

A.

• Exercise 11.45

(a)

1.

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_A: \text{At least one } \beta_i \neq 0$$

2.

$$F = \frac{(R_1^2 - R_2^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.211 - 0.063)/3}{(1 - 0.211)/218} = \frac{0.148/3}{0.789/218} = \boxed{13.63}$$

$$\implies P\text{-value} = \boxed{0}$$

3. Reject H_0 since $P\text{-value} = 0$.

4. High school grades are important for predicting GPA, after accounting for SAT scores.

(b) The test for SAT scores is

$$H_0: \beta_4 = \beta_5 = 0$$

$$H_A: \beta_4 \neq 0 \text{ or } \beta_5 \neq 0$$

The F statistic for this model is 7.48 with $P\text{-value}$ 0.001. Therefore SAT scores are important (if used by themselves.)

B.

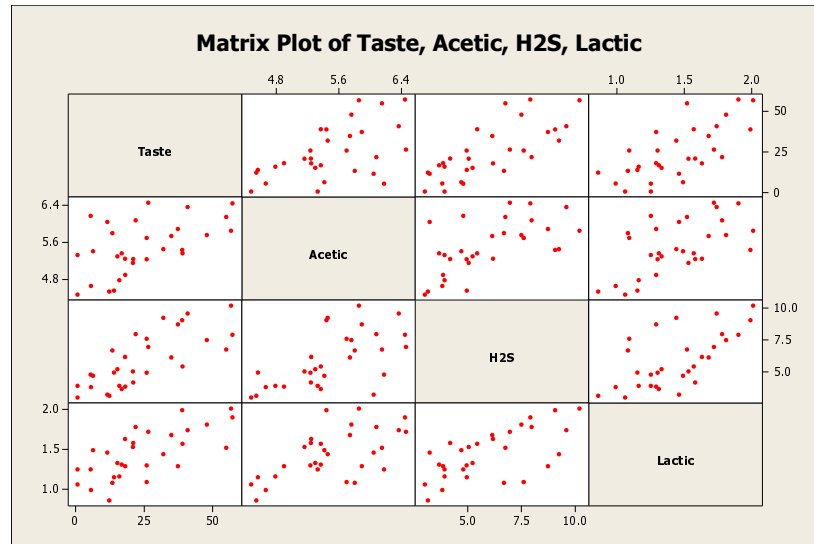
• Exercise 11.108

(a) The formula for a prediction interval is (see Exam 3 formula sheet)

$$\begin{aligned} \hat{y} \pm t^* \cdot SE_{\hat{y}} &= 2.136 \pm (1.984)(0.013) \\ &= 2.136 \pm 0.026 \\ &= (\$2.110, \$2.162) \end{aligned}$$

(b) Since \$2.13 is contained in the 95% prediction interval, a plausible explanation of the price is provided by the three “supply-and-demand explanatory variables.” So there is not strong evidence of price fixing.

C.



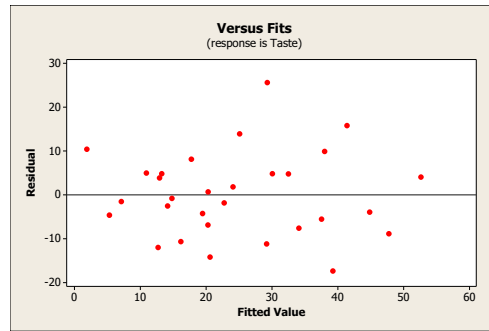
- (a) The response Taste appears to be positively correlated with all three predictors Acetic, H2S, and Lactic. The predictors also appear to all be positively correlated with each other.
- (b) The predictor variable H2S is most highly-correlated with response Taste ($r = 0.756$.) H2S and Lactic are the two most highly-correlated predictors ($r = 0.645$.)
- (c) The F test indicates that at least one of the three predictor variables is linearly related to Taste ($F = 16.22$, P -value = 0.)
- (d) Acetic is not significant when combined with any other predictor.
- (e) The candidate models are the ones whose predictors are all significant at the 10% level:

Model	Variables	R^2	t test
1	Acetic	30.2%	P -value = 0.002
2	H2S	57.1%	P -value = 0
3	Lactic	49.6%	P -value = 0
4	H2S, Lactic	65.2%	P -value = 0.002, P -value = 0.019

- (f) The best conservative model is Model 4 above since it has highest R^2 among the candidate models.

$$\text{Taste} = -27.6 + 3.95 \text{ H2S} + 19.9 \text{ Lactic}$$

- (g) The residuals plot appears to be fairly random. So regression assumptions appear to be satisfied and the regression can be safely used.



(h)

- Taste is unrelated to acetic acid, after accounting for H2S and lactic acid.
- Taste improves by 3.95 points, on average, for each additional one-percent concentration of H2S, when lactic acid is held constant.
- Taste improves by 19.9 points, on average, for each additional one-percent concentration of lactic acid, when H2S is held constant.

- (i) Yes! Even though Acetic is not a variable in the “best conservative” model, Model 1 above shows that Acetic is significant if used by itself (P -value= 0.002.)

Interpret:

Taste improves by 15.6 points on average for each additional one-percent concentration of acetic acid.

- (j) We are 99% certain that Taste for the batch of cheese is between 1.36 and 57.57 points.

(k)

$$\text{Taste} = -27.6 + 3.95 \text{ H2S} + 19.9 \text{ Lactic}$$

(**Note:** This model is the same as the *best conservative* model: H2S is already a variable in that model, so no modification is necessary!)

(l)

$$\text{Taste} = -28.9 + 0.33 \text{ Acetic} + 19.7 \text{ Lactic} + 3.91 \text{ H2S}$$

D.

(a)

$$\begin{aligned}x_2 = 0 \implies \mu_{\text{Wages}} &= \beta_0 + \beta_1 \times \text{LOS} + \beta_2(0) \\ &= \beta_0 + \beta_1 \times \text{LOS}\end{aligned}$$

(b) The intercept in the equation above is β_0 .

(c)

$$\hat{\beta}_0 = b_0 = \boxed{\$302.54}$$

(d)

$$\begin{aligned}x_2 = 1 \implies \mu_{\text{Wages}} &= \beta_0 + \beta_1 \times \text{LOS} + \beta_2(1) \\ &= (\beta_0 + \beta_2) + \beta_1 \times \text{LOS}\end{aligned}$$

(e) $(\beta_0 + \beta_2)$

(f)

$$\hat{\beta}_0 + \hat{\beta}_2 = 302.54 + 71.78 = \boxed{\$374.32}$$

E.(a) $n = \boxed{1745}$

(b)

$$\mu_{\text{Earnings}} = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Status} + \beta_3 \text{Race}$$

(c)

$$\text{Race} = 1 \implies \mu_{\text{Earnings}} = (\beta_0 + \beta_3) + \beta_1 \text{Gender} + \beta_2 \text{Status}$$

(d)

$$\text{Race} = 0 \implies \mu_{\text{Earnings}} = \beta_0 + \beta_1 \text{Gender} + \beta_2 \text{Status}$$

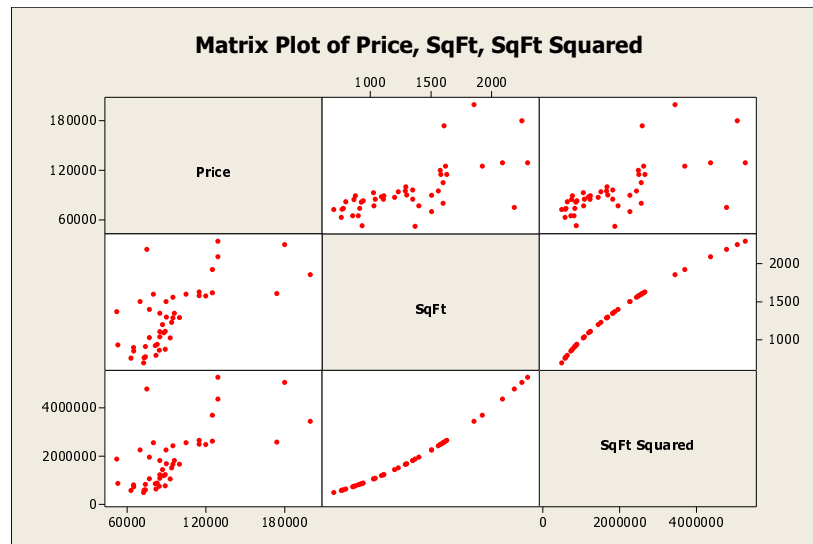
(e) $\beta_3 = (\text{Mean earnings for whites}) - (\text{Mean earnings for nonwhites})$ when Gender and Status are held constant.(f) No. ($t = -1.38$, $P\text{-value} = 0.167$) for Gender in the full model. Gender is not significant, after adjusting for Race and Status.(g) Yes. Race is significant in the full model, after adjusting for Gender and Status. ($t = 13.59$, $P\text{-value} = 0$)(h) $\text{Earnings} = 16943 + 2390 \text{ Status} + 2909 \text{ Race}$

- (i)
- Earnings average \$2909 more for whites than nonwhites, after adjusting for job status.
 - Earnings average \$2390 more for full-time workers than for part-time workers, after adjusting for race.
 - Earnings are unrelated to gender, after adjusting for race and job status.

(j) (\$16,461.40, \$17,425.30)

(k) (\$21,975.70, \$22,507.60)

F.



(a) It's hard to say from the plot but it looks like there might be a curve.

(b) The two variables are highly correlated.

(c)

- Quadratic regression passes the F test ($F = 16.14$, P -value = 0) but neither t test: ($t = 0.73$, P -value = 0.469 for SqFt), ($t = 0.19$, P -value = 0.854 for SqFt Squared)
- Linear regression passes the t test ($t = 5.74$, P -value = 0)
- The two models have identical $R^2 = 44.0\%$.
- The residual plots for both models look similar.
- So it appears that adding SqFt Squared to the model really adds nothing of value. In this case linear regression is a better choice since it is a simpler model.

(d) Price = 39798 + 39.3 SqFt - 11616 BedRooms + 24545 Baths

(e) 14.3%

(f)

- Price increases by \$39.30 on average for each additional square foot of area, when number of bedrooms and number of bathrooms are held constant.
- Price decreases by \$11,616 on average for each additional bedroom, when area and number of bathrooms are held constant.
- Price increases by \$24,545 on average for each additional bathroom, when area and number of bedrooms are held constant.
- Price is unrelated to garage capacity, after accounting for area, number of bedrooms, and number of bathrooms.

(g) Estimate = \$100,485 With 90% certainty the estimate is (\$91,965, \$109,004)

G.

(a) Variables are dropped from the full model in this order:

$$C4 \rightarrow C6 \rightarrow C3 \rightarrow SC \rightarrow C1 \rightarrow C5 \rightarrow \text{Age}$$

(b) The chosen model uses the remaining predictor variables: IQ, C2, and Sex.
Let $y = \text{GPA}$. The theoretical equation is

$$\mu_y = \beta_0 + \beta_1 \text{IQ} + \beta_2 \text{C2} + \beta_3 \text{Sex}$$

(c) The fitted equation is

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 \text{IQ} + \hat{\beta}_2 \text{C2} + \hat{\beta}_3 \text{Sex} \\ &= -1.90 + 0.0769 \text{IQ} + 0.193 \text{C2} - 0.909 \text{Sex}\end{aligned}$$

(d) No, since the variable SC was dropped from the chosen model: SC is not significant, after accounting for IQ, C2, and Sex.

(e) Slopes:

- Mean GPA increases by 0.0769 points for every one-point increase in IQ score, when C2 and Sex are held constant.
- Mean GPA increases by 0.193 points for every one-point increase in C2 score, when IQ and Sex are held constant.
- Mean GPA for males is 0.909 points lower than mean GPA for females, when IQ and C2 are held constant.
- The slopes for the variables Age, SC, C1, C3, C4, C5, C6 are all interpreted as 0: These variables are not linearly related to mean GPA, after accounting for the variables IQ, C2, and Sex.

(f) We are 95% certain that such a student's GPA will be between 3.718 and 9.644 points.

(g)

$$\text{GPA} = -4.07 + 0.0797 \text{IQ} - 0.0685 \text{SC} + 0.226 \text{C1} + 0.156 \text{C2} + 0.230 \text{C5}$$

(h) Yes since all variables in the new model are significant at the 10% level and R^2 increases from 53.8% to 55.3%.

H.

- (a) The *full* model is recommended for the most accurate (unbiased) slopes. The slope for Weight in the full model is 2.1145.

Interpret:

Systolic blood pressure increases by 2.1145 points per extra kilogram of weight on average, when holding all other factors constant.

(b)

1.

$$H_0: \beta_{\text{Chin}} = \beta_{\text{Forearm}} = \beta_{\text{Calf}} = 0$$

$$H_A: \beta_{\text{Chin}} \neq 0 \quad \text{or} \quad \beta_{\text{Forearm}} \neq 0 \quad \text{or} \quad \beta_{\text{Calf}} \neq 0$$

2. $F =$ $P\text{-value} =$

3. Fail to Reject H_0 since

$$P\text{-value} = 0.29259 > 0.05 = \alpha$$

4. There is insufficient evidence to show that the skin-fold measurements cannot be dropped. So the three measurements can be dropped from the full model.

(c)

1.

$$H_0: \beta_{\text{Height}} = \beta_{\text{Weight}} = 0$$

$$H_A: \beta_{\text{Height}} \neq 0 \quad \text{or} \quad \beta_{\text{Weight}} \neq 0$$

2. $F =$ $P\text{-value} =$

3. Reject H_0 since

$$P\text{-value} = 0.000158 < 0.05 = \alpha$$

4. There is sufficient evidence to show that the body measurements cannot be dropped. So the two measurements cannot be dropped from the full model.

(d) Variables are dropped from the full model in this order:

Pulse \longrightarrow Calf \longrightarrow Forearm \longrightarrow Age \longrightarrow Height \longrightarrow Chin

(e) The fitted equation of the chosen model is

$$\text{Systolic} = 50.3 - 0.572 \text{ Years} + 1.35 \text{ Weight}$$

(f) Slopes in the chosen model:

- Systolic blood pressure decreases by 0.572 points on average for every additional year since migration, for all Peruvian Indians of the same weight.
- Systolic blood pressure increases by 1.35 points on average for each additional kilogram of weight, for all Peruvian Indians with the same number of years since migration.
- None of the other variables — Age, Height, Pulse Rate, and skin-fold measurements — are important for predicting systolic blood pressure, after accounting for Weight and Years since migration.

(g) We are 95% certain that the systolic blood pressure of an individual Peruvian Indian who fits the profile is between 109.04 and 151.50 points.

I.

(a) The model using both Year and SoyBeanYield is the best model since both Year and SoyBeanYield are highly significant. (The P -values for t tests in the full model are 0.)

$$\text{CornYield} = -1553 + 0.789 \text{ Year} + 2.91 \text{ SoyBeanYield}$$

$$R^2 = \boxed{92.2\%}$$

(b) $\boxed{1.85}$ bushels/year (This is the “total effect” of Year on CornYield.)

(c) $\boxed{0.789}$ bushels/year (This is the “direct effect” of Year on CornYield.)

(end of solution)