3.48 **Learning the mechanics**.
(a) Explain why for a particular x-value, the prediction interval for an individual y-value will always be wider than the confidence interval for a mean value of y.

Ans: Confidence interval: $\hat{y} \pm \left(t_{\alpha/2}\right)s\sqrt{\frac{1}{n}+\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}}$

Prediction interval: $\hat{y} \pm \left(t_{\alpha/2}\right)s\sqrt{1+\frac{1}{n}+\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}}$.

Because $\sqrt{\frac{1}{n}+\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}} < \sqrt{1+\frac{1}{n}+\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}}$, the prediction interval must be wider than the confidence interval. The error in estimating the mean value of y, $E(y)$, for a given value of x, is the distance between the least squares line and the true line of means. The error $\left(y_{\mathrm{p}}-\hat{y}\right)$ is the sum of two errors: the error of estimating the mean of y, $E(y)$, plus the random error that is a component of the value of y to be predicted. Consequently, the error of predicting a particular value of y will be larger than the error of estimating the mean value of y for a particular value of x.

(b) Explain why the confidence interval for the mean value of y for a particular x-value, say, x_{p}, gets wider the farther x_{p} is from $\bar{x}$. What are the implications of this phenomenon for estimation and prediction?

Since the standard error contains the term $\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}$, the further x_{p} is from, the larger the standard error. This causes the confidence intervals to be wider for values of x_{p} further from. The implication is our best confidence intervals (narrowest) will be found when $x_{\mathrm{p}} = \bar{X}$.

3.49 **Learning the mechanics**. A simple linear regression analysis for $n = 20$ data points produce the following results:

$\hat{y} = 2.1 + 3.4x$	$SS_{xx} = 4.77$
$\bar{x} = 2.5$	$SS_{yy} = 59.21$
$\bar{y} = 10.6$	$SS_{xy} = 16.22$

(a) Find SSE and s^2.

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 59.21 - 3.4(16.22) = 4.062;\ s^2 = \frac{SSE}{n-2} = \frac{4.062}{20-2} = 0.226$$

Ans: SSE = 4.062, s^2 = 0.226

(b) Find a 95% confidence interval for $E(y)$ when $x = 2.5$. Interpret this interval.

$$\hat{y} \pm \left(t_{\alpha/2}\right)s\sqrt{\frac{1}{n}+\frac{(x_{\mathrm{p}}-\bar{x})^2}{\mathrm{SS}_{xx}}} = 2.1 + 3.4(2.5) \pm 2.101\left(\sqrt{0.226}\right)\sqrt{\frac{1}{20}+\frac{(2.5-2.5)^2}{4.77}}$$
$$= 10.6 \pm 0.223 = (10.377, 10.823)$$

Ans: $\mathbf{0.6 \pm 0.223}$ or $\mathbf{(10.377, 10.823)}$

(c) Find a 95% confidence interval for $E(y)$ when $x = 2.0$. Interpret this interval.

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}} = 2.1 + 3.4(2.0) \pm 2.101(\sqrt{0.226})\sqrt{\frac{1}{20}+\frac{(2.0-2.5)^2}{4.77}}$$
$$= 8.1 \pm 0.320 = (7.780, 8.420)$$

Ans: 8.1 ± 0.320 or $(7.780, 8.420)$

(d) Find a 95% confidence interval for $E(y)$ when $x = 3.0$. Interpret this interval.

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}} = 2.1 + 3.4(3.0) \pm 2.101(\sqrt{0.226})\sqrt{\frac{1}{20}+\frac{(3.0-2.5)^2}{4.77}}$$
$$= 12.3 \pm 0.320 = (11.980, 12.620)$$

Ans: 12.3 ± 0.320 or $(11.980, 12.620)$

(e) Examine the widths of the confidence intervals obtained in parts (b), (c), and (d). What happens to the width of the confidence interval for $E(y)$ as the value of x moves away from the value of $\bar{x}$? The width of the confidence interval becomes wider as x moves away from $\bar{x}$.

Ans: wider

(f) Find a 95% prediction interval for a value of y to be observed in the future when $x = 3.0$. Interpret its value.

$$\hat{y} \pm (t_{\alpha/2})s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}} = 2.1 + 3.4(3.0) \pm 2.101(\sqrt{0.226})\sqrt{1+\frac{1}{20}+\frac{(3.0-2.5)^2}{4.77}}$$
$$= 12.3 \pm 1.049 = (11.251, 13.349)$$

Ans: 12.3 ± 1.049 or $(11.251, 13.349)$

4.6 **Earnings of Mexican street vendors.** Detailed interviews were conducted with over 1,000 street vendors in the city of Puebla, Mexico, in order to study the factors influencing vendors' incomes (*World Development*, February 1998). Vendors were defined as individuals working in the street, and included vendors with carts and stands on wheels and excluded beggars, drug dealers, and prostitutes. The researchers collected data on gender, age, hours worked per day, annual earnings, and education level. A subset of these data appears in the accompanying table.

STREETVEN

VENDOR NUMBER	ANNUAL EARNINGS y	AGE x_1	HOURS WORKED PER DAY x_2
21	$2841	29	12
53	1876	21	8
60	2934	62	10
184	1552	18	10
263	3065	40	11
281	3670	50	11
354	2005	65	5
401	3215	44	8
515	1930	17	8
633	2010	70	6
677	3111	20	9
710	2882	29	9
800	1683	15	5
914	1817	14	7
997	4066	33	12

Source: Adapted from Smith, P. A., and Metzger, M. R. "The return to education: Street vendors in Mexico," *World Development*, Vol. 26, No. 2, Feb. 1998, pp. 289–296.

(a) Write a first-order model for mean annual earnings, $E(y)$, as a function of age (x_1) and hours worked (x_2).

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Ans: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

(b) The model was fit to the data using SAS. Find the least squares prediction equation on the printout shown below.

$\hat{y} = -20.35201 + 13.35045x_1 + 243.71446x_2$

Ans: $\hat{y} = -20.35201 + 13.35045x_1 + 243.71446x_2$

SAS output for Exercise 4.6

Dependent Variable: EARNINGS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5018232	2509116	8.36	0.0053
Error	12	3600196	300016		
Corrected Total	14	8618428			

Root MSE	547.73748	R-Square	0.5823
Dependent Mean	2577.13333	Adj R-Sq	0.5126
Coeff Var	21.25375		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-20.35201	652.74532	-0.03	0.9756	-1442.56189	1401.85787
AGE	1	13.35045	7.67168	1.74	0.1074	-3.36470	30.06559
HOURS	1	243.71446	63.51174	3.84	0.0024	105.33428	382.09465

(c) Interpret the estimated β coefficients in your model.

For every 1-year increase in age (x_1), annual income increases by \$13.35045; for every increase in hour worked per day (x_2), annual income increases by \$243.71446.

(d) Conduct a test of the global utility of the model (at $\alpha = .01$). Interpret the result.

$n - (k + 1) = 14 - (2 + 1) = 11$, $F = 8.36 > 7.21$ (Table D6, Appendix D, $v_1 = 2$, $v_2 = 11$), p-value = 0.0053. Since $\alpha = .01$ exceeds the observed significance level, $p = 0.0053$, the data provide strong evidence that at least one of the model coefficients is nonzero. The overall model appears to be statistically useful for predicting annual income.

(e) Find and interpret the value of R_a^2.

$R_a^2 = 0.5823$. 58.23% of the sample variation in annual income is explained by the model.

Ans: 0.5823

(f) Find and interpret s, the estimated standard deviation of the error term.

Root MSE = 547.73748; $2s = 547.73748(2) = 1095.47496$. About 95% of sample annul income fall within 1095.47496 of model predicted values.

Ans: s = 547.73748

(g) Is age (x_1) a statistically useful predictor of annual earnings? Test using $\alpha = .01$.

$H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$. $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{7.67168}{1.74} = 4.409$. p-value = 0.1074. Fail to reject H_0.

Ans: Fail to reject H_0

(h) Find a 95% confidence interval for β_2. Interpret the interval in the words of the problem.
(105.33428, 382.09465). We are 95% confident that β_2 falls between 105.33428 and 382.09465.

Ans: (105.33428, 382.09465)

4.28 **Earnings of Mexican street vendors.** Refer to the *World Development* (February 1998) study of street vendors in the city of Puebla, Mexico, Exercise 4.6 (p. 184). Recall that the vendors' mean annual earnings, $E(y)$, was modeled as a first-order function of age (x_1) and hours worked (x_2). Now consider the interaction model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. The SAS printout for the model is displayed in the next column.

The REG Procedure
Dependent Variable: EARNINGS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5287427	1762476	5.82	0.0124
Error	11	3331000	302818		
Corrected Total	14	8618428			

Root MSE	550.28921	R-Square	0.6135
Dependent Mean	2577.13333	Adj R-Sq	0.5081
Coeff Var	21.35276		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1041.89440	1303.59326	0.80	0.4411
AGE	1	-13.23762	29.23395	-0.45	0.6595
HOURS	1	103.30564	162.01356	0.64	0.5368
AGEHRS	1	3.62096	3.84044	0.94	0.3660

(a) Give the least squares prediction equation.
$\hat{y} = 1041.89440 - 13.23762x_1 + 103.30564x_2 + 3.62096x_1x_2$

Ans: $\hat{y} = 1041.89440 - 13.23762x_1 + 103.30564x_2 + 3.62096x_1x_2$

(b) What is the estimated slope relating annual earning (y) to age (x_1) when number of hours worked (x_2) is 10? Interpret the result.
Estimated x_1 slope = $\hat{\beta}_1 + \hat{\beta}_3 x_2 = -13.23762 + 3.62096 \cdot 10 = 22.97198$
The annual income will increase by about \$22.97198 for every additional year in age.

Ans: 22.97198

(c) What is the estimated slope relating annual earnings (y) to hours worked (x_2) when age (x_1) is 40? Interpret the result.
Estimated x_2 slope = $\hat{\beta}_2 + \hat{\beta}_3 x_1 = 103.30564 + 3.62096 \cdot 40 = 248.14404$
The annual income will increase by about \$248.14404 for every additional hour worked per day.

Ans: 248.14404

(d) Give the null hypothesis for testing whether age (x_1) and hours worked (x_2) interact.
$H_0: \beta_3 = 0; H_1: \beta_3 > 0.$

Ans: $H_0: \beta_3 = 0; H_1: \beta_3 > 0.$

(e) Find the *p*-value of the test, part (d).
$t = 0.94$, *p*-value = 0.3660.

Ans: 0.3660

(f) Refer to part (e). Give the appropriate conclusion in the words of the problem.
Failed to reject null hypothesis. The interaction term does not need to be included in the model.

Ans: Fail to reject H_0

4.29 **Psychology of waiting in line.** While waiting in a lone line for service (e.g., to use an ATM or at the post office), at some point you may decide to leave the queue. The *Journal of Consumer Research* (November 2003) published a study of consumer behavior while waiting in a queue. A sample of $n = 148$ college students were asked to imagine that they were waiting in line at a post office to mail a package and that the estimated waiting time is 10 minutes or less. After a 10-minute wait, students were asked about their level of negative feelings (annoyed, anxious) on a scale of 1 (strongly disagree) to 9 (strongly agree). Before answering, however, the students were informed about how many people were ahead of them and behind them in the line. The researchers used regression to relate negative feelings score (y) to number ahead in line (x_1) and number behind in line (x_2).
(a) The researchers fit an interaction model to the data. Write the hypothesized equation of this model.
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Ans: $\boldsymbol{E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$

(b) In the words of the problem, explain what it means to say that "x_1 and x_2 interact to effect y."
Ans: Linear relationship between negative feelings score and number ahead in the line depends on number behind in line

(c) A *t*-test for the interaction β in the model resulted in a *p*-value greater than .25. Interpret this result.
$H_0: \beta_3 = 0;\ H_a: \beta_3 \neq 0$. Fail to reject H_0.

Ans: Fail to reject H_0

(d) From their analysis, the researchers concluded that "the greater than number of people ahead, the higher the negative feelings score" and "the greater the number of people behind, the lower the negative feeling score." Use this information to determine the signs of β_1 and β_2 in the model.

Ans: $\boldsymbol{\beta_1 > 0, \beta_2 < 0}$