Simple Linear Regression
- Example - mammals
- Response variable: gestation (length of pregnancy) days
- Explanatory: brain weight

“Man”
- Extreme negative residual but that residual is not statistically significant.
- The extreme brain weight of “man” creates high leverage that is statistically significant.

“Man”
- Is the point for “Man” influencing where the simple linear regression line is going?
- Is this influence statistically significant?
Simple Linear Regression

- Predicted Gestation = 85.25 + 0.30*Brain Weight
- $R^2 = 0.372$, so only 37.2% of the variation in gestation is explained by the linear relationship with brain weight.

Exclude “Man”

- What happens to the simple linear regression line if we exclude “Man” from the data?
- Do the estimated intercept and estimated slope change?
Simple Linear Regression
- Predicted Gestation = 62.05 + 0.634*Brain Weight
- \( R^2 = 0.600 \), 60% of the variation in gestation is explained by the linear relationship with brain weight.

Changes
- The estimated slope has more than doubled once “Man” is removed.
- The estimated intercept has decreased by over 20 days.
Influence

- It appears that the point associated with “Man” influences where the simple linear regression line goes.
- Is this influence statistically significant?

Influence Measures

- Quantifying influence involves how much the point differs in the response direction as well as in the explanatory direction.
- Combine information on the residual and the leverage.

Cook’s D

\[ d = \left( \frac{h}{p+1} \right) \left( \frac{z}{(1-h)} \right)^2 \]

- where \( z \) is the standardized residual and \( p \) is the number of explanatory variables in the model.
Cook’s D

- If \( d > 1 \), then the point is considered to have high influence.

Cook’s D for “Man”

\[
d = \left( \frac{h}{p+1} \right) \left( \frac{z}{1-h} \right)^2
\]

\[
d = \left( \frac{0.6612}{2} \right) \left( \frac{-2.516}{1-0.6612} \right)^2
\]

\[
d = 18.23
\]
Cook’s D

- There are no other mammals with a value of \( d \) greater than 1.
- The okapi has \( d = 0.30 \)
- The Brazilian Tapir has \( d = 0.10 \)

Studentized Residuals

- The studentized residual is the standardized residual adjusted for the leverage.

\[
    r_s = \frac{z}{\sqrt{1 - h}}
\]

<table>
<thead>
<tr>
<th></th>
<th>( z )</th>
<th>( h )</th>
<th>( r_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>3.010</td>
<td>0.0217</td>
<td>3.043</td>
</tr>
<tr>
<td>“Man”</td>
<td>-2.516</td>
<td>0.6612</td>
<td>-4.323</td>
</tr>
<tr>
<td>Okapi</td>
<td>2.443</td>
<td>0.0839</td>
<td>2.552</td>
</tr>
</tbody>
</table>
Studentized Residuals

- If the conditions for the errors are met, then studentized residuals have an approximate $t$-distribution with degrees of freedom equal to $n - p - 1$.

Computing a P-value

- **JMP - Col - Formula**
- $(1 - t \text{ Distribution}(|r_s|, n-p-I)) \times 2$
- For our example
  - $r_s = 3.043, n-p-I=48$
  - P-value = 0.0038

<table>
<thead>
<tr>
<th></th>
<th>$z$</th>
<th>$h$</th>
<th>$r_s$</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>3.010</td>
<td>0.0217</td>
<td>3.043</td>
<td>0.0038</td>
</tr>
<tr>
<td>“Man”</td>
<td>-2.516</td>
<td>0.6612</td>
<td>-4.323</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Okapi</td>
<td>2.443</td>
<td>0.0839</td>
<td>2.552</td>
<td>0.0139</td>
</tr>
</tbody>
</table>
Conclusion - “Man”
- The P-value is much less than 0.001 (the Bonferroni corrected cutoff), therefore “Man” has statistically significant influence on where the regression line is going.

Other Mammals
- The Brazilian Tapir has the most extreme standardized residual but not much leverage and so is not influential according to either Cook’s D or the Studentized Residual value.

Other Mammals
- The Okapi has high leverage, greater than 0.08, but it’s standardized residual is not that extreme and so is not influential according to either Cook’s D or the Studentized Residual value.