3.6 Confidence and Prediction Intervals

► A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y.

- Let X_h denote the level of X for which we wish to estimate the mean response.

- Then mean response when $X = X_h$ is denoted by $E(Y_h)$ or $\hat{Y}_h$and the point estimate of $E(Y_h)$ is:

$$\hat{Y}_h = b_0 + b_1 X_h$$

- The sampling distribution of $\hat{Y}_h$ is

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h \ , \ \sigma\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\right).$$

- 100(1-α)% confidence intervals for the mean response when $X = X_h$ is:

$$\left(\hat{Y}_h - t_{[\alpha/2]}^{(n-2)} s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\ ,\right.$$

$$\left.\hat{Y}_h + t_{[\alpha/2]}^{(n-2)} s\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}\right)$$

- To obtain 99% CI for the mean response:

```
> muresp3.1 <- predict(results3.1, interval="confidence",
  level=.99 )

> muresp3.1
           fit       lwr       upr
  1  12.256049 10.912788 13.599311
  2  12.256049 10.912788 13.599311
  3  11.680402 10.545978 12.814825
  4  10.848911  9.932931 11.764890
  5   9.966251  9.099728 10.832774
  6   8.443982  7.204640  9.683324
  7   8.405606  7.152173  9.659038
  8   7.842750  6.368684  9.316817
```

- Plotting a scatter plot with fitted line and confidence interval:

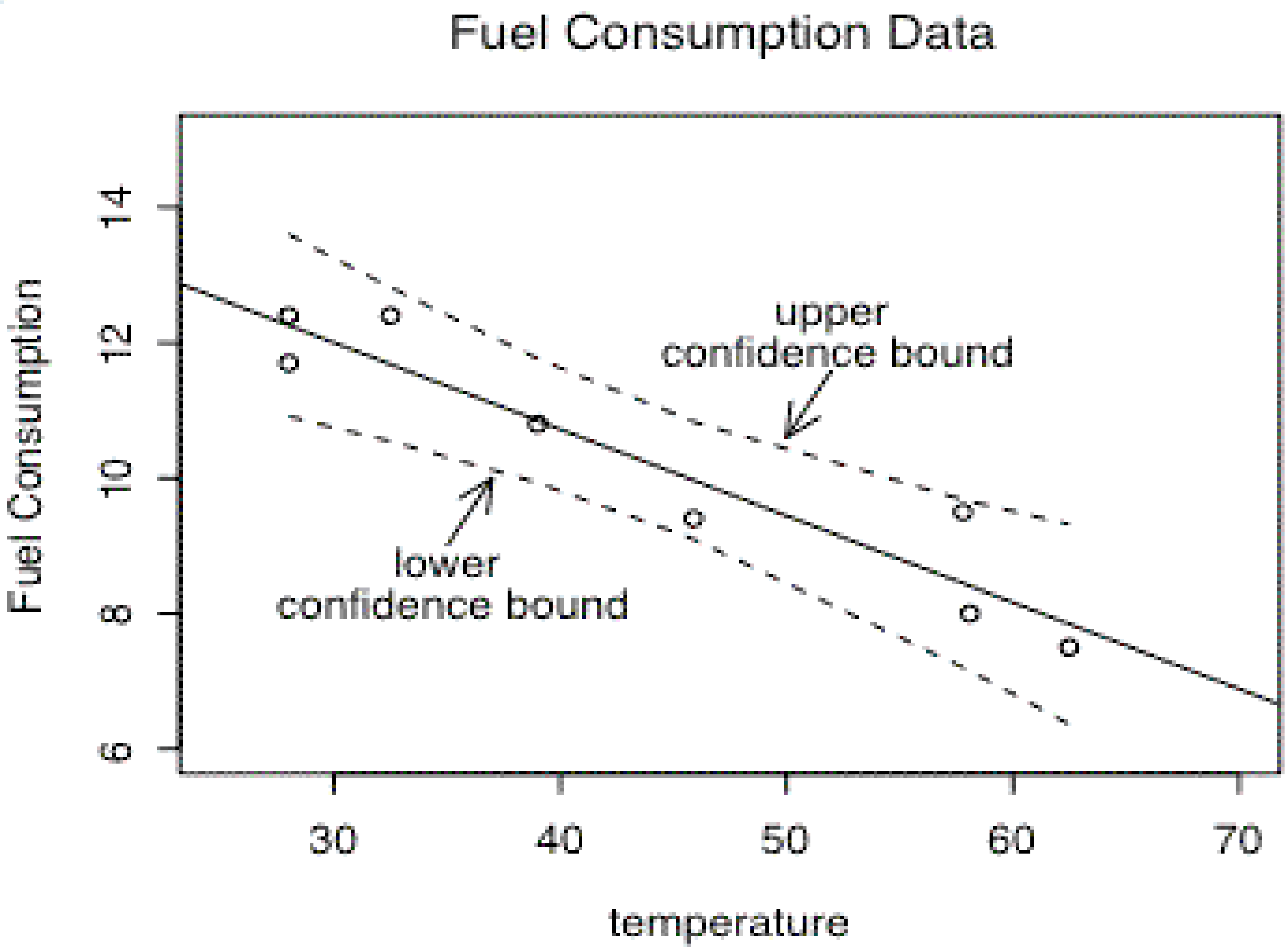
```
> plot(data3.1, main="Fuel Consumption Data ",
  xlab="temperature", ylab="Fuel Consumption")
```

```
> abline(coef(results))
```

  - abline(a,b): Add a line with intercept a and slope b to an existing plot

```
> lines(data3.1[,"Temp"], muresp3.1[,2])
```

```
> lines(data3.1[,"Temp"], muresp3.1[,3])
```


Fuel Consumption Data
upper
confidence bound
lower
confidence bound
Fuel Consumption
6
8
10
12
14
30
40
50
60
70
temperature

► The prediction of a new observation Y corresponding to a given level X of the predictor variable is viewed as the result of a new trial, independent of the trials on which the regression analysis is based.

- In the estimation of the mean response, we estimate the mean of the distribution of Y. In the prediction of a new observation , we predict an individual outcome drawn from the distribution of Y.

- Then prediction of a new observation when $X = X_h$ is denoted by $E(Y_h(new))$ or $\hat{Y}_h(new)$ and the point estimate of $E(Y_h(new))$ is:

$$\hat{Y}_h(new) = b_0 + b_1 X_h$$

- The sampling distribution of $\hat{Y}_h$ is

$$\hat{Y}_h \sim N\left(\beta_0 + \beta_1 X_h\ ,\ \sigma\sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right).$$

- Two sources of variations in the standard deviation of the prediction:

1. Variation in possible location of the distribution of Y
2. Variation within the probability distribution of Y

    ▷ Note that the first source is the only source of variations for estimating the mean response.

- 100(1-α)% prediction interval for an individual value of Y when $X = X_h$ is:

$$\left(\hat{Y}_h - t_{[\alpha/2]}^{(n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\,, \right.$$

$$\left. \hat{Y}_h + t_{[\alpha/2]}^{(n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \right)$$

- To obtain 99% PI for a new observation:

```
> pred3.1 <- predict(results3.1, interval="prediction",
  level=.99 )

> pred3.1
           fit      lwr      upr
   1 12.256049 9.483493 15.02861
   2 12.256049 9.483493 15.02861
   3 11.680402 9.002784 14.35802
   4 10.848911 8.256279 13.44154
   5  9.966251 7.390677 12.54182
   6  8.443982 5.720256 11.16771
   7  8.405606 5.675439 11.13577
   8  7.842750 5.004513 10.68099
```

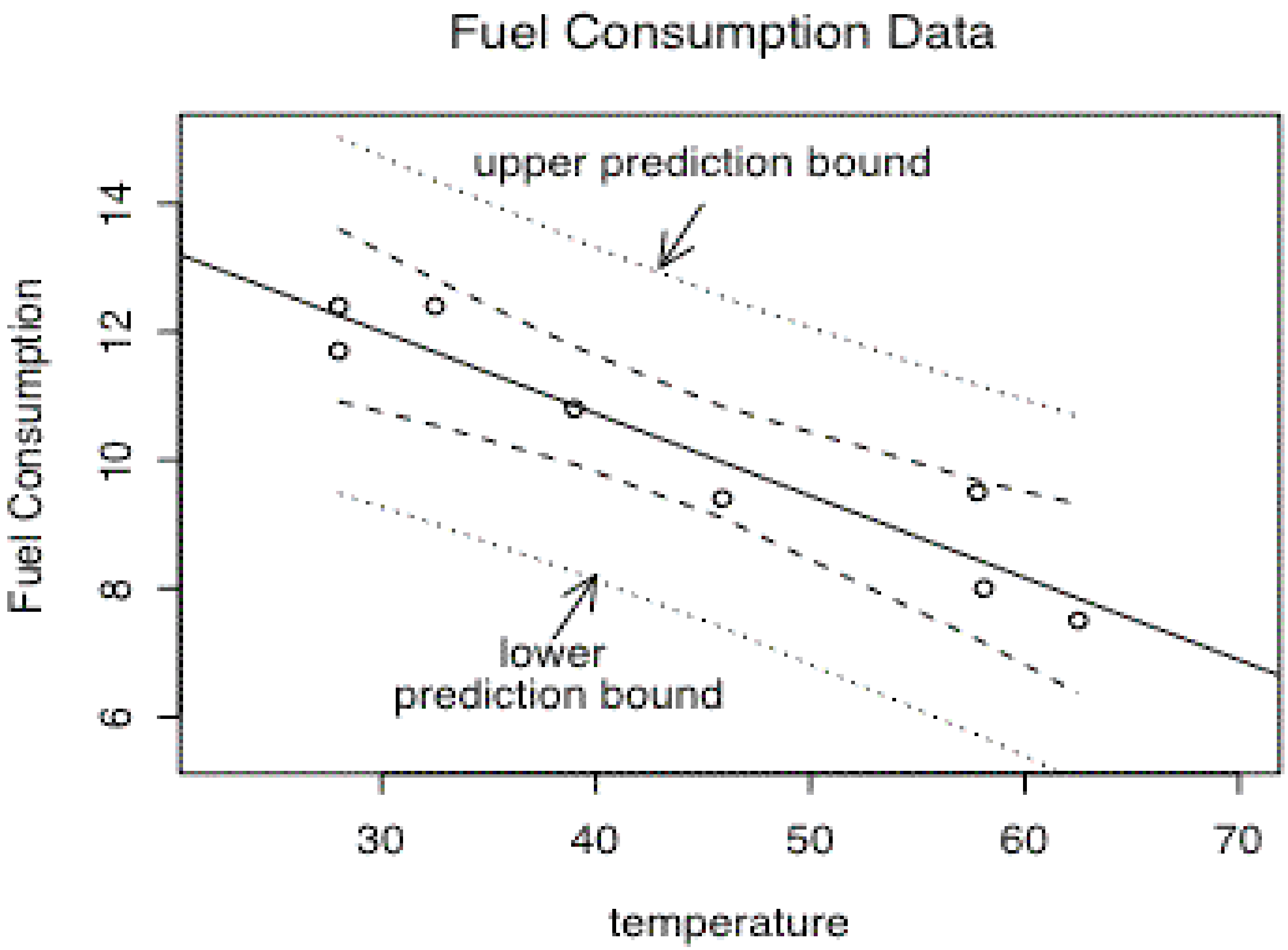
- Plotting a scatter plot with fitted line and prediction interval:

```
> plot(data3.1, main="Fuel Consumption Data ",
  xlab="temperature", ylab="Fuel Consumption")

> abline(coef(results3.1))

> lines(data3.1[,"Temp"], pred3.1[,2], lty=3 )

> lines(data3.1[,"Temp"], pred3.1[,3], lty=3)
```


Fuel Consumption Data
upper prediction bound
lower
prediction bound
Fuel Consumption
6
8
10
12
14
30
40
50
60
70
temperature

3.7 Coefficients of Determination and Correlation

▶ There are times when the degree of linear association is of interest in its own right. We now discuss two descriptive measures to describe the degree of linear association between X and Y.

▶ Partitioning of Total Sum of Squares:

- Total Sum of Squares (SST) $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - Error Sum of Squares (SSE) $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
  - Regresion Sum of Squares (SSR) $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

- $SST = SSE + SSR$

- The coefficient of determination r^2 is defined as

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Since $0 \leq SSE \leq SST$, it follows that

$$0 \leq r^2 \leq 1$$

- Interpret r^2 as the proportionate reduction of total variation associated with the use of the predictor variable X.

- The limiting values of r^2 occurs as follows:

1. When all observations fall on the fitted regression line, then $SSE = 0$ and $r^2 = 1$.

2. When the fitted regression line is horizontal so that b_0 and $\hat{Y}_i \equiv \bar{Y}$, then $SSE = SST$ and $r^2 = 0$.

- Thecorrelation coefficient r is the square root of r^2:

$$r = \pm\sqrt{r^2}$$

A plus of minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative.

- Since $r^2 \in [0, 1]$, it follows that

$$-1 \le r \le 1$$

- If a value of r is close to 1 then X and Y are said to be strongly positively correlated
- If a value of r is close to -1 then X and Y are said to be strongly negatively correlated

▶ A common misunderstanding:

- A correlation coefficient near zero indicates that X and Y are *not related*.

▶ A direct computational formula for r, which automatically furnishes the proper sign, is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right\}^{1/2}}$$

3.8 An *F*-test for the Model

- *F*-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

- test statistic: F-score $= \dfrac{SSR/1}{SSE/(n-2)}$

- P-value: $P(F^{(1,n-2)} \geq$ F-score$)$

- Reject H_0 if P-value $\leq \alpha$, and fail to reject H_0 otherwise

► *F*-test, *SST*, *SSR*, and *SSE*.

```
> anova(results3.1)
```

```
Analysis of Variance Table

Response: data3.1[, "Fuelcons"]
                 Df  Sum Sq Mean Sq F value    Pr(>F)
data3.1[, "Temp"]  1 22.9808 22.9808  53.695 0.0003301 ***
Residuals         6  2.5679  0.4280
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

► For the simple linear regression model, the *F*-test and the *t*-test for β_1 are equivalent. The *F*-test and the *t*-test are *NOT* equivalent in multiple regression model.