Week 7 Inference for regression

1. Inference for Linear Regression

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

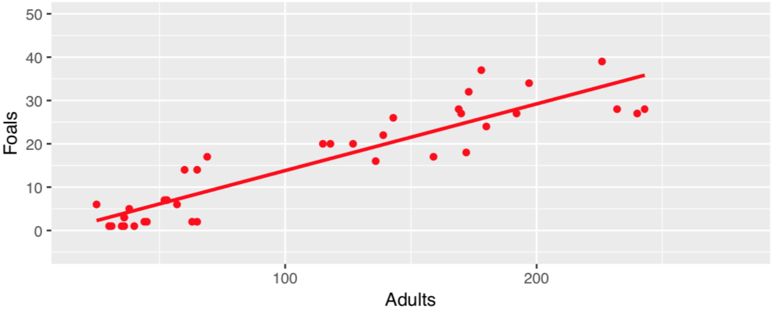
Slides posted at *http://sshanshans.github.io/stat140*

What is the relationship between the size of a herd of horses and the number of foals (baby horses!!) that are born to that herd in a year?

Warm up questions

- What are the variable data types (categorical or quantitative)?
- Which of these variables is the explanatory variable and which is the response?

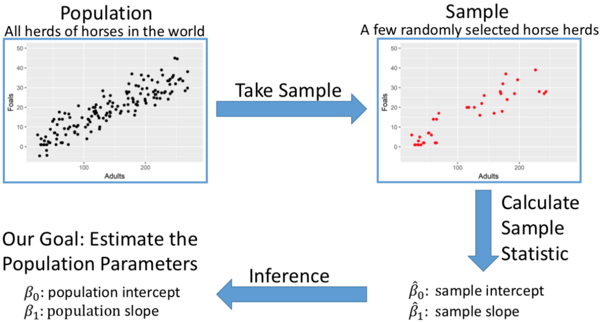
Fit linear regression to describe the relationship between number of adults and number of foals in the sample.



Estimated line: $\hat{y} = b_0 + b_1 x = 0.1540 + -1.5784x$

b_0 and b_1 are sample statistics: they describe the data in our sample

- Everything we have done so far is based solely on sample data
- Now, we will extend from the sample to the population

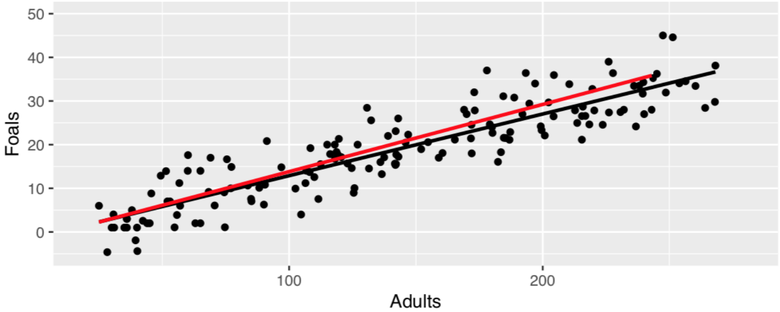


Simple Linear Model

$$y = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} x + \underbrace{\varepsilon}_{\text{Random error}}$$

- β_0 and β_1, are unknown parameters
- Can use familiar inference methods

Imagine all herds of horses in the world. Use data from this sample to learn about the relationship between number of adults and number of foals in the population



Confidence intervals and hypothesis tests for the slope can be done using the familiar formulas:

$$\text{sample statistic} \pm t^* \times \text{SE}$$

$$t^* = \frac{\text{sample statistics - null value}}{\text{SE}}$$

- Population Parameter: β_1, Sample Statistic: $\hat{\beta}_1$
- Use t-distribution with n – 2 degrees of freedom

```r
lm_fit <- lm(Foals ~ Adults, data = horses)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Foals ~ Adults, data = horses)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -8.374 -3.312 -0.965  3.686 11.172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5784     1.4916   -1.06      0.3
## Adults        0.1540     0.0114   13.49  1.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.94 on 36 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.83
## F-statistic:  182 on 1 and 36 DF,  p-value: 1.19e-15
```

$\hat{\beta}_0$, estimated intercept

$\hat{\beta}_1$, estimated slope

Standard Error for $\hat{\beta}_1$: an estimate of the variability in values of b1 we will obtain from different samples

t statistic for a test of whether β1 = 0

p value for a test of whether β1 = 0

Residual standard deviation

Degrees of freedom: n - 2

R^2

Give a 95% confidence interval for the true slope.

Poll question

Is the slope significantly different from 0?

a. Yes
b. No

Give a 95% confidence interval for the true slope.

Poll question

Is the slope significantly different from 0?

(a) Yes
(b) No

Another way to compute CI in R

```
confint(lm_fit, level = 0.95)
```

Set up the hypotheses

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

p-value

probability of observing a slope at least as different from 0 as the one observed if in fact there is no relationship between x and y

```r
lm_fit <- lm(Foals ~ Adults, data = horses)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Foals ~ Adults, data = horses)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -8.374 -3.312 -0.965   3.686 11.172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.5784     1.4916   -1.06      0.3
## Adults        0.1540     0.0114   13.49  1.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.94 on 36 degrees of freedom
## Multiple R-squared: 0.835,  Adjusted R-squared:  0.83
## F-statistic:  182 on 1 and 36 DF,  p-value: 1.19e-15
```

$\hat{\beta}_0$, estimated intercept

$\hat{\beta}_1$, estimated slope

Standard Error for $\hat{\beta}_1$: an estimate of the variability in values of b1 we will obtain from different samples

t statistic for a test of whether β1 = 0

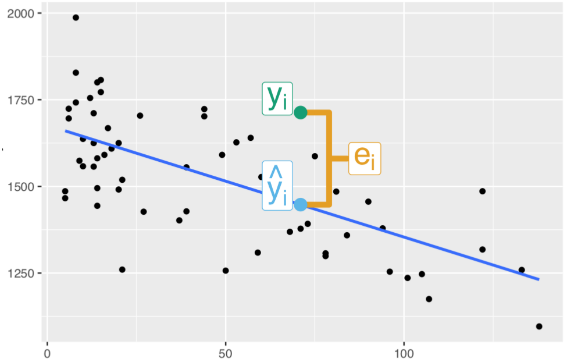
p value for a test of whether β1 = 0

Residual standard deviation

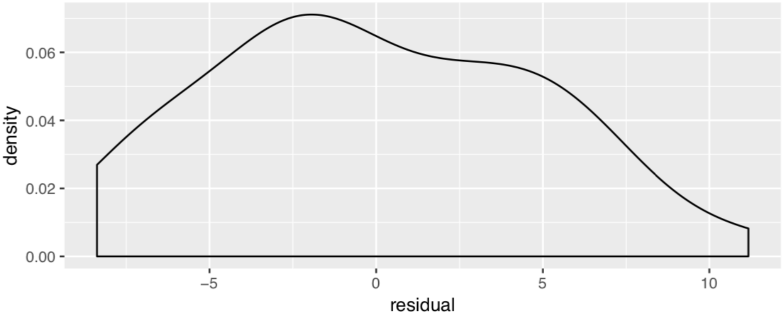
Degrees of freedom: n - 2

R^2

- No outliers (points that don't fit the trend)
- Straight enough?
- Does the plot thicken?
- **New:** Sample representative of population
- **New:** Independence
- **New:** Normally distributed residuals (or large enough sample size)



- Residuals give the vertical distance between a data point and the line of best fit
- Positive if point above line, negative otherwise
- Residual = Observed - Predicted

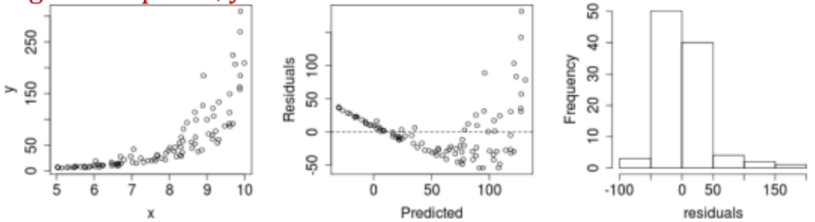

0.06
0.04
0.02
0.00
density
−5
0
5
10
residual

What if the conditions for inference aren't met???

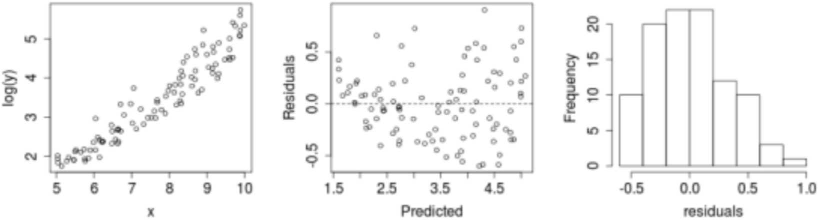
- Option 1: Take 200 level Stat classes and learn more about modeling!
- Option 2: Try a transformation...
  You can take any function of y and use it as the response, but the most common are
  - $\log(y)$ (natural logarithm - ln)
  - $\sqrt{y}$ (square root)
  - y^2 (squared)
  - e^y (exponential))

log(y)

Original Response, y:

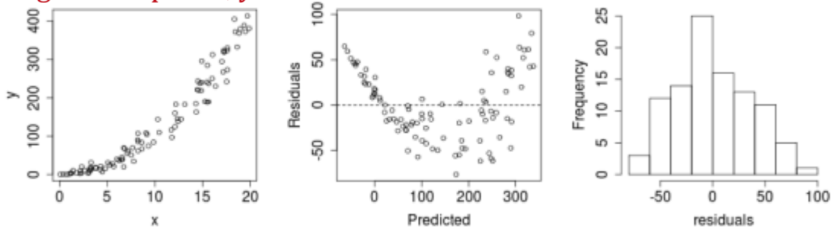


Logged Response, log(y):

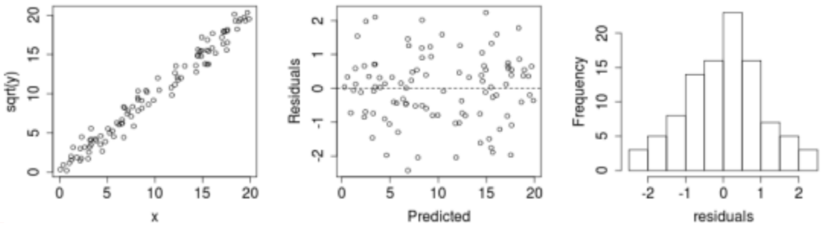


$\sqrt{y}$

Original Response, y:

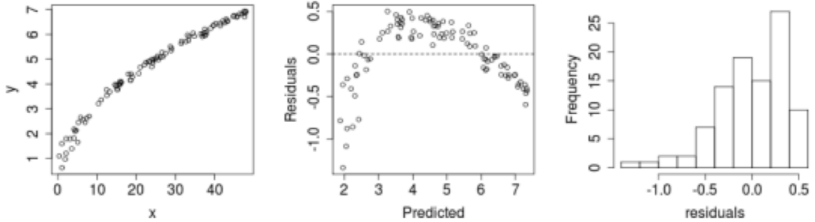


Square root of Response, $\sqrt{y}$:

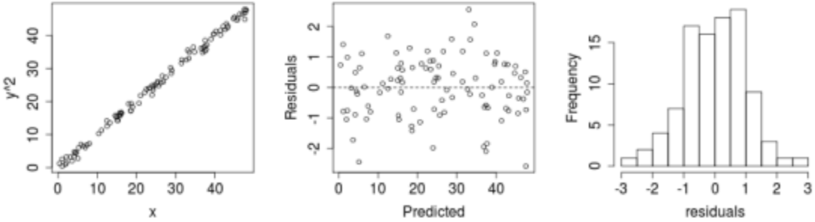


y^2

Original Response, y:

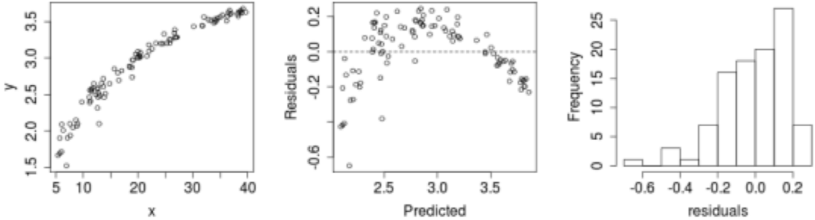


Squared response, y^2:

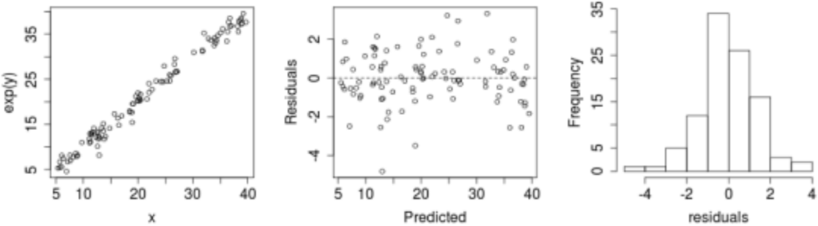


e^y

Original Response, y:



Exponentiated Response, e^y:



- Interpretation becomes a bit more complicated if you transform the response – it should only be done if it clearly helps the conditions to be met
- If you transform the response, be careful when interpreting coefficients and predictions
- You do NOT need to know which transformation would be appropriate for given data in this class, but they may help if conditions are for future data you may want to analyze