HOMEWORK 3

Handed out: Friday, Sept. 20 2017 Due: Wednesday, Oct. 4 midnight

Solution 1. Load the 33×11 matrix contained in the 'caterpillar.mat' file, in which the first 10 columns are the 10 explanatory variables $x_i, i = 1, \ldots, 10$, and the last column is the scalar output y. The semi-log-y plots are shown in Figure 1. For implementation details consult the MatLab or Python implementations.

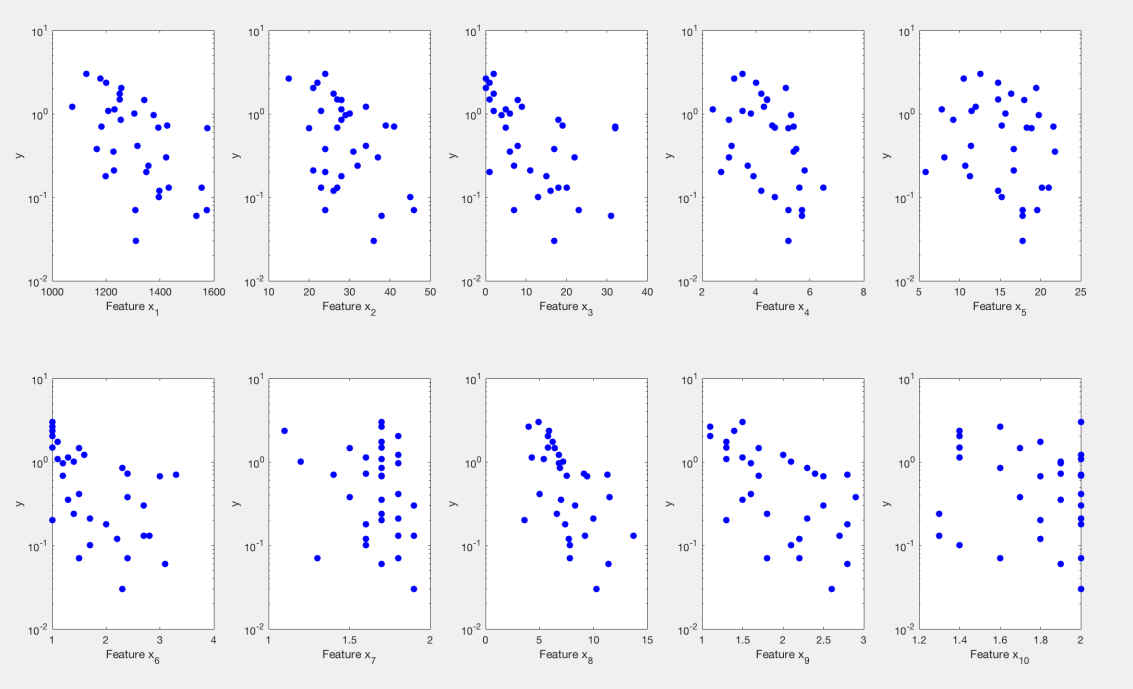

Figure 1: Problem 1 - Semi-log-y plots.

Solution 2. The likelihood for the *ordinary normal linear model* is:

$$y|\beta, \sigma^2, X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n),$$

or explicitly in terms of the parameters β, σ^2 under this model can be written as

$$l(\beta, \sigma^2|y, X) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right].$$

Observe the expression and it is clear that the MLE estimate of β is the one that minimizes the term $(y - X\beta)^T(y - X\beta)$, which can be obtained by the orthogonal projection of y to the

linear subspace spanned by the columns of X, i.e., $\hat{\beta} = (X^T X)^{-1} X^T y$. Plug in the estimated $\hat{\beta}$ back into the likelihood expression and minimize w.r.t. σ^2. The unbiased estimator of σ^2 can be similarly obtained as $\hat{\sigma^2} = \frac{s^2}{n-k-1}$.

The t-values $t_{values(i)}$ can be then defined as

$$t_{values(i)} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 \omega_{(i,i)}}}$$

where $\omega_{(i,i)} = (X^T X)^{-1}_{(i,i)}$. Then the null sampling distribution of this t-statistics is $T_i \sim \mathcal{T}(n-k-1, 0, 1)$. We can then compute the p-value as the areas under the null sampling distribution more extreme (farther from zero) than the t-values. The smaller the p-value is, the stronger we have evidence against the null hypothesis, so we can typically reject the null hypothesis if p exceeds 0.05. The area can be computed with the CDF of the standard t-distribution:

$$\begin{aligned} p_{values(i)} =& \mathbb{P}\big(|T_i| > |t_{values(i)}|\big) \\ =& \mathbb{P}\big(T_i > |t_{values(i)}|\big) + \mathbb{P}\big(T_i < -|t_{values(i)}|\big) \\ =& 2[1 - \mathbb{F}_{T_i}\big(|t_{values(i)}|\big)], \end{aligned}$$

where $\mathbb{F}_{T_i}$ is the CDF of the respective distribution, and $|.|$ is the absolute value. The table should be identical as the one shown in Figure 2 (Fig. 3.2 of the provided reference).

Solution 3. In this problem, we employ the conjugate prior:

$$\begin{aligned} \beta|\sigma^2, X &\sim \mathcal{N}_{k+1}\big(\tilde{\beta}, \sigma^2 M^{-1}\big) \\ \sigma^2|X &\sim \mathcal{IG}(a, b), \end{aligned}$$

where $\tilde{\beta} = 0_{k+1}$ thus every term involving its multiplication is neglected, $M = I_k/c$ with c $= 100$, $a = 2.1$, $b = 2.0$. Since we have 10 columns for X, $k = 10$. The prior is conjugate, leading us to the following normal-inverse-gamma posterior distribution:

```
Residuals:
      Min         1Q     Median         3Q        Max
-1.6989839 -0.2731726 -0.0003620  0.3246311  1.7304969

Coefficients:
            Estimate  Std. Error  t-value Pr(>|t|)
intercept  10.998412   3.060272   3.594  0.00161 **
XV1        -0.004431   0.001557  -2.846  0.00939 **
XV2        -0.053830   0.021900  -2.458  0.02232 *
XV3         0.067939   0.099472   0.683  0.50174
XV4        -1.293636   0.563811  -2.294  0.03168 *
XV5         0.231637   0.104378   2.219  0.03709 *
XV6        -0.356800   1.566464  -0.228  0.82193
XV7        -0.237469   1.006006  -0.236  0.81558
XV8         0.181060   0.236724   0.765  0.45248
XV9        -1.285316   0.864847  -1.486  0.15142
XV10       -0.433106   0.734869  -0.589  0.56162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 3.2. Dataset caterpillar: R output providing the maximum likelihood estimates of the regression coefficients and their standard significance analysis.

Figure 2: Table of Problem 2.

$$\beta|\sigma^2, y, X \sim \mathcal{N}_{k+1}\left((M + X^TX)^{-1}X^TX\hat{\beta}, \sigma^2[M^{-1} + (X^TX)^{-1}]^{-1}\right)$$
$$\sigma^2|y, X \sim \mathcal{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{\hat{\beta}^T[M^{-1} + (X^TX)^{-1}]^{-1}\hat{\beta}}{2}\right).$$

The needed developments are as follows:

$$\begin{aligned}
\mathbb{E}_\pi(\sigma^2|y, X) &= \int \sigma^2 \cdot \pi(\sigma^2|y, X)\mathrm{d}\sigma^2 \\
&= \frac{2b + s^2 + \hat{\beta}^T[M^{-1} + (X^TX)^{-1}]^{-1}\hat{\beta}}{n + 2a - 2}, \\
\mathbb{E}_\pi(\beta|y, X) &= \int_{\sigma^2} \mathbb{E}_\pi(\beta|\sigma^2, y, X)\pi(\sigma^2|y, X)\mathrm{d}\sigma^2 \\
&= (M + X^TX)^{-1}X^TX\hat{\beta} \cdot \int_{\sigma^2} \pi(\sigma^2|y, X)\mathrm{d}\sigma^2 \\
&= (M + X^TX)^{-1}X^TX\hat{\beta}.
\end{aligned}$$

For $\mathbb{V}_\pi$, we have to resort to $\pi(\beta|y,X)$:

$$\begin{aligned}\pi(\beta|y,X) &= \int_{\sigma^2} \pi(\beta|\sigma^2,y,X)\pi(\sigma^2|y,X)\mathrm{d}\sigma^2 \\ &= \int_{\sigma^2} \left[\mathcal{N}_{k+1}\left((M+X^TX)^{-1}X^TX\hat{\beta}, \sigma^2[M^{-1}+(X^TX)^{-1}]^{-1}\right)\right. \\ &\left.\times\mathcal{IG}\left(\frac{n}{2}+a, b+\frac{s^2}{2}+\frac{\hat{\beta}^T[M^{-1}+(X^TX)^{-1}]^{-1}\hat{\beta}}{2}\right)\right]\mathrm{d}\sigma^2.\end{aligned}$$

Organizing the terms inside, it can be written as a t-distribution

$$\beta|y,X \propto \mathcal{T}_{2k+1}\left(n+2a, \hat{\mu}, \hat{\Sigma}\right),$$

with

$$\begin{aligned}\hat{\mu} =& (M+X^TX)^{-1}(X^TX\hat{\beta}), \\ \hat{\Sigma} =& \frac{2b+s^2+\hat{\beta}^T[M^{-1}+(X^TX)^{-1}]^{-1}\hat{\beta}}{n+2a}(M+X^TX)^{-1}.\end{aligned}$$

Since the posterior distribution $\pi(\beta|y,X)$ is a student-t distribution, its posterior variance is simply $\mathbb{V}_\pi(\beta|y,X) = \frac{n+2a}{n+2a-2}\hat{\Sigma}$. The table should be identical as the one shown in Figure 3. See Section 3.2.1 on the provided 'Bayesian Core' reference for a more detailed procedure.

Solution 4. Consider the Zellner's informative G-prior:

$$\begin{aligned}\beta|\sigma^2,X &\sim \mathcal{N}_{k+1}\big(\tilde{\beta}, c\sigma^2(X^TX)^{-1}\big), \\ \sigma^2|X &\propto \sigma^{-2},\end{aligned}$$

where $\tilde{\beta} = 0_{k+1}$ with $c = 100$ or 1000. The posterior distribution can be easily derived as:

$$\begin{aligned}\pi(\beta,\sigma^2|y,X) \propto& f(y|\beta,\sigma^2,X)\pi(\beta,\sigma^2|X) \\ \propto& (\sigma^2)^{-n/2-1}\exp\left[-\frac{1}{2\sigma^2}(y-X\hat{\beta})^T(y-X\hat{\beta}) - \frac{1}{2\sigma^2}(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})\right] \\ &\cdot(\sigma^2)^{-k/2}\exp\left[-\frac{1}{2c\sigma^2}(\beta^TX^TX\beta)\right].\end{aligned}$$

Conduct marginalization and we can derive the conditional and marginal posteriors on β

Table 3.1. Dataset caterpillar: Influence of the prior scale c on the Bayes estimates of σ^2 and β_0.

c	$\mathbb{E}(\sigma^2\|\mathbf{y}, X)$	$\mathbb{E}(\beta_0\|\mathbf{y}, X)$	$\mathbb{V}(\beta_0\|\mathbf{y}, X)$
.1	1.0044	0.1251	0.0988
1	0.8541	0.9031	0.7733
10	0.6976	4.7299	3.8991
100	0.5746	9.6626	6.8355
1000	0.5470	10.8476	7.3419

Table 3.2. Dataset caterpillar: Bayes estimates of β for $c = 100$.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y}, X)$	$\mathbb{V}(\beta_i\|\mathbf{y}, X)$
β_0	9.6626	6.8355
β_1	–0.0041	2×10^{-6}
β_2	–0.0516	0.0004
β_3	0.0418	0.0076
β_4	–1.2633	0.2615
β_5	0.2307	0.0090
β_6	–0.0832	1.9310
β_7	–0.1917	0.8254
β_8	0.1608	0.0046
β_9	–1.2069	0.6127
β_{10}	–0.2567	0.4267

Figure 3: Table of Problem 3.

and σ^2:

$$\beta|\sigma^2, y, X \sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}\hat{\beta}, \frac{\sigma^2 c}{c+1}(X^T X)^{-1}\right),$$
$$\sigma^2|y, X \sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(c+1)}\hat{\beta}^T X^T X \hat{\beta}\right).$$

Further integrating out σ^2 on the conditional posterior of β, we can show that

$$\beta|y, X \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\hat{\beta}, \frac{c\left(s^2 + \frac{\hat{\beta}^T X^T X \hat{\beta}}{c+1}\right)}{n(c+1)}(X^T X)^{-1}\right),$$

which implies that we can use the analytical mean/variance expression of student-t distribution to compute the marginal moments for β:

$$\begin{aligned}\mathbb{E}[\beta|y,X] =& \frac{c}{c+1}\hat{\beta},\\ \mathbb{V}[\beta|y,X] =& \frac{c\big(s^2+\frac{\hat{\beta}^T X^T X\hat{\beta}}{c+1}\big)}{(n-2)(c+1)}(X^TX)^{-1},\end{aligned} \tag{1}$$

where we can see the interplay between c and the posterior moments of β.

For the Bayes factor, we need to compute the evidence $f(y|X,H_1 \text{ or } X_0,H_0)$. Here H_0 is the null hypothesis. Observe the prior

$$\beta|\sigma^2,X \sim \mathcal{N}_{k+1}\big(\tilde{\beta},c\sigma^2(X^TX)^{-1}\big),$$

multiplying X on both sides, it's a linear transformation therefore

$$X\beta|\sigma^2,X \sim \mathcal{N}_n\big(X\tilde{\beta},c\sigma^2X(X^TX)^{-1}X^T\big),$$

we already know

$$y|\beta,\sigma^2,X \sim \mathcal{N}_n(X\beta,\sigma^2I_n) \to y-X\beta|\sigma^2,X \sim \mathcal{N}_n(0_n,\sigma^2I_n).$$

Hence

$$y|\sigma^2,X \sim \mathcal{N}_n\big(X\tilde{\beta},c\sigma^2[X(X^TX)^{-1}X^T+I_n]\big).$$

Integrating out σ^2 with its prior $\propto 1/\sigma^2$ yields

$$f(y|X,H_1) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(n/2)\left[y^Ty-\frac{c}{c+1}y^TX(X^TX)^{-1}X^Ty\right]^{-n/2}.$$

For X_0 and H_0 which correspond to the null model, the prior changes to

$$\beta^0|\sigma^2,X_0 \sim \mathcal{N}_{k+1-q}\big(\tilde{\beta}^0,c_0\sigma^2(X_0^TX_0)^{-1}\big).$$

Similarly we get

$$f(y|X_0,H_0) = (c_0+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2)\left[y^Ty-\frac{c_0}{c_0+1}y^TX_0(X_0^TX_0)^{-1}X_0^Ty\right]^{-n/2}.$$

The null model (H_0) effectively eliminates q explanatory variable comparing with the full model H, and only relies on $k+1-q$ variables, denoted as X_0. Therefore the Bayes factor

can be given a closed form:

$$B_{10}^{\pi} = \frac{f(y|X, H_1)}{f(y|X_0, H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \times \left[\frac{y^T y - \frac{c_0}{c_0+1} y^T X_0 (X_0^T X_0)^{-1} X_0^T y}{y^T y - \frac{c}{c+1} y^T X (X^T X)^{-1} X^T y} \right]^{n/2}.$$

We consider to compare the full model with all variables (k=10) to the null models with one of them removed ($q = 1$, and c_0 is chosen to be equal to c). Doing so for each of the explanatory variables, we can get Table 3.3−4 of the provided reference as shown in Figure 4. See Section 3.2.2 on 'Bayesian Core' for a more detailed procedure.

Table 3.3. Dataset caterpillar: Posterior mean and variance of β for $c = 100$ using Zellner's G-prior.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y}, X)$	$\mathbb{V}(\beta_i\|\mathbf{y}, X)$
β_0	10.8895	6.4094
β_1	–0.0044	2×10^{-6}
β_2	–0.0533	0.0003
β_3	0.0673	0.0068
β_4	–1.2808	0.2175
β_5	0.2293	0.0075
β_6	–0.3532	1.6793
β_7	–0.2351	0.6926
β_8	0.1793	0.0383
β_9	–1.2726	0.5119
β_{10}	–0.4288	0.3696

Table 3.4. Dataset caterpillar: Same legend as Table 3.3 for $c = 1000$.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y}, X)$	$\mathbb{V}(\beta_i\|\mathbf{y}, X)$
β_0	10.9874	6.2604
β_1	–0.0044	2×10^{-6}
β_2	–0.0538	0.0003
β_3	0.0679	0.0066
β_4	–1.2923	0.2125
β_5	0.2314	0.0073
β_6	–0.3564	1.6403
β_7	-0.2372	0.6765
β_8	0.1809	0.0375
β_9	–1.2840	0.5100
β_{10}	–0.4327	0.3670

Figure 4: Table of Problem 4.

Solution 5. Inheriting the same prior framework as in Solution 4, for the HPD regions,

notice that the marginal posterior distribution:

$$\beta|y,X \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1}\hat{\beta}, \frac{c\left(s^2+\frac{\hat{\beta}^T X^T X\hat{\beta}}{c+1}\right)}{n(c+1)}(X^TX)^{-1}\right).$$

As the standard student-t is one symmetrical distribution, we can analytically calculate its 90% HPD region as the two points whose respective CDF is 5% and 95%. Denote them as δ_5 and δ_{95}. The corresponding HPD of β would then be

$$\beta_5 = \frac{c}{c+1}\hat{\beta} + \delta_5 \cdot \frac{c\left(s^2+\frac{\hat{\beta}^T X^T X\hat{\beta}}{c+1}\right)}{n(c+1)}(X^TX)^{-1},$$
$$\beta_{95} = \frac{c}{c+1}\hat{\beta} + \delta_{95} \cdot \frac{c\left(s^2+\frac{\hat{\beta}^T X^T X\hat{\beta}}{c+1}\right)}{n(c+1)}(X^TX)^{-1}.$$

β_i	HPD Interval
β_0	[6.60,15.2]
β_1	[-0.0066,-0.0022]
β_2	[-0.084,-0.023]
β_3	[-0.072,0.207]
β_4	[-2.07,-0.491]
β_5	[0.0831,0.376]
β_6	[-2.54,1.84]
β_7	[-1.64,1.17]
β_8	[-0.152,0.511]
β_9	[-2.48,-0.0607]
β_{10}	[-1.46,0.60]

Note that this table does not match Table 3.5 or the one generated by the reference code, because we adopt the Zellner's informative G-prior in this problem, while in the book the table is generated under the Jeffery's prior. See Section 3.3.1 on 'Bayesian Core' for a more

detailed procedure on reproducing the table with Jeffery's prior.

Solution 6. In the non-informative Zellner's G-prior, we adopt the same setting as the informative Zellner's G-prior, but now we treat c as unknown and assign it a prior, $\pi(c) \propto c^{-1}$. The marginal (w.r.t. c) posterior of β, σ^2 is written as:

$$\pi(\beta, \sigma^2|y, X) = \sum_c \pi(\beta, \sigma^2|y, X, c)\pi(c|y, X).$$

For the last term, using Bayes' rule to re-write $\pi(c|y, X) \propto f(y|X, c)\pi(c)$, we get

$$\pi(\beta, \sigma^2|y, X) \propto \sum_c \pi(\beta, \sigma^2|y, X, c)f(y|X, c)\pi(c)\mathrm{d}c.$$

Conditioned on a specific c, the term $f(y|X, c)$ is similar to what we'll get using Zellner's informative G-prior:

$$f(y|X, c) \propto (c+1)^{-\frac{k+1}{2}} \left[y^T y - \frac{c}{c+1} y^T X (X^T X)^{-1} X^T y \right]^{-\frac{n}{2}},$$

which implies that

$$\begin{aligned} f(y|X) &\propto \sum_c f(y|X, c)\pi(c) \\ &\propto \sum_c c^{-1}(c+1)^{-\frac{k+1}{2}} \left[y^T y - \frac{c}{c+1} y^T X (X^T X)^{-1} X^T y \right]^{-\frac{n}{2}}. \end{aligned}$$

For the posterior mean and variance of β, we integrate out c by summation:

$$\mathbb{E}[\beta|y, X] = \sum_c \mathbb{E}[\beta|y, X, c]\pi(c|y, X).$$

Plug in $\pi(c|y, X) \propto f(y|X, c)\pi(c)$, and account for the normalization:

$$\begin{aligned} \mathbb{E}[\beta|y, X] =& \sum_c \frac{c}{c+1}\hat{\beta}\pi(c|y, X) \\ =& \hat{\beta} \sum_c \left[\frac{c}{c+1} \frac{\pi(c|y, X)}{\sum_c \pi(c|y, X)} \right] = \hat{\beta} \sum_c \left[\frac{c}{c+1} \frac{f(y|X, c)\pi(c)}{\sum_c f(y|X, c)\pi(c)} \right]. \end{aligned}$$

For $\mathbb{V}(\beta|y,X)$, we rely on the variance decomposition or the law of total variance:

$$\mathbb{V}(\beta|y,X) = \mathbb{E}[\mathbb{V}(\beta|y,X,c)|y,X] + \mathbb{V}(\mathbb{E}[\beta|y,X,c]|y,X).$$

We derived the formula for both $\mathbb{V}(\beta|y,X,c)$ and $\mathbb{E}[\beta|y,X,c]$ in Problem 4. Plugging them back into the equation we can expand to an explicit solvable formula. See Section 3.3.2 of 'Bayesian Core' for a more detailed procedure.

Table 3.6. Dataset caterpillar: Posterior mean and variance of β under the noninformative Zellner's G-prior.

β_i	$\mathbb{E}(\beta_i\|\mathbf{y},X)$	$\mathbb{V}(\beta_i\|\mathbf{y},X)$
β_0	9.2714	9.1164
β_1	–0.0037	2×10^{-6}
β_2	–0.0454	0.0004
β_3	0.0573	0.0086
β_4	–1.0905	0.2901
β_5	0.1953	0.0099
β_6	–0.3008	2.1372
β_7	–0.2002	0.8815
β_8	0.1526	0.0490
β_9	–1.0835	0.6643
β_{10}	–0.3651	0.4716

Figure 5: Table of Problem 6.

Solution 7. The answer is given as follows. We inherit the terminology of the reference book: each model $\mathfrak{M}_\gamma$ contains unique combination of explanatory variables, and it is associated with a binary indicator vector $\gamma \in \{0,1\}^{10}$, where the i-th component γ_i=1 means variable included in the model and 0 otherwise. Similarly, X_γ and β_γ indicates the subset of explanatory variables/coefficients used in the model, hence

$$y|\gamma,\beta_\gamma,\sigma^2,X \sim \mathcal{N}_n\Big(X_\gamma\beta_\gamma,\sigma^2 I_n\Big).$$

(a) For Table 3.7, Zellner's G-prior ($c = 100$) is employed. As we have 10 explanatory

variables (and one constant term that is always included), we end up with $2^{10} = 1024$ models. To compute the posterior distribution of model $\pi(\gamma|y, X)$, we start from

$$\pi(\gamma|y, X) \propto f(y|\gamma, X)\pi(\gamma|X).$$

We opt for the uniform prior thus $\pi(\gamma|X) = 2^{-10}$

$$\begin{aligned}\pi(\gamma|y, X) \propto & f(y|\gamma, X). \\ & \propto \int \left(\int f(y|\gamma, \beta, \sigma^2)\pi(\beta|\gamma, \sigma^2)\mathrm{d}\beta \right) \pi(\sigma^2|X)\mathrm{d}\sigma^2 \\ & \propto (c+1)^{-(q_\gamma+1)/2} \left[y^T y - \frac{c}{c+1} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y \right]^{-n/2}.\end{aligned}$$

where q_γ is the number of explanatory variables used in the model $\mathfrak{M}_\gamma$.

Table 3.7. Dataset caterpillar: Most likely models ordered by decreasing posterior probabilities under Zellner's G-prior ($c = 100$).

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0, 1, 2, 4, 5	0.2316
0, 1, 2, 4, 5, 9	0.0374
0, 1, 9	0.0344
0, 1, 2, 4, 5, 10	0.0328
0, 1, 4, 5	0.0306
0, 1, 2, 9	0.0250
0, 1, 2, 4, 5, 7	0.0241
0, 1, 2, 4, 5,8	0.0238
0, 1, 2, 4, 5, 6	0.0237
0,1, 2, 3, 4, 5	0.0232
0, 1, 6, 9	0.0146
0, 1, 2, 3, 9	0.0145
0, 9	0.0143
0, 1, 2, 6, 9	0.0135
0, 1, 4, 5, 9	0.0128
0, 1, 3, 9	0.0117
0, 1, 2, 8	0.0115
0, 1, 8	0.0095
0, 1, 2, 3, 4, 5, 9	0.0090
0, 1, 2, 4, 5, 6, 9	0.0090

Figure 6: Table of Problem 7(a).

(b) For Table 3.8. Zellner's non-informative G-prior ($c = 100$) is employed. Similar to Solution 6, we have to integrate out the unknown c:

$$\pi(\gamma|y, X) \propto \sum_c \pi(\gamma|y, X, c)\pi(c).$$

Plugging in the $\pi(\gamma|y, X, c)$ we derived at part (a)

$$\pi(\gamma|y, X) \propto \sum_c c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[y^T y - \frac{c}{c+1} y^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y\right]^{-n/2}.$$

Table 3.8. Dataset caterpillar: 20 most likely models under Zellner's noninformative G-prior.

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0, 1, 2, 4, 5	0.0929
0, 1, 2, 4, 5, 9	0.0325
0, 1, 2, 4, 5, 10	0.0295
0, 1, 2, 4, 5, 7	0.0231
0, 1, 2, 4, 5, 8	0.0228
0, 1, 2, 4, 5, 6	0.0228
0, 1, 2, 3, 4, 5	0.0224
0, 1, 2, 3, 4, 5, 9	0.0167
0, 1, 2, 4, 5, 6, 9	0.0167
0, 1, 2, 4, 5, 8, 9	0.0137
0, 1, 4, 5	0.0110
0, 1, 2, 4, 5, 9, 10	0.0100
0, 1, 2, 3, 9	0.0097
0, 1, 2, 9	0.0093
0, 1, 2, 4, 5, 7, 9	0.0092
0, 1, 2, 6, 9	0.0092
0, 1, 4, 5, 9	0.0087
0, 1, 2, 3, 4, 5, 10	0.0079
0, 1, 2, 4, 5, 8, 10	0.0079
0, 1, 2, 4, 5, 7, 10	0.0079

Figure 7: Table of Problem 7(b).

(c) For Table 3.9 – 11, Gibbs sampling is employed to sample from $\pi(\gamma|y, X)$, as it is impossible to compute posterior probabilities for the whole 2^k model space when k is large. The key idea here is we can sample the posterior distribution, Gibbs sampler can be naturally applied in such setting. The key idea is that the conditional distribution

$\pi(\gamma_i|y,\gamma_{-i},X)$ is proportional to $\pi(\gamma,|y,X)$. Since y_i is binary, the conditional distribution is easy to be evaluated. Starting from some initial sample γ^0, the Gibbs sampler iteratively generates samples, while in each iteration (t), the sample component $\gamma_i^{(t)}$ is generated according to $\pi(\gamma_i|y,\gamma^{(t)}_{j\in[1,i-1]},\gamma^{(t-1)}_{j\in[i+1,k]},X)$, and the single iteration finishes when all components have been updated. After a large number of iterations (that is, when the sampler is supposed to have converged or, more accurately, when the sampler has sufficiently explored the support of the target distribution), its output can be used to approximate the posterior probabilities empirically:

$$\hat{\pi}(\gamma|y,X)=\left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T}\mathbb{I}_{\gamma^{(t)}=\gamma},$$
$$\hat{\mathbb{P}}(\gamma_i=1|y,X)=\left(\frac{1}{T-T_0+1}\right)\sum_{t=T_0}^{T}\mathbb{I}_{\gamma_i^{(t)}=\gamma},$$

where the first T_0 samples are discarded (also known as burn-in period) as the Gibbs chain might not achieve convergence (stationary) stage yet. We use $T_0=10000$ samples and $T=20000$ samples. See Section 3.5 on 'Bayesian Core' for a more detailed procedure.

Table 3.9. Dataset caterpillar: First-level G-prior model choice with $\tilde{\beta}=0_{11}$ and $c=100$, compared with the Gibbs estimates of the top ten posterior probabilities.

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y},X)$	$\widehat{\pi}(\gamma\|\mathbf{y},X)$
0, 1, 2, 4, 5	0.2316	0.2208
0, 1, 2, 4, 5, 9	0.0374	0.0375
0, 1, 9	0.0344	0.0358
0, 1, 2, 4, 5, 10	0.0328	0.0344
0, 1, 4, 5	0.0306	0.0313
0, 1, 2, 9	0.0250	0.0268
0, 1, 2, 4, 5, 7	0.0241	0.0260
0, 1, 2, 4, 5, 8	0.0238	0.0251
0, 1, 2, 4, 5, 6	0.0237	0.0244
0, 1, 2, 3, 4, 5	0.0232	0.0224

Table 3.10. Dataset caterpillar: Noninformative G-prior model choice compared with Gibbs estimates of the top ten posterior probabilities.

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$	$\widehat{\pi}(\gamma\|\mathbf{y}, X)$
0, 1, 2, 4, 5	0.0929	0.0929
0, 1, 2, 4, 5, 9	0.0325	0.0326
0, 1, 2, 4, 5, 10	0.0295	0.0272
0, 1, 2, 4, 5, 7	0.0231	0.0231
0, 1, 2, 4, 5, 8	0.0228	0.0229
0, 1, 2, 4, 5, 6	0.0228	0.0226
0, 1, 2, 3, 4, 5	0.0224	0.0220
0, 1, 2, 3, 4, 5, 9	0.0167	0.0182
0, 1, 2, 4, 5, 6, 9	0.0167	0.0171
0, 1, 2, 4, 5, 8, 9	0.0137	0.0130

Table 3.11. Dataset caterpillar: First-level ($\tilde{\beta} = 0_{11}$ and $c = 100$) and noninformative G-prior variable inclusion estimates (based on the same Gibbs output as Table 3.9).

γ_i	$\widehat{\mathbb{P}}_c(\gamma_i = 1\|\mathbf{y}, X)$	$\widehat{\mathbb{P}}(\gamma_i = 1\|\mathbf{y}, X)$
γ_1	0.8624	0.8844
γ_2	0.7060	0.7716
γ_3	0.1482	0.2978
γ_4	0.6671	0.7261
γ_5	0.6515	0.7006
γ_6	0.1678	0.3115
γ_7	0.1371	0.2880
γ_8	0.1555	0.2876
γ_9	0.4039	0.5168
γ_{10}	0.1151	0.2609

Figure 8: Tables of Problem 7(c).