Correlation, Regression, and Analysis of Variance

Article

1 author:

John Burkett
University of Rhode Island
27 PUBLICATIONS **238** CITATIONS

SEE PROFILE

Correlation, Regression, and Analysis of Variance

John P. Burkett
burkett@uri.edu

October 28, 2009

Outline

Correlation

- Definitions
- Using R to calculate r
- Exact posterior density of ρ
- Approximate posterior density of ρ
- Hyperbolic-tangent substitution
- Inferences about ρ based on more than one sample

Regression

- Normal linear regression with reference prior
- Sampling from posterior distributions

Correlation I

Definitions

A relationship between two random variables has several aspects—e.g., sign, noisiness, slope, and curvature. Sign and noisiness are expressed by various correlation concepts. Slope and curvature are ignored for now.

One of the most commonly used concepts of correlation is the Pearson product-moment correlation coefficient, which Lee calls simply "the correlation coefficient." If x and y are random variables, their correlation coefficient can be denoted by $\rho(x, y)$ and defined by

$$\rho(x, y) := \frac{\mathcal{C}(x, y)}{\sqrt{\mathcal{V}(x)\mathcal{V}(y)}},$$

where $\mathcal{C}(x, y)$ is the (population) covariance and $\mathcal{V}(x)$ and $\mathcal{V}(y)$ are the (population) variances for x and y.

Correlation II

Definitions

A corresponding *sample* correlation coefficient can be denoted by r and defined by

$$r := \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means. The distribution of the sample correlation coefficient for given values of the population correlation coefficient is tabulated by David (1954) and corrected by Boomsma (1975).

Using R to calculate r I

The R fn `cor` can be used to calculate one or more sample correlation coefficients. It can take one or two arguments. If only one argument is given, it must be a matrix. In this case the fn calculates correlations among columns of the matrix. For example the following code generates a 30×3 matrix and calculates correlation coefficients among columns. The coefficients are displayed in a matrix, the diagonal elements equaling one because each column is perfectly correlated with itself. The matrix is symmetric because correlation is commutative.

```
> X <- cbind(rnorm(30), rnorm(30), rnorm(30))
> cor(X)
```

```
            [,1]       [,2]       [,3]
[1,]  1.0000000 -0.1682179 -0.3255620
[2,] -0.1682179  1.0000000 -0.2819254
[3,] -0.3255620 -0.2819254  1.0000000
```

Using R to calculate *r* II

If the fn `cor` is given two arguments, they may be either vectors or matrices. In either case, the number of rows must be the same for the both arguments. In the case of two vectors, a single correlation coefficient is calculated, as in the following example, where the value of *r* is calculated from 30 observations on each of two variables.

```
> x <- rnorm(30)
> y <- rnorm(30)
> cor(x,y)
```

```
[1] 0.2533441
```

Using R to calculate *r* III

Given two matrices, `cor` calculates the sample correlations between columns of one matrix and columns of the other. The number of columns in the two matrices may differ, as in the following example involving one 30×2 matrix and one 30×3 matrix.

```
> X <- cbind(rnorm(30), rnorm(30))
> Y <- cbind(rnorm(30), rnorm(30), rnorm(30))
> cor(X,Y)
```

```
            [,1]        [,2]        [,3]
[1,] -0.3027033  0.23038758 -0.10563162
[2,] -0.1070201 -0.08059259 -0.01193728
```

Exact posterior density of ρ I

Suppose we have n independent draws (x_i, y_i) from a bivariate normal distribution with parameters $\lambda = E(x)$, $\mu = E(y)$, $\phi = \mathcal{V}(x)$, $\psi = \mathcal{V}(y)$, and $\rho = \rho(x, y)$. A likelihood fn based on the bivariate normal density can be multiplied by a prior joint density $p(\lambda, \mu, \phi, \psi, \rho)$ to obtain a posterior joint density $p(\lambda, \mu, \phi, \psi, \rho|\mathbf{x}, \mathbf{y})$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Integrating this joint density over λ, μ, ϕ, and ψ produces a posterior marginal density $p(\rho|\mathbf{x}, \mathbf{y})$.

The form of $p(\rho|\mathbf{x}, \mathbf{y})$ is derived by Lee (2004) for the case in which $p(\lambda, \mu, \phi, \psi, \rho) \propto p(\rho)/(\phi\psi)$—that is, the joint prior density is the product of five marginal prior densities, including four reference priors ($p(\lambda) = p(\mu) = 1$, $p(\phi) \propto 1/\phi$, and $p(\psi) \propto 1/\psi$), and a general prior density for the correlations coefficient, $p(\rho)$. Given this prior, the posterior density is

$$p(\rho|\mathbf{x}, \mathbf{y}) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty \omega^{-1}(\omega + \omega^{-1} - 2\rho r)^{(n-1)} d\omega.$$

Plotting the posterior density for ρ I

A fn proportional to the posterior density shown on the previous slide can be evaluated for specific choices of $p(\rho)$ and values of n, r, and ρ. As an example, let $p(\rho) \propto 1$, $n = 7$, and $r = .4$. SAGE can be used to evaluate this fn for a range of values for ρ, say from -.95 to .95. The SAGE code to evaluate the fn at $\rho = -.95$ using a symbolic integration procedure is as follows:

```
r = .4
n = 7
f = lambda x: (x^-1)*(x + x^-1 -2*rho*r)^-(n-1)
rho = -.95
post = (1-rho^2)^((n-1)/2)*integral(f(x), x, 0, 1)
round(post, 10)
```

Plotting the posterior density for ρ II

The integral from 0 to 1 (used above) is half that from 0 to ∞ and thus proportional to it, as noted by Lee (2004, p. 162). The code returns the fn value 0.0000013029.

To double check the results, we can ask SAGE to redo the calculations using a numerical rather than symbolic integration procedure. Code for doing that is as follows:

```
numint = numerical_integral(f(x),0,1)
rhofn = (1-rho^2)^((n-1)/2)
rhofn*numint[0]
```

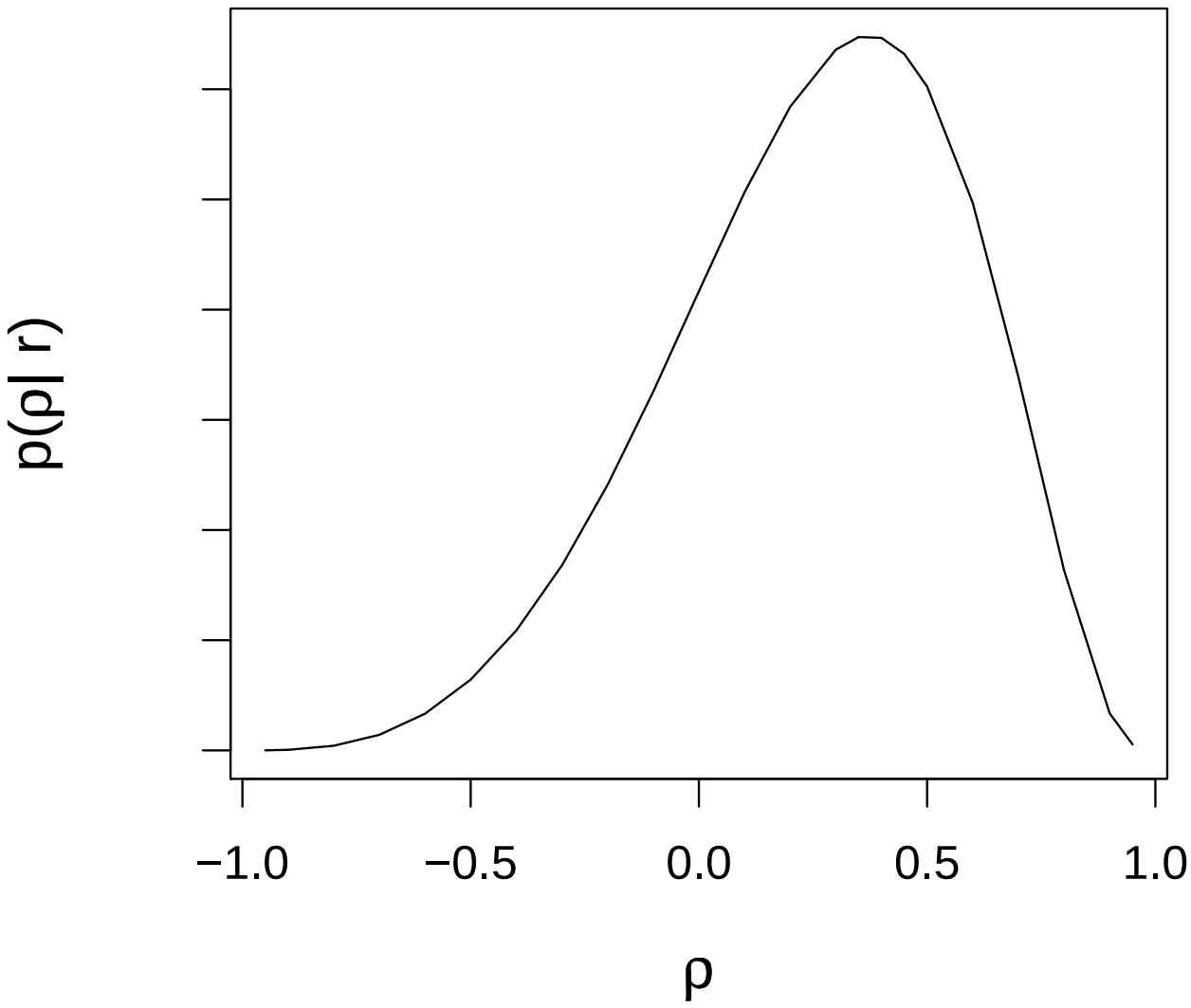
The results of symbolic and numerical integration are the same out to at least 10 decimal places for all 23 values of ρ that I checked.

After evaluating the fn at other values of ρ, we can feed the ρ and fn values to R and use it to plot our results as follows:

Plotting the posterior density for ρ III

```
> rhopost <- read.table("postrho.txt")
> rho <- rhopost[,1]
> fun <- rhopost[,2]
> par(cex=1.3, cex.lab=1.3)
> plot(rho, fun, type="l", axes=FALSE, main=
+  "Posterior density for correlation
+  coefficient when n=7 and r=.4",
+      xlab=expression(rho), ylab=expression(paste(
+     "p(",rho, "| r)")))
> axis(1)
> axis(2, labels=FALSE)
> box()
```

The resulting plot is shown in the next frame.


Posterior density for correlation coefficient when n=7 and r=.4
p(ρ| r)
−1.0
−0.5
0.0
0.5
1.0
ρ

Approximate posterior density of ρ I

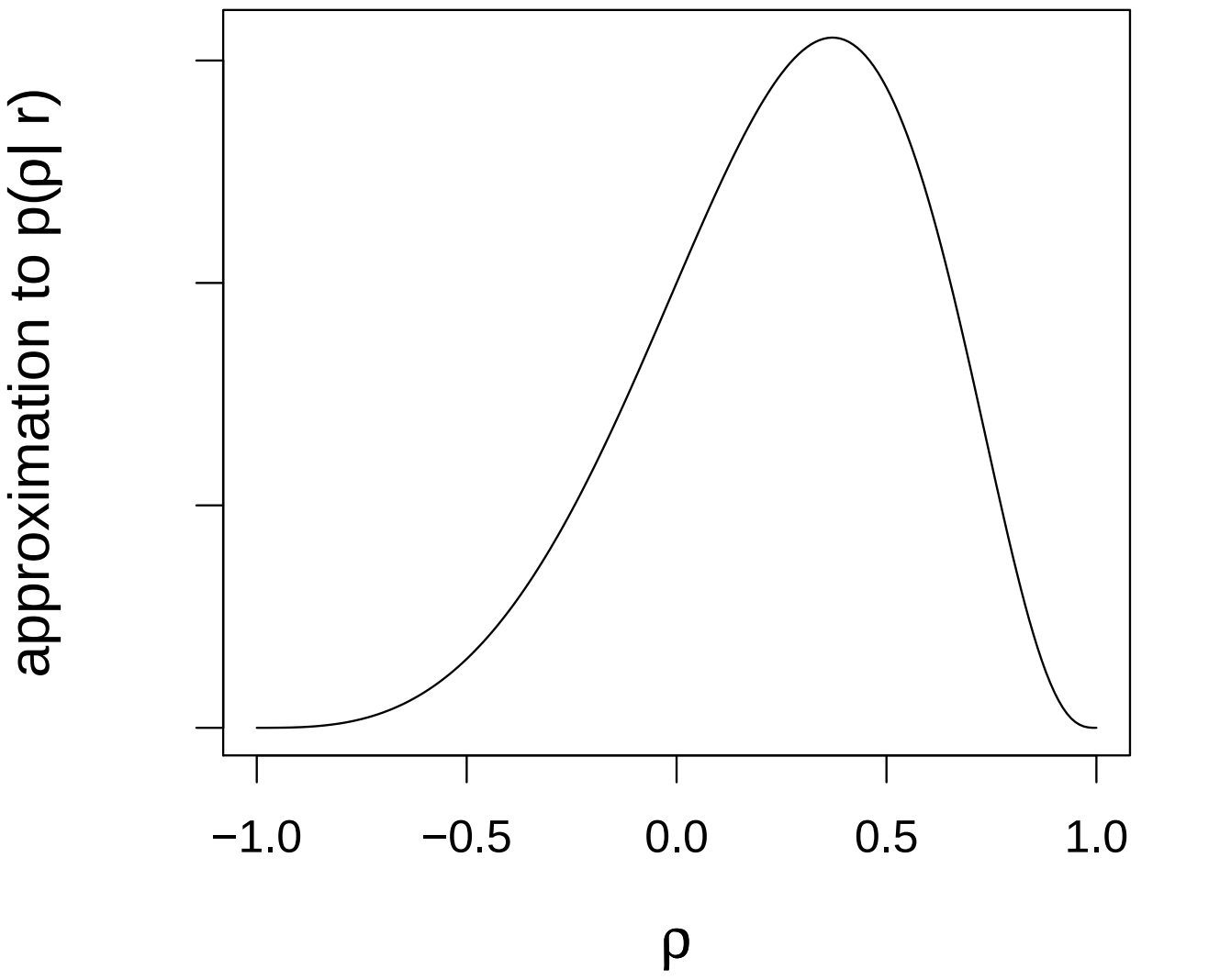
Lee shows that the posterior density of ρ is approximately proportional to

$$p(\rho)\frac{(1-\rho^2)^{(n-1)/2}}{(1-\rho r)^{n-(3/2)}}$$

This approximate density can be plotted with the following R code:

```
> rho <- (seq(0:200)-101)/100
> p <- ((1-rho^2)^3)/(1-rho*.4)^5.5
> par(cex=1.3, cex.lab=1.3)
> plot(rho, p, type='l', axes = FALSE, main=
+  "Approximate posterior density for
+  correlation coefficient when  n=7 and r=.4",
+  xlab=expression(rho), ylab=expression(
+  paste("approximation to p(",rho, "| r)")))
> axis(1); axis(2, labels = FALSE); box()
```

The resulting figure is shown in the next slide.


Approximate posterior density for
correlation coefficient when n=7 and r=.4
approximation to p(ρ| r)
−1.0
−0.5
0.0
0.5
1.0
ρ

Hyperbolic-tangent substitution

Lee suggests transformations of ρ and r that allow inference based on the normal distribution: $\zeta := \text{atanh}(\rho)$ and $z := \text{atanh}(r)$, where $\text{atanh}(\cdot)$ is the inverse of the hyperbolic tangent fn. As ρ ranges over $(-1, 1)$, ζ ranges of $(-\infty, \infty)$. Similarly, as r ranges over$(-1, 1)$, z ranges over $(-\infty, \infty)$.

He further shows that if $p(\rho)$ is uniform, the posterior distribution of ζ can be roughly approximated by

$$N(z, 1/n) \tag{1}$$

and somewhat more closely approximated by

$$N\left[z - \frac{5r}{2n}, \left(n - \frac{3}{2} + \frac{5}{2}(1 - r^2)\right)^{-1}\right] \tag{2}$$

Inferences about ρ based on more than one sample I

The hyperbolic-tangent substitution is particularly useful when we want to use two or more sample correlation coefficients as a basis for inference about ρ. Lee shows how such inference may be based on the rough approximation (1). For purposes of comparison, let us reanalyze Lee's data, using the closer approximation (2). To do this, we can use the following R code:

```
> r <- c(.7, .9)
> n <- c(19, 25)
> z <- atanh(r)
> meanzeta <- z - 5*r/(2*n)
> varzeta <- (n - 3/2 + (5/2)*(1-r^2))^-1
> postvz <- (1/varzeta[1] + 1/varzeta[2])^-1
> postmz <- postvz*sum(meanzeta/varzeta)
> rhohat <- tanh(postmz)
> rhohat
```

Inferences about ρ based on more than one sample II

```
[1] 0.806042
```

```
> ci95zeta <- c(postmz - 1.96*sqrt(postvz), postmz +
+                 1.96*sqrt(postvz))
> ci95rho <- c(tanh(ci95zeta[1]), tanh(ci95zeta[2]))
> ci95rho
```

```
[1] 0.6728080 0.8886345
```

Our point estimate 0.8060 is not very different from Lee's 0.8367. Thus in this case, at least, the two approximations yield similar results.

An application of the rougher approximation to epidemiological research is contained in Schisterman et al. (2003).

Regression

A multiple linear regression indicates how a dependent (response) variable y varies as a fn of k covariates (a.k.a. predictors or explanatory variables) $x_1, \ldots x_k$. If we observe all $k+1$ variables in each of n cases, we can organize our data into a column vector of observations on the dependent variable and a **design matrix** of observations on the covariates:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

The regression model expressed in matrix notation is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters and $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of errors (disturbances).

Normal linear regression with reference prior I

The elements of $\boldsymbol{\varepsilon}$, denoted $\varepsilon_i, i = 1, \ldots, n$ are assumed to be independently and identically distributed (iid) with mean 0 and variance σ^2. The dependent variable, conditional on $\mathbf{X}$ and the parameters, inherits the distribution of the errors. Thus we can write the model as $\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$, where N_n denotes a multivariate normal distribution of dimension n and I denotes an identity matrix.

Bayesians are interested in the posterior distribution of $\boldsymbol{\beta}$ and σ^2. By Bayes's thm, $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2)$, where $p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ is a likelihood fn and $p(\boldsymbol{\beta}, \sigma^2)$ is a prior density. The likelihood implied by the normal linear regression model is based on the $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ distribution. Thus a likelihood fn is

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= (2\pi)^{-n/2}\sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\ &\propto \sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \qquad (3) \end{aligned}$$

Normal linear regression with reference prior II

In terms of the sufficient statistics $\hat{\boldsymbol{\beta}} := (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $s^2 := (n-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, we may write the likelihood fn as

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)s^2 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\right\}.$$

The prior $p(\boldsymbol{\beta}, \sigma^2)$ should reflect any knowledge we have of the parameters. At a minimum, we know that $\sigma^2 > 0$. An otherwise uninformative prior is uniform on $(\boldsymbol{\beta}, \ln \sigma)$ or equivalently $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$. Even if we know more, we may use this prior when $n - k$ is large, or our audience may not share our other prior beliefs, or as a first step in analysis. Following Albert, let us, for the time being, adopt $p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ as our reference prior.

Normal linear regression with reference prior III

Combining a normal likelihood and our reference prior gives us the following posterior pdf:

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) \propto \sigma^{-(n+2)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-k)s^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\}.$$

This joint posterior pdf contains all our postsample information on $\boldsymbol{\beta}$ and σ^2. Understanding this pdf and drawing samples from it is facilitated by factoring it into a conditional density for $\boldsymbol{\beta}$ and a marginal density for σ^2:

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = p(\boldsymbol{\beta}|\sigma^2, \mathbf{y})p(\sigma^2|\mathbf{y}). \tag{4}$$

Normal linear regression with reference prior IV

The posterior predictive density for a future observation $\tilde{y}$ of the dependent variable is

$$\begin{aligned} p(\tilde{y}|\mathbf{y}) &= \int p(\tilde{y}, \boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int p(\tilde{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\boldsymbol{\beta} d\sigma^2. \end{aligned} \tag{5}$$

Equations (4) and (5) are the conceptual basis for Albert's method of sampling from the posterior distributions of the parameters and of future observations.

Sampling from posterior distributions I

A draw from the the posterior distributions of the parameters can be done in two steps:

1. draw a value of σ^2 from its marginal posterior distribution, which is an inverse gamma distribution
2. draw a value of $\boldsymbol{\beta}$ from its posterior distribution conditional on σ^2, which is a multivariate normal distribution.

This sampling procedure can be implemented by using the `blinreg` fn in the LearnBayes package (Albert 2009). To see an example, we can type *?blinreg* and scroll to the bottom of the output to find the following code:

Sampling from posterior distributions II

```
> library(LearnBayes)
> chirps=c(20,16.0,19.8,18.4,17.1,15.5,14.7,17.1,15.4,
+          16.2,15,17.2,16,17,14.1)
> temp=c(88.6,71.6,93.3,84.3,80.6,75.2,69.7,82,69.4,
+        83.3,78.6,82.6,80.6,83.5,76.3)
> X=cbind(1,chirps)
> m=1000
> s=blinreg(temp,X,m)
```

Running this code generates 1000 draws from the posterior distribution of the parameters of a regression of `temp` on `chirps`. To summarize these draws, we can calculate the means and standard deviations for the intercept and slope as follows:

```
> apply(s$beta,2,mean)
```

```
        X   Xchirps
25.901547  3.250042
```

Sampling from posterior distributions III

```
> apply(s$beta,2,sd)
```

```
         X    Xchirps
11.1063882  0.6647408
```

Similarly, the factorization of the posterior predictive density specified in (5) suggests a sampling procedure: Letting $\tilde{y}$ denote a future observation of the dependent variable and letting $\mathbf{x}^*$ denote a row vector whose elements are covariate values corresponding with $\tilde{y}$, we can outline the predictive sampling procedure as follows:

1. draw a value of $(\boldsymbol{\beta}, \sigma^2)$ from the joint posterior distribution
2. draw a value of $\tilde{y}$ from its distribution conditional on $(\boldsymbol{\beta}, \sigma^2)$, which is normal with mean $\mathbf{x}^*\boldsymbol{\beta}$.

The sampling procedures for both parameters and future observations are more fully described in (Gelman et al. 2004, ch. 14).

References

Albert, J. (2009). *Bayesian Computation with R*. Springer, New York, second edition.

Boomsma, A. (1975). An error in F. N. David's tables of the correlation coefficient. *Biometrika*, 62(3):711.

David, F. N. (1954). *Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*. Cambridge University Press, Cambridge.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, London, second edition.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction*. Arnold, London, third edition.

Schisterman, E. F., Moysich, K. B., England, L. J., and Rao, M. (2003). Estimation of the correlation coefficient using the Bayesian approach and its applications for epidemiologic research. *BMC Medical Research Methodology*, 3(5):1–4.