Introduction: Mathematically, regression is like ANOVA. The result is a ratio of two variances using the F test. If the F ratio is so large that it is “improbable”, then we reject the null hypothesis. The big difference between regression and ANOVA is that the independent variable in ANOVA is a category. In the example for ANOVA, the engineer had four pressure settings. In regression, the independent variable is not (and often can not be) controlled.

Example:
What variables affect the mileage of a car? The design of the engine, the number of cylinders, the weight. We decide to find out about the effect of weight on miles per gallon. We collected the following data on 10 cars.

<table>
<thead>
<tr>
<th>X: Weight (1000 lbs.)</th>
<th>2.5</th>
<th>3.0</th>
<th>4.0</th>
<th>3.5</th>
<th>2.7</th>
<th>4.5</th>
<th>3.8</th>
<th>2.9</th>
<th>5.0</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y: Miles per gallon</td>
<td>20</td>
<td>21.5</td>
<td>15</td>
<td>17.5</td>
<td>21</td>
<td>10</td>
<td>16</td>
<td>19.5</td>
<td>7.5</td>
<td>22</td>
</tr>
</tbody>
</table>

The numbers on the graph are the observation numbers. As an example, “10” indicated the tenth element of the sample $X_{10}$, not ten observations.

1 Simple Linear Regression
PURPOSE. The purpose of simple linear regression is to find an equation for a linear relationship between an independent variable $X$ and a dependent variable $Y$.

MODEL FOR THE REGRESSION OF $Y$ ON $X$. The model predicts the value of the dependent variable $Y$ from the value of the independent variable $X$:

$$y = \alpha + \beta x + \varepsilon,$$

where $\varepsilon$ is the only random variable in the equation.

Warning: note that $\alpha$ is not the confidence level.

ASSUMPTIONS.
1. $X$ values are assumed to be fixed and measured with minimal error.
2. The relationship between $X$ and $Y$ is linear.
3. $Y$ values at each value of $X$ are assumed to be independent and to be from normally distributed subpopulations with equal variances. From this, $\varepsilon$ (in the model) is normally distributed with mean 0 and variance $\sigma^2$.

LINE is an often-used mnemonic (Linear, Independent, Normal, Equal variances).
Sample regression line
The sample regression equation is obtained by the least squares method:
\[ \hat{y}_i = a + bx_i, \]
where \( b = \beta \) is the estimate of the population slope \( \beta \) and \( a = \alpha \) is the estimate of the population y-intercept.

The Method of Least Squares: Deriving the Coefficients \( a \) and \( b \)
We find \( a \) and \( b \) by the method of least squares which minimizes the sum of the squared deviations. The deviation for the \( i \)th observation is the difference between an observed \( y_i \) value and the predicted value \( \hat{y}_i \): \( e_i = y_i - \hat{y}_i \). This difference or error, \( e_i = y_i - \hat{y}_i \), is the vertical distance from the observed \( y \) value to the regression line.

The sum of the squared deviations is
\[
\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2
\]

As the term least squares implies, we find the minimum of this sum by taking the partial derivatives with respect to \( a \) and \( b \). When we do this, we obtain the Normal Equations:
\[
2 \sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0
\]
\[
2 \sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0
\]

Solving the Normal Equations for \( a \) and \( b \), we obtain,
\[
b = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad a = \bar{y} - b \bar{x}
\]

The following formulae include the basic “intuitive” formulae and equivalent formulae that are more efficient for computation.
Math 308 SP 2020: Introduction to simple linear regression and correlation

$$SS_{xx} = \sum_{k=1}^{n} (x_k - \bar{x})^2 = \sum_{k=1}^{n} x_k^2 = \frac{\left(\sum_{k=1}^{n} x_k\right)^2}{n} = \frac{\sum_{k=1}^{n} x_k^2 - n\bar{x}^2}{n}$$

$$SS_{yy} = \sum_{k=1}^{n} (y_k - \bar{y})^2 = \sum_{k=1}^{n} y_k^2 = \frac{\left(\sum_{k=1}^{n} y_k\right)^2}{n} = \frac{\sum_{k=1}^{n} y_k^2 - n\bar{y}^2}{n}$$

$$SS_{xy} = \sum_{k=1}^{n} (y_k - \bar{y})(x_k - \bar{x}) = \sum_{k=1}^{n} x_k y_k - \left(\frac{\sum_{k=1}^{n} x_k}{n}\right) \left(\frac{\sum_{k=1}^{n} y_k}{n}\right) = \sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}$$

**Example: Weight and MPG**

a = 34.8534  
b = -5.2356

So our best fitting line is:  
$$\hat{y} = 34.8534 - 5.2356x$$

2. Hypothesis testing: Evaluating the Regression Equation

1. Hypotheses are two-tailed:
   - H₀:  β = 0  (there is no linear relationship between y and x)
   - Hₐ:  β ≠ 0  (there is a linear relationship between y and x)

2. The F statistic with 1 and n – 2 degrees of freedom is used to test the null hypothesis at the stated significance level α.

3. The test for rejecting H₀ is a one tailed test: we reject H₀ when the F(obtained) > Fₐ.

4. The F(obtained) is calculated by the following ANOVA method.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Variance Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>MSR = SSR/1</td>
<td>F = MSR/MSE</td>
</tr>
<tr>
<td>Residual</td>
<td>n - 2</td>
<td>SSE</td>
<td>MSE = s² = SSE/n - 2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n - 1</td>
<td>SST</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Residual sum of squares:

$$SSE = \sum_{k=1}^{n} e_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 = \sum_{k=1}^{n} (y_k - a - b x_k)^2 = SS_{yy} - SS_{xy}/SS_{xx}$$

Total Sum of Squares: $SST = \sum_{k=1}^{n} (y_k - \bar{y})^2 = SS_{yy}$

Sum of Squares Regression: $SSR = SST - SSE$

Intuitive Understanding: The F ratio is (the variance in Y accounted for by the model based on X) divided by (the variance not explained by model based on values of X). The null hypothesis is that there is no relation between Y and X. So, when the null hypothesis is true, the numerator should not be very large because the model does not
account for any variability in Y. But, when the null hypothesis is false (and there is a relationship between X and Y), the numerator should be large and, consequently, the value of the obtained F will be large and, consequently again, we will be more likely to reject the null hypothesis.

**Example: Weight and MPG**

1. **Hypotheses:**
   - H₀: \( \beta = 0 \)
   - Hₐ: \( \beta \neq 0 \)

2. If \( \alpha = 0.05 \), the F statistic with 1 and 8 degrees of freedom is 5.32 from p. 589.

3. **Reject H₀ when F(obtained) > 5.32.**

4. **Computation of F(obtained).**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>Variance Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>204.188</td>
<td>204.188</td>
<td>97.166</td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>16.812</td>
<td>2.101</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. **Reject H₀.** We conclude that \( \beta \neq 0 \) and that there is a statistically significant relationship between Weight and MPG.

**3. Simple Linear Regression in R of MPH on Weight for the Car data**

```r
> # Reading in the data to a data frame named cars.
> cars=data.frame(
+   xweight=c(2.5,3.0,4.0,3.5,2.7,4.5,3.8,2.9,5.0,2.2),
+   ymph=c(20,21.5,15,17.5,21,10,16,19.5,7.5,22)
+ )
> #
> # Loading the the add-on package "lattice" which has the scatterplot routine xyplot
> library(lattice)
> xyplot(ymph~xweight, data=cars,
+   xlab="Weight in 1000 pound units",
+   ylab="Miles per hour",
+   main="Car data: mph on weight"
+ )
```
```r
> # Executing the regression analysis
> car.mod1=lm(ymph~xweight, data=cars)
> summary(car.mod1)

Call:
  lm(formula = ymph ~ xweight, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76440 -1.26374  0.05628  1.02421  2.35340

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.8534     1.8683  18.655 7.04e-08 ***
xweight     -5.2356     0.5311 -9.857 9.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.45 on 8 degrees of freedom
Multiple R-squared:  0.9239 ,  Adjusted R-squared:  0.9144
F-statistic: 97.17 on 1 and 8 DF,  p-value: 9.449e-06
```

> # Generating the ANOVA table for the regression
> anova(car.mod1)
Analysis of Variance Table

<p>| Response: ymph |</p>
<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>xweight</td>
<td>1</td>
<td>204.188</td>
<td>204.188</td>
<td>97.166</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>16.812</td>
<td>2.101</td>
<td></td>
</tr>
</tbody>
</table>

---

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

4. Correlation

The correlation coefficient is related to the regression coefficient and is a single number that can be used to measure the degree to which the points of the scatterplot clustered about a straight line. If $r = -1$ or $r = +1$, the points plotted in a scatterplot lie on a straight line of either positive or negative slope; therefore, in such a case, the relationship is perfectly linear and apparently without a random component. Values of $r$ between $+1$ and $-1$ indicate lesser degrees of clustering about a straight line pattern, with value of $r$ near zero indicating little or no discernable linear pattern to the points.

Sample Correlation Coefficient. The formula for the correlation coefficient in terms of sums of squares is:

$$ r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} $$

Assumptions: Primarily bivariate normal distribution

H$_0$: $\rho = 0$ (there is no linear relationship between $y$ and $x$)

H$_A$: $\rho \neq 0$ (there is a linear relationship between $y$ and $x$)

Test of rho equaling other non-zero values are also possible.

One important aspect of the correlation coefficient: $r^2$ is the proportion of the total variance explained by the regression.

Warning: You cannot infer causation from correlation alone. The correlation does not reflect the units of measurement whereas the regression coefficient $b$ does reflect the units of measurement.