Chapter 11.7 Bayesian Inferences with Simple Linear Regression

Jim Albert and Monika Hu

Chapter 11 Simple Linear Regression

Simulate fits from the regression model

- The intercept β_0 and slope β_1 determine the linear relationship between the mean of the response Y and the predictor x

$$E(Y) = \beta_0 + \beta_1 x \tag{1}$$

- Each pair of values (β_0, β_1) corresponds to a line $\beta_0 + \beta_1 x$ in the space of values of x and y
- Posterior means: $\tilde{\beta}_0$ and $\tilde{\beta}_1$
- The line

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

  corresponds to a “best" line of fit through the data

Simulate fits from the regression model cont'd

- This best line represents a most likely value of the line $\beta_0 + \beta_1 x$ from the posterior distribution
- How about the uncertainty of this line estimate?
- We can draw a sample of J rows from the matrix of posterior draws of (β_0, β_1) and collecting the line estimates

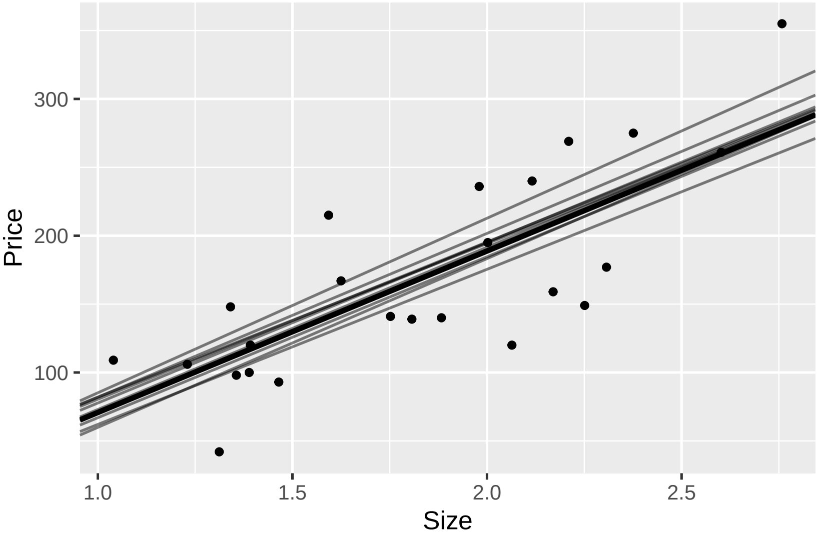
$$\tilde{\beta}_0^{(j)} + \tilde{\beta}_1^{(j)} x,$$

where $j = 1, ..., J$

Simulate fits from the regression model cont'd

```
post <- as.mcmc(posterior)
post_means <- apply(post, 2, mean)
post <- as.data.frame(post)
ggplot(PriceAreaData, aes(newsize, price)) +
  geom_point(size=3) +
  geom_abline(data=post[1:10, ],
              aes(intercept=beta0, slope=beta1),
              alpha = 0.5) +
  geom_abline(intercept = post_means[1],
              slope = post_means[2],
              size = 2) +
  ylab("Price") + xlab("Size") +
  theme_grey(base_size = 18, base_family = "")
```

Simulate fits from the regression model cont'd



- Variation among the ten fits
- What happens with a larger sample size?

Learning about the expected response

- Learn about the expected response $E(Y)$ for a specific value of the predictor x
- How?
- We can obtain a simulated sample from the posterior of $\beta_0 + \beta_1 x$ by computing this linear function, $E(Y) = \beta_0 + \beta_1 x$, on each of the simulated pairs from the posterior of (β_0, β_1)

Learning about the expected response cont'd

- Suppose we are interested in the expected price $E(Y)$ for a house with a size of 1, i.e. $x = 1$ (1000 sq feet)

```
size <- 1
mean_response <- post[, "beta0"] +
  size * post[, "beta1"]
```

Learning about the expected response cont'd

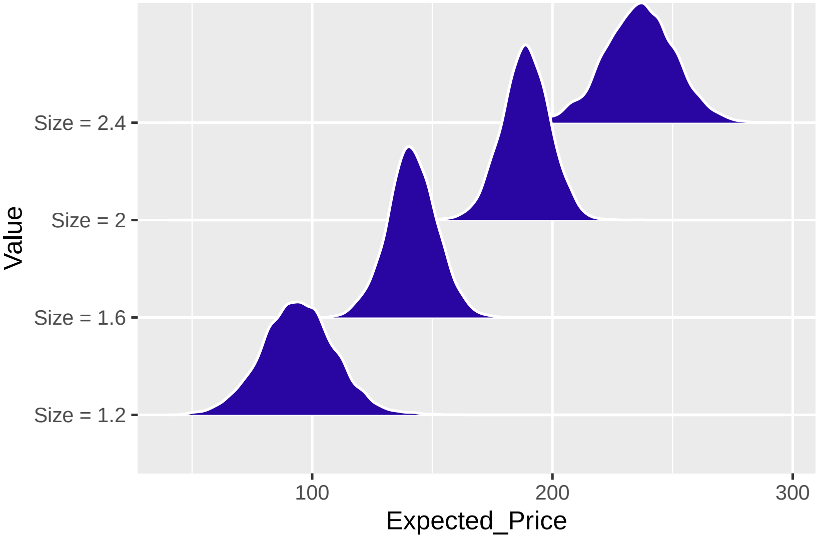
```
one_expected <- function(x){
  lp <- post[ , "beta0"] +  x * post[ , "beta1"]
  data.frame(Value = paste("Size =", x),
             Expected_Price = lp)
}
df <- map_df(c(1.2, 1.6, 2.0, 2.4), one_expected)

ggplot(df, aes(x = Expected_Price, y = Value)) +
  geom_density_ridges(fill = crcblue,
                      color = "white") +
  theme_grey(base_size = 18, base_family = "")
```

- Density plots of the simulated posterior samples for the expected prices $E(Y \mid 1.2)$, $E(Y \mid 1.6)$, $E(Y \mid 2.0)$, $E(Y \mid 2.4)$ for these four house sizes.

Learning about the expected response cont'd

```
## Picking joint bandwidth of 2.03
```



Learning about the expected response cont'd

```
df %>% group_by(Value) %>%
  summarize(P05 = quantile(Expected_Price, 0.05),
            P50 = median(Expected_Price),
            P95 = quantile(Expected_Price, 0.95))
```

```
## # A tibble: 4 x 4
##   Value          P05   P50   P95
##   <chr>        <dbl> <dbl> <dbl>
## 1 Size = 1.2  69.6  94.2  120.
## 2 Size = 1.6 125.  141.   159.
## 3 Size = 2   172.  189.   205.
## 4 Size = 2.4 211.  236.   260.
```

Prediction of future response

- So far, we have seen
  - the variability among the fitted lines
  - the variability among the simulated house price for fixed size (reflects the variability in the posterior draws of β_0 and β_1)
- To predict future values for a house sale price Y given its size x, we also need to incorporate the sampling model in the simulation process

$$Y_i \mid \beta_0, \beta_1, \sigma \overset{ind}{\sim} \mathrm{Normal}(\beta_0 + \beta_1 x_i, \sigma) \tag{2}$$

Prediction of future response cont'd

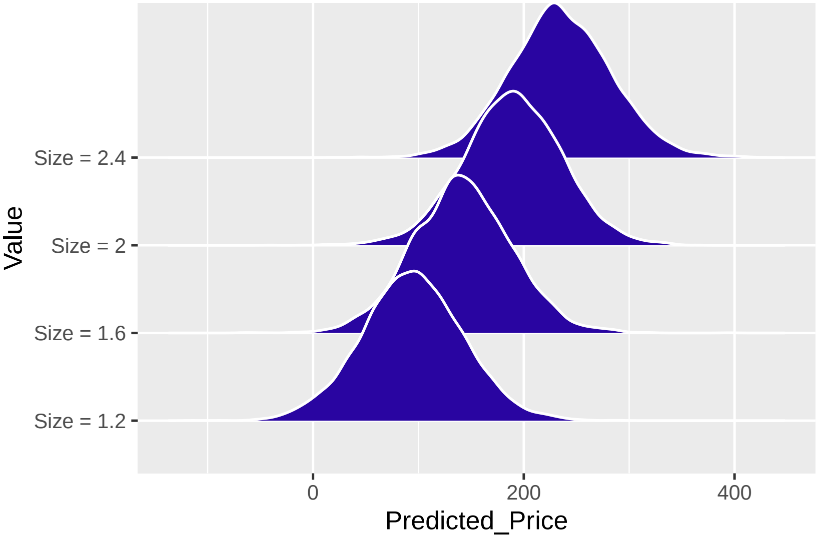
$$
\begin{array}{rcl}
\text{simulate } E[y]^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} x & \rightarrow & \text{sample } \tilde{y}^{(1)} \sim \mathrm{Normal}(E[y]^{(1)} \\
\text{simulate } E[y]^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} x & \rightarrow & \text{sample } \tilde{y}^{(2)} \sim \mathrm{Normal}(E[y]^{(2)} \\
 & \vdots & \\
\text{simulate } E[y]^{(S)} = \beta_0^{(S)} + \beta_1^{(S)} x & \rightarrow & \text{sample } \tilde{y}^{(S)} \sim \mathrm{Normal}(E[y]^{(S)}
\end{array}
$$

Prediction of future response cont'd

```r
one_predicted <- function(x){
  lp <- post[ , "beta0"] +  x * post[ , "beta1"]
  y <- rnorm(5000, lp, post[, "sigma"])
  data.frame(Value = paste("Price =", x),
             Predicted_Price = y)
}
```

Prediction of future response cont'd

```
## Picking joint bandwidth of 7.68
```



Prediction of future response cont'd

```
df %>% group_by(Value) %>%
  summarize(P05 = quantile(Predicted_Price, 0.05),
            P50 = median(Predicted_Price),
            P95 = quantile(Predicted_Price, 0.95))
```

```
## # A tibble: 4 x 4
##   Value          P05   P50   P95
##   <chr>        <dbl> <dbl> <dbl>
## 1 Size = 1.2  14.3  94.0  173.
## 2 Size = 1.6  64.8 141.   219.
## 3 Size = 2   112.  189.   266.
## 4 Size = 2.4 157.  234.   314.
```

- The prediction interval is substantially wider than the posterior interval - why?

Posterior predictive model checking

- Review:
  - helpful in judging the suitability of the linear regression model
  - the observed response values should be consistent with predicted responses generated from the fitted model

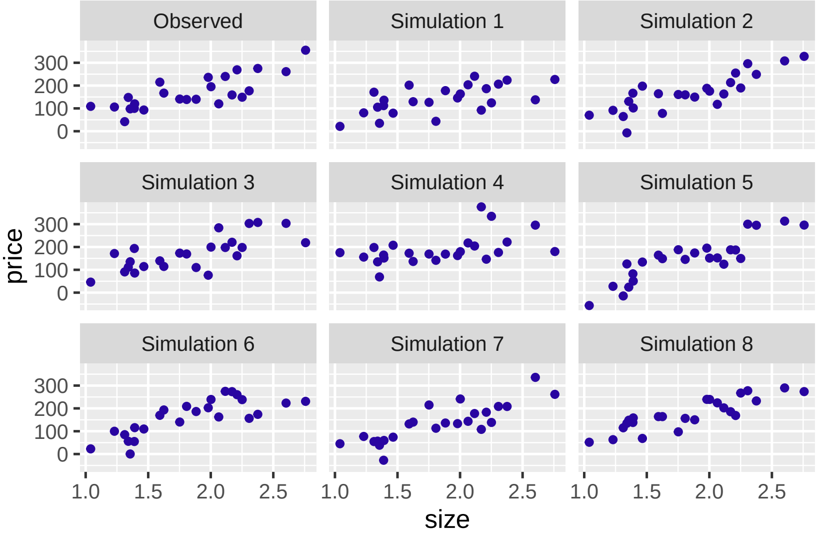
Posterior predictive model checking

- Review:
  - helpful in judging the suitability of the linear regression model
  - the observed response values should be consistent with predicted responses generated from the fitted model

- Two steps to get a replicated sample (same sample size):

1. Values of the parameters $(\beta_0, \beta_1, \sigma)$ are simulated from the posterior distribution – call these simulated values $(\beta_0^*, \beta_1^*, \sigma^*)$
2. A sample $\{y_1^R, ..., y_n^R\}$ is simulated where the sample size is $n = 24$ and y_i^R is Normal(μ_i^*, σ^*), where $\mu_i^* = \beta_0^* + \beta_1^* x_i$.

Posterior predictive model checking cont'd



- Your conclusion?

Predictive residuals

- Consider the observed point (x_i, y_i)
- Is the observed response value y_i consistent with predictions $\tilde{y}_i$ of this observation from the fitted model?

Predictive residuals

- Consider the observed point (x_i, y_i)
- Is the observed response value y_i consistent with predictions $\tilde{y}_i$ of this observation from the fitted model?

- We can simulate predictions $\tilde{y}_i$ from the posterior predictive distribution in two steps:

1. One simulates $(\beta_0, \beta_1, \sigma)$ from the posterior distribution
2. One simulates $\tilde{y}_i$ from a normal distribution with mean $\beta_0 + \beta_1 x_i$ and standard deviation σ

- By repeating this process many times, we have a sample of values $\{\tilde{y}_i\}$ from the posterior predictive distribution

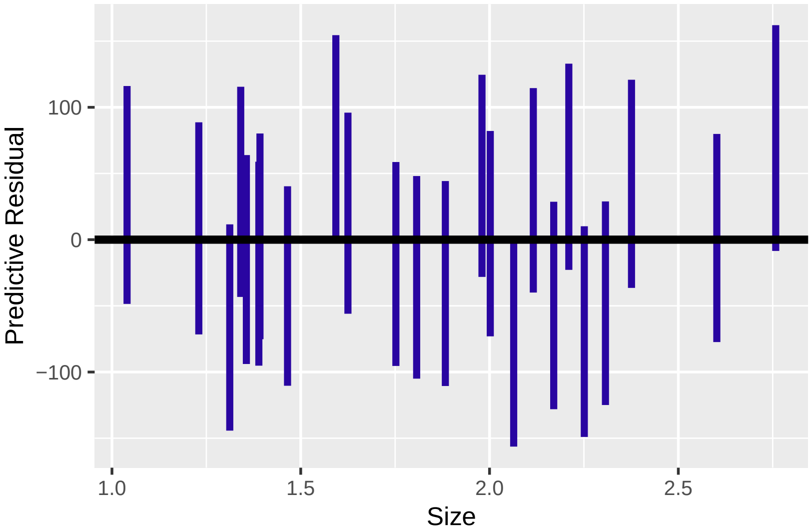
Predictive residuals cont'd

- Compute the predictive residual

$$r_i = y_i - \tilde{y}_i \tag{3}$$

- If this predictive residual is away from zero, that indicates that the observation is not consistent with the linear regression model

Predictive residuals cont'd



- Your conclusion?