Homework 2

MATH 4387/5387 Fall 2020

Subrata Paul

8/24/2020

Problem 1

The linear regression model can be written in matrix notation as $y = X\beta + \epsilon$. Create the table shown below and describe what each term (y, X, β, ϵ) represents (interpretation), specify the dimension of each term (size), indicate whether we model the term as random or non-random, and whether the term is observed or unobserved.

Term	Size	Interpretation	Random?	Observable?
y	$n \times 1$	The vector of responses	Yes	Yes
β	$p \times 1$	The vector of regression coefficients	No	No
X	$n \times p$	The matrix of regressor values	No	Yes
ϵ	$n \times 1$	The vector of errors	Yes	No

Problem 2

Assuming a simple linear regression model, derive the ordinary least squares estimators of β_0 and β_1. Do not use matrix notation in deriving your solution.

Solution

Under the ordinary least squares procedure we want to minimize

$$Q = \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Setting partial derivative of Q with respect to β_0 to zero,

$$\begin{aligned}
&\frac{\partial Q}{\partial \beta_0} = 0 \\
\Rightarrow &-2\sum (y_i - \beta_0 - \beta_1 x_i) = 0 \\
\Rightarrow &\sum y_i - \beta_1 \sum x_i = n\beta_0 \\
\Rightarrow &\beta_0 = \overline{y} - \overline{x}\beta_1
\end{aligned}$$

Setting partial derivative of Q with respect to β_1 to zero,

$$
\begin{aligned}
&\frac{\partial Q}{\partial \beta_1} = 0 \\
\Rightarrow & -2\sum(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \\
\Rightarrow & \sum x_i y_i - (\overline{y} - \beta_1 \overline{x})\sum x_i - \beta_1 \sum x_i^2 = 0 \\
\Rightarrow & \beta_1 = \frac{\sum x_i y_i - \overline{y}\sum x_i}{\sum x_i^2 - \overline{x}\sum x_i} \\
\Rightarrow & \beta_1 = \frac{\sum(x_i y_i - x_i \overline{y}) - \overline{x}\sum(y_i - \overline{y})}{\sum(x_i^2 - \overline{x}x_i) - \overline{x}\sum(x_i - \overline{x})} \qquad \left[\sum(x_i - \overline{x}) = \sum(y_i - \overline{y}) = 0\right] \\
\Rightarrow & \frac{\sum(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{xy})}{\sum(x_i^2 - 2x_i\overline{x} + \overline{x}^2)} \\
\Rightarrow & \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}
\end{aligned}
$$

Problem 3

Let $H = X(X^T X)^{-1} X^T$ is the hat matrix. Prove that $I - H$ is a projection matrix (Symmetric + Idempotent).

Solution

$I - H$ is symmetric because,

$$(I - H)^T = I^T - H^T = I - H, \qquad [H \text{ is symmetric}]$$

$I - H$ is idempotent because,

$$(I - H)^2 = (I - H)(I - H) = I - 2H - HH = I - 2H - H = I - H$$

where $HH = H$ since H is idempotent.

Problem 3

While proving that $\hat{\beta}_1$ is an unbiased estimator of β_1, we represented the OLS estimate as $\hat{\beta}_1 = \sum k_i Y_i$, where $k_i = \frac{X_i - \overline{X}}{\sum(X_i - \overline{X})^2}$. Use it with the properties of k_i, that we already have proved, to derive the variance of $\hat{\beta}_1$. (Hint: in linear regression framework we assume $Var(\epsilon_i) = \sigma^2$ and $Cov(\epsilon_i, \epsilon_j) = 0$ when $i \neq j$)

Solution

$$Var(\hat{\beta}_1) = Var(\sum k_i Y_i) = \sum k_i^2 Var(Y_i) = \sum k_i^2 \sigma^2 = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}$$

Problem 4

Under simple linear regression model, the Mean Squared Error (MSE) is defined as $\frac{\sum(y_i - \hat{y}_i)^2}{n-2}$ where n is the number of observations. MSE is an unbiased estimator of σ^2, where σ^2 is the variance of ϵ_i. What is an unbiased estimator of the variance of $\hat{\beta}_1$?

Solution

$$\frac{MSE}{\sum(x_i - \overline{x})^2}$$

Because

$$E\left[\frac{MSE}{\sum(x_i - \overline{x})^2}\right] = \frac{\sigma^2}{\sum(x_i - \overline{x})^2} = Var(\hat{\beta}_1)$$

Problem 5 The square root of the variance of an estimator is the standard error (SE). You can derive the SE ($\hat{\beta}$) from problem 4. According to theory,

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

where t_{n-2} represents a Student's t distribution with $n-2$ degrees of freedom. Find an expression for 95% confidence interval of β_1.

Solution

95% confidence interval

$$\hat{\beta}_1 \pm t_{0.025,n-2} SE(\hat{\beta}_1)$$

Problem 6

If $\hat{\beta}_1 = 2$, $SE(\hat{\beta}_1) = 0.02$, and $n = 50$ calculate 95% confidence interval for β_1.

Solution

```r
2 + qt(c(0.025,0.975),50-2) * 0.02
```

```
## [1] 1.959787 2.040213
```

Problem 7

Based on the confidence interval on problem 6, perform the hypothesis test,

$$H_0: \ \beta_1 = 0 \quad \text{Vs.} \ H_1: \ \beta_1 \neq 0$$

Since the 95% confidence interval does not contains 0, we reject the null hypothesis with 5% level of significance.

Problem 8

Use the `simu_hw1.txt` data and fit a multiple linear regression model with `response` as the response variable and `pred1`, `pred2`, and `pred3` as predictors. Write down the equation of the fitted line (fitted model).

Solution

```r
library(xtable)
dat = read.table('../data/simu_hw1.txt', header = T)
lmod = lm(response ~ pred1 + pred2 + pred3, data = dat)
```

Model Equation: `E[response]` = -2.73 + 2.01 `pred1` + 3 `pred2` - -0.25 `pred 3`

```r
options(xtable.comment = FALSE)
xtable(summary(lmod))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.7287	0.0649	-42.06	0.0000
pred1	2.0115	0.0101	199.61	0.0000
pred2	2.9980	0.0063	477.19	0.0000
pred3	-0.2510	0.0101	-24.81	0.0000