7.2 Multicollinearity.

(a) Discuss the problems that result when multicollinearity is present in a regression analysis.

Ans: High correlations among the independent variables increase the likelihood of rounding errors in the calculations of the β estimates, standard errors, and so forth. The regression results may be confusing and misleading. Multicollinearity can also have an effect on the signs of the parameter estimates.

(b) How can you detect multicollinearity?

Ans: 1. Significant correlations between pairs of independent variables in the model; 2. nonsignificant *t*-test for all (or nearly all) the individual β parameters when the *F*-test for overall model adequacy $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is significant; 3. opposite signs (from what is expected) in the estimated parameters; 4. variance inflation factor (VIF) for a β parameter greater than 10.

(c) What remedial measures are available when multicollinearity is detected?

Ans: 1. Drop one or more of the correlated independent variables form the final model. A screening procedure such as stepwise regression is helpful in determining which variables to drop; 2. if you decide to keep all the independent variables in the model: (i) avoid making inferences about the individual β parameters (such as establishing a cause-and-effect relationship between y and the predictor variables); (b) restrict inferences about $E(y)$ and future y-values to values of the independent variables that fall within the experimental region; 3. if your ultimate objective is to establish a cause-and-effect relationship between y and the predictor variables, use a designed experiment; 4. to reduce rounding errors in polynomial regression models, code the independent variables so that first-, second-, and higher-order terms for a particular x variables are not highly correlated; 5. to reduce rounding errors and stabilize the regression coefficients, use ridge regression to estimate the β parameters.

7.11 FTC cigarette study. Refer to the FTC cigarette data of Example 7.5 (p. 365). The data are saved in the FTCCIGAR file.

(a) Fit the model $E(y) = \beta_0 + \beta_1 x_1$ to the data. Is there evidence that tar content x_1 is useful for predicting carbon monoxide content y?

```
> y<-c(13.6,16.6,23.5,10.2,5.4,15.0,9.0,12.3,16.3,15.4,13.0,14.4,10.0,10
.2,9.5,1.5,18.5,12.6,17.5,4.9,15.9,8.5,10.6,13.9,14.9)
> x1<-c(14.1,16.0,29.8,8.0,4.1,15.0,8.8,12.4,16.6,14.9,13.7,15.1,7.8,11.
4,9.0,1.0,17.0,12.8,15.8,4.5,14.5,7.3,8.6,15.2,12.0)
> fit_x1<-lm(y~x1)
> fit_x1

Call:
lm(formula = y ~ x1)

Coefficients:
(Intercept)          x1
      2.743       0.801

> summary(fit_x1)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1124 -0.7167 -0.3754  1.0091  2.5450

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.74328    0.67521   4.063 0.000481 ***
x1           0.80098    0.05032  15.918 6.55e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.397 on 23 degrees of freedom
Multiple R-squared:  0.9168,    Adjusted R-squared:  0.9132
F-statistic: 253.4 on 1 and 23 DF,  p-value: 6.552e-14
```

Ans: $\hat{y} = 2.74328 + .80098x_1$; yes, $t = 15.918$

(b) Fit the model $E(y) = \beta_0 + \beta_1 x_2$ to the data. Is there evidence that nicotine content x_2 is useful for predicting carbon monoxide content y?

```
> x2<-c(.86,1.06,2.03,.67,.40,1.04,.76,.95,1.12,1.02,1.01,.90,.57,.78,.7
4,.13,1.26,1.08,.96,.42,1.01,.61,.69,1.02,.82)
> fit_x2<-lm(y~x2)
> fit_x2

Call:
lm(formula = y ~ x2)

Coefficients:
(Intercept)          x2
      1.665      12.395

> summary(fit_x2)

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3273 -1.2228  0.2304  1.2700  3.9357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6647     0.9936   1.675    0.107
x2           12.3954     1.0542  11.759 3.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.828 on 23 degrees of freedom
Multiple R-squared:  0.8574,    Adjusted R-squared:  0.8512
F-statistic: 138.3 on 1 and 23 DF,  p-value: 3.312e-11
```

Ans: $\hat{y} = 1.6647 + 12.3954x_2$; yes, $t = 11.759$

(c) Fit the model $E(y) = \beta_0 + \beta_1 x_3$ to the data. Is there evidence that weight x_3 is useful for predicting carbon monoxide content y?

```
> x3<-c(.9853,1.0938,1.1650,.9280,.9462,.8885,1.0267,.9225,.9372,.8858,.
9643,.9316,.9705,1.1240,.8517,.7851,.9186,1.0395,.9573,.9106,1.0070,.980
6,.9693,.9496,1.1184)
> fit_x3<-lm(y~x3)
> fit_x3

Call:
lm(formula = y ~ x3)

Coefficients:
(Intercept)          x3
     -11.80        25.07

> summary(fit_x3)

Call:
lm(formula = y ~ x3)

Residuals:
   Min     1Q Median     3Q    Max
-6.524 -2.533  0.622  2.842  7.268

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -11.795      9.722  -1.213   0.2373
x3            25.068      9.980   2.512   0.0195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.289 on 23 degrees of freedom
Multiple R-squared:  0.2153,    Adjusted R-squared:  0.1811
F-statistic: 6.309 on 1 and 23 DF,  p-value: 0.01948
```

Ans: $\hat{y} = -11.795 + 25.068x_3$; yes, $t = 2.512$

(d) Compare the signs of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ in the models of parts (a), (b), and (c), respectively, to the signs of the $\hat{\beta}$'s in the multiple regression model fit in Example 7.5. Is the fact that the $\hat{\beta}$'s change dramatically when the independent variables are removed from the model an indication of a serious multicollinearity problem?

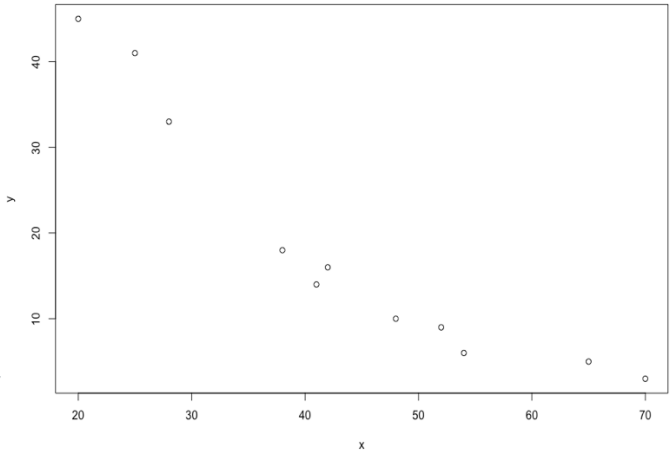
```
> X<-cbind(x1,x2,x3)
> library(mctest)
> imcdiag(X,y)

Call:
imcdiag(x = X, y = y)


All Individual Multicollinearity Diagnostics Result

       VIF    TOL       Wi       Fi Leamer    CVIF Klein
x1 21.6307 0.0462 226.9378 474.5062 0.2150 -1.7798     1
x2 21.8999 0.0457 229.8991 480.6981 0.2137 -1.8020     1
x3  1.3339 0.7497   3.6724   7.6788 0.8659 -0.1098     0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

x2 , x3 , coefficient(s) are non-significant may be due to multicollinea
rity

R-square of y on all x: 0.9186

* use method argument to check which regressors may be the reason of col
linearity
===================================
```

Ans: yes

7.20 Log–log transformation. Consider the data shown in the data below.

x	54	42	28	38	25	70	48	41	20	52	65
y	6	16	33	18	41	3	10	14	45	9	5

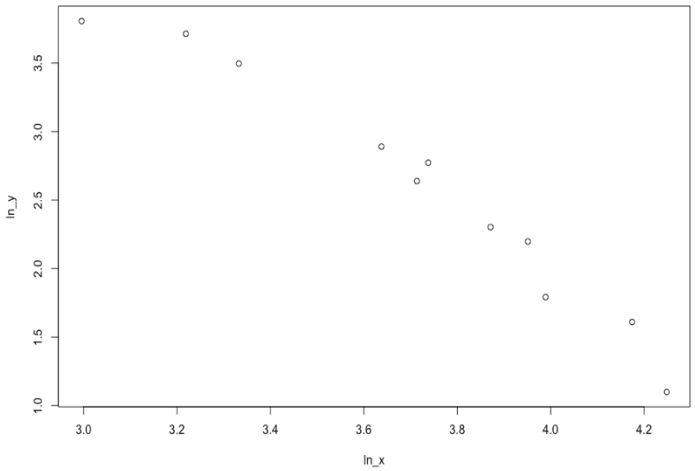
(a) Plot the points on a scatterplot. What type of relationship appears to exist between x and y?




```
> x<-c(54,42,28,38,25,70,48,41,20,52,65)
> y<-c(6,16,33,18,41,3,10,14,45,9,5)
> plot(y~x)
> fit<-lm(y~x)
> fit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    56.1573      -0.8649

> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-6.698 -4.238 -2.184  5.599  7.383

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.1573     5.2031  10.793 1.89e-06 ***
x            -0.8649     0.1120  -7.723 2.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.643 on 9 degrees of freedom
Multiple R-squared:  0.8689,    Adjusted R-squared:  0.8543
F-statistic: 59.65 on 1 and 9 DF,  p-value: 2.929e-05
```

Ans: negative nonlinear relationship between x and y

(b) For each observation calculate ln x and ln y. Plot the log-transformed data points on a scatterplot. What type of relationship appears to exist between ln x and ln y?



Ans: negative linear relationship between ln x and ln y

(c) The scatterplot from part (b) suggests that the transformed model

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

may be appropriate. Fit the transformed model to the data. Is the model adequate? Test using $\alpha = .05$.

```
> ln_x<-log(x)
> ln_y<-log(y)
> plot(ln_y~ln_x)
> fit_ln<-lm(ln_y~ln_x)
> fit_ln

Call:
lm(formula = ln_y ~ ln_x)

Coefficients:
(Intercept)         ln_x
      10.64        -2.17

> summary(fit_ln)

Call:
lm(formula = ln_y ~ ln_x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32942 -0.07912  0.06168  0.11249  0.24640

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6364     0.6028   17.64 2.73e-08 ***
ln_x         -2.1699     0.1614  -13.44 2.91e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2021 on 9 degrees of freedom
Multiple R-squared:  0.9526,    Adjusted R-squared:  0.9473
F-statistic: 180.7 on 1 and 9 DF,  p-value: 2.911e-07
```

Ans: adequate because p-value is very small

(d) Use the transformed model to predict the value of y when x = 30. [*Hint*: Use the inverse transformation $y = e^{\ln y}$.]

```
> exp(predict(fit_ln, newdata=data.frame(ln_x=log(30))))
       1
25.95282
```

Ans: y = 25.95282