Natalie Heineman
9:30-11:00

February 5th, 2014 Scribing

Prediction Intervals

1. See pick up truck sales example (pickup.R).
   - 1 standard deviation from the regression line will cover about 65% of the data
   - 2 standard deviations from the regression line will cover about 95% of the data

2. See mammalsleep.R
   - We fit the model on the log-log scale and then predicted the interval (standard deviations of residuals) on the log scale. However, we want the interval on the original scale
   - The intervals curve just like the line of best fit after undoing the transformation because of the power law
   - Key feature: the interval fans out- the greater x, the greater the absolute error. This happens because the errors on the original scale or multiplicative, while the errors on the log-log scale are additive.
   - When calculating prediction intervals, undo the transformation **last**!

3. Equations

original scale =

$$y_i = \hat{B}_0 + \hat{B}_1 x_i + e_i$$

future x : x^*

point estimate: $\hat{y}^* = \hat{B}_0 + \hat{B}_1 X^*$

$$y^* = \hat{y}^* + e^*$$

$\hat{\sigma}_e$ = st. dev. of residuals   a = conservatism parameter

$$y^* = \hat{y}^* \pm a \cdot \hat{\sigma}_e$$

transformation:

$$\log y_i = \hat{B}_0 + \hat{B}_1 \log x_i + e_i$$

$$\updownarrow$$

$$y_i = e^{B_0} + x_i^{B_1} e^{\varepsilon_i} \rightarrow$$

lower line: $\log y_i = \hat{B}_0 + \hat{B}_1 \log x_i - \hat{\sigma}_e$

upper line: $\log y_i = \hat{B}_0 + \hat{B}_1 \log x_1 + \hat{\sigma}_e$

$$\downarrow$$

original scale:

lower: $y_i = e^{\hat{B}_0} x_i^{B_1} e^{-\hat{\sigma}_\varepsilon}$

upper: $y_i = e^{\hat{B}_0} x_i^{B_1} e^{\hat{\sigma}_\varepsilon}$

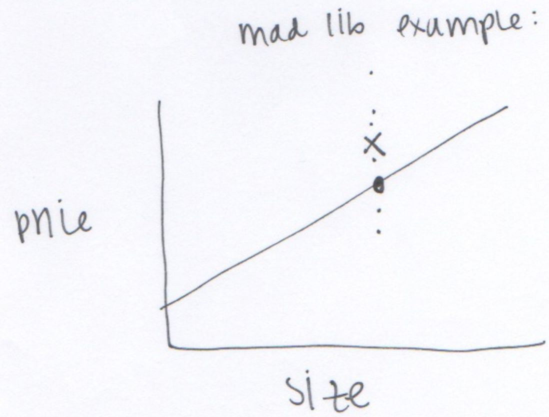
Natalie Heineman
9:30-11:00

Normal (Gaussian) Linear Regression Model

1. Key concepts
   - Sampling distribution
   - Standard error
   - Bootstrapping
     - Fishing example from Monday: our sample acts as the entire population
     - Sample from sample with replacement, which allows for ties and omissions
     - Always use the same sample size as the original sample, because the variation will mimic real samples of the population.

2. Normality assumption

Normality Assumption:
normal regression model:
$y_i = B_0 + B_1 x_i + \varepsilon_i$
(linearity)
$\varepsilon_i \sim N(0, \sigma^2)$
(normality)



- Mad Lib house example- small factors (smell, view, etc) affect the price of the house, causing the price to "nudge" up or down
  - The aggregation of these "nudges" are the residual
  - Each "nudge" has the same probability of moving up or down, so can be considered a binomial distribution
  - Binomial distribution look like normal distributions according to the normality assumption and the central limit theorem
- Example of the normal (Gaussian) distribution: minnows vs. shark
  - A school of minnows is an aggregation of each individual minnow's movements

3. Using the normal distribution to model prediction intervals

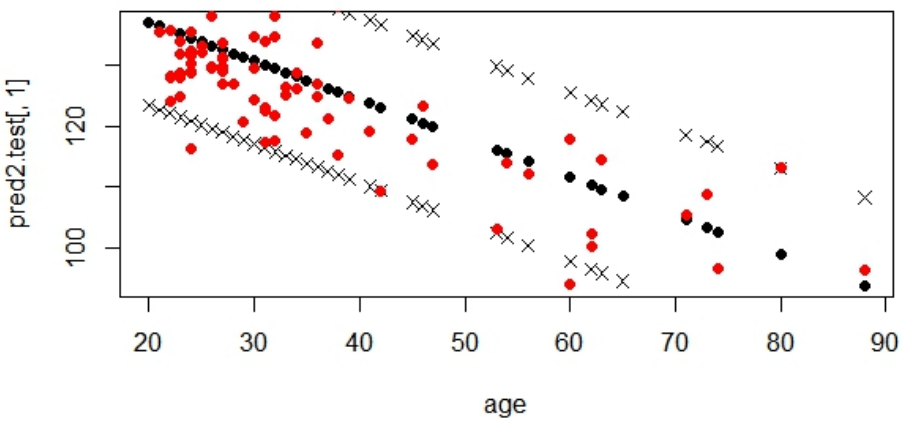
$y_i = B_0 + B_1 x_i + \varepsilon_i$ $\quad\sigma^2$: residual variance
$\varepsilon_i \sim N(0, \sigma^2)$ $\quad\sigma$: residual st. dev.

$$\left.\begin{matrix}(x_1, y_1)\\(x_2, y_2)\\\vdots\\(x_N, y_N)\end{matrix}\right\} \hat{B}_0, \hat{B}_1 \text{ (least squares est.)}$$

sampling dist: $\hat{B}_1 \sim N(B_1, \sigma^2_{\hat{B}_1})$

Natalie Heineman
9:30-11:00

- See creatinine.R and creatinine.csv
    - Find the standard error using bootstrapping or the summary function, which uses the normality assumption
    - Prediction interval:



***Important take-away from today**: assumption of a normal distribution yields uncertainty estimates, which give us an idea about residuals and prediction intervals.