How does a quantity \( y \), vary as a function of another quantity, or vector of quantities \( x \)? We are interested in \( p(y | \theta, x) \) under a model in which \( n \) observations \((x_i, y_i)\) are exchangeable.

**NOTATION.**

- \( y \) (continuous) is the *response* or *outcome* variable;
- \( x = (x_1, \ldots, x_k) \) (discrete or continuous) are the *explanatory variables*;
- We will denote \( y = (y_1, \ldots, y_n) \) the vector of outcomes and \( X \) the \( n \times k \) matrix of explanatory variables.
• The *normal linear model* is a model such that the distribution of $y \mid x$ is a normal whose mean is a linear function of $x$

$$E(y_i \mid \beta, X) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \ i = 1 : n.$$ 

Usually $x_{i1} = 1$. 
Pines example, BUGS vol 2. 42 specimens of radiate pine (Carlin & Chib, 1995 and Williams 1995). For each specimen the maximum compressive strength $y_i$ was measured, with its density $x_i$ and its density adjusted for resin content $z_i$. 

![Graph of strength vs density](image1)

![Graph of adjusted strength vs adjusted density](image2)
Two models can be considered in this case

\[ M_1 := E(y_i|\beta^{(1)}, x) = \beta^{(1)}_1 + \beta^{(1)}_2 x_i \]

\[ M_2 := E(y_i|\beta^{(2)}, z) = \beta^{(2)}_1 + \beta^{(2)}_2 z_i \]

In the previous notation, for model \( M_1 \) we have \( n = 42, k = 2, x_{1i} = 1, x_{2i} = x_i, \beta_1 = \beta^{(1)}_1 \) and \( \beta_2 = \beta^{(1)}_2 \). Similarly, for model \( M_2 \), \( n = 42, k = 2, x_{1i} = 1, x_{2i} = z_i, \beta_1 = \beta^{(2)}_1 \) and \( \beta_2 = \beta^{(2)}_2 \).

We consider ordinary linear regressions in which the conditional variances are equal, i.e., \( V(y_i|\beta^{(1)}, x) = \sigma^2_1 \), \( V(y_i|\beta^{(2)}, z) = \sigma^2_2 \), and the observations are conditionally independent given \( \beta^{(i)}, x \) and \( z \).
**Classical regression analysis.** Let’s work with $M_2$. If $y_i \sim N(\beta_1^{(2)} + \beta_2^{(2)} z_i, \sigma_2^2)$, the maximum likelihood estimator of $\beta^{(2)}$ is given by the solution of $Z^T Z \hat{b} = Z^T Y$, i.e.

$$
\hat{\beta}^{(2)} = (Z^T Z)^{-1} Z^T Y.
$$

Furthermore, $\hat{\beta}^{(2)} \sim N(\beta^{(2)}, \sigma_2^2 (Z^T Z)^{-1})$. The MLE of $\sigma_2^2$ is given by,

$$
\tilde{\sigma}_2^2 = (y - Z\hat{\beta}^{(2)})^T (y - Z\hat{\beta}^{(2)}) / n,
$$

however, this estimator is not unbiased, so an unbiased estimator is given by

$$
\hat{\sigma}_2^2 = (y - Z\hat{\beta}^{(2)})^T (y - Z\hat{\beta}^{(2)}) / (n - k).
$$
We can fit regression models in R.

```r
> pines.linear <- lm(strength ~ adjusted)
Call:
  lm(formula = strength ~ adjusted)

Residuals:
         Min          1Q      Median          3Q         Max
-623.907 -188.821   4.951   197.334   619.691

Coefficients:
             Estimate Std. Error    t value  Pr(>|t|)  
(Intercept)  -1917.639    252.874    -7.583 < 2e-16 ***
adjusted      183.273     9.304     19.698 < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 276.9 on 40 degrees of freedom
Multiple R-Squared: 0.9065,    Adjusted R-squared: 0.9042
F-statistic: 388 on 1 and 40 DF,  p-value: < 2.2e-16
```
Results for linear regression models.

\[ y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I). \]

The MLE of \( \beta \) is given by

\[ \hat{\beta} = (X^T X)^{-1} X^T y. \]

In the example we obtain \( \hat{\beta}^{(2)} = (\beta_1^{(2)}, \beta_2^{(2)}) \) in R as

```r
> coefficients(pines.linear)(Intercept) adjusted
   -1917.6391   183.2733
```

or equivalently,

```r
>X <- matrix(c(rep(1, 42), adjusted), ncol=2, byrow=F)
>y <- strength
>solve(t(X) %*% X) %*% t(X) %*% y
   [,1]
[1,] -1917.6391
[2,]  183.2733
```
\[ \hat{\sigma}^2 = \frac{[(y - X\hat{\beta})^T(y - X\hat{\beta})]}{(n - k)} \]

\[
\begin{align*}
&> \text{beta<-coefficients(pines.linear)} \\
&> \text{sigma.2<-(t(y-X*beta)*%t(y-X*beta))/40} \\
&> \text{sigma.2} \\
&\quad [,1] \\
&\quad [1,] 76661.48 \\
&> \text{sigma<-sqrt(sigma.2)} \\
&> \text{sigma} \\
&\quad [,1] \\
&\quad [1,] 276.8781
\end{align*}
\]

\( \hat{\sigma}^2 \) also corresponds to the MSE in the ANOVA table

\[
> \text{anova(pines.linear)} \\
\text{Analysis of Variance Table} \\
\text{Response: strength} \\
\begin{array}{rrrr}
\text{Df} & \text{Sum Sq} & \text{Mean Sq} & \text{F value} \\
\hline
\text{adjusted} & 1 & 29745989 & 29745989 & 388.02 \\
\text{Residuals} & 40 & 3066459 & 76661 \\
\end{array}
\]
...and $\hat{\sigma}$ is the Residual Standard Error (RSE):

```r
> summary(pines.linear)
...
Residual standard error: 276.9 on 40 degrees of freedom

$\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$. This justifies the following 100$(1-\alpha)$% C.I. for the regression coefficients $\beta_i$,

$$\hat{\beta}_i \pm t_{\alpha/2, n-k} \hat{\sigma} \sqrt{(X^T X)_{ii}^{-1}}$$

A 95% C.I. for $\beta_2^{(2)}$ is given by (164.5, 202.1)

```r
> se.beta.2 <- sigma * sqrt(solve(t(X)%*%X)[2,2]); se.beta.2
   [,1]
[1,] 9.304086

> beta[2] + qt(0.975, 40) * se.beta.2
   [,1]
[1,] 202.0776

> beta[2] - qt(0.975, 40) * se.beta.2
   [,1]
[1,] 164.4691
> summary(pines.linear)

...  
Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -1917.639 | 252.874    | -7.583  | 2.93e-09 |
| adjusted       | 183.273   | 9.304      | 19.698  | < 2e-16  |

An important specific value for the regression coefficients is 0, then, we can test the following hypothesis on each \( \beta_i \)

\[
H_0 : \ \beta_i = 0 \ \ vs \ \ H_1 : \ \beta_i \neq 0
\]

The test statistics is given by

\[
t = \frac{\hat{\beta}_i}{\hat{\sigma} \sqrt{(X^T X)^{-1}}_{ii}},
\]
Rejection region: $|t| > t_{\alpha/2, n-k}$

For $\beta_2^{(2)}$ we have that $t = 19.698$ and we reject $H_0$.

**Prediction.**
The C.I. for the mean response at a point $x = x_0 = (x_1,0, \ldots, x_k,0)^T$ is given by

$$x_0^T \hat{\beta} \pm t_{\alpha/2, n-k} \hat{\sigma} \ast \sqrt{x_0^T (X^T X)^{-1} x_0}.$$
For example, assume you want to predict the mean response of the strength for an adjusted density of 20.5. Then, \( x_0 = (1, 20.5) \) and a 95\% C.I. for the mean response is given by

\[
\begin{align*}
\text{Y ou can also use the function } & \text{predict.lm with} \\
\text{se.fit=TRUE, interval="confidence", type="response"}
\end{align*}
\]
**Prediction.** If, given $x_0$, a specific value needs to be predicted, the confidence interval is given by

$$x_0^T \hat{\beta} \pm t_{\alpha/2, n-k} \hat{\sigma} \times \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

```r
> t(x.0) %*% beta - qt(0.975, 40) * sigma * 
+ sqrt(1.0 + t(x.0) %*% solve(t(X) %*% X) %*% x.0) 
[1,] 1261.036
[1,] 2417.893
```

You can also use the function `predict.lm` with `se.fit=TRUE, interval="prediction", type="response"`
$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ is the proportion of total variation in the data which is explained by the model. $R^2 = 1$ corresponds to a ‘perfect’ fit. $R^2$ can be inflated by a large number of regressors, then an adjusted version is preferred

$$R_a^2 = 1 - \frac{(n - 1)(1 - R^2)}{n - k}.$$ 

We can also compare two nested models using an F-test.

$H_0 : \beta = \beta_0 = (\beta_1, \ldots, \beta_q)^T$

$H_1 : \beta = \beta_1 = (\beta_1, \ldots, \beta_q, \ldots, \beta_p)^T$

where $q < p < n$. 
Let $x_0$ and $x_1$ denote the corresponding design matrices and $\hat{\beta}_0$, $\hat{\beta}_1$ the maximum likelihood estimators. If $H_0$ is correct, then

$$f = \frac{(\hat{\beta}_1^T x_1^T y - \hat{\beta}_0^T x_0^T y) / (p - q)}{(y^T y - \hat{\beta}_1^T x_1^T y) / (n - p)} \sim F_{p-q,n-p}$$

Therefore, values of $f$ that are large relative to the $F_{p-q,n-p}$ provide evidence against $H_0$. 
> anova(pines.linear)
Analysis of Variance Table
Response: Specimen

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model with $\beta_0$</td>
<td>$q$</td>
<td>$\hat{\beta}_0 X_0^T y$</td>
<td></td>
</tr>
<tr>
<td>Improvement due to $\beta_1$</td>
<td>$p - q$</td>
<td>$\hat{\beta}_1 X_1^T y - \hat{\beta}_0 X_0^T y$</td>
<td>$\frac{\hat{\beta}_1 X_1^T y - \hat{\beta}_0 X_0^T y}{p-q}$</td>
</tr>
<tr>
<td>Residual</td>
<td>$n - p$</td>
<td>$y^T y - \hat{\beta}_1 X_1^T y$</td>
<td>$\frac{y^T y - \hat{\beta}_1 X_1^T y}{n-p}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
<td>$y^T y$</td>
<td></td>
</tr>
</tbody>
</table>

In this case, R produces a table to compare the current model $M_2$ with the default model that contains only the intercept.
We will start by considering the standard case of non-informative prior distributions. In this case the Bayesian estimates and the standard errors coincide with the classical results. Posterior simulations are useful for predictive inference and model checking.

\[ y | \beta, \sigma^2, X \sim N(X\beta, \sigma^2 I) \]

\[ p(\beta, \sigma^2 | X) \propto \sigma^{-2} \]

If there are many data points and a few parameters, this prior is useful (it gives reasonable results and takes less effort than specifying prior knowledge). The posterior distribution.

\[ p(\beta, \sigma^2 | y) = p(\beta | \sigma^2, y)p(\sigma^2 | y) \]
Conditional posterior of $\beta$. 

$$\beta | \sigma^2, y \sim N(\hat{\beta}, V_\beta \sigma^2)$$

with $\hat{\beta} = (X^T X)^{-1} X^T y$ and $V_\beta = (X^T X)^{-1}$.

Marginal posterior of $\sigma^2$. 

$$p(\sigma^2 | y) = \frac{p(\beta, \sigma^2 | y)}{p(\beta | \sigma^2, y)}$$

$$\sigma^2 | y \sim Inv - \chi^2(n - k, \hat{\sigma}^2),$$

Checking that the posterior is proper. $p(\beta, \sigma^2 | y)$ is proper if 

1. $n > k$
2. the rank of $X$ equals $k$ (i.e. columns of $X$ are l.i.)
Sampling from the posterior

1. Compute $\hat{\beta}$ and $V_\beta$. These quantities can be computed using a standard software. If computational efficiency is important, then

   (a) Compute the $QR$ factorization of $X = QR$, where $Q$ is an $n \times k$ matrix of orthonormal columns and $R$ is a $k \times k$ upper triangular matrix.

   (b) Compute $R^{-1}$ (easy). $R^{-1}$ is a Cholesky factor (square root) of the covariance matrix $V_\beta$, since $R^{-1}(R^{-1})^T = V_\beta$.

   (c) Compute $\hat{\beta}$ by solving the linear system $R\hat{\beta} = QTy$, using the fact that $R$ is upper triangular.

2. Compute $\hat{\sigma}^2$. Draw $\sigma^2$ from $p(\sigma^2|y)$

3. Draw $\beta$ from $N(\hat{\beta}, \sigma^2 V_\beta)$
Suppose we have observed a new set of explanatory variables $\tilde{X}$ and we want to predict the outcomes $\tilde{y}$ using the regression model.

Components of uncertainty in $p(\tilde{y}|y)$

- variability of the model, represented by $\sigma^2$ and not accounted for by $X\beta$
- posterior uncertainty in $\beta$ and $\sigma^2$ due to the finite sample size of $y$. As $n \to \infty$ this uncertainty decreases to zero.

Drawing a sample $\tilde{y}$ from its posterior predictive distribution can be done as follows

1. draw $(\beta, \sigma^2)$ from $p(\beta, \sigma^2|y)$
2. draw $\tilde{y} \sim N(X\beta, \sigma^2I)$
Given $\sigma^2$, the future observation $\tilde{y}$ has a normal distribution and the mean and the variance are given by

$$E(\tilde{y}|y, \sigma^2) = E(E(\tilde{y} | \beta, \sigma^2, y) | \sigma^2, y)$$
$$= E(\tilde{X}\beta | \sigma^2, y)$$
$$= \tilde{X}\hat{\beta}$$

and

$$V(\tilde{y}|\sigma^2, y) = E[V(\tilde{y} | \beta, \sigma^2, y) | \sigma^2, y]$$
$$+ V[E(\tilde{y} | \beta, \sigma^2, y) | \sigma^2, y]$$
$$= E[\sigma^2 I | \sigma^2, y] + V[\tilde{X}\beta | \sigma^2, y]$$
$$= (I + \tilde{X}V\beta\tilde{X}^T)\sigma^2$$
To determine $p(\tilde{y}|y)$ we must average over the marginal posterior of $\sigma^2$, then, $p(\tilde{y}|y)$ is multivariate $t$ with center $\hat{\beta}$, squared scale matrix $\hat{\sigma}^2(I + \tilde{X}V\hat{\beta}\tilde{X}^T)$. 
Example 8.1

The table gives short-term radon measurements for a sample of houses in three counties in Minnesota. All the measurements were recorded on the basement level of the houses, except for those indicated with *, which were recorded on the first floor.

<table>
<thead>
<tr>
<th>County</th>
<th>Radon measurements (pCi/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Earth</td>
<td>5.0, 13.0, 7.2, 6.8, 12.8, 5.8*, 9.5, 6.0, 3.8, 14.3*, 1.8, 6.9, 4.7, 9.5</td>
</tr>
<tr>
<td>Clay</td>
<td>0.9*, 12.9, 2.6, 3.5*, 26.6, 1.5, 13.0, 8.8, 19.5, 2.5*, 9.0, 13.1, 3.6, 6.9*</td>
</tr>
<tr>
<td>Goodhue</td>
<td>14.3, 6.9*, 7.6, 9.8*, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7</td>
</tr>
</tbody>
</table>
We want to fit a linear regression to the data (or to a transformation of the data) with indicator variables for the three counties and for whether a measurement was recorded on the first floor. In addition, we would like to be able to make predictions.
We can define a model in terms of indicator variables as follows,

\[
x_2 = \begin{cases} 
1 & y_{i,j} \in B.E. \\
0 & y_{i,j} \notin B.E.
\end{cases}, \quad x_3 = \begin{cases} 
1 & y_{i,j} \in C' \\
0 & y_{i,j} \notin C'
\end{cases}, \quad z = \begin{cases} 
1 & y_{i,j} \in F.F \\
0 & y_{i,j} \notin F.F
\end{cases},
\]

for all \( i = 1, 2, 3, \) and \( j = 1, \ldots, n_i \). In addition, \( R \) will automatically add an indicator variable for the intercept, i.e. \( x_1 = 1 \) for all the observations. Then, the model can be written in the following form,

\[
\log(y_{i,j}) = \mu + \alpha_i + \delta + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim N(0, \sigma^2),
\]

with \( \mu \) the mean effect for Goodhue, \( \alpha_1 \) the effect of Blue Earth over \( \mu \), \( \alpha_2 \) the effect of Clayton over \( \mu \), \( \alpha_3 = 0 \), and \( \delta \) the effect of the first floor.
Equivalently, we can write

\[ y = [x_1, x_2, x_3, z]\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I), \]

with \( \beta = (\beta_1, \beta_2, \beta_3, \beta_4) = (\mu, \alpha_1, \alpha_2, \delta) \). Here is another way of writing a regression model,

\[ x_1 = \begin{cases} 
1 & y_{i,j} \in B.E. \\
0 & y_{i,j} \notin B.E. 
\end{cases}, \quad x_2 = \begin{cases} 
1 & y_{i,j} \in C \\
0 & y_{i,j} \notin C 
\end{cases}, \]

\[ x_3 = \begin{cases} 
1 & y_{i,j} \in G. \\
0 & y_{i,j} \notin G. 
\end{cases}, \quad x_4 = \begin{cases} 
1 & y_{i,j} \in F.F. \\
0 & y_{i,j} \notin F.F. 
\end{cases}, \]

\[ i = 1, 2, 3, \quad j = 1, \ldots, n_i. \] In this case we don’t need the indicator variable for the intercept.
This model can be written as

$$\log(y_{i,j}) = \theta_i + \delta + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim N(0, \sigma^2),$$

with $\theta_i$ the mean effects for the three counties and $\delta$ the effect of the first floor. Equivalently, we can write

$$y = [x_1, x_2, x_3, x_4] \beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I),$$

with $\beta = (\beta_1, \beta_2, \beta_3, \beta_4) = (\theta_1, \theta_2, \theta_3, \delta)$. 
#The file radon.data has the following form
#5.0  1  0  0
#13.0 1  0  0
#7.2  1  0  0
#6.8  1  0  0
#12.8 1  0  0
#5.8  1  0  1
#9.5  1  0  0
#.
#.
#.
#0.9  0  1  1
#12.9 0  1  0
#2.6  0  1  0
#3.5  0  1  1
#.
#.
#.
#3.9  0  0  0
#6.7  0  0  0
# 1. read in the data
data <- matrix(scan("radon.data"), ncol=4, byrow=t)
x <- cbind(data[,2], data[,3], data[,4])  # design matrix
# note: x1 = intercept (not in design matrix - splus will
#       add it automatically!)
n
# x2 = indicator for blue earth
# x3 = indicator for clay
# x4 = indicator for first floor
k <- ncol(x)+1  # number pars (+1 for intercept)
n <- length(y)  # number of observations

### 2. get the posterior moments, i.e. beta-hat, v-beta and s2
lf <- lsfit(x, y)  # get the m.l.e. beta-hat which is also
#                   # the posterior mean for (beta|sigma,y)
bhat <- lf$coef  # beta hat
print(bhat)
# Intercept    X1    X2    X3
# 1.918    0.038  -0.042  -0.328
The file radon.data.2 has the following form:

5.0 1 0 0 0
13.0 1 0 0 0
7.2 1 0 0 0
6.8 1 0 0 0
12.8 1 0 0 0
5.8 1 0 0 1
9.5 1 0 0 0
0.9 0 1 0 1
12.9 0 1 0 0
2.6 0 1 0 0
3.5 0 1 0 1
3.9 0 0 1 0
6.7 0 0 1 0
# 1. read in the data
data <- matrix(scan("radon.data.2"), ncol=5, byrow=T)
y <- log(data[,1])    # y is the last data column
X <- cbind(data[,2], data[,3], data[,4], data[,5])  # design matrix
#    x1 = indicator for Blue Earth
#    x2 = indicator for Clay
#    x3 = indicator for Goodhue
#    x4 = indicator for first floor
k <- ncol(X)     # number pars
n<- length(y)    # number of observations
# 2. get the posterior moments, i.e. beta-hat, V-beta and s2
lf <- lsfit(y ~ X, intercept=FALSE)  # m.l.e. beta-hat
bhat <- lf$coef    # beta hat
print(bhat)
#      X1       X2       X3       X4
# 1.9561320 1.8763268 1.9181916 -0.3283398
Example

# 3. Get V-beta
R <- qr.R(1f$qr)  # QR decomp of X
L <- solve(R)     # chol decomp of V_beta

# 4. now generate (sig2,beta) ~ p(sig2,beta|y)
#               =p(sig2|y) * p(beta|sig2,y)
#
# generate sig2 ~ Inv-Chi(n-k,s2)
sig2 <- (n-k)*s2/rchisq(1000,n-k)
# and beta ~ N(bhat, sig2*LL’)
# (i) z[j] ~ N(0,sig2[j]*I), columns j=1...1000
z <- matrix(rnorm(4000,sd=sqrt(sig2)),byrow=T,ncol=1000)
# (ii) b[j] = bhat+L*z[j], columns j=1...1000
b <- bhat + L %*% z
### 5. various summaries and plots

```r
apply(b,1,summary)

Min.   1.223 -0.99520 -1.16500 -1.46700
1st Qu. 1.770 -0.16910 -0.25240 -0.53940
Median 1.917  0.05053 -0.03672 -0.33980
Mean   1.918  0.04653 -0.03667 -0.33410
3rd Qu. 2.068  0.25640  0.17590 -0.12780
Max.   2.694  0.99890  0.84530  1.09800
```

```r
boxplot(b[1,],b[2,],b[3,],b[4,],
names=c("CONST","BLUE E","CLAY","1ST FL"))
abline(h=0)
boxplot(b[1,]+b[2,],b[1,]+b[3,],b[1,],
names=c("BLUE","CLAY","GOODH"))
c(t <- cbind(b[1,]+b[2,], b[1,]+b[3,], b[1,])
sum(b[2,]>0)/1000  # = 0.552
sum(b[3,]>0)/1000  # = 0.447
sum(b[2,]-b[3,]>0)/1000 # 0.597
```
Prediction

Assume another house is sampled at random from Blue Earth County. We have two scenarios depending on whether the measurement we want to predict will be recorded on the basement or on the first floor. If we want to predict a basement measurement, we need to sample $y_{rep}$ from the posterior predictive distribution $N(\mu + \alpha_1, \sigma^2)$. If we want a prediction for a first-floor measurement, then we need to sample $y_{rep}$ from the posterior predictive distribution $N(\mu + \alpha_1 + \delta, \sigma^2)$.

<table>
<thead>
<tr>
<th>Location</th>
<th>95% P.I</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basement</td>
<td>(0.526,29.663)</td>
<td>7.152</td>
</tr>
<tr>
<td>First floor</td>
<td>(0.266,20.994)</td>
<td>5.012</td>
</tr>
</tbody>
</table>