Regression analysis_Homework Assignment 1

心理所碩二 R08227112 林子堯

2020/09/21

Douglas C. Montgomery, (2012). Introduction to Linear Regression Analysis, 5th ed.

1. (#2.23) Consider the simple linear regression model $y = 50 + 10x + \varepsilon$ where ε is NID(Normally and Independently Distributed)(0, 16). Suppose that $n = 20$ pairs of observations are used to fit this model.

Generate 500 samples of 20 observations, drawing one observation for each level of $x = 0.5, 1, 1.5, \ldots, 10$ for each sample.

```r
library(tidyverse)
```

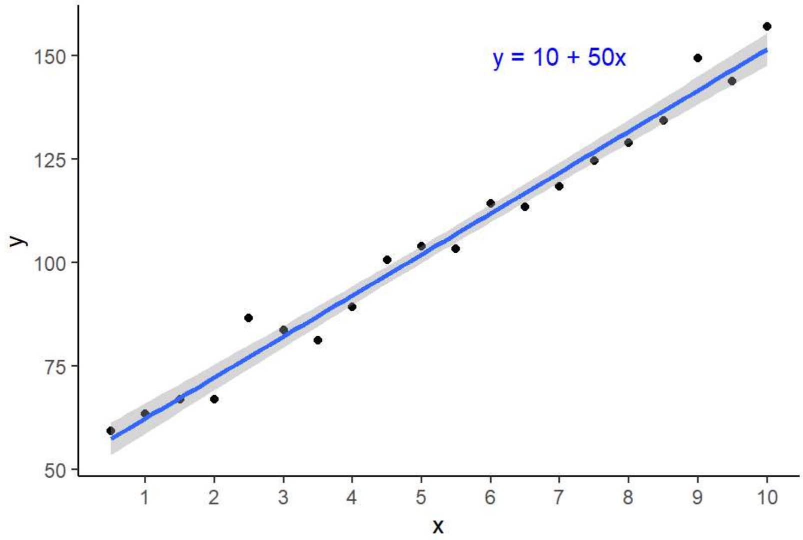
```r
# generate 500 samples with 19 observations in each sample
set.seed(9999)

nReplicate <- 500

getObservation <- function(){
  beta <- c(50, 10)
  sigma <- 4
  x <- seq(0.5, 10, by = 0.5)
  y <- rnorm(n = length(x),
             mean = beta[1]+beta[2]*x,
             sd = sigma)
  sample <- tibble(x = x, y = y)
}

samples <- replicate(nReplicate, getObservation(), simplify = FALSE)

# draw one observation
samples[[1]] %>%
  ggplot(aes(x, y)) +
    geom_point() +
    geom_smooth(method = "lm") +
    annotate("text", x = 7, y = 150,
             label = "y = 10 + 50x", color = "blue") +
    scale_x_continuous(breaks = 1:10) +
    theme_classic()
```



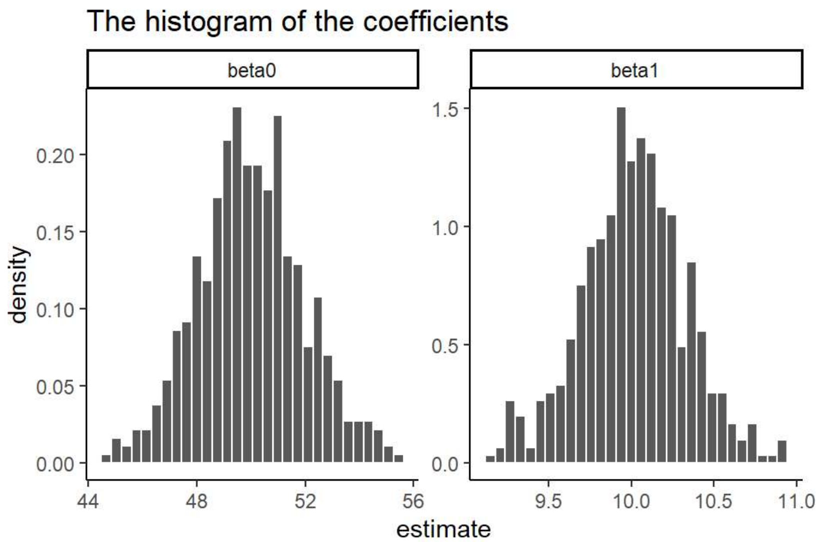
a. For each sample compute the least - squares estimates of the slope and intercept. Construct histograms of the sample values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Discuss the shape of these histograms.

```
fitlms ← lapply(samples, function(.sample){lm(y ~ 1 + x, data = .sample)})

coefs ← lapply(fitlms, broom::tidy) %>% bind_rows(.id = "replicate")

coefs_labeller ← labeller(term = c(`(Intercept)` = "beta0", x = "beta1"))

coefs %>%
  ggplot(aes(x = estimate)) +
    geom_histogram(aes(y = ..density..), color = "white") +
    facet_wrap(~ term, scales = "free", labeller = coefs_labeller) +
    labs(title = "The histogram of the coefficients") +
    theme_classic()
```

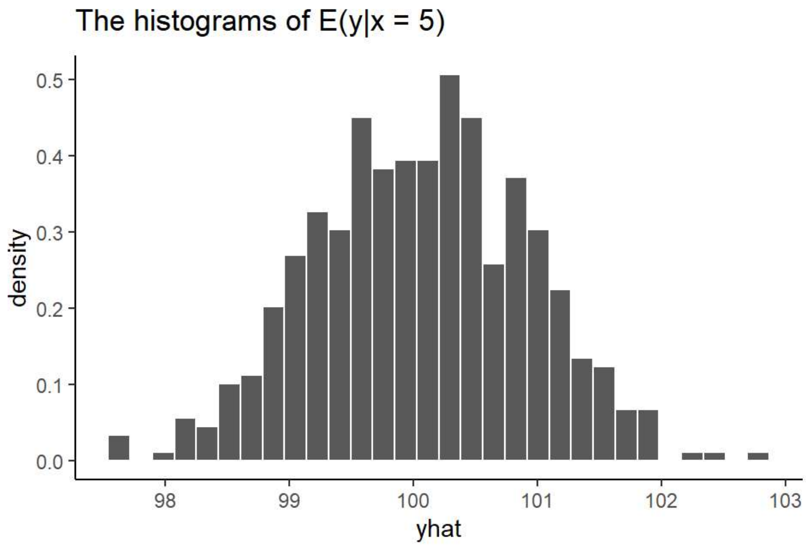


The both histograms are bell shape. The distribution of $\hat{\beta}_0$ is centered around 50 and the one of $\hat{\beta}_1$ is centered around 10.

b. For each sample, compute an estimate of $E(y|x = 5)$. Construct a histogram of the estimates you obtained. Discuss the shape of the histogram.

```r
predictsAt5 <- lapply(fitlms, function(.fitlm){
  data.frame(yhat = predict(.fitlm, data.frame(x = 5)))
}) %>%
  bind_rows(.id = "replicate")

ggplot(predictsAt5, aes(x = yhat)) +
  geom_histogram(aes(y = ..density..), color = "white") +
  labs(title = expression(paste("The histograms of E(y|x = 5)"))) +
  theme_classic()
```



The histogram of the conditional expectation estimator $\hat{E}(y|x=5) \equiv \hat{\mu}_{y|x=5}$ is also like bell shape, and it is centered around 100.

c. For each sample, compute a 95% CI on the slope. How many of these intervals contain the true value $\beta_1 = 10$? Is this what you would expect?

```r
coef_cis <- lapply(fitlms, function(.fitlm){
  as.data.frame(confint(.fitlm, "x", level = 0.95))
}) %>%
  bind_rows(.id = "replicate")

coef_cis %>%
  mutate(contain10 = (`2.5 %` <= 10) & (10 <= `97.5 %`)) %>%
  summarise(confidence = mean(contain10))
```

```
  confidence
1       0.94
```

It finds that 94% of our 500 CI's, which were constructed from samples respectively, for β_1 cover 10. This result is close to our expectation at 95%.

d. For each estimate of $E(y|x=5)$ in part b, compute the 95% CI. How many of these intervals contain the true value of $E(y|x=5) = 100$? Is this what you would expect?

```
predictsAt5_cis ← lapply(fitlms, function(.fitlm){
  data.frame(predict(.fitlm, data.frame(x = 5), interval = "confidence"))
}) %>%
  bind_rows(.id = "replicate")

predictsAt5_cis %>%
  mutate(contain100 = (lwr ⩽ 100) & (100 ⩽ upr)) %>%
  summarise(confidence = mean(contain100))
```

```
  confidence
1      0.974
```

It finds that 97.4% in our 500 CI's for $\hat{E}(y|x=5)$ cover 100. This result is close to our expectation at 95%.

2. (#2.25) Consider the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, with $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, and ε uncorrelated.

a. Show that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -x\sigma^2/S_{xx}$

$$\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) \\
&= Cov(\bar{y}, \hat{\beta}_1) - Cov(\hat{\beta}_1\bar{x}, \hat{\beta}_1) \\
&= 0 - \bar{x}Var(\hat{\beta}_1) && \text{(by part b)} \\
&= -\bar{x}\sigma^2/S_{xx}
\end{aligned}$$

b. Show that $Cov(\bar{y}, \hat{\beta}_1) = 0$

$$\begin{aligned}
Cov(\bar{y}, \hat{\beta}_1) &= Cov(\sum_{i=1}^{n} \frac{y_i}{n}, \sum_{j=1}^{n} \frac{(x_j - \bar{x})y_j}{S_{xx}}) \\
&= \frac{1}{nS_{xx}} \sum_{i,j} (x_j - \bar{x}) Cov(y_i, y_j)
\end{aligned}$$

Since $y_i' s$ are mutually independent, $Cov(y_i, y_j) = 0$ if $i \neq j$ and $Cov(y_i, y_i) = Var(y_i) = Var(\varepsilon_i) = \sigma^2$. Therefore, we have

$$\begin{aligned}
Cov(\bar{y}, \hat{\beta}_1) &= \frac{\sigma^2}{nS_{xx}} \sum_{j=1}^{n} (x_j - \bar{x}) \\
&= 0
\end{aligned}$$

3. (#2.27) Suppose that we have fit the straight-line regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ but the response is affected by a second variable x_2 such that the true regression function is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

a. Is the least-squares estimator of the slope in the original simple linear regression model unbiased?

$$\begin{aligned}
E(\hat{\beta}_1) &= E(\sum_{i=1}^{n} \frac{(x_{1i} - \bar{x}_1)y_i}{S_{xx}}) \\
&= \sum_{i=1}^{n} \frac{(x_{1i} - \bar{x}_1)}{S_{xx}} E(y_i) \\
&= \sum_{i=1}^{n} \frac{(x_{1i} - \bar{x}_1)}{S_{xx}} (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)x_{2i}}{S_{xx}}
\end{aligned}$$

where $\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)=0$ and $\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)x_{1i}=\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{1i}-\bar{x}_1)=S_{xx}$. In this situation, the original estimator $\hat{\beta}_1$ is a biased estimator for β_1.

b. Show the bias in $\hat{\beta}_1$

$$\begin{aligned} Bias_{\beta_1}(\hat{\beta}_1) &= E(\hat{\beta}_1)-\beta_1 \\ &= (\beta_1+\beta_2\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)x_{2i}}{S_{xx}})-\beta_1 \\ &= \beta_2\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)x_{2i}}{S_{xx}} \end{aligned}$$

4. (#2.32) Consider the simple linear regression model $y=\beta_0+\beta_1 x+\varepsilon$ where the intercept β_0 is known.

a. Find the least-squares estimator of β_1 for this model. Does this answer seem reasonable?

The least-squares criterion is

$$LS(\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

The least-square estimators of β_1, say $\hat{\beta}_1$, must satisfy

$$\begin{aligned} 0 &= \frac{d}{d\beta_1}LS(\beta_1)\Big|_{\hat{\beta}_1} \\ &= -2\sum_{i=1}^{n}(y_i-\beta_0-\hat{\beta}_1 x_i)x_i \end{aligned}$$

one has

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(y_i-\beta_0)x_i}{\sum_{i=1}^{n}x_i^2}$$

It seems reasonable because $\hat{\beta}_1$ is depended on x_i and pure $y_i^*=y_i-\beta_0$ effect, which minus the intercept effect. And this regression line must go through the point $(0,\beta_0)$.

b. What is the variance of the slope $\left(\hat{\beta}_1\right)$ for the least-squares estimator found in part a?

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^{n}(y_i-\beta_0)x_i}{\sum_{i=1}^{n}x_i^2}\right) \\ &= \frac{1}{(\sum_{i=1}^{n}x_i^2)^2}\sum_{i=1}^{n}x_i^2 Var(y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}x_i^2} \end{aligned}$$

c. Find a $100(1-\alpha)$ percent CI for β_1. Is this interval narrower than the estimator for the case where both slope and intercept are unknown?

We can get that $E(SS_{RES})=E(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2)=(n-1)\sigma^2$, so let $MS_{RES}=\frac{SS_{RES}}{n-1}$ be an unbiased estimator of σ^2. If we assume $\varepsilon_i's$ are independently and normally distributed with mean 0 and variance σ^2, and it can be shown that $\frac{(n-1)MS_{RES}}{\sigma^2}\sim\chi^2_{n-1}$.

Furthermore, $\hat{\beta}_1$ is a linear combination of $\{y_i\}$, so $\hat{\beta}_1\sim N(E(\hat{\beta}_1)=\beta_1, Var(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2})$.

Therefore, the test statistic

$$
\begin{aligned}
T &= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{RES} / \sum_{i=1}^n x_i^2}} \\
&= \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / \sum_{i=1}^n x_i^2}}}{\sqrt{\frac{(n-1)MS_{RES}/\sigma^2}{n-1}}} \\
&\sim \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}
\end{aligned}
$$

follows a t_{n-1} distribution. One has

$$
\begin{aligned}
1-\alpha &= \Pr(t_{1-\alpha/2,n-1} < T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{RES} / \sum_{i=1}^n x_i^2}} < t_{\alpha/2,n-1}) \\
&= \Pr(\hat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{MS_{RES}}{\sum_{i=1}^n x_i^2}} < \beta_1 < \hat{\beta}_1 - t_{1-\alpha/2,n-1}\sqrt{\frac{MS_{RES}}{\sum_{i=1}^n x_i^2}}) \\
&= \Pr(\hat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{MS_{RES}}{\sum_{i=1}^n x_i^2}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-1}\sqrt{\frac{MS_{RES}}{\sum_{i=1}^n x_i^2}})
\end{aligned}
$$

So the $100(1-\alpha)\%$ CI for β_1 is $\hat{\beta}_1 \pm t_{\alpha/2,n-1}\sqrt{\frac{MS_{RES}}{\sum_{i=1}^n x_i^2}}$.

On the other hand, the $100(1-\alpha)\%$ CI of β_1 for the case where both unknown slope and intercept is $\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MS^*_{RES}}{\sum_{i=1}^n (x_i-\bar{x})^2}}$, where $MS^*_{RES} = \frac{SS_{RES}}{(n-2)} > \frac{SS_{RES}}{(n-1)} = MS_{RES}$. One can check that $t_{\alpha/2,n-2}\sqrt{\frac{1}{n-2}} > t_{\alpha/2,n-1}\sqrt{\frac{1}{n-1}} \quad \forall \alpha, n \geq 2$.

Finally, the $100(1-\alpha)\%$ CI for β_1 when β_0 is known is narrower than one when β_0 & β_1 are unknown.

5. (#2.33) Consider the least-squares residuals $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$, from the simple linear regression model. Find the variance of the residuals $Var(e_i)$. Is the variance of the residuals a constant? Discuss.

$$
\begin{aligned}
Var(e_i) &= Var(y_i - \hat{y}_i) \\
&= Var(y_i) + Var(\hat{y}_i) - 2Cov(y_i, \hat{y}_i) \\
&= \sigma^2 + Var(\hat{\beta}_0 + \hat{\beta}_1 x_i) - 2Cov(y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sigma^2 + Var(\bar{y} + \hat{\beta}_1(x_i - \bar{x})) - 2Cov(y_i, \bar{y} + \hat{\beta}_1(x_i - \bar{x})))
\end{aligned}
$$

Since from Exercise 2.25 part (b), we know $Cov(\bar{y}, \hat{\beta}_1) = 0$. Therefore, the variance of the residual e_i is

$$
\begin{aligned}
Var(e_i) &= \sigma^2 + Var(\bar{y}) + (x_i - \bar{x})^2 Var(\hat{\beta}_1) - 2[Cov(y_i, \bar{y}) + (x_i - \bar{x})Cov(y_i, \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{S_{xx}})] \\
&= \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}) - 2[\frac{1}{n}Var(y_i) + \frac{(x_i - \bar{x})^2}{S_{xx}}Var(y_i)] \qquad (\text{since } y_i's \text{ are independent}) \\
&= \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}) - 2\sigma^2(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}) \\
&= \sigma^2(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}})
\end{aligned}
$$

We can find that the variance of the residual e_i which depends on the x_i is not a constant. The $Var(e_i)$ is decreasing as the distance $|x_i - \bar{x}|$ increases.