Regression

Multiple Linear Regression

File: **bwght mod learn.sav**[1]

Sampling:
Data from an observational study at a hospital in 1989
All births of a year were recorded (n=1388)

Objective:
Clarify the influence of smoking on the birth weight of a child.

codebook:

Variable	Label	Measurement Level
faminc	1988 family income, $1000s	Scale
cigtax	cig. tax in home state, 1988	Scale
cigprice	cig. price in home state, 1988	Scale
bwght	birth weight, ounces	Scale
fatheduc	father's yrs of educ	Scale
motheduc	mother's yrs of educ	Scale
parity	birth order of child	Scale
male	=1 if male child	Nominal
white	=1 if white	Nominal
cigs	cigs smoked per day while preg	Scale
packs	packs smoked per day while preg	Scale

Value		Label
male	0	female child
	1	male child
white	0	non white
	1	white

1. Create new variables:

Variable	Label	Measurement Level
bwgr	birth weight, gramm	Scale
smoked	mother smoked while preg	Nominal
lowbwg	low birth weight (<=2500 gr)	Nominal

Value		Label
smoked	0	no
	1	yes
lowbwg	0	no
	1	yes

2. Explore whether smokers and non smokers are comparable in respect to other factors that might influence the birth weight
3. Explore the correlations between all factors
4. Perform an ANOVA
5. Perform an Multiple Regression (include collinearity tests)

[1] Modified example from: Wooldridge, J. (2003) Introductory econometrics: a modern approach. South Western College Publishing.
1 ounce = 28,349523 gr
Normal birthweight values (male av 3400gr; female av 3200 gr)
Low birthweight: 2500gr

Monotone Linear Nonlinear (simple)

Perform:

1. Scatter
2. Linear and monotone bivariate correlation
3. Curve Estimation
4. Regression Linear / Regression Nonlinear
5. Interactive Graph Scatter with individual and mean confidence

for

YX_1.sav

YX_2.sav

USAPop.sav

For USAPop.sav :
use demographic saturation model

$$\text{Population} = \frac{c}{1 + e^{a + b \bullet \text{Decade}}}$$

Restrict your regression knowledge to data from 1630 - 1960
Discuss extrapolation and real values 1970 - 2000

Linear Regression (multiple)

Perform:

Regression Linear
Method : Enter
Case Labels : Name
Statistics : all except Covariance, choose case wise diagnostics for outliers ≥ 2
Plot : Dependent vs. ZResiduals
Options : no changes

Same but Method: Stepwise

both for

YX1X2X3X4_1.sav

and

YX1X2X3X4_2.sav

Observational cross sectional sample survey (hypothetical data)

Sex and salaries[2]

Background: There is concern about sex equity in salaries (wage gap)[3]

Population: Software engineers from a large software company

Hypothesis: Mean annual salary for male and female differ

H0: There is no difference in mean annual salary between sexes.

Sampling Design: A stratified (38 male and 38 female) sample of total 76 employees was drawn randomly from 1220 engineers of promotion level A

Recorded: Actual annual salary (salary) , Sex (sex) and Job experience (exp) in years.

Exercise

Part 1

1. Open **Salary Sex JobExperience ND.sav**, label, label value
2. Start with **Boxplots** (salary vs. sex ; salary vs. job experience; job experience vs. sex)

---

3. Choose a simple one factorial model: **salary sex**
4. Formulate the model and the null hypothesis
5. Test the null hypothesis

---

6. Repeat 3-5 for two simple one factorial models: **salary exp** and **exp sex**
7. Choose an appropriate model to **adjust for exp**
8. Multiple linear regression
9. Formulate the model and the null hypothesis
10. Test the null hypotheses
11. What conclusions can be made concerning the population

Part 2

12. Repeat the analysis with data from **Salary Sex JobExperience CD.sav**

---

[2] Modified example from Raabe-Hesketh S (2008). Multilevel and Longitudinal Modeling Using Stata. Stata Press, pp. 20-25.
[3] Background: e.g. Brown E et al. (2007). Sex and salaries at the University of Manitoba. http://www.cerforum.org/conferences/200705/papers/BrownPrenticeTroutt.pdf.