



Universidad Simón Bolívar.

Departamento de Computo Científico y Estadística.

Estadística para Ingenieros CO-3321 y para Matemáticos CO-3322

Enero-Marzo 2012

Proyecto II (10 puntos)

Considere el conjunto de datos *swiss* que está disponible en R. Estos datos constan de medidas de fertilidad e indicadores socio-económicos para cada una de las 47 provincias de habla francesa de Suiza en el año 1888.

Los indicadores socio-económicos incluidos en el estudio son:

- **Agriculture:** porcentaje de la población dedicado a la agricultura.
- **Examination:** porcentaje de reclutas que obtuvieron el máximo puntaje en los exámenes del ejército.
- **Education:** porcentaje de la población que ha recibido educación más allá de la escuela primaria.
- **Catholic:** porcentaje de católicos.
- **Infant.Mortality:** Mortalidad infantil (en porcentaje).

La variable respuesta es un índice ajustado de fertilidad, **Fertility**.

1. Realice un análisis descriptivo de los datos e indique los hallazgos más importantes. **(0.5 puntos)**.
2. Calcule e interprete la matriz de correlación de los datos, así como los diagramas de dispersión de las variables. **(0.5 puntos)**.
3. Formule y estime el (o los) modelo(s) lineal(es) que Usted propone para predecir a **Fertility**, compárelos de ser el caso. Realice un análisis de residuos del modelo que haya seleccionado. **(1.5 puntos)**
4. Escriba la ecuación e interprete los coeficientes del modelo seleccionado en la parte 3. **(0.5 punto)**
5. Supongamos que en el estudio no fueron incluidas cuatro provincias que, para el momento, presentaban los siguientes indicadores socio-económicos:

Nº provincia	Agriculture	Examination	Education	Catholic	Infant.Mortality
1	60	23	50	70	11
2	70	25	40	60	15
3	10	24	10	4	20
4	20	20	9	10	25

Para los datos anteriores, ¿cuáles serían las predicciones de **Fertility**? Grafique las bandas del intervalo de confianza y de predicción a un nivel de significancia de 1%. Use el (o los) modelo(s) propuestos. **(1 punto)**

6. Implemente una función, en el software estadístico R, que realice una prueba de hipótesis con muestras pequeñas para $\mu_1 - \mu_2$. Esta función debe incluir los tres tipos de hipótesis posibles y algunos valores por defecto. Utilice el código para responder lo siguiente:

Un investigador afirma que el índice de fertilidad de las provincias con un porcentaje de educación (porcentaje de la población que ha recibido educación más allá de la escuela primaria) mayor e igual a 9.6% es menor a, en promedio, a los índices de fertilidad de las provincias con un porcentaje de educación (porcentaje de la población que ha recibido educación más allá de la educación primaria)

menor a 9.6%. ¿Los datos muestrales aportan suficiente evidencia para contradecir lo afirmado por el investigador en el nivel de significancia de 0.025? Interprete y concluya a partir de la salida de su código. **(2 puntos)**

7. Construya una tabla de distribución de frecuencias observadas para la variable **“Infant.Mortality”**. Pruebe la hipótesis de que estos datos se ajustan a una distribución Gaussiana. Use $\alpha=0.05$ (**1.5 puntos**)
8. Supongamos que, las 47 provincias de Suiza, se pueden clasificar en cuatro grupos. Para construir los grupos utilice los cuartiles de la variable **Education** de la siguiente forma:
 - Grupo I: provincias cuyo porcentaje de personas con educación más allá de la escuela primaria es menor que el cuartil 1.
 - Grupo II: provincias cuyo porcentaje de personas con educación más allá de la escuela primaria está en el intervalo [cuartil 1, cuartil 2).
 - Grupo III: provincias cuyo porcentaje de personas con educación más allá de la escuela primaria está en el intervalo [cuartil 2, cuartil 3).
 - Grupo IV: provincias cuyo porcentaje de personas con educación más allá de la escuela primaria es mayor e igual al cuartil 3.

Considere para cada provincia de cada grupo, el porcentaje de mortalidad infantil y construya una tabla con el siguiente formato:

Grupo I	Grupo II	Grupo III	Grupo IV
Mortalidad provincia 1 del Grupo I	Mortalidad provincia 1 del Grupo II	Mortalidad provincia 1 del Grupo III	Mortalidad provincia 1 del Grupo IV
Mortalidad provincia 2 del Grupo I	Mortalidad provincia 2 del Grupo II	Mortalidad provincia 2 del Grupo III	Mortalidad provincia 2 del Grupo IV
⋮	⋮	⋮	⋮

Responda justificadamente: ¿los datos de la tabla anterior presentan suficiente evidencia para indicar una diferencia en el porcentaje promedio de mortalidad infantil para los cuatro grupos de provincias? Reporte su conclusión en base al nivel de significancia alcanzado (p-valor). **(1.5 puntos)**

Condiciones de entrega:

- I. El proyecto debe entregarse impreso en semana 11.
- II. Se aceptan únicamente grupos de mínimo 2 personas y máximo 3, de la misma sección del laboratorio. En caso de que no se cumpla esta condición el proyecto no será corregido.
- III. Recuerde que en la evaluación se hará énfasis sobre el seguimiento de las recomendaciones para elaborar informes que está disponible en la página del curso en Aula Virtual. La estructura del proyecto vale **1 punto**.