Stats: Problem Set 12

Q.3 (Ex 15.7 Q.8)

Ans. Since the scatterplots of the 2 midterms are ellipsoidal & also there is decent correlatⁿ of 0.5 b/w the 2 exam scores, we can find a linear regression line that predicts the scores of a test based on the other midterm test

$$\text{Slope of line} = \text{cor}(X, Y) * \frac{\text{S.d}(Y)}{\text{S.d}(X)}$$

$$= 0.5 * \frac{12}{10}$$

$$= \boxed{0.6}$$

Line 2

$$\text{Slope} = \text{Cor}(X, Y) . \frac{\cancel{\text{Slope}}\ \text{S.d}(X)}{\text{S.d}(Y)}$$

$$= 0.5 * \frac{10}{12}$$

$$= \boxed{0.4167}$$

$$\text{Intercept of regression line} = \text{mean}(Y) - (\text{slope} * \text{mean}(X))$$

$$= 64 - 0.6 * 75$$

$$= 64 - 45$$

$$= \boxed{19}$$

∴ Regression line is:

$$\boxed{Y = 0.6x + 19}$$

Line 2.

Intercept:

$$= 75 - 0.41667 * 64$$

$$= \boxed{48.333}$$

(a) Jill scored 80pts on Test 1

Her predicted score in Test 2 is:

$$Y = 0.6 \times 80 + 19$$

$$Y = 48 + 19 = \boxed{67\text{ pts}}$$

∴ Jill's suggestion of replacing her 2nd test score with 80 is overestimated. Her predicted score for test 2 is 67pts.

(b) Jack scored 76pts on Test 2

His expected score in Test 1 is:

~~$76 = 0.6x + 19$~~
~~$57 = 0.6x$~~
~~$\boxed{95 = x}$~~

$y = 0.4167 * 76 + 48.333$
$\boxed{y = 80.0022 \text{ pts.}}$

$\therefore$ Jack's estimate of his 1st midterm test score is also overvalued ~~undervalued~~.
It should be 80pts instead of 85pts

Ans. 4.

	Height	Log (Weight)
Mean	162.1cm	4.289
S.d	7.3 cm	0.253

Cor (Height, log (Weight)) = 0.3075

(a) Slope of Regression line $= \text{Cor}(x, y) * \dfrac{\text{S.d}(y)}{\text{S.d}(x)}$

$= 0.3075 * \dfrac{0.253}{7.3}$

$= \boxed{0.0107}$

Intercept of Regression line :
mean(y) − (Slope * mean(x))
$= 4.289 - (0.0107 * 162.1)$
$= \boxed{2.5545}$

Eq

Eq^n of Regression Line : $\boxed{y = 0.0107x + 2.5545}$
Here y is log(weight) & x is Height of adult women in U.S

(b) For adult women who are 165 cm tall, y = log(weight) is :
$y = 0.0107 * 165 + 2.5545$
$y = 4.32$
$\therefore$ mean log weight is $\boxed{4.32}$ ~~& S.D of log weight remains 0.253~~

(b) S.d of log weight becomes : $S_y(\sqrt{1-r^2})$ [Predict^n error]

$= 0.253 * \sqrt{1-(0.3075)^2}$

$= \boxed{0.2407417}$

(c) $Y = \log_e(80) = 4.382027.$

```
# Rcode
P(Y < 4.382027)
= pnorm(4.382027, mean = 4.32, s.d = 0.2407417)
= 0.6016613
```

Ans.5 (a) The Correlat^n b/w team's wins last season & this season is given as 0.54. This is not a perfect correlat^n despite being a decent enough correlat^n to conclude that these 2 variables ARE related. So, we cannot conclude that if a team won 'n' matches in 2014, it will win the same 'n' matches in 2015. If it were so, the correlat^n should have been a perfect one of 1

(b) Slope of Regress^n line $= \text{Cor}(X,Y) * \frac{S.d(Y)}{S.d(X)}$

$= 0.54 * \frac{11.7}{11.7}$

$= 0.54.$

Intercept $= \bar{Y} - (\text{slope} * \bar{X})$

$= 81 - (0.54 * 81)$

$= 81 - 43.74$

$\boxed{\text{Intercept} = 37.26}$

$\therefore \ y = 0.54x + 37.26$

For LA angels $x = 98$, $\therefore$ wins in 2015 prediction or y is :

$y = 0.54 * 98 + 37.26$

$y = 90.18 \sim 90$ matches

$\therefore$ Predicted wins for LA Angels in 2015 are $\boxed{90}$

(C) The regression predictions are always going to not model the outliers pretty well simply because there isn't much data to model. So the prediction for the outlier is always going to regress towards the mean & be less than the actual value. That is why predicted max games won is less than 96 or 91 to be precise.