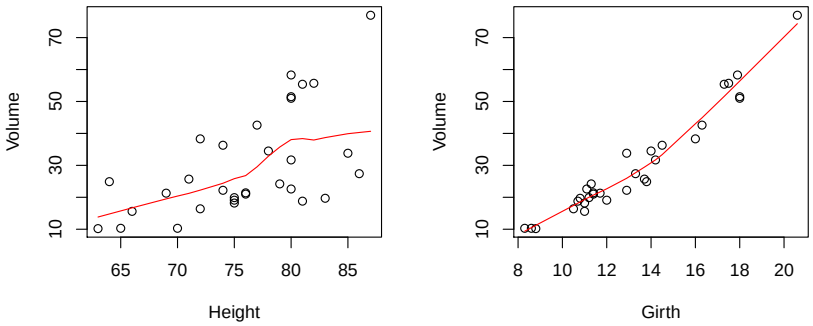
Homework 1 STAT 505 Fall 2014 Solutions

My audience is not just a forestry student – it's also Stat 505 folks – and I'll try to explain the choices I would make. Many times I'll be choosing one path of several justifiable paths.

Data Exploration

Data has been collected on dimensions of 31 black cherry logs. We have three variables: height in feet, DBH (the foresters call it 'Diameter at Breast Height', which the R help mistakenly calls girth) in inches, and volume in ft^3. I would guess that these trees are a convenience sample from one region of the US.

Any analysis starts with plots of the data. We want to examine the relationships between two predictors (Height and Girth) and the response: Volume, so a scatterplot with smoother is a good starting point.

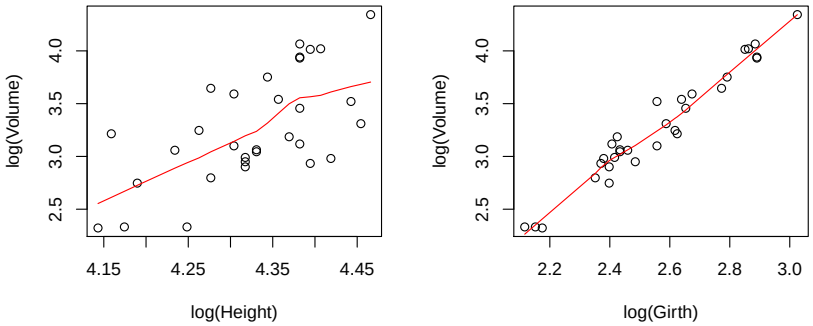


One could explain that these plots exhibit spread increasing with mean volume and possibly non-linear relationships.

Alternatively, we could actually avoid plots in the raw dimensions and instead discuss the geometry of Volume. If logs were perfect cylinders, volume would be $\pi r^2 h = \pi(\frac{g}{2})^2 h$ where r is radius, h is height, and g is girth. We can't very well model multiplicative relationships, so we would take logs to get a nice additive model:

$$\log(V) = \log(\pi/4) + 2\log(g) + \log(h)$$

These plots are then of interest:



I would emphasize that the reason for taking logs is to allow use of our knowledge about the geometry (not due to a problem with diagnostics, because I would not even fit a model in the original scale.). If you did plot both, then you should note that plots in the log scale do show a more linear relationship and seem to fit better with a 'constant variance' assumption. I see no reason to include the coplots in the report. Looking at them convinced me that we do not need to include an interaction term between height and girth in the model. The girth effect seems to be the same (same slope) for all levels of height (when height is cut into six shingles), but with the geometric argument, interaction is not considered.

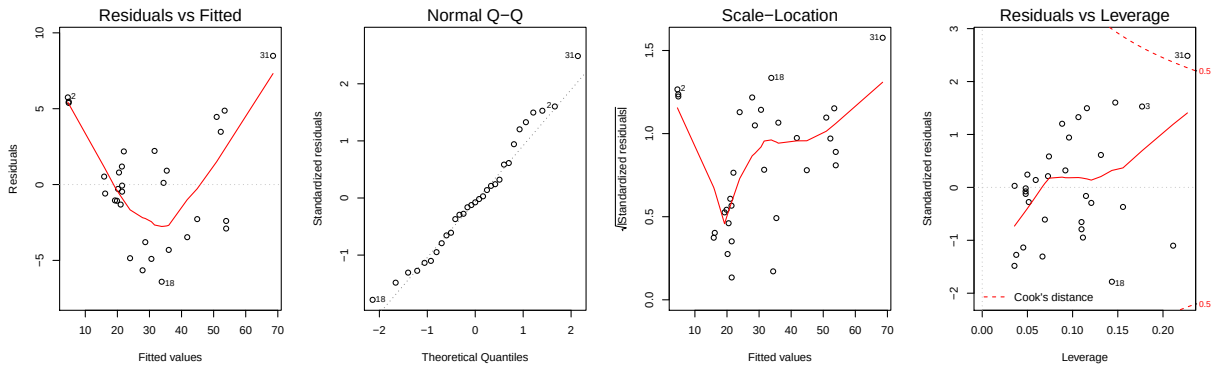
One person observed that we also could obtain a linear model by creating the variable g^2h and use it as the sole predictor. That also makes sense, and would perhaps give a model which is easier to interpret. In that case, one could replace the last two plots with a plot of Volume (not logged) versus a newly created variable, g^2h.

Model Fitting and Interpretation

I wish now that I hadn't forced you to fit the model to volume in original scale. Here it is with coefficients and diagnostic plots.

$$V_i = \beta_0 + \beta_1 h_i + \beta_2 g_i + \epsilon_i \; i = 1, 2, \ldots, 31$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.9877	8.6382	-6.71	0.0000
Height	0.3393	0.1302	2.61	0.0145
Girth	4.7082	0.2643	17.82	0.0000



I don't think this model is very useful because it does not take the geometry into account. It says that a one foot increase in Height increases Volume by .34 ft^3 (SE = 0.13), and a one inch increase in girth increases volume by 4.7 ft^3 (SE = 0.26). Tests for slopes being zero have small p-values, and the model explains 94% of the variation in Volume. The major problem in the diagnostic plots is a strong "V" shape in the residuals versus fitted values plot. This indicates that we have missed some important aspect of the data. I did try adding a Girth^2 term which removes the problem in the first plot, but then the scale – location plot shows strong trend of increasing variance. There are no points with large leverage, and residuals seem quite close to having a normal distribution.

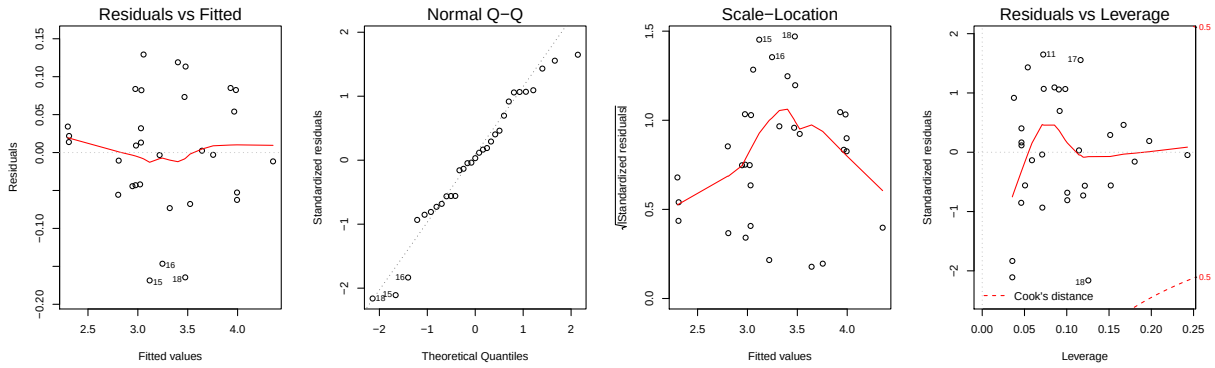
Lesson learned: we either get an odd curve in the residual versus fitted plot or we find that variance is not constant.

We'll now see that the 'log both sides' model solves both problems.

$$\log(V_i) = \beta_0 + \beta_1 \log(h_i) + \beta_2 \log(g_i) + \epsilon_i \; i = 1, 2, \ldots, 31$$

The output for this linear model is in this table:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.6316	0.7998	-8.29	0.0000
log(Height)	1.1171	0.2044	5.46	0.0000
log(Girth)	1.9826	0.0750	26.43	0.0000



I want to know that Stat 505 folks can interpret the output. We do not expect the formula for volume of a cylinder (or a cone) to hold exactly for these data. Tree trunks are bumpy and have some taper, and are not circular in cross section. Therefore we do not expect to obtain estimates which match the cylindrical volume formula exactly. I'm sure that a forestry student knows that, so I might say this:

The intercept in the model estimates the multiplicative factors needed to model the geometry (π) and for unit conversion (girth in inches, height in feet, volume in cubic feet). Our estimate is $\widehat{\beta}_0$ = -6.632 with an approximate 95% confidence interval of (-8.27, -4.993). The test associated with this line of output is not of interest because we do not expect this factor to be zero. The second line of output shows an estimate of $\widehat{\beta}_1$ =1.117 as the exponent for log(height) with a 95% CI of (0.698, 1.536). It is interesting to note that the interval contains 1, which is consistent with the formula for area which contains height as a simple factor. Volume does indeed increases in direct proportion to height. Finally, the coefficient for log girth is estimated as $\widehat{\beta}_2$ =1.983 with a 95% CI of (1.829, 2.136). The interval contains 2, as it should because girth (or radius) enters the volume formula with a coefficient of 2. The tests in the summary table each have small p-values, indicating strong evidence that the exponents each are non–zero, so each term is improving the fit of the model substantially. The test for height effect is conditional on having girth in the model and the test for girth is conditional on having height in the model.

To look at the relative strength of the two variables, we have to imagine holding one fixed and changing the other by a factor, say of 10% (or doubling). If Height is increased by 10% with Girth held constant, then volume increases by a factor of $1.1^{1.12} = 1.11$, and if doubled by a factor of $2^{1.12} = 2.17$. If we hold Height constant and increase Girth by

10%, we see volume increase by a factor of $1.1^{1.98} = 1.21$, or if doubled, by a factor of $2^{1.98} = 3.94$.

The plots of diagnostics show no problems (Residuals seem randomly scattered over the fitted values, residuals have a close to normal distribution, no evidence of a trend in variances, and no large leverage points).

Discussion

I assume that these trees are a convenience sample of logs cut in one area of eastern USA, not a random sample from all areas in which black cherry grows. Statistics does not allow us to project inference to a larger population. The data are from an observational study, and no treatment was applied, however, there is a very convincing geometric argument that volume is proportional to height times girth^2, so the geometry provides causal explanation of the relationship between height, girth, and volume.

The estimated formula

$$\widehat{Volume} = \exp(-6.632) \times \text{height}^{1.117} \times \text{girth}^{1.983}$$

may not hold exactly for other trees in other areas of the country. We should discuss with the student how the estimated formula might be used. I would be comfortable in using it to predict volume of trees cut in the same area over the same years as the data we observed, but we should revalidate it before applying it to other areas and time periods.

R Code Appendix

```
par(mfrow = c(1, 2))
plot(Volume ~ Height, trees)
lines(with(trees, lowess(x = Height, y = Volume)), col = 2)
plot(Volume ~ Girth, trees)
lines(with(trees, lowess(x = Girth, y = Volume)), col = 2)
```

```
coplot(log(Volume) ~ log(Girth) | Height, data = trees, panel = panel.smooth)
coplot(log(Volume) ~ log(Height) | Girth, data = trees, panel = panel.smooth)
```

```
par(mfrow = c(1, 2))
plot(log(Volume) ~ log(Height), trees)
lines(with(trees, lowess(x = log(Height), y = log(Volume))), col = 2)
plot(log(Volume) ~ log(Girth), trees)
lines(with(trees, lowess(x = log(Girth), y = log(Volume))), col = 2)
```

```
xtable(rawCoef <- summary(rawFit <- lm(Volume ~ Height + Girth, trees)))
par(mfrow = c(1, 4))
plot(rawFit)
```

```
xtable(summary(logFit <- lm(log(Volume) ~ log(Height) + log(Girth), trees)))
fitCoefTable <- summary(logFit)$coef
CI <- cbind(fitCoefTable[, 1] - qt(0.975, 28) * fitCoefTable[, 2], fitCoefTable[, 1] +
    qt(0.975, 28) * fitCoefTable[, 2])
par(mfrow = c(1, 4))
plot(logFit)
```