Assignment #11 Solutions – M. Lacey, 05/01/02

Chapter 14, Problem #12:

(a) The variance of $\hat{Y}_0 - Y_0$ is equal to

$$Var(\hat{Y}_0) + Var(Y_0) - 2Cov(\hat{Y}_0, Y_0) = Var(\hat{b}_0 + \hat{b}_1 x_0) + Var(b_0 + b_1 x_0 + e_0))$$

$$= s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] + s^2 = s^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right].$$

The first term was derived in Problem #11, and we're given that e_0 is independent of the original observations and has variance s^2.

(b) Assuming that e_0 is normally distributed, the distribution of $\hat{Y}_0 - Y_0$ is normal with mean 0 and the variance calculated in part (a). Using the estimated variance *s*, a 100(1- a)% prediction interval will be given by $\hat{Y}_0 \pm t^*_{n-2} s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$, where $t^*_{(n-2)}$ is the (1- a /2) critical value from the *t* distribution with (*n*-2) degrees of freedom.

Chapter 14, Problem #19:

To choose *n* points on the interval [-1,1] to minimize $Var(\hat{b}_1) = \frac{ns^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}$, we need to maximize the denominator $n\sum x_i^2 - \left(\sum x_i\right)^2$. This is achieved by minimizing $\left(\sum x_i\right)^2$ while maximizing $n\sum x_i^2$, and we can easily see that $\sum x_i^2 \le n$ for $x_i \in [-1,1]$ while $\left(\sum x_i\right)^2 \ge 0$. By setting half of the values to +1 and the other half to –1, the denominator is maximized and $Var(\hat{b}_1) = \frac{ns^2}{n(n) - 0} = \frac{s^2}{n}$.

Chapter 14, Problem #25:

Show that $s^2 I = \Sigma_{\hat{Y}\hat{Y}} + \Sigma_{\hat{e}\hat{e}}$. Let $P = X(X^T X)^{-1} X^T Y$. From Lemma A on p. 539, we know that $P = P^T = P^2$ and $(I - P) = (I - P)^T = (I - P)^2$. Since $\hat{Y} = PY$ and $\Sigma_{YY} = s^2 I$, Theorem B on p.534 gives $\Sigma_{\hat{Y}\hat{Y}} = P\Sigma_{YY}P^T = s^2 PP^T = s^2 P$. Also, since $\hat{e} = Y - \hat{Y} = Y - PY = (I - P)Y$, another application of Theorem B gives $\Sigma_{\hat{e}\hat{e}} = (I - P)\Sigma_{YY}(I - P)^T = s^2(I - P)$. Then $\Sigma_{\hat{Y}\hat{Y}} + \Sigma_{\hat{e}\hat{e}} = s^2 P + s^2(I - P) = s^2 I$. It follows that $\sum Var(\hat{Y}_i) + \sum Var(\hat{e}_i) = Trace(s^2 I) = ns^2$.

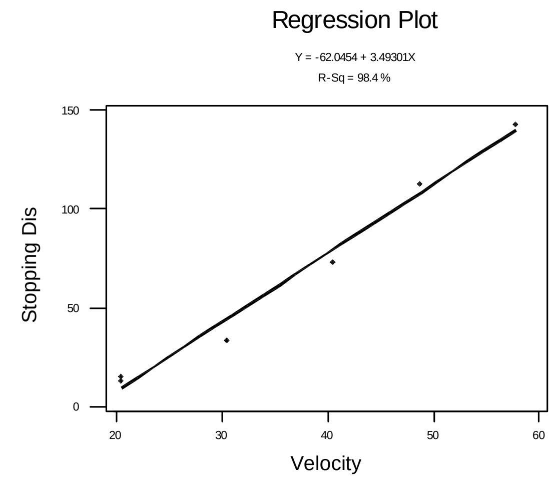
Chapter 14, Problem #38:

Modeling stopping distance (or the square root of stopping distance) as a function of velocity for 6 observations.

Model 1: `Stopping Distance = - 62.0 + 3.49 Velocity`

Predictor	Coef	StDev	T	P
Constant	-62.045	8.631	-7.19	0.002
Velocity	3.4930	0.2212	15.79	0.000

S = 7.563   R-Sq = 98.4%   R-Sq(adj) = 98.0%



Model 2: Sqrt(Stopping Distance) `= - 0.878 + 0.228 Velocity`

Predictor	Coef	StDev	T	P
Constant	-0.8776	0.3673	-2.39	0.075
Velocity	0.227724	0.009415	24.19	0.000

S = 0.3219   R-Sq = 99.3%   R-Sq(adj) = 99.2%

While both models provide a very good fit to the data, the R^2 is better for the second model. Because the laws of physics tell us that the stopping distance of an object is proportional to the square of its initial velocity, it seems logical that Model 2 would be the better choice.

Chapter 14, Problem #44:

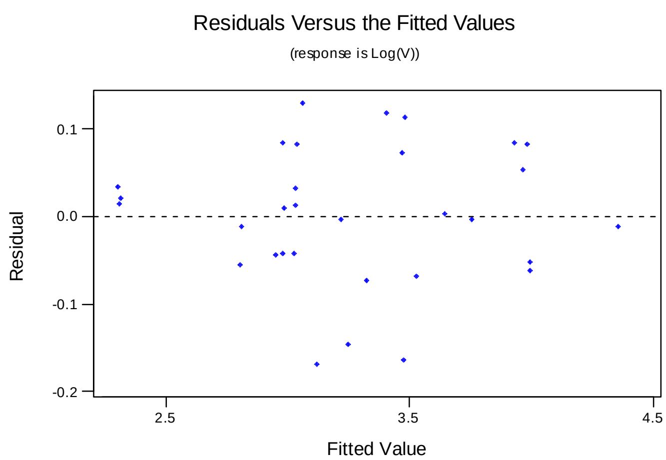
Developing a model relating the volume of 31 black cherry trees to their height and diameter.

There are a lot of different models to try, but thinking about the geometry of the relationship between height (H), diameter (D), and volume (V) for cylindrical objects suggests a model based on the equation $V=H\pi D^2/4$. For a linear model, this is equivalent to $\log(V) = b_0 + b_1 (\log(H)) + b_2 (\log(D))$. Taking the appropriate transformations of the variables and fitting the model, we get

```
Log(V) = - 6.63 + 1.98 Log(D) + 1.12 Log(H)

Predictor        Coef       StDev          T        P
Constant      -6.6316      0.7998      -8.29    0.000
Log(D)        1.98265     0.07501      26.43    0.000
Log(H)         1.1171      0.2044       5.46    0.000

S = 0.08139     R-Sq = 97.8%     R-Sq(adj) = 97.6%
```



There's no evidence of any patterns in the residuals, both regression coefficients are significant, and the R^2 value is very high, so we can conclude that this model provides a good fit to the data.