STA 250: Statistics

HW 10

Due Wed Dec 04 2013

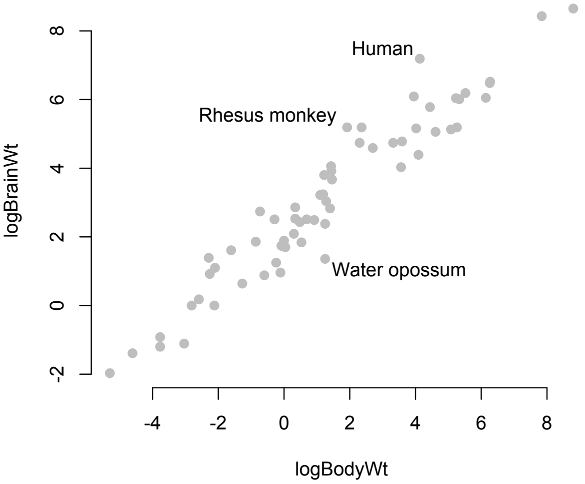
1. Consider the linear model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma^2)$ where (x_i, Y_i), $i = 1, \cdots, n$ are data on X = the boiling point of water (in degrees F) and Y = atmospheric pressure (in inches-Mercury). For observed data we have

$$n = 17, \ \ \bar{x} = 202.95, \ \ s_x^2 = 33.18, \ \ \bar{y} = 25.06, \ \ s_{xy} = 17.35, \ \ s_{y|x}^2 = 0.0542.$$

   (a) Report a 95% predictive interval for atmospheric pressure when the water boiling point is recorded to be 190 degrees F.

   (b) Atmospheric pressure at altitude h (in thousand feet) above the sea level equals $29.92(1 - 0.074401 \cdot h)^{5.25588}$ inches-Mercury. Give a 95% predictive interval for altitude when the water boiling point is measured to be 190 degrees F.

2. The figure below shows log brain weight (in log-grams) against log body weight (in log kilograms) for 62 mammals (three of them are marked). The actual measurements are given below. Remember that these are paired observations, so the first numbers from the two cells (1.22, 3.8) give the body and brain weights (in logarithm) of a single mammal, and so on.



Log Body weight	1.22, -0.73, 0.3, 6.14, 3.59, 3.32, 2.7, 0.04, 1.43, -0.86, -2.29, -0.08, 0, -5.3, -2.81, 1.25, 0.69, 0.53, 7.84, -3.77, 5.23, 6.26, -0.24, 2.3, 1.19, -1.61, 0.34, 6.27, 5.33, 4.44, -0.29, 4.13, 8.8, 1.25, 1.92, 3.56, 1.4, -2.12, -3.77, -4.61, 0.34, 5.52, 0.92, 4.02, 4.61, 3.95, 2.36, -0.6, 4.09, 1.28, 1.46, -1.27, -2.59, -2.1, -3.04, 5.26, 1.1, 5.08, -0.11, 0.48, -2.26, 1.44
Log Brain weight	3.8, 2.74, 2.09, 6.05, 4.78, 4.74, 4.59, 1.7, 4.06, 1.86, 1.39, 1.74, 1.89, -1.97, 0, 2.38, 2.51, 1.84, 8.43, -1.2, 6.04, 6.48, 1.25, 4.74, 3.24, 1.61, 2.86, 6.52, 6.01, 5.78, 2.51, 7.19, 8.65, 1.36, 5.19, 4.03, 2.83, 0, -0.92, -1.39, 2.53, 6.19, 2.49, 5.16, 5.06, 6.09, 5.19, 0.88, 4.39, 3.04, 3.67, 0.64, 0.18, 1.1, -1.11, 5.19, 3.22, 5.13, 0.96, 2.43, 0.92, 3.92

Test how well does the simple linear regression model, with log body weight as the explanatory variable and log brain weight as the response, fit the observed measurements? Follow the binning protocols we have discussed before for implementing Pearson's chi-square goodness-of-fit test.

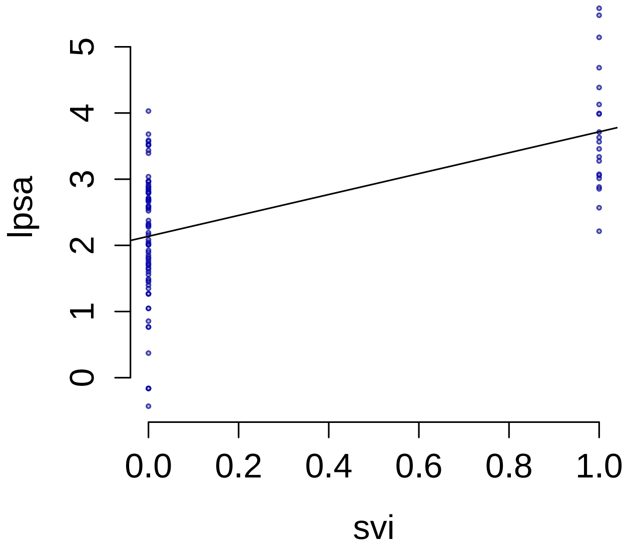
3. The table below shows data on 97 prostate cancer patients with measurements on SVI status (SVI is present/absent) and log-PSA.

SVI status	Log-PSA
'present' (21 individuals)	2.21, 2.57, 2.85, 2.88, 3.01, 3.06, 3.08, 3.28, 3.34, 3.46, 3.57, 3.63, 3.71, 3.98, 3.99, 4.13, 4.39, 4.68, 5.14, 5.48, 5.58
'absent' (76 individuals)	-0.43, -0.16, -0.16, -0.16, 0.37, 0.77, 0.77, 0.85, 1.05, 1.05, 1.27, 1.27, 1.27, 1.35, 1.4, 1.45, 1.47, 1.49, 1.56, 1.6, 1.64, 1.66, 1.7, 1.71, 1.73, 1.77, 1.8, 1.82, 1.85, 1.89, 1.92, 2.01, 2.01, 2.02, 2.05, 2.09, 2.16, 2.19, 2.28, 2.3, 2.31, 2.33, 2.37, 2.52, 2.55, 2.57, 2.59, 2.59, 2.66, 2.68, 2.68, 2.69, 2.7, 2.72, 2.79, 2.79, 2.81, 2.81, 2.84, 2.85, 2.88, 2.89, 2.92, 2.96, 2.96, 2.97, 3.04, 3.39, 3.44, 3.51, 3.52, 3.53, 3.57, 3.59, 3.68, 4.03

For patient i let x_i denote a binary encoding of his SVI status, $x_i = 1$ if 'present' and $x_i = 0$ if 'absent', and let Y_i denote hist log-PSA measurement. Consider the following linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma^2)$$

with unknown parameters $-\infty < \beta_0, \beta_1 < \infty$, $\sigma > 0$. The figure below shows a plot of the data, overlaid with the least squares line fit of the above model:



Here are some relevant summaries of the observed data:

$$n = 97, \quad \bar{x} = 0.216, \quad \bar{y} = 2.48, \quad s_x = 0.414, \quad s_{xy} = 0.271$$
$$\hat{\beta}_0 = 2.137, \quad \hat{\beta}_1 = 1.579, \quad \hat{\sigma} = 0.9565$$

(a) Calculate the p-value for testing $H_0 : \beta_1 = 0$.

(b) Suppose we denote the log-PSA measurements of the SVI-present patients as $Y_1^1, \cdots, Y_{21}^1$ and those of the SVI-absent patients as $Y_1^0, \cdots, Y_{76}^0$ and assume our usual two sample normal model: $Y_i^1 \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma^2)$, $Y_j^0 \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_0, \sigma^2)$. For these data and model, the ML p-value for testing $H_0 : \mu_1 = \mu_0$ (which you could obtain by using `t.test()`) turns out to be identical to the p-value you calculated in part (a). Explain why it should be expected that the two p-values are equal [present a brief but logical argument by comparing the two models, and use the fact that we are looking at ML tests under either setting.]

(c) Let Y^{*1} and Y^{*0} denote the log-PSA measurements on two future prostate cancer patients who, respectively, are SVI-present and SVI-absent. Give a 95% prediction interval for $D^* = Y^{*1} - Y^{*0}$. Justify your calculations.