Key Points of Chapter 2

- A major function of statistics methods is *prediction*
- Purpose: Predict a variable Y in terms of other variables $X_1, \ldots, X_p$
- *Regression function* $\mu_Y(x_1, \ldots, x_p)$ is expected value of Y when $X_1 = x_1, \ldots, X_p = x_p$.
- Subpopulation approach
- Graphical approaches, e.g. histograms and scatterplots
- Sampling methods, whether to use X values in constructing the sample
- Linear and nonlinear regression functions

Prediction

- A major function of statistics methods is *prediction*
- Predict a variable Y as a function of $X_1, \ldots, X_p$

EXAMPLE 2.2.5

An investigator wants to study the pattern of associations among the following variables for U.S.-born individuals who are at least 18 years old now.

Y = height of the individual at age 18
X_1 = length of the individual at birth
X_2 = mother's height at age 18
X_3 = father's height at age 18
X_4 = paternal grandmother's height at age 18
X_5 = paternal grandfather's height at age 18
X_6 = maternal grandmother's height at age 18
X_7 = maternal grandfather's height at age 18

The investigator may not actually be interested in predicting what an individual's height will be at age 18, but if a good prediction function is found, then this function may yield information regarding what the predominant determinant of an individual's height is—the heights of his maternal ancestors, the heights of his paternal ancestors, both, or neither. ■

Purpose of Prediction

- Intrinsic interest in predicting something
  - Cars example — predicting maintenance costs
- Prediction for the purpose of understanding relationships
  - Blood pressure as a function of height and weight — real interest in control?
- Prediction as an alternative to taking a direct measurement
  - Tree volume as a function of width and height — avoid chopping down the tree!
- One predictor (X) variable or several

TABLE 2.2.1

A Schematic Representation of a Bivariate Population with Response Variable Y and Predictor Variable X

Item Number I	Response Variable Y	Predictor Variable (Explanatory Variable) X
1	Y_1	X_1
2	Y_2	X_2
⋮	⋮	⋮
I	Y_I	X_I
⋮	⋮	⋮
N	Y_N	X_N

TABLE 2.2.2

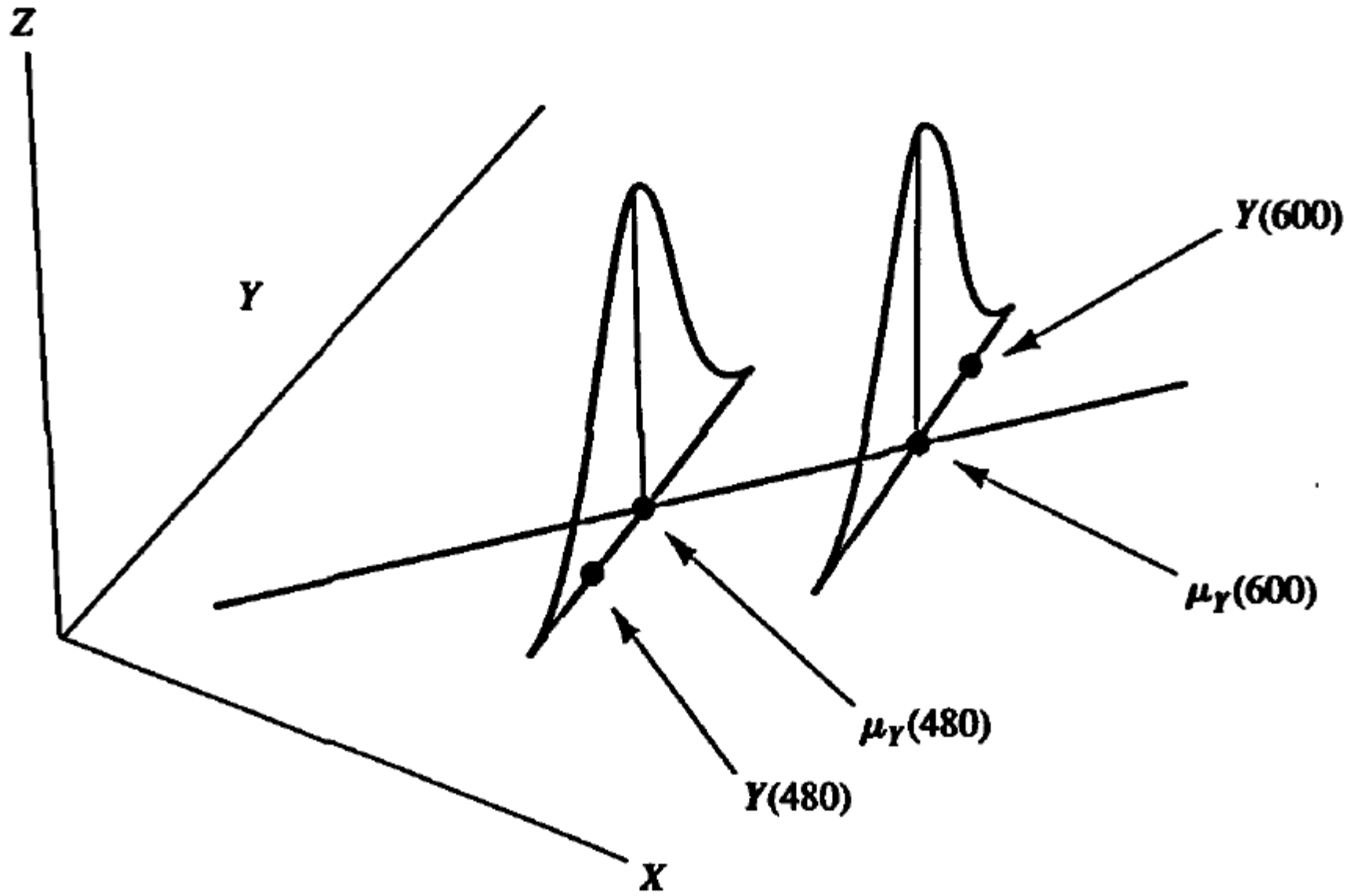
A Schematic Representation of a Trivariate Population with Response Variable Y and Predictor Variables X_1 and X_2

Item Number I	Response Variable Y	Predictor Variable 1 (Explanatory Variable 1) X_1	Predictor Variable 2 (Explanatory Variable 2) X_2
1	Y_1	X_{11}	X_{12}
2	Y_2	X_{21}	X_{22}
⋮	⋮	⋮	⋮
I	Y_I	X_{I1}	X_{I2}
⋮	⋮	⋮	⋮
N	Y_N	X_{N1}	X_{N2}

Subpopulations

- Idea of restricting to a specific *subpopulation* to learn more about a quantity of interest
- Example of car maintenance costs
  - I drive my car 14,000 miles in its first year
  - Other car owners have driven much more or much less
  - If I want to compare my maintenance costs with those of other owners, it makes sense to restrict to a subpopulation of like users
    - Reality check: unlikely to find many drivers with *exactly* the same mileage as me
    - In practice, would restrict to a range of nearby values
    - So far, this assumes *nothing* about the relationship being linear

FIGURE 2.3.1



Calculations for Car Costs Example

- Among all drivers: mean maintenace cost is $526, SD is $106
- Among drivers with 14,000 miles: mean maintenace cost is $621, SD is $23
- By restricting to a relevant subpopulation, I get a better estimate for my car, with much smaller standard deviation
- However (point added): in practice I would have to decide how wide an interval to take (around my car's mileage) — bias/variance trade-off
- To discuss: implementation in R

FIGURE 2.3.3

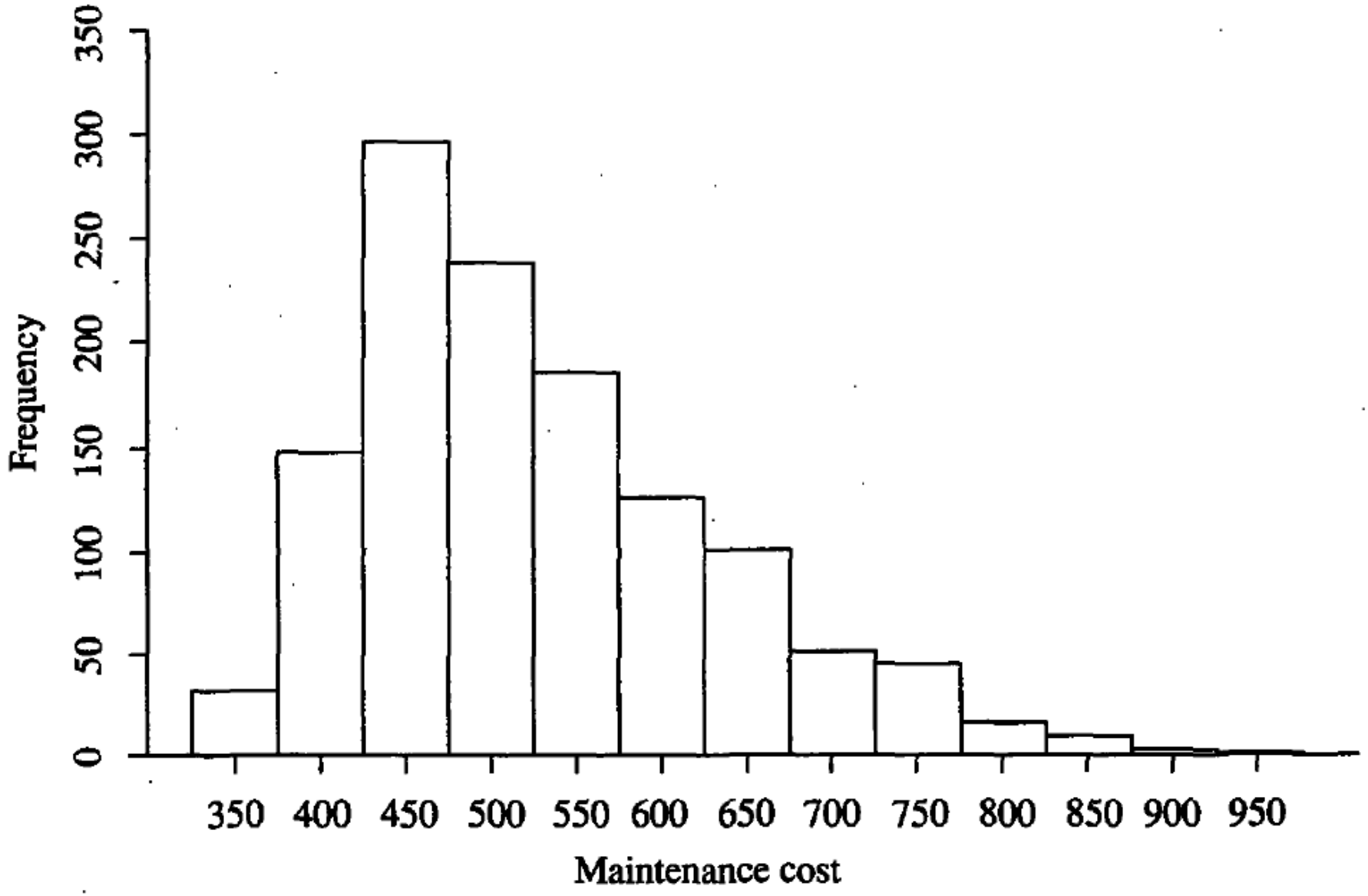
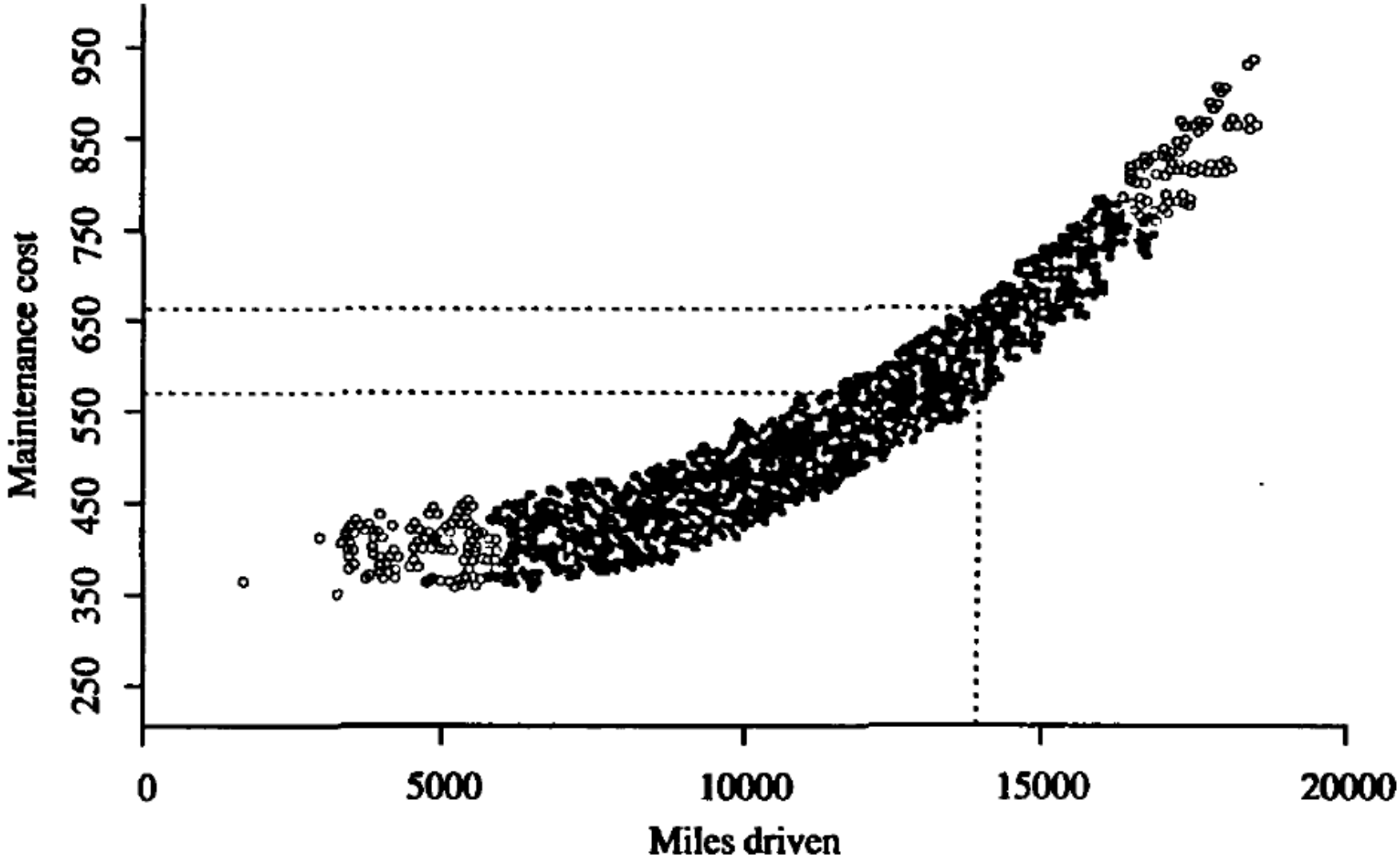
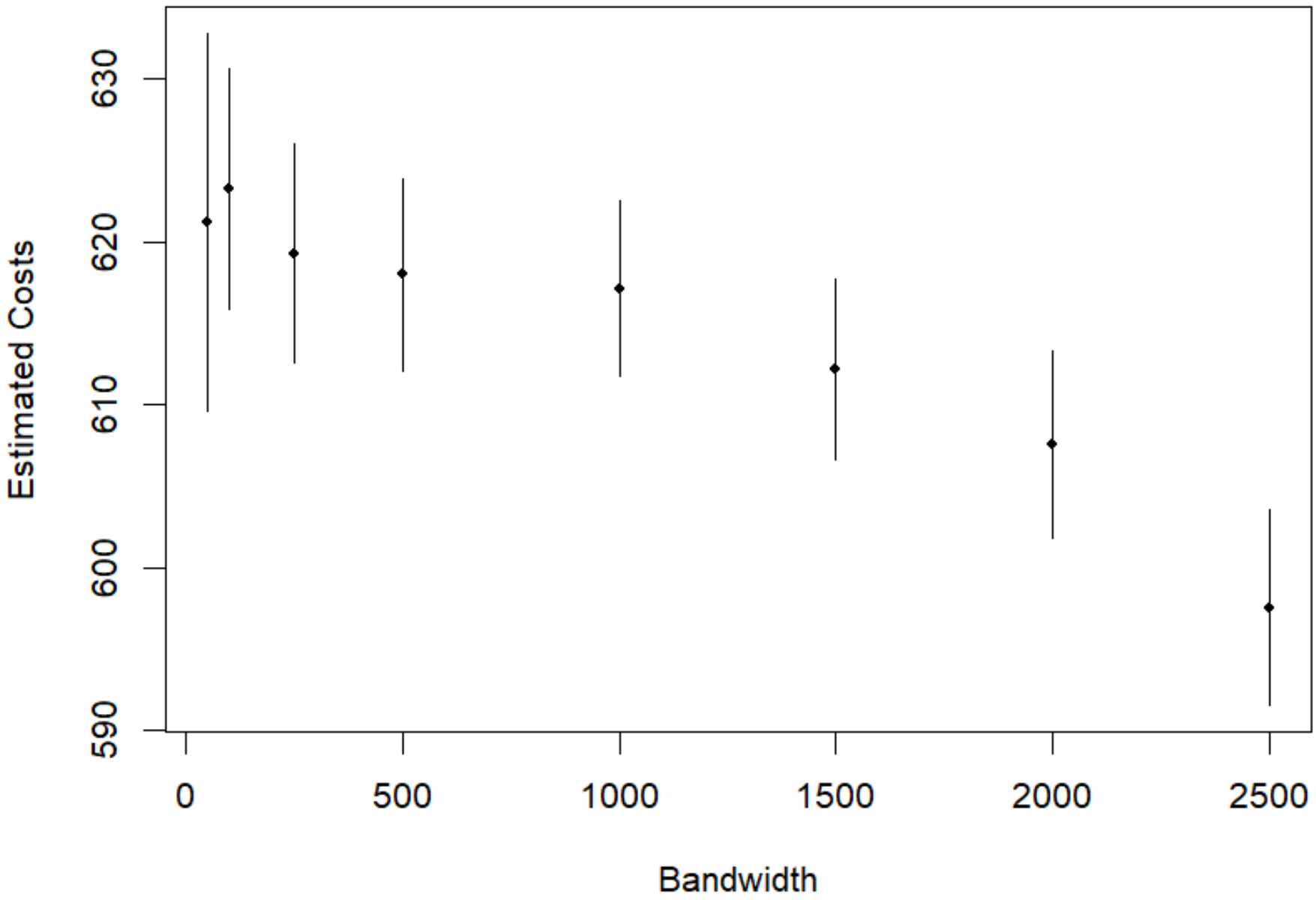

Frequency
0
50
100
150
200
250
300
350
350
400
450
500
550
600
650
700
750
800
850
900
950
Maintenance cost

FIGURE 2.3.4


Maintenance cost
250
350
450
550
650
750
850
950
0
5000
10000
15000
20000
Miles driven

Some R Code and Extensions

```
# read data and basic calculations
CAR=read.table('.../car.txt',header=T)
plot(CAR$Miles,CAR$Maint,ylab='Maintenance Cost',xlab='Miles Driven')
hist(CAR$Maint)
u=which(CAR$Miles==14000)
length(u)
print(c(mean(CAR$Maint),sqrt(var(CAR$Maint))))
print(c(mean(CAR$Maint[u]),sqrt(var(CAR$Maint[u]))))
#
# here's an extension
#
# consider various "bandwidths" of possible intervals round 14000 miles
#
# for each bandwidth, compute mean, SE and 95% confidence interval for
# predicted maintenance costs
X=matrix(nrow=8,ncol=3)
bw=c(50,100,250,500,1000,1500,2000,2500)
for(i in 1:8){
X[i,1]=bw[i]
X[i,2]=mean(CAR$Maint[abs(CAR$Miles-14000)<=bw[i]])
X[i,3]=sqrt(var(CAR$Maint[abs(CAR$Miles-14000)<=bw[i]])/sum(abs(CAR$Miles-14000)<=bw[i]))
}
ymax=max(X[,2]+2*X[,3])
ymin=min(X[,2]-2*X[,3])
par(cex=1.3)
plot(X[,1],X[,2],pch=20,xlab='Bandwidth',ylab='Estimated Costs',ylim=c(ymin,ymax))
for(i in 1:8){lines(c(X[i,1],X[i,1]),c(X[i,2]-2*X[i,3],X[i,2]+2*X[i,3]))}
```


Estimated Costs
590
600
610
620
630
0
500
1000
1500
2000
2500
Bandwidth

Conclusion from this example

- As the bandwidth increases, the confidence interval gets narrower, but the estimated mean also changes as we include more and more cars
- Visually, it looks as though we should use a bandwidth of 1,000 miles or less

Sampling Methods (p. 93–96 of text)

- Simple random sampling (SRS)
- Sampling with pre-selected variables
  - What's the more familiar name for that?
- Which is better? Contrast between *observational studies* and *experiments*

Linear v. Nonlinear Models (p. 96–97 of text)

- A model is *linear* if the regression function is linear in the *unknown parameters* (usually written β_0, β_1, etc.
- It doesn't really matter whether it's linear or nonlinear in the x variables
- Most of this course is about *linear* models, for which the theory and methods are much better developed
- If there's time, I'll do a bit about nonlinear models at the end.

$$
\left.
\begin{aligned}
\mu_Y(x) &= \beta_0 \\
\mu_Y(x) &= \beta_0 + \beta_1 x \\
\mu_Y(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\mu_Y(x_1) &= \beta_0 + \beta_1 x_1^2 + \beta_2 x_1^{3/2} + \beta_3 / \ln|x_1| \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{x_1} + \beta_2 x_2 + \beta_3 e^{x_1 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 e^{-2x_1} + \beta_2 \sin(x_1 x_2) + \beta_3 x_1 \ln(x_2^2) \tan(x_3) \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_1 x_3^2
\end{aligned}
\right\} \quad (2.3.1)
$$

$$
\left.
\begin{aligned}
\mu_Y(x_1) &= \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1) &= \beta_0 + \beta_1 e^{\beta_2 x_1} \\
\mu_Y(x_1, x_2) &= \beta_0 + \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2} \\
\mu_Y(x_1, x_2, x_3) &= \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \\
\mu_Y(x_1, x_2) &= \beta_1 x_1 / (\beta_2 e^{\beta_3 x_2})
\end{aligned}
\right\} \quad (2.3.2)
$$

2.4.12 Which of the following regression functions are (simultaneously) linear in the unknown parameters (the symbols $\beta_0, \beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \gamma_3$ refer to unknown parameters)?

a $\mu_Y(x) = \beta_0 + \beta_1 x^4$.

b $\mu_Y(x_1, x_2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1 x_2$.

c $\mu_Y(x) = \beta_0 + \beta_1 x^{\beta_2}$.

d $\mu_Y(x_1, x_2) = \gamma_1 \sqrt{\gamma_2 x_1 + \gamma_3 x_2}$.

e $\mu_Y(x) = \beta_0 + \beta_1 x^{1/2} + \beta_2 / x + \beta_3 e^{-2x}$.

f $\mu_Y(x) = \beta_0 + \sin(\beta_1 x)$.

g $\mu_Y(x_1, x_2, x_3) = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3}$.

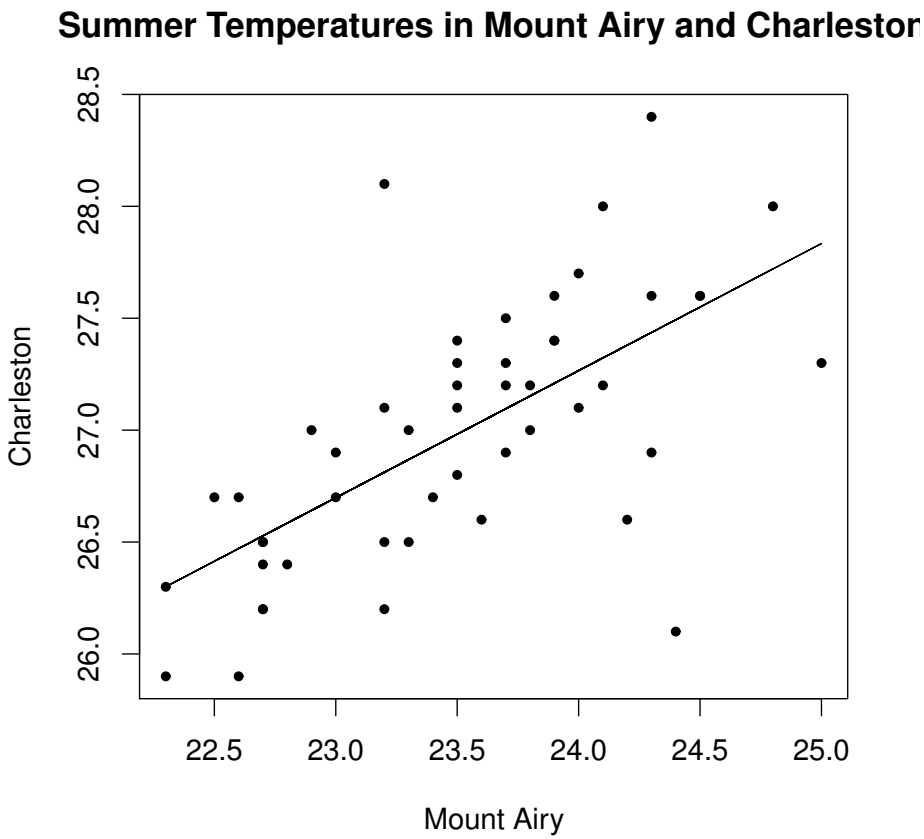
Comment about the Last Example in the Previous Class

- If you took log μ_Y and rewrote log β_0 as β_0, it would be a linear model
- But log of the expected value is not the same as expected value of the logarithm
- Question of whether the same model would be justified when applied to log Y instead of Y

Chapter 3: Simple Linear Regression

- Sample (Y_i, X_i), $i = 1, \ldots, n$, X_i a predictor of Y_i.
- "Simple Linear Regression" means there is only one X variable (more than one: Multiple Linear Regression comes later)
- *Linear* means $\mu_Y(x) = \beta_0 + \beta_1 x$ for parameters β_0 and β_1.
- Standard deviation $\sigma_Y(x)$ could also depend on x. The most common assumption is that $\sigma_Y(x)$ is a constant σ, but we shouldn't assume this automatically!
- In addition:
  - Observations independent (or uncorrelated). The text doesn't state this explicitly but if the sampling is truly random (SRS or stratified) this would be automatically satisfied
  - Gaussian distributions? — also a very common assumption, but some of the theory is valid without that
  - Assumes X_i and Y_i are measured without error — more on this later

Example: Mount Airy and Charleston



- Straight line regression?
- Constant variances?
- Gaussian distributions?
- Independent?

Assumptions of Simple Linear Regression I

Different authors adopt slightly different assumptions — mine differ a bit from the text's.

Assumptions A:

- $y_i = \beta_0 + \beta_1 x_i + e_i$
- For each i, e_i has mean 0 and standard deviation σ (same for all i)
- $e_1, \ldots, e_n$ are uncorrelated

Assumptions of Simple Linear Regression II

Assumptions B:

- $y_i = \beta_0 + \beta_1 x_i + e_i$
- For each i, e_i has mean 0 and standard deviation σ (same for all i)
- $e_1, \ldots, e_n$ are *independent*
- In addition, each of the e_i has a *Gaussian distribution*
- B assumes a little bit more than A — some of the theoretical results do in fact require B (clarify this later)
- In practice, it is rather hard to distinguish the two sets of assumptions

Method of Least Squares

Suppose we have a statistical model with

- Observations $y_1, \ldots, y_n$
- $\mathrm{E}\{y_i\} = f(x_i, \theta)$ possibly depending on additional known *covariates* x_i and unknown parameter θ, with $f(\cdot, \cdot)$ a known function of x and θ
- Uncorrelated observations with a common variances (this assumption will be relaxed later)

The *method of least squares* chooses the parameter θ to minimize

$$S = \sum_{i=1}^{n} \{y_i - f(x_i; \theta)\}^2 .$$

Application 1

Suppose $\mathrm{E}\{y_i\} = \mu$ for a fixed constant. This is the $y_i \sim N[\mu, \sigma]$ form in a different guise. Let $\bar{y} = \frac{1}{n}\sum y_i$. Then

$$\begin{aligned} S &= \sum_{i=1}^{n} (y_i - \mu)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \mu)^2 \\ &= \sum (y_i - \bar{y})^2 - 2(\bar{y} - \mu)\sum (y_i - \bar{y}) + n(\bar{y} - \mu)^2 \\ &= \sum (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \end{aligned} \tag{1}$$

since $\sum (y_i - \bar{y}) = 0$. But the first term in (2) does not depend on μ while the second is minimized when $\mu = \bar{y}$. Therefore, the least squares estimate of μ is

$$\hat{\mu} = \bar{y}.$$

Application 2

Now suppose $\mathrm{E}\{y_i\} = \beta_0 + \beta_1(x_i - \bar{x})$ where $x_1, ..., x_n$ are known (scalar) covariates, $\bar{x} = \frac{1}{n}\sum x_i$ and β_0 and β_1 are unknown parameters. This is the classical *simple linear regression* problem, where "simple" means that there is only a single covariate.

Comment: Centering the x_i's about their mean $\bar{x}$ simplifies the math, but the model is essentially the same without that.

In this context, the *principle of least squares* chooses the parameters β_0 and β_1 to minimize

$$S = \sum \{y_i - \beta_0 - \beta_1(x_i - \bar{x})\}^2 .$$

Solution to Least Squares Regression I

First, consider quadratic expressions of the form

$$\begin{aligned} S &= A - 2B\beta + C\beta^2 \\ &= C\left(\beta - \frac{B}{C}\right)^2 + A - \frac{B^2}{C}. \end{aligned}$$

This is minimized with respect to β when

$$\beta = \frac{B}{C}$$

and leads to the expression

$$S = A - \frac{B^2}{C}.$$

Solution to Least Squares Regression II

Next, consider β_0.

$$\begin{aligned} S &= \sum [\{y_i - \beta_1(x_i - \bar{x})\} - \beta_0]^2 \\ &= \sum \{y_i - \beta_1(x_i - \bar{x})\}^2 - 2\beta_0 \sum \{y_i - \beta_1(x_i - \bar{x})\} + n\beta_0^2 \\ &= \sum \{y_i - \beta_1(x_i - \bar{x})\}^2 - 2n\bar{y}\beta_0 + n\beta_0^2 \\ &= \sum \{y_i - \beta_1(x_i - \bar{x})\}^2 + n(\bar{y} - \beta_0)^2 - n\bar{y}^2. \end{aligned}$$

The first and third terms do not depend on β_0 while the middle term is minimized when $\beta_0 = \bar{y}$. Therefore, the least squares estimator of β_0 is

$$\hat{\beta}_0 = \bar{y}$$

and this substitution also leads to

$$S = \sum \{y_i - \bar{y} - \beta_1(x_i - \bar{x})\}^2 .$$

Solution to Least Squares Regression III

Now write

$$\begin{aligned} S &= \sum \{y_i - \bar{y} - \beta_1(x_i - \bar{x})\}^2 \\ &= \sum (y_i - \bar{y})^2 - 2\beta_1 \sum (y_i - \bar{y})(x_i - \bar{x}) + \beta_1^2 \sum (x_i - \bar{x})^2. \end{aligned}$$

Now apply the result on slide I. The least squares estimate for β_1 is

$$\widehat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

and this leads to

$$S = \sum (y_i - \bar{y})^2 - \frac{\{\sum (y_i - \bar{y})(x_i - \bar{x})\}^2}{\sum (x_i - \bar{x})^2}.$$

Summary

The *least squares estimators* for a simple linear regression are

$$\begin{aligned} \widehat{\beta}_0 &= \bar{y}, \\ \widehat{\beta}_1 &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \end{aligned}$$

and lead to

$$\begin{aligned} S &= \sum \left\{ y_i - \widehat{\beta}_0 - \widehat{\beta}_1 (x_i - \bar{x}) \right\}^2 \\ &= \sum (y_i - \bar{y})^2 - \frac{\{\sum(y_i - \bar{y})(x_i - \bar{x})\}^2}{\sum(x_i - \bar{x})^2}. \end{aligned}$$

This also leads to an estimator for the variance ($\widehat{\sigma}^2$ or s^2),

$$\widehat{\sigma}^2 = \frac{S}{n-2}.$$

Question: why is the divisor $n-2$ and not n or $n-1$?

Standard Errors

- If $Y_1, \ldots, Y_n$ are independent (or uncorrelated) random variables and $a_1, \ldots, a_n$ are constants, then $\text{Var}\{\sum a_i Y_i\} = \sum \text{Var}(a_i Y_i) = \sum a_i^2 \sigma_i^2$. If all variances are the same, then $\text{Var}\{\sum a_i Y_i\} = \sigma^2 \sum a_i^2$.

- Application 1: suppose $a_i = \frac{1}{n}$. Then $\sum a_i^2 = n \cdot \left(\frac{1}{n}\right)^2 = \frac{1}{n}$. So $\text{Var}\{\bar{Y}\} = \frac{\sigma^2}{n}$.

- Application 2: suppose $a_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$. Then $\sum a_i^2 = \sum \left[\frac{(x_i - \bar{x})^2}{\{\sum (x_i - \bar{x})^2\}^2}\right] = \frac{1}{\sum (x_i - \bar{x})^2}$.

- Hence if $\hat{\sigma}^2$ is an unbiased estimator of σ^2, we call $\frac{\hat{\sigma}}{\sqrt{n}}$ and $\frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$ the *standard errors* of $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.

Relation to Textbook Discussion

- The text (pp. 112–114) defines $\mu_Y(x) = \beta_0 + \beta_1 x$ (without subtracting $\bar{x}$) and then gives the estimates

$$\widehat{\beta}_1 \;=\; \frac{\sum\{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum\{(x_i - \bar{x})^2\}},\; \widehat{\beta}_0 \;=\; \bar{y} - \widehat{\beta}_1 \bar{x}. \qquad (2)$$

- If you write $\mu_Y(x) = \beta_0 + \beta_1(x_i - \bar{x})$ as I did, with $\widehat{\beta}_1$ as above and $\widehat{\beta}_0 = \bar{y}$, then

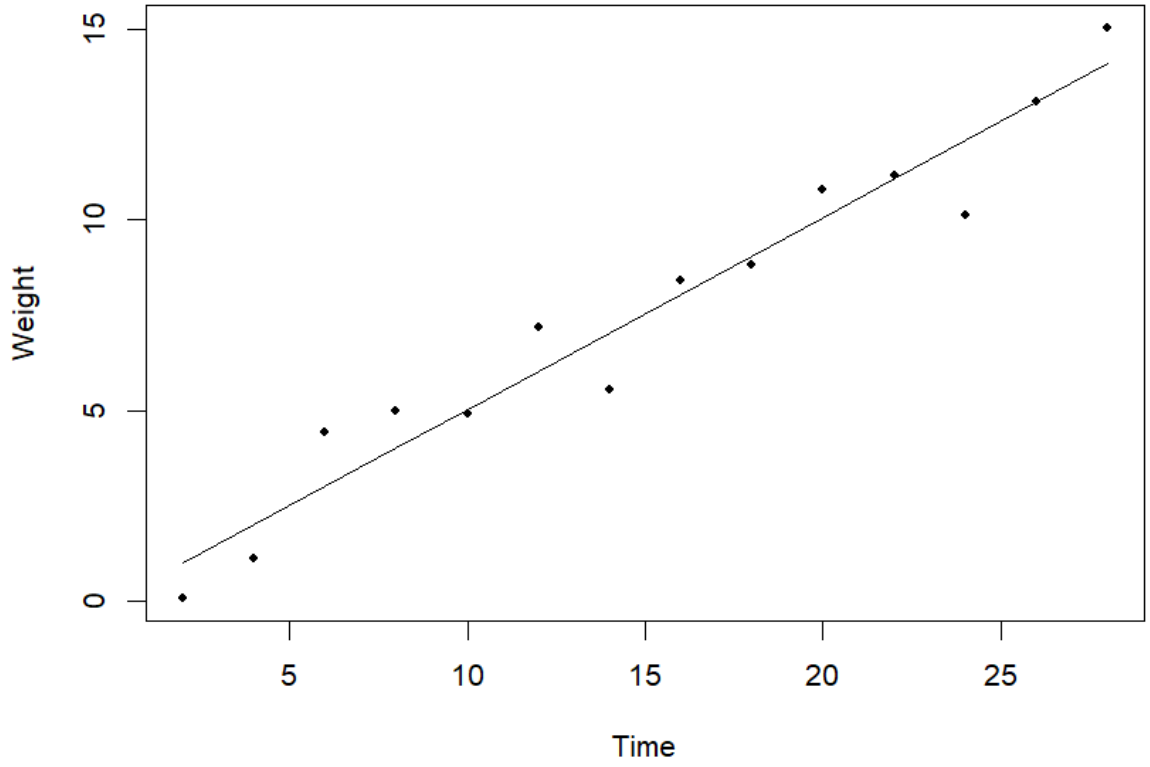
$$\widehat{\mu}_Y(x) \;=\; \bar{y} + \widehat{\beta}_1(x_i - \bar{x}) \;=\; \bar{y} - \widehat{\beta}_1 \bar{x} + \widehat{\beta}_1 x_i$$

  This is equivalent to what you get from (2).

- The text (p. 114) says "It can be mathematically proven..." but doesn't say how. Well, now you know how.

Example 1

"Crystal" data from the text (p. 119). Predict weight as a function of time.



Estimating the Parameters

```
> Cry=read.table('C:/Users/rls/aug20/UNC/STOR455/Data/Crystal.txt',header=T)
> x=Cry$Time
> y=Cry$Weight
> n=length(y)
> SSX=sum((x-mean(x))^2)
> SSY=sum((y-mean(y))^2)
> SXY=sum((x-mean(x))*(y-mean(y)))
> print(c(mean(y),SXY/SSX,mean(y)-mean(x)*SXY/SSX,sqrt((SSY-SXY^2/SSX)/(n-2))))

[1] 7.552857143 0.503428571 0.001428571 1.061766946
```

In algebra:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum\{(x_i-\bar{x})(y_i-\bar{y})\}}{\sum\{(x_i-\bar{x})^2\}} = 0.5034, \\ \bar{y} &= 7.5529, \\ \bar{y}-\widehat{\beta}_1\bar{x} &= 0.0014, \\ \widehat{\sigma} &= 1.0618. \end{aligned}$$

See the text, pp. 119 and 121.

Also...

Still writing the model as $\mu_Y(x) = \beta_0 + \beta_1(x - \bar{x})$:

$$SE(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{1.0618}{\sqrt{14}} = 0.2838,$$

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum\{(x_i - \bar{x})^2\}}} = \frac{1.0618}{\sqrt{910}} = 0.0352.$$

```
> # R code for the above
> x1=x-mean(x)
> lm1=lm(y~x1)
> summary(lm1)

Call:
lm(formula = y ~ x1)
Residuals:
     Min       1Q   Median       3Q      Max
-1.96371 -0.73464  0.05629  0.89193  1.40800
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5529     0.2838   26.62 4.85e-12 ***
x1            0.5034     0.0352   14.30 6.69e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.062 on 12 degrees of freedom
Multiple R-squared:  0.9446,Adjusted R-squared:   0.94
F-statistic: 204.6 on 1 and 12 DF,  p-value: 6.688e-09
```

Rest of this problem (pp. 120–121)

2 **If a crystal is allowed to grow for 15 hours, what is its predicted weight?**

$$\begin{aligned} \widehat{\mu}_Y(x) &= 0.0014 + 0.5034x, \\ \widehat{\mu}_Y(15) &= 0.0014 + 0.5034 \times 15 = 7.5524. \end{aligned}$$

Rest of this problem (pp. 120–121)

3 The crystals are priced depending on the time taken to grow them as well as their actual weight. Crystals that are grown for 8 hours or less are priced at \$2 per gram, those that are grown between 8 hours and 16 hours are priced at \$10 per gram, and those that are grown for more than 16 hours are priced at \$16 per gram. These prices reflect the additional amount of operator intervention necessary to grow crystals for longer periods. Estimate the additional dollars that a crystal will sell for if it is allowed to grow for 24 hours instead of 12 hours.

First do

$$\begin{aligned} \widehat{\mu}_Y(12) &= 0.0014 + 0.5034 \times 12 = 6.0422, \\ \widehat{\mu}_Y(24) &= 0.0014 + 0.5034 \times 24 = 12.0830. \end{aligned}$$

Estimated price of first crystal is $\$6.0422 \times 10 = \$60.42$.
Estimated price of second crystal is $\$12.0830 \times 16 = \$193.33$.
Difference is \$132.91.

Rest of this problem (pp. 120–121)

4 **An electronic components manufacturer places an order for 100 crystals weighing 12 grams each with a tolerance of ±0.5 gram, i.e., weighing between 11.5 and 12.5 grams. How long should the crystals be allowed to grow? If 100 crystals are grown for this amount of time, how many crystals may be expected to meet the specifications?**

1. Solve $0.0014 + 0.5034x = 12$, $x = 23.84$ hours.

2. If the weight of the crystal Y has a mean of $\mu = 12$ and a standard deviation $\sigma = 1.062$, then

$$\Pr\{11.5 \le Y \le 12.5\} = \Pr\left\{-\frac{0.5}{1.062} \le \frac{Y-\mu}{\sigma} \le \frac{0.5}{1.062}\right\}$$
$$= \texttt{pnorm}(0.4708) - \texttt{pnorm}(-0.4708) = 0.3622.$$

About 36 crystals will meet the specifications.

Mount Airy and Charleston Dataset

(a) What is the regression equation for predicting Charleston summer mean temperature from that in Mount Airy, and what is the estimated standard deviation?

(b) The mean summer temperature for one year in Mount Airy is 24^{o}C. Predict the mean summer temperature in Charleston, and calculate the probability that this is above 28^{o}C.

(c) My summer AC bill is \$100 per month if the average temperature is below 27^{o}C, \$120 per month if the average temperature is between 27^{o}C and 28^{o}C, and \$150 per month if the average temperature is above 28^{o}C. If the mean summer temperature in Mount Airy is 23^{o}C and I live in Charleston, what is my expected AC bill for the summer?

(d) If the mean summer temperature in Charleston is 27.5^{o}C, what is the expected mean summer temperature in Mount Airy?

Residuals

- Assume $y_i = \beta_0 + \beta_1 x_i + e_i$ or, alternatively, $y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + e_i$ (both models lead to the same e_i)

- Use first model: we have seen that we can estimate β_0 and β_1 by the *least squares estimators* $\hat{\beta}_0$ and $\hat{\beta}_1$, and this also leads to $\hat{\sigma}$ for the residual standard deviation

- Hence, we can estimate:

$$\hat{e}_i \;=\; y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

  and these $\hat{e}_i$'s are called *residuals*

- Residuals have many uses, but especially as a *diagnostic* for whether the model assumptions are correct.

Standardized Residuals

- Since we are assuming the e_i's have common standard deviation σ, the same will be approximately true of the $\widehat{e}_i$'s.
- Therefore, a natural step would be to divide each e_i by $\widehat{\sigma}$ so that they have approximate standard deviation 1.
- A more accurate approximation is to calculate

$$r_i = \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_i}}$$

where h_i is called the ith *hat value* (the text uses $h_{i,i}$ rather than h_i, but both notations are in common use).

- For a simple linear regression, a formula for h_i is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}$$

where $SSX = \sum(x_i - \bar{x})^2$ as in our earlier notation. Note that $\sum_i h_i = 2$.

Calculating Residuals and Standardized Residuals

- "by hand"..

```
lm1=lm(y~x)
res=y-lm1$coef[1]-lm1$coef[2]*x
sighat=sqrt(sum(res^2)/(n-2))
hat=rep(1/n,n)+(x-mean(x))^2/sum((x-mean(x))^2)
sres=res/(sighat*sqrt(1-hat))
```

- or use the tools built into R...

```
residuals(lm1)
summary(lm1)$sigma
hatvalues(lm1)
rstandard(lm1)
```

Plots

- What features are we looking for in a regression analysis?
  - Linear relationship
  - Constant variance
  - Normal distribution of e_i
  - No outliers (we hope ..)
- *Plotting the data* is a key way to assess these properties
  - Plot y against x
  - Plot the (standardized) residuals against x

FIGURE 3.5.1

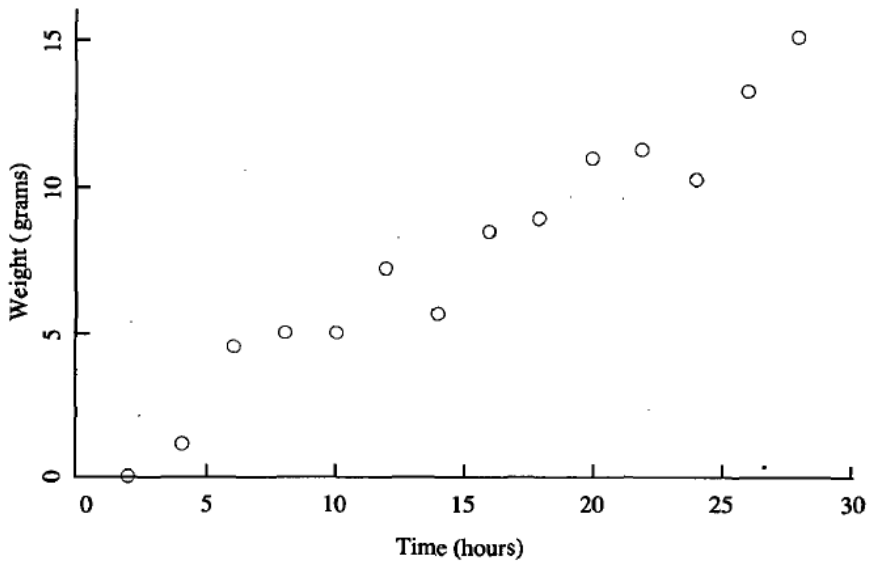


FIGURE 3.5.2

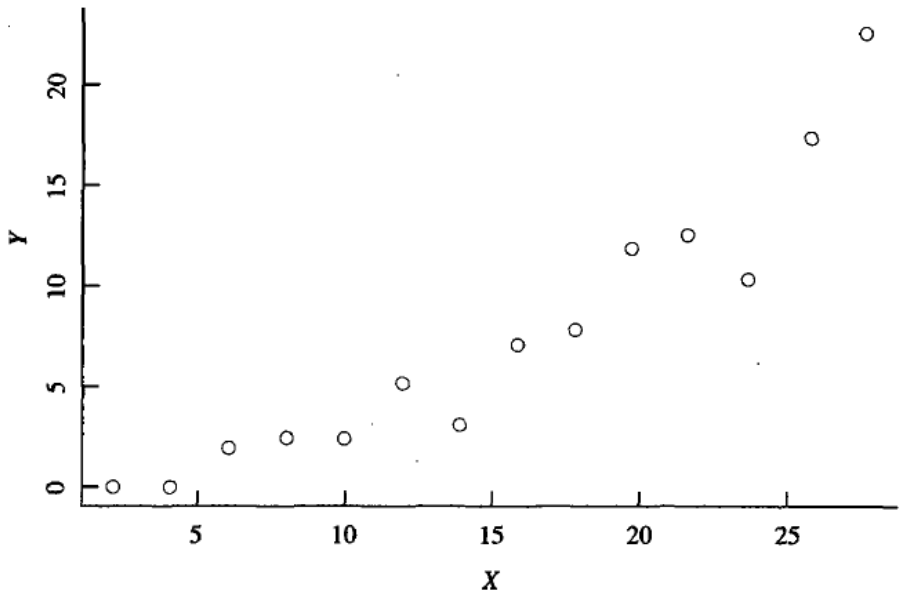


FIGURE 3.5.3

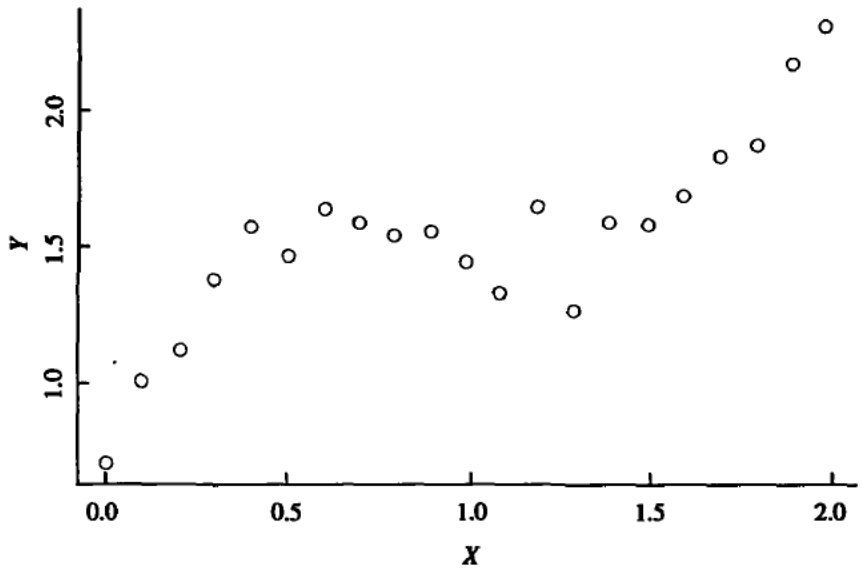


FIGURE 3.5.4

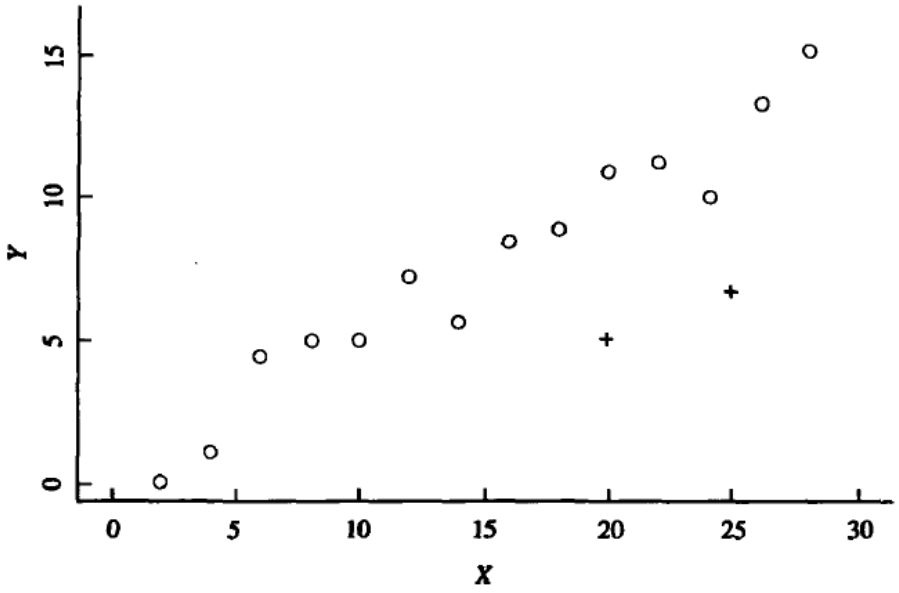


FIGURE 3.5.6

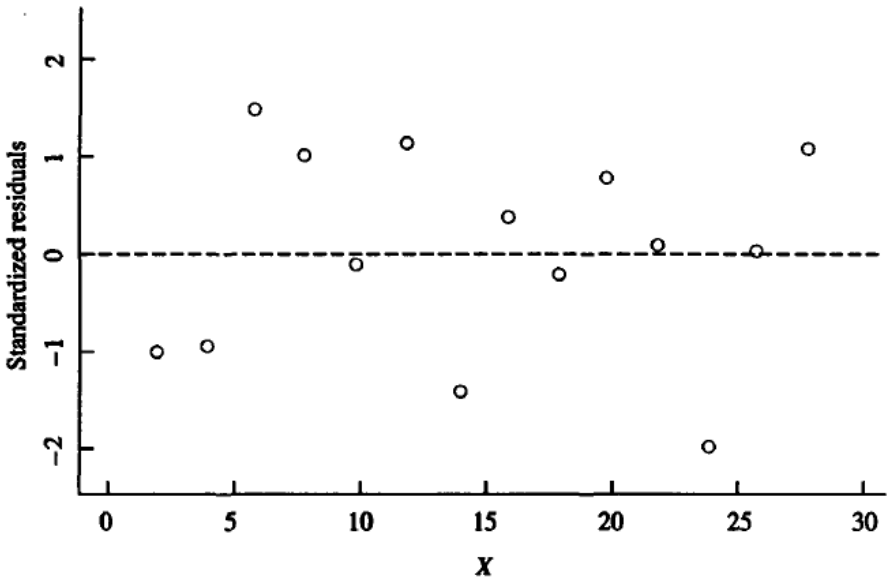


FIGURE 3.5.7

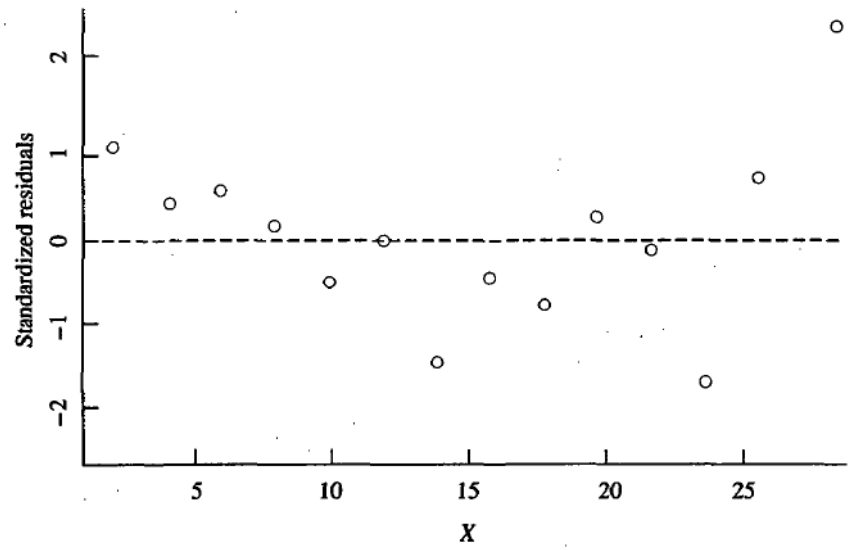


FIGURE 3.5.8

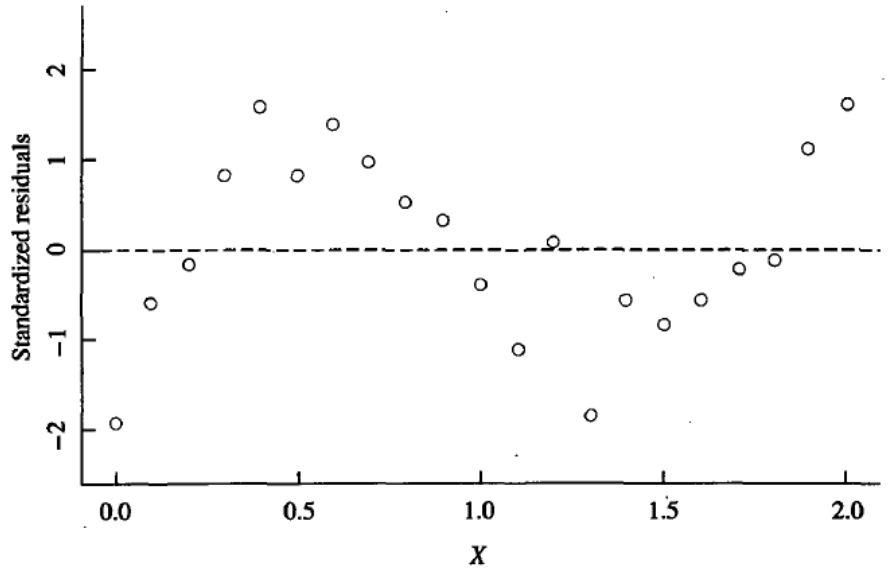
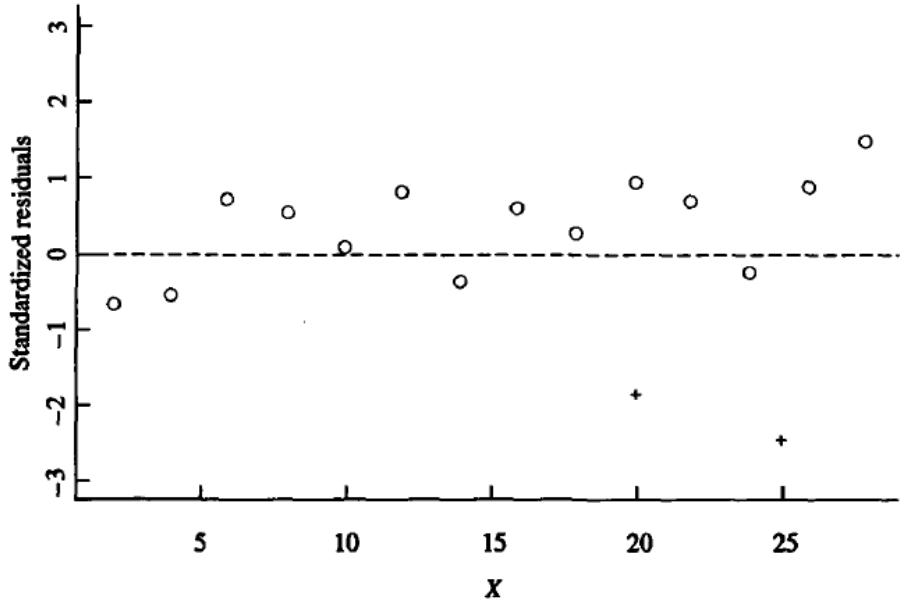


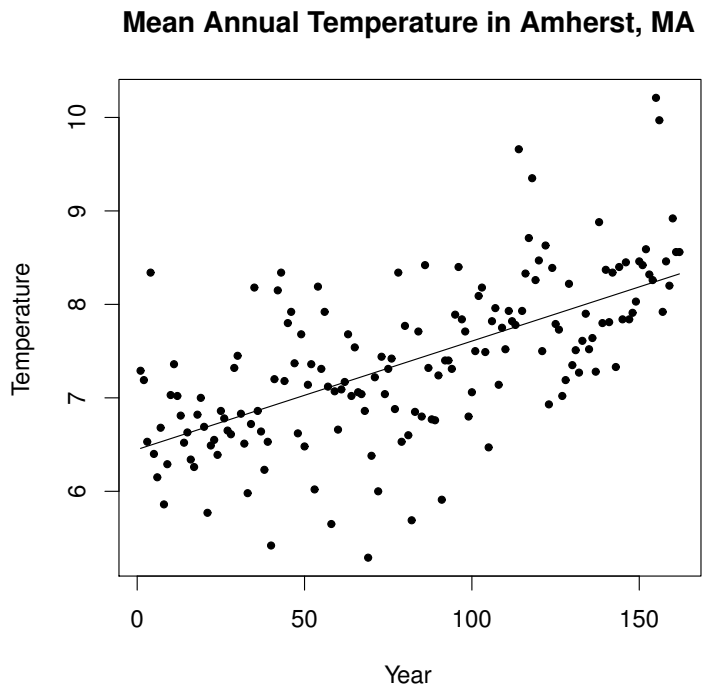
FIGURE 3.5.9

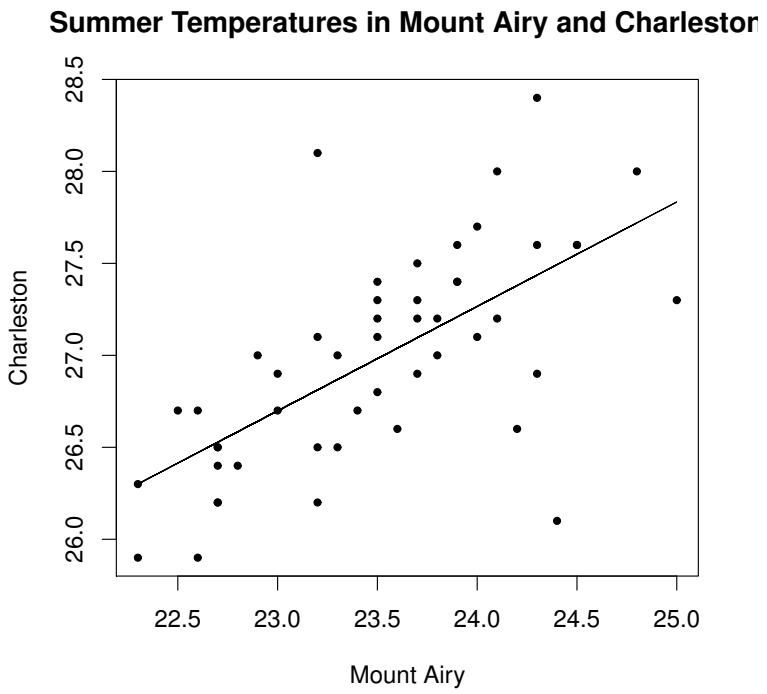


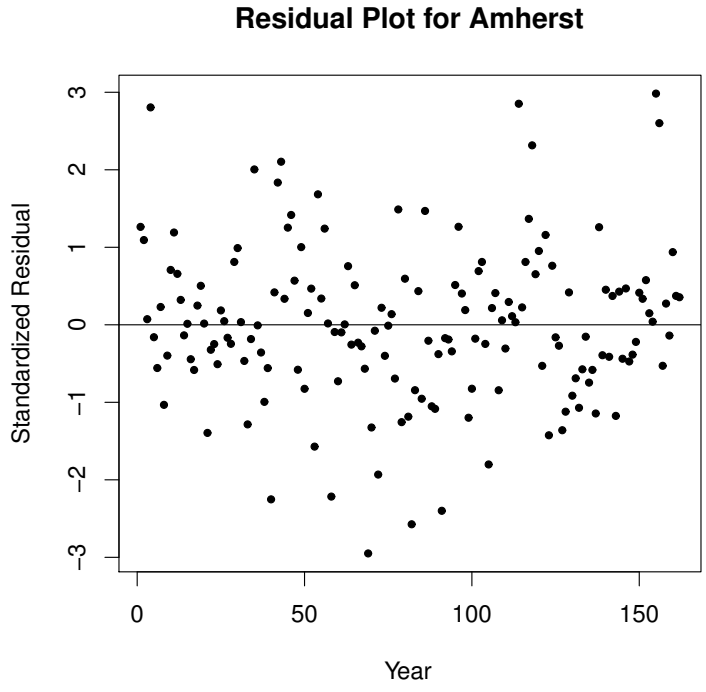
Conclusions from these Plots

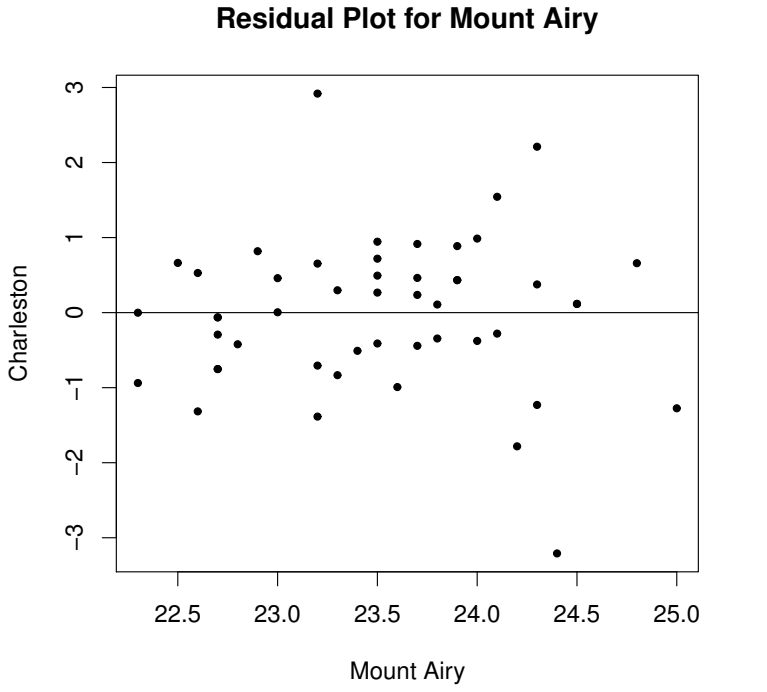
- Figs. 3.5.1/3.5.6 do look like a linear relationship
- Figs. 3.5.2/3.5.7 look nonlinear
- Figs. 3.5.3/3.5.8 look nonlinear
- Figs. 3.5.4/3.5.9 either show increasing variance or two clear outliers...
- In each case the shape is more clearly seen in the resodual plot than the original x, y plot
- Now let's look at the Amherst and Mount Airy datasets, e.g.

```
plot(Mta$MtAiry,rstandard(lm(Mta$Charleston~Mta$MtAiry)),xlab='Mount Airy',
ylab='Charleston',pch=20,main='Residual Plot for Mount Airy')
abline(0,0)
```


Mean Annual Temperature in Amherst, MA
Temperature
Year
6
7
8
9
10
0
50
100
150

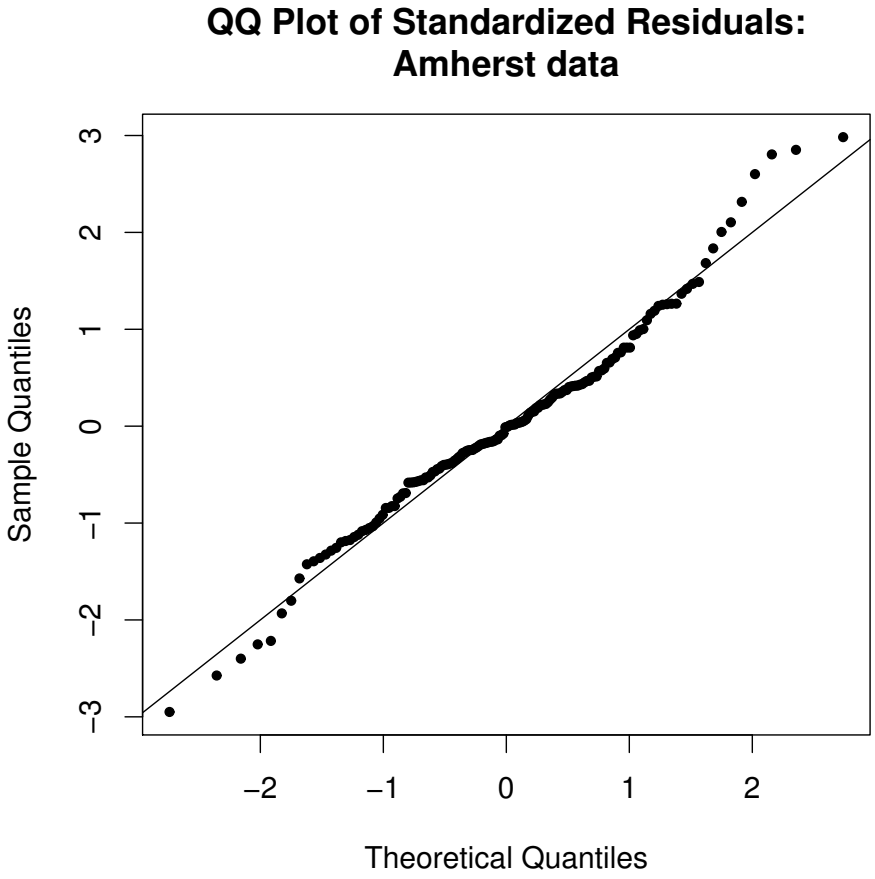

Summer Temperatures in Mount Airy and Charleston
Charleston
Mount Airy
26.0
26.5
27.0
27.5
28.0
28.5
22.5
23.0
23.5
24.0
24.5
25.0

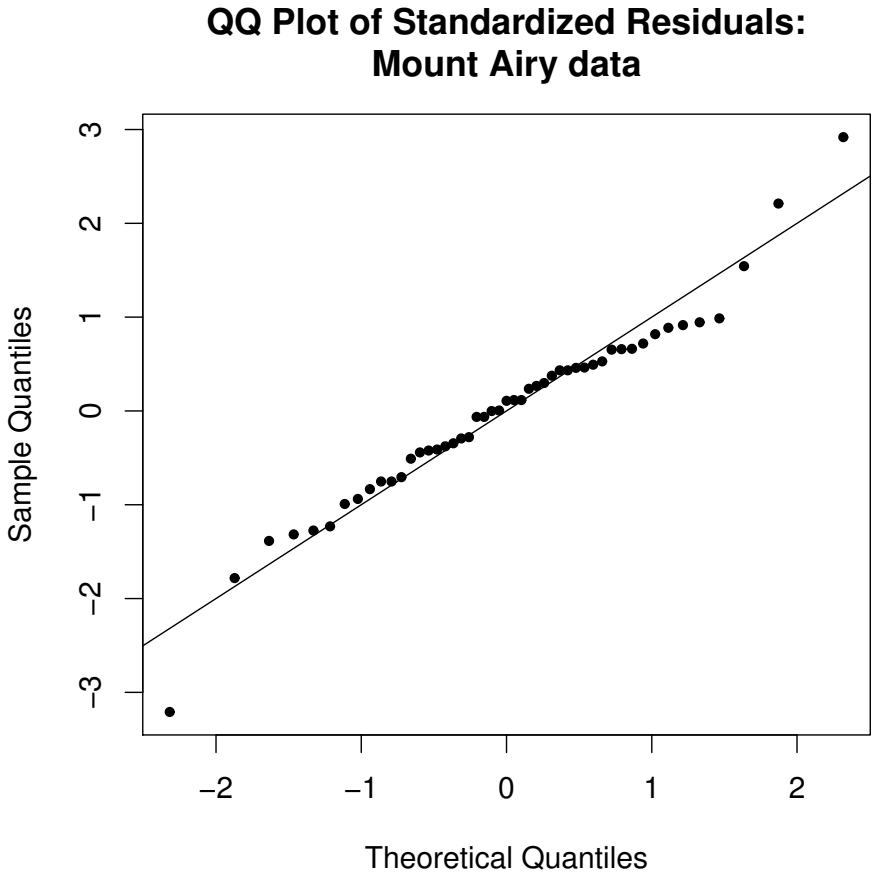

Residual Plot for Amherst
Standardized Residual
Year
−3
−2
−1
0
1
2
3
0
50
100
150


Residual Plot for Mount Airy
Charleston
Mount Airy
−3
−2
−1
0
1
2
3
22.5
23.0
23.5
24.0
24.5
25.0

Other Plotting Techniques

- Plot residuals against *fitted values* (`lm(...)$fitted`)
  - With only one x variable, this is no different conceptually from plotting against x
  - However, with multiple x variables, this technique will be very useful
- QQ plots of (standardized) residuals
  - Also called normal probability plots, rankit plots, and various other names
  - For formal definition, see p. 146 of text
  - `qqplot` in R
  - An extension (p. 147 of text): do this for various combinations of x and y values to test for a bivariate normal distribution

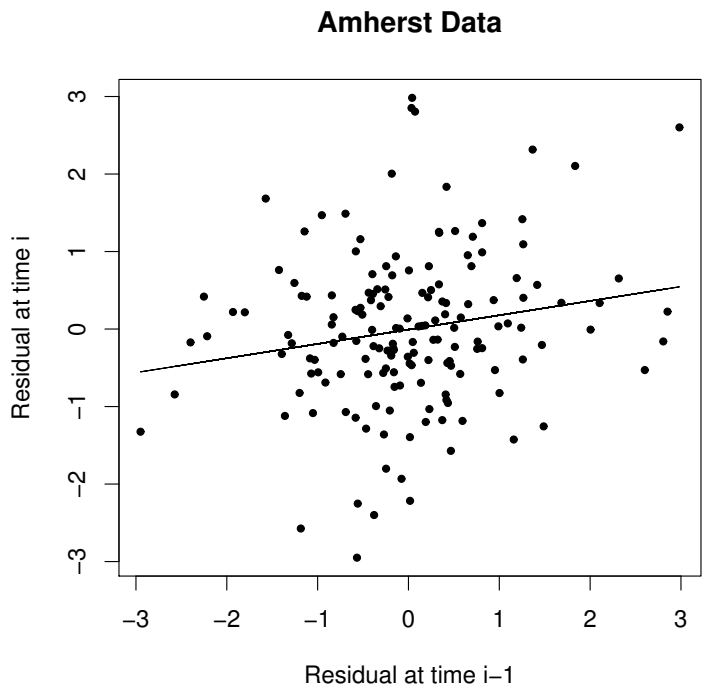

QQ Plot of Standardized Residuals:
Amherst data
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2

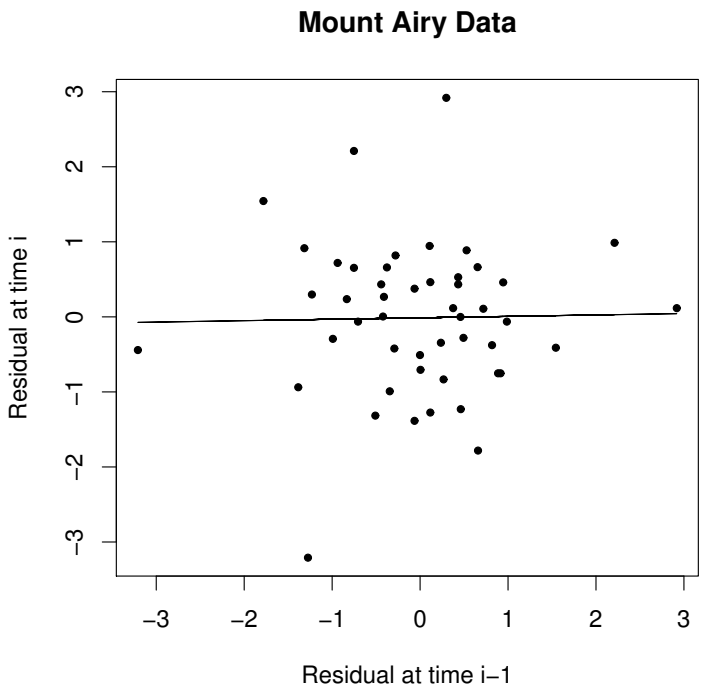

QQ Plot of Standardized Residuals:
Mount Airy data
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2

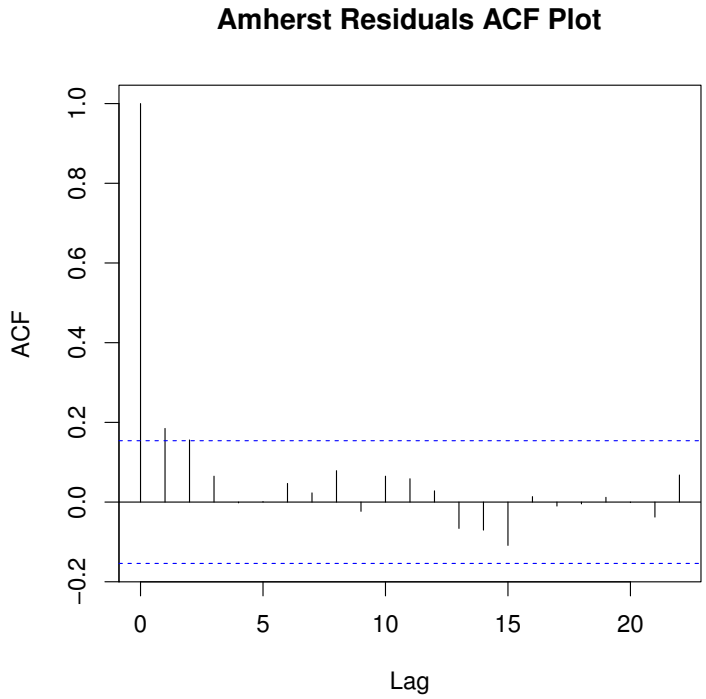
Other Plotting Techniques (continued)

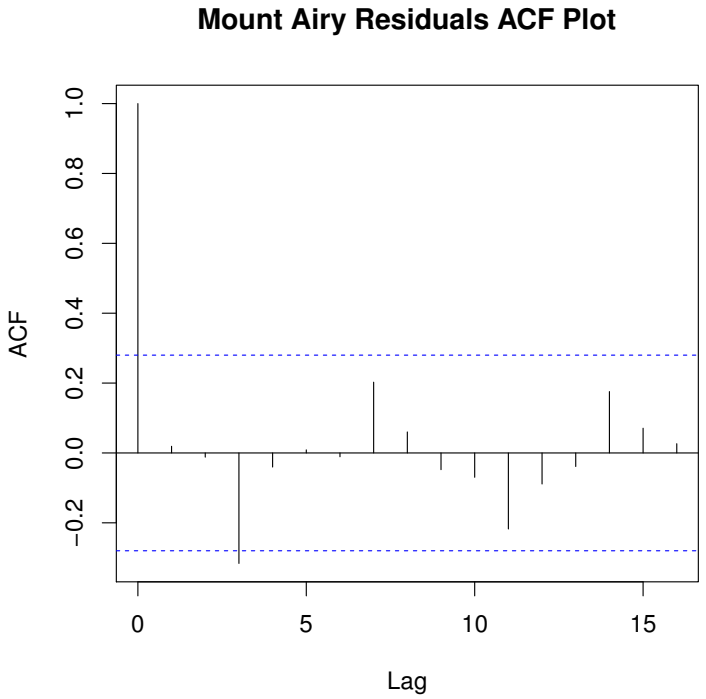
Not mentioned in the text, but these are also common plots people make:

- Plotting the residual at time i against the residual at time $i-1$ (looking for autocorrelation)
- The full "ACF" plot at many lags (R function `acf`)
- Durbin-Watson test (function `dwtest` within library `lmtest`)
- For Amherst data, the ACF plot is not strong but the DW test clearly rejects the null hypothesis of no autocorrelation (p=0.006)
- The Mount Airy data shows no evidence of autocorrelation
- These are really "time series" datasets. You'll learn much more about time series if you take STOR 556.


Amherst Data
Residual at time i
Residual at time i−1
−3
−2
−1
0
1
2
3


Mount Airy Data
Residual at time i
Residual at time i−1
−3
−2
−1
0
1
2
3


Amherst Residuals ACF Plot
ACF
Lag
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


Mount Airy Residuals ACF Plot
ACF
Lag
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15

Confidence Intervals, Prediction Intervals and Hypothesis Tests

- Write model in form $y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + e_i$, assumes least squares estimates $\widehat{\beta}_0^* = \bar{y},\ \widehat{\beta}_1^* = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$.
- $\mathrm{Var}(\widehat{\beta}_0^*) = \frac{\sigma^2}{n}$, $\mathrm{Var}(\widehat{\beta}_1^*) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$.
- It's also possible to show that $\widehat{\beta}_0^*$ and $\widehat{\beta}_1^*$ are *uncorrelated*
- Hence, for any a and b, $\mathrm{Var}(a\widehat{\beta}_0^* + b\widehat{\beta}_1^*) = \sigma^2 \left(\frac{a^2}{n} + \frac{b^2}{\sum(x_i - \bar{x})^2} \right)$.
- If σ is unknown, then

$$\frac{a\widehat{\beta}_0^* + b\widehat{\beta}_1^*}{\widehat{\sigma}\sqrt{\frac{a^2}{n} + \frac{b^2}{\sum(x_i - \bar{x})^2}}} \sim t_{n-2}.$$

Application to Confidence Intervals

- Example: Find a $100(1-\alpha)\%$ confidence interval for $\mu_Y(x) = \beta_0 + \beta_1 x$ for a given value of x.
- This problem is actually easier if you write it in the alternative format: $\mu_Y(x) = \beta_0^* + \beta_1^*(x - \bar{x})$
- $\widehat{\mu}_Y(x) = \widehat{\beta}_0^* + \widehat{\beta}_1^*(x - \bar{x})$ and

$$\frac{\widehat{\mu}_Y(x) - \mu_Y(x)}{s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}} \sim t_{n-2}$$

- Therefore, an appropriate confidence band for $\mu_Y(x)$ is

$$\widehat{\mu}_Y(x) \pm qt\left(1 - \frac{\alpha}{2}, n-2\right)\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}$$

- This is on p. 161 of the text.

Prediction Intervals

- I've been asked a number of times by students in the class: what is the distinction between $Y(x)$ and $\mu_Y(x)$? (see, e.g., question 3.4.4(b) from one of the homework assigments)
- The distinction is whether we are talking about an individual observation or an average over many observations
  - In the question about car maintenance costs:
  - $Y(13000)$ is the maintenance expense for *my* car, if I drive it 13,000 miles in the first year
  - $\mu_Y(13000)$ is the average maintenance expense over *all* cars that are driven 13,000 miles in the first year
  - My interest is surely in $Y(13000)$, not $\mu_Y(13000)$ (unless I want the latter number for comparison)
- Both quantities have the same point estimate
  - In this analysis, we write the model as $\mu_Y(x) = \beta_0^* + \beta_1^*(x - \bar{x})$
  - $\widehat{Y}(x)$ and $\widehat{\mu}_Y(x)$ are both $\widehat{\beta}_0^* + \widehat{\beta}_1^*(x - \bar{x})$ where $\widehat{\beta}_0, \widehat{\beta}_1$ are the least squares estimates. Note that $\bar{x}$ refers to the mean x values of the observations that were used to form the estimates — it's not updated to include the new x
  - However the *variability* of $Y(x)$ and $\mu_Y(x)$ are very different — this is our main focus here

Computing Interval Estimates

- $\widehat{\mu}_Y(x) - \mu_Y(x) = (\widehat{\beta}_0^* - \beta_0^*) + (\widehat{\beta}_1^* - \beta_1^*)(x - \bar{x})$
- Variance is $\sigma^2 \left\{ \frac{1}{n} + \frac{(x-\bar{x})^2}{SSX} \right\}$
- Therefore, a $100(1-\alpha)\%$ confidence interval for $\mu_Y(x)$ is

$$\widehat{\mu}_Y(x) \pm qt\left(1 - \frac{\alpha}{2}, n-2\right) \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}$$

- $\widehat{Y}(x) - Y(x) = (\widehat{\beta}_0^* - \beta_0^*) + (\widehat{\beta}_1^* - \beta_1^*)(x - \bar{x}) + e$
- e has mean 0 and SD= σ, ind. of past $e_1, \ldots, e_n$.
- Variance is $\sigma^2 \left\{ \frac{1}{n} + \frac{(x-\bar{x})^2}{SSX} + 1 \right\}$
- Therefore, a $100(1-\alpha)\%$ *prediction* interval for $Y(x)$ is

$$\widehat{\mu}_Y(x) \pm qt\left(1 - \frac{\alpha}{2}, n-2\right) \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX} + 1}$$

Example

- Redo crystal example
- What are the confidence and prediction interval for Weight if Time is (a) 9 hours, (b) 15 hours, (c) 21 hours?

```
Cry=read.table('.../Crystal.txt',header=T)
lm1=lm(Weight~Time,Cry)
Time=c(9,15,21)
Weight=rep(NA,3)
Crynew=data.frame(Weight,Time)
# Crynew has same structure as Cry
predict.lm(lm1,newdata=Crynew,interval='confidence',level=0.95)
predict.lm(lm1,newdata=Crynew,interval='prediction',level=0.95)
predict.lm(lm1,newdata=Crynew,interval='prediction',level=0.99)
```

Results

```
> predict.lm(lm1,newdata=Crynew,interval='confidence',level=0.95)
       fit      lwr       upr
1  4.532286 3.761579  5.302992
2  7.552857 6.934577  8.171137
3 10.573429 9.802722 11.344135
> predict.lm(lm1,newdata=Crynew,interval='prediction',level=0.95)
       fit      lwr       upr
1  4.532286 2.093891  6.970680
2  7.552857 5.158270  9.947445
3 10.573429 8.135034 13.011823
> predict.lm(lm1,newdata=Crynew,interval='prediction',level=0.99)
       fit      lwr       upr
1  4.532286 1.113831  7.950741
2  7.552857 4.195817 10.909898
3 10.573429 7.154974 13.991883
```

Simultaneous Confidence/Prediction Intervals

- Suppose we have K values of x, denoted $x_1, \ldots, x_K$.
- Simultaneous CIs: find bounds $L_k, U_k,\ 1 \le k \le K$, such that

$$\Pr\{L_k \le \mu_Y(x_k) \le U_k \text{ for } k = 1, \ldots, K\} \ge 1 - \alpha.$$

- Expect solution of form

$$\left\{ \begin{array}{c} U_k \\ L_k \end{array} \right\} = \hat{\mu}_Y(x_k) \pm t_K^* \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{SSX}}$$

- Simultaneous PIs: find bounds $L_k, U_k,\ 1 \le k \le K$, such that

$$\Pr\{L_k \le Y(x_k) \le U_k \text{ for } k = 1, \ldots, K\} \ge 1 - \alpha.$$

- Expect solution of form

$$\left\{ \begin{array}{c} U_k \\ L_k \end{array} \right\} = \hat{\mu}_Y(x_k) \pm t_K^* \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{SSX} + 1}$$

- What should we use for t_K^*?

Formulas for t_K^*

- Bonferroni: $t_K^* = qt\left(1 - \frac{\alpha}{2K}, n-2\right)$
- For *prediction intervals*, this is (nearly always) the best theoretical formula, though it might be possible to do better by simulation
- For *confidence intervals*, there is (nearly always) a better result: $t_K^* = \sqrt{2 \cdot qf(1-\alpha, 2, n-2)}$ where qf is a quantile of the *F distribution*
- The formal definition of $F_{m,n}$ is that it is the distribution of $\frac{U/m}{V/n}$ where $U \sim \chi^2_m$, $V \sim \chi^2_n$ are independent chi-square
- In R: functions $pf(x, m, n)$ or $qf(p, m, n)$ for the distribution and quantile functions
- This is known as the *Working-Hotelling* procedure. It's a special case of *Scheffé's method.*

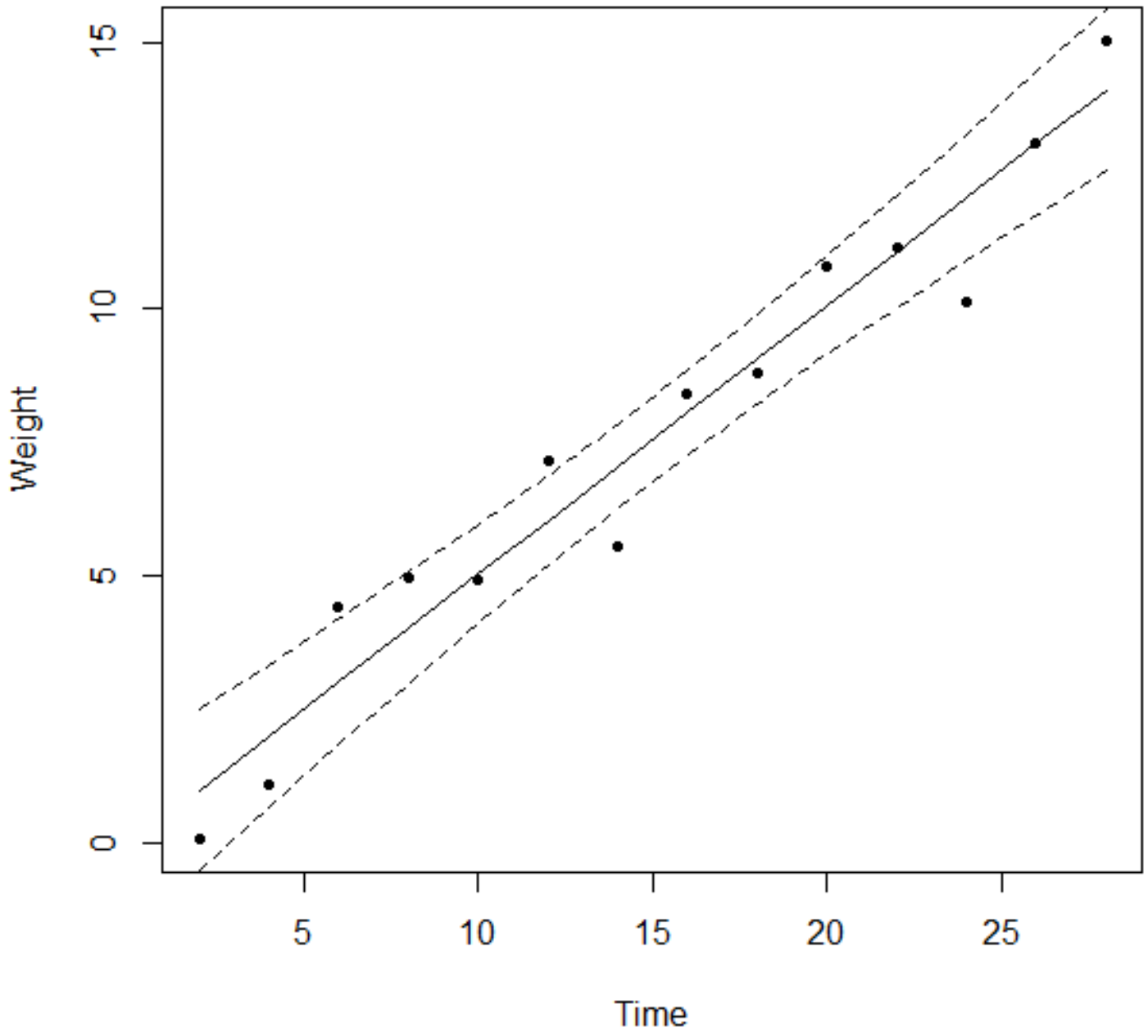
Quick Note in Passing

If $T \sim t_n$ then $T^2 \sim F_{1,n}$.

We'll use this later (Section 3.8)

Example based on Crystal Dataset

```
Cry=read.table('.../Crystal.txt',header=T)
lm1=lm(Weight~Time,Cry)
n=nrow(Cry)
tstar=sqrt(2*qf(0.95,2,n-2))
SSX=sum((Cry$Time-mean(Cry$Time))^2)
y1=lm1$coef[1]+lm1$coef[2]*Cry$Time
y2=y1+tstar*summary(lm1)$sigma*sqrt((Cry$Time-mean(Cry$Time))^2/SSX+1/n)
y3=y1-tstar*summary(lm1)$sigma*sqrt((Cry$Time-mean(Cry$Time))^2/SSX+1/n)
plot(Cry$Time,Cry$Weight,pch=20,ylab='Weight',xlab='Time')
lines(Cry$Time,y1)
lines(Cry$Time,y2,lty=2)
lines(Cry$Time,y3,lty=2)
```


Weight
0
5
10
15
5
10
15
20
25
Time

Alternatively ...

```
library(investr)
plotFit(lm1, interval = 'confidence', adjust = 'Scheffe',
main = 'Working-Hotelling Procedure for Crystal Data')
plotFit(lm1, interval = 'prediction', adjust = 'Bonferroni', k=14,
main = 'Simultaneous PIs for Crystal Data')
```

Note: there seems to be an error on the webpage of Aaron Schlegel (where I found this) — wrong value of k or K

Who were Working and Hotelling?

Holbrook Working
(1895–1985)

Harold Hotelling
(1895–1973)

Hypothesis Testing (Section 3.7)

All but one of the examples can be subsumed in the following:

- Write the model in the form $\mu_Y(x) = \beta_0^* + \beta_1^*(x - \bar{x})$
- $\widehat{\mu}_Y(x) = \widehat{\beta}_0^* + \widehat{\beta}_1^*(x - \bar{x})$, $\mathsf{SE}(\widehat{\mu}_Y(x)) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}}$
- $\frac{\widehat{\mu}_Y(x) - \mu_Y(x)}{SE(\widehat{\mu}_Y(x))} \sim t_{n-2}$
- To test $H_0: \ \mu_Y(x) = \mu_0$ versus $H_1: \ \mu_Y(x) \neq \mu_0$ at significance level α,
  - *Either:* Calculate $C = qt\left(1 - \frac{\alpha}{2}, n-2\right) \cdot SE(\widehat{\mu}_Y(x))$, reject H_0 if $|\widehat{\mu}_Y(x) - \mu_Y(x)| > C$,
  - *Or:* Calculate $P = 2 \cdot pt\left(\frac{|\widehat{\mu}_Y(x) - \mu_Y(x)|}{SE(\widehat{\mu}_Y(x))}, n-2, \texttt{lower.tail} = \texttt{F}\right)$, reject H_0 if $P < \alpha$.

Comments and Extensions

- The text again emphasizes that confidence intervals are more useful than hypothesis tests and I'd (broadly) agree with that. Nevertheless, you need to know how to do hypothesis tests.
- A particular special case: $H_0: \ \beta_1 = 0$ versus $H_1: \ \beta_1 \neq 0$
- In that case you reject H_0 when

$$\left| \frac{\widehat{\beta}_1^* \sqrt{SSX}}{\widehat{\sigma}} \right| \ > \ qt\left(1 - \frac{\alpha}{2}, n-2\right).$$

  We'll see this again in Section 8.
- One-sided tests: what do we do differently if H_1 is either $\mu_Y(x) > \mu_0$ or $\mu_Y(x) < \mu_0$?

(Q3.7.1, HW6)

The regression function of shelf-life Y on storage temperature X is assumed to be a straight line $\mu_Y(x) = \beta_0 + \beta_1 x$ for values of x in the range 10°C to 35°C, and assumptions (A) are presumed to be satisfied.

a Define an appropriate target population for this investigation.

b Define an appropriate study population for this investigation.

c Are the data in this investigation obtained by simple random sampling or by sampling with preselected X values?

d Plot y_i versus x_i. Examine this plot and decide whether a straight line regression model seems reasonable.

e The director of the laboratory wants to determine whether the data provide evidence (at the 0.05 level) indicating that shelf-life does indeed depend on storage temperature, so he decides to use a statistical test. State an appropriate pair of hypotheses, suitably designating one as the null hypothesis and the other as the alternative hypothesis, and calculate the P-value for this test. What is your conclusion?

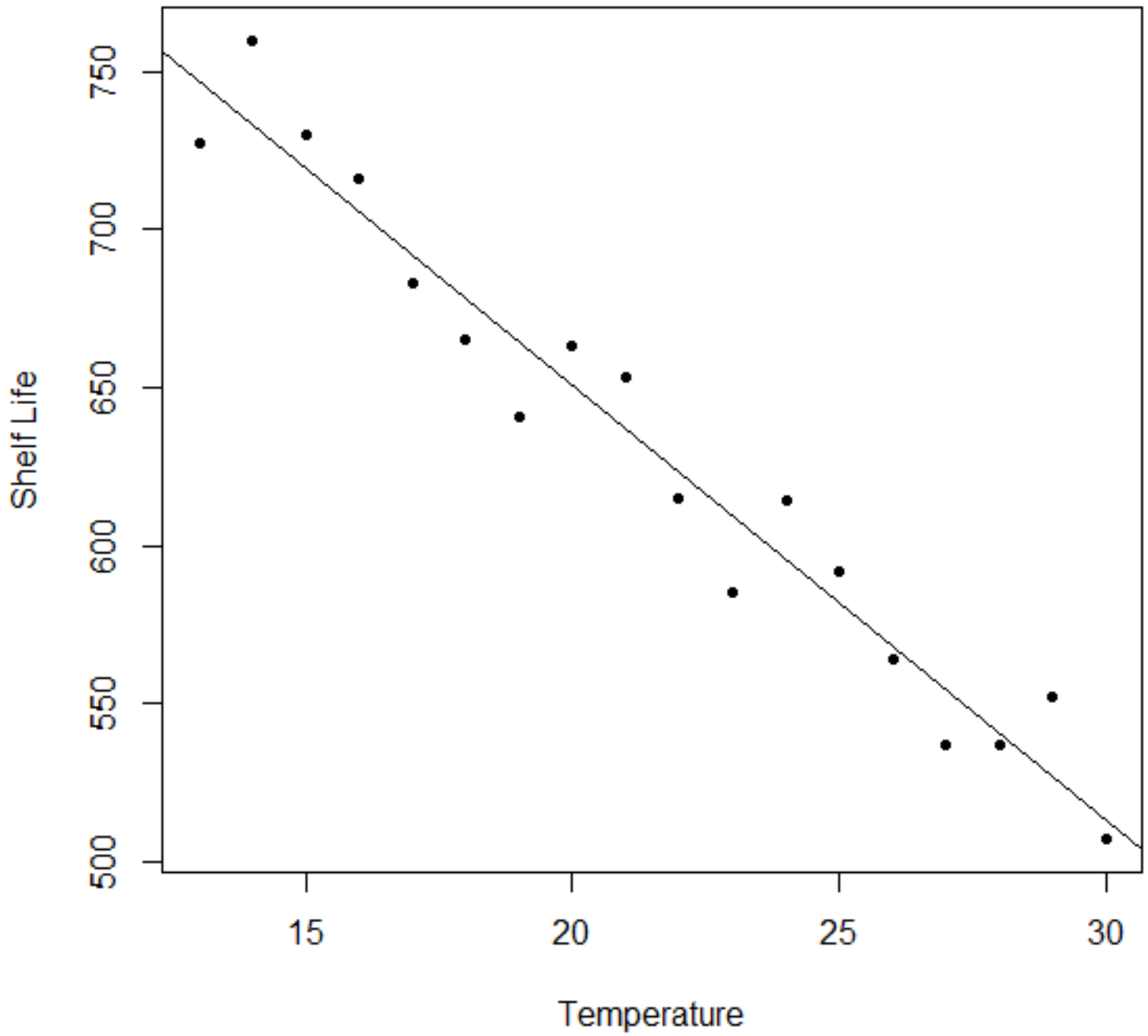
f Estimate, if possible, the average shelf-life for this cough syrup if it is to be stored at 0°C.

g Estimate the average shelf-life for this cough syrup if it is to be stored at 15°C. Also compute a 95% confidence interval for this quantity.

h Answer part (e) using an appropriate confidence interval instead of a hypothesis test.

i Do the data provide evidence (at the 0.05 level) indicating that the average shelf-life for bottles of cough syrup stored at 13°C is at least 650 days? Carry out an appropriate statistical test and state your conclusions.

j Construct an appropriate confidence interval to answer part (i).


Shelf Life
750
700
650
600
550
500
15
20
25
30
Temperature

Analysis of Variance (Section 1.8)

- Recall from Slide 28:

$$\sum\{y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \bar{x})\}^2 \;=\; SSY - \frac{SXY^2}{SSX} \;=\; SSY - \hat{\beta}_1^2 \cdot SSX.$$

- The expression $\sum\{y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - \bar{x})\}^2$ is called the *error sum of squares*, abbreviated SSE. Also, $\hat{\beta}_1^2 \cdot SSX$ is called the *sum of squares due to regression*, abbreviated SSR. Therefore, we have shown

$$SSY \;=\; SSR + SSE.$$

The Analysis of Variance Table

TABLE 3.8.1
ANOVA for Straight Line Regression

Source	Degrees of Freedom *df*	Sum of Squares *SS*	Mean square *MS*	Computed *F*-Value
Regression	1	*SSR*	*MSR*	$F_C = \frac{MSR}{MSE}$
Error	$n-2$	*SSE*	*MSE*	
Total	$n-1$	*SSY*	*MSY*	

- The statistic F_c has an $F_{1,n-2}$ distribution when $\beta_1 = 0$.
- Reject H_0 at significance level α when $F_c > qf(1-\alpha, 1, n-2)$.

Connecting the Dots

- Slide 64: Reject H_0 when $\left|\frac{\hat{\beta}_1^* \sqrt{SSX}}{\hat{\sigma}}\right| > qt\left(1-\frac{\alpha}{2}, n-2\right)$.
- $F_C = \frac{MSR}{MSE} = \frac{SSR}{(SSE/n-2)} = \frac{(\hat{\beta}_1^*)^2 SSX}{\hat{\sigma}^2}$.
- When $H_0: \ \beta_1 = 0$ is true,
  (a) $\frac{\hat{\beta}_1^* \sqrt{SSX}}{\hat{\sigma}} \sim t_{n-2}$,
  (b) $F_C = \frac{(\hat{\beta}_1^*)^2 SSX}{\hat{\sigma}^2} \sim F_{1,n-2}$.
  (c) But we already noted (slide 58) that the square of a t_{n-2} distribution is $F_{1,n-2}$.
  (d) They are the same test!

(Q3.8.1, HW6)

The following questions refer to the shelf-life data in Table 3.7.1, which are also stored in the file **shelflif.dat** on the data disk.

a Present an analysis of variance table.

b Use F_C from the ANOVA table in part (a) to test NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$. What is the P-value for this test? Interpret the result.

c Calculate t_C for testing NH against AH in part (b). What is the P-value for this test? Interpret the result.

d Verify that the square of t_C in part (c) is equal to F_C in (b). Further verify that the P-value calculated from the t statistic in part (c) is the same as that calculated from the F statistic in part (b).

e What conclusion do you draw regarding β_1 based on the test in part (b)?

f Compute a 99% confidence interval for β_1. How will you use this confidence interval to *decide* whether or not β_1 is close enough to zero to be considered negligible for this problem?

g Write a short paragraph outlining your conclusions in parts (b)–(f) and give reasons for your statements.