PSTAT 120C HW 1

TJ Sipin

2022-08-10

Reading

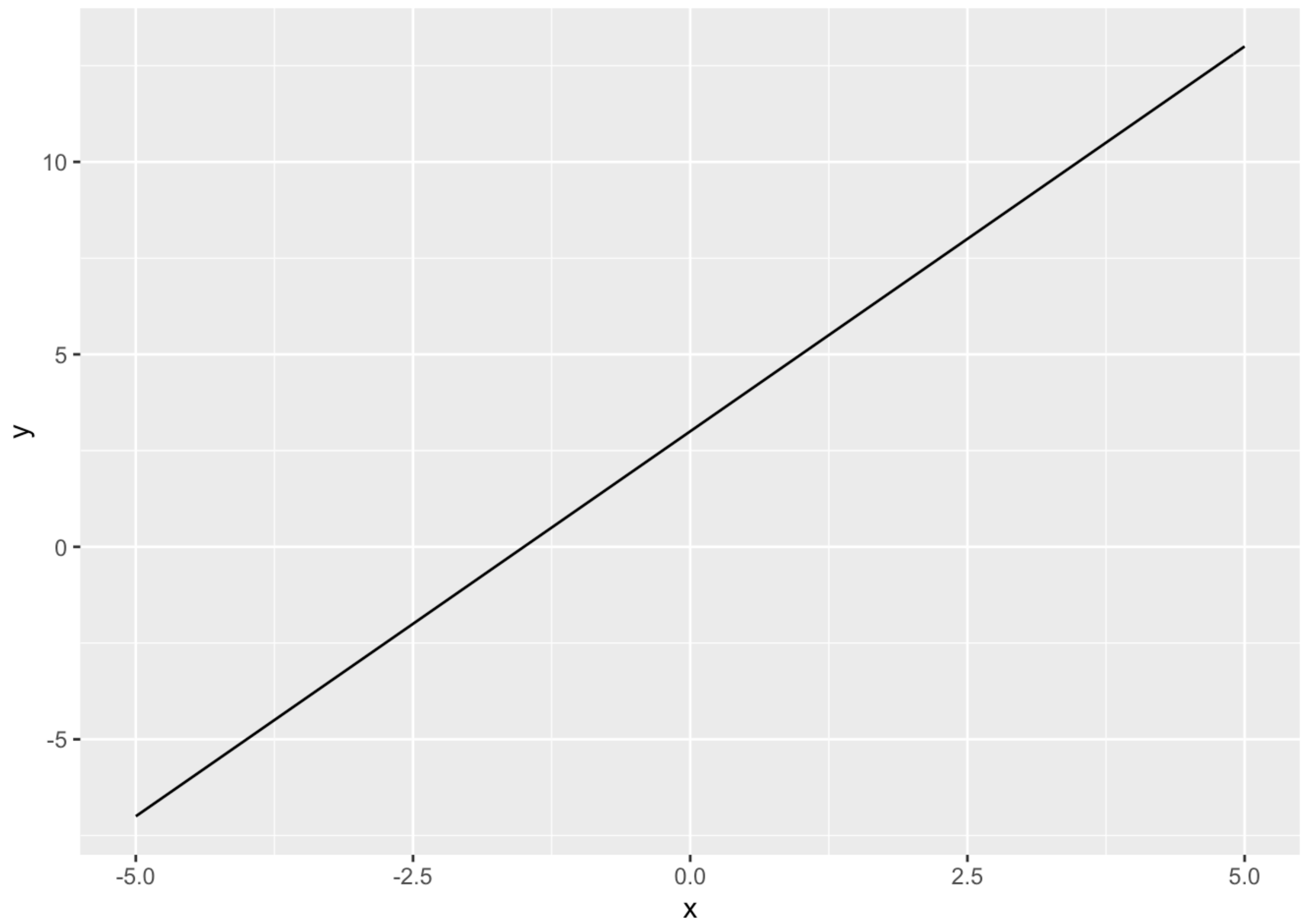
Define deterministic and probabilistic mathematical models. Give an example of each.

A deterministic model does not have an error component when predicting a response variable y as a function of a set of explanatory variables. On the other hand, a probabilistic model is one that has an error component ϵ, such that it produces a random variable Y. It must be noted that a probabilistic model can be interpreted as a sum of a deterministic component $\mathbb{E}(Y)$ and a random component ϵ.

An example of a deterministic model:

$$y = 2x + 3$$

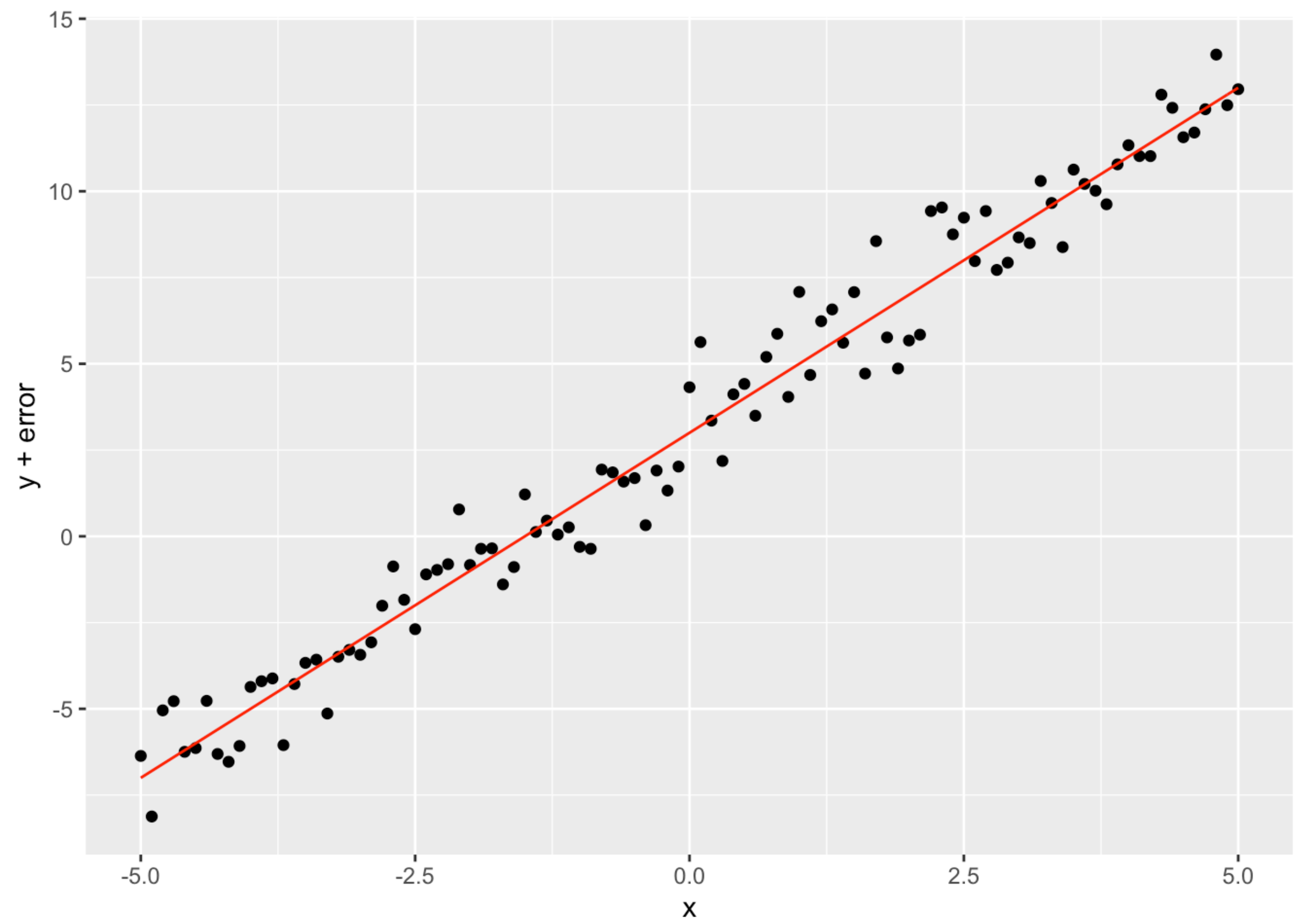
```
x = seq(-5, 5, by = 0.1)
y = 2*x + 3

ggplot() +
  geom_line(aes(x = x,
                y = y))
```



An example of a probabilistic model:

$$Y = 3 + 2x + \epsilon, \qquad \epsilon \sim N(0, 1)$$

```
error = rnorm(n = length(x), mean = 0, sd = 1)

ggplot() +
  geom_point(aes(x = x,
                 y = y + error)) +
  geom_line(aes(x = x,
                y = y),
            col = 'red')
```



Write the general equation for a simple linear regression model.

The general equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Describe, in your own words, the overall concept of the method of least squares.

The idea is to find a set of parameter estimates $\beta_i, \quad i = 0, \ldots, n$, where $n = 1$ for the simple linear regression model case, such that the error $y_i - \hat{y}_i$ is minimized. In other words, we want to minimize $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

State the least-squares estimators for the simple linear regression model.

Derived from the lecture demonstration, the OLS estimators for the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

State the means and variances of the least-squares estimators in simple linear regression.

Means

Since the estimators $\hat{\beta}_0, \hat{\beta}_1$ are unbiased, then

$$\mathbb{E}(\hat{\beta}_0) = \beta_0$$
$$\mathbb{E}(\hat{\beta}_1) = \beta_1$$

Variances

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2}{nS_{xx}} \sum x_i^2$$
$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

State a pair of null and alternative hypotheses for making inferences about single regression parameters and linear functions of the parameters.

Testing for $\theta = a_o \beta_0 + a_1 \beta_1$

$$H_0 : \theta = \theta_0$$
$$H_a : \begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

Practice

1. Auditors are often required to compare the audited (or current) value of an inventory item with the book (or listed) value. If a company is keeping its inventory and books up to date, there should be a strong linear relationship between the audited and book values. A company sampled ten inventory items and obtained the audited and book values given in the accompanying code.

a. Fit the model $Y = \beta_0 + \beta_1 x + \epsilon$ to the data, using least squares.

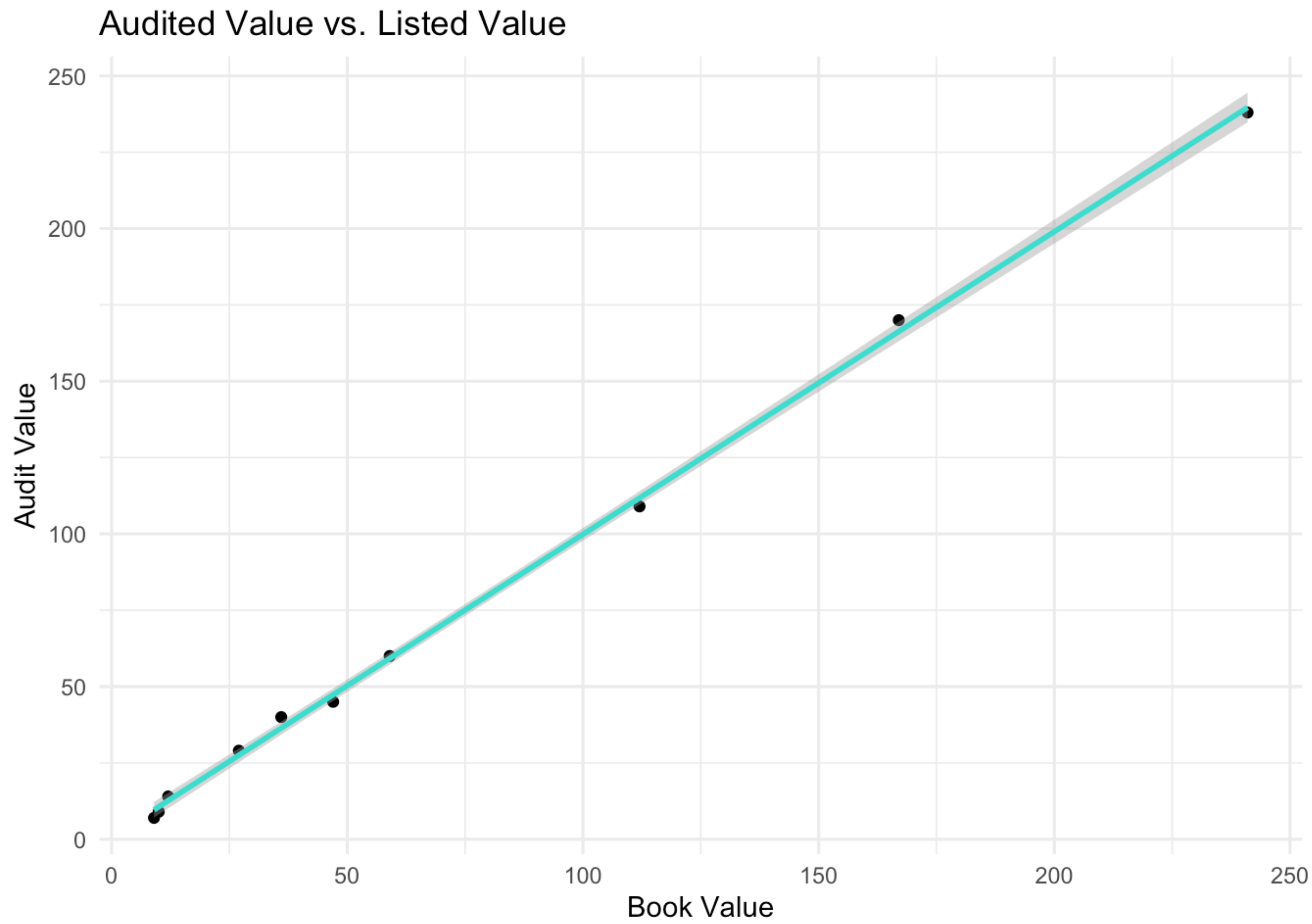
```
audit <- c(9,14,7,29,45,
           109,40,238,60,170)
book <- c(10,12,9,27,47,
          112,36,241,59,167)
audit_df <- data.frame(audit = audit,
                       book = book)

lm_audit <- lm(audit ~ book, data = audit_df)

lm_audit %>% summary
```

```
##
## Call:
## lm(formula = audit ~ book, data = audit_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7557 -2.1477 -0.4228  1.4803  3.7178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7198     1.1764   0.612    0.558
## book          0.9914     0.0114  86.994  3.4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.666 on 8 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9988
## F-statistic:  7568 on 1 and 8 DF,  p-value: 3.401e-13
```

b. Plot the 10 data points and graph the line representing the model.

```
ggplot(data = audit_df,
       aes(x = book,
             y = audit)) +
  geom_point() +
  geom_smooth(method = 'lm',
              color = 'turquoise') +
  theme_minimal() +
  labs(x = 'Book Value',
       y = 'Audit Value') +
  ggtitle('Audited Value vs. Listed Value')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



c. Calculate SSE and S^2

```
sse = sum(lm_audit$residuals^2)
sse
```

```
## [1] 56.84544
```

```
var(lm_audit$residuals) # sse/(n-1), n = 10
```

```
## [1] 6.31616
```

$$SSE \approx 56.8$$
$$S^2 \approx 6.3$$

d. Do the data present sufficient evidence to indicate that the slope β_1 differs from zero? Conduct a hypothesis test at the 5% significance level.

The p-value for β_1 is 3.401×10^{-13}, which is much smaller than 0.05, thus the data presents sufficient evidence to indicate that the slope differs from zero at a 5% significance level.

e. What is the model's estimate from the expected change in audited value per one-unit change in book value?

There is about a one-unit (0.9914) change in audited value per one-unit change in book value.

f. What does the model predict the audited value to be for an item with a book value of $100?

$$\begin{aligned} y &= 0.7198 + 0.9914(100) \\ &= 99.8598 \end{aligned}$$

2. Let β_0 and β_1 be the least-squares estimates for the intercept and slope in a simple linear regression model. Show that the least-squares equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ will always go through the point $(\bar{x}, \bar{y})$.

Since

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

then we have

$$\begin{aligned} \hat{y} &= \bar{y} - \hat{\beta}_1 \bar{x} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} x \\ &= \bar{y} - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \bar{x} + \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} x \end{aligned}$$

To show that this line always goes through the point $(\bar{x}, \bar{y})$, then we substitute $\bar{x}$ for x and $\bar{y}$ for $\hat{y}$.

$$
\begin{aligned}
\bar{y} &= \bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} \\
&= \bar{y}
\end{aligned}
$$

We see that the left- and right-hand sides of the equation are indeed equal with this substitution, so the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ will always go through the point $(\hat{x}, \hat{y})$.

Another approach would be

$$
\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i + \epsilon_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \qquad (1) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0) + \sum_{i=1}^{n}(\beta_1 x_i) \\
&= \frac{1}{n}n\beta_0 + \frac{1}{n}n\beta_1\sum_{i=1}^{n} x_i \\
&= \beta_0 + \beta_1\bar{x} \\
\implies \bar{y} &= \beta_0 + \beta_1\bar{x}
\end{aligned}
$$

The ϵ_i disappears in (1) because of the assumption that $\mathbb{E}(\epsilon) = 0$. Since $\bar{y} = \beta_0 + \beta_1\hat{x}$ is of the same form as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \beta_0 + \beta_1 x$, then we can conclude that the equation always goes through $(\bar{x}, \bar{y})$.

3. Suppose that the model $y = \beta_0 + \beta_1 x + \epsilon$ is fit to the n data points $(y_1, x_1), \ldots, (y_n, x_n)$. At what value of x will the length of the prediction interval for y be minimized?

The length of a prediction interval for an actual value of Y when $x = x^*$ is given by the equation

$$
2 \times t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.
$$

So what we want is the choice of x^* that minimizes this equation. In other words,

$$
\min \arg_{x^*} 2 \times t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.
$$

The choice of x^* that minimizes the length of the prediction interval is $\bar{x}$.

```r
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(tidymodels)
x = seq(-5, 5, by = 0.1)
y = 2*x + 3

ggplot() +
  geom_line(aes(x = x,
                y = y))
error = rnorm(n = length(x), mean = 0, sd = 1)

ggplot() +
  geom_point(aes(x = x,
                 y = y + error)) +
  geom_line(aes(x = x,
                y = y),
            col = 'red')

audit <- c(9,14,7,29,45,
           109,40,238,60,170)
book <- c(10,12,9,27,47,
          112,36,241,59,167)
audit_df <- data.frame(audit = audit,
                       book = book)

lm_audit <- lm(audit ~ book, data = audit_df)

lm_audit %>% summary
ggplot(data = audit_df,
       aes(x = book,
             y = audit)) +
  geom_point() +
  geom_smooth(method = 'lm',
              color = 'turquoise') +
  theme_minimal() +
  labs(x = 'Book Value',
       y = 'Audit Value') +
  ggtitle('Audited Value vs. Listed Value')
sse = sum(lm_audit$residuals^2)
sse

var(lm_audit$residuals) # sse/(n-1), n = 10
```