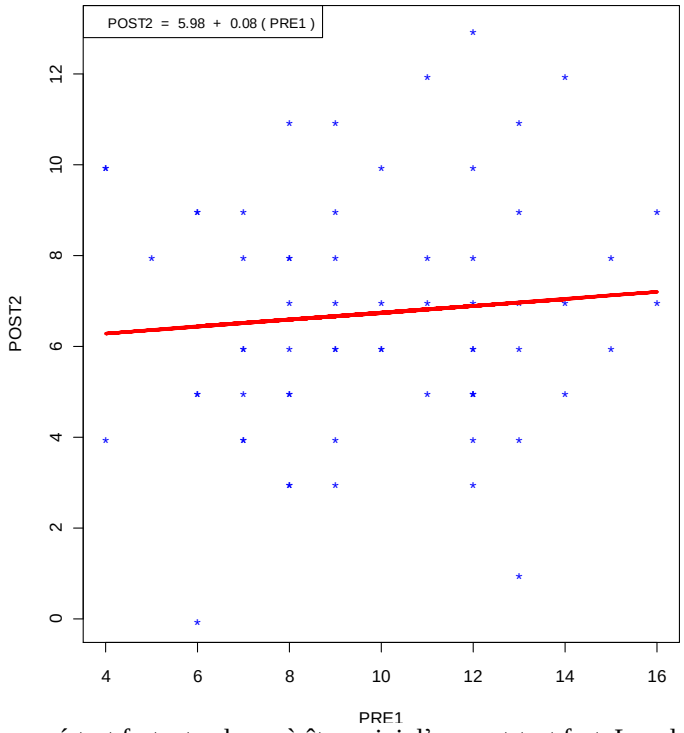
MAT7381 Solution #4.28

4.28 a) , -b)



La relation est positive : un pré-test fort a tendance à être suivi d'un post-test fort. La relation, cependant semble très faible (r = 0,089) et pourrait s'avérer non significative. La droite pré-test = $\beta_0+\beta_1 x$ qui exprime la relation entre le pré-test est le post-test est estimée par Post-test = 5,9534+ 0,07752(pré-test).

4.28-c)

Voici un tableau qui résume les résultats de la régression.

```
> summary(a)
Coefficients:
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept)   5.95340    1.11212     5.353   1.25e-06 ***
x             0.07752    0.10864     0.714      0.478

Residual standard error: 2.646 on 64 degrees of freedom
Multiple R-squared: 0.007892, Adjusted R-squared: -0.00761
F-statistic: 0.5091 on 1 and 64 DF, p-value: 0.4781
```

L'estimation $\hat{\beta}_1$ = 0,07752 de β_1 est est sujette à des variations aléatoire, mesurées par un écart-type σ estimé par $\hat{\sigma}_{\hat{\beta}_1}$ = 0,10864. Par conséquent, la valeur observée de $\hat{\beta}_1$ centrée-réduite correspond à une cote de $T = \dfrac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$ = 0,714, ce qui n'a rien d'improbable sous l'hypothèse H que $\beta_1 = 0$ (en d'autres termes, qu'il n'y a pas réellement de relation entre le pré-test et le post-test. Cette conclusion est confirmée par la valeur p, qui représente la probabilité que la variable T prenne, en valeur absolu, une valeur aussi élevée que celle qu'elle a prise. La valeur p de 0,4781 montre que le résultat observé n'a rien d'improbable sous l'hypothèse H. Nous ne pouvons donc pas conclure avec confiance qu'il y a une relation entre pré-test et post-test.

4.28-d)

Voici la table d'analyse de variance :

```
> anova(a)
            Df    Sum Sq   Mean Sq   F value   Pr(>F)
x            1      3.56    3.5634    0.5091   0.4781
Residuals   64    447.97    6.9995
```

Le tableau présente deux sommes de carrés, la somme des carrés expliquée et la somme des carrés résiduelle (2e colonne). La somme des carrés expliquée (SCE) constitue une partie de la variation totale des scores au post-test, celle qu'on peut en partie expliquer par les variations des pré-tests. La somme des carrés résiduelle (SCR) estime essentiellement la variation entre les post-tests d'une population de personnes ayant le même pré-test. Cette variation ne peut s'expliquer par des variations de pré-test. SCE et SCR sont normalisées pour tenir compte du nombre de termes qui constituent ces sommes (le nombre de degrés de liberté, 1ère colonne). Les sommes normalisées sont les moyennes de carrés MCE et MCR (4e colonne) et la statistique *F* (5e colonne) est le rapport des deux, *F* = MCE/MCR. MCE et MCR devraient avoir à peu près la même valeur lorsque le pré-test n'est pas lié au post test et dans ce cas, *F* devrait être proche de 1. Si, au contraire, la relation entre pré-test et post-test est forte, MCE aura tendance à prendre une valeur plus élevée que celle de MCR et *F* a tendance à prendre une valeur supérieure à 1 — d'autant plus supérieure que la dépendance est forte. On rejettera l'hypothèse H, donc, si *F* prend une valeur peu probable. La valeur p, donc, est la probabilité (sous l'hypothèse H) que *F* prenne une valeur aussi élevée que celle qu'elle a prise. Si elle est très faible, on rejette H. Elle ne l'est pas, ce qui signifie que nous ne pouvons pas conclure avec confiance qu'il y a une relation entre pré-test et post-test. La conclusion est la même qu'au numéro précédent car en fait les deux tests sont équivalents. Il est important de réaliser que nous n'affirmons *pas* qu'il n'y a pas de relation entre les deux variables.

3.28-e)

On estime le score moyen au post-test de ceux dont le score au pré test est 10 est de 6,728 et pour tenir compte de l'erreur de cette estimation, on l'entoure d'une marge d'erreur qui nous permet d'affirmer avec 75 % de confiance que cette moyenne se situe entre 6,35 et 7,11.
Pour un pré-test de 18, la moyenne estimée du post-test est 7,35 et on affirme avec 75 % de confiance que cette moyenne se situe entre 6,25 et 8,45. Nos estimations sont toujours moins fiables lorsqu'on estime une moyenne pour un pré-test éloigné du centre des données. C'est ce qui explique que l'intervalle est plus large, puisque 10 est plus proche de la moyenne (9,79) que 18.

```
> predict.lm(a,data.frame(x=c(10,18)),interval="confidence",level=.75)
fit lwr upr
1 6.728564 6.349568 7.107561
2 7.348697 6.246143 8.451251
```

```
> xbar<-mean(x)
> xbar
[1] 9.787879
> (10-xbar)
[1] 0.2121212
> 18-xbar
[1] 8.212121
```

3.28-f)

On affirme avec 75 % de confiance que le score au post-test d'un nouvel individu ayant un score au pré-test de 10 au pré-test aura un score entre 3,63 et 9,82 au post-test.
La prédiction pour une personne ayant un pré-test de 18 est que son post-test est entre 4,09 et 10,61.
Ces limites de prédiction sont naturellement beaucoup plus larges que les intervalles de confiance pour les moyennes. Nous avons affirmé que la moyenne de ceux qui ont pré-test de 10 se situe entre 6,35 et 7,11. Un individu parmi ceux-là va normalement s'écarter de la moyenne et pourrait bien avoir un score inférieure à 6,35 ou supérieure à 7,11.

```
> predict.lm(a,data.frame(x=c(10,18)),interval="prediction",level=.75)
fit lwr upr
1 6.728564 3.633966 9.823162
2 7.348697 4.085489 10.611905
```

[NB Il est important de réaliser que nous effectuons ces calculs pour l'exercice et seraient difficiles à justifier en pratique, puisqu'elles ont pour prémisse que la relation entre les deux variables est réelle, chose que nous n'avons pas pu conclure. Si on admet que le pré-test n'est nullement lié au post-test, alors un intervalle de confiance pour la moyenne est donné par $1;/2\ 1;/2[\ ;\ ]\ nynyyytsnytsn\ -\alpha-\alpha-+$ ce qui donne [6,34 ; 7,09], ce qui est pratiquement le même que celui obtenu plus haut, une autre indication de l'absence de relation entre le pré-test et le post-test : on n'améliore pas l'estimation du post-test en se servant de l'information sur le pré-test].