# Lab 1: Topics in Regression, Fall 2015
# We demonstrate the methods on the Galapagos Islands data. Using the least squares first:

```
> data(gala)
> g1<-lm(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> summary(g1)
```
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data = gala)

Residuals:
  Min     1Q Median     3Q    Max
-111.679 -34.898  -7.862  33.460 182.584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.068221  19.154198   0.369 0.715351
Area       -0.023938   0.022422  -1.068 0.296318
Elevation   0.319465   0.053663   5.953 3.82e-06 ***
Nearest     0.009144   1.054136   0.009 0.993151
Scruz      -0.240524   0.215402  -1.117 0.275208
Adjacent  -0.074805   0.017700  -4.226 0.000297 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-Squared: 0.7658,   Adjusted R-squared: 0.7171
F-statistic: 15.7 on 5 and 24 DF,  p-value: 6.838e-07

# Least squares works well when there are normal errors, but performs poorly for long-tailed errors.
# The Huber method is the default choice of the “rlm ()” function,
>
> ##Huber method for M-estimation
> library(MASS)
> gr<-rlm(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> summary(gr)

Call: rlm(formula = Species ~ Area + Elevation + Nearest + Scruz +
         Adjacent, data = gala)
Residuals:
       Min      1Q  Median      3Q     Max

Coefficients:
        Value     Std. Error   t value
(Intercept)  6.3611     12.3897     0.5134
Area       -0.0061      0.0145    -0.4214
Elevation    0.2476      0.0347     7.1320
Nearest     -0.3592      0.6819    -0.5267
Scruz     -0.1952      0.1393    -1.4013
Regression output:

\[
\begin{array}{l}
\text{Adjacent} & -0.0546 & 0.0114 & -4.7648 \\
\end{array}
\]

Residual standard error: 29.73 on 24 degrees of freedom

#The R2 and F-statistics are not given because they can not be calculated (at least not in the same way). Similarly, p-values are not given although we can use the asymptotic normality of the estimator to make appropriate inferences using the t-values. The numerical values of the coefficients have changed a small amount, but the general significance of the variables remains the same and our substantive conclusion would not be altered.

#We can also do LAD regression using the “quantreg” package. The default option does LAD while other options allow for quantile regression.

```r
> detach(gala)
> gq<-rq(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> summary(gq)
```

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Coefficients</th>
<th>lower bd</th>
<th>uppp bd</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.3144484</td>
<td>-19.8777769</td>
<td>24.374115</td>
</tr>
<tr>
<td>Area</td>
<td>-0.0030600</td>
<td>-0.031851</td>
<td>0.527999</td>
</tr>
<tr>
<td>Elevation</td>
<td>0.2321147</td>
<td>0.124526</td>
<td>0.501955</td>
</tr>
<tr>
<td>Nearest</td>
<td>0.1636558</td>
<td>-3.163385</td>
<td>2.988965</td>
</tr>
<tr>
<td>Scruz</td>
<td>-0.1231408</td>
<td>-0.479867</td>
<td>0.134763</td>
</tr>
<tr>
<td>Adjacent</td>
<td>-0.0518535</td>
<td>-0.104577</td>
<td>0.017394</td>
</tr>
</tbody>
</table>

#Again, there is some change in the coefficients. The confidence intervals now suggest that adjacent is not significant.
#For this example, we do not see any big qualitative difference in the coefficients and ofr want of evidence to the contrary, we might stick with least squares as the easiest to work with. Had we seen something different, we would need to find out the cause. Perhaps some group of observations were not being fit well and the robust regression excluded these points.

Another popular method is least trimmed squares (LTS). This method has a high breakdown point because it can tolerate a large number of outliers depending on how $q$ is chosen. The huber and L1 methods will fail if some epsilon goes inifty. LTS is an example of a resistant regression method. Resistant methods are good for dealing with data where we expect there to be a certain number of bad observations that we want to have no weight in the analysis.

```r
> g<-ltsreg(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> coef(g)
   (Intercept)       Area       Elevation       Nearest       Scruz       Adjacent
10.58601026  1.52482792  0.01966708  0.64976689 -0.08244073 -0.17039517
> g<-ltsreg(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> coef(g)
   (Intercept)       Area       Elevation       Nearest       Scruz       Adjacent
  8.83938017  1.54437983  0.02196670  0.76490136 -0.09526354 -0.19781076
> g<-ltsreg(Species ~Area+Elevation+Nearest+Scruz+Adjacent, gala)
> coef(g)
   (Intercept)       Area       Elevation       Nearest       Scruz       Adjacent
  9.22463773  1.58979518  0.02110958  0.80342306 -0.11680384 -0.20332699
```
I repeated the command three times and you will notice that the results are somewhat different. This is because the default genetic algorithm used to compute the coefficient is nondeterministic. An exhaustive search method can be used;