Simple Linear Regression  Here are key things to recall.

- Regression line minimizes sum of squared residuals.
- Response $Y$ and explanatory variable $X$ are treated differently; analysis is made conditional on the values of $X$.
- Slope is $\hat{\beta}_1 = r \frac{s_y}{s_x}$.
- Intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.
- Assumptions are (LINC):
  1. Linearity ($E(Y \mid X) = \beta_0 + \beta_1 X$,)
  2. Independent (deviations around true regression line are independent of each other and $x$ values).
  3. Normality (distribution of deviations around true regression line are normally distributed).
  4. Constant Variance (same standard deviation of distribution of deviations around the mean for all $X$).
- Know how to find confidence intervals and to test hypotheses for the intercept and slope.
- Know how to find confidence intervals and prediction intervals for $\mu_{Y \mid X}$ and $\hat{Y}$ respectively.
- Know how to interpret computer output about regression.
- Beware of extrapolation! (Linear relationship may fit data okay, but trend may not extend beyond range of $X$ data.)
- Graphical examination of residuals can help detect deviations from assumptions.

**Example problem:** Use first round score at Masters to predict final score at end of the tournament. Data is a sample of 20 golfers from the 2011 tournament.

```r
require(ggplot2)
require(Lock5Data)
data(MastersGolf)
ggplot(MastersGolf, aes(x=First,y=Final)) + geom_point() +
  geom_smooth(method="lm",se=FALSE)
```
Summary output.

```r
fit = lm(Final ~ First, data=MastersGolf)
summary(fit)

##
## Call:
## lm(formula = Final ~ First, data = MastersGolf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##    -6.259   -2.113    0.076   1.338    9.314
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.162    0.817    0.20   0.85
## First        1.476    0.262    5.64  2.4e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.6 on 18 degrees of freedom
## Multiple R-squared:  0.638, Adjusted R-squared:  0.618
## F-statistic: 31.8 on 1 and 18 DF,  p-value: 2.38e-05
```

```r
with( MastersGolf, c(mean(First),sd(First),mean(Final),sd(Final),cor(First,Final)) )
```

```
## [1] -0.550  3.154  -0.650  5.824  0.799
```
Questions:
1. Write the regression line.
2. Test if \( \beta_1 = 0 \) versus the alternative \( \beta_1 \neq 0 \). Explain in context.
3. What is the predicted final score of a golfer that shoots even par (score equals 0) in the first round? Answer with a 95% prediction interval.
4. How is a confidence interval for the mean final score of all golfers that shoot even par in the first round different?
5. Comment on the linearity, normality, and constant variance assumptions.

ANOVA  Here are key things to recall.

• (One-way) ANOVA is used when there is a quantitative response variable and a single categorical explanatory variable.

• The ANOVA table is a structured form used to compute a test statistic.

• The null hypothesis is that all population means are equal; the alternative hypothesis is that at least one is different from the rest—they are not all equal.

• The idea behind the test is to compare variation among sample means to variation within samples. If the sample means vary more than can be explained by chance (as measured by variation within samples), then there is evidence that the null hypothesis is false.
• The total sum of squares may be partitioned into a sum of squares among group means and a sum of squares within groups (error). (In this notation, let $i$ range over all $n$ observations, let $j[i]$ be the group of the $i$th observation, let $k$ be the number of groups, and let $n_j$ be the number of observations in the $j$th group. In addition, let $\bar{y}$ be the grand mean of all $n$ observations and let $\bar{y}_j$ and $s_j$ be the mean and standard deviation of the observations in the $j$th group.)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k} (n_j - 1)s_j^2$$

(Total Sum of Squares) = (Group Sum of Squares) + (Error Sum of Squares)

• In the ANOVA table, the MSE (mean square error) is an estimate of individual variance, so $\sqrt{\text{MSE}}$ is an estimate of $\sigma$, the standard deviation of the distribution of how individual measurements differ from their population means.

• In the ANOVA table, each mean square (MS) is found by taking the ratio between a sum of squares (SS) and a degrees of freedom value (df).

• The $F$ statistic is a ratio of mean squares.

• Each row of the ANOVA table has degrees of freedom. If there are $k$ groups and $n$ total observations, there are $k - 1$ degrees of freedom for the groups, $n - k$ degrees of freedom for error (residuals in R), and $n - 1$ total degrees of freedom.

• Model assumptions for inference include normal deviations from means for individual observations and constant variance across groups.

• After fitting an ANOVA model, there are usually scientifically interesting questions about comparing specific groups which can be tested or estimated.

• These following tests use an estimate of $\sigma$ based on data from all samples which changes the degrees of freedom and SE estimate as compared to two-sample problems.

• Beware of the issue of multiple testing.

**Example Problem**  Cuckoo birds have a behavior in which they lay their eggs in other birds nests. The other birds then raise and care for the newly hatched cuckoos. Cuckoos return year after year to the same territory and lay their eggs in the nests of a particular host species. Furthermore, cuckoos appear to mate only within their territory. Therefore, geographical sub-species are developed, each with a dominant foster-parent species. A general question is, are the eggs of the different sub-species adapted to a particular foster-parent species? Specifically, we can ask, are the mean lengths of the cuckoo eggs the same in the different sub-species?

Here is a graph of the table.
Here is the ANOVA analysis and table.

```r
fit = lm(eggLength ~ hostSpecies, data = cuckoo)
anova(fit)
```

```text
## Analysis of Variance Table
##
## Response: eggLength
## Df    Sum Sq Mean Sq F value    Pr(>F)
## hostSpecies 5     42.9  8.590 10.400 3.2e-08 ***
## Residuals 114     94.2  0.825
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual plots are also useful for ANOVA analyses.

```r
d = data.frame(residuals = residuals(fit), hostSpecies=cuckoo$hostSpecies)
ggplot(d, aes(x=hostSpecies,y=residuals)) +
  geom_point(position=position_jitter(w=0.2,h=0)) +
  geom_hline(yintercept=0)
```
Some more summary data.

```r
n = with(cuckoo, as.vector(by(eggLength, hostSpecies, length)))
m = with(cuckoo, as.vector(by(eggLength, hostSpecies, mean)))
s = with(cuckoo, as.vector(by(eggLength, hostSpecies, sd)))
out = rbind(n, m, s)
rownames(out) = c("Sample Size", "Sample Mean", "Sample SD")
colnames(out) = levels(cuckoo$hostSpecies)
print(round(out, 2))
```

<table>
<thead>
<tr>
<th></th>
<th>HedgeSparrow</th>
<th>MeadowPipet</th>
<th>PiedWagtail</th>
<th>Robin</th>
<th>TreePipet</th>
<th>Wren</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Size</strong></td>
<td>14.00</td>
<td>45.00</td>
<td>15.00</td>
<td>16.00</td>
<td>15.00</td>
<td>15.00</td>
</tr>
<tr>
<td><strong>Sample Mean</strong></td>
<td>23.12</td>
<td>22.30</td>
<td>22.90</td>
<td>22.57</td>
<td>23.09</td>
<td>21.13</td>
</tr>
<tr>
<td><strong>Sample SD</strong></td>
<td>1.07</td>
<td>0.92</td>
<td>1.07</td>
<td>0.68</td>
<td>0.90</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Questions:

6. State the null and alternative hypotheses.

7. Interpret the result of the test in context.

8. What is the pooled estimate for the standard deviation of egg length?

9. Find a confidence interval for the difference in mean egg length for cuckoo bird eggs laid in robin and wren nests.

10. Comment on data conforming to model assumptions.

**Categorical Data Analysis**  Here are some key ideas.

- We have examined two situations: (1) do the counts from a single variable match expected values; and (2) are two categorical variables independent?
• In each case, inference is based on a test statistic that compares observed and expected counts in each “cell” of a table and totals the measure of discrepancy by summing over all cells.

• Expected counts in each cell are the total sample size times the null probability that an observation falls into the cell.

• For a single categorical variable, the null hypothesis specifies a probability \( p_i \) for each cell and the expected count is \( np_i \).

• For two categorical variables, one has probabilities \( a_1, \ldots, a_r \) and the other has probabilities \( b_1, \ldots, b_c \). Under the null hypothesis of independence, the probability that an observation is in categories \( i \) from the first and \( j \) from the second is \( p_{ij} = a_i b_j \). Individual probabilities are estimated by row (or column) total over the total sample size. Hence, the expected counts are found by

\[
np_{ij} = na_i b_j = n \times \left( \frac{n_i}{n} \right) \times \left( \frac{n_j}{n} \right) = \frac{n_i n_j}{n}
\]

where \( n_i \) is the total in the \( i \)th row, \( n_j \) is the total in the \( j \)th column and \( n \) is the total sample size.

• The traditional formula is

\[
X^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}
\]

• A potentially better test statistic is based on a likelihood ratio test.

\[
G = \sum_i \text{Observed}_i \ln \left( \frac{\text{Observed}_i}{\text{Expected}_i} \right)
\]

• The test statistic is compared to a chi-square distribution with some number of degrees of freedom.

• For a single categorical variable with \( k \) groups, there are \( k - 1 \) degrees of freedom.

• For testing the relationship between two categorical variables with \( r \) and \( c \) groups, the degrees of freedom is \( (r - 1)(c - 1) \).

• The p-value is the area to the right of the test statistic under a chi-square density with the appropriate number of degrees of freedom.

Example Problem  When asked to identify a favorite Skittles candy, is there a difference in response proportions when asked to pick a favorite by color (Green, Orange, Purple, Red, or Yellow) or by flavor (Lime, Orange, Grape, Strawberry, Lemon)?

Here is some data.
x = matrix(c(18,13,9,16,15,19,13,34,11,9),nrow=2,ncol=5)
rownames(x) = c("Color","Flavor")
colnames(x) = c("Green/Lime","Orange","Purple/Grape","Red/Strawberry", "Yellow/Lemon")
print(x)

Compute row and column totals and expected counts.

row.sums = apply(x, 1, sum)
col.sums = apply(x, 2, sum)
n = sum(x)
expected = (row.sums %o% col.sums)/n
print(expected)

Questions:
11. Find the observed proportions for each category by color or flavor. Where are the largest differences?
13. Explain how to find a p-value for the test.
14. Interpret the result of the test in context.
Solutions to all problems

1. 
\[(\text{Final Score}) = 0.162 + 1.476(\text{Score in First Round})\]
where scores are relative to par.

2. \(t = 5.64\) corresponds to a two-sided p-value of about \(2.4 \times 10^{-5}\) from a \(t\) distribution with 18 degrees of freedom, which is very strong evidence against the null hypothesis. In context:

There is extremely strong evidence that final scores relative to par in the Masters tournament are related to scores in the first round of the tournament.

3. The predicted score of a golfer who shoots par in the first round is 0.162. This could be interpreted that most probable score is even par and that on average, the final scores will be a little higher than par. To complete a 95% prediction interval, we would need the information to find

\[t^*s\sqrt{1 + \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}\]
where \(t^*\) is the 0.975 quantile from a \(t\) distribution with 18 degrees of freedom, \(s = 3.6\) is the estimated residual standard error \(\sqrt{RSS/(n-2)}\), and we would need to know the mean and sd of the first round scores.

4. The confidence interval will be narrower. Drop the \(1+\) under the square root from the prediction error formula.

5. There is no pattern of +/- residuals to suggest nonlinearity. There is no fan shape pattern of residuals to suggest a consistent trend in variation with changes in \(x\). There are no extreme outliers or skewness to suggest strong nonnormality.

6. The null hypothesis is that all six population means are equal. The alternative is that at least one is different.

7. The tiny p-value allows one to conclude that there is overwhelming evidence that the mean size of cuckoo bird eggs laid in nests of different species are different.

8. The pooled estimate for standard deviation of bird egg length is the square root the of the MSE.

\[\sqrt{0.83} = 0.91\ mm\]

This numerical value must be within the range of the sample sds as its square (the variance) is a weighted average of the sample variances, weighted by degrees of freedom.

9. The confidence interval will be a point estimate plus or minus a margin of error where the \(t^*\) multiplier will be based on degrees of freedom pooled from all samples.

\[(22.57 - 21.13) \pm (1.98)\sqrt{0.83}\sqrt{\frac{1}{16} + \frac{1}{15}}\]
10. The residual plot shows data that is fairly evenly spread within each sample with similar standard deviations. Assuming good representative sampling, inference is valid.

11. Here are observed proportions within each row.

```r
y = x
y[1, ] = x[1, ]/sum(x[1, ])
y[2, ] = x[2, ]/sum(x[2, ])
print(round(y, 3))
```

<table>
<thead>
<tr>
<th></th>
<th>Green/Lime</th>
<th>Orange</th>
<th>Purple/Grape</th>
<th>Red/Strawberry</th>
<th>Yellow/Lemon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.273</td>
<td>0.136</td>
<td>0.227</td>
<td>0.197</td>
<td>0.167</td>
</tr>
<tr>
<td>Flavor</td>
<td>0.143</td>
<td>0.176</td>
<td>0.209</td>
<td>0.374</td>
<td>0.099</td>
</tr>
</tbody>
</table>

The largest difference is the proportion that select red is much smaller than the proportion that select strawberry.

12. The null hypothesis is that the proportions selecting each skittle color/flavor will be the same when requesting by color as when selecting by flavor. The alternative is some difference among these probabilities will be true.

13. The p-value will be the area to the right of the test statistic under a chi-square density with 4 degrees of freedom. This is the area to the right of 9.069 which is about 0.06.

14. There is marginal evidence that people’s stated choice of favorite Skittles depends on whether the question is asked by color or flavor.