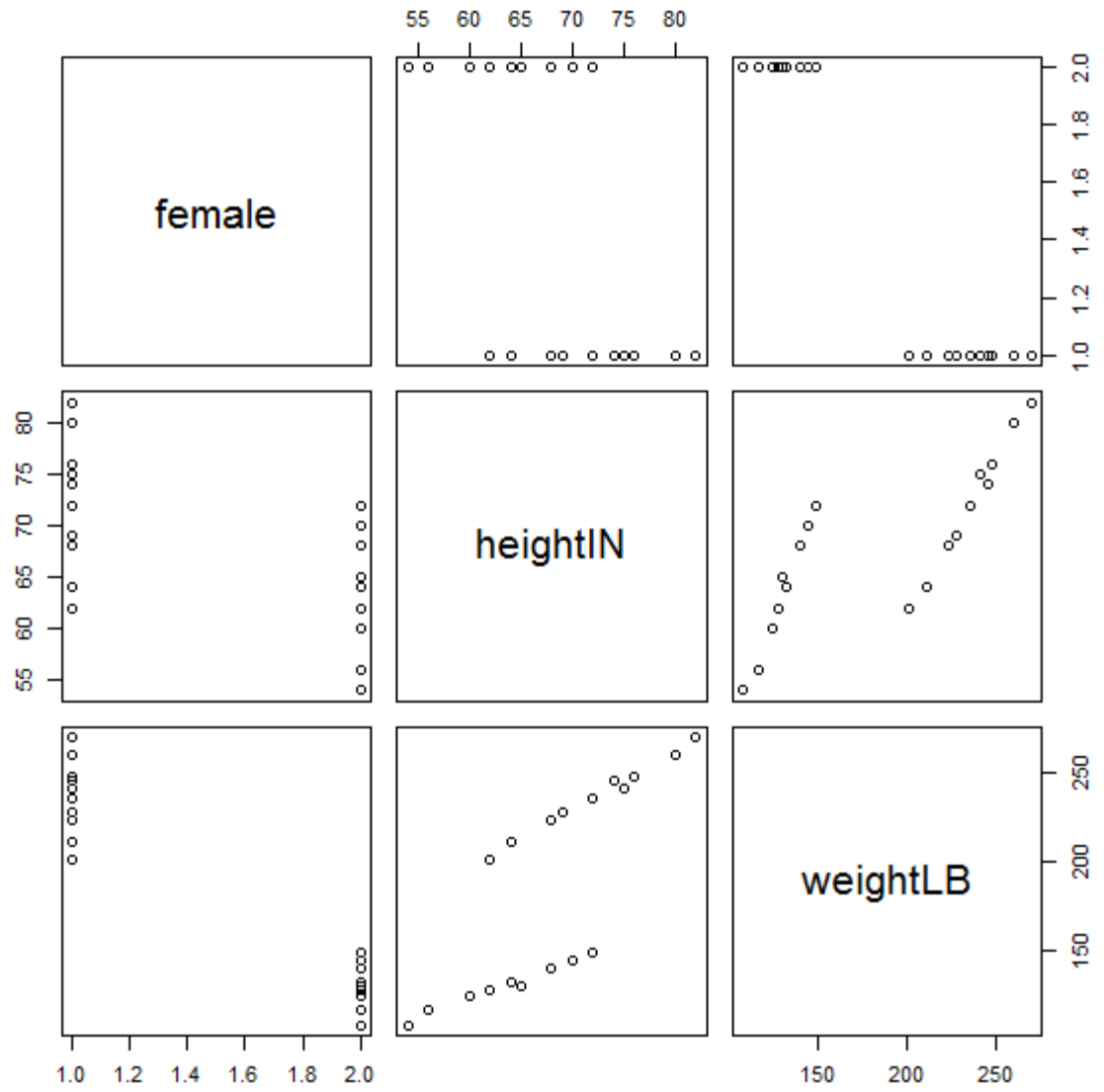
Height in inches

Weight in pounds

The end point of our lecture will be to build a **linear model** that predicts a person's weight

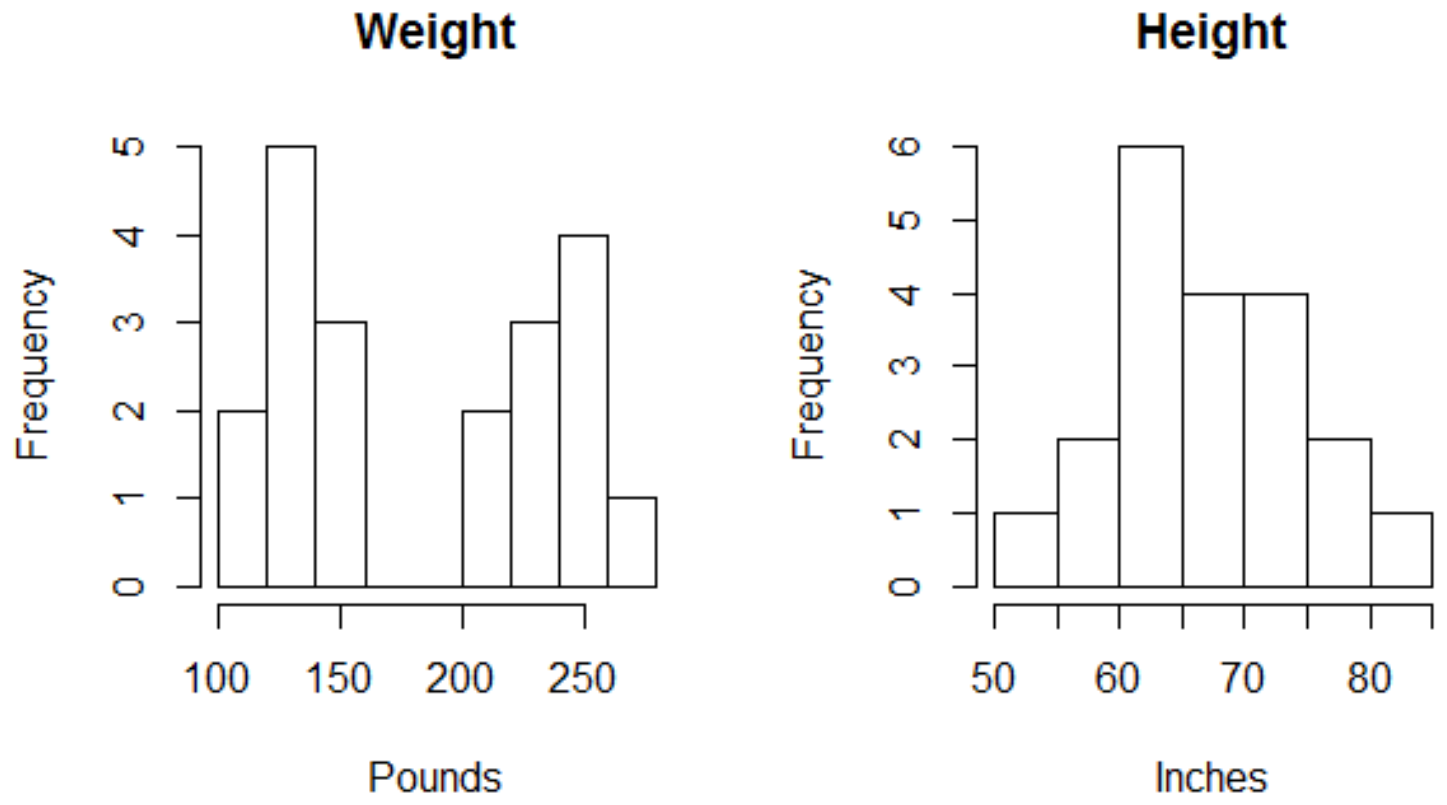
Linear model: a statistical model for an outcome that uses a linear combination (a weighted sum) of one or more predictor variables to produce an estimate of an observation's predicted value

What you will learn is that models underlie all statistics


female
heightIN
weightLB
55
60
65
70
75
80
1.0
1.2
1.4
1.6
1.8
2.0
150
200
250

Histograms of Height and Weight

The weight variable seems to be bimodal – does that worry you?



Descriptive Statistics

We can summarize each variable **marginally** through a set of descriptive statistics

Marginal: one variable by itself

Common marginal descriptive statistics:

Central tendency: *Mean*, Median, Mode

Variability: *Standard deviation (variance)*, range

We can also summarize the **joint** (bivariate) **distribution** of two variables through a set of descriptive statistics:

Joint distribution: more than one variable simultaneously

Common bivariate descriptive statistics:

Correlation and covariance

Variable	Mean	SD	Variance
Height	67.9	7.44	55.358
Weight	183.4	56.383	3,179.095
Female	0.5	0.513	0.263

Above Diagonal: Covariance

Diagonal: Variance

Correlation /Covariance	Height	Weight	Female
Height	55.358	334.832	-2.263
Weight	.798	3,179.095	-27.632
Female	-.593	-.955	.263

Below Diagonal: Correlation

Re-examining the Concept of Variance

Variability is a central concept in advanced statistics

In multivariate statistics, covariance is also central

Two formulas for the variance
(about the same when N is large):

Unbiased or "sample"

$$S_{Y_1}^2 = \frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)^2$$

Biased/ML or "population"

$$S_{Y_1}^2 = \frac{1}{N}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)^2$$

Here: p = person; 1 = variable number one

Interpretation of Variance

The variance describes the spread of a variable in squared units (which come from the $\left(Y_{1p} - \bar{Y}_1\right)^2$ term in the equation)

Variance: **the average ***squared*** distance of an observation from the mean**

- Variance of Height: 55.358 inches squared
- Variance of Weight: 3,179.095 pounds squared
- Variance of Female – not applicable in the same way!

Because squared units are difficult to work with, we typically use the standard deviation – which is reported in units

Standard deviation: **the average distance of an observation from the mean**

- SD of Height: 7.44 inches
- SD of Weight: 56.383 pounds

Variance/SD as a More General Statistical Concept

Variance (and the standard deviation) is a concept that is applied across statistics – not just for data

Statistical parameters have variance

e.g. The sample mean $\bar{Y}_1$ has a "standard error" (SE) of $S_{\bar{Y}} = \frac{S_Y}{\sqrt{N}}$

The standard error is another name for standard deviation

So "standard error of the mean" is equivalent to "standard deviation of the mean"

Usually "error" refers to parameters; "deviation" refers to data

Variance of the mean would be $S^2_{\bar{Y}} = \frac{S^2_Y}{N}$

More generally, variance = error

You can think about the SE of the mean as telling you how far off the mean is for describing the data

Correlation of Variables

Moving from marginal summaries of each variable to joint (bivariate) summaries, the Pearson correlation is often used to describe the association between a pair of variables:

$$r_{Y_1,Y_2} = \frac{\frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)}{S_{Y_1}S_{Y_2}}$$

The correlation is **unitless** as it ranges from -1 to 1 for continuous variables, regardless of their variances

- Pearson correlation of binary/categorical variables with continuous variables is called a point-biserial (same formula)
- Pearson correlation of binary/categorical variables with other binary/categorical variables has bounds within -1 and 1

More on the Correlation Coefficient

The Pearson correlation is a **biased** estimator

Biased estimator: the expected value differs from the true value for a statistic

Other biased estimators: Variance/SD when $\frac{1}{N}$ is used

The unbiased correlation estimate would be:

$$r^{U}_{Y_1,Y_2} = r_{Y_1,Y_2}\left[1 + \frac{\left(1 - r^2_{Y_1,Y_2}\right)}{2N}\right]$$

As N gets large bias goes away; Bias is largest when $r_{Y_1,Y_2} = 0$

Pearson is an underestimate of true correlation

If it is biased, then why does everyone use it anyway?

Answer: forthcoming when we talk about (ML) estimation

Covariance of Variables: Association with Units

The numerator of the correlation coefficient is the covariance of a pair of variables:

$$S_{Y_1,Y_2} = \frac{1}{N-1}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)$$

Unbiased or "sample"

$$S_{Y_1,Y_2} = \frac{1}{N}\sum_{p=1}^{N}\left(Y_{1p} - \bar{Y}_1\right)\left(Y_{2p} - \bar{Y}_2\right)$$

Biased/ML or "population"

The covariance uses the units of the original variables (but now they are multiples):

- Covariance of height and weight: 334.832 inch-pounds

The covariance of a variable with itself is the variance

The covariance is often used in multivariate analyses because it ties directly into multivariate distributions

- But…covariance and correlation are easy to switch between

Going from Covariance to Correlation

If you have the covariance matrix (variances and covariances):

$$r_{Y_1,Y_2} = \frac{S_{Y_1,Y_2}}{S_{Y_1} S_{Y_2}}$$

If you have the correlation matrix and the standard deviations:

$$S_{Y_1,Y_2} = r_{Y_1,Y_2} S_{Y_1} S_{Y_2}$$

REVIEW OF LINEAR MODELS

Learning Objectives

Types of distributions:

- Conditional distributions

The General Linear Model

- Regression
- Analysis of Variance (ANOVA)
- Analysis of Covariance (ANCOVA)
- Beyond – Interactions

The General Linear Model

The general linear model incorporates many different labels of analyses under one unifying umbrella:

	Categorical Xs	Continuous Xs	Both Types of Xs
Univariate Y	ANOVA	Regression	ANCOVA
Multivariate Ys	MANOVA	Multivariate Regression	MANCOVA

The typical assumption is that error is normally distributed – meaning that the data are **conditionally** normally distributed

Models for non-normal outcomes (e.g., dichotomous, categorical, count) fall under the *Generalized* Linear Model, of which the GLM is a special case (i.e., for when model residuals can be assumed to be normally distributed)

General Linear Models: Conditional Normality

$$Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p$$

Model for the Means (Predicted Values):

- Each person's expected (predicted) outcome is a function of his/her values on x and z (and their interaction)
- y, x, and z are each measured only once per person (*p* subscript)

Model for the Variance:

- $e_p \sim N(0, \sigma_e^2)$ → **ONE** residual (unexplained) deviation
- e_p has a mean of 0 with some estimated constant variance σ_e^2, is normally distributed, is unrelated to x and z, and is unrelated across people (across all observations, just people here)

We will return to the normal distribution in a few weeks – but for now know that it is described by two terms: a mean and a variance

Building a Linear Model for Predicting a Person's Weight

We will now build a linear model for predicting a person's weight, using height and sex as predictors

Several models we will build are done for didactic reasons – to show how regression and ANOVA work with the GLM

You wouldn't necessarily run these models in this sequence

Our beginning model is that of an **empty model** – no predictors for weight (an **unconditional model**)

Our ending model is one with both predictors and their interaction (a **conditional model**)

Model 1: The Empty Model

Linear model: $Weight_p = \beta_0 + e_p$

where $e_p \sim N(0, \sigma_e^2)$

Estimated Parameters: [ESTIMATE (STANDARD ERROR)]

- $\beta_0 = 183.4\ (12.61)$
  - Overall intercept – the "grand" mean of weight across all people
  - Just the mean of weight
  - SE for β_0 is standard error of the mean for weight $\frac{S_{Weight}}{\sqrt{N}}$
- $\sigma_e^2 = 3{,}179.1$ (SE not given)
  - The (unbiased) variance of weight:

$$e_p = Weight_p - \beta_0 = Weight_p - \overline{Weight}_p$$

$$S_e^2 = \frac{1}{\mathrm{N} - 1} \sum_{p=1}^{N} \left(Weight_p - \overline{Weight}_p\right)^2$$

Note: Classically, this comes from the Mean Square Error of an F-table (but may not in models estimated with different methods)

Model 2: Predicting Weight from Height ("Regression")

Linear model: $Weight_p = \beta_0 + \beta_1 Height_p + e_p$

where $e_p \sim N(0, \sigma_e^2)$

Estimated Parameters:

- $\beta_0 = -227.292\ (73.483)$
  - Predicted value of Weight for a person with Height = 0
  - Nonsensical – but we could have centered Height
- $\beta_1 = 6.048\ (1.076)$
  - Change in predicted value of Weight for every one-unit increase in height (weight goes up 6.048 pounds per inch)
- $\sigma_e^2 = 1{,}218$ (SE not given)
  - The residual variance of weight
  - Height explains $\frac{3{,}179.1 - 1{,}218}{3{,}179.1} = 61.7\%$ of variance of weight

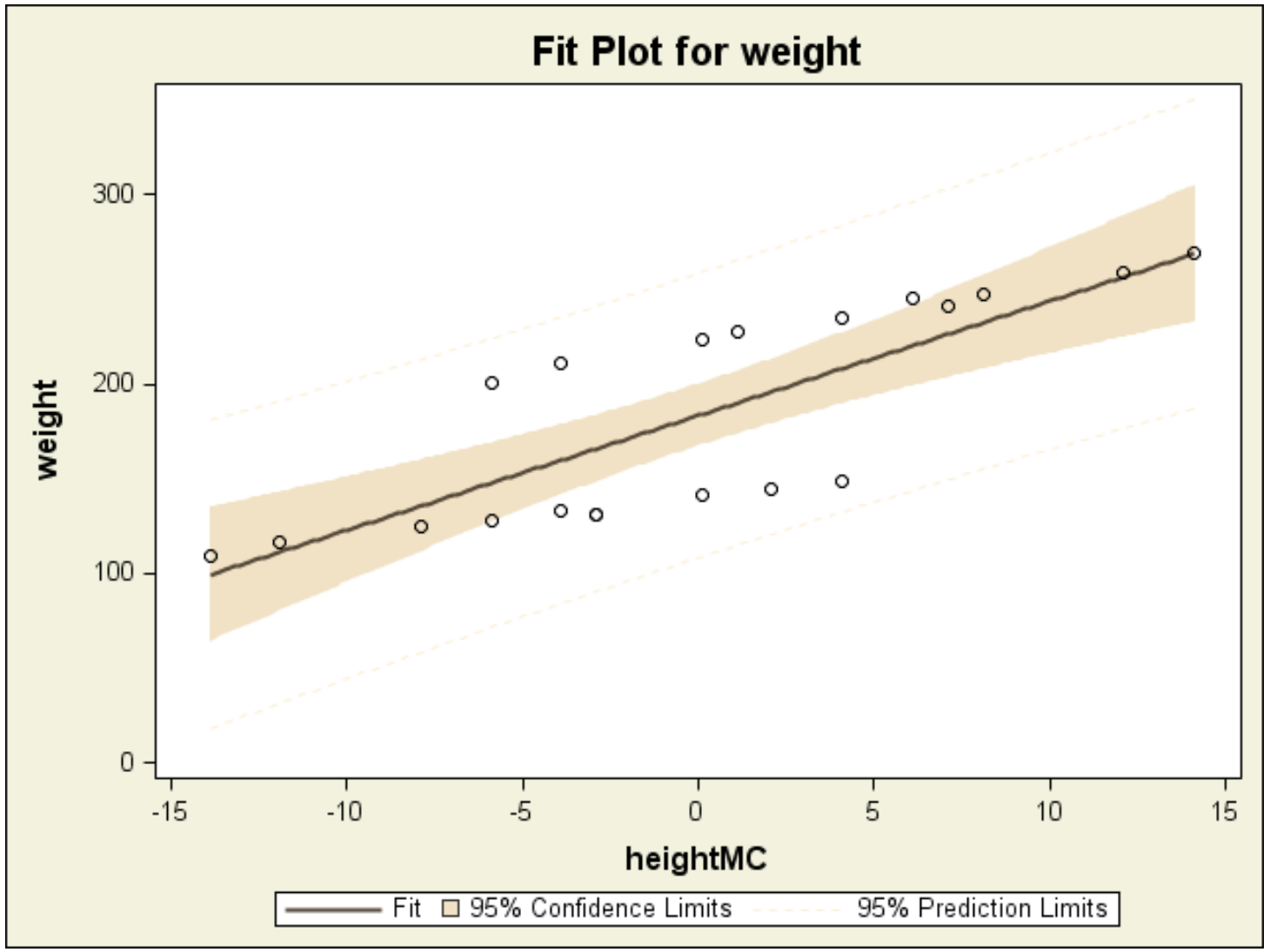
Model 2a: Predicting Weight from Mean-Centered Height

Linear model: $W_p = \beta_0 + \beta_1\left(H_p - \bar{H}\right) + e_p$

where $e_p \sim N(0, \sigma_e^2)$

Estimated Parameters: [ESTIMATE (STANDARD ERROR)]

- $\beta_0 = 183.4\ (7.804)$
  - Predicted value of Weight for a person with Height = Mean Height
  - Is the Mean Weight (regression line goes through means)
- $\beta_1 = 6.048\ (1.076)$
  - Change in predicted value of Weight for every one-unit increase in height (weight goes up 6.048 pounds per inch)
  - Same as previous
- $\sigma_e^2 = 1{,}218$ (SE not given)
  - The residual variance of weight
  - Height explains $\frac{3{,}179.1 - 1{,}218}{3{,}179.1} = 61.7\%$ of variance of weight
  - Same as previous


Fit Plot for weight
weight
300
200
100
0
-15
-10
-5
0
5
10
15
heightMC
Fit
95% Confidence Limits
95% Prediction Limits

Hypothesis Tests for Parameters

To determine if the regression slope is significantly different from zero, we must use a hypothesis test:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

We have two options for this test (both are same here)

- Use ANOVA table: sums of squares – F-test
- Use "Wald" test for parameter: $t = \frac{\beta_1}{se(\beta_1)}$
- Here $t^2 = F$

Wald test: $t = \frac{\beta_1}{se(\beta_1)} = \frac{6.048}{1.076} = 5.62; p < .001$

Conclusion: reject null (H_0); slope is significant

Model 3: Predicting Weight from Sex ("ANOVA")

Linear Model: $Weight_p = \beta_0 + \beta_2 Female_p + e_p$

where $e_p \sim N(0, \sigma_e^2)$

Note: because sex is a categorical predictor, we must first code it into a number before entering it into the model (typically done automatically in software)

Here we use Female = 1 for females; Female = 0 for males

Estimated Parameters: [ESTIMATE (STANDARD ERROR)]

- $\beta_0 = 235.9\ (5.415)$
  - Predicted value of Weight for a person with Female=0 (males)
  - Mean weight of males
- $\beta_2 = -105.0\ (7.658)$

$$t = -\frac{105}{7.658} = -13.71; p < .001$$

  - Change in predicted value of Weight for every one unit increase in female
  - In this case, the difference between the mean for males and the mean for females
- $\sigma_e^2 = 293$ (SE not given)
  - The residual variance of weight
  - Sex explains $\frac{3{,}179.1-293}{3{,}179.1} = 90.8\%$ of variance of weight

Model 3: More on Categorical Predictors

Sex was coded using what is called reference or dummy coding:

- Intercept becomes mean of the "reference" group (the 0 group)
- Slopes become the difference in the means between reference and other groups
- For C categories, C-1 predictors are created

All coding choices can be recovered from the model:

Predicted Weight for Females (mean weight for females):

$$W_p = \beta_0 + \beta_2 = 235.9 - 105 = 130.9$$

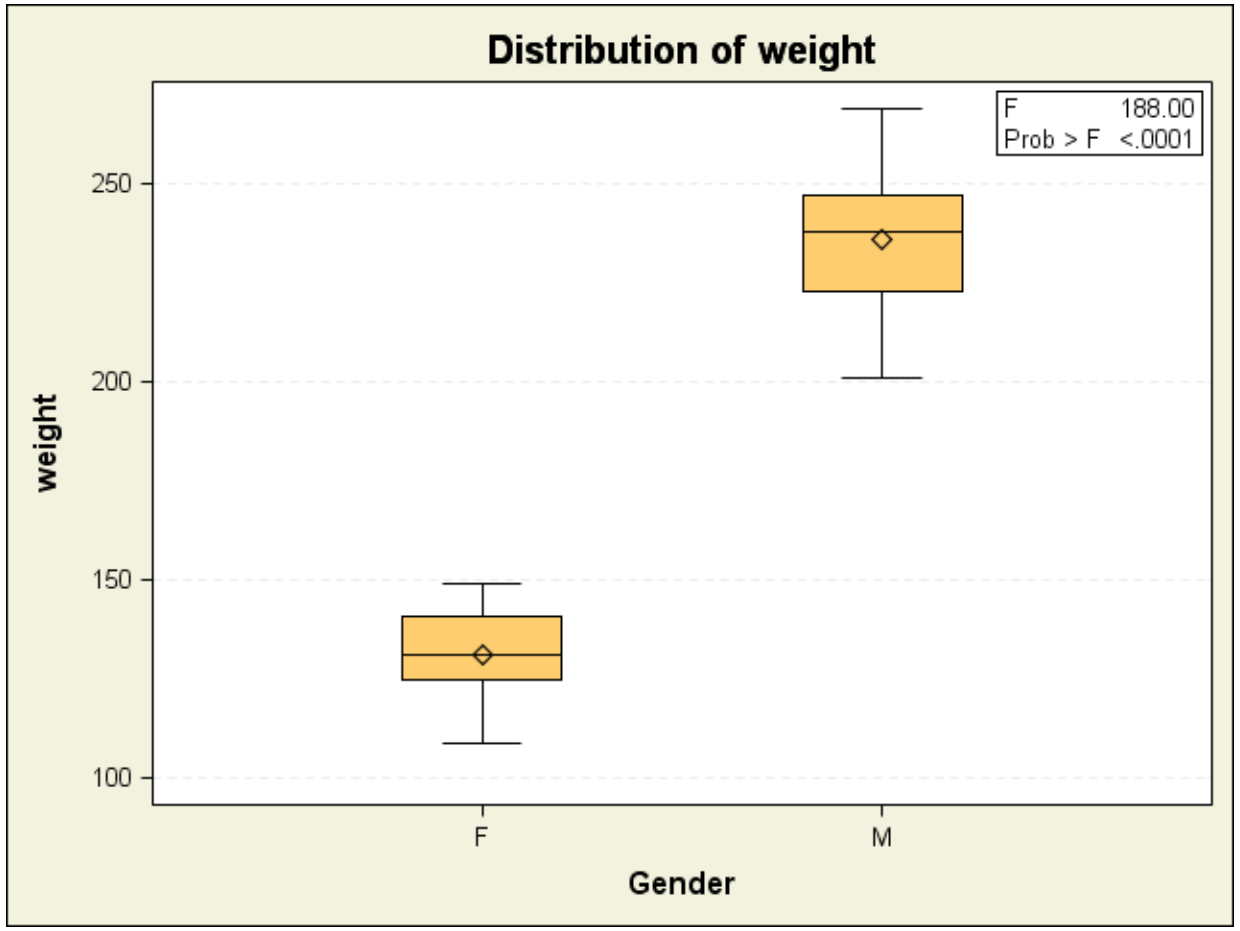
Predicted Weight for Males:

$$W_p = \beta_0 = 235.9$$

What would β_0 and β_2 be if we coded Male = 1?

Super cool idea: what if you could do this in software all at once?

Model 3: Predictions and Plots



Model 4: Predicting Weight from Height and Sex (w/o Interaction); ("ANCOVA")

Linear Model: $W_p = \beta_0 + \beta_1\left(H_p - \bar{H}\right) + \beta_2 F_p + e_p$

where $e_p \sim N(0, \sigma_e^2)$

Estimated Parameters: [ESTIMATE (STANDARD ERROR)]

- $\beta_0 = 224.256\ (1.439)$
  - Predicted value of Weight for a person with Female=0 (males) and has Height = Mean Height $\left(H_p - \bar{H}\right) = 0$
- $\beta_1 = 2.708\ (0.155)$

$$t = \frac{2.708}{0.155} = 17.52; p < .001$$

  - Change in predicted value of Weight for every one-unit increase in height (holding sex constant)

$$\beta_2 = -81.712\ (2.241)$$

$$t = -\frac{81.712}{2.241} = -36.46; p < .001$$

  - Change in predicted value of Weight for every one-unit increase in female (holding height constant)
  - In this case, the difference between the mean for males and the mean for females holding height constant
- $\sigma_e^2 = 16$ (SE not given)
  - The residual variance of weight

Model 4: By-Sex Regression Lines

Model 4 assumes identical regression slopes for both sexs but has different intercepts

This assumption is tested statistically by model 5

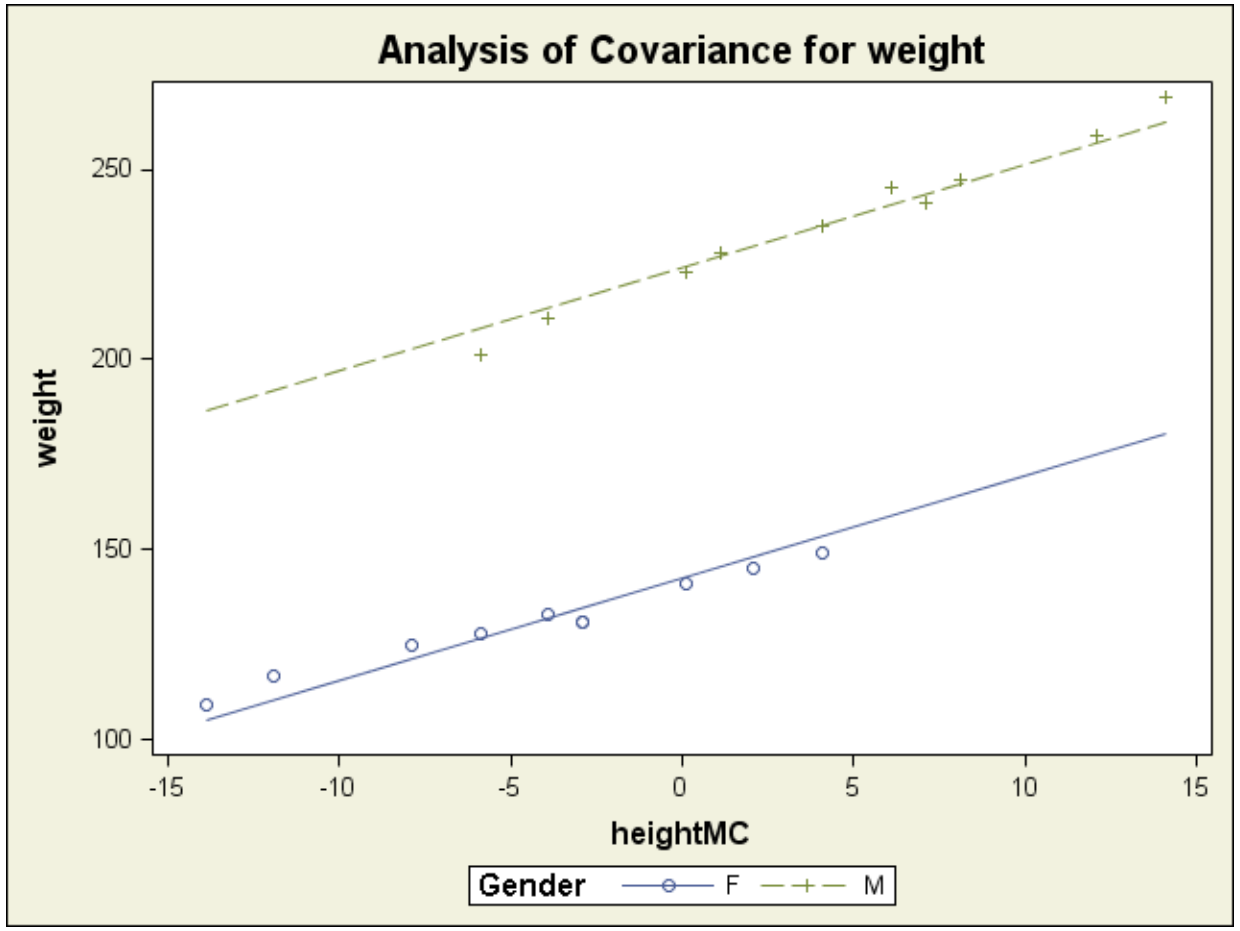
Predicted Weight for Females:

$$W_p = 224.256 + 2.708\left(H_p - \bar{H}\right) - 81.712F_p = 142.544 + 2.708\left(H_p - \bar{H}\right)$$

Predicted Weight for Males:

$$W_p = 224.256 + 2.708\left(H_p - \bar{H}\right) - 81.712F_p = 224.256 + 2.708\left(H_p - \bar{H}\right)$$

Model 4: Predicted Value Regression Lines



Model 5: Predicting Weight from Height and Sex (with Interaction); ("ANCOVAish")

Linear Model:

$$W_p = \beta_0 + \beta_1\left(H_p - \bar{H}\right) + \beta_2 F_p + \beta_3\left(H_p - \bar{H}\right)F_p + e_p$$

where $e_p \sim N(0, \sigma_e^2)$

Estimated Parameters: [ESTIMATE (STANDARD ERROR)]

- $\beta_0 = 222.184\ (0.838)$
  - Predicted value of Weight for a person with Female=0 (males) and has Height = Mean Height $\left(H_p - \bar{H}\right) = 0$

$$\beta_1 = 3.190\ (0.111)$$

$$t = \frac{3.190}{0.111} = 28.65; p < .001$$

- **Simple main effect of height**: Change in predicted value of Weight for every one-unit increase in height (for males only)
  - A conditional main effect: when interacting variable (sex) = 0

Model 5: Estimated Parameters

Estimated Parameters:

- $\beta_2 = -82.272\ (1.211)$

$$t = -\frac{82.272}{1.211} = -67.93; p < .001$$

  - **Simple main effect of sex:** Change in predicted value of Weight for every one unit increase in female, for height = mean height
  - Sex difference at 67.9 inches

- $\beta_3 = -1.094\ (0.168)$

$$t = -\frac{1.094}{0.168} = -6.52; p < .001$$

  - **Sex-by-Height Interaction**: Additional change in predicted value of weight for change in either sex or height
  - Difference in slope for height for females vs. males
  - Because Female = 1, it modifies the slope for height for females (here the height slope is *less positive* than for females than for males)

- $\sigma_e^2 = 5$ (SE not given)

Model 5: By-Sex Regression Lines

Model 5 does not assume identical regression slopes

Because β_3 was significantly different from zero, the data supports different slopes

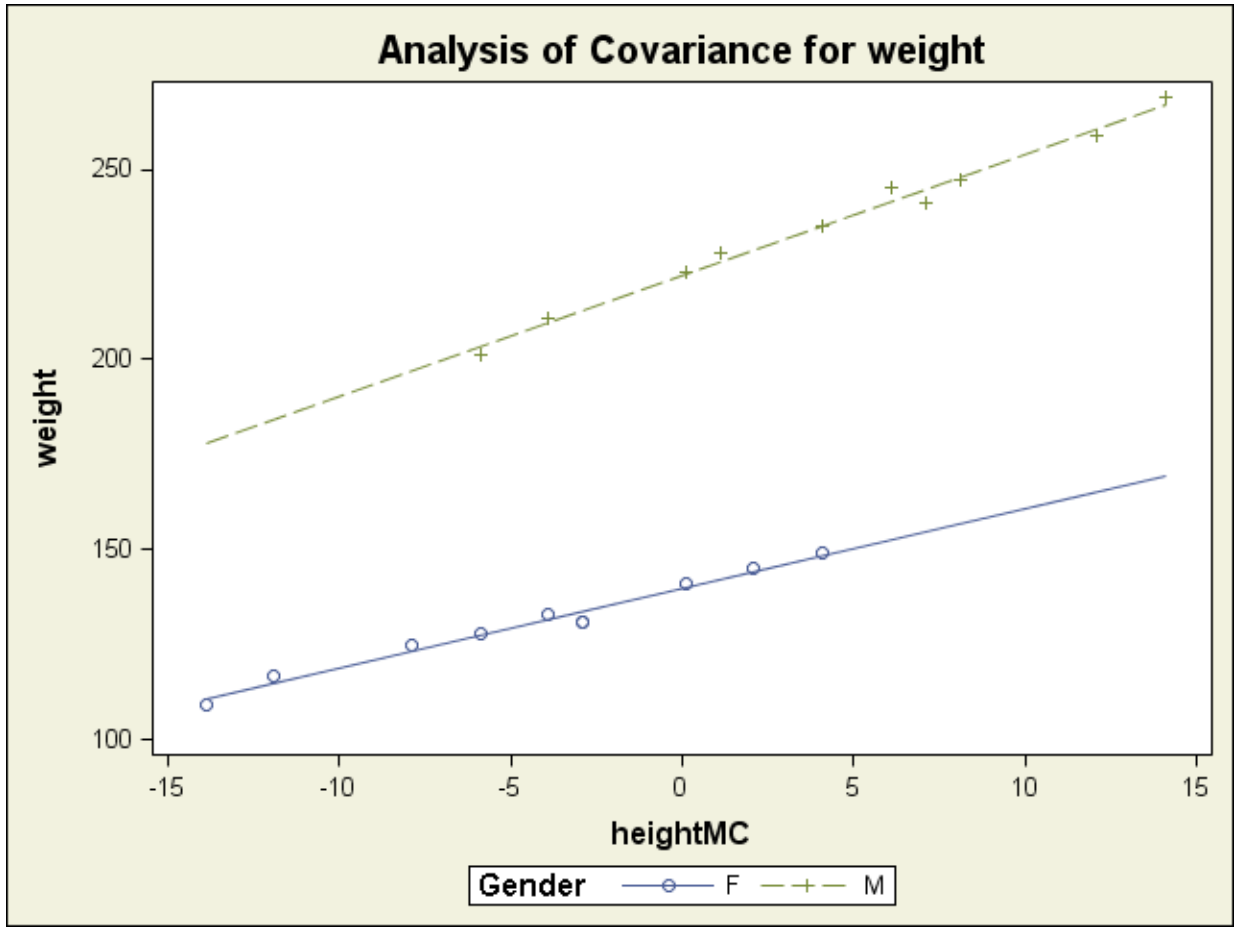
Predicted Weight for Females:

$$W_p = 222.184 + 3.190(H_p - \bar{H}) - 82.272F_p - 1.094(H_p - \bar{H})F_p$$
$$= 139.912 + 2.096(H_p - \bar{H})$$

Predicted Weight for Males:

$$W_p = 222.184 + 3.190(H_p - \bar{H}) - 82.272F_p - 1.094(H_p - \bar{H})F_p$$
$$= 222.184 + 3.190(H_p - \bar{H})$$

Model 5: Predicted Value Regression Lines



Comparing Across Models

Typically, the empty model and model #5 would be the only models run

The trick is to describe the impact of all and each of the predictors – typically using variance accounted for (explained)

All predictors:

Baseline: empty model #1; $\sigma_e^2 = 3{,}179.095$

Comparison: model #5; $\sigma_e^2 = 4.731$

All predictors (sex, height, interaction)explained $\frac{3{,}179.095 - 4.731}{3{,}179.095} = 99.9\%$ of variance in weight

R^2 hall of fame worthy

Comparing Across Models

The total effect of height (main effect and interaction):

Baseline: model #3 (sex only); $\sigma_e^2 = 293.211$

Comparison: model #5 (all predictors); $\sigma_e^2 = 4.731$

Height explained $\frac{293.211-4.731}{293.211} = 98.4\%$ of variance in weight *remaining after sex*

98.4% of the 100-90.8% = 9.2% left after sex

True variance accounted for is 98.4%*9.2% = 9.1%

The total effect of sex (main effect and interaction):

Baseline: model #2a (height only); $\sigma_e^2 = 1{,}217.973$

Comparison: model #5 (all predictors); $\sigma_e^2 = 4.731$

Sex explained $\frac{1{,}217.973-4.731}{1{,}217.973} = 99.6\%$ of variance in weight *remaining after height*

99.6% of the 100-61.7% = 38.3% left after height

True variance accounted for is 99.6%*38.3% = 38.1%

About Weight…

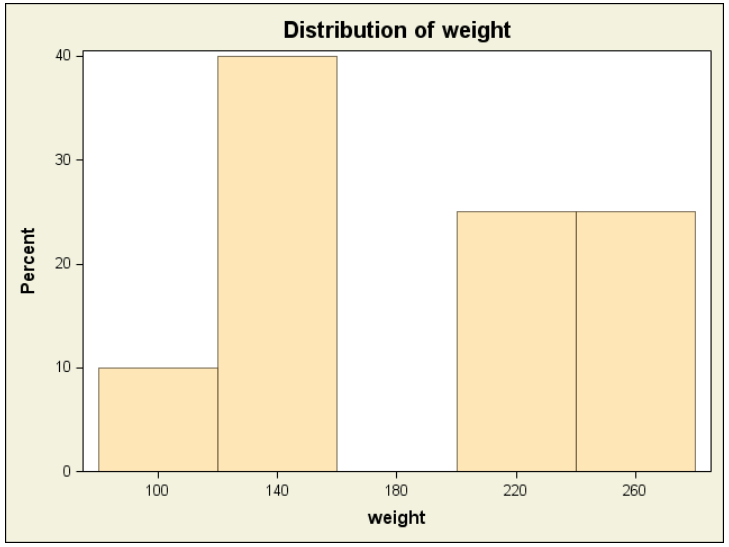
The distribution of weight was bimodal (shown in the beginning of the class)

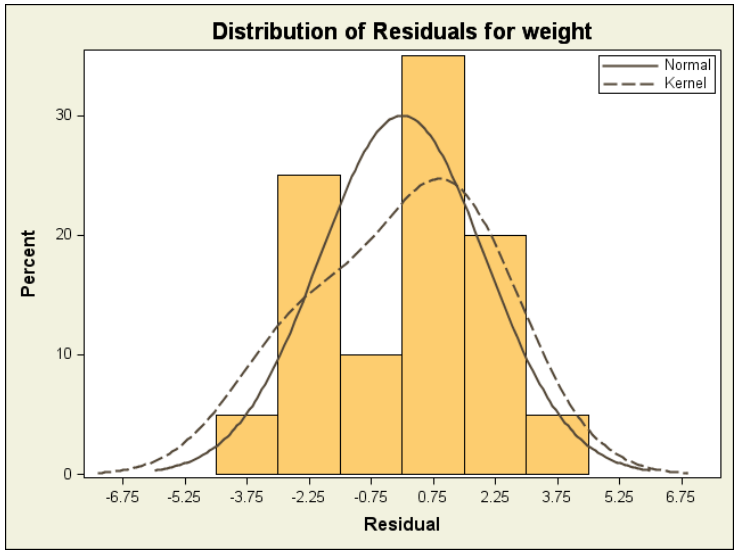
However, the analysis only called for the residuals to be normally distributed – not the actual data

$$e_p = Weight_p - \widehat{Weight}_p$$
$$= Weight_p - \left[\beta_0 + \beta_1\left(H_p - \bar{H}\right) + \beta_2 F_p + \beta_3\left(H_p - \bar{H}\right)F_p\right]$$

This is the same as saying the **conditional distribution** of the data given the predictors must be normal

Residual:





Wrap Up

Today was a brief review of basic statistics and linear models

Going forward we will

- Learn more about R and integrated development environments
- Drill into interaction terms