Code listing formatting

Applied Biostatistics with R and rk.Teaching

Alfredo Sánchez Alberca (asalber@ceu.es)

Department of Applied Math and Statistics
CEU San Pablo

March 7, 2016

CEU
Universidad San Pablo

Contents

Practice 1

Linear regression

1 Solved exercises

1. The values of two variables X and Y measured in a sample of 10 individuals are:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Do the following operations:

(a) Create a data set with the variables X and Y amd enter the data.

(b) Construct the scatter plot of X and Y.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, select the variable Y in the field Variable Y, and the variable X in the field Variable X, and click the button Submit.

According to the point cloud, what type of regression model explains better the relation between X and Y?

(c) Compute the regression line of Y on X.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, insert the variable Y in the field Dependent variable and the variable X in the field Independent variable, and click the button Submit.

(d) Plot the regression line on the scatter plot.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, insert the variable Y in the field Variable Y and the variable X in the field Variable X.
3. In the Fitted line tab, check the box Linear and click the button Submit.

(e) Compute la regression line of X on Y and plot it on the scatter plot.

Repeat the steps of the previous part but inserting the variable X in the field Dependent variable and the variable Y in the field Independent variable.

(f) How are the residuals? Comment the results.

2. En una licenciatura se quiere estudiar la relación entre el número medio de horas de estudio diarias and el número de asignaturas suspensas. Para ello se obtuvo la siguiente muestra:

Horas	Suspensos	Horas	Suspensos	Horas	Suspensos
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

Do the following operations:

(a) Create a data set con las variables horas.estudio and suspensos e introducir estos datos.

(b) Construir la frequency table bidimensional de las variables horas.estudio and suspensos.

1. Select the menu Teaching » Frequency distribution » Tabla de frecuencias bidimensional.
2. In the dialog displayed, select the variable horas.estudio in the field Variable a tabular en filas, la variable suspensos in the field Variable a tabular en columnas, and hacer clic sobre el botón Submit.

(c) Compute la regression line of suspensos sobre horas.estudio and dibujarla.

Para calcular la recta de regresión:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, select the variable suspensos in the field Variable dependiente and la variable horas.estudio in the field Independent variable, seleccionar Guardar el modelo, introducir un nombre para el modelo and click the button Submit.

Para dibujar la recta de regresión:

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, select the variable suspensos in the field Variable Y y la variable horas.estudio in the field Variable X.
3. En la solapa Fitted line, seleccionar Lineal and click the button Submit.

(d) Indicar el coeficiente de regresión de suspensos sobre horas.estudio. ¿Cómo lo interpretarías?

El coeficiente de regresión es la pendiente de la recta de regresión.

(e) La relación lineal entre estas dos variables, ¿es mejor o peor que la del ejercicio anterior? Comentar los resultados a partir las gráficas de las rectas de regresión and sus residuos.

(f) Compute los coeficientes de correlación and de determinación lineal. ¿Es un buen modelo la recta de regresión? ¿Qué porcentaje de la variabilidad del número de suspensos está explicada por el modelo?

El coeficiente de determinación aparece en la ventana de resultados como R^2, and el coeficiente de correlación es su raíz cuadrada.

(g) Utilizar la recta de regresión para predecir el número de suspensos correspondiente a 3 horas de estudio diarias. ¿Es fiable esta predicción?

1. Select the menu Teaching > Regression > Predicciones.
2. In the dialog displayed seleccionar como modelo de regresión la recta calculada en el segundo apartado, introducir los valores para los que se desea la predicción in the field Predicciones para and hacer clic sobre el botón Submit.

(h) Según el modelo lineal, ¿cuántas horas diarias tendrá que estudiar como mínimo un alumno si quiere aprobarlo todo?

Seguir los mismos pasos de los apartados anteriores, pero escogiendo como variable dependiente horas.estudio, y como independiente suspensos, and haciendo la predicción para 0 suspensos.

3. Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180	210
Concentración (gramos/litro)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

Do the following operations:

(a) Crear las variables tiempo and alcohol e introducir estos datos.

(b) Compute el coeficiente de correlación lineal entre el alcohol and el tiempo e interpretarlo. ¿Es bueno el modelo lineal?

1. Select the menu Teaching > Regression > Linear regression.
2. In the dialog displayed, select the variable alcohol in the field Variable dependiente and la variable tiempo in the field Independent variable, and click the button Submit.

(c) Dibujar la regression line ofl alcohol sobre el tiempo. ¿Existe algún individuo con un residuo demasiado grande? Si es así, eliminar dicho individuo de la muestra and volver a calcular el coeficiente de correlación. ¿Ha mejorado el modelo?

1. Select the menu Teaching > Charts > Scatter plot.
2. In the dialog displayed, select the variable alcohol in the field Variable Y y la variable tiempo in the field Variable X.
3. En la solapa Fitted line, seleccionar Lineal and click the button Submit.

Se observa que hay un residuo atípico para el punto que corresponde al los 210 minutos. Para eliminarlo: En la ventana de edición del conjunto de datos hacer clic con el botón derecho del ratón sobre la fila correspondiente al dato con el residuo atípico and seleccionar Borrar esta fila.

(d) Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0.3 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley? ¿Es fiable esta predicción?

Para construir la recta de regresión:

1. Select the menu Teaching > Regression > Linear regression.
2. In the dialog displayed, select the variable tiempo in the field Variable dependiente and la variable alcohol in the field Independent variable.

3. Seleccionar Guardar el modelo, introducir un nombre para el modelo and click the button Submit.

Para hacer la predicción:

1. Select the menu Teaching » Regression » Predicciones.
2. In the dialog displayed seleccionar como modelo de regresión la recta calculada e introducir los valores para los que se desea la predicción in the field Predicciones para and click the button Submit.

4. El conjunto de datos edad.estatura del paquete rk.Teaching contine la edad and la estatura de 30 personas. Do the following operations:

   (a) Cargar datos del conjunto de datos edad.estatura desde el paquete rk.Teaching.

   (b) Compute la regression line of la estatura sobre la edad. ¿Es un buen modelo la recta de regresión?

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, select the variable estatura in the field Variable dependiente and la variable edad in the field Independent variable, and click the button Submit.

   (c) Dibujar el diagrama de dispersión de la estatura sobre la edad. ¿Alrededor de qué edad se observa un cambio en la tendencia?

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, select the variable estatura in the field Variable Y, la variable edad in the field Variable X, and click the button Submit.

   (d) Recodificar la variable edad en dos grupos para mayores and menores de 20 años.

1. Select the menu Teaching » Datos » Recodificar variable.
2. In the dialog displayed seleccionar in the field Variable a recodificar la variable edad.
3. En el campo Reglas de recodificación introducir

```
lo:20 = "menores"
20:hi = "mayores"
```

4. En el cuadro Guardar nueva variable click the button Cambiar.
5. In the dialog displayed seleccionar como objeto padre la el conjunto de datos edad_estatura and click the button Aceptar.
6. Introducir el nombre de la nueva variable grupo.edad and click the button Submit.

   (e) Compute la regression line of la estatura sobre la edad para cada grupo de edad. ¿En qué grupo explica mejor la recta de regresión la relación entre la estatura and la edad? Justificar la respuesta.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, select the variable estatura in the field Variable dependiente and la variable edad como Independent variable.
3. Seleccionar la opición Ajuste por grupos, introducir la variable grupo.edad in the field Grouping variable(s), and hacer clic en el Submit.

   (f) Dibujar las rectas de regresión anteriores.

1. Select the menu Teaching > Charts > Scatter plot.
2. In the dialog displayed, select the variable estatura in the field Variable Y y la variable edad in the field Variable X.
3. Seleccionar la opción Plot by groups e introducir la variable grupo.edad in the field Grouping variable(s).
4. En la solapa Fitted line, seleccionar Lineal and click the button Submit.

(g) ¿Qué estatura se espera que tenga una persona de 14 años? ¿Y una de 38?

Para predecir la estatura de la persona de 14 años:

1. Select the menu Teaching > Regression > Predicciones.
2. In the dialog displayed seleccionar como modelo de regresión la recta calculada para los menores e introducir 14 in the field Predicciones para and click the button Submit.

para predecir la estatura de la persona de 38 años, repetir lo mismo pero seleccionando la recta de regresión para los mayores e introducidento 38 in the field Predicciones para.

largo

5. La siguiente tabla recoge la información de las calificaciones obtenidas por un grupo de alumnos en dos asignaturas X e Y.

Alumno	1	2	3	4	5	6	7	8	9	10	11	12
X	NT	AP	SS	SS	AP	AP	SS	NT	SB	SS	AP	AP
Y	SB	SS	AP	SS	AP	NT	SS	NT	NT	AP	AP	NT

Do the following operations:

(a) Create a data set con las variables X e Y and enter the data.

(b) ¿Existe relación entre las calificaciones de X e Y? Justificar la respuesta.

1. Select the menu Teaching > Regression > Correlación.
2. In the dialog displayed select the variables X e Y in the field Variables.
3. En la solapa Opciones de correlación seleccionar el método de Ro de Spearman and hacer clic sobre el botón Submit.

2 Proposed exercises

1. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

Se desea calcular:

(a) La relación fundamental (recta de regresión) entre actividad restante and tiempo transcurrido.

(b) ¿En qué porcentaje disminuye la actividad cada año que pasa?

(c) ¿Cuándo tiempo debe pasar para que el fármaco tenga una actividad del 80%? ¿Cuándo será nula la actividad? ¿Son igualmente fiables estas predicciones?

2. Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg and otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, and 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días and 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días and 2 al cabo de 4 días. Do the following operations:

   (a) Compute la regression line ofl tiempo de curación con respecto a la dosis suministrada.

   (b) Compute el coeficiente de regresión del tiempo de curación con respecto a la dosis e interpretarlo.

   (c) Compute el coeficiente de correlación lineal e interpretarlo.

   (d) Determinar el tiempo esperado de curación para una dosis de 5 mg diarios. ¿Es fiable esta predicción?

   (e) ¿Qué dosis debe aplicarse si queremos que el paciente tarde 4 días en curarse? ¿Es fiable la predicción?

3. El fichero estaturas.pesos.alumnos del paquete rk.Teaching, contiene la estatura, el peso y el sexo de una muestra de alumnos universitarios. Do the following operations:

   (a) Cargar el conjunto de datos estaturas.pesos.alumnos desde el paquete rk.Teaching.

   (b) Compute la regression line ofl peso sobre la estatura and dibujarla.

   (c) Compute las rectas de regresión del peso sobre la estatura para cada sexo and dibujarlas.

   (d) Compute los coeficientes de determinación de ambas rectas. ¿Qué recta es mejor modelo? Justificar la respuesta.

   (e) ¿Qué peso tendrá un hombre que mida 170 cm? ¿Y una mujer de la misma estatura?

4. El conjunto de datos neonatos del paquete rk.Teaching, contiene información sobre una muestra de 320 recién nacidos en un hospital durante un año que cumplieron el tiempo normal de gestación. Do the following operations:

   (a) Construir la frequency table bidimensional del Agpar al minuto de nacer frente a si la madre ha fumado o no durante el embarazo. ¿Qué conclusiones se pueden sacar?

   (b) Construir la frequency table bidimensional del peso de los recién nacidos frente a la edad de la madre. ¿Qué conclusiones se pueden sacar?

   (c) Construir la regression line ofl peso de los recién nacidos sobre el número de cigarros fumados al día por las madres. ¿Existe una relación lineal fuerte entre el peso and el número de cigarros?

   (d) Dibujar la recta de regresión calculada en el apartado anterior. ¿Por qué la recta no se ajusta bien a la nube de puntos?

   (e) Compute and dibujar la regression line ofl peso de los recién nacidos sobre el número de cigarros fumados al día por las madres en el grupo de las madres que si fumaron durante el embarazo. ¿Es este modelo mejor o pero que la recta de los apartados anteriores?

   Según este modelo, ¿cuánto disminuirá el peso del recién nacido por cada cigarro más diario que fume la madre?

   (f) Según el modelo anterior, ¿qué peso tendrá un recién nacido de una madre que ha fumado 5 cigarros diarios durante el embarazo? ¿Y si la madre ha fumado 30 cigarros diarios durante el embarazo? ¿Son fiables estas predicciones?

   (g) ¿Existe la misma relación lineal entre el peso de los recién nacidos and el número de cigarros fumados al día por las madres que fumaron durante el embarazo en el grupo de las madres menores de 20 and en el grupo de las madres mayores de 20? ¿Qué se puede concluir?