Partial answer key

11.27

The scatter plot with the sample regression model line is:



> ex11.21<-read.table(file = "ex11-21.txt", header=TRUE, sep = "")

> head(ex11.21)

 Runsize Total\_cost

1 2.6 230

2 5.0 341

3 10.0 629

4 2.0 187

5 0.8 159

6 4.0 327

> plot(x = ex11.21$Runsize, y = ex11.21$Total\_cost, xlab = "Run size", ylab =

 "Total cost", main = "Automobile bumper stickers", xlim = c(0,20),

 ylim = c(100,1200), col = "red",

 pch = 1, cex = 1.0, panel.first = grid(col = "gray", lty = "dotted"))

> mod.fit<-lm(formula = Total\_cost ~ Runsize, data = ex11.21)

> curve(expr = mod.fit$coefficients[1] + mod.fit$coefficients[2]\*x,

 from = min(ex11.21$Runsize), to = max(ex11.21$Runsize), col = "blue",

 add = TRUE, n = 1000, lwd = 1)

The sample regression model is:

= 99.68 + 51.93x

where y = total cost and x = run size.

> mod.fit<-lm(formula = TOTCOST ~ Runsize, data = ex11.27)

> mod.fit

Call:

lm(formula = TOTCOST ~ Runsize, data = ex11.27)

Coefficients:

(Intercept) Runsize

 99.68 51.93

The level of significance is α = 0.05.

Confidence interval method:

1. Ho: β1=0
Ha: β1≠0
2. The 95% C.I. for β1 is (50.74, 53.12).

> confint(object = mod.fit, level = 0.95)

 2.5 % 97.5 %

(Intercept) 93.92137 105.4290

Runsize 50.73621 53.1232

1. Since 0 is outside of this interval, reject Ho.
2. There is sufficient evidence to show that run size is linearly related to total cost.

The predicted mean total costs for these four printing runs are 515.1, 774.7, 930.6 and 1034.4, respectively.

> more.Runsize<-data.frame(Runsize = c(8, 13, 16, 18))

> predict(object = mod.fit, newdata = more.Runsize, se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 515.1129 507.5857 522.6400

2 774.7614 761.9565 787.5662

3 930.5505 914.3476 946.7534

4 1034.4099 1015.9033 1052.9165

$se.fit

 1 2 3 4

3.674614 6.251134 7.910021 9.034621

$df

[1] 28

$residual.scale

[1] 12.12725

At x = 8, the 95% confidence interval for E(Yh) is (507.6, 522.6).

At x = 13, the 95% confidence interval for E(Yh) is (762.0, 787.6).

At x = 16, the 95% confidence interval for E(Yh) is (914.3, 946.8).

At x = 18, the 95% confidence interval for E(Yh) is (1015.9, 1052.9).

At x = 8, the 95% prediction interval for Yh is (489.1, 541.1).

At x = 13, the 95% prediction interval for Yh is (746.8, 802.7).

At x = 16, the 95% prediction interval for Yh is (900.9, 960.2).

At x = 18, the 95% prediction interval for Yh is (1003.4, 1065.3).

> predict(object = mod.fit, newdata = more.Runsize, se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 515.1129 489.1560 541.0697

2 774.7614 746.8138 802.7090

3 930.5505 900.8918 960.2092

4 1034.4099 1003.4326 1065.3873

$se.fit

 1 2 3 4

3.674614 6.251134 7.910021 9.034621

$df

[1] 28

$residual.scale

[1] 12.12725

The coefficient of determination is 0.9964.

* 99.64% of the variation in total cost can be explained by using run size to predict total cost.
* The model fits the data very well.

> summary(mod.fit)

Call:

lm(formula = TOTCOST ~ Runsize, data = ex11.27)

Residuals:

 Min 1Q Median 3Q Max

-23.500 -6.767 1.895 7.647 19.606

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 99.6752 2.8089 35.48 <2e-16 \*\*\*

Runsize 51.9297 0.5826 89.13 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.13 on 28 degrees of freedom

Multiple R-squared: 0.9965, Adjusted R-squared: 0.9964

F-statistic: 7944 on 1 and 28 DF, p-value: < 2.2e-16

The correlation coefficient is 0.9982.

> sqrt(0.9965)

[1] 0.9982485

11.44

The scatter plot with the sample regression model line and 95% CI and PI band is:



> ex11.44 <- read.table("ex11-44.txt", header=T, sep="")

> head(ex11.44)

 EXPER SALARY

1 8 113.9

2 5 112.5

3 5 109.0

4 11 125.1

5 4 111.6

6 3 112.7

> plot(x = ex11.44$EXPER, y = ex11.44$SALARY, xlab = "years of experience",

 ylab = "first year salary", main = "salary vs years of experience", col =

 "red", xlim = c(0,14), ylim = c(90,130), pch = 1, cex = 1.0, panel.first

 = grid(col = "gray", lty = "dotted"))

> mod.fit<-lm(formula = SALARY ~ EXPER, data = ex11.44)

> curve(expr = mod.fit$coefficients[1] + mod.fit$coefficients[2]\*x,

 from = min(ex11.44$EXPER), to = max(ex11.44$EXPER),

 col = "blue", add = TRUE, n = 1000, lwd = 1)

> curve(expr = predict(object = mod.fit, newdata = data.frame(EXPER = x),

 interval = "confidence", level = 0.95)[,2],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.44$EXPER), to = max(ex11.44$EXPER))

> curve(expr = predict(object = mod.fit, newdata = data.frame(EXPER = x),

 interval = "confidence", level = 0.95)[,3],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.44$EXPER), to = max(ex11.44$EXPER))

> curve(expr = predict(object = mod.fit, newdata = data.frame(EXPER = x),

 interval = "prediction", level = 0.95)[,2],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.44$EXPER), to = max(ex11.44$EXPER))

> curve(expr = predict(object = mod.fit, newdata = data.frame(EXPER = x),

 interval = "prediction", level = 0.95)[,3],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.44$EXPER), to = max(ex11.44$EXPER))

The sample regression model is:

= 100.62 + 1.49x

where y = first-year salary and x = years of prior work experience.

> mod.fit<-lm(formula = SALARY ~ EXPER, data = ex11.44)

> mod.fit

Call:

lm(formula = SALARY ~ EXPER, data = ex11.44)

Coefficients:

(Intercept) EXPER

 100.616 1.491

The level of significance is α = 0.05.

P-value method:

1. Ho: β1=0
Ha: β1≠0
2. The p-value is 1.1×10-8.

> summary(mod.fit)

Call:

lm(formula = SALARY ~ EXPER, data = ex11.44)

Residuals:

 Min 1Q Median 3Q Max

-23.5844 -1.6891 0.2953 2.3335 10.7499

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 100.6157 1.2918 77.886 < 2e-16 \*\*\*

EXPER 1.4906 0.2183 6.828 1.11e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.556 on 50 degrees of freedom

Multiple R-squared: 0.4825, Adjusted R-squared: 0.4722

F-statistic: 46.63 on 1 and 50 DF, p-value: 1.114e-08

1. α = 0.05
2. Because p-value = 1.1×10-8 < 0.05, reject H0.
3. There is sufficient evidence to show that years of prior work experience is linearly related to first year salary.

The predicted mean first-year salary for 10 years of prior work experience is 115.5.

> predict(object = mod.fit, newdata = more.EXPER, se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 115.5219 112.7481 118.2957

$se.fit

[1] 1.380985

$df

[1] 50

$residual.scale

[1] 5.555951

The 95% confidence interval for E(Yh) is (112.7, 118.3).

The 95% prediction interval for Yh is (104.0, 127.0).

> predict(object = mod.fit, newdata = more.EXPER, se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 115.5219 104.0229 127.0209

$se.fit

[1] 1.380985

$df

[1] 50

$residual.scale

[1] 5.555951

The coefficient of determination is 0.4825. Thus, 48.25% of the variation in first-year salary can be explained by using years of prior work experience to predict first-year salary.

The correlation coefficient is 0.6946.

> sqrt(.4825)

[1] 0.6946222

11.65

The scatter plot with the sample regression model line and 95% CI and PI band is:



> ex11.65 <- read.table("ex11-65.txt", header=T, sep="")

> head(ex11.65)

 Mileage FuelUse

1 530 382

2 533 257

3 536 376

4 569 290

5 580 416

6 603 362

> plot(x = ex11.65$Mileage, y = ex11.65$FuelUse, xlab = "Mileage", ylab =

 "Fuel Use", main = "Fuel USe vs Mileage", col = "red",

 pch = 1, cex = 1.0, panel.first = grid(col = "gray", lty = "dotted"))

> mod.fit<-lm(formula = FuelUse ~ Mileage, data = ex11.65)

> curve(expr = mod.fit$coefficients[1] + mod.fit$coefficients[2]\*x,

 from = min(ex11.65$Mileage), to = max(ex11.65$Mileage),

 col = "blue", add = TRUE, n = 1000, lwd = 1)

> curve(expr = predict(object = mod.fit, newdata = data.frame(Mileage = x),

 interval = "confidence", level = 0.95)[,2],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.65$Mileage), to = max(ex11.65$Mileage))

> curve(expr = predict(object = mod.fit, newdata = data.frame(Mileage = x),

 interval = "confidence", level = 0.95)[,3],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.65$Mileage), to = max(ex11.65$Mileage))

> curve(expr = predict(object = mod.fit, newdata = data.frame(Mileage = x),

 interval = "prediction", level = 0.95)[,2],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.65$Mileage), to = max(ex11.65$Mileage))

> curve(expr = predict(object = mod.fit, newdata = data.frame(Mileage = x),

 interval = "prediction", level = 0.95)[,3],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.65$Mileage), to = max(ex11.65$Mileage))

The sample regression model is:

= 123.90 + 0.43x

where y = Fuel Use and x = Mileage.

> mod.fit<-lm(formula = FuelUse ~ Mileage, data = ex11.65)

> mod.fit

Call:

lm(formula = FuelUse ~ Mileage, data = ex11.65)

Coefficients:

(Intercept) Mileage

 123.8969 0.4319

The level of significance is α = 0.05.

Test statistic method:

1. Ho: β1=0
Ha: β1≠0
2. The test statistic is t = 27.48.

> summary(mod.fit)

Call:

lm(formula = FuelUse ~ Mileage, data = ex11.65)

Residuals:

 Min 1Q Median 3Q Max

-97.11 -39.85 15.57 40.58 73.94

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 123.89689 19.77667 6.265 9.87e-08 \*\*\*

Mileage 0.43192 0.01572 27.476 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.51 on 48 degrees of freedom

Multiple R-squared: 0.9402, Adjusted R-squared: 0.939

F-statistic: 754.9 on 1 and 48 DF, p-value: < 2.2e-16

1. Critical Value: ± t0.05/2, 48 = ±2.01

 > qt(0.975, df = 48)

[1] 2.010635

1. Because 27.48 > 2.01, reject H0.
2. There is sufficient evidence to show that mileage is linearly related to fuel use.

The predicted mean transformed fuel use for mileage level of 500, 1000 and 1500 are 339.86, 555.82 and 771.78, respectively.

> more.mileage.<-data.frame( Mileage= c(500, 1000, 1500))

> predict(object = mod.fit, newdata = more.mileage., se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 339.8583 314.3220 365.3947

2 555.8198 540.9007 570.7389

3 771.7812 754.6765 788.8860

$se.fit

 1 2 3

12.700645 7.420109 8.507144

$df

[1] 48

$residual.scale

[1] 48.50595

At x = 500, the 95% confidence interval for E(Yh) is (314.32,365.39).

At x = 1000, the 95% confidence interval for E(Yh) is (540.90,570.74).

At x = 1500, the 95% confidence interval for E(Yh) is (754.68, 788.89).

At x = 500, the 95% prediction interval for Yh is (239.01,440.67).

At x = 1000, the 95% prediction interval for Yh is (457.16,654.48).

At x = 1500, the 95% prediction interval for Yh is (672.77,870.80).

> predict(object = mod.fit, newdata = more.mileage., se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 339.8583 239.0428 440.6739

2 555.8198 457.1575 654.4821

3 771.7812 672.7649 870.7976

$se.fit

 1 2 3

12.700645 7.420109 8.507144

$df

[1] 48

$residual.scale

[1] 48.50595

The coefficient of determination is 0.9408.

* 94.08% of the variation in Fuel use can be explained by using Mileage to predict Fuel use.
* The model fits the data relatively well.

The correlation coefficient is 0.9699.

> sqrt(.9408)

[1] 0.9699485

11.82

The scatter plot with the sample regression model line and 95% CI and PI band is:



> ex11.82 <- read.table("ex11-82.txt", header=T, sep="")

> head(ex11.82)

 Patient x y

1 1 70.00 18.88

2 2 55.43 7.26

3 3 18.87 6.50

4 4 40.41 9.83

5 5 57.43 46.05

6 6 31.14 20.10

> plot(x = ex11.82$x, y = ex11.82$y,

 xlab = "Pellet Method", ylab = "Homogenate Method",

 main = "Homogenate Method vs Pellet Method", col = "red", xlim =

 c(10,330), ylim = c(-20,140),

 pch = 1, cex = 1.0, panel.first = grid(col = "gray", lty = "dotted"))

> mod.fit<-lm(formula = y ~ x, data = ex11.82)

> curve(expr = mod.fit$coefficients[1] + mod.fit$coefficients[2]\*x,

 from = min(ex11.82$x), to = max(ex11.82$x),

 col = "blue", add = TRUE, n = 1000, lwd = 1)

> curve(expr = predict(object = mod.fit, newdata = data.frame(x = x),

 interval = "confidence", level = 0.95)[,2],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.82$x), to = max(ex11.82$x))

> curve(expr = predict(object = mod.fit, newdata = data.frame(x = x),

 interval = "confidence", level = 0.95)[,3],

 col = "darkgreen", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.82$x), to = max(ex11.82$x))

> curve(expr = predict(object = mod.fit, newdata = data.frame(x = x),

 interval = "prediction", level = 0.95)[,2],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.82$x), to = max(ex11.82$x))

> curve(expr = predict(object = mod.fit, newdata = data.frame(x = x),

 interval = "prediction", level = 0.95)[,3],

 col = "blue", lty = "dashed", lwd = 1, add = TRUE,

 from = min(ex11.82$x), to = max(ex11.82$x))

The sample regression model is:

= 10.33 + 0.27x

where y = sucrose activity measured by homogenate method and x = sucrose activity measured by pellet method.

> mod.fit<-lm(formula = y ~ x, data = ex11.82)

> mod.fit

Call:

lm(formula = y ~ x, data = ex11.82)

Coefficients:

(Intercept) x

 10.3348 0.2669

The level of significance is α = 0.05.

P-value method:

1. Ho: β1=0
Ha: β1≠0
2. The p-value is 3.8×10-8.

> summary(mod.fit)

Call:

lm(formula = y ~ x, data = ex11.82)

Residuals:

 Min 1Q Median 3Q Max

-23.5773 -10.9652 -0.2517 9.4272 30.0099

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.33483 5.99543 1.724 0.0988 .

x 0.26694 0.03251 8.210 3.83e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.62 on 22 degrees of freedom

Multiple R-squared: 0.7539, Adjusted R-squared: 0.7428

F-statistic: 67.41 on 1 and 22 DF, p-value: 3.827e-08

1. α = 0.05
2. Because p-value = 3.8×10-8 < 0.05, reject H0.
3. There is sufficient evidence to show that the pellet method is linearly related to the homogenate method.

When the sucrose activity measured by the pellet method is 80, the mean sucrose activity measured by homogenate method is 31.7.

> more.Pellet\_Method\_x<-data.frame(x = 80)

> predict(object = mod.fit, newdata = more.Pellet\_Method\_x, se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 31.69002 23.31816 40.06188

$se.fit

[1] 4.036823

$df

[1] 22

$residual.scale

[1] 15.61692

The 95% confidence interval for E(Yh) is (23.3, 40.1).

The 95% prediction interval for Yh is (-1.8, 65.1).

> predict(object = mod.fit, newdata = more.Pellet\_Method\_x, se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 31.69002 -1.762025 65.14207

$se.fit

[1] 4.036823

$df

[1] 22

$residual.scale

[1] 15.61692

The coefficient of determination is 0.7539. Thus, 75.39% of the variation in the sucrose activity measured by the homogenate method can be explained by the sucrose activity measured by the pellet method.

The correlation coefficient is 0.868.

> sqrt(.7539)

[1] 0.8682742

Additional problem:

The scatter plot with the sample regression model line is:



> ex.add<-read.csv(file = "C:\\sdata\\hometax.csv", header=TRUE)

> head(ex.add)

 Price Tax

1 2050 1639

2 2080 1088

3 2150 1193

4 2150 1635

5 1999 1732

6 1900 1534

> plot(x = ex.add$Price, y = ex.add$Tax,

 xlab = "Resale price", ylab = "Annual tax",

 main = "Annual tax vs resale price", col = "red",

 pch = 1, cex = 1.0, panel.first = grid(col = "gray", lty = "dotted"))

> mod.fit<-lm(formula = Tax ~ Price, data = ex.add)

> curve(expr = mod.fit$coefficients[1] + mod.fit$coefficients[2]\*x,

 from = min(ex.add$Price), to = max(ex.add$Price),

 col = "blue", add = TRUE, n = 1000, lwd = 1)

The sample regression model is:

= 36.34 + 0.70x

where y = annual tax and x = resale price.

> mod.fit<-lm(formula = Tax ~ Price, data = ex.add)

> mod.fit

Call:

lm(formula = Tax ~ Price, data = ex.add)

Coefficients:

(Intercept) Price

 36.3444 0.7028

The level of significance is α = 0.05.

Test statistic method:

1. Ho: β1=0
Ha: β1≠0
2. The test statistic is t = 18.581.

> summary(mod.fit)

Call:

lm(formula = Tax ~ Price, data = ex.add)

Residuals:

 Min 1Q Median 3Q Max

-417.28 -74.28 1.12 72.48 463.64

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 36.34435 43.23741 0.841 0.402

Price 0.70278 0.03782 18.581 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 149.5 on 105 degrees of freedom

Multiple R-squared: 0.7668, Adjusted R-squared: 0.7646

F-statistic: 345.2 on 1 and 105 DF, p-value: < 2.2e-16

1. Critical Value: ± t0.05/2, 22 = ±1.98

> qt(0.975, df = 105)

[1] 1.982815

Because 18.58 > 1.98, reject H0

1. There is sufficient evidence to show that the resale price is linearly related to the annual tax.

The mean annual tax for a house with resale price of 1005 is 742.6.

> more.Price<-data.frame(Price = 1005)

> predict(object = mod.fit, newdata = more.Price, se.fit =

 TRUE, interval = "confidence", level = 0.95)

$fit

 fit lwr upr

1 742.6425 713.47 771.815

$se.fit

[1] 14.71264

$df

[1] 105

$residual.scale

[1] 149.5333

The 95% confidence interval for E(Yh) is (713.5, 771.8).

The 95% prediction interval for Yh is (444.7, 1040.6).

> predict(object = mod.fit, newdata = more.Price, se.fit =

 TRUE, interval = "prediction", level = 0.95)

$fit

 fit lwr upr

1 742.6425 444.7139 1040.571

$se.fit

[1] 14.71264

$df

[1] 105

$residual.scale

[1] 149.5333

The coefficient of determination is 0.7668. Thus, 76.68% of the variation in the annual tax can be explained by using the resale price to predict the annual tax.

The correlation coefficient is 0.876.

> sqrt(.7668)

[1] 0.8756712