Chapter 3 Solutions

3.1 a. If a student scored 100 on the midterm and 30 on the project, then we predict a final exam score of $\widehat{Final} = 11.0 + 0.53(100) + 1.20(30) = 11 + 53 + 36 = 100$

b. The predicted final exam score for Michael is $11.0 + 0.53(87) + 1.20(21) = 82.31$. Thus the residual for Michael is $80 - 82.31 = -2.31$. Michael scored 2.3 points lower on the final exam than he was expected to score based on his midterm and project scores.

3.2 a. If a cereal has 1 gram of fiber and 11 grams of sugar per serving, the model predicts the number of calories to be $\hat{Y} = 109.3 + 1.0(11) - 3.7(1) = 116.6$ calories.

b. The residual for Frosted Flakes is $y - \hat{y} = 110 - 116.6 = -6.6$ calories. Frosted Flakes has 6.6 fewer calories than the model predicts based on the amount of fiber and sugar in each serving.

3.3 The coefficient of the project score is greater than the coefficient of the midterm score, but that does not indicate a stronger relationship. To determine which predictor has a stronger relationship with the response, we need to know what the standard errors are of each predictor, which depend in part on how much each predictor varies. It might be that the correlation between project score and final exam score is smaller than the correlation between midterm score and final exam score.

3.4 The coefficient of sugar is smaller than the coefficient of fiber, but that does not indicate a weaker relationship. To determine which predictor has a weaker or stronger relationship with the response, we need to know what the standard errors are of each predictor, which depend in part on how much each predictor varies. It might be that the correlation between sugar and calories is larger than the correlation between fiber and calories.

3.5 As the score on the project goes up by 1, after accounting for the midterm grade, the average final score goes up by 1.2 points.

3.6 As the number of grams of fiber per serving goes up by 1, after accounting for the amount of sugar, the average number of calories goes down by 3.7.

3.7 a. True. Because $n - 1 > n - k - 1$, we have $\frac{SSE/(n-k-1)}{SSTotal/(n-1)} > \frac{SSE}{SSTotal}$. Thus

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SStotal/(n-1)} < R^2 = 1 - \frac{SSE}{SSTotal}$$

b. False. If a new predictor is added to a model, but that predictor explains very little extra variability in the response Y, in the presence of the other predictors, then SSE only decreases by a small amount, while $n - k - 1$ could decrease more when k goes up. This means that $SSE/(n - k - 1)$ can increase, which causes R^2_{adj} to go down.

3.8 a. This is true. Adding a new predictor to a multiple regression model can never decrease the percentage of variability explained by that model.

b. This is false. When a weak predictor is added to a multiple regression model, the R^2(adj) can decrease if the decrease in the SSE is not enough to offset the decrease in the error degrees of freedom. If the second variable (the one with the lower original R^2) is a very weak predictor, it may actually decrease the R^2(adj).

3.9 a. One would expect that men with large waist sizes have a high level of body fat, so the two variables should be positively correlated.

b. A taller man with the same waist size as a shorter man is likely to have a trimmer physique and a *lower* percentage of $BodyFat$. Thus for a fixed waist size, $BodyFat$ would be negatively correlated with $Height$.

c. If from part (b) we expect a negative correlation between $Height$ and $BodyFat$ at each given $Waist$ level, then the multiple regression coefficient of $Height$ should be negative (even if $Height$ and $BodyFat$ have zero correlation overall), because in the multiple regression we are accounting for the presence of $Waist$.

3.10 a. Positively. Every year people drive their cars, so each year the car is older, the more total miles the car will have been driven.

b. Negatively. The more miles the car has been driven, the lower the price of the used car will be.

3.11 a. A negative residual indicates that the asking price for that car is less than would be expected by the model based on year and mileage, so he would prefer dealerships with more negative residuals to get a better deal.

b. The 2-predictor model using $Year$ and $Mileage$ to predict $Price$ of cars is

$$Price = \beta_0 + \beta_1 Year + \beta_2 Mileage + \epsilon$$

c. The model could include interaction because there might be newer cars with very high mileage, which will reduce the price, and older cars with very low mileage, which will make them more attractive. This term would probably have a negative coefficient because the combination high year and high mileage would tend to mean a lower than usual price, while low mileage for a low year (older car) would tend to boost the price over what would be expected based just on the year alone.

3.12 a. The model to compare two regression lines for the relationship between metabolic rate and body size that accounts for the free growth period is

$$Mrate = \beta_0 + \beta_1 BodySize + \beta_2 Ifgp + \beta_3 BodySize \cdot Ifgp + \epsilon$$

b. If the rate of change (slope) is the same for free growth and no free growth periods (but the intercepts might differ), the appropriate model is

$$Mrate = \beta_0 + \beta_1 BodySize + \beta_2 Ifgp + \epsilon$$

c. To see if one or two different regression lines are needed, the full model is as in part (a)

$$Mrate = \beta_0 + \beta_1 BodySize + \beta_2 Ifgp + \beta_3 BodySize \cdot Ifgp + \epsilon$$

and the reduced model (which does not distinguish between the growth stages at all) is

$$Mrate = \beta_0 + \beta_1 BodySize + \epsilon$$

3.13 a. To predict $Arsenic$ using $Year$, $Miles$, and their interaction, we use

$$Arsenic = \beta_0 + \beta_1 Year + \beta_2 Miles + \beta_3 Year \cdot Miles + \epsilon$$

b. To predict $Lead$ based on $Year$ with two different lines depending on whether or not the well has been cleaned ($Iclean$), we use

$$Lead = \beta_0 + \beta_1 Year + \beta_2 Iclean + \beta_3 Year \cdot Iclean + \epsilon$$

c. To predict $Titanium$ based on a possible quadratic relationship with $Miles$, we use

$$Titanium = \beta_0 + \beta_1 Miles + \beta_2 Miles^2 + \epsilon$$

d. To predict $Sulfide$ based on $Year$, $Miles$, $Depth$, and any pairwise interactions, we use

$$\begin{aligned} Sulfide &= \beta_0 + \beta_1 Year + \beta_2 Miles + \beta_3 Depth + \\ &\quad \beta_4 Year \cdot Miles + \beta_5 Year \cdot Depth + \beta_6 Miles \cdot Depth + \epsilon \end{aligned}$$

3.14 In each case we find the degrees of freedom for the error term by subtracting the predictors in the model (plus one for the intercept) from the sample size, error df $= n - k - 1$. For each of these models, the sample size is $n = 53$.

a. $k = 3$ predictors $\Rightarrow 53 - 3 - 1 = 49$ df

b. $k = 2$ predictors $\Rightarrow 53 - 2 - 1 = 50$ df

3.15 In each case we find the degrees of freedom for the error term by subtracting the predictors in the model (plus one for the intercept) from the sample size, error df $= n - k - 1$. For each of these models, the sample size is $n = 198$.

a. $k = 3$ predictors $\Rightarrow 198 - 3 - 1 = 194$ df

b. $k = 3$ predictors $\Rightarrow 198 - 3 - 1 = 194$ df

c. $k = 2$ predictors $\Rightarrow 198 - 2 - 1 = 195$ df

d. $k = 6$ predictors $\Rightarrow 198 - 6 - 1 = 191$ df

3.16 a. To predict *Salary* based on *Age*, *Seniority*, *Pub*, *IGender*, and any pairwise interactions, we use

$$\begin{aligned} Salary &= \beta_0 + \beta_1 Age + \beta_2 Seniority + \beta_3 Pub + \beta_4 IGender + \\ &\quad \beta_5 Age \cdot Seniority + \beta_6 Age \cdot Pub + \beta_7 Age \cdot IGender + \\ &\quad \beta_8 Seniority \cdot Pub + \beta_9 Seniority \cdot IGender + \beta_{10} Pub \cdot IGender + \epsilon \end{aligned}$$

b. Yes, *Age* and *Seniority* should be positively correlated, because older faculty members would tend to have more seniority.

c. Yes, *Pub* and *Seniority* should be positively correlated, because more senior faculty members members would tend to have more publications.

d. No, the dean does not want to see significant differences in salary based on gender, especially after accounting for other factors (like age, seniority, and number of publications) that are likely to be related to salaries.

3.17 a. The test statistic is $t = 0.03420/0.03173 = 1.08$. The P-value is given in the regression output as 0.282, which is rather large. We do not have evidence that weight is associated with active pulse after accounting for resting pulse rate and exercise.

b. There are $232-4 = 228$ degrees of freedom for a t procedure, which means that the t multiplier is $t^* = 1.65$. A 90% confidence interval is given by

$$0.0342 \pm 1.65(0.03173) = 0.0342 \pm 0.0524 = (-0.0182, 0.0866)$$

We are 90% confident that, as weight increases by 1 pound, after adjusting for simultaneous linear change in resting pulse and in exercise, the average active pulse changes by between (approximately) -0.02 and 0.09 beat per minute. That is, active pulse might go down slightly or it might increase slightly. We note that zero is in the confidence interval, which means that the data are consistent with the claim that weight is not associated with active pulse after accounting for resting pulse and exercise.

c. The model predicts an active pulse rate of $\widehat{Active} = 11.8 + 1.12(76) + 0.0342(200) - 1.09(7) = 96.13$ beats per minute.

3.18 a. The simple linear regression is summarized in the following output. Each increase of one mile of distance is associated with about \$54,427 decrease in selling price. The relationship is statistically significant (P-value $= 1.57 \times 10^{-7}$), and a modest 23.74% of variation in selling price is explained by the regression. The regression standard error is \$92,130.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.204     14.052  27.626  < 2e-16 ***
distance     -54.427      9.659  -5.635 1.56e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 92.13 on 102 degrees of freedom
Multiple R-squared: 0.2374,  Adjusted R-squared: 0.2299
F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

b. The summary table from the output follows. Both explanatory variables seem important, showing P-values each well under 0.01. The effect of distance on price adjusted for the presence of *squarefeet* is -16.486, so controlling for the size of the home decreases our estimate of distances effect. Still, the relationship is significant. The R^2 has increased substantially—from 23.74% to 76.55%—so including *squarefeet* improves the models fit considerably. This is also reflected in the reduction of the regression standard error from 92.13 (thousand dollars) down to 51.34, nearly a reduction in half.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  109.742     20.057   5.472 3.25e-07 ***
distance     -16.486      5.942  -2.775  0.00659 **
squarefeet   150.780      9.998  15.080  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 51.34 on 101 degrees of freedom
Multiple R-squared: 0.7655,Adjusted R-squared: 0.7608
F-statistic: 164.8 on 2 and 101 DF,  p-value: < 2.2e-16
```

c. The following R output shows a 95% confidence interval from the simple linear model. We are 95% confident that the increase in price for each mile closer to a trail is between -73.586 and -35.269 (thousand dollars). The following R output also shows the 95% confidence interval for the distance coefficient adjusting for the presence of *squarefeet* in the model. In this case we are 95% confident that the true coefficient is between -28.273 and -4.699 thousand dollars. The two answers are quite different in magnitude and the latter one is somewhat narrower (the margins of error are 19.2 and 11.8 for the one-predictor and two-predictor models, resp.)

```
> moe = qt(.975,102)*9.659
> -54.4272 - moe #LCL
[1] -73.58578
```

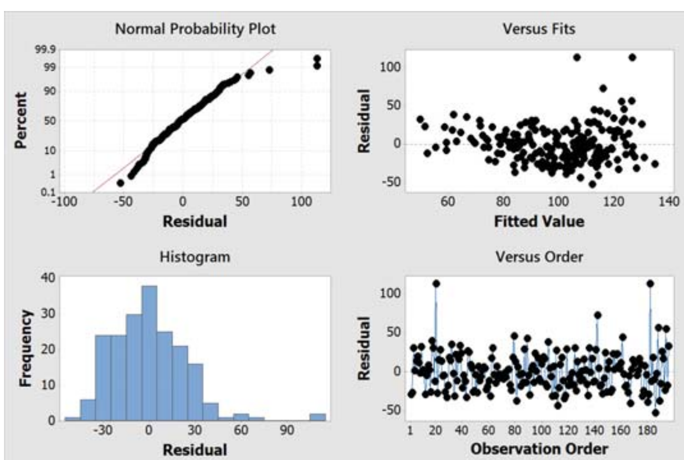
```
> -54.4272 + moe #UCL
[1] -35.26862

> moe = qt(.975,101)*5.942
> -16.486 - moe
[1] -28.27333
> -16.486 + moe
[1] -4.69867
```

d. Plugging in a distance of 0.5 mile and 1500 squarefeet of space gives $\widehat{adj2007} = 109.742 - 16.486(0.5) + 150.78(1500) = 226,271$. We predict the price to be \$226,217.

3.19 a. The following computer output shows that the predicted equation is $\widehat{Animus} = 124.3 + 0.564Black - 0.578Hispanic - 2.054BachPlus + 1.519Age65Plus$

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     124.3     20.1     6.19    0.000
Black        0.564    0.182     3.10    0.002  1.23
Hispanic    -0.578    0.129    -4.48    0.000  1.28
BachPlus    -2.054    0.344    -5.97    0.000  1.24
Age65Plus    1.519    0.889     1.71    0.089  1.55

Regression Equation
Animus = 124.3 + 0.564Black - 0.578Hispanic - 2.054BachPlus + 1.519Age65Plus
```

b. The residual plots show that there may be a problem with normality and there is a problem with constant variance.

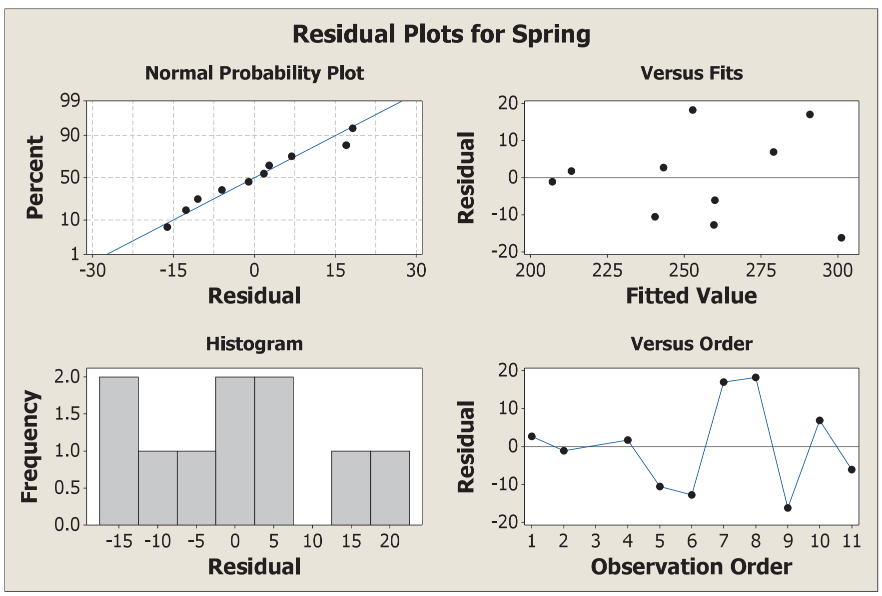


3.20 a. Here is some output for fitting a multiple regression model to predict *Spring* enrollment using *Fall* enrollment and *Ayear* (after deleting the data for 2003).

```
The regression equation is Spring = - 11716 - 1.01 Fall + 6.11 Ayear

Predictor      Coef  SE Coef      T      P
Constant     -11716     2686  -4.36  0.003
Fall        -1.0069   0.2041  -4.93  0.002
Ayear         6.107    1.337   4.57  0.003

S = 13.3668   R-Sq = 87.1%   R-Sq(adj) = 83.4%
```

The fitted prediction equation is $\widehat{Spring} = -11716 - 1.0069 Fall + 6.107 Ayear$.

b. Some plots of the residuals follow. They reveal no significant problems with the conditions for a regression model. The residual versus fits plot shows a good random scatter above and below the zero line. The normal probability plot tracks along a straight line indicating appropriate normality, even if the small sample size prevents us from seeing a clear bell curve in the histogram. In particular, the problem with residuals increasing over time that we saw in the model without a predictor for academic year is not longer an issue.



3.21 Here is some computer output for fitting a multiple regression model to predict *Spring* enrollment using *Fall* enrollment and *Ayear* (after deleting the data for 2003).

```
The regression equation is Spring = - 11716 - 1.01 Fall + 6.11 Ayear
```

```
Predictor      Coef  SE Coef      T      P
Constant     -11716     2686  -4.36  0.003
Fall        -1.0069   0.2041  -4.93  0.002
Ayear         6.107    1.337   4.57  0.003

S = 13.3668   R-Sq = 87.1%   R-Sq(adj) = 83.4%
```

a. The value of `R-Sq = 87.1%` tells us that 87.1% of the variability in spring math enrollments is explained by this combination of fall enrollments and academic year.

b. The size of a typical error from the actual spring enrollments is reflected in the regression standard error, $\hat{\sigma}_\epsilon = 13.4$.

c. Here is the ANOVA table for assessing the effectiveness of this model.

```
Source           DF      SS      MS      F      P
Regression        2  8446.9  4223.4  23.64  0.001
Residual Error    7  1250.7   178.7
Total             9  9697.6
```

We are testing $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The F-statistic is 23.64, which gives a P-value of 0.001 using an F-distribution with 2 and 7 degrees of freedom. This small P-value gives evidence to reject H_0 and conclude that at least one of the *Fall* and *Ayear* predictors is helpful to explain spring enrollments.

d. To test the effectiveness of *Fall* in this model, the hypotheses are $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. The t-statistic for this coefficient in the output is $t = -4.93$, which gives a small P-value of 0.002. This indicates that fall enrollment is a useful predictor for spring enrollments in this model.

To test the effectiveness of *Ayear* in this model, the hypotheses are $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$. The t-statistic for this coefficient in the output is $t = 4.57$, which gives a small P-value of 0.003. This indicates that the academic year is also a useful predictor for spring enrollments in this model.

3.22 a. $R^2 = \frac{SSModel}{SSTotal} = \frac{9350}{17{,}190} = 0.544$. We can explain 54.4% of the variability in calories in these cereals by using grams of sugar and grams of fiber in a multiple regression model.

b. The regression standard error is

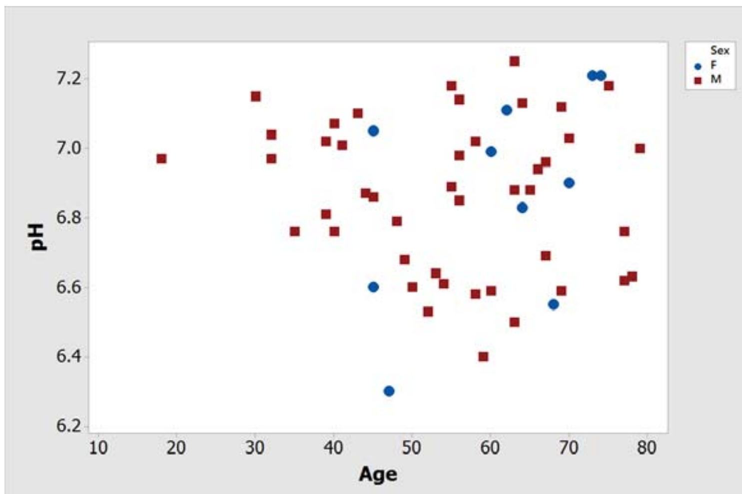
$$\hat{\sigma}_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{7840}{36-2-1}} = \sqrt{237.58} = 15.4$$

c. The F-test statistic is

$$F = \frac{MSModel}{MSE} = \frac{9350/2}{7840/33} = \frac{4675}{237.58} = 19.7$$

d. The Anova F-ratio is testing $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 = 0$ or $\beta_2 = 0$. Using technology, the P-value for $F = 19.7$ using an the upper tail of an F-distribution with 2 and 33 degrees of freedom is 0.00002. This small P-value gives very strong evidence to reject this null hypothesis; thus either the amount of sugar or fiber (or both) are probably related to calories in cereals.

3.23 a. The plot follows. There does not seem to be much relationship between pH and Age.



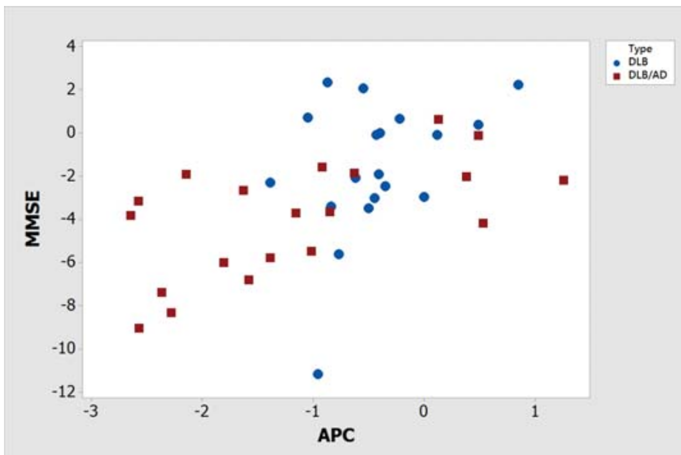
b. The output follows. With a $t = -0.17$, and a P-value $= 0.866$, we do not reject the null hypothesis of no linear relationship.

```
Coefficients
Term           Coef   SE Coef  T-Value  P-Value   VIF
Constant      6.888     0.132    52.14    0.000
Age        -0.00039   0.00229    -0.17    0.866  1.00

Regression Equation
pH = 6.888 - 0.00039Age
```

c. The following output gives the fitted prediction model as $\widehat{pH} = 6.903 - 0.00045Age - 0.0134Sex_M$. Thus for males the prediction model is $\widehat{pH} = 6.89 - 0.00045Age$ and for females the prediction model is $\widehat{pH} = 6.903 - 0.00045Age$.

```
Coefficients
Term           Coef  SE Coef  T-Value  P-Value
Constant      6.903    0.161    42.77    0.000
Age        -0.00045  0.00235    -0.19    0.848
Sex_M       -0.0134   0.0843    -0.16    0.874

Regression Equation
pH = 6.903 - 0.00045Age - 0.0134Sex_M
```

3.24 a. The following scatterplot shows the DLB points as blue circles and the DLB/AD points as red squares, which tend to be slightly lower on the APC scale. Overall, there is an upward trend between APC and MMSE.



b. The following output shows that the t-statistic is 3.97 and the P-value is approximately 0. Yes, there is a linear association between the two variables.

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant    -1.421    0.553    -2.57    0.014
APC          1.746    0.440     3.97    0.000  1.00

Regression Equation
MMSE = -1.421 + 1.746APC
```

c. The following output gives the fitted prediction model as $\widehat{MMSE} = -0.94 + 1.5APC - 1.31TypeDLB/AD$. Thus when *Type* is DLB the prediction model is $\widehat{MMSE} = -0.94 + 1.5APC$ and when *Type* is DLB/AD the prediction model is $\widehat{MMSE} = -2.25 + 1.5APC$.

```
Coefficients

Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant     -0.943    0.636    -1.48    0.147
APC           1.501    0.465     3.23    0.003  1.15
Type_DLB/AD  -1.314    0.902    -1.46    0.154  1.15


Regression Equation

MMSE = -0.943 + 1.501APC - 1.314Type_DLB/AD
```

3.25 a. The following output shows that the coefficient of *PaperTrail* is -16.6 and the P-value for the t-test is 0.003. In states with a paper trail Clinton did worse than in states without a paper trail. On average the difference was 16.6 delegates.

```
Coefficients
Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant      65.13     3.87    16.83    0.000
Paper Trail  -16.60     5.08    -3.27    0.003  1.00

Regression Equation
Delegates = 65.13 - 16.60PaperTrail
```

b. The following output shows that the coefficient of *PaperTrail* is -6.15 and the P-value for the t-test is 0.13. Controlling for the percentage of African Americans in each state, the effect of having a paper trail is negative but is not statistically significantly different from zero. The effect of *AfAmPercent* is highly significant: The higher the percentage of African Americans in a state, the higher the percentage of delegates won by Clinton.

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     42.17     4.75     8.88    0.000
Paper Trail  -6.15     3.91    -1.57    0.127  1.27
AfAmPercent  1.167    0.200     5.83    0.000  1.27

Regression Equation
Delegates = 42.17 - 6.15PaperTrail + 1.167AfAmPercent
```

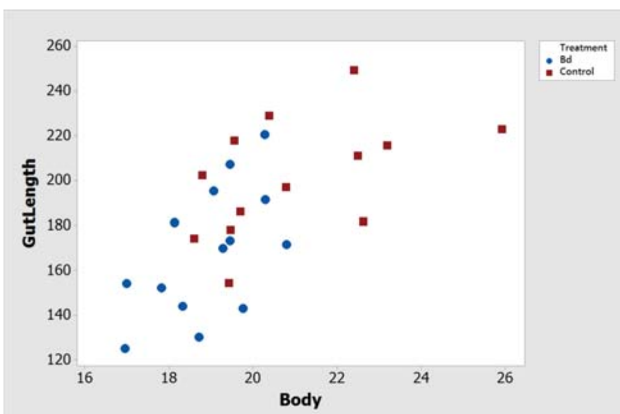
c. The output follows. The first set of output shows that when *PaperTrail* is used as the only predictor it is highly significant; the P-value is 0.001 for the t-test of the null hypothesis that there is no linear relationship between *PopularVote* and *PaperTrail*.

```
Coefficients
Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant      64.16     3.32    19.33    0.000
Paper Trail  -15.74     4.36    -3.61    0.001  1.00

Regression Equation
PopularVote = 64.16 - 15.74PaperTrail
```

When both *PaperTrail* and *AfAmPercent* are used as predictors, *AfAmPercent* has a highly significant relationship with *PopularVote*, but the coefficient of *PaperTrail* has a t-test P-value of 0.053. (The output follows.)

```
Coefficients

Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     43.04     3.68    11.69    0.000
Paper Trail  -6.13     3.03    -2.02    0.053  1.27
AfAmPercent  1.073    0.155     6.91    0.000  1.27


Regression Equation

PopularVote = 43.04 - 6.13PaperTrail + 1.073AfAmPercent
```

3.26 a. The following graph shows blue circles as the Bd points and red squares as the control points. The control points tend to be slightly higher on the *GutLength* scale. Overall, there is an upward trend between *Body* and *GutLength*.



b. The following output shows that the t-statistic is 4.20 and the P-value is approximately 0. Yes, there is a linear association between the two variables.

```
Coefficients
Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  -20.8     49.0    -0.42    0.675
Body      10.28     2.45     4.20    0.000  1.00

Regression Equation
GutLength = -20.8 + 10.28Body
```

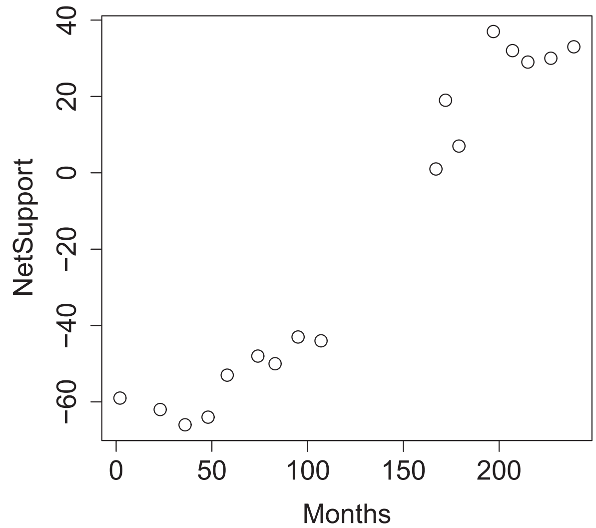
c. The following output gives the fitted prediction model as $\widehat{GutLength} = 31.7 + 8.07Body - 16.3Treatment$. Thus when *Treatment* is Bd, the prediction model is $\widehat{GutLength} = 15.4 + 8.07Body$ and when *Treatment* is Control, the prediction model is $\widehat{GutLength} = 31.7 + 8.07Body$. (The coefficient of *Treatment* is not statistically significantly different from zero, but with a small sample size we don't expect statistical significance. The important thing is the negative sign of the coefficient.)

```
Coefficients
Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant       31.7     59.7     0.53    0.599
Body           8.07     2.82     2.86    0.009  1.39
Treatment     -16.3     11.0    -1.48    0.153  1.39

Regression Equation
GutLength = 31.7 + 8.07Body - 16.3Treatment
```

d. The following output gives the fitted prediction model as $\widehat{GutLength} = 5.2 + 6.44Body - 25.4TreatmentBd + 96.8MouthpartDamage$. The fitted model says that Bd is associated with shorter intestinal length, of about 25.4 mm, at a given Body and MouthpartDamage. This is the opposite of what the biologists expected.

```
Coefficients
Term              Coef  SE Coef  T-Value  P-Value   VIF
Constant           5.2     57.2     0.09    0.929
Body              6.44     2.75     2.35    0.028  1.51
Treatment_Bd     -25.4     11.2    -2.27    0.033  1.64
MouthpartDamage   96.8     45.8     2.11    0.046  1.18

Regression Equation
GutLength = 5.2 + 6.44Body - 25.4Treatment_Bd + 96.8MouthpartDamage
```

3.27 a. The scatterplot shows a clear upward trend in *NetSupport* over time, with two clusters of points one prior to month 110 and one after month 150.



b. Here is computer output for fitting the model $NetSupport = \beta_0 + \beta_1 Months + \epsilon$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.31441    5.01290  -16.22 6.40e-11 ***
Months        0.50791    0.03415   14.88 2.19e-10 ***
```

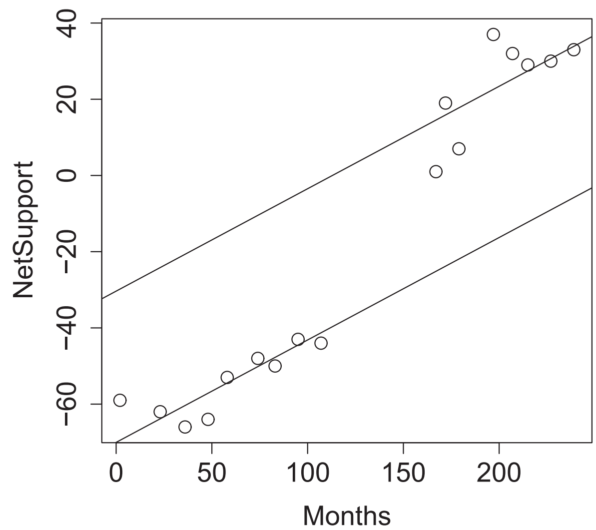
We see a very large test statistic ($t = 14.88$) and small P-value (2.19×10^{-10}) for testing the coefficient of $Months$. This provides strong evidence that the slope is not zero and $NetSupport$ tends to increase as $Months$ beyond August 1975 increase.

c. The model for parallel lines is $NetSupport = \beta_0 + \beta_1 Months + \beta_2 Late + \epsilon$, where $Late$ is the indicator for months after August 1984. Some computer output for fitting this model follows.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -70.04051    4.39829 -15.924  2.3e-10 ***
Months        0.26875    0.06208   4.329 0.000693 ***
Late         39.68893    9.52692   4.166 0.000951 ***
```

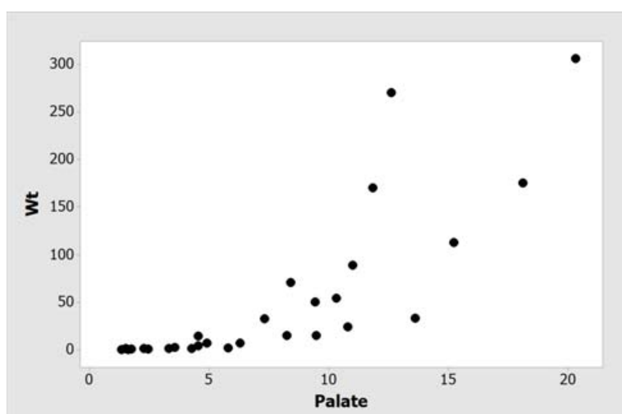
The fitted prediction equation is $\widehat{NetSupport} = -70.04 + 0.26875 Months + 39.689 Late$.

A plot of the two regression lines (which is not asked for in the exercise) follows.

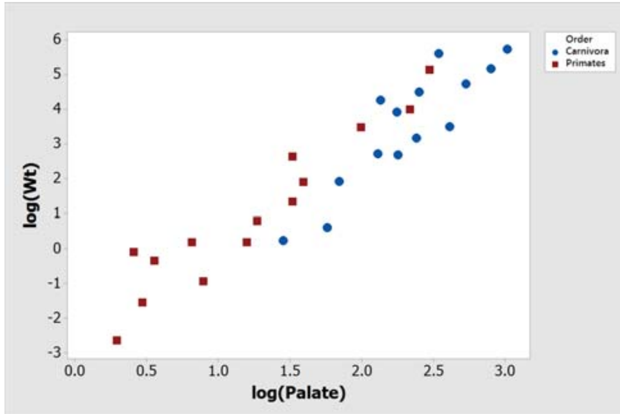


d. To test if we need two parallel lines rather than a single line for both time periods, we consider $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 = 0$, where β_2 is the coefficient of the *Late* indicator in the model of part (c). From that output, we see the P-value of this test is 0.000951, which is very small, so we have strong evidence that the intercepts of the two lines should be different and we need to use the two parallel lines to describe the relationship more adequately.

3.28 a. The scatterplot is:



b. The following scatterplot shows the Carnivora points as blue circles and the Primates points as red squares, which tend to be higher on the $\log(Wt)$ scale. Overall, there is a linear, upward trend between $\log(Palate)$ and $\log(Wt)$.



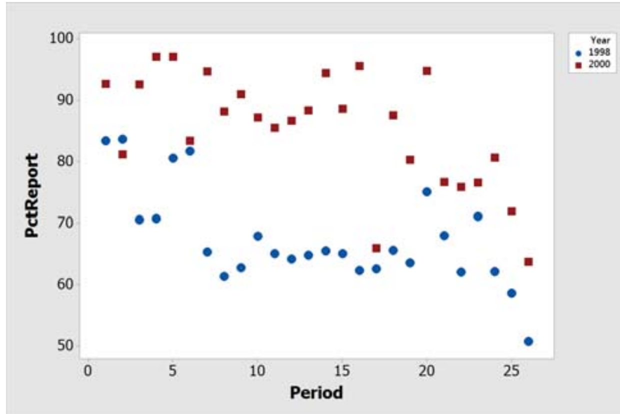
c. The following output shows that the t-statistic is 14.02 and the P-value is essentially zero. Yes, there is a linear association between the two variables.

```
Coefficients
Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant      -2.608    0.378    -6.90    0.000
log(Palate)    2.737    0.195    14.02    0.000  1.00

Regression Equation
log(Wt) = -2.608 + 2.737log(Palate)
```

d. The following output gives the fitted prediction model as $\widehat{\log(Wt)} = -3.738 + 3.126 \log(Palate) + 0.884 OrderPrimates$. Thus when Order is Carnivora, the prediction model is $\widehat{\log(Wt)} = -3.738 + 3.126 \log(Palate)$, and when Order is Primates, the prediction model is $\widehat{\log(Wt)} = -3.738 + 3.126 \log(Palate) + 0.884 = -2.857 + 3.126 \log(Palate)$.

```
Coefficients

Term              Coef  SE Coef  T-Value  P-Value   VIF
Constant        -3.738    0.610    -6.13    0.000
log(Palate)      3.126    0.250    12.52    0.000  1.89
OrderPrimates    0.884    0.390     2.26    0.032  1.89


Regression Equation

log(Wt) = -3.738 + 3.126log(Palate) + 0.884OrderPrimates
```

3.29 a. The following scatterplot shows that over the year, the percentage of potential jurors reporting for duty decreases. But the points for the year 2000 are, indeed, higher than the points for 1998, which suggests that the new methods are working.



b. The following output gives a slope of -0.717, which is significant with $t = -3.44$ and P-value $= 0.001$.

```
Coefficients
Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   86.00     3.22    26.74    0.000
Period    -0.717    0.208    -3.44    0.001  1.00

Regression Equation
PctReport = 86.00 - 0.717Period
```

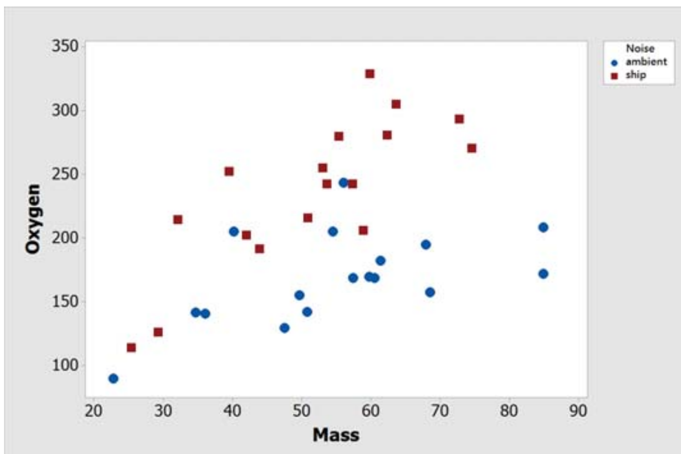
c. The following output provides a model of $\widehat{PctReport} = 77.08 - 0.717 Period + 17.83 I2000$. This results is a model of $\widehat{PctReport} = 77.08 - 0.717 Period$ for 1998, and a model of $\widehat{PctReport} = 94.91 - 0.717 Period$ for 2000. Since the intercept is significantly larger for 2000, it appears that the methods are working.

```
Coefficients
Term         Coef  SE Coef  T-Value  P-Value
Constant   77.08     2.13    36.19    0.000
Period    -0.717    0.124    -5.78    0.000
I2000      17.83     1.86     9.58    0.000

Regression Equation
PctReport = 77.08 - 0.717Period + 17.83I2000
```

d. The following output provides a model of $\widehat{PctReport} = 76.43 - 0.668 Period + 19.15 I2000 - 0.097 Period \cdot I2000$. This results is a model of $\widehat{PctReport} = 76.43 - 0.668 Period$ for 1998, and a model of $\widehat{PctReport} = 95.58 - 0.765 Period$ for 2000. The interaction term is not significant, so we do not have enough evidence to suggest that the slopes are different in the two different years.

```
Coefficients
Term            Coef  SE Coef  T-Value  P-Value
Constant       76.43     2.73    27.96    0.000
Period        -0.668    0.177    -3.78    0.000
I2000          19.15     3.87     4.95    0.000
Period*I2000  -0.097    0.250    -0.39    0.699

Regression Equation
PctReport = 76.43 - 0.668Period + 19.15I2000 - 0.097Period*I2000
```

3.30 a. The following scatterplot shows the ambient points as blue circles and the ship points as red squares, which tend to be higher on the *Oxygen* scale. Overall, there is an upward trend between *Mass* and *Oxygen*.



b. The following output shows that the t-statistic is 2.94 and the P-value is 0.006. Yes, there is a linear association between the two variables.

```
Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant   108.4     33.3     3.26    0.003
Mass       1.767    0.601     2.94    0.006  1.00

Regression Equation
Oxygen = 108.4 + 1.767Mass
```

c. The following output gives the fitted prediction model as $\widehat{Oxygen} = 54.4 + 2.07Mass + 75.3Noiseship$. Thus when *Noise* is "ambient" the prediction model is $\widehat{Oxygen} = 54.4 + 2.07Mass$ and when *Noise* is "ship" the prediction model is $\widehat{Oxygen} = 129.7 + 2.07Mass$.

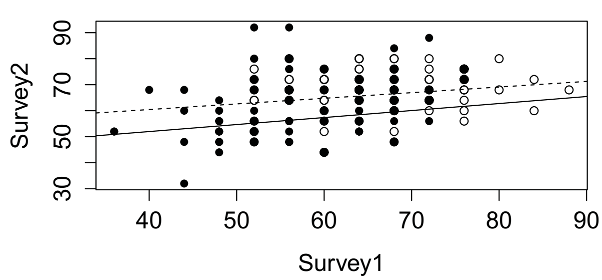
```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant      54.4     25.0     2.18    0.037
Mass         2.073    0.423     4.90    0.000
Noiseship     75.3     12.8     5.88    0.000

Regression Equation
Oxygen = 54.4 + 2.073Mass + 75.3Noise_ship
```

d. The following output gives the fitted prediction model as $\widehat{Oxygen} = 103.3 + 1.19Mass - 34.4Noiseship + 2.07Mass \cdot Noiseship$. Thus when *Noise* is "ambient," the prediction model is $\widehat{Oxygen} = 103.3 + 1.19Mass$, and when *Noise* is "ship," the prediction model is $\widehat{Oxygen} = 68.9 + 3.26Mass$.

```
Coefficients
Term              Coef  SE Coef  T-Value  P-Value
Constant         103.3     29.4     3.51    0.001
Mass             1.187    0.512     2.32    0.027
Noiseship        -34.4     43.1    -0.80    0.431
Mass*Noiseship   2.071    0.783     2.65    0.013

Regression Equation
Oxygen = 103.3 + 1.187Mass - 34.4Noise_ship + 2.071Mass*Noise_ship
```

e. The interaction term in the part (d) model is statistically significant (P-value $= 0.013$), so this most general model is the one that should be used.

3.31 a. The *Survey1* vs. *Survey2* relationship is similar for both men and women, except that the women's intercept is higher, reflecting their overall better success rate with identifying gender of author. The slopes are very similar. In the graph, the open circles are for women and filled circles are for men.



b. The model summary is given here:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.63051    4.59354   9.498  < 2e-16 ***
Survey1      0.31330    0.07759   4.038 7.79e-05 ***
Gender       3.46231    1.48740   2.328    0.021 *
---
Residual standard error: 9.233 on 192 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1649,Adjusted R-squared:  0.1562
F-statistic: 18.96 on 2 and 192 DF,  p-value: 3.058e-08
```

The estimated slope in this model is 0.31330 and assumed under the model the same for both men and women. But this assumption of parallel regressions seems consistent with the scatterplot in part (a). The significant *Gender* coefficient reflects a statistically significantly greater intercept for women than for men, by about 3.46.

c. We next fit a model with both *Gender* and $Gender \cdot Survey1$ terms. The summary is here:

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     37.4606     6.3529   5.897 1.66e-08 ***
Survey1          0.4200     0.1085   3.870 0.000149 ***
Gender          16.9725     9.7494   1.741 0.083316 .
Survey1:Gender  -0.2171     0.1548  -1.402 0.162516
---
Residual standard error: 9.21 on 191 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1734,Adjusted R-squared:  0.1605
F-statistic: 13.36 on 3 and 191 DF,  p-value: 5.942e-08
```

This fit of this model is similar to the previous one, with a R^2 gain of only 16.49% to 17.34%. The significance of the intercept difference has, perhaps, lost its statistical significance, at the expense of fitting a different slope for the men and women, even though this slope difference is small and nonsignificant.

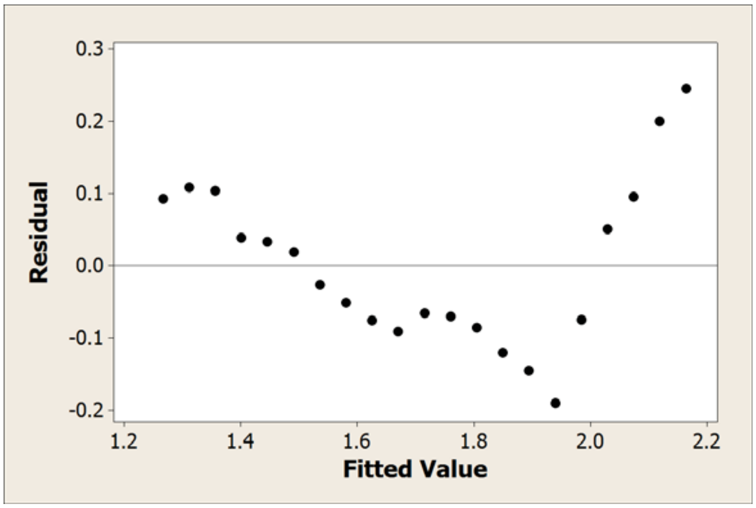
3.32 a. Here is computer output for fitting $FatalityRate = \beta_0 + \beta_1 Year + \epsilon$

```
The regression equation is
FatalityRate = 91.3 - 0.0449 Year

Predictor       Coef   SE Coef       T      P
Constant      91.321     8.374   10.90  0.000
Year       -0.044870  0.004193  -10.70  0.000
```

```
S = 0.116362   R-Sq = 85.8%   R-Sq(adj) = 85.0%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  1.5503  1.5503  114.49  0.000
Residual Error  19  0.2573  0.0135
Total           20  1.8075
```

The slope of the regression line is $\hat{\beta}_1 = -0.0449$, indicating that fatality rates decline, on average, by about 0.0449 death per 100 million vehicle miles each year.

b. A plot of residuals versus fitted values shows a distinct V shape.



c. Here is computer output for fitting $FatalityRate = \beta_0 + \beta_1 Year + \beta_2 StateControl + \beta_3 Year \cdot StateControl + \epsilon$

```
The regression equation is
FatalityRate = 216 - 0.108 Year - 161 StateControl + 0.0810 Year*StateControl

Predictor                Coef   SE Coef       T      P
Constant               216.23     13.03   16.59  0.000
Year                -0.107619  0.006548  -16.44  0.000
StateControl          -161.38     14.47  -11.15  0.000
Year*StateControl    0.080971  0.007264   11.15  0.000

S = 0.0424331   R-Sq = 98.3%   R-Sq(adj) = 98.0%
```

```
Analysis of Variance
Source          DF       SS       MS       F      P
Regression       3  1.77691  0.59230  328.95  0.000
Residual Error  17  0.03061  0.00180
Total           20  1.80752
```

The t-tests for all of the coefficients have very small P-values. In particular, the significant coefficients for $StateControl$ and the interaction term indicate that we have strong evidence that the relationship between fatality rate and year is different before and after 1995.

d. For the years before 1995, $StateControl = 0$, so the prediction equation from the interaction model reduces to

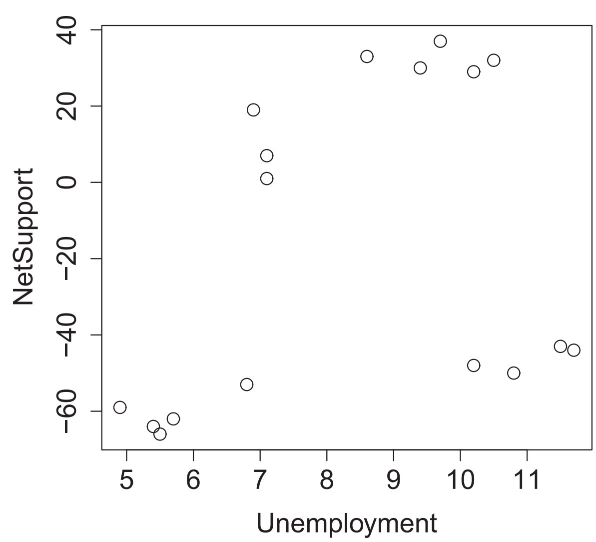
$$\widehat{FatalityRate} = 216.23 - 0.1076Year$$

For the years after (and including) 1995, the value of $StateControl = 1$, so the interaction model gives

$$\widehat{FatalityRate} = 216.23 - 0.1076Year - 161.38(1) + 0.08097Year(1) = 54.85 - 0.02663Year$$

Note that the fatality rate was dropping much more sharply before 1995 than after.

3.33 a. The plot shows several clusters of points, but no consistent linear relationship between $NetSupport$ and $Unemployment$.



b. Here is computer output for fitting $NetSupport = \beta_0 + \beta_1 Unemployment + \epsilon$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -67.660     37.862  -1.787   0.0942 .
Unemployment    5.980      4.379   1.366   0.1921
```

From the regression output, the test statistic is 1.366 and the P-value is 0.1921, which is larger than 0.10. We do not have sufficient evidence to conclude that $Unemployment$ is linearly related to $NetSupport$.

c. Here is computer output for fitting $NetSupport = \beta_0 + \beta_1 Unemployment + \beta_2 Months + \epsilon$.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -65.51220    9.27541  -7.063 5.66e-06 ***
Unemployment  -2.35767    1.20207  -1.961     0.07 .
Months         0.53898    0.03508  15.362 3.71e-10 ***
```

Using the regression output, to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 = 0$ the test statistic is -1.961 and the P-value is 0.07, which is smaller than 0.10. At this significance level, we reject the hypothesis that $Unemployment$ is not linearly related to $NetSupport$ when controlling for $Months$ being in the model. Thus when $Months$ and $Unemployment$ are both used, each of them has a significant relationship with $NetSupport$.

d. In part (b), the coefficient of $Unemployment$ is positive—showing that $NetSupport$ is slightly higher when unemployment is higher—but in part (c), the coefficient is negative—showing that, after adjusting for $Months$, higher rates of $Unemployment$ are associated with lower $NetSupport$ for British unions.

3.34 a. From computer output,

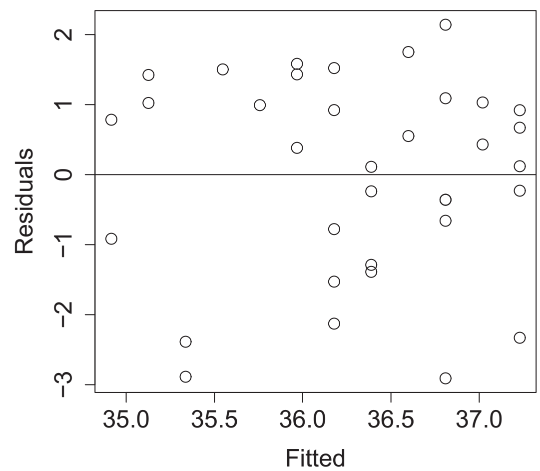
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.70206    0.86783  44.596   <2e-16 ***
Age         -0.21033    0.07313  -2.876    0.007 **

Residual standard error: 1.426 on 33 degrees of freedom
Multiple R-squared: 0.2004,Adjusted R-squared: 0.1762
F-statistic: 8.272 on 1 and 33 DF,  p-value: 0.007001
```

we have $\widehat{PctDM} = 38.70206 - 0.21033Age$. The fitted model indicates that as Age increases, $PctDM$ decreases.

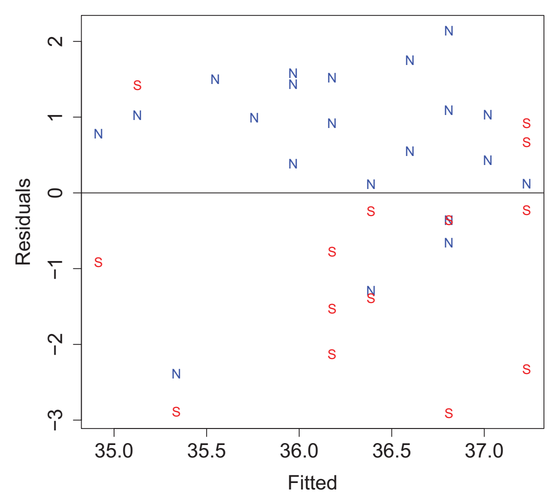
b. For the output, we have $R^2 = 0.2004$, which means that 20% of the variability in $PctDM$ is explained by Age.

c. The P-value for the t-test is 0.007 (which agrees with the P-value for the F-test). Since this value is small, the relationship between Age and $PctDM$ is statistically significant.

d. Following is a plot of residuals versus fitted values for this model. There is no evident pattern here.



e. Here is a plot with September points (plotting symbol "S") and November points (plotting symbol "N"). The September residuals tend to be negative and the November residuals tend to be positive. Now fit a multiple regression model, using an indicator ($Sept$) for the month and interaction product, to compare the regression lines for September and November.



The fitted model is $\widehat{PctDM} = 39.40 - 0.218Age - 1.276Sept - 0.0214Age \cdot Sept$, where $Sept$ is 1 for September and 0 for November.

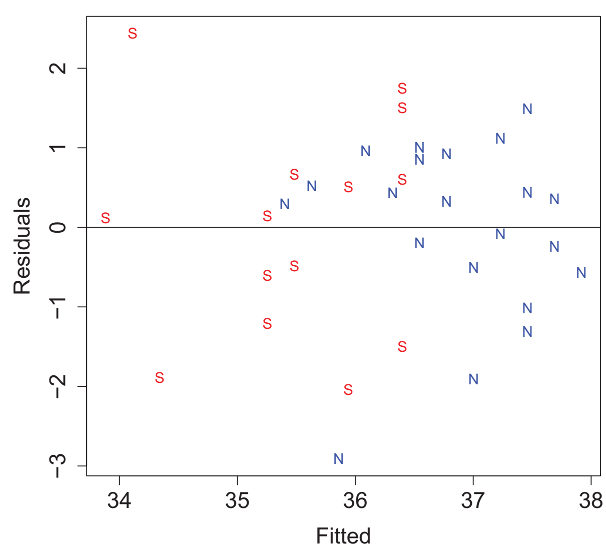
f. The *Sept* indicator and interaction term from part (e) are both not significant (P-value $= 0.4051$ and P-value $= 0.8679$). However, deleting the interaction term and fitting the reduced model gives

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.51922    0.77827  50.778  < 2e-16 ***
Age         -0.22870    0.06292  -3.635 0.000965 ***
Sept        -1.51929    0.42342  -3.588 0.001096 **

Residual standard error: 1.223 on 32 degrees of freedom
Multiple R-squared: 0.4298,Adjusted R-squared: 0.3942
F-statistic: 12.06 on 2 and 32 DF,  p-value: 0.0001248
```

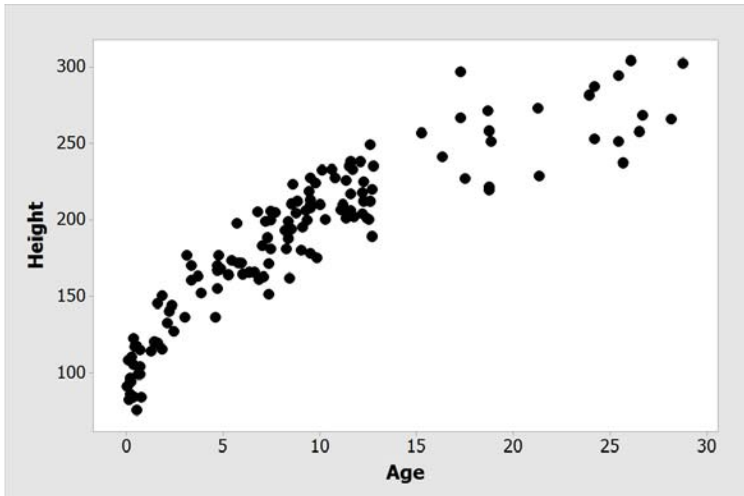
Both the *Age* and *Sept* effects are significant in this model.

g. R^2 is 0.4298, so 42.98% of the variability in $PctDM$ is explained by the regression model from (f) that uses *Age* and *Sept*.

h. The following new residual plot does show improvement over the plot from part (e). Now the S residuals are fairly well balanced between positive and negative; likewise for the N residuals.



3.35 a. The plot follows. There is substantial curvature in the plot. As *Age* increases, *Height* increases. This increase is rapid when *Age* is less than 5 years, but the pattern of increase tapers off at higher ages.

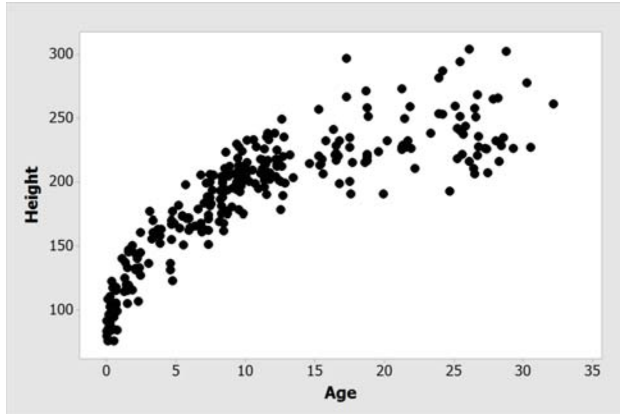


b. The following output gives the fitted model as $\widehat{Height} = 100.2 + 13.383Age - 0.2643Age^2$

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant    100.21     3.16    31.69    0.000
Age         13.383    0.615    21.78    0.000
Age*Age    -0.2643   0.0237   -11.17    0.000

Regression Equation
Height = 100.21 + 13.383Age - 0.2643Age*Age
```

c. Plugging in$Age = 15$ we get $100.20 + 13.383(15) - 0.2643(15^2) = 241.5$ cm.

3.36 a. The plot follows. There is substantial curvature in the plot. As *Age* increases, *Height* increases. This increase is rapid when *Age* is less than 5 years, but the pattern of increase tapers off at higher ages.

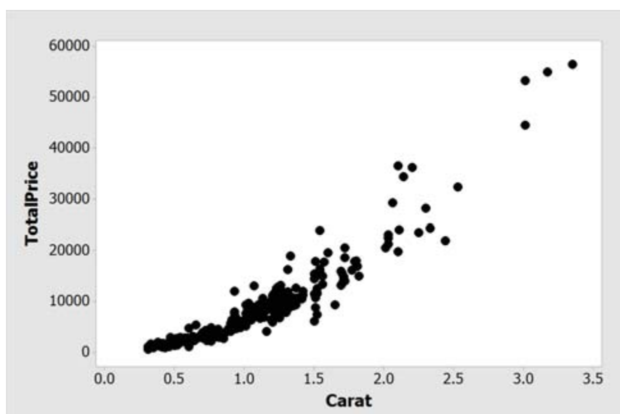


b. The following output gives the fitted model as $\widehat{Height} = 102.5 + 12.566Age - 0.2763Age^2$

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant    102.48     2.55    40.27    0.000
Age         12.566    0.452    27.80    0.000
Age*Age    -0.2763   0.0158   -17.47    0.000

Regression Equation
Height = 102.48 + 12.566Age - 0.2763Age*Age
```

c. Plugging in $Age = 10$ we get $102.5 + 12.566(10) - 0.2763(10^2) = 200.5$ cm.

3.37 a. The following scatterplot shows clear curvature in this relationship.



b. The following output shows that the fitted model is $\widehat{TotalPrice} = -523 + 2386Carat + 4498Carat^2$. Also $R^2 = 0.9257$ and the adjusted $R^2 = 0.9253$.

S	R-sq	R-sq(adj)	R-sq(pred)
2126.76	92.57%	92.53%	92.30%

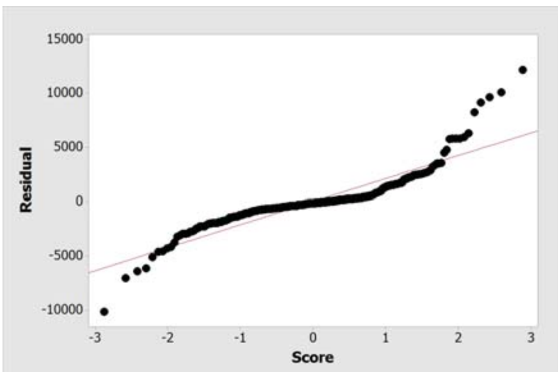
Coefficients

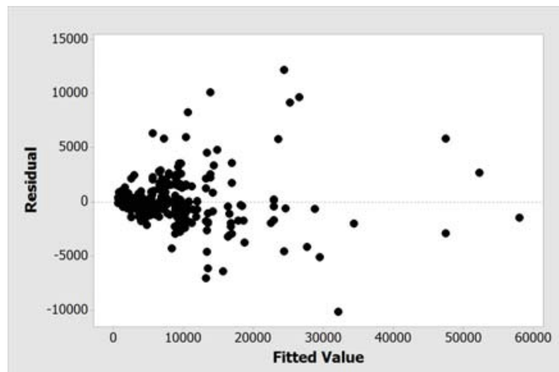
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-523	466	-1.12	0.263	
Carat	2386	753	3.17	0.002	10.69
Carat*Carat	4498	263	17.10	0.000	10.69

Regression Equation

```
TotalPrice = -523 + 2386Carat + 4498Carat*Carat
```

c. The following plots show that the residuals do not appear to come from a normal distribution. Also, they do not exhibit constant variance.





d. The following output shows that the fitted model is $\widehat{TotalPrice} = -723 + 2942Carat + 4078Carat^2 + 88Carat^3$. $R^2 = 0.9257$, adjusted $R^2 = 0.9251$.

S	R-sq	R-sq(adj)	R-sq(pred)
2129.60	92.57%	92.51%	92.16%

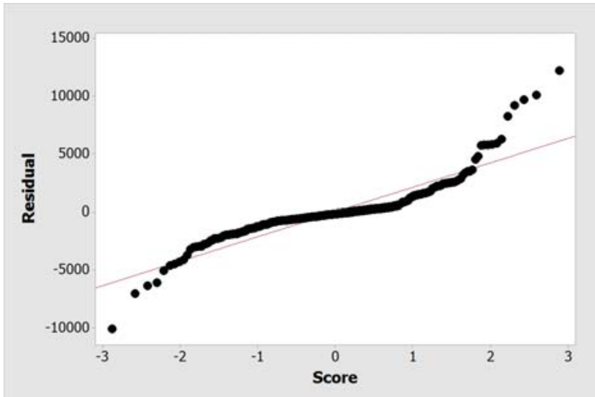
Coefficients

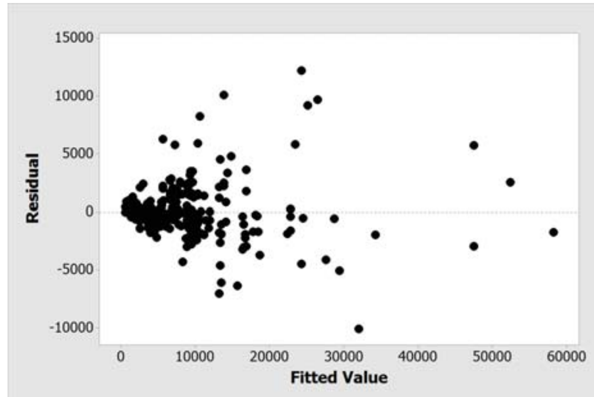
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-723	876	-0.83	0.409	
Carat	2942	2185	1.35	0.179	89.90
Carat*Carat	4078	1574	2.59	0.010	381.61
Carat*Carat*Carat	88	324	0.27	0.787	124.74

Regression Equation

```
TotalPrice = -723 + 2942Carat + 4078Carat*Carat + 88Carat*Carat*Carat
```

e. The plots that follow show that the residuals do not appear to come from a normal distribution. Also, they do not exhibit constant variance.





3.38 a. For the model using *Depth* and $Depth^2$:

```
Predictor     Coef  SE Coef      T      P
Constant    -28407   112212  -0.25  0.800
Depth          766     3353   0.23  0.819
DepthSq      -3.23    24.87  -0.13  0.897

S = 7615.74   R-Sq = 4.7%   R-Sq(adj) = 4.2%
```

According to the individual t-tests, neither of these terms is important (given the other) in this model.

b. For the models using *Carat* and *Depth*:

```
Predictor      Coef  SE Coef      T      P
Constant       1059     1918   0.55  0.581
Carat       15087.0    321.0  47.01  0.000
Depth       -134.94    30.92  -4.36  0.000

S = 2809.30   R-Sq = 87.0%   R-Sq(adj) = 87.0%
```

According to the individual t-tests, each of these terms is important (given the other) in this model.

c. For the model using *Carat*, *Depth*, and $Carat \cdot Depth$:

Predictor	Coef	SE Coef	T	P
Constant	31171	4220	7.39	0.000
Carat	-11828	3436	-3.44	0.001
Depth	-598.18	65.47	-9.14	0.000
Carat*Depth	408.45	51.96	7.86	0.000

S = 2591.98   R-Sq = 89.0%   R-Sq(adj) = 88.9%

According to the individual t-tests, each of these terms is important (given the others) in this model.

d. For a complete second order model using *Carat* and *Depth*:

Predictor	Coef	SE Coef	T	P
Constant	24339	30298	0.80	0.422
Carat	7574	3041	2.49	0.013
Depth	-728.7	904.4	-0.81	0.421
CaratSq	4761.6	330.2	14.42	0.000
DepthSq	5.276	6.727	0.78	0.433
Carat*Depth	-83.89	53.53	-1.57	0.118

S = 2053.42   R-Sq = 93.1%   R-Sq(adj) = 93.0%

According to the t-tests, only the terms involving $Carat$ and $Carat^2$ are important in this model.

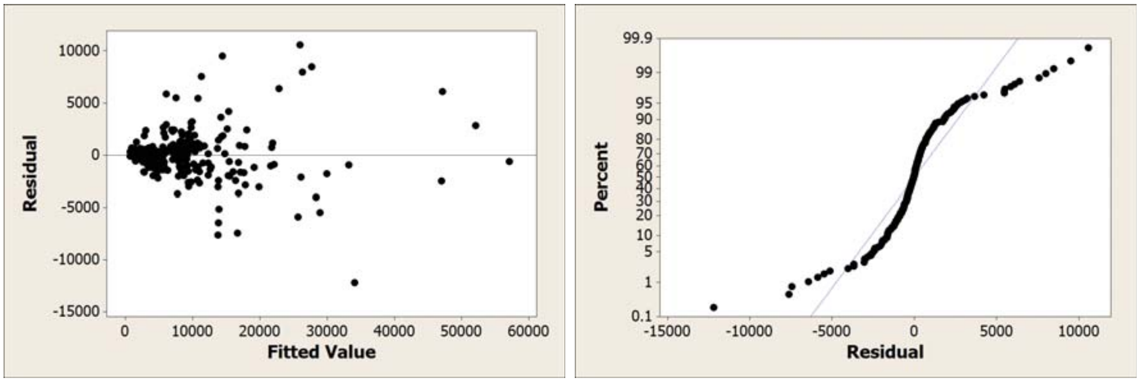
Here is output for the quadratic model using $Carat$:

Predictor	Coef	SE Coef	T	P
Constant	-522.7	466.3	-1.12	0.263
Carat	2386.0	752.5	3.17	0.002
CaratSq	4498.2	263.0	17.10	0.000

S = 2126.76   R-Sq = 92.6%   R-Sq(adj) = 92.5%

If we use adjusted R^2 as a criteria, the best model among these would be the complete second-order model (d). However, since each of the three coefficients in that model involving *Depth* have P-values over 0.10, we might also consider just using the quadratic model based on $Carat$ (which has nearly as large an adjusted R^2 and small P-values for the coefficients of each predictor). If we were not limited to just these models, we might also try a model that adds just *Depth* to the quadratic model based on $Carat$ to see if it would be significant without the other two strongly related predictors ($Depth^2$ and $Carat \cdot Depth$).

3.39 a. Using the complete second-order model with $Carat$ and *Depth* to predict $TotalPrice$ of diamonds, we obtain the following plots for the residuals versus fits and a normal probability



plot of the residuals. The residuals versus fits plot shows that the variability of the residuals increases as the fitted values increase, indicating a problem with the equal variance condition. The normal probability plot shows a consistent curve away from a straight line, indicating a problem with the normality condition.

b. Computer output for fitting the complete second-order model to predict $ln(TotalPrice)$ follows.

Predictor	Coef	SE Coef	T	P
Constant	13.505	3.402	3.97	0.000
Carat	2.5863	0.3414	7.57	0.000
Depth	-0.2028	0.1016	-2.00	0.047
CaratSq	-0.57141	0.03708	-15.41	0.000
DepthSq	0.0013384	0.0007553	1.77	0.077
Carat*Depth	0.009594	0.006011	1.60	0.111

S = 0.230571   R-Sq = 93.0%   R-Sq(adj) = 92.9%

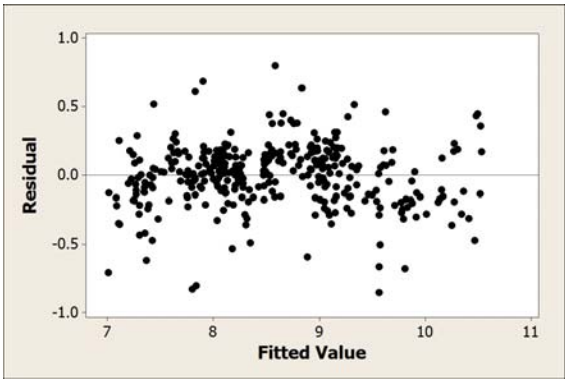
This set of predictors is still a reasonable choice. All but the interaction term is significant at a 10% level and no smaller model using these predictors has a larger adjusted R^2 than 92.9%. As an alternate model, we might also drop the second-order terms involving $Depth$ (that is $Depth^2$ and $Carat \cdot Depth$) to obtain the fitted model shown as follows.

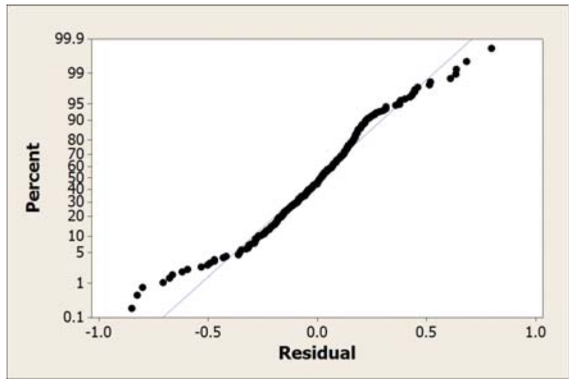
Predictor	Coef	SE Coef	T	P
Constant	6.8085	0.1621	42.01	0.000
Carat	3.11534	0.08304	37.52	0.000
Depth	-0.011267	0.002557	-4.41	0.000
CaratSq	-0.53400	0.02873	-18.59	0.000

S = 0.231950   R-Sq = 92.9%   R-Sq(adj) = 92.8%

This smaller model has very strong evidence for the importance of each of its predictors and has only a slightly smaller adjusted R^2 than the complete second-order model.

c. The plots that follow are for the residuals of the complete second-order model to predict $ln(TotalPrice)$. They are very similar to the corresponding plots for the model using just $Carat$, $Carat^2$, and $Depth$.





The residuals versus fits plot shows that the problem with increasing variance has been eliminated and a condition of equal variance of the residuals is quite reasonable. There is still a small amount of wiggle in the very tails for the normal probability plot, but overall the normality condition is now much more appropriate than in the model without the log transformation.

3.40 a. Here is some computer output for fitting $TotalPrice = \beta_0 + \beta_1 Carat + \beta_2 Carat^2 + \epsilon$.

```
Predictor    Coef  SE Coef      T      P
Constant   -522.7    466.3  -1.12  0.263
Carat      2386.0    752.5   3.17  0.002
CaratSq    4498.2    263.0  17.10  0.000
```

The predicted average price when $Carat = 0.5$ is

$$\widehat{TotalPrice} = -522.7 + 2386.0(0.5) + 4498.2(0.5^2) = 1794.85$$

or about \$1795.

b. Here is some computer output for confidence and prediction intervals from the quadratic model for $TotalPrice$ when $Carat = 0.5$.

```
Predicted Values for New Observations
NewObs   Fit  SE Fit      95% CI          95% PI
     1  1795     188  (1424, 2165)  (-2404, 5994)
```

```
Values of Predictors for New Observations
NewObs  Carat  CaratSq
     1  0.500    0.250
```

Based on this model, we are 95% confident that the average price of all 0.5-carat diamonds is between \$1424 and \$2165.

c. The prediction interval from the output above is $(-2404, 5994)$. Of course, a negatively priced diamond is not feasible, so we can adjust the lower bound to zero. We expect that 95% of all 0.5-carat diamonds will cost between \$0 and \$5994.

d. Here is some computer output for confidence and prediction intervals from the complete second-order model to predict $ln(TotalPrice)$ when $Carat = 0.5$ and $Depth = 62$.

```
Predicted Values for New Observations
NewObs     Fit  SE Fit       95% CI             95% PI
     1  7.5260  0.0210  (7.4847, 7.5673)  (7.0706, 7.9814)

Values of Predictors for New Observations
NewObs  Carat  Depth  CaratSq  DepthSq  Carat*Depth
     1  0.500   62.0    0.250     3844         31.0
```

The predicted $logPrice$ is 7.5260, so the predicted $TotalPrice$ is $e^{7.5260} = \$1856$.
We exponentiate the 95% confidence interval for average $ln(TotalPrice)$, $(7.4847, 7.5673)$, to obtain a confidence interval for average $TotalPrice$.

$$(e^{7.4847}, e^{7.5673}) = (1781, 1934)$$

Thus we are 95% sure that the average price of all 0.5-carat diamonds with a depth of 62% is between \$1781 and \$1934.
We exponentiate the 95% prediction interval for $ln(TotalPrice)$, $(7.0706, 7.9814)$, to obtain a prediction interval for $TotalPrice$.

$$(e^{7.0706}, e^{7.9814}) = (1177, 2926)$$

We expect that 95% of all 0.5 carat diamonds with 62% depth will cost between \$1177 and \$2926.

3.41 Here is some output for fitting the model $ProteinProp = \beta_0 + \beta_1 Calcium + \beta_2 Calcium^2 + \epsilon$.

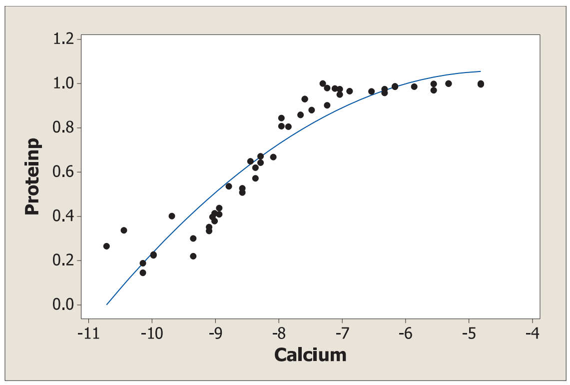
```
The regression equation is ProteinProp = 0.480 - 0.253 Calcium - 0.0278 Calciumsq

Predictor       Coef   SE Coef      T      P
Constant      0.4799    0.3179   1.51  0.138
```

```
Calcium      -0.25319   0.08410  -3.01  0.004
Calciumsq  -0.027788  0.005425  -5.12  0.000

S = 0.0973831   R-Sq = 89.4%   R-Sq(adj) = 89.0%
```

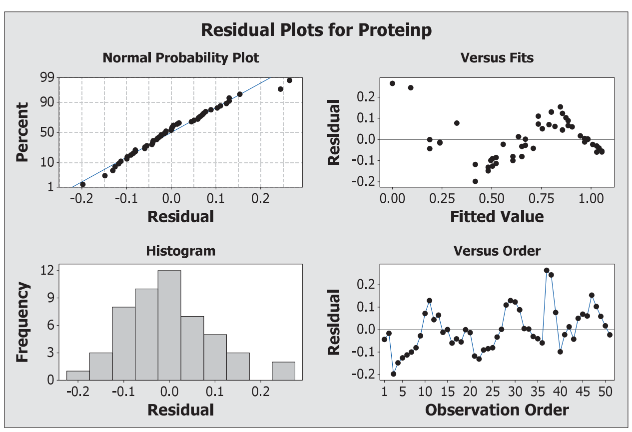
a. From the output, we see that the fitted quadratic regression model is

$$\widehat{ProteinProp} = 0.48 - 0.2532Calcium - 0.0278Calcium^2$$

b. Here is a scatterplot of *ProteinProp* versus *Calcium* with the quadratic fit, which captures some of the curvature in the relationship.



c. Here are some plots of the residuals for the quadratic model. The normal probability plot shows a linear trend, so the normality condition is fine. The histogram of the residuals is centered at zero but shows one large value. The plot of the residuals versus the fitted values shows a nonrandom pattern (decreasing, increasing, then decreasing again), which indicates that a higher-order term might be useful.



d. To assess the importance of the quadratic term we test $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 = 0$. The t-statistic for this coefficient in the output is $t = -5.12$, which gives a P-value that is very close to zero. This indicates that the quadratic term ($Calcium^2$) is useful in this model for the proportion of protein bound to calcium.

e. In the output, we see `R-Sq = 89.4%`, which indicates that 89.4% of the variation in $ProteinProp$ for this sample is explained by the quadratic model based on $Calcium$.

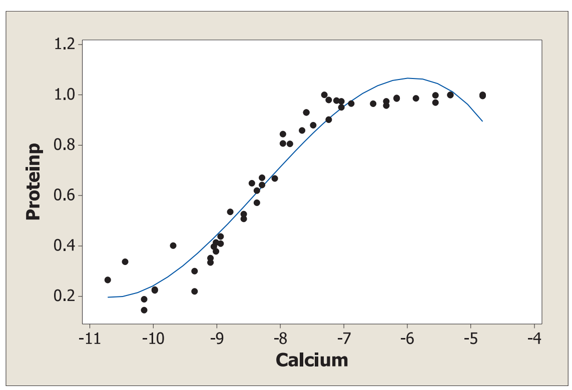
3.42 Here is some output for fitting the cubic model $ProteinProp = \beta_0+\beta_1 Calcium+\beta_2 Calcium^2+ \beta_3 Calcium^3 + \epsilon$.

```
The regression equation is
ProteinProp = - 6.52 - 3.14 Calcium - 0.411 Calciumsq - 0.0165 Calcium3

Predictor        Coef   SE Coef       T      P
Constant       -6.524     1.089   -5.99  0.000
Calcium       -3.1384    0.4426   -7.09  0.000
Calciumsq    -0.41134   0.05840   -7.04  0.000
Calcium3    -0.016515  0.002509   -6.58  0.000

S = 0.0709873   R-Sq = 94.5%   R-Sq(adj) = 94.1%
```

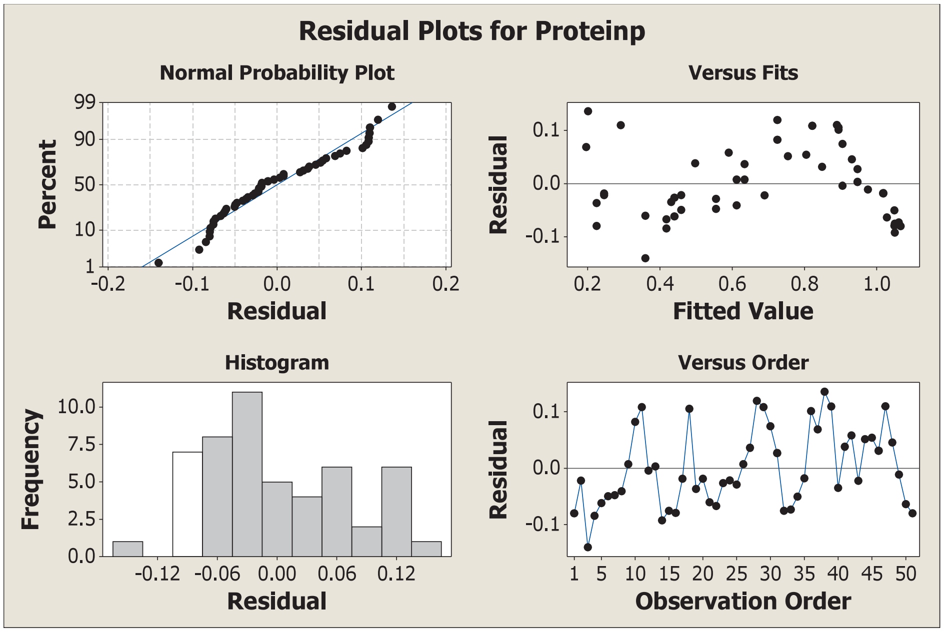
a. From the output, we see that the fitted cubic regression model is

$$\widehat{ProteinProp} = -6.52 - 3.138 Calcium - 0.4113 Calcium^2 - 0.0165 Calcium^3$$

b. Here is a scatterplot of $ProteinProp$ versus $Calcium$ with the cubic fit, which captures some of the curvature in the relationship. However, we might wonder whether the trend should really be starting to decrease for large values of $Calcium$ at the right of the graph.



c. Here are some plots of the residuals for the cubic model. The normal probability plot shows some slight curvature in the tails. The plot of the residuals versus fitted values still contains patterns, but not as strong as the quadratic model.



d. To assess the importance of the cubic term, we test $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 = 0$. The t-statistic for this coefficient in the output is $t = -6.58$, which gives a P-value that is very close to zero. This indicates that the cubic term ($Calcium^3$) is useful in this model for the proportion of protein bound to calcium.

e. In the output, we see `R-Sq = 94.5%`, which indicates that 94.5% of the variation in $ProteinProp$ for this sample is explained by the cubic model based on $Calcium$. This is a fairly good improvement over the quadratic model of the previous exercise ($R^2 = 89.4\%$).

3.43 a. Following is some output for fitting the model $Margin = \beta_0 + \beta_1 Days + \beta_2 Days^2 + \epsilon$.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.477958   1.095676   4.087 8.89e-05 ***
Days        -0.604426   0.138598  -4.361 3.18e-05 ***
I(Days^2)    0.021129   0.003776   5.595 1.97e-07 ***

Residual standard error: 3.014 on 99 degrees of freedom
Multiple R-squared: 0.3495,     Adjusted R-squared: 0.3363
F-statistic: 26.59 on 2 and 99 DF,  p-value: 5.711e-10
```

The prediction equation is $\widehat{Margin} = 4.478 - 0.60443 Days + 0.021129 Days^2$. $R^2 = 34.95\%$ and $SSE = 99(3.014^2) = 899$.

b. Here is some output for fitting the interaction model $Margin = \beta_0 + \beta_1 Days + \beta_2 Charlie + \beta_2 Days \cdot Charlie + \epsilon$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.5656     1.0885   5.113 1.57e-06 ***
Days          -0.5984     0.1206  -4.960 2.96e-06 ***
Charlie      -10.1117     1.9251  -5.253 8.74e-07 ***
Days:Charlie   0.9207     0.1364   6.752 1.04e-09 ***

Residual standard error: 2.868 on 98 degrees of freedom
Multiple R-squared: 0.417,      Adjusted R-squared: 0.3992
F-statistic: 23.37 on 3 and 98 DF,  p-value: 1.712e-11
```

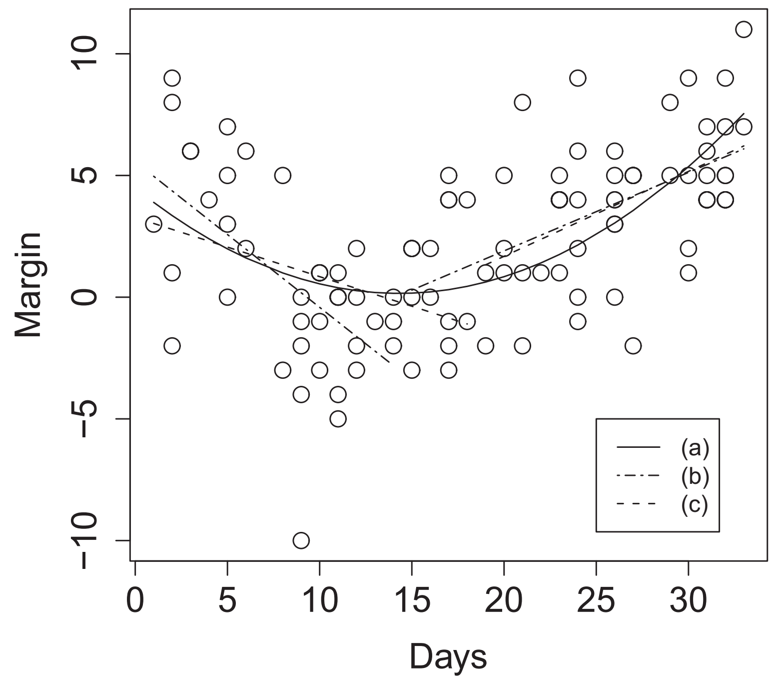
The prediction equation is $\widehat{Margin} = 5.566 - 0.5984 Days - 10.112 Charlie + 0.9207 Days \cdot Charlie$. $R^2 = 41.7\%$ and $SSE = 98(2.868^2) = 806.1$.

c. Here is some output for fitting the interaction model $Margin = \beta_0 + \beta_1 Days + \beta_2 Meltdown + \beta_2 Days \cdot Meltdown + \epsilon$.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.2725     0.9933   3.295  0.00137 **
Days           -0.2429     0.0863  -2.815  0.00590 **
Meltdown       -8.5701     2.9390  -2.916  0.00439 **
Days:Meltdown   0.5917     0.1343   4.406  2.7e-05 ***

Residual standard error: 3.088 on 98 degrees of freedom
Multiple R-squared: 0.3239,     Adjusted R-squared: 0.3032
F-statistic: 15.65 on 3 and 98 DF,  p-value: 2.162e-08
```

The prediction equation is $\widehat{Margin} = 3.273 - 0.2429 Days - 8.57 Meltdown + 0.5917 Days \cdot Meltdown$. $R^2 = 32.4\%$ and $SSE = 98(3.088^2) = 934.5$.

d. The three fitted models for parts (a), (b), and (c) are shown on the following scatterplot of $Margins$ versus $Days$.



The model for part (b), involving the $Charlie$ indicator, has the highest R^2 (41.7%), adjusted R^2 (39.9%), smallest SSE (806), and all significants terms, so it would be the best choice among these models for explaining the polling $Margin$ between Obama and McCain.

3.44 a. Here are correlations between each of the variables.

	SqrtMDs	Hospitals
Hospitals	0.923	
Beds	0.949	0.909

Beds had a stronger correlation with $SqrtMDs$ ($r = 0.949$) than does $Hospitals$ ($r = 0.923$), so it would be a stronger predictor by itself.

b. We square each correlation with $SqrtMDs$ to find the portion of variability that each predictor explains. The portion of variability in $SqrtMDs$ that is explained by $Hospitals$ is $R^2 = 0.923^2 = 0.852$, or 85.2%. The portion of variability in $SqrtMDs$ that is explained by $Beds$ is $R^2 = 0.949^2 = 0.901$, or 90.1%.

c. Here is some output for fitting the model $SqrtMDs = \beta_0 + \beta_1 NumHospitals + \beta_2 NumBeds + \epsilon$.

```
Coefficients
Term        Coef    SE Coef T-Value P-Value
Constant    3.58    1.74    2.06    0.045
Hospitals   2.579   0.709   3.64    0.001
Beds        0.01231 0.00185 6.67    0.000

Model Summary
S         R-sq  R-sq(adj)    R-sq(pred)
6.36171 92.17%     91.86%        90.41%
```

The amount of variability in $SqrtMDs$ that is explained by this two predictor model is $R^2 = 91.86\%$.

d. When testing either the individual correlations with $SqrtMDs$ or the slopes in separate single predictor regression models, the P-values for $Hospitals$ and $Beds$ are both very close to zero. Both predictors have strong relationships with $SqrtMDs$ on their own.

e. From the multiple regression output in part (c), the P-value for testing the importance of $Beds$ is very small (0.000), so that is an important predictor of $SqrtMDs$ in this model. The P-value for testing $Hospitals$ is also very small (0.001), so $Hospitals$ is useful for helping predict $SqrtMDs$ if $Beds$ is also in the model.

3.45 a. Here is computer output for fitting $NetSupport = \beta_0 + \beta_1 Months + \beta_2 Late + \beta_3 Months \cdot Late + \epsilon$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -66.62827    4.94880 -13.464  5.2e-09 ***
Months        0.21037    0.07392   2.846   0.0138 *
Late         13.11464   21.57377   0.608   0.5537
Months:Late   0.17398    0.12761   1.363   0.1959
```

The fitted prediction equation is $Net\widehat{Support} = -66.628 + 0.2104 Months + 13.115 Late + 0.1740 Months \cdot Late$.

b. We test $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 = 0$. The test statistic from the computer output in (a) is 1.363 and the P-value is 0.1959, which is not small. We fail to reject the hypothesis that parallel lines are adequate and can drop the interaction term from this model to describe $NetSupport$.

Note: Even though the coefficient of $Late$ in the computer output also has a large P-value, we should not automatically assume that predictor is not important for modeling $NetSupport$. As is shown in the previous exercise, the $Late$ term is valuable on its own when the interaction term is not also in the model.

c. We use a nested F-test for the hypotheses $H_0 : \beta_2 = \beta_3 = 0$ versus $H_a : \beta_2 = 0$ or $\beta_3 = 0$. The following computer output shows the sum of squared errors (labeled as RSS) for the full interaction model (Model 2 with 13 error d.f.) versus the reduced model with $Months$ alone (Model 1 with 15 error d.f).

```
Model 1: NetSupport ~ Months
Model 2: NetSupport ~ Months * Late
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     15 1744.73
2     13  681.56  2    1063.2 10.139 0.002221 **
```

The test statistic is

$$F = \frac{(1744.73 - 681.56)/2}{681.56/13} = 10.139$$

We compare this to an F-distribution with 2 and 13 degrees of freedom to get a P-value of 0.002221. Since this is a very small P-value, we have strong evidence that at least one of the terms involving the *Late* indicator substantially improves the fit for predicting $NetSupport$.

3.46 The full model is $TotalPrice = \beta_0 + \beta_1 Carat + \beta_2 Depth + \beta_3 Carat^2 + \beta_4 Depth^2 + \beta_5 Carat \cdot Depth + \epsilon$. The ANOVA table for fitting the model follows.

Source	DF	SS	MS	F	P
Regression	5	19735107439	3947021488	936.08	0.000
Residual Error	345	1454702094	4216528		
Total	350	21189809533			

To test $H_0 : \beta_2 = \beta_4 = \beta_5 = 0$ versus H_a : At least one of these $\beta_i = 0$, we fit the reduced model $TotalPrice = \beta_0 + \beta_1 Carat + \beta_3 Carat^2 + \epsilon$.

Source	DF	SS	MS	F	P
Regression	2	19615765122	9807882561	2168.39	0.000
Residual Error	348	1574044410	4523116		
Total	350	21189809533			

The drop in $SSModel$ for eliminating these three predictors is 19,735,107,439 − 19,615,765,122 = 119,342,317. The F-ratio is

$$F = \frac{119{,}342{,}317/3}{1{,}454{,}702{,}094/345} = 9.43$$

We compare this to an F-distribution with 3 and 345 degrees of freedom to find a P-value that is very close to zero. This gives strong evidence that at least one of the terms involving *Depth* should be included in the model and that dropping all three would significantly impair the effectiveness for predicting $TotalPrice$.

Note: If we had coded diamond prices in \$100s or \$1000s, the sums of squares in the ANOVA table would be more manageable without changing the effectiveness of the models.

3.47 a. Here is some computer output for fitting the full model, $FatalityRate = \beta_0 + \beta_1 Year + \beta_2 StateControl + \beta_3 Year \cdot StateControl + \epsilon$.

```
Predictor                  Coef   SE Coef       T      P
Constant                 216.23     13.03   16.59  0.000
Year                  -0.107619  0.006548  -16.44  0.000
StateControl            -161.38     14.47  -11.15  0.000
Year*StateControl      0.080971  0.007264   11.15  0.000

S = 0.0424331   R-Sq = 98.3%   R-Sq(adj) = 98.0%

Analysis of Variance
Source          DF       SS        MS       F      P
Regression       3  1.77691   0.59230  328.95  0.000
Residual Error  17  0.03061   0.00180
Total           20  1.80752
```

To test $H_0 : \beta_2 = \beta_3 = 0$ versus $H_a : \beta_2 = 0$ or $\beta_3 = 0$ we consider the reduced model $FatalityRate = \beta_0 + \beta_1 Year + \epsilon$. Here is an ANOVA table for that model.

```
Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  1.5503  1.5503  114.49  0.000
Residual Error  19  0.2573  0.0135
Total           20  1.8075
```

Comparing the $SSModel$ amounts for these two models, we compute a test statistic

$$F = \frac{(1.77691 - 1.5503)/2}{0.03061/17} = \frac{0.1133}{0.0018} = 62.9$$

We use an F-distribution with 2 and 17 degrees of freedom to find the P-value as the area beyond 62.9 to be around 10^{-8}. This extremely small P-value says that we have strong evidence for a difference in slope, intercept, or both in the relationship between $FatalityRate$ and $Year$ before and after states assumed control of speed limits.

b. From the computer output in part (a) for the full interaction model, we see the test statistic for a t-test of $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 = 0$ is $t = 11.15$ and the P-value is 0.000. This gives strong evidence that the slopes of the two lines differ.

To approach this question with a nested F-test, we consider the reduced model without the interaction term, $FatalityRate = \beta_0 + \beta_1 Year + \beta_2 StateControl + \epsilon$. Here is an ANOVA table for that model.

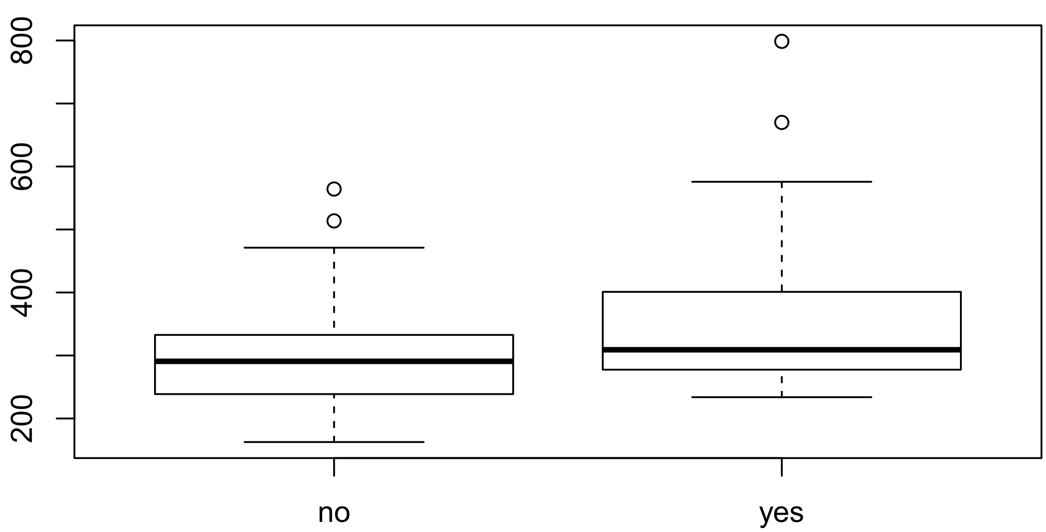
```
Analysis of Variance
Source          DF       SS       MS      F      P
Regression       2  1.55318  0.77659  54.96  0.000
Residual Error  18  0.25434  0.01413
Total           20  1.80752
```

Comparing the $SSModel$ amounts for this model to the full model, we compute a test statistic

$$F = \frac{(1.77691 - 1.55318)/1}{0.03061/17} = \frac{0.22373}{0.0018} = 124.3$$

We use an F-distribution with 1 and 17 degrees of freedom to find the P-value as the area beyond 124.3 to be around 3×10^{-9}. Although the t-test in part (a) does not show this many decimal places, the P-value for the t-test is the same as for the nested F-test. In fact, the F-statistic, 124.3, is the square of the t-statistic, $11.15^2 = 124.3$.

3.48 a. The boxplots show that houses with some garage space tend to have higher selling prices. The plots overlap considerably and both distributions are skewed slightly toward the higher prices. The t-test results (following) show the mean differential is about \$53,000 and the difference is deemed statistically significant (P-value 0.007896).



```
Welch Two Sample t-test

data:  adj2007 by garagegroup
t = -2.7145, df = 94.013, p-value = 0.007896
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -93.36936 -14.48237
sample estimates:
 mean in group no mean in group yes
         300.0728          353.9987
```

b. The simple linear regression output shows a statistically significant negative relationship between *price* and *distance*. Each mile farther from a trail corresponds to about a \$54,000 decrease in selling price. (coefficient $= -54.427$.)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  388.204     14.052  27.626  < 2e-16 ***
distance     -54.427      9.659  -5.635 1.56e-07 ***
---
Residual standard error: 92.13 on 102 degrees of freedom
Multiple R-squared:  0.2374,Adjusted R-squared:  0.2299
F-statistic: 31.75 on 1 and 102 DF,  p-value: 1.562e-07
```

c. Fitting the two-predictor model leads to the following results. While both *distance* and *garagegroup* are significant predictors, the estimated rate of change between price and distance changes by only about \$3,000 when controlling for the presence of garage space. The two-predictor model increases the R^2 from 23.74% to 26.93%, a modest improvement.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     365.103     17.661  20.673   <2e-16 ***
distance        -51.025      9.638  -5.294    7e-07 ***
garagegroupyes   37.892     18.032   2.101   0.0381 *
---
Residual standard error: 90.62 on 101 degrees of freedom
Multiple R-squared:  0.2693,Adjusted R-squared:  0.2549
F-statistic: 18.62 on 2 and 101 DF,  p-value: 1.311e-07
```

d. The interaction term has coefficient -9.878 (see summary table that follows), and this value represents the estimated difference in rates of change in price relative to distance for homes with and without garage space. The rate of change is estimated to be -46.302 for garage-less homes, and $-46.302 - 9.878 = -56.18$ for homes with garages. But the P-value (0.611) is too large to deem this difference statistically significant.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              359.083     21.295  16.862  < 2e-16 ***
distance                 -46.302     13.391  -3.458 0.000802 ***
garagegroupyes            48.862     28.108   1.738 0.085222 .
distance:garagegroupyes   -9.878     19.366  -0.510 0.611125
---
Residual standard error: 90.96 on 100 degrees of freedom
Multiple R-squared:  0.2712,Adjusted R-squared:  0.2494
F-statistic: 12.41 on 3 and 100 DF,  p-value: 5.785e-07
```

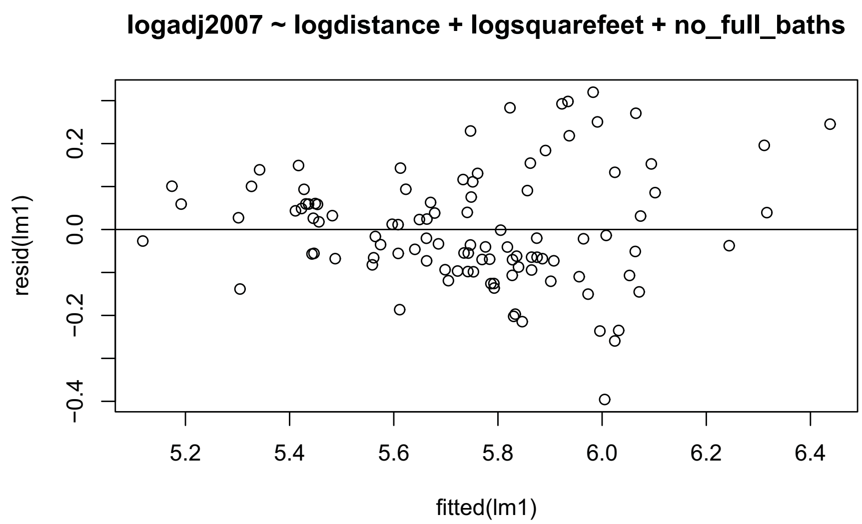
e. Using the anova function in R, we can perform a nested F-test and we discover that the additional information of garage space does not add significantly to the estimation of the price-versus-distance relationship. The P-value is 0.1034, so bordering on significance, but not so very.

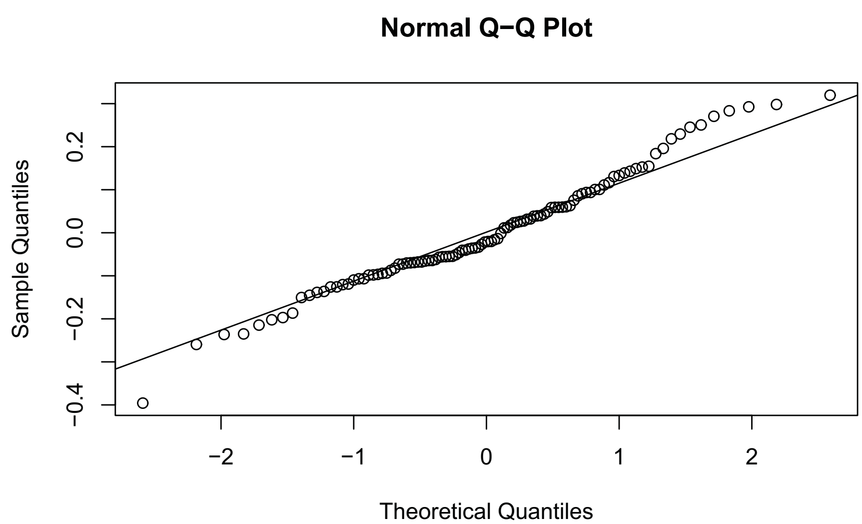
```
> anova(lm1, lm3)
Analysis of Variance Table

Model 1: adj2007 ~ distance
Model 2: adj2007 ~ distance + garagegroup + distance:garagegroup
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    102 865718
2    100 827301  2     38417 2.3218 0.1034
```

3.49 a. The following summary table suggests that all three predictors are significant in the model. The only variable that is a close call is *no_full_baths* with a P-value of 0.0255; still significant. As distance from trails increases, price of home goes down (-0.04883). Homes with more square footage sell for more (0.59328). Homes with more full baths sell for more (0.05667). These are not easy to interpret since things are on a logged scale.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.41777    0.03368 160.870  < 2e-16 ***
logdistance   -0.04883    0.01245  -3.922 0.000161 ***
logsquarefeet  0.59328    0.04567  12.991  < 2e-16 ***
no_full_baths  0.05667    0.02500   2.267 0.025548 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1344 on 100 degrees of freedom
Multiple R-squared:  0.7834,  Adjusted R-squared:  0.7769
F-statistic: 120.6 on 3 and 100 DF,  p-value: < 2.2e-16
```

b. The residuals-versus-fits plots and the normal plot of residuals suggest a good adherence to model conditions.





c. The summary table of the complicated model is not easily interpreted term by term. Our main goal is to assess whether interactions are necessary. Note first that the R^2 values went from 78.34% for the simpler model to 80.07% for the interaction model. That is a modest gain.

```
Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                            5.545207   0.058168  95.331  < 2e-16 ***
logdistance                           -0.040887   0.045200  -0.905 0.367955
logsquarefeet                          0.355179   0.102008   3.482 0.000751 ***
no_full_baths                         -0.048636   0.047595  -1.022 0.309413
logdistance:logsquarefeet             -0.024984   0.083870  -0.298 0.766428
logdistance:no_full_baths             -0.009463   0.034035  -0.278 0.781580
logsquarefeet:no_full_baths            0.172022   0.064910   2.650 0.009410 **
logdistance:logsquarefeet:no_full_baths 0.018293  0.054586   0.335 0.738263
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1316 on 96 degrees of freedom
Multiple R-squared:  0.8007,Adjusted R-squared:  0.7861
F-statistic: 55.09 on 7 and 96 DF,  p-value: < 2.2e-16
```

d. We use a nested F-test to decide if the complexity is needed. Using the following R code, we get the necessary information:

```
anova(lm1, lm2 )
```

we obtain these results:

```
Model 1: logadj2007 ~ logdistance + logsquarefeet + no_full_baths
Model 2: logadj2007 ~ logdistance + logsquarefeet + no_full_baths +
    logdistance:logsquarefeet + logdistance:no_full_baths +
    logsquarefeet:no_full_baths + logdistance^2 + logsquarefeet^2 +
    no_full_baths^2 + logdistance:logsquarefeet:no_full_baths
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    100 1.8051
2     96 1.6614  4   0.14373 2.0763 0.08986 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The P-value is 0.09, suggesting the gain in complexity is probably not worth it.

3.50 a. The following output gives a fitted model of $\widehat{Height} = 100.21 + 13.383Age - 0.2643Age^2$.

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant    100.21     3.16    31.69    0.000
Age         13.383    0.615    21.78    0.000  8.58
```

```
Age*Age   -0.2643   0.0237   -11.17    0.000  8.58

Regression Equation
Height = 100.21 + 13.383Age - 0.2643Age*Age
```

b. The following output gives a fitted *Firstborn* coefficient of -11.68, a t-statistic of -3.0, and a P-value of 0.003, so yes, the effect is statistically significant. Controlling for the quadratic relationship between age and height, being firstborn decreases height by 11.68 cm on average.

```
Coefficients

Term           Coef  SE Coef  T-Value  P-Value   VIF
Constant     102.02     3.13    32.58    0.000
Age          13.339    0.597    22.33    0.000  8.59
Firstborn    -11.68     3.90    -3.00    0.003  1.02
Age*Age     -0.2596   0.0231   -11.26    0.000  8.63


Regression Equation

Height = 102.02 + 13.339Age - 11.68Firstborn - 0.2596Age*Age
```

c. We can fit the model for part (c) with the R command

```
ModelC <- lm(Height~Age*Firstborn+I(Age^2)*Firstborn, data=ElephantsFB).
```

Then the command

```
anova(ModelB, ModelC)
```

gives an F-statistic of 0.29 and a P-value of 0.75. Thus we choose the model from part (b).

3.51 a. The following output gives a fitted model of $\widehat{Height} = 102.48 + 12.566Age - 0.2763Age^2$.

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant    102.48     2.55    40.27    0.000
Age         12.566    0.452    27.80    0.000
Age*Age    -0.2763   0.0158   -17.47    0.000

Regression Equation
Height = 102.48 + 12.566Age - 0.2763Age*Age
```

b. The following output gives a fitted *SexM* coefficient of 13.46, a t-statistic of 6.07, and a P-value that is essentially zero, so yes, the effect is statistically significant. Controlling for the quadratic relationship between age and height, being male increases height by 13.5 cm on average.

```
Coefficients
Term          Coef  SE Coef  T-Value  P-Value
Constant     94.50     2.74    34.53    0.000
Age         12.625    0.426    29.62    0.000
SexM         13.46     2.22     6.07    0.000
Age*Age    -0.2716   0.0149   -18.20    0.000

Regression Equation
Height = 94.50 + 12.625Age + 13.46SexM - 0.2716Age*Age
```

c. We can fit the model for part (c) with the R command

```
ModelCnew <- lm(Height~Sex*Firstborn+I(Sex^2)*Firstborn, data=ElephantsMF).
```

Then the command

```
anova(ModelBnew, ModelCnew)
```

gives an F-statistic of 18 and a P-value that is essentially zero. [Notice that the $Age \cdot Sex$ interaction term is significant.] Thus we choose the model from part (c) and conclude that the quadratic relationship between *Age* and *Height* is different for males than for females. Although the male and female trends have similar curvature, the male curve rises more sharply than does the female curve.

3.52 a. The following output gives the fitted prediction model as $\widehat{MMSE} = -0.59 + 2.32APC - 1.85TypeDLB/AD - 0.97APC \cdot TypeDLB/AD$. Thus when *Type* is DLB, the prediction model is $\widehat{MMSE} = -0.59 + 2.32APC$, and when *Type* is DLB/AD, the prediction model is $\widehat{MMSE} = -1.42 + 1.26APC$.

```
Coefficients
Term               Coef  SE Coef  T-Value  P-Value   VIF
Constant         -0.585    0.793    -0.74    0.466
APC                2.32     1.16     1.99    0.054  7.12
TypeDLB/AD        -1.85     1.15    -1.61    0.116  1.84
APC*TypeDLB/AD    -0.97     1.27    -0.77    0.449  9.13

Regression Equation
MMSE = -0.585 + 2.32APC - 1.85TypeDLB/AD - 0.97APC*TypeDLB/AD
```

b. From the output for part (a), the test statistic is $t = -0.77$ and the P-value is 0.449. The interaction term is not needed.

c. The R command anova(model1,model2) gives the nested F-statistic as 1.34 and the P-value as 0.27. We retain the null hypothesis and conclude that a common regression line is adequate.

3.53 a. Here is some output for fitting the model $GPA = \beta_0 + \beta_1 HSGPA + \beta_2 SATV + \beta_3 HU + \beta_4 White + \epsilon$.

```
Predictor        Coef     SE Coef      T       P
Constant       0.6410      0.2788   2.30   0.022
HSGPA         0.47620     0.07109   6.70   0.000
SATV        0.0007372   0.0003417   2.16   0.032
HU           0.015057    0.003638   4.14   0.000
White         0.21212     0.06862   3.09   0.002
```

The predicted GPA when $HSGPA = 3.20$, $SATV = 600$, $HU = 10$, and $White = 1$ is

$$\widehat{GPA} = 0.641 + 0.4762(3.20) + 0.000737(600) + 0.01506(10) + 0.212(1) = 2.97$$

b. Here is some additional computer output to give confidence and prediction intervals for these predictors.

```
Predicted Values for New Observations
NewObs     Fit  SE Fit        95% CI              95% PI
     1  2.9698  0.0361  (2.8987, 3.0409)  (2.2127, 3.7269)

Values of Predictors for New Observations
NewObs  HSGPA  SATV    HU  White
     1   3.20   600  10.0   1.00
```

For predicting the GPA of an individual student, we use the prediction interval, thus we are 95% sure that a student with these characteristics will have a GPA between 2.213 and 3.727.

c. Adding SS to the multiple regression model, we obtain the following output for confidence and prediction intervals (when $SS = 10$ and the other characteristics are unchanged).

```
Predicted Values for New Observations
NewObs     Fit  SE Fit        95% CI              95% PI
     1  2.9851  0.0376  (2.9111, 3.0592)  (2.2295, 3.7407)

Values of Predictors for New Observations
NewObs  HSGPA  SATV    HU  White    SS
     1   3.20   600  10.0   1.00  10.0
```

Now the predicted GPA for a student with these characteristics is 2.985 and we are 95% sure that such a student will have a GPA between 2.230 and 3.741.

3.54 a. $Y = 0.5X_1 + 5 = 0.5(2X_2 - 4) + 5 = X_2 + 3$, The association between X_2 and Y in this equation is positive.

b. Adding the equations gives

$$Y + X_1 = (0.5X_1 + 5) + (2X_2 - 4) = 0.5X_1 + 2X_2 + 1$$

Simplifying, we get $Y = -0.5X_1 + 2X_2 + 1$. The coefficient of X_2 is still positive, but the coefficient of X_1 in the new equation switches to negative.

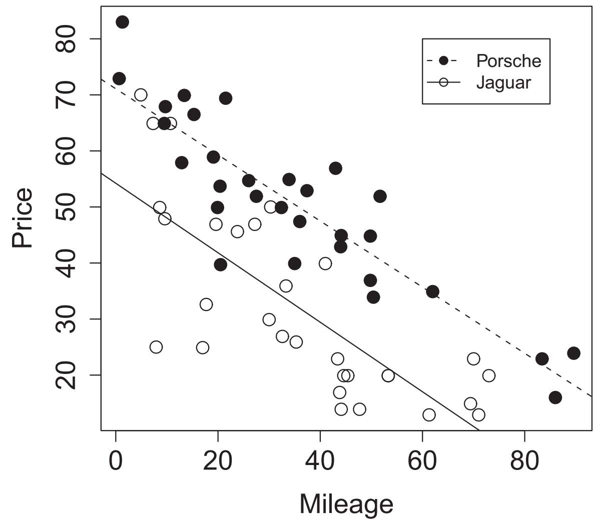
3.55 To compare lines for the two car models, we consider a multiple regression model, $Price = \beta_0 + \beta_1 Mileage + \beta_2 IPorsche + \beta_3 Mileage \cdot IPorsche + \epsilon$. Some output from fitting this model to the data in **PorscheJaguar** is given here.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     54.22746    3.41097  15.898  < 2e-16 ***
Mileage         -0.62030    0.08254  -7.515 4.88e-10 ***
Porsche         16.86299    4.58044   3.682 0.000523 ***
Mileage:Porsche  0.03090    0.11024   0.280 0.780302
```

Based on the P-values for the individual terms in this model, there would appear to be a significant difference in the intercepts P-value $= 0.000523$ for testing $H_0 : \beta_2 = 0$), but not the slopes P-value $= 0.780303$ for testing $H_0 : \beta_3 = 0$).

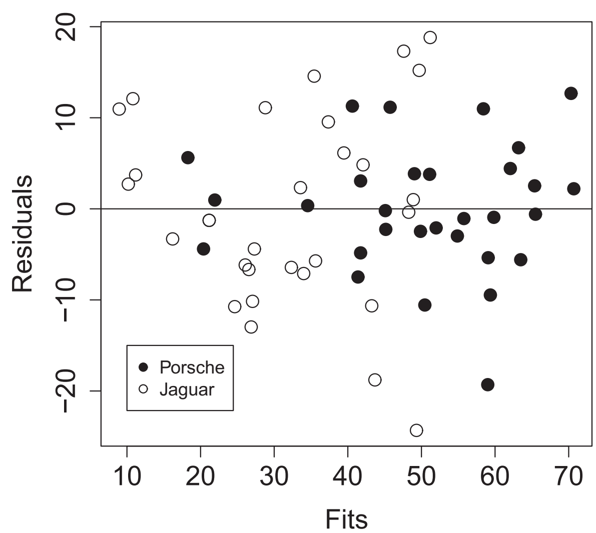
From the fitted model, we can determine least squares lines for each car model.
Jaguar: $\widehat{Price} = 54.23 - 0.62 Mileage$
Porsche: $\widehat{Price} = (54.23 + 16.86) + (-0.62 + 0.03) Mileage = 71.09 - 0.59 Mileage$

The following scatterplot shows these two lines, with the open dots corresponding to Jaguars in the sample and the filled dots corresponding to Porsches.



This plot reinforces the conclusion that the slopes are similar for the two car models, but the prices for Porsches (given the same mileage) tend to be higher than Jaguars.

The plot of residuals versus fitted values for this model shows a random scatter for both Porsche and Jaguar residuals above and below the zero line. This raises no concerns about the usual regression conditions.



3.56 With more than a dozen potential predictors of $WinPct$ in the **MLBStandings2016** data, there are lots of potential models to consider. If we want the adjusted R^2 to decrease when new predictors are added, we should start with a relatively effective model, then add either some weak predictors or predictors that are strongly correlated with those already in the model. This should cause the degrees of freedom for the error term to decrease with little corresponding decrease in the SSE.

One way to start is to consider correlations of several of the predictors with $WinPct$.

	BattingAvg	Runs	Hits	HR	Doubles	Triples	RBI	SB	ERA	Strikeouts
WinPct	0.343	0.540	0.292	0.364	0.092	-0.266	0.544	-0.254	-0.798	0.556

A model using RBI and ERA to predict $WinPct$ should be fairly effective as the following output shows. Both predictors have small P-values for their individual t-tests and the adjusted R^2 is 79.68%.

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.6039	0.0918	6.58	0.000
ERA	-0.1059	0.0123	-8.61	0.000
Runs	0.000468	0.000094	4.99	0.000

S	R-sq	R-sq(adj)
0.0298	81.08%	79.68%

To get the adjusted R^2 to decrease, we can try adding a weak predictor like *Doubles*—but that improves the adjusted R^2 in this situation. We could also try adding predictors that are strongly related to ones already in the model, such as *RBI* ($r = 0.994$ with *Runs*), *HitsAllowed* ($r = 0.873$ with *ERA*), and *WHIP* ($r = 0.911$ with *ERA*). Adding these three predictors gives the following output.

Coefficients

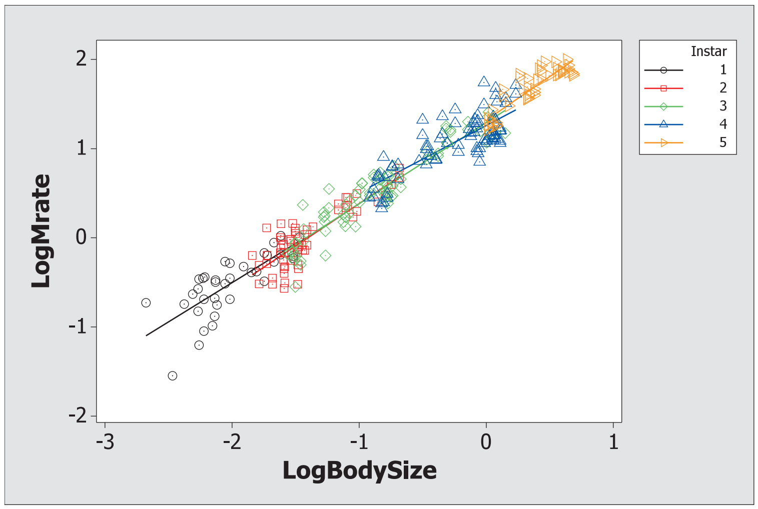
Term	Coef	SE Coef	T-Value	P-Value
Constant	0.750	0.149	5.03	0.000
RBI	0.00069	0.00101	0.68	0.504
ERA	-0.0674	0.0332	-2.03	0.054
Runs	-0.000202	0.000971	-0.21	0.837
HitsAllowed	-0.000121	0.000156	-0.77	0.447
WHIP	-0.095	0.191	-0.50	0.624

S	R-sq	R-sq(adj)
0.03056	82.35%	78.68%

Although the unadjusted R^2 increases slightly when adding these three new predictors (from 81.08% to 82.35%), the adjusted R^2 decreases to 78.68%. Note that none of these new predictors has a small P-value for its individual t-test. In fact, the P-value for *Runs* is now the largest in the model (since *Runs* is highly related to *RBI*), and the P-value for *ERA* is now only marginally significant.

Note: There are other combinations of models that will show a similar decrease in adjusted R^2. For example, using just *Runs* and then adding *RBI*.

3.57 Here is a scatterplot with five separate regression lines, one for each *Instar*. Notice that it is hard to distinguish the separate regression lines, because of the consistent linear pattern.



Only four indicator variables are needed because the fifth one can be determined from knowledge about the other four. That is, these five variables are linearly dependent. If we know the values for four of the indicators, the fifth offers no new information.

Here is some output for fitting a model to predict *LogMrate* with *LogBodySize*, the first four *Instar* indicators, and the product terms for each of these indicators with *LogBodySize*.

```
The regression equation is
LogMrate = 1.32 + 0.980 LogBodySize - 0.066 Instar_1 + 0.016 Instar_2
           - 0.0257 Instar_3 - 0.0607 Instar_4 - 0.101 LBSI1int - 0.029 LBSI2int
           - 0.068 LBSI3int - 0.227 LBSI4int

Predictor        Coef   SE Coef      T      P
Constant      1.31892   0.04740  27.82  0.000
LogBodySize    0.9800    0.1135   8.63  0.000
Instar_1      -0.0657    0.2090  -0.31  0.753
Instar_2       0.0161    0.1203   0.13  0.894
Instar_3     -0.02575   0.06135  -0.42  0.675
Instar_4     -0.06071   0.05317  -1.14  0.254
LBSI1int      -0.1010    0.1512  -0.67  0.505
```

```
LBSI2int      -0.0293   0.1371  -0.21  0.831
LBSI3int      -0.0679   0.1205  -0.56  0.573
LBSI4int      -0.2271   0.1267  -1.79  0.074

S = 0.173855   R-Sq = 95.0%   R-Sq(adj) = 94.8%
```

Together, these variables explain 95.0% of the variability in the $Mrate$ values. How does this compare to the simple linear model based on just the *LogBodySize* alone? Here is some output for fitting the single predictor model.

```
The regression equation is LogMrate = 1.31 + 0.916 LogBodySize

Predictor       Coef  SE Coef      T      P
Constant     1.30655  0.01356  96.33  0.000
LogBodySize  0.91641  0.01235  74.20  0.000

S = 0.175219   R-Sq = 94.8%   R-Sq(adj) = 94.8%
```

We see that the simple linear model using just body size explains 94.8% of the variability for $Mrate$ all by itself. Is it really worth it to add all of the information from the $Instar$ categories? Probably not, but let's test it anyway.

The null hypothesis is $H_0 : \beta_2 = \beta_3 = \cdots = \beta_9 = 0$, and the alternative is that at least one of the coefficients based on the $Instar$ indicators is different from zero. This calls for a nested F-test where the full model is the one at the start of this solution, and the reduced model is the single predictor model using *LogBodySize* alone. Here are the ANOVA tables for those two models.
Full model (9 predictors):

```
Source           DF       SS      MS       F      P
Regression        9  169.406  18.823  622.75  0.000
Residual Error  295    8.916   0.030
Total           304  178.322
```

Reduced model (1 predictor):

```
Source           DF      SS      MS        F      P
Regression        1  169.02  169.02  5505.26  0.000
Residual Error  303    9.30    0.03
Total           304  178.32
```

We find the F-statistic by seeing how much new variability is explained by the 8 predictors being tested, dividing by the number of terms being tested, and then dividing the result by the mean square error for the full model.

$$F = \frac{(169.406 - 169.02)/8}{8.916/295} = \frac{0.04825}{0.0302} = 1.60$$

We compare this to an F-distribution with 8 and 295 degrees of freedom to find a P-value = 0.124. This is not a small P-value, so we do not reject H_0 and fail to find evidence that the terms based on $Instar$ are useful in this model.

3.58 Here is some output for fitting a model to predict $LogNassim$ with $LogMass$, the $Ifpg$ indicator, the first four $Instar$ indicators, and the product terms for each of these indicators with $LogMass$.

```
The regression equation is
LogNassim = - 1.97 + 0.354 LogMass + 0.346 IFpg - 0.555 Instar1 - 0.558 Instar2
            - 0.213 Instar3 - 0.250 Instar4 + 0.183 LMxIFpg - 0.320 LMxInstar1
            - 0.329 LMxInstar2 - 0.060 LMxInstar3 - 0.211 LMxInstar4

Predictor       Coef  SE Coef       T      P
Constant    -1.97488  0.09177  -21.52  0.000
LogMass      0.35426  0.09869    3.59  0.000
IFpg         0.34564  0.03538    9.77  0.000
Instar1      -0.5548   0.2477   -2.24  0.026
Instar2      -0.5578   0.2207   -2.53  0.012
Instar3      -0.2131   0.1338   -1.59  0.113
Instar4     -0.24977  0.08173   -3.06  0.002
LMxIFpg      0.18259  0.02937    6.22  0.000
LMxInstar1   -0.3199   0.1542   -2.07  0.039
LMxInstar2   -0.3292   0.1653   -1.99  0.048
LMxInstar3   -0.0603   0.1384   -0.44  0.663
LMxInstar4   -0.2110   0.1182   -1.79  0.075

S = 0.169894   R-Sq = 89.2%   R-Sq(adj) = 88.7%
```

Together, these variables explain 89.2% of the variability in the $LogNassim$ values. How does this compare to the simple linear model based on just $LogMass$ alone? Following is some output for fitting the single predictor model.

```
The regression equation is LogNassim = - 1.89 + 0.371 LogMass

Predictor      Coef  SE Coef        T      P
Constant   -1.88738  0.01841  -102.53  0.000
LogMass     0.37096  0.01332    27.85  0.000

S = 0.250145   R-Sq = 75.5%   R-Sq(adj) = 75.5%
```

We see that the simple linear model using just $LogMass$ explains 75.5% of the variability for $LogNassim$ on its own. Is that difference from the model with indicators statistically significant? Looks like a question for the nested F-test.

The null hypothesis is $H_0 : \beta_2 = \beta_3 = \cdots = \beta_{11} = 0$, and the alternative is that at least one of the coefficients based on the indicators or their interactions with $LogMass$ is different from zero. The full model is the one at the start of this solution, and the reduced model is the single predictor model using $LogMass$ alone. Here are the ANOVA tables for those two models.

Full model (11 predictors):

Source	DF	SS	MS	F	P
Regression	11	57.2776	5.2071	180.40	0.000
Residual Error	241	6.9562	0.0289		
Total	252	64.2338			

Reduced model (1 predictor):

Source	DF	SS	MS	F	P
Regression	1	48.528	48.528	775.55	0.000
Residual Error	251	15.706	0.063		
Total	252	64.234			

We find the F-statistic by seeing how much new variability is explained by the 10 terms being tested, dividing by the number of terms being tested, and then dividing the result by the mean square error for the full model.

$$F = \frac{(57.2776 - 48.528)/10}{6.9562/241} = \frac{0.875}{0.02886} = 30.32$$

We compare this to an F-distribution with 10 and 241 degrees of freedom to find a P-value that is essentially zero. This is gives strong evidence that one or more of the indicator or interaction terms are useful in this model.

We can also check the residual plots for both models, single predictor on the right and model with indicators on the left. We see considerable improvement in the normality condition and the residual versus fits plot with the more complicated model (although there are still some issues with both plots).