PubH 7405:

REGRESSION ANALYSIS

SLR: PARAMETER ESTIMATION

REGRESSION MODEL

- Model: $\mathbf{Y = \beta_0 + \beta_1 x + \varepsilon}$ where β_0 and β_1 are two new parameters called regression coefficients, the Intercept and the Slope, respectively. The last term, ε, is the "error" representing the random fluctuation of y-values around their mean, $\beta_0 + \beta_1 x$, when X=x.
- **The presence of the error term is an important** characteristic of a statistical relationship; **the points on a scatter diagram do not fall perfectly on the line.**
- The scatter diagram is an useful **diagnostic tool** for **checking out the Model (e.g. to see if it is linear)**.

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

The "normal" assumption can sometimes be weakened to $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$

The Normal Error Regression Model

REGRESSION COEFFICIENTS

- **The error term ε - with variance σ^2 -would tell how spread the dots are around the regression line.**
- The regression coefficients, β_0 and β_1, determine the position of the line and are important quantities in the analysis process. In "correlation analysis", we need to know only the coefficient of correlation r which is proportional to the slop β_1 (we'll see); but in a "regression analysis", with **new emphasis on prediction** , so we need them both, β_0 and β_1.
- **As parameters, both β_0 and β_1 are unknown; but they can be "estimated" by statistics from data**

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \in \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$$

The Variance σ^2 (around the regression line) is the third parameter : It is hidden, but has a specific role & very important too!

THE INTERCEPT

- If the scope of the model include X = 0, **β_0 gives the Mean of Y when X = 0**; otherwise, it does not have any particular meaning as a separate term.
- If the scope of the model does not include X = 0, we may choose a "transformation" such as:
  (New) $x = x - \bar{x}$
  Under this transformation, β_0 gives the Mean of Y when $X = \bar{x}$, i.e. a "**<u>typical</u>**" subject (value = $\bar{x}$)

Original Model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

$$E(Y \mid X = 0) = \beta_0$$

Transformed Model:

$$X^* = X - \bar{X}$$

$$E(Y \mid X^* = x^*) = \beta_0^* + \beta_1^* x^*$$

$$E(Y \mid X^* = 0) = \beta_0^*$$

$$E(Y \mid X^* = 0) = E(Y \mid X = \bar{x})$$

THE SLOPE

- The Slope is a more important parameter:
- (i) **If X is binary (=0/1) representing an exposure, β_1 represents the increase in the mean of Y associated with the exposure (or a decrease if β_1 is negative);**
- (ii) **If X is on a continuous scale, β_1 represents the increase in the mean of Y associated with one unit increase in the value of X, X=x+1 vs. X=x, (or a decrease if β_1 is negative).**

Binary Independent Variable X:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

$$E(Y \mid X = 0) = \beta_0$$

$$E(Y \mid X = 1) = \beta_0 + \beta_1$$

$$\mathbf{E(Y \mid X = 1) - E(Y \mid X = 0) = \beta_1}$$

The <u>change</u> in the mean of Y associated with the <u>exposure</u>

Continuous Independent Variable X:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

$$E(Y \mid X = x + 1) = \beta_0 + \beta_1 (x + 1)$$

$$E(Y \mid X = x + 1) = \beta_0 + \beta_1 x + \beta_1$$

$$\mathbf{E(Y \mid X = x + 1) - E(Y \mid X = x) = \beta_1}$$

The <u>change</u> in the mean of Y associated with one unit <u>increase</u> in the value of X

EXAMPLE

- For example, let X be a mother's weight gain during her pregnancy and Y the birth weight of the newborn. When X=x, the birth weights (BW) of all infants form certain normal distribution.
- The Mean of that Normal Distribution depends on the weight gain:
  Mean (of BW) = Intercept + (Slope)(x)
- **The "slope" represents the "average increase in birth weight for <u>every pound</u> the mother gained"; In this case, slope>0**

The Regression Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

THE NEED

Use the "observed data": $\{(x_i, y_i)\}_{i=1}^{n}$ to estimate the "model" three parameters: β_0, β_1, and σ^2

SUM OF SQUARED ERRORS

- By the Model, when X=x, the Mean of Y is $\beta_0 + \beta_1 x$.
- Let b_0 and b_1 are estimates of β_0 and β_1, respectively; an estimate of $(\beta_0 + \beta_1 x)$ is y – considered as a sample (of size 1). The **error** of that estimate is $[y - (\beta_0 + \beta_1 x)]$ so that $Q = \Sigma\, [y - (\beta_0 + \beta_1 x)]^2$ represents a form of the "total errors" (not distinguishing an under-estimation from an over-estimation); called "**the sum of squared errors**"
- The **method of least squares** requires that we find "good estimates" of β_0 and β_1 the **values of b_0 and b_1 so as to minimize this "sum of squared deviations".**

METHOD OF LEAST SQUARES

<u>PROCESS</u>: We take the two "partial derivatives" of Q with respect to β_0 and β_1, set each equal to zero, and solve a system of two equations for two unknowns β_0 and β_1; the solutions are b_0 and b_1.

$$\text{Data}: \{(x_i, y_i)\}_{i=1}^{n}$$

$$\mathbf{Q} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \boldsymbol{\beta}_0 - \boldsymbol{\beta}_1 \mathbf{x_i})^2$$

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\delta Q}{\delta \beta_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\delta Q}{\delta \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i) = 0$$

called "**Normal Equations**" (page 17):

$$\sum \mathbf{y_i} = \mathbf{nb_0} + \mathbf{b_1}\sum \mathbf{x_i}$$

$$\sum \mathbf{x_i y_i} = \mathbf{b_0}\sum \mathbf{x_i} + \mathbf{b_1}\sum \mathbf{x_i^2}$$

Results: Point estimators/estimates

RESULTS

- The "Least Squares Estimates" are:

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

- Given the estimates "b_0" of the Intercept and "b_1" of the Slope, Estimate of y (for the mean or a "new" value x of X) is $\mathbf{\hat{Y} = b_0 + b_1 x}$; this is called "**fitted value**"

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$b_1 = r\frac{s_y}{s_x}$$

From this simple result, we can see that "b_1" and "r" are of the same sign – and both are equal to zero at the same time; **they measure the same thing but on different "scale".**

Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Method:

$$Minimize\ \mathrm{Q} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Note:

We **do not need** the "normal" assumption

to obtain the estimates b_0 and b_1

However, later, **we do need it for inferences**

concerning the parameters $\boldsymbol{\beta_0, \beta_1, \sigma^2}$

EXAMPLE #0

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

	x	y	x^2	y^2	xy
	1	3	1	9	3
	2	5	4	25	10
	6	7	36	49	42
Totals	9	15	41	83	55

the estimates of the Slope and the Intercept are:

$$b_1 = \frac{55 - \frac{(9)(15)}{3}}{41 - \frac{(9)^2}{3}} = .714$$

$$b_0 = \frac{15}{3} - (.714)(\frac{9}{3}) = 2.858$$

For example, for new subject with X=5, it is predicted that its y-value would be:
2.858 + (.714)(5) = 6.428

EXAMPLE #1:

Birth weight data:

x (oz)	y (%)
112	63
111	66
107	72
119	52
92	75
80	118
81	120
84	114
118	42
106	72
103	90
94	91

$$b_1 = \frac{94{,}322 - \frac{(1{,}207)(975)}{12}}{123{,}561 - \frac{(1{,}207)^2}{12}} = -1.74$$

$$b_0 = \frac{975}{12} - (-1.74)\left(\frac{1{,}207}{12}\right) = 256.3$$

Note: if the birth weight is 95 ounces, it is predicted that the increase between days 70 & 100 would be 256.3 + (-1.74)(95) = 90.1%

EXAMPLE #2: Age and SBP

Age (x)	SBP (y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

$$b_1 = \frac{146{,}260 - \frac{(984)(2{,}193)}{15}}{67{,}954 - \frac{(984)^2}{15}} = .71$$

$$b_0 = \frac{2{,}193}{15} - (.71)(\frac{984}{15}) = 99.6$$

Note: for a 60-year-old woman, it is predicted that her systolic blood pressure would be 99.6 + (.71)(60) = 142.2 mmHg.

EXAMPLE #3: Toluca Company Data

(Description on page 19 of Text)

LotSize	WorkHours
80	399
30	121
50	221
90	376
70	361
60	224
120	546
80	352
100	353
50	157
40	160
70	252
90	389
20	113
110	435
100	420
30	212
50	268
90	377
110	421
30	273
90	468
40	244
80	342
70	323

$b_0 = 62.37$

$b_1 = 3.57$

Suppose we are interested in the mean number of work hours required when the lot size is X = 65; our point estimate is:

62.37 + (3.57)(65) = 294.4 hours

(**See textbook, page 21**)

SCOPE OF THE MODEL

In formulating a regression model, we need to restrict the “coverage” of the model to some interval of values of the independent variable X; this is determined either by the design or the availability of data at hand. **The shape of the regression function outside this range would be in doubt** because the investigation provided no evidence as to the nature of the statistical relation outside this range. In short, **one should not do any extrapolation**.

(Observed) **SUM OF SQUARED ERRORS**

- **$Q = \Sigma [y - (\beta_0 + \beta_1 x)]^2$** is "**the sum of squared errors**"
- **Since $(b_0 + b_1 x)$ is an estimate of the mean of Y, "$e = [y - (b_0 + b_1 x)]$" represents the "error" of our prediction; so that $SSE = \Sigma e^2 = \Sigma [y - (b_0 + b_1 x)]^2$ is the (observed) "sum of squared errors" very much like the numerator of the sample variance s^2.**

ESTIMATING THE VARIANCE

- In the Regression Model, the error term ε is assumed to have a Normal Distribution with mean 0 and variance σ^2.
- ε is like a “variable” of which we have a sample with sample mean zero: $\{e_i\}$; $i = 1,\ldots,n$
- Variance σ^2 is estimated by MSE=SSE/(n-2); **two degrees of freedom were lost due to the need to estimate the intercept and slope**.

$$e_i = y_i - \hat{y}_i$$

$$= y_i - (b_0 + b_1 x_i)$$

$$MSE = \frac{\sum e_i^2}{(n-2)} = \hat{\sigma}^2$$

CHARACTERISTICS OF PREDICTION ERRORS

$$\frac{\delta Q}{\delta \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \Rightarrow \sum \mathbf{e_i} = \mathbf{0}$$

$$\frac{\delta Q}{\delta \beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \Rightarrow \sum \mathbf{x_i e_i} = \mathbf{0}$$

INTERPRETATION

$$\sum e_i = 0$$

$$\sum x_i e_i = 0$$

(1) Average error is zero,

(2) Error & Predictor are not correlated

(3) As a result of (2), error and fitted value are not correlated

$$\sum e_i = 0$$

$$\sum x_i e_i = 0$$

$$r = \frac{\sum xe - \frac{(\sum x)(\sum e)}{n}}{\sqrt{[\sum x^2 - \frac{(\sum x)^2}{n}][\sum e^2 - \frac{(\sum e)^2}{n}]}} = 0$$

Implication: Dots on scatter diagram form a band with **constant width** around the regression line

UNBIASED ESTIMATES

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

$$E(MSE) = \sigma^2$$

They are correct on the average; we'll prove at least the first two - later

MORE ON THE SLOPE

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$= \frac{\sum (x-\bar{x})y - \bar{\mathbf{y}} \sum (\mathbf{x}-\bar{\mathbf{x}})}{\sum (x-\bar{x})^2}$$

$$= \frac{\sum (x-\bar{x})y}{\sum (x-\bar{x})^2}$$

Data points with x-values at both ends are influential

$$\mathrm{b}_1 = \frac{\sum (x - \bar{x}) y}{\sum (x - \bar{x})^2}$$

$$Var(b_1) = \frac{\sum (x - \bar{x})^2 Var(y)}{\{\sum (x - \bar{x})^2\}^2}$$

$$= \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$Var(b_1) = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$\hat{=} \frac{MSE}{\sum (x - \bar{x})^2}$$

$$SE(b_1) = \sqrt{\frac{MSE}{\sum (x - \bar{x})^2}}$$

$$= \left(\sqrt{\frac{MSE}{n-1}} \right) \left(\frac{1}{s_x} \right)$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$Var(b_0) = Var(\bar{y}) + (\bar{x})^2 Var(b_1)$$

$$= \frac{\sigma^2}{n} + (\bar{x})^2 \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

$$= \sigma^2 \{ \frac{1}{n} + \frac{(\bar{x})^2}{\sum (x - \bar{x})^2} \}$$

$$SE(b_0) = \left(\sqrt{\frac{MSE}{n-1}}\right)\left(\sqrt{1+\frac{\bar{x}^2}{s_x^2}}\right)$$

DESIGN IMPLICATION

$$\sigma^2(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\sigma^2(\mathrm{b}_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{(\bar{x})^2}{\sum (x - \bar{x})^2} \right\}$$

These variances, for given n and σ^2, are affected by the spacing of the X's levels in the data. **The larger the sum of squares of X, the more precise the estimates of the Slope and the Intercept.**

FITTED VALUE & RESIDUAL

From the model :

$$y_i = \beta_0 + \beta_1 x_i$$

Fitted value :

$$\hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{b_0} + \mathbf{b_1 x_i}$$

Residual :

$$e_i = y_i - \hat{y}_i$$

Across the sample, we have:

$$Var(\hat{Y}) = b_1^2 s_x^2$$

$$= \left(r^2 \frac{s_y^2}{s_x^2} \right) s_x^2$$

$$= r^2 s_y^2$$

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

$$Var(e) = Var(Y) - Var(\hat{Y})$$

$$= s_y^2 - r^2 s_y^2$$

$$= (1 - r^2) s_y^2 \geq 0$$

Result:

$$\mathbf{r^2 \leq 1}$$

$$\mathbf{-1 \leq r \leq 1}$$

Recall that, across the sample, we have:

$$
\begin{aligned}
Var(\hat{Y}) &= b_1^2 s_x^2 \\
&= \left(r^2 \frac{s_y^2}{s_x^2} \right) s_x^2 \\
&= r^2 s_y^2 \\
&= r^2 Var(Y)
\end{aligned}
$$

$Var(\hat{Y})$ is the explained variance; so r^2 is the fraction or **proportion of the total variance that is "explained"** by the regression. We call it "**Coefficient of Determination**".

Besides “Least Squares”, parameters can be estimated using the method of “Maximum Likelihood”; results are called “MLE” – maximum likelihood estimators/estimates.

MAXIMUM LIKELIHOOD ESTIMATION

Suppose that we can assume a (parametric) Model for the Dependent Variable Y which is characterize by a Density Function f(t; θ) – say, normal distribution - involving a parameter or parameters θ (which is fixed but unknown). Given a random sample {yi}i=1n; the **Likelihood Function** for θ is given by: **L = Πf(yi; θ)**, and data can be analyzed by standard methods associated with large-sample Maximum Likelihood Theory (Maximum Likelihood Estimator- MLE- and its asymptotic normality, Score statistic, Likelihood Ratio statistic)

Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Density Function for Y:

$$\mathrm{f(y)} = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2 \right\}$$

Density Function for Y :

$$f(y) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y - \beta_0 - \beta_1 x)^2 \right\}$$

Likelihood Function :

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

RESULTING MLEs

- **The maximum likelihood estimates of the Intercept & Slope are identical to the Least Squares estimates.**
- The **Variance Estimate is slightly difference**; since the MLE variance estimator is biased, we prefer and use the Least Squares estimator (MSE).
- The MLE variance estimator is:

$$\frac{\sum (y - \hat{y})^2}{n} = \frac{n-2}{n} MSE$$

INTERCEPT & SLOPE

- Since the MLEs of Intercept and Slope are the same as the least squares estimates, they have the properties of least squares estimates: (1) they are **unbiased**, and (2) they are "**minimum variance unbiased**" (that is, **they have minimum variance in the class of all unbiased estimators**).
- In addition, as MLEs for the normal error regression model: (3) they are consistent, and (4) they are sufficient.

Readings & Exercises

- **Readings: A thorough reading of the text's "Chapter 1" is highly recommended.**
- **Exercises: The following exercises are good for practice, all from chapter 1 of text: 1.19, 1.20, 1.21, 1.22, 1.27, 1.32, and 1.35.**

Due As Homework

#5.1 We have a data set on 86 smokers (File: Cigarettes); three outcome or response variables are Carbon monoxide, Cotinine (a derivative of Nicotine), and NNAL (a derivative of NNN, a toxin only comes from tobacco products). Data for 3 other explanatory variables are also included: Age, Gender (1=female), and Cigarettes per Day (CPD). Let Y= log(NNAL) & X=CPD:

a) Obtain Least Squares estimates of β_0 and β_1, then state/express the estimated regression function (i.e. the mean of the dependent variable, the fitted value).

b) Plot the estimated regression function on the same plot with your scatter diagram; does the linear relationship appear to fit the data? Does plot support the anticipation that the average urine log(NNAL) increases with increasing CPD? Is the linear relationship strong?

c) Give an estimate of mean NNAL when CPA = 30.

d) What is the point estimate of the change in the mean log(NNAL) when CPD increases by 1 cig? by 10 cigs?

e) Does any data point appear to be out of its place?

#5.2 It has been generally known that respiratory function may decline with age. To study this possibility, We consider a data set consisting of age (years) and vital capacity (VC, liters) for each of 44 men working in the cadmium industry but have not been exposed to cadmium fumes (File: Vital Capacity). Let X = Age and Y = (100)(Vital Capacity):

a) Obtain Least Squares estimates of β_0 and β_1, then state/express the estimated mean of the dependent variable.

b) Plot the estimated regression function on the same plot with your scatter diagram; does the linear relationship appear to fit the data? Does plot support the anticipation that the average vital capacity decreases with increasing Age? Is the linear relationship strong?

c) Give an estimate of mean E(Y) when Age = 35 years.

d) What is the point estimate of the change in the mean E(Y) when Age increases by 1 year? by 10 years?

e) What would be the values of the Intercept, Slope, and MSE if Vital Capacity is use as the dependent variable (instead of Y; not run the computer program with the new response variable)