HW1

Jinhong Du - 12243476

2019/9/26

Contents

1. The `gala` data set from the Faraway textbook counts the number of tortoise species on different Galapagos islands. Each data point is one island. We will be interested in how the number of endemic species (meaning, species that live only on that island) relates to the size of the island (its area). Here is code to load this data:

```
> library(faraway)
> data(gala)
> x = gala$Area
> y = gala$Endemic
```

If you've never used the `faraway` package in R, you will need to install it first before you can run the code above:

```
> install.packages('faraway')
```

(a) Make a scatterplot of the X and Y values. Do you feel that a linear model is appropriate for this data set? Are there any features of this data set that would make you question this model?

From the scatter plot Figure 1 (a), a linear model is not appropriate for this data set. It is because that most points with $Y < 50$ are near the line $X = 0$ while some points get away from $X = 0$ as Y increases ($Y \geq 50$). The plotted points seems to be near a log function of X.

```
library(faraway)
data(gala)
x = gala$Area
y = gala$Endemic

library(ggplot2)
df <- data.frame(
  X = x,
  Y = y
)
ggplot(df, aes(X, Y))+ geom_point() + ggtitle('Scatter plot of Y against X')
```

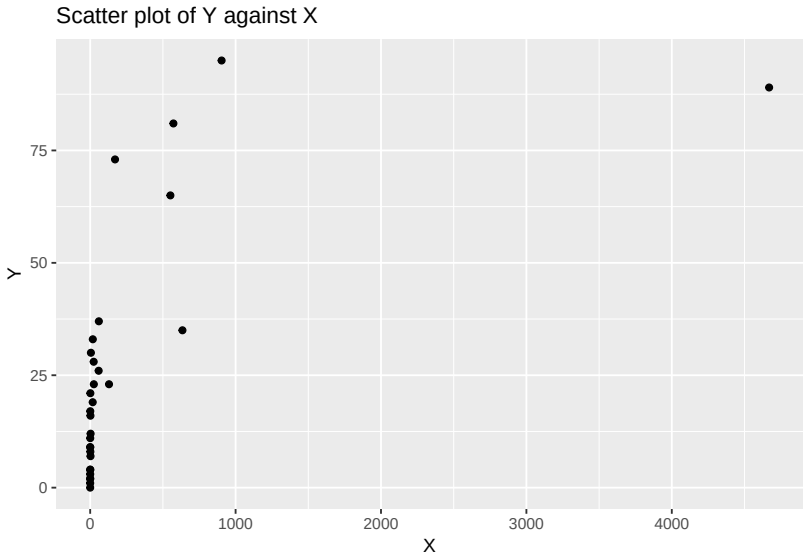


Figure 1 (a)

(b) Now try replacing X with the log-area:

```
> x = log(gala$Area)
```

Make a new scatterplot. Do you feel that a linear model is appropriate for this data set? Are there any features of this data set that would make you question this model?

From the scatter plot Figure 1 (b), when $\log X < 5$, the basic trend is that Y increases slowly as $\log X$ increases. While when $\log X \geq 5$, Y increases more as $\log X$ increases. It may be not appropriate, but it is better than that in the first plot.

```
df <- data.frame(
  X = log(x),
  Y = y
  )
  ggplot(df, aes(X, Y))+ geom_point() + xlab('log X') +
      ggtitle('Scatter plot of Y against log X')
```

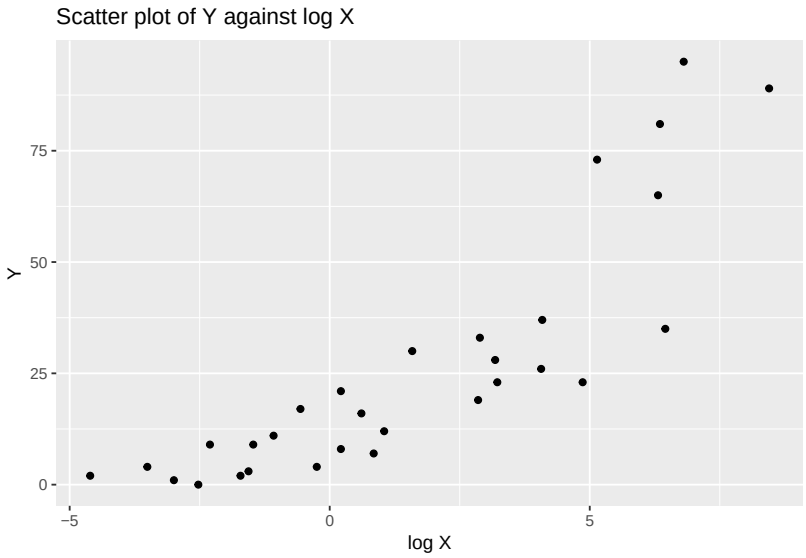


Figure 1 (b)

(c) From this point on we'll use log-area as our X, and we will proceed as though the assumptions for linear regression are satisfied. Compute the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and the variance estimator $\hat{\sigma}^2$ without using the `lm` command or any other regression commands, i.e. show the raw calculations for computing $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$. (It's of course fine to use vector multiplication and other elementary operations, you don't need to add up numbers individually.)

Let

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

be the respone vector, coefficients vector, error vector and the design matrix respectively, then the underlying model is given by

$$Y = X\beta + e$$

The OLS is given by

$$\hat{\beta} \triangleq \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^\top X)^{-1} X^\top Y.$$

The hat matrix is $H = X(X^\top X)^{-1} X^\top$ and

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}(Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= \frac{1}{n-2} Y^\top (I - H)(I - H) Y \\ &= \frac{1}{n-2} Y^\top (I - H) Y. \end{aligned}$$

```
n <- length(x)
X <- cbind(matrix(1, n, 1), matrix(log(x), n, 1))
Y <- matrix(y, n, 1)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y
H <- X %*% solve(t(X) %*% X) %*% t(X)
sigma2 <- t(Y) %*% (diag(n)-H) %*% Y /(n-2)
cat("beta_hat_0 = ", beta[1],", beta_hat_1 = ", beta[2],", sigma_hat^2 = ",sigma2)
```

```
## beta_hat_0 =  15.69099 , beta_hat_1 =  6.697806 , sigma_hat^2 =  204.2586
```

(d) According to your answer above, what's the predicted number of endemic species for an island whose area is 5.0**?**

```
yhat <- matrix(c(1, log(5.0)), 1, 2) %*% beta
cat('The predicted number of endemic species is ', yhat,
    ' for an island whose area is 5.0')
```

```
## The predicted number of endemic species is  26.47069  for an island whose area is 5.0
```

2. This problem continues with the least squares regression of Y = number of endemic species on X = log-area, from the previous problem.

(a) In R, compute the correlations between the vector of residuals and (i) the vector X; (ii) the vector Y ; and (iii) the vector of fitted values. Briefly explain each of the three answers you see—how do they relate to the definition & properties of least squares?

(i) The correlation between the vector of residuals and X is almost 0 ($2.000445e-16$). Since least squares assume that the random errors are independent, the residuals will be uncorrelated with the predictors. To see this,

$$\begin{aligned}
\bar{\hat{\epsilon}} &= \frac{1}{n}\hat{\epsilon}^\top \mathbf{1} \\
&= \frac{1}{n}(Y-\hat{Y})^\top \mathbf{1} \\
&= \frac{1}{n}Y^\top (I-H)\mathbf{1} \\
&= \frac{1}{n}Y^\top (\mathbf{1}-H\mathbf{1}) \\
&= 0 \\
Cov(\hat{\epsilon}, X) &= \frac{1}{n}(\hat{\epsilon}-\bar{\epsilon})^\top (X-\overline{X}) \\
&= \frac{1}{n}Y^\top (I-H)(X-\frac{1}{n}X^\top \mathbf{1})
\end{aligned}$$

Notice that $HX = X$, we have $Cov(\hat{\epsilon}, X) = 0$ and $Cor(\hat{\epsilon}, X) = 0$.

(ii) The correlation between the vector of residuals and Y is 0.513822. As we pointed out in problem 1 (b), Y increases more as $\log Y$ increases when $\log X \geq 5$, which means when Y gets too much large, the linear model becomes less reliale, and the residual $\hat{e}_i$ tends to be larger. So they may have positive correlation.

(iii) The correlation between the vector of residuals and the vector of fitted values is almost 0 ($1.867528e-16$). Since $\hat{Y}$ is a linear function of X, from the properties of covariance,

$$Cov(\hat{\epsilon}, \hat{Y}) = Cov(\hat{\epsilon}, X)H^\top = 0$$

Thus, $Cor(\hat{\epsilon}, \hat{Y}) = 0$.

```
res <- Y - X %*% beta
Y_hat <- X %*% beta
cat('(i) ',cor(res, log(x)),'; (ii) ',cor(res, Y),'; (iii) ',cor(res, Y_hat))
```

```
## (i)  2.000445e-16 ; (ii)  0.5138222 ; (iii)  1.867528e-16
```

(b) In R, compute the correlation between the vector of residuals and the variable gala$Nearest, which is the distance from each island to its nearest island (i.e. it's large if the island is far from any other island). Explain the answer you see—why does it make sense in the context of the data?

The correlation between the vector of residuals and the variable gala$Nearest is -0.409941. The larger the distance from each island to its nearest island, the number of endemic species in this island will be less effected by other islands. Since the assumptions for linear regression are satisfied, the fitted values in the island with large value of gala$Nearest will be more near to the true value. So the residual tends to be smaller.

```
cat('The correlation between the vector of residuals and ',
    'the variable gala$Nearest is ', cor(res, gala$Nearest))
```

```
## The correlation between the vector of residuals and  the variable gala$Nearest is  -0.4099413
```

3. Consider a data set consisting of X values $X_1, \cdots, X_n$ and Y values $Y_1, \cdots, Y_n$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ and $\hat{\sigma}$ be the output of OLS on this data set. Now define a transformation of the covariate:

$$\tilde{X}_i = c \cdot (X_i + d)$$

for each $i = 1, \cdots, n$, where $c \neq 0$ and d are arbitrary constants. Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ and $\tilde{\sigma}$ be the output of OLS run on data $\tilde{X}_1, \cdots, \tilde{X}_n$ and $Y_1, \cdots, Y_n$. Write equations for $\tilde{\beta}_0$ and $\tilde{\beta}_1$ and $\tilde{\sigma}$ in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$ and $\hat{\sigma}$ (and in terms of the constants c and d), and justify your answer.

Let

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & \vdots \\ 1 & X_n \end{pmatrix} \qquad \tilde{X} = \begin{pmatrix} 1 & \tilde{X}_1 \\ 1 & \tilde{X}_2 \\ 1 & \vdots \\ 1 & \tilde{X}_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \beta' = \begin{pmatrix} \beta_0' \\ \beta_1' \end{pmatrix} \qquad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad \tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}$$

then the underlying linear models are expressed as

$$\begin{aligned} Y &= X\beta + e \\ Y &= \tilde{X}\beta' + e' \end{aligned}$$

where β, β' and e, e' are the linear coefficients and random errors.

Notice that

$$\begin{aligned} \tilde{X}\tilde{\beta} &= \begin{pmatrix} \tilde{\beta}_0 + \tilde{\beta}_1 c(\tilde{X}_1 + d) \\ \tilde{\beta}_0 + \tilde{\beta}_1 c(\tilde{X}_2 + d) \\ \vdots \\ \tilde{\beta}_0 + \tilde{\beta}_1 c(\tilde{X}_n + d) \end{pmatrix} \\ &= \begin{pmatrix} (\tilde{\beta}_0 + \tilde{\beta}_1 cd) + \tilde{\beta}_1 c\tilde{X}_1 \\ (\tilde{\beta}_0 + \tilde{\beta}_1 cd) + \tilde{\beta}_1 c\tilde{X}_2 \\ \vdots \\ (\tilde{\beta}_0 + \tilde{\beta}_1 cd) + \tilde{\beta}_1 c\tilde{X}_n \end{pmatrix} \\ &= X \begin{pmatrix} \tilde{\beta}_0 + \tilde{\beta}_1 cd \\ \tilde{\beta}_1 c \end{pmatrix} \end{aligned} \tag{1}$$

Therefore, the two groups of estimators have the following relationship,

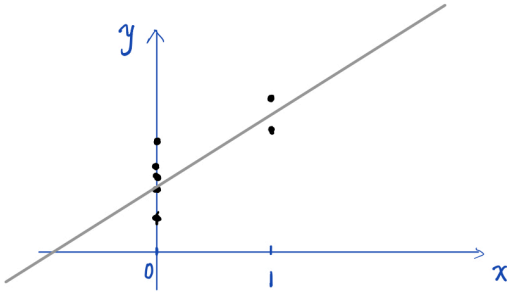
$$\hat{\beta} = \begin{pmatrix} \tilde{\beta}_0 + \tilde{\beta}_1 cd \\ \tilde{\beta}_1 c \end{pmatrix} \qquad \Longrightarrow \qquad \begin{cases} \tilde{\beta}_1 = \frac{1}{c}\hat{\beta}_1 \\ \tilde{\beta}_0 = \hat{\beta}_0 - d\hat{\beta}_1 \end{cases} \tag{2}$$

Recall the definition of $\hat{\sigma}^2$,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}(Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= \frac{1}{n-2} Y^\top (I - H)(I - H) Y \\ &= \frac{1}{n-2} Y^\top (I - H) Y \\ &= \frac{1}{n-2} Y^\top (Y - X\hat{\beta}) \end{aligned}$$

where $H = X(X^\top X)^{-1} X^\top$. Similarly, we have $\tilde{\sigma}^2 = \frac{1}{n-2} Y^\top (Y - \tilde{X}\tilde{\beta}) = \frac{1}{n-2} Y^\top (Y - X\hat{\beta}) = \hat{\sigma}^2$ from (1) and (2).

4. Suppose you have a data set where X takes only two values while Y can take arbitrary real values. To consider a concrete example, consider a clinical trial where $X_i = 0$ indicates that patient i received the placebo, while $X_i = 1$ indicates that patient i received the treatment, and Y_i is the real-valued outcome for patient i, e.g. blood pressure. Let $\overline{Y}_P$ and $\overline{Y}_T$ indicate the sample mean outcome values for the placebo group and for the treatment group, respectively. What will be the values of the OLS coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of these group means? Justify your answer.



Intuitively, the regression line might go through points $(0, \overline{Y}_P)$ and $(1, \overline{Y}_T)$. So the regression function will be $Y = (\overline{Y}_T - \overline{Y}_P)X + \overline{Y}_P$. Now we give a proof below.

Suppose the sample size is $n \in \mathbb{Z}_+$. Without loss of generality, we rearange the sample points so that the first n_P samples are in the placebo group and let $n_T = n - n_P$. Then we have

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{1}_P & \mathbf{0}_P \\ \mathbf{1}_T & \mathbf{1}_T \end{pmatrix} \qquad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_P \\ Y_T \end{pmatrix}$$

where $\mathbf{1}_l \in \mathbb{R}^l$ and $\mathbf{0}_l \in \mathbb{R}^l$ are the l-vectors with all elements equal to 1 and 0 respectively, and $Y_P \in \mathbb{R}^{n_P}$ and $Y_T \in \mathbb{R}^{n_T}$ are the subvectors of Y. Let $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 & \hat{\beta}_1 \end{pmatrix}^\top$, then

$$\begin{aligned}
\hat{\beta} &= (X^\top X)^{-1} X^\top Y \\
&= \left[\begin{pmatrix} \mathbf{1}_P^\top & \mathbf{1}_T^\top \\ \mathbf{0}_P^\top & \mathbf{1}_T^\top \end{pmatrix} \cdot \begin{pmatrix} \mathbf{1}_P & \mathbf{0}_P \\ \mathbf{1}_T & \mathbf{1}_T \end{pmatrix} \right]^{-1} X^\top Y \\
&= \begin{pmatrix} n & n_T \\ n_T & n_T \end{pmatrix}^{-1} X^\top Y \\
&= \begin{pmatrix} \frac{1}{n_P} & -\frac{1}{n_P} \\ -\frac{1}{n_P} & -\frac{1}{n_T} + \frac{1}{n_P} \end{pmatrix} X^\top Y \\
&= \begin{pmatrix} \frac{1}{n_P}\mathbf{1}_P^\top & \mathbf{0}_T^\top \\ -\frac{1}{n_P}\mathbf{1}_T^\top & \frac{1}{n_T}\mathbf{1}_T^\top \end{pmatrix} \begin{pmatrix} Y_P \\ Y_T \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n_P}\mathbf{1}^\top Y_P \\ \frac{1}{n_T}\mathbf{1}^\top Y_T - \frac{1}{n_P}\mathbf{1}^\top Y_P \end{pmatrix} \\
&= \begin{pmatrix} \overline{Y}_P \\ \overline{Y}_T - \frac{1}{n_P}\mathbf{1}^\top Y_P \end{pmatrix}
\end{aligned}$$

5. In this problem we'll examine the effect of the assumptions of the linear model, on the validity of inference.

(a) Generate a simulated data set of size $n = 100$ as follows:

- **Draw $X_i \overset{iid}{\sim} N(0, 1)$**
- **Set $\beta_0 = \beta_1 = \sigma^2 = 1$ and $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$**

Run an ordinary least squares regression, and compute a 90% prediction interval at two new X values, $X_{\text{new},1} = -0.5$ and $X_{\text{new},2} = 2$. (In this problem you are now free to use `lm` and related commands to calculate the estimated coefficients, $\hat{\sigma}^2$, etc.) Then generate new Y values, $Y_{\text{new},1}$ and $Y_{\text{new},2}$, which are generated from the same model as the Y_i's (but using the new X values), and check if the two prediction intervals cover their respective targets. Repeat this simulation 1000 times. What coverage rate do you observe empirically for each of the two prediction problems?

Since

$$\begin{aligned} Y_{new} &= \beta_0 + \beta_1 X_{new} + \epsilon_{new} \sim N(\beta_0 + \beta_1 X_{new}, \sigma^2) \\ \hat{Y}_{new} &= \hat{\beta}_0 + \hat{\beta}_1 X_{new} \\ &\sim N\left(\beta_0 + \beta_1 X_{new}, \sigma^2 \left(\frac{1}{n} + \frac{(X_{new} - \overline{X})^2}{SS_{XX}}\right)\right) \end{aligned}$$

we have

$$\frac{Y_{new} - \hat{Y}_{new}}{s_{pred}} \sim t(n-2)$$

where

$$\begin{aligned} s^2_{pred} &= MSE\left(1 + \frac{1}{n} + \frac{(X_{new} - \overline{X})^2}{SS_{XX}}\right) \\ SS_{XX} &= \sum_{i=1}^{n} (X_i - \overline{X})^2 \end{aligned}$$

Therefore, the 90% prediction interval of Y_{new} is given by $(\hat{Y}_{new} - s_{pred} \cdot t(0.95, n-2), \hat{Y}_{new} + s_{pred} \cdot t(0.95, n-2))$.

Both of the coverage rates of the two prediction problems are near 90%.

```
simulation <- function(X_new_list) {
    # Generate dataset
    n <- 100
    X <- rnorm(n, mean = 0, sd = 1)
    beta_0 <- 1
    beta_1 <- 1
    sigma <- 1
    epsilon <- rnorm(n, mean = 0, sd = sigma)
    Y <- beta_0 + beta_1 * X + epsilon

    # Fit lm model
    df <- data.frame(
      X = X,
      Y = Y
    )
    model <- lm(Y~X, df)
    out <- summary(model)
    X_bar <- mean(X)
    SS_XX <- sum(X^2) - n*X_bar^2
```

```
    # Prediction interval of Y_new
    result <- c(0,0)
    for(j in 1:2){
        X_new <- X_new_list[j]
        Y_new <- beta_0 + beta_1 * X_new + rnorm(1, mean = 0, sd = sigma)
        Y_new_hat <- model$coefficients[1] + model$coefficients[2] * X_new
        s_pred <- out$sigma * sqrt(1+1/n+(X_new-X_bar)^2/SS_XX)
        lwr <- Y_new_hat - s_pred * qt(0.95, n-2)
        upr <- Y_new_hat + s_pred * qt(0.95, n-2)

        # Prediction interval by predict.lm
        # predict(model, newdata=data.frame(X=c(X_new)),
        #         interval='prediction', level=0.9)

        if(Y_new >= lwr && Y_new <= upr){result[j]=1}
    }
    return(result)
}

set.seed(0)
X_new_1 <- -0.5
X_new_2 <- 2
m <- 1000
result <- matrix(0, m, 2)
for(i in 1:m){
    result[i, ] <- simulation(c(X_new_1, X_new_2))
}
cat("The coverage rate of Y_new_1 is ", mean(result[,1]),
    "\nThe coverage rate of Y_new_2 is ", mean(result[,2]))
```

```
## The coverage rate of Y_new_1 is  0.909
## The coverage rate of Y_new_2 is  0.909
```

(b) Now repeat the same experiment, but now the data is generated in a way that violates the model assumptions: draw $Y_i = \beta_0 + \beta_1 X_i + e^{X_i} + \epsilon_i$. Then construct 90% prediction intervals at the same two new X values, and as before generate the new Y values from the same model as the Y_i's, and repeat 1000 times. What coverage rate do you observe empirically for each of the two prediction problems?

The coverage rate of the first prediction problem is near 100%, while the coverage of the second one is about 50%.

```
simulation <- function(X_new_list) {
    # Generate dataset
    n <- 100
    X <- rnorm(n, mean = 0, sd = 1)
    beta_0 <- 1
    beta_1 <- 1
    sigma <- 1
    epsilon <- rnorm(n, mean = 0, sd = sigma)
    Y <- beta_0 + beta_1 * X + exp(X) + epsilon

    # Fit lm model
```

```
    df <- data.frame(
      X = X,
      Y = Y
    )
    model <- lm(Y~X, df)
    out <- summary(model)
    X_bar <- mean(X)
    SS_XX <- sum(X^2) - n*X_bar^2

    # Prediction interval of Y_new
    result <- c(0,0)
    for(j in 1:2){
        X_new <- X_new_list[j]
        Y_new <- beta_0 + beta_1 * X_new + exp(X_new) + rnorm(1, mean = 0, sd = sigma)
        Y_new_hat <- model$coefficients[1] + model$coefficients[2] * X_new
        s_pred <- out$sigma * sqrt(1+1/n+(X_new-X_bar)^2/SS_XX)
        lwr <- Y_new_hat - s_pred * qt(0.95, n-2)
        upr <- Y_new_hat + s_pred * qt(0.95, n-2)

        # Prediction interval by predict.lm
        # predict(model, newdata=data.frame(X=c(X_new)),
        #         interval='prediction', level=0.9)

        if(Y_new >= lwr && Y_new <= upr){result[j]=1}
    }
    return(result)
}

set.seed(0)
X_new_1 <- -0.5
X_new_2 <- 2
m <- 1000
result <- matrix(0, m, 2)
for(i in 1:m){
    result[i, ] <- simulation(c(X_new_1, X_new_2))
}
cat("The coverage rate of Y_new_1 is ", mean(result[,1]),
    "\nThe coverage rate of Y_new_2 is ", mean(result[,2]))
```

```
## The coverage rate of Y_new_1 is  0.983
## The coverage rate of Y_new_2 is  0.497
```

(c) Explain the results you observe in parts (a) and (b). In particular, you should explain why you observe opposite trends for the two points in part (b).

When data satisfy the linear regression model assumptions, the meaning of the 90% prediction interval for Y_{new} is that the true value of Y_{new} lies in this interval with probability 0.9, which accounts for the result in part (a). However, in part (b), the assumptions are violated. As X inceases, the term e^X increases and violates the linear assumption even more. Y_{new_2} becomes even more unpredictable with respect to the large predictor $X_{new,2}$, so the prediction interval becomes more "permissive" and the coverage rate of the prediction interval decreases. Relatively, the prediction interval of $Y_{new,1}$ becomes more "conservative" and include more points.