Bayesian Inference for Simple Linear Regression

Al Nosedal.
University of Toronto.

November 29, 2015

Sometimes we want to model a relationship between two variables, X and Y. We might want to find an equation that describes the relationship. Often we plan to use the value of X to help predict Y using that relationship. The data consist of n ordered pairs of points (x_i, y_i) for $i = 1, 2, 3, ..., n$. We think of x as the predictor variable (independent variable) and consider that we know it without error. We think y is a response variable that depends on x in some unknown way, but that each observed y contains an error term as well.

A linear relationship is the simplest equation relating two variables. This would give a straight line relationship between the predictor x and the response y. We leave the parameters of the line, the slope β, and the y-intercept α_0 unknown, so all lines are possible. Then we determine the best estimates of the unknown parameters by some criterion. The criterion that is most frequently used is **least squares**. This is where we find the parameter values that minimize the sum of squares of the **residuals**, which are the vertical distances of the observed points to the fitted equation.

Sum of Squares of the Residuals

The sum of squares of the residuals from line $y = \alpha_0 + \beta x$ is

$$SS_{res} = \sum_{i=1}^{n} [y_i - (\alpha_0 + \beta x_i)]^2.$$

To find values of α_0 and β that minimize SS_{res} using calculus, take derivatives with respect to each α_0 and β and set equal to 0, and solve the resulting set of simultaneous equations.

Least squares line

The equation of the least squares line is

$$\hat{y} = A_0 + Bx$$

where

$$B = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - (\bar{x})^2} = r\frac{S_y}{S_x}.$$

$$A_0 = \bar{y} - B\bar{x}.$$

(where r = sample correlation, S_y = sample standard deviation of y, and S_x = sample standard deviation of x)

Alternative equation

The slope and any other point besides y-intercept also determines the line. Say the point $A_{\bar{x}}$, where the least squares line intercepts the vertical line at $\bar{x}$:

$$A_{\bar{x}} = A_0 + B\bar{x} = \bar{y}.$$

Thus the least squares line goes through the point $((\bar{x}, \bar{y})$. An alternative equation for the least squares line is

$$\hat{y} = A_{\bar{x}} + B(x - \bar{x}) = \bar{y} + B(x - \bar{x}),$$

which is particularly useful.

Estimating the Variance around the Least Squares Line

The estimate of the variance around the least squares line is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[y_i - (A_{\bar{x}} + B(x_i - \bar{x}))]^2}{n-2}$$

which is the sum of squares of the residuals divided by $n-2$. The reason we use $n-2$ is that we have used two estimates, $A_{\bar{x}}$ and B in calculating the sum of squares. The general rule for finding an unbiased estimate of the variance is that the sum of squares is divided by the degrees of freedom, and we lose a degree of freedom for every estimated parameter in the sum of squares formula.

Simple Linear Regression Assumptions

1. Mean assumption. The conditional mean of y given x is an unknown linear function of x.

$$\mu_{y|x} = \alpha_0 + \beta x,$$

where β is the unknown slope and α_0 is the unknown y intercept, the intercept of the vertical line x=0. In the alternate parameterization

$$\mu_{y|x} = \alpha_{\bar{x}} + \beta(x - \bar{x}),$$

where $\alpha_{\bar{x}}$ is the unknown intercept of the vertical line $x = \bar{x}$.

Simple Linear Regression Assumptions

2. Error assumption. Observation equals mean plus error, which is Normally distributed with mean 0 and **known** variance σ^2. All errors have equal variance.

Simple Linear Regression Assumptions

3. Independence Assumption. The errors for all the observations are independent of each other.

Using the alternate parameterization

$$y_i = \alpha_{\bar{x}} + \beta \times (x_i - \bar{x}) + e_i,$$

where $\alpha_{\bar{x}}$ is the mean value for y given $x = \bar{x}$, and β is the slope. Each e_i is Normally distributed with mean 0 and known variance σ^2. The e_i are all independent of each other. Therefore $y_i|x_i$ is Normally distributed with mean $\alpha_{\bar{x}} + \beta \times (x_i - \bar{x})$ and variance σ^2 and all the $y_i|x_i$ are all independent of each other.

Bayes' Theorem for the Regression Model

Bayes' theorem is always summarized by

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

so we need to determine the likelihood and decide on our prior for this model.

The Joint Likelihood for β and $\alpha_{\bar{x}}$

The likelihood of observation i is:

$$likelihood_i(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2},$$

since we can ignore the part not containing the parameters.

The Joint Likelihood for β and $\alpha_{\bar{x}}$

The observations are all independent, so the likelihood of the whole sample of all the observations is the product of the individual likelihoods:

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}[y_i-(\alpha_{\bar{x}}+\beta(x_i-\bar{x}))]^2\right]}.$$

The Joint Likelihood for β and $\alpha_{\bar{x}}$

The term in brackets in the exponent equals

$$\sum_{i=1}^{n}[y_i - \bar{y} + \bar{y} - (\alpha_{\bar{x}} + \beta(x_i - \bar{x}))]^2$$

Breaking this into three sums and simplifying gives us

$$SS_y - 2\beta SS_{xy} + \beta^2 SS_x + n(\alpha_{\bar{x}} - \bar{y})^2,$$

The Joint Likelihood for β and $\alpha_{\bar{x}}$

where $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$, and $SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, and $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Thus the joint likelihood can be written as

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}\left[SS_y - 2\beta SS_{xy} + \beta^2 SS_x + n(\alpha_{\bar{x}} - \bar{y})^2\right]}.$$

The Joint Likelihood for β and $\alpha_{\bar{x}}$

Writing this as the a product of two exponentials gives

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{1}{2\sigma^2}\left[SS_y - 2\beta SS_{xy} + \beta^2 SS_x\right]} e^{-\frac{1}{2\sigma^2}\left[n(\alpha_{\bar{x}} - \bar{y})^2\right]}.$$

We factor out SS_x in the first exponential, complete the square, and absorb the part that doesn't depend on any parameter into the proportionality constant. This gives us

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto e^{-\frac{SS_x}{2\sigma^2}\left[\beta - \frac{SS_{xy}}{SS_x}\right]^2} e^{-\frac{n}{2\sigma^2}\left[(\alpha_{\bar{x}} - \bar{y})^2\right]}.$$

The Joint Likelihood for β and $\alpha_{\bar{x}}$

Note that $B = \frac{SS_{xy}}{SS_x}$, the least squares slope, and $\bar{y} = A_{\bar{x}}$, the least squares estimate of the intercept of the vertical line $x = \bar{x}$. We have factored out the joint likelihood into the product of two individual likelihoods

$$likelihood_{sample}(\alpha_{\bar{x}}, \beta) \propto likelihood_{sample}(\alpha_{\bar{x}}) \times likelihood_{sample}(\beta),$$

Since the joint likelihood has been factored into the product of the individual likelihoods we know that the individual likelihoods are independent.

The Joint Prior for β and $\alpha_{\bar{x}}$

If we multiply the joint likelihood by a joint prior, it is proportional to the joint posterior. We will use independent priors for each parameter. The joint prior of the two parameters is the product of the two individual priors:

$$g(\alpha_{\bar{x}}, \beta) = g(\alpha_{\bar{x}}) \times g(\beta).$$

We can either use Normal priors, or flat priors.

The Joint Posterior for β and $\alpha_{\bar{x}}$

The joint prior and the joint likelihood both factor into a part depending on $\alpha_{\bar{x}}$ and a part depending on β. Rearranging them gives the joint posterior factored into the marginal posteriors

$$g(\alpha_{\bar{x}}, \beta | \ \textit{data}) \propto g(\alpha_{\bar{x}} | \ \textit{data}) \times g(\beta | \ \textit{data}).$$

The Joint Posterior for β and $\alpha_{\bar{x}}$

Since the joint posterior is the product of the marginal posteriors, they are independent. Each of these marginal posteriors can be found by using the simple updating rules for Normal distributions, which works for Normal and flat priors. For instance, if we use a Normal(m_β, s_β^2) prior for β, we get a Normal$(m_\beta^{'}, (s_\beta^{'})^2)$, where

$$\frac{1}{(s_\beta^{'})^2} = \frac{1}{s_\beta^2} + \frac{SS_x}{\sigma^2}$$

and

$$m_\beta^{'} = \frac{\frac{1}{s_\beta^2}}{\frac{1}{(s_\beta^{'})^2}} \times m_\beta + \frac{\frac{SS_x}{\sigma^2}}{\frac{1}{(s_\beta^{'})^2}} \times B.$$

The Joint Posterior for β and $\alpha_{\bar{x}}$

Similarly if we use a Normal$(m_{\alpha_{\bar{x}}}, s^2_{\alpha_{\bar{x}}})$ prior for $\alpha_{\bar{x}}$, we get a Normal$(m'_{\alpha_{\bar{x}}}, (s'_{\alpha_{\bar{x}}})^2)$, where

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{s^2_{\alpha_{\bar{x}}}} + \frac{n}{\sigma^2}$$

and

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{s^2_{\alpha_{\bar{x}}}}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} \times m_{\alpha_{\bar{x}}} + \frac{\frac{n}{\sigma^2}}{\frac{1}{(s'_{\alpha_{\bar{x}}})^2}} \times A_{\bar{x}}.$$

Bayesian Credible Interval for Slope

The posterior distribution of β summarizes our entire belief about it after examining the data. We may want to summarize it by a $(1-\alpha)\times 100\%$ Bayesian credible interval for slope β. This will be

$$m'_\beta \pm z_{\alpha/2}\sqrt{(s'_\beta)^2}.$$

Bayesian Credible Interval for Slope

More realistically, we don't know σ^2. A sensible approach in that instance is to use the estimate calculated from the residuals ($\hat{\sigma}^2$). We have to widen the confidence interval to account for the increased uncertainty due to not knowing σ^2. We do this by using a Student's t critical value with $n-2$ degrees of freedom**. The credible interval becomes

$$m'_\beta \pm t_{\alpha/2}\sqrt{(s'_\beta)^2}.$$

** *Actually we are treating the unknown parameter σ^2 as a nuisance parameter and using the prior $g(\sigma^2) \propto (\sigma^2)^{-1}$. The marginal posterior of β is found by integrating σ^2 out of the joint posterior.*

Testing One-sided Hypothesis about Slope

Often we want to determine whether or not the amount of increase in y associated with one unit increase in x is greater than some value, β_0. We can do this by testing

$$H_0 : \beta \leq \beta_0 \quad \text{vs} \quad H_1 : \beta > \beta_0$$

at the α level of significance in a Bayesian manner. To do the test in a Bayesian manner, we calculate the posterior probability of the null hypothesis.
(If we used the estimate of the variance, then we would use a Student's t with n-2 degrees of freedom instead of the standard Normal Z.)

Testing Two-sided Hypothesis about Slope

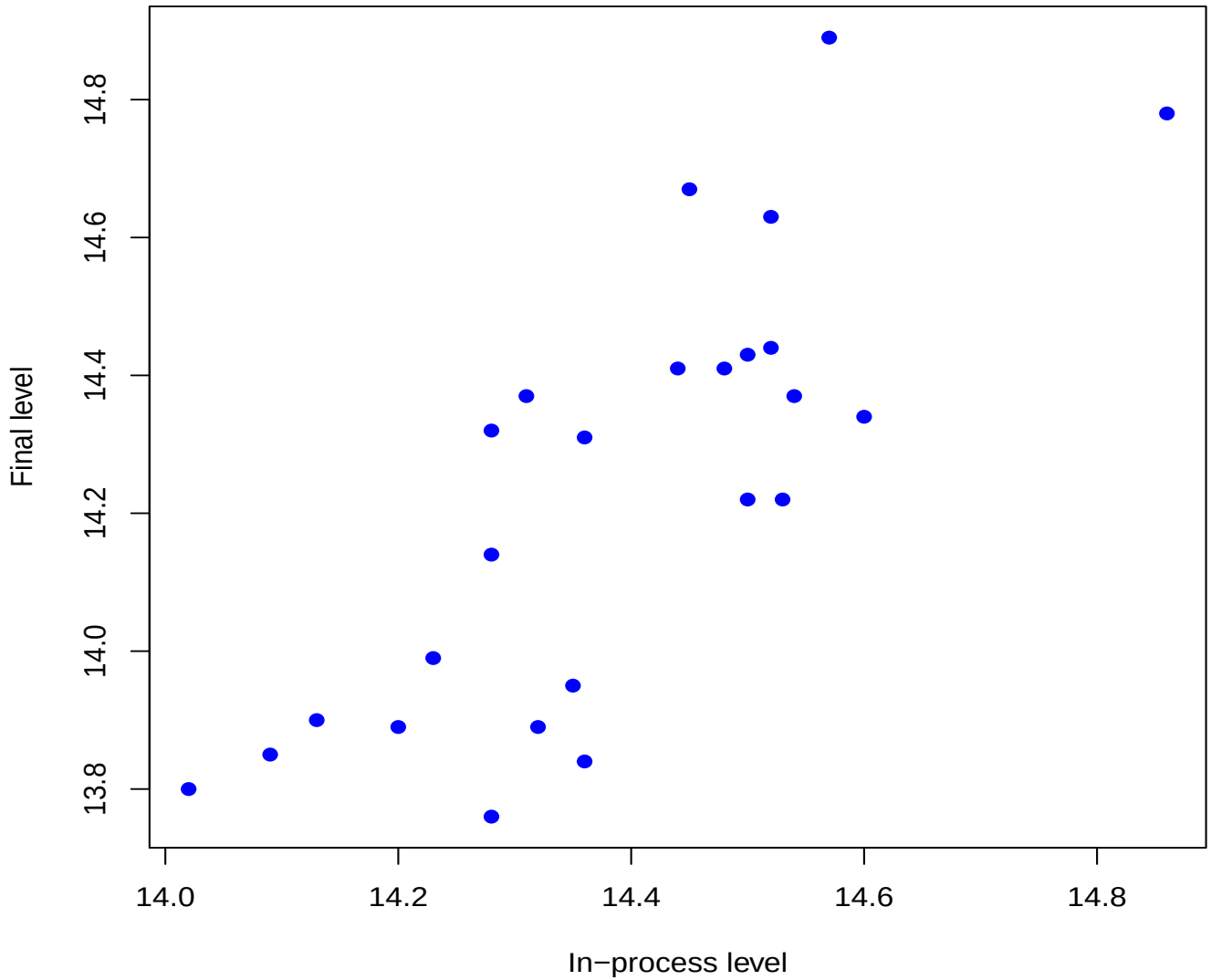
If $\beta = 0$, then the mean of y does not depend on x at all. We really would like to test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ in a Bayesian manner, before we use the regression model to make predictions. To do the test in a Bayesian manner, look where 0 lies in relation to the credible interval. If it lies outside interval, reject H_0. Otherwise, we can't reject the null hypothesis, and we should not use the regression model to help with predictions.

Example

A company is manufacturing a food product, and must control the moisture level in the final product. It is cheaper (and hence preferable) to measure the level at an in-process stage rather than in the final product. The company statistician recommends to the engineers running the process that a measurement of the moisture level at an in-process stage may be a good prediction of what the final moisture level will be. He organizes the collection of the data from 25 batches, giving the moisture level at the in-process stage and the final moisture level for each batch.

Summary statistics for these data are: $\bar{x} = 14.389$, $\bar{y} = 14.221$, $\bar{x^2} = 207.0703$, $\bar{y^2} = 202.3186$, and $\bar{xy} = 204.6628$.

Scatterplot



Example

He then calculates the least squares line relating the final moisture level to the in-process moisture level:

$$B = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - (\bar{x})^2} = \frac{204.6628 - 14.389(14.221)}{207.0703 - (14.389)^2} = \frac{0.042569}{0.032755} = 1.29963$$

The equation of the least squares line is

$$\hat{y} = 14.221 + 1.29963(x - 14.389)$$

Example

Then, he calculates the least squares fitted values $\bar{y} + B(x_i - \bar{x})$, the residuals, and the squared residuals. The estimated variance about the least squares line is

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \frac{0.80188}{23} = 0.0320753.$$

To find the estimated standard deviation about the least squares line, he takes the square root:

$$\hat{\sigma} = \sqrt{0.0320753} = 0.179096.$$

Example (cont.)

The statistician decides that he will use a Normal$(1, 0.3^2)$ prior for β and a Normal$(15, 1^2)$ prior for $\alpha_{\bar{x}}$. Since he doesn't know the true variance, he will use the estimated variance about the least squares regression line $\hat{\sigma^2} = 0.0320753$. Note that $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n(\bar{x^2} - (\bar{x})^2) = 25(207.0703 - 14.389^2) = 0.674475$

Example (cont.)

The posterior precision of β is

$$\frac{1}{(s'_\beta)^2} = \frac{1}{0.3^2} + \frac{25}{0.674475} = 48.177,$$

so the posterior standard deviation of β is

$$s'_\beta = \sqrt{48.177} = 0.144$$

Example (cont.)

The posterior mean of β is

$$m'_{\beta} = \frac{\frac{1}{0.3^2}}{48.177} \times 1 + \frac{\frac{25}{0.64775}}{48.177} \times 1.29963 = 1.231.$$

Example (cont.)

Similarly the posterior precision of $\alpha_{\bar{x}}$ is

$$\frac{1}{(s'_{\alpha_{\bar{x}}})^2} = \frac{1}{1^2} + \frac{25}{0.674475} = 38.066,$$

so the posterior standard deviation of $\alpha_{\bar{x}}$ is

$$s'_{\alpha_{\bar{x}}} = \sqrt{38.066} = 0.162.$$

Example (cont.)

The posterior mean of $\alpha_{\bar{x}}$ is

$$m'_{\alpha_{\bar{x}}} = \frac{\frac{1}{1^2}}{38.066} \times 15 + \frac{\frac{25}{0.64775}}{38.066} \times 14.221 = 14.242.$$

Example (cont.)

Since he used the estimated variance in place of the unknown true variance, he used $m'_\beta \pm t_{\alpha/2}\sqrt{(s'_\beta)^2}$ to find the 95% Bayesian credible interval where there are 23 degrees of freedom. The interval is

$$1.231 \pm 2.069(0.144)$$

$$(0.933, 1.529).$$

Example

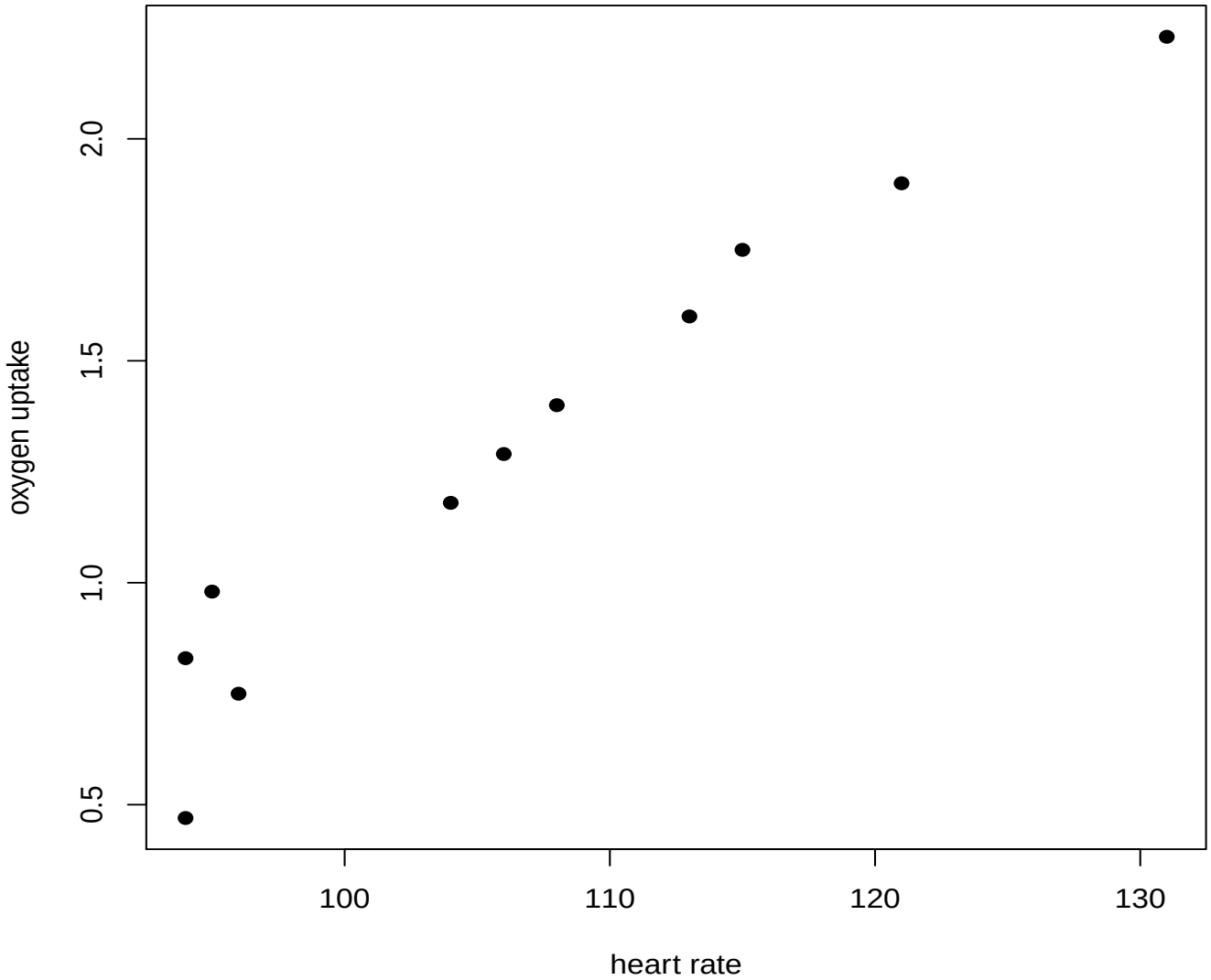
A researcher measured heart rate (x) and oxygen uptake (y) for one person under varying exercise conditions. He wishes to determine if heart rate which is easier to measure can be used to predict oxygen uptake. If so, then the estimated oxygen uptake based on the measured heart rate can be used in place of the measured oxygen uptake for later experiments on the individual.

Heart Rate x	Oxygen Uptake y
94	0.47
96	0.75
94	0.83
95	0.98
104	1.18
106	1.29
108	1.40
113	1.60
115	1.75
121	1.90
131	2.23

Example

a) Plot a scatterplot of oxygen uptake y versus heart rate x.
b) Calculate the parameters of the least squares line.
c) Graph the least squares line on your scatterplot.
d) Calculate the estimated variance about the least squares line.
e) Suppose that we know that oxygen uptake given the heart rate is Normal$(\alpha_0 + \beta x, \sigma^2)$, where $\sigma^2 = 0.13^2$ is **known**. Use a Normal$(0, 1^2)$ prior for β. What is the posterior distribution of β?
f) Find a 95% credible interval for β.
g) Perform a Bayesian test of $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ at the 95% level of significance.

Solution a



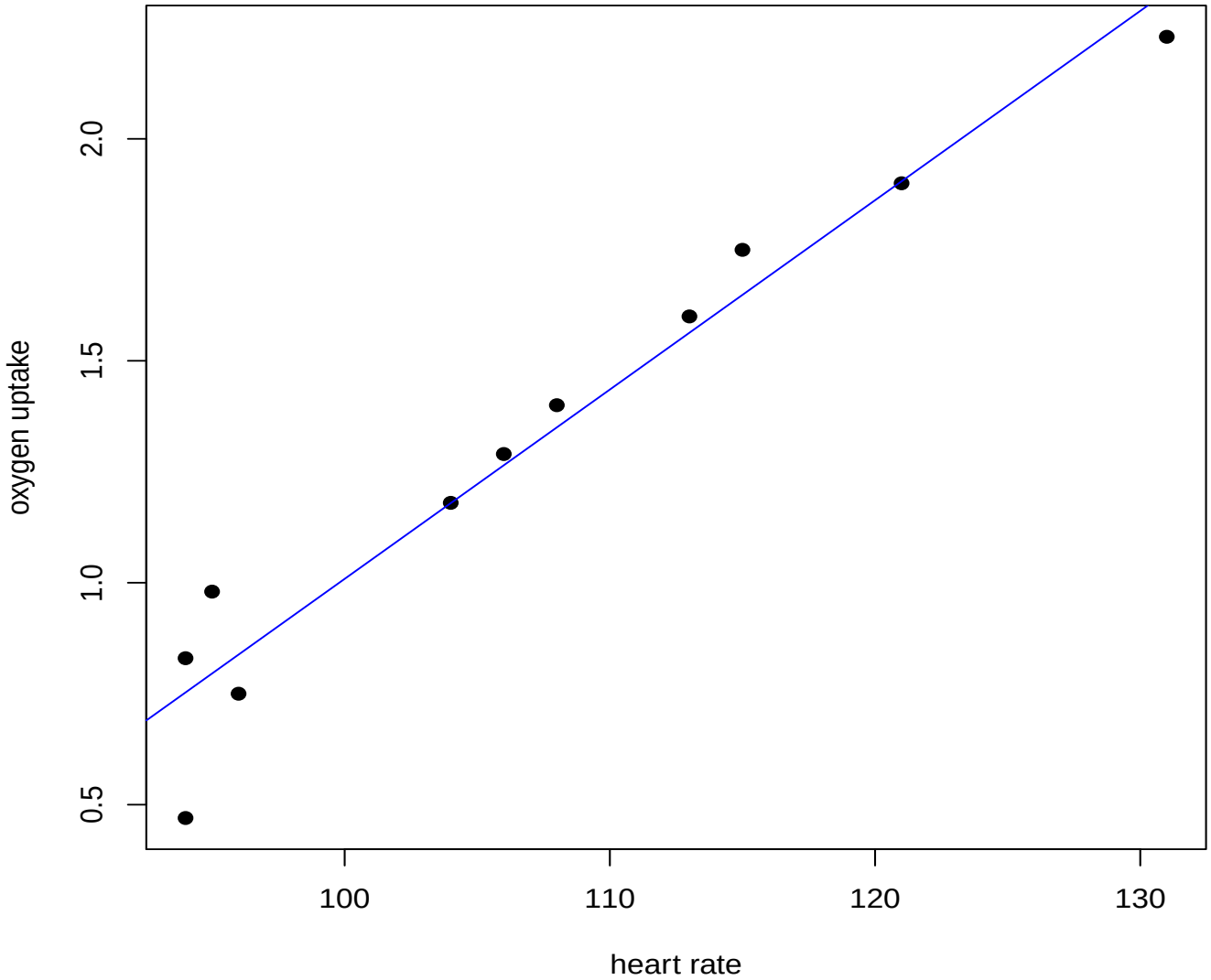
Solution b

The least squares slope

$$B = \frac{145.64 - (107)(1.307273)}{11584.09 - (107)^2} = \frac{5.761818}{135.0909} = 0.04265141$$

$$A_0 = 1.307273 - 0.04265141(107) = -3.256428$$

Solution c



Solution d

The estimated variance about the least squares line is found by taking the sum of squares of residuals and dividing by $n-2$ and equals $\hat{\sigma}^2 = 0.1302867^2$.

Solution e

The likelihood of β is proportional to a Normal$(B, \frac{\sigma^2}{SS_x})$ where B is the least squares slope and $SS_x = 1486$ and $\sigma^2 = 0.13^2$.
The prior for β is Normal$(0, 1^2)$. The posterior precision will be

$$\frac{1}{(s')^2} = \frac{1}{1^2} + \frac{SS_x}{0.13^2} = 87930$$

the posterior variance will be $(s')^2 = \frac{1}{87930} = 0.000011$ and the posterior mean

$$m' = \frac{1/1^2}{87930} \times 0 + \frac{SS_x/0.13^2}{87930} \times 0.04265141 = 0.0426509$$

The posterior distribution of β is Normal$(0.0426, 0.0033^2)$.

Solution f and g

$$m'_\beta \pm z_{\alpha/2}\sqrt{(s'_\beta)^2}.$$

$$0.0426 \pm 1.96(0.0033).$$

A 95% Bayesian Credible Interval for β is (0.036, 0.049).

We observe that the null value 0 lies outside the credible interval, so we reject the null hypothesis.

```
# R Code;

# Data;

heart=c(94,96,94,95,104,106,108,113,115,121,131);

oxygen=c(0.47,0.75,0.83,0.98,1.18,1.29,1.40,1.60,
1.75,1.90,2.23);

mean(heart);

mean(oxygen);

mean(heart*oxygen);

mean(heart*heart);
```

```
# R Code;

# Scatterplot;

plot(heart,oxygen,pch=19,xlab="heart rate",
ylab="oxygen uptake");

# pch=19 tells R to draw solid circles;
```

```
# R Code;

lin.reg=lm(oxygen~heart);

names(lin.reg);

lin.reg$res;

# gives you the residuals;

n=length(lin.reg$res);

sigma2.hat= sum(lin.reg$res*lin.reg$res)/(n-2);

sigma.hat=sqrt(sigma2.hat);
```

```
# Scatterplot with least-squares line;

plot(heart,oxygen,pch=19,xlab="heart rate",
ylab="oxygen uptake");

abline(lin.reg,col="blue");

# abline tells R to add least-squares line;
```

A few comments

- Bayesian statistics does inference using the rules of probability directly
- Bayesian statistics is based on a single tool, Bayes' theorem, which finds the posterior density of the parameters, given the data. It combines both the prior information we have given in the prior $g(\theta_1, ..., \theta_p)$ and the information about the parameters contained in the observed data given in the likelihood $f(y_1, ..., y_n|\theta_1, ..., \theta_p)$

A few comments

- It is easy to find the unscaled posterior by posterior proportional to prior times likelihood. The unscaled posterior has all the shape information. However, it is not the exact posterior density. It must be divided by its integral to make it exact.
- Evaluating the integral may be very difficult, particularly if there are lots of parameters. It is hard to find the exact posterior except in a few special cases.

A few comments

Computational Bayesian Statistics is based on developing algorithms that we can use to draw samples from the true posterior, even when we only know the unscaled version. There are two types of algorithms we can use to draw a sample from the true posterior, even when we only know it in the unscaled form. The first type are direct methods, where we draw a random sample from an easily sampled density, and reshape this sample by only accepting some of the values into the final sample, in such a way that the accepted values constitute a random sample from the posterior. These methods quickly become inefficient as the number of parameters increase.

A few comments

The second type is where we set up a Markov chain that has the posterior as its long-run distribution, and letting the chain run long enough so a random draw from the Markov chain is a random draw from the posterior. These are known as Markov chain Monte Carlo (MCMC) methods. The Metropolis-Hastings algorithm and the Gibbs sampling algorithm are the two main Markov chain Monte Carlo methods. The Markov chain Monte Carlo samples will not be independent. There will be serial dependence due to the Markov property. Different chains have different mixing properties. That means they move around the parameter space at different rates. However, a MCMC sample provides an approximation to a random sample from the posterior, one that can be used for inference.