

## NOTES ON HETEROSCEDASTICITY

Recall that the second Gauss-Markov Assumption requires that the variances of all regression residuals  $(\varepsilon_1, \dots, \varepsilon_n)$  be the same, i.e. that the Homoscedasticity Condition

$$(1) \quad \text{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \dots, n$$

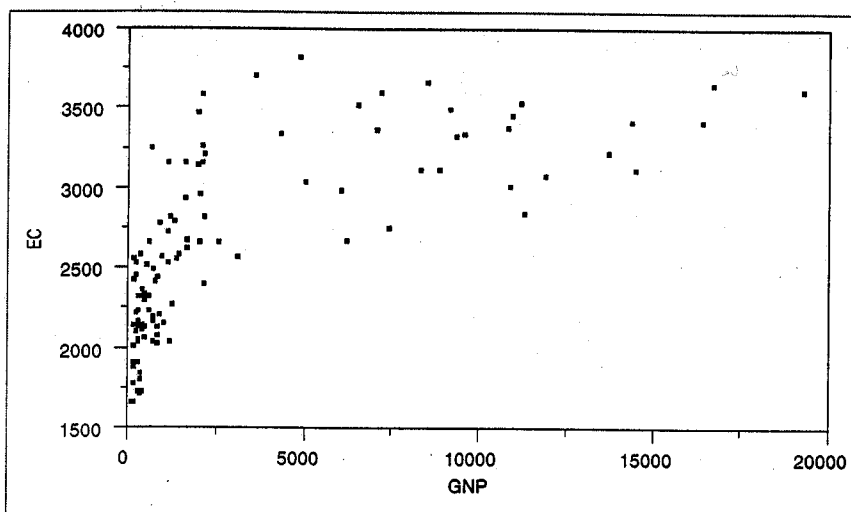
hold. Violations of this condition are called heteroscedasticity. In these notes, we consider two approaches to eliminating heteroscedasticity effects. The first method is designated as variance-stabilizing transformations and the second is known as weighted least squares.

### 1. Variance-Stabilizing Transformations

To illustrate this method, we start with an example involving national energy consumption. Energy consumption has long been regarded as one of the best indicators of standard of living. This can be tested by regressing Energy Consumption on Real GNP for a selection of countries.

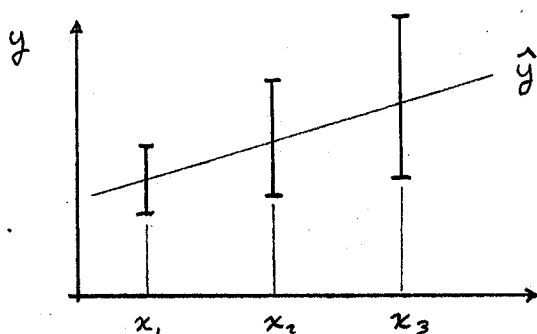
A 1965 study of  $n = 109$  countries included data on annual per capita energy consumption ( $EC_i; i = 1, \dots, n$ ) and annual per capita GNP ( $GNP_i; i = 1, \dots, n$ ). [Note that per capita data is used in order to avoid the obvious "size" effects of big versus small countries, which can lead to violations of (1)]. This data is contained in the class file energy.jmp, and is plotted below:

2



Notice that while there does appear to be a positive trend, the "scatter" of points becomes more extreme at higher levels of GNP.

Empirical Observation: Larger values of  $x$  are often associated with larger variances of  $y$



In the above example, there may be more variations in life styles and energy-consumption patterns among higher GNP countries than lower GNP countries.

3

Log-Log Specification. Consider the power-model

$$Y = ax^b u \quad [P(u > 0) = 1]$$

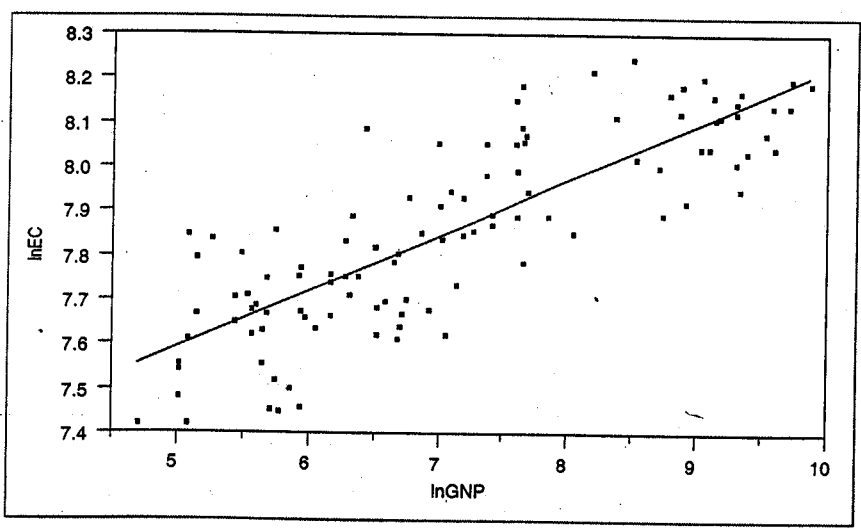
and notice that

$$\text{var}(Y) = (ax^b)^2 \text{var}(u)$$

so if  $a, b > 0$  then  $x \uparrow \Rightarrow \text{var}(Y) \uparrow$ .

$\Rightarrow$  This is a natural model for treating the "increasing variance" problem.

Try Log-Log Regression =



**Summary of Fit**

RSquare	0.684427
RSquare Adj	0.68145
Root Mean Square Error	0.122141
Mean of Response	7.858763
Observations (or Sum Wgts)	108

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	6.9643783	0.060147	115.79	<.0001
lnGNP	0.126016	0.008311	15.16	<.0001

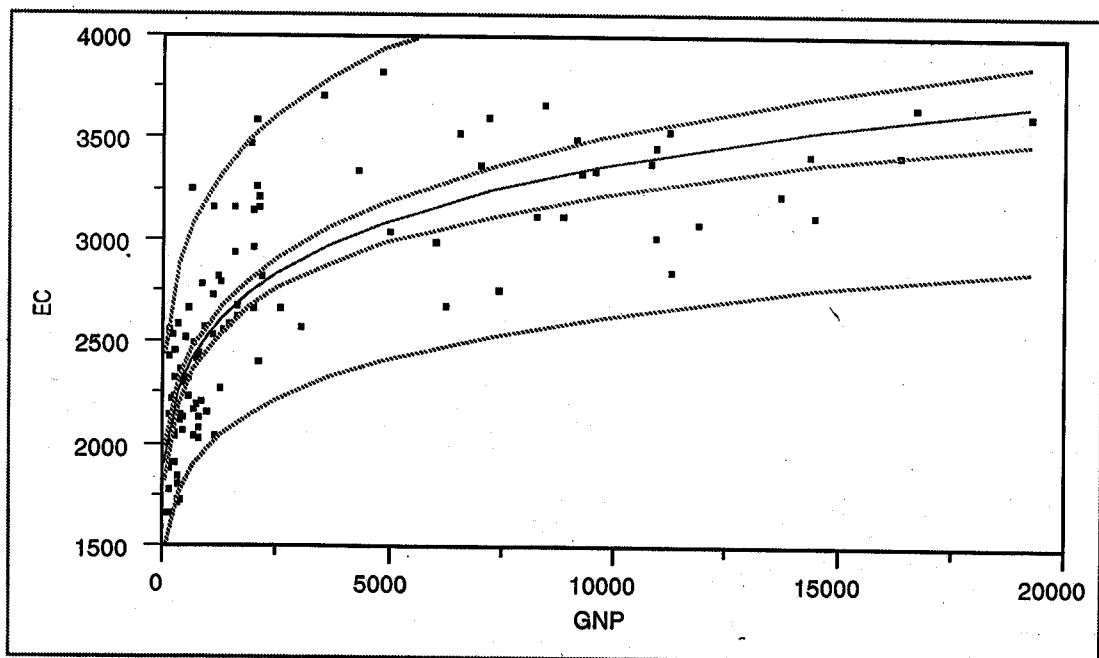
(4)

Note that the overall fit is now quite good ( $R^2 = .68$ ). More importantly, the residual variance now appears to be uniform. Hence the Gauss-Markov assumption of constant variance seems to be satisfied for this transformed model.

If we transform back, by setting

$$\boxed{\text{pred} = e^{\hat{y}}}, \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

and transform the corresponding Confidence Bands for both mean and individual values [done automatically by the 'Fit Transformed' option under 'Fit Y by X' in JMP IN], then the final regression result appears as follows:



(5)

This specific model transformation represents a type of variance-stabilizing transformation. Others include transformations of the  $Y$ -variable only, such as the log transformation

$$(2) \quad Y \rightarrow \log(Y)$$

and the square-root transformation

$$(3) \quad Y \rightarrow \sqrt{Y}$$

All are designed to handle violations of (1) involving variance inflation for larger values of  $Y$ .

## 2: Weighted Least Squares

While such techniques can be very effective, they are only applicable to a limited range of situations. In particular, all transformations above require that  $Y$  be nonnegative (positive in the case of logs). The second method we consider is much more general (but also more complicated to apply).

To motivate this approach, consider the linear model

$$(24) \quad Y_i = \beta_0 + \sum_{j=1}^K \beta_j X_{ij} + \varepsilon_i, \quad i=1, \dots, n$$

with independent residuals  $(\varepsilon_1, \dots, \varepsilon_n)$  distributed as:

(6)

$$(5) \quad \varepsilon_i \sim N(0, \alpha_i \sigma^2), \quad i=1, \dots, n$$

Where the variances,  $\alpha_i \sigma^2$ , differ from sample to sample. This model clearly satisfies all assumptions of the standard multiple regression sampling model except for homoscedasticity.

Now in this context, suppose that we know the  $\alpha$ -factors,  $\alpha_i$ ,  $i=1, \dots, n$ . Then we could easily transform this model into a homoscedastic model by dividing both sides of (4) by the square root of  $\alpha_i$ , to obtain a new model:

$$(6) \quad \frac{1}{\sqrt{\alpha_i}} Y_i = \beta_0 \left( \frac{1}{\sqrt{\alpha_i}} \right) + \sum_{j=1}^K \beta_j \left( \frac{1}{\sqrt{\alpha_i}} X_{ij} \right) + \frac{1}{\sqrt{\alpha_i}} \varepsilon_i$$

which can be written as

$$(7) \quad \tilde{Y}_i = \sum_{j=0}^K \beta_j \tilde{X}_{ij} + \tilde{\varepsilon}_i, \quad i=1, \dots, n$$

where

$$(8) \quad \tilde{Y}_i = \frac{1}{\sqrt{\alpha_i}} Y_i, \quad i=1, \dots, n$$

$$(9) \quad \tilde{\varepsilon}_i = \frac{1}{\sqrt{\alpha_i}} \varepsilon_i, \quad i=1, \dots, n$$

and

$$(10) \quad \tilde{X}_{ij} = \begin{cases} \frac{1}{\sqrt{\alpha_i}}, & j=0 \\ \frac{1}{\sqrt{\alpha_i}} X_{ij}, & j=1, \dots, K \end{cases}$$

(7)

This model is very similar to (4), except that the intercept  $\beta_0$  is now the beta coefficient of a new variable,  $\tilde{x}_{i0}$ .

Hence this new model has  $(k+1)$  variables and no intercept.

This is called a no-intercept regression model (recall Problem 1 in the Additional Practice Problems for Exams 1). Most important however is the fact that the new residuals  $(\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n)$  satisfy all conditions of the standard model. Since  $E(\tilde{\epsilon}_i) = \frac{1}{\sqrt{\alpha_i}} E(\epsilon_i) = 0$  and since independent normality of  $(\epsilon_1, \dots, \epsilon_n)$  implies that  $(\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n)$  are also independent normal, we need only check homoscedasticity.

But since

$$(11) \quad \text{var}(\tilde{\epsilon}_i) = \left(\frac{1}{\sqrt{\alpha_i}}\right)^2 \text{var}(\epsilon_i) = \frac{\alpha_i}{\alpha_i} \sigma^2 = \sigma^2, \quad i=1, \dots, n$$

we see the new residuals satisfy the desired condition:

$$(12) \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i=1, \dots, n$$

$\Rightarrow$  Hence if we knew  $(\alpha_i)$  then this new model [(7), (12)] could be used to estimate  $(\beta_0, \beta_1, \dots, \beta_k)$  by means of a no-intercept regression.

Of course we don't know these  $\alpha$ -factors, but by using the residuals  $(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)$  of the standard multiple regression, we can often estimate the  $\alpha_i$ 's as functions of the explanatory variables. We shall illustrate this procedure in terms of the following Housing Example with data given in Housing - Wtd. jmp.

8

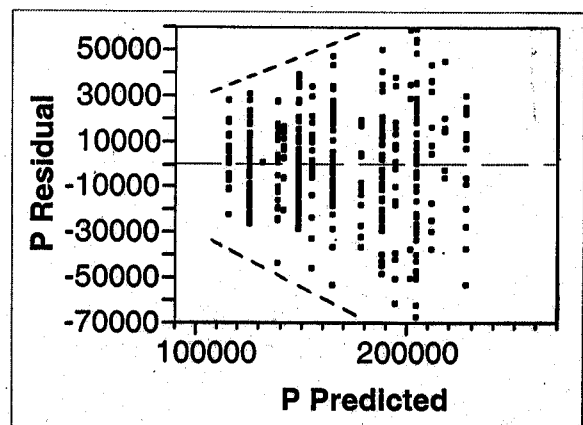
This suburban housing data is similar to the Chicago data, and gives the sales price ( $P$ ) of  $n = 397$  houses, along with three attribute variables: number of bedrooms (BDR), location (LOC), and garage (GAR). [Here  $LOC = 1$  if the house is in the more "desirable" suburban area and  $LOC = 0$  otherwise. Also  $GAR = 1, 2$  depending on whether the garage is one-car or two-car, and  $GAR = 0$  for no garage.]

A multiple regression of  $P$  on  $[BDR, LOC, GAR]$  yields the following results:

Term	Estimate	Std Error	t Ratio	Prob> t	RSquare	0.685775
Intercept	91525.956	3629.44	25.22	<.0001	RSquare Adj	0.683376
BDR	23252.91	1727.006	13.46	<.0001	Root Mean Square Error	21571.53
LOC	9878.8079	2426.347	4.07	<.0001	Mean of Response	164717.6
GAR	16144.337	2044.333	7.90	<.0001	Observations (or Sum Wgts)	397

All coefficients are extremely significant, and taken together, these three variables account for more than 68% of the variance. A Normal Quantile plot confirms normality of the residuals.

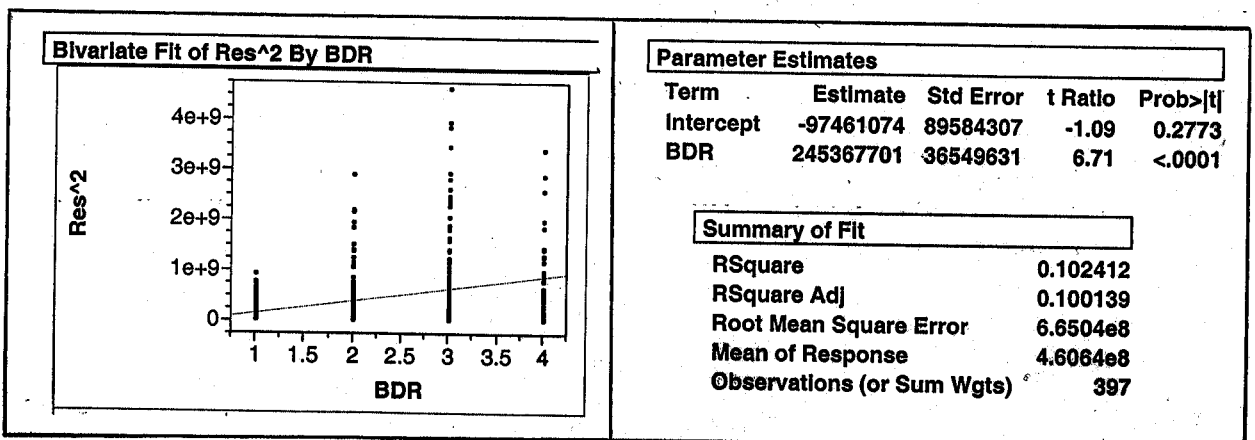
But unfortunately the residual plot shows a clear inflation of residual variance at higher predicted price levels (as is emphasized graphically by the dashed lines added to the plot).



(9)

One could try to eliminate this effect by a variance-stabilizing transformation of  $P$ . [Here a log transformation of  $P$  is actually quite successful, as can be verified by a regression of  $\log(P)$  on (BDR, LUC, GAR).] However, such transformations in this case would destroy the natural interpretation of the beta coefficients as implicit prices ("Hedonic Prices") of these housing attributes.

Hence a weighted least squares approach is preferable since the betas estimated are precisely those of the original linear model. In order to estimate the  $\alpha_i$ 's for this model, the simplest approach (adopted here) is to treat the squared residuals,  $\hat{\epsilon}_i^2$ , as (single-sample) estimates of the variances,  $\text{var}(\epsilon_i) = E(\epsilon_i^2)$ , and to regress each explanatory variable against these squared residuals to see which (if any) best accounts for this variance-inflation effect. Both BDR and GAR are very significant factors. Since they are highly correlated ( $\text{corr} = .73$ ), only BDR is used, as shown below:



Note that  $R^2$  is low, but the  $P$ -value on BDR is very significant.

Note also that (as with the original plot of residuals), variance actually diminishes at the highest level ( $BDR = 4$ ). So perhaps a quadratic fit might be better here. But since our main objective is to diminish heteroscedastic effects in the simplest possible way, we adopt the "linearity" assumption that the  $d$ -factors are proportional to BDR, i.e. that

$$(13) \quad \text{var}(\varepsilon_i) = (BDR_i) \sigma^2, \quad i=1, \dots, n$$

(where the constant of proportionality is included in  $\sigma^2$ ).

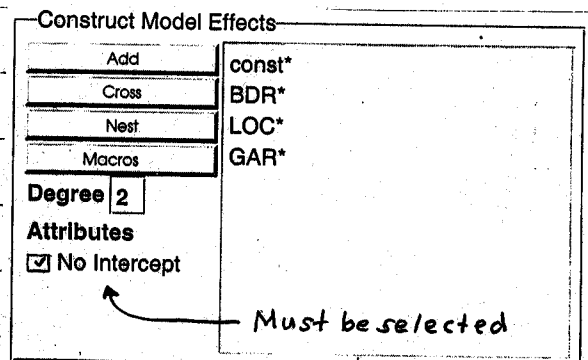
This implies that we may take  $d_i \equiv BDR_i$ , and construct a new regression model as in (7) - (10). To do so, let

$W = 1/\sqrt{BDR}$ , and construct the new variables (new columns in JMPIN):

$$(14) \quad P^* = W \cdot P, \quad BDR^* = W \cdot BDR, \quad LOC^* = W \cdot LOC, \quad GAR^* = W \cdot GAR$$

Also a new variable must be added for the intercept,  $\beta_0$ , which is identical with  $W$ , say  $\text{const}^* \equiv W$ .

To run this regression, use Fit Model to regress  $P^*$  against the four variables ( $\text{const}^*$ ,  $BDR^*$ ,  $LOC^*$ ,  $GAR^*$ ) as shown to the right. Here the "No Intercept" option must be turned on to ensure the proper beta estimates.



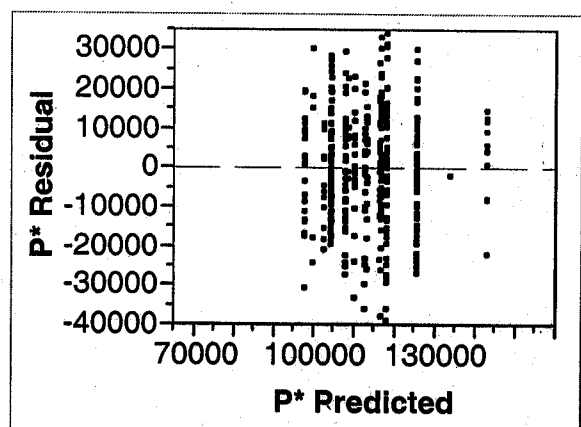
The results of this regression are shown below.

Term	Estimate	Term	Std Error	t Ratio	Prob> t	RSquare	
Intercept Zeroed	0	Intercept	0			RSquare Adj	
const*	94049.313	const*	2985.021	31.51	<.0001	Root Mean Square Error	13998.63
BDR*	21919.337	BDR*	1578.357	13.89	<.0001	Mean of Response	112671.3
LOC*	9178.7246	LOC*	2165.516	4.24	<.0001	Observations (or Sum Wgts)	397
GAR*	17395.322	GAR*	1940.192	8.97	<.0001		

Notice first that  $R^2$  is not reported. The reason for this is that the basic variance decomposition used to motivate  $R^2$  is no longer valid for the "no-intercept" form of regression. [Some packages compute it anyway, but JMPIN does not do so since the value is not comparable with that of the "intercept" form.] Next observe that all betas are still significant (in fact slightly more significant) than in the original regression. Notice also that the beta values are quite close to the old values (with differences less than 10%).

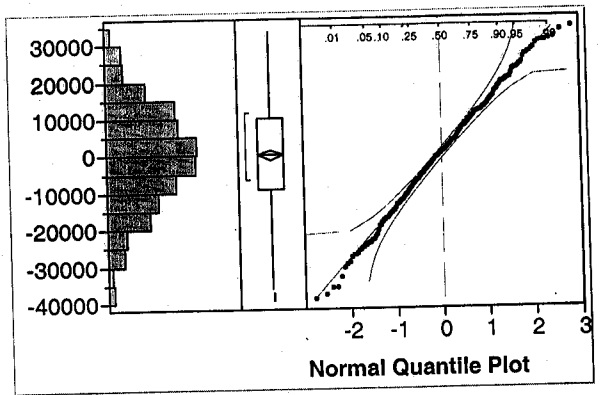
Finally, an examination of the new residuals shows that the inflation pattern has all but disappeared. There is still a noticeable decrease in variation for high values [that was not corrected by the simple linearity assumption used in (13)]

Note that the range and spacing of  $P^*$ -Predicted values is very different from the  $P$ -Predicted values. These new values have little meaning, but the residual pattern does.



(12)

To complete the analysis, one must check for normality of the residuals. This is confirmed by a Normal Quantile plot of the residual estimates ( $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ ) shown to the right.



→ In summary, this weighted least squares procedure has eliminated heteroscedasticity, and thereby produced beta estimates that are more in accord with the Gauss-Markov assumptions. This ensures that these estimates are BLUE, and hence can be expected to be more reliable than the original estimator.