# Section VIII

# Correlation &Linear regression for continuous outcomes

## VIIIa – simple bivariate regression

## VIIIb-multiple regression

# bivariate & multivariate continuous data- regression

## Ex: Riddle, J. of Perinatology (2006) 26, 556–561

**50th percentile for birth weight (BW) in g as a function of gestational age**

$$BW(g)=42\exp(0.1155 \text{ gest age})$$

**Or**

$$\text{Log}_e(BW)=3.74 + 0.1155 \text{ gest age}$$

**In general:**

$$BW = A \exp(B \text{ gest age}),$$

**A & B change for different percentiles**

# Section VIIIa

# Simple bivariate regression

**Statistics for** bivariate **continuous data –** regression and correlation

Measures of association
correlation coefficient (r in sample, ρ in population)
slope (b in sample, β in population)

Measure of location
intercept (a or $b_0$ in sample, α or $\beta_0$ in population)

Measures of fit
Squared correlation ($R^2$)
Residual SD ($SD_e$ in sample, $\sigma_e$ in population)

---

## Statistics for Describing a Bivariate (two variable) Relationship between two continuous variables

We first consider the simplest case where we relate a continuous measured variable X to another continuous measured variable Y where both X and Y are measured on the same persons.

In the example below, X is age in years and Y is systolic blood pressure (SBP) in mm Hg for adult females. In examining the relationship between X and Y, the first step is to make a scatter plot (also called a scattergram).

Now it is often (but not always) the case, that there is a roughly **linear** relationship between X and Y. That is, as X doubles, Y may double (or -Y may double). By a linear relationship we mean a relationship of the form

$$Y = a + b\,X + \text{error}$$

That is, the relationship is expressed with an **equation w**here a and b are constants estimating population values α and β. This equation says that, for every unit X increases, Y increases by an amount b. When X is zero, Y is equal to a. The constant b is called the **slope or the rate**, and the constant a is called the **intercept.**

If the relationship between X and Y is (at least approximately) linear, then we can summarize the relationship by four statistics:
the slope, b
the intercept, a (sometimes denoted $b_0$)
the (Pearson) correlation coefficient, r or the squared correlation ($R^2$)
the residual standard deviation denoted $S_e$ or $SD_e$ (also called the root mean square error)

The correlation r and $SD_e$, the residual SD, are defined below.

By definition, r is defined as

```
r =    Σ (Y deviations from mean)(X deviations from mean)
```

$$(n-1)\ SD_y\ SD_x$$

$$= \frac{\Sigma\ (Y - \overline{Y})(X - \overline{X})}{(n-1)\ SD_y\ SD_x}$$

where the subscripted SDs refer to the standard deviations of y and x respectively. This correlation coefficient is called the Pearson correlation coefficient or the product moment correlation coefficient.

If most of the XY products are positive, r is positive and, on average, Y **increases** as X increases.
If most of the XY products are negative, r is negative and, on average, Y **decreases** as X increases.

Not surprisingly, r and b are related by the formula

$b = r\ SD_y/SD_x$ or $r = b\ SD_x/SD_y$ (r is a “slope” in SD units)

Note that b and r have the same sign. If r is zero, b is also zero.

**How the slope and intercept are estimated** (short version)
**Definition of the residual standard deviation** ($SD_e$)

For every X value, there is a corresponding Y value. If we draw a straight lie through the scatter plot, for every X value there will also be a value on the line which we will denote **$\hat{Y}$** ("Y hat"). $\hat{Y}$ is the predicted (not actually observed) value of Y based on the line. The residual error, denoted "e" is the difference between the observed and predicted (or "expected") Y values.

residual error = $e = Y - \hat{Y}$

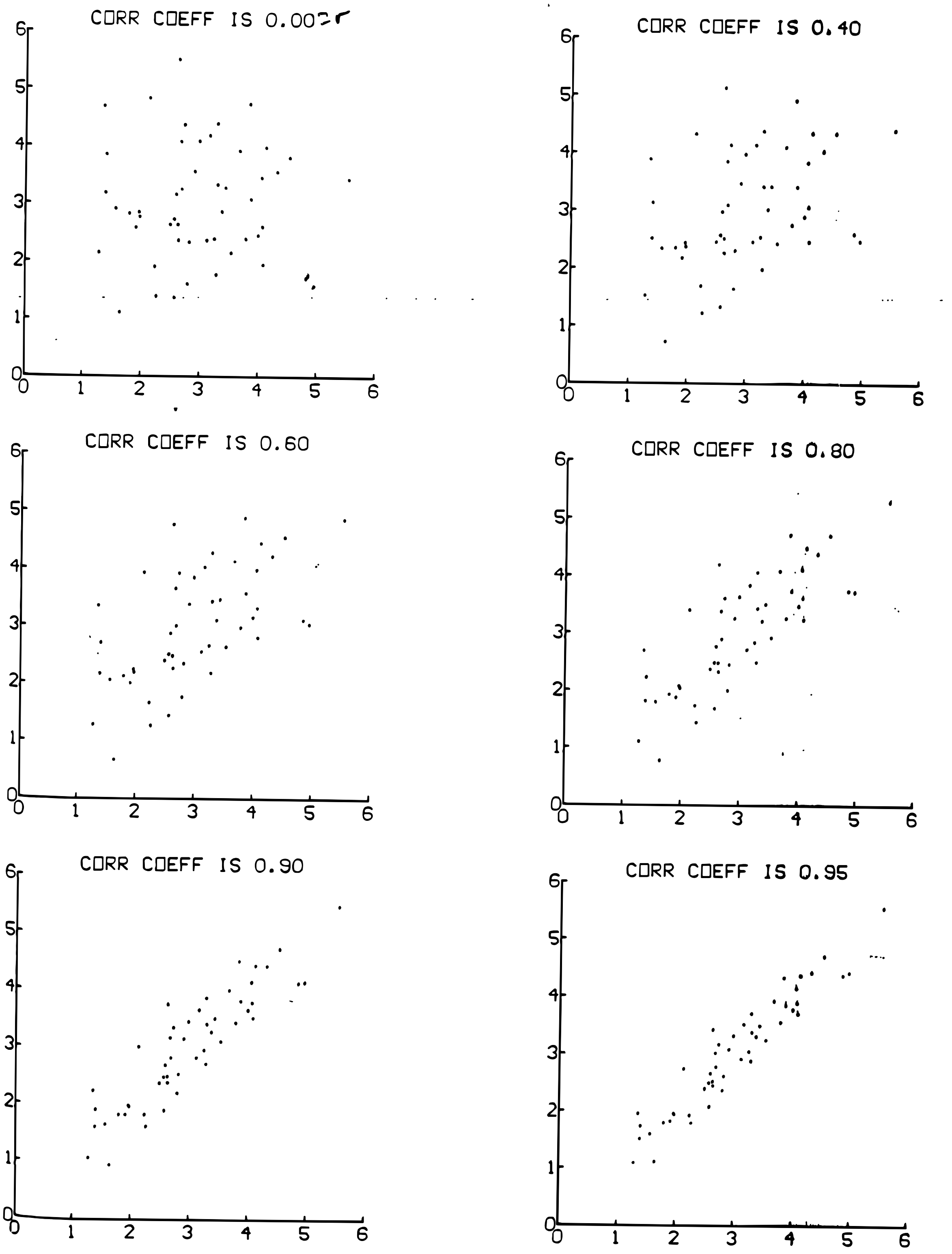
The slope, b and the intercept, a, are chosen such that the quantity

RSS = residual sum of squares = $\Sigma\ e^2 = \Sigma\ (Y - \hat{Y})^2$

is minimized. That is, a and b are chosen so that, on average, the line is as close to the observations as possible.

When the slope and intercept are chosen this way, the average value of e (the average residual error) is zero and the standard deviation of the residual errors is given by
$SD_e = \sqrt{}RSS/(n-2)$= SD of the residual errors, e = Root mean square error =RMSE

Figure 6. The correlation coefficient—six positive values..The diagrams are scaled so that the average equals 3 and the SD equals 1, horizontally and vertically. The clustering around a line is measured by the correlation coefficient.



## Data for the simple regression example: age vs SBP

**Age vs SBP in women**

**predicted SBP (mmHg) = 81.5 + 1.22 age, r=0.72, $R^2$=0.515**

| | **X** | **Y** | **Predicted $Y=\hat{Y}$** | **e = error** |
|---|---|---|---|---|
| **patient** | **age (yrs)** | **SBP (mmHg)** | **predicted SBP (mmHg)** | **residual error=e (mmHg)** |
| 1 | 22 | 131 | 108.42 | 22.58 |
| 2 | 23 | 128 | 109.65 | 18.35 |
| 3 | 24 | 116 | 110.87 | 5.13 |
| 4 | 27 | 106 | 114.53 | -8.53 |
| 5 | 28 | 114 | 115.76 | -1.76 |
| 6 | 29 | 123 | 116.98 | 6.02 |
| 7 | 30 | 117 | 118.20 | -1.20 |
| 8 | 32 | 122 | 120.64 | 1.36 |
| 9 | 33 | 99 | 121.87 | -22.87 |
| 10 | 35 | 121 | 124.31 | -3.31 |
| 11 | 40 | 147 | 130.42 | 16.58 |
| 12 | 41 | 139 | 131.64 | 7.36 |
| 13 | 41 | 171 | 131.64 | 39.36 |
| 14 | 46 | 137 | 137.75 | -0.75 |
| 15 | 47 | 111 | 138.97 | -27.97 |
| 16 | 48 | 115 | 140.20 | -25.20 |
| 17 | 49 | 133 | 141.42 | -8.42 |
| 18 | 49 | 128 | 141.42 | -13.42 |
| 19 | 50 | 183 | 142.64 | 40.36 |
| 20 | 51 | 130 | 143.86 | -13.86 |
| 21 | 51 | 133 | 143.86 | -10.86 |
| 22 | 51 | 144 | 143.86 | 0.14 |
| 23 | 52 | 128 | 145.08 | -17.08 |
| 24 | 54 | 105 | 147.53 | -42.53 |
| 25 | 56 | 145 | 149.97 | -4.97 |
| 26 | 57 | 141 | 151.19 | -10.19 |
| 27 | 58 | 153 | 152.42 | 0.58 |
| 28 | 59 | 157 | 153.64 | 3.36 |
| 29 | 63 | 155 | 158.53 | -3.53 |
| 30 | 67 | 176 | 163.41 | 12.59 |
| 31 | 71 | 172 | 168.30 | 3.70 |
| 32 | 77 | 178 | 175.63 | 2.37 |
| 33 | 81 | 217 | 180.52 | 36.48 |
| | | | | |
| mean | 46.7 | 138.6 | 138.6 | 0.0 |
| SD | 15.5 | 26.4 | 18.9 | **18.3** |
| | $SD_x$ | $SD_y$ | | **$SD_e$ = $S_e$=Root MSE** |

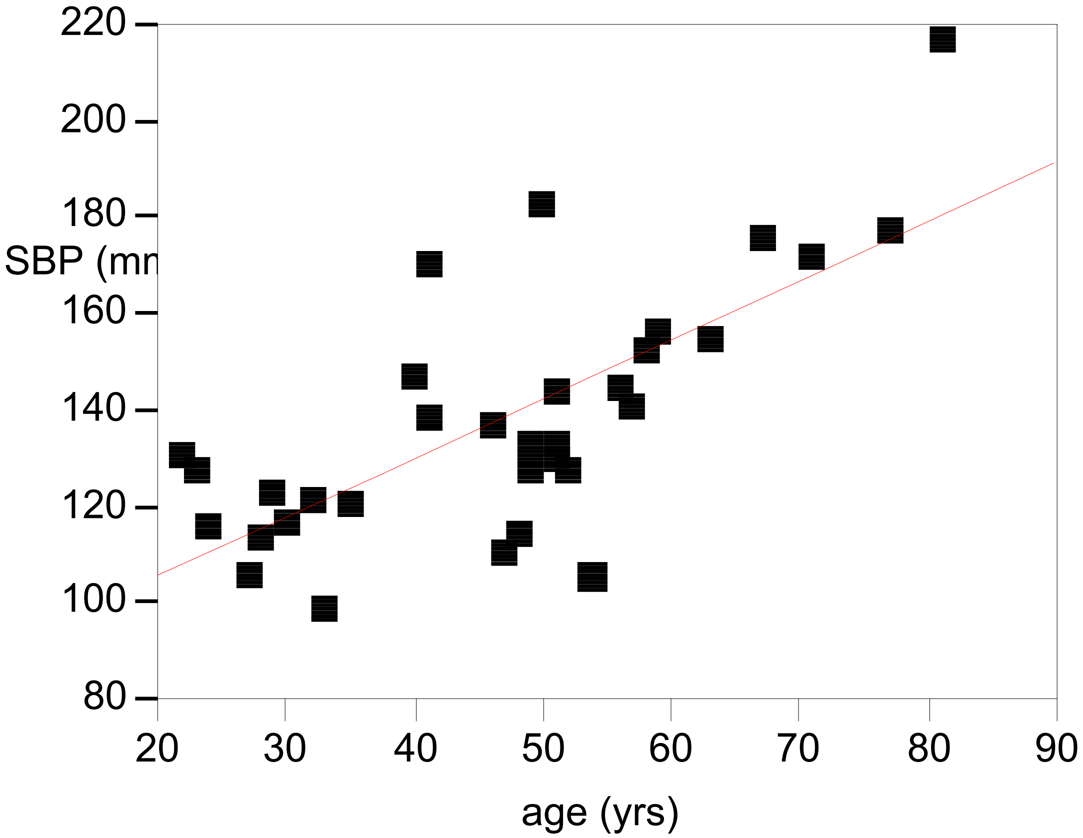
Mean error is always zero

## Regression Example: age versus systolic blood pressure (SBP)

In this example SBP = 81.5 + 1.22 age + error

**For every year increase in age, SBP increases on average by 1.22 mm Hg/year**

**Bivariate Fit of *y=*SBP (mmHg) By *x=*age (yrs) *(adult females)***



(Sample) Intercept = a = 81.5 mm Hg (intercept sometimes denoted $b_0$, not a)
(Sample) Slope = b= 1.22 mm Hg / year

(Sample) Residual error SD = $SD_e$ = $S_e$ = 18.6 mmHg
(Also called the RMSE = root mean square error)

Sample squared correlation = $R^2$ = 0.515
Sample correlation = r = √0.515 = 0.718

| Variable | SD |
|---|---|
| age | 15.5 years |
| SBP | 26.4 mm Hg |

*JMP output*

```
Linear Fit
   SBP = 81.516752 + 1.2224041 age

Summary of Fit
                   Rsquare=R^2                  0.515307
                   RSquare Adj                  0.499671
     Root Mean Square Error=SD_e                18.63894
          Mean of Response=Y                    138.6364
   Observations (or Sum Wgts)=n                   33
```

$R^2$=51% of the variation in SBP is accounted for by age.

```
Lack Of Fit
   Source       DF   Sum of Squares   Mean Square    F Ratio
  Lack Of Fit   27     10136.544        375.428      2.3717
  Pure Error     4       633.167        158.292     Prob > F
  Total Error   31     10769.710                     0.2084
```

Reject linear assumption if this p is small

```
Analysis of Variance
  Source    DF   Sum of Squares   Mean Square   F Ratio
  Model      1     11449.926        11449.9     32.9580
  Error     31     10769.710          347.4     Prob > F
 C. Total   32     22219.636                    <.0001
```

$SD_e^2$

```
Parameter Estimates
   Term        Estimate    Std Error   t Ratio   Prob>|t|
  Intercept     81.517      10.465      7.79     <.0001
  age (yrs)      1.222      0.2129      5.74     <.0001
```

*p values*

*Slope = b*

*P value for age vs SBP correlation*

$r = \sqrt{0.5153} = 0.7178$

## Slope , correlation & $SD_e$ – key facts

*** The slope b is the rate of change in Y for a unit change in X. It has units of y/x. The correlation (r) is dimensionless and is the change in Y in SD units for a one SD change in X. $SD_e$ has units of Y.**

When r =1.0 or r = -1.0, $SD_e$ is zero (perfect fit)

The intercept, slope and correlation are not very meaningful when the relation between X and Y is systematically nonlinear (see below)

*** Slope = correlation x $(SD_y/SD_x)$**

$$b = r\,(SD_y/SD_x) \qquad 1.22=0.7178(26.4/15.5)$$

where $SD_y$ is the SD of the y variable, $SD_x$ is the SD of the X variable.

$$* \qquad r = b\,(SD_x/SD_y) \qquad 0.7178=1.22(15.5/26.4)$$

$$r = b\,SD_x/\sqrt{b^2\,SD_x^2 + SD_e^2}$$

where $SD_e$ is the residual error and $SD_x$ is the SD of the x variable

*** $R^2$ is the proportion of the total (squared) variation in Y that is "accounted for" by X.**

$$R^2 = r^2 = (SD_y^2 - SD_e^2)/SD_y^2 = 1 - (SD_e^2/SD_y^2)$$

$$SD_y\sqrt{(1-r^2)} = SD_e \qquad 26.4\sqrt{(1-0.5153)} = (18.64)$$

$$\text{If } Y = \hat{Y} + e, \quad Var(Y)=Var(\hat{Y}+e)= Var(\hat{Y}) + Var(e)$$

$$\text{So, } Var(\hat{Y}) = Var(Y) - Var(e) = SD_y^2 - SD_e^2$$

$$R^2 = Var(\hat{Y})/Var(Y) = (SD_y^2 - SD_e^2)/ SD_y^2$$

*** Under Gaussian theory, 95% of the errors are within +/- 2 $SD_e$ of their corresponding predicted Y value.**

# Sums of Squares (SS)

**Most regression software also prints out a table such as the one below, the "summary analysis of variance table".**

```
   Summary Analysis of Variance table
 Source     DF   Sum of Squares  Mean Square   F Ratio    p value
 Model       1      11449.926      11449.9     32.9580    0.0001
 Error      31      10769.710        347.4
C. Total    32      22219.636
```

$SD_e^2$

**This table shows how much of the variation in the outcome Y (SBP in this example) is accounted for by the "model", that is, the predictor X variable(s) and how much variation in Y is not accounted for, the "error" variation.**

**For a given dataset, the SD of Y ($SD_y$) and the variance of Y ($=SD_y^2$) is fixed. So the sum of squares (SS) for Y is defined as the sample size (minus 1) times the variance, is also fixed. The SS is the numerator of the variance formula and is a measure of how much Y varies.**

**The table is shown below, for k predictor variables. (In our example above, k=1).**

| | df | Sum of Squares=SS | Mean Square=MS=SS/df |
|---|---|---|---|
| **Model** | **k** | $b^2\sum(x-\bar{x})^2$ *(for k=1)* | $b^2\sum(x-\bar{x})^2/k$ |
| **Error** | **n-k-1** | $\sum e^2 =(n-k-1)SD_e^2$ | $SD_e^2$ |
| **Total** | **n-1** | $\sum(y-\bar{y})^2 =(n-1)SD_y^2$ | $SD_y^2$ |

**In the above, $\bar{y}$ is the mean Y and $\bar{x}$ is the mean x.**

**The $R^2$ value = Model SS /Total SS = 11450/22220=0.515.**
**F = Model SS/ Error SS, the corresonding p value tests that the true β=0.**

## Confidence intervals and prediction intervals from regression models

As previously studied, confidence intervals and prediction intervals are not the same.

Example: In our model:
Predicted SBP = 81.52 + 1.222 age (SD$_e$= 18.6 mm Hg)

For a 50 year old, the predicted SBP is 81.5 + 1.22(50) = **142.6** mm Hg = $\hat{Y}$.

The standard error for this $\hat{Y}$= 142.6 is SE=3.3 mm Hg, so a 95% **confidence interval** for the average SBP in a 50 year old is (136.0 mm Hg, 149.2 mm Hg).

But, the individual standard deviation is 18.9 mm Hg (similar to SD$_e$= 18.6 mm Hg). So a 95% **prediction interval** for individuals is (104.8 mm Hg, 180.4 mm Hg).

The **142.6** is both the estimated mean for all women age 50 and the predicted value for each individual age 50!

The confidence interval (CI) indicates the uncertainty (assuming the model is correct) in estimating the population **mean** SBP for all women age 50 in the target population. The prediction interval (PI) indicates where the middle 95% of individual SBP values will fall for all women age 50 in the target population. **The PI may be more clinically relevant as it gives the uncertainty in the prediction for one individual.**

For CI For PI

| patient | age | SBP | Predicted SBP=Ŷ | **SE** for **Pred** SBP | **SD** for Pred SBP |
|---|---|---|---|---|---|
| | (yrs) | (mmHg) | (mmHg) | (mmHg) | (mmHg) |
| 1 | 22 | 131 | 108.4 | 6.2 | 19.6 |
| 2 | 23 | 128 | 109.6 | 6.0 | 19.6 |
| 3 | 24 | 116 | 110.9 | 5.8 | 19.5 |
| 4 | 27 | 106 | 114.5 | 5.3 | 19.4 |
| 5 | 28 | 114 | 115.7 | 5.1 | 19.3 |
| 6 | 29 | 123 | 117.0 | 5.0 | 19.3 |
| 7 | 30 | 117 | 118.2 | 4.8 | 19.3 |
| 8 | 32 | 122 | 120.6 | 4.5 | 19.2 |
| 9 | 33 | 99 | 121.9 | 4.4 | 19.1 |
| 10 | 35 | 121 | 124.3 | 4.1 | 19.1 |
| 11 | 40 | 147 | 130.4 | 3.5 | 19.0 |
| 12 | 41 | 139 | 131.6 | 3.5 | 19.0 |
| 13 | 41 | 171 | 131.6 | 3.5 | 19.0 |
| 14 | 46 | 137 | 137.7 | 3.2 | 18.9 |
| 15 | 47 | 111 | 139.0 | 3.2 | 18.9 |
| 16 | 48 | 115 | 140.2 | 3.3 | 18.9 |
| 17 | 49 | 133 | 141.4 | 3.3 | 18.9 |
| 18 | 49 | 128 | 141.4 | 3.3 | 18.9 |
| 19 | 50 | 183 | 142.6 | 3.3 | 18.9 |
| 20 | 51 | 130 | 143.9 | 3.4 | 18.9 |
| 21 | 51 | 133 | 143.9 | 3.4 | 18.9 |
| 22 | 51 | 144 | 143.9 | 3.4 | 18.9 |
| 23 | 52 | 128 | 145.1 | 3.4 | 19.0 |
| 24 | 54 | 105 | 147.5 | 3.6 | 19.0 |
| 25 | 56 | 145 | 150.0 | 3.8 | 19.0 |
| 26 | 57 | 141 | 151.2 | 3.9 | 19.0 |
| 27 | 58 | 153 | 152.4 | 4.0 | 19.1 |
| 28 | 59 | 157 | 153.6 | 4.2 | 19.1 |
| 29 | 63 | 155 | 158.5 | 4.7 | 19.2 |
| 30 | 67 | 176 | 163.4 | 5.4 | 19.4 |
| 31 | 71 | 172 | 168.3 | 6.1 | 19.6 |
| 32 | 77 | 178 | 175.6 | 7.2 | 20.0 |
| 33 | 81 | 217 | 180.5 | 8.0 | 20.3 |

# How big should $R^2$ be?

While $R^2$ is the proportion of Ys variation accounted for by the Xs (in an observational study), there is no universal rule saying that $R^2$ must be at least 0.30 or 0.40 or 0.9. How big $R^2$ "needs" to be can sometimes be determined by how small $S_e$ needs to be using $\mathbf{R^2 \approx 1 - (S_e/S_y)^2}$

Example-Predicting SBP:

The SBP SD=26.4 mm Hg. From the model with age, $SD_e$=18.6 mm Hg. The 95% PI is $\hat{Y} \pm$ 37.2 mm Hg. Not very precise.

Q-How large does $R^2$ have to be for a 95% prediction interval width of about ±10 mm Hg?

A-If the 95% PI is ±10 mm Hg wide, $2S_e \approx 10$ mm Hg (Gaussian theory). So $S_e$=5 mm Hg.

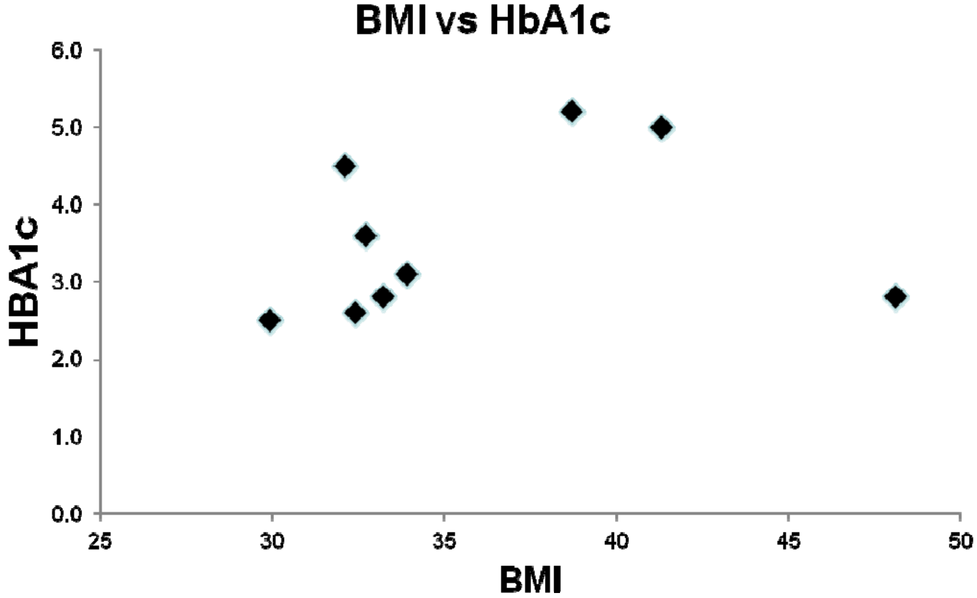
$R^2 = 1 - (SD_e/SD_y)^2 =$

So $R^2 = 1-(5/26.4)^2 = 1-0.036 = 0.964 = 96.4\%$

# Pearson (r) vs Spearman ($r_s$) correlation

Pearson r – Assumes relationship between Y and X is linear except for noise. "parametric" (inspired by bivariate normal model). Strongly affected by outliers.

Spearman $r_s$ – Based on **ranks** of Y and X. Assume relation between Y and X is monotone (non increasing, non decreasing). "Non parametric". Less affected by outliers.



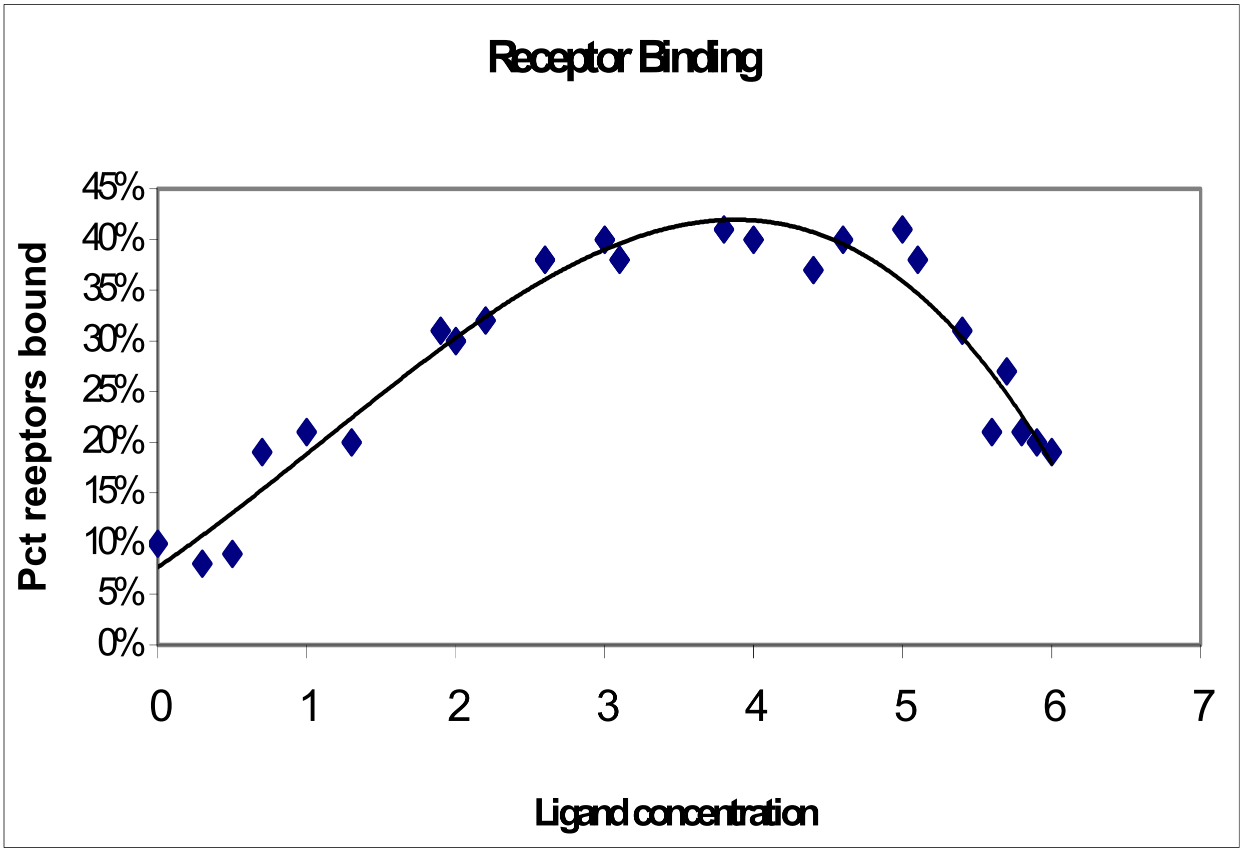
r =0.25, $r_s$ = 0.48

## Limitations of Correlation and Linear Regression Statistics

The slope (b), the intercept (a), the (Pearson) correlation coefficient (r) and the residual $SD_e$ are only useful when there is (at least approximately) a linear association between x and y.

Often there are systematic relationships in nature that are not linear. Quoting linear regression statistics (and not showing a picture) for these relationships can be misleading.

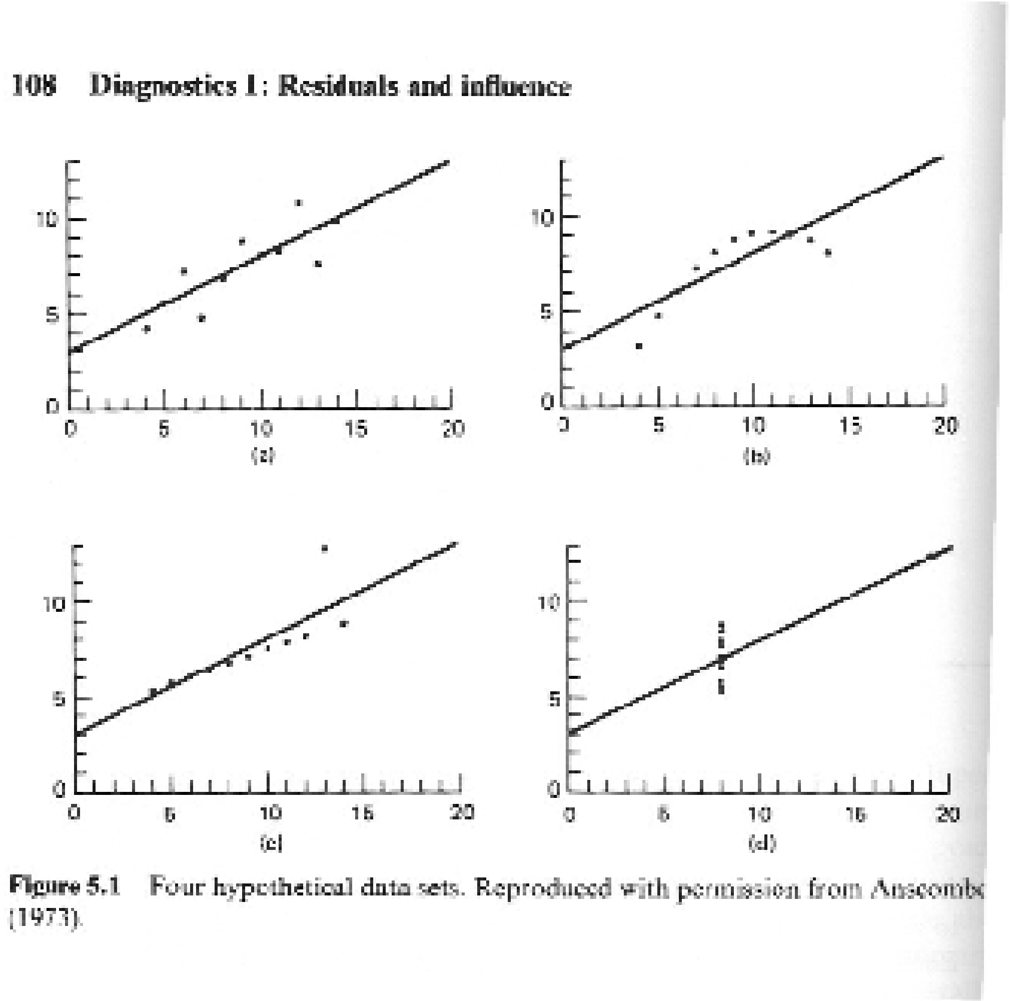
Example - In biochemistry, there are definite, well know relations between y= receptor binding versus x= ligand concentration. However, this association is not linear and is not described well by correlations or slopes.



## Pathological behavior - For all four datasets below

$$\hat{Y} = 3 + 0.5\,X, \quad r = 0.817, \quad SD_e = 13.75, \quad n=11$$

**Would not know they are different if you only saw the statistics**



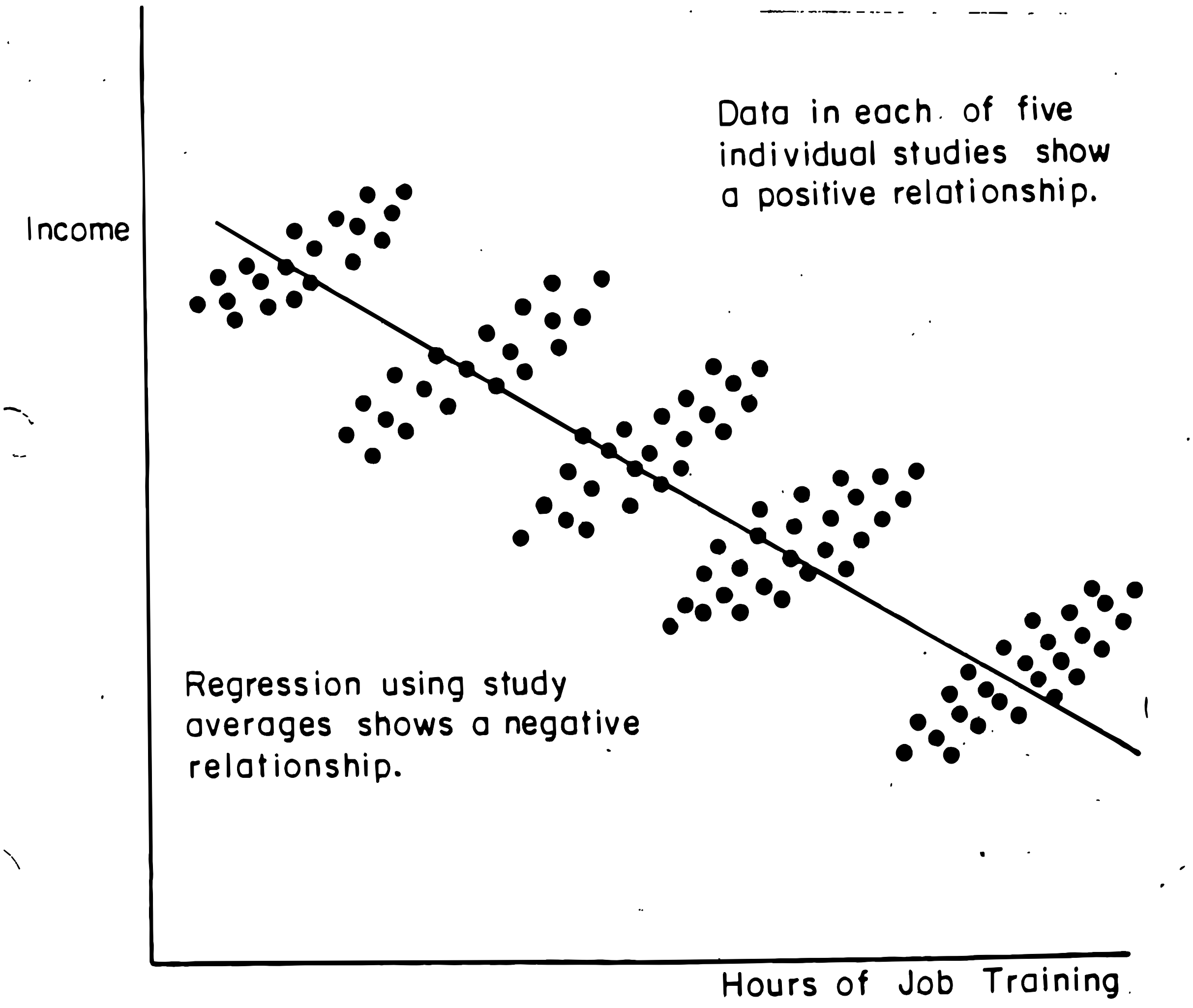
108 Diagnostics I: Residuals and influence

**Figure 5.1** Four hypothetical data sets. Reproduced with permission from Anscombe (1973).

Weisberg, Applied Linear Regression, p 108

## Ecologic fallacy

The figure below illustrates another situation where regression can be misleading if not applied carefully. If one looks at the relationship between mean income and mean hours of job training in five different cities one might get the impression that there is a negative relationship between these two measures. However, if one looks within any one city, one can see a positive relationship! Clearly, using city instead of person as the unit of analysis can completely change the impression one can get from the data.



[illegible]. Relationship between income and time spent in job training (hypothetical data from five studies).

# Interpreting "correlation" in experiments

While $R^2$ always has the interpretation of the proportion of the sample variation in Y "accounted for" by the model, r, the correlation coefficient, is not always interpretable as a measure of correlation.

When both Y and X are **observed** without any restraint or sampling bias, then X and Y are truly "random" variables from a representative population sample and r can be unbiasedly interpreted as the estimated correlation coefficient. (Also assumes X and Y have an intrinsic linear relation).

However, in many planned **experiments**, the range or values of X may be prespecified and therefore may not vary the same way as in the population. In an experiment, the X values may be restricted or fixed at certain values of interest and then Y measured at these X values (such as in a dose response experiment). When X is fixed and not allowed to vary "naturally", then r is no longer interpretable as a valid measure of correlation, even if the relation between Y and X is intrinsically linear. However, b, the estimate of β (the slope) will still be valid/unbiased since b only depends on the conditional distribution of Y given the Xs.

Algebraically, since $\mathbf{r = b\ SD_x / \surd b^2\ SD_x^2 + SD_e^2}$, if the X values are "manipulated", $SD_x$ is no longer correct (representative of the population $SD_x$) so r is no longer correct. In particular, if the X values are truncated, $SD_x$ is too small so r and $R^2$ will be too small.

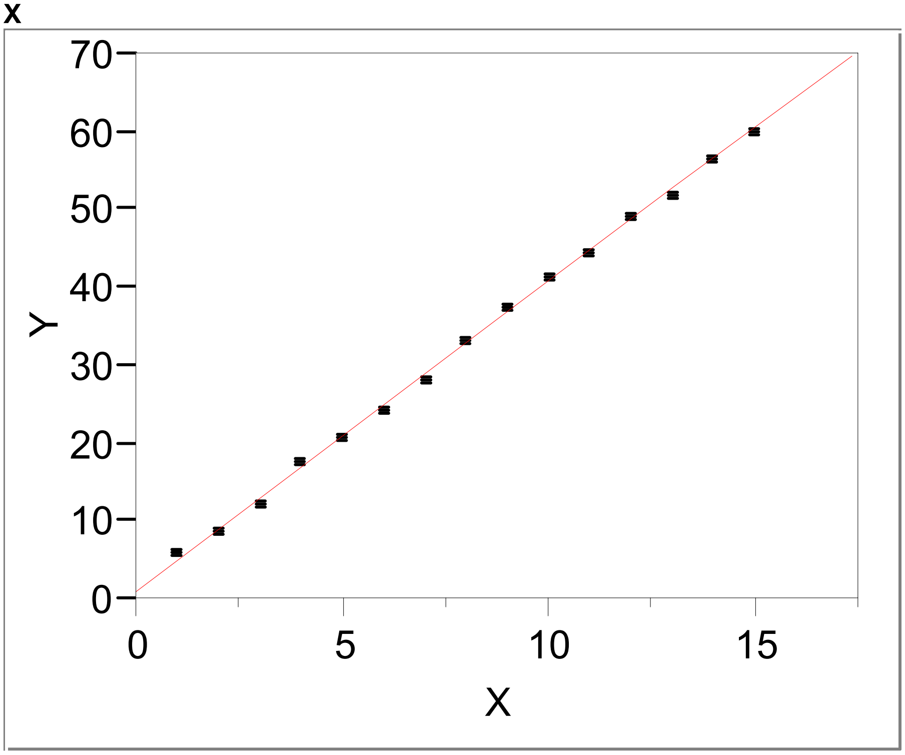
$R^2$, b and $SD_e$ when X is systematically changed

| Data | $R^2$ | b | $SD_e$ |
|---|---|---|---|
| **Complete data**<br>("truth") | **0.81** | **0.90** | **0.43** |
| Truncated<br>(X < -1 SD deleted) | 0.47 | 1.03 | 0.43 |
| center deleted<br>( -1 SD< X < 1 SD deleted) | 0.91 | 0.90 | 0.45 |
| extremes deleted<br>(X < -1 SD deleted, X > 1 SD deleted) | 0.58 | 0.92 | 0.42 |

# Attenuation of regression coefficients (estimated βs) when there is error in X

The usual regression models are forced to assume that the X values are measured without error. When the X values are in fact measured with (random) error, the resulting β estimates are too small. They are "attenuated" toward the null (zero) value.

Negligible errors in X, estimated β = 3.96 (true β is 4.0)



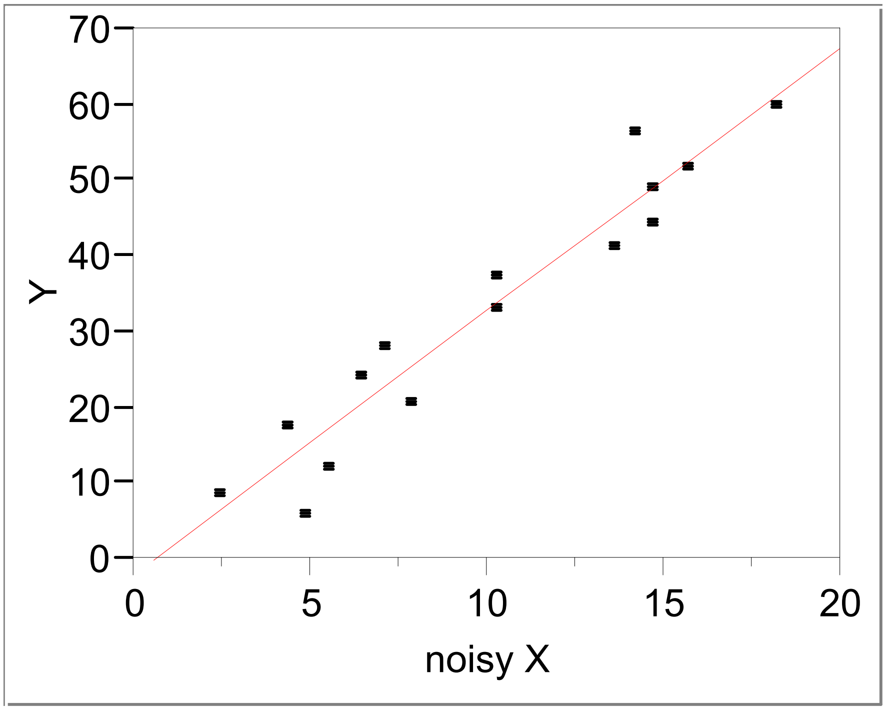
Y = 1.1490652 + 3.9591393

| | |
|---|---|
| RSquare | 0.998825 |
| RSquare Adj | 0.998734 |
| Root Mean Square Error | 0.630241 |
| Mean of Response | 32.82218 |
| Observations (n) | 15 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | p value |
|---|---|---|---|---|
| Intercept | 1.1490652 | 0.342447 | 3.36 | 0.0052 |
| X | **3.9591393** | 0.037664 | 105.12 | <.0001 |

Errors in X, estimated β value of 3.49 is too small



Y = -2.131676 + 3.4865502 noisy X

| | |
|---|---|
| Rsquare | 0.924914 |
| RSquare Adj | 0.919138 |
| Root Mean Square Error | 5.037825 |
| Mean of Response | 32.82218 |
| Observations (n) | 15 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | p value |
|---|---|---|---|---|
| Intercept | -2.131676 | 3.053135 | -0.70 | 0.4974 |
| noisy X | **3.4865502** | 0.27552 | 12.65 | <.0001 |

Any statistic that measures relationships including regression coefficients, correlation coefficients, risk ratios, odds ratios and mean differences can be attenuated in the presence of measurement noise. Random “noise” tends to make the estimates closer to their null value.

# Checking for linearity – smoothing & splines

Smoothing is a method for deciding if the relationship between Y and X is intrinsically linear (or monotone). Can suggest the proper transformation to make a linear relationship between Y and X.

Basic idea: In a plot of Y vs X, also plot $\hat{Y}$ vs X where

$$\hat{Y}_i = \sum W_{ni} Y_i \qquad \& \sum W_{ni}=1,\ W_{ni}>0.$$

The "weights" $W_{ni}$, are larger near $Y_i$ and smaller far from $Y_i$.

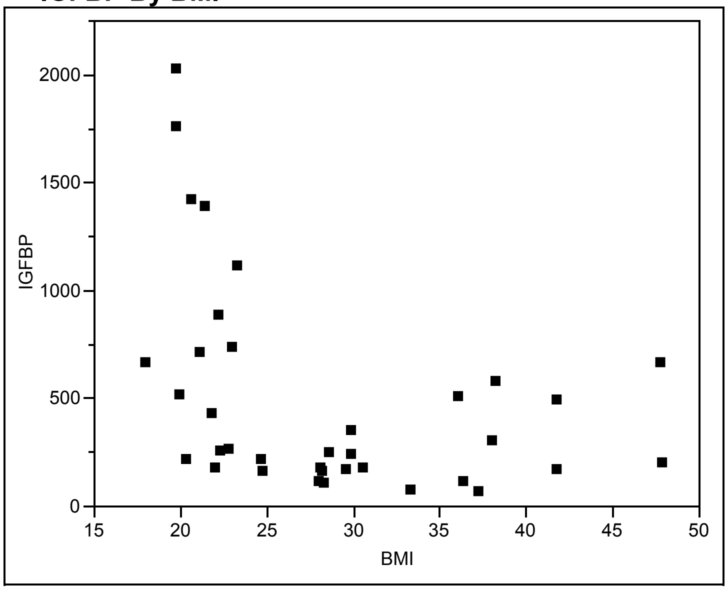
Smooth – Define a moving "window" of a given width centered around the i$^{th}$ data point. Fit a mean (moving average) or a linear or quadratic function in this window. The smoothed value is the predicted value ($\hat{Y}_i$) of the fitted function at i. Move the window over one point (to i+1) and repeat. Then connect the $\hat{Y}_i$ values across the windows.

Spline- Break the X axis into equally spaced non overlapping windows. Fit a polynomial (usually a quadratic or cubic) within each bin such that the "ends" all "match" (are piecewise continuous and their first derivative is also continuous) from window to adjacent window.

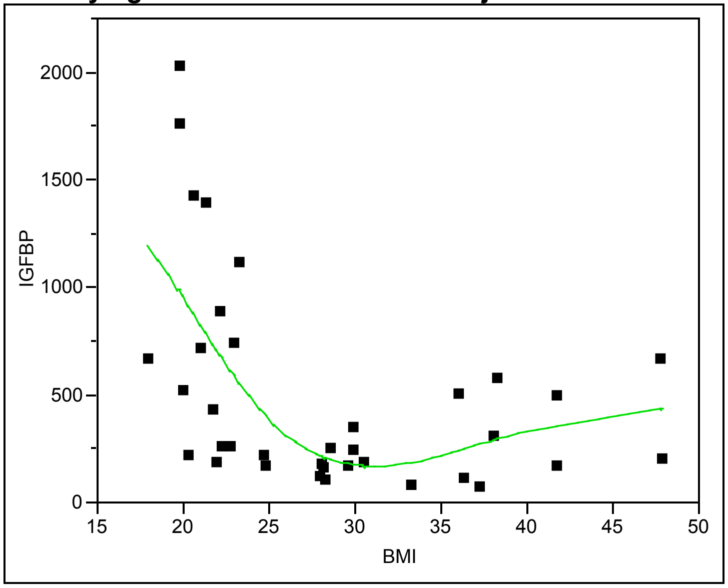
.

The size of the window controls the amount of smoothing. The bigger the window, the greater the smoothing. Maximum smoothing occurs when there is only one window covering the range of the X data. This usually produces a straight line. While exactly how much smoothing to do is somewhat subjective, **the rule is to smooth until all the "small bumps" are gone, making a smooth curve. But one smooths no farther than this.**

IGFBP By BMI



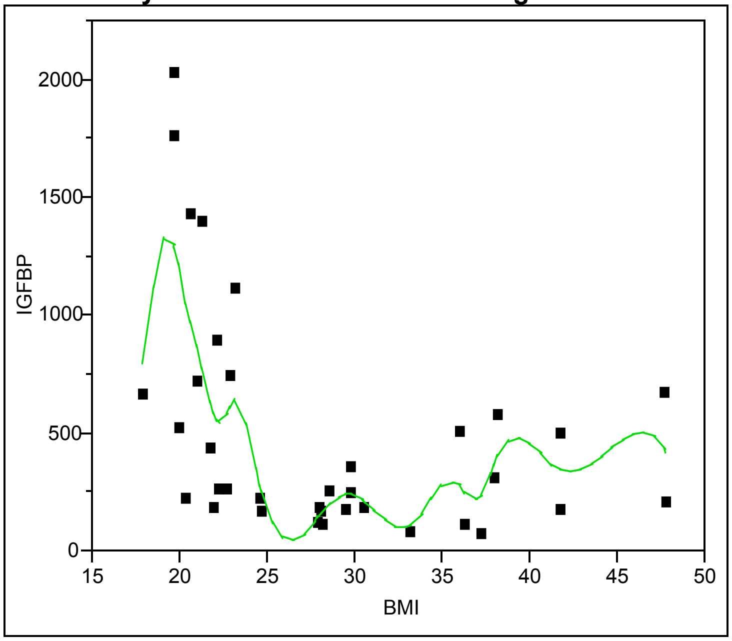
**Underlying relation is not linear – not just because of random noise**



**IGFBP By BMI with smoothing (PROC LOESS) – monotone curve**

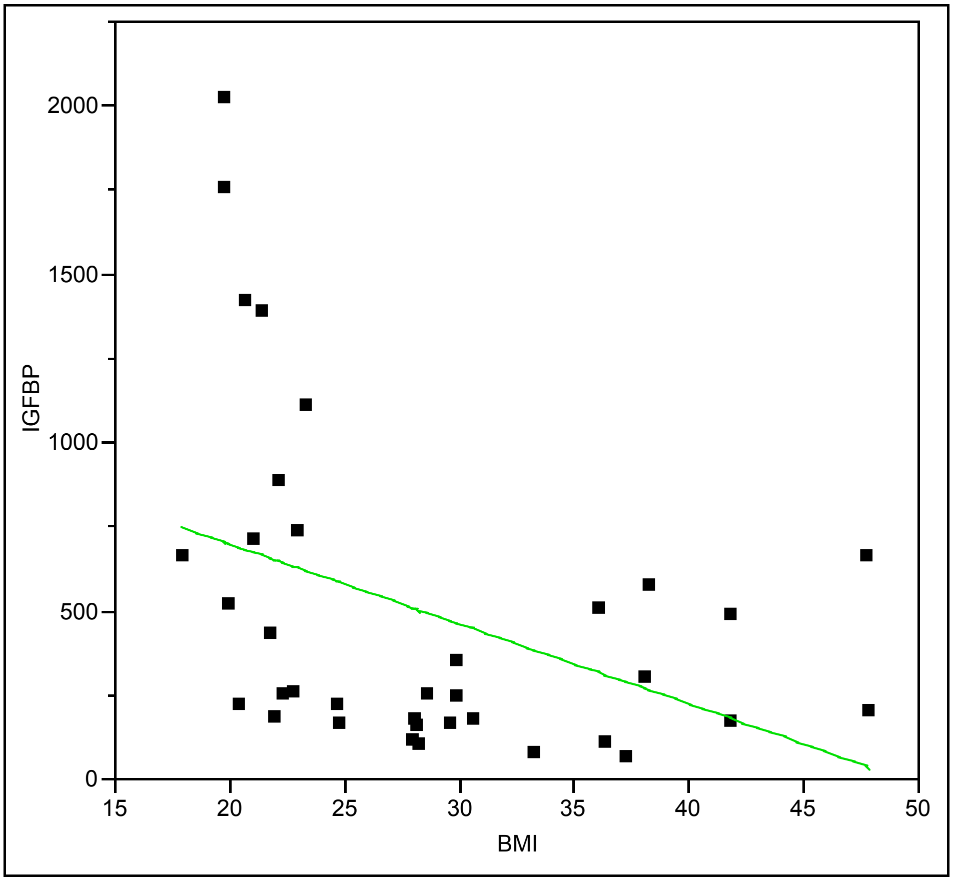
| | |
|---|---|
| R-Square | 0.416263 |
| Sum of Squares Error | 4881719 |

**IGFBP By BMI - insufficient smoothing – not a monotone curve**



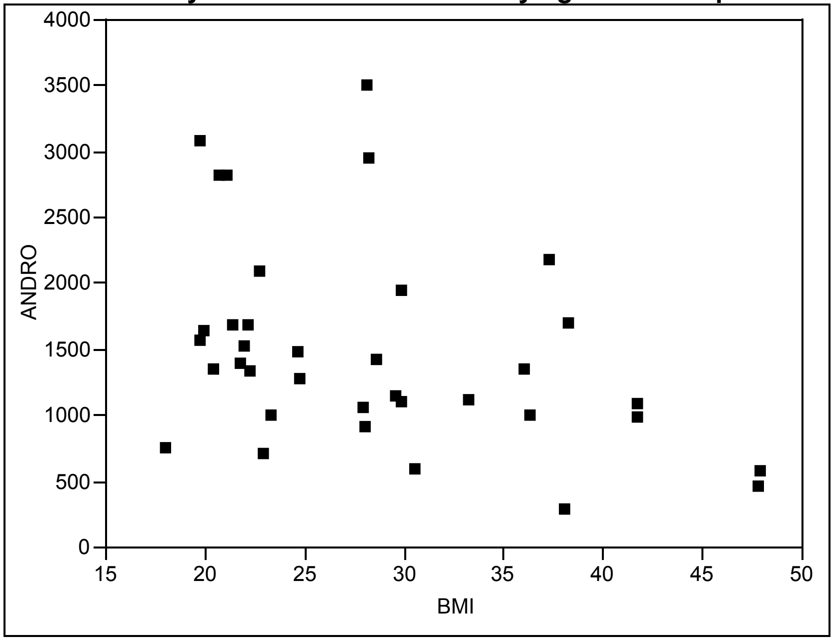
| | |
|---|---|
| R-Square | 0.565795 |
| Sum of Squares Error | 3631203 |

## IGFBP By BMI – over smoothing – almost always produces a straight line

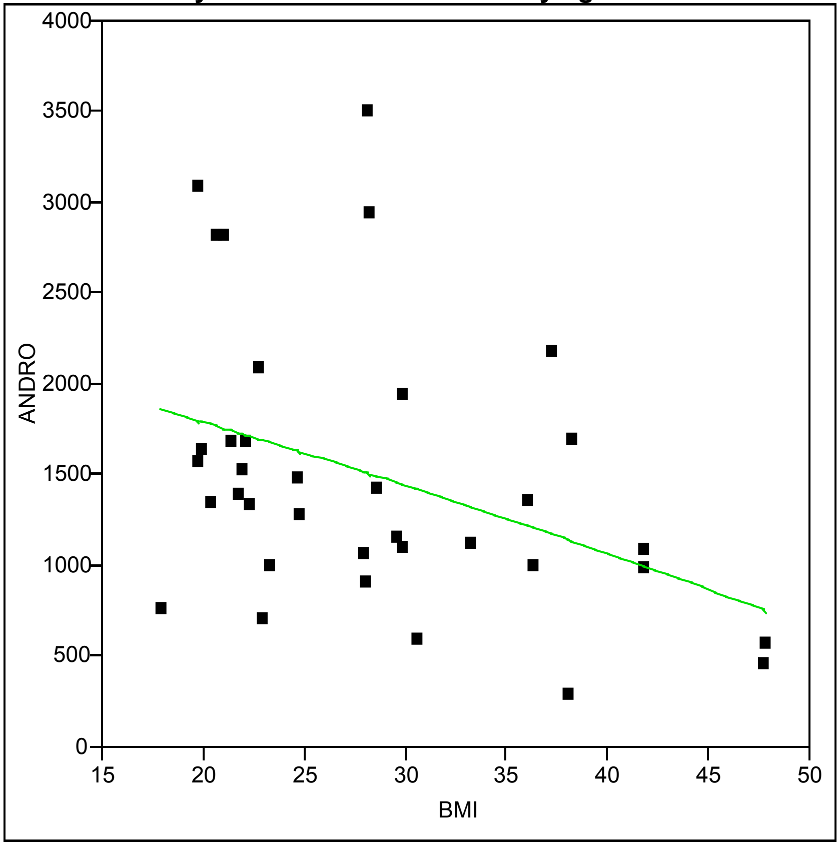


| | |
|---|---|
| R-Square | 0.159773 |
| Sum of Squares Error | 7026708 |

**ANDRO By BMI     not clear if underlying relationship is linear**



**ANDRO By BMI  - can see that underlying relation here is linear**



```
R-Square                              0.156004
Sum of Squares Error                 17458751
```

# Sec VIIIb- Multiple Regression - Overview

Multiple Regression in statistics is the science and art of creating an **equation** that relates an outcome Y, to one or more predictors, $X_1$, $X_2$, $X_3$, .. $X_k$. The predictors can be continuous variables such as age or weight or they can be discrete variables such as treatment or gender.
In the case of "c" treatments, c-1 "dummy" X variables must be made. For example, if there are c=3 treatment groups, A, B and C, where C might be the referent (or control) group, a dummy X variable is made for A vs C and another is made for B vs C. The predictors can also be interactions among variables or non linear transformations of variables.

Regression is a powerful tool for describing the multiple, simultaneous influences of many factors on Y. It is also can be very misleading if applied carelessly.

There are many types of regression. One of the most important considerations is the nature of the outcome variable, Y.

### Multiple linear regression

If Y is continuous over a large range, it is modeled as a linear function of the Xs. This is called linear regression and is a model of the form

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k + e = \hat{Y} + e$$

where "e" is the residual error between the observed Y and the prediction ($\hat{Y}$). In this and all other regression models, $b_1$, $b_2$, … $b_k$ are called **regression coefficients** but their interpretation is somewhat different for each type of regression. In linear regression, $b_i$ is the average change in Y for a one unit change in X. That is, $b_i$ is the rate of change in Y per X. In all regression models, if the $b_i$ is positive, Y increases as X increases and if the $b_i$ is negative, Y decreases as X increases.

### Multiple Logistic regression

When Y is binary (coded 0 for negative and 1 for positive), Y itself cannot be a linear function of the Xs. Instead, let P = mean Y. P is the proportion of persons with a given set of X values (covariate pattern) who have Y=1. If Y is disease or no disease, P is the risk. We define the **logit** of P as

Logit(P) = ln( P/(1-P)). "Logit" is short for log of the odds since P/(1-P) is the odds.

In multiple logistic regression, the logit of P, not P, is a linear function of the Xs

$$\text{Logit}(P) = \ln(P/(1-P)) = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k$$

There is no error term ("e") in logistic regression.

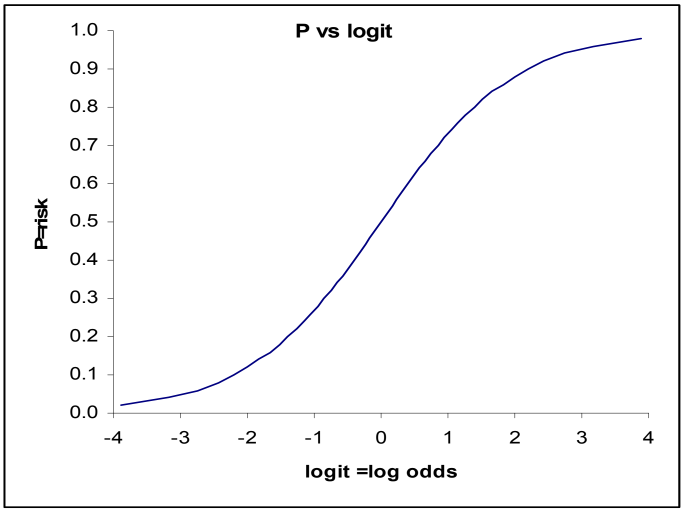
The above equation implies that the odds is given by

$odds=\exp(a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k)$,

$risk = P = odds/(odds+1)$

In logistic regression, each b is the change in the logit for a unit change in X. Therefore $e^b = \exp(b)$ is the odds ratio for a unit change in X. The odds ratio for a change of ΔX is exp(b ΔX).

Logit function P versus logit(P)=ln[P/(1-P)]



**Example - Predictors of in hospital infection**

| Characteristic | Odds Ratio (95% CI) | | p value |
|---|---|---|---|
| Incr APACHE score | 1.15 | (1.11-1.18) | <.001 |
| Transfusion (y/n) | 4.15 | (2.46-6.99) | <.001 |
| Increasing age (yr) | 1.03 | (1.02-1.05) | <.001 |
| Malignancy | 2.60 | (1.62-4.17) | <.001 |
| Max Temperature | 0.70 | (0.58-0.85) | <.001 |
| Adm to treat>7 d | 1.66 | (1.05-2.61) | 0.03 |
| Female (y/n) | 1.32 | (0.90-1.94) | 0.16 |

*APACHE = Acute Physiology & Chronic Health Evaluation Score

**Multiple Poisson regression**

When Y is a positive integer that is zero or larger (0,1,2,3…), it is also not advisable to model Y as a linear function of the Xs. Instead, it is better to model the log of Y as a linear function of the Xs. So the multiple Poisson regression model is given by

$$\ln(\text{mean } Y) = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k$$

This implies that $\text{mean } Y = \exp(a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k)$

In this model, Y cannot be negative.

In the Poisson model, the regression coefficient bi is the rate of change in ln(Y) per unit change in X. Also, 100 $b_i$ is the **percent change** in Y per one unit change in X. For example, if the regression coefficient for age in years is b=0.057, then Y changes 5.7% per year.

**Multiple proportional hazards regression (Cox model) for time dependent events**

For time dependent outcomes (ie time to death), we often with to model the hazard, h, instead of the mean Y. The hazard is the event rate per unit time (ie for death, it is the mortality rate). Since $h > 0$, we model the log of the hazard as a linear function of the Xs (similar to Poisson regression)

$$\ln(h) = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k$$

so $h = \exp(a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k)$

If $h_o=\exp(a)$ is the 'baseline' hazard, (that is, $a=\log(h_0)$) the hazard ratio is

$HR = h/h_0 = \exp(b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k)$ **no 'a'.**

If $S_0(t)$ is the 'baseline' Survival curve (survival function) corresponding to the baseline hazard, then the survival curve for particular values of $X_1, X_2, \ldots X_k$ is given by

$S(t) = S_0(t)^{HR}$ where HR is first computed by plugging the Xs into the above equation.

In this model, $\exp(b_i)$ is the hazard rate ratio for a unit change in $X_i$.

.

Example: Busuttil et. a. 2005 - *Annals of Surgery* • Volume 241, Number 6, June

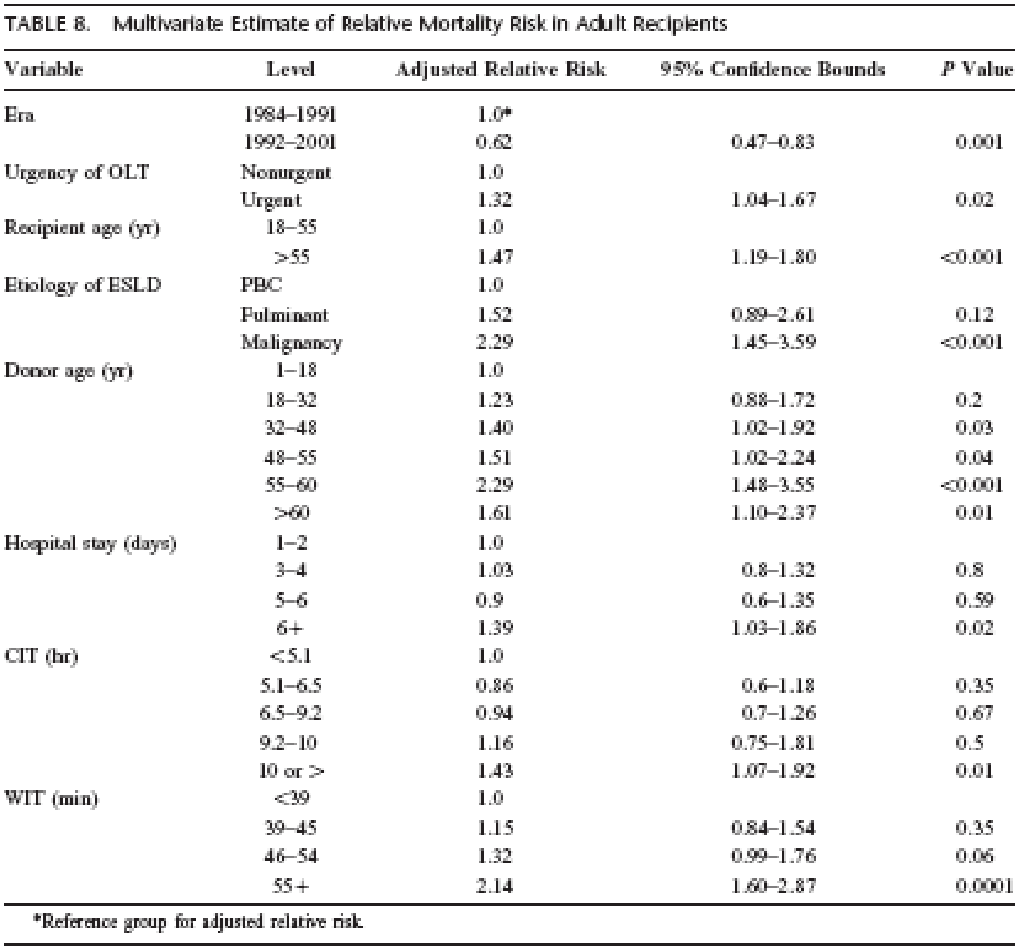
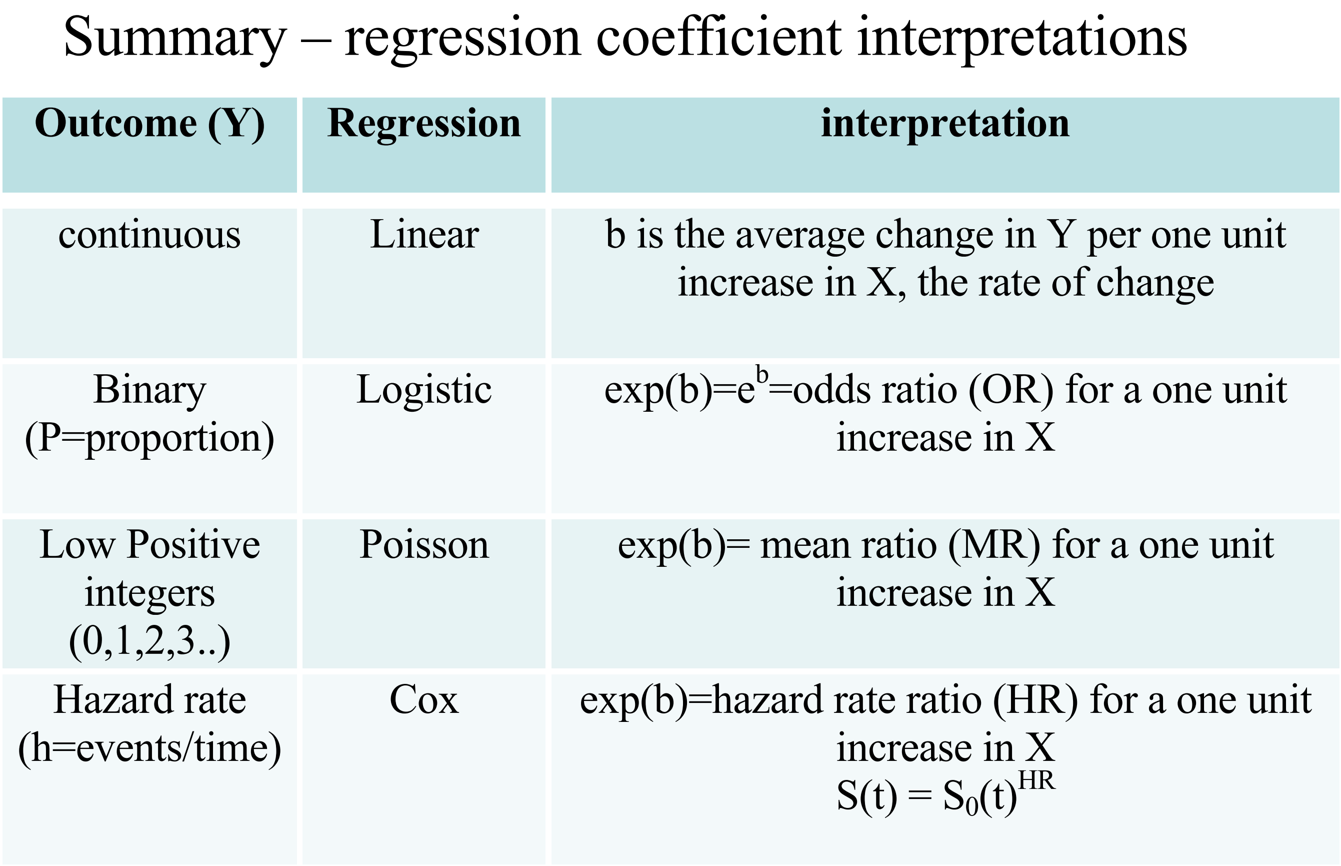


TABLE 8. Multivariate Estimate of Relative Mortality Risk in Adult Recipients

| Variable | Level | Adjusted Relative Risk | 95% Confidence Bounds | *P* Value |
|---|---|---|---|---|
| Era | 1984–1991 | 1.0* | | |
| | 1992–2001 | 0.62 | 0.47–0.83 | 0.001 |
| Urgency of OLT | Nonurgent | 1.0 | | |
| | Urgent | 1.32 | 1.04–1.67 | 0.02 |
| Recipient age (yr) | 18–55 | 1.0 | | |
| | >55 | 1.47 | 1.19–1.80 | <0.001 |
| Etiology of ESLD | PBC | 1.0 | | |
| | Fulminant | 1.52 | 0.89–2.61 | 0.12 |
| | Malignancy | 2.29 | 1.45–3.59 | <0.001 |
| Donor age (yr) | 1–18 | 1.0 | | |
| | 18–32 | 1.23 | 0.88–1.72 | 0.2 |
| | 32–48 | 1.40 | 1.02–1.92 | 0.03 |
| | 48–55 | 1.51 | 1.02–2.24 | 0.04 |
| | 55–60 | 2.29 | 1.48–3.55 | <0.001 |
| | >60 | 1.61 | 1.10–2.37 | 0.01 |
| Hospital stay (days) | 1–2 | 1.0 | | |
| | 3–4 | 1.03 | 0.8–1.32 | 0.8 |
| | 5–6 | 0.9 | 0.6–1.35 | 0.59 |
| | 6+ | 1.39 | 1.03–1.86 | 0.02 |
| CIT (hr) | <5.1 | 1.0 | | |
| | 5.1–6.5 | 0.86 | 0.6–1.18 | 0.35 |
| | 6.5–9.2 | 0.94 | 0.7–1.26 | 0.67 |
| | 9.2–10 | 1.16 | 0.75–1.81 | 0.5 |
| | 10 or > | 1.43 | 1.07–1.92 | 0.01 |
| WIT (min) | <39 | 1.0 | | |
| | 39–45 | 1.15 | 0.84–1.54 | 0.35 |
| | 46–54 | 1.32 | 0.99–1.76 | 0.06 |
| | 55+ | 2.14 | 1.60–2.87 | 0.0001 |

*Reference group for adjusted relative risk.

| Donor age | HR | 95% CI | p value |
|---|---|---|---|
| 1-18 | 1.00 (ref) | -- | -- |
| 18-32 | 1.23 | 0.88-1.72 | 0.20 |
| 32-48 | 1.40 | 1.02-1.92 | 0.03 |
| 48-55 | 1.51 | 1.02-2.24 | 0.04 |
| 55-60 | 2.29 | 1.48-3.55 | < 0.001 |
| 60+ | 1.61 | 1.10-2.37 | 0.01 |



## Summary – regression coefficient interpretations

| Outcome (Y) | Regression | interpretation |
|---|---|---|
| continuous | Linear | b is the average change in Y per one unit increase in X, the rate of change |
| Binary (P=proportion) | Logistic | exp(b)=$e^b$=odds ratio (OR) for a one unit increase in X |
| Low Positive integers (0,1,2,3..) | Poisson | exp(b)= mean ratio (MR) for a one unit increase in X |
| Hazard rate (h=events/time) | Cox | exp(b)=hazard rate ratio (HR) for a one unit increase in X<br>$S(t) = S_0(t)^{HR}$ |

## Multiple linear regression

Linear regression is where a continuous Y is modeled by

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k + e$$

where "e" is the residual error between the observed Y and the prediction based on the Xs.

In this model, the $i^{th}$ regression coefficient, $b_i$, is the average rate of (assumed linear) change in the predicted Y for a **unit change** in the ith predictor, $X_i$, given that all of the other covariates are **held constant**.

As an example, consider predictors of Y=Bilirubin (mg/dl) in liver transplant candidates. Two predictors are $X_1$=Prothombin time (PT in seconds) and $X_2$=ALT (alanine aminotransferase in U/L).

A multiple regression equation (on the log scale) is

$\hat{Y}$ = (predicted) log Bilirubin = -3.96 + 3.47 log PT + 0.21 log ALT

This equation says that, holding the (linear) influence of log ALT constant, for every 1 log second increase in PT, there is an average 3.47 log mg/dl increase in log Bilirubin. Holding log PT constant, there is an average 0.21 log mg/dl increase in log Bilirubin for a 1 log U/L increase in log ALT.

The correlation between the observed log Bilirubin (Y) and the predicted log Bilirubin ($\hat{Y}$) is $r = \sqrt{0.448} = 0.67$. If $SD_y$ is the SD of log Bilirubin ignoring log PT and log ALT and $SD_e$ is the SD of the residual errors ($e = Y - \hat{Y}$), as before,

$$SD_e^2 = SD_y^2(1 - r^2) \text{ or } \mathbf{r^2 = (SD_y^2 - SD_e^2)/SD_y^2.}$$

That is, $r^2$ is the amount that the variation (variance) in Y is **reduced** by knowledge of the Xs. In this example, since $r^2 = 0.448$. We say that $X_1$ and $X_2$ (log PT, log ALT) "account for" **45%** of the observed variation in log Bilirubin, leaving $1-r^2$ = 55% not accounted for. $SD_e^2$ is 55% as big as $SD_y^2$. So in this example, much of the observed variation in log Bilirubin is still not accounted for.

**Response Log Bilirubin*=y*** ***JMP output***

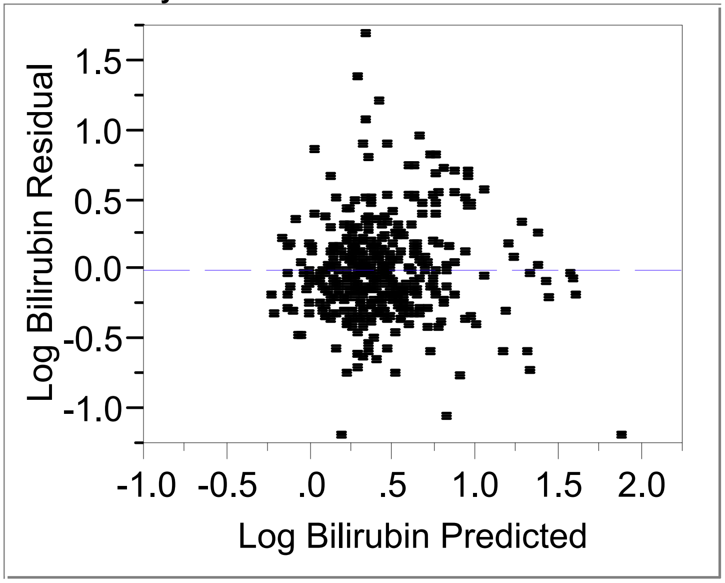
```
Summary of Fit
RSquare                          0.447812
RSquare Adj                       0.44477
Root Mean Square Error           0.358133  <---- SDe
Mean of Response                 0.438745
Observations (or Sum Wgts)          366

Analysis of Variance
 Source    DF   Sum of Squares  Mean Square   F Ratio
 Model      2     37.757715       18.8789     147.1926
 Error    363     46.558206        0.1283     Prob > F
C. Total  365     84.315922                    <.0001

Lack Of Fit                                                   Fit not
  Source        DF   Sum of Squares  Mean Square   F Ratio    rejected
Lack Of Fit    354     43.888595       0.123979     0.4180
 Pure Error      9      2.669612       0.296624    Prob > F
Total Error    363     46.558206                    0.9878

Parameter Estimates
   Term      Estimate    Std Error   t Ratio   Prob>|t|
 Intercept  -3.960849    0.257399    -15.39     <.0001
  log PT     3.4714393   0.214307     16.20     <.0001
  log ALT    0.210873    0.05515       3.82     0.0002
```

**Residual by Predicted Plot**



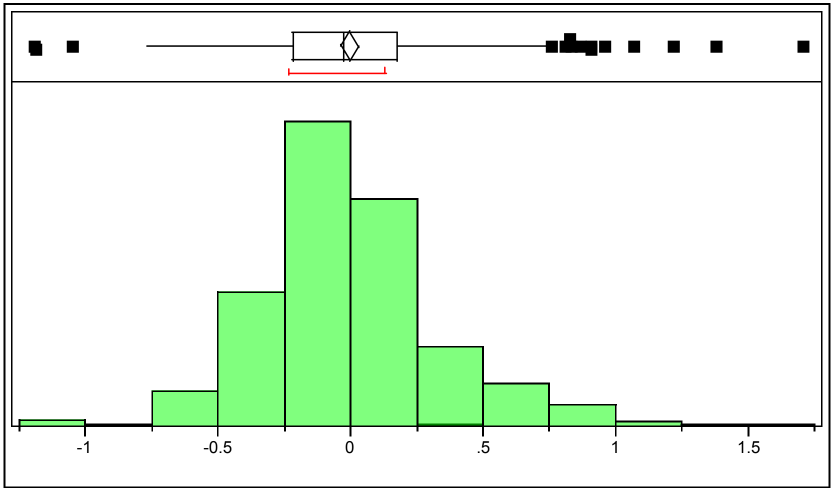
***Residual error plot***

***When the model is valid, this plot should look like a circular cloud if the errors have constant variance. The example above is a "good" result.***

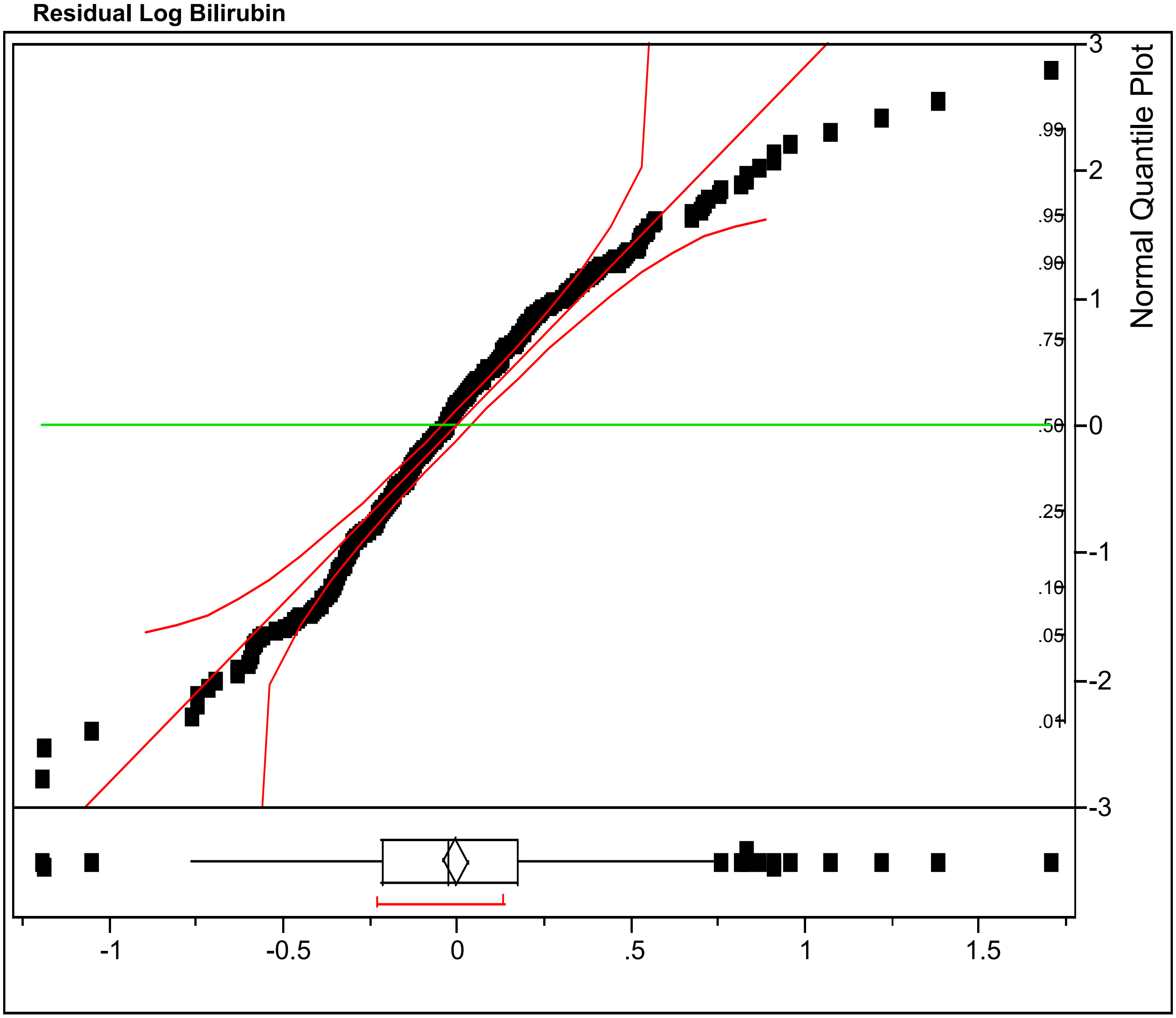
Example of a "good" residual error histogram

**Distributions**

**Residual Log Bilirubin** *(residual errors = e)*



## Normal quantile plot
## Should be (at least approximately) a straight line if the residual error data is Gaussian



residual error (e)

# Interpretation of multiple regression coefficients (cont.)

The multiple regression will not in general be the same as the individual regression coefficient for each variable one at a time, even though the same Y is being modeled.

| variable | Simple (one Y, one X) regression | simultaneous multiple regression ($b_1X_1+b_2X_2$) |
|---|---|---|
| Log PT | 3.560 | 3.470 |
| Log ALT | 0.310 | 0.211 |

Log Bilirubin = - 3.70 + **3.56** log PT, $R^2 = 0.425$

Log Bilirubin = -.105 + **0.310** log ALT, $R^2 = 0.049$

Log Bilirubin = -3.96 + **3.47** log PT + **0.211** log ALT , $R^2 = 0.448$

**The simple & multple regression coeffs for log PT don't match**

### Rare special case – orthogonality

However, if all of the Xs have zero correlation with each other (but not with Y), then the simple "bivariate" regression coefficients for the regression of Y on each $X_j$ (ignoring all the other Xs) will be the same as the multiple regression coefficients for Y regressed on all of the Xs. (Y regressed on $X_1$, $X_2$, … $X_{k-1}$). When all of the Xs are uncorrelated with each other they are said to be **orthogonal.** This usually only happens in designed experiments, not in observational studies. (Collinearity is the "opposite" when the X variables are strongly correlated with each other).

Since we usually do NOT have orthogonality, evaluating a set of k-1 variables one at a time will NOT generally give the same results as evaluating all k-1 variables simultaneously in a multiple regression model.

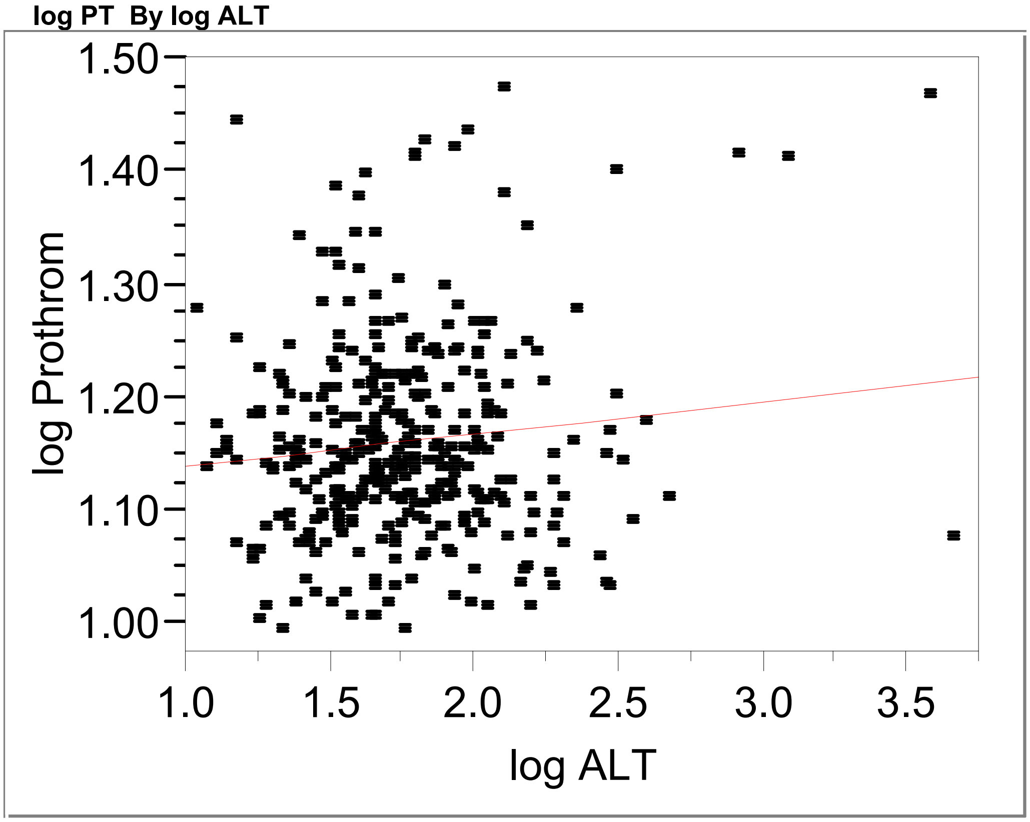
While multiple regression is useful for simultaneously evaluating all the factors that affect an outcome, and so can be an important tool for controlling for confounding, artifacts/bias can arise if two assumptions are not verified.

1. When an X is continuous or interval, the relation between X and Y is assumed linear. Sometimes this is true on a transformed scale. If this is not true on any scale, then X must be polychotomized into groups.

2. By default, the effects of the X’s are assumed additive. This can be checked by adding interaction terms (ie $X_3 = X_1 \text{ x } X_2$). Sometimes interactions are very important.

3 Also, in linear regression, prefer residual errors to have a Gaussian distribution with a constant variance that is independent of Y. But additivity and linearity are more important since lack of additivity and linearity lead to bias.

Correlation of $X_1$=log PT with $X_2$=log ALT



$r_{12}$ = 0.111, $R^2$ = 0.0123

Since the correlation between log PT ($X_1$) and log ALT ($X_2$) is low, the simple versus multiple regression results are similar.

# Interaction effects (& subgroups)- definition

The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

implies that change in Y due to $X_1$ $(=\beta_1)$ is the same (constant) for <u>all</u> values of $X_2$.

In the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{\beta_3} X_1 X_2 + \varepsilon$$

the $\beta_3$ term is an **interaction** term. Change in Y for a unit change in $X_1$ is $(\beta_1+\beta_3 X_2)$ and is therefore not constant.

Positive $\beta_3$ is often termed a "synergism"

Negative $\beta_3$ is often termed an "antagonism"

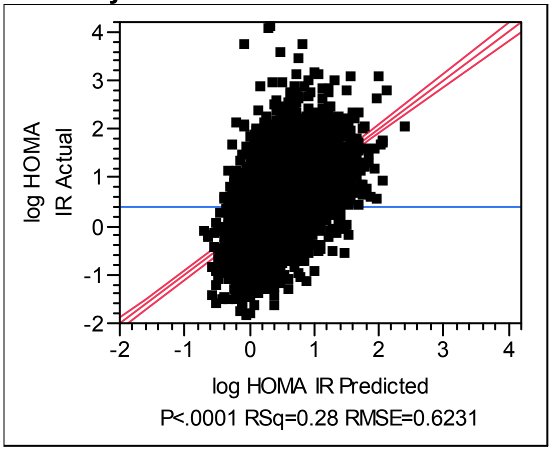
How to implement in software? Make new variable $W = X_1 X_2$.

This is a way to test for additivity.

# Interaction effects example

Response: **Y= log HOMA IR** (MESA study, *output from JMP software*)

**Actual by Predicted Plot**



**Summary of Fit**

| | |
|---|---|
| RSquare | 0.28044 |
| RSquare Adj | 0.280122 |
| Root Mean Square Error | 0.623101 |
| Mean of Response | 0.395153 |
| Observations (n) | 6782 |

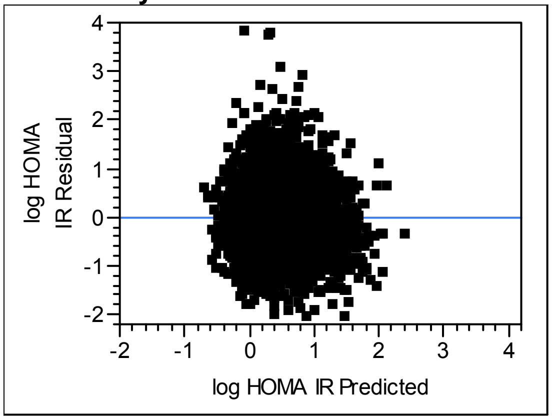
**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | p value |
|---|---|---|---|---|
| Intercept | -1.388058 | 0.049285 | -28.16 | <.0001 |
| gender | -0.668769 | 0.085421 | -7.83 | <.0001 |
| BMI | 0.0610655 | 0.001675 | 36.45 | <.0001 |
| gender*BMI | 0.0279925 | 0.002986 | 9.38 | <.0001 |

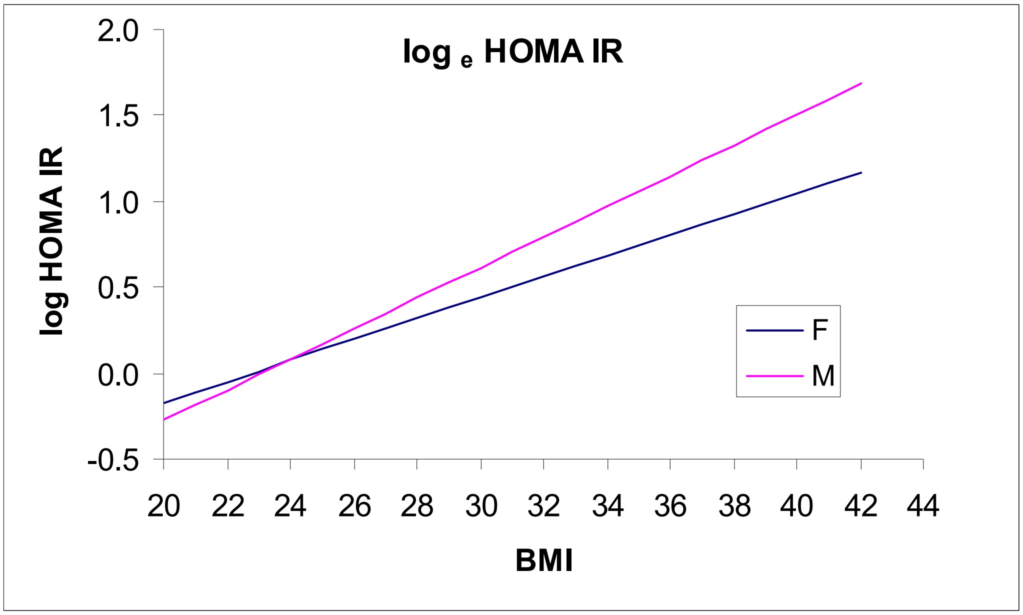
*Predicted log HOMA IR = -1.39 – 0.669 gender + 0.061 BMI + 0.028 gender * BMI*

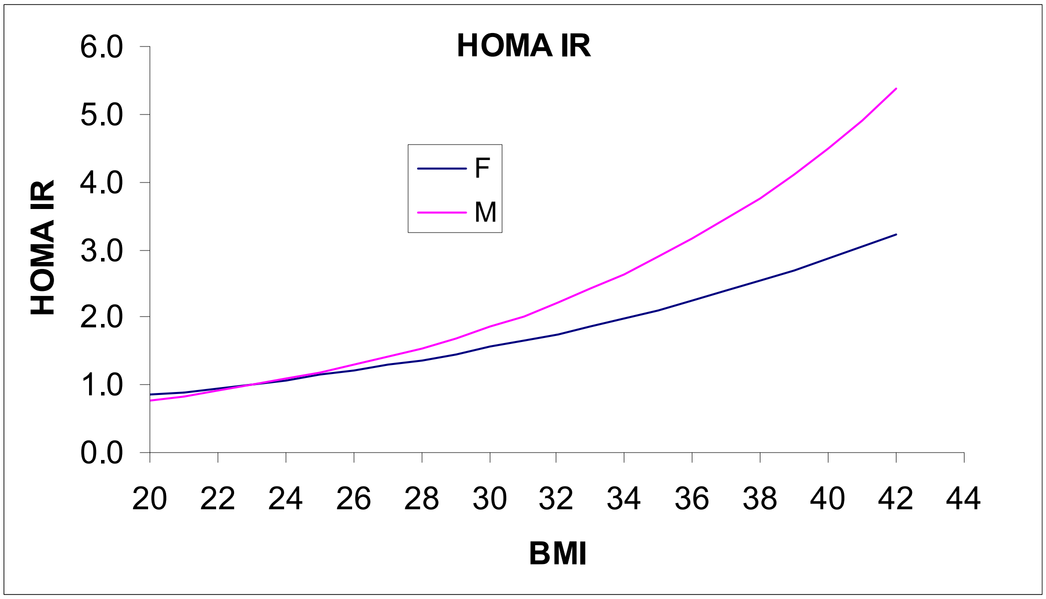
*(gender is coded 0 for female and 1 for male)*

**Residual by Predicted Plot**



# Gender x BMI interaction- non additivity





# Hierarchially well formulated (HWF) regression models

HWF Rule – To correctly evaluate the $X_1$*$X_2$ interaction, must also have $X_1$ and $X_2$ in the model. In general, one must include the lower order terms in order to correctly evaluate the higher order terms.

**Non HWF:** Model: chol = $a_0$ + $a_1$ smoke x age

**If model is not HWF, significance of interaction depends on coding (bad!)**

**0, 1 (dummy) coding:** smoke=0 or 1, smokeage = smoke x age

| Variable | DF | Estimate | std error | t | p value |
|---|---|---|---|---|---|
| INTERCEP | 1 | 156.863323 | 3.99284362 | 39.286 | 0.0001 |
| SMOKEAGE | 1 | 0.360968 | 0.18161802 | 1.988 | **0.0567** |

---

**-1, 1 (effect) coding:** smoke=-1 or 1, smokeage = smoke x age

| Variable | DF | Estimate | std error | t | p value |
|---|---|---|---|---|---|
| INTERCEP | 1 | 162.277848 | 3.10164240 | 52.320 | 0.0001 |
| SMOKEAGE | 1 | 0.054653 | 0.09975929 | 0.548 | **0.5881** |

p value for 'smokeage' has changed from 0.0567 to 0.5881

**HWF:**Model: chol =$b_0$ + $b_1$ smoke + $b_2$ age + $b_3$ smoke x age

**For HWF, significance is the same regardless of coding**

0, 1 (dummy) coding: smoke=0 or 1, smokeage = smoke x age

| Variable | DF | Estimate | std error | t | p value |
|---|---|---|---|---|---|
| INTERCEP | 1 | 100.220801 | 1.10981217 | 90.304 | 0.0001* |
| SMOKE | 1 | 3.812141 | 1.56951142 | 2.429 | 0.0224 |
| AGE | 1 | 2.009533 | 0.03569531 | 56.297 | 0.0001* |
| SMOKEAGE | 1 | -0.009001 | 0.05048079 | -0.178 | **0.8599** |

-1, 1 (effect) coding: smoke=-1 or 1, smokeage =smoke x age

| Variable | DF | Estimate | std error | t | p value |
|---|---|---|---|---|---|
| INTERCEP | 1 | 102.126872 | 0.78475571 | 130.138 | 0.0001** |
| SMOKE | 1 | 1.906070 | 0.78475571 | 2.429 | 0.0224 |
| AGE | 1 | 2.005033 | 0.02524039 | 79.437 | 0.0001** |
| SMOKEAGE | 1 | -0.004501 | 0.02524039 | -0.178 | **0.8599** |

*Testing in non smokers, ** testing overall

# Non linear regression

## The model

Log(Bilirubin)= -3.96 + **3.47** log(PT) + **0.211** log(ALT)

is a non linear model in terms of PT and ALT but is a linear model in terms of log PT, log ALT and the regression coefficients $b_0$=-3.96, $b_1$=3.47 and $b_2$=0.211.

We can still use linear regression to fit this model by making new variables $X_1$=log PT, $X_2$=log ALT. Model is **linear in the coefficients** $b_0$, $b_1$ and $b_2$.

Consider a model of the form: $\hat{Y}$= Drug conc = $b_1\ 10^{b2\ x}$

This is nonlinear in $b_2$ but can be made linear with a transformation. ( $\log_{10}(\text{conc})=\log_{10}(b_1) + b_2\ x$ )

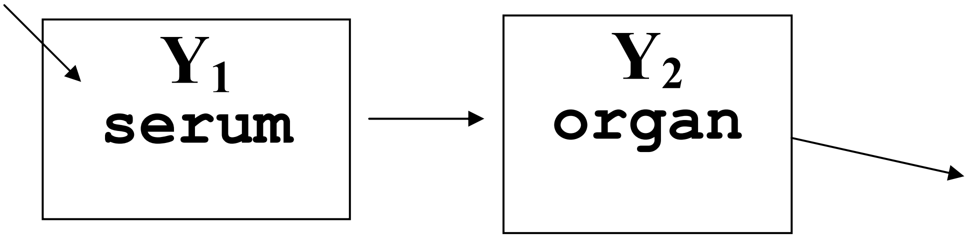
What about: Drug conc = $b_0 + b_1\ 10^{b2\ x}$

This model is **non linear in $b_2$** and can't be transformed. It requires non linear regression software to estimate $b_0$, $b_1$ and $b_2$, giving "diagnostics" ($R^2$, $SD_e$) that are the same as in linear regression. Main difference from the usual linear regression software is one needs a starting "guess" for the values $b_0$, $b_1$ and $b_2$ in order to run the non linear analysis.

# Example: Compartmental drug models

Model of how drug (or any chemical) is metabolized by an organism.

$Y_1$=conc in serum, $Y_2$=conc in organ, x=time



$d (Y_1)/dx = -b_1 Y_1$

$d (Y_2)/dx = b_1 Y_1 - b_2 Y_2$

$b_1 > b_2 > 0$

solutions:

$Y_1 = \text{const } e^{-b1\ x}$

$Y_2 = (b_1/(b_1-b_2))\ [e^{-b2x} - e^{-b1x}]$ <-*fit model*

$Y_2$ takes on a maximum value when

$x = \ln(b_1/b_2)/[b_1-b_2]$

$Y_2$ is zero when x=0 or x is very large

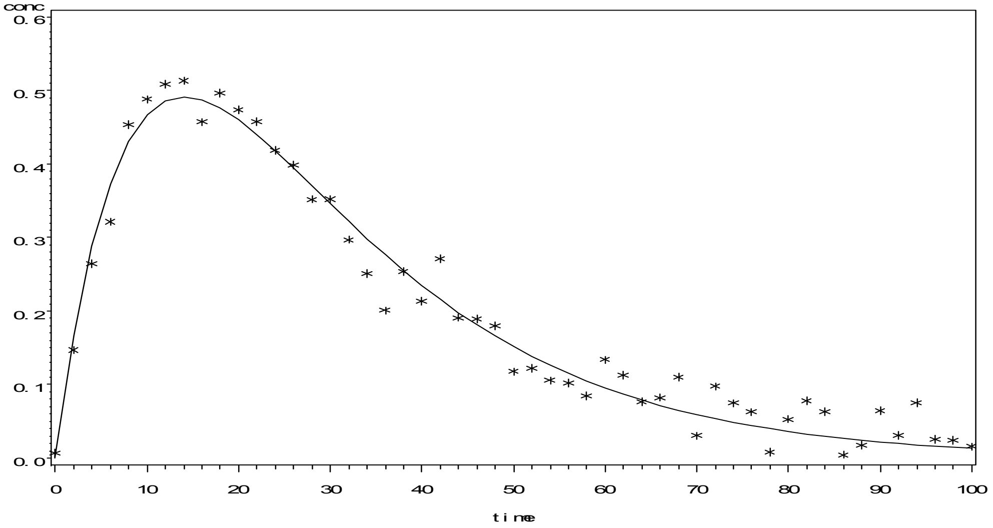
The constants $b_1$ and $b_2$ are rates. They are in units of 1/x (i.e 1/time).

If we can estimate $b_1$ and $b_2$ we can then compute other important pharmacokinetic parameters such as the mean residence time, the peak concentration, the time of the peak concentration and the area under the concentration curve after some (relatively) long time such as 24 hours. This can be important if we wish to be sure we are giving an adequate (therapeutic) and non toxic dose.

In this example

$\hat{Y}= [0.0967/(0.0967-0.0506)]*[\exp(-0.0506*t)-\exp(-0.0967*t)]$

at peak, $t = 14$ and $\hat{Y} =0.49$ conc units



## Residual diagnostics and “model criticism”

Assumptions of linear regression:

1. Linear relation between Y and each X except for random “noise” (but can transform X).

2. Effect of each X is additive (but can make interaction terms)

3. Errors (e) have constant variance and come from a Gaussian distribution

4. All observations from the same population

5. All observations independent (usually ok)

A plot of $\hat{Y}$ versus e, called a residual error (diagnostic) plot, can help verify if these assumptions are met.

# Residual diagnostic plot – good plot

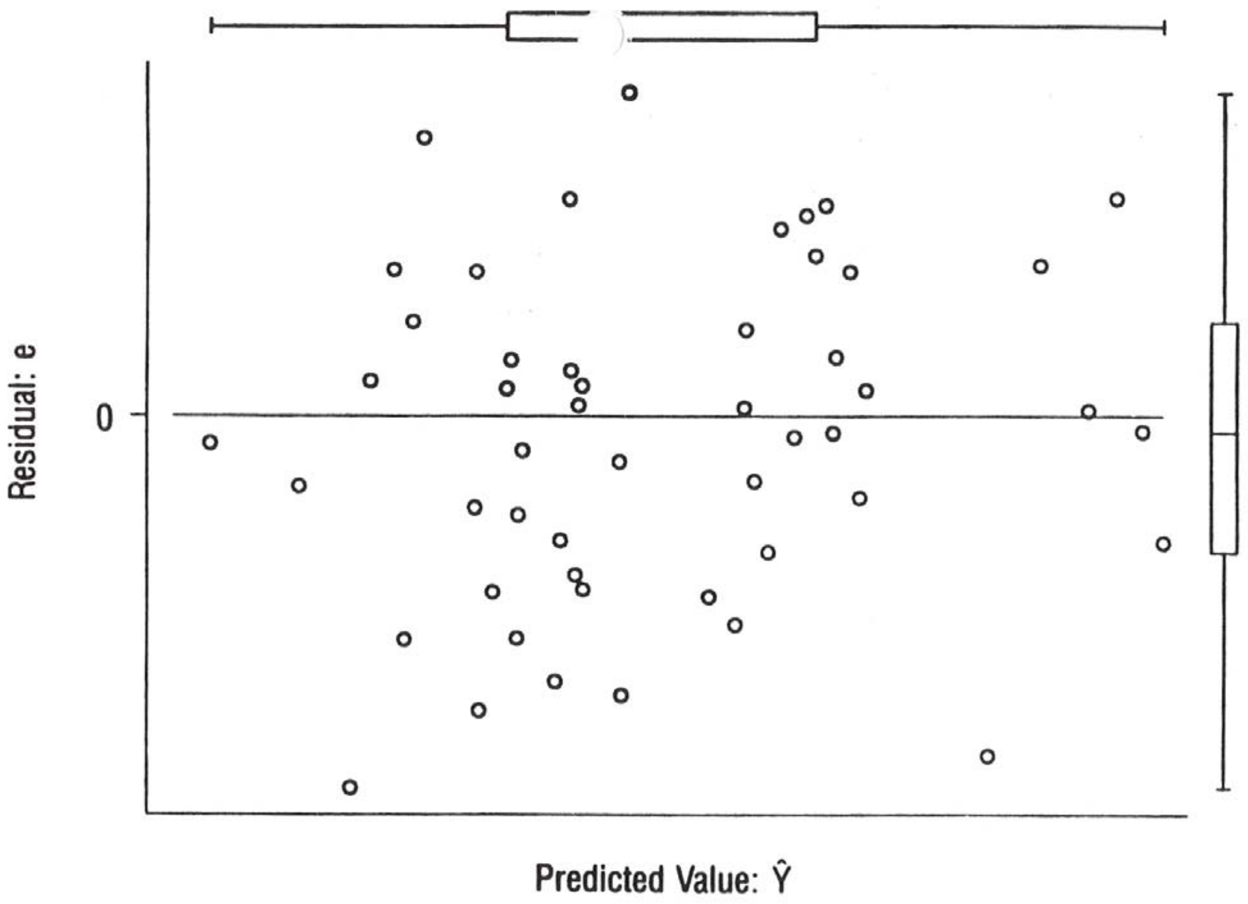
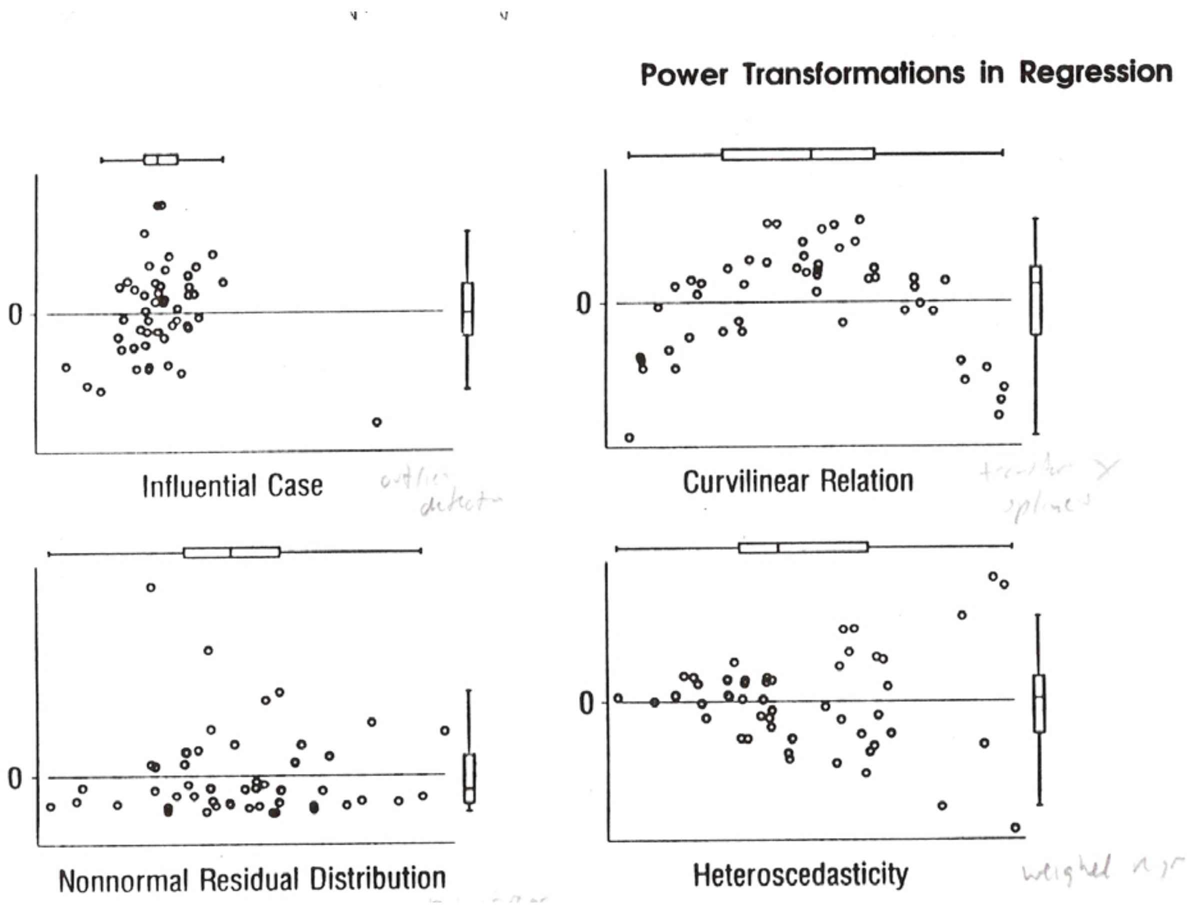


**Figure 2.10** "All clear" *e*-versus-$\hat{Y}$ plot (artificial data).

# Residual diagnostic plots – “bad” plots



Influential Case

Curvilinear Relation

Nonnormal Residual Distribution

Heteroscedasticity

## Regression model diagnostics

Residual error plots

Problem – Influential case(s) / "outliers"

Solution – If there are only a few, find them (on the residual error plot) and remove them. Determine why they are different. They often belong to a different population (ie children, not adults).

Problem – Curvilinear trend in error

Solution – Add "non linear terms" to model equation. Most common are squared terms (ie $\text{Age}^2$ as well as age), log terms and antilog (exp) terms.

Problem – Non constant error (e) variance – Heteroscedasity

Solution –Find out how the variance changes as a function of the predicted Y, $\hat{Y}$. Create "weights" that are inversely proportional to the variance.

Most common example: $SD_e$ increases as $\hat{Y}$ increases. So variance of e increases as $\hat{Y}^2$ increases. Make weight = $1/(\hat{Y}^2)$.

When $SD_e$ is not a constant, but depends on $\hat{Y}$, if the weighting is not done, the **prediction intervals** in particular based on a constant $SD_e$ may be very misleading!

# **Adjusting means - simple case** (ANCOVA)

The point $\overline{X}, \overline{Y}$ is always on the regr. line

For **each** group, the equation $Y = b_0 + b_1 X$ can be rewritten

$Y = \overline{Y} + b_1 (X - \overline{X})$ i.e. $b_0 = \overline{Y} - b_1 \overline{X}$

Let $\overline{X}_g$, $\overline{Y}_g$ be the means in the gth group
Let $\overline{X}$ be the overall mean (the mean of the means)

Where to adjust - adjust at the overall mean

The adjusted Y mean is given by

$\overline{Y}_{g\text{-adj}} = \overline{Y}_g + b_1 (\overline{X} - \overline{X}_g)$

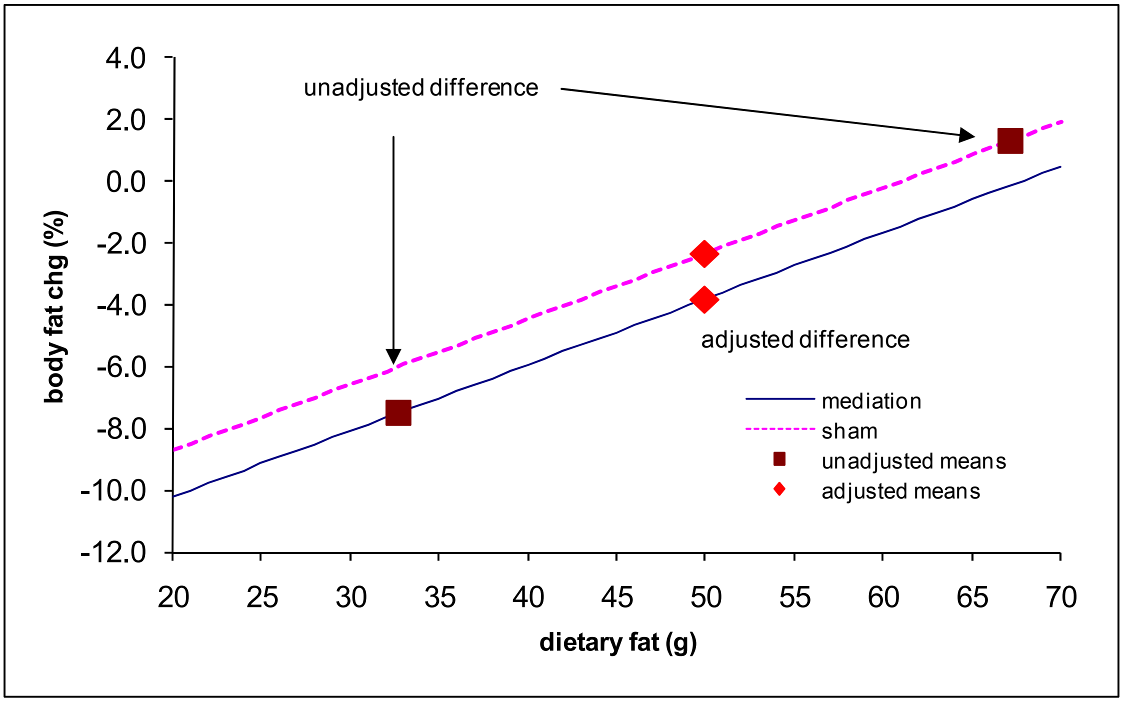
Get by plugging overall mean into regr eqn.

Assumptions:

The <u>**slope**</u> (b) must be the same in all groups! (parallelism)

Usually, we also use the $S_e$ pooled from all groups.

# Example: Meditation and change in percent body fat



# Example:
# Meditation and change in percent body fat

Two groups of overweight persons chose a meditation program or a “sham” (lectures) as part of a weight loss effort. They were NOT randomized.

**Change in percent body fat by treatment group (mediation or sham) over three months**

**Unadjusted Means**

| Level | n | Mean pct body fat change | SEM | Mean dietary fat (gm) |
|---|---|---|---|---|
| 1-meditate | 439 | **-7.51%** | 0.47% | 32.7 g |
| 2-sham | 704 | **1.34%** | 0.35% | 67.1 g |

Unadjusted Mean difference (sham minus meditation) = **8.85%**

SE of the difference = $SE_{diff}$ = $\sqrt{0.472^2 + 0.353^2} = 0.586\%$

t = mean diff/$SE_{diff}$ = 8.85% / 0.586% = 15.1, $p < 0.0001$

Overall unweighted mean dietary fat = 49.9g

**“ Regression” – Y=change in percent body fat vs X=“sham”**

(variable “sham”=0 for meditation or “sham”=1 for sham)

| | |
|---|---|
| RSquare | 0.168 |
| Root Mean Square Error | 9.57 |
| Mean of Response | -2.06 |
| Observations | 1143 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | p value |
|---|---|---|---|---|
| Intercept | -7.51 | 0.457 | -16.43 | <.0001 |
| Sham | **8.85** | 0.582 | 15.20 | <.0001 |

Chg in pct body fat = -7.51% + 8.85% sham.

# Regression controlling for dietary fat and computing adjusted means

### Y=change in percent body fat versus $X_1$=sham, $X_2$=dietary fat

| | |
|---|---|
| RSquare | 0.366 |
| Root Mean Square Error | 8.358 |
| Mean of Response | -2.06 |
| Observations =n | 1143 |

**Parameter Estimates**

| Term | Estimate | Std Error | t Ratio | p value |
|---|---|---|---|---|
| Intercept | -14.47 | 0.543 | -26.64 | <.0001 |
| Sham | 1.51 | 0.558 | 2.71 | 0.007 |
| Diet fat | 0.213 | 0.0113 | 18.88 | <.0001 |

$\hat{Y}$=Chg in pct body fat =-14.5 + 1.51 sham + 0.213 dietary fat

Adjusted means

Meditation: -14.5 + 1.51 (0) + 0.213 (**49.9**) = -3.84%
Sham: -14.5 + 1.51 (1) + 0.213 (**49.9**) = -2.33%

Overall: -14.5 + 1.51 (0.5) + 0.213 (49.9) = -3.09%

**Summary**

| Group | Dietary fat | Unadjusted mean body fat chg | Adjusted mean body fat chg* |
|---|---|---|---|
| 1-meditate | 32.7 g | -7.51% | -3.84% |
| 2-sham | 67.1 g | 1.34% | -2.33% |
| Overall | 49.9 g | -3.09% | -3.09% |
| **difference** | **34.4 g** | **8.85%** | **1.51%** |
| **p value** | **< 0.01** | **< 0.0001** | **0.007** |

* adjusted to overall mean dietary fat (X) of 49.9 gm