MATH 4720 / MSSC 5720

Instructor: Mehdi Maadooliat

Chapter 11

Department of Mathematical and Statistical Sciences

REGRESSION

- A regression function describes how a response variable y changes as an explanatory variable x changes.

- We often use a regression line to predict the value of y for a given value of x.

- Example: How much should you pay for a house?
- What factors are important in determining a reasonable price?
  - Amenities
  - Location
  - Square footage

- To determine a price, you might consider a model of the form:

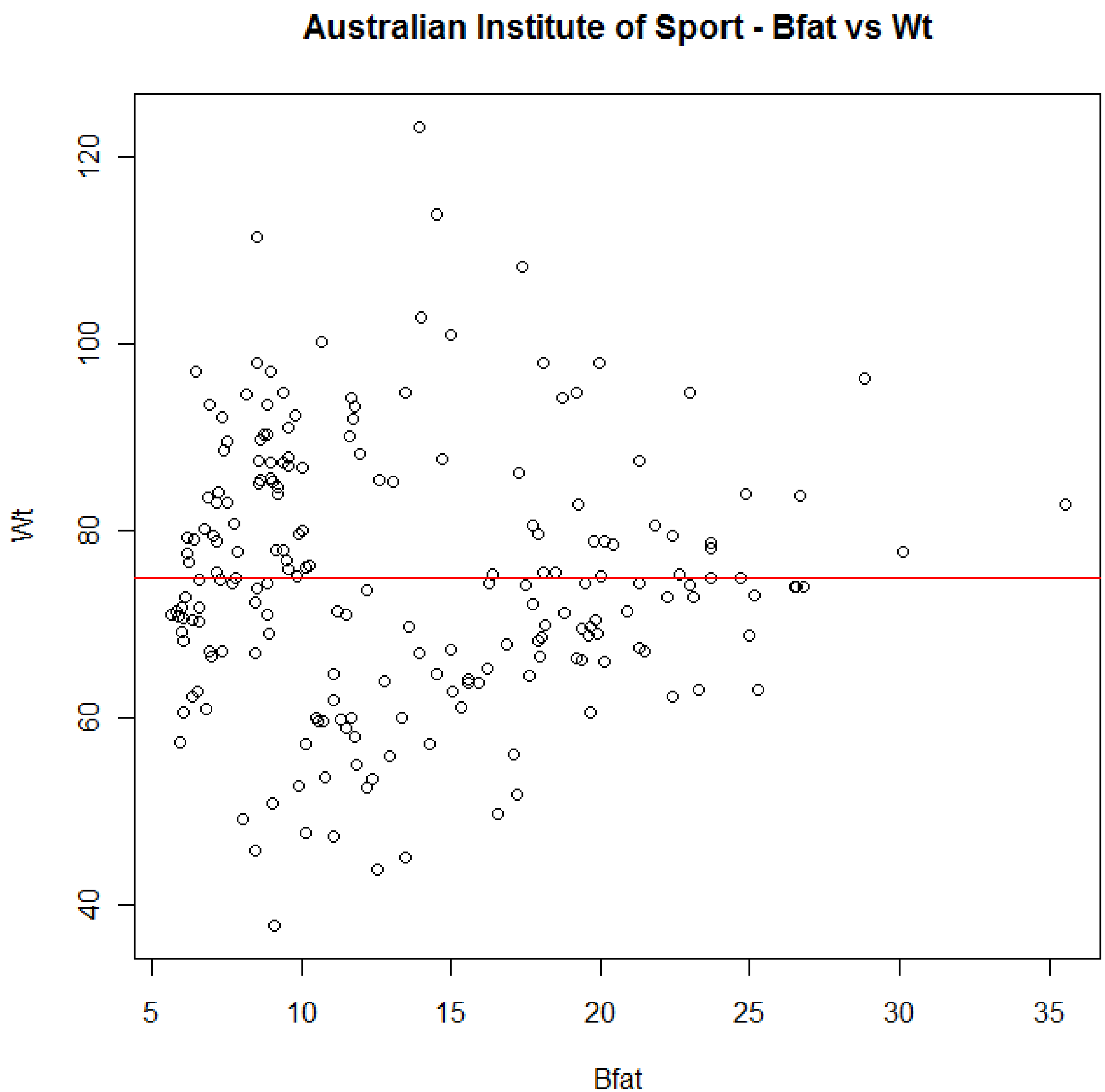
$$\text{Price} = f(\text{square footage}) + \epsilon$$

EXAMPLE – AUSTRALIAN INSTITUTE OF SPORT

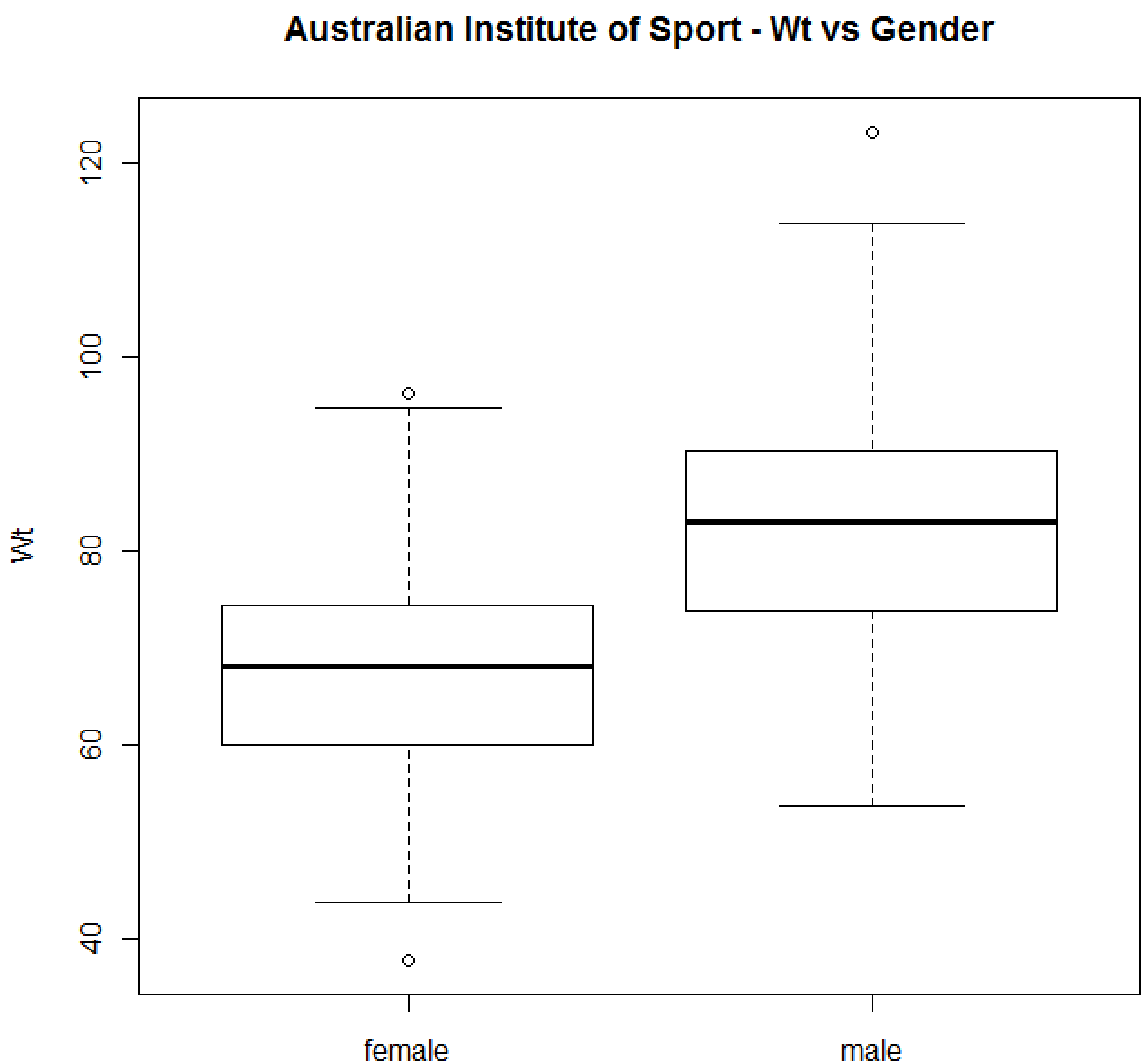
- **Data on 102 male and 100 female athletes collected at the Australian Institute of Sport, (courtesy of Richard Telford and Ross Cunningham.)**

	Gender	Bfat	Wt
1	female	19.75	78.9
2	female	21.30	74.4
3	female	19.88	69.1
4	female	23.66	74.9
5	female	17.64	64.6
	:	:	:
198	male	11.79	93.2
199	male	10.05	80.0
200	male	8.51	73.8
201	male	11.50	71.1
202	male	6.26	76.7

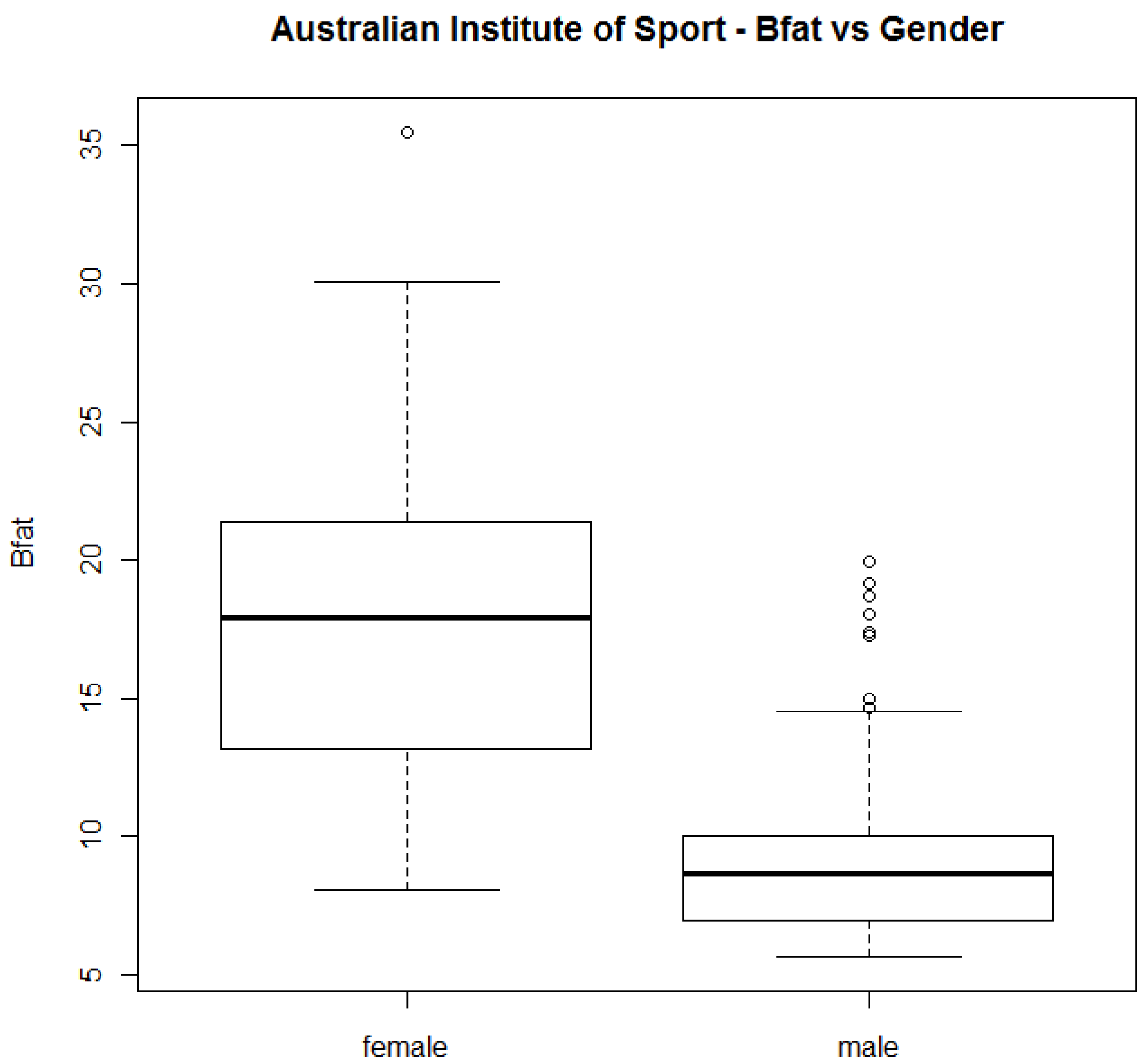
EXAMPLE CONT'D



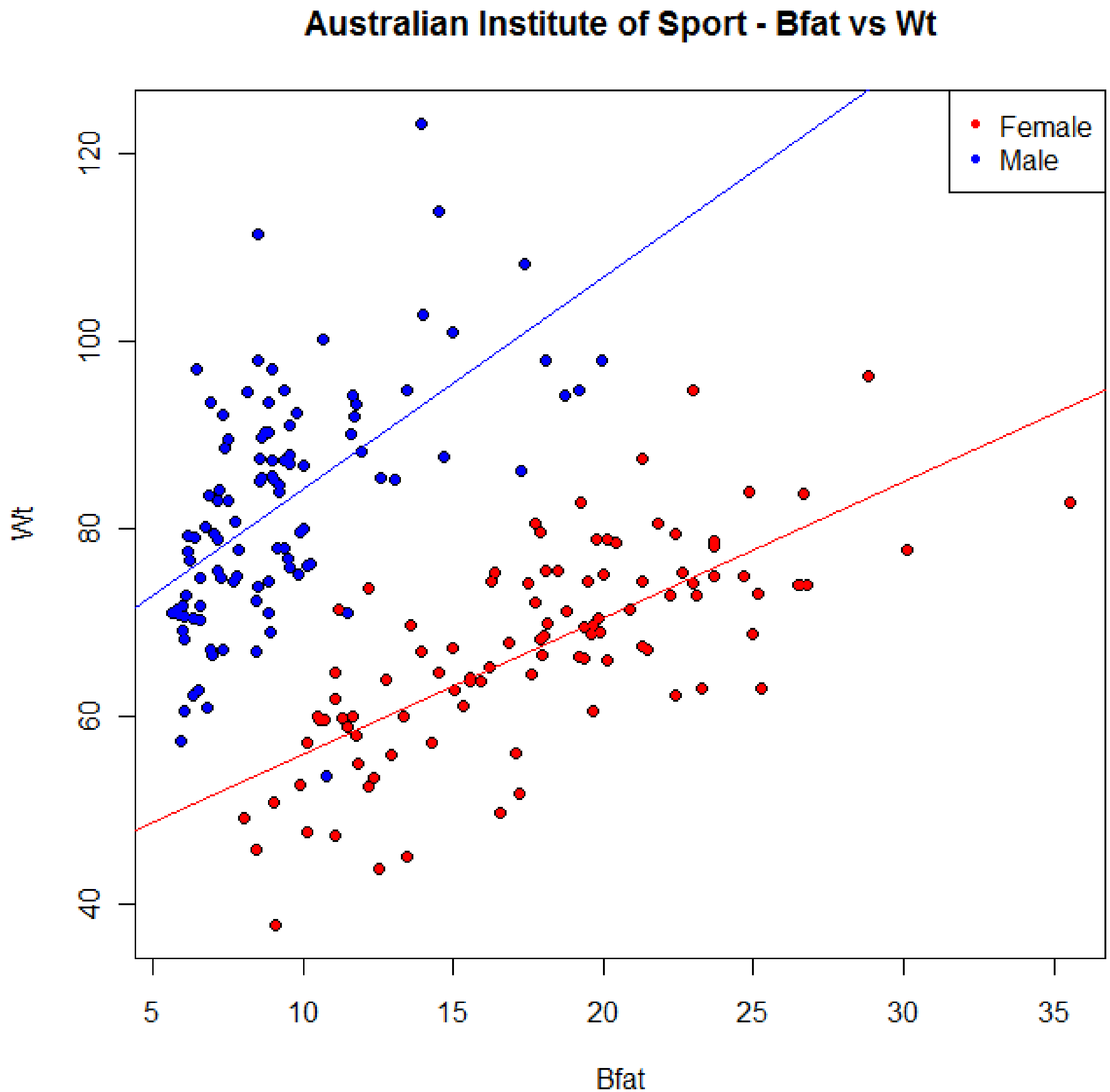
EXAMPLE CONT'D



EXAMPLE CONT'D



EXAMPLE CONT'D



REGRESSION LINE

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.

- We often use a regression line to predict the value of y for a given value of x.

- Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

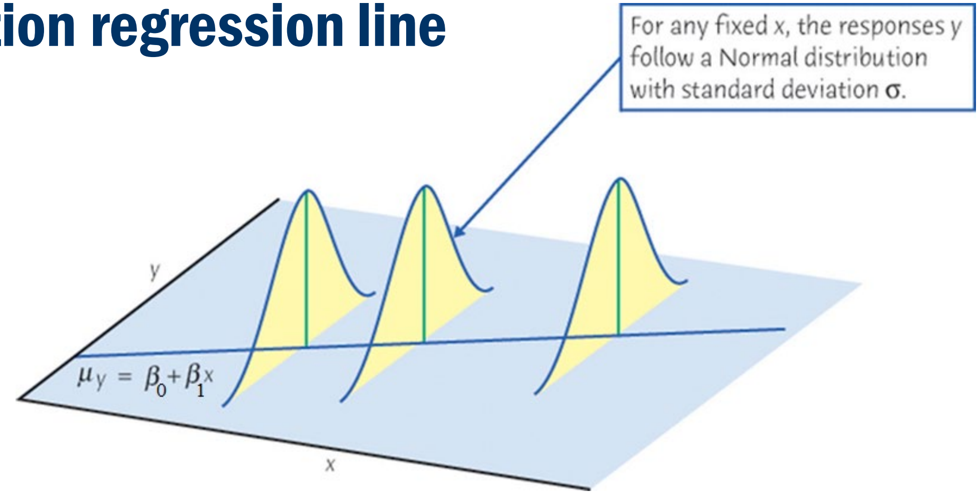
- In this equation, b is the **slope**, the amount by which y changes when x increases by one unit. The number a is the **intercept**, the value of y when $x = 0$.

POPULATION REGRESSION LINE

MARQUETTE UNIVERSITY
Be The Difference.

- We often assume that for any fixed value of x, the response y varies according to a **Normal distribution**. Additionally, we assume that repeated responses y are independent of each other. The mean response μ_y has a straight-line relationship with x given by a population regression line

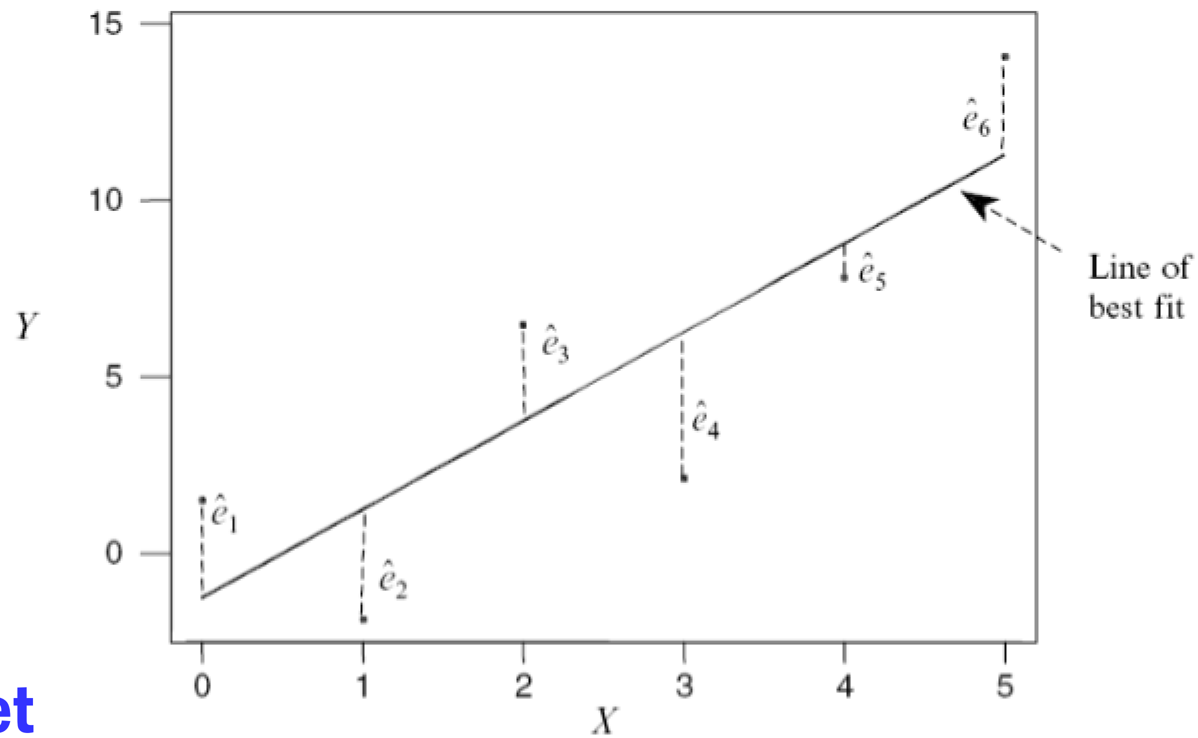
$$\mu_y = \beta_0 + \beta_1 x$$



- The **slope** β_1 and **intercept** β_0 are unknown parameters. The standard deviation of y at a given x (call it σ) is assumed to be the same for all values of x. The value of σ is unknown. There are thus three population parameters that we must estimate from the data: β_0, β_1 and σ.

LEAST SQUARES REGRESSION LINE

- The least-squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- **Residual:** $\hat{e}_i = y_i - \hat{y}_i$



- [Applet]

SIMPLE LINEAR REGRESSION

- The simplest model form to consider is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i is called the **dependent variable** or **response**.

- X_i is called the **independent variable** or **predictor**.

- ϵ_i is the random error term which is typically assumed to have
  - a Normal distribution with mean 0 and variance σ^2.
  - We also assume that error terms are independent of each other.

LEAST SQUARES CRITERION

- If the simple linear model is appropriate then we need to estimate the values β_0 and β_1.

- To determine the line that best fits our data, we choose the line that minimizes the sum of squared vertical deviations from our observed points to the line.

- In other words, we minimize

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

LEAST SQUARES ESTIMATORS

- **Objective Function:**
  - $Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$

- **Take the derivatives with respect to β_0:**
  - $\frac{\mathrm{d}Q}{\mathrm{d}\beta_0} = \sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\beta_0}(Y_i - \beta_0 - \beta_1 X_i)^2$
    $= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$

- **Setting the derivative equal to zero:**
  - $\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$
  - $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

- **To estimate β_1, likewise:**
  - $\frac{\mathrm{d}Q}{\mathrm{d}\beta_1} = \sum_{i=1}^{n}\frac{\mathrm{d}}{\mathrm{d}\beta_1}(Y_i - \beta_0 - \beta_1 X_i)^2$
    $= -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i)$

LEAST SQUARES ESTIMATORS

- **Setting the derivative $\frac{\mathrm{d}Q}{\mathrm{d}\beta_1}$ equal to zero and plugging in $\hat{\beta}_0$:**
  - $\sum_{i=1}^{n} X_i Y_i - (\overline{Y} - \hat{\beta}_1 \overline{X}) n \overline{X} - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}.\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$
- **Letting**
  - $S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$
  - $S_{xx} = \sum_{i=1}^{n}(X_i - \overline{X})^2$
  - **We have:** $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
- **How about error variance σ^2?**
  - **Use the sum of squared deviations of the points from the regression line:**
  - $\hat{\sigma}^2 = MSE = \frac{SS_E}{df_E} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$
  - **where** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ **and**
  - n **is the # of pairs** (X_i, Y_i)

LEAST SQUARES ESTIMATORS

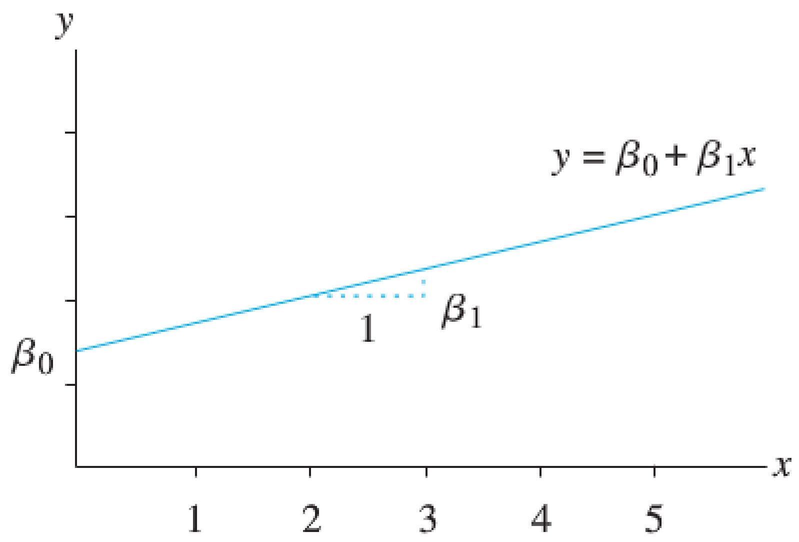
- **For** n **pairs of** (X_i, Y_i)**s,** **where** $i = 1, \dots, n$
- **If** $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, **where** ϵ_i**s are independent** $N(0, \sigma^2)$
- **Let's define** $\overline{X} = {}^1/_n \sum_i X_i$ **and** $\overline{Y} = {}^1/_n \sum_i Y_i$
  - $S_{xx} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$
  - $S_{xy} = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$
- **We can estimate the regression coefficients as:**
  - $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
  - $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$
- **Furthermore, let's define the predicted value,** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
  - $SS_E = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ **and** $S_{yy} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$
- **We have the following:**
  - $\hat{\sigma}^2 = \frac{SS_E}{n-2}$
  - $r_{XY} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, r_{XY} **is correlation coefficient (were discussed in chapter 3)**

RELATIONSHIPS BETWEEN 2 NUMERIC VARIABLES

- **Correlation** or ***r*** : measures the direction and strength of the **linear** relationship between two numeric variables
  - General Properties
    - It must be between -1 and 1, or ($-1 \leq r \leq 1$).
    - If r is negative, the relationship is negative.
    - If $r = -1$, there is a perfect negative linear relationship (extreme case).
    - If r is positive, the relationship is positive.
    - If $r = 1$, there is a perfect positive linear relationship (extreme case).
    - If r is 0, there is no **linear** relationship.
    - r measures the strength of the **linear** relationship.
    - If explanatory and response are switched, r remains the same.
    - r has no units of measurement associated with it
    - Scale changes do not affect r
- Correlation Applet

INFERENCE

- Interpretation of the parameters for $Y = \beta_0 + \beta_1 X_i$
- Often times, inference for the slope parameter, β_1, is most important.
- β_1 tells us the expected change in Y per unit change in X.



- Note that
  - β_0 is the expected value of y when $x = 0$
  - β_1 is the expected rate of change,
  - $\beta_0 = E(y|x = 0)$, and $\beta_1 = \frac{d}{dx} E(y|x)$

INFERENCE CONT'D

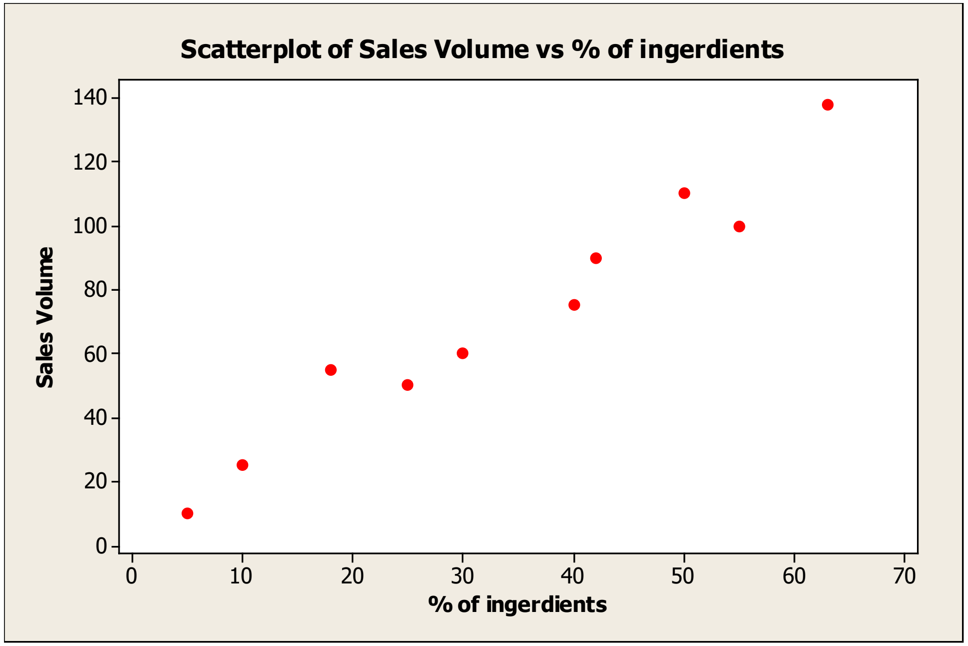
- **In practice, it may be of interest to estimate β_0 and β_1. Estimation of β_1 is more important.**
  - **If we conclude that β_1 equals 0, then we are concluding that there is no linear relationship between Y and X.**
  - **If we conclude that β_1 equals 0, then it makes no sense to use our linear model with X to predict Y.**

- **The Confidence intervals of β_0 and β_1 are given by**
  - $\hat{\beta}_1 \pm t_{\alpha/2} se(\hat{\beta}_1)$ **(df = n-2)**
    - **where** $se(\hat{\beta}_1) = \sqrt{\frac{MSE}{S_{xx}}}$
  - $\hat{\beta}_0 \pm t_{\alpha/2} se(\hat{\beta}_0)$ **(df = n-2)**
    - **where** $se(\hat{\beta}_0) = \sqrt{MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right)}$

BOOK EXAMPLE 11.2:

- **Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and the percentage of prescription ingredients purchased directly from the supplier.**

Pharmacy	Sales Volume, y (in \$1,000)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55



BOOK EXAMPLE 11.2 CONT'D

MARQUETTE UNIVERSITY
Be The Difference.

- $\overline{X} = {}^1/_n \sum_i X_i = 33.8$
- $\overline{Y} = {}^1/_n \sum_i Y_i = 71.3$
- $S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2 = 3407.6$
- $S_{xy} = \sum_i (X_i - \overline{X})(Y_i - \overline{Y}) = 6714.6$

	y	x	$y - \bar{y}$	$x - \bar{x}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Totals	713	338	0	0	6,714.60	3,407.60
Means	71.3	33.8				

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6714.6}{3407.6} = 1.97$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 71.3 - 1.97(33.8) = 4.70$
- **Prediction formula for X:** $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 4.70 + 1.97(X)$

➢ **In R:** summary(lm(y ~ x, data=exmp11.2))

Confidence Intervals:

- $\hat{\beta}_0 \pm t_{\alpha/2} se(\hat{\beta}_0)$ **(df = 8)**
  - $4.70 \pm 2.306\ (5.95)$
- $\hat{\beta}_1 \pm t_{\alpha/2} se(\hat{\beta}_1)$ **(df = 8)**
  - $1.97 \pm 2.306\ (0.15)$

Regression Analysis: Sales Volume versus % of ingerdients

```
The regression equation is
Sales Volume = 4.70 + 1.97 % of ingerdients

Predictor            Coef   SE Coef      T      P
Constant            4.698     5.952   0.79  0.453
% of ingerdients   1.9705    0.1545  12.75  0.000
```

T-statistics and p-value for testing
$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

HYPOTHESIS TESTING

- $H_0: \beta_1 = 0$ **(This means that y does not depend on x)**
- $H_a: \beta_1 \neq 0$ **(This means that y depends on x)**

- **T.S.** $t = \dfrac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}$
- **Reject H_0 in favor of H_a if $|t| > t_{\alpha/2}$ (df = n-2)**

- $H_0: \beta_0 = 0$ **(This means that $\mathrm{E}(y|x = 0) = 0$)**
- $H_a: \beta_0 \neq 0$ **(This means that $E(y|x = 0) \neq 0$)**

- **T.S.** $t = \dfrac{\widehat{\beta}_0}{se(\widehat{\beta}_0)}$
- **Reject H_0 in favor of H_a if $|t| > t_{\alpha/2}$ (df = n-2)**

BACK TO EXAMPLE 11.2

- **Suppose, we want to test if the Sale volume depend on the % of ingredients:**

- **Minitab Output**

Regression Analysis: Sales Volume versus % of ingerdients

```
The regression equation is
Sales Volume = 4.70 + 1.97 % of ingerdients


Predictor           Coef   SE Coef      T      P
Constant           4.698     5.952   0.79  0.453
% of ingerdients  1.9705    0.1545  12.75  0.000
```

- $H_0: \beta_1 = 0$ **vs.** $H_a: \beta_1 \neq 0$
- **T.S.** $T = 12.75$ **with** $\text{p}-\text{value} = 0.000$

- **Conclusion: Is p-value < 0.05? Yes. We reject H_0 in favor of H_a. We have sufficient evidence to conclude that the Sale volume depend on the % of ingredients.**

ASSUMPTIONS

- For both confidence interval and hypothesis testing problems, we make assumptions on the model
- $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1,2,\ldots,n$$
- Assumption 1: y and x are linearly related. If not, some transformation is needed.
  - y and x are linearly related can be checked through scatter plot.
- Book Example 11.1:
- Free Flights

FIGURE 11.7
Frequent flyer free flights by month

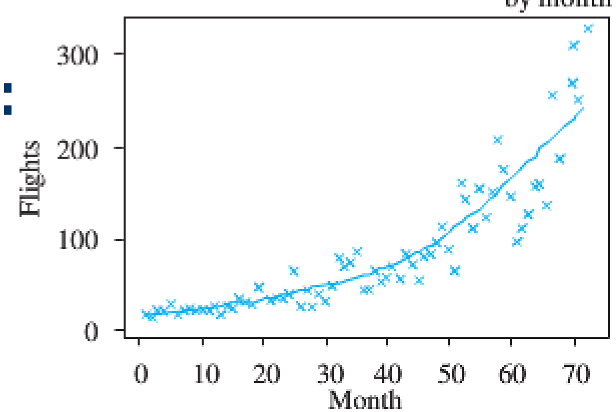
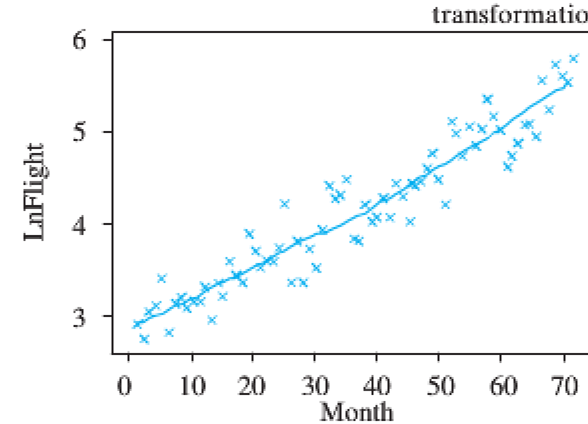


FIGURE 11.8
Result of logarithm transformation



HOW TO CHECK THE ASSUMPTIONS

- **Assumption 2: The error terms $\epsilon_i, i = 1,2, \ldots, n$ are independent and identically distributed as normal.**
  - **Second assumption that errors are independent and identically distributed can be checked through the normal probability plot of the residual, and the residual plots.**

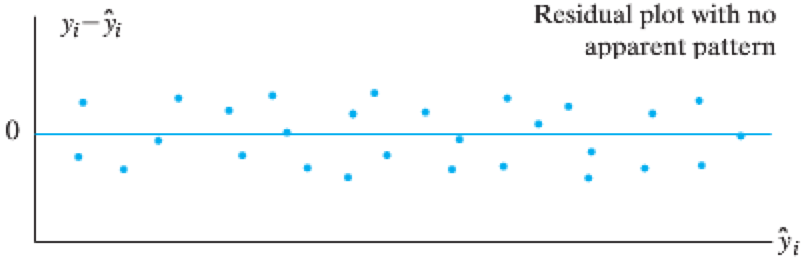


FIGURE 11.15 Residual plot with no apparent pattern

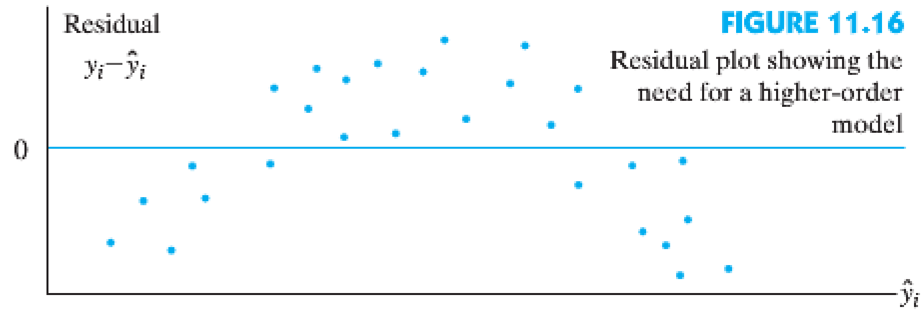


FIGURE 11.16 Residual plot showing the need for a higher-order model

- **Assumption 3:** $Var(\epsilon_i) = constant$

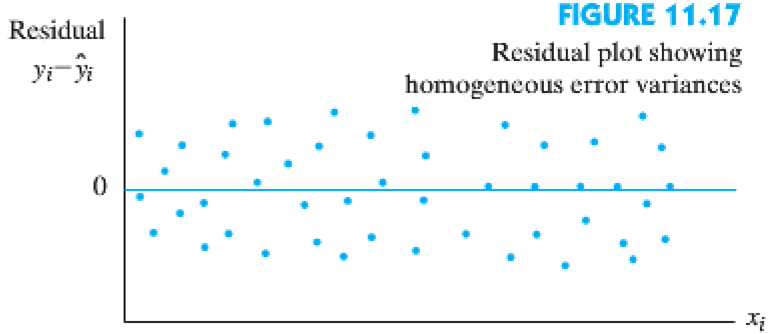


FIGURE 11.17 Residual plot showing homogeneous error variances

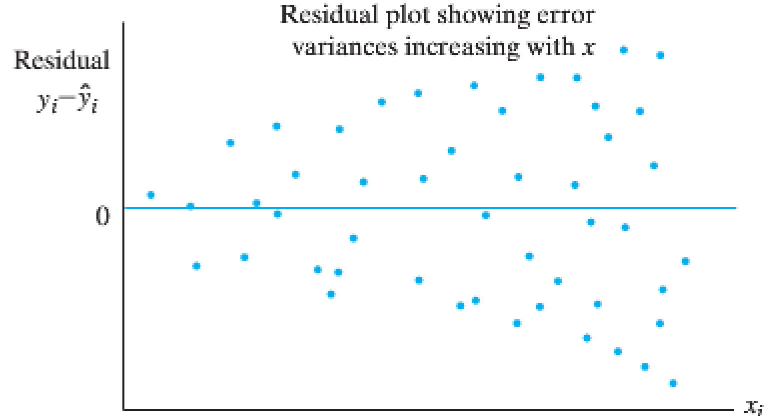


FIGURE 11.18 Residual plot showing error variances increasing with x

4 DATASETS WITH THE SAME LEAST SQUARES REGRESSION LINE

- **4 Scatterplot with same Regression Estimates:**

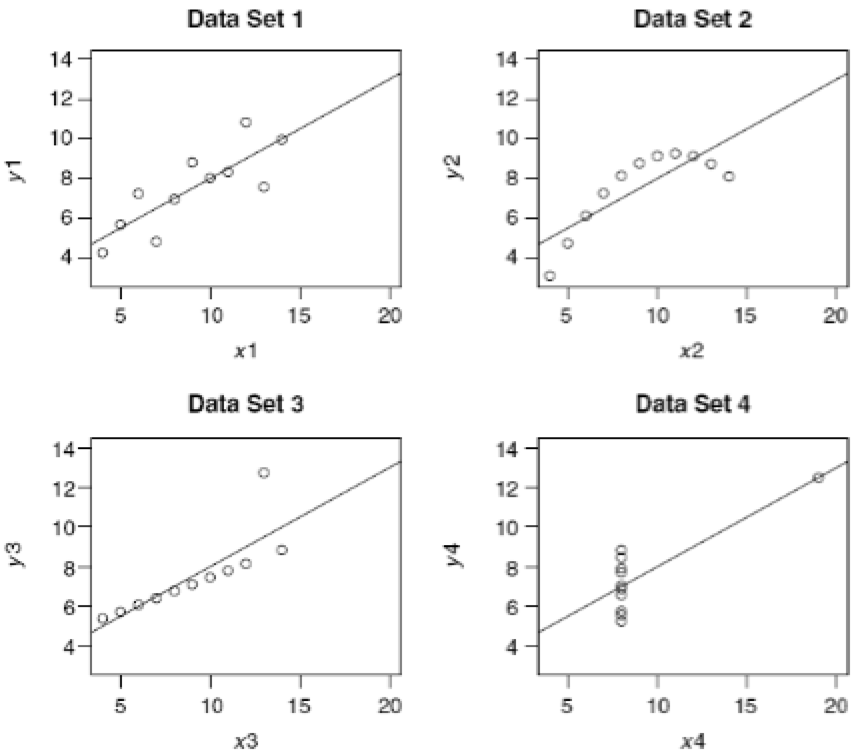
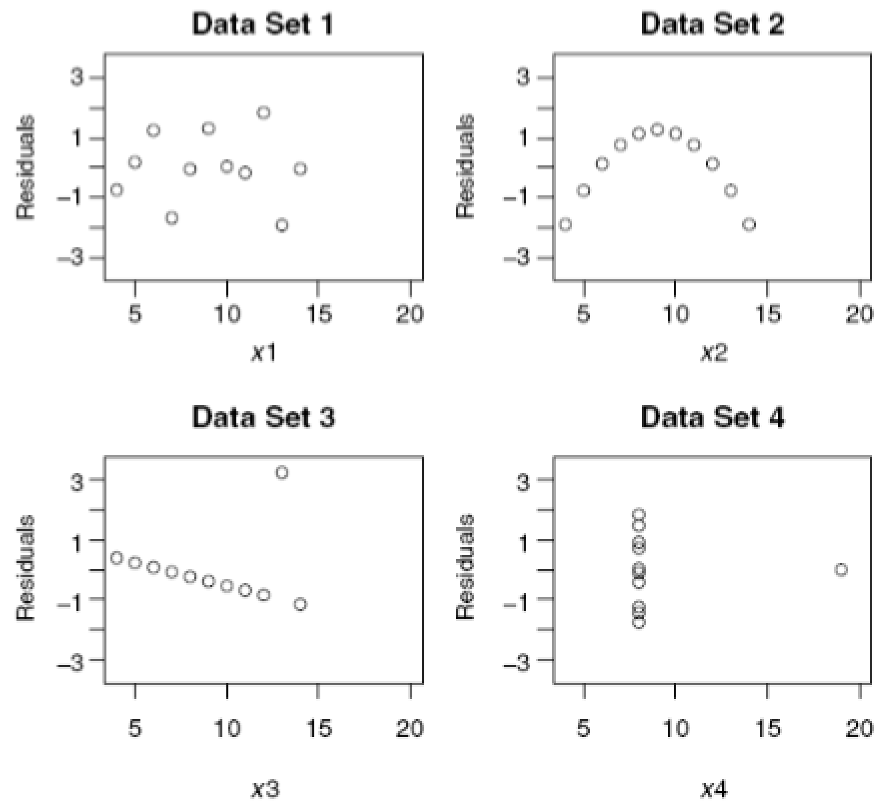


Figure 3.1 Plots of Anscombe's four data sets

- **4 Residual Plots:**



- **Source: Sheather, S.J. (2009) A Modern Approach to Regression with R, Springer, New York**

REGRESSION ANALYSIS USING ANOVA

- **Note that there is variation in y:**
  - some of it is due to regression: SS_{Reg}
  - and some due to error: SS_E

- $$SS_{Total} = SSReg + SS_E$$
  $$df_{Total} = df_{Reg} + df_E$$

- **If Regression is significant, then**

$$F = \frac{SS_{Reg}/df_{Reg}}{SS_E/df_E} > F_\alpha(df_1 = 1, df_2 = n-2)$$

- **R Output:**

Source	DF	SS	MS	F-ratio
Regression/ Explained	1	SS_{Reg}	$MS_{Reg} = SS_{Reg}/1$	$F = MS_{Reg}/MSE$
Residual/Error /Unexplained	n-2	SS_E	$MSE = SS_E/(n-2)$	
Total	n-1 = $df_{Reg} + df_E$	SST = $SS_{Reg} + SS_E$		

COEFFICIENT OF DETERMINATION

MARQUETTE UNIVERSITY
Be The Difference.

- **One way to answer the question of reliability is through a** **coefficient of determination (R^2)**
- $R^2 = \text{Proportion of Variability in Y due to Regression}$

$$= \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_E}{SS_{Tot}}$$

- **Reminder:** $SS_{Total} = SS_{Reg} + SS_E$
- R **is also the correlation in simple linear regression.**
- **If** $R^2 \approx 1$**, then most of the variability can be attributed to Regression. In this case, prediction is reliable.**
- **If** $R^2 \approx 0$**, then most of the variability is due to error. In this case prediction is not reliable.**

PREDICTION

- **Probably the most important objective of regression is to predict Y for a given X.**

- **Based on the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, Y can be predicted as**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- **Back to Example 11.2: Suppose we want to predict sale volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.**

- **Prediction formula for X:**
  - $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 4.70 + 1.97(X)$
  - $\hat{Y} = 4.70 + 1.97(15) = 34.26$

CONFIDENCE INTERVAL AND PREDICTION INTERVAL

- **100% Confidence Interval of** $E(y|x^*) = \beta_0 + \beta_1 x^*$
  - $\widehat{E(y|x^*)} \pm t_{\alpha/2} * se\left(\widehat{E(y|x^*)}\right)$, **where**
    - $\widehat{E(y|x^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^*$
    - $se\left(\widehat{E(y|x)}\right) = \sqrt{MSE\left\{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right\}}$

- **100% Prediction Interval of** $y = \beta_0 + \beta_1 x^* + \epsilon$
  - $\hat{y} \pm t_{\alpha/2} * se(\hat{y})$, **where**
    - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$
    - $se(\hat{y}) = \sqrt{MSE\left\{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right\}}$

- **Main Difference:**
  - **the confidence interval** $E(y|x)$ **provides the interval of the average of** y**, while**
  - **the prediction interval of** y **provides the interval of the individual** y**.**

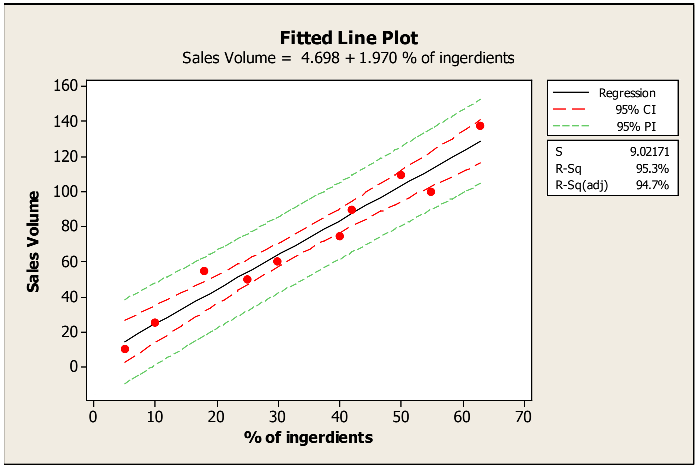
BACK TO BOOK EXAMPLE 11.2:

- **Suppose we want to obtain the Confidence and Prediction intervals of sale volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.**
- **In R:**
  - new <- data.frame(x = 15)
  - predict(model, new, interval=c(“confidence"))

```
       fit      lwr      upr
1 34.25502 24.86499 43.64505
```

  - predict(model, new, interval=c("prediction"))

```
       fit      lwr      upr
1 34.25502 11.42996 57.08008
```



- **Prediction band in R using ggplot:**
  - pred.int <- predict(model, interval = "prediction")
  - mydata <- cbind(exmp11.2, pred.int)

2. Regression line + confidence intervals

  - library("ggplot2")
  - p <- ggplot(mydata, aes(x, y)) + geom_point() + stat_smooth(method = lm)

3. Add prediction intervals

  - p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y = upr), color = "red", linetype = "dashed")

INFLUENTIAL POINTS – BAD LEVERAGE POINTS

- An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation.
- Points that are outliers in the x direction are often influential for the least-squares regression line.

- [Applet]

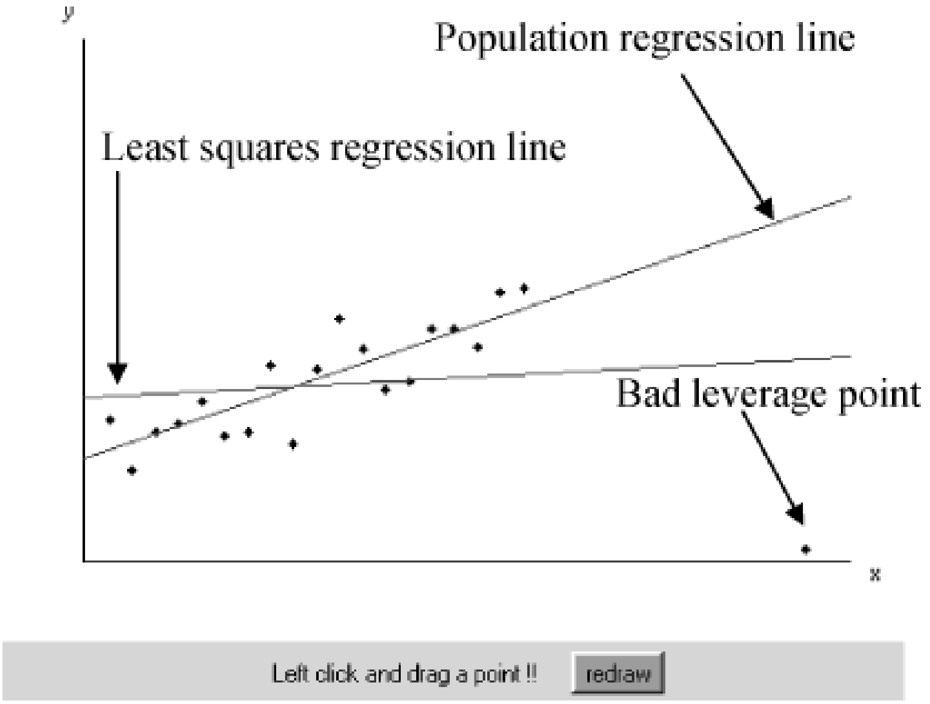
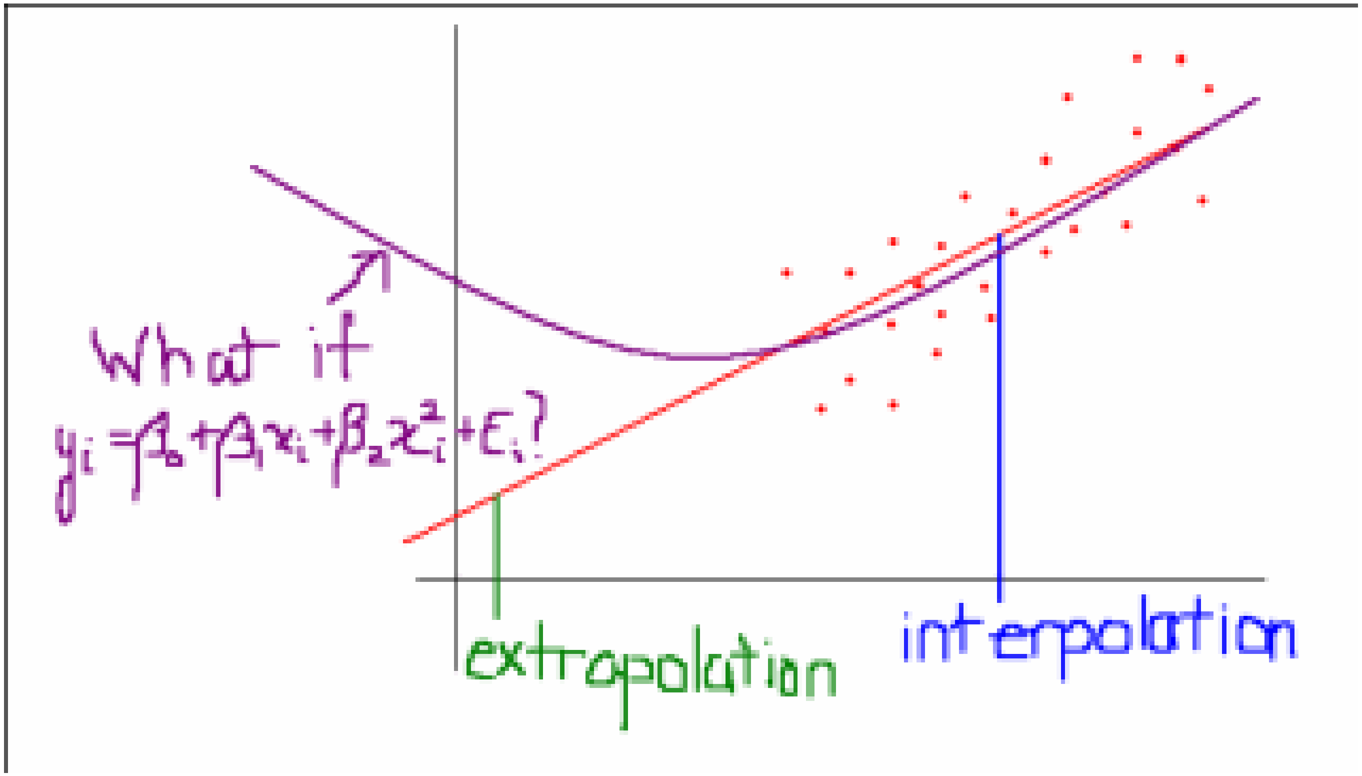


Figure 3.5 A plot showing a bad leverage point

EXTRAPOLATION VS INTERPOLATION



GOING BACK TO AIS DATA

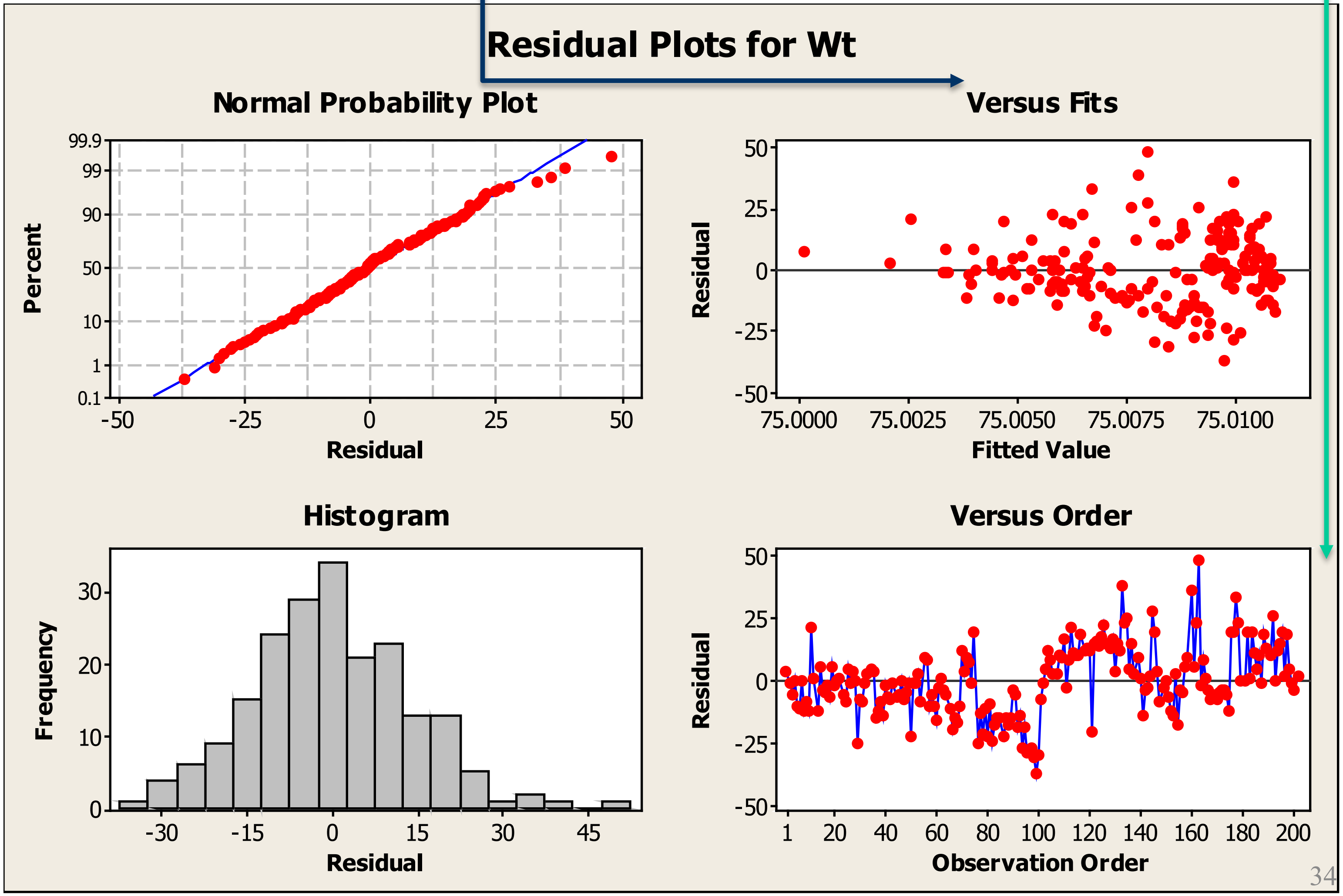
- **Wrong Analysis:**
- **Fail to Reject**
  - $H_0: \beta_1 = 0$
- $R^2 \approx 0$,
  - **Most of the variability is due to Error**
- **Regression Model is NOT significant**

Regression Analysis: Wt versus Bfat

```
The regression equation is
Wt = 75.0 - 0.000 Bfat


Predictor       Coef   SE Coef      T      P
Constant      75.013     2.363  31.75  0.000
Bfat         -0.0004    0.1591  -0.00  0.998


S = 13.9603   R-Sq = 0.0%   R-Sq(adj) = 0.0%


Analysis of Variance

Source          DF       SS     MS     F      P
Regression       1      0.0    0.0  0.00  0.998
Residual Error 200  38978.2  194.9
Total          201  38978.2
```

MODEL ASSUMPTIONS (WRONG ANALYSIS)

- The error terms $\epsilon_i, i = 1,2,\ldots,n$ are NOT independent
- $Var(\epsilon_i)$ is NOT constant



GOING BACK TO AIS DATA

MARQUETTE UNIVERSITY
Be The Difference.

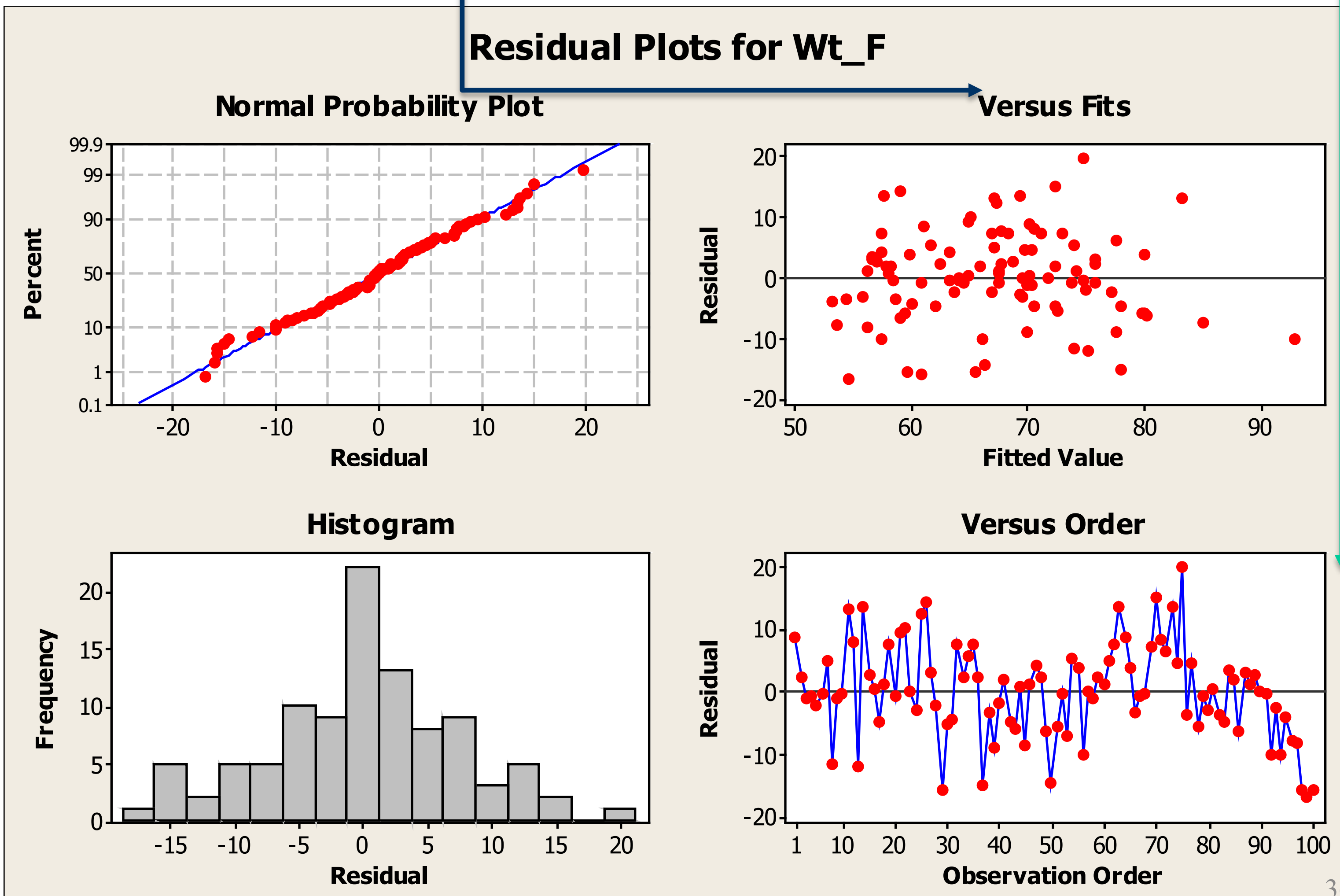
- **Correct Analysis for "Gender=Female"**
- **Reject**
  - $H_0: \beta_1 = 0$
- $R^2 \approx 53\%$,
  - **53% of the variability is due to Model**
- **Regression Model is significant**
- $F = T^2$

Regression Analysis: Wt_F versus Bfat_F

```
The regression equation is
Wt_F = 41.4 + 1.45 Bfat_F


Predictor      Coef   SE Coef      T      P
Constant     41.443     2.599  15.95  0.000
Bfat_F       1.4510    0.1393  10.42  0.000


S = 7.55769   R-Sq = 52.5%   R-Sq(adj) = 52.1%


Analysis of Variance

Source           DF       SS      MS       F      P
Regression        1   6197.9  6197.9  108.51  0.000
Residual Error   98   5597.6    57.1
Total            99  11795.6
```

MODEL ASSUMPTIONS

- **The error terms $\epsilon_i, i = 1,2, \ldots, n$ are independent**
- $Var(\epsilon_i)$ **is constant**



CAUTIONS ABOUT REGRESSION

MARQUETTE UNIVERSITY
Be The Difference.

- Regression is a powerful tool for describing the relationship between two variables. When you use these tools, you must be aware of their limitations.
- **Regression lines describe only linear relationships**. You can do the calculations for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern.
- **Least-squares regression lines are not resistant**. Always plot your data and look for observations that may be influential.
- **Beware extrapolation**. Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.