Statistical linear models.

Assignment 1.

毕圣杰 11910901

1.(a). simple linear regression model:

Assume that the relationship between X and Y is linear.

satisfy $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

then we have $P(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)}{2\sigma^2}\right\}$

thus, we have

the likelihood function: $L = L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} P(y_i | \beta_0, \beta_1, \sigma^2)$

log-likelihood function: $l = \log L = \sum_{i=1}^{n} \log P(y_i | \beta_0, \beta_1, \sigma^2)$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \beta_0 - \beta_1 x_i)}{2\sigma^2}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

By differentiating $l(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2 and letting them equal 0

$\Rightarrow$ we have

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)}{\sigma^2} = 0. \quad \text{and} \quad \cdots 1^\circ$$

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i}{\sigma^2} = 0 \quad \text{and} \quad \cdots 2^\circ$$

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^4} = 0 \quad \cdots 3^\circ$$

from 1° we have $\bar{y} = \beta_0 + \beta_1 \bar{x}$

from 2° we have $\sum_{i=1}^{n} x_i y_i - n\bar{x}\beta_0 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$.

$$\xrightarrow{\beta_0 = \bar{y} - \beta_1 \bar{x}} \sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - \beta_1 \bar{x}) - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \beta_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$\Rightarrow (\hat{\beta}_1)_{MLE} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \overset{d}{=} \frac{S_{xr}}{S_{xx}} \quad \cdots 4^\circ$$

then we have $(\hat{\beta}_1)_{MLE} = \frac{S_{XY}}{S_{XX}} \Rightarrow (\hat{\beta}_0)_{MLE} = \bar{y} - \frac{S_{XY}}{S_{XX}}\bar{x}$ ... 5°.

from 4°. 5°. $\Rightarrow$ equation 3°. we have

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - (\hat{\beta}_0)_{MLE} - (\hat{\beta}_1)_{MLE}x_i)^2}{n} \xrightarrow{\text{Since } \begin{cases}(\hat{\beta}_0)_{MLE} = \hat{\beta}_0 \\ (\hat{\beta}_1)_{MLE} = \hat{\beta}_1\end{cases}} \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

thus, from above, we have

$$\begin{cases} (\hat{\beta}_0)_{MLE} = \hat{\beta}_0 = \bar{y} - \frac{S_{XY}}{S_{XX}}\bar{x} \\ (\hat{\beta}_1)_{MLE} = \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \\ (\hat{\sigma}^2)_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{cases}$$

(b). Difference:

LSE of σ^2: $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is unbiased

$E(S^2) = \sigma^2$ $(df = n-2)$

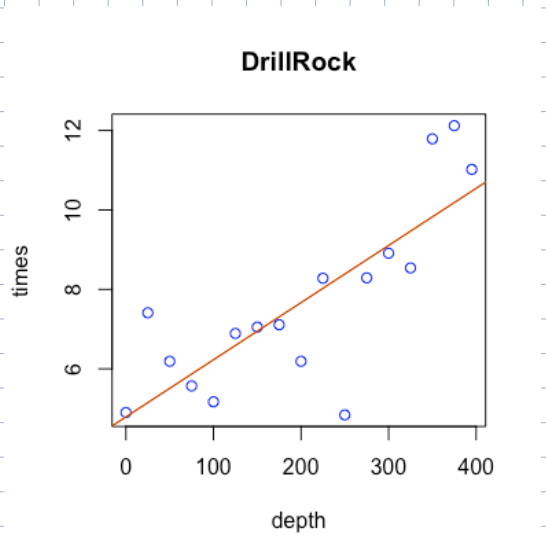
MLE of σ^2: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is asymptotically unbiased.

$$E(\hat{\sigma}^2)_{MLE} = \frac{1}{n}E\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)$$

$$= \frac{n-2}{n}E\left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}\right)$$

$$= \frac{n-2}{n}E(S^2) = \frac{n-2}{n}\sigma^2 \to \sigma^2 \text{ as } n \to +\infty$$

2.(a). See the figure



(b) See the figure By the conclusion of LSE, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

where $\bar{X} = \frac{1}{17}(0+25+50+\cdots+395) = 199.7059$

$\bar{y} = \frac{1}{17}(4.9+7.41+\cdots+11.02) = 7.663$

$S_{XY} = \sum_{i=1}^{n}(X_i-\bar{X})(y_i-\bar{y}) = (0-\bar{X})(4.9-\bar{y})+\cdots+(395-\bar{X})(11.02-\bar{y}) = 3640.465$

$S_{XX} = \sum_{i=1}^{n}(X_i-\bar{X})^2 = (0-\bar{X})^2+\cdots+(395-\bar{X})^2 = 253023.5$

$\Rightarrow$ thus we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(y_i-\bar{y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_{XX}} = \frac{3640.465}{253023.5} = 0.014388$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 7.663 - 199.7059 \times 0.014388$$
$$= 4.7896$$

$\Rightarrow \hat{y}_i = 0.014388\,X_i + 4.7896$

(c) simple linear regression model.

Assumptions: the relationship between X and Y is linear, satisfy $Y_i = \beta_0+\beta_1X_i+\varepsilon_i$

and $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ $i=1,2,\cdots,n$

(d) From the problem, we have that

$H_0: \beta_1=0$ against $H_1: \beta_1\neq 0$.

And under the assumption, we have $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX})$ where $S_{XX} = \sum_{i=1}^{n}(X_i-\bar{X})^2$ and $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2)$

$\Rightarrow$ then we have the test statistics

$$t = \frac{\hat{\beta}_1-\beta_1}{\sigma/S_{XX}^{\frac{1}{2}}} \Big/ \left(\frac{(n-2)S^2}{\sigma^2}/n-2\right)^{\frac{1}{2}} = \frac{\hat{\beta}_1-\beta_1}{S/S_{XX}^{\frac{1}{2}}} \sim t_{15}$$

where $S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \frac{1}{15}\left[(4.9-\hat{y}_1)^2+(7.41-\hat{y}_2)^2+\cdots(11.02-\hat{y}_{17})^2\right] = 2.051179$

$S_{XX} = 253023.5$ (See (b))

Since we want to test $H_0: \beta_1=0$, then we have

$$t = \frac{\hat{\beta}_1}{(S^2/S_{XX})^{\frac{1}{2}}} = \frac{0.014388}{(2.051179/253023.5)^{\frac{1}{2}}} = 5.053346 > t(0.0025, 15) = 2.131$$

$\Rightarrow$ thus we need to reject H_0: the depth of the rock provides information for the prediction of the time required a distance of 5 feet.

(e). From (d) we consider the test statistics $\frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} \sim t_{15}$

$\Rightarrow$ then we have

$$Pr\left(-t_{0.025,15} \le \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} \le t_{0.025,15}\right) = 95\%.$$

$\Rightarrow$ 95% CI of β_1 satisfy:

$$-2.131 \le \frac{\hat{\beta}_1 - \beta_1}{S/S_{xx}^{\frac{1}{2}}} \le 2.131 \quad \text{where } S/S_{xx}^{\frac{1}{2}} = (S^2/S_{xx})^{\frac{1}{2}} = (2.051179/253023.5)^{\frac{1}{2}} = 0.0028472$$

$$\Rightarrow \quad -2.131 \le \frac{0.014388 - \beta_1}{0.0028472} \le 2.131$$

$\Rightarrow$ $\beta_1 \in [0.008321, 0.020455]$ is the 95% confidence interval for β_1.

Interpretion: there is 95% probability that β_1 is in $[0.008321, 0.020455]$

(f). the coefficient of determination for the linear regression model is

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \cdots (*)$$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 30.76769$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 83.14165$$

$$\Rightarrow (*) = 1 - \frac{30.76769}{83.14165} = 0.62994 \simeq 0.63$$

Interpletion: 63% of the variation of the time required to drill a distance of 5 feet with depth of rock is explained by the model.

(g). Regression prediction equation: $\hat{y}_h = 0.014388 x_h + 4.7896$

Since $x_h = 6$. $E(y_h) = E(\beta_0 + \beta_1 x_h + \varepsilon_i) = \beta_0 + \beta_1 x_h$.

$\Rightarrow \hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = 0.014388 \times 6 + 4.7896 = 4.875928$

Note that two-sided $100(1-\alpha)\%$ CI for $E(Y_h)$ is

$$\left[\hat{y}_h - t_{\frac{\alpha}{2},n-2}\, S\sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ \hat{y}_h + t_{\frac{\alpha}{2},n-2}\, S\sqrt{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]$$

from above, we have

$S^2 = 2.051179$. $t_{0.025,15} = 2.131$. $(x_h - \bar{x})^2 = 37251.97$

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{xx} = 253023.5$

$\Rightarrow S\sqrt{\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = 0.6501132$, $t_{0.025,13}\, S\sqrt{\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = 1.385391$

$\Rightarrow$ thus we have that

95% interval for the mean amount of time to drill a distance of 5 feet when depth is 6 feet is $[3.490537, 6.261319]$

(h). Note that the two-sided $100(1-\alpha)\%$ CI for $Y_{h(new)}$ is

$$\left[\hat{y}_h - t_{\frac{\alpha}{2},n-2}\, S\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} \;,\; \hat{y}_h + t_{\frac{\alpha}{2},n-2}\, S\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right]$$

from above. we have

$S\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = 1.57284$, $t_{0.025,13}\, S\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = 3.351722$

$\Rightarrow$ thus we have that

95% interval for the single drill of time to drill a distance of 5 feet when depth is 6 feet is $[1.524206, 8.22765]$

(i) From above. we have the ANOVA table:

	SS	df	MS	F
regression	52.37396	1	52.37396	25.5335
error	30.76769	15	2.05118	
total	83.14165	16		

Consider the hypothesis: $H_0 = 0$ against $H_1: \beta_1 \neq 0$ and at $100(1-\alpha)\%$ interval.

We will reject H_0 if $F: 25.5335 > F(\alpha, 1, n-2)$