Covariance, correlation and linear regression

* We are often interested in studying if/how one continuous variable **covaries** with another – i.e. whether they change in concert

* Covariance between two such variables, Y_1 and Y_2:

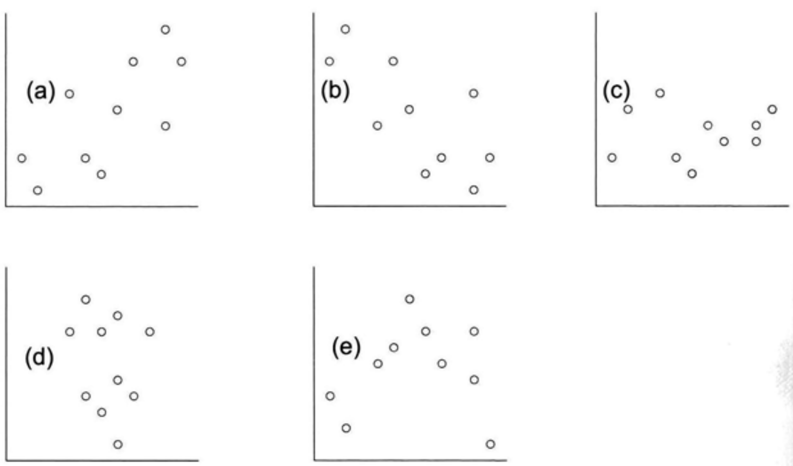
Parameter	Estimate	Standard error
Covariance: $\sigma_{Y_1Y_2}$	$s_{Y_1Y_2} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{n-1}$	n/a
Correlation: $\rho_{Y_1Y_2}$	$r_{Y_1Y_2} = \frac{\sum_{i=1}^{n}[(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)]}{\sqrt{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2\sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2}}$	$s_r = \sqrt{\frac{(1-r^2)}{(n-2)}}$

* Standardized to range between -1 over 0 to 1 → the Pearson correlation coefficient (*r*).

* Tested with a t-test: t = *r* / standard error of *r* (with df = n-2).

* Assumes that Y_1 and Y_2 are both continuous and normally distributed variables (also independence of observations and random sampling).

* Note: only measures strength of **linear** relationships!



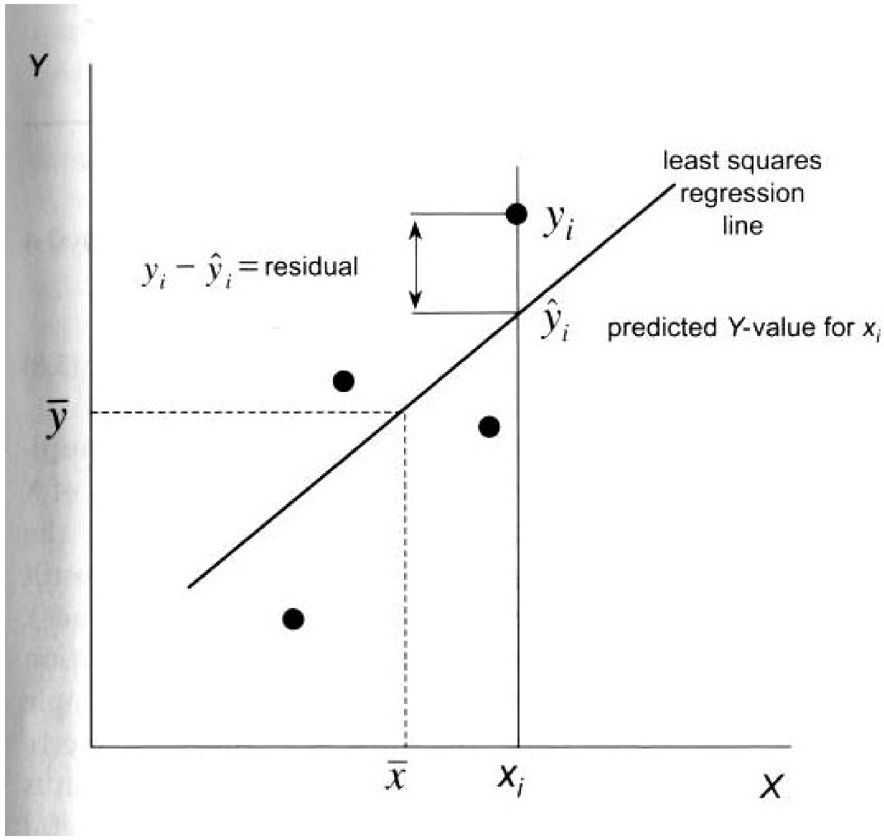
Linear regression

- Whenever you have a continuous variable ***Y*** that you suspect is a response or function of variation in another variable (a predictor variabe) ***X***, you model this with a linear regression (Y: dependent variable; X: independent variable).
- If you want to fit a line in a bivariate space, this line needs a “height” (i.e., an intercept) and a slope....
- $Y = \beta_0 + \beta_1 X + \varepsilon$
- What is the best fit of our model to data?

  1) Least squares minimization – (O)LS (sums of squares of deviations) – simple analytical solution...

  2) Maximum likelihood – ML (maximizes the likelihood of observing our data) – iterative procedure...

Parameter	OLS estimate	Standard error
β_1	$b_1 = \dfrac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$s_{b_1} = \sqrt{\dfrac{MS_{Residual}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
β_0	$b_0 = \bar{y} - b_1\bar{x}$	$s_{b_0} = \sqrt{MS_{Residual}\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
ε_i	$e_i = y_i - \hat{y}_i$	$\sqrt{MS_{Residual}}$ (approx.)



- What can we do with our model? Most commonly....

  1) Test the null hypotheses that our parameters (slope and intercept) are equal to zero: $t = b / s_b$ with $df = n - 2$.

  2) Predict values for Y given a known value of X: $\hat{y} = b_0 + b_1X$

  3) Calculate confidence intervals for our predicted $\hat{y}$

  4) r^2 (or R^2) expresses the proportion of variance in Y that is explained by variation in X (measure of the strength of an association).

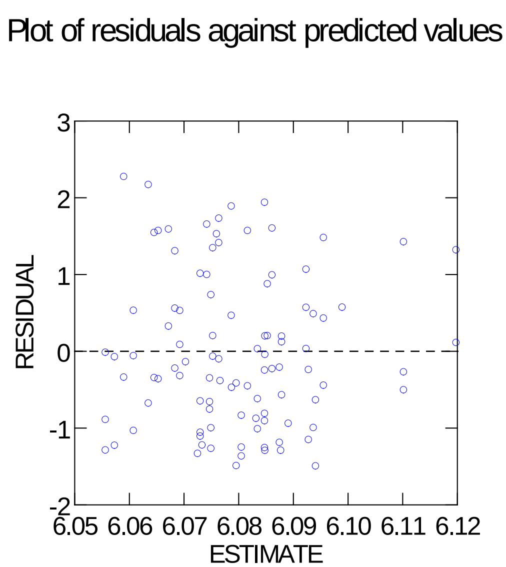
- When fitting a model and testing its parameters, we make five important **assumptions** about our data. These should be checked!

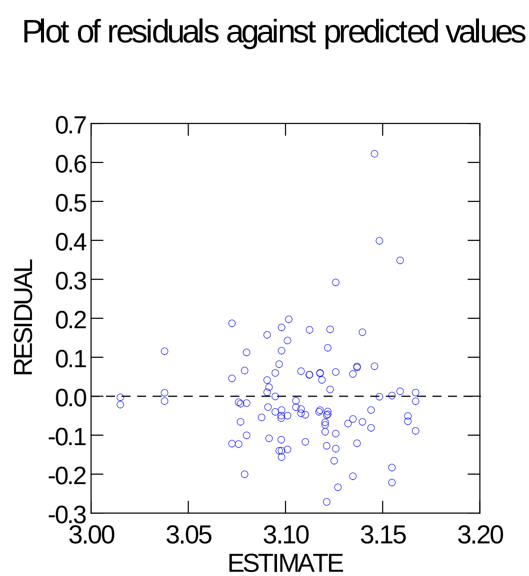
  1) Observations are statistically independent from one another. If not, another model should be employed.

2) The response variable is **continuous**, i.e. can assume any value (so, not a ratio, count data or other forms of discrete variables). If not, use a generlized linear model. Predictor variable (X) can be of any kind.

3) The error (ε) has a normal distribution – test residual distribution with e.g. Shapiro-Wilk's or Kolmogorov-Smirnov test. ***Note that it is the error, not the response variable, that is assumed to be normal.*** If not, easiest to test parameter estimates with a randomization test.

4) Homogeneity of variance (homoscedasticity): Y should have the same variance for all values of X – otherwise a problem with heteroscedasticity! Check best by plotting residuals against predicted value of Y (below). If heteroscedastic, try transforming Y (e.g. a Y' = log [1+Y] transformation) or else test parameter estimates with a randomization test.



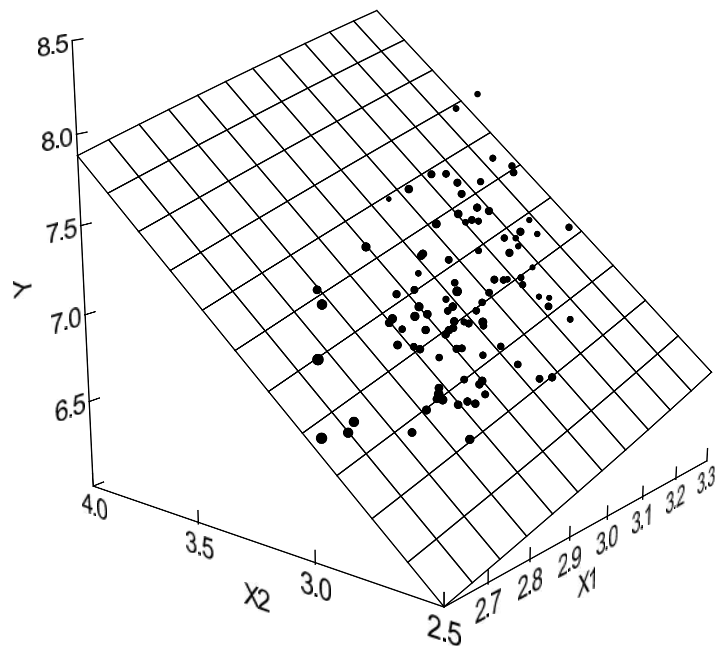


5) The response variable is measured with error (ε) but the independent variable is fixed and measured without error...
- if both X and Y are random variables with similar error, then use type II regression (major axis or reduced major axis regression).

- Common misunderstandning: what does **linear** mean? As in, for example, *linear* regression, general *linear* models, generlized *linear* models... Does **not** necessarily require a linear relationship between response variable and independent variables/factors, but refer to the linear and additive structure of the model ($Y = \beta_0 + \beta_1 X + \varepsilon$) – i.e. is linear in a mathematical sense (as opposed to e.g. $Y = \beta_1 * X^{\beta_2}$).

Multiple regression

- We often wish to simultaneously analyze the independent effects of two (or more) predictor variables on a single continuous response variable:
  1) If all predictors are categorical variables → **ANOVA**
  2) If all predictors are continuous variables → **multiple regression**
  3) A mix of categorical and continuous predictor variables → **ANCOVA** (analysis of covariance)
- All are sometimes referred to collectively as **general linear models**



- If we want to study the independent effect of two or more (correlated) variables on a third → multiple regression!
- The example above – linear regressions:

Effect	Coefficient	Std Error	Std Coef	t	P(2 Tail)
CONSTANT	3.92991	0.79287	0.00000	4.95654	0.00000
X1	1.04283	0.26422	0.37199	3.94688	**0.00015**

and...

Effect	Coefficient	Std Error	Std Coef	t	P(2 Tail)
CONSTANT	2.65455	0.76233	0.00000	3.48216	0.00075
X2	1.42213	0.24608	0.50609	5.77921	**0.00000**

- However, X1 and X2 are correlated (r = 0.22)...one could have an effect through the other → known as a confounding effect.....
- For their independent effects, extend the linear regression model to:

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

Effect	Coefficient	Std Error	Std Coef	t	P(2 Tail)
CONSTANT	0.85202	0.91809	0.00000	0.92803	0.35572
X2	1.25536	0.24061	0.44675	5.21740	**0.00000**
X1	0.77341	0.24004	0.27588	3.22194	**0.00174**

OLS estimates of slopes

SE of the slopes

t = slope / SE of slope

t test of H0: β=0

Standardized partial regression coeff/slopes: "How many SD_Y does Y change when you move one SD_X up along the X-axis..."

In R: model <- lm(scale(Y) ~ scale(X1) + scale(X2), data=file)

How does Y change if we change X1 but hold the other predictors constant? Statistically independent effects...

- In this case; both X variables had ***independent*** effects on Y (not always the case....!!)
- Standardized partial regression coefficients (where variables are put on a common scale – standardized) tells us that the effects X2 is larger than that of X1.
- Model easily extended to *ij* predictor variables;

$$Y = \beta_0 + \beta_i X_i + \ldots\ldots + \beta_j X_j + \varepsilon$$

- Same **five assumptions** apply as for linear regression – should be checked!

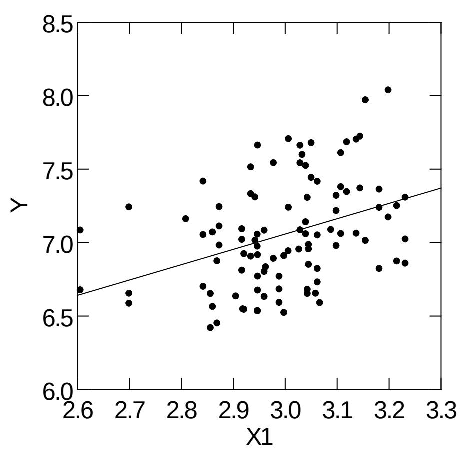
- Additional problem: if the predictor variables are too tightly corrlelated, estimates become biassed and/or model becomes totally inestimable → **collinearity (or, multicollinearity)**. If correlation coefficients between X's are higher than about 0.5, there may be cause for concern. Problem can be assessed with several indicies (see book). If a problem, drop X variables that are highly corrlelated with others (or reduce the number of X-variables by a PCA).

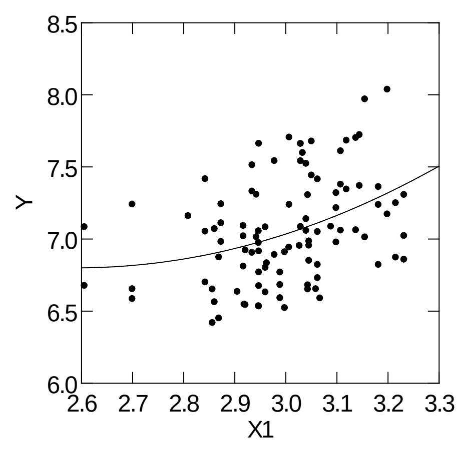
- Common form of multiple regression: **polynomial regression**

  Second order: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ (*i.e., quadratic regression*)

  Third order: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

  Allows tests of non-linear relationships between X and Y (despite being a linear model!). ***NOTE*: independent variable has to be standardized (i.e., $\bar{X} = 0$, SD=1) prior to polynomial regression**!





Linear regression model **Quadratic regression model**

- An unrelated note: always ask yourself if your regression must pass through the origin! If so, drop the intercept, which forces the model through the origin (i.e.; $Y = \beta_1 X + \varepsilon$)

Questions we ask in multiple regression: **comparing hierarchical models**!

- Relies on analysis of variance (next week) – so will be dealt with only in principle here.
- *Is our entire model significant, i.e. does it explain a significant amount of variance in Y?*
  Equals comparing the models

  (A) $Y = \beta_0 + \beta_i X_i + ... + \beta_j X_j + \varepsilon$ *and* (B) $Y = \beta_0 + \varepsilon$

  with a **partial F-test** – is fit to data significatly better in A than in B? This is the overall ANOVA that is often given in stat programs after a regression model has been fitted to data.
- *Is the relationship between X and Y significantly non-linear?*
  Equals comparing the models

  (A) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ *and*

  (B) $Y = \beta_0 + \beta_1 X + \varepsilon$

  with a **partial F-test** – is fit to data significatly better in A than in B?
- *Does the addition of X_{kl} significantly improve a model with X_{ij} already in it?*
  Equals comparing the models

  (A) $Y = \beta_0 + \beta_i X_i + ... + \beta_j X_j + \beta_k X_k + ... + \beta_l X_l + \varepsilon$ *and*

  (B) $Y = \beta_0 + \beta_i X_i + ... + \beta_j X_j + \varepsilon$

  with a **partial F-test** – is fit to data significatly better in A than in B?

When many X's are at hand: which is the best model?

- Model building strategies is a a bit of a shadow land...few simple answers...little consensus...three principal strategies:
- When many X's: **step-wise multiple regression** – comes in two forms:

  1) Forward selection (inclusion): starting with the simplest model with only an intercept, add X's one at the time, starting with the one with the greatest regression coefficient → retain if P is larger than some criteria, drop if it is not.

  2) Backward selection (elimination): start with the full model, drop terms one by one, starting with the least significant, stop when all remaining X's have effects larger than some criteria.

  Main problem: increases the rate of type I errors. If you have 100 X's, you are likely to end up with a nice model with about 5 significant variables... For this reason, mostly used for hypothesis building, rather than testing…
- Run all possible models, then compare fit to data using some criterion:

  1) R^2 doesn't work....more X's means higher R^2...look at **adjusted R^2** (adjusted for the number of X's).

  2) Akaike Information Criterion (AIC) - minimized

  3) Bayesian Information Criterion (BIC)

  4) Mallow's C_p
- Build and compare groups of hierarchical models that make "biological sense", using partial F-tests...