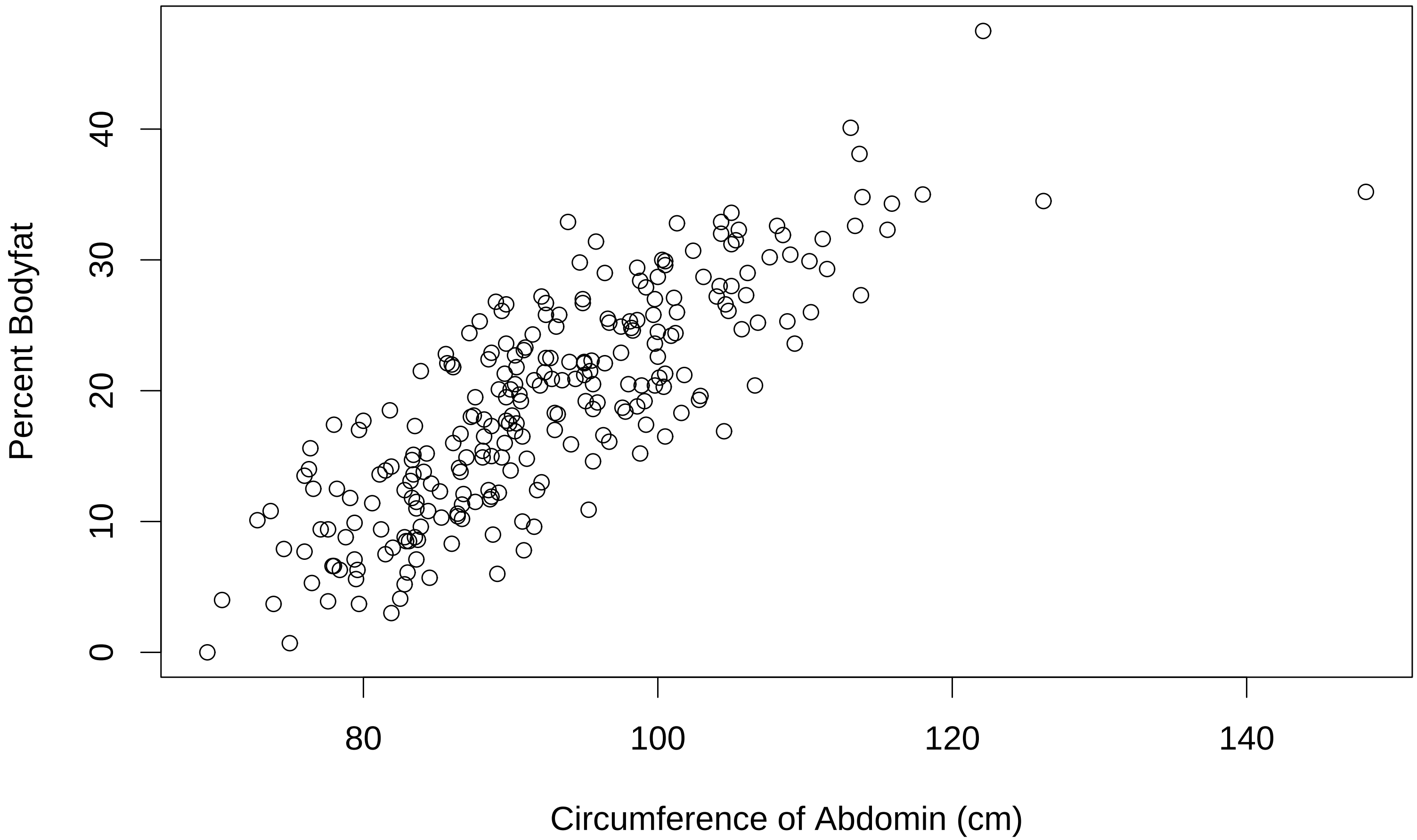
Simple Linear Regression

March 16, 2009

Reading Lee Ch 6

BodyFat Data



Body Fat Example

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-39.2802	2.6603	-14.77	2.210^{-16}
Abdomen	0.6313	0.0286	22.11	2.210^{-16}

95% HPD interval

$$0.6313 \pm \texttt{qt}(.025, 250) * 0.0286 = (0.57, 0.69)$$

For every additional cm of abdominal circumference, percent bodyfat increases by 0.57 to 0.69 percent with probability 0.95.

Interpretation

- For every additional centimeter of abdominal circumference, percent body fat increases by 0.67 percent $(0.57, 0.69)$
- For every additional inch of abdominal circumference, percent body fat increases by $2.54 * .67 = 1.7$ percent $(1.45, 1.74)$
- Abdominal circumference explains roughly 68% of the variation in bodyfat
- Percent Body fat for 34 inch abdomin $-42.96 + 34 * 2.54 * .67 = 14.9\%$

Significance of the Regression

Question: How probable is $\beta = 0$ under the posterior?
Informal Answer: Compute posterior probability on β values with lower posterior density than $\beta = 0$

- "Measures" probability of β "less likely" than $\beta = 0$
- Informal "test": Probability in tails = significance level = (Bayesian) p-value

$$\text{p-value} = P(|t| > |\hat{\beta}/s_\beta|) = P(|t| > |\hat{\beta}/\text{SE}(\beta)|)$$

- Classical testing terminology:
  "The regression on x is significant at the 5% level (or 1%, etc) if the p-value is smaller than 0.05 (or 0.01, etc)"

Lindley's Method

Lindley suggested rejecting the hypothesis that $\beta = 0$ at the $\alpha 100\%$ level of significance if the $(1-\alpha)100\%$ HPD region does not include 0.

$$0 \notin (\hat{\beta} - t_{1-\alpha/2} s_{\beta}, \hat{\beta} + t_{1-\alpha/2} s_{\beta})$$

Equivalent to comparing the p-value to α and concluding that the regression is significant if the p-value is less than α.

Alternative approach is to compute a Bayes Factor.

Bayes Factors

Testing $H_o : \beta = 0$ versus $H_a : \beta \neq 0$

- Assign prior probabilities to H_o and H_a
- Find $P(H_i \mid Y)$ via Bayes Theorem

Bayes Factor for comparing evidence in favor of H_o

$$\text{BF}[H_o : H_a] = \frac{p(H_o \mid Y)/p(H_o)}{p(H_a \mid Y)/p(H_a)}$$

Often difficult to calculate, instead use lower bound based on p-values (Berger, Selke and Bayarri)

$$\text{BF}[H_o : H_a] = -ep\log(p)$$

Bodyfat Example

- $P(|t| > 22.11) = 2.2 \times 10^{-16}$
- The regression of bodyfat on abdominal circumference is highly significant (p-value = 2.2×10^{-16}).
- Lower bound on Bayes Factor
  BF$[H_0 : H_a] = 2.15 \times 10^{-14}$

$$-2.2 \times 10^{-16} \exp(1) \log(2.2 \times 10^{-16}) = 2.156043 \times 10^{-14}$$

- Approximate posterior probability of H_0 = 2.15×10^{-14}

$$P(H_0 \mid Y) = \frac{\text{BF}[H_0 : H_a] O[H_0 : H_a]}{1 + \text{BF}[H_0 : H_a] O[H_0 : H_a]}$$

Jeffreys Scale of Evidence

Bayes Factor	Interpretation
$B \geq 1$	H_0 supported
$1 > B \geq 10^{-\frac{1}{2}}$	minimal evidence against H_0
$10^{-\frac{1}{2}} > B \geq 10^{-1}$	substantial evidence against H_0
$10^{-1} > B \geq 10^{-2}$	strong evidence against H_0
$10^{-2} > B$	decisive evidence against H_0

$B = \text{BF}[H_o : B_a]$

Decisive evidence against hypothesis that bodyfat is not associated with abdominal circumference

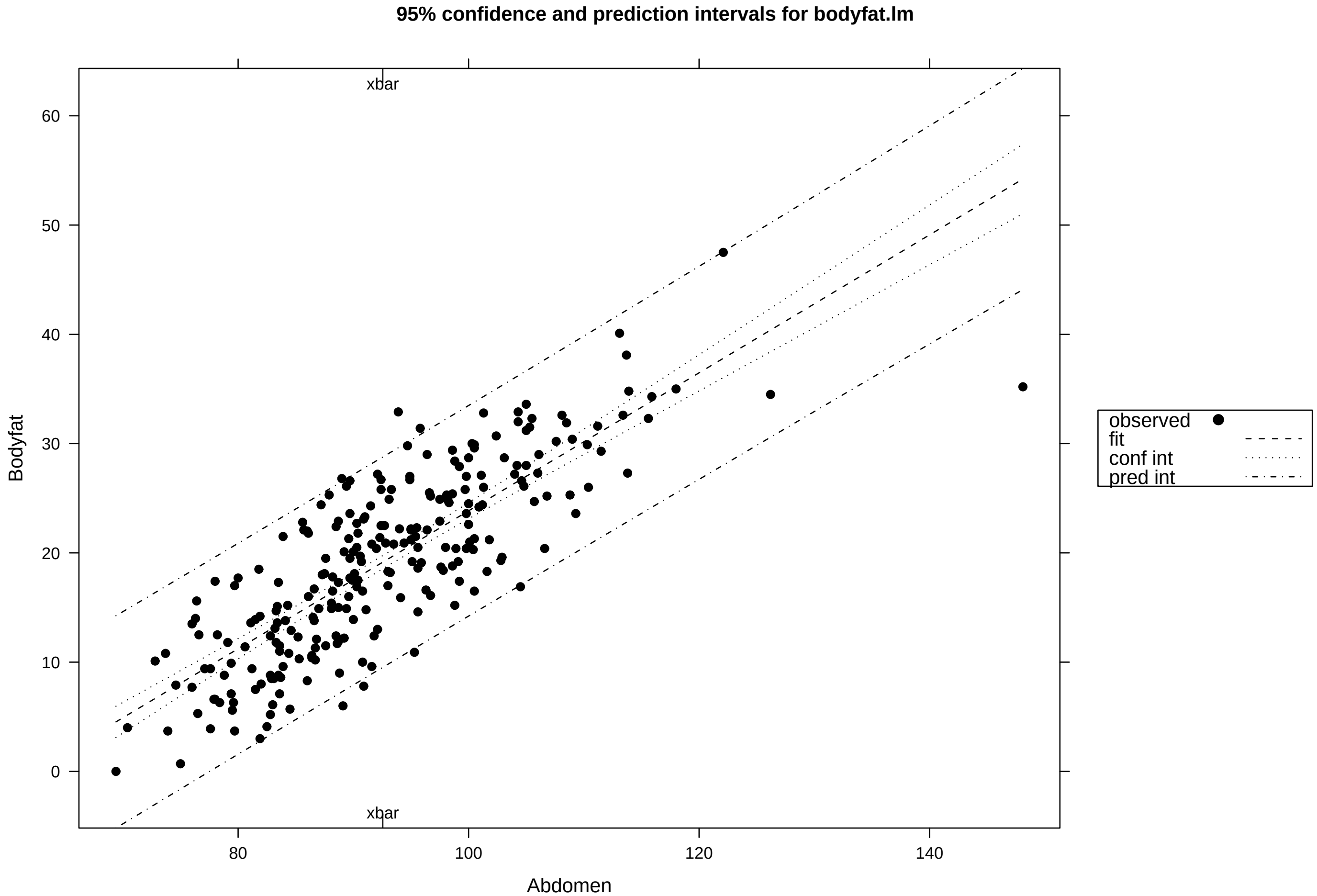
Predictions

The (posterior) predictive distribution for a new case, $y_{n+1} = \alpha + \beta x_{n+1} + \varepsilon_{n+1}$ is also a Student t distribution with $n-2$ df.

$$
\begin{aligned}
y_{n+1} | y_1, \ldots y_n &\sim t_{n-2}(\hat{y}, s^2_{y_{n+1}}) \\
\hat{y} &= \hat{\alpha} + \hat{\beta} x_{n+1} \\
s^2_{y_{n+1}} &= s^2_{Y|X}(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}})
\end{aligned}
$$

- posterior uncertainty about $\alpha + \beta x_{n+1}$
- depends on x_{n+1} spread is higher for x_{n+1} far from $\bar{x}$
- additional variability $+s^2_{Y|X}$ due to ε_{n+1}

Intervals: `ci.plot(bodyfat.lm)`



Residual Analysis