CA09 - STAT100

Fábio Rodrigues Pereira

Oppgave 1

(Fra eksamen høst 2018. Det betyr at det nå er din tur å skrive den samme rapporten som du nettopp har lest eksempelbesvarelser av. Prøv å gjøre dette uten å se på de andre besvarelsene underveis, og sammenlign din/deres rapport med de gode besvarelsene etterpå.)

De pågående klimaendringene er et til stadig tilbakevendende tema og gjenstand for bekymring, både for vanlige mennesker, politikere og forskere. Hvordan vil en øking i temperatur påvirke naturtyper og økosystem?

Korallrev inneholder noen av de mest divergente økosystemer på planeten, de er habitat og ly for mange marine organismer. Revene har også en viktig rolle i å beskytte kystlinjer ved å ha en dempende effekt på bølgekraft og tropiske stormer, derfor er det ønskelig å beskytte korallrev så mye som mulig.
Det er gjort en studie på hvordan en økning i vanntemperaturer vil påvirke vekst i korallrev. Flere år på rad samlet forskere inn data på vanntemperaturer og hvor mye koraller hadde vokst i mm per år i Rødehavet.

Data som følger:

(X_i) Temperatur(°C)	Vekst(mm/år) (Y_i)
29,7	2,63
29,9	2,57
30	2,67
30,2	2,6
30,5	2,47
30,7	2,39
30,9	2,25
31,2	2,24
31,5	2,15
28,9	2,77
29,4	2,53

Det ble utført en lineær regresjonsanalyse. Noen resultater som kan være til hjelp:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	1.06	9.35	0.0000062
Temp	-0.23	0.03	-6.91	0.0000698

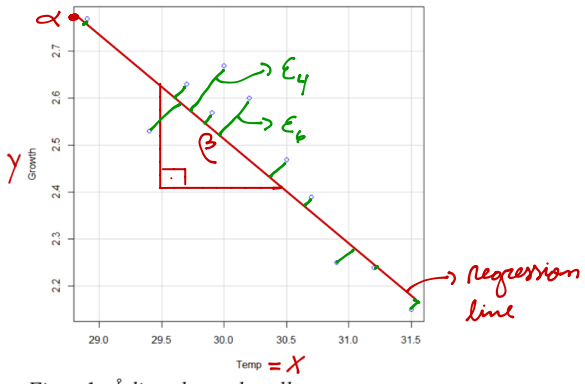
s: 0.083 on 9 degrees of freedom
Multiple R-squared: 0.8415

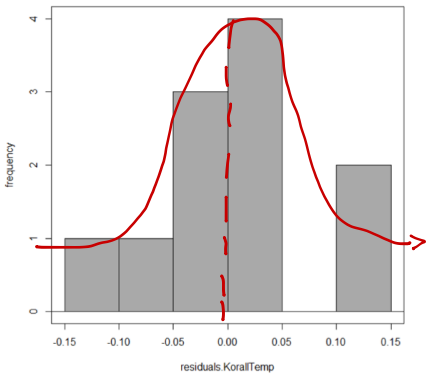
Mean Temp
30.26

*Skriv en **kort** rapport over analysen som er gjort (ta med modeller, antagelser, parametere med tolkning, estimater med usikkerhet, tester som kan være nyttige). Gi konkrete forklaringer på resultatene av forsøkene til en ikke-statistiker (for eksempel en klimaforsker).*

→ It was performed a linear regression analysis where it is aimed to know how the increase of the global temperature can impact the ecosystem. The statistical model chosen is $Y_i = \alpha + \beta X_i + \varepsilon_i$, where Y_i is the response variable, which represents the growth of the ocean reefs by mm/år, and correlated to the explanatory variable X_i which represents the independent yearly average temperature in °C. α is the expected value of Y_i when the value of X_i is zero, also called the y-intercept of the regression line. β is another parameter which is the expected increase of Y_i when X_i increases 1 unit, also called slope of the regression line. ε_i is the residual or error of the model, for i observations between 1 to n, and which is independent normally distributed with mean 0 and standard deviation represented by σ. This first introduction can be technically illustrated as:

→ Model: $Y_i = \alpha + \beta X_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma)$, independents for $i = 1, \cdots, n = 11$, such that:





Not completely symmetrical but near to normal distribution.

Figur 1: Årlig vekst av koraller mot vanntemperatur

Now, we are interested to estimate the unknown parameters α, β and σ of the model. For that, we use the technique called least squares to find unbiased estimates $\hat{\alpha}$ and $\hat{\beta}$, such that:

1.) Since $\bar{Y} = \sum_{i=1}^{11} Y_i \tilde{=} 2,48$ and $\bar{X} = \sum_{i=1}^{11} X_i \tilde{=} 30,26$, Then $\hat{\beta} = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{11}(Y_i - \bar{Y})\cdot(X_i - \bar{X})}{\sum_{i=1}^{11}(X_i - \bar{X})^2} \cdot \frac{(n-1)}{(n-1)} = -0,23$ and

$\hat{\alpha} = \bar{Y} - \hat{\beta}\cdot\bar{X} = 2,48 + 0,23 \cdot 30,26 \tilde{=} 9,50$. We got $\hat{\alpha} = 9,5$ and $\hat{\beta} = -0,23$. It is still missing $\hat{\sigma}$, which we will calculate below:

2.) We know that the unbiased sample estimator S^2 for the estimate σ^2 is $S^2 = \frac{SSE}{n-2} = MSE$. Then, we need to split the variance of the model in order to find the SSE, such that:

↳ Since
$$E[Y_i \mid X_i] = E[\alpha + \beta X_i + \varepsilon_i] = \alpha + \beta X_i + \underbrace{E[\varepsilon_i]}_{0} = \alpha + \beta X_i = \hat{Y}_i \quad (*)$$
$$Var[Y_i \mid X_i] = Var[\alpha + \beta X_i + \varepsilon_i] = 0 + 0 + \sigma^2 = \hat{\sigma}^2$$
$$SD[Y_i \mid X_i] = \sqrt{\hat{\sigma}^2} = \hat{\sigma}$$

and $Y_i = \hat{\alpha} + \hat{\beta} X_i + \varepsilon_i \rightarrow \varepsilon_i = Y_i - \underbrace{(\hat{\alpha} + \hat{\beta} X_i)}_{\hat{Y}_i} = Y_i - \hat{Y}_i$

→ explained variance within the quantity values of the coral reefs

and $Y_i = \alpha + \beta X_i + \varepsilon_i \rightarrow Y_i - \bar{Y} = \underbrace{(\hat{\alpha} + \hat{\beta} X_i)}_{\hat{Y}_i} - \bar{Y} + \underbrace{Y_i - (\hat{\alpha} + \hat{\beta} X_i)}_{\hat{Y}_i} = \underbrace{\sum_{i=1}^{11}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{11}(\hat{Y}_i - \bar{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{11}(Y_i - \hat{Y}_i)^2}_{SSE}$

SSE → unexplained variance

(*) where SST is the total variance, SSR is the regression variance and SSE is the error/residual variance.

(*) Then, the unbiased sample standard variation estimator S for the estimate $\hat{\sigma}$ is the square root of the mean of the error sum of squares, also called mean square error, such that:

↳ $S = \sqrt{S^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{11}\left[Y_i - (\hat{\alpha} + \hat{\beta} X_i)\right]^2}{9}} \tilde{=} 0,083$

Now, we want to calculate how good is the model, where we use the coefficient of determination $R^2 = \frac{SSR}{SST} = 0,8415$. This score, which has range between $0 \leq R^2 \leq 1$, indicates that our regression has a very good precision, however we still need to perform the following tests to be secure of that.

We proceed a statistical test to check if the temperature has any effect to the coral growth, such that:

↳ hypotesis claims: $H_0 := \beta = 0$ where explanatory variable X (temperature) has no impact on Y;

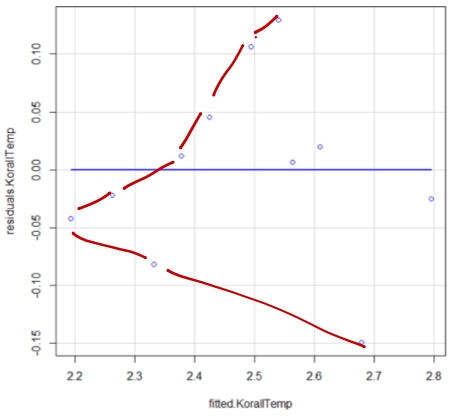
$H_1 := \beta \neq 0$ where // // // // has impact on Y;

↳ We choose a significance level $\alpha = 0{,}05$;

↳ Test statistic: $T^* = \frac{\hat{\beta} - 0}{SE[\hat{\beta}]}$ where $SE[\hat{\beta}] = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \overset{N}{=} 0{,}03$ $\therefore$ $T^* = \frac{-0{,}23}{0{,}03} \overset{N}{=} -7{,}6$ // $\therefore |T^*| = 7{,}6 > 2{,}262 = t_{9,0{,}025}$

↳ Reject H_0

* Conclusion here: we are 95% secure that the temperature has impact on the coral reefs growth.
* Also, p-value $P(T \geq T^* \mid H_0 \text{ is true}) \overset{N}{=} 0{,}00007 < 0{,}05 = \alpha$ which also shows that the claim H_0 is rejected, in other words, we are more than 95% secure that the temperature has a negative effect ($\hat{\beta}$ is negative) on the growth of the coral reefs.



Figur 2. Residualer mot tilpassede verdier

→ Despite of our conclusion that the temperature has a linear combination which a regression model shows the negative impact of the temperature on the coral reefs growth, the diagnose plot of the residual values shows a problem. This plot should have gotten points near the 0 x-axis to confirm that the regression model is linear. However, its plot shows a expansion pattern which indicates that the model is not linear. This could have happened because of the scarce number of samples and would've been fixed by applying a transformation technique.

En klimaforsker er bekymret for hvordan det vil gå med veksten i korallrev når temperaturen fortsetter å øke. Han ønsket å vite hva du anslår veksten av koraller vil være i mm per år når/hvis vanntemperaturen når 33 °C. Hva sier du til han?

we can use $\hat{Y} = \hat{\alpha} + \hat{\beta} X_i$ to predic the value of the weight of the coral given $X_i = 33°C$, such that:

↳ $\hat{Y} = 9{,}5 - (0{,}23 \cdot 33) \overset{N}{=} 1{,}90$ mm per year.

Also, it is possible to give a prediction interval, with 95% of security, such that:

↳ 95% PI for $\hat{Y} = (\hat{\alpha} + \hat{\beta} X_i) \pm t_{df, \frac{\alpha}{2}} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \left(\frac{x_i - \bar{x}}{S/SE[\hat{\beta}]}\right)^2}$

↳ $= 1{,}90 \pm 2{,}262 \cdot 0{,}083 \sqrt{1 + \frac{1}{11} + \left(\frac{33 - 30{,}26}{0{,}083 / 0{,}03}\right)^2}$

↳ $= [1{,}63 \; ; \; 2{,}17]$

$\therefore$ we are 95% secure that the prediction interval will be between 1,63 to 2,17 mm/year given 33°C.