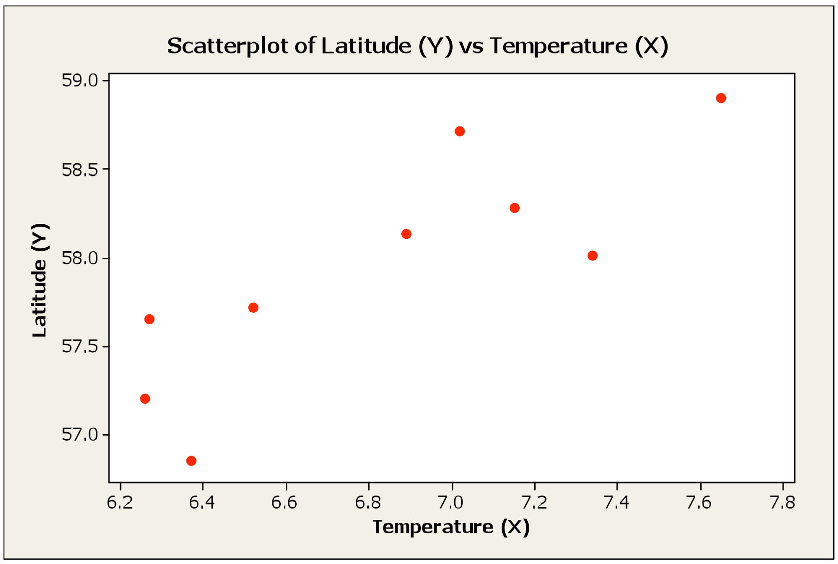
Regression

Question: With predictions for a warmer climate, will species shift their distribution (to maintain their preferred temperature range)?

Data: The data below comes from a study by Allison Perry and co-workers (Perry et al. 2005. Climate change and distribution shifts in marine fishes. *Science* 308: 1912-1915). Researchers used survey data to determine centers of abundance for bottom-dwelling fish in the North Sea (= latitude) over a 25-year period. **They asked whether mean winter temperatures at the bottom of the North Sea (degrees C) were predictive of fish species distribution (latitude).**

We will be considering some of the data for anglerfish (*Lophius piscatorius)* from this study.

1) Examine the data - *scatterplot* (2 numerical variables)



2) Determine the *least squares regression line* (prediction equation)

Response variable: latitude (Y) Explanatory variable: temperature (X)

Temp (X_i)	*Lat* (Y_i)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
6.26	57.20	-0.57	-0.74	0.3249	0.5476	0.4218
6.27	57.65	-0.56	-0.29	0.3136	0.0841	0.1624
6.37	56.85	-0.46	-1.09	0.2116	1.1881	0.5014
6.52	57.72	-0.31	-0.22	0.0961	0.0484	0.0682
6.89	58.13	0.06	0.19	0.0036	0.0361	0.0114
7.02	58.71	0.19	0.77	0.0361	0.5929	0.1463
7.15	58.28	0.32	0.34	0.1024	0.1156	0.1088
7.34	58.01	0.51	0.07	0.2601	0.0049	0.0357
7.65	58.90	0.82	0.96	0.6724	0.9216	0.7872

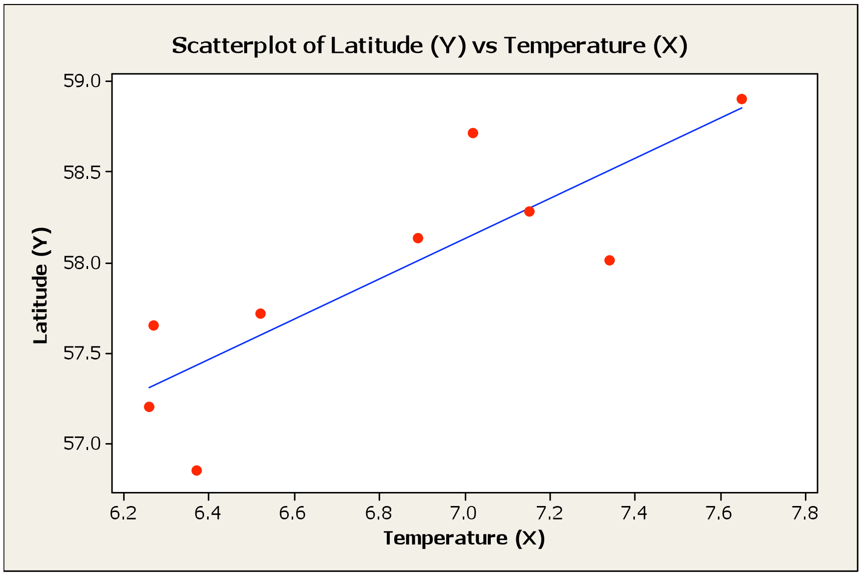
Average temperature $(\bar{X})$ = 6.83 Sum of columns: 2.02 3.54 2.24

Average latitude $(\bar{Y})$ = 57.94

Equation for a line: $Y = a + b\,X$ a = *Y- intercept* b = *slope*

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{2.24}{2.02} = 1.109 \qquad a = \bar{Y} - b\,\bar{X} = 57.94 - 1.109\,(6.83) = 50.37$$

Least squares regression line: $Y = 50.37 + 1.109\,X$



3. Statistical inference about the slope (β)

A) Hypothesis Test

Step 1: Does knowledge of the ocean temperature predict species distribution (latitude)?

Step 2: H_0: The slope of the regression of latitude on temperature is zero ($\beta = 0$).
H_A: The slope of the regression of latitude on temperature is not zero ($\beta \neq 0$).

Steps 3 and 4: Collect a random sample and calculate descriptive statistics (see previous page).

Step 5: Determine how compatible data are with results expected under H_0 (calculate test statistic, compare to null distribution, and assign *P*-value).

2 methods: *t*-test of β or ANOVA approach [you are responsible for doing tests using the *t*-test approach and for understanding the ANOVA table (but not creating it)]

test statistic for *t*-test: $t = \dfrac{b - \beta_0}{SE_b}$ $df = n\text{-}2$

$$SE_b = \sqrt{\frac{MS_{residual}}{\sum (X_i - \bar{X})^2}} \qquad MS_{residual} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2}$$

n = number of (X,Y) pairs

Lat (Y_i)	$\hat{Y}_i$	$(Y_i - \hat{Y}_i)$	$(Y_i - \hat{Y}_i)^2$
57.20	57.3123	-0.11234	0.012620
57.65	57.3234	0.32657	0.106648
56.85	57.4343	-0.58433	0.341442
57.72	57.6007	0.11932	0.014237
58.13	58.0110	0.11899	0.014159
58.71	58.1552	0.55482	0.307825
58.28	58.2994	-0.01935	0.000374
58.01	58.5101	-0.50006	0.250060
58.90	58.8539	0.04615	0.002130
			SUM: 1.0495

$MS_{residual} = 1.0495/7 = 0.1499$

$SE_b = \sqrt{\frac{.1499}{2.02}} = 0.272$ $t = \frac{1.109 - 0}{0.272} = 4.077 \approx 4.08$ df = 9 – 2 = 7

With t distribution and df = 7, $0.001 < P < 0.01$

Step 6: Make a decision regarding H_0. P-value < α, we reject H_0.

Step 7: Conclusions.

With $\alpha = 0.05$, we have evidence ($t_7 = 4.081$, $P < 0.01$) from a sample of 9 centers of abundance for the anglerfish *Lophius piscatorius* that the population slope of latitude on temperature is positive. Knowledge of the ocean temperature is predictive of anglerfish distribution, with greater temperatures leading to higher latitudes (latitude = 50.37 + 1.109 temperature).

ANOVA (from Minitab)

Source of variation	**SS**	**df**	**MS**	**F**	***P*-value**
Regression	2.4901	1	2.4901	16.61	0.005
Residual (error)	1.0492	7	0.1499		
Total	3.5393	8			

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{2.4901}{3.5393} = 0.7036$$

Approximately 70.4% of the variation in latitude can be explained by regression on temperature.

B. Confidence interval for the slope, (1 – α)% CI for β: $b \pm t_{\alpha(2),df} SE_b$

For 95% CI and $df = 7$, $t_{\alpha(2),df} = 2.36$ $1.109 \pm 2.36(0.272) = 1.109 \pm 0.642$ (0.467, 1.751)

For 99% CI and $df = 7$, $t_{\alpha(2),df} = 3.50$ $1.109 \pm 3.50(0.272) = 1.109 \pm 0.952$ (0.157, 2.061)