Unusual Points

- Outlier box plot identifies potential outliers.
- How do we determine if a potential outlier identified on the box plot is statistically significant?

Unusual Points in Regression

- Outlier – a point with an unusually large residual.
- High leverage point – a point with an extreme value for one, or more, of the explanatory variables

Influential Points

- Does a point influence where the regression line goes?
  - An outlier can.
  - A high leverage point can.
- Is that point statistically significant in terms of influence?
Simple Linear Regression

- Example - 50 mammals
- Response variable: gestation (length of pregnancy) days
- Explanatory: brain weight

Distributions

Gestation

- Skewed to the right.
- Several potential outliers.
- Mean = 117.4 days
- Median = 65.5 days
- Values from 12 days to 440 days.
Brain Weight

- Highly skewed to the right with several mounds.
- Six potential outliers.
- Mean = 107.25 g
- Median = 16.25 g
- Values from 0.14 g to 1320 g

Simple Linear Regression

- Trying to explain variation in the response (gestation) by relating the response to the explanatory variable (brain weight).
Regression Residuals

\[ \text{residual} = y - \hat{y} \]

- Those observations that do not follow the general trend will have residuals that are far from zero, either positive or negative.

Regression Outlier

- A residual far from zero, either negative or positive, will be called an outlier for regression.
- An outlier for regression corresponds to a value of the response that does not match the overall trend.

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.
- RMSE = 85.1 days
Simple Linear Regression

- The model is useful.
- $F = 28.49$, $P$-value < 0.0001
- This also indicates that there is a statistically significant linear relationship between brain weight and gestation.

Unusual Points

- The mammal with a brain weight around 1300 g has the residual furthest from zero on the negative side.
- There are other mammals with residuals of the same magnitude on the positive side.
Outlier Box Plot

- Start with five number summary
  - Minimum = -214.1
  - 25% Quartile = -57.9
  - 50% Median = -31.1
  - 75% Quartile = 36.7
  - Maximum = 256.1

InterQuartile Range (IQR)

- IQR = 75% Quart - 25% Quart
  - IQR = 36.7 - (-57.9) = 94.6
- Upper = 75% Quart + 1.5*IQR
  - Upper = 36.7 + 141.9 = 178.6
- Lower = 25% Quart - 1.5*IQR
  - Lower = -57.9 - 141.9 = -199.8

Outlier Box Plot

- Any point above the Upper or below the Lower will be flagged as a potential outlier.
- Lines extend to the most extreme points inside the Lower and Upper bounds.
Regression Outliers

<table>
<thead>
<tr>
<th></th>
<th>Brain Weight</th>
<th>Gestation</th>
<th>Pred</th>
<th>Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>169 g</td>
<td>392 days</td>
<td>135.9</td>
<td>256.1</td>
</tr>
<tr>
<td>“Man”</td>
<td>1320 g</td>
<td>267 days</td>
<td>481.1</td>
<td>-214.1</td>
</tr>
<tr>
<td>Okapi</td>
<td>490 g</td>
<td>440 days</td>
<td>232.2</td>
<td>207.8</td>
</tr>
</tbody>
</table>

Comments

- The residual for “Man” is not the most extreme.
- The residual for the Brazilian Tapir is the furthest from zero.
- Are any of these residuals statistically significant?
Standardized Residual

\[ z = \frac{\text{residual}}{\text{RMSE}} \]

- A standardized residual should follow a standard normal distribution.

Computing a P-value

- **JMP – Col – Formula**
- \( (1 - \text{Normal Distribution}(|z|)) \times 2 \)

where \(|z|\) is the absolute value of \(z\).

<table>
<thead>
<tr>
<th>Residual</th>
<th>z</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>256.1</td>
<td>3.01</td>
</tr>
<tr>
<td>“Man”</td>
<td>-214.1</td>
<td>-2.52</td>
</tr>
<tr>
<td>Okapi</td>
<td>207.8</td>
<td>2.44</td>
</tr>
</tbody>
</table>
Caution
- We are essentially doing 50 tests of hypothesis.
- If each test has a chance of error of 5%, then one would expect to see some P-values less than 0.05 just by chance.

Bonferroni Correction
- Adjust what is a small P-value.
  \[
  \frac{0.05}{\text{# of residuals}} = \frac{0.05}{50} = 0.001
  \]
- If a P-value is less than 0.001, then the standardized residual is statistically significant.

Conclusion
- Although some of the residuals were flagged on the outlier box plot, none were deemed statistically significant once we corrected for doing 50 simultaneous tests.
Comment

- Because the 3 potential outliers have residuals far from zero, they inflate the value of RMSE.
- Is there a way to evaluate these potential outliers using a value of RMSE more representative of the remaining values?

Adjust RMSE

\[ RMSE = \sqrt{\frac{\text{Var(Residuals)}(n-1)}{n-(k+1)}} \]

- \( \text{Var(Residuals)} = 7091.274 \)
- Adjust by subtracting off the squared residuals for the potential outliers.

### Adjust RMSE

<table>
<thead>
<tr>
<th></th>
<th>Residual</th>
<th>Residual^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>256.1</td>
<td>65587.21</td>
</tr>
<tr>
<td>“Man”</td>
<td>-214.1</td>
<td>45838.81</td>
</tr>
<tr>
<td>Okapi</td>
<td>207.8</td>
<td>43180.84</td>
</tr>
</tbody>
</table>
Adjust RMSE

- $7091.2704(49) = 347472.2496$
- Subtract off $(65587.21+45838.81+43180.84) = 154606.86$
- Adjusted Sum of Squares Residual $347472.25 - 154606.86 = 192865.39$

Adjust RMSE

- $\text{Adj RMSE} = \sqrt{\frac{\text{Adj SS Residual}}{n-(k+1)-0}} = \sqrt{\frac{192865.39}{50-(1+1)-3}} = 65.4668$
- Calculate New Standardized Residuals

New Standardized Residual

<table>
<thead>
<tr>
<th>Residual</th>
<th>z</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazilian Tapir</td>
<td>256.1</td>
<td>3.91</td>
</tr>
<tr>
<td>“Man”</td>
<td>-214.1</td>
<td>-3.27</td>
</tr>
<tr>
<td>Okapi</td>
<td>207.8</td>
<td>3.17</td>
</tr>
</tbody>
</table>