Linear Regression and ANOVA

NRES 710

Fall 2020

Download the R code for this lecture!

To follow along with the R-based lessons and demos, right (or command) click on this link and save the script to your working directory

Overview: Linear Regression

Classical linear regression involves testing for a relationship between a continuous response variable (dependent variable) and a continuous predictor variable (independent variable).

You can have multiple explanatory variables… hence you can have multiple linear regression. We will focus on *simple linear regression* here.

The null hypothesis is that there is no relationship between the response variable and the predictor variable in your population of interest. That is, observations with larger values for your predictor variable are not expected to be associated with larger or smaller values of the response variable on average.

Simple example

Imagine we are testing for a relationship between the brightness of artificial lighting at long stretch of beach (e.g., from hotels and other forms of development) and the total number of hatchling sea turtles per nest that successfully make it to the ocean.

Population: All nests in this particular stretch of beach
Parameter(s): The mean number of hatchlings per nest that successfully travel from their nest to the ocean and how this changes as a function of the brightness of artificial lighting.
Sample: All monitored nests
Statistic(s): Slope and intercept of the linear relationship between the measured response variable (number of successful ocean-arrivers per nest) and the predictor variable (brightness of artificial lighting)

Some more specifics!

We assume there is some true model out there describing the expected (mean) value of our response variable y as a linear function of our predictor variable x:

$E(y) = \beta_0 + \beta_1 \cdot x$

To interpret this equation: the true mean of our response variable ($E(y)$) is computed by taking the true intercept (β_0) and adding the product of the true slope term (β_1) and our predictor variable. This is just another way of saying that the expected value of the response variable is computed as a linear function of the predictor variable. β_0 and β_0 are both *parameters* that we wish to estimate!

To complete the thought, we also assume that there is some "noise" in the system. The "noise" term in regression and ANOVA is also known as the *residual error*. Specifically, we assume that the noise (residual error) is normally distributed with mean of zero and standard deviation of σ.

Mathematically, we are assuming that our data/sample was generated from this process:

$y = \beta_0 + \beta_1 \cdot x + \epsilon$

OR:

$y = E(y) + \epsilon$

WHERE:

$\epsilon \equiv Normal(0, \sigma)$

This is actually the same assumption we made for a one-sample t-test!

For a t-test, we assume that there is a true population mean μ (equivalent to $E(y)$) and that the true "noise" is normally distributed with standard deviation of σ.

As with a t-test, where we can only approximate the true population mean by computing the sample mean, we can only approximate the linear relationship between our response and predictor variables:

$\bar{y} = \hat{B}_0 + \hat{B}_0 \cdot x$

Just like any other statistical test, we assume that our observed linear relationship (defined by test statistics $\hat{B}_0$ and $\hat{B}_1$) is just one of many such possible relationships that *could have been derived* from random sampling from our population of interest. If we collected a different sample, we would get a different linear relationship.

NOTE: in linear regression we are generally far more interested in the slope of the linear relationship ($\hat{B}_1$ rather than the intercept). So for now, we assume $\hat{B}_1$ (slope between response and predictor, computed from the sample) is the main test statistic of interest!

So.. what is the sampling distribution for our test statistic $\hat{B}_1$ under the null hypothesis in this case? Well, the answer is that it (when converted to units of standard error) is t-distributed! Let's look into this a bit more.

Regression and t-tests- the link!

Our discussion of t-tests actually rolls us straight into linear regression. Why? How?

In a one-sample t-test we are interested in estimating the true population mean, and we assume that our t-statistic (i.e., deviation of the sample mean from the null mean, in units of standard error) is t-distributed with degrees of freedom of one less than the sample size.

What is the hypothesis of a typical one-sample t-test? ($\mu = 0$ - that is, the true mean is zero!)

What is the hypothesis of a linear regression? (Slope = 0 - that is, the true relationship is zero).

So already we are seeing a bit of a similarity.

t-test recap In a t-test we assume that the population mean is equal to the null mean and that the data are normally distributed. We could write this as (using regression notation):

$y = \beta_0 + \epsilon$

WHERE:

$\epsilon \equiv Normal(0, \sigma)$

In the above equation, β_0 represents the population mean under the null hypothesis.

We approximate our population mean using the sample mean $\bar{\beta_0}$ (formerly known as $\bar{x}$) and we use the CLT and other statistical theories to show that the t-statistic:

$t = \frac{\bar{\beta_0} - \beta_0}{StdErr(\beta_0)}$

Is t-distributed with df computed as the sample size minus the number of parameters estimated in the model (there is only one estimated parameter- the sample mean $\bar{\beta_0}$).

linear regression version In linear regression we assume that the mean of our response is determined by two parameters- the intercept and the slope (linear relationship with the predictor variable). The null hypothesis (usually) is that the true mean is defined only by the intercept term and there is no relationship with the predictor variable (slope term is equal to zero).

The slope term of the linear regression can be computed as:

$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i-1}^{n}(x_i - \bar{x})^2}$

And the intercept term can be computed as:

$\hat{\beta_0} = \bar{y} - \beta_1 * \bar{x}$

Simple linear regression models (that is, the slope and intercept terms above) are fitted using “least squares”. The best fit model minimizes the sum of the squared residuals. DRAW THIS OUT.

The standard error of the slope term (as opposed to the standard error of the mean) is computed as:

$std.err_{\hat{\beta_1}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Where the $\hat{\epsilon}_i$ refers to the residual errors.

We can then compute a t-statistic for the slope term (difference from the sample slope term and the null slope term in units of standard error):

$t = \frac{\hat{\beta_1} - \beta_{1null}}{std.err_{\hat{\beta_1}}}$

Just like with the t-test, we assume that this t-statistic is t-distributed under the null hypothesis. This time the degrees of freedom is 2 less than the sample size (since computing the residual error requires computing two parameters: mean and slope).

Simple linear regression: examples

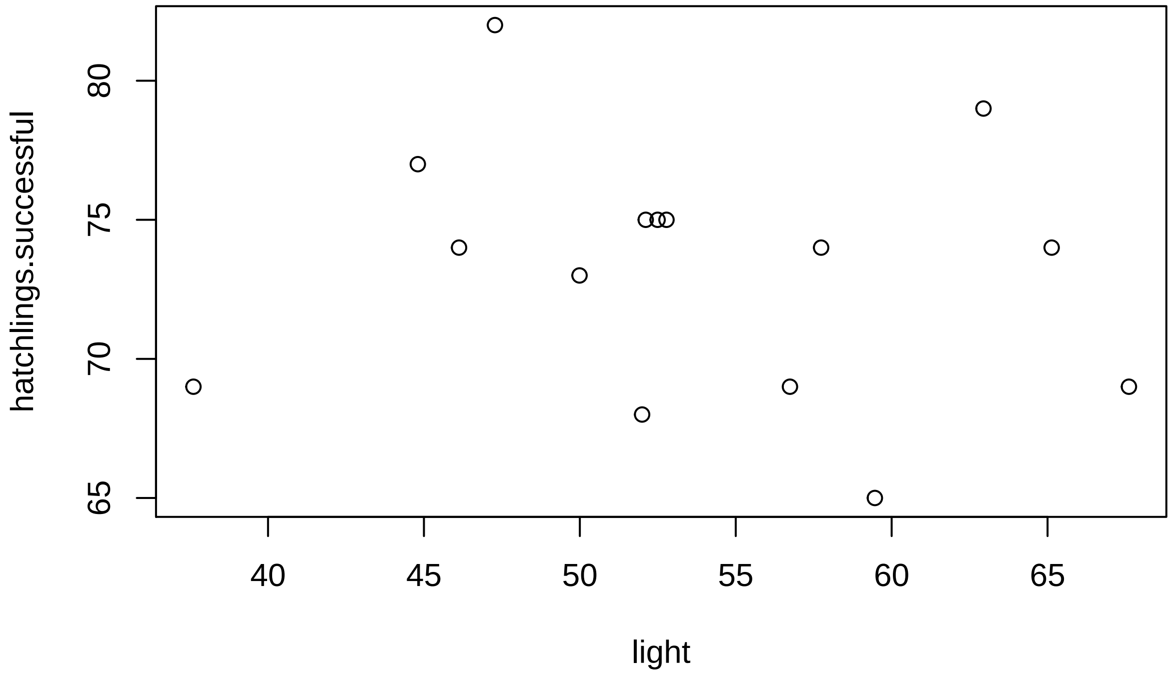
Okay let’s consider the sea turtle example from the beginning of lecture:

Imagine we are testing for a relationship between the brightness of artificial lighting at long stretch of beach (e.g., from hotels and other forms of development) and the total number of hatchling sea turtles per nest that successfully make it to the ocean.

Population: All nests in this particular stretch of beach
Parameter(s): The mean number of hatchlings per nest that successfully travel from their nest to the ocean and how this changes as a function of the brightness of artificial lighting.
Sample: All monitored nests
Statistic(s): Slope and intercept of the linear relationship between the measured response variable (number of successful ocean-arrivers per nest) and the predictor variable (brightness of artificial lighting)

First we will simulate some data under a known process model:

```
eggs.per.nest <- 100
n.nests <- 15
light <- rnorm(n.nests,50,10)   # make up some light pollution values (predictor var)

probsucc <- function(light){    # egg success as a function of light pollution
  plogis(1.5-0.01*light)
}

hatchlings.successful <- rbinom(n.nests,eggs.per.nest,probsucc(light))   # determine number of successf

#curve(probsucc,0,100)

plot(hatchlings.successful~light)  # plot the data
```



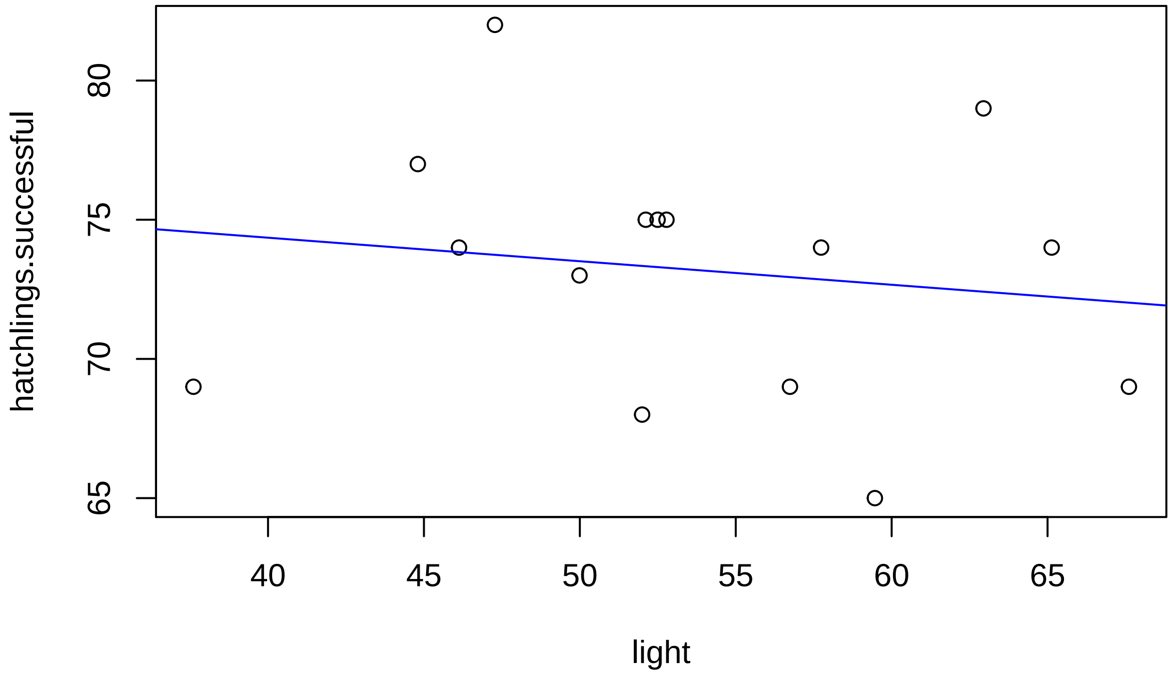
Now that we have data, let's run a linear regression!

```
slope <- sum((light-mean(light))*(hatchlings.successful-mean(hatchlings.successful)))/sum((light-mean(li
intercept <- mean(hatchlings.successful) - slope*mean(light)

exp.successful <- intercept+slope*light # expected number of eggs for each observation
residuals <- hatchlings.successful-exp.successful

stderr <- sqrt(((1/(n.nests-2))*sum(residuals^2))/(sum((light-mean(light))^2)))    # standard error

t.stat <- (slope-0)/stderr    # t statistic
```

```
pval <- 2*pt(t.stat,n.nests-2)    # p value


############
# use lm function instead (easy way!)

model <- lm(hatchlings.successful~light)

summary(model)   # get the same t stat and p-value hopefully!
```

```
## 
## Call:
## lm(formula = hatchlings.successful ~ light)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -7.709 -3.480  1.145  1.748  8.261 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 77.73473    8.23497   9.440  3.5e-07 ***
## light       -0.08452    0.15186  -0.557    0.587    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.623 on 13 degrees of freedom
## Multiple R-squared:  0.02327,    Adjusted R-squared:  -0.05186 
## F-statistic: 0.3097 on 1 and 13 DF,  p-value: 0.5873
```

```
############
# plot regression line!

plot(hatchlings.successful~light)  # plot the data
abline(intercept,slope,col="blue")
```



```
mod <- lm(Volume~Girth,data=trees)
summary(mod)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Assumptions of simple linear regression

Normal distribution

Simple linear regression assumes that the model **residuals** are normally distributed.

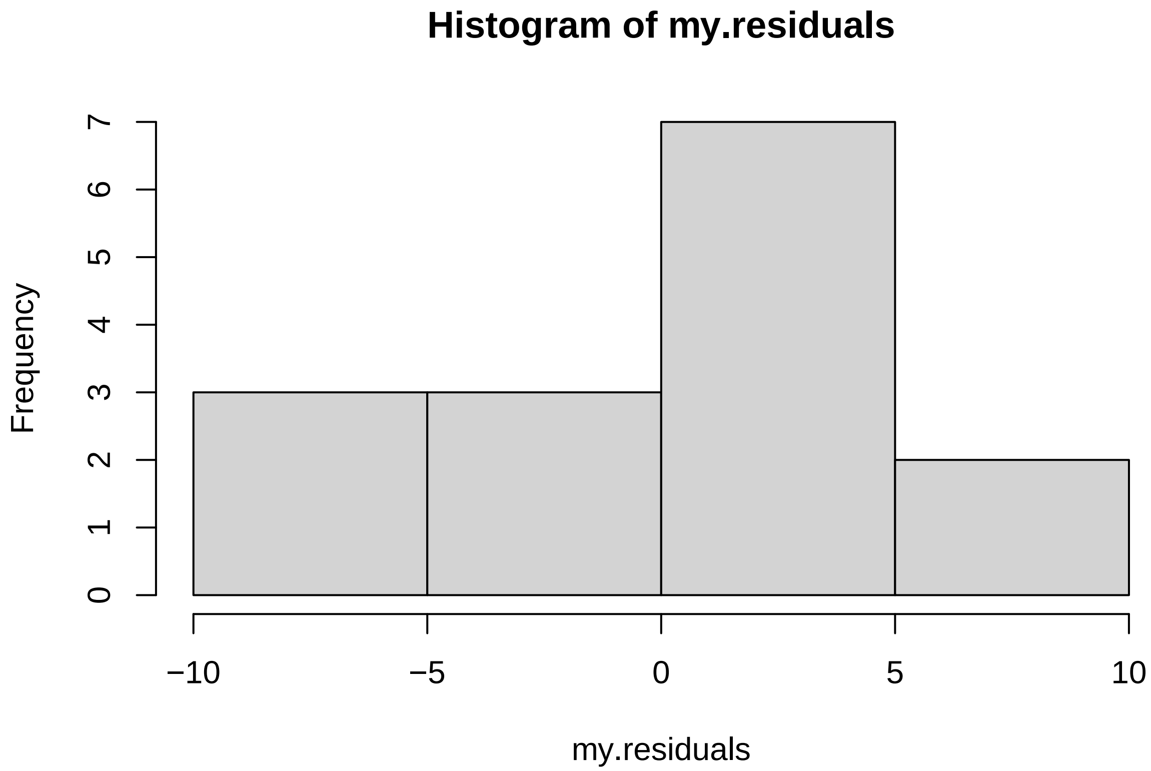
Let's quickly review what residuals are (see whiteboard exercise)

Let's look at the residuals for our sea turtle example:

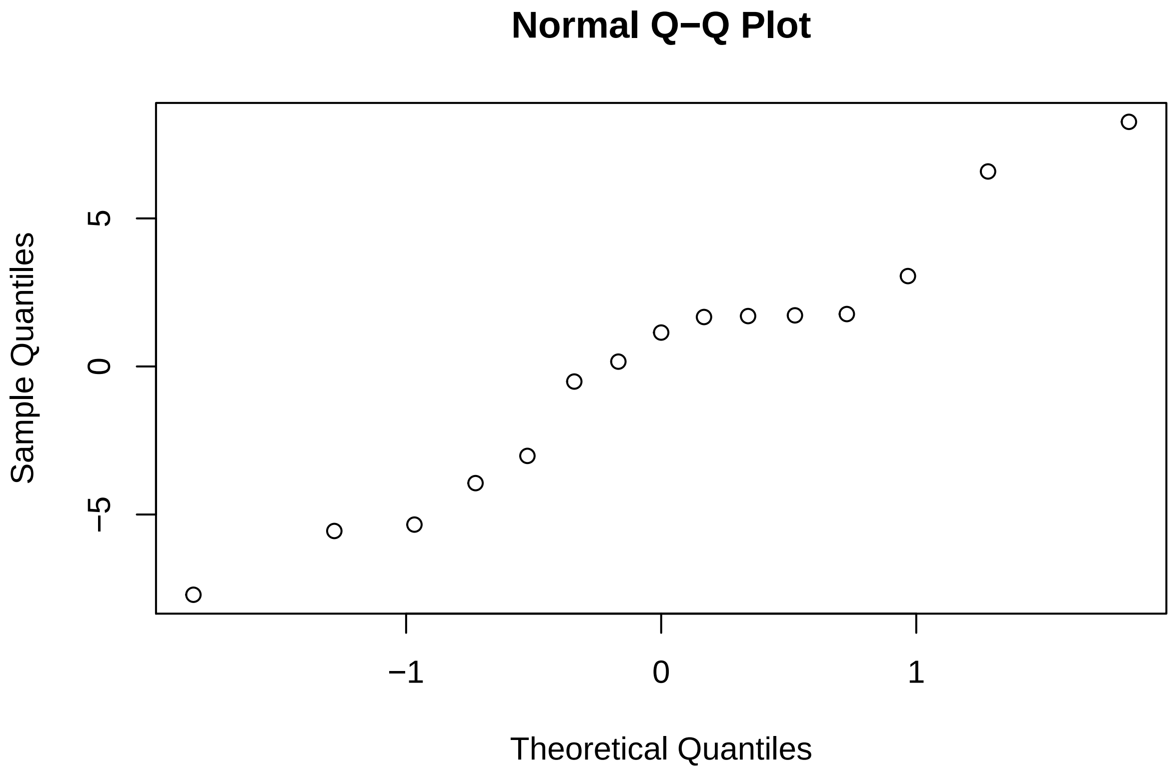
```r
my.intercept <- model$coefficients["(Intercept)"]
my.slope <- model$coefficients["light"]
expected.vals <- my.intercept+my.slope*light
my.residuals <- hatchlings.successful-expected.vals
my.residuals
```

```
##  [1]  1.7260254  8.2610947 -3.9392937  6.5852377  0.1637050 -3.0205154  3.0520308 -0.5097154
##  [9] -7.7091633 -5.3400802  1.7700709  1.7019382 -5.5562116  1.6697528  1.1451240
```

```r
### alternative way of getting residuals (best way!)

my.residuals2 <- model$residuals

### alternative way using predict function

my.residuals3 <- hatchlings.successful-predict(model)

### histogram of residuals

hist(my.residuals)
```



```
### test for normality

qqnorm(my.residuals)
```



```r
shapiro.test(my.residuals)
```

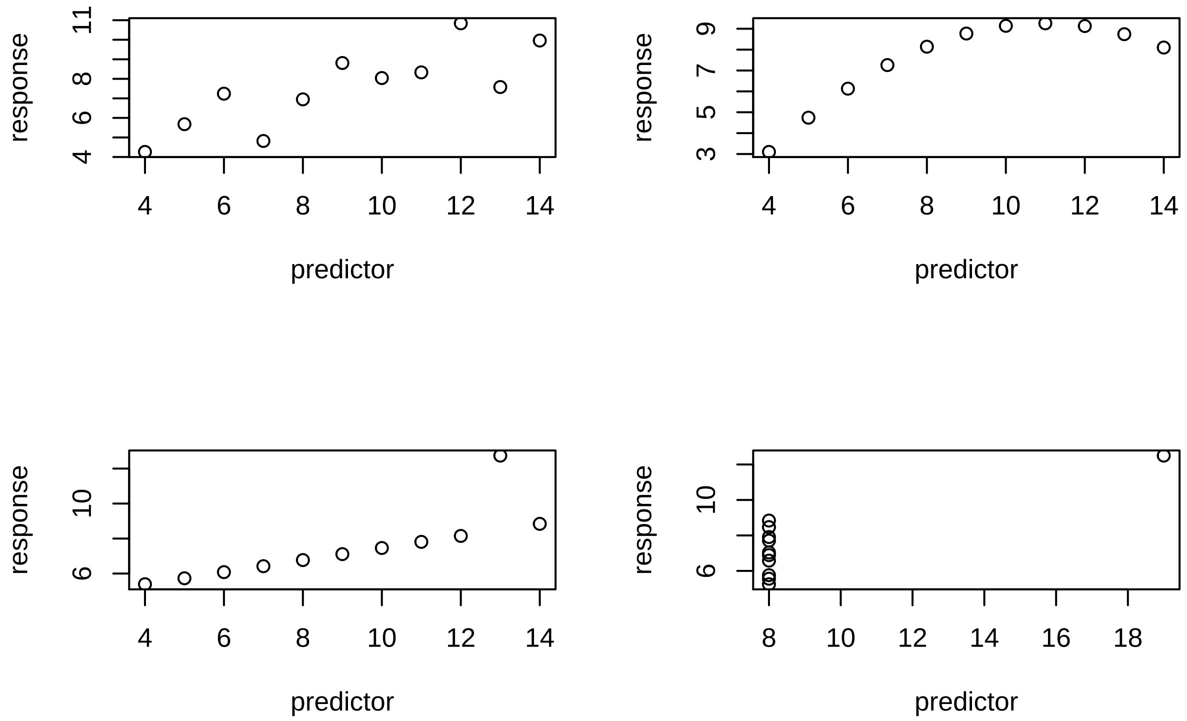
```
##
##  Shapiro-Wilk normality test
##
## data:  my.residuals
## W = 0.95736, p-value = 0.6467
```

Linear model

The true relationship between the response and predictor variables is linear!

Which one of the following plots violates this assumption?

```r
layout(matrix(1:4,nrow=2,byrow = T))
plot(anscombe$y1~anscombe$x1,ylab="response",xlab="predictor")
plot(anscombe$y2~anscombe$x2,ylab="response",xlab="predictor")
plot(anscombe$y3~anscombe$x3,ylab="response",xlab="predictor")
plot(anscombe$y4~anscombe$x4,ylab="response",xlab="predictor")
```



Independence of observations

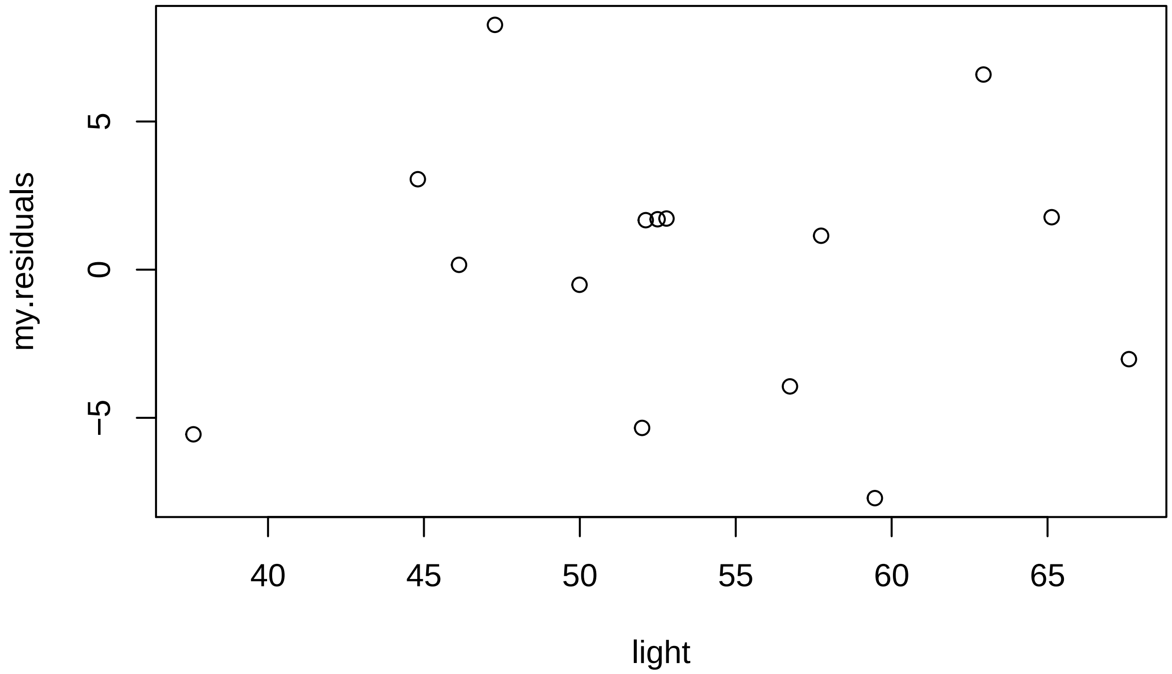
Of course! All classical analyses make this assumption!

Equal variance (homoscedasticity, or lack of heteroscedasticity)

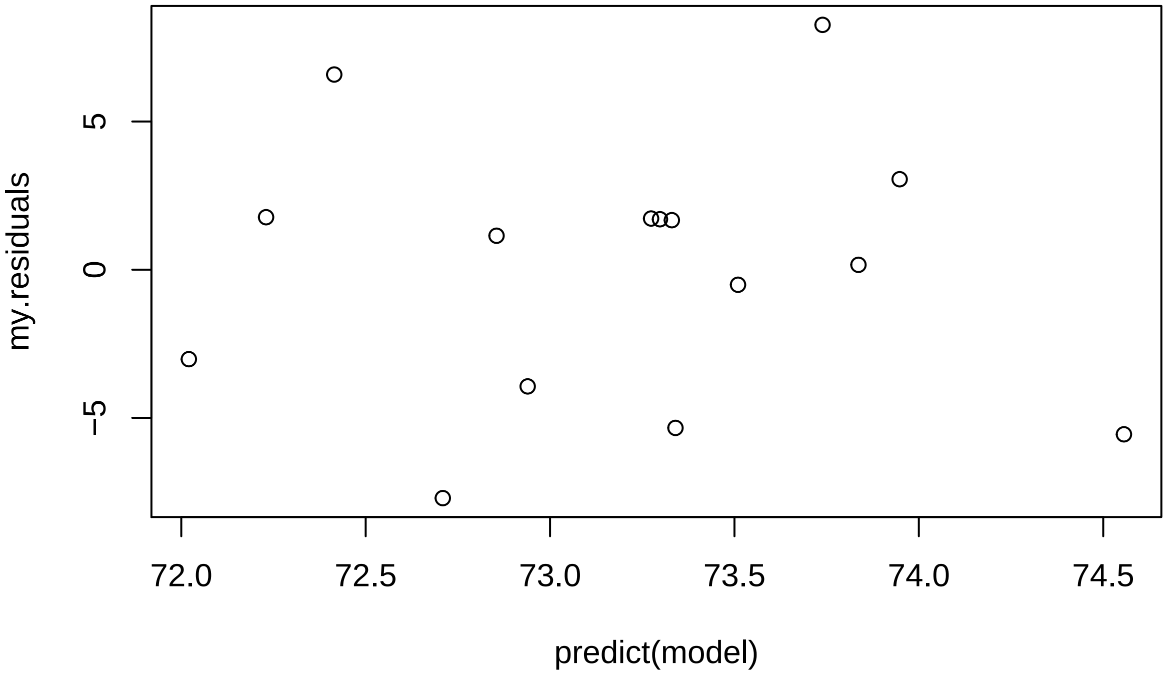
In simple linear regression, we assume that the model residuals are normally distributed– and that the spread of that distribution does not change as your predictor variable or response variable goes from small to large. That is, the residuals are **homoskedastic** – which means they have equal variance across the range of the predictor and response variables.

Let's look at the sea turtle example:

```r
my.residuals <- model$residuals

plot(my.residuals~light)
```



```
plot(my.residuals~predict(model))
```



Do you see any evidence for **heteroskedasticity**? Heteroskedasticity means non-homogeneity of variance across the range of the predictor variable. Serious violations of the equal variance assumption may warrant a **transformation** of your data, or you may choose to use an alternative analysis like a **generalized linear model**.

Predictor variable is known with certainty

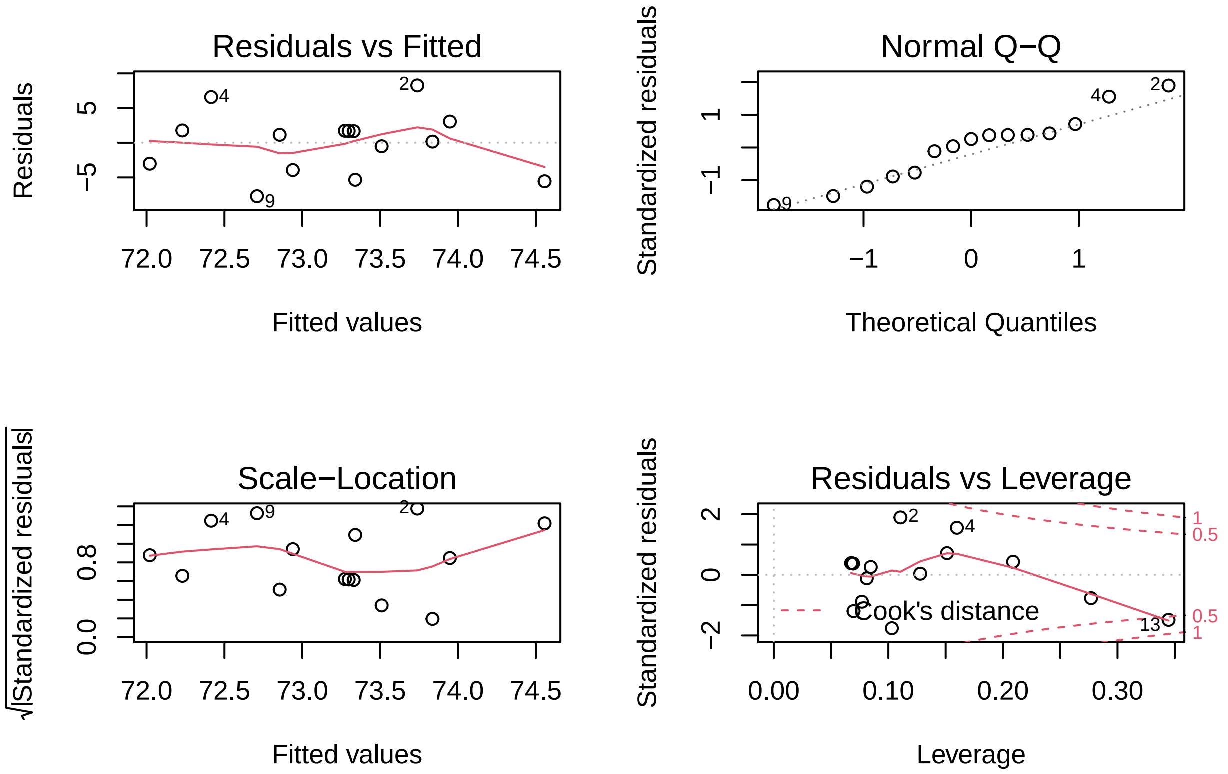
That is, the observed (e.g., measured) values for your predictor variable are correct and known with perfect precision. In regression, the randomness in linear regression is associated only with the response variable - and that randomness is normally distributed.

Diagnostic plots

Simple linear regression analyses are generally accompanied by 'diagnostic plots', which are intended to diagnose potential violations of key assumptions, or other potential pitfalls of regression.

When you use the 'plot' function in R to evalate a model generated with the 'lm' function, R returns four diagnostic plots:

```r
layout(matrix(1:4,2,byrow = T))
plot(model)
```



The first diagnostic plot (residuals vs fitted) lets you check for possible non-linear patterns. This plot should have no obvious pattern to it- it should look like random noise across the range of fitted values.

The second diagnostic plot (normal Q-Q) lets you check for normality of residuals. You already know how to do this.

The third diagnostic plot (scale-location) lets you check for homoskedascitiy. This plot should have no obvious pattern to it- it should look like random noise across the range of fitted values.

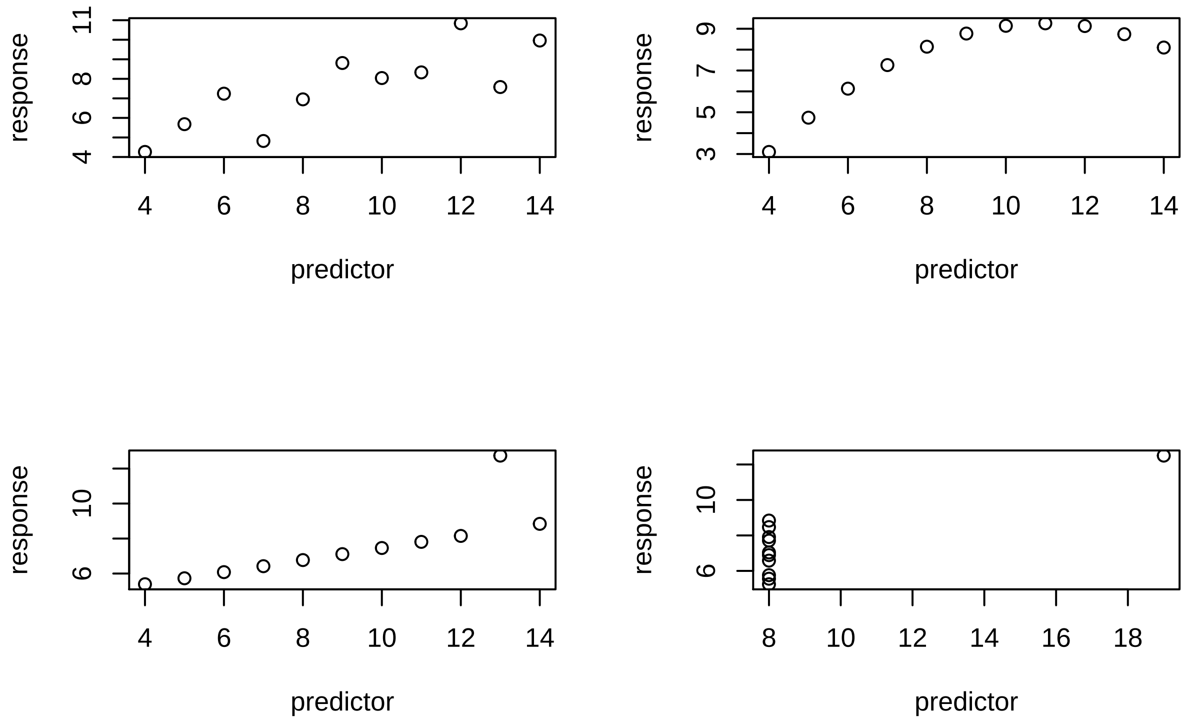
The fourth diagnostic plot (residuals vs leverage) lets you check to make sure your regression model isn't driven by a few **influential points**. Look for points that fall far to the right of the other points- these are high leverage points. Also look specifically at the upper and lower right hand side of this figure- points in these regions (with large values for "Cook's distance") have the property that if they were removed from the analysis the results would change substantially.

These four figures, taken together, should be a guide to interpreting how robust your regression is. You need to be the scientist. Analyze these plots and make decisions about what you're willing to accept.

Influential points

Which of the following plots has a highly influential point?

```
layout(matrix(1:4,nrow=2,byrow = T))
plot(anscombe$y1~anscombe$x1,ylab="response",xlab="predictor")
plot(anscombe$y2~anscombe$x2,ylab="response",xlab="predictor")
plot(anscombe$y3~anscombe$x3,ylab="response",xlab="predictor")
plot(anscombe$y4~anscombe$x4,ylab="response",xlab="predictor")
```



Can you try to draw a regression line if this point was removed?

A deeper exploration

R-squared

The coefficient of determination, also known as R-squared, is a statistic that is commonly used to indicate the performance of a regression model. Specifically, R-squared tells you how much of the total variance in your response variable is explained by your predictor variable.

R-squared can be computed as:

$R_2 = 1 - \frac{SS_{res}}{SS_{tot}}$

where $SS_{tot} = \sum(y_i - \bar{y})^2$ and $SS_{res} = \sum(y_i - y_{pred})^2$.

The maximum value for R-squared is 1- values close to 1 indicate a very "good" model!

Regression outcomes

Let's explore some possible outcomes of linear regression:

Try to come up with scenarios (with plots) for each of the following:

1. Non-significant p, high R2
2. Significant p/ low R2

3. Significant p/high R2
4. Non-significant p, low R2

–go to next lecture–