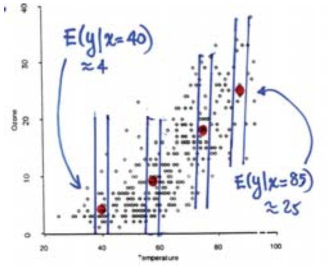
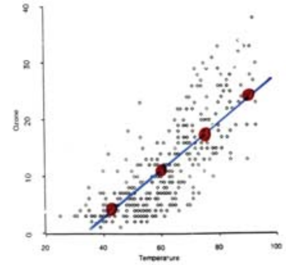
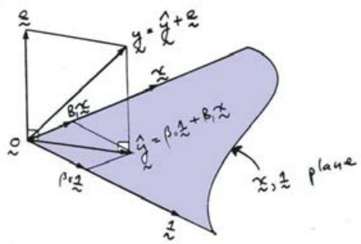
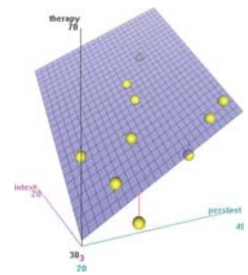




Regression: Introduction

Psychology 6140

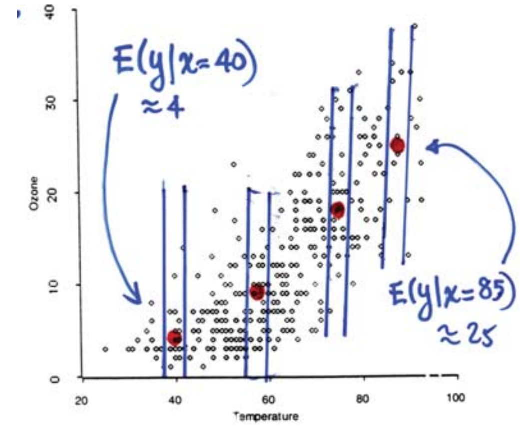




Prototype example: Ozone in LA

How does atmospheric ozone in LA depend on temperature?

- Consider many small slices (x)
- For each x, find average y
- Call this E(y | x)



In general, we want to be able to describe / predict how a response (y) is related to one (or more) explanatory variables (x)

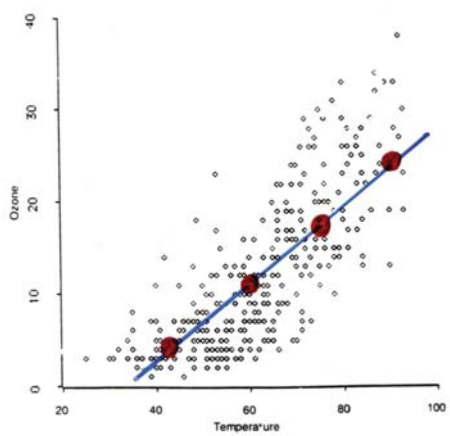
But --- possibly different goals:

- simple description (given data)
- prediction (future data)
- causal explanation – mechanism?

Prototype example: Ozone in LA

If the averages, E(y | x) can be assumed to be linearly related to x, we have a simple **linear** regression model,

$$E(y\,|\,x) = \beta_0 + \beta_1 x$$



Such a description is **always** approximate:

- the true relation of y to x may not be exactly linear (as here)
- y may also depend on other x's

Nevertheless, this is a model we can extend

Linear regression model

- Model: $y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{fixed}} + \underbrace{\varepsilon_i}_{\text{random}}$, where $\begin{cases} \beta_0, \beta_1 : \text{fixed, unknown} \\ x_i : \text{fixed, known} \end{cases}$
- Assumptions:
  - Unbiased: $E(\varepsilon_i) = E(y|x) = 0 \rightarrow$ only x matters
  - Independence: $\text{cov}(\varepsilon_i, \varepsilon_j) = \sigma(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ independent sampling
  - Homogeneity of variance: $\text{var}(\varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2$
- Implies:
  - For each x_i there is a (hypothetical) distribution of y_i values, with

    $E(y_i) = \beta_0 + \beta_1 x_i$ (linear regression)

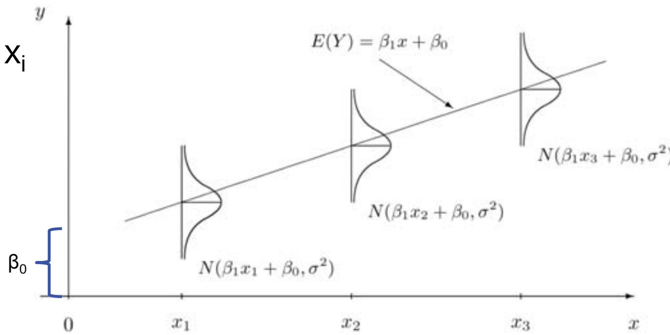
    $\sigma^2(y_i\,|\,x_i) = \sigma^2$ (constant error variance)

In application, assumption of fixed **x** is unrealistic and not necessary. OK as long as residuals meet the assumptions.

Linear regression model

$E(y_i) = \beta_0 + \beta_1 x_i$

$\sigma^2 (y_i \mid x_i) = \sigma^2$



Thus, for a given value of X, we assume that there is a distribution of Y values with constant variance and means linearly related to X

The assumption of a normal distribution is only used for statistical inference

Least squares estimation

- In the linear regression model,

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

For a sample, (x_i, y_i), i=1,2,…n, find estimates, b_0, b_1, which minimize the sum of squared errors

$$SSE = Q(\beta_0, \beta_1) = \sum \epsilon_i^2$$
$$= \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Least squares estimation

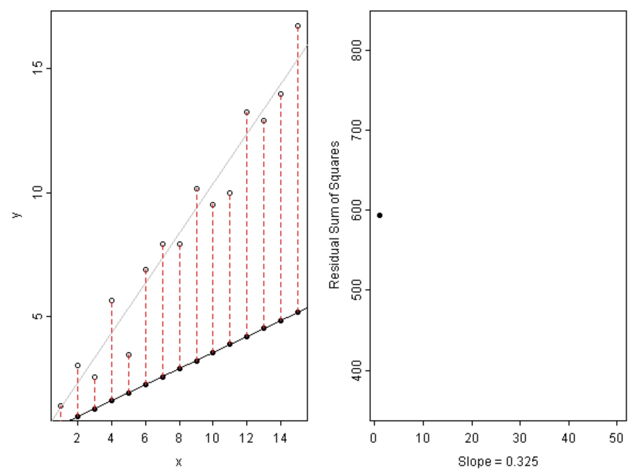
This animation varies the slope of the line and plots the SSE in the panel at the right

For any fixed value of b_0, the SSE is a quadratic function with some minimum – the value of b_1

That's what we want!

Could do the same, varying b_0

-- or better yet, calculus



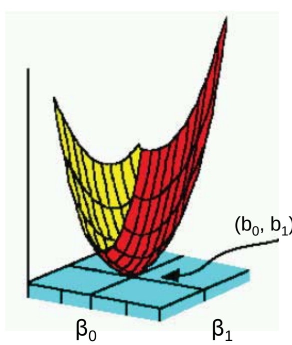
Least squares estimation

- Least squares solution:
  - By calculus, the function $Q(\beta_0, \beta_1)$ has min (or max) where

$$\frac{\partial Q}{\partial \beta_1} = \text{slope of Q} \mid \beta_0 \text{ fixed} = 0$$
$$\frac{\partial Q}{\partial \beta_0} = \text{slope of Q} \mid \beta_1 \text{ fixed} = 0$$

  - Derivatives of SSE = $Q(\beta_0, \beta_1)$

$$\left.\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2\sum (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} &= -2\sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}\right\} \text{for min (or max)}$$



Least squares estimation

- Simplifying → Normal equations

$$\sum y_i = nb_0 + b_1 \sum x_i \qquad \sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 \Rightarrow \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

Two equations in 2 unknowns

- Solve for b_0, b_1:

$$b_0 = (\Sigma y_i - b_1 \Sigma x_i)/n = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n \Sigma x_i^2 - (\Sigma x_i)^2} \Rightarrow \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Solution exists if $(X^T X)$ is non singular

Regression: Matrix notation

- Model: $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ $\qquad \underset{n\times 1}{\mathbf{y}} = \underset{n\times 2}{\mathbf{X}} \; \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$
- Assumptions: $\boldsymbol{\varepsilon} \underset{iid}{\sim} \mathbb{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ "iid": independent and identically distributed
- Least squares: $\min_{\boldsymbol{\beta}} Q = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
- Normal eqns: $\begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{pmatrix}$ or, $(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{y}$
- LS solution: $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

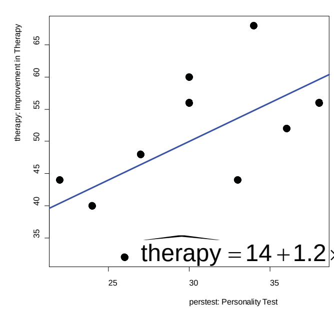
Regression: Matrix notation

- Fitted values: $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$
- Residuals: $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
- Residual SS: $SSE = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}$
- Std errors:

$$s^2 \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = MSE \;\; (\mathbf{X}^T\mathbf{X})^{-1} = \frac{MSE}{\Sigma(x - \bar{x})^2} \begin{pmatrix} 1/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Example: Improvement in Therapy

NAME	SEX	PERSTEST (x)	THERAPY (y)	INTEXT	SX
John	M	26	32	3	0
Susan	F	24	40	4	1
Mary	F	22	44	8	1
Paul	M	33	44	4	0
Jenny	F	27	48	6	1
Rick	M	36	52	4	0
Cathy	F	30	56	10	1
Robert	M	38	56	4	0
Lisa	F	30	60	12	1
Tina	F	34	68	15	1



$\widehat{\text{therapy}} = 14 + 1.2 \times \text{perstest}$

```
proc reg data=therapy;
  model therapy = perstest;
  run;
```

```
lm(therapy ~ perstest,
data = therapy)
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.000	17.204	0.81	0.4393
PERSTEST	1	1.200	0.566	2.12	0.0667

Statistical Inference: Regression

Data : $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$

Assume: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; x's fixed

$\epsilon_i \sim N(0, \sigma^2)$ $\quad$ $N.I.D(0,\sigma^2)$

Sample Estimates
$$\begin{cases} b_1 = \dfrac{n\sum x_i y_i - (\sum x)(\sum y)}{n\sum x_i^2 - (\sum x_i)^2} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}$$

$$\sigma^2 \xrightarrow{est} MSE = \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{(\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i)}{n-2}$$

Statistical Inference: Regression

Classical statistical inference: Use the sample estimate (b_1) to draw a conclusion about population value (β_1)

- Two types: (a) hypothesis tests; (b) confidence intervals

Hypothesis test: $\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}$

"Is there evidence that the true slope is different from 0?"

Confidence interval: Find *c* such that

"What range around b_1 includes the true value β_1 with probability 1-α?"

$$\Pr[\beta_1 \in b_1 \pm c] = \Pr[b_1 - c \le \beta_1 \le b_1 + c] \ge 1 - \alpha$$

These are equivalent, in the sense that if the CI includes 0, the hypothesis test will not reject H_0.

Statistical Inference: Regression

How to go from our single sample estimate (b_1) to the population value (β_1)?

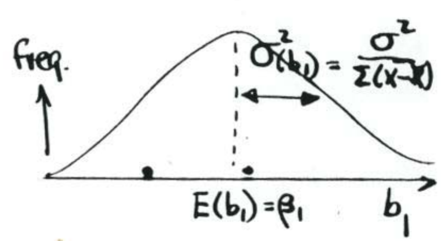
The key idea was that of the sampling distribution of a statistic like b_1.

Sampling distribution: Over many repeated samples of y's, (x's fixed) the dist^y of b_1

(a)* $b_1 \sim N(\cdot,\cdot)$

(b) $E(b_1) = \beta_1$

(c) $\sigma^2(b_1) = \dfrac{\sigma^2}{\sum(x-\bar{x})^2}$



*$b_1 = \sum k_i Y_i \Rightarrow$ Normal by CLT

Statistical Inference: Regression

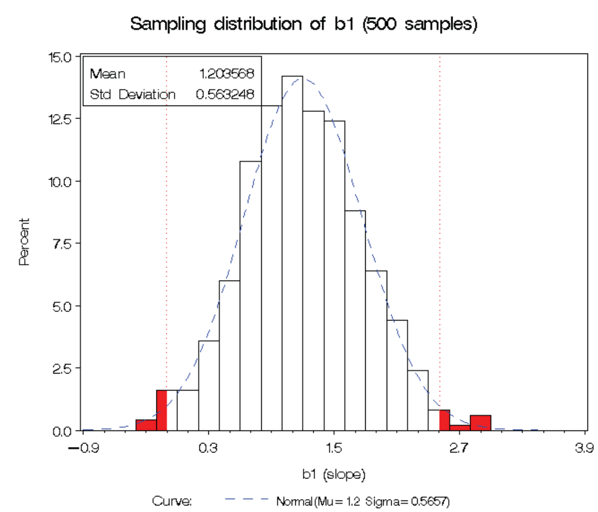
Here we simulated 500 samples from a linear regression in which

$y_i = 14 + 1.2\, x + \varepsilon_i$

and $\varepsilon_i \sim N(0, 80)$

mean $b_1 \approx \beta_1 = 1.2$

std dev $b_1 \approx \sigma(\beta_1) = 0.566$



The theoretical 95 % CI for β_1 is shown by the dotted red lines

Statistical Inference: Regression

<u>Inference</u>

σ^2 unknown $\rightarrow$ estimate it by $s^2_{y\cdot x} \equiv MSE$

then est $\sigma^2(b_1) \equiv s^2(b_1) = \dfrac{MSE}{\Sigma(X_i - \bar{X})^2}$ sampling variance b_1

& $\dfrac{b_1 - \beta_1}{s(b_1)} \sim t$ with $n-2$ df

std error $\rightarrow s(b_1)$

(a) Hypothesis test: $H_0: \beta_1 = 0$ Decision rule

$t^* = \dfrac{b_1 - 0}{s(b_1)}$ If $|t^*| > t_{1-\alpha/2,\, n-2}$ reject H_0 at signif level α

(b) C.I.

$Pr\{b_1 - t_{1-\alpha/2} \cdot s(b_1) \leq \beta_1 \leq b_1 + t_{1-\alpha/2} \cdot s(b_1)\} \geq 1-\alpha$

Regression with SAS: therapy data

```
proc reg data=therapy;
   model therapy = perstest / p;
   output out=results p=fitted r=residual;
   id name;
   run;
```

options

output stats

```
The REG Procedure
Dependent Variable: THERAPY
```

overall model: H_0: $R^2 = 0$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	360.00000	360.00000	4.50	0.0667
Error	8	640.00000	80.00000		
Corrected Total	9	1000.00000			

Root MSE	8.94427	R-Square	0.3600
Dependent Mean	50.00000	Adj R-Sq	0.2800
Coeff Var	17.88854		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.00000	17.20465	0.81	0.4393
PERSTEST	1	1.20000	0.56569	2.12	0.0667

Confidence bands

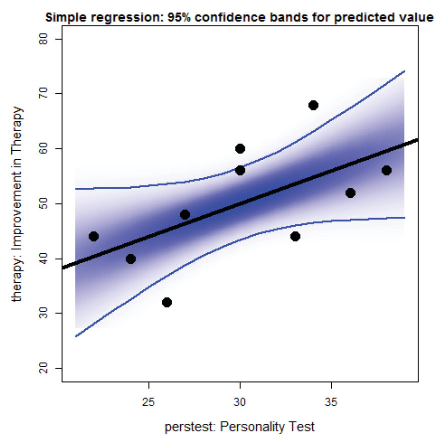
- To understand uncertainty in predicted y, it is useful to calculate and display confidence bands
- For a given value, $x = x_h$

$$\hat{\mathbf{y}}_h = \mathbf{x}_h^T \mathbf{b} \quad \text{where} \quad \mathbf{x}_h^T = (1 \quad x_h)$$

$$s^2(\hat{y}_h) = MSE \times \mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h^T$$

- In SAS, the option is CLM

```
proc reg data=therapy;
    model therapy = perstest / CLM;
```



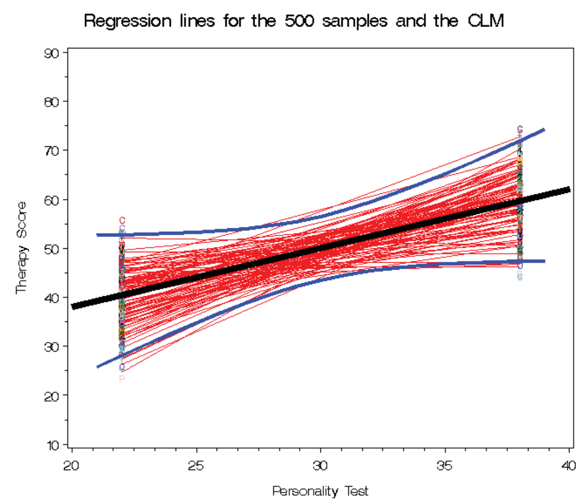
NB: CI gets larger as we move away from mean of X

Confidence bands

The simulation results show why uncertainty increases with distance^2 from the mean of x

$$s^2(\hat{y}_h) = MSE \times \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \right\}$$

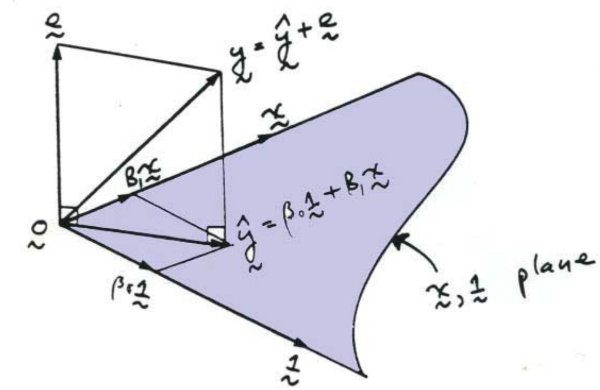
Note that these are limits for the **mean** predicted value (CLM), not for any individual (CLI)



Vector geometry of least squares fit

Model:

$$\mathbf{y} = \hat{\mathbf{y}} + e$$
$$= \beta_0 \mathbf{1} + \beta_1 \mathbf{x} + \mathbf{e}$$



Minimizing $\Sigma e_i^2 = ||\mathbf{e}||^2$

$\rightarrow \hat{\mathbf{y}}$ is orthogonal projection of y onto plane of **x** and **1**

In matrix form:

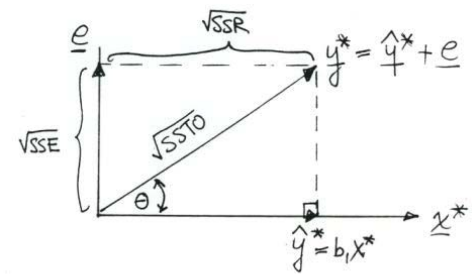
$$\underset{(n\times 1)}{\hat{\mathbf{y}}} = \underset{(n\times n)}{\mathbf{H}} \underset{(n\times 1)}{\mathbf{y}} \quad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Diagonal elements, h_{ii} of the "hat" matrix are measures of "leverage"

Vector geometry of least squares fit

- The vector geometry of regression can be shown in 2D by expressing variables in mean deviation form
- Original model: $y_i = b_0 + b_1 x + e_i$
- Deviation form: $\underbrace{(y_i - \bar{y})}_{y_i^*} = b_1 \underbrace{(x_i - \bar{x})}_{x_i^*} + e_i$
- Then,

$$\mathbf{y}^* = \widehat{\mathbf{y}^*} + \mathbf{e} = b_1 \mathbf{x}^* + \mathbf{e}$$

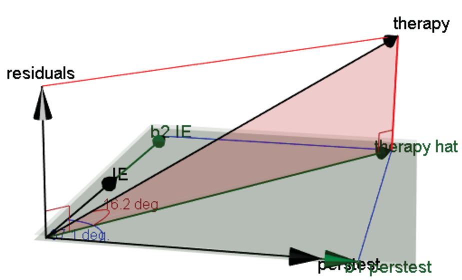


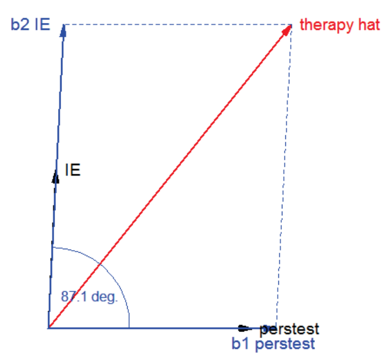
Example: regvec3d

The matlib function `regvec3d()` extends this idea to two predictors, calculating a 3D vector representation of the model y ~ x1 + x2, in deviation form.

The result can be viewed in 2D or 3D accurately reflecting the partial relations of y to x1 and x2.

```
therapy.vec <- regvec3d(therapy ~ perstest + IE, data=therapy)
plot(therapy.vec)
plot(therapy.vec, dimension=2)
```





Vector geometry: ANOVA sums of squares

The ANOVA sums of squares are just the squared lengths of these vectors

Source	DF	Sum of Squares
Model	1	360.00000
Error	8	640.00000
Corrected Total	9	1000.00000

ANOVA:

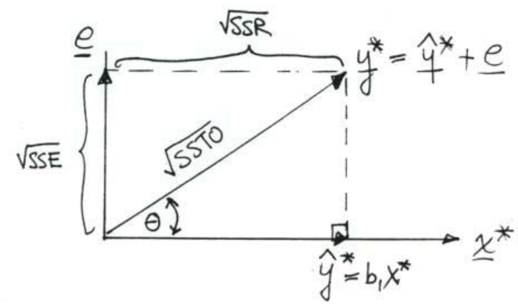
$$||\mathbf{y}^*||^2 = ||\hat{\mathbf{y}}^*||^2 + ||\mathbf{e}||^2$$
$$SSTO = SSR + SSE$$

df: # of dimensions

(n-1) = 1 + (n-2)

R squared:

R^2 = SSR / SSTO

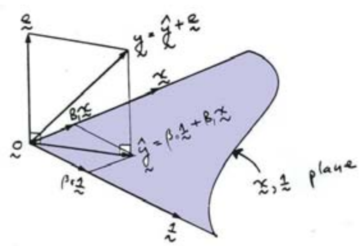


correlation: r = cos (θ)

Vector geometry: Derivation of LS fit

- In the model **y** = (**1**, **x**) **b** + **e** = **X b** + **e**, the residual vector, **e**, is orthogonal to plane of (**1**, **x**)
- This provides another derivation of the LS solution

$$(\mathbf{1}, \mathbf{x})^T \mathbf{e} = \mathbf{X}^T \mathbf{e} = 0$$
$$\rightarrow \mathbf{X}^T (\mathbf{y}-\mathbf{X}\mathbf{b}) = 0$$
$$\rightarrow \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X}\, \mathbf{b} = 0$$
$$\rightarrow \mathbf{X}^T \mathbf{X}\, \mathbf{b} = \mathbf{X}^T \mathbf{y}$$
$$\rightarrow \mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



Multiple regression

Linear model with two independent variables

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

intercept

slope for X_1 with X_2 fixed

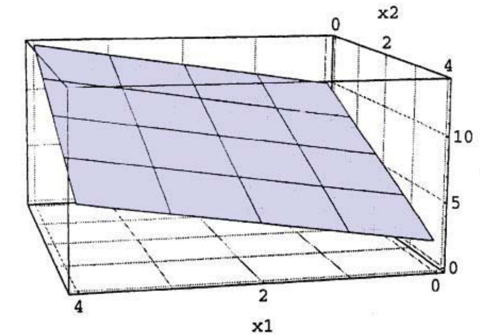
slope for X_2 with X_1 fixed

e.g.,

$$E(Y) = 10 + 1X_1 - 2X_2$$

$\beta_1 = 1$ = change in Y per unit change in X_1 with ***X_2 fixed***.

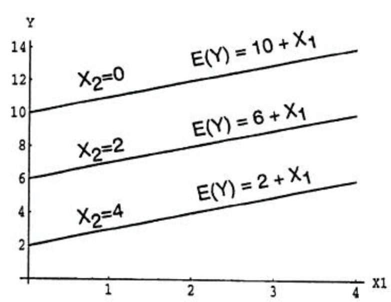
$\beta_2 = -2$ = change in Y per unit change in X_2 with ***X_1 fixed***.



Multiple regression

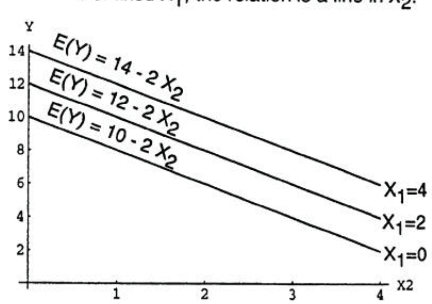
Linear in x_1 and x_2 means:
- we can interpret the slopes b_1 and b_2 w/o regard for the other variable
- at the same time, we are controlling for the other variable

For fixed X_2, the relation is a line in X_1.



The slope for X_1 is the same for all X_2.

For fixed X_1, the relation is a line in X_2.



The slope for X_2 is the same for all X_1.

Multiple regression: therapy data

```
proc reg data=therapy;
   model therapy = perstest intext;
   run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	922.42744	461.21372	41.62	0.0001
Error	7	77.57256	11.08179		
Corrected Total	9	1000.00000			

Overall model test

Root MSE	3.32893	R-Square	0.9224
Dependent Mean	50.00000	Adj R-Sq	0.9003
Coeff Var	6.65787		

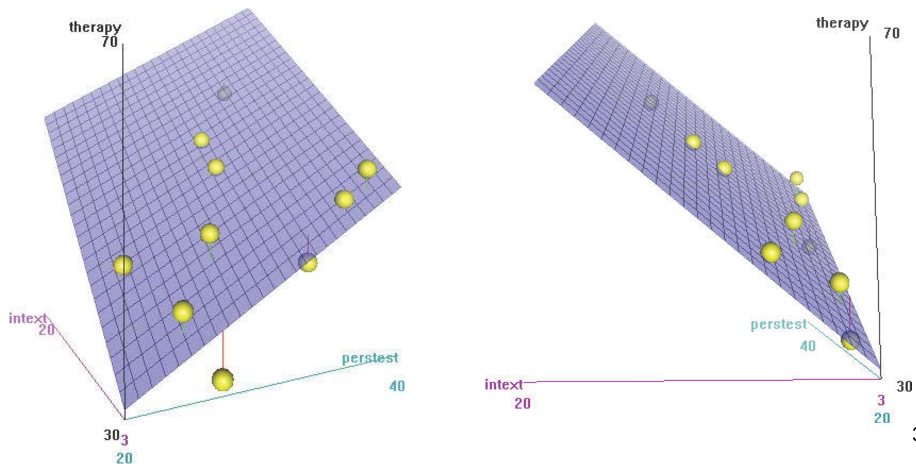
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.82850	6.59255	0.43	0.6808
PERSTEST	1	1.12296	0.21082	5.33	0.0011
INTEXT	1	1.92612	0.27037	7.12	0.0002

Partial tests (more later)

Multiple regression: therapy data

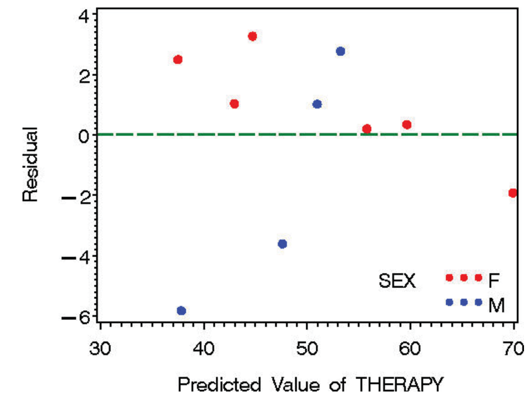
Fitted response surface: $\widehat{therapy} = 2.83 + 1.12\ perstest + 1.92\ intext$



Multiple regression: therapy data

What about sex? (or other x's)

- Residual plots should show no systematic structure
- Here, females tend to have + residuals, suggesting an additional effect of sex on therapy outcome



Multiple regression: therapy data

```
proc reg data=therapy;
   model therapy = perstest intext sx;
   run;
```

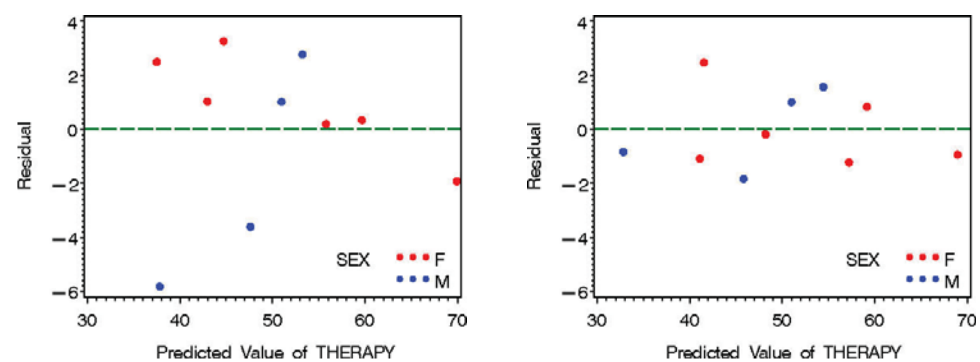
Dummy (0/1) for sex

```
 Analysis of Variance

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

Model                     3          982.05152      327.35051     109.43    <.0001
Error                     6           17.94848        2.99141
Corrected Total           9         1000.00000


Root MSE              1.72957    R-Square     0.9821
Dependent Mean       50.00000    Adj R-Sq     0.9731
Coeff Var             3.45914


                          Parameter Estimates

                        Parameter       Standard
Variable        DF       Estimate          Error    t Value    Pr > |t|

Intercept        1      -14.79157        5.22575      -2.83      0.0299
PERSTEST         1        1.71897        0.17268       9.95      <.0001
INTEXT           1        0.96956        0.25620       3.78      0.0091
SX               1       10.72600        2.40251       4.46      0.0043
```


Multiple regression: therapy data

Model 2: ignoring Sex

Model 3: including Sex

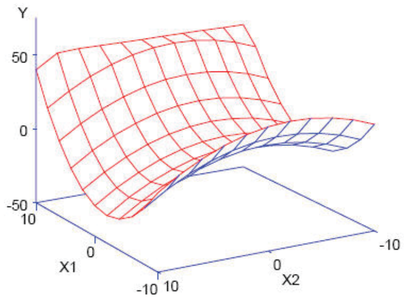


Benefits: Residuals no longer associated with sex
Residual SSE now considerably smaller: smaller std errors

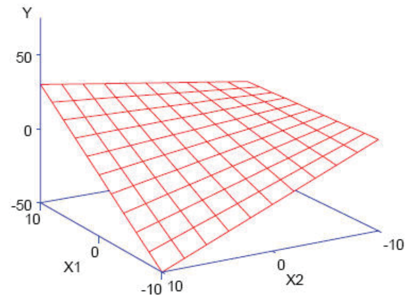
More general linear models…

- Response surface models:
  - Q: Is the relation of y to x_1 and x_2 *linear*?
  - M: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_2 x_{i2} + \beta_{22} x_{i2}^2 + \epsilon_i$
- Models with interactions:
  - Q: Is the relation of y to x_1 the *same* for all x_2 ?
  - M: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i$

Quadratic response surface
Y = 2*X1 + -1*X2 +0.5*X1*X1+-0.2*X2*X2



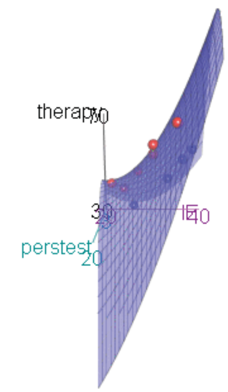
X1 * X2 Interaction
Y = 2*X1 + -1*X2 +0.20*X1*X2



Therapy data: Quadratic response surface model

```
mod3 <- lm(therapy ~ poly(perstest, IE, degree=2), data=therapy)
mod3 <- lm(therapy ~ (perstest + IE)^2 + I(perstest^2) + I(IE^2))
```

```
> anova(mod3)
Analysis of Variance Table

Response: therapy
               Df Sum Sq Mean Sq F value  Pr(>F)
perstest        1    360     360   60.12 0.00149 **
IE              1    562     562   93.93 0.00063 ***
I(perstest^2)   1     14      14    2.27 0.20638
I(IE^2)         1     23      23    3.76 0.12455
perstest:IE     1     18      18    2.93 0.16228
Residuals       4     24       6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



More general linear models…

In each case, we can represent the model in the *same* form:

y_i =	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$	$+ \epsilon_i$
response =	wtd. sum of predictors	+ residual
data =	explained (partial summary)	+ unexplained

where the xs can be:

- Quantitative regressors: age, income, education
- Transformed regressors: $\sqrt{\text{age}}$, log(income)
- Polynomial regressors: age^2, age^3, $\cdots$
- Categorical predictors: treatment, sex— coded as “dummy” (0/1) variables
- Interaction regessors: treatment $\times$ age, sex $\times$ age
- Any combinations of the above $\Rightarrow$ the General Linear Model

“Linear model” $\longrightarrow$ linear in the *parameters*, $\beta_1, \beta_2, \beta_3, \ldots$, e.g.,

$$y_i = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \beta_3 \log(\text{income}) + \beta_4 (\text{sex='F'}) + \epsilon_i$$

More general linear models…

All of these can be represented in matrix form,

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

or,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (2)$$

In all cases,

- Parameter estimates: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$
- Residuals = estimated errors = $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$
- Residual variance: MSE $\equiv \widehat{\text{Var}}(\epsilon) = (\boldsymbol{e}^{\mathsf{T}}\boldsymbol{e})/(n-p-1)$
- Standard errors: Var$(\widehat{\boldsymbol{\beta}})$ = MSE$(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}$
- Parameter tests: $H_0 : \beta_i = 0 \Rightarrow t = \widehat{\beta}_i / \sqrt{\text{Var}(\widehat{\beta}_i)} \sim t(n-p-1)$

Fitting linear models in SAS: PROC REG

- PROC REG
  - One (or more) *quantitative* response variable(s)
  - + Extensive facilities for regression diagnostics
  - + Model selection methods: stepwise, forward, backward
  - + PLOT statement → plots of *any* data or computed variables

```
proc reg data=...;
  model y = X1 X2 X3 /          /* MRA, influence stats */
        influence partial;
  plot nqq. * r.;               /* Normal QQ plot       */
  model y = X1-X5 /             /* MRA, model selection */
        selection = stepwise sle=0.10;
```

  - + V9.1.3: ODS GRAPHICS → easy plots, automatically

- — no CLASS statement— must create dummy variables (DUMMY macro)
- — no | notation— must create interaction terms (INTERACT macro)

```
data test;
   input x y group $ sex $ @@;
cards;
 5 10  A M   8  12  A F   9 13  A M  10 18  B M   16 19  B M
10 16  B F  15  21  C M  13 19  C F  15 20  C M
;
  *-- Dummy variables for Sex and Group;
%dummy (data=test, var =sex group, prefix=Sex_ Gp_);
  *-- Interaction of X * Sex;
%interact(data=test, v1=x, v2=Sex_F, names=XSex);
proc print noobs; run;
```

Produces:

x	y	group	sex	SEX_F	GP_A	GP_B	XSex
5	10	A	M	0	1	0	0
8	12	A	F	1	1	0	8
9	13	A	M	0	1	0	0
10	18	B	M	0	0	1	0
16	19	B	M	0	0	1	0
10	16	B	F	1	0	1	10
15	21	C	M	0	0	0	0
13	19	C	F	1	0	0	13
15	20	C	M	0	0	0	0

Fitting linear models in SAS: PROC GLM

- PROC GLM
  - One (or more) *quantitative* response variable(s)
  - Multiple response variables → multivariate analyses or repeated measures
  - GLM model syntax: regression effects (covariates)

```
proc glm data=...;
  model y = X1;                /* simple linear regression   */
  model y = X1 X2 X3;          /* multiple linear regression */
  model y = X1-X5;             /* multiple linear regression */
  model y = wages--education;  /* multiple linear regression */

  model y = X1 X1*X1 X1*X1*X1;     /* polynomial regression  */
  model y = X1 X2 X1*X2;           /* interaction model      */
  model y = X1 X2 X1*X1 X2*X2 X1*X2;  /* response surface    */
```

  - Bar notation: A | B | C → A B C A*B A*C B*C A*B*C

```
proc glm data=...;
    *-- same, using '|' notation;
  model y = X1 | X1 | X1;         /* polynomial regression  */
  model y = X1 | X2;              /* interaction model      */
  model y = X1 | X1 | X2 | X2 @2; /* response surface       */
```


Fitting linear models in R: lm()

- In R, much simpler: **lm()** for everything
  - Regression models (X1, ... quantitative)

```
lm(y ~ X1, data=dat)           # simple linear regression
lm(y ~ X1+X2+X3, data=dat)     # multiple linear regression
lm(y ~ (X1+X2+X3)^2, data=dat) # all two-way interactions
lm(log(y) ~ poly(X,3), data=dat) # arbitrary transformations
```

  - ANOVA/ANCOVA models (A, B, ... factors)

```
lm(y ~ A)               # one way ANOVA
lm(y ~ A*B)             # two way: A + B + A:B
lm(y ~ X + A)           # one way ANCOVA
lm(y ~ (A+B+C)^2)       # 3-way ANOVA: A, B, C, A:B, A:C, B:C
```

Fitting linear models in R: lm()

- Multivariate models: **`lm()`** for everything
  - Multivariate regression

```
lm(cbind(y1, y2) ~ X1 + X2 + X3)                 # std MMreg: all linear
lm(cbind(y1, y2) ~ poly(X1,2) + poly(X2,2))     # response surface
```

  - MANOVA/MANCOVA models

```
lm(cbind(y1, y2, y3) ~ A * B)     # 2-way MANOVA: A + B + A:B
lm(cbind(y1, y2, y3) ~ X + A)     # MANCOVA (equal slopes)
lm(cbind(y1, y2) ~ X + A + X:A)   # heterogeneous slopes
```


Working with lm() objects

- R functions → objects, which have methods
- **`print(obj)`** gives just basic output

```
> # fit some models
> mod1 <- lm(therapy ~ perstest, data= therapy)
> print(mod1)

Call:
lm(formula = therapy ~ perstest, data = therapy)

Coefficients:
(Intercept)     perstest
       14.0          1.2
```


Working with lm() objects

- **`summary(obj)`** gives more detailed results

```
> summary(mod1)

Call:
lm(formula = therapy ~ perstest, data = therapy)

Residuals:
   Min     1Q Median     3Q    Max
 -13.2   -4.8   -0.6    5.4   13.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.0000    17.2047   0.814   0.4393
perstest      1.2000     0.5657   2.121   0.0667 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.944 on 8 degrees of freedom
Multiple R-squared:  0.36,        Adjusted R-squared:  0.28
F-statistic:   4.5 on 1 and 8 DF,  p-value: 0.06669
```


Working with lm() objects

- **`plot(model)`** gives diagnostic plots

```
> plot(mod1)
```

These show possible problems in the residuals:

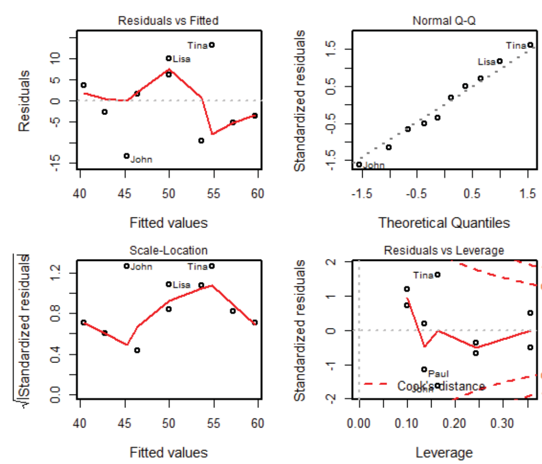
(a) Systematic pattern?

(b) Normal?

(c) Constant variance?

(d) Influential points?

Better versions in many R packages (car)



Working with lm() objects

- anova() tests differences among nested models

```
> mod2 <- lm(therapy ~ perstest + intext, data=therapy)
> mod3 <- lm(therapy ~ perstest + intext + sex, data=therapy)
> anova(mod1, mod2, mod3)
Analysis of Variance Table

Model 1: therapy ~ perstest
Model 2: therapy ~ perstest + intext
Model 3: therapy ~ perstest + intext + sex
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1      8 640.00
2      7  77.57  1    562.43 188.014 9.352e-06 ***
3      6  17.95  1     59.62  19.932  0.004262 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note: these are so-called “Type I” (sequential) tests, testing the additional contribution of each new predictor. Other (“Type II”) tests are more generally useful.

Summary, to here

- Simple linear regression:
  - Fit a model predicting $E(y \mid x) = \beta_0 + \beta_1 x$
  - Use least squares to find estimates, b_0, b_1
  - Matrix solution: $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- Multiple regression:
  - Include any number of linear predictors
  - $E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$
  - Partial coefficients: Effect of x_i controlling for others
  - Can include terms like x^2,x^3, x_1*x_2, factor variables, etc.
  - For all, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$s^2(\mathbf{b}) = MSE \quad (\mathbf{X}^T\mathbf{X})^{-1}$$

What we still have to learn

- Model assessment
  - How to judge the contributions of different Xs?
    - Type I (sequential) and Type II (partial) tests
    - Principle of marginality (main effects & interactions)
  - Ordered (“hierarchical”) tests
- Model diagnosis
  - How to see and test for violations of assumptions
  - Regression diagnostics: influential observations???
  - Detecting and dealing with collinearity
- Model building/selection strategies
  - How to select an adequate/optimal subset of predictors
  - Dangers of “stepwise” selection
  - Cross-validation, shrinkage, LASSO methods