statweb.stanford.edu/~owen/courses/305 + axess

- Linear Models

- Applied Statistics

Goals

- linear model in depth

- prepare for applied statistics

- bridge to research

Prerequisites
Matrix algebra: eigenvalue, rank, orthogonal matrices
Probability: normal, t, χ^2, F, CLT, covariance
Statistics: p-value, confidence interval, hypothesis testing, regression
Computing: R, python, matlab, C
Experience: fitting models, applying methods

Predictive Statistics: predict Y from X:

- a value

- a distribution

- an interval

		Y				
		$\mathbb{R}$	$[0,1]$	k groups	ordered groups	$\mathbb{R}^p$
	1 group					
	2 groups					
X	k groups					
	$\mathbb{R}$					
	$\mathbb{R}^p$					

Statistics is almost but not quite math;
Statistics is almost but not quite computing.

Modeling is tricky.

- hard to choose a model, easy to work with it.
- wrong assumptions can lead to right answers.
- cannot quite prove things about the world.

Linear Models: have X predict $Y \in \mathbb{R}$
X arbitrary data $(X_i, Y_i), i = 1, \cdots, n$
"Best" predictor of Y is
for $X = x, \mu(x) = \mathbb{E}(Y|X = x)$
$\mu(x)$ minimizes $\mathbb{E}((Y - m(X))^2|X = x)$.

proof

$$\begin{aligned}
\mathbb{E}((Y - m(X))^2|X = x) &= \mathbb{E}([Y - \mu(X) + \mu(X) - m(X)]^2|X = x) \\
&= \mathbb{E}((Y - \mu(X))^2|X = x) + 2\mathbb{E}([Y - \mu(X)][\mu(X) - m(X)]|X = x) \\
&\quad + \mathbb{E}((\mu(X) - m(X))^2|X = x) \\
&= \mathrm{Var}(Y|X = x) + 0 + [\mu(X) - m(X)]^2 \\
&\geqslant \mathrm{Var}(Y|X = x)
\end{aligned}$$

For loss $= \mathbb{E}(|Y - m(X)||X = x)$
Take $m(X) = \mathrm{median}(Y|X = x)$
This is called "quantile regression".

Alternative proof (sketch)
Set $\frac{\mathrm{d}}{\mathrm{dm}}\mathbb{E}((Y - m)^2|X = x) = 0$

$$\Rightarrow \mathbb{E}(\frac{\mathrm{d}}{\mathrm{dm}}(Y - m)^2|X = x) = 0$$

Linear Model Examples

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)(\text{maybe normal})$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$y = \text{fuel consumption}$$

$$x_1 = \text{temp}$$

$$\vdots$$

$$x_k = \text{wind speed}$$

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

polynomial, linear in β not x.

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$
$$\mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
$$\mathbb{E}(Y) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$$

2 groups

$$X_i = \begin{cases} 1 & i \in \text{ group 1} \\ 0 & i \in \text{ group 0} \end{cases}$$

e.g., Male versus Female, Ni versus Cu, Treatment versus Control

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x = \begin{cases} \beta_0 + \beta_1 & x = 1 \\ \beta_0 & x = 0 \end{cases}$$

$k \geqslant 2$ groups

$$X_1 = \begin{cases} 1 & \text{if group 2} \\ 0 & \text{else} \end{cases}$$
$$\vdots$$
$$X_{k-1} = \begin{cases} 1 & \text{if group k} \\ 0 & \text{else} \end{cases}$$

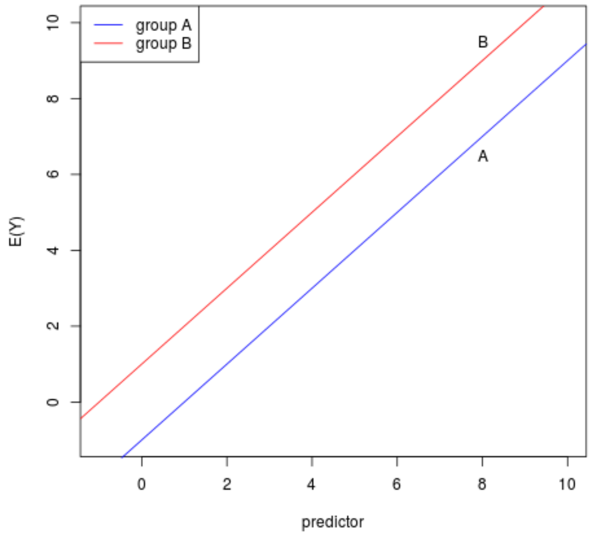
$$\begin{aligned} \mathbb{E}(Y) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} \\ &= \begin{cases} \beta_0 & \text{group 1} \\ \beta_0 + \beta_1 & \text{group 2} \\ \vdots & \\ \beta_0 + \beta_{k-1} & \text{group k} \end{cases} \end{aligned}$$

Choice of group 1 **matters**!
versus

$$\mathbb{E}(Y) = \beta_1 x_1 + \cdots + \beta_k x_k$$
$$x_j = \begin{cases} 1 & \text{group j} \\ 0 & \text{else} \end{cases}$$

no intercept
cell mean model

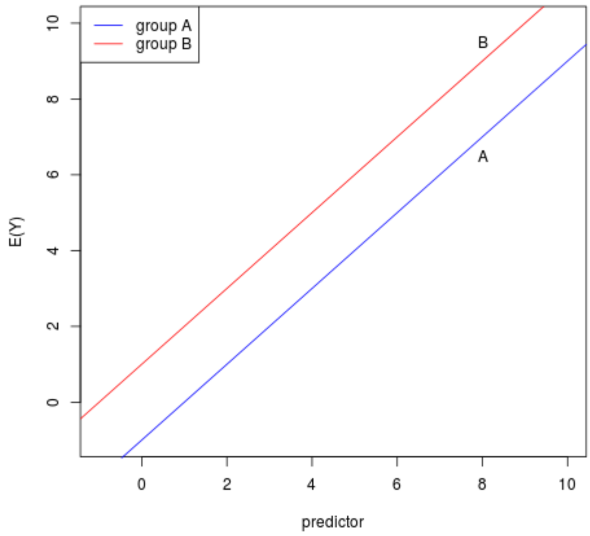
β_1	β_2	$\cdots$	β_k



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

$$z_i = \begin{cases} 1 & \text{group B} \\ 0 & \text{group A} \end{cases}$$

OR



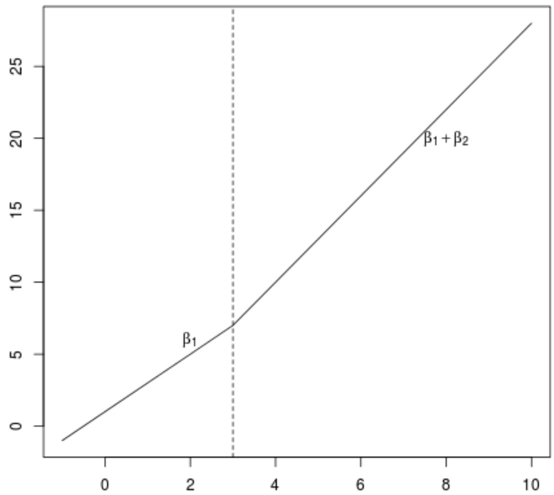
$$\beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i$$

z_i: dummy variable

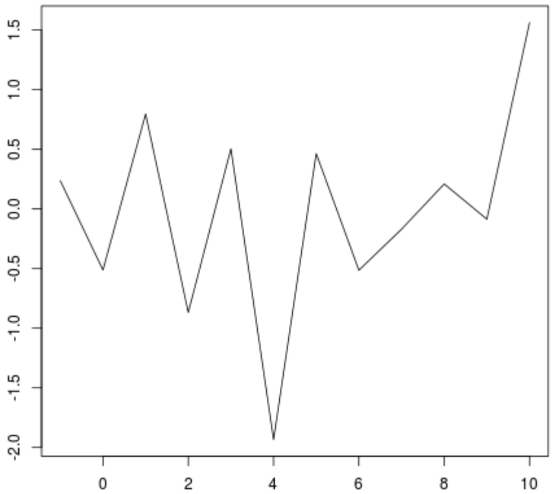
Simple Interaction

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i_2} + \beta_3 x_{i1} x_{i2}$$

Two phase regression



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - t)_+ + \varepsilon_i$$
$$Z_+ = \max(Z, 0)$$
$$= \begin{cases} Z & Z \geqslant 0 \\ 0 & Z < 0 \end{cases}$$



Add more pieces $\beta_l(x-t_l)_+$, add$(x-t_l)_+^2$ Fourier $0 \leqslant x \leqslant 1$

$$\beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \cdots$$

$$y_i = \beta_0 + \varepsilon_i$$

Haar wavelets

Notation

$$x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

$$y_i = \sum_{j=1}^{p} z_{ij}\beta_j + \varepsilon_i$$

$$z_{ij} = j \text{ th feature of } x_i = (x_{i1}, \cdots, x_{id})$$

$$Y = Z\beta + \varepsilon$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, Z = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ & \ddots & \\ z_{n1} & \cdots & z_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Tasks:

- estimate β

- test $\beta_j = 0$

- confidence intervals, p-values

- predict y at new $x_0 = (x_{01}, \cdots, x_{od})$

Distribution assumptions
2 Models:
Correlation Model:

$$(x_i, y_i) \text{ i.i.d. } i = 1, \cdots, n$$

Regression Model:
x_i fixed, $i = 1, \cdots, n$
y_i independent $\mathcal{L}(Y|X = x_i)$.

Moments

- random $X \in \mathbb{R}$

$$
\begin{aligned}
\mu &= \mathbb{E}(X) \\
\sigma^2 &= \mathrm{Var}(X) \\
&= \mathbb{E}((X-\mu)^2) \\
\gamma &= \mathbb{E}((X-\mu)^3)/\sigma^3 \text{ - skewness} \\
\kappa &= \mathbb{E}((X-\mu)^4)/\sigma^4 - 3 \text{ - kurtosis} \\
\gamma &= \kappa = 0 \text{ for } N(\mu, \sigma^2)
\end{aligned}
$$

- $X_1, \cdots, X_n$ i.i.d.

$$
\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\mathbb{E}(\overline{X}) &= \mu \\
\mathrm{Var}(\overline{X}) &= \frac{\sigma^2}{n} \\
\gamma(\overline{X}) &= \frac{\gamma}{\sqrt{n}} \\
\kappa(\overline{X}) &= \frac{\kappa}{n}
\end{aligned}
$$

Random vectors and matrices

- expectation componentwise

$$
\mathbb{E}\begin{pmatrix} X_{11} & \cdots & X_{1n} \\ & \vdots & \\ X_{m1} & \cdots & X_{mn} \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_{11}) & \cdots & \mathbb{E}(X_{1n}) \\ & \vdots & \\ \mathbb{E}(X_{m1}) & \cdots & \mathbb{E}(X_{mn}) \end{pmatrix}
$$

- X random, A, B fixed

$$
\mathbb{E}(AX) = A\mathbb{X}, \mathbb{E}(XB) = \mathbb{E}(X)B, \mathbb{E}(AXB) = A\mathbb{E}(X)B
$$

-

$$
\begin{aligned}
X \in \mathbb{R}^{n\times 1}&, Y \in \mathbb{R}^{m\times 1} \\
\mathrm{Cov}(X,Y) &= \mathbb{E}\left((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))^T\right) \in \mathbb{R}^{n\times m} \\
\mathrm{Cov}(Y,Y) &= \mathrm{Var}(Y) \in \mathbb{R}^{m\times m} \\
\mathrm{Cov}(AX, BY) &= A\mathrm{Cov(X, Y)}B^T \\
\mathrm{Var}(AX+b) &= A\mathrm{Var}(X)
\end{aligned}
$$

- $\mathrm{Var}(X)$ positive semi-definite (symmetric) given $C \in \mathbb{R}^{n\times 1}$

$$0 \leqslant \mathrm{Var}(C^T X) = C^T \mathrm{Var}(X) C$$

positive definite unless $\mathrm{Var}(C^T X) = 0$ for some $c \neq 0$.

Quadratic Forms
$A \in \mathbb{R}^{n\times n}$ non-random

$$X^T A X = \sum_i \sum_j A_{ij} X_i X_j$$

quadratic form
w.l.o.g. $A = A^T, (A_{ij} = \frac{A^*_{ij}+A^*_{ji}}{2}, X^T A^* X = X^T A X)$
Variance estimation

$$\text{e.g. } \sum_{i=1}^n (y_i - \overline{y})^2 = y^T \begin{pmatrix} 1-\frac{1}{n} & & -\frac{1}{n} \\ & 1-\frac{1}{n} & \\ -\frac{1}{n} & & 1-\frac{1}{n} \end{pmatrix} y$$

$$A = I - \frac{1}{n}J, \mathrm{Var}(y) = y^T A y$$

Quadratic Forms

$$\mathbb{E}(Y) = \mu, \mathrm{Var}(Y) = \Sigma$$

$$\mathbb{E}(Y^T A Y) = \mu^T A \mu + \mathrm{tr}(A\Sigma)$$

$$\begin{aligned} Y^T A Y &= [\mu + (Y-\mu)]^T A [\mu + (Y-\mu)] \\ &= \mu^T A \mu + \mu^T A (Y-\mu) + (Y-\mu)^T A \mu + (Y-\mu)^T A (Y-\mu) \\ \Rightarrow \mathbb{E}(Y^T A Y) &= \mu^T A \mu + \mathbb{E}((Y-\mu)^T A (Y-\mu)) \end{aligned}$$

Trace trick: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ when both exist.

$$\begin{aligned} (Y-\mu)^T A (Y-\mu) &= \mathrm{tr}[A(Y-\mu)(Y-mu)^T] \\ \Rightarrow \mathbb{E}[(Y-\mu)^T A (Y-\mu)] &= \mathbb{E}(\mathrm{tr}[A(Y-\mu)(Y-\mu)^T]) \\ &= \mathrm{tr}(\mathbb{E}[A(Y-\mu)(A-\mu)^T]) \\ &= \mathrm{tr}(A\mathbb{E}[(Y-\mu)(Y-\mu)^T]) \\ &= \mathrm{tr}(A\Sigma) \end{aligned}$$

$$\mathrm{Var}(Y^TAY)$$

tedious computation, involves $\mathbb{E}(y_{i1}y_{i2}y_{i3}y_{i4})$
If $\mathbb{E}(Y^TA_1Y) = \mathbb{E}(Y^TA_2Y) = \sigma^2$, which better?
For $Y \sim N(0, \sigma^2 I), \mathrm{Var}(Y^TAY) = 2\sigma^4\mathrm{tr}(A^2)$

Friends of normal distribution

$$Z_i \overset{i.i.d.}{\sim} N(0,1)$$

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$$

$$\frac{Z_{k+1}}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} Z_i^2}} \overset{d}{=} \frac{N(0,1)}{\sqrt{\frac{1}{k}\chi^2(k)}} \sim t(k)$$

$$\frac{\frac{1}{n}\sum_{i=1}^{n} Z_i^2}{\frac{1}{d}\sum_{i=1}^{n+d} Z_i^2} \overset{d}{=} \frac{\frac{1}{n}\chi^2(n)}{\frac{1}{d}\chi^2(d)} \sim F_{n,d}$$

Multivariate Normal
Let $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}, Z_i \overset{i.i.d.}{=} N(0,1)$
multivariate normal is distribution of $Y = AZ + b$

$$Y \sim N(\mu, \Sigma)$$

$$\mu = A\mathbb{E}(Z) + b = b$$

$$\Sigma = A\mathrm{Var}(Z)A^T = AA^T$$

Characteristic function

$$\phi_Y(t) = \mathbb{E}(e^{it^TY}) = e^{it^T\mu - \frac{1}{2}t^T\Sigma t}, t \in \mathbb{R}^n$$

If

$$\Sigma^{-1}$$

exists, Y has density

$$(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(Y-\mu)^T\Sigma^{-1}(Y-\mu)}$$

$$Y \sim N(\mu, \Sigma)$$

$$e_i = Y_i - \overline{Y}$$

$$\sum e_i = 0 \text{ w.p. } 1$$

Partitioned Normal

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Y_1, Y_2 independent iff $\Sigma_{12} = 0$
Let $Y \sim N(\mu, \Sigma), Y \in \mathbb{R}^n, |\Sigma| \neq 0$, then $(Y - \mu)^T \Sigma^{-1}(Y - \mu) \sim \chi^2(n)$
proof Σ positive definite symmetric $\Rightarrow \Sigma = P^T \Lambda P$
P orthogonal $n \times n, \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n), \lambda_i > 0$

$$\Sigma^{-1} = P^T \Lambda^{-1} P$$

Let $Z = \Lambda^{-\frac{1}{2}} P(Y - \mu)$ (like $\frac{y-\mu}{\sigma} = [\sigma^2]^{-\frac{1}{2}}(y - \mu)$)

$$\begin{aligned} \Rightarrow (Y - \mu)^T \Sigma^{-1}(Y - \mu) &= (Y - \mu)^T P^T \Lambda^{-1} P(Y - \mu) \\ &= Z^T Z \\ &= \sum_i z_i^2 \end{aligned}$$

$$Z \sim N(0, \Lambda^{-\frac{1}{2}} P(P^T \Lambda P) P^T \Lambda^{-\frac{1}{2}}$$

$$Z \sim N(0, I_n), z_i \overset{i.i.d.}{\sim} N(0, 1)$$

Also, $Z \sim N(0, I), Q$ orthogonal

$$Y = QZ \sim N(0, QQ^T) = N(0, I)$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N(\mu, \Sigma)$$

$\Rightarrow Y_1$ also normal, Y_1 independent of $Y_2 \Leftrightarrow \Sigma_{12} = 0$
$\Rightarrow g_1(Y_1)$ independent of $g_2(Y_2)$, Y_1 independent of $Y_2^T Y_2$ ($\Rightarrow$ t-test)

$$y_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \overline{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$\begin{pmatrix} \overline{y} \\ y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \quad \cdots \quad \frac{1}{n} \\ I - \frac{1}{n} J \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & (I - \frac{1}{n})J \end{pmatrix}\right)$$

$$J = 11^T (n \times n \ 1s)$$

$$\Rightarrow \overline{y} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\sum(y_i - \overline{y})^2 \sim \sigma^2\chi^2(n-1)$$

$$\mathcal{L}(Y_2|Y_1 = y_1) = N(\underbrace{\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1 - y_1)}_{\text{linear shift in mean}}, \quad \underbrace{\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}_{\text{constant reduction in variance}} \quad)$$

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\right)$$

Get

$$\begin{aligned}\mathcal{L}(Y|X = x) &= N(\mu_y + \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \rho\sigma_x\sigma_y\sigma_x^{-2}\rho\sigma_x\sigma_y) \\ &= N(\mu_y + \rho\frac{x - \mu_x}{\sigma_x}\sigma_y, \sigma_y^2(1 - \rho^2))\end{aligned}$$

For $X_i \sim N(a_i, 1), X \sim N(a, I), i = 1, \cdots, n$
Let $\lambda = \sum a_i^2 = \|a\|^2$
then $\sum x_i^2 \sim \chi_n^2(\lambda)$ noncentral χ^2, n degrees of freedom, noncentrability λ
Used in power calculation.
Noncentral F

$$\frac{\frac{\chi_n^{2'}(\lambda)}{n}}{\frac{\chi_d^2}{d}} \sim F'_{n,d}(\lambda)$$

doubly noncentral

$$F'_{nd}(\lambda_1, \lambda_2) = \frac{\frac{\chi_n^{2'}(\lambda_1)}{n}}{\frac{\chi_d^{2'}(\lambda_2)}{d}}$$

noncentral t

$$\frac{N(0,1)}{\sqrt{\frac{1}{n}\chi_n'^2(\lambda)}}$$

Least Squares
"best" β minimizes $\mathbb{E}((y - 2\beta)^T(y - 2\beta))$

$$Y = Z\beta + \varepsilon$$

$$\left.\begin{array}{l}\text{probability} \\ \text{algebra} \\ \text{calculus} \\ \text{geometry} \\ \text{computation}\end{array}\right\}\text{once for all models}$$

Statistics — case by case

Sample least squares:
pick $\hat{\beta} \in \mathbb{R}^p$ to $\underbrace{\text{minimize}}_{\beta} \frac{1}{2} \sum_{i=1}^{n} (y_i - \overline{z_i}\beta)^2$

$$H(I-H)=0$$

$$\begin{aligned}(I-H)(I-H)&=I-H-H+H^2\\&=I-H\end{aligned}$$

Let $y=2\beta+\varepsilon, \varepsilon \sim N(0,\sigma^2 I)$
Z fixed full rank $p<n$
Then $\hat{\beta} \sim N(\beta,(Z^TZ)^{-1}\sigma^2)$

$$\hat{y} \sim N(Z\beta, H\sigma^2)$$

independent of $\hat{\varepsilon}=N(0,(I-H)\sigma^2)$ and $\sum_{i=1}^n \hat{\varepsilon_i}^2 \sim \sigma^2\chi^2_{n-p}$

$$\begin{pmatrix}\hat{\beta}\\ \hat{y}\\ \hat{\varepsilon}\end{pmatrix}=\begin{pmatrix}(Z^TZ)^{-1}Z^T\\ H\\ I-H\end{pmatrix} y \in \mathbb{R}^{2n+p}$$

$$\begin{aligned}\mathrm{Cov}(\hat{y},\hat{\varepsilon})&=\mathrm{Cov}(H\varepsilon,(I-H)\varepsilon)\\&=H\mathrm{Cov}(\varepsilon,\varepsilon)(I-H)^T\\&=H(\sigma^2 I)(I-H)\\&=0\end{aligned}$$

$$\mathrm{Cov}(\hat{\beta},\hat{\varepsilon})=0$$

$I-H$ symmetric idempotent.

$$\begin{aligned}\hat{\varepsilon}^T\hat{\varepsilon}&=[(I-H)\varepsilon]^T(I-H)\varepsilon\\&=\varepsilon^T(I-H)\varepsilon\end{aligned}$$

$$I-H=P\Lambda P^T$$

P orthogonal, $\Lambda=\mathrm{diag}(\lambda_i)$
So $\Lambda^2=\Lambda$

$$\begin{aligned}&\Rightarrow \lambda_i^2=\lambda_i, \lambda_i \in \{0,1\}\\ \sum \hat{\varepsilon_i}^2&=(P^T\varepsilon)^T\Lambda(P^T\varepsilon)\\&\overset{d}{=}\varepsilon^T\Lambda\varepsilon\\&=\sum \lambda_i\varepsilon_i^2\\&\sim \sigma^2\chi^2_{\sum\lambda_i}\end{aligned}$$

$$\sum \lambda_i = n-p$$

Why? $H = P^T(I - \Lambda)P$

$$\begin{aligned}\sum \lambda_i &= \mathrm{tr}(I - H) \\ &= n - \mathrm{tr}(H) \\ &= n - \mathrm{tr}(Z(Z^TZ)^{-1}Z^T) \\ &= n - p\end{aligned}$$

t-tests in the linear model
Linear combination of $\beta_j : C\beta = \sum_{j=1}^{p} C_j\beta_j$
e.g. $C = (0 \cdots 0 \underbrace{1}_{j\text{ th}} 0 \cdots 0)$

$$\Rightarrow C\beta = \beta_j$$

$$C = (0 \cdots \underbrace{1}_{i\text{ th}} \cdots \underbrace{1}_{j\text{ th}} \cdots 0)$$

$$C\beta = \beta_i - \beta_j$$

or, $C = (z_{01}, \cdots, z_{op}) = \vec{z_0} = z(\vec{x_0})$