


# Stat 401 B – Lecture 28



## Outliers

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

1

---

---

---


---

---

---

---

---



## 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

2

---

---

---


---

---

---

---

---



## 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?

3

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 28

## Simple Linear Regression

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

4

---

---

---

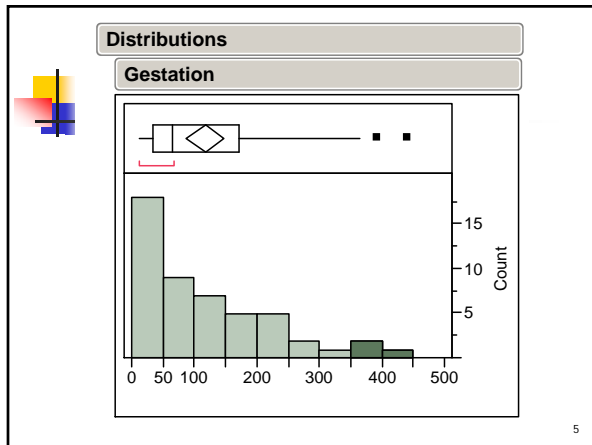
---

---

---

---

---



---

---

---

---

---

---

---

---

## Gestation (days)

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

6

---

---

---

---

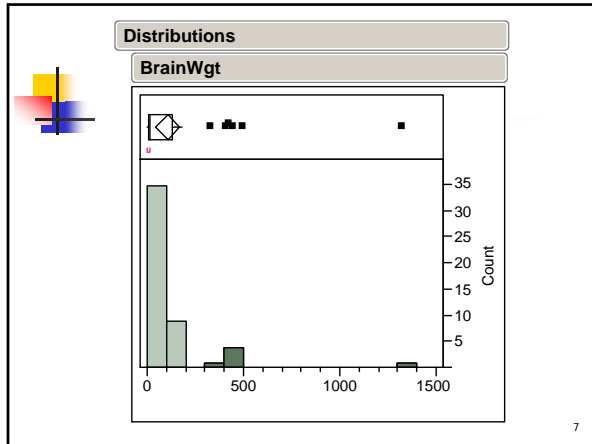
---

---

---

---

# Stat 401 B – Lecture 28



---

---

---

---

---

---

---

---

## Brain Weight

- Highly skewed to the right with several mounds.
- Six potential outliers.
- Mean = 107.25 g
- Median = 16.3 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).

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 28

## 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.

10

---

---

---

---

---

---

---

---

## 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.

11

---

---

---

---

---

---

---

---

## Simple Linear Regression

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

12

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 28

## Simple Linear Regression

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

13

---

---

---

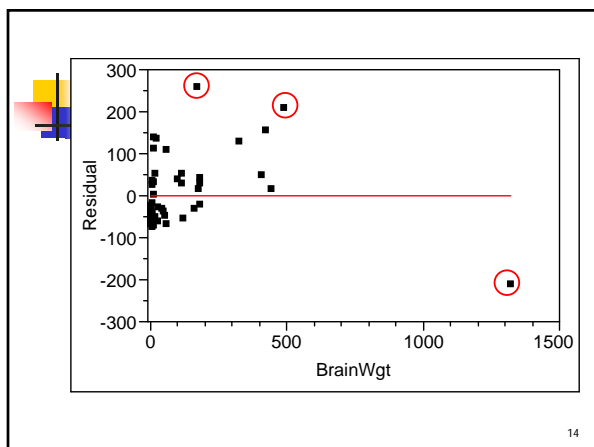
---

---

---

---

---



---

---

---

---

---

---

---

---

## 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.

15

---

---

---

---


---

---

---

---

# Stat 401 B – Lecture 28



## 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

16

---

---

---


---

---

---

---

---



## 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
  - Upper =  $-57.9 - 141.9 = -199.8$

17

---

---

---


---

---

---

---

---



## 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.

18

---

---

---

---

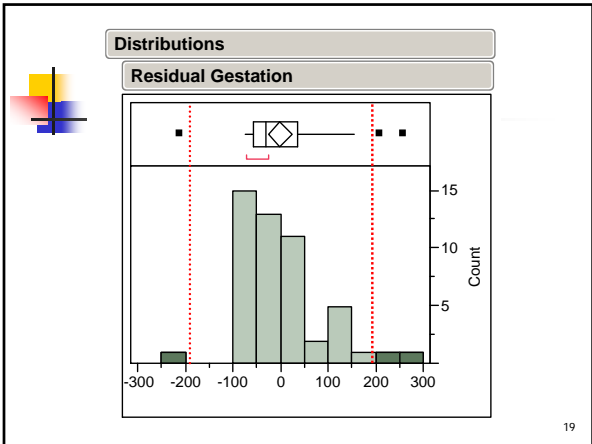
---

---

---

---

# Stat 401 B – Lecture 28




---

---

---

---

---

---

---

---

## Regression Outliers

	Brain Weight	Gestation	Pred	Resid
Brazilian Tapir	169 g	392 days	135.9	256.1
"Man"	1320 g	267 days	481.1	-214.1
Okapi	490 g	440 days	232.2	207.8

---

---

---

---

---

---

---

---

- ## 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?

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 28

## Standardized Residual

$$z = \frac{\text{residual}}{\text{RMSE}}$$

- A standardized residual should follow a standard normal distribution.

22

---

---

---

---

---

---

---

---

## Computing a P-value

- JMP – Col – Formula
- $(1 - \text{Normal Distribution}(|z|)) * 2$
- Where  $|z|$  is the absolute value of  $z$ .

23

---

---

---

---

---

---

---

---

## Standardized Residual

	Residual	z	P-value
Brazilian Tapir	256.1	3.01	0.0026
"Man"	-214.1	-2.52	0.0119
Okapi	207.8	2.44	0.0146

24

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 28

## Caution

- We are essentially doing 50 tests of hypothesis.
- If each test has a chance of error of 5%, then I would expect to see some P-values less than 0.05 just by chance.

25

---

---

---

---

---

---

---

---

## 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.

26

---

---

---

---

---

---

---

---

## 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.

27

---

---

---

---

---

---

---

---