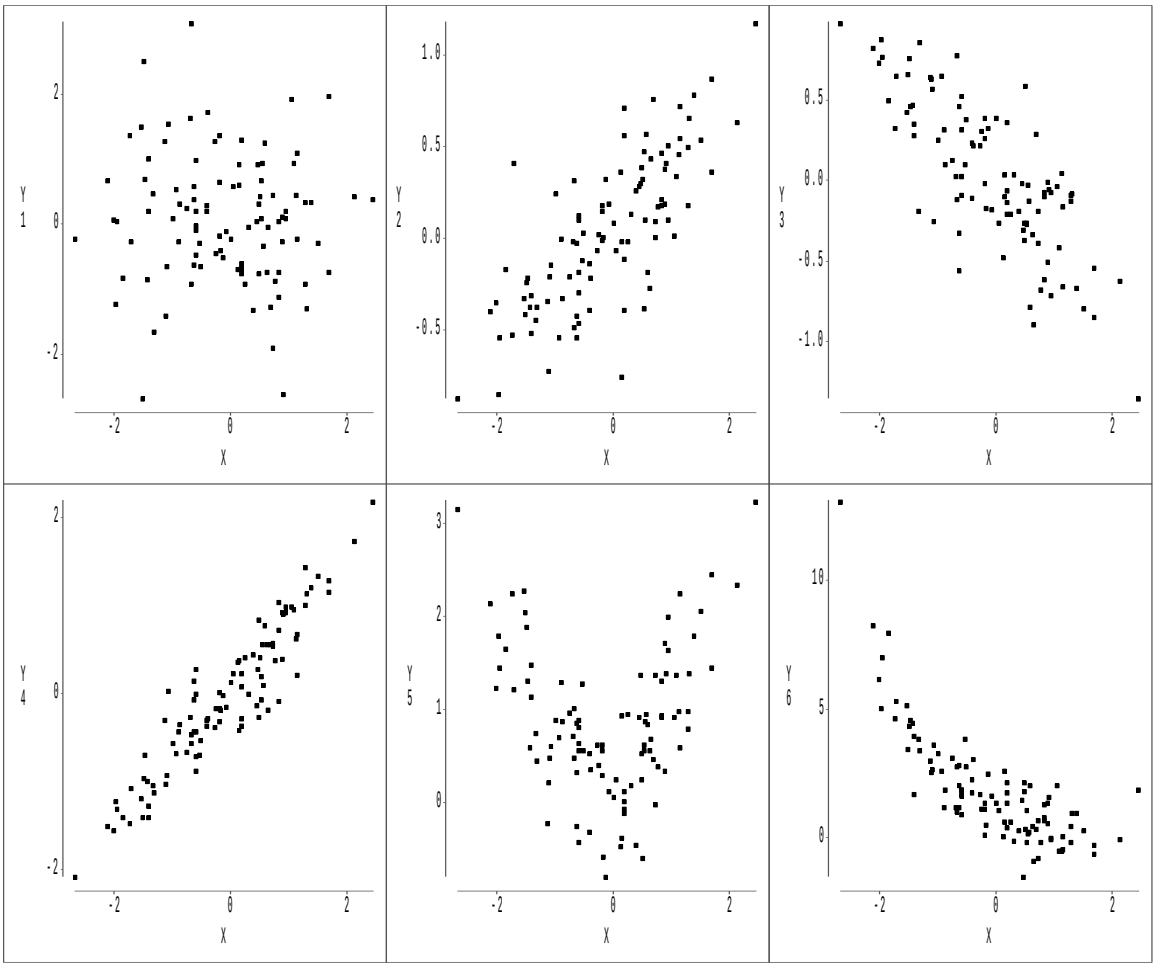
Test Your Understanding 6

Summarize the pattern of variation for each of the six scatterplots shown.



Solution:

In clockwise order from top left:

1. *Randomly scattered.*
2. *Positive linear reasonably strong association.*
3. *Negative linear reasonably strong association.*
4. *Negative nonlinear strong association Increasing variation.*
5. *Nonlinear reasonably strong association. Negative for $x < 0$, positive for $x > 0$.*
6. *Positive linear strong association.*

Test Your Understanding 7

The Pearson correlation between X and Y is 0.51. An understanding of Pearson correlation, but no calculation, is needed to answer the following questions.

a. If $W = 6 \times X$, what is the Pearson correlation between W and Y?

Solution: 0.51

b. If $U = X - 8$, what is the Pearson correlation between U and Y?

Solution: 0.51

c. If $V = -6 \times Y$, what is the Pearson correlation between V and Y?

Solution: -1.0

Test Your Understanding 8

Obtain the least squares estimates of the slope and intercept for the data

X	-1	0	1
Y	-2	1	1

Solution: *Since* $\overline{y} = \overline{x} = 0$, *The LSE of slope is*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{3}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{3}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{3} x_i y_i}{\sum_{i=1}^{3} x_i^2} = \frac{(-1)(-2) + (0)(1) + (1)(1)}{(-1)^2 + 0^2 + 1^2} = \frac{3}{2},$$

and the LSE of intercept is

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 0.$$

Test Your Understanding 9

SAS regression output for the regression of highway miles per gallon on engine displacement for a sample of new cars, is shown in the figure on the reverse.

a. Evaluate the fit of the model in terms of the residuals.

Solution: *The model is a linear function and does not seem adequate to model the nonlinear trend.*

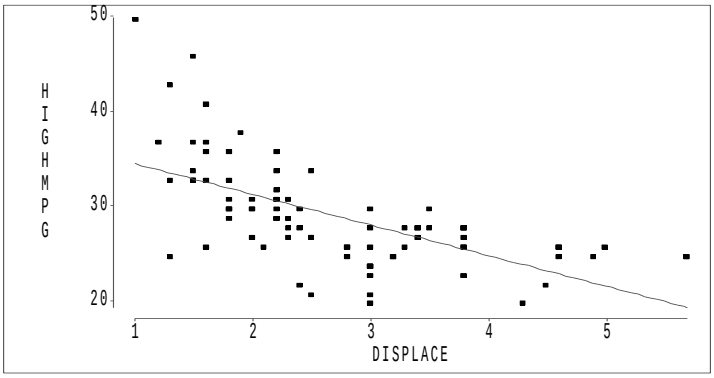
b. Evaluate the fit of the model in terms of r^2. Interpret r^2.

Solution: $r^2 = 0.3929$, *meaning that the model (i.e., the fitted linear function of DISPLACE) accounts for 39.29% of the variation in HIGHMPG. The model does not seem very successful in accounting for the variation in HIGHMPG.*

c. What is the Pearson correlation between the response and the predictor?

Solution: $r = \sqrt{0.3929} = 0.6268$.

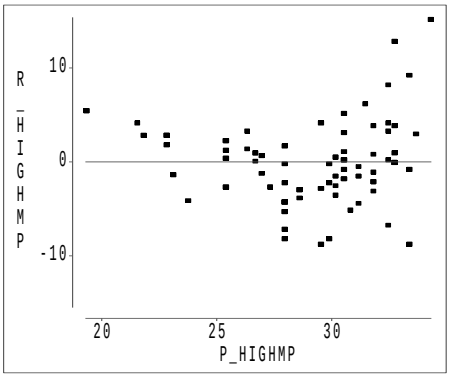
Model Equation				
HIGHMPG	=	37.6802	-	3.2215 DISPLACE



Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	1	1	1027.4813	91	17.4487	0.3929	58.8859	0.0001

Summary of Fit			
Mean of Response	29.0860	R-Square	0.3929
Root MSE	4.1772	Adj R-Sq	0.3862

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	37.6802	1.2008	31.3793	0.0001	.	0
DISPLACE	1	-3.2215	0.4198	-7.6737	0.0001	1.0000	1.0000



Test Your Understanding 10

Refer again to the regression output from TYU 9 (shown on the reverse).

a. What is the fitted model?

Solution: HIGHMPG= $37.6802 - 3.2215$ DISPLACE

b. Does the fitted intercept have a meaning of its own? Why or why not?

Solution: *No, since* $DISPLACE = 0$ *is outside the range of the data, and indeed, makes no sense.*

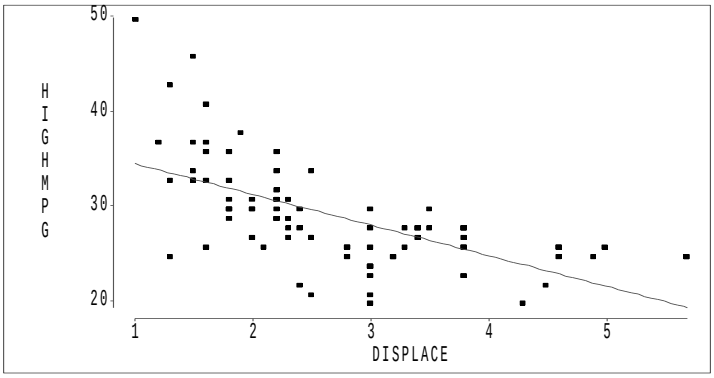
c. Interpret the slope of this line in terms that have meaning for car shoppers.

Solution: *The model estimates that average highway MPG is lower for cars with higher engine displacements, and that the rate of decrease is 3.2215 per 1 cubic inch increase in displacement.*

d. Estimate the variance of the random errors.

Solution: MSE=17.4487

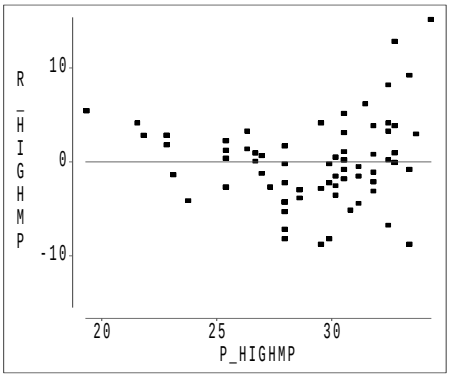
Model Equation				
HIGHMPG	=	37.6802	-	3.2215 DISPLACE



Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	1	1	1027.4813	91	17.4487	0.3929	58.8859	0.0001

Summary of Fit			
Mean of Response	29.0860	R-Square	0.3929
Root MSE	4.1772	Adj R-Sq	0.3862

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	37.6802	1.2008	31.3793	0.0001	.	0
DISPLACE	1	-3.2215	0.4198	-7.6737	0.0001	1.0000	1.0000



Test Your Understanding 11

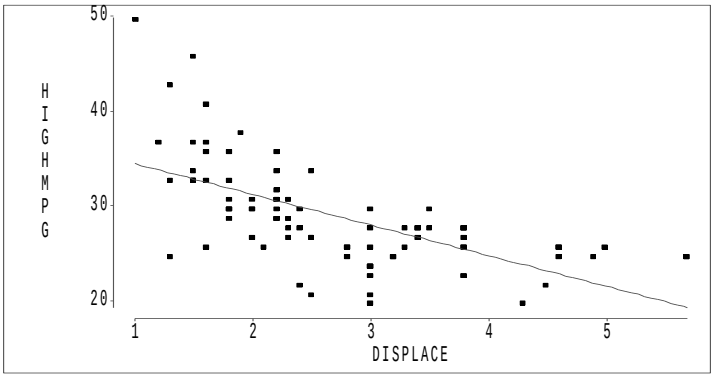
Using the regression output from TYU 9 (shown on the reverse), obtain 95% confidence intervals for β_0 and β_1. Interpret these intervals in terms of the model.

Solution: *Since* $t_{91,0.975} = 1.9864$ *is not tabled, we will use* $t_{90,0.975} = 1.9867$ *as an approximation (Note that we could also have used* $z_{0.975} = 1.96$*). Then the confidence intervals are:*

$$\begin{aligned}
\beta_0: \quad \hat{\beta}_0 \pm \hat{\sigma}(\hat{\beta}_0) t_{91,0.975} &\approx \quad 37.6802 \pm (1.2008)(1.9867) = (35.2946, 40.0658) \\
\beta_1: \quad \hat{\beta}_1 \pm \hat{\sigma}(\hat{\beta}_1) t_{91,0.975} &\approx \quad -3.2215 \pm (0.4198)(1.9867) = (-4.0555, -2.3875)
\end{aligned}$$

With 95% confidence, we estimate that the true intercept lies between 35.2946 *and* 40.0658*. With 95% confidence, we estimate that the true slope lies between* -4.0555 *and* -2.3875*. Of particular note, is that we are 95% confidence there is a negative relationship between displacement and mean highway mpg, and we estimate the decline in mean highway mpg to be between* 2.3875 *and* 4.0555 *per cubic inch increase in displacement.*

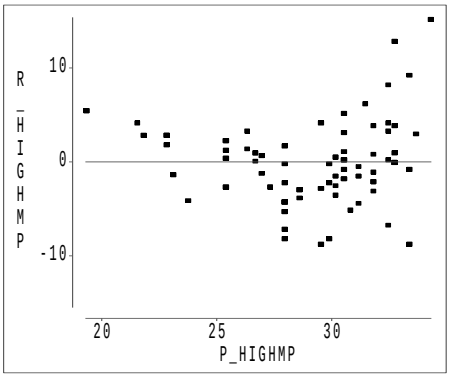
Model Equation				
HIGHMPG	=	37.6802	-	3.2215 DISPLACE



Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Prob > F
———	1	1	1027.4813	91	17.4487	0.3929	58.8859	0.0001

Summary of Fit			
Mean of Response	29.0860	R-Square	0.3929
Root MSE	4.1772	Adj R-Sq	0.3862

Parameter Estimates							
Variable	DF	Estimate	Std Error	T Stat	Prob >|T|	Tolerance	Var Inflation
INTERCEPT	1	37.6802	1.2008	31.3793	0.0001	.	0
DISPLACE	1	-3.2215	0.4198	-7.6737	0.0001	1.0000	1.0000



Test Your Understanding 12

Scientists conducted a study to see whether a specific genetic mutation in humans is associated with increased risk of colon cancer. In the study, two random samples were taken: one of individuals with colon cancer, and another of individuals without colon cancer. All individuals were then tested for the presence of the genetic mutation. The results are shown in the following 2×2 table:

		Mutation		
	Frequency Percent Row Pct. Col Pct.	Absent	Present	Total
Cancer	Absent	42 38.18 79.25 58.33	11 10.00 20.75 28.95	53 48.18
	Present	30 27.27 52.63 41.67	27 24.55 47.37 71.05	57 51.82
	Total	72 65.45	38 34.55	110 100.00

Fill in the overall, row, column, and marginal percents, and describe what these tell about the association between colon cancer and the genetic mutation.

Solution: *The values are filled-in above.*

Test Your Understanding 13

Recall TYU 13: *Scientists conducted a study to see whether a specific genetic mutation in humans is associated with increased risk of colon cancer. In the study, two random samples were taken: one of individuals with colon cancer, and another of individuals without colon cancer. All individuals were then tested for the presence of the genetic mutation. The following* 2×2 *table summarizes the results:*

		Mutation		
	Frequency	Absent	Present	Total
Cancer	Absent	42	11	53
	Present	30	27	57
	Total	72	38	110

Conduct a χ^2 test for the independence of genetic mutation and occurrence of colon cancer.

Solution:

$$E_{11} = \frac{(72)(53)}{110} = 36.691,\ E_{12} = \frac{(38)(53)}{110} = 18.309,\ E_{21} = \frac{(72)(57)}{110} = 37.309,\ E_{22} = \frac{(38)(57)}{110} = 19.691.$$

So the test statistic is

$$\chi^2 = \frac{(42-36.691)^2}{36.691} + \frac{(11-18.309)^2}{18.309} + \frac{(30-37.309)^2}{37.309} + \frac{(27-19.691)^2}{19.691} = 8.6028.$$

The p-value is $P(\chi^2_1 \geq 8.6028) = 0.0034$. *Since the p-value is less than* 0.05, *we reject the null hypothesis of independence between cancer and mutation. (If you use the table, argue that since 8.6028 exceeds* $\chi^2_{1,0.95} = 3.841$, *we reject the null hypothesis of independence between cancer and mutation.)*

Since there are only two populations sampled here (those with colon cancer and those without), we could have used the methods of chapter 6 to conduct the test. The hypotheses would be $H_0: \ p_C - p_N = 0$, *versus* $H_a: \ p_C - p_N \neq 0$, *where* p_C *is the population proportion of those with colon cancer who have the mutation and* p_N *is the population proportion of those without colon cancer who have the mutation. The standardized test statistic is*

$$Z = \frac{\hat{p}_C - \hat{p}_N}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_C} + \frac{1}{n_N}\right)}},$$

where $\hat{p}$ *is the pooled estimate of the common value of* p_C *and* p_N *under* H_0. *For these data,* $\hat{p}_C = 27/57 = 0.4737$, $\hat{p}_N = 11/53 = 0.2075$, *and* $\hat{p} = (11+27)/(53+57) = 0.3455$. *The observed value of* Z *is*

$$z^* = \frac{0.4737 - 0.2075}{\sqrt{0.3455(1-0.3455)\left(\frac{1}{57} + \frac{1}{53}\right)}} = 2.934.$$

$p^+ = P(Z \geq 2.934) = 0.0017$, $p_- = P(Z \leq 2.934) = 0.9983$, *and the p-value for this test is* $p\pm = 2\ min(p_-, p^+) = 0.0034$.