Chapter 3 – Diagnostics and Remedial Measures

Diagnostics for the Predictor Variable (*X*)

Levels of the independent variable, particularly in settings where the experimenter does not control the levels, should be studied. Problems can arise when:

- One or more observations have X levels far away from the others
- When data are collected over time or space, *X* levels that are close together in time or space are "more similar" than the overall set of *X* levels

Useful plots of X levels include: histograms, box-plots, stem-and-leaf diagrams, and sequence plots (versus time order). Also, a useful measure is simply the *z*-score for each observation's *X* value. We will later discuss remedies for these problems in **Chapter 9**???.

Residuals

"True" Error Term: $\varepsilon_i = Y_i - E\{Y_i\} = Y_i - (\beta_0 + \beta_1 X_i)$

Observed Residual: $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$

Recall the assumption on the "true" error terms: they are independent and normally distributed with mean 0, and variance σ^2 ($\varepsilon_i \sim NID(0,\sigma^2)$). The residuals have mean 0, since they sum to 0, but they are not independent since they are based on the fitted values from the same observations, but as *n* increases, this becomes less important. Ignoring the non-independence for now, we have, concerning the residuals ($e_1,\ldots,e_n$):

$$\bar{e} = \frac{\sum e_i}{n} = \frac{0}{n} = 0 \qquad s^2\{e_i\} = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum(e_i - 0)^2}{n-2} = \frac{\sum e_i^2}{n-2} = MSE$$

Semi-studentized Residuals

We are accustomed to standardizing random variables by centering them (subtracting off the mean) and scaling them (dividing through by the standard deviation), thus creating a *z*-score.
While the theoretical standard deviation of e_i is a complicated function of the entire set of sample data (we will see this after introducing the matrix approach to regression), we can approximate the standardized residual as follows, which we call the **semi-studentized residuals**:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

In large samples, these can be treated approximately as *t*-statistics, with *n*-2 degrees of freedom.

Diagnostic Plots for Residuals

The major assumptions of the model are: (i) the relationship between the mean of Y and X is linear, (ii) the errors are normally distributed, (iii) the mean of the errors is 0, (iv) the variance of the errors is constant and equals σ^2, (v) the errors are independent, (vi) the model contains all predictors related to E{Y}, and (vii) the model fits for all data observations. These can be visually investigated with various plots.

Linear Relationship Between E{Y} and X

Plot the residuals versus either X or the fitted values. This will appear as a random cloud of points centered at 0 under linearity, and will appear U-shaped (or inverted U-shaped) if the relationship is not linear.

Normally Distributed Errors

Obtain a histogram of the residuals, and determine whether it is approximately mound shaped. Alternatively, a normal probability plot can be obtained as follows (Note that in R, this is trivial with the **qqnorm** and **qqline** commands):

1. Order the residuals from smallest (large negative values) to largest (large positive values). Assign the ranks as k.
2. Compute the percentile for each residual (this is one of several versions): $\frac{k-0.375}{n+0.25}$
3. Obtain the z value from the standard normal distribution corresponding to these percentiles: $z\left(\frac{k-0.375}{n+0.25}\right)$
4. Multiply the z values by $s=\sqrt{MSE}$ these are the "expected" residuals for the k^{th} smallest residuals under the normality assumption
5. Plot the observed residuals on the vertical axis versus the expected residuals on the horizontal axis. This should be approximately a straight line with slope 1.

Errors have Mean 0

Since the residuals sum to 0, and thus have mean 0, we have no need to check this assumption.

Errors have Constant Variance

Plot the residuals versus X or the fitted values. This should appear as a random cloud of points, centered at 0, if the variance is constant. If the error variance is not constant, this may appear as a funnel shape.

Errors are Independent (When Data Collected Over Time)

Plot the residuals versus the time order (when data are collected over time). If the errors are independent, they should appear as a random cloud of points centered at 0. If the errors are positively correlated they will tend to approximate a smooth (not necessarily monotone) functional form.

No Predictors Have Been Omitted

Plot residuals versus omitted factors, or against *X* seperately for each level of a categorical omitted factor. If the current model is correct, these should be random clouds of points centered at 0. If patterns arise, the omitted variables may need to be included in model (Multiple Regression).

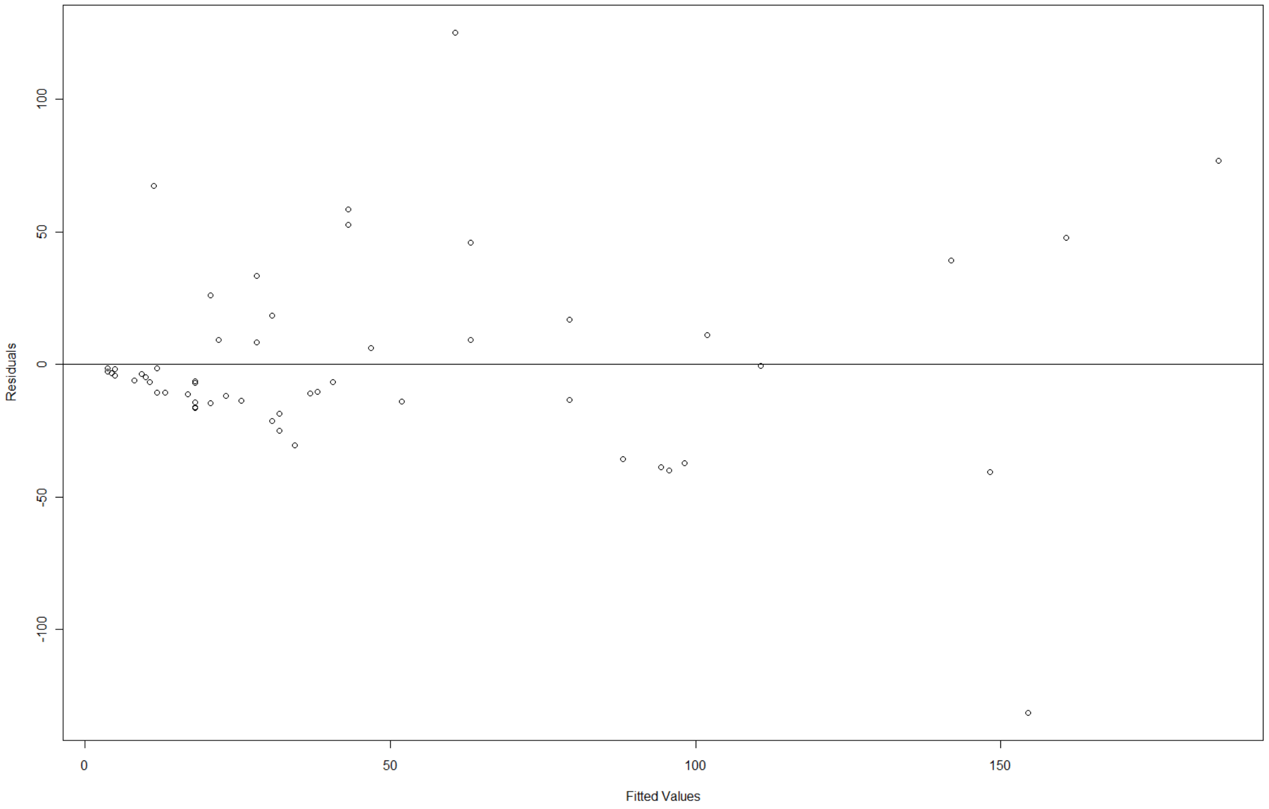
Model Fits for All Observations

Plot Residuals versus fitted values. As long as no residuals stand out (either much higher or lower) from the others, the model fits all observations. Any residuals that are very extreme, are evidence of data points that are called *outliers*. Any outliers should be checked as possible data entry errors. We will cover this problem in detail in Chapter 10.

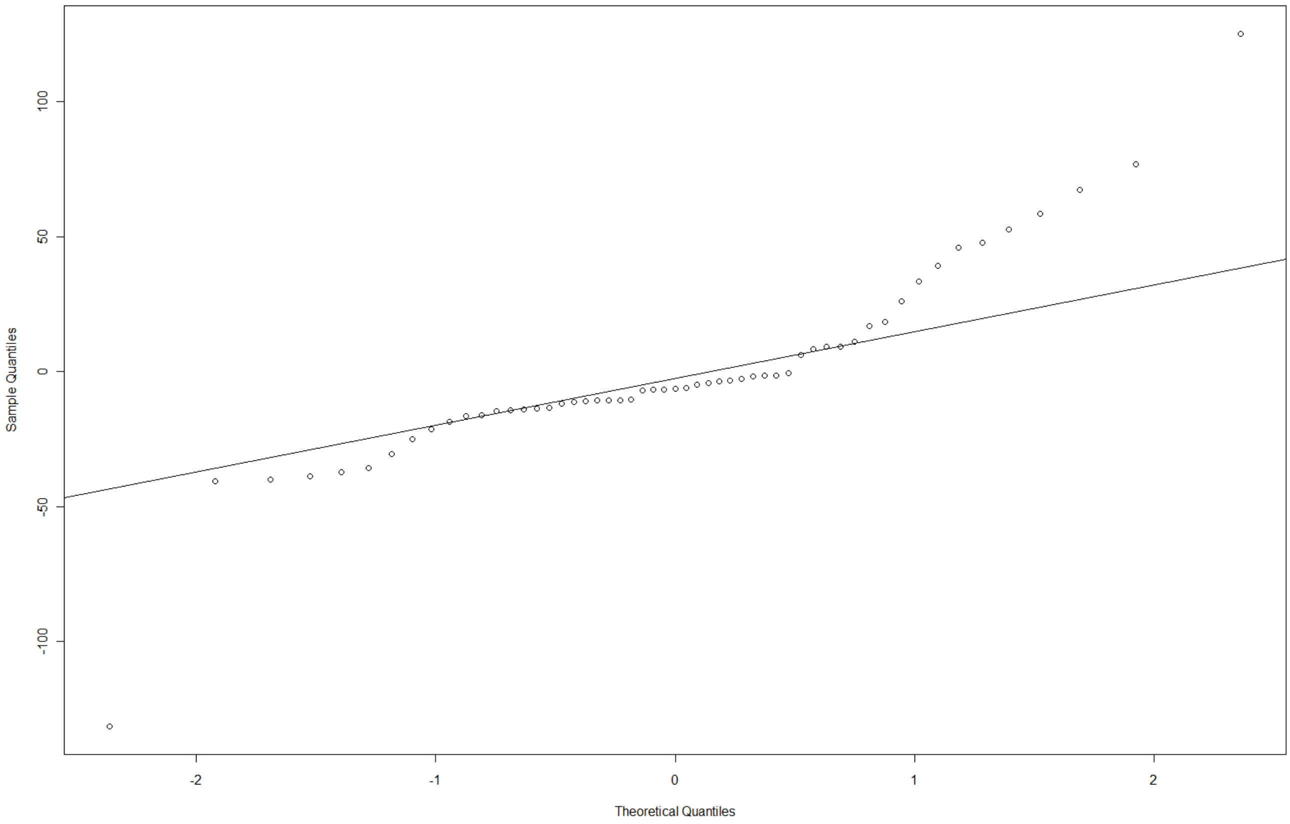
Example: Bollywood Box Office Data

```
bbo <-
read.csv("http://www.stat.ufl.edu/~winner/sta4210/mydata/bollywood_boxoffice.csv",
      header=T)

attach(bbo)
names(bbo)

bbo.reg1 <- lm(Gross ~ Budget)
summary(bbo.reg1)
e1 <- residuals(bbo.reg1)
yhat1 <- predict(bbo.reg1)

plot(yhat1,e1,main="Bollywood Regression - Residuals vs Fitted Values",
xlab="Fitted Values",
     ylab="Residuals")
abline(h=0)

qqnorm(e1); qqline(e1)
```

The plots below appear to make the constant variance assumption and the normality assumption seem unreasonable. We will conduct formal tests below.



Bollywood Regression - Residuals vs Fitted Values



Normal Q-Q Plot

Tests Involving Residuals

Several of the assumptions stated above can be formally tested based on statistical tests.

Normally Distributed Errors (Correlation Test)

Using the expected residuals (denoted $e_i *$) obtained to construct a normal probability plot, we can obtain the correlation coefficient between the observed residuals and their expected residuals under normality:

$$\boxed{r_{ee*} = \frac{\sum ee *}{\sqrt{\sum e^2 \sum (e*)^2}}}$$

The test is conducted as follows:

- H_0 :Error terms are normally distributed
- H_A :Error terms are not normally distributed
- *TS*: r_{ee*}
- *RR*: $r_{ee*} \leq$ Tabled values in Table B.6, Page 1329 (indexed by α and n)

Note this is a test where we do not wish to reject the null hypothesis. Another test that is more complex to manually compute, but is automatically reported by several software packages is the Shapiro-Wilks test. It's null and alternative hypotheses are the same as for the correlation test, and *P*-values are computed for the test.

Example: Bollywood Box Office Data

The residuals and their expected values under normality are given below. The correlation between the actual residuals and their expected values is $r_{ee*} = 0.9220$. From Table B.6, we have the following critical values for sample sizes of n=50 and n=60 for various α levels.

	α=.10	α=.05	α=.01
n=50	**0.981**	**0.977**	**0.966**
n=60	**0.984**	**0.980**	**0.971**

Clearly, r_{ee*} is well below all of the critical values. Normality assumption is rejected. Shapiro-Wilk test:

```
bbo <- read.csv("http://www.stat.ufl.edu/~winner/sta4210/mydata/bollywood_boxoffice.csv",
      header=T)

attach(bbo)
names(bbo)

bbo.reg1 <- lm(Gross ~ Budget)
summary(bbo.reg1)
e1 <- residuals(bbo.reg1)
yhat1 <- predict(bbo.reg1)

shapiro.test(e1)

############## R Output

> shapiro.test(e1)

        Shapiro-Wilk normality test

data:  e1
W = 0.87, p-value = 2.627e-05
```

Movie Name	Y	X	Y-hat	e	rank_e	pctile	z(pct)	e*
Ek Villain	95.64	36.00	43.08	52.56	51	0.9163	1.3805	50.41
Humshakals	55.65	77.00	94.37	-38.72	4	0.0656	-1.5093	-55.11
Holiday	110.01	90.00	110.63	-0.62	38	0.6810	0.4705	17.18
Fugly	11.16	16.00	18.06	-6.90	25	0.4457	-0.1365	-4.99
City Lights	5.19	9.50	9.93	-4.74	30	0.5362	0.0909	3.32
Kuku Mathur Ki Jhand Ho Gayi	2.23	4.50	3.67	-1.44	36	0.6448	0.3713	13.56
Heropanti	49.07	26.00	30.57	18.50	45	0.8077	0.8694	31.75
2 States	101.61	36.00	43.08	58.53	52	0.9344	1.5093	55.11
Main Tera Hero	53.04	39.00	46.83	6.21	39	0.6991	0.5218	19.05
Ragini MMS 2	46.59	18.00	20.56	26.03	46	0.8258	0.9377	34.24
Queen	61.47	24.00	28.07	33.40	47	0.8439	1.0106	36.90
Gunday	72.26	52.00	63.10	9.16	41	0.7353	0.6289	22.96
Jai Ho	107.71	120.00	148.16	-40.45	2	0.0294	-1.8895	-68.99
Kochadaiiyaan (All Languages India)	23.07	125.00	154.42	-131.35	1	0.0113	-2.2797	-83.24
The Xpose	11.69	16.00	18.06	-6.37	28	0.5000	0.0000	0.00
Hawa Hawaai	10.42	11.00	11.81	-1.39	37	0.6629	0.4204	15.35
Mastram	3.36	5.50	4.93	-1.57	35	0.6267	0.3231	11.80
Koyelaanchal	2.17	8.00	8.05	-5.88	29	0.5181	0.0454	1.66
Yeh Hain Bakrapur	0.97	4.50	3.67	-2.70	34	0.6086	0.2757	10.07
Manjunath	1.02	5.00	4.30	-3.28	33	0.5905	0.2288	8.36
Purani Jeans	1.28	11.00	11.81	-10.53	23	0.4095	-0.2288	-8.36
Kya Dilli Kya Lahore	0.63	5.50	4.93	-4.30	31	0.5543	0.1365	4.99
Revolver Rani	9.44	26.00	30.57	-21.13	9	0.1561	-1.0106	-36.90
Kaanchi	3.93	29.00	34.32	-30.39	7	0.1199	-1.1754	-42.92
Samrat & Co	2.10	16.00	18.06	-15.96	12	0.2104	-0.8050	-29.39
Bhootnath Returns	34.03	34.00	40.58	-6.55	27	0.4819	-0.0454	-1.66
Youngistaan	6.76	27.00	31.82	-25.06	8	0.1380	-1.0893	-39.78
Dishkiyaoon	5.79	9.00	9.30	-3.51	32	0.5724	0.1825	6.66
O Teri	3.72	16.00	18.06	-14.34	14	0.2466	-0.6852	-25.02
Gang Of Ghosts	1.55	16.00	18.06	-16.51	11	0.1923	-0.8694	-31.75
Bewakoofiyaan	12.06	22.00	25.57	-13.51	16	0.2828	-0.5745	-20.98
Gulaab Gang	13.32	27.00	31.82	-18.50	10	0.1742	-0.9377	-34.24
Total Siyappa	5.91	18.00	20.56	-14.65	13	0.2285	-0.7438	-27.16
Shaadi ke Side Effects	37.95	43.00	51.84	-13.89	15	0.2647	-0.6289	-22.96
Highway	27.71	32.00	38.08	-10.37	24	0.4276	-0.1825	-6.66
Darr @ Mall	5.70	15.00	16.81	-11.11	19	0.3371	-0.4204	-15.35
Hasee Toh Phasee	36.52	24.00	28.07	8.45	40	0.7172	0.5745	20.98
Heartless	1.16	11.00	11.81	-10.65	22	0.3914	-0.2757	-10.07
One By Two	2.41	12.00	13.06	-10.65	21	0.3733	-0.3231	-11.80
Yaariyan	31.04	19.00	21.81	9.23	42	0.7534	0.6852	25.02
Dedh Ishqiya	25.87	31.00	36.83	-10.96	20	0.3552	-0.3713	-13.56
Sholay 3D	11.25	20.00	23.06	-11.81	18	0.3190	-0.4705	-17.18
Joe B Carvalho	3.83	10.00	10.55	-6.72	26	0.4638	-0.0909	-3.32
Dhoom 3 (Hindi)	262.58	150.00	185.69	76.89	54	0.9706	1.8895	68.99
Chennai Express	208.44	130.00	160.67	47.77	50	0.8982	1.2713	46.42
Krrish 3 (Hindi)	181.11	115.00	141.91	39.20	48	0.8620	1.0893	39.78
Yeh Jawani Hain Deewani	185.83	50.00	60.59	125.24	55	0.9887	2.2797	83.24
R Rajkumar	66.10	65.00	79.36	-13.26	17	0.3009	-0.5218	-19.05
Ram Leela	112.96	83.00	101.88	11.08	43	0.7715	0.7438	27.16
Boss	52.38	72.00	88.12	-35.74	6	0.1018	-1.2713	-46.42
Besharam	55.79	78.00	95.62	-39.83	3	0.0475	-1.6695	-60.96
OUATIMD	60.93	80.00	98.12	-37.19	5	0.0837	-1.3805	-50.41
Bhag Milkha Bhag	109.18	52.00	63.10	46.08	49	0.8801	1.1754	42.92
Race 2	96.34	65.00	79.36	16.98	44	0.7896	0.8050	29.39
Aashiqui 2	78.42	10.50	11.18	67.24	53	0.9525	1.6695	60.96

Errors have Constant Variance

1. Brown-Forsyth (aka Modified Levene) Test

There are several ways to test for equal variances. One simple (to describe) approach is a modified version of Levene's test, which tests for equality of variances, without depending on the errors being normally distributed. Recall that due to Central Limit Theorems, lack of normality causes us no problems in large samples, as long as the other assumptions hold. The procedure can be described as follows:

1. Split the data into 2 groups, one group with low X values containing n_1 of the observations, the other group with high X values containing n_2 observations (n_1=n_2=n).
2. Obtain the medians of the residuals for each group, labeling them $\tilde{e}_1$ and $\tilde{e}_2$, respectively.
3. Obtain the absolute deviations for each residual from its group median:

   $$d_{i1} = |e_{i1} - \tilde{e}_1| \qquad i = 1,\ldots,n_1 \qquad\qquad d_{i2} = |e_{i2} - \tilde{e}_2| \qquad i = 1,\ldots,n_2$$

4. Obtain the sample mean absolute deviation from the median for each group: $\bar{d}_1 = \frac{\sum_{i=1}^{n_1} d_{i1}}{n_1}$, $\bar{d}_2 = \frac{\sum_{i=1}^{n_2} d_{i2}}{n_2}$
5. Obtain the pooled variance of the absolute deviations: $s^2 = \frac{\sum_{i=1}^{n_1}(d_{i1} - \bar{d}_1)^2 + \sum_{i=1}^{n_2}(d_{i2} - \bar{d}_2)^2}{n-2}$
6. Compute the test statistic: $t_{BF}^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
7. Conclude that the error variance is not constant if $|t_{BF}^*| \geq t(1-\alpha/2; n-2)$, otherwise conclude the error variance is constant.

Example: Bollywood Box Office Data

By using a split-point at X=25, we have n_1=27 films with budgets below 25, and n_2=28 films with budgets above 25. Using the AVERAGE and DEVSQ functions in EXCEL, we obtain the following test:

$$n_1 = 27 \quad \tilde{e}_1 = -5.8830 \quad \bar{d}_1 = 10.0418 \quad \sum_{i=1}^{n_1}\left(d_{i1} - \bar{d}_1\right)^2 = 6256.945$$

$$n_2 = 28 \quad \tilde{e}_2 = -8.4577 \quad \bar{d}_2 = 34.5666 \quad \sum_{i=1}^{n_2}\left(d_{i2} - \bar{d}_2\right)^2 = 31342.34$$

$$s^2 = \frac{6256.945 + 31342.34}{55-2} = 709.4205 \quad \Rightarrow \quad s = \sqrt{709.4205} = 26.6350$$

$$t_{BF}^* = \frac{10.0418 - 34.5666}{26.6350\sqrt{\frac{1}{27} + \frac{1}{28}}} = \frac{-24.5248}{7.1841} = -3.4138 \qquad t\left(.975, 53\right) = 2.0057$$

We conclude that there is evidence of non-constant error variance.

2. Breusch-Pagan (aka Cook-Weisberg) Test

Breusch-Pagan (aka Cook-Weisberg) Test:

$$H_0: \text{ Equal Variance Among Errors } \quad \sigma^2\{\varepsilon_i\} = \sigma^2 \ \forall\, i$$

$$H_A: \text{Unequal Variance Among Errors} \quad \sigma_i^2 = \sigma^2 h\left(\gamma_1 X_{i1} + \ldots + \gamma_p X_{ip}\right)$$

1) Let $SSE = \sum_{i=1}^{n} e_i^2$ from original regression

2) Fit Regression of e_i^2 on $X_{i1}, \ldots X_{ip}$ and obtain $SS(\text{Reg}^*)$

$$\text{Test Statistic: } X_{BP}^2 = \frac{SS(\text{Reg}^*)/2}{\left(\sum_{i=1}^{n} e_i^2 \Big/ n\right)^2} \quad \overset{H_0}{\sim} \quad \chi_p^2$$

$$\text{Reject H}_0 \text{ if } X_{BP}^2 \geq \chi^2(1-\alpha; p) \qquad p = \#\text{ of predictors}$$

<u>Example: Bollywood Box Office Data</u>

Regressing the squared residuals of Budget (*X*) results in *SS(Reg*)* = 106398914.7. From the original regression, we have *SSE* = 70664.39. Thus, the test statistic and rejection region are:

$$X_{BP}^2 = \frac{106398914.7/2}{(70664.39/55)^2} = \frac{53199457.35}{1650729.11} = 32.2279$$

$$\chi^2(0.95,1) = 3.841$$

Thus, there is strong evidence against the assumption of normality.
R code for computing the Breusch-Pagan test (it utilizes the **lmtest** package):

```
install.packages("lmtest")
library(lmtest)

bptest(Gross ~ Budget,studentize=FALSE)

### Output ####
> bptest(Gross ~ Budget,studentize=FALSE)

        Breusch-Pagan test

data:  Gross ~ Budget
BP = 32.2279, df = 1, p-value = 1.371e-08
```

Errors are Independent (When Data Collected Over Time)

When data are collected over time, one common departure from independence is that error terms are positively autocorrelated. That is, the errors that are close to each other in time are similar in magnitude and sign. This can happen when learning or fatigue is occuring over time in physical processes or when long-term trends are occuring in social processes. A test that can be used to determine whether positive autocorrelation (non-independence of errors) exists is the Durbin-Watson test (see Section **12.3**, we will consider it in more detail later). The test can be conducted as follows:

- H_0 : The errors are independent
- H_A : The errors are not independent (positively autocorrelated)
- $TS: D = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$
- Decision Rule: (i) Reject H_0 if $D \le d_L$ (ii) Accept H_A if $D \ge d_U$ (iii) withhold judgment if $d_L < D < d_U$ where d_L, d_U are bounds indexed by: α, n, and p-1 (the number of predictors, which is 1 for now). These bounds are given in Table B.7, pages 1330-1331.

The Bollywood data is not a time series. We will cover an example of this test later in the course.

F Test for Lack of Fit to Test for Linear Relation Between E{*Y*} and *X*

A test can be conducted to determine whether the true regression function is that which is being currently specified. For the test to be conducted, we must have the following conditions hold. The observations *Y*, conditional on their *X* level are independent, normally distributed, and have the same variance σ^2. Further, the *X* levels in the sample must have repeat observations at a minimum (preferably more) of one *X* level. Repeat trials at the same level(s) of the predictor variable(s) are called *replications*. The actual observations are referred to as *replicates*.

The null and alternative hypotheses for the simple linear regression model are stated as:

$$H_0 : E\{Y \mid X\} = \beta_0 + \beta_1 X \qquad\qquad H_A : E\{Y \mid X\} = \mu_X \ne \beta_0 + \beta_1 X$$

The null hypothesis states that the mean structure is a linear relation, the alternative says that the mean structure is any structure except linear (this is not simply a test of whether β_1=0). The test (which is a special case of the general linear test) is conducted as follows:

1. Begin with *n* total observations at *c* distinct levels of *X*. There are n_j observations at the j^{th} of *X*. $n_1 + \cdots + n_c = n$
2. Let Y_{ij} be the i^{th} replicate at the j^{th} level of *X* $\quad j = 1, \ldots, c \quad i = 1, \ldots, n_j$
3. Fit the Full model (H_A): $Y_{ij} = \mu_j + \varepsilon_{ij}$ The least squares estimate of μ_j is $\hat{\mu}_j = \overline{Y}_j$
4. Obtain the error sum of squares for the Full model, also known as the Pure Error sum of squares.

$$SSE(F) = SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

5. The degrees of freedom for the Full model is df_F= *n-c*. This is from the fact that the j^{th} level of *X*, we have n_j-1 degrees of freedom, and they sum up to *n-c*. Also, we have estimated *c* parameters ($\mu_1,\ldots,\mu_c$).
6. Fit the Reduced model (H_0): $Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}$ The least squares estimate of $\beta_0 + \beta_1 X_j$ is $\hat{Y}_j = b_0 + b_1 X_j$
7. Obtain the error sum of squares for the Reduced model, also known as the Error sum of squares. $$SSE(R) = SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$$
8. The degrees of freedom for the Reduced model is df_R=*n*-2. We have estimated two parameters in this model (β_0, β_1)
9. Compute the *F* statistic: $$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}} = \frac{\dfrac{SSE - SSPE}{(n-2)-(n-c)}}{\dfrac{SSPE}{n-c}} = \frac{\dfrac{SSE - SSPE}{c-2}}{MSPE}$$
10. Obtain the rejection region: $RR: F^* \geq F(1-\alpha; c-2, n-c)$

Note that the numerator of the *F* statistic is also known as the **Lack of Fit** sum of squares:

$$SSLF = SSE - SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\overline{Y}_j - \hat{Y}_j)^2 = \sum_{j=1}^{c} n_j(\overline{Y}_j - \hat{Y}_j)^2 \qquad df_{LF} = c-2$$

The degrees of freedom can be intuitively thought of as being a result of us fitting a aimple linear regression model of *c* sample means on *X*. Note then that the *F* statistic can be written as:

$$\boxed{F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}} = \frac{\dfrac{SSE - SSPE}{(n-2)-(n-c)}}{\dfrac{SSPE}{n-c}} = \frac{\dfrac{SSE - SSPE}{c-2}}{MSPE} = \frac{\dfrac{SSLF}{c-2}}{MSPE} = \frac{MSLF}{MSPE}}$$

Thus, we have partitioned the Error sum of squares for the linear regression model into Pure Error (based on deviations from individual responses to their group means) and Lack of Fit (based on deviations from group means to the fitted values from the regression model).

The expected mean squares for *MSPE* and *MSLF* are as follows:

$$E\{MSPE\} = \sigma^2 \qquad E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2}$$

Under the null hypothesis (relationship is linear), the second term for the lack of fit mean square is 0. Under the alternative hypothesis (relationship is not linear), the second term is positive. Thus large values of the *F* statistic are consistent with the alternative hypothesis.

<u>Example: Bollywood Box Office Data</u>

There are n=55 movies with c=40 distinct budgets. The following EXCEL spreadsheet has results:

Movie Name	Y	X	Y-hat	Group	Ybar(grp)	LackFit	PureError
Ek Villain	2.23	4.50	3.67	1.00	1.60	-2.07	0.63
Humshakals	0.97	4.50	3.67	1.00	1.60	-2.07	-0.63
Holiday	1.02	5.00	4.30	2.00	1.02	-3.28	0.00
Fugly	3.36	5.50	4.93	3.00	2.00	-2.93	1.37
City Lights	0.63	5.50	4.93	3.00	2.00	-2.93	-1.37
Kuku Mathur Ki Jhand Ho Gayi	2.17	8.00	8.05	4.00	2.17	-5.88	0.00
Heropanti	5.79	9.00	9.30	5.00	5.79	-3.51	0.00
2 States	5.19	9.50	9.93	6.00	5.19	-4.74	0.00
Main Tera Hero	3.83	10.00	10.55	7.00	3.83	-6.72	0.00
Ragini MMS 2	78.42	10.50	11.18	8.00	78.42	67.24	0.00
Queen	10.42	11.00	11.81	9.00	4.29	-7.52	6.13
Gunday	1.28	11.00	11.81	9.00	4.29	-7.52	-3.01
Jai Ho	1.16	11.00	11.81	9.00	4.29	-7.52	-3.13
Kochadaiiyaan (All Languages India)	2.41	12.00	13.06	10.00	2.41	-10.65	0.00
The Xpose	5.70	15.00	16.81	11.00	5.70	-11.11	0.00
Hawa Hawaai	11.16	16.00	18.06	12.00	6.04	-12.02	5.12
Mastram	11.69	16.00	18.06	12.00	6.04	-12.02	5.65
Koyelaanchal	2.10	16.00	18.06	12.00	6.04	-12.02	-3.94
Yeh Hain Bakrapur	3.72	16.00	18.06	12.00	6.04	-12.02	-2.32
Manjunath	1.55	16.00	18.06	12.00	6.04	-12.02	-4.49
Purani Jeans	46.59	18.00	20.56	13.00	26.25	5.69	20.34
Kya Dilli Kya Lahore	5.91	18.00	20.56	13.00	26.25	5.69	-20.34
Revolver Rani	31.04	19.00	21.81	14.00	31.04	9.23	0.00
Kaanchi	11.25	20.00	23.06	15.00	11.25	-11.81	0.00
Samrat & Co	12.06	22.00	25.57	16.00	12.06	-13.51	0.00
Bhootnath Returns	61.47	24.00	28.07	17.00	49.00	20.93	12.48
Youngistaan	36.52	24.00	28.07	17.00	49.00	20.93	-12.48
Dishkiyaoon	49.07	26.00	30.57	18.00	29.26	-1.32	19.82
O Teri	9.44	26.00	30.57	18.00	29.26	-1.32	-19.82
Gang Of Ghosts	6.76	27.00	31.82	19.00	10.04	-21.78	-3.28
Bewakoofiyaan	13.32	27.00	31.82	19.00	10.04	-21.78	3.28
Gulaab Gang	3.93	29.00	34.32	20.00	3.93	-30.39	0.00
Total Siyappa	25.87	31.00	36.83	21.00	25.87	-10.96	0.00
Shaadi ke Side Effects	27.71	32.00	38.08	22.00	27.71	-10.37	0.00
Highway	34.03	34.00	40.58	23.00	34.03	-6.55	0.00
Darr @ Mall	95.64	36.00	43.08	24.00	98.63	55.54	-2.99
Hasee Toh Phasee	101.61	36.00	43.08	24.00	98.63	55.54	2.99
Heartless	53.04	39.00	46.83	25.00	53.04	6.21	0.00
One By Two	37.95	43.00	51.84	26.00	37.95	-13.89	0.00
Yaariyan	185.83	50.00	60.59	27.00	185.83	125.24	0.00
Dedh Ishqiya	72.26	52.00	63.10	28.00	90.72	27.62	-18.46
Sholay 3D	109.18	52.00	63.10	28.00	90.72	27.62	18.46
Joe B Carvalho	66.10	65.00	79.36	29.00	81.22	1.86	-15.12
Dhoom 3 (Hindi)	96.34	65.00	79.36	29.00	81.22	1.86	15.12
Chennai Express	52.38	72.00	88.12	30.00	52.38	-35.74	0.00
Krrish 3 (Hindi)	55.65	77.00	94.37	31.00	55.65	-38.72	0.00
Yeh Jawani Hain Deewani	55.79	78.00	95.62	32.00	55.79	-39.83	0.00
R Rajkumar	60.93	80.00	98.12	33.00	60.93	-37.19	0.00
Ram Leela	112.96	83.00	101.88	34.00	112.96	11.08	0.00
Boss	110.01	90.00	110.63	35.00	110.01	-0.62	0.00
Besharam	181.11	115.00	141.91	36.00	181.11	39.20	0.00
OUATIMD	107.71	120.00	148.16	37.00	107.71	-40.45	0.00
Bhag Milkha Bhag	23.07	125.00	154.42	38.00	23.07	-131.35	0.00
Race 2	208.44	130.00	160.67	39.00	208.44	47.77	0.00
Aashiqui 2	262.58	150.00	185.69	40.00	262.58	76.89	0.00

$$SSLF = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\overline{Y}_j - \hat{Y}_j\right)^2 = 67402.17 \qquad df_{LF} = c-2 = 40-2 = 38 \qquad MSLF = \frac{67402.17}{38} = 1773.74$$

$$SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 = 3262.22 \qquad df_{PE} = n-c = 55-40 = 15 \qquad MSPE = \frac{3262.22}{15} = 217.48$$

$$F^* = \frac{MSLF}{MSPE} = \frac{1773.74}{217.48} = 8.156 \qquad F\left(0.95;38,15\right) = 2.211 \qquad P\text{-value} = 0.00004$$

There is strong evidence for lack-of-fit in this data. Note: typically this test is more useful when there are few groups, with multiple replicates at each group level. The test is widely conducted when response surfaces are fit to optimize processes.

Remedial Measures

Nonlinearity of Regression Function

Several options apply:

Quadratic Regression Function: $E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$ (Places a bend in the data)

Exponential Regression Function: $E\{Y\} = \beta_0 \beta_1^X$ (Allows for multiplicative increases)

Nonlinear Regression Function: Mathematical form typically generated from differential equations, with parameters to be estimated.

Nonconstant Error Variance

Often transformations can solve this problem. Another option is weighted least squares.

Nonindependent Error Terms

One option is to work with a model permitting correlated errors. Other options include working with differenced data or allowing for previously observed *Y* values as predictors.

Nonnormality of Errors

Non-normal errors and errors with non-constant variances tend to occur together. Some of the transformations used to stabilize variances often normalize errors as well. The Box-Cox transformation often can (but not necessarily) cure both problems.

Omission of Important Variables

When important predictors have been omitted, they can be added in the form of a multiple linear regression model (Chapter 6).

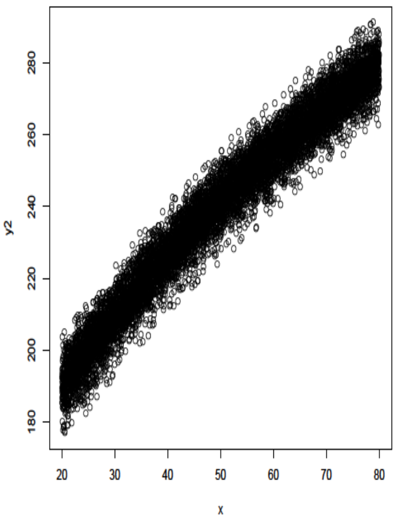
Outliers

When an outlier has been determined to be not due to data entry or recording error and should not be removed from model due to other reasons, indicator variables may be used to classify these observations away from others, or use of robust methods that decrease the effect of the outlying observation on the regression estimates.

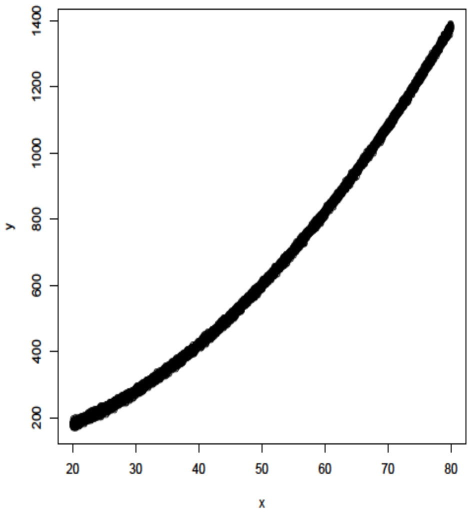
Transformations

Prototype plots and transformations of *Y* and/or *X* that are useful in linearizing the relation and/or stabilizing the variance are given below. Many times simply taking the logarithm of *Y* and/or *X* can solve the problems, as we will see below for the Bollywood data (where the distributions of both *Y* and *X* are highly skewed).

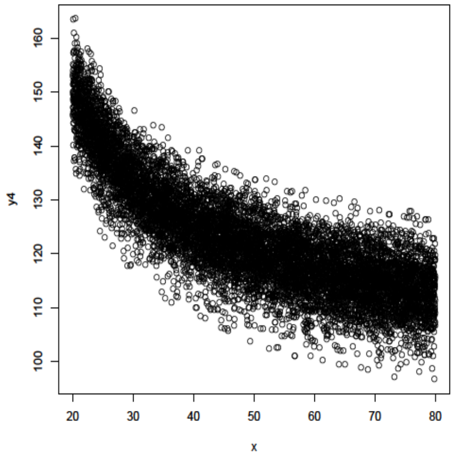
Transformations on *X* when the relation is nonlinear, with constant variance.



$X' = \sqrt{X}$ $X' = \ln(X)$

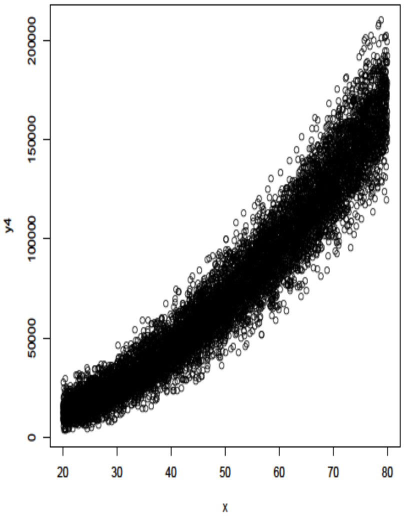


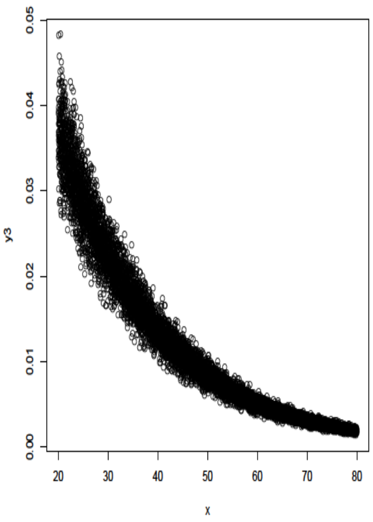
$X' = X^2$ $X' = e^X$

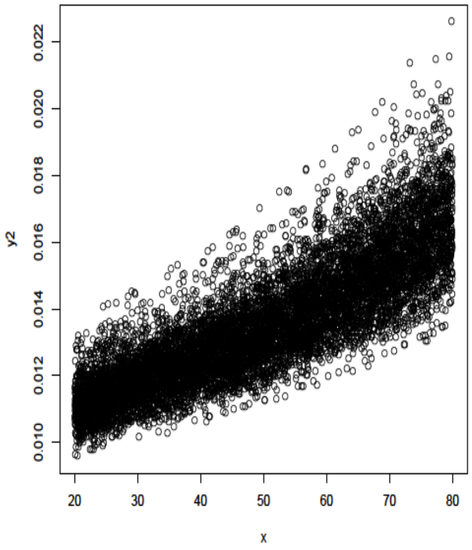


$X' = 1/X$ $X' = e^{-X}$

Transformations on *Y* when the relation is nonlinear, with non-constant variance.







Common transformations on Y: $Y' = \ln(Y)$ $\quad Y' = \sqrt{Y}$ $\quad Y' = \dfrac{1}{Y}$

Often (as in the Bollywood data), simultaneous transformations of *Y* and *X* will work.

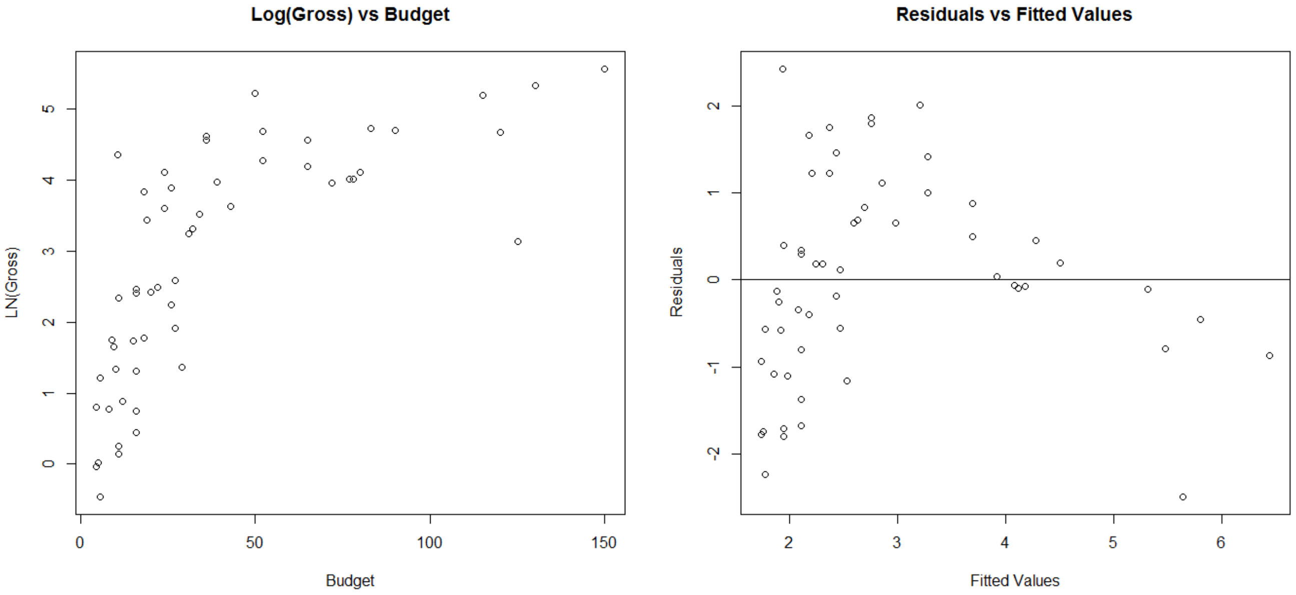
<u>Example: Bollywood Box Office Data</u>

First, we try taking logarithm of *Y*, leaving *X* untransformed.

```
################ LN transformation on Y
bbo.reg2 <- lm(log(Gross) ~ Budget)
summary(bbo.reg2)
e2 <- residuals(bbo.reg2)
yhat2 <- predict(bbo.reg2)
plot(yhat2,e2,main="Bollywood Regression - Residuals vs Fitted Values", xlab="Fitted Values", ylab="Residuals")
abline(h=0)
qqnorm(e2); qqline(e2)
shapiro.test(e2)
# install.packages("lmtest")
library(lmtest)
bptest(log(Gross) ~ Budget,studentize=FALSE)

##### R Output
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.59676    0.23114   6.908 6.34e-09 ***
Budget       0.03229    0.00434   7.439 8.86e-10 ***

Residual standard error: 1.166 on 53 degrees of freedom
Multiple R-squared:  0.5108,    Adjusted R-squared:  0.5016
F-statistic: 55.34 on 1 and 53 DF,  p-value: 8.859e-10
```

```
####  R Output Continued

> shapiro.test(e2)

        Shapiro-Wilk normality test

data:  e2
W = 0.9886, p-value = 0.8803

> bptest(log(Gross) ~ Budget,studentize=FALSE)

        Breusch-Pagan test

data:  log(Gross) ~ Budget
BP = 0.6532, df = 1, p-value = 0.419
```

Based on the Shapiro-Wilk test (Normality) and the Breusch-Pagan test (Constant Variance), the new model appears to be better. However, see the plot of *Y*’ versus *X*, and the residuals versus fitted values plot below. The relation is clearly not linear.

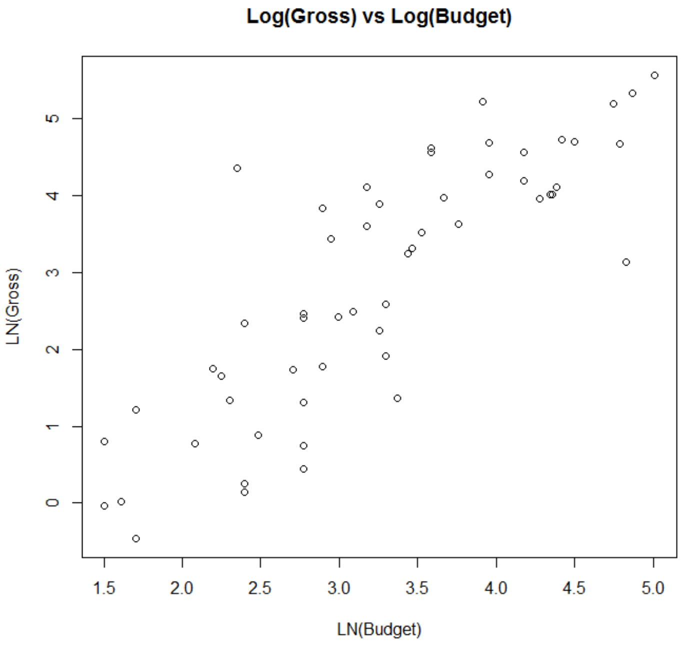


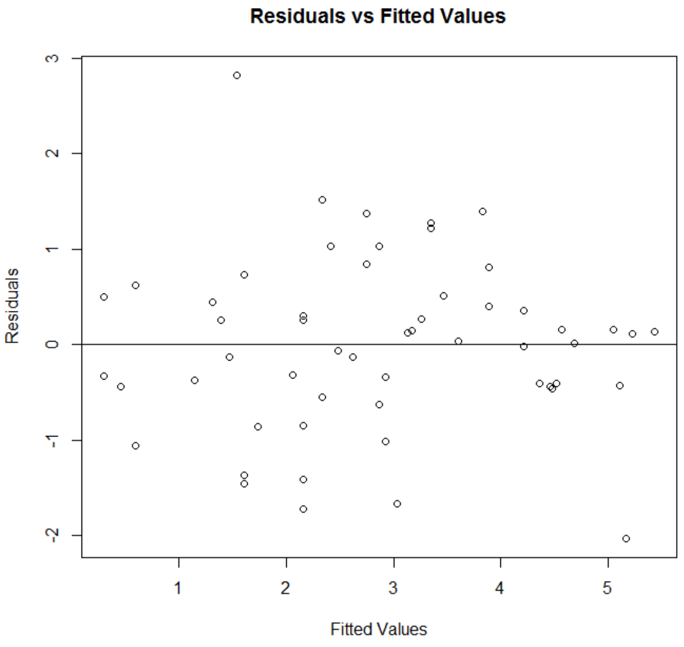
Now, consider the model where both Gross Revenues and Budget have been long transformed.

```
################ LN transformation on Y and X

bbo.reg3 <- lm(log(Gross) ~ log(Budget))
summary(bbo.reg3)
e3 <- residuals(bbo.reg3)
yhat3 <- predict(bbo.reg3)
par(mfrow=c(1,2))
plot(Budget,log(Gross),main="Log(Gross) vs Budget",xlab="Budget",ylab="LN(Gross)")
plot(yhat3,e3,main="Residuals vs Fitted Values", xlab="Fitted Values", ylab="Residuals")
abline(h=0)
qqnorm(e3); qqline(e3)
shapiro.test(e3)
library(lmtest)
bptest(log(Gross) ~ log(Budget),studentize=FALSE)
```

```
### R output

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.9038     0.4489  -4.241 8.95e-05 ***
log(Budget)   1.4645     0.1327  11.034 2.41e-15 ***

Residual standard error: 0.9181 on 53 degrees of freedom
Multiple R-squared:  0.6967,    Adjusted R-squared:  0.691
F-statistic: 121.7 on 1 and 53 DF,  p-value: 2.411e-15

> shapiro.test(e3)

        Shapiro-Wilk normality test

data:  e3
W = 0.98, p-value = 0.4866

> bptest(log(Gross) ~ log(Budget),studentize=FALSE)

        Breusch-Pagan test

data:  log(Gross) ~ log(Budget)
BP = 1.1056, df = 1, p-value = 0.293
```

Now, looking at the plot of transformed Y versus transformed X, and the residuals versus fitted values plot, we see that the model appears to meet the normality and constant variance assumptions.





For the model, with both Y and X log transformed, the interpretation of the slope is the **elasticity** between Y and X. As X increases 1%, Y increases by b_1%. For this data, as Budget increases 1%, Gross Revenue increases 1.4645%.

Box-Cox Transformations

Procedure to choose a transformation on *Y* (not *X*) with goal of choosing a power of *Y* that meets the model assumptions.

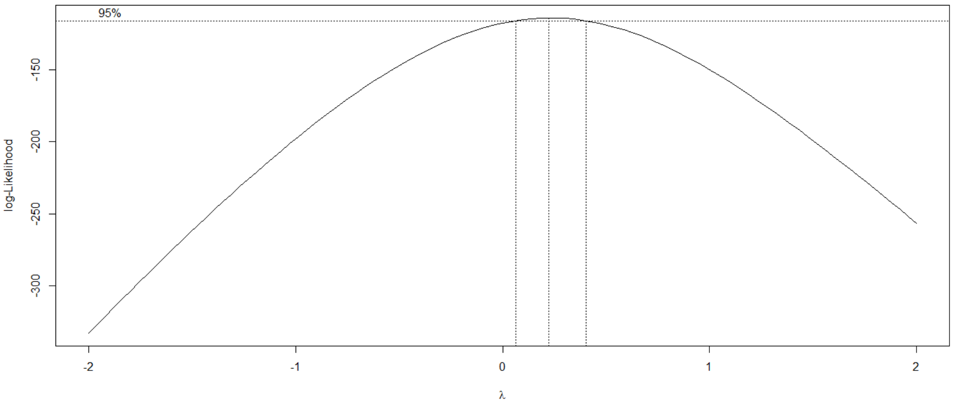
- Automatically selects a transformation from power family with goal of obtaining: normality, linearity, and constant variance (not always successful, but widely used)
- Goal: Fit model: $Y' = b_0 + b_1X + e$ for various power transformations on *Y*, and selecting transformation producing minimum SSE (maximum likelihood)
- Procedure: over a range of l from, say -2 to +2, obtain W_i and regress W_i on *X* (assuming all $Y_i > 0$, although adding constant won't affect shape or spread of *Y* distribution)
- When the power (λ) is 0, this implies a logarithmic transformation.

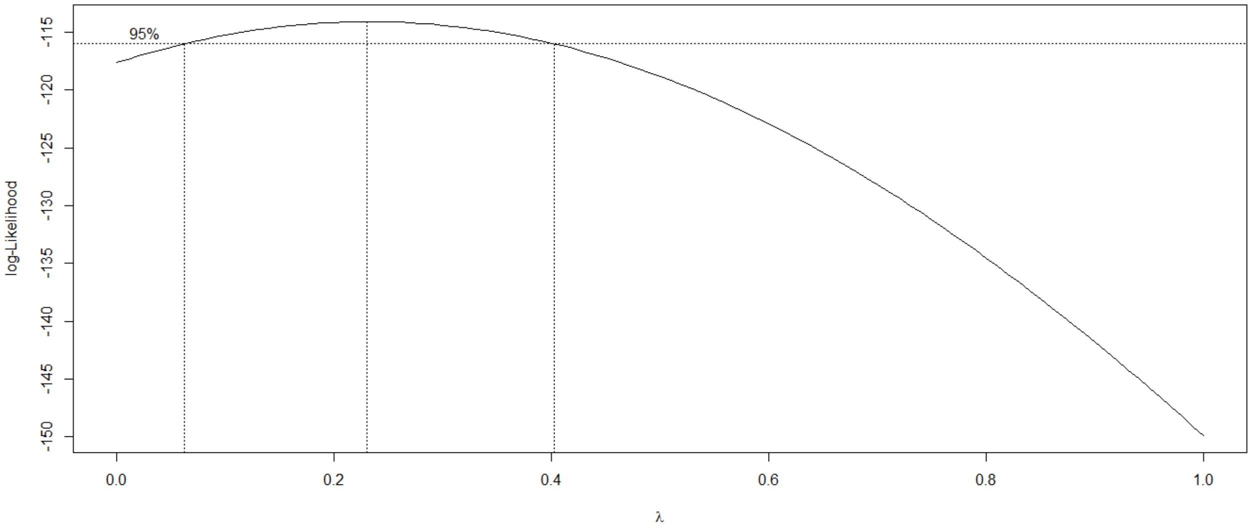
$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) & \lambda \neq 0 \\ K_2 \ln\left(Y_i\right) & \lambda = 0 \end{cases} \qquad K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \qquad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Example: Bollywood Box Office Data

The **boxcox** procedure has a default range for λ of -2 to 2. The second command "blows up" the plot to show the range better that contains the 95% Confidence Interval for λ.

```
### Box-Cox Transformation (must load MASS library first)

library(MASS)

bbo.reg4 <- lm(Gross ~ Budget)

boxcox(bbo.reg4,plotit=T)
boxcox(bbo.reg4,lambda=seq(0,1,0.01),plotit=T)
```



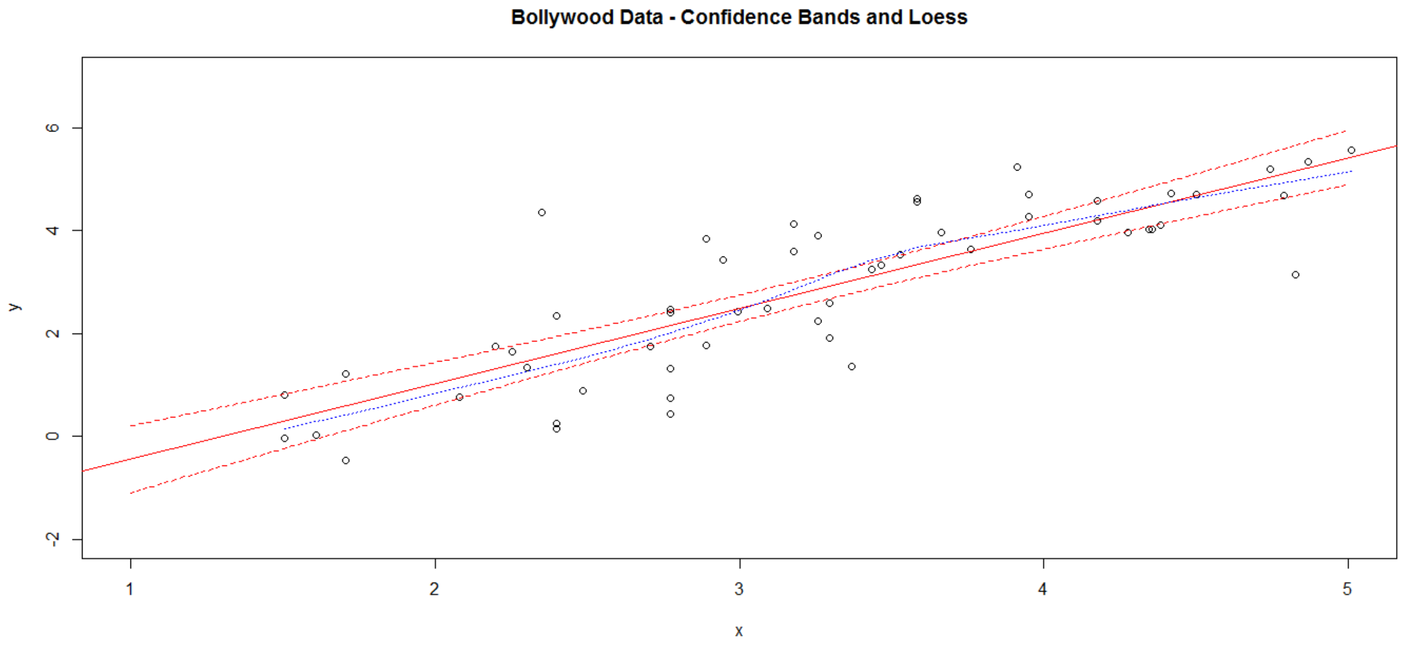


The procedure chooses a “quarter root” transformation for *Y*. We will not pursue that here, as we have seen that log transformations of *Y* and *X* work quite well.

Lowess (Smoothed) Plots

- Nonparametric method of obtaining a smooth plot of the regression relation between *Y* and *X*
- Fits regression in small neighborhoods around points along the regression line on the X axis
- Weights observations closer to the specific point higher than more distant points
- Re-weights after fitting, putting lower weights on larger residuals (in absolute value)
- Obtains fitted value for each point after “final” regression is fit
- Model is plotted along with linear fit, and confidence bands, linear fit is good if lowess lies within bands

```
#############  Loess  Plot
x <- log(Budget);  y <- log(Gross)
par(mfrow=c(1,1))
plot(x,y,xlim=c(1.0,5.0),ylim=c(-2,7),
main="Bollywood Data - Confidence Bands and Loess")
bbo.reg6 <- lm(y~x)
abline(bbo.reg6,col="red")
xh <- seq(1.0,5.0,0.01)
yhatci <-predict(bbo.reg6,list(x=xh),interval = c("confidence"),  level = 0.95,type="response")
lines(xh,yhatci[,2],col="red",lty=2)
lines(xh,yhatci[,3],col="red",lty=2)
lines(lowess(x,y),col="blue",lty=3)
```



Note that for the log transformed variables, the loess curve lies within the 95% Confidence Lines for the mean, confirming the linear fit is good for these data.

Chapter 4 – Simultaneous Inference and Other Topics

Joint Estimation of β_0 and β_1

We've obtained (1-α)100% confidence intervals for the slope and intercept parameters in Chapter 2. Now we'd like to construct a range of values (β_0, β_1) that we believe contains BOTH parameters with the same level of confidence. One way to do this is to construct each individual confidence interval at a higher level of confidence, namely:
(1-(α/2))100% confidence intervals for β_0 and β_1 seperately. The resulting ranges are called **Bonferroni Joint (Simultaneous) Confidence Intervals**.

Joint Confidence Level (1-α)100%	Individual Confidence Level (1-(α/2))100%
90%	95%
95%	97.5%
99%	99.5%

The resulting simultaneous confidence intervals, with a joint confidence level of at least (1-α)100% are:

$$b_0 \pm Bs\{b_0\} \qquad b_1 \pm Bs\{b_1\} \qquad B = t(1-(\alpha/4); n-2)$$

Example: Bollywood Box Office Data

Simultaneous 95% Confidence Intervals for β_0, β_1 for the log transformed model:

$$t\left(1-.05/4; 55-2\right) = t\left(.9875; 53\right) = 2.3069$$

$$b_1 = 1.4645 \quad s\left\{b_1\right\} = 0.1327 \quad b_0 = -1.9038 \quad s\left\{b_0\right\} = 0.4489$$

Simultaneous 95% CIs:

$$\beta_1 : 1.4645 \pm 2.3069(0.1327) \equiv 1.4645 \pm 0.3061 \equiv \left(1.1584, 1.7706\right)$$

$$\beta_0 : -1.9038 \pm 2.3069(0.4489) \equiv -1.9038 \pm 1.0356 \equiv \left(-2.9394, -0.8682\right)$$

Simultaneous Estimation of Mean Responses

Case 1: Simultaneous (1-α)100% Bounds for the Regression Line (Working-Hotelling's Approach)

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \qquad W = \sqrt{2F(1-\alpha; 2, n-2)}$$

Case 2: Simultaneous (1-α)100% Bounds at g Specific X Levels (Bonferroni's Approach)

$$\hat{Y}_h \pm Bs\{\hat{Y}_h\} \qquad B = t(1-(\alpha/2g); n-2)$$

Simultaneous Prediction Intervals for New Observations

Sometimes we wish to obtain simultaneous prediction intervals for g new outcomes.

Scheffe's Method:

$$\hat{Y}_h \pm Ss\{pred\} \qquad S = \sqrt{gF(1-\alpha; g, n-2)}$$

where $s\{pred\} = \sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_h-\overline{X})^2}{\sum(X_i-\overline{X})^2}\right)}$ is the estimated standard error of the prediction.

Bonferroni's Method:

$$\hat{Y}_h \pm Bs\{pred\} \qquad B = t(1-\alpha/(2g); n-2)$$

Both S and B can be computed before observing the data, and the smallest of the two should be used.

Regression Through the Origin

Sometimes it is desirable to have the mean response be 0 when the predictor variable is 0 (this is not the same as saying Y must be 0 when X is 0). Even though it can cause extra problems, it is an interesting special case of the simple regression model, and is also used in various tests/procedures.

$$Y_i = \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim NID(0, \sigma^2)$$

We obtain the least squares estimate of β_1 (which also happens to be maximum likelihood) as follows:

$$Q = \sum \varepsilon_i^2 = \sum (Y_i - \beta_1 X_i)^2 \quad \Rightarrow \quad \frac{\partial Q}{\partial \beta_1} = 2\sum (Y_i - \beta_1 X_i)(-X_i)$$

$$\Rightarrow \quad -2\left[\sum X_i Y_i - b_1 \sum X_i^2\right] = 0 \quad \Rightarrow \quad \sum X_i Y_i = b_1 \sum X_i^2 \quad \Rightarrow \quad b_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

The fitted values and residuals (which no longer necessarily sum to 0) are:

$$\hat{Y}_i = b_1 X_i \qquad e_i = Y_i - \hat{Y}_i$$

An unbiased estimate of the error variance σ^2 is:

$$\boxed{s^2 = MSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1} = \frac{\sum e_i^2}{n-1}}$$

Note that we have only estimated one parameter in this regression function.

Note that the following are linear functions of $Y_1, \ldots, Y_n$:

$$\boxed{\begin{aligned}
&b_1 = \frac{\sum X_i Y_i}{\sum X_i^2} = \sum \frac{X_i}{\sum X_i^2} Y_i = \sum a_i Y_i \qquad a_i = \frac{X_i}{\sum X_i^2} \\
&\Rightarrow \quad E\{b_1\} = E\left\{\sum a_i Y_i\right\} = \sum a_i E\{Y_i\} = \sum a_i \beta_1 X_i = \sum \frac{X_i}{\sum X_i^2} \beta_1 X_i = \beta_1 \frac{\sum X_i^2}{\sum X_i^2} = \beta_1 \\
&\Rightarrow \quad \sigma^2\{b_1\} = \sigma^2\{\sum a_i Y_i\} = \sum a_i^2 \sigma^2\{Y_i\} = \sum \left[\frac{X_i}{\sum X_i^2}\right]^2 \sigma^2 = \sigma^2 \frac{\sum X_i^2}{\left[\sum X_i^2\right]^2} = \frac{\sigma^2}{\sum X_i^2}
\end{aligned}}$$

Thus, b_1 is an unbiased estimate of the slope parameter β_1, and its variance (and thus standard error) can be estimated as follows:

$$\boxed{s^2\{b_1\} = \frac{s^2}{\sum X_i^2} = \frac{MSE}{\sum X_i^2} \qquad \Rightarrow \qquad s\{b_1\} = \sqrt{\frac{MSE}{\sum X_i^2}}}$$

This can be used to construct confidence intervals for or conduct tests regarding β_1.

<u>Example: Bollywood Box Office Data</u>

For the original (non-transformed) data, we obtain the following quantities and estimates:

$$\boxed{\begin{aligned}
&\sum X_i Y_i = 190927.5 \quad \sum X_i^2 = 155976.5 \quad \Rightarrow \quad b_1 = \frac{190927.5}{155976.5} = 1.2241 \\
&\Rightarrow \quad \hat{Y}_i = 1.2241 X_i \quad \Rightarrow \quad SSE = \sum \left(Y_i - \hat{Y}_i\right)^2 = 70761.64 \quad \Rightarrow \quad s^2 = MSE = \frac{70761.64}{55-1} = 1310.40 \\
&\Rightarrow \quad s\left\{b_1\right\} = \sqrt{\frac{1310.40}{155976.5}} = 0.0917 \qquad t\left(0.975; 54\right) = 2.0049 \\
&95\%\text{CI for } \beta_1: \quad 1.2241 \pm 2.0049(.0917) \equiv 1.2241 \pm 0.1838 \equiv (1.0403, 1.4079)
\end{aligned}}$$

The mean response at X_h for this model is: $E\{Y_h\} = \beta_1 X_h$ and its estimate is $\hat{Y}_h = b_1 X_h$, with mean and variance:

$$\boxed{\begin{array}{l} E\{\hat{Y}_h\} = E\{b_1 X_h\} = X_h E\{b_1\} = X_h \beta_1 \\ \sigma^2\{\hat{Y}_h\} = \sigma^2\{b_1 X_h\} = X_h^2 \sigma^2\{b_1\} = \sigma^2 \dfrac{X_h^2}{\sum X_i^2} \quad \Rightarrow \quad s^2\{\hat{Y}_h\} = MSE \dfrac{X_h^2}{\sum X_i^2} \end{array}}$$

This can be used to obtain a confidence interval for the mean response when $X=X_h$.

The estimated prediction error for a new observation at $X=X_h$ is:

$$\boxed{s^2\{pred\} = s^2\{Y_{h(new)} - \hat{Y}_h\} = s^2\{Y_{h(new)}\} + s^2\{\hat{Y}_h\} = s^2 + \frac{s^2 X_h^2}{\sum X_i^2} = MSE\left[1 + \frac{X_h^2}{\sum X_i^2}\right]}$$

This can be used to obtain a prediction interval for a new observation at this level of *X*.

Comments Regarding Regression Through the Origin:

- You should test whether the true intercept is 0 when $X=0$ before proceeding.
- Remember the notion of constant variance. If you are forcing *Y* to be 0 when *X* is 0, you are saying that the variance of *Y* at $X=0$ is 0.
- If $X=0$ is not an important value of *X* in practice, there is no reason to put this constraint into the model.
- r^2 is no longer constrained to be between 0 and 1, the error sum of squares from the regression can exceed the total corrected sum of squares. The coefficient of determination loses its interpretation of being the proportion of variation in *Y* that is "explained" by *X*.

Effects of Measurement Errors

Measurement errors can take on one of three forms. Two of the three forms cause no major problems, one does.

Measurement Errors in *Y*

This causes no problems as the measurement error in *Y* becomes part of the random error term, which represents effects of many unobservable quantities. This is the case as long as the random errors are independent, unbiased, and not correlated with the level of *X*.

Measurement Errors in *X*

Problems do arise when the measurement of the predictor variable is measured with error. This is particularly the case when the observed (reported) X_i^* level is the true level X_i plus a random error term. In this case the random error terms are not independent of the reported levels of the predictor variable, causing the estimated regression coefficients to be biased and not consistent. See textbook for a mathematical development. Certain methods have been developed for particular forms of measurement error. See *Measurement Error Models* by W.A. Fuller for a theoretical treatment of the problem or *Applied Regression Analysis* by J.O. Rawlings, S.G. Pantula, and D.A. Dickey for a brief description.

Measurement Errors with Fixed Observed *X* Levels

When working in engineering and behavioral settings, a factor such as temperature may be set by controlling a level on a thermostat. That is, you may set an oven's cooking temperature at 300, 350, 400, etc. When this is the case and the actual physical temperatures vary at random around these actual observed temperatures, the least squares estimators are unbiased. Further when normality and constant variance assumptions are applied to the "new errors" that reflect the random actual temperatures, the usual tests and confidence intervals can be applied.

Inverse Predictions

Sometimes after we fit (or calibrate) a regression model, we can observe *Y* values and wish to predict the *X* levels that generated the outcomes. Let $Y_{\text{h(new)}}$ represent a new value of *Y* we have just observed, or a desired level of *Y* we wish to observe. In neither case, was this observation part of the sample. We wish to predict the *X* level that led to our observation, or the *X* level that will lead to our desired level. Consider the estimated regression function:

$$\hat{Y} = b_0 + b_1 X$$

Now we observe a new outcome $Y_{\text{h(new)}}$ and wish to predict the *X* value corresponding to it, we can use an estimator that solves the previous equation for *X*. The estimator and its (approximate) estimated standard error are:

$$\boxed{\hat{X}_{h(new)} = \frac{Y_{h(new)} - b_0}{b_1} \qquad s\{predX\} = \sqrt{\frac{MSE}{b_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{X}_{h(new)} - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]}}$$

Then, an approximate (1-α)100% Prediction Interval for $X_{\text{h(new)}}$ is:

$$\boxed{\hat{X}_{h(new)} \pm t(1-\alpha/2; n-2) s\{predX\}}$$

<u>Example: Bollywood Box Office Data</u>

Suppose a new movie was released, and we observed that the log of its box-office gross was $Y'_{\text{h(new)}}$ = ln(30) = 3.4012. We want to obtain a 95% Prediction Interval for $X'_{\text{h(new)}}$

$$b_0 = -1.9038 \quad b_1 = 1.4565 \quad MSE = 0.8429 \quad \overline{X'} = 3.2509 \quad \sum\left(X_i' - \overline{X'}\right)^2 = 47.8452$$

$$\Rightarrow \quad \hat{X}'_{h(new)} = \frac{3.4012 - (-1.9038)}{1.4565} = 3.6423$$

$$s\{predX\} = \sqrt{\frac{0.8429}{1.4565^2}\left[1 + \frac{1}{55} + \frac{(3.6423 - 3.2509)^2}{47.8482}\right]} = \sqrt{0.3973(1.0214)} = 0.6370$$

$$\text{95\%PI for } X'_{h(new)}: \quad 3.6423 \pm 2.0057(.6370) \equiv 3.6423 \pm 1.2777 \equiv (2.3646, 4.9200)$$

$$\text{Back-transforming to obtain PI for Budget in original units: } \left(e^{2.3646}, e^{4.9200}\right) \equiv (10.6398, 137.0026)$$

Choosing *X* Levels

Issues arising involving choices of *X* levels and sample sizes include:

- The “range” of *X* values of interest to experimenter
- The goal of research: inference concerning the slope, predicting future outcomes, understanding the shape of the relationship (linear, curved,…)
- The cost of collecting measurements

Note that all of our estimated standard errors depend on the number of observations and the spacing of *X* levels. The more spread out, the smaller the standard errors, generally. However, if we wish to truly understand the shape of the response curve, we must space the observations throughout the set of *X* values. See quote by D.R. Cox on page 171 of textbook.