ENTMLGY 6707 Entomological Techniques and Data Analysis

Simple linear regression

Regression analysis is a statistical technique used to quantify a linear relationship between a response/dependent variable and one or more predictor/independent variables. Many of you might remember the equation of a line as $y = mx + b$, where m = slope (the change in y expected with a one unit change in x; "rise over run") and $b = y$-intercept (where the line crosses the y-axis, which by definition is where $x = 0$). It is effectively the same thing as simple linear regression: we are estimating the line that best predicts values of our response variable (Y) as a function of our predictor (X) BUT we are also interested in quantifying the leftover, unexplained variation (i.e., the variability in Y that our fit line fails to predict):

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

$y \ = b \ + \ mx$

where β_0 is the y-intercept ($= b$), β_1 is the slope ($= m$), and ϵ_i is the residual (more on residuals below). Regression is a correlative analysis, so often remind yourself of the adage "correlation $\neq$ causation" when interpreting regression models. When we make a prediction (i.e., produce an estimate using a regression model), we are estimating the expected mean of Y, represented using $\hat{Y}$, for a given value of X. The absolute values of the residuals (i.e., the differences between each observed value and the corresponding predicted value ($\hat{Y}_i$) for given values of X) provides insight on how much of the variation in Y is explained by X.

Here are some famous quotes to keep in mind when using regression analysis:

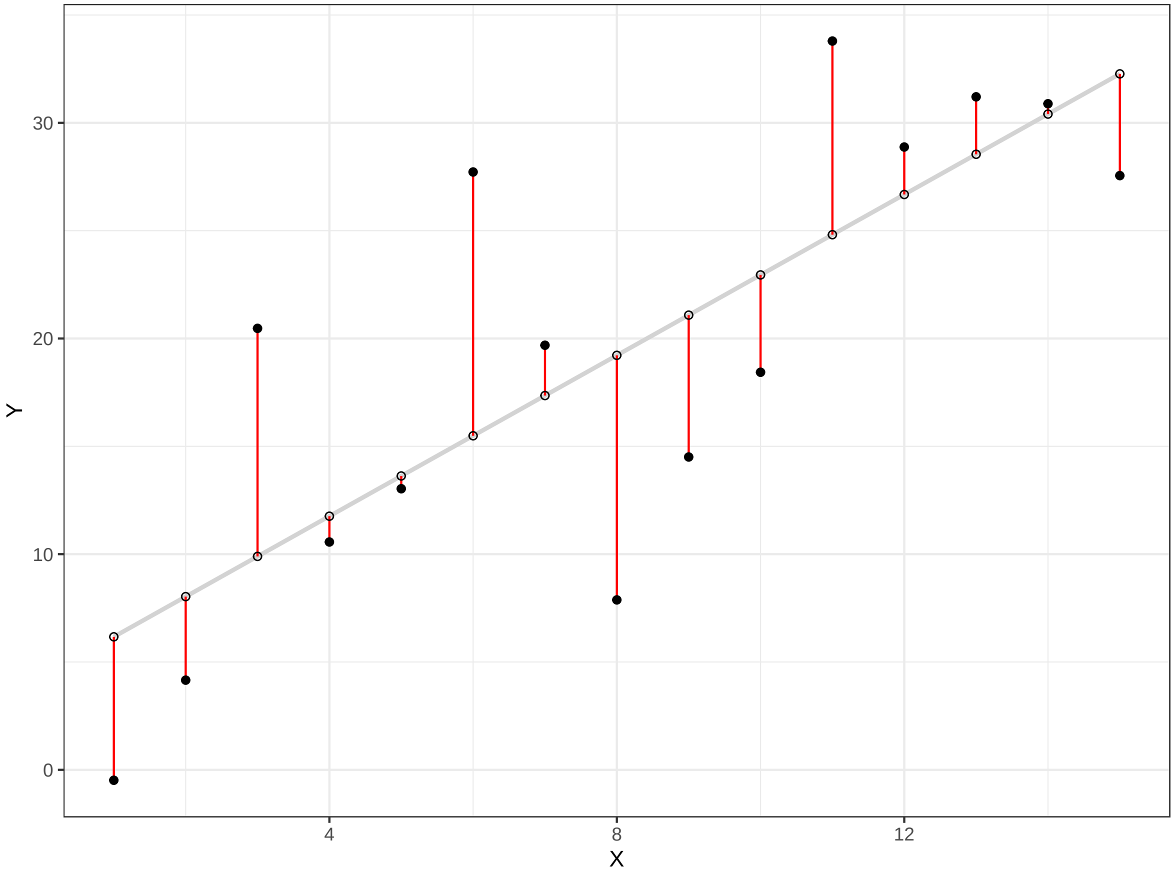
- "Essentially, all models are wrong, but some models are useful." - George Box
- "Regression, without first plotting the data, is truly a regression." - Not sure…it was some famous statistician and my stats adviser always said it.

Assumptions

1. The relationship between response and predictor is linear
2. Residuals are independent
3. Residuals are normally distributed
4. Residuals are homoscedastic (i.e., the variance in Y does not increase or decrease as X increases or decreases)

Residuals

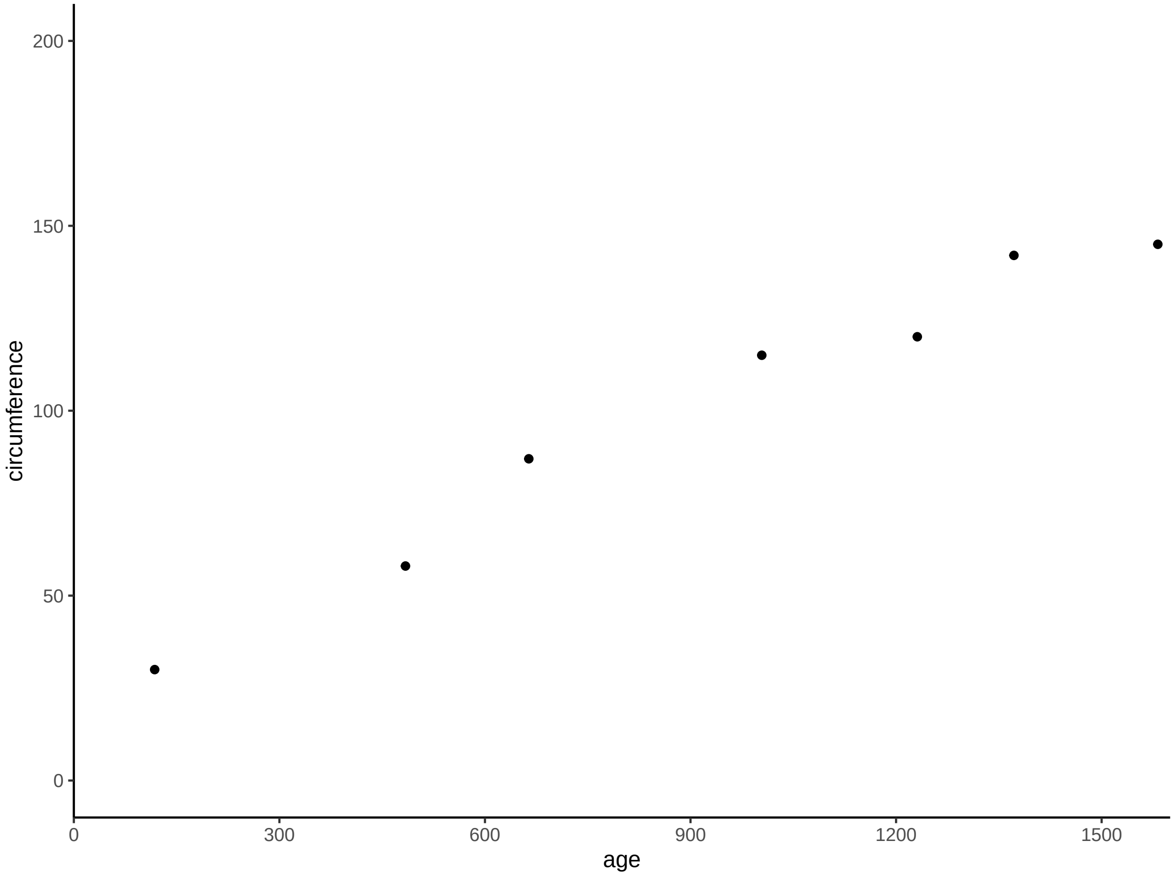
When fitting regression models, there is a lot we assume about the residuals. Thus, understanding how we estimate residuals is key. A residual is the observed value minus the expected (=predicted) value according to our best fit line ($r_i = y_i - \hat{y}$), colored red in the below graph. For the assumptions of a regression, you will often hear or read that the observations should be normal and independent, but the assumptions really pertain to the residuals.



Fitting the model

We will use the `Orange` data from the `datasets` package, which reports the circumferences of orange trees as they age. Specifically, we will look at the increase in circumference with age for a single orange tree. Before you analyze data, plot it first! This will provide some insight into the nature of the relationship.

```
OT <- Orange[Orange$Tree == 1,]
ggplot(data=OT, mapping=aes(x=age , y=circumference)) +
        geom_point()+theme_classic() +
        scale_y_continuous(limits = c(0, 200), breaks = seq(0, 200, by = 50)) +
        scale_x_continuous(limits = c(0, 1600), breaks = seq(0, 1500, by = 300),
                           expand=c(0,0))
```



Fit a simple linear regression using `lm()`, the same command we fit ANOVAs with in R.

```
fit_oranges_1 <- lm(circumference ~ age, data = OT)
summary(fit_oranges_1)
```

```
Call:
lm(formula = circumference ~ age, data = OT)

Residuals:
     1      2      3      4      5      6      7 
-4.052 -5.873  8.461  8.759 -4.736  5.775 -8.335 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.437847   6.543311   3.735   0.0135 *  
age          0.081477   0.006281  12.973 4.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.056 on 5 degrees of freedom
Multiple R-squared:  0.9711,	Adjusted R-squared:  0.9654 
F-statistic: 168.3 on 1 and 5 DF,  p-value: 4.852e-05
```

Intercept and Slope

The test-statistics (for simple linear regression they are t-values) in linear models are testing if each component (intercept and slope) is statistically different from 0 (Null hypotheses: both equal 0). That is, we are testing (i) if the line crosses the y-intercept at a value that differs from 0 and (ii) if the slope is different from 0 (i.e., different from a completely flat line). Using a type I error rate of $\alpha = 0.05$, we reject the null hypothesis when p<0.05.

If you look at the above output, we can see the intercept (B_0) is 24.4 and slope (B_1) is 0.08 (both differ significantly from 0). This means that at age 0 this tree was ~24.4 units in circumference (which is somewhat ridiculous, yes?). Be careful when extrapolating beyond the observed values in your data, and note that the y-intercept might not be biologically interesting/realistic.

As for the slope, we might write something like the following in a manuscript: "The circumference of our study tree increased with age such that a 1 unit increase in age was associated with

~0.08 unit increase in circumference.” The bottom line: try to write in biological/ecological terms, and not statistical terms.

The bottom three lines of summary(fit_oranges_1)

Residual standard error

The bottom three lines of the `summary(fit_oranges_1)` output provide some useful information. The first of those three lines provides an estimate for the total amount of variation that our model “failed” to explain, the `Residual standard error`. This value is equal to the square root of “residual sums of squares divided by the degrees freedom”…the same residual sums of squares from our ANOVA table! That is, $\sqrt{residual\ mean\ square = 324.5/5 = 64.9} = 8.06$, $\sqrt{324.5/5} = 8.06$.

```r
anova(fit_oranges_1)
```

```
Analysis of Variance Table

Response: circumference
          Df  Sum Sq Mean Sq F value    Pr(>F)    
age        1 10921.2 10921.2  168.29 4.852e-05 ***
Residuals  5   324.5    64.9                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R^2

You will next notice values for the Multiple R^2 and Adjusted R^2. These are both R^2 values indicating that our model explains 97.1% and 96.5% of the variation, respectively, in our response variable. The Multiple R^2 will always increase as we add predictors: in the worst case scenario, a predictor would explain 0% of the variation, but almost always *some* additional variation is explained by adding more predictors. The Adjusted R^2 includes a penalty for each predictor added to the model (hence, it is always smaller the Multiple R^2). We have one predictor above, so the Adjusted R^2 includes a small penalty.

F-statistic

The last row of the output provides the overall summary statistics for the fit model. Since we only have one predictor here, $12.973^2 = 168.3$. That is, when the numerator degree of freedom in the F-ratio is 1, $t^2 = F$. Take a look at the ANOVA table above and you will see these exact statistics provided.

Checking assumptions and fit of a regression

Checking assumption 1

As noted above, create a scatter plot of $Y \sim X$ and assess the relationship. Is it linear (i.e., would a line fit through the data)? You might see curvature in the relationship, and that is fine, BUT we will need to transform our response and/or predictor to account for such curvature. More is provided in the "Transformations and Curvilinear Models" tutorial.

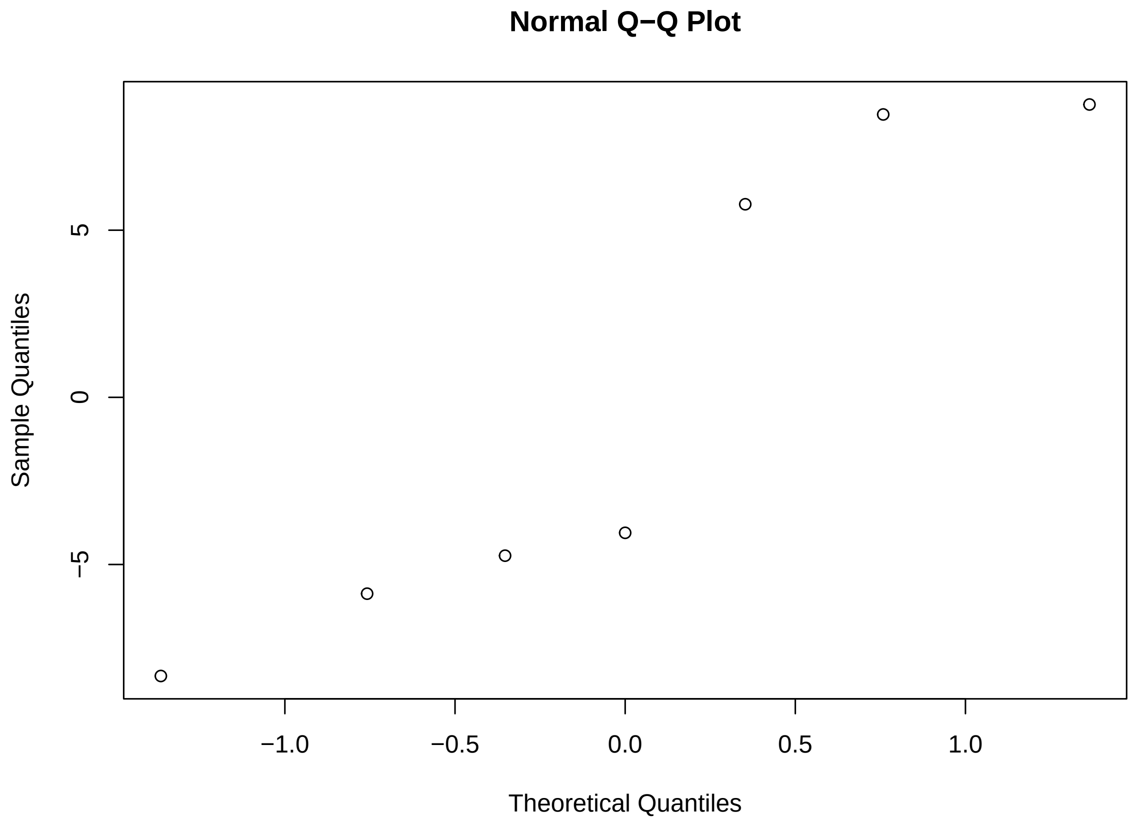
Checking assumption 2

We typically need to know something about the experimental design and nature of data collection to assess independence of residuals. There are formal statistical tests to assess this assumption when analyzing time-series data (e.g., Durbin-Watson statistic) or spatial data (e.g., Moran's I). Things that are closer together in time and/or space tend to be more similar than things that are far apart. We often have to contend with such challenges when analyzing ecological data! Mixed-effects models (which we will cover) are frequently used to analyze data in which residuals are expected or known to be correlated.

Checking assumption 3

Just like with our t-tests and ANOVAs, we can check the normality of our residuals using a Q-Q plot. This one looks quite poor, and we could try some transformations to improve the fit. The supplemental "Transformations and Curvilinear Models" tutorial provides further reading on this topic if you are interested.

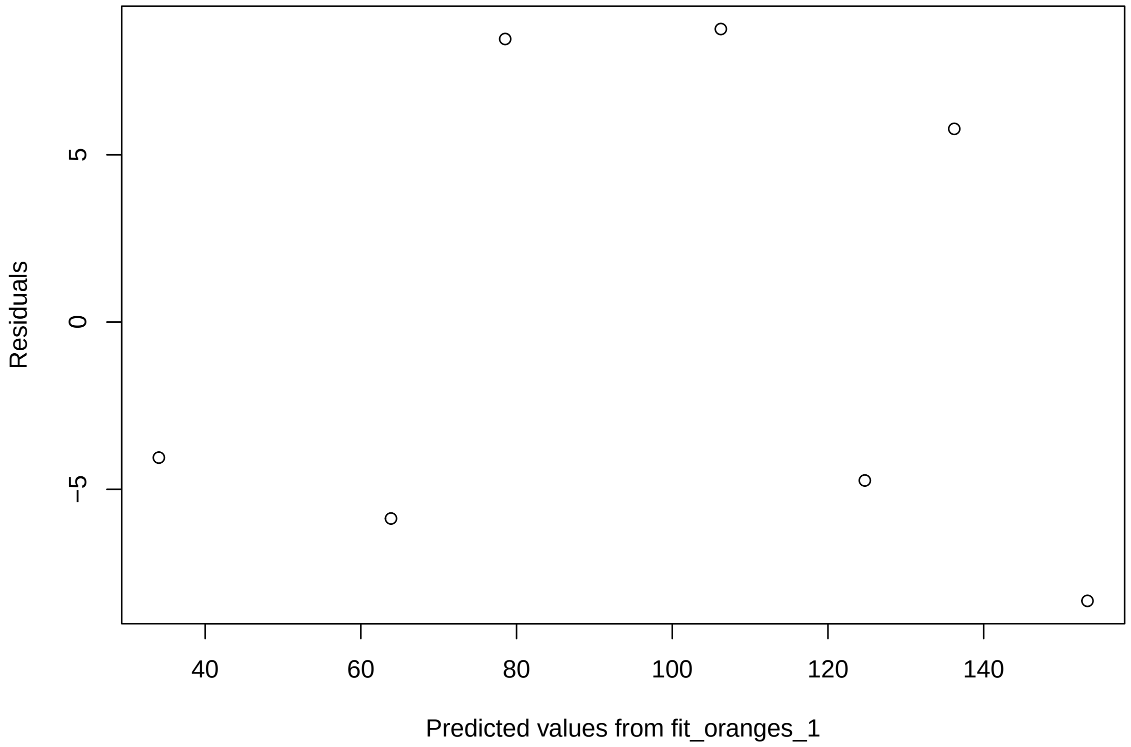
```r
qqnorm((residuals(fit_oranges_1)))
```



Checking assumption 4

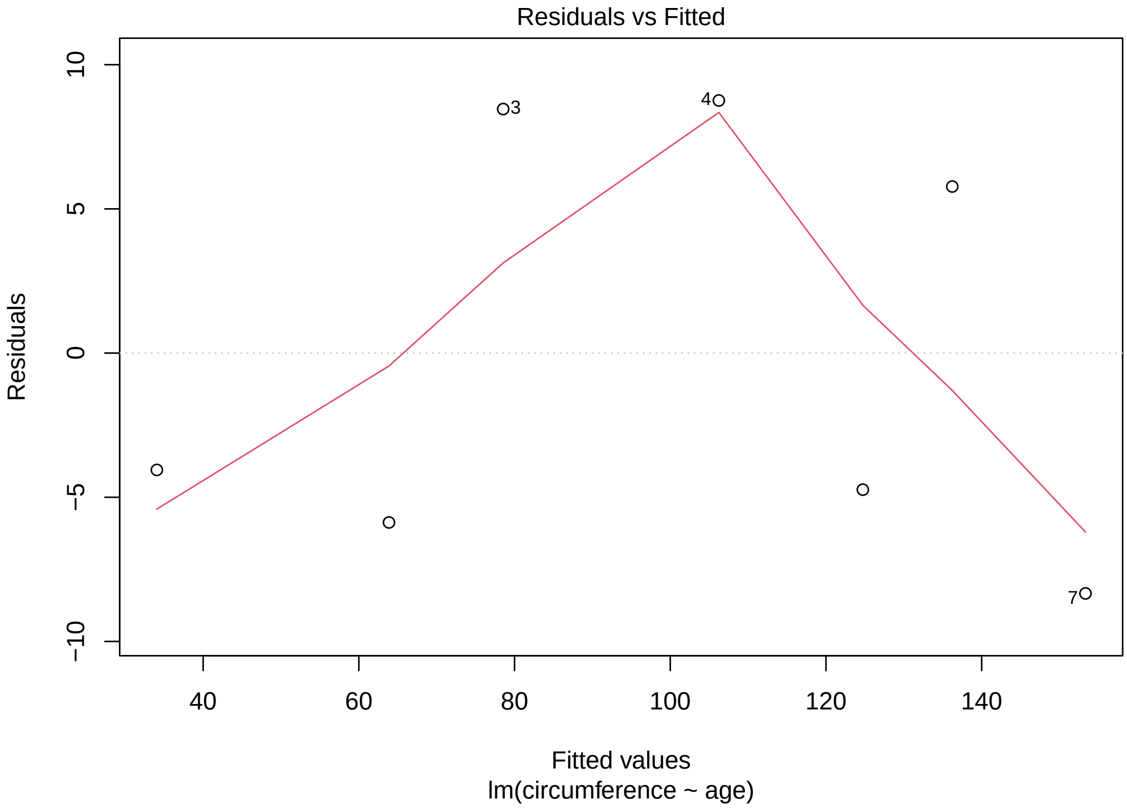
Residual plots ($Residuals \sim Predicted\ values$) are most commonly used to check the assumption of homoscedasticity (= equal variance across values of $\hat{Y}$). A residual plot can also help with assumption 1: if you see curvature or any pattern in the residuals (e.g., if values are more spread out on one side of the graph than the other), you have some problems that need fixing. Indeed, a residual plot should appear as a random scatter of points. Here, there appears to be some issues with heteroscedasticity - the points get more spread out as the predicted values increase.

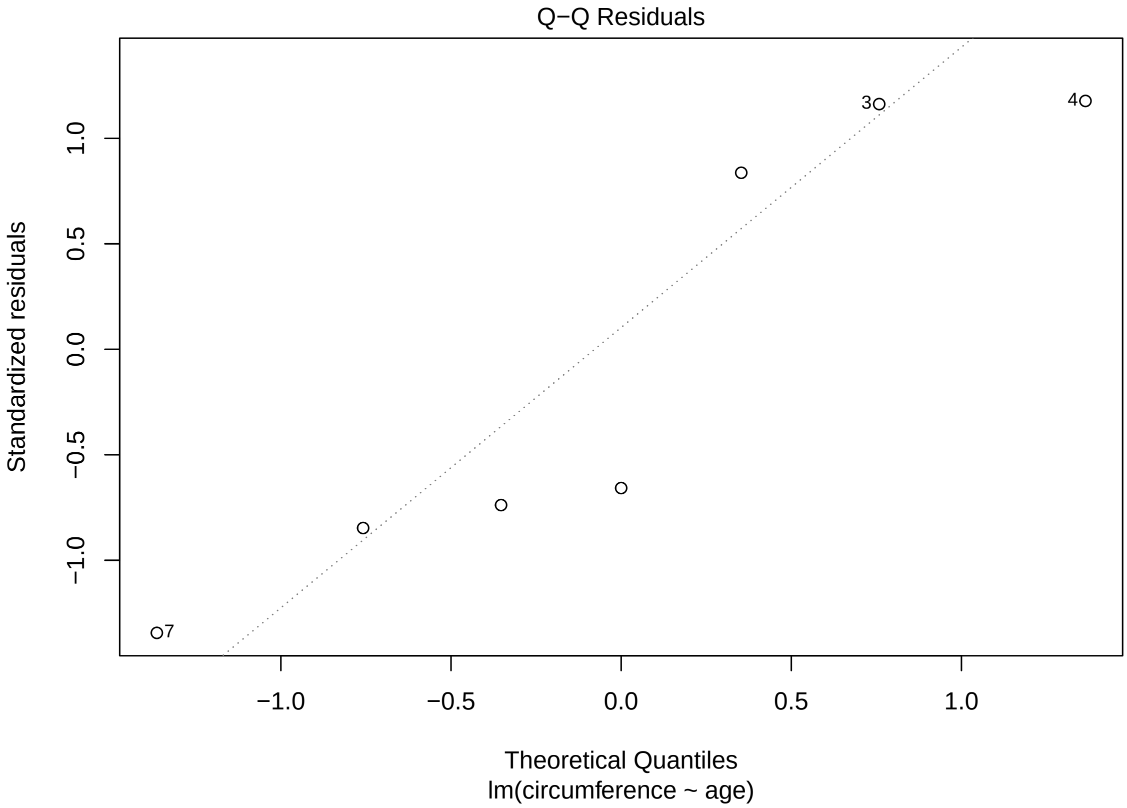
```
plot(residuals(fit_oranges_1) ~ fitted.values(fit_oranges_1),
     xlab = "Predicted values from fit_oranges_1", ylab = "Residuals")
```

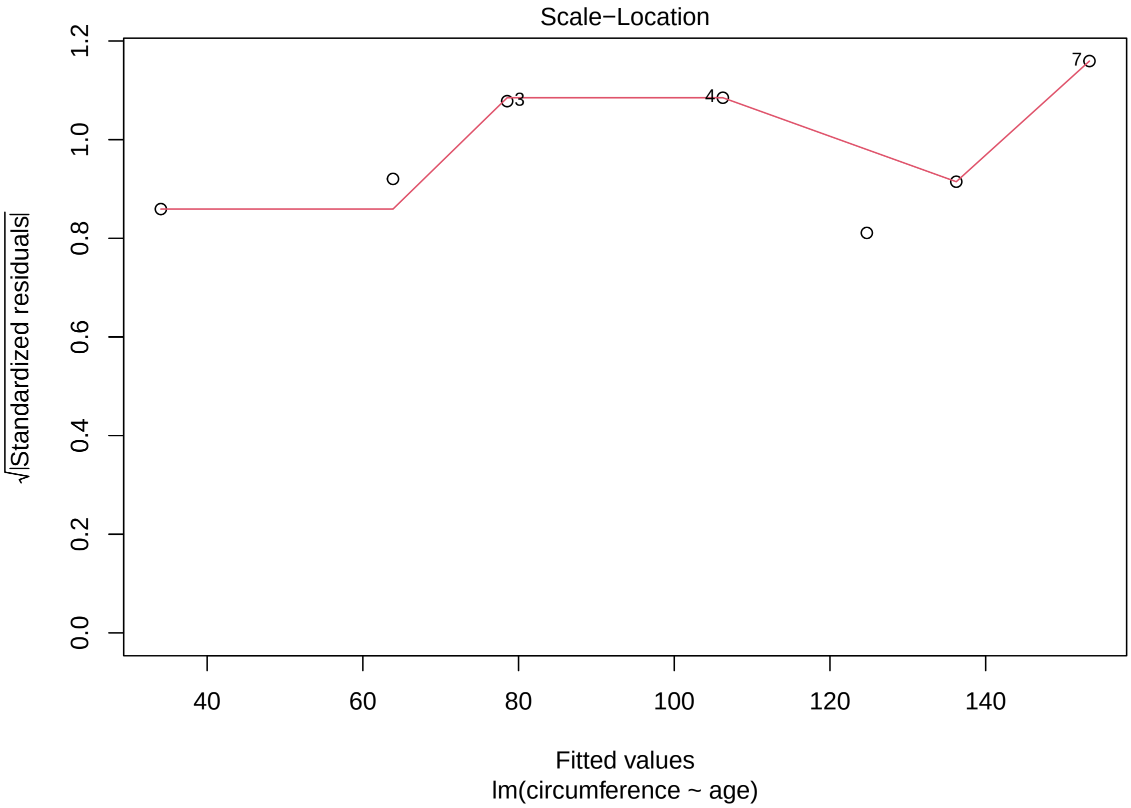


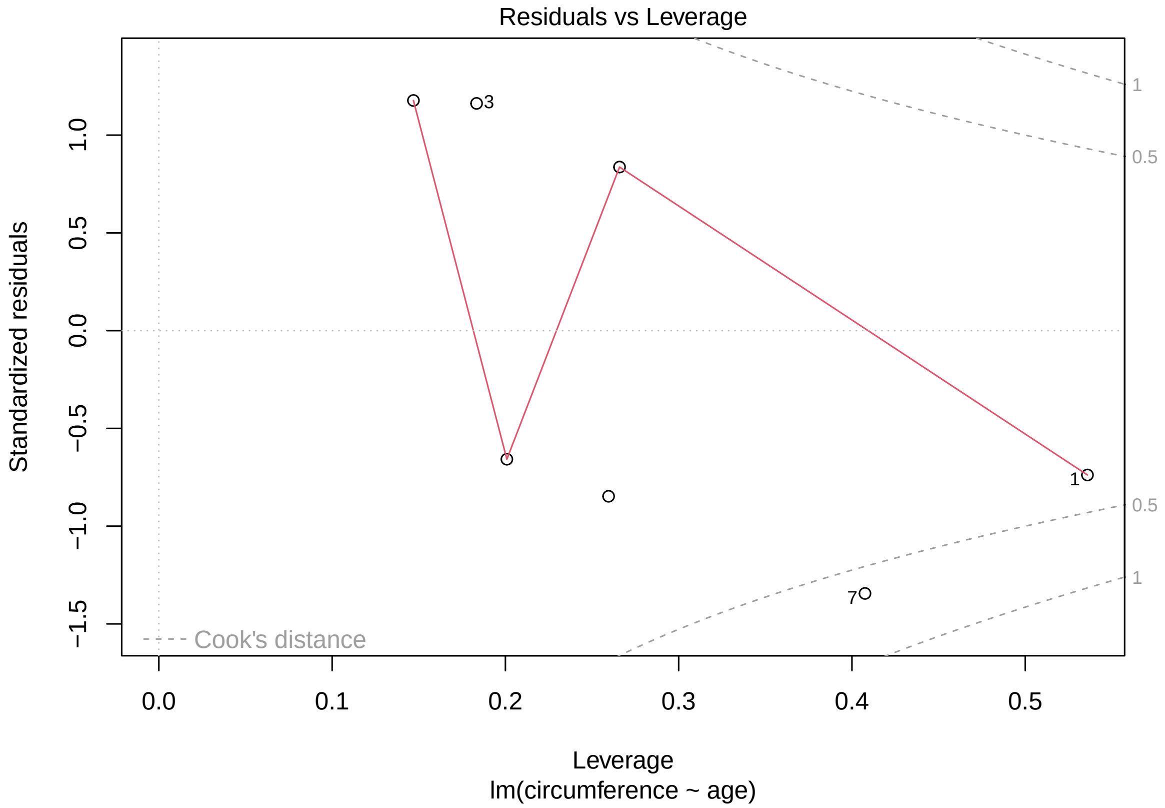
A one-stop-shop for checking assumptions of regressions is available by inputting the model into the `plot()` command. As noted above, we need to try a few things to fix these issues.

```r
plot(fit_oranges_1)
```


Residuals vs Fitted
Residuals
10
5
0
−5
−10
40
60
80
100
120
140
Fitted values
lm(circumference ~ age)
3
4
7


Q−Q Residuals
Standardized residuals
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
3
4
7
Theoretical Quantiles
lm(circumference ~ age)


Scale−Location
√|Standardized residuals|
0.0
0.2
0.4
0.6
0.8
1.0
1.2
40
60
80
100
120
140
Fitted values
lm(circumference ~ age)
3
4
7



Extrapolation warning

Be careful when extrapolating (i.e., making predictions or extending inferences beyond the observed values of your data (Y and X)). You may have an exceptional model, but if the relationship changes as values of X get bigger or smaller, your predictions won't be accurate.

In thinking about orange trees, for example, older trees will add less height per year compared with younger trees, which would mean the slope becomes less steep.

Adding the fit line to a graph

You should always plot your estimated line over your raw data to get a sense of how the model performs. We usually do this after we know the assumptions are met (but this whole process can be quite iterative!). In this example, the estimated line provides the expected tree circumference at each corresponding value on the x-axis, `age`. Thus, we could input totally new values of age and predict the circumference at specific ages. For making predictions from a linear model, R has a very flexible command called predict(). For example, let's say we want to predict the circumference at 500 days old (but again, see my note on extrapolation above).

```
new_data <- data.frame(age = 500)
predicted_val <- predict(fit_oranges_1, newdata = new_data);predicted_val
```

```
       1
65.17643
```

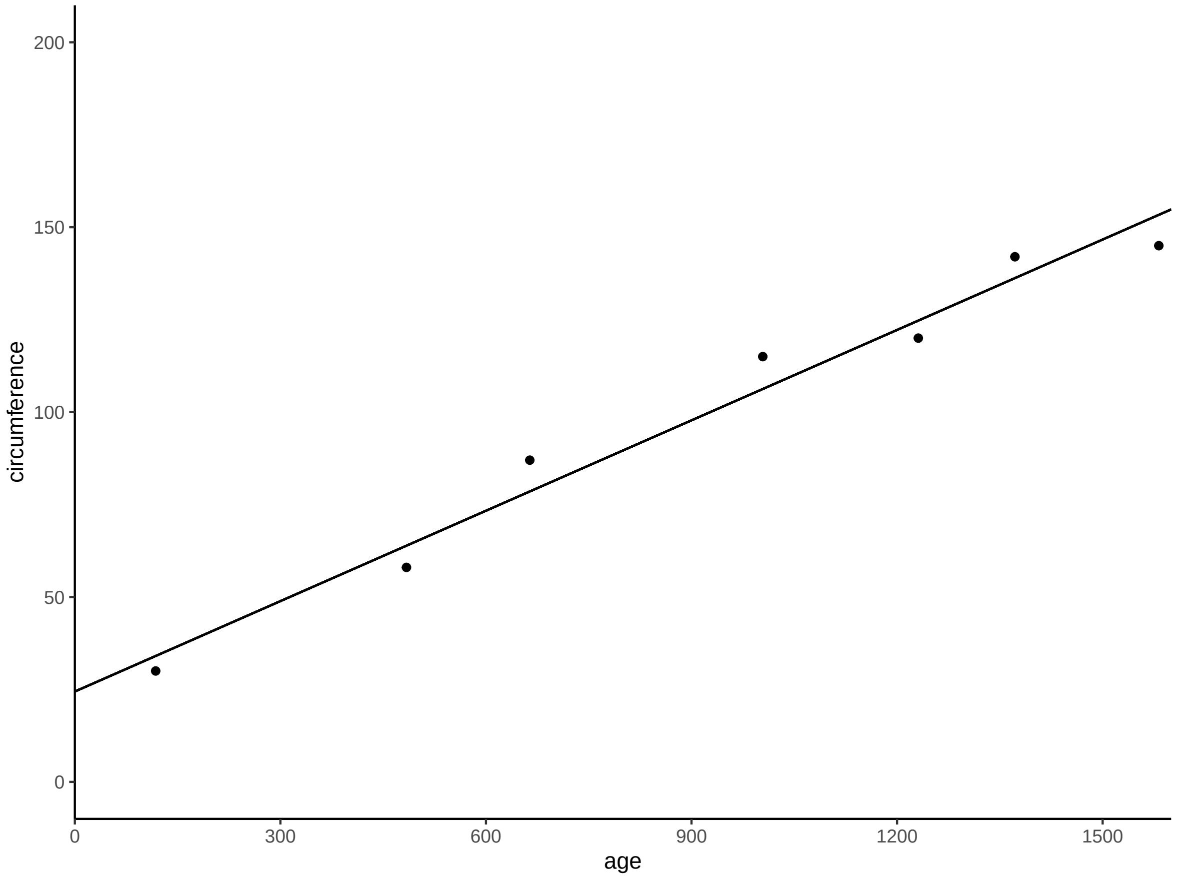
To plot a line over a ggplot() graph, first create a new data frame with really high-resolution values for the predictor age. Then use the predict() command to predict values of circumference for each input value of age. We usually start by looking at the minimum and maximum values of the predictor to avoid over-extrapolating - making predictions outside of the observed range of our predictor.

```
new_data <- data.frame(age = seq(0, 1600, 0.01))
new_data$predicted_values  <- predict(fit_oranges_1, newdata = new_data)
head(new_data)
```

```
   age predicted_values
1 0.00         24.43785
2 0.01         24.43866
3 0.02         24.43948
4 0.03         24.44029
5 0.04         24.44111
6 0.05         24.44192
```

Create a graph of circumference as a function of age, and then use geom_line to fit a line through the high resolution data frame we just created.

```
ggplot(data=OT, mapping=aes(x=age, y=circumference)) +
        geom_point()+theme_classic() +
        geom_line(data=new_data, aes(x=age, y=predicted_values)) +
        scale_y_continuous(limits = c(0, 200), breaks = seq(0, 200, by = 50)) +
        scale_x_continuous(limits = c(0, 1600), breaks = seq(0, 1500, by = 300),
                           expand=c(0,0))
```


200
150
100
50
0
circumference
0
300
600
900
1200
1500
age

R activity

1. Download the `loblolly_pines` data set and load it into R.
2. Create a scatter plot of `height` (feet) as a function of `age` (years) using `ggplot2`. IF you fit a regression of `height` as a function of `age`, what would be your guesses for the `Estimates` of the `Intercept` and slope for `Age`? (i.e., please complete this step without doing any formal analyses).
3. Fit a simple linear regression of `height` as a function of `age`. Name this model `fit_linear`.
4. According to `fit_linear`, how tall is the average loblolly pine at 0 years old? What about 15 years old? What does the model tell you about the average height gained per year by loblolly pines?
5. Provide a residual and a Q-Q plot for `fit_linear`. Do the residuals look normally distributed and homoscedastic? Are you happy with how well the model fits the data? Explain your reasoning.
6. Fit the same model as `fit_linear` but add a polynomial term for `age` (i.e., a quadratic or squared version of `age`). Name this new model `fit_poly` (coding hint: see the supplemental tutorial).
7. Look at the residuals of `fit_poly`. Do they look normally distributed and homoscedastic? You don't need to provide diagnostic graphs.
8. Reproduce the plot you created above of `height` as a function of `age` and add fit lines from `fit_linear` and `fit_poly` to the graph (i.e., your graph should have two lines overlaid on the cloud of raw data points). Color the line from `fit_poly` as your favorite color.