# Lecture 14: Introduction to Regression

## • 14. 1 Regression Fundamentals

- Regression analysis models the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables) (usually named $Y$), and the predictors (also called independent variables, explanatory variables, control variables, or regressors,) usually named $X_1,..., X_p$). *Multivariate regression* describes models that have more than one response variable

- The Regression method was discovered independently by Carl Friedrich Gauss (Germany) in 1795 and by Adrien Marie Legendre (France) around 1805. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun

## • 14. 2 Types of Regression

- **Simple and Multiple Linear Regression Models** are related statistical methods for modeling the relationship between two or more random variables using a linear equation. Simple linear regression refers to a regression on two variables while multiple regression refers to a regression on more than two variables. Linear regression assumes the best estimate of the response is a linear function of some parameters

- **Nonlinear Regresssion Models** are used in situations when the relationship between the variables being analyzed is not linear in parameters. A number of nonlinear regression techniques may be used to obtain a more accurate regression in these situations

- **Other Models**. Although these three types are the most common, there also exist Poisson regression, supervised learning, and unit-weighted regression

## • 14. 3 Linear Models

- Predictor variables may be defined quantitatively (i.e., continuous) or qualitatively (i.e., *categorical*). Categorical predictors are sometimes called factors. Although the method of estimating the model is the same for each case, different situations are sometimes known by different names for historical reasons:

  - If the predictors are all quantitative, we speak of multiple regression

  - If the predictors are all categorical, one performs analysis of variance

  - If some predictors are quantitative and some categorical, one performs an analysis of covariance

- The linear model usually assumes that the dependent variable is continuous. If least squares estimation is used, then if it is assumed that the error component is normally distributed, the model is fully parametric. If it is not assumed that the data are normally distributed, the model is semi-parametric. If the data are not normally distributed, there are often better approaches to fitting than least squares. In particular, if the data contain outliers, robust regression might be preferred

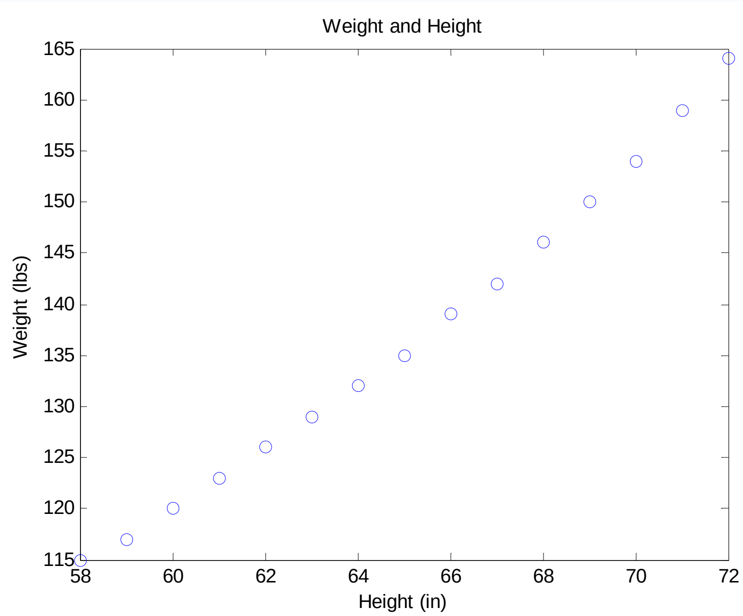
- If the regression error is not normally distributed but is assumed to come from an exponential family, generalized linear models should be used. For example, if the response variable can take only binary values (for example, a Boolean or Yes/No variable), logistic regression is preferred. The outcome of this type of regression is a function which describes how the probability of a given event (e.g. probability of getting "yes") varies with the predictors

- Data of average heights and weights for American women aged 30-39

    - Height: [58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72]

    - Weight: [115, 117, 120, 123, 126, 129, 132, 135, 139, 142, 146, 150, 154, 159, 164]

- We would like to examine whether the weight of these women depends on their height. We are looking for a function β such that $Y = \beta(X) + \varepsilon$, where $Y$ is the weight of the women and $X$ their height



## • 14. 4 Simple Linear Regression Model

- The simple linear regression model is typically stated in the form

$$Y = \alpha + \beta X + \varepsilon$$
$$\varepsilon \sim \text{NID}(0, \sigma_e^2)$$

- The right hand side may take more general forms, but generally comprises a linear combination of the parameters, here denoted α and β. The term ε represents the unpredicted or unexplained variation in the response variable; it is conventionally called the "error", whether it is really a measurement error or not, and is assumed to be independent of *X*. The error term is conventionally assumed to have expected value equal to zero and variance $\sigma^2$

- Regression is the process of defining an equation that predicts values of one variable conditional on values of another variable

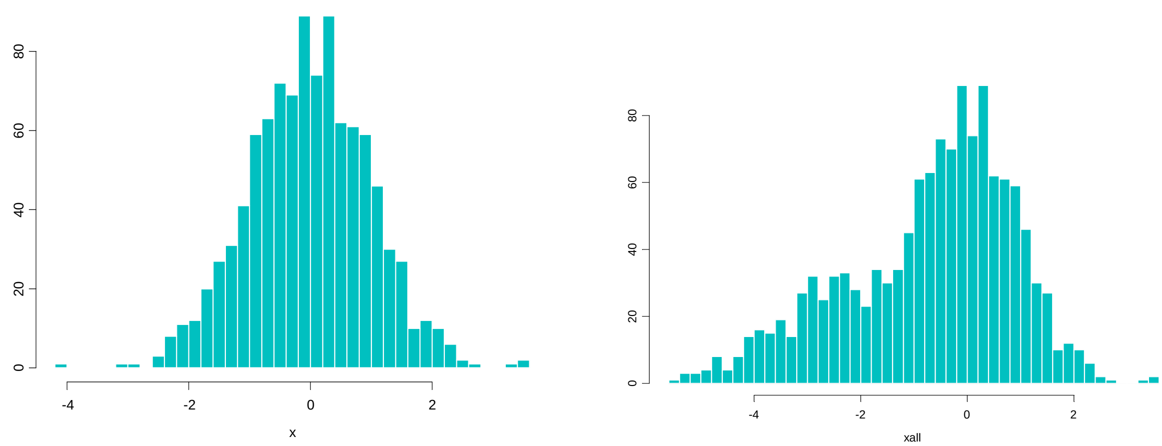
  - An equivalent formulation that explicitly shows the linear regression as a model of conditional expectation is

$$E(y \mid x) = \alpha + \beta x$$

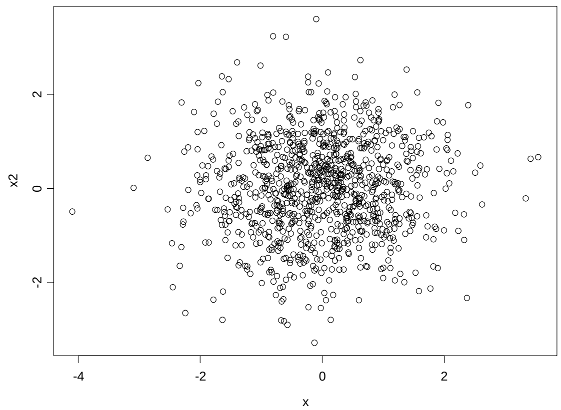
- Often in linear regression problems statisticians rely on the Gauss-Markov assumptions:

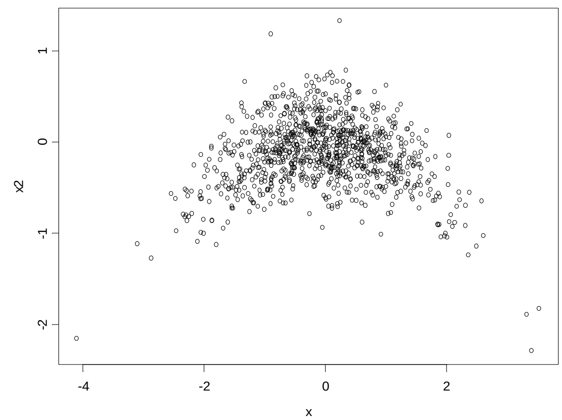
  - The random errors $\varepsilon_i$ have expected value 0
  - The random errors $\varepsilon_i$ are independent (or uncorrelated)
  - The random errors $\varepsilon_i$ are homoscedastic, i.e., they all have the same variance
  - They are normally distributed

- ***Normal*** distribution: the errors should be normally distributed – Gauss posits that normality is an optimal distributional form for the least squares method



- ***Independence*** of residuals: the error terms are ***uncorrelated*** among themselves as well as with the predictors

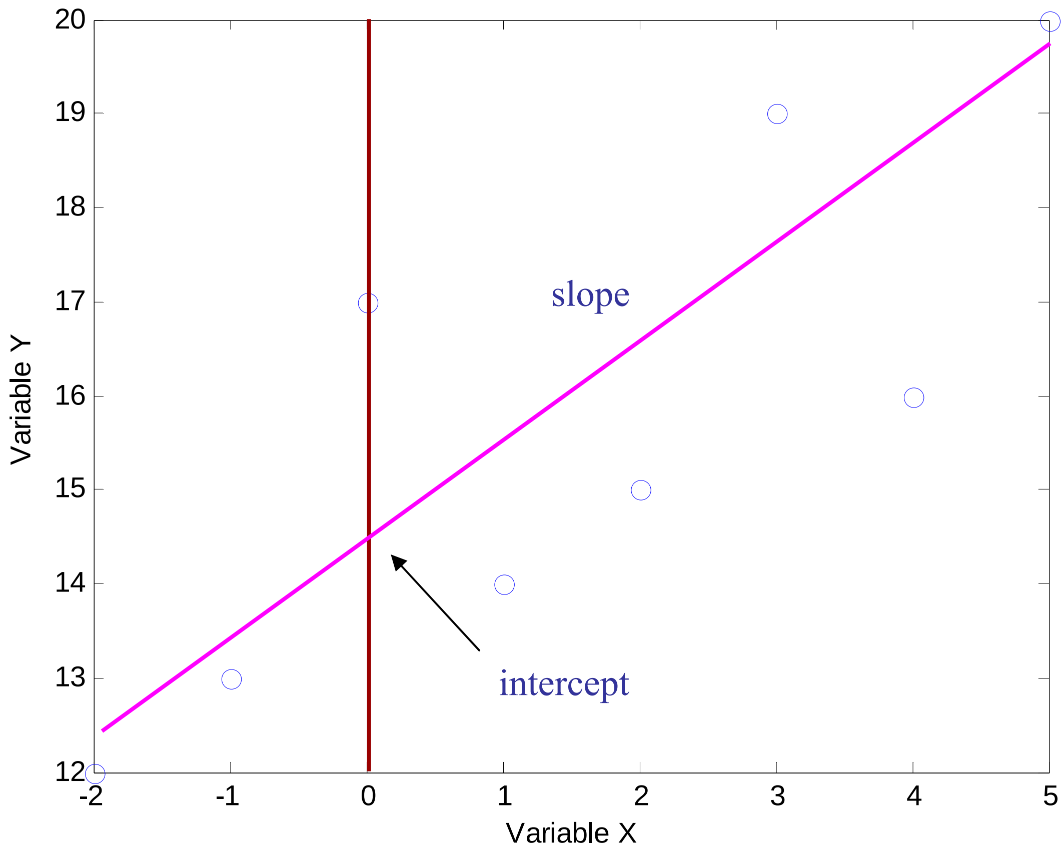



- Violation of normality leads to biases in inferential statistics – parameter estimates, confidence interval, prediction interval, hypothesis testing, other parametric derivatives

## • 14. 5 Method of Least Squares and Model Parameters

- The simple linear regression $Y = \alpha + \beta X + \varepsilon$ (or $Y = a + bX + e$; or $Y = b_0 + b_1 X + e$)




$b_0$ = intercept; mean response when $x = 0$

$b_1$ = slope; change in mean response in y when $x$ increases by 1 unit. It describes the linear relationship between $x$ and $y$, can be positive or negative, and increases with magnitude as the linear relationship becomes stronger

$b_0$, $b_1$ are unknown parameters ($\alpha$, $\beta$)

$b_0 + b_1 x$ = mean response when explanatory variable takes on the value $x$

- The equation defined by these values represents the "line of best fit," which minimizes *residuals* or *errors of prediction*.

- Goal: Choose values (estimates) that minimize the sum of squared errors (*SSE*) of observed values to the straight-line:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_i$$

$$\sum_{i=1}^{N} (y_i - \hat{y})^2 = \sum_{i=1}^{N} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2$$

- Compute the slope first

$$b_1 = \frac{\mathrm{cov}_{x,y}}{s_x^{\,2}}, \text{ also } b_1 = r\frac{s_y}{s_x}$$

- Compute then the intercept

$$b_0 = \bar{y} - b_1 \bar{x}$$

- The regression equation is

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

## • 14. 6 Method of Least Squares (cont.)

- A statistician will usually estimate the unobservable values of the parameters α and β by the method of least squares, which consists of finding the values of *a* and *b* that minimize the sum of squares of the residuals

$$\hat{\varepsilon}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i) \text{ ; or } \hat{\varepsilon}_i = y_i - (\hat{b}_0 + \hat{b}_1 x_i)$$

- The residual is the vertical distance from the estimated regression line to the data point $(x_i, y_i)$

- Those values of $\hat{\alpha}$ and $\hat{\beta}$ are the "least-squares estimates" of α and β respectively. The residuals may be regarded as estimates of the errors.

- Notice that, whereas the errors are independent, the residuals cannot be independent because the use of least-squares estimates implies that the sum of the residuals must be 0, and the scalar product of the vector of residuals with the vector of *x*-values must be 0, i.e., we must have

$$\hat{\varepsilon}_1 + \ldots + \hat{\varepsilon}_n = 0$$

and

$$\hat{\varepsilon}_1 x_1 + \ldots + \hat{\varepsilon}_n x_n = 0$$

- These two linear constraints imply that the vector of residuals must lie within a certain $(n - 2)$ dimensional subspace of $R^n$; hence we say that there are "$n - 2$ degrees of freedom for error". If one assumes the errors are normally distributed and independent, then it can be shown to follow that 1) the sum of squares of residuals

$$\hat{\varepsilon}_1^2 + \ldots + \hat{\varepsilon}_n^2 = 0 \text{ is distributed as } \sigma^2 \chi^2_{n-2}$$

- So we have :

  - the sum of squares divided by the error-variance $\sigma^2$, has a $\chi^2$ distribution with $n - 2$ degrees of freedom

  - the sum of squares of residuals is actually probabilistically independent of the estimates $\hat{\alpha}, \hat{\beta}$ of the parameters α and β

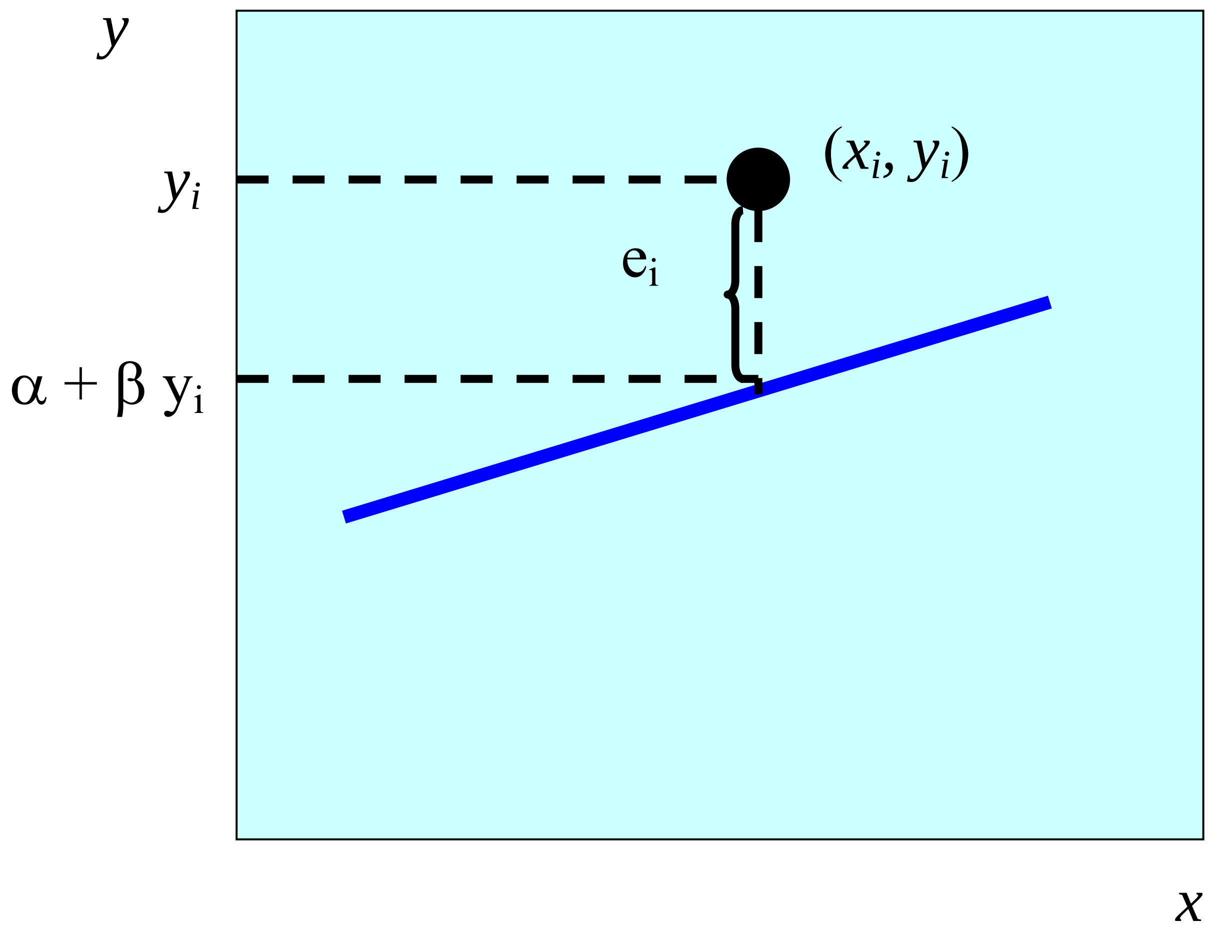
- These facts make it possible to use Student's t-distribution with $n - 2$ degrees of freedom to find confidence intervals for α and β

- Minimize sum of squared residuals – Thus, named "***Least Squares***" Method

$$B = (XX)^{-1}\,(XY)' \quad \text{--->} \quad E\{Y\} = BX + e$$

$$\sum \left(y_i - \hat{y}\right)^2$$

- The residual is the vertical distance from the estimated regression line to the data point ($x_i$, $y_i$)



- ## 14. 7 Example

- You want to examine whether or not High School GPA is related to the number of years of education of the student's mother

  - Mother's education: $X$ = [0, 1, 3, 4]

  - HS GPA: $Y$ = [3.0, 3.2, 3.3, 3.7]

- Slope: $b_1 = \dfrac{\mathrm{cov}_{x,y}}{s_x^{\,2}}$, for this we need the variance of $x$ and the covariance of $x$ with $y$

$$s_x^{\,2} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}, \text{ and } \mathrm{cov}(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$\text{Thus, } b_1 = \frac{\mathrm{cov}_{x,y}}{s_x^{\,2}} = \frac{\dfrac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}}{\dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\bar{x} = 2;\ \bar{y} = 3.3$$

$$b_1 = \frac{[(-2)(-.3)] + [(-1)(-.1)] + [(1)(0)] + [(2)(.4)]}{(-2)^2 + (-1)^2 + (1)^2 + (2)^2} = \frac{1.5}{10} = .15$$

- Intercept: $b_0 = \bar{y} - b_1\bar{x} = 3.3 - .15(2) = 3$

- So, $b_0$ = 3 and $b_1$ = .15

- The regression equation is

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = 3 + .15x$$

GPA = 3 + .15 (Mother's education)

## • 14. 8 Accuracy of Prediction

- The standard deviation as a measure of error

  - The best prediction of $\hat{y}$ is $\bar{y}$ and the error associated with that prediction is the SD of $Y$ (deviations from the mean)

$$s_Y = \sqrt{\frac{\sum_{i=1}^{N}(Y-\bar{Y})^2}{N-1}} \text{ and the variance } s_Y^{\,2} = \frac{\sum_{i=1}^{N}(Y-\bar{Y})^2}{N-1} = \frac{SS_Y}{df}$$

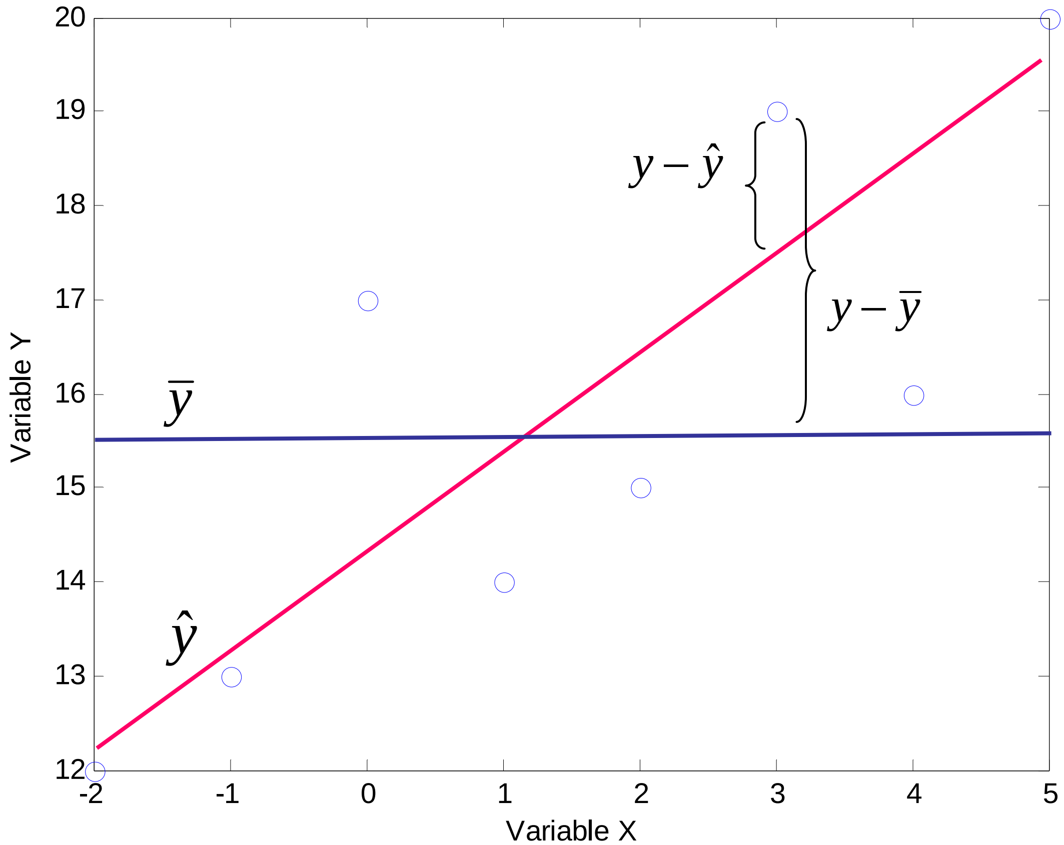
- The standard error of estimate

$$s_{Y\cdot X} = \sqrt{\frac{\sum_{i=1}^{N}(Y-\hat{Y})^2}{N-2}} = \sqrt{\frac{SS_{residual}}{df}}$$

(df = $N$ – 2 because we estimated $b_0$ and $b_1$ from data to obtain the regression line)

$s^2_{Y\cdot X}$ is the residual variance or error variance

$s_{Y\cdot X}$ conveys information about the variability of residuals


20
19
18
17
16
15
14
13
12
-2
-1
0
1
2
3
4
5
Variable Y
Variable X
$y - \hat{y}$
$y - \bar{y}$
$\bar{y}$
$\hat{y}$

- $r^2$ and the Standard Error of Estimate

$r^2$ is easier to understand and doesn't depend on the units of *x* and *y*

$$s^2_{Y \cdot X} = \frac{\sum_{i=1}^{N}(Y-\hat{Y})^2}{N-2} = \frac{SS_{residual}}{df}$$

$$s_{Y \cdot X} = s_Y\sqrt{(1-r^2)\frac{N-1}{N-2}}$$ For large samples $\frac{N-1}{N-2} \approx 1$, thus

$s^2_{Y \cdot X} = s^2_Y(1-r^2)$, or

$$s_{Y \cdot X} = s_Y\sqrt{(1-r^2)}$$

- $r^2$ as a Measure of Predictable Variability

$$SS_{residual} = SS_Y(1-r^2) = SS_Y - SS_Y(r^2)$$

$$r^2 = \frac{SS_Y - SS_{residual}}{SS_Y} = \frac{SS_{\hat{Y}}}{SS_Y}$$

$$r^2 = \frac{SS_{\hat{Y}}}{SS_Y} = \frac{\sum_{i=1}^{N}(\hat{y}-\bar{y})^2}{\sum_{i=1}^{N}(y-\bar{y})^2} = \frac{\dfrac{\sum_{i=1}^{N}(\hat{y}-\bar{y})^2}{N-1}}{\dfrac{\sum_{i=1}^{N}(y-\bar{y})^2}{N-1}} = \frac{s_{\hat{y}}^{\,2}}{s_y^{\,2}}$$

To the degree that $\hat{y} = \bar{y}$, $r^2$ approaches 1

## • 14. 9 Hypothesis Testing

- The standard deviation as a measure of error

$H_0$: $b_1 = 0$

$H_1$: $b_1 \neq 0$

- We can divide $b_1$ by its standard error and use a *t*-test.

- The standard error of $b_1$ ($SE_{b1}$) is the standard deviation of the distribution of *B* if we were to take a large number of samples and compute *B* for each sample

$$SE_{b1} = \frac{s_y\sqrt{(1-r)^2\left(\frac{N-1}{N-2}\right)}}{s_x\sqrt{N-1}}$$

$$t = \frac{b_1}{\frac{s_y\sqrt{(1-r)^2\left(\frac{N-1}{N-2}\right)}}{s_x\sqrt{N-1}}} = \frac{b_1 s_x\sqrt{N-1}}{s_y\sqrt{(1-r)^2\left(\frac{N-1}{N-2}\right)}}$$

$$t = \frac{b_1 s_x\sqrt{1}}{s_y\sqrt{(1-r)^2\left(\frac{1}{N-2}\right)}} = \frac{b_1}{\frac{s_y}{s_x}\sqrt{\frac{1-r^2}{N-2}}}$$

- *t* can now be compared to a critical *t*, where $df = N - 2$

- When the scales of $x$ and $y$ are the same (for instance, when $x$ and $y$ are standardized), two things happen:

1. $b_1$ (or β, beta) is now a correlation coefficient, and
2. $(s_y / s_x) = 1.0$, so it disappears, leaving

$$t = \frac{b_1}{\frac{s_y}{s_x}\sqrt{\frac{1-r^2}{N-2}}} = \frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$$

which is the $t$-test for a correlation coefficient

- Standardized Regression Weights

- Why is the standardized intercept always zero, and the standardized slope always $r$?

$$b_1 = r\frac{s_y}{s_x}$$

- For standardized variables, the variance and standard deviation are always 1.0, and the mean is always 0. Therefore, for the slope

$$b_1 = r\frac{s_y}{s_x} = r\frac{1}{1} = r$$

- and for the intercept

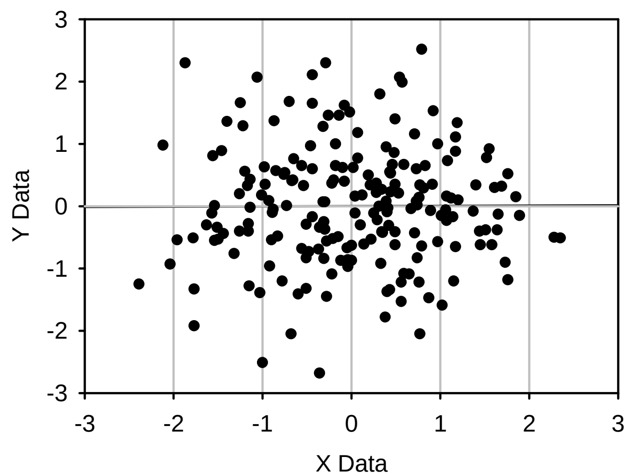
$$b_0 = \bar{y} - b_1\bar{x} = 0 - r0 = 0$$

$$\hat{z}_y = b_0 + b_1 z_x$$

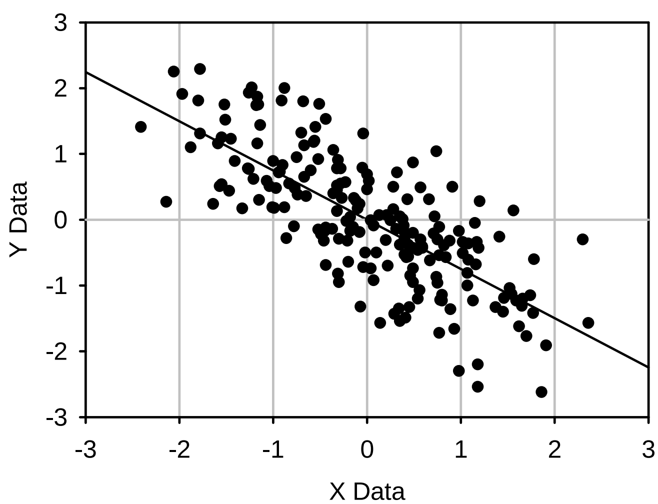
$$\hat{z}_y = 0 + rz_x = rz_x$$

- Standardized Regression Plot

a) For $r = 0$




b) For $r = -.75$



## • 14. 10 Multiple Regression Model

- *Multiple correlation* is the association between an outcome (or criterion) variable and two or more predictor variables

- Making predictions in this situation is called *multiple regression*

- The goal of *multiple regression* is to find a regression equation to predict $Y$ on the basis of $p$ predictors

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

where $b_0$ is the intercept and $b_1, b_2, \ldots, b_p$ are the regression coefficients for the predictors $X_1, X_2, \ldots, X_p$, respectively

- To estimate the parameters $b_{01}, b_2, \ldots, b_p$ we also use the least squares method

$$\sum_{i=1}^{N} (y_i - \hat{y})^2 = \min$$

## • 14. 11 Single vs. Multiple Regression

- In single (bivariate) regression, the regression slope (or regression coefficient) using standardized variables is the correlation coefficient

- In multiple regression, the standardized regression coefficient (β) for each predictor variable is not the same as the correlation coefficient between that predictor and the outcome variable

- Usually, the β will be smaller (in absolute value) than $r$ because of the overlap between the association between the predictor and the outcome with the associations for the other predictors (*multicollinearity*)

- In multiple regression, the correlation between the criterion and all the predictors is called the *multiple correlation coefficient* (***R***)

- ***R*** is typically smaller than the sum of all the individual correlations

- $R^2$ (*squared multiple correlation*) is the proportion of variance accounted for in the criterion variable by all the predictors together

- Tests in multiple regression
  - Overall test for the entire set of predictors (test of multiple correlation)
  - Significance tests for each of the predictors (whether each of the predictors adds more than zero to the prediction beyond what the other predictors in the model already predict)

- **14. 12 Example**

Sample Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| relsat1 | 262 | 6.18764 | 0.70537 | 1621 | 2.83300 | 7.00000 |
| age1 | 261 | 22.73245 | 7.61973 | 5933 | 17.91700 | 74.25000 |
| involv1 | 260 | 2.76151 | 5.12021 | 717.99300 | 0.04200 | 35.08300 |
| avoid1 | 262 | 2.51373 | 0.78461 | 658.59800 | 1.27800 | 4.88900 |
| anxiety1 | 262 | 3.44296 | 0.95706 | 902.05500 | 1.33300 | 6.16700 |
| balance1 | 262 | 3.18893 | 0.53748 | 835.50000 | 1.00000 | 4.00000 |

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

| | relsat1 | age1 | involv1 | avoid1 | anxiety1 | balance1 |
|---|---|---|---|---|---|---|
| relsat1 | 1.00000 | -0.07397 | -0.05898 | -0.44573 | -0.09907 | 0.15164 |
| | | 0.2337 | 0.3435 | <.0001 | 0.1096 | 0.0140 |
| age1 | -0.07397 | 1.00000 | 0.90240 | 0.07595 | -0.08612 | 0.02987 |
| | 0.2337 | | <.0001 | 0.2214 | 0.1654 | 0.6310 |
| involv1 | -0.05898 | 0.90240 | 1.00000 | -0.00499 | -0.10895 | 0.03831 |
| | 0.3435 | <.0001 | | 0.9362 | 0.0795 | 0.5386 |
| avoid1 | -0.44573 | 0.07595 | -0.00499 | 1.00000 | 0.12613 | -0.07495 |
| | <.0001 | 0.2214 | 0.9362 | | 0.0413 | 0.2266 |
| anxiety1 | -0.09907 | -0.08612 | -0.10895 | 0.12613 | 1.00000 | -0.04569 |
| | 0.1096 | 0.1654 | 0.0795 | 0.0413 | | 0.4615 |
| balance1 | 0.15164 | 0.02987 | 0.03831 | -0.07495 | -0.04569 | 1.00000 |
| | 0.0140 | 0.6310 | 0.5386 | 0.2266 | 0.4615 | |

**- Simple Regression Models**

| | | | |
|---|---|---|---|
| Root MSE | 0.69834 | R-Square | 0.0055 |
| Dependent Mean | 6.19410 | Adj R-Sq | 0.0016 |
| Coeff Var | 11.27419 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 6.34833 | 0.13625 | 46.59 | <.0001 | 0 |
| age1 | 1 | -0.00678 | 0.00568 | -1.19 | 0.2337 | -0.07397 |

| | | | |
|---|---|---|---|
| Root MSE | 0.69524 | R-Square | 0.0035 |
| Dependent Mean | 6.19934 | Adj R-Sq | -0.0004 |
| Coeff Var | 11.21476 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 6.22145 | 0.04901 | 126.94 | <.0001 | 0 |
| involv1 | 1 | -0.00801 | 0.00844 | -0.95 | 0.3435 | -0.05898 |

| | | | |
|---|---|---|---|
| Root MSE | 0.63264 | R-Square | 0.1987 |
| Dependent Mean | 6.18764 | Adj R-Sq | 0.1956 |
| Coeff Var | 10.22429 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 7.19492 | 0.13141 | 54.75 | <.0001 | 0 |
| avoid1 | 1 | -0.40071 | 0.04991 | -8.03 | <.0001 | -0.44573 |

| | | | |
|---|---|---|---|
| Root MSE | 0.70325 | R-Square | 0.0098 |
| Dependent Mean | 6.18764 | Adj R-Sq | 0.0060 |
| Coeff Var | 11.36543 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 6.43903 | 0.16251 | 39.62 | <.0001 | 0 |
| anxiety1 | 1 | -0.07302 | 0.04548 | -1.61 | 0.1096 | -0.09907 |

| | | | |
|---|---|---|---|
| Root MSE | 0.69856 | R-Square | 0.0230 |
| Dependent Mean | 6.18764 | Adj R-Sq | 0.0192 |
| Coeff Var | 11.28955 | | |

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 5.55303 | 0.26015 | 21.35 | <.0001 | 0 |
| balance1 | 1 | 0.19900 | 0.08045 | 2.47 | 0.0140 | 0.15164 |

## - Multiple Regression Model

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 26.09998 | 5.22000 | 13.39 | <.0001 |
| Error | 254 | 99.04216 | 0.38993 | | |
| Corrected Total | 259 | 125.14215 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.62444 | R-Square | 0.2086 |
| Dependent Mean | 6.19934 | Adj R-Sq | 0.1930 |
| Coeff Var | 10.07275 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate |
|---|---|---|---|---|---|---|
| Intercept | 1 | 6.58860 | 0.36839 | 17.88 | <.0001 | 0 |
| age1 | 1 | 0.00968 | 0.01201 | 0.81 | 0.4209 | 0.10604 |
| involv1 | 1 | -0.02252 | 0.01782 | -1.26 | 0.2077 | -0.16585 |
| avoid1 | 1 | -0.37929 | 0.05099 | -7.44 | <.0001 | -0.42558 |
| anxiety1 | 1 | -0.02917 | 0.04118 | -0.71 | 0.4794 | -0.04008 |
| balance1 | 1 | 0.15779 | 0.07222 | 2.18 | 0.0298 | 0.12244 |

## - Stepwise Regression

```
                         The STEPWISE Procedure
                             Model: MODEL1
                      Dependent Variable: relsat1

       Variable avoid1 Entered: R-Square = 0.1859 and C(p) = 5.2618

                          Analysis of Variance
                                Sum of          Mean
Source                DF       Squares        Square    F Value    Pr > F

Model                  1      23.26838      23.26838      58.93    <.0001
Error                258     101.87377       0.39486
Corrected Total      259     125.14215


                 Parameter     Standard
      Variable    Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    7.16170      0.13128   1175.05474  2975.88  <.0001
      avoid1      -0.38430      0.05006     23.26838    58.93  <.0001

--------------------------------------------------------------------------------

                        Forward Selection: Step 2


      Variable balance1 Entered: R-Square = 0.2006 and C(p) = 2.5550


                          Analysis of Variance

                                Sum of          Mean
Source                DF       Squares        Square    F Value    Pr > F

Model                  2      25.10370      12.55185      32.25    <.0001
Error                257     100.03845       0.38925
Corrected Total      259     125.14215


                 Parameter     Standard
      Variable    Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    6.64236      0.27239    231.47755   594.67  <.0001
      avoid1      -0.37623      0.04984     22.17735    56.97  <.0001
      balance1     0.15650      0.07207      1.83533     4.71  0.0308

--------------------------------------------------------------------------------
```

Forward Selection: Step 3

Variable involv1 Entered: R-Square = 0.2049 and C(p) = 3.1662

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 25.64526 | 8.54842 | 21.99 | <.0001 |
| Error | 256 | 99.49689 | 0.38866 | | |
| Corrected Total | 259 | 125.14215 | | | |

Forward Selection: Step 3

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.65704 | 0.27246 | 232.01743 | 596.97 | <.0001 |
| involv1 | -0.00894 | 0.00757 | 0.54156 | 1.39 | 0.2389 |
| avoid1 | -0.37635 | 0.04981 | 22.19209 | 57.10 | <.0001 |
| balance1 | 0.15973 | 0.07207 | 1.90925 | 4.91 | 0.0275 |

---

Forward Selection: Step 4

Variable age1 Entered: R-Square = 0.2070 and C(p) = 4.5018

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 25.90433 | 6.47608 | 16.64 | <.0001 |
| Error | 255 | 99.23782 | 0.38917 | | |
| Corrected Total | 259 | 125.14215 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.49015 | 0.34084 | 141.10713 | 362.59 | <.0001 |
| age1 | 0.00979 | 0.01200 | 0.25907 | 0.67 | 0.4153 |
| involv1 | -0.02207 | 0.01780 | 0.59886 | 1.54 | 0.2159 |
| avoid1 | -0.38349 | 0.05060 | 22.35345 | 57.44 | <.0001 |
| balance1 | 0.15938 | 0.07212 | 1.90077 | 4.88 | 0.0280 |

---

Forward Selection: Step 5

Variable anxiety1 Entered: R-Square = 0.2086 and C(p) = 6.0000

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 26.09998 | 5.22000 | 13.39 | <.0001 |
| Error | 254 | 99.04216 | 0.38993 | | |
| Corrected Total | 259 | 125.14215 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.58860 | 0.36839 | 124.72426 | 319.86 | <.0001 |
| age1 | 0.00968 | 0.01201 | 0.25342 | 0.65 | 0.4209 |
| involv1 | -0.02252 | 0.01782 | 0.62223 | 1.60 | 0.2077 |
| avoid1 | -0.37929 | 0.05099 | 21.57172 | 55.32 | <.0001 |
| anxiety1 | -0.02917 | 0.04118 | 0.19566 | 0.50 | 0.4794 |
| balance1 | 0.15779 | 0.07222 | 1.86114 | 4.77 | 0.0298 |

-------------------------------------------------------------------------------

All variables have been entered into the model.

Summary of Forward Selection

| Step | Variable Entered | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
|---|---|---|---|---|---|---|---|
| 1 | avoid1 | 1 | 0.1859 | 0.1859 | 5.2618 | 58.93 | <.0001 |
| 2 | balance1 | 2 | 0.0147 | 0.2006 | 2.5550 | 4.71 | 0.0308 |
| 3 | involv1 | 3 | 0.0043 | 0.2049 | 3.1662 | 1.39 | 0.2389 |
| 4 | age1 | 4 | 0.0021 | 0.2070 | 4.5018 | 0.67 | 0.4153 |
| 5 | anxiety1 | 5 | 0.0016 | 0.2086 | 6.0000 | 0.50 | 0.4794 |

**- Stepwise Regression (Alternative)**

```
                            The STEPWISE Procedure
                                Model: MODEL1
                         Dependent Variable: relsat1

             Number of Observations Read                          262
             Number of Observations Used                          260
             Number of Observations with Missing Values             2

                      Maximum R-Square Improvement: Step 1
          Variable avoid1 Entered: R-Square = 0.1859 and C(p) = 5.2618

                              Analysis of Variance
                                     Sum of           Mean
 Source                    DF       Squares         Square    F Value    Pr > F
 Model                      1      23.26838       23.26838      58.93    <.0001
 Error                    258     101.87377        0.39486
 Corrected Total          259     125.14215

                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F
       Intercept     7.16170      0.13128   1175.05474  2975.88  <.0001
       avoid1       -0.38430      0.05006     23.26838    58.93  <.0001

              The above model is the best  1-variable model found.

--------------------------------------------------------------------------------
                      Maximum R-Square Improvement: Step 2
         Variable balance1 Entered: R-Square = 0.2006 and C(p) = 2.5550

                              Analysis of Variance
                                     Sum of           Mean
 Source                    DF       Squares         Square    F Value    Pr > F
 Model                      2      25.10370       12.55185      32.25    <.0001
 Error                    257     100.03845        0.38925
 Corrected Total          259     125.14215

                   Parameter     Standard
       Variable     Estimate        Error   Type II SS  F Value  Pr > F

       Intercept     6.64236      0.27239    231.47755   594.67  <.0001
       avoid1       -0.37623      0.04984     22.17735    56.97  <.0001
       balance1      0.15650      0.07207      1.83533     4.71  0.0308

             The above model is the best  2-variable model found.
--------------------------------------------------------------------------------
```

Maximum R-Square Improvement: Step 3
Variable involv1 Entered: R-Square = 0.2049 and C(p) = 3.1662

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 25.64526 | 8.54842 | 21.99 | <.0001 |
| Error | 256 | 99.49689 | 0.38866 | | |
| Corrected Total | 259 | 125.14215 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.65704 | 0.27246 | 232.01743 | 596.97 | <.0001 |
| involv1 | -0.00894 | 0.00757 | 0.54156 | 1.39 | 0.2389 |
| avoid1 | -0.37635 | 0.04981 | 22.19209 | 57.10 | <.0001 |
| balance1 | 0.15973 | 0.07207 | 1.90925 | 4.91 | 0.0275 |

The above model is the best 3-variable model found.

---

Maximum R-Square Improvement: Step 4
Variable age1 Entered: R-Square = 0.2070 and C(p) = 4.5018

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 25.90433 | 6.47608 | 16.64 | <.0001 |
| Error | 255 | 99.23782 | 0.38917 | | |
| Corrected Total | 259 | 125.14215 | | | |

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 6.49015 | 0.34084 | 141.10713 | 362.59 | <.0001 |
| age1 | 0.00979 | 0.01200 | 0.25907 | 0.67 | 0.4153 |
| involv1 | -0.02207 | 0.01780 | 0.59886 | 1.54 | 0.2159 |
| avoid1 | -0.38349 | 0.05060 | 22.35345 | 57.44 | <.0001 |
| balance1 | 0.15938 | 0.07212 | 1.90077 | 4.88 | 0.0280 |

The above model is the best 4-variable model found.

---

## • 14. 13 Testing the Significance of $R^2$

- The critical question is, Does the set of variables taken together predict $Y$ at better-than-chance levels?

$$F = \frac{(N - p - 1)R^2}{p(1 - R^2)}$$

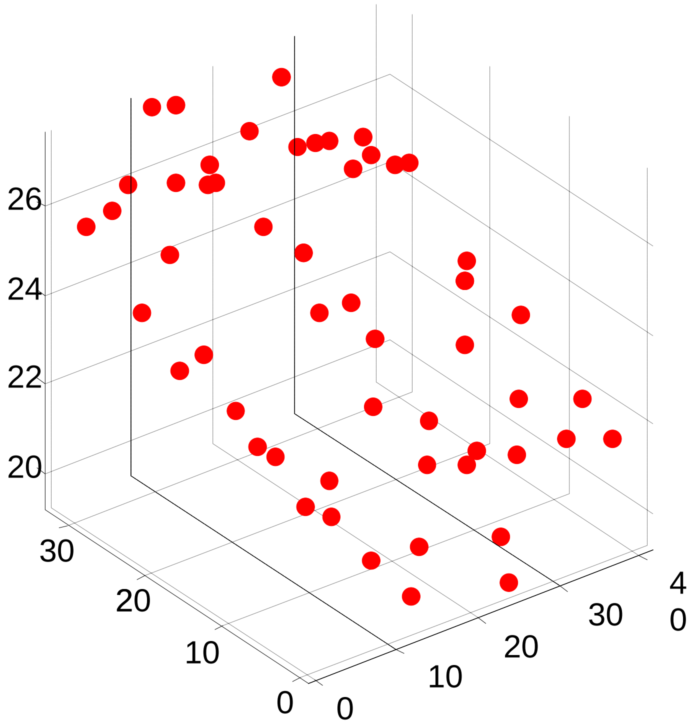
that can be evaluated in a $F$ distribution on $p$ and $N - p - 1$ degrees of freedom

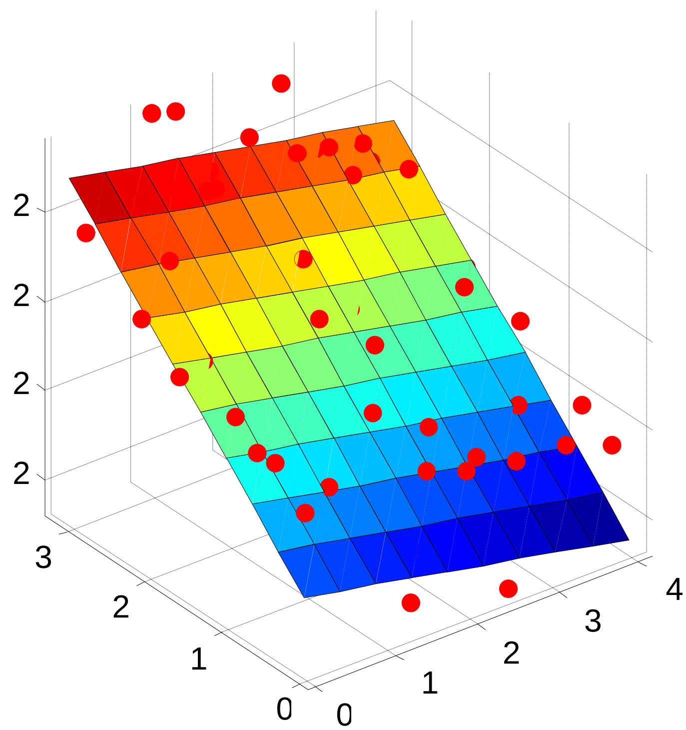
## • 14. 14 Interpretation of Multiple Regression

- The regression coefficient for each predictor represents the amount of variance in $Y$ explained above and beyond the previous predictors

- The regression coefficient for each predictor represents the amount of explained variance in $Y$ that is leftover (residual) after being regressed on the previous predictors

- A *multiple correlation* can be thought of as a simple Pearson correlation between the criterion (outcome) variable and the best linear combination of predictors


26
24
22
20
30
20
10
0
0
10
20
30
4
0


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

## • 14. 15 Partial and Semipartial Correlation

- A partial correlation is the correlation $r_{yx.z}$ between two variables (*X* and *Y*) with one (*Z*) or more variables partialled out of both *X* and *Y*. Also, it is the correlation between the two sets of residuals formed from the prediction of the original variables by one or more other variables

- A semipartial correlation is the correlation $r_{y(x.z)}$ between the criterion and a partialled predictor variable

- Whereas the partial correlation $r_{yx.z}$ has variable *Z* partialled out of both the criterion *Y* and predictor *X*, the semipartial correlation $r_{y(x.z)}$ has variable *Z* partialled out of the predictor *X* only

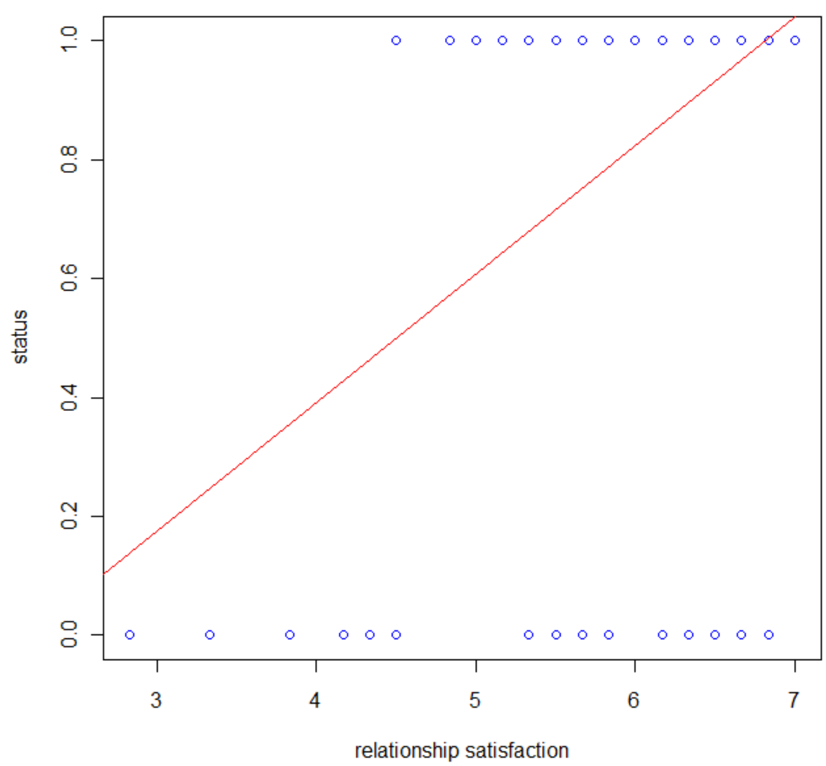
## • 14. 16 Logistic Regression

- *Logistic regression* is the technique for fitting a regression surface to data in which the dependent variable is dichotomus (e.g., absent vs. present, married vs. divorced, graduated vs. dropout)

- Logistic regression describes the relationship between a *dichotomous* response variable and a set of explanatory variables. The explanatory variables may be continuous or (with dummy variables) discrete

- *Conditional means* are the means of the outcome score (the mean of 0s and 1s) associated with each value of the predictor. They represent the proportion of people with a given value of $X$ who have a value of 1 in $Y$. In the example (see graph), they would be the proportion of people with a given value of *relationship satisfaction* who stayed together ($Y$ =1)

- *Relationship Status* as a function of *Relationship Satisfaction*



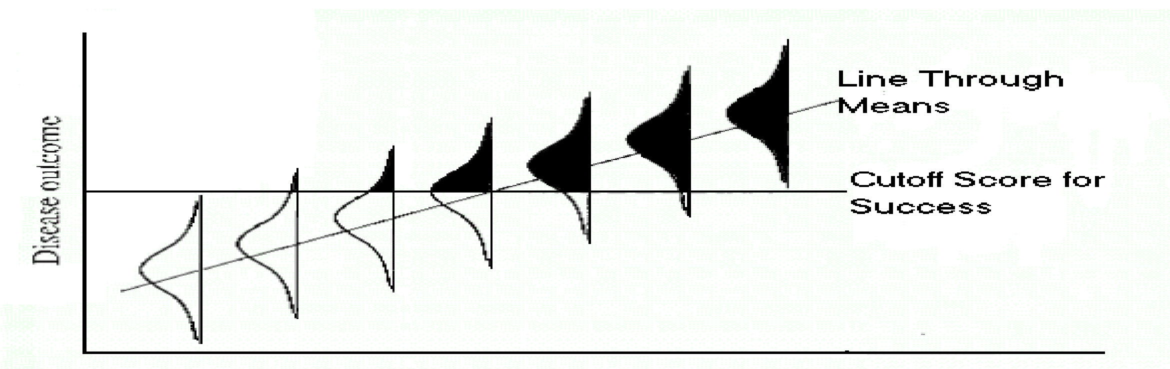

- It seems that the proportion of people who stay together ($Y$ =1) is much higher when the relationship satisfaction is high (as one would expect)
- The standard regression line represents the regression line that fits the *probability* of *staying together* as a function of *relationship satisfaction*. However,

- For many values of *relationship satisfaction*, the predicted probability would be outside the range 0-1, which is impossible. Thus, standard linear regression is not optimal

- There is a violation of homogeneity of variance (e.g., low or high values of *X* are associated with either 0 or 1 in *Y*, but mid values of *X* are split between 0 and 1)

- The true relationship between *X* and *Y* is not likely linear (e.g., differences in *X* near the center of the scale will lead to noticeably larger differences in *Y*, compared with differences in *X* at either end of the scale). An *S*-shaped (*sigmoidal*) curve is more likely

- Logistic regression can be thought of as applying linear regression to *censored data* (data that take values of 0 or 1 when the scores are below or above a given cutoff



## • 14. 17 Odd Ratios

- One useful way to think about data like these is in terms of ***probabilities***. We talk about the probability of staying together. But, it is equally possible to think in terms of the ***odds*** of staying together, and it works much better (in statistically terms)

odds *staying together* = Number stayed together /Number broke up

or, equivalently,

odds *staying together* = $p$(*staying together*)/(1–$p$(*staying together*))

- If odds *staying together* are given as above, then

$p$(*staying together*) = 1/(1+ odds)

- One reason why it is useful to work with odds is because odds are a logarithmic function of $X$, whereas, probabilities are a sigmoidal function of $X$

- Advantages of a logarithmic function are that it can increase without a ceiling, makes computations easy, and if we plot the *log* of the odds, the relationship will be linear

- ***log odds*** allow the relationship to become linear

*log odds staying together* = ln(odds) = ln($p$/1–$p$)

where ln is the natural logarithm ($\ln(x) = \log_e(x)$, rather than, say, $\log_{10}$)

- the log odds will be positive for odds greater than 1/1 and negative for odds less than 1/1 (undefined for odds = 0)

- this is often called the **logit** or the **logit transform**

- we will work with the logit, and will solve for the equation

**log($p$/(1-$p$)) = log(odds) = logit = $b_0$ + $b_1$ $X$**, or

**log($p$/(1-$p$)) = log(odds) = logit = $b_0$ + $b_1$ *relationship satisfaction***

- This is now a linear equation because we are using logs. The equation would not be linear in terms of odds or probabilities

$b_0$ is the intercept (not very meaningful here)

$b_1$ is the slope, and is the change in the *log* odds for a one unit change in $X$

- ## 14. 18 Example

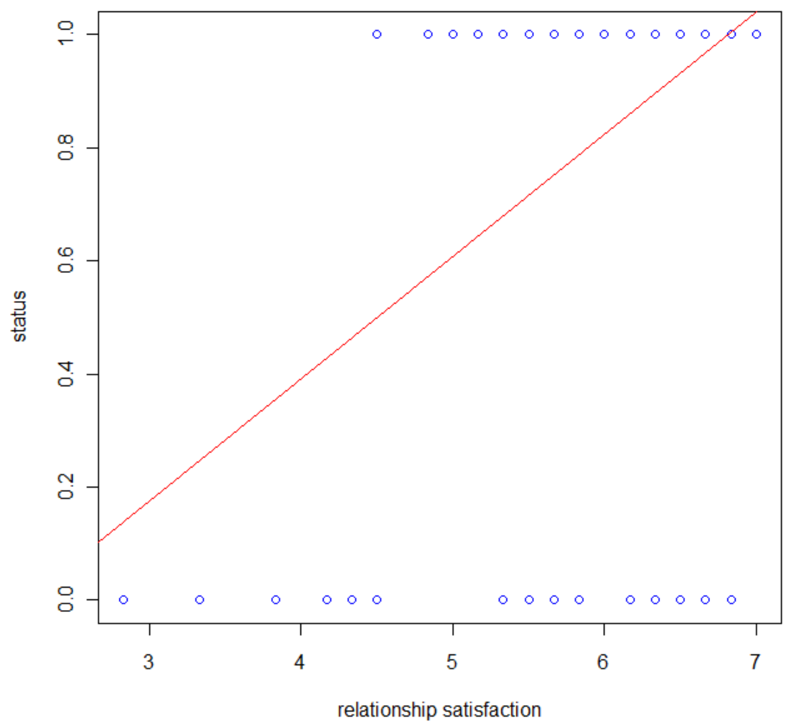
newstatus2=0

| Variable | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|
| age1 | 22 | 19.9393182 | 1.6256914 | 18.2500000 | 24.2500000 |
| involv1 | 22 | 0.7083182 | 0.8558104 | 0.0420000 | 3.0830000 |
| avoid1 | 22 | 2.8711818 | 0.7208063 | 1.6670000 | 4.1670000 |
| anxiety1 | 22 | 3.6364091 | 0.9315474 | 1.8330000 | 5.2220000 |
| relsat1 | 22 | 5.3635909 | 1.0834194 | 2.8330000 | 6.8330000 |
| balance1 | 22 | 3.1250000 | 0.7144345 | 1.2500000 | 4.0000000 |

newstatus2=1

| Variable | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|
| age1 | 139 | 21.9373597 | 7.1130440 | 18.0000000 | 74.2500000 |
| involv1 | 138 | 2.3610652 | 4.7877981 | 0.1670000 | 35.0830000 |
| avoid1 | 140 | 2.4496000 | 0.7992280 | 1.2780000 | 4.8890000 |
| anxiety1 | 140 | 3.5095214 | 1.0655886 | 1.3330000 | 6.1670000 |
| relsat1 | 140 | 6.3214000 | 0.5510158 | 4.5000000 | 7.0000000 |
| balance1 | 140 | 3.1857143 | 0.5392603 | 1.0000000 | 4.0000000 |

- *Relationship Status* as a function of *Relationship Satisfaction*



- Let's start with the simple prediction of *relationship status* as a function of *relationship satisfaction*. Let $p$ = probability of staying together and $1 - p$ = the probability of breaking up, we will solve for the equation

$$\log(p/(1-p)) = \log(\text{odds}) = \text{logit} = b_0 + b_1 \textit{ relationship satisfaction}$$

- This is a linear equation because we are using logs

$b_0$ is the intercept (not very meaningful here)

$b_1$ is the slope, and is the change in the *log* odds for a one unit change in $X$ (amount of increase in the *log odds* for a one unit increase in *relationship satisfaction*)

- The estimation of parameters in logistic regression is more complicated than in simple regression. For logistic regression, the typical method is maximum likelihood, solving for the regression coefficients iteratively (difficult to do by hand)

## - Logistic Regression Model (one predictor)

Logistic Regression (one predictor)

Response Profile

| Ordered Value | newstatus2 | Total Frequency |
|---|---|---|
| 1 | 1 | 140 |
| 2 | 0 | 22 |

Probability modeled is newstatus2=1.

Model Fit Statistics

| Criterion | Intercept Only | Intercept and Covariates |
|---|---|---|
| AIC | 130.715 | 104.567 |
| SC | 133.803 | 110.742 |
| -2 Log L | 128.715 | 100.567 |

R-Square 0.1595 Max-rescaled R-Square 0.2909

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 28.1484 | 1 | <.0001 |
| Score | 33.5204 | 1 | <.0001 |
| Wald | 19.2969 | 1 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -7.9982 | 2.2062 | 13.1425 | 0.0003 |
| relsat1 | 1 | 1.6568 | 0.3772 | 19.2969 | <.0001 |

**Equation: log(odds) = -7.9982 + 1.6568 * *relationship satisfaction***
**odds = $e^{b1}$; $e^{1.6568}$ = 5.243**

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| relsat1 | 5.243 | 2.503 | 10.980 |

Classification Table

| | Correct | | Incorrect | | Percentages | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prob Level | Event | Non-Event | Event | Non-Event | Correct | Sensi-tivity | Speci-ficity | False POS | False NEG |
| 0.500 | 138 | 6 | 16 | 2 | 88.9 | 98.6 | 27.3 | 10.4 | 25.0 |

### - Logistic Regression Model (multiple predictors)

Response Profile

| Ordered Value | newstatus2 | Total Frequency |
|---|---|---|
| 1 | 1 | 138 |
| 2 | 0 | 22 |

| Criterion | Intercept Only | Intercept and Covariates |
|---|---|---|
| AIC | 130.128 | 102.518 |
| SC | 133.203 | 124.044 |
| -2 Log L | 128.128 | 88.518 |

R-Square 0.2193 Max-rescaled R-Square 0.3980

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 39.6097 | 6 | <.0001 |
| Score | 40.4877 | 6 | <.0001 |
| Wald | 20.3302 | 6 | 0.0024 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -7.5709 | 4.6702 | 2.6280 | 0.1050 |
| age1 | 1 | -0.0017 | 0.1697 | 0.0001 | 0.9921 |
| involv1 | 1 | 0.7221 | 0.4048 | 3.1825 | 0.0744 |
| avoid1 | 1 | -0.2368 | 0.3811 | 0.3860 | 0.5344 |
| anxiety1 | 1 | -0.0746 | 0.2910 | 0.0657 | 0.7977 |
| relsat1 | 1 | 1.7995 | 0.4949 | 13.2227 | 0.0003 |
| balance1 | 1 | -0.3916 | 0.5329 | 0.5400 | 0.4624 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| age1 | 0.998 | 0.716 | 1.392 |
| involv1 | 2.059 | 0.931 | 4.551 |
| avoid1 | 0.789 | 0.374 | 1.666 |
| anxiety1 | 0.928 | 0.525 | 1.642 |
| relsat1 | 6.047 | 2.292 | 15.949 |
| balance1 | 0.676 | 0.238 | 1.921 |

Classification Table

| | Correct | | Incorrect | | Percentages | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Prob Level | Event | Non-Event | Event | Non-Event | Correct | Sensi-tivity | Speci-ficity | False POS | False NEG |
| 0.500 | 133 | 6 | 16 | 5 | 86.9 | 96.4 | 27.3 | 10.7 | 45.5 |