

Statistics 210 – Part 2
Statistical Methods

Hal S. Stern

Department of Statistics
University of California, Irvine
sternh@uci.edu

- Thus far:
 - design of experiments
 - two sample methods
 - one factor ANOVA
 - pairing/blocking
- To come:
 - simple regression/correlation
 - multiple regression
 - regression and ANOVA
 - factorial ANOVA
 - random/mixed effects

Simple linear regression

Introduction

- Methods for studying the relationship of two or more quantitative variables
- Examples:
 - predict salary from education, years of experience, age
 - find effect of lead exposure on school performance
- Functional or mathematical relation:
 $Y = f(X)$ (deterministic)
- Structural or statistical relation:
 $Y = f(X) + \text{error}$ (stochastic)
- Additional reference:
Applied Linear Regression by S. Weisberg

Simple linear regression

Linear regression model

- The basic model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$$

- Y_i is the response or dependent variable
 - x_i is the predictor, explanatory variable, independent variable
 - x_i is treated as a fixed quantity
(or if random it is conditioned on)
 - ϵ_i is the error term or individual variation
 - ϵ_i are iid $N(0, \sigma^2)$ random variables
- Key assumptions (will check these later)
 - x 's are fixed (or conditioned upon)
 - correct mean/model specification (linearity)
 - independent (uncorrelated) errors
 - constant variance errors
 - normally distributed errors

Simple linear regression

Interpreting the model

- Model can also be written as

$$Y_i | X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

- mean of Y given $X = x$ is $\beta_0 + \beta_1 x$
(known as the conditional mean)
 - β_0 is conditional mean when $x = 0$
 - if we replace x by $x - x_0$ then β_0 is interpreted as conditional mean when $x = x_0$
 - β_1 = slope, change in mean of Y per 1 unit change in x
 - σ^2 = variation of responses about the mean
- Relationship to ANOVA
 - ANOVA - each group has its own mean
 - Each x_i in regression defines its own group
 - But ... too many groups with too few observations per group
 - Therefore stronger assumption about the means (linear structure)

Simple linear regression

A bit of history

- Galton's sweet pea seeds (1870s)
 - experiment:
 - * created seven groups of seeds
(lightest seeds to heaviest seeds)
 - * planted seeds and grew pea plants
 - * studied seeds from these “offspring”
 - offspring seeds were normal in each group with the same spread
 - total distn of offspring in the seven groups looked normal with same variance as in parents
 - why didn't variance grow?
 - offspring seeds' means had “regressed” towards the population mean
- Galton studied human inheritance in 1880s
(parent and child heights .. a symmetric case)

Simple linear regression

Estimation

- Least squares estimation

- choose b_o, b_1 to minimize

$$g(b_o, b_1) = \sum_{i=1}^n (Y_i - (b_o + b_1 x_i))^2$$

- why squared errors? (convenient math)
- why vertical distances? (Y is response)
- result:

$$b_o = \bar{Y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- these are best linear unbiased estimates
- predicted (fitted) value $\hat{Y}_i = b_o + b_1 x_i$
- residual $e_i = Y_i - \hat{Y}_i$

- Maximum likelihood estimation

- we can write down joint distn of all of the Y 's, sometimes known as the likelihood function

$$L(b_o, b_1, \sigma^2) = \prod_{i=1}^n N(Y_i | b_o + b_1 x_i, \sigma^2)$$

- gives same estimates for b_o, b_1

Simple linear regression

Estimation - some details

- Least squares estimation:
choose b_o, b_1 to minimize

$$g(b_o, b_1) = \sum_{i=1}^n (Y_i - (b_o + b_1 x_i))^2$$

- Taking derivatives and setting them equal to zero yields normal equations

$$\begin{aligned} b_o n + b_1 \sum x_i &= \sum Y_i \\ b_o \sum x_i + b_1 \sum x_i^2 &= \sum x_i Y_i \end{aligned}$$

- Normal equations are equivalent to

$$\begin{aligned} \sum e_i &= 0 \\ \sum e_i x_i &= 0 \end{aligned}$$

- Second derivatives guarantee we have minimum
- Note: estimation (and interpretation) are different if there is no intercept in the model

Simple linear regression

Estimation of error variance

- Common estimate of σ^2 is

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (b_0 + b_1 x_i))^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

- The $\frac{1}{n-2}$ makes this an unbiased estimate
- Maximum likelihood estimate, $\frac{1}{n} \sum_i e_i^2$, is too small on average

Simple linear regression

Inference - ANOVA table

- Recall the ANOVA decomposition

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..})$$

- For regression we can write

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

- In terms of sums of squares

$$\begin{aligned} SS(\text{total, corrected}) &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\ &= SS(\text{residuals}) + SS(\text{model}) \end{aligned}$$

- Cross product term is

$$2 \sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 2 \sum_i e_i(b_0 + b_1 x_i - \bar{Y}) = 0$$

because $\sum_i e_i = \sum_i e_i x_i = 0$

- $SS(\text{residuals})$ is also known as $SS(\text{error})$, SS_{error} , SSE
- $SS(\text{model})$ is also known as SS_{model} , $SS_{\text{regression}}$

Simple linear regression

Inference - ANOVA table

- Collect sums of squares in ANOVA table

source of variation	degrees of freedom	sums of squares	mean square
model	1	SS_{model}	SS/df
error	$n - 2$	SS_{error}	SS/df
total	$n - 1$	SS_{total}	

- Note that $MS_{error} = MSE$ is usual variance estimate s_e^2
- $R^2 = SS_{model}/SS_{total}$ is the amount of variation in the response that is explained by the model
- $E(MSE) = \sigma^2$
 $E(MS_{model}) = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$
- $F = MS_{model}/MSE$ will have $F_{1,n-2}$ distn if $\beta_1 = 0$ and tend to be large otherwise, which leads to F -test for $H_o : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$
- Note similarity with ANOVA ... $\beta_1 = 0$ implies all “groups” (defined by x) have the same mean

Simple linear regression

Inference for β_1

- Many quantities of interest in regression analysis:
 $\beta_0, \beta_1, \hat{Y}$ at a given x , and more
- Discuss β_1 in detail (then summarize the rest)

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- b_1 is a linear combination of normal random variables (the Y_i 's) so b_1 is normally distributed

$$E(b_1) = \beta_1 \quad \text{Var}(b_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

- σ^2 is unknown; plug in estimate s_e^2
(s_e^2 has chi-squared distn)
- Standard error of b_1 is $s_{b_1} = \sqrt{s_e^2 / \sum_i (x_i - \bar{x})^2}$
- $(b_1 - \beta_1) / s_{b_1}$ has t -distn with $n - 2$ d.f.
- $100(1 - \alpha)\%$ CI for β_1 is $b_1 \pm t_{n-2, 1-\alpha/2} s_{b_1}$
- Test $H_0 : \beta_1 = 0$ using $t = b_1 / s_{b_1}$ vs t_{n-2} distn
- Note that t^2 is the ANOVA F -statistic

Simple linear regression

Inference for other quantities

- β_o
 - $b_o \sim N(\beta_o, \sigma_{b_o}^2 = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}))$
 - b_o has std err $s_{b_o} = \sqrt{s_e^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2})}$
 - t inference ($n - 2$ d.f.)
- $E(Y|X = x) = \beta_o + \beta_1 x$
 - estimate is $\hat{Y} = b_o + b_1 x$
 - std. err. is $s_{\hat{y}} = \sqrt{s_e^2(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2})}$
 - t inference
 - use multiple comparisons proc to get CIs for several x 's or the whole line: $\hat{Y} \pm \sqrt{2F_{2, n-2, 1-\alpha}} s_{\hat{y}}$

Simple linear regression

Inference for other quantities (cont'd)

- Y at given x (prediction) = $Y_{new} = \beta_o + \beta_1 x + \epsilon$
 - estimate is still $\hat{Y} = b_o + b_1 x$
 - std. err. is $s_{pred} = s_p = \sqrt{s_e^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}$
 - t inference
- X_{new} corresponding to given new Y_{new}
(the calibration problem)
 - estimate is $\hat{X} = (Y_{new} - b_o)/b_1$
 - approx std. err. is $s_{\hat{x}} = s_p/|b_1|$
 - approx t distn with $n - 2$ d.f.
 - only works if b_1 is highly significant

Simple linear regression

Model diagnostics (short version)

- Recall the key assumptions
 - x fixed (or conditioned upon)
 - correct mean/model specification
 - independent (uncorrelated) errors
 - constant variance errors
 - normal distn for errors
- Here: effects, diagnosis, a small amount on remedies
- More diagnostics (especially remedies) later with multiple regression discussion

Simple linear regression

Model diagnostics - effects of violations

- Random X's/Measurement error (next slide)
- Incorrect model specification
 - linear model will fit poorly
 - parameter estimates are biased/meaningless
 - includes case of linear model with important variables omitted
- Non-independence
 - parameter estimates are unbiased
 - standard errors are a problem and thus so is inference
- Nonconstant variance
 - parameter estimates are unbiased
 - standard errors are a problem
- Nonnormality
 - least important
 - inference is fairly robust to nonnormality
 - important effects on prediction intervals

Simple linear regression

Model diagnostics - effects of violations

- Random x values
 - problem occurs if: $Y = \beta_o + \beta_1 X + \epsilon$, we observe $x = X + \delta$, and we regress Y on x
 - known as measurement error problem
 - “true” model can be written in terms of x as $Y = \beta_o + \beta_1 x + (\epsilon - \delta\beta_1)$
 - $\text{Cov}(x, \epsilon - \delta\beta_1) < 0$ which causes bias in b_1
 - $E(b_1) \approx \frac{\beta_1}{1 + \frac{n\sigma_{\epsilon}^2}{\sum_i (x_i - \bar{x})^2}} < \beta_1$
 - there exist measurement error models for handling situations like this (not in Stat 210)

Simple linear regression

Model diagnostics - residuals

- For other assumptions (model spec., indep. errors, const. var., normality), model checking uses residuals
- Recall $e_i = Y_i - \hat{Y}_i$
- In regression $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 x_i$
- e_i is an approximation of ϵ_i
- Important properties
 - sum of residuals is zero
 - $\sum_i x_i e_i = \sum_i \hat{Y}_i e_i = 0$
 - $e_i \sim N\left(0, \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)\right)$
(residuals are not constant variance)
 - e_i 's are negatively correlated (sum is zero)
- Sometimes use $r_i = e_i / \sqrt{MSE \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}$
(known as studentized residuals)

Simple linear regression

Diagnosing violations with residual plots

- Plot residuals versus predicted values
 - detect nonconstant variance
 - detect nonlinearity
 - detect outliers
- Plot residuals versus X
 - in simple linear regression this is same as above
 - in multiple regression will be useful
- Plot residuals versus other possible predictors (e.g., time)
 - detect important missing variable
- Plot residuals vs lagged residuals
 - detect correlated errors
- Normal probability plot of residuals
 - detect nonnormality

Simple linear regression

Remedies for violated assumptions

- Transformation of Y
- Adding/modifying predictors
- More sophisticated models/estimation
 - weighted least squares for nonconstant variance
 - time series models for correlated errors
 - robust regression methods for nonnormality
- These will be described more fully under multiple regression

Simple linear regression

Test for lack of fit

- One last (clever) idea for model checking, a statistical test for lack of fit (Sect 3.7-3.9)
- Suppose we have multiple observations at one or more of the x_i 's
- Then we can compare the ANOVA model (separate mean at each x_i) with the regression model (all means are $\beta_0 + \beta_1 x_i$)
- Need ANOVA notation: Y_{ij} is j th obs at x_i

$$\begin{aligned}SS_{error} &= \sum_i \sum_j (Y_{ij} - \hat{Y}_i)^2 \\ &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \hat{Y}_i)^2 \\ &= SS_{pure-error} + SS_{lack-of-fit}\end{aligned}$$

- $SS_{pure-error}$ is a “pure” estimate of σ^2 because all observations at x_i have the same mean
- $SS_{lack-of-fit}$ is an estimate of error variance only if the linear model is correct

Simple linear regression

Test for lack of fit (cont'd)

- Let $r =$ number of distinct x values
- New and improved ANOVA table

source of variation	degrees of freedom	sums of squares
model	1	SS_{model}
lack-of-fit	$r - 2$	$SS_{lack-of-fit}$
error	$n - r$	$SS_{pure-error}$
total	$n - 1$	SS_{total}

- $E(MS_{pure-error}) = \sigma^2$
 $E(MS_{lack-of-fit}) = \sigma^2 +$ lack of fit term
- $F = MS_{lack-of-fit} / MS_{pure-error}$ will have $F_{r-2, n-r}$ distn if model fits and tend to be large otherwise
- Carry out F -test of $H_o : E(Y|X = x) = \beta_o + \beta_1 x$ vs $H_a :$ model doesn't fit
- If reject - need a new model
- If don't reject - can pool the two SSE terms to get back to original regression ANOVA table

Correlation

Introduction

- Correlation is closely related to regression
- Regression analysis
 - one variable is the response Y
 - one variable is the predictor X
 - model distn of Y given X
 - distn of X is not relevant
- Correlation analysis
 - both variables are of interest
 - want to measure association

Correlation

Bivariate normal distribution

- The bivariate normal distribution is an example of a joint distribution for two continuous random variables
- We say X and Y have the bivariate normal distribution with parameters $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho$
- The probability density is

$$f(x, y) = \frac{1}{2\pi\sqrt{\sigma_x^2\sigma_y^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)} \cdot \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \frac{(y-\mu_y)^2}{\sigma_y^2} \right]\right)$$

- Interpretation of parameters
 - μ_x = mean of X
 - μ_y = mean of Y
 - σ_x^2 = variance of X
 - σ_y^2 = variance of Y
 - ρ = correlation of X and Y (see next slide)

Correlation

Definition of the correlation coefficient

- The parameter ρ is called the correlation coefficient
- Recall that $\text{Var}(X) = E(X - \mu_x)^2$ and $\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$
- Definition: $\rho = \text{Cov}(X, Y) / (\sigma_x \sigma_y)$
- Properties:
 - $-1 \leq \rho \leq 1$
 - $|\rho| = 1$ implies that X and Y are linearly related
 - $\rho = 0$ implies independence (of two normals)
 - $\rho = 0$ implies no linear pattern

Correlation

Properties of bivariate normal distn

- Density is constant on ellipses (contour plots)
- Marginal distns are normal
 $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$
- Conditional distns are normal, e.g.,

$$Y \mid X = x \sim N\left(\mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x), \sigma_y^2 (1 - \rho^2)\right)$$

- Note similarity of conditional distn to simple linear regression model
- Two motivations for simple linear regression
 - bivariate normal observations
 - x fixed (not necessarily normal), $Y|x$ is normal
- We can relate the bivariate normal and simple linear regression parameterizations:

$$\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x = \beta_0$$

$$\rho \frac{\sigma_y}{\sigma_x} = \beta_1$$

$$(1 - \rho^2) \sigma_y^2 = \sigma^2$$

$$\mu_x = \mu_x$$

$$\sigma_x^2 = \sigma_x^2$$

Correlation

Inference for correlation coefficient

- The maximum likelihood estimate is the sample correlation coefficient (Pearson correlation)

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

- the obvious estimate, replacing population quantities by sample quantities
- biased (some corrections are available but not usually applied)
- relation to regression:
 - * $r^2 = \frac{SS_{model}}{SS_{total}}$ is proportion of variance explained
 - * $r = \text{corr}(Y, \hat{Y})$
- Test of $H_o : \rho = 0$
 - this is a special inference question
 - equivalent to testing $\beta_1 = 0$ in regression
 - $t = r\sqrt{n-2}/\sqrt{1-r^2}$ has t -distn with $n-2$ d.f.
 - in fact this is exactly the t -test for $H_o : \beta_1 = 0$

Correlation

Inference for correlation coefficient

- Remaining inference procedures rely on Fisher's z-transformation

$$Z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

- Z_r is approx $N\left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3}\right)$
- Transformation creates approx normality and variance indep of ρ
- Confidence interval for ρ
 - 100(1 - α)% CI for $\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$ is $Z_r \pm z_{1-\alpha/2} \sqrt{\frac{1}{n-3}}$
 - invert transformation using $r = (e^{2Z_r} - 1)/(e^{2Z_r} + 1)$
- Testing $H_o : \rho = \rho_o$
 - test statistic $z = \sqrt{n-3} \left(Z_r - \frac{1}{2} \log \left(\frac{1+\rho_o}{1-\rho_o} \right) \right)$
 - P -value from standard normal distn

Multiple regression

Introduction

- Multiple regression is in Chapters 5-12 of text, begin with Chapters 5-7
 - Chapter 5 is matrix theory
 - Chapters 6-7 are basics of multiple regression
- Basic idea: regression models containing more than one explanatory variable
- Data:
 - Each experimental unit, each run of expt, each individual provides response Y_i and vector of k explanatory variables $x_{i1}, x_{i2}, \dots, x_{ik}$
 - $i = 1, \dots, n$ indexes cases
 - when we wish to refer to individual i 's vector of explanatory variables we will use \mathbf{x}_i
 - to refer to the value of a single explanatory variable for all n cases we will use \mathbf{X}_i
 - more on matrix notation later

Multiple regression

The model

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- Model
 - ϵ_i are iid $N(0, \sigma^2)$
 - k predictors implies $k + 1$ coefficients (incl β_o)
 - text takes $k = p - 1$ with a total of p coefficients
 - a key point: β_j is the change in the mean of Y for a unit change in X_j **with all other variables held constant**
- Assumptions - same as in simple case
 - x 's are fixed (or conditioned upon)
 - correct model specification
 - independent (uncorrelated) errors
 - constant variance errors
 - normally distributed errors

Multiple regression

Linear and nonlinear models

- Linear models are linear in the parameters

- Examples of linear models:

$$- Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$- Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$$

$$- Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 \\ + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$

- Examples of nonlinear models:

$$- Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2}^{\beta_3} + \epsilon_i$$

$$- Y_i = \alpha_o z_{i1}^{\alpha_1} z_{i2}^{\alpha_2} \eta_i$$

- Some nonlinear models can be made linear by transformation, e.g., in 2nd nonlinear example

$$\log Y_i = \log \alpha_o + \alpha_1 \log z_{i1} + \alpha_2 \log z_{i2} + \log \eta_i$$

$$Y_i^* = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

(such models are known as intrinsically linear)

- Other nonlinear models can be addressed using Chap 13 of text
- A subtle issue is error specification (additive?)

Multiple regression

A quick matrix review

- Matrix M : a rectangular array of numbers m_{ij}
- If r rows and c cols then we say an $r \times c$ matrix
- Includes vectors as a special case (r or c equal one)
- We will take boldface for vectors and capital letters for matrices
- Some terminology/definitions
 - matrix is square if # of rows equals # of columns
 - matrix is symmetric if $m_{ij} = m_{ji}$
 - transpose $N = M^T$ is the $c \times r$ matrix with $n_{ij} = m_{ji}$
 - identity matrix I is square with ones on the diagonal and zeros elsewhere
 - rank of M is the number of linearly independent (nonredundant) columns (rows)
 - trace of square matrix M is denoted $tr(M)$; it is the sum of the diagonal elements
 - determinant of square matrix M is denoted $|M|$; it is a measure of "size" or "volume"

Multiple regression

A quick matrix review

- Matrix arithmetic
 - addition/subtraction
 - * need same size matrices
 - * perform elementwise
 - multiply by a constant - perform elementwise
 - multiplication
 - * can calculate $C = AB$ only if
 $p = (\# \text{ cols in } A) = (\# \text{ rows in } B)$
 - * $c_{ij} = \sum_{k=1}^p a_{ik}b_{kj}$
 - * note $AB \neq BA$
 - inverse - If $AB = I$ then $B = A^{-1}$
(full rank square matrices have inverses)

Multiple regression

Multivariate random variables review

- Suppose \mathbf{x} is a $n \times 1$ random vector
(or random variable)
- $E\mathbf{x} = \boldsymbol{\mu} = (EX_1, EX_2, \dots, EX_k)^T$
- $\text{Var } \mathbf{x} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \Sigma$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \vdots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix}$$

- $\sigma_{jj} = \text{Var}(x_j)$
- $\sigma_{ij} = \text{Cov}(x_i, x_j)$

Multiple regression

Multiple linear regression in matrix notation

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$
- Assume $\text{rank}(X) = k + 1$ (all columns are linearly independent)
- We can write $\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n)$

Multiple regression

Least squares estimation

- As in simple regression find \mathbf{b} (sometimes known as $\hat{\boldsymbol{\beta}}$) that minimizes

$$q(\mathbf{b}) = \sum_{i=1}^n (Y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2 = (\mathbf{Y} - X\mathbf{b})^T (\mathbf{Y} - X\mathbf{b}) = \mathbf{e}^T \mathbf{e}$$

where \mathbf{e} is the vector of residuals

- This is minimized by solving the set of equations $(X^T X)\mathbf{b} = X^T \mathbf{Y}$ (note that the simple regression normal equations are of this form)
- Solution: assuming X is of full rank $(k + 1)$

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$$

- Note: if X is not of full rank, then one or more columns is redundant and no unique solution
- Note: this is also the MLE for $\boldsymbol{\beta}$
- Unbiased estimate of σ^2 is

$$s_e^2 = \frac{(\mathbf{Y} - X\mathbf{b})^T (\mathbf{Y} - X\mathbf{b})}{n - (k + 1)}$$

Multiple regression

Towards inference

- The “hat matrix”
 - $\hat{Y}_i = \mathbf{x}_i^T \mathbf{b} = b_o + b_1 x_{i1} + \cdots + b_k x_{ik}$ is the fitted value or predicted value
 - $e_i = Y_i - \hat{Y}_i$ is the residual
 - We can write $\hat{\mathbf{Y}} = X\mathbf{b} = X(X^T X)^{-1} X^T \mathbf{Y} = H\mathbf{Y}$ where $H = X(X^T X)^{-1} X^T$ is the “hat” matrix or the projection matrix
 - Then $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (I - H)\mathbf{Y}$
- \mathbf{b} as a point estimate
 - $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$ is best linear unbiased estimate for $\boldsymbol{\beta}$
 - $E(\mathbf{b}) = \boldsymbol{\beta}$ (unbiased)
 - $\text{Var}(\mathbf{b}) = (X^T X)^{-1} X^T \text{Var}(\mathbf{Y}) X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$
 - replace σ^2 by s_e^2 to estimate $\text{Var}(\mathbf{b})$
 - these results do not require normal distn (normality required for inference procedures)

Multiple regression

Inference: ANOVA table

- As in simple linear regression we find

$$\begin{aligned}
 SS(\text{total, corrected}) &= \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
 &= \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \\
 &= SS(\text{residuals}) + SS(\text{model})
 \end{aligned}$$

- Collect sums of squares results in ANOVA table

source of variation	degrees of freedom	sums of squares	mean square
model	k	SS_{model}	SS/df
error	$n - (k + 1)$	SS_{error}	SS/df
total	$n - 1$	SS_{total}	

- Note that $MS_{error} = MSE$ is usual variance estimate s_e^2
- $R^2 = SS_{model}/SS_{total}$ is the amount of variation in the response that is explained by the model
- $R = \sqrt{R^2}$ is known as the multiple correlation coefficient

Multiple regression

Inference: F -test

- $E(MSE) = \sigma^2$
 $E(MS_{model}) = \sigma^2 + \text{term involving } \beta \text{ and } X$
- $F = MS_{model}/MSE$ will have $F_{k, n-k-1}$ distn if $\beta_1 = \dots = \beta_k$ and tend to be large otherwise
- This leads to F -test for
 $H_o : \beta_1 = \dots = \beta_k = 0$ vs $H_a : \text{not } H_o$
- Note F -test really compares two models:
 - reduced model: $Y_i = \beta_o + \epsilon_i$
 - full model: $Y_i = \beta_o + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$

Multiple regression

Inference for a single coefficient

- Recall that we have point estimates $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$ with $\text{Var}(\mathbf{b}) = \sigma^2 (X^T X)^{-1}$ and $\widehat{\text{Var}}(\mathbf{b}) = \text{MSE}(X^T X)^{-1}$
- Inference for a single coefficient
 - b_j = estimate of β_j
 - s_{b_j} = std error of $b_j = \sqrt{(j+1)^{\text{st}} \text{ diagonal of } \widehat{\text{Var}}(\mathbf{b})}$
 - $(b_j - \beta_j)/s_{b_j} \sim t_{n-k-1}$
 - $b_j \pm t_{n-k-1, 1-\alpha/2} s_{b_j}$ is 100(1 - α)% CI for β_j
 - $t = b_j/s_{b_j}$ is compared to t_{n-k-1} to test $H_0 : \beta_j = 0$
 - Note this test compares two models:
reduced model is $Y_i = \beta_0 + \dots + \beta_{j-1} x_{i,j-1} + \beta_{j+1} x_{i,j+1} + \dots + \beta_k x_{ik} + \epsilon_i$
full model is $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$

Multiple regression

Inference for other quantities

- Let's suppose there is a vector \mathbf{x} of interest
- Inference for $E(Y|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$
 - estimate is $\hat{Y} = \mathbf{x}^T \mathbf{b}$
 - std error is $s_{\hat{y}} = \sqrt{MSE \mathbf{x}^T (X^T X)^{-1} \mathbf{x}}$
 - term inside square root is a measure of distance from \mathbf{x} to center of data
 - CI is $\hat{Y} \pm t_{n-k-1, 1-\alpha/2} s_{\hat{y}}$
 - CIs for multiple points (or the whole regression line) is $\hat{Y} \pm \sqrt{(k+1) F_{k+1, n-k-1, 1-\alpha}} s_{\hat{y}}$ (Scheffe)
- Inference/prediction for new response Y_{new} at given \mathbf{x}
 - estimate is $\hat{Y} = \mathbf{x}^T \mathbf{b}$ (estimate of ϵ_{new} is zero)
 - std error is $s_p = \sqrt{MSE + s_{\hat{y}}^2}$
 - extra variance for unobserved error
 - CI is $\hat{Y} \pm t_{n-k-1, 1-\alpha/2} s_p$

Multiple regression

Geometry of regression - projection matrices

- $H = X(X^T X)^{-1} X^T$ is known as a projection matrix
- For today we will call it P
- Interpretation: $\hat{\mathbf{Y}} = P\mathbf{Y}$ is the orthogonal projection of \mathbf{Y} onto the linear space spanned by the columns of X
- Defining properties of projection matrices
 - P is symmetric ($P^T = P$)
 - P is idempotent ($P^2 = P$)
- More properties
 - eigenvalues are 0 or 1 with $\text{rank}(X) = \text{rank}(P) = k + 1 = \text{number of nonzero eigenvalues}$
 - $I - P$ is also a projection matrix
 - $P(I - P) = 0$ (two projections are orthogonal)
 - if $\mathbf{w} = X\mathbf{c}$ (\mathbf{w} is a linear combination of the columns of X), then $P\mathbf{w} = \mathbf{w}$ and $(I - P)\mathbf{w} = \mathbf{0}$

Multiple regression

Geometry of regression

- \mathbf{Y} is a point in n -dimensional space
- $\hat{\mathbf{Y}} = P\mathbf{Y}$ is a point in $(k + 1)$ -dimensional subspace spanned by columns of X (it is the nearest point in that space to \mathbf{Y})
- $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (I - P)\mathbf{Y}$ is point in the $n - k - 1$ dimensional space orthogonal to the column space of X
- This means that \mathbf{e} is perpendicular to all vectors that are linear combinations of the columns of X
- Note that $\mathbf{Y}^T \mathbf{Y} = (\hat{\mathbf{Y}} + \mathbf{e})^T (\hat{\mathbf{Y}} + \mathbf{e}) = \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} + \mathbf{e}^T \mathbf{e}$
 - cross term is $2\hat{\mathbf{Y}}^T \mathbf{e} = 2\mathbf{Y}^T P(I - P)\mathbf{Y} = 0$
 - length of $\mathbf{Y} =$ length of $\hat{\mathbf{Y}}$ + length of \mathbf{e}
(a form of Pythagoras' thm)

Multiple regression

Geometry of regression - submodels

- Suppose we consider a subset of the columns of X , say X_s (assume X_s includes the column of 1's that give us the intercept)
- New projection matrix $P_s = X_s(X_s^T X_s)^{-1} X_s^T$
- This defines a $\text{rank}(X_s)$ -dimensional subspace of the original predictor space
- We can perhaps compare the “fit” of \mathbf{Y} to these two subspaces by looking at the length of the residual vectors
- Example
 - take $X_s = \mathbf{1}$
 - P_s is matrix containing $1/n$ in each location (call this P_1)
 - $\mathbf{Y}^T P_1 \mathbf{Y}$ is $n\bar{Y}^2$
 - $\mathbf{Y}^T (I - P_1) \mathbf{Y}$ is $SS_{total,corrected}$

Multiple regression

Quadratic forms

- Some of the quantities we have looked at can be written as matrix products
 - $SS_{total} = \mathbf{Y}^T \mathbf{I} \mathbf{Y}$
 - $SS_{total,corrected} = \mathbf{Y}^T (\mathbf{I} - P_1) \mathbf{Y}$
 - $SS_{residuals} = \mathbf{Y}^T (\mathbf{I} - P) \mathbf{Y}$
 - $SS_{model} = \mathbf{Y}^T (P - P_1) \mathbf{Y}$
- These are known as quadratic forms
- In general $Q_A = \mathbf{Y}^T A \mathbf{Y} = \sum_i \sum_j a_{ij} Y_i Y_j$ is a quadratic form
- Often (including in our examples) A is symmetric

Multiple regression

Quadratic forms and chi-squares

- If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$, then we can find the mean and variance of the quadratic form $Q_A = \mathbf{Y}^T A \mathbf{Y}$
 - $E(Q_A) = \boldsymbol{\mu}^T A \boldsymbol{\mu} + \text{trace}(A \Sigma)$
(also true w/out normality)
 - $\text{Var}(Q_A) = 4\boldsymbol{\mu}^T A \Sigma A \boldsymbol{\mu} + 2\text{trace}(A \Sigma A \Sigma)$
- If $A \Sigma A \Sigma = A \Sigma$ then Q_A has the non-central χ^2 distn with non-centrality parameter $0.5\boldsymbol{\mu}^T A \boldsymbol{\mu}$ and d.f. equal to $\text{rank}(A)$
- Two quadratic forms Q_A and Q_B are independent if $A \Sigma B = 0$
- Can use above results to create F -tests
- Note often $\Sigma = I$ which makes things easier

Multiple regression

More general hypothesis tests

- So far:
 - test $H_o : \beta_1 = \dots = \beta_k = 0$ with ANOVA F -test
 - test $H_o : \beta_j = 0$ with univariate t -test
- There are other possible hypotheses, e.g.,
 - $H_o : \beta_1 + \beta_2 + \beta_3 = 1$
 - $H_o : \beta_{k-1} = \beta_k = 0$
- General approach is to compare the full model to the reduced model implied by the null hypothesis
- Let the number of “restrictions” in the reduced model be r
- This is essentially the difference in the number of parameters (and can be found as the number of equal signs in H_o)

Multiple regression

More general hypothesis tests (cont'd)

- Consider two ANOVA tables

Reduced model

source	d.f.	SS
model	$k - r$	$SS_{model,red}$
error	$n - k + r - 1$	$SS_{error,red}$
total	$n - 1$	SS_{total}

Full model

source	d.f.	SS
model	k	$SS_{model,ful}$
error	$n - k - 1$	$SS_{error,ful}$
total	$n - 1$	SS_{total}

Multiple regression

More general hypothesis tests (cont'd)

- These can be combined in a single table

source	d.f.	SS
reduced model	$k - r$	$SS_{reducedmodel}$
extra for full model	r	$SS_{full reduced}$
error	$n - (k + 1)$	SS_{error}
total	$n - 1$	SS_{total}

- The middle row represents the extra sum of squares explained when expanding from the reduced model to the full model
- Equivalently it represents the reduction in the sum of squared errors when going from reduced model to the full model
- Test significance of this extra sum of squares using F-test (details follow)

Multiple regression

More general hypothesis tests (cont'd)

- Let model A denote reduced model (H_o is true)
- Let model B denote the full model (H_o is false)
- We use $SS_{error}(A)$ to denote the SS_{error} for model A, etc.
- Fit both models
- Test reduced model using

$$F = \frac{(SS_{error}(A) - SS_{error}(B)) / (df_{error}(A) - df_{error}(B))}{SS_{error}(B) / df_{error}(B)}$$

which has $F_{df_{error}(A)-df_{error}(B), df_{error}(B)}$ if H_o is true

- This test includes F -test and t -test as special cases

Multiple regression

More general hypothesis tests (cont'd)

- Can write arbitrary sets of linear restrictions as
 $H_o : C\boldsymbol{\beta} = \mathbf{m}$ (C an $r \times (k + 1)$ matrix of rank r)
- To test $H_o : C\boldsymbol{\beta} = \mathbf{m}$ compare

$$F = \frac{(C\mathbf{b} - \mathbf{m})^T [C(X^T X)^{-1}C^T]^{-1} (C\mathbf{b} - \mathbf{m})}{rMS_{error}(\text{Full model})}$$

with $F_{r,df_{error.full}}$ distn

- One other possibility in this notation
 - solve for r of the β_j 's in terms of the others and fit the reduced model directly
 - e.g., if $Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ and $C\boldsymbol{\beta} = \mathbf{m}$ is $\beta_1 + \beta_2 = 0$, then $Y_i = \beta_o + \beta_1(x_{i1} - x_{i2}) + \epsilon_i$ is the reduced model

Multiple regression

Interpreting regression coefficients

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- β_j is the j th regression coefficient or the j th partial regression coefficient
- β_j is the change in the mean of Y for a unit change in X_j **with all other variables held constant**
- Often one can't change X_j without changing other predictors (e.g., polynomial terms (X_j, X_j^2) or highly correlated predictors)
- A slightly different interpretation: β_j is the linear effect of X_j on Y after adjusting for the linear effect of the other predictors on Y and the linear effects of the other predictors on X_j
- If let P_{-x_j} represent the projection matrix without variable X_j , then $\hat{\beta}_j$ is found from the simple regression of $(I - P_{-x_j})Y$ on $(I - P_{-x_j})X_j$

Multiple regression

Direct and indirect effects

- A complicated but interesting expression

$$r_{x_j,y} = b_j \frac{SD(X_j)}{SD(Y)} + \sum_m r_{x_j,x_m} b_m \frac{SD(X_m)}{SD(Y)}$$

- left-hand-side (LHS) is simple correlation of X_j and Y
- first term on RHS can be considered “direct effect of X_j on Y ”
- second term on RHS is the sum of “indirect effects”
- related to the idea of path diagrams
- if all X_j 's are uncorrelated then second term on right is zero
- this expression can explain why the regression coefficient of a variable doesn't match the sign of its simple correlation with the response

Multiple regression

Comparing coefficients

- How can we compare the relative contribution of different X_j 's
 - magnitude of b_j 's? NO because b_j is $\Delta Y/\Delta X_j$ so that changing the units of X_j will change b_j
 - magnitude of t-statistic b_j/s_{b_j} ? NO because this measures certainty that β_j is not zero but doesn't give the true effect
 - standardized coefficients
 - * $b_{std,j} = b_j \frac{SD(X_j)}{SD(Y)}$ (a unitless quantity)
 - * measures expected change in Y (in s.d. units) per 1 s.d. change in X_j
 - * can get standardized coefficients (B-weights?) from above or by regression using standardized variables

Multiple regression

Multicollinearity

- Multicollinearity (sometimes known as collinearity) is a problem that occurs sometimes in multiple regression
- It is not a violation of any assumptions
- Consider what happens if $X_2 = aX_1 + b$ (i.e., two predictors are linearly related)?
 - X is not of full rank
 - many solutions provide the same fit
- Multicollinearity is the situation where one or more predictor variables are “nearly” linearly related to the others
- Problem can be a pair of highly correlated variables or a large group of moderately correlated variables
- We treat this like a model assumption violation: effects, diagnostics, remedies

Multiple regression

Multicollinearity (cont'd)

- Effects of multicollinearity
 - fitted values are probably OK
 - b_j 's have high std errs
 - difficult to interpret the b_j 's
 - great sensitivity to minor changes in model/data (e.g., if we remove a variable or case)
- Diagnosis of collinearity
 - examine pairwise correlations
 - look for b_j 's with unusual signs
 - notice great sensitivity
 - $VIF_j = (1 - R_j^2)^{-1}$ (where R_j^2 is the R^2 for a regression of X_j on the other predictors)
 - VIF measures increase in standard error of b_j due to the presence of other variables
 - $VIF > 10$ indicates a problem ($VIF > 100$ is a big problem)

Multiple regression

Multicollinearity (cont'd)

- Remedies for collinearity
 - use the model only for prediction (huh???)
 - drop variables that are highly correlated (need to be careful)
 - create composite (combined) variables (e.g., using principal components)
 - find new cases that “break” the observed correlation (i.e., have a different pattern)
 - ridge regression $\mathbf{b}(c) = (X^T X + cI)^{-1} X^T \mathbf{Y}$ (biased estimates with smaller standard errors; corresponds to a Bayesian procedure)

Multiple regression

Case diagnostics - outliers

- Individual cases can be outliers in Y , outliers in \mathbf{x} , or regression outliers (outliers in Y for a given \mathbf{x})
- Y outliers – detected as usual in a single sample, not usually a problem
- Regression outliers – outliers in Y for a given \mathbf{x}
 - residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (I - H)\mathbf{Y}$
 - * $\text{Var}(\mathbf{e}) = \sigma^2(I - H)$
 - * $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$ and $\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2$
 - internally studentized resid $r_i = e_i / \sqrt{MSE(1 - h_{ii})}$
 - * have mean zero and approx equal variance
 - * but outlier will inflate MSE
 - externally studentized resid $t_i = e_i / \sqrt{MSE_{(i)}(1 - h_{ii})}$
 - * $MSE_{(i)}$ is MSE without the i th case
 - * alternative form $t_i = e_i \left[\frac{n - (k + 1) - 1}{SSE(1 - h_{ii}) - e_i^2} \right]^{0.5}$
 - * test for outlier by comparing t_i to t_{n-k-2} distn
 - * need to use multiple comparisons procedure

Multiple regression

Case diagnostics - high leverage cases

- Outliers in \mathbf{x} are called high leverage cases because they exert a large “pull” on the fitted regression
- Can write $\hat{Y}_i = \sum_{j=1}^n h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$
- h_{ii} measures the extent to which i th response dictates its own fitted value
- h_{ii} is called the leverage
 - $0 \leq h_{ii} \leq 1$
 - $\sum_{i=1}^n h_{ii} = k + 1$
 - h_{ii} measures distance between the i th case and the center (mean) of the cases
- Often use $2(k + 1)/n$ or $3(k + 1)/n$ as a guide for determining large h_{ii}
- In addition to an absolute cutoff, I look for high h_{ii} by examining the distn of h_{ii} values across cases
- Leverage is a measure of “potential” influence

Multiple regression

Case diagnostics - influential cases

- Sometimes a case doesn't look too unusual but has a major influence (effect) on the regression fit
- Easiest way to check is to delete the case, reanalyze data and examine the change
- Fortunately this can be done analytically
- We use subscript (i) to indicate quantities calculated without case i
- DFFITS - effect of i th case on fitted value for Y_i

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{0.5}$$

- $DFFITS > 1$ is considered large in small or medium sized samples
- $DFFITS > 2\sqrt{\frac{k+1}{n}}$ is considered large in big samples

Multiple regression

Case diagnostics - influential cases (cont'd)

- COOKSD - effect on all fitted values

$$COOKSD_i = \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(k+1)MSE} = \frac{e_i^2 h_{ii}}{(k+1)MSE(1-h_{ii})^2}$$

- equiv to measuring effect on estimates of β
 - $D_i < F_{k+1, n-k-1, 0.2}$ (i.e., 20th%ile)
is no concern
 - $D_i > F_{k+1, n-k-1, 0.5}$ is substantial influence
 - can also judge D_i relative to other D_j 's
- DFBETAS - effect on a single estimated coef.

$$DFBETAS_{k,i} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

- c_{kk} is relevant element from $(X^T X)^{-1}$
- DFBETA bigger than 1 (small or med samples)
or $2n^{-1/2}$ (large samples) is a problem

Multiple regression

Model checking

- We briefly discussed model checking in simple linear regression case
- Now revisit in more detail
- Recall the key assumptions
 - x fixed (or conditioned upon)
 - correct mean/model specification
 - independent (uncorrelated) errors
 - constant variance errors
 - normal distn for errors
- Diagnostics often based on residuals e_i or r_i

Multiple regression

Model checking: mean specification

- Effects of incorrect specification
 - if model is incorrect than \mathbf{b} and $\hat{\mathbf{Y}}$ are biased
 - inferences are not valid
- Diagnosing incorrect specification
 - residual plots - expect random scatter if model is correct, patterns indicate potential problems
 - * plot residuals (e_i or r_i) vs fitted values (\hat{Y}_i)
 - * plot residuals vs predictors
 - * plot residuals vs hypothetical new predictors
 - partial regression leverage plots
 - * offer slight improvement on r_i vs x_{ij}
 - * \mathbf{e}_j = resids regressing Y on X 's except X_j
 - * \mathbf{u}_j = resids regressing X_j on X 's except X_j
 - plot \mathbf{e}_j vs \mathbf{u}_j
 - * shows relationship of Y and X_j after adjusting for other variables
 - add quadratic terms (e.g., X_3^2 or X_3X_2) and see if they help
 - lack-of-fit test described in simple regression

Multiple regression

Model checking: mean specification

- Corrections/remedies for misspecification
 - Nonlinear regression models (chapters 13, 14) can be used if an appropriate nonlinear model is identified
 - Add polynomial terms to accommodate unidentifiable non-linear relationships
 - Transformations of Y and/or X
 - * transformations of Y :
trial and error (using R^2 to compare),
Box-Cox algorithm to be discussed later
 - * transformations of X : trial and error,
polynomial terms
 - * transformations of X : Box-Tidwell algorithm
 - idea is to use x^α in place of x
 - note that $\beta x^\alpha \approx \beta x + \beta(\alpha - 1)x \ln x$
 - use x and $x \ln x$ as predictors
 - if coefficient of $x \ln x$, say γ , is significant then
 $\hat{\alpha} = (\hat{\gamma}/\hat{\beta}) + 1$

Multiple regression

Model checking: mean specification

- Corrections/remedies for misspecification (cont'd)
 - Non parametric methods
 - * Loess - locally weighted regression
scatterplot smoothing
 - to get fit at \mathbf{x} use following steps
 - identify all pts within a given distance of \mathbf{x}
 - weight points according to distance from \mathbf{x}
 - use weighted LS (more later) to fit linear (or quad.) model for Y using points near \mathbf{x}
 - evaluate this fitted regression at \mathbf{x}
 - repeat for other points
 - widely used in 1-dimension (Splus), a bit harder in ≥ 2 -dimensions
 - * Generalized additive models
(Tibshirani and Hastie book)
 - $Y = g_1(X_1) + g_2(X_2) + \cdots + g_k(X_k) + e$
 - iteratively fit (perhaps using 1-dim loess)
 - * Regression trees (CART)

Multiple regression

Model checking: constant variance

- Effects of nonconstant variance
 - \mathbf{b} is still unbiased
 - $\text{Var}(\mathbf{b})$ is larger than necessary (i.e., \mathbf{b} is not the best estimate)
 - estimate of $\text{Var}(\mathbf{b})$ is biased (likely too small) so inferences are not valid
- Diagnosing nonconstant variance
 - residual plots (as before) - look for incr/decr spread as a function of \hat{Y} or a predictor
 - can compute s.d. of residuals in groups defined by \hat{Y} or a predictor
 - score test: hypothesize $\sigma_i^2 = \sigma^2 e^{\lambda^T \mathbf{z}_i}$
 - * regress ne_i/SSE on \mathbf{z}_i
 - * test for nonconstant variance by comparing $\frac{1}{2}SS_{model}$ to $\chi_{dim(\mathbf{z})}^2$ distn

Multiple regression

Model checking: constant variance

- Corrections/remedies for nonconstant variance
 - Transformation of Y
 - * trial and error
 - * rules of thumb (square root of counts, log of positive numbers with large range, logit of proportions)
 - * if we can write $\text{Var}(Y_i) \approx g(EY_i)$, then $h(Y) = \int^Y \frac{1}{\sqrt{g(v)}} dv$ has constant variance
 - * Box-Cox approach: $\mathbf{Y}^\lambda = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with λ treated as a parameter
 - $L(\lambda) = -\frac{n}{2} \ln SSE_\lambda + (\lambda - 1) \sum_i \ln Y_i + n \ln |\lambda|$
 - evaluate for many λ 's and choose best value
 - some related approaches use a different parameterization than Y^λ

Multiple regression

Model checking: constant variance

- Corrections/remedies for nonconstant variance
 - Generalized least squares (GLS)
 - * model is $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma)$
 - * if Σ is known, then $\hat{\boldsymbol{\beta}} = (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} \mathbf{Y})$
 - * usually Σ depends on parameters – can use maximum likelihood to estimate these parameters and $\boldsymbol{\beta}$
 - * the special case with Σ diagonal is very useful, known as weighted least squares
 - Weighted least squares (WLS)
 - * suppose that e_i in usual model is $N(0, \sigma_i^2 \kappa)$ (each case has diff't variance which we write as a multiplier of a common variance κ)
 - * define $w_i = 1/\sigma_i^2$
 - * write $\sqrt{w_i} Y_i = \sqrt{w_i} X \boldsymbol{\beta} + \sqrt{w_i} e_i$ and note that $\text{Var}(\sqrt{w_i} e_i) = \kappa$ (is constant!)

Multiple regression

Model checking: constant variance

- Corrections/remedies for nonconstant variance
 - Weighted least squares (WLS) (cont'd)
 - * fitting weighted model is equivalent to using *GLS* results with $\Sigma = W^{-1}$
 - $\mathbf{b}_{wls} = (X^T W X)^{-1} (X^T W \mathbf{Y})$
 - $\text{Var}(\mathbf{b}_{wls}) = \kappa (X^T W X)^{-1}$
 - estimate κ with $MSE_{wls} = \frac{\mathbf{e}_{wls}^T W \mathbf{e}_{wls}}{(n-k-1)}$
where $\mathbf{e}_{wls} = \mathbf{Y} - X \mathbf{b}_{wls}$
 - * problem – σ_i^2 's are usually unknown
 - there are some common situations (e.g., if obs i is an average of n_i values, then $w_i = n_i$)
 - estimate a variance function from the LS residuals, $\sigma_i^2 = f(\mathbf{x}_i \text{ or } \hat{Y}_i)$
 - iterative approach:
fit LS, estim variance fn from LS residuals,
fit WLS, estim variance fn from WLS resids,
etc. (how do we do inference???)

Multiple regression

Model checking: indep/uncorrelated errors

- Effects of correlated errors
 - \mathbf{b} is still unbiased
 - $\text{Var}(\mathbf{b})$ is larger than necessary (i.e., \mathbf{b} is not the best estimate)
 - estimate of $\text{Var}(\mathbf{b})$ is biased (likely too small) so inferences are not valid
- Diagnosing correlated errors
 - residual plots
 - * plot residuals vs time (or other factor inducing correlation)
 - * plot residuals vs lagged residuals (r_i vs r_{i-1}) if sequence is informative
 - * examine residuals in clusters (e.g., family, school) if such structure exists

Multiple regression

Model checking: indep/uncorrelated errors

- Diagnosing correlated errors (cont'd)
 - statistical tests (mainly detect time dependence)
 - * Durbin-Watson test (Table B.7)
 - $DW = \sum_i (e_i - e_{i-1})^2 / \sum_i e_i^2 = 2 - 2\text{Corr}(e_i, e_{i-1})$
 - reject if DW is far from 2 ($DW < D_l$); don't reject if $DW > D_u$; inconclusive otherwise
 - if negative correlation, then use 4 - DW in test
 - * runs test
 - count R = number of runs of positive and negative residuals
 - let N_+, N_- represent # of pos and neg resid
 - if $n \geq 20$ then can use normal approx
 - $E(R) = \frac{2N_+N_-}{N_++N_-} + 1$ and
 - $\text{Var}(R) = \frac{2N_+N_-(2N_+N_- - N_+ - N_-)}{(N_++N_-)^2(N_++N_- - 1)}$

Multiple regression

Model checking: indep/uncorrelated errors

- Corrections/remedies for correlated errors
 - add predictor to explain time or space correl
 - model clusters (if present)
 - * e.g., suppose $Y_i = \beta_o + \beta_1 x_i + \epsilon_i$
with obs correlated within clusters
 - * define $z_{ij} = 1$ if obs i is in cluster j (else 0)
 - * rewrite model as $Y_i = \beta_o + \beta_1 x_i + \sum_j \alpha_j z_{ij} + \epsilon_i$
 - * if many clusters with few observations let α_j 's be random effects (which leads to)
 - generalized least squares (GLS)
 - * write $\mathbf{Y} \sim N(X\boldsymbol{\beta}, \Sigma)$ where Σ describes correlation structure
 - * e.g., time series correlation, block correlation
 - * Σ generally includes unknown parameters
 - estimate by maximum likelihood
(or use Bayesian methods)
 - alternative:
fit LS, estim params of Σ from resids
use GLS, re-estim parameters from GLS resids,
etc.

Multiple regression

Model checking: indep/uncorrelated errors

- Corrections/remedies for correlated errors (cont'd)

- time series models

- * consider a simple case:

- the first order autoregressive (AR) model

- $Y_t = \beta_o + \beta_1 x_t + \epsilon_t$ (could be multiv too)

- with $\epsilon_t = \rho\epsilon_{t-1} + u_t$ and u_t 's iid $N(0, \sigma^2)$

- this makes ϵ_t id $N(0, \frac{\sigma^2}{1-\rho^2})$ but not indep

- $\text{corr}(\epsilon_t, \epsilon_{t-j}) = \rho^j$

- * suppose we know ρ , then we can create model with iid errors

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \beta_o + \beta_1 x_t - \rho\beta_o - \rho\beta_1 x_{t-1} + \epsilon_t - \rho\epsilon_{t-1} \\ &= \beta_o(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + u_t \\ &= \alpha_o + \alpha_1(x_t - \rho x_{t-1}) + u_t \end{aligned}$$

- * but ρ is unknown

- Cochran-Orcutt procedure:

- fit LS, estimate ρ from resids,

- fit above model using LS

- check resids for correl (iterate as needed)

- Hildreth-Lu: estimate ρ using ML

- (as in Box-Cox)

- use $\rho = 1$... crude but often works

Multiple regression

Model checking: normally dist errors

- Effects of nonnormal errors
 - inference is still OK in large samples
(why? $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$ is an avg of Y_i 's)
 - can't believe inferences with small samples
 - can't believe prediction intervals
 - regression is sensitive to outliers
- Diagnosing nonnormal errors
 - normal probability plot of residuals
 - tests of normality on residuals
 - case diagnostics

Multiple regression

Model checking: normally dist errors

- Corrections/remedies for nonnormal errors
 - transformation (Box-Cox procedure)
 - squared-error is quite convenient and optimal for normal data but there are alternatives
 - * generalized linear models (Poisson regression, logistic regression) are discussed in Chap 14 of the text and in Stat 211 next quarter (hopefully!)
 - * robust regression, e.g.,
 - least absolute deviations (*LAD* or L_1) regr.
 - find \mathbf{b} to minimize $\sum_i |Y_i - \mathbf{x}_i^t \mathbf{b}|$
 - this can be solved by linear programming (but where are standard errors)
 - can solve with iteratively reweighted least-squares (IRLS):
 - rewrite as minimize $\sum_i \frac{(Y_i - \mathbf{x}_i^t \mathbf{b})^2}{|Y_i - \mathbf{x}_i^t \mathbf{b}|}$
 - use LS to get first estimate of \mathbf{b}
 - use $1/|Y_i - \mathbf{x}_i^t \mathbf{b}|$ as weights in WLS
 - repeat above steps until convergence (?)

Multiple regression

Model building: types of models

- “Linear” regression models are quite flexible
- Briefly review some situations that can be addressed in linear regression
 - polynomial regression models
 - categorical predictors (including ANOVA)
 - change points/piecewise linear models

Multiple regression

Model building: polynomial regression

- Can use polynomial regression as true model or to approximate a nonlinear relationship (square root and logarithm functions both look quadratic)
- Polynomials in one variable
 - model

$$Y_i = \beta_o + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

- collinearity is often a problem;
situation improves if we use $(x_i - \bar{x})$ in place of x_i or if we use orthogonal polynomials
- interpretation (assuming we use $(x_i - \bar{x})$):
 $\beta_o = E(Y|\bar{x})$, $\beta_1 =$ linear effect, etc.
- note β_1 is not effect of a one unit change in x
- big danger is extrapolation
(true for all regressions, especially true here)

Multiple regression

Model building: polynomial regr (cont'd)

- Polynomials in two variables

- model

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$

- the term $x_{i1} x_{i2}$ is known as an interaction, the change in Y expected for a one unit change in X_1 depends on the value of X_2
- interactions of contin. variables can be difficult to interpret (sensitive to scales of X_1, X_2)

- Generalizes to more variables

Multiple regression

Model building: categorical variables

- We have emphasized the use of regression with quantitative variables
- May also have categorical variables (as response or predictors)
- Binary response (we define outcomes as 1 and 0)
 - can use LS but assumptions are clearly violated
 - seems to work OK but wrong in principle
 - can use logistic regression (in Chap 14 and in Stat 211/212)
- Ordinal response (two or more categories)
 - can use LS after constructing quantitative response variable
 - often not a realistic model
 - multinomial regression (in Chap 14 and in Stat 211/212)
- Categorical (non-ordinal) response
 - multinomial regression

Multiple regression

Model building: categorical predictors

- Can integrate categorical predictors by constructing artificial variables (known also as dummy variables, indicator variables)
- Illustrate here with a binary predictor (e.g., M/F)
- Pick one category as the default (say M)
- Define $z_i = 1$ if obs i is in the other (say F) and $z_i = 0$ otherwise
- A simple model

$$Y_i = \beta_o + \beta_1 x_i + \beta_2 z_i + \epsilon_i$$

where Y_i, x_i are continuous variables

– note: $E(Y|x) = \beta_o + \beta_1 x$ (for M)

$$E(Y|x) = (\beta_o + \beta_2) + \beta_1 x \text{ (for F)}$$

– both groups' data are used to estimate β_1, σ^2

– β_o is Y -intercept for M

– β_2 is difference in expected value of Y for identical units (same x) in the two groups

Multiple regression

Model building: categorical predictors (cont'd)

- A slightly fancier model

$$Y_i = \beta_o + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i$$

– note: $E(Y|x) = \beta_o + \beta_1 x$ (for M)

$$E(Y|x) = (\beta_o + \beta_2) + (\beta_1 + \beta_3)x \text{ (for F)}$$

– x, z interaction yields different slopes for the two groups

– separate regression lines for the two groups

- What if there are more than g categories/groups?
 - pick one category as the default
 - define $g - 1$ indicator variables for other groups
 - note this is related to ANOVA (more on this soon)
- Categorical vs continuous predictors
 - we always have the option of chopping a continuous predictor into categories
 - advantage: accommodates non-linear relations
 - disadvantage: categorical predictors use many parameters

Multiple regression

Model building: fun with dummy variables

- If we are creative, can use dummy variables to build many different models
- Suppose we are relating Y and x and expect a change in slope at $x = 100$. A possible model is

$$Y_i = \beta_o + \beta_1 x_i + \beta_2 (x_i - 100) z_i + \epsilon_i$$

where $z_i = 1$ if $x_i > 100$ and 0 otherwise

- $E(Y|x) = \beta_o + \beta_1 x$ (for $x \leq 100$)
 $E(Y|x) = \beta_o + \beta_1 x + \beta_2 (x - 100)$ (for $x > 100$)
slope changes from β_1 to $\beta_1 + \beta_2$ at $x = 100$
- can allow for jumps at $x = 100$ also (use $\beta_3 z$)
- if change point is unknown then can replace 100 by parameter τ but it is no longer a linear model

Multiple regression

Model building: model selection

- Regression can be applied in many contexts
 - controlled randomized experiment (continuous treatment x and continuous response y)
 - controlled randomized experiment with supplemental variables (other predictors)
 - confirmatory observational study - following up on a previous study
 - exploratory observational study - trying to find valuable predictors
- Especially in last context there is wide latitude for picking explanatory variables
- This is often called the model selection or variable reduction problem
- Need to remember that this is exploratory/hypothesis generating

Multiple regression

Model building: model selection

- Model selection or variable reduction
 - too few variables leads to biased models
 - too many variables leads to excess variability
- Different methods for searching among models
 - consider all possible subsets of a given group of predictors
 - stepwise model searching (modify models one step at a time)
- Different criteria are possible, broadly speaking we may want to
 - estimate $E(Y|x)$
 - predict out of sample
 - or do both of the above

Multiple regression

Model building: model selection theory

- Consider two models
 - model B (“true”) $\mathbf{Y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
 - model A (“fit”) $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
 - we have omitted the variables in Z
- Fit model A: $\hat{\mathbf{Y}} = P_x \mathbf{Y} = X(X^T X)^{-1} X^T \mathbf{Y}$
 - results for future reference:
 - * $E(\mathbf{b}) = \boldsymbol{\beta} + (X^T X)^{-1} X^T Z\boldsymbol{\gamma}$
 - * $E(MS_{error}) = \sigma^2 + \boldsymbol{\gamma}^T Z^T (I - P_x) Z \boldsymbol{\gamma} / (n - k - 1)$
(see next slide for details)
 - bias in fit $\hat{\mathbf{Y}}$
 $\boldsymbol{\delta} = E(\hat{\mathbf{Y}}) - E(\mathbf{Y}) = (P_x - I)(X\boldsymbol{\beta} + Z\boldsymbol{\gamma}) = (P_x - I)Z\boldsymbol{\gamma}$
 - zero bias if $\boldsymbol{\gamma} = \mathbf{0}$ or Z is in col space of X
 - variance in fit $\hat{\mathbf{Y}}$
 $\sum_i \text{Var}(\hat{Y}_i) = \text{trace}(\text{Var}(\hat{\mathbf{Y}})) = \text{trace}(\text{Var}(P_x \mathbf{Y})) = \text{trace}(\sigma^2 P_x P_x^T) = \sigma^2 \text{trace}(P_x) = \sigma^2(k + 1)$

Multiple regression

Model selection: theory (cont'd)

- Note that adding a predictor to A
(taking a column from Z and adding it to X)
 - decreases bias (or may leave it the same)
 - increases variance
- In the limit if we fit the “true” model
 - bias = 0
 - variance = $\sigma^2(k + 1 + \dim(Z))$
- Here are details from calculation of $E(MS_{error})$ when we fit model A

$$\begin{aligned}E(MS_{error}) &= \frac{1}{n - k - 1} E[\mathbf{Y}^T (I - P_x) \mathbf{Y}] \\&= \frac{1}{n - k - 1} E[(X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon})^T (I - P_x) (X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon})] \\&= \frac{1}{n - k - 1} [(X\boldsymbol{\beta} + Z\boldsymbol{\gamma})^T (I - P_x) (X\boldsymbol{\beta} + Z\boldsymbol{\gamma}) \\&\quad + E(\boldsymbol{\epsilon}^T (I - P_x) \boldsymbol{\epsilon})] \\&= \frac{1}{n - k - 1} [\boldsymbol{\gamma}^T Z^T (I - P_x) Z \boldsymbol{\gamma} + \sigma^2(n - k - 1)] \\&= \sigma^2 + \frac{1}{n - k - 1} \sum_i \delta_i^2\end{aligned}$$

Multiple regression

Model selection: theory (cont'd)

- How many predictors?
- One criterion is to minimize $E[\sum_i (\hat{Y}_i - E(Y_i))^2]$, the mean squared error of the predictions (*MSEP*)

$$\begin{aligned}MSEP &= E\left[\sum_i (\hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - E(Y_i))^2\right] \\&= E\left[\sum_i (\hat{Y}_i - E(\hat{Y}_i))^2\right] + E\left[\sum_i (E(\hat{Y}_i) - E(Y_i))^2\right] \\&= \sum_i \text{Var}(\hat{Y}_i) + \sum_i \text{Bias}(\hat{Y}_i)^2 \\&= \sigma^2(k+1) + \sum_{i=1}^n \delta_i^2 \\&= \sigma^2(k+1) + E(SS_{error}) - \sigma^2(n-k-1) \\&= E(SS_{error}) - \sigma^2(n-2(k+1))\end{aligned}$$

(second to last line uses the previously obtained relationship of bias and $E(MS_{error})$)

- This leads to Mallows's C_p statistic, a popular model selection criterion

Multiple regression

Model selection: Mallows's C_p

- Mallows's C_p statistic estimates $MSEP/\sigma^2$
- The p in C_p is the number of parameters, our $k + 1$. We switch to this notation for discussing model selection criteria
- Let m denote the size of biggest possible model ($m - 1$ predictors)
- Definition

$$C_p = \frac{SS_{error}}{\hat{\sigma}^2} - (n - 2p)$$

where SS_{error} is for the fitted model and $\hat{\sigma}^2$ is the MSE for the “full” (biggest possible model)

- want C_p to be small
- but also want $C_p \approx p$
(why? with no bias $MSEP/\sigma^2 = p$)
- note $C_m = m$ (for biggest model)
- C_p is related to the F -test that the submodel p is acceptable

$$C_p = (m - p)(F - 1) + p$$

- * $F < 2$ then $C_p < m$ and submodel accepted
- * $F < 1$ then $C_p < p$ (negative bias???)

Multiple regression

Model selection: other criterion

- We stay with modified notation: p is number of parameters (our $k + 1$)
- Criteria:
 - R^2
 - * bigger is better
 - * problem - this never decreases when adding a variable
 - * useful for comparing two models of same size but not otherwise
 - Adjusted $R^2 = R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p}$
 - * also have $R_{adj}^2 = 1 - \frac{MS_{error}}{MS_{total}}$
 - * bigger is better
 - * equivalent to minimizing MS_{error}
 - * not crazy but not the best criterion

Multiple regression

Model selection: other criterion

- We stay with modified notation: p is number of parameters (our $k + 1$)
- Criteria (cont'd):
 - AIC (Akaike information criterion)
 - * $AIC = n \ln(SS_{error}/n) + 2p$ (small is better)
 - * based on asymptotic theory
 - * generally favors big models (not consistent)
 - BIC (Schwarz' Bayesian info crit)
 - * $BIC = n \ln(SS_{error}/n) + p \ln n$ (small is better)
 - * based on asymptotic theory
 - PRESS (predicted sum of squares)
 - * predict each observation using other $n - 1$
 - * $PRESS = \sum_i (Y_i - \hat{Y}_{i(i)})^2 = \sum_i \left(\frac{e_i}{1-h_{ii}} \right)^2$
 - * nice idea

Multiple regression

Model selection: search strategies

- All subsets
 - given a set of m predictors
 - fit all $2^m - 1$ possible models
(leaves out model with no predictors)
 - compare based on some criterion (like C_p)
 - works up to about $m = 15$ or 20
(i.e., takes a reasonable amount of time)
- Stepwise methods - build one step at a time
 - many algorithms, we describe a basic one on the next slide
 - note that a series of local steps is not guaranteed to find a global optimum

Multiple regression

Model selection: search strategies

- Stepwise methods (cont'd) - a basic algorithm
 1. begin with empty model
 2. fit all one variable models, select best predictor (F -test or t -test) as long as its P -value is less than α_{entry}
 3. try all unincluded variables with current model (one at a time)
 4. add most significant of these if $P < \alpha_{entry}$
 5. now examine all current variables (t -stat for testing $\beta_j = 0$)
 6. delete “least” significant if $P > \alpha_{remove}$
 7. repeat steps 3-6 until no changes
 - note: need $\alpha_{entry} \leq \alpha_{remove}$ to avoid loops
 - note: other algorithms include forward selection (no removals) and backward elimination (no additions)

Multiple regression

Model selection: discussion ??

- Is model selection the right thing to do?
(What about model averaging)
- If we do model selection
still need to examine case diagnostics
still need to examine model assumptions
- How can we possibly do inference given all of this searching?
 - adjust significance for all models analyzed?
 - ignore it (most common practice but VERY MISLEADING)
 - explore conclusions from several top models
- Have we overfit to this particular data set?
 - model validation

Multiple regression

Model validation

- Validation idea: split data into two parts
 - fitting sample (perhaps 2/3 of the data)
 - testing sample (the remainder of the data)
 - do what ever you want on fitting sample
 - test resulting regr. model on testing sample
 - * compute $MSEP = \frac{1}{n_{test}} \sum(\text{pred.err.})^2$
 - * should be approximately the same as s_e^2 from selected model
 - * will be very large if model is over fit to the training sample
- Is there enough data to do this?
- Can be used as a model selection technique (fit many models to fitting sample, compute MSEP on test sample for each, then choose best) but that is not the spirit in which it's offered here
- PRESS is doing this over-and-over with the size of the test sample equal to one case