hw1

Thibaud Bruyelle

9/15/2020

Question 2

(a)

(b)

The estimates are given by the following formula:

$$\boxed{\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} \widetilde{x}_i (y_i - \bar{y})}{\sum_{i=1}^{n} \widetilde{x}_i^2}}$$

and

$$\boxed{\hat{\alpha}_0 = \bar{y}}$$

because $\frac{1}{n}\sum_{i=1}^{n} \widetilde{x}_i = 0$

(c)

We can write:

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\sum_{i=1}^{n} \widetilde{x}_i (y_i - \bar{y})}{\sum_{i=1}^{n} \widetilde{x}_i^2} \\
&= \frac{(1/s_X) \times \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(1/s_X)^2 \times \sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= s_X \times \frac{n \times s_{XY} \times s_Y}{n \times s_X^2 \times s_Y}
\end{aligned}$$

Finally :

$$\boxed{\hat{\alpha}_1 = s_Y \times \widehat{\rho(X, Y)}}$$

(d)

By noting that $var(Y) = \sigma^2$, the sampling variance of these estimates is given by:

$$\begin{aligned} var(\hat{\alpha}_1) &= s_X^2 \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \times \sigma^2 \\ &= s_X^2 \times \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Finally :

$$\boxed{var(\hat{\alpha}_1) = \frac{\sigma^2}{n}}$$

Then :

$$var(\hat{\alpha}_0) = var(\bar{y})$$

So then

$$\boxed{var(\hat{\alpha}_0) = \frac{\sigma^2}{n}}$$

Furthermore,

$cov(\hat{\alpha}_0, \hat{\alpha}_1) = \frac{\sum \widetilde{x_i}}{\sum \widetilde{x_i^2}} cov(y_i - \bar{y}, \bar{y})$ with $cov(y_i, \bar{y}) = \frac{1}{n}\sigma^2$ as $cov(y_i, y_k) = 0$ when $i \neq k$ because samples are *iid* and also $cov(\bar{y}, \bar{y}) = \frac{1}{n}\sigma^2$.

So finally:

$$cov(\hat{\alpha}_0, \hat{\alpha}_1) = \sigma^2 \times \frac{\sum \widetilde{x}_i}{\sum \widetilde{x}_i^2} \times (\frac{1}{n} - \frac{1}{n}) \implies \boxed{cov(\hat{\alpha}_0, \hat{\alpha}_1) = 0}$$

.

(e)

Using the results from question (c), we get:

$$\boxed{\hat{\beta}_1 = \frac{\hat{\alpha}_1}{s_X}} \quad \text{and} \quad \boxed{\hat{\beta}_0 = \hat{\alpha}_0 - \frac{\hat{\alpha}_1}{s_X}\bar{x}}$$

(f)

Then, we can compute the variances :

$$var(\hat{\beta}_1) = \frac{1}{s_X^2} \times var(\hat{\alpha}_1) \implies \boxed{var(\hat{\beta}_1) = \frac{\sigma^2}{s_X^2 n}}$$

.

And also :

$var(\hat{\beta}_0) = var(\hat{\alpha}_0) + (\frac{\bar{x}}{s_X})^2 \times var(\hat{\alpha}_1)$ since we previouslsy showed that $cov(\hat{\alpha}_0, \hat{\alpha}_1) = 0$.

So finally :

$$\boxed{var(\hat{\beta}_0) = \frac{\sigma_2}{n} \times (1 + (\frac{\bar{x}}{s_X})^2)}$$

Question (3)

(a)

We pick β_1 in order to minimize the residual sum of squares $RSS(\beta) = Y^T Y + \beta^2 X^T X + 2X^T Y$ where $X = (x_1, \ldots, x_n)^T$ and $Y = (y_1, \ldots, y_n)^T$:

$$\frac{\partial RSS(\beta)}{\partial \beta} = 2\beta \times \sum x_i^2 - 2\sum x_i y_i \implies \boxed{\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}}$$

Furthermore :

$$\mathbb{E}(\hat{\beta}_1) = \frac{\sum x_i \mathbb{E}(y_i)}{\sum x_i^2} = \beta_1 \frac{\sum x_i \times x_i}{\sum x_i^2} = \beta_1$$

so the estimator is unbiased and its variance is given by:

$$\boxed{var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2}}$$

Finally, σ^2 can be estimated by the sample variance : $\frac{1}{n-1} \sum_i^n (y_i - \beta_1 x_i)^2$. So in order to estimate σ^2, we could use the plug-in estimator :

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

.

Besides, it has $n - 1$ degrees of freedom.

(b)

Let us assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then we have :

$$\boxed{\hat{\beta}_1 \sim \mathcal{N}(\beta_1, var(\hat{\beta}_1))}$$

. So under the null hypothesis $\mathcal{H}_0 : \beta_1 = 0$, we could write the t-statistic:

$$\frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim \mathcal{N}(0,1)$$

.

However, since we do not know $var(\hat{\beta}_1)$, we have to estimate $\widehat{var(\hat{\beta}_1)}$ by using the estimate of $\widehat{\sigma^2}$ of the error's variance. Finally:

$$\boxed{\frac{\hat{\beta}_1}{\sqrt{var(\hat{\beta}_1)}} \sim t_{n-1}}$$

where t_{n-1} is the Student distribution with $n-1$ degrees of freedom. Now we can compute the t-statistic, and then reject/accept the null hypothesis with a given level of confidence.

Question 4

```
linear.regression <- function(x, y, intercept = TRUE){

  if (intercept == T){
    X = cbind(1,x)
  }
  else if (intercept == F){
    X = x
  }

  beta = solve(t(X) %*% X) %*% t(X) %*% y
  estimated_cov_matrix = t(X) %*% X
  fitted_values = X %*% beta
  residuals = y - fitted_values
  R_squared = sum((fitted_values - mean(y))^2) / sum((y - mean(y))^2)

  res = list("beta" = beta,
             "estimated_cov_matrix" = estimated_cov_matrix,
             "fitted_values" = fitted_values,
             "residuals" = residuals,
             "R_squared" = R_squared)

  return(res)
}

# For testing
# abalone <- read.csv("/Users/thibaudbruyelle/Documents/Stanford/fall2020/stats305A/datasets/abalone_30
# lm(formula = Length ~ Diameter + Height, data = abalone)
# y = as.matrix(abalone$Length)
# x = as.matrix(abalone[,c(3,4)])
```

Question 5

```
path = "/Users/thibaudbruyelle/Documents/Stanford/fall2020/stats305A/datasets/abalone_305a.csv"
abalone <- read.csv(path, header = T)
```

(a)

```
y = as.matrix(abalone$Rings)
x = as.matrix(abalone$Length)
model <- linear.regression(x, y, intercept = F)
print("beta_1")
```

```
## [1] "beta_1"
```

```
print(model$beta[1])
```

```
## [1] 18.6427
```

```
print("residuals variance")
```

```
## [1] "residuals variance"
```

```
print(var(model$residuals)[1,])
```

```
## [1] 7.764876
```

Eventually, we get:

$$\boxed{\hat{\beta}_1 = 18.6427}$$

$$\boxed{\widehat{\sigma^2} = 7.7649}$$

A 95% confidence interval for $\hat{\beta}_1$:

$$\boxed{\hat{\beta}_1 \in [a, b]}$$

(c)