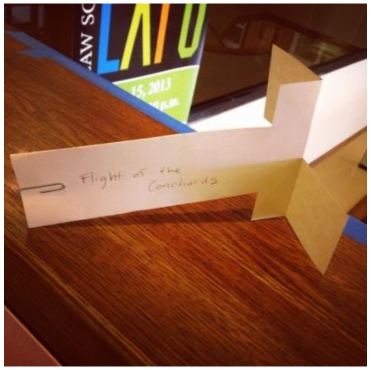
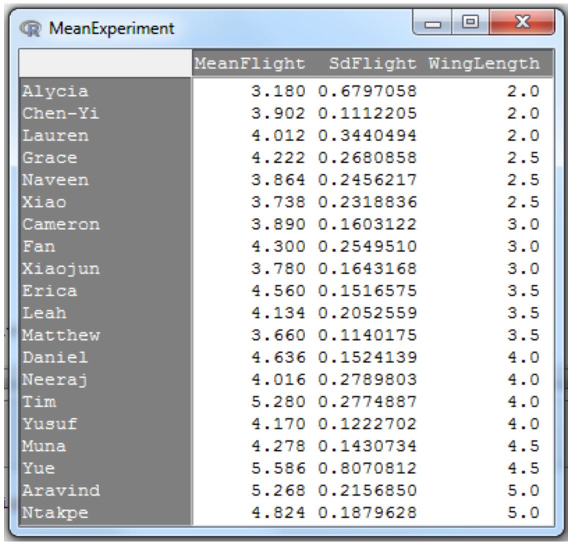
The Bivariate Regression Model

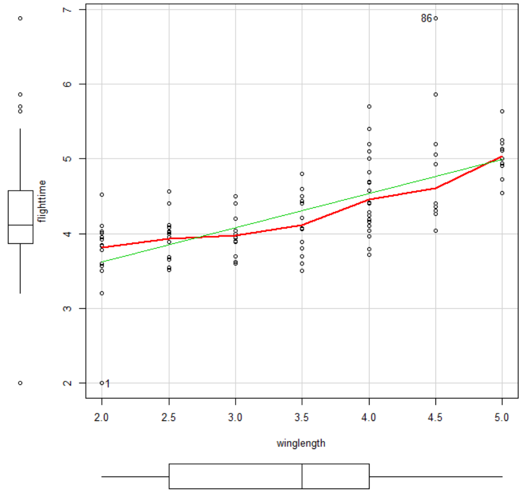
- Bivariate Regression analysis assumes a clear distinction between a cause and an effect variable:
  - One variable is given ***exogenously***. This variable is called the independent variable or cause and denoted by X . In theory, the independent variable can be ***controlled*** and its ***measurements are replicated***.
  - The response, ***endogenous***, effect or dependent variable Y is linked by a linear function to the independent variable $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$ where ε_i is the ***random disturbance*** term. It captures the random variation $\varepsilon_i = \hat{y}_i - y_i$ of y_i around its predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$.
- Regression traces the conditional distribution of *Y* given any feasible value of $X = x_i$.

Data from the Students' Helicopter Experiments:





	MeanFlight	SdFlight	WingLength
Alycia	3.180	0.6797058	2.0
Chen-Yi	3.902	0.1112205	2.0
Lauren	4.012	0.3440494	2.0
Grace	4.222	0.2680858	2.5
Naveen	3.864	0.2456217	2.5
Xiao	3.738	0.2318836	2.5
Cameron	3.890	0.1603122	3.0
Fan	4.300	0.2549510	3.0
Xiaojun	3.780	0.1643168	3.0
Erica	4.560	0.1516575	3.5
Leah	4.134	0.2052559	3.5
Matthew	3.660	0.1140175	3.5
Daniel	4.636	0.1524139	4.0
Neeraj	4.016	0.2789803	4.0
Tim	5.280	0.2774887	4.0
Yusuf	4.170	0.1222702	4.0
Muna	4.278	0.1430734	4.5
Yue	5.586	0.8070812	4.5
Aravind	5.268	0.2156850	5.0
Ntakpe	4.824	0.1879628	5.0



Notational considerations

- For the i^{th} observation the ***population model*** is $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$.
    - Neither the parameters β_0, β_1 nor the error term ε_i are directly observable.
    - The parameter β_0 is called the ***intercept*** and the parameter β_1 is the ***slope***.
    - The parameters β_0 and β_1 are constant for all observations. They denote a part of the ***model structure***.
    - The disturbance ε_i is directly associated with the i^{th} observation.
    - The disturbances have an underlying random distribution, which is also part of the ***model structure***.
- For the ***sample estimates*** the predicted value of the model becomes $\hat{y}_i = b_0 + b_1 \cdot x_i$ with the estimated ***residual*** $e_i = y_i - \hat{y}_i$.
    - *Note* that some books use α instead of β_0 for the population intercept and a instead of b_0 for the estimated intercept (for instance, Burt and Barber).
    - Note some people also write $\hat{\beta}_0$ and $\hat{\beta}_1$ for the estimated parameters b_0 and b_1 as well as $\hat{\varepsilon}_i$ for the residuals.
    - Bivariate regression has two estimated parameters *K* = 2: one for the intercept β_0 and one for the slope β_1.

Basic Interpretation

- Intercept and slope:

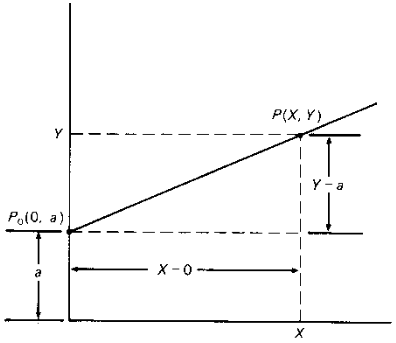
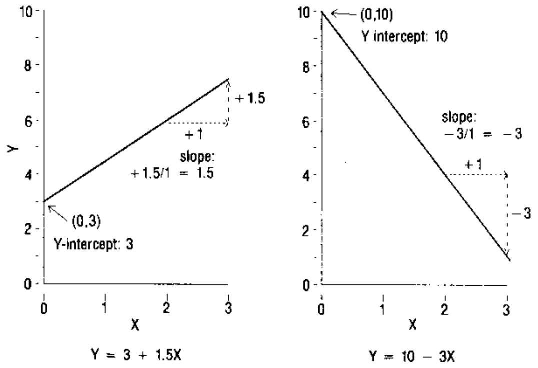


FIGURE 13-20. Derivation of equation for a line.



Y = 3 + 1.5X

Y = 10 − 3X

Figure 2.1 Two lines.

- The estimate b_1 slope is interpreted as: If we change X by one unit Y will change by b_1 units.
- Questions:

  What happens to the intercept if the slope equals zero?
  What happens to two regression lines if they only differ in their intercepts?
  What happens to two regression lines if they only differ in their slopes?
  Can we make certain statements outside the observed support of the variable X?

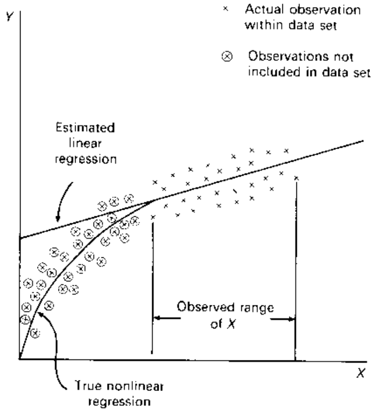


FIGURE 13-13. Regression model.

- The regression line ***summaries*** the data by a line (uses only two parameters for the n observed data pairs of observations).

- ***Parsimony Principle:*** *reduce the number of data points by replacing them with simple but general rules (summarize the data points).*
- Linear regression analysis separates
    - the systematic component which is the ***conditional expectations*** (conditional on the observation x_i) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$ from
    - the random component which are the ***disturbances*** ε_i.
- The residuals are ***unique*** for each observation. They deviate from the general rule $\beta_0 + \beta_1 \cdot x_i$.

Key Assumptions of Regression Analysis

1. Linear Relationship: the relationship between the independent variable and the dependent variable is ***linear*** (or can be transformed to linearity)
2. Non-randomness of X: The independent variable X is in theory deterministic variables.
   Note: should it for some reason be influenced by randomness, then we need to assume at least that it is uncorrelated with the population disturbances.
3. Disturbances: the disturbances have an ***identical distribution***, with zero mean and constant variance, for every observation i.
   They just deviate randomly with a given variance from their fixed mean of zero.
4. Independence: Disturbances are in general assumed to be independent (uncorrelated) among each other.

5. Normality: The additional assumption of *normality* of the disturbances allows exact statistical significance testing in the estimated regression model.
   This will be discussed later in the semester.

   Notes:

   ○ The independence and identical distribution assumption of the disturbances is abbreviated by ***i.i.d.*** (**i**ndependently **i**dentically **d**istributed)

   ○ Only the disturbances are required to be normal i.i.d. Neither *Y* nor *X* need to follow necessarily a normal distribution.

   ○ However, a joint normal distribution of *Y* and *X* is highly desirable to approach a ***linear relationship*** and a ***balanced distribution*** of all data points in the scatterplot.

Ordinary Least Squares Estimation and Variance decomposition

- While it is possible to draw many different regression lines through the cloud of data points the method of ***ordinary least squares*** calculates a line that minimizes the ***sum of the squared regression residuals*** *(RSS)*.
- We want to minimize the sum of squared residuals RSS in ordinary least squares regression

$$\min_{b_0,b_1} RSS = \min_{b_0,b_1} \sum (y_i - \hat{y}_i)^2$$

  with the residual being written as $e_i = y_i - \underbrace{(b_0 + b_1 \cdot x_i)}_{\hat{y}_i}$.

Minimizing this function is done by setting its first derivatives with regards to b_0 an b_1 equal to zero:

$$\frac{\partial RSS}{\partial b_0} = -\sum y_i + nb_0 + b_1 \sum x_{i1} \equiv 0$$

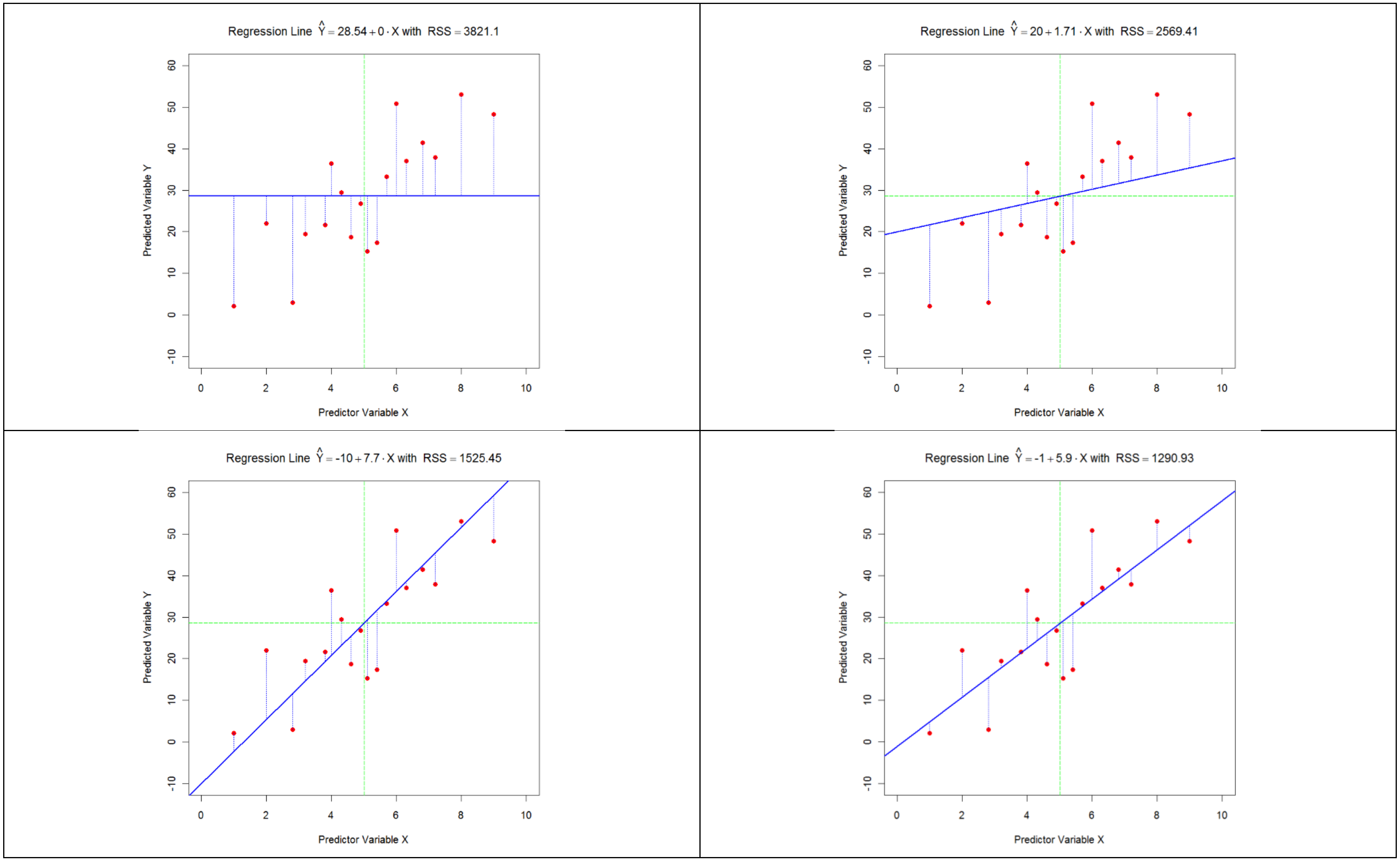
$$\frac{\partial RSS}{\partial b_1} = -\sum x_i \cdot y_i + b_0 \sum x_{i1} + b_1 \sum x_{i1}^2 \equiv 0$$

and solving for the unknown regression parameters b_0 and b_1.

- <u>Important property:</u> The first equation shows that the regression line must ***go through the means*** of the independent and the dependent variables $(\bar{Y}, \bar{X})$.

$$\begin{aligned} & -\sum y_i + nb_0 + b_1 \sum x_{i1} \equiv 0 \\ \Rightarrow\ & \sum y_i = nb_0 + b_1 \sum x_{i1} \quad | \div n \\ \Rightarrow\ & \bar{y} = b_0 + b_1 \cdot \bar{x} \end{aligned}$$

- See below four possible lines through the same point cloud and the joint-mean $(\bar{Y}, \bar{X})$. The line with the lowest *RSS* is the ***optimal*** regression line:



- Questions:
  - Which optimality condition did the arithmetic mean satisfy?
  - Why do we focus on the ***sum of the squared differences*** rather than the simple differences $e_i = \hat{y}_i - y_i$
  - What ***impact*** may ***outliers*** and extreme cases have on the estimated regression line?
  - What happens if the best fitting regression line is the horizontal line?
  - Do the estimated intercept and the estimated slope parameters covary?
- Important properties:
  1. As long as we have an intercept b_0 in the model, the ***sum of the residuals*** is always zero:
     $$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$$
  2. One also can show that the estimated residuals are always ***uncorrelated*** with the independent variable as well as with the predicted value:
     $Corr(\hat{y}, e) = Corr(x, e) = 0$

     $\Rightarrow$ Consequently, the independent variable in the regression model ***cannot be used to explain any additional variation*** in the regression residuals. Its explanatory power is exhausted.
- Important notational note: Burt, Barber and Rigby's notation differs from the commonly used notation. BBR use ***RSS*** are *"**R**egression **S**um of **S**quares"* and ***ESS*** means *"**E**rrors **S**um of **S**quares"*. The lecture sticks to the ***standard notation*** where ***RSS*** stands for "*__R__esidual __S__um of __S__quares"* and ***ESS*** stands for "*__E__xplained __S__um of __S__quares"*.

- ***Decomposition*** of the total sum of squares into *TSS = ESS + RSS*: since $y = \hat{y} + e$

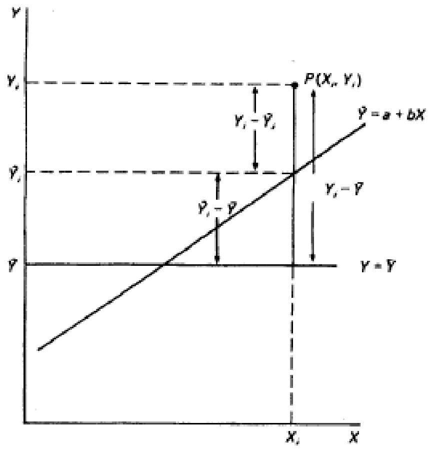


FIGURE 13-11. Decomposition of the total variation.

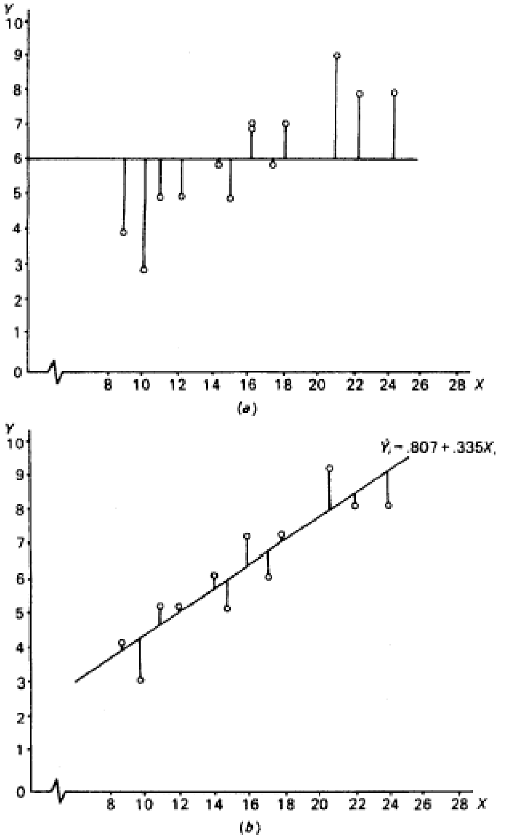


FIGURE 13-12. Geometrical representation of (a) total variation, (b) residual variation, and (c) explained variation.

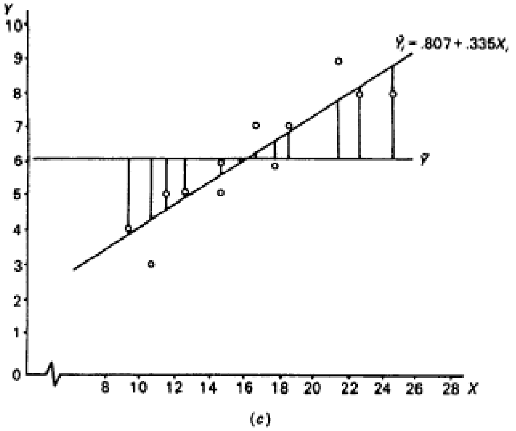


FIGURE 13-12 (continued).

- Definition of variance terms ***Total Sum of Squares*** (TSS), ***Residual Sum of Squares*** (RSS) and ***Explained Sum of Squares*** (ESS):

$$TSS = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \text{ with } df = n-1$$

$$RSS = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \text{with } df = n-K$$

$$ESS = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \text{with } df = K-1$$

- $K = 2$ is the number of the estimated regression coefficients b_0 and b_1.
- The values *TSS*, *RSS* and *ESS* with their degrees of freedom can be found in the so called regression ANOVA table of the standard regression output.

Coefficient of Determination: R^2 and Adjusted R^2_{adj}

- The goodness of fit measure is defined as $R^2 \equiv \dfrac{ESS}{TSS} = 1 - \dfrac{RSS}{TSS}$
- It measures what ***percentage of the total variation in the dependent variable is explained*** by the regression model.

  Note: a regression line based on just ***two observations*** will always ***fit the data perfectly***. Therefore, the ***adjusted goodness of fit*** takes the degrees of freedom into account.

$$R^2_{adj} \equiv 1 - \frac{RSS/(n-K)}{TSS/(n-1)}$$

  When *n* is large relative to *K* then the difference between the adjusted and the ordinary R^2 becomes negligible.

The more variables we enter into the regression equation, the better the fit of the model becomes.

Root Mean Square Error (Standard Error of Estimate)

- The root mean square error measures the standard deviation of the residuals:

$$s_e = \sqrt{Var(e)} = \sqrt{\frac{RSS}{n-K}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-K}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)_i^2}{n-K}}$$

(Note: Burt and Barber use the notation $s_{Y|X}$)

- We are losing $K = 2$ degrees freedom because the regression line is constrained to go thru the point $(\overline{x}, \overline{y})$. Alternative explanation: We need two points to define a line.

Helicopter Example Continued

```
> LinearModel.1 <- lm(flighttime ~ winglength, data=Helicopter)
> summary(LinearModel.1)

Call: lm(formula = flighttime ~ winglength, data = Helicopter)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71151    0.18479  14.674  < 2e-16 ***
winglength   0.45691    0.05237   8.725 7.05e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4941 on 98 degrees of freedom
Multiple R-squared:  0.4372,  Adjusted R-squared:  0.4314
F-statistic: 76.12 on 1 and 98 DF,  p-value: 7.047e-14
```

Multiple Regression

Introduction

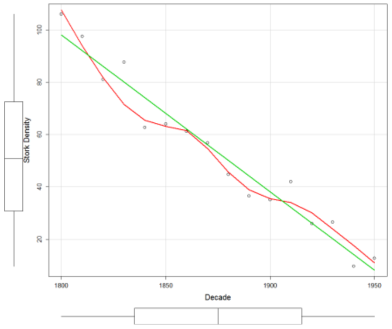
- Multiple regression is a logical extension of bivariate regression analysis. It combines several independent variables through a linear combination in the regression coefficients and allows performing predictions of the dependent variable.

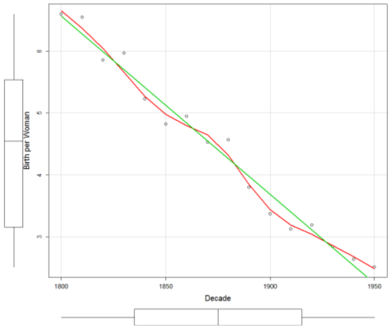
  Example: Prediction equation with three independent variables

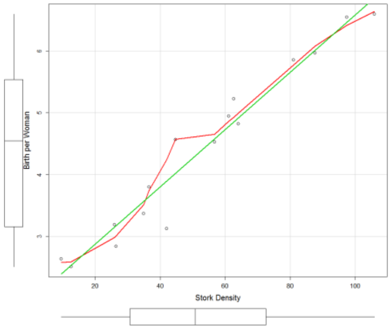
$$\boxed{\hat{Y} \leftarrow b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3}$$

- Multiple regression allows to ***control*** for any confounding effects that a variable may have on the dependent and other independent variables in the model.
- Multiple regression combines the influences of several independent variables on a dependent variable.
  - The measurement units of the independent variables can differ from each other and the dependent variable.
  - The sign and magnitude of the associate weighting factors, i.e., the regression coefficient b_k, simultaneously
    - compensate for the differences in the measurement scales and
    - captures any influence of the independent variable on the dependent variable.

- Control for Confounding: Revisit the Stork example







Bivariate Regression Model

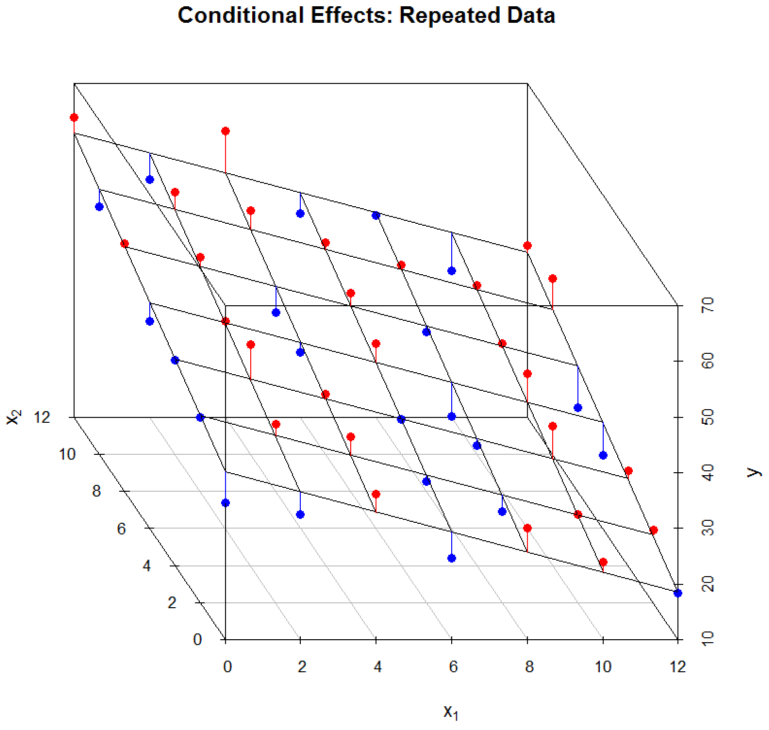
```
lm(formula = birth ~ stork, data = myStork)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.939020   0.168849   11.48 1.64e-08 ***
stork       0.046458   0.002807   16.55 1.37e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.317 on 14 degrees of freedom
Multiple R-squared:  0.9514,    Adjusted R-squared:  0.9479
F-statistic:   274 on 1 and 14 DF,  p-value: 1.372e-10
```

Multiple Regression Model Controlling for Decade

```
lm(formula = birth ~ stork + decade, data = myStork)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.640053  10.678527   4.274 0.000906 ***
stork        0.010918   0.008895   1.227 0.241429
decade      -0.022300   0.005449  -4.093 0.001270 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2175 on 13 degrees of freedom
Multiple R-squared:  0.9788,    Adjusted R-squared:  0.9755
F-statistic: 299.5 on 2 and 13 DF,  p-value: 1.338e-11
```

Interpretation of the Regression Coefficient in Multiple Regression

- For two independent variables the ***predicted values*** $\hat{Y} = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$ will be on a ***linear plane***.
  - This surface is anchored in the ***mean point*** of $\bar{Y}$, $\bar{X}_1$ and $\bar{X}_2$, which is always in the center of the surface.
  - The plane will pass through ***intercept*** b_0 when $X_1 = 0$ and $X_2 = 0$.
  - The slope along the X_1-axis is b_1 and along the X_2-axis b_2, respectively.
- An individual regression coefficient associated with a variable X_1 or X_2 is interpreted as if (***conditionally on***) the values of the other variable in the model remain fixed at a given level.
  - *Ceteris paribus* mental experiment of repeated sampling at fixed levels of X_2: Estimate the regression lines $Y \leftarrow f(X_1 | X_2 = x_2)$ at each level of x_2 of X_2.
- By holding the level of X_2 fixed, the model is estimated just for the subset of observations that satisfy $X_2 = x_2$. $\Rightarrow$ The slope b_1 does not change.
  - Therefore, the relationship between $Y \leftarrow f(X_1 | X_2 = x_2)$ is no longer influenced by X_2.



  - For linear relationships this implies that the regression lines at any level of $X_2 = const$ have the same slope. That is, the slope b_1 does not change.
- Question: Do the two variables X_1 and X_2 in the plot appear to be correlated?

SAT Score Example

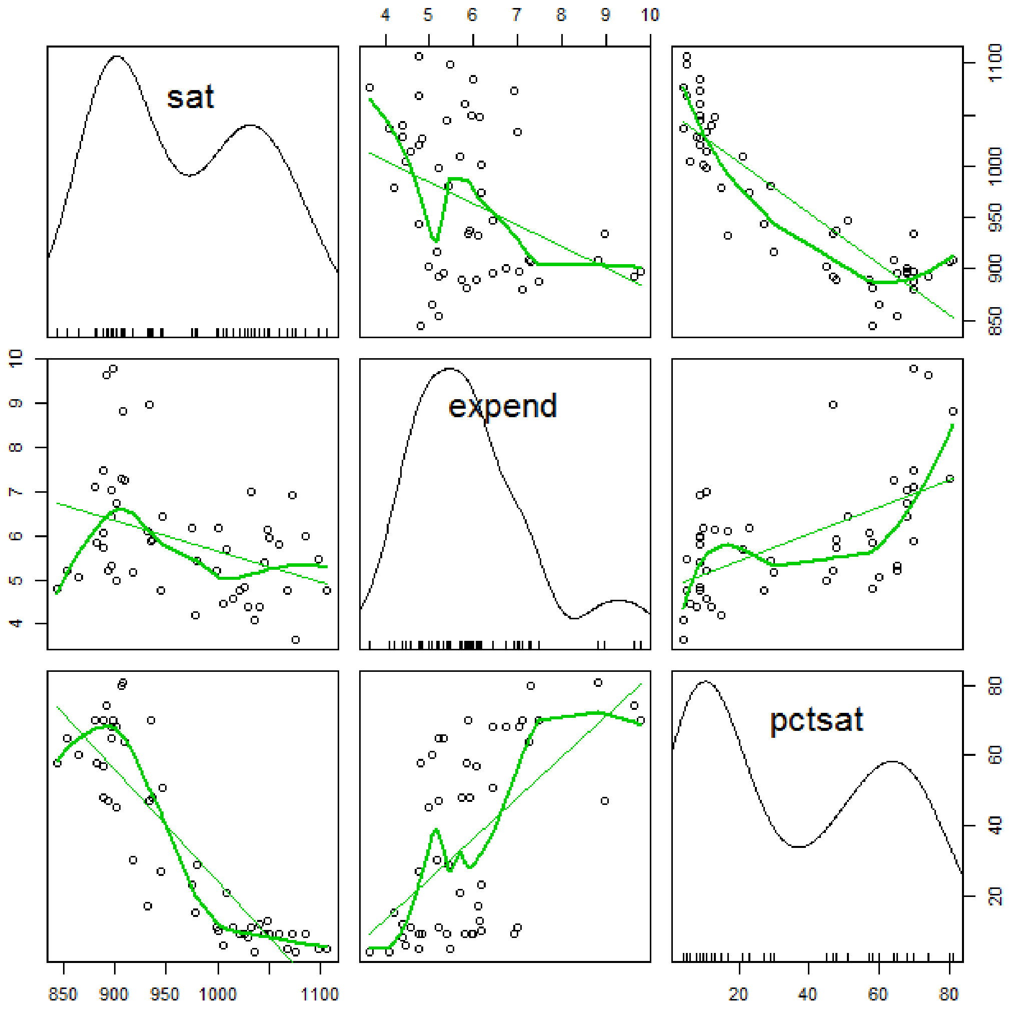
The data for this example can be found in the SPSS file **STATESCHOOL.SAV**

- Discuss the variables in the model $SAT = f(Expend, PctSAT)$:
  - SAT: The state's average SAT score for all those student who took the exam.
  - $Expend$: The average per student state's educational expenditure.
  - $PctSAT$: Percentage of student within a state who took the SAT exam.
- Comments: We can assume that only students aiming at attending prestigious universities prefer to take the exam unless they are forced to do so by state law.
  Therefore we have a ***selection bias***:
  [a] a low the percentage of state's participation means that predominately students with better academic achievements will take the test leading to a higher test score;
  [b] whereas, a high percentage of state's participation implies that larger fraction of mediocre performing students will also take the test most likely leading to a lower state’s score.

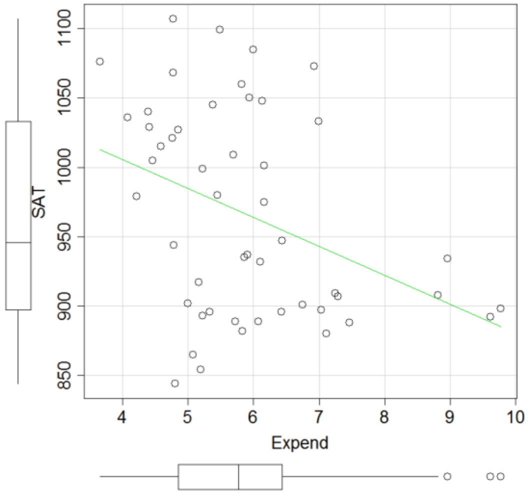
Univariate and Bivariate Data Description:

```
rcorr.adjust(StateSchool[ ,
c("Expend","PctSAT","SAT")],
type="pearson")
```

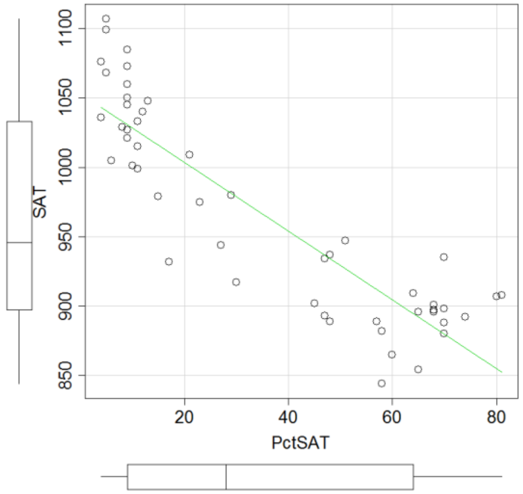
```
       Expend PctSAT   SAT
Expend   1.00   0.59 -0.38
PctSAT   0.59   1.00 -0.89
SAT     -0.38  -0.89  1.00

n= 50

Two-Sided Prob-Value
       Expend PctSAT SAT
Expend        0.0000 0.0064
PctSAT 0.0000        0.0000
SAT    0.0064 0.0000
```



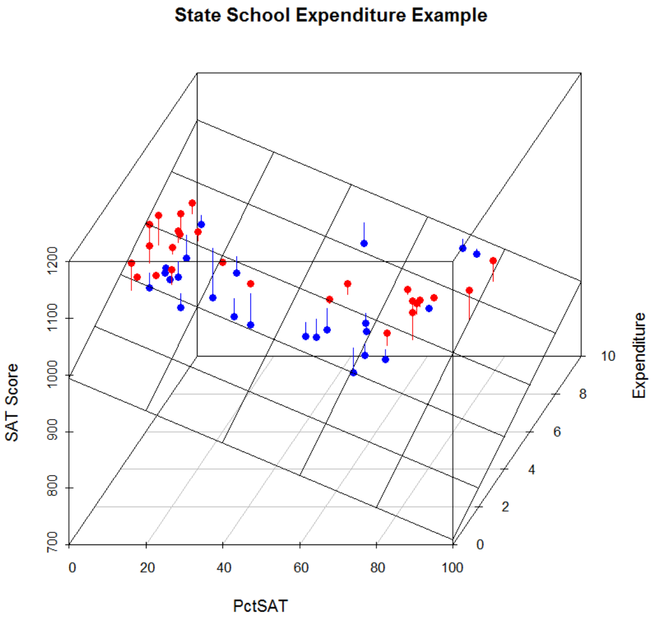
From Bivariate to Multivariate Analysis



$$\widehat{SAT} = 1089.3 - 20.9 \cdot Expend$$



$$\widehat{SAT} = 1053.3 - 2.5 \cdot PctSAT$$



- ***Counterintuitive*** bivariate result: We expect that the expenditures in primary education will have a positive impact on the SAT scores but both variables are negatively correlated
- In the multivariate model the variable **EXPEND** has a positive impact on the **SAT** scores.
- The variable **PCTSAT** can be viewed as a confounder because it is jointly correlated with the dependent variable **SAT** and the independent variable **EXPEND**.

- Results of the multivariate regression model:

```
lm(formula = SAT ~ Expend + PctSAT, data = StateSchool)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
Expend       12.2865     4.2243   2.909  0.00553 **
PctSAT       -2.8509     0.2151 -13.253  < 2e-16 ***

Residual standard error: 32.46 on 47 degrees of freedom
Multiple R-squared: 0.8195,  Adjusted R-squared: 0.8118
F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

- In the multivariate model the confounding effect is clearly controlled and the regression coefficient $b_1 = 12.3$ of the independent variable **EXPEND** expresses its sole effect on the **SAT** score without interference from the confounder **PCTSAT**.
- The intercept just guarantees that the prediction surface goes through the mean point of **SAT**, **EXPEND** and **PCTSAT**.

  It is meaningless here, because for zero participation rate, a SAT score would not exist.
- See Gruber (1999) http://www.amstat.org/publications/jse/v7n2_abstracts.html for more information about this example.

Multiple Correlation R^2

- As in bivariate regression analysis, the estimated model in multiple regression analysis is the ***best fitting model***.

  $\Rightarrow$ Any other combination of regression coefficients will lead to larger squared residuals.
- The ***multiple correlation coefficient*** R^2 measures the squared correlation between the observed dependent variable Y and the predicted dependent variable $\hat{Y}$.

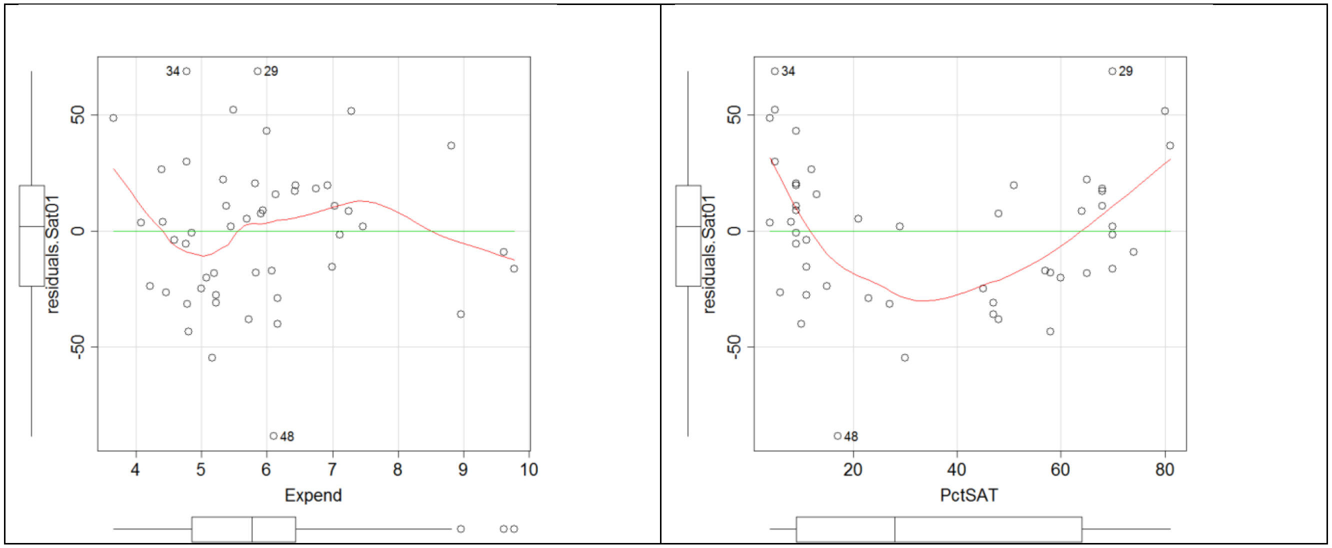
  $\Rightarrow$ No other linear regression model using same variables will lead to a larger correlation.

Analysis of Regression Residuals

- Residuals, which are that part of the dependent variable that is ***not explained*** by the regression model, play an important part in statistical analysis.
- Should we uncover a so far ***unknown pattern*** in the regression residuals, then we accomplished scientific progress.
- More ***advance courses*** will explore regression residual quite extensively to check for potential model violations and to identify outliers and influential observations.
- Remember: The residuals are by design ***linearly uncorrelated*** with any independent variable in the multivariate regression model as well as with the predicted value of the dependent variable.

  $\Rightarrow$ the information with regards to the independent variables has been fully exploited and became solely part of the predicted variable.

- That is, the residuals are free of any linear effect of (no information about) the independent variables. They may, however, still hold information about so far unconsidered variables.
- For instance, the curvilinear relationship between the residuals and **PctSAT** indicates that a model with $\widehat{SAT} = b_0 + b_1 \cdot Expend + b_2 \cdot PctSAT + b_3 \cdot PctSAT^2$ may be more appropriate.



- The observations 29, 34 and 48 are have large ***residuals*** of ***opposite*** sign.
  $\Rightarrow$ This requires further investigation. Why are both states performing so different?
  $\Rightarrow$ If we find a yet unknown variable explaining this difference then we can improve the regression model and have enhanced our knowledge.