INFOST691 Data Analysis for Data Science

Advanced Bivariate Analysis

Eddie Chapman

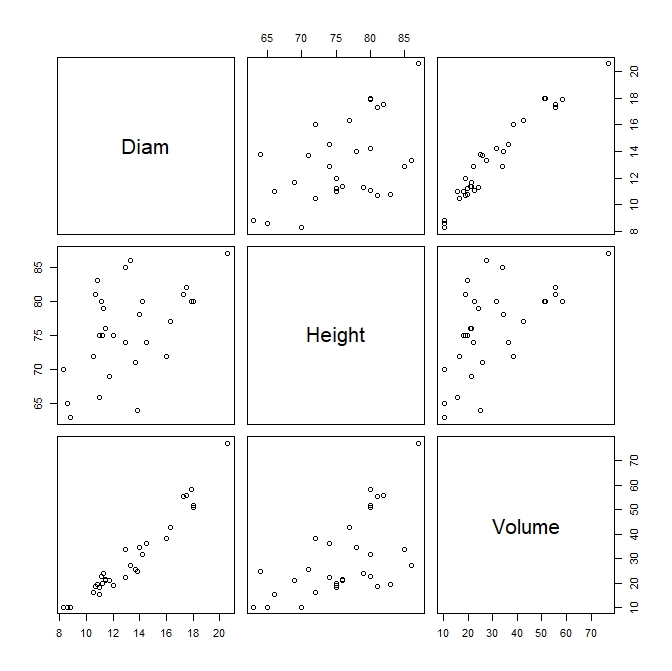
March 2nd, 2018

I used R for the following exercises.

**Regression - Cherry tree data set**

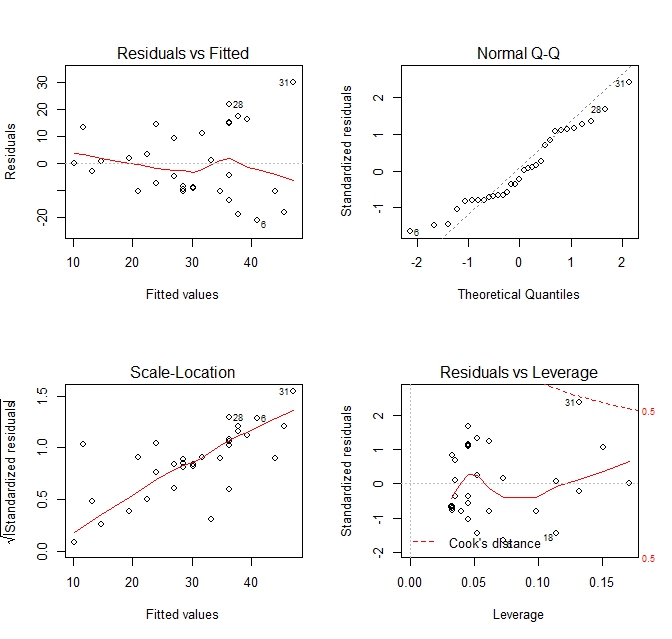
I’m looking to predict the volume of a cherry tree based on diameter, height, or both.

I start by plotting the entire dataset to get a look at variable relationships. It seems like there are correlations worth examining, especially between volume and diameter.



Next, I run separate linear regressions for volume as a product of height, and volume as a product of diameter.

Before examining the descriptive statistics for the regression, I inspect the diagnostic visualization of the fit to confirm that it meets the assumptions of a linear regression.



The Residuals vs Fitted plot can show if residuals have non-linear patterns, indicating a non-linear relationship between variables. Residuals should be equally spread around the red line, which should be roughly horizontal.

The Normal Q-Q plot can show if residuals are normally distributed. They don’t all hug the straight line, but they seem to follow a linear trend.

The Scale-Location plot (aka Spread-Location plot) checks the assumption of equal variance. The red line should be horizontal and the points should be equally spread out. I’m a little concerned about the tilt to the line, but points seem randomly distributed.

The Residuals vs Leverage plot helps identify influential cases. None of the values fall outside of the dotted line for Cook’s distance, but row 31 is close.

I notice that row 31 is identified as an outlier on all four charts. It happens to be the largest tree in the sample. I decide to get rid of it.

Rerunning the linear regression with row 31 removed, I inspect the following descriptive statistics:

Call:

lm(formula = Volume ~ Height, data = data)

Residuals:

Min 1Q Median 3Q Max

-17.992 -8.436 -3.615 10.400 24.306

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -64.6377 28.0369 -2.305 0.02876 \*

Height 1.2329 0.3695 3.336 0.00241 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

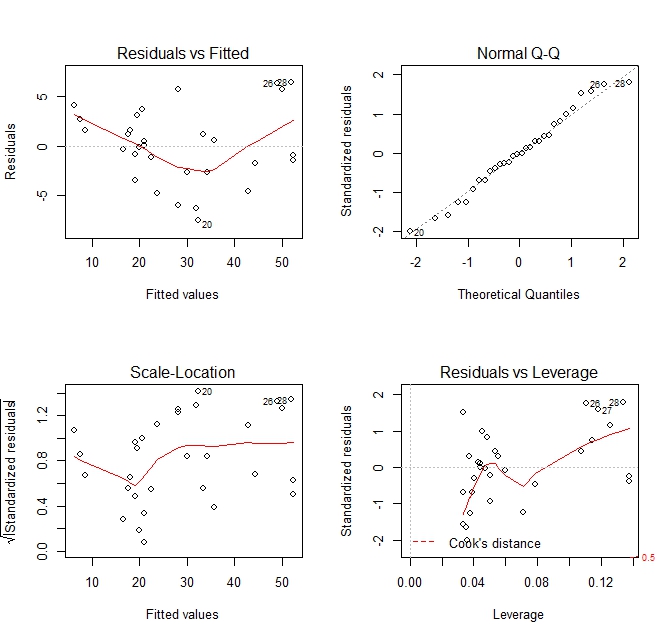
Residual standard error: 12.22 on 28 degrees of freedom

Multiple R-squared: 0.2845, Adjusted R-squared: 0.2589

F-statistic: 11.13 on 1 and 28 DF, p-value: 0.002405

I notice that the residuals are skewed to the right. This might indicate that the model is less effective at predicting the volume of larger trees. Both coefficients have a significant p-value of the t-value, confirming there is some relationship exists between the volume and height. The R-squared values indicate that only a quarter of the variance in volume can be explained by height. That doesn’t seem like much. However, the F-statistic confirms that it is a statistically significant model.

Repeating these steps for volume as a product of diameter:



The residuals vs. fitted plot makes it look like there’s a non-linear relationship between the variables. Looking back at the scatter plot of volume vs. diameter, I wonder if removing that “outlier” is responsible. The distribution of residuals look normal. Again, a few at the top and bottom start to stray. The scale location plot indicates that the data is homoscedastic from fitted value 30-50 but uneven before that. I’m going to proceed anyway. The Residuals vs. Leverage plot doesn’t indicate any points with undue leverage.

Call:

lm(formula = Volume ~ Diam, data = data)

Residuals:

Min 1Q Median 3Q Max

-7.5036 -2.3834 -0.0489 2.3951 6.3726

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -33.3104 3.2784 -10.16 0.0000000000676 \*\*\*

Diam 4.7619 0.2464 19.33 < 0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.813 on 28 degrees of freedom

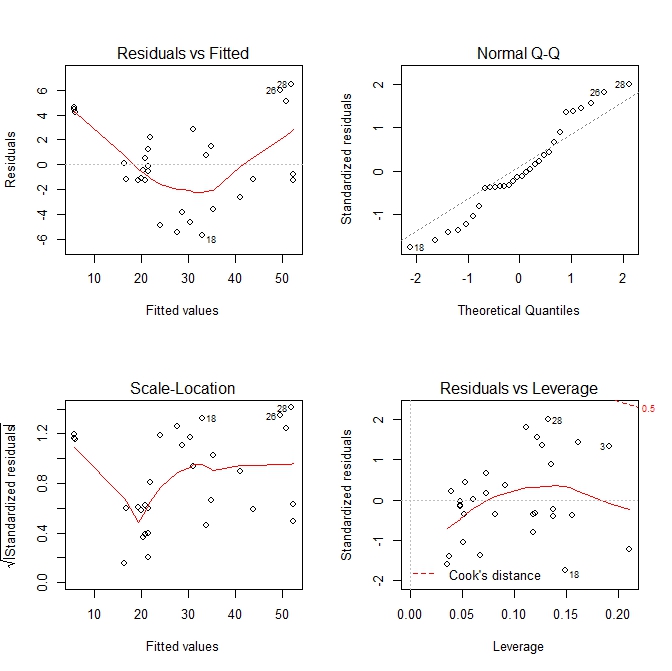
Multiple R-squared: 0.9303, Adjusted R-squared: 0.9278

F-statistic: 373.6 on 1 and 28 DF, p-value: < 0.00000000000000022

Here we see a more symmetrically distributed set of residuals, with a median near 0. The standard error is very low for the Diameter coefficient, a sign of a good fit. Again, the t-value of the coefficient is statistically significant, but to a much higher degree. The R-squared values are very close to 1.0, meaning nearly all of the variance in volume can be explained by diameter. The F-statistic for this fit is many times larger than the fit for height and volume, using the same degrees of freedom. The corresponding p-value adds confidence to this comparison.

As far picking a single variable to predict volume with simple linear regression, the choice is clear: diameter outperforms height in every category. However, to accurately predict the volume of a tree, I would think height would be a useful factor as well. I’m guessing that a taller tree has more branches, meaning more volume compared to a shorter tree with the same diameter. Plus, the regression summary showed that height was a significant variable in the correlation (although multiple regression does not require significant p-values for every coefficient).

So, repeating the steps with multiple regression of volume as a product of diameter and height:



Once again, the residuals vs. fitted plot looks concerningly quadratic, but maybe that’s my doing? The normal q-q plot also looks less normally distributed than the previous regressions. The scale-location plot looks less homoscedastic than either, as well. The residuals vs. leverage plot seems OK.

Call:

lm(formula = data$Volume ~ data$Diam + data$Height, data = data)

Residuals:

Min 1Q Median 3Q Max

-5.6396 -1.2786 -0.4938 2.0262 6.4599

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -52.2362 8.0390 -6.498 0.000000576 \*\*\*

data$Diam 4.4773 0.2518 17.781 < 0.0000000000000002 \*\*\*

data$Height 0.2992 0.1179 2.538 0.0172 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.49 on 27 degrees of freedom

Multiple R-squared: 0.9437, Adjusted R-squared: 0.9395

F-statistic: 226.3 on 2 and 27 DF, p-value: < 0.00000000000000022

These results are encouraging. The residuals are the lowest yet. They’re normally distributed but skewed to the right, just like the volume vs. height simple linear regression. The standard errors for the coefficients are very low, giving a residual standard error that is lower than either of the previous regressions. The height coefficient is significant but has a lower p-value than in the individual regression due to its decreased coefficient estimate. The R-squared values have increased even closer to 1. The F-statistic is not as high as in the simple diameter regression, but there are more variables to consider so it’s still a successful fit.

That said, there is the concern that the predictor variables are correlated, which would throw off the multiple regression. Running a Pearson’s correlation test on height and diameter reveals a significant correlation:

Pearson's product-moment correlation

data: data$Height and data$Diam

t = 2.6326, df = 28, p-value = 0.01364

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1014357 0.6942823

sample estimates:

cor

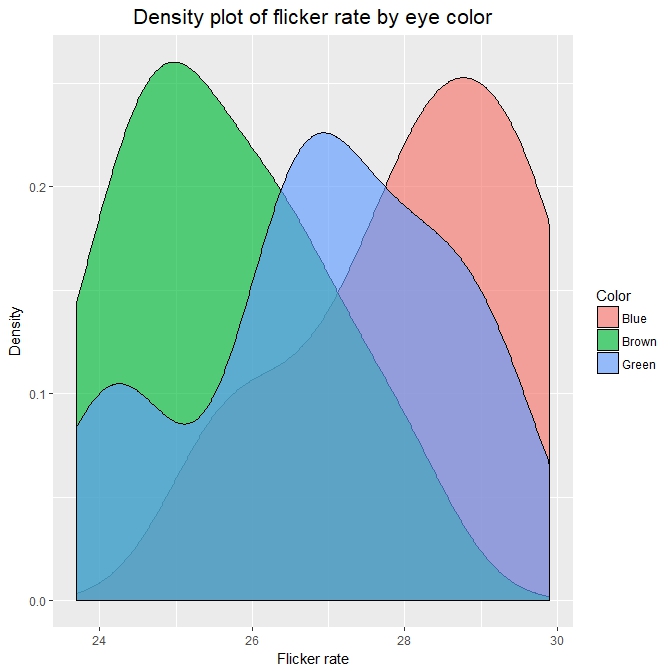
0.4454272

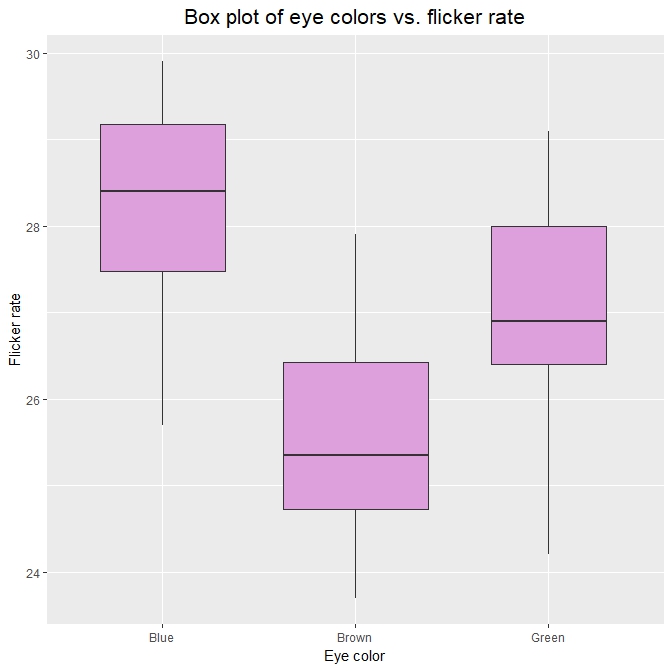
I’m not sure that it’s really changing anything, though. The results of the multiple regression aren’t drastically different than either of the simple linear regressions. The strengths of both simple models show up in the multiple regression.

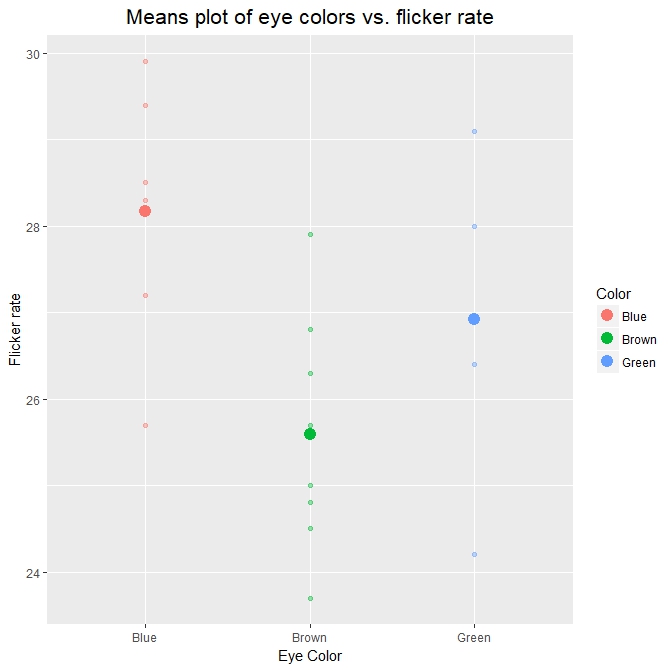
Assuming this correlation isn’t problematic, I’ll take the multiple linear regression. Height isn’t nearly as influential as diameter, but it seems to have a role in determining volume. I would be curious to see how this would play out in a more stalk-like plant that doesn’t branch out as it grows. In that case, height might not be a significant factor. Also, I bet these cherry trees were grown on a farm where there is ample room to grow. I wonder how fighting for space in a forest would change the weighted values of either variable. Maybe the correlation between height and diameter would decrease as tall trees get more sun, and shorter trees have to grow wider to catch more distributed sunlight.

**ANOVA – Eye Color & Flicker Frequency**

For this data set, I am comparing three samples based on eye color for differences in the perception of flickering light. I will be testing the variance in the means between the three groups with an ANOVA test. First, I check the distributions with a density plot and box plot. They appear to be normal and symmetrical. A point chart to visualize variation and difference in means shows does not indicate any outliers. Some points have a large gap between them, but the range and location of the mean in the three samples is comparable.







Looking at the descriptive statistics by eye color, the differences in means seem larger than the differences in standard deviation.

summarise(group\_by(data, Color), count = n(),

m = mean(Flicker), sd = sd(Flicker))

# A tibble: 3 x 4

Color count m sd

<fct> <int> <dbl> <dbl>

1 Blue 6 28.2 1.53

2 Brown 8 25.6 1.37

3 Green 5 26.9 1.84

To test this, I run an ANOVA test on the differences in means of flicker rate between eye color groups. The null hypothesis is that the means of the groups are equal. The alternate hypothesis is that the mean of one or more groups differ.

fit <- aov(Flicker ~ Color, data = data)

summary(fit)

Df Sum Sq Mean Sq F value Pr(>F)

Color 2 23.00 11.499 4.802 0.0232 \*

Residuals 16 38.31 2.394

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The results show a significant difference in the means of one of the samples.

To determine which sample differs from which, I need to run a post-hoc test comparing the differences in means. First I have to test the sample for equality of variance using Levene’s test. In the lecture notes and in-class exercise answers, it seems like the null hypothesis for this test is that the variance is not equal, and that only a significant p-value would indicate equality in variance. But when I looked up the test online (for general use, which might be different than in R), the null hypothesis is that the variance is equal. I might be interpreting these results incorrectly, but I decided to read it that way.

# Checking for equality of variance

# H0 - all group variances are equal

# H1 - at least one group variance differs

leveneTest(Flicker ~ Color, data = data)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 2 0.0912 0.9133

16

So, in this case, the p-value is not low enough to reject the null hypothesis of equal variance. Also, I feel like I’ve only used the “alternate” tests for when data isn’t normal etc. throughout the class so far and I just wanted to try the more powerful post-hoc test. In this case, that is the Tukey test.

# TukeyHSD to find which group differs

# H0 - there is no difference in means between the two samples

# H1 - there is a difference in means between the two samples

TukeyHSD(fit)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Flicker ~ Color, data = data)

$Color

diff lwr upr p adj

Brown-Blue -2.579167 -4.7354973 -0.422836 0.0183579

Green-Blue -1.246667 -3.6643959 1.171063 0.3994319

Green-Brown 1.332500 -0.9437168 3.608717 0.3124225

Correctly or incorrectly applied, it indicates a significant difference in means between the brown and blue eye colors. This is clearly visible in the distribution plots due to the small sample size. In a larger, more complicated experiments, it would be more important to apply the correct tests.