Chapter 2

Simple Linear Regression

Summary

Hypotheses testing and confidence interval of σ^2 (pages 3-4)

Example – Rocket propellant data (pages 5-6)

Confidence interval of the mean response *E(y)* at x_0 (pages 7-9)

Example – Rocket propellant data (pages 10-16)

Prediction of new observation at x_0 (page 17)

Example – Rocket propellant data (pages 18-21)

Prediction of the mean of *n* new observations at x_0 (page 22)

Coefficient of determination (R-squared) (page 23)

Example – Rocket propellant data (pages 24-25)

Hypotheses testing and confidence interval of σ^2

1. The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

for variation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $i = 1, 2, ..., n$ and ϵ_i's are independent.

what we are interested to estimate

Assume further that ϵ_i's are normally distributed.

2. We want to test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

If H_0 is true, $X = \frac{(n-2)MS_{Res}}{\sigma_0^2}$ follows the χ^2 distribution with $n-2$ degrees of freedom according to linear model theory.

3. For testing $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$, we reject H_0 if

$$\frac{(n-2)MS_{Res}}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-2} \quad \text{or} \quad \frac{(n-2)MS_{Res}}{\sigma_0^2} > \chi^2_{\alpha/2,n-2}$$

where α is the level of significance.

4. A $100(1-\alpha)\%$ confidence interval for σ^2 is given as

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}$$

The Rocket Propellant Data

$$\frac{(n-2)MS_{\text{Res}}}{\chi^2_{0.025,n-2}} \le \sigma^2 \le \frac{(n-2)MS_{\text{Res}}}{\chi^2_{0.975,n-2}}$$

$$\frac{18(9244.59)}{\chi^2_{0.025,18}} \le \sigma^2 \le \frac{18(9244.59)}{\chi^2_{0.975,18}}$$

$$\frac{18(9244.59)}{\mathbf{31.52638}} \le \sigma^2 \le \frac{18(9244.59)}{\mathbf{8.230746}}$$

$$5282.62 \le \sigma^2 \le 20,219.03$$

```r
#extract MSRes
anova(fitted.model)
MSRes <- anova(fitted.model)$"Mean Sq"[2]
MSRes

#95% confidence interval of sigma^2
(n-2)*MSRes/qchisq(0.025, df=n-2)
(n-2)*MSRes/qchisq(0.975, df=n-2)

```

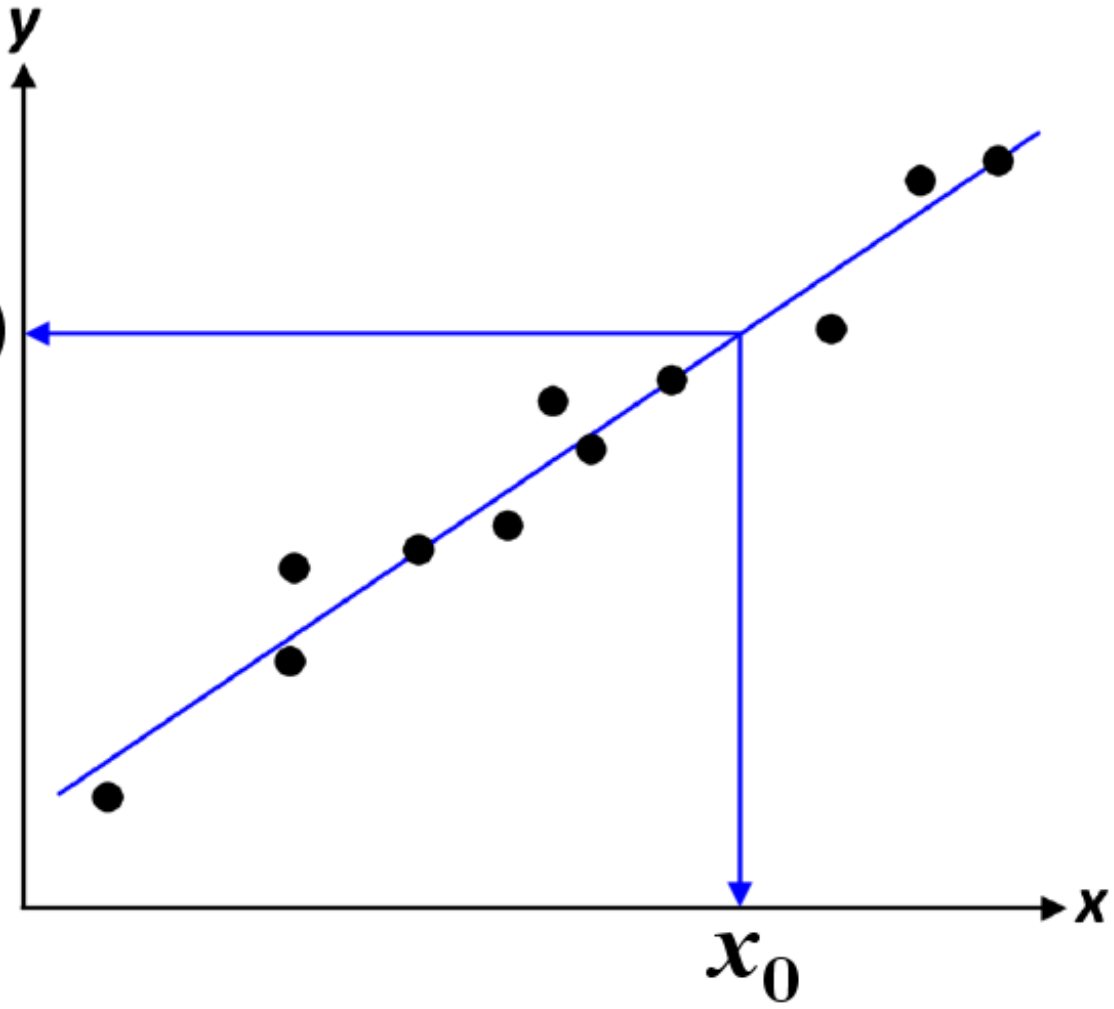
97:1 (Top Level)

Console | Background Jobs

R 3.4.1 · ~/

```
Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 1527483 1527483  165.38 1.643e-10 ***
Residuals 18  166255    9236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> MSRes <- anova(fitted.model)$"Mean Sq"[2]
> MSRes
[1] 9236.381
>
> #95% confidence interval of sigma^2
> (n-2)*MSRes/qchisq(0.025, df=n-2)
[1] 20199.24
> (n-2)*MSRes/qchisq(0.975, df=n-2)
[1] 5273.516
```

mean response of a particular value of x



parameter

$\mathbf{E(y|x_0)}$

expected value

$\Rightarrow$ predicted would be

$\widehat{E(y|x_0)}$

(need a hat)

a random variable

Confidence interval of the mean response $E(y)$ at the point x_0

$$E(y|x_0) = \mu_{y|x_0} = \beta_0 + \beta_1 x_0$$

a constant

Constants have no variance

$$\mathrm{Var}(\mu_{y|x_0}) = 0$$

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$E(y|x_0) = \beta_0 + \beta_1 x_0 + \underbrace{E(\varepsilon)}_{0}$$

a random variable

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\mathrm{Var}(\hat{\mu}_{y|x_0}) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= \mathrm{Var}[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})]$$

$$\boxed{\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0}$$

$$= \mathrm{Var}(\bar{y}) + (x_0 - \bar{x})^2 \mathrm{Var}(\hat{\beta}_1) + \underbrace{2\,\mathrm{cov}(\bar{y}, \hat{\beta}_1 (x_0 - \bar{x}))}_{0}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}}$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

$\hat{\mu}_{y|x_0}$ is a normally distributed random variable because it is a linear combination of the observations y_i.

$$\frac{\hat{\mu}_{y|x_0} - E(y|x_0)}{\sqrt{MS_{\text{Res}}\left(1/n + (x_0 - \bar{x})^2 / S_{xx}\right)}}$$ is t with $n-2$ degrees of freedom.

$\hat{\mu}_{y|x_0} \sim N$

$MS_{\text{Res}} \sim \chi^2$

$\frac{z}{\sqrt{\frac{\chi}{\nu}}} \sim t$ (ν: degree of freedom)

100(1 – α) percent CI on the mean response at the point $x = x_0$

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

$$\leq E(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Example Rocket Propellant Data

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$\leq E(y|x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$\hat{\mu}_{y|x_0} - (2.101)\sqrt{9244.59\left(\frac{1}{20}+\frac{(x_0-13.3625)^2}{1106.56}\right)}$$

$$\leq E(y|x_0) \leq \hat{\mu}_{y|x_0} + (2.101)\sqrt{9244.59\left(\frac{1}{20}+\frac{(x_0-13.3625)^2}{1106.56}\right)}$$

if $x_0 = \bar{x} = 13.3625$, $\hat{\mu}_{y|x_0} = 2131.40$, and the CI becomes

$$2086.230 \leq E(y|13.3625) \leq 2176.571$$

Construct confidence interval for *E(y)* at x0 using R

```
x0 <- mean(x)
newdata <- data.frame(x=x0)
predict.lm(fitted.model, newdata, interval="confidence", level=0.95)
```

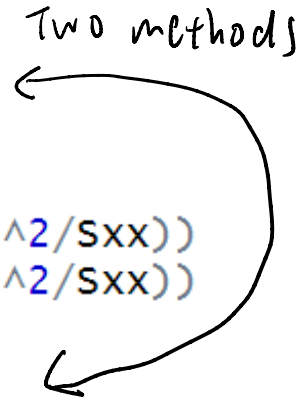
```r
MSRes <- anova(fitted.model)$"Mean Sq"[2]
MSRes

#Sxx
n <- length(x)
Sxx <- sum(x^2)-sum(x)^2/n

#95% confidence interval of mean of y at x0 = mean of x
x0 <- mean(x)
mean.y <- beta0 + beta1*x0
mean.y - qt(0.975,df=n-2)*sqrt(MSRes*(1/n + (x0-mean(x))^2/Sxx))
mean.y + qt(0.975,df=n-2)*sqrt(MSRes*(1/n + (x0-mean(x))^2/Sxx))

#95% confidence interval of mean of y at x0 = mean of x
x0 <- mean(x)
newdata <- data.frame(x=x0)
predict.lm(fitted.model, newdata, interval="confidence", level=0.95)

#95% confidence interval of mean of y at all x0
predict.lm(fitted.model, interval="confidence", level=0.95)
```



```
> #Sxx
> n <- length(x)
> Sxx <- sum(x^2)-sum(x)^2/n
>
> #95% confidence interval of mean of y at x0 = mean of x
> x0 <- mean(x)
> mean.y <- beta0 + beta1*x0
> mean.y - qt(0.975,df=n-2)*sqrt(MSRes*(1/n + (x0-mean(x))^2/Sxx))
[1] 2086.209
> mean.y + qt(0.975,df=n-2)*sqrt(MSRes*(1/n + (x0-mean(x))^2/Sxx))
[1] 2176.506
>
> #95% confidence interval of mean of y at x0 = mean of x
> x0 <- mean(x)
> newdata <- data.frame(x=x0)
> predict.lm(fitted.model, newdata, interval="confidence", level=0.95)
       fit      lwr      upr
1 2131.357 2086.209 2176.506
```

```
> #95% confidence interval of mean of y at all x0
> predict.lm(fitted.model, interval="confidence", level=0.95)
        fit      lwr      upr
1  2051.942 2004.966 2098.918
2  1745.425 1667.877 1822.973
3  2330.594 2274.935 2386.252
4  1996.211 1945.953 2046.470
5  2423.478 2357.782 2489.174
6  1921.904 1865.253 1978.555
7  1736.136 1657.349 1814.923
8  2534.938 2455.029 2614.848
9  2349.170 2291.684 2406.656
10 2219.133 2171.762 2266.504
11 2144.826 2099.623 2190.028
12 2488.496 2414.722 2562.271
13 1698.983 1615.149 1782.816
14 2265.575 2215.383 2315.767
15 1810.443 1741.255 1879.632
16 1959.058 1905.853 2012.263
17 2404.901 2341.375 2468.426
18 2163.402 2117.951 2208.854
19 2553.515 2471.083 2635.947
20 1829.020 1762.102 1895.939
```

TABLE 2.6 Confidence Limits on $E(y|x_0)$ for Several Values of x_0

Lower Confidence Limit	x_0	Upper Confidence Limit
2438.919	3	2593.821
2341.360	6	2468.481
2241.104	9	2345.836
2136.098	12	2227.942
2086.230	$\bar{x} = 13.3625$	2176.571
2024.318	15	2116.822
1905.890	18	2012.351
1782.928	21	1912.412
1657.395	24	1815.045

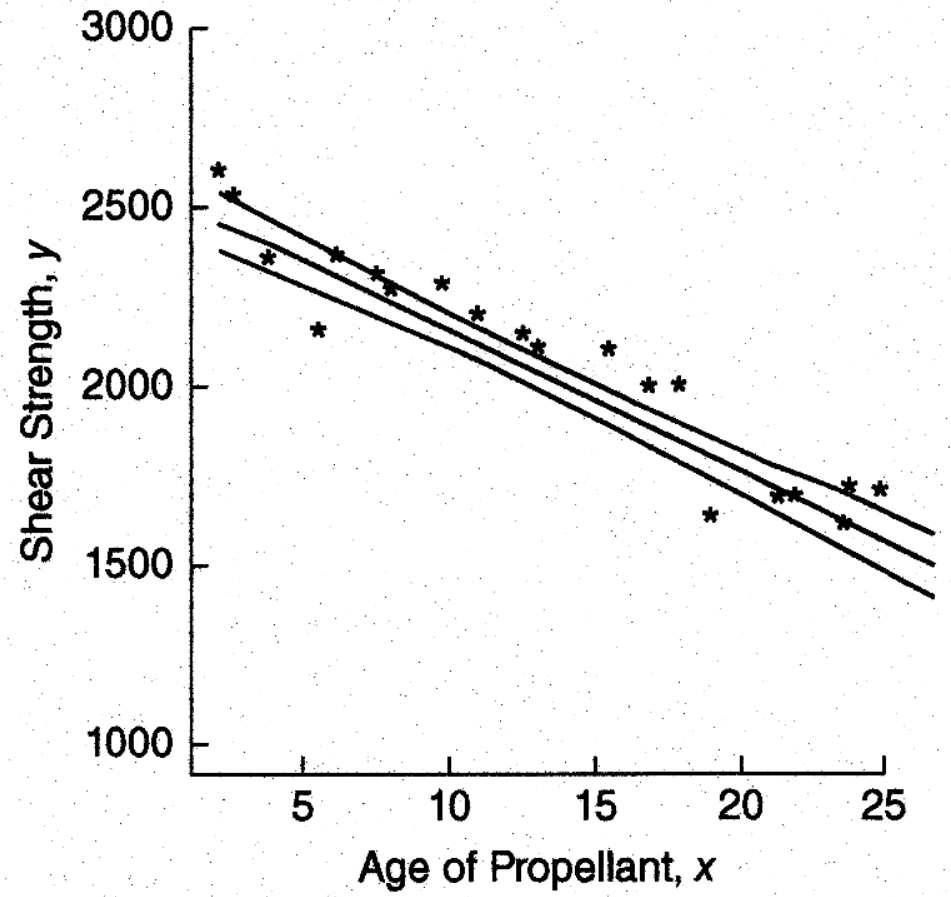


Figure 2.4 The upper and lower 95% confidence limits for the propellant data.

PREDICTION OF NEW OBSERVATIONS

$$y_0 = \beta_0 + \beta_1 x_0$$

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Prediction interval for the future observation y_0

$\psi = y_0 - \hat{y}_0$ is normally distributed with mean zero and variance

$$\mathrm{Var}(\psi) = \mathrm{Var}(y_0 - \hat{y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

no covariance

the future observation y_0 is independent of $\hat{y}_0$

measured

predicted

$100(1-\alpha)$ percent prediction interval on a future observation y_0

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\mathrm{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

$$\leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\mathrm{Res}}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Example The Rocket Propellant Data

95% prediction interval on a future value of propellant shear strength in a motor made from a batch of sustainer propellant that is 10 weeks old

≠ CI (used for estimating a population parameter → a constant)

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$2256.32-(2.101)\sqrt{9244.59\left(1+\frac{1}{20}+\frac{(10-13.3625)^2}{1106.56}\right)} \le y_0 \le 2256.32+(2.101)\sqrt{9244.59\left(1+\frac{1}{20}+\frac{(10-13.3625)^2}{1106.56}\right)}$$

$$2048.32 \le y_0 \le 2464.32$$

interval is wider than CI from before (more uncertainty)

Construct prediction interval for *y0* at x0=10 using R

```
x0 <- 10
newdata <- data.frame(x=x0)
predict.lm(fitted.model, newdata, interval="prediction", level=0.95)
```

```
> #95% prediction interval of y at x0
> x0 <- 10
> newdata <- data.frame(x=x0)
> predict.lm(fitted.model, newdata, interval="prediction", level=0.95)
       fit      lwr      upr
1 2256.286 2048.385 2464.188
```

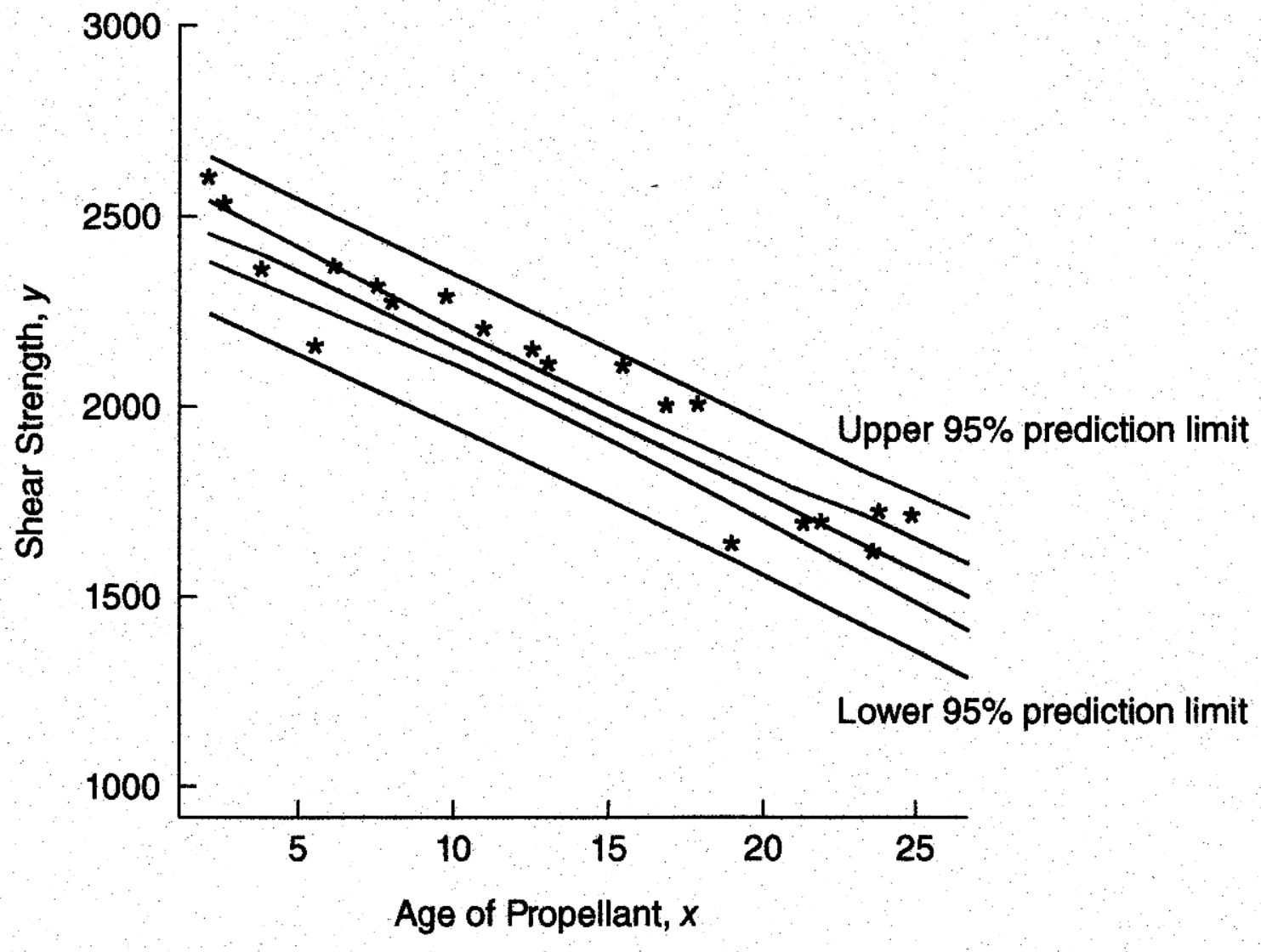


Figure 2.5 The 95% confidence and prediction intervals for the propellant data.

$100(1-\alpha)$ percent prediction interval on the **mean** of m future observations on the response at $x = x_0$

Let $\bar{y}_0$ be the mean of m future observations at $x = x_0$.

A point estimator of $\bar{y}_0$ is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

$100(1-\alpha)\%$ prediction interval on $\bar{y}_0$ is

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{m}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

$$\leq y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left(\frac{1}{m}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

Proportion of variation explained by the fitted line

Coefficient of determination (R-squared)

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

(constant)

$$SS_T = SS_R + SS_{Res}$$

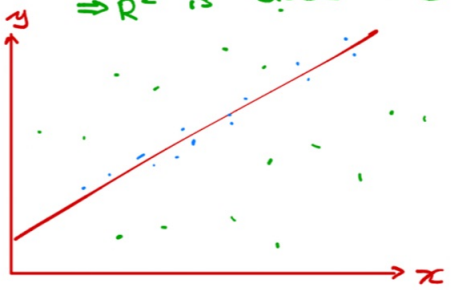
const const = small large + large small

$SS_T = SS_R + SS_{Res}$

$\Rightarrow R^2$ is close to 1

$\Rightarrow R^2$ is close to 0

for case (a) (b)



SS_T is a measure of the variability in y without considering the effect of the regressor variable x

SS_{Res} is a measure of the variability in y remaining after x has been considered

R^2 is often called the proportion of variation explained by the regressor x.

$0 \le R^2 \le 1$ $\qquad$ $0 \le SS_{Res} \le SS_T$

The statistic R^2 should be used with caution.

Example Rocket propellant data

$$R^2 = \frac{SS_R}{SS_T} = \frac{1,527,334.95}{1,693,737.60} = 0.9018$$

```
> #r square
> summary(fitted.model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10

> names(summary(fitted.model))
 [1] "call"          "terms"         "residuals"     "coefficients"  "aliased"
 [6] "sigma"         "df"            "r.squared"     "adj.r.squared" "fstatistic"
[11] "cov.unscaled"
> summary(fitted.model)$r.squared
[1] 0.9018414
```

The End