

# PubH 7405: REGRESSION ANALYSIS



**SLR: GOODNESS-OF-FIT & REMEDIES**

# ONGOING QUESTION

**Does the Regression Model fit the data?**

**Then what if the Regression Model, or certain part of the Regression Model, does not fit the data ? i.e. (1) If it does not fit, could we do something to make it fit? And (2) Does it matter if it still does not fit?**

In doing statistical analyses, a “**statistical model**” – such as the “normal error regression model”- is **absolutely necessary**.

However, a “model” is just an assumption or a set of assumptions; **they may or may not fit the observed data**. Certain part or parts of a model may be violated and, as a consequence, the **results may not be valid**.

# POSSIBLE DEPARTURES FROM THE NORMAL REGRESSION MODEL

- The regression function is **not linear**
- Variance (of error terms) is **not constant**
- Model fits all but a few “**outliers**”
- Responses (error terms) are **not independent**
- Responses terms are **not normally distributed**
- An important predictor (independent variable)  
– including **time** - has been **omitted**.

Diagnostics could be informal using plots/graphs or could be based on **formal application of statistical tests**; graphical method is more popular and, most of the times, **would be sufficient**.

In general, I'm not an enthusiastic supporter of “tests of goodness-of-fit”; We need to “**accept**” a model but statistical tests only allow us to reject or not to reject the Null Hypothesis under investigation, the model. **We can only tell when a model does not fit the data – if we have enough information; we cannot formally tell when a model fits the data.**

# TESTS FOR NORMALITY

- Goodness-of-fit tests – such as the Kolmogorov-Smirnov test – can be used for examining the normality of the error terms; but they are a bit advanced for first-year students.
- A more simple – but also formal – test for normality can be conducted by calculating the coefficient of correlation between the residuals and their expected values under normality. **High** value of the coefficient of correlation is indicative of normality. This is a supplement to Q-Q plot.
- “Critical value” for various sample sizes are in Appendix Table B6 (page 673).

When the distribution (of the response) is only near normal, **most of the dots (on the Q-Q plot” are already very close to a straight line; the “cut-point” for rejection is quite high.** Again, as mentioned, a formal statistical test may not really be needed here; but could use to supplement the Q-Q plot – **more valuable when sample size  $n$  is small.**

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

## EXAMPLE #3: Toluca Company Data (Description on page 19 of Text)

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	62.3658586	26.17743389	2.382428	0.025851	8.21371106	116.518006
X Variable 1	3.57020202	0.346972157	10.28959	4.45E-10	2.85243543	4.28796861

→ Residuals

### Results:

Correlation  $r = .991$ ,  $n = 25$

Critical value = **.959**

(from Table B6, p.673);

**Rejection when  $r$  is small!**

**No departure from normality**

Expected  $k^{\text{th}}$  residual:

$$\sqrt{MSE} \left[ z \left( \frac{k - .375}{n + .25} \right) \right]$$

# DEPARTURE FROM NORMALITY

If the probability distributions of  $Y$  are not exactly normal but do not depart seriously, the sampling distributions of  $b_0$  and  $b_1$  would still be approximately normal.

**Even if the probability distributions of  $Y$  are far from normal, the effects are still minimal provided that the samples sizes are sufficiently large; i.e. the sampling distributions of  $b_0$  and  $b_1$  are asymptotically normal.**

However, lack of normality and non-constant variances are often go together, so we still to learn how to deal with it.

# TESTS FOR CONSTANT VARIANCE

- There are many but two are often mentioned
- The Breusch-Pagan test assumes normality of error terms but the test follows the usual regression methodology – not hard to do.
- The Brown-Forsythe test does not depend on normality of error terms; this is desirable because non-constant variance and non-normality tend to go together. This test is easy.

# BROWN-FORSYTHE TEST

- The Brown-Forsythe test is used to ascertain whether the error terms have constant variance; especially when the **variance of the error terms either increases or decreases with the independent variable X.**
- The Test: divide the data into 2 groups, say **half with larger values of X** and **half with smaller values of X**;  
(1) calculating the “absolute deviations” of the residuals around their group mean (or median);  
(2) applying the two-sample t-test.
- Test statistic follows approximately the t-distribution when the variance of the error terms is constant (under the Null Hypothesis) and the sizes of the two group are not extremely small.

# BROWN-FORSYTHE: RATIONALE

- If the error variance is either increasing or decreasing with  $X$ , the residuals in one group tend to be more variable than those residuals in the other.
- The Brown-Forsythe test does not assume normality of error terms; this is desirable because non-constant variance and non-normality tend to go together.
- It's is very similar to “**Levine's test**” to compare any two variances – instead of forming the ratio of two sample variances (& use “F-test”).

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

## EXAMPLE #3: Toluca Company Data (Description on page 19 of Text)

Group 1: n = 13 with lot sizes from 20 to 70; median residual = -19.88

Group 2: n = 12 with lot sizes from 80 to 120; median residual = -2.68

Mean of absolute residuals :

Group 1: 44.815

Group 2: 28.450

Pooled Variance: 964.21;  $s_p = 31.05$

$$t = \frac{44.815 - 28.450}{31.05 \sqrt{\frac{1}{13} + \frac{1}{12}}}$$

$$= 1.32$$

two – sided p – value = .20

This example shows that **the half with smaller X's has larger residuals** – and vice versa; the pattern of an inverse mega phone – but it's “not significant”, a case that makes me uneasy with statistical tests: I want to assume that the variance is constant, **it only says that we do not have enough data to conclude that the variance is not constant!**

# BREUSCH-PAGAN TEST

- This is another test to see if the error variance is either increasing or decreasing with  $X$
- Alternative: The Breusch-Pagan test assumes that the error terms are independent and normally distributed but the “log of variance” is linearly related to  $X$  level (Alternative  $H_A$ ).
- Two “runs” of regression are needed: (1)  $Y$  against  $X$ , and (2) Squared residuals against  $X$ .

**Model :  $\ln \sigma_i^2 = \gamma_0 + \gamma_1 x_i$**

**Null Hypothesis  $H_0 : \gamma_1 = 0$**

**Test Statistic :  $X_{BP}^2 = \frac{SSR^*}{2} \div \left( \frac{SSE}{n} \right)^2$**

**It is distributed as Chi – square, 1df, under  $H_0$**

**$SSR^*$  is the SSR for regressing  $e^2$  against X,**

**SSE is the SSE for regressing Y against X**

**Logic: If  $H_A$  is true,  $SSR^*$  is large –  
as compared to (a function of) SSE**

Large values of the Test Statistic lead to the conclusion that the error variance is not constant (note: a very specific Alternative).

Reference:

Breusch T.S. & Pagan A.R.;

Econometrica 47: 1287-1294, 1979

Also referred to as the Cook-Weisberg test

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

## EXAMPLE #3: Toluca Company Data (Description on page 19 of Text)

After 2 runs of regressions :

$$SSR^* = 7,896,128$$

$$SSE = 54,825$$

$$X_{BP}^2 = \frac{7,896,128}{2} \div \left( \frac{54,825}{25} \right)^2$$

$$= .821$$

$$p - \text{value} = .64$$

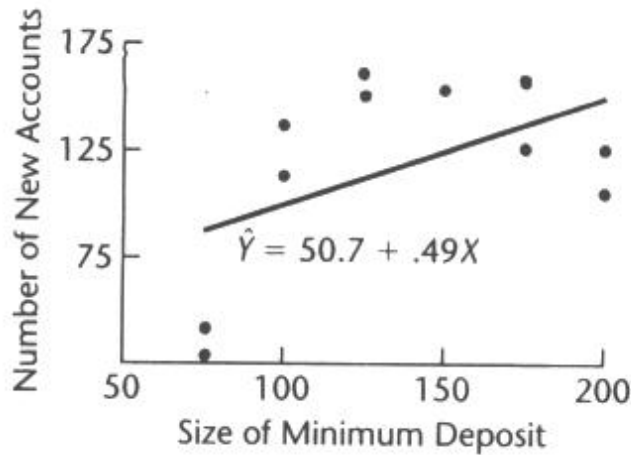
# TEST FOR LINEARITY

- This is a formal test to determine whether a specific type of regression function adequately fits the data ; in our case the linear function.
- It's an F test for goodness-of-fit following the usual regression methodology.
- Assumptions for observations are: (1) independent, (2) normally distributed, and (3) the distributions (of Y's) have the same variance.
- In other words, **we assume all aspects of the linear regression model except for the linearity part.**
- **It does require repeat trials (replications)** for some levels of the predictor variable X; kind of “rare”.

# EXAMPLE #5: GIFT FOR DEPOSIT

- Banks often offer gifts for setting up accounts; the value of the gift was directly proportional to the required **minimum deposit (X)**
- The aim is to attract more account; **Y is the number of new accounts.**
- In this “experiment”, six levels of X were used but, for some reason, there was one missing piece of data so that  $n = 11$ .
- **Result:  $Y = 50.7225 + .4867 * X$**
- Does the “**linear**” regression model fits?

**FIGURE 3.11**  
Scatter Plot  
and Fitted  
Regression  
Line—Bank  
Example.



**TABLE 3.5**  
Data Arranged  
by Replicate  
Number and  
Minimum  
Deposit—Bank  
Example.

Replicate	Size of Minimum Deposit (dollars)					
	$j = 1$ $X_1 = 75$	$j = 2$ $X_2 = 100$	$j = 3$ $X_3 = 125$	$j = 4$ $X_4 = 150$	$j = 5$ $X_5 = 175$	$j = 6$ $X_6 = 200$
$i = 1$	28	112	160	152	156	124
$i = 2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

It's rather obvious that the relationship is not linear

The question is how to prove it.

# NOTATIONS

$X_j$  : Level "j" of X; there are "c" levels

$n_j$  : Number of replicates at jth level of X

$Y_{ij}$  : The ith observation at the jth level

$$n = \sum n_j$$

$$\bar{Y}_j = \frac{\sum_i Y_{ij}}{n_j}$$

# THE FULL MODEL

When we make no assumption on the linear relationship; total “unexplained” variation of  $Y$  is measured by  $SSE(F)$  (F: for “full”):

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2$$

$$\begin{aligned} df(F) &= \sum_j (n_j - 1) \\ &= n - c \end{aligned}$$

$$SST = \sum_j \sum_i (Y_{ij} - \bar{Y})^2$$

$$df(T) = n - 1$$

“Different levels of  $X$ ”  
are considered in  
 $SSE(F)$  but not in  $SST$

# THE REDUCED MODEL

This is when we impose the linear relationship between  $X$  and  $Y$ ; the total “unexplained” variation (of  $Y$ ) is measured by (R: for “reduced”):

$$SSE(R) = \sum_j \sum_i [Y_{ij} - (b_0 + b_1 x_j)]^2$$

$$df(R) = n - 2$$

SSE(R) is different from SSE(F) because a “linear function of  $X$ ” is considered

Usual Regression



# EVIDENCE OF “LACK OF FIT”

This is the “difference” between “making no assumption” and “imposing a linear relationship between X and Y”; the total variation (of Y) that is accountable for by the lack of fit is measured by (LF: “lack of fit”):

$$\text{SSE(LF)} = \text{SSE(R)} - \text{SSE(F)}$$

$$\begin{aligned} \text{df(LF)} &= \text{df(R)} - \text{df(F)} \\ &= (n - 2) - (n - c) \\ &= c - 2 \end{aligned}$$

SSE(F): only levels of X

SSE(R): levels & linear

**SSE(R)-SSE(F): for “Linear”**

# TEST STATISTIC

$$\begin{aligned} F^* &= \frac{\text{SSE(LF)}}{c-2} \div \frac{\text{SSE(F)}}{n-c} \\ &= \frac{\text{MSE(LF)}}{\text{MSE(F)}} \end{aligned}$$

Under the Null Hypothesis that a linear relationship fits,  $F^*$  is distributed as F distribution with  $df = (c-2, n-c)$ . The error under the Full Model is sometimes referred to as “**pure error**”,  $\text{MSE(F)} = \text{MSPE}$ .

$$E\{\text{MSE(F)}\} = \sigma^2$$

$$E\{\text{MSE(LF)}\} = \sigma^2 + \frac{\sum n_j [\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2}$$

# ANOVA FORMAT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F Statistic	p-value
Regression	SSR	1	MSR		
Error	SSE(R)	n-2	MSE(R)		
Lack of Fit	SSE(LF)	c-2	MSE(LF)	$F^* = \text{MSE(LF)} / \text{MSE(F)}$	p-value
Pure Error	SSE(F)	n-c	MSE(F)		
<b>Total</b>	<b>SST</b>	<b>n-1</b>			

**Approach:** to split the Error/Residual into “Pure Error” and “Lack of Fit/Linearity”

# EXAMPLE #5: GIFT FOR DEPOSIT

Source of Variation	Sum of Squares	Degrees of Freedom	Mean of Squares	F Statistic	p-value
Regression	5,141.3	1	5141.3		
Error	14,741.6	9	1638.0		
Lack of Fit	13,593.6	4	3398.4	$F^* = 3,398.4 / 229.6 = 14.8$	0.006
Pure Error	1,148.0	5	229.6		
<b>Total</b>	<b>19,882.9</b>	<b>10</b>			

# REMEDIAL MEASURES

- If a SLR model is found not appropriate for the data at hand, there are **two basic choices**:
  - (1) **Abandon** it and search for a suitable one, or
  - (2) Use **some transformation** on the data to create a fit for the transformed data
- **Each has advantages & disadvantages**: first approach may yield better insights but may lead to more technical difficulties; transformations are more simple but may obscure the fundamental real relationship; sometimes **it's hard to explain.**

# LOG TRANSFORMATIONS

- Typical:  $Y^* = \text{Log}(Y)$ , turns a multiplicative model into an additive model – **for linearity**.
- Residuals should be used to check if model fits transformed data: normality, independence, and constant variance because the distribution changes the distribution and the variance of the error terms.
- **Others:** (1)  $X^* = \text{Log}(X)$ ,  
(2)  $X^* = \text{Log}(X)$  and  $Y^* = \text{Log}(Y)$ ;  
**Example:** Model (2) is used to study “demand” (Y) versus “price of commodity” (X) in economics.

When the distribution of the error terms is close to normal with an approximately constant variance, and a transformation is needed only for linearizing a non-linear regression relation, only transformations on  $X$  should be attempted.

# RECIPROCAL TRANSFORMATIONS

- Also aimed for linearity
- Possibilities are:
  - (1)  $X^* = 1/X$ ,
  - (2)  $Y^* = 1/Y$ ,
  - (3)  $X^* = 1/X$  and  $Y^* = 1/Y$
- **Example**: Models (1) and (2) are useful when it seems that  $Y$  has a lower or upper “asymptote” (e.g. hourly earning)

Logarithmic and Reciprocal Transformations can be employed together to linearize a regression function. For example, the “Logistic Regression Model” (with  $Y = \text{probability/proportion “p”}$ ):

$$Y = \ln\left(\frac{p}{1-p}\right)$$
$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$$

# NON-CONSTANT VARIANCE & NON-NORMALITY

- Non-normality alone needs not be fixed because method/results are robust but **non-normality and non-constant variance often go together**
- Constant variance = Homoscedasticity
- Non-constant variance = **Heteroscedasticity**
- Most often: Variance is functionally related to the mean; e.g. (1) Y is Poisson distributed, or (2) Standard Deviation is proportional to X

**If  $Y$ , certain count, is distributed as Poisson, try  $Y = \sqrt{Y}$ ; it both stabilizes the variance and improves normality.**

When Standard Deviation is proportional to  $X$ ,  
try:  $X^* = 1/X$  and  $Y^* = Y/X$ ; it would lead to a  
simple linear model with a constant variance

Model #1 :

$$Y = \beta_0 + \beta_1 X + \varepsilon;$$

$$\sigma^2 = kX^2$$

**Model #1 :**

$$Y = \beta_0 + \beta_1 X + \varepsilon; \sigma^2 = kX^2$$

$$\frac{Y}{X} = \frac{\beta_0}{X} + \beta_1 + \frac{\varepsilon}{X}$$

**Model #2 :**

$$Y^* = \beta_1 + \beta_0 X^* + \varepsilon^*$$

$$\begin{aligned} \text{Var}(\varepsilon^*) &= \frac{1}{X^2} \text{Var}(\varepsilon) \\ &= k \end{aligned}$$

Same coefficients, in swapped roles, but they should be estimated using transformed data in order to obtain “minimum variance unbiased estimates”. The alternative is performing “**weighted**” least-squares estimation instead of “ordinary” (i.e. un-weighted) least-squares.

With ordinary least squares, estimators for regression coefficients are obtained by minimizing the quantity  $Q$ ; setting the partial derivatives equal to zero to have the “normal equations”:

$$Q = \sum (Y - \beta_0 - \beta_1 X)^2$$

With **weighted least squares**, estimators for regression coefficients are obtained by minimizing the quantity  $Q$  where “ $w$ ” is a “**weight**” (associated with the error term); setting the partial derivatives equal to zero to have the “normal equations”:

$$Q = \sum w(Y - \beta_0 - \beta_1 X)^2$$

The optimal choice for the weight is the inverse of the variance; when the variance is constant, ordinary and weighted least squares estimators are identical. For example, **when standard deviation is proportional to X** (variance is  $kX^2$ ), we minimize:

$$Q = \sum \frac{1}{kX^2} (Y - \beta_0 - \beta_1 X)^2$$

You can verify that, when Standard Deviation is proportional to  $X$  and for the transformation ( $X^* = 1/X$  and  $Y^* = Y/X$ ), the ordinary least squares estimators using transformed data and the weighted least squares using original data are identical.

# Readings & Exercises

- Readings: A thorough reading of the text's sections 3.5-3.7 (pp. 115-127) and 3.9 (pp. 129-137) is highly recommended.
- Exercises: The following exercises are good for practice, all from chapter of text: 3.3(d-e), 3.7(d-e), 3.8(d-e), 3.9, 3.11(b), and 3.18.

# Due As Homework

**10.1** Refer to dataset “Cigarettes”, let  $X = \text{CPD}$  &  $Y = \log(\text{NNAL})$ :

a) Prepare a normal plot of the residuals against their expected values under normality. Test for the normality assumption using  $\alpha = .05$

b) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of  $X$ ; Does your conclusion support the graphical findings in Exercise 9.1?

c) Does the result of the Bruesch-Pagan test agree with that of the Brown-Forsythe test? Which test requires stronger assumption?

d) Prepare the scatter plots:  $\text{NNAL}$  vs.  $\text{CPD}$ ,  $\log(\text{NNAL})$  vs.  $\text{CPD}$ ,  $\log(\text{NNAL})$  vs.  $\text{SQRT}(\text{CPD})$ ; which appears to be the best fit for a linear regression function?

**10.2** Answer the 4 questions of Exercise 10.1 using dataset “Infants” with  $X = \text{Gestational Weeks}$  and  $Y = \text{Birth Weight}$ .

**Only #10.2 is required**