REGRESSION

|  |  |  |
| --- | --- | --- |
| **QUESTION 1** | |  |
| **Pour la régression simple**, | Formule | |
| Total de la variabilité à expliquer pour Y= | variabilité de Y expliquée par la régression + variabilité de Y non expliquée par la régression. = SCR+SCE = SCT | |
| Écart-type de l’estimateur | S = SQRT(MSE) mesure la fiabilité des estimateurs produits par la régression. s représente la précision des prévisions générées par l’équation de la régression estimée. Plus la valeur de s est faible mieux c’est  , moyenne de la somme des erreurs au carré mesure la variabilité de la variable cible laissée non expliquée par la régression. | |
|  |  | |
| **QUESTION 2** | | / |

Le fichier examRevision.xlsx contient 20 observations pour un examen. Chaque observation concerne un étudiant et reprend :

1) le score qu’il a obtenu à cet examen (cote sur 100)  
2) le nombre d’heures de révision   
3) le niveau d’anxiété  
4) des points représentant son niveau au départ.

A partir de ces données , les résultats ci-dessous ont été obtenus (cfr tableaux )

1. Quelle régression simple vous paraît la plus pertinente ? Justifiez.  
     
   Variable cible : Exam score  
   Variable explicative : Level Entry  
   Justification : Je remarque que la plus grande corrélation dans le tableau entre ces deux valeurs est celle de level-entry & Exam score
2. Expliquez et interprétez contextuellement le paramètre R-square du tableau model summary de la régression score – anxiété.  
   Explication : Cette valeur représente le taux d’explication de l’anxiété par rapport à notre taux de score aux examens pour les élèves.

Interprétation contextuelle : L’anxiété n’explique qu’a 1,4% le taux score

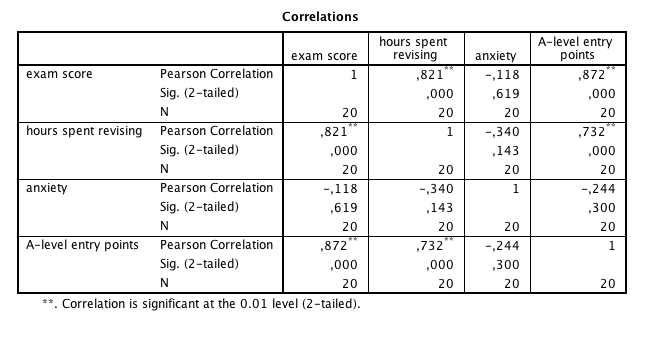
1. Pour cette même régression, expliquez et interprétez contextuellement le test F.  
     
   Le test F permet de valider ou non le modèle de regression appliqué. Le test F verifiera si ce qu’on explique est significativement > que ce qu’on explique pas.

H0 : variance Expliquée / VarianceNon Expliqué =1 --> on explique pas plus

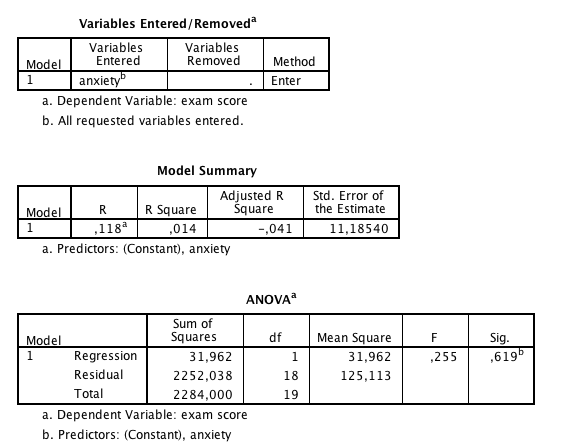
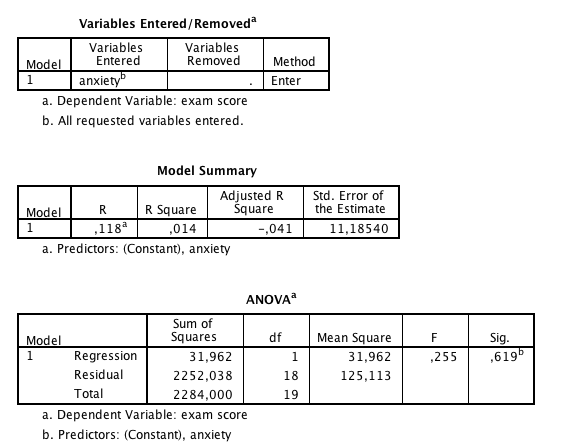
H1 : variance Expliquée / VarianceNon Expliqué > 1 🡪 on explique plus que ce qu’on explique pas

Dans le tableau on remarque que la p-value du test F est de 61,9% et donc on ne rejette pas H0 on conclu donc qu’on explique pas plus que ce qu’on explique pas. Et donc le modèle de regression ‘nest pas bon.

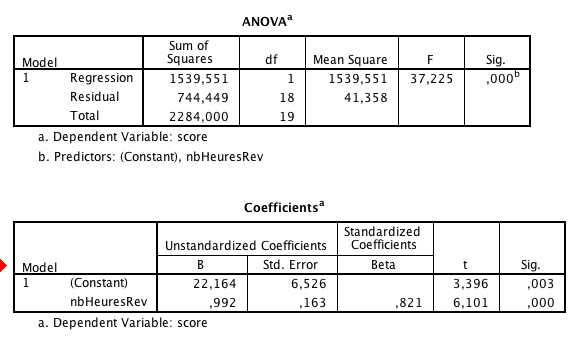
1. Quelle est l’équation de la régression simple score – nombre d’heures de révision ?  
     
   y = 0,992\* X + 22,164
2. Selon le modèle proposé ci-dessous, combien d’heures de révision seraient nécessaires pour passer d’un score de 60 à un score de 70 ? Justifiez.  
     
   x = (60-22 ,164)/0,992 et de même pour 70 ainsi la différence des deux résultat = 10,08 donc 10,08H



Quelques résultats score-anxiété



Quelques résultats score – nombre d’heures de révision

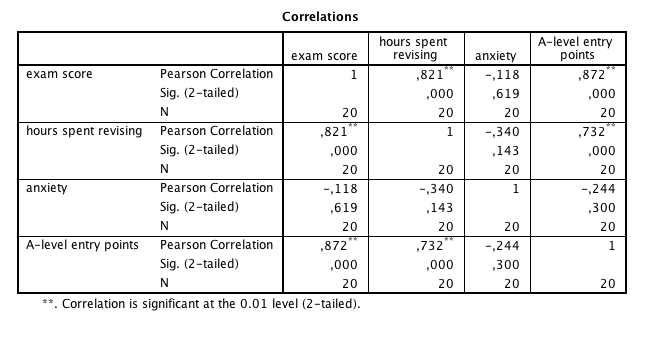
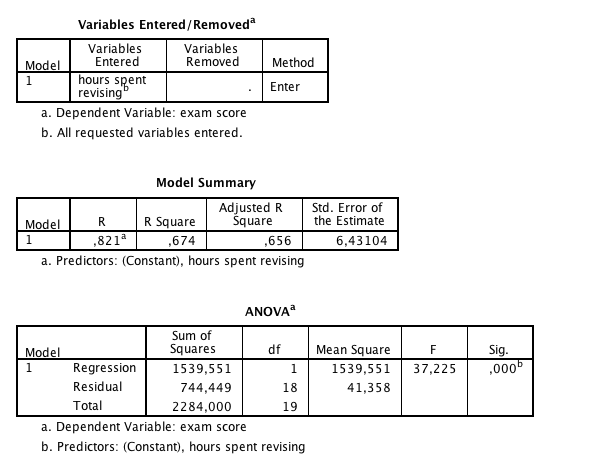
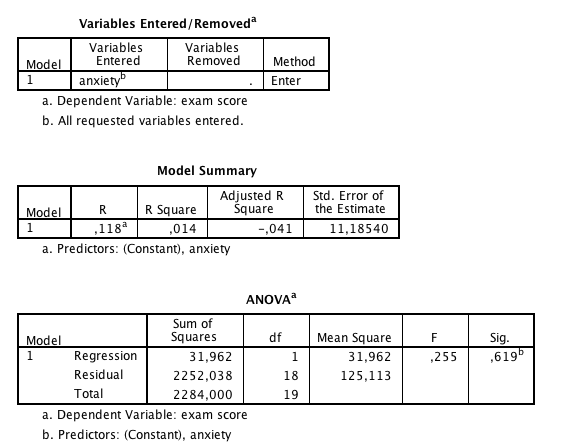


|  |  |
| --- | --- |
| **QUESTION 2 (régression linéaire et multiple)** |  |

Le fichier examRevision.xlsx contient 20 observations pour un examen. Chaque observation concerne un étudiant et reprend :

1) le score qu’il a obtenu à cet examen   
2) le nombre d’heures de révision   
3) le niveau d’anxiété  
4) des points représentant son niveau au départ.

A partir des données du fichier « examRevision.xlsx » et des résultats consignés dans le fichier Resultats-Reg.docx, on souhaite trouver un modèle de régression linéaire

* quelle sera la variable cible ?  
  Exam score
* quelles pourraient être les variables explicatives ? Justifiez
* pour chacune des variables explicatives, testez la pertinence de la régression correspondante en calculant les paramètres les plus adéquats. Notez vos solutions et argumentez ;
* trouvez la régression qui vous paraît la plus significative et argumentez;
* calculez le SSE de cette régression et précisez ce qu’il signifie ;
* que signifie l’indicateur s (std error of the estimate) dans une régression linéaire simple ?
* combien de point(s) supplémentaire(s) à l’examen peut-on espérer en étudiant cette matière une heure de plus ? Justifiez.
* quelle est la différence principale entre l’information donnée par les tests T et celle donnée par le test F ?
* 
* 
* 

|  |  |
| --- | --- |
| **QUESTION 2 (régression linéaire)** | / 10 |

A partir des données du fichier « interro.xlsx », on souhaite trouver un modèle de régression linéaire  
  
- quelle sera la variable cible ?

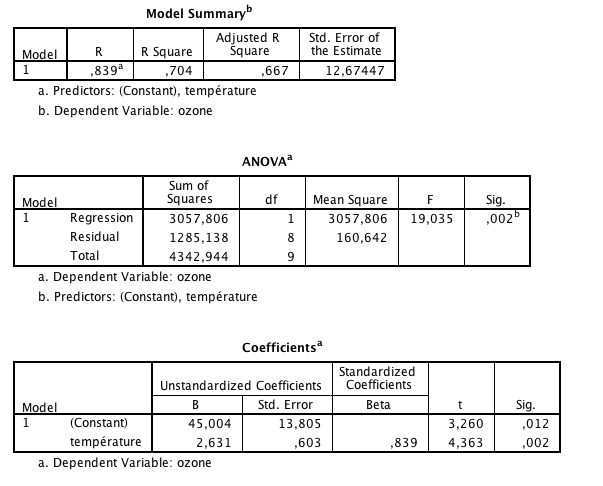
- quelles pourraient être les variables explicatives ?  
  
  
- pour chacune des variables explicatives, testez les différentes régressions possibles en calculant ;   
notez ici les solutions obtenues :

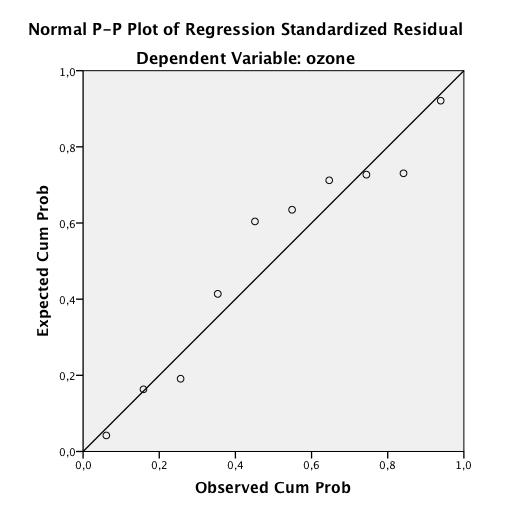
- trouvez celle qui vous paraît la plus significative et justifiez ;

- quel indicateur supplémentaire faudrait-il ajouter pour comparer les régressions ? Expliquez.

|  |  |
| --- | --- |
| **QUESTION 3 (régression linéaire)** | / 10 |

Repartons de l’exercice qui expliquait la concentration d’ozone en fonction de la température (les données se trouvent aussi dans le fichier interro.xlsx). Voici l’analyse de la régression proposée par un logiciel statistique :





* Expliquez de manière générale ce que représentent les valeurs fléchées (il y en a 3) ainsi que le graphique ci-dessus et interprétez-les dans le contexte de l’exercice température-ozone
* Les données température-ozone présente une valeur qui pourrait avoir un levier élevé, laquelle ? Expliquez.