# MA 214: Applied Statistics

## Instructor: Ashis Gangopadhyay

**Regression Analysis**
**Part II**



## Where We've Been

- Discussed the meaning of Regression Model
- Introduced model fitting via least square
- Learnt how and why we carry out statistical inference of model parameters.
- Discussed how to use the fitted model for prediction.

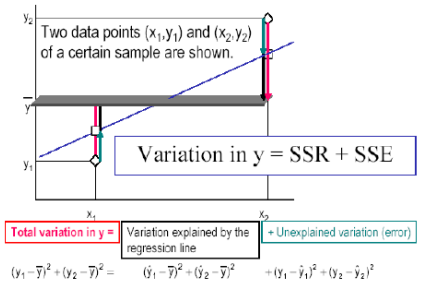


## Where We're Going

- To quantify how well the model fits the data.
- To judge if the fitted regression model is an adequate description of data (model adequacy check).
- To identify the remedial measures if the model turns out to be not a good fit to the data.

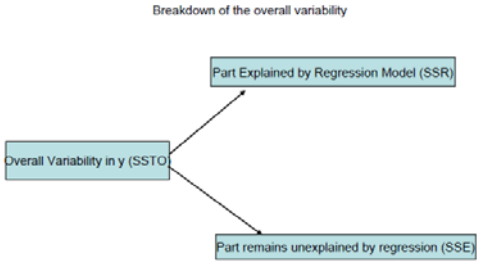


## Analysis of Variance Approach to Regression






## Analysis of Variance Approach to Regression






## Analysis of Variance Approach to Regression

**Data:** $\{(x_1,y_1),\ldots,(x_n,y_n)\}$

A measure of overall variability of the observations on the dependent variable:

Total Sum of Squares (SSTO) = $\sum_{i=1}^{n}(y_i-\bar{y})^2$

Residual variability after fitting the regression model:

Error Sum of Squares (SSE) = $\sum_{i=1}^{n}(y_i-\hat{y})^2=\sum_{i=1}^{n}e_i^2$

Thus, the part of the total variability explained by regression:

Regression Sum of Squares (SSR) = *SSTO – SSE*

## Analysis of Variance Approach to Regression

**Analysis of Variance Table for Simple Linear Regression**

| Source of Variability | SS | DF | MS | F-Statistic |
|---|---|---|---|---|
| Regression | SSR | 1 | MSR | $F = \frac{MSR}{MSE}$ |
| Error | SSE | n-2 | MSE | |
| Total | SSTO | n-1 | | |

**Abbreviations:**

SS = Sum of Squares
SSR = Regression Sum of Squares
SSE = Error Sum of Squares
DF = Degrees of Freedom
MS = Mean Squares = SS/DF
MSR = Mean Squares Regression
MSE = Mean Square Error



## Analysis of Variance Approach to Regression

**Computational Formulas for Sum of Squares**

$$SSE = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 \left[ \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right]$$

$$SSTO = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

**Also note the relationship:** *SSTO = SSR + SSE*



## Salary Example

$$SSE = 88021 - (-0.675)931 - (14.229)5960 = 3845$$

$$SSR = (14.229^2)\left[422 - \frac{(66)^2}{12}\right] = 11945$$

$$SSTO = 88021 - \frac{(931)^2}{12} = 15790$$

**Note the relationship:** *SSTO = SSR + SSE*



## Salary Example

**ANOVA**

| Source | SS | DF | MS | F |
|---|---|---|---|---|
| Regression | 11945 | 1 | 11945 | 31.06 |
| Error | 3845 | 10 | 384.5 | |
| Total | 15790 | | | |



## Coefficient of Determination ($R^2$)

$$R^2 = \frac{SSR}{SSTO} = \frac{\text{Variation in Y explained by regression}}{\text{Total observed variation in Y}}$$

=Proportion of total variability explained by regression

**Note:** $0 \le R^2 \le 1$

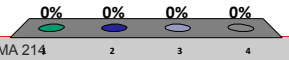
**Example:**

$$R^2 = \frac{SSR}{SSTO} = \frac{11945}{15790} = 0.756$$



## In this example, $R^2$ =0.756 means…

1. 75.6% of the variability of *Experience* is explained by the model.
2. 75.6% of the variability of *Salary* is explained by the model.
3. 75.6% of the time covariate *Experience* can predict *Salary* correctly.
4. 24.4% of the variability of *Salary* is explained by the model.



## Coefficient of Determination ($R^2$)

**1.1.1 How to interpret $R^2$ value?**

| | |
|---|---|
| *** | $R^2$ "large": Covariate(s) explains almost all of the variability of the dependent variable. Thus the model is a very good fit to the data. |
| ** | $R^2$ "moderate": Covariate(s) explains a part of the total variability of the dependent variable. Need to develop better model possibly incorporating additional covariates. |
| * | $R^2$ "low": Covariate(s) explains only a very small part of the total variability, suggesting a reinvestigation of the problem with a new set of covariates. |



## Coefficient of Correlation

**Definition 1:**

$$r = \sqrt{R^2} \text{ if } \hat{\beta}_1 \geq 0; \text{ or } -\sqrt{R^2} \text{ if } \hat{\beta}_1 < 0$$

**Definition 2:**

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)^2}}$$

- Correlation coefficient ($r$) is a measure of the strength of the linear relationship between the dependent variable and the independent variable in a linear regression model.



## Coefficient of Correlation

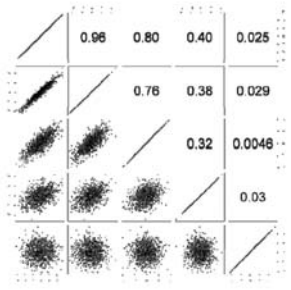
**1.2.1 Interpretation of correlation coefficient $r$**

- $-1 \leq r \leq 1$
- If the two variables are strongly positively related, then the value of the correlation coefficient $r$ is close to 1.
- If the variables are strongly negatively related, then the value of the correlation coefficient $r$ is close to -1.
- Correlation coefficient $r$ close to 0 suggests a lack of a linear relationship between the variables.



## Coefficient of Correlation

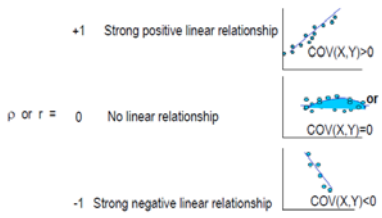
**Correlation**






## Coefficient of Correlation

**The coefficient of correlation**






## 1.2 Coefficient of Correlation

**Example 1: Compute the correlation coefficient $r$ for**

Since $\hat{\beta}_1 = 14.229 > 0,$ we have that

$$r = \sqrt{R^2} = \sqrt{0.756} = 0.869,$$

suggesting a strong linear relationship between years of experience and annual salary of graduates from this school!

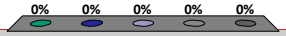
**If you believe strongly that the more hours students work at a job, lower their GPA would be, then which correlation you would expect to find?**

✓1. -0.7
2. 0.02
3. 0.5
4. 0.7




**Let's play the correlation game ☺. The correct answer is…..**

1. A
2. B
3. C
4. D
5. E






## Coefficient of Correlation

**Some things to remember!**

- Note that $|r| > R^2$ (unless $R^2$ is 0 or 1, when equality holds). Thus the use of correlation coefficient as a measure of performance of a linear regression model may be misleading, as it may provide a false sense of overall fit of the model to that data. $R^2$ is a better measure of goodness of fit of a regression model, and it has a clear interpretation in terms of the proportion of total variability explained by the covariates in the model.



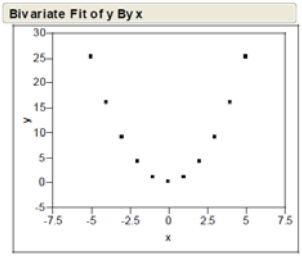
## 1.2 Coefficient of Correlation

**Some things to remember!**

- It is important to note that if $r = 0$, it only means that there is no linear relationship between covariate $X$ and dependent variable $Y$; however, that doesn't necessarily imply that there is no relationship between these variables! Consider the following example, where $Y$ is perfectly related to $X$ according to the functional relationship $Y = X^2$; however, it is easy to check that $r = 0$.



## Coefficient of Correlation

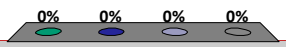



$Y = X^2$, but $r = 0$



**A researcher found that r = +.92 between the high temperature of the day and the number of ice cream cones sold at Boston Common. This result tells us that**

1. High temperature causes people to buy ice cream.
2. Buying ice cream causes temperature to go up.
3. Some extraneous variable causes both temperature and ice cream sale to go up.
4. ✓ Temperature and ice cream sales have a strong positive linear relationship.



## Correlation does not imply Causation

- Lawlor DA, Davey Smith G, Ebrahim S (June 2004). "Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology?". *Int J Epidemiol* **33** (3): 464–7
- Numerous epidemiological studies showed that women who were taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading doctors to propose that HRT was protective against CHD.



## Correlation does not imply Causation

- However, a careful analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups, with better than average diet and exercise regimens.
- The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socioeconomic status), rather than cause and effect as had been supposed.
- As a matter of fact, randomized controlled trials showed that HRT caused a small but statistically significant increase in risk of CHD.



## F-test under ANOVA for simple linear regression model

**The F-statistic in the ANOVA table tests the following hypothesis:**

$H_0$: $\beta_1 = 0$ vs. $H_a$: $\beta_1 \neq 0$

**Test Statistic:** $F = \dfrac{MSR}{MSE}$

**Decision Rule:** Reject $H_0$ if $F > F_\alpha$, where $F_\alpha$ is based on $(1, n-2)$ degrees of freedom



## F-test under ANOVA for simple linear regression model

**Example 1: Carry out the F-test at 5% level of significance in the**

$H_0$: $\beta_1 = 0$ vs. $H_a$: $\beta_1 \neq 0$

$F = 31.06$

$F_{0.05} = 4.965$ with (1,10) degrees of freedom

Since $F > F_{0.05}$, reject the null hypothesis.



## F-test under ANOVA for simple linear regression model

**Remark:** It should be noted that earlier we discussed an alternative approach to test the hypothesis $H_0$: $\beta_1 = 0$ which is based on the t-distribution. Referring to the result, we

$$|t| > t_{\frac{\alpha}{2}} \text{ (with } n-2 = 10 \text{ df)}$$
$$\Leftrightarrow 5.57 > 2.228$$
$$\Leftrightarrow (5.57)^2 > (2.228)^2$$
$$\Leftrightarrow 31.06 > 4.965$$
$$\Leftrightarrow F > F_\alpha$$

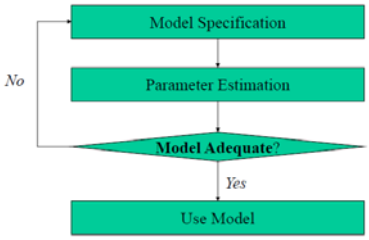


## F-test under ANOVA for simple linear regression model

- Thus both the t-test and the F-test are equivalent! So why do we need two different ways of testing the same hypothesis!
- The answer is that this equivalence is only true for simple linear regression with one covariate in the model. In multiple regression models (more than one covariate), which we will be discussing later in this course, the F-test in ANOVA plays a crucial role in understanding the validity of such models.

## Statistical Modeling

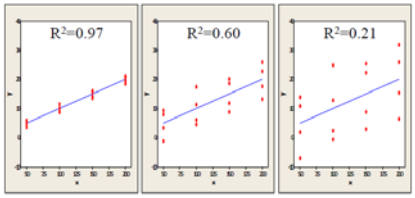
Statistical Modeling Procedure






## F-test under ANOVA for simple linear regression model

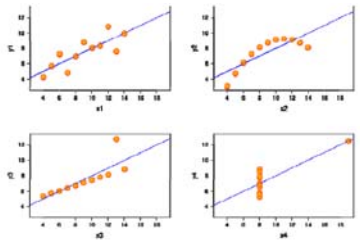
R-Square & Prediction Power




**All three cases have significant slopes and the same regression line function.**



## F-test under ANOVA for simple linear regression model



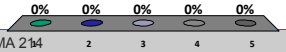
**All four cases have significant slopes and the same R-square (0.7), and regression line ($y = 3 + 0.5x$).**

Source: Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17–21.



## Model is a good fit to the data in…..

1. All four cases
2. ✔ Only for case 1
3. Only for case 2
4. Only for case 3
5. Only for case 4

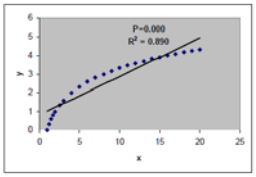





## F-test under ANOVA for simple linear regression model

### Caution

- Regression analysis is perhaps the most widely used statistical technique, and probably the most widely misused.
- Just because you *can* fit a linear model to a set of data, does not mean you should.






## Verification of model adequacy

Let's consider the simple linear regression model given by

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

where the random errors $\{\varepsilon_i, i = 1, 2, ..., n\}$ are **independent** random variables that are **normally** distributed with mean 0 and variance $\sigma^2$.

## Verification of model adequacy

**Note that the underling assumptions of the model are:**

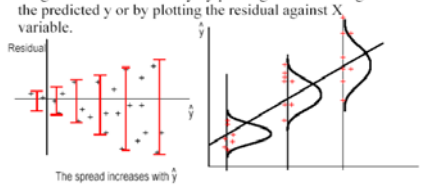
- **Linearity:** The dependent variable *Y* and covariate *X* are linearly related.
- **Normality:** Random errors are normally distributed.
- **Constant variance:** Error variance is constant across all values of the covariate *X*.
- **Independence:** Errors associated with the observations are mutually independent.



## Verification of model adequacy

Heteroscedasticity

- When the requirement of a constant variance is violated we have a condition of heteroscedasticity.
- Diagnose heteroscedasticity by plotting the residual against the predicted y or by plotting the residual against X variable.






## Verification of model adequacy

**1.4.1 Residual Analysis**

- Residual Analysis is a technique that allows us to detect violations of the assumptions of a regression model by examining the residuals.
- Residual Plot
  - Plot of residuals against the corresponding values of the covariate, i.e., plot of $\{(e_i, x_i), i = 1, 2, ..., n\}$

**OR**

  - Plot of residuals against fitted (predicted) values; i.e., plot of $\{(e_i, \hat{y}_i), i = 1, 2, ..., n\}$



## Verification of model adequacy

**1.4.1 Residual Analysis**

- **Detection of nonlinearity and heteroscedasticity**
  Violations of linearity and constant variance assumptions can be detected by examining the residual plots.
- **Scenario 1:** In a residual plot, the residuals are randomly scattered around the zero level within two parallel bands.

  **No apparent violation of either linearity or constant variance assumptions.**



## Verification of model adequacy

**1.4.1 Residual Analysis**

- **Detection of nonlinearity and heteroscedasticity**
  Violations of linearity and constant variance assumptions can be detected by examining the residual plots.
- **Scenario 2:** In a residual plot, the residuals exhibit distinct nonlinear pattern.

  **Linearity assumption is violated, and the dependent variable *Y* is related to the covariate *X* via a nonlinear model.**



## Verification of model adequacy

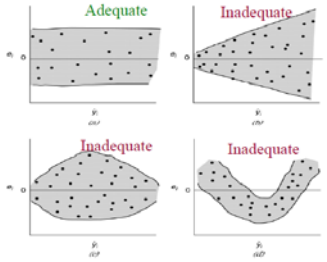
**1.4.1 Residual Analysis**

- **Detection of nonlinearity and heteroscedasticity**
  Violations of linearity and constant variance assumptions can be detected by examining the residual plots.
- **Scenario 3:** The residual plot fans out or fans in.

  **Variance is not constant. If the residual plot fans out, then the variance is increasing as a function of the covariate *X*; if the plot fans in, then the variance is a decreasing function of *X*.**

## Verification of model adequacy

Residual Plot: Residual vs. Fitted Value




Check constant variance and linearity, and **look for potential outliers**



## Verification of model adequacy

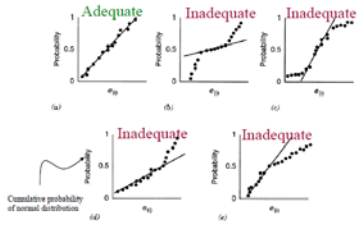
**Residual Analysis**

- **Detection of nonnormality**
  Nonnormality of the error term in the regression model can be detected easily by examining the normal probability plot of the residuals. If the plot is approximately a straight line, then the normality assumption is true, otherwise the assumption is false.



## Verification of model adequacy

Residual Plot: Normal Probability Plot



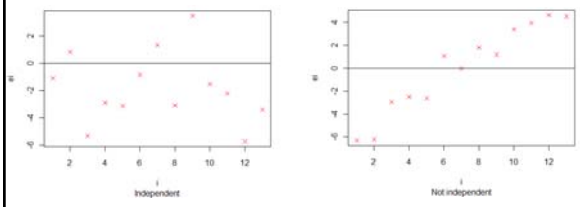

Check the normality assumption



## Verification of model adequacy

- Residual Analysis to verify Independence

  Plot the residuals against the order of data collection.






## For the job training data, the residual analysis shows that the fitted model is…

1. Adequate
2. Not adequate due to heteroscedasticity.
3. ✓ Not adequate due to nonlinearity.
4. Not adequate due to nonnormality

0% 0% 0% 0%
1 2 3 4



## For the blood plasma data, the resiidual analysis shows that the fitted model is….

1. Adequate
2. Not adequate due to heteroscedasticity.
3. Not adequate due to nonlinearity.
4. ✓ Both 2 and 3.

0% 0% 0% 0%
1 2 3 4

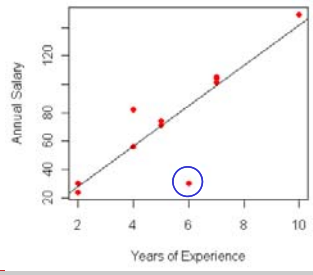
## Detection of outliers

Frequently, in regression analysis, the data may contain observation(s) that are outlying or extreme. These observations are called outliers. Outliers may have a dramatic effect on regression analysis.



## 1.6 Detection of outliers

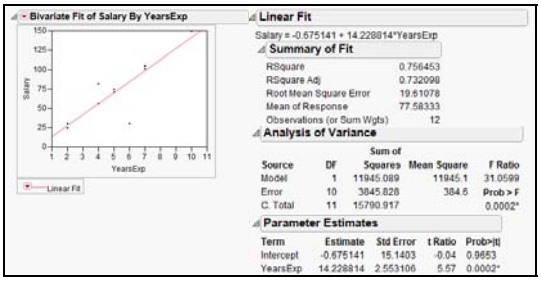
Influential Observation






## 1.6 Detection of outliers

**Regression Output**



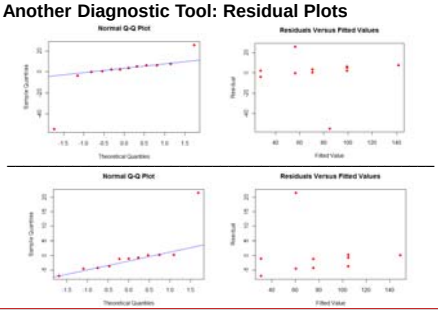



## 1.6 Detection of outliers

**Another Diagnostic Tool: Residual Plots**

**Original Model**

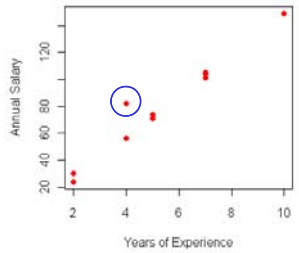
**Revised Model**






## 1.6 Detection of outliers

Another Influential Observation?



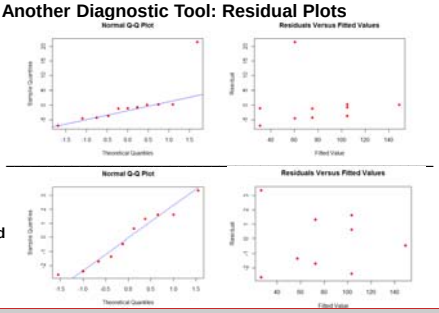



## 1.6 Detection of outliers

**Another Diagnostic Tool: Residual Plots**

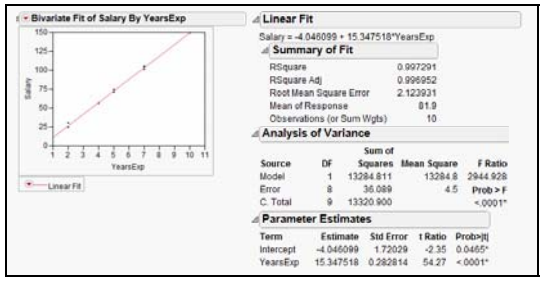
**Revised Model**

**Further Revised Model**



## 1.6 Detection of outliers

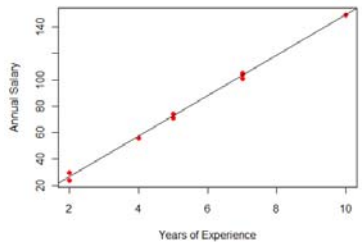
**Re-estimate parameters**






## 1.6 Detection of outliers

Final Revised Model






## Detection of outliers

**Cook's distance**

- It is a measure of influence of a data point; i.e., the effect that omitting an observation has on the estimated regression parameters.
- *Rule of thumb:* Observations with cook's distance $D_i >$ 2 have high influence. You may also be concerned about any observation that has $D_i < 2$, but has a much bigger $D_i$ than other observations.



## Detection of outliers

**How to deal with possible outliers in the data?**

Remove the observations suspected as outliers from the data. Do the estimates of the parameters change when observations are deleted?

- *NO.* Then proceed with the observations included in the data. Study the observations to see if anything can be learned.
- *YES.* Then, is there reason to believe that the observation belongs to a population other than the one being investigated?



## Detection of outliers

**How to deal with possible outliers in the data?**

Remove the observations suspected as outliers from the data. Do the estimates of the parameters change when observations are deleted?

***YES***. Then, is there reason to believe that the observation belongs to a population other than the one being investigated?

- *YES.* Delete the observation and proceed.
- *NO.* Then, does the observation have an unusually distant independent variable?



## Detection of outliers

**1.6.2 How to deal with possible outliers in the data?**

Remove the observations suspected as outliers from the data. Do the estimates of the parameters change when observations are deleted?

***YES***. Then, is there reason to believe that the observation belongs to a population other than the one being investigated?
***NO***. Then, does the observation have an unusually distant independent variable?

- *YES.* Delete the case and proceed.
- *NO.* Not much can be said. More data needed to resolve the question.

## Remedial Measures

It may be possible to fix some of the violations of the assumptions of the linear regression model by applying certain transformations to the data. These transformations can be applied either to the dependent variable *Y* or to the independent variable *X* or both.

**Most common transformations**

| Transformations | on dep. var. | on ind. var. |
|---|---|---|
| Square root | $\sqrt{Y}$ | $\sqrt{X}$ |
| Log | $\log(Y)$ | $\log(X)$ |
| Inverse | $1/Y$ | $1/X$ |



## Remedial Measures

**Some remarks regarding these transformations:**

- These classes of transformations are generally called *power transformations*. The transformations are written in the increasing order of their "strength." Thus, square root transformation is the weakest of the group, logarithmic transformation is the second strongest, etc. It is important to note that we need to apply the weakest transformation that gets the job done, and it is a bad idea to apply a transformation stronger than necessary.



## Remedial Measures

**Some remarks regarding these transformations:**

- Unfortunately, it is difficult to guess the correct transformation that may fix the problem associated with the model. Thus we need to try a range of transformations, and check the corresponding residuals to verify if the transformation has been successful in remedying the problem. However, there are some rule of thumbs that one can use in deciding a candidate transformation that is most likely to be successful.



## Remedial Measures

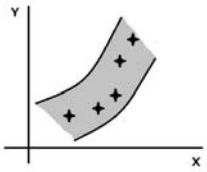
**Some remarks regarding these transformations:**

- It is important to note that although the transformation technique can often remedy the problems associated with linear regression, there are situations where suitable power transformations may not be available, and more complex techniques may be needed to remedy the problem.



## Remedial Measures

**Selection of transformation to remedy non-linearity: If the scatterplot looks like**

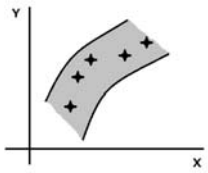



**Then try transforming Y first.**



## Remedial Measures

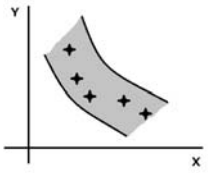
**Selection of transformation to remedy non-linearity: If the scatterplot looks like**




**Then try transforming X first.**

## Remedial Measures

**Selection of transformation to remedy non-linearity: If the scatterplot looks like**




**Then try transforming X and/or Y, but choose the variable with wider range first.**



## Remedial Measures

**Selection of transformation to remedy non-constant variance**

- Note that non-constant variance is a property of the dependent variable *Y*, as it is the conditional variance of *Y* given *X* = *x* that is not a constant. This suggests that the remedial measure must involve a transformation of the dependent variable *Y*. Frequently either $\sqrt{Y}$ or $\log(Y)$ transformations are effective in fixing non-constant variance.



## Remedial Measures

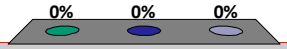
**Selection of transformation to remedy non-normality**

- Usually non-normality and non-constant variance occur together, and the transformation applied to remedy non-constant variance is all you need to remedy non-normality.



## For the job training data, is the model adequate after square root transformation?

1. Yes, adequate.
2. No, consider another transformation
3. Not sure.






## For the blood plasma data, is the model adequate after square root transformation?

1. Yes, adequate.
2. No, consider another transformation.
3. Not sure.