0.1 Matrix Form Linear Model

Form: $Y = X\beta, X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n, e \in \mathbb{R}^n, \beta \in \mathbb{R}^p$
Assumptions: [A1] $Y = X\beta + e$ [A2] $\text{E}(e|X) = 0$ [A3] $\text{Var}(e_i|X) = \sigma^2$ [A4] $\text{Cov}(e_i, e_j|X) = 0$ [A5] $e \sim N(0, \sigma I_n)$
Normal Equations:

$$RSS(\beta) = \|(Y - X\beta)\|^2, SXX = \|x - \bar{x}\|^2, SXY = \langle x - \bar{x}, y - \bar{y}\rangle$$

$$\hat{\beta} = (X^TX)^{-1}X^TY, H = X(X^TX)^{-1}X^T, \hat{Y} = X\hat{\beta} = HY$$

$$\hat{\beta}(e) = \beta + (X^TX)^{-1}X^Te$$

Properties of H: (i) $\hat{\epsilon} = (I-H)Y$ (ii) $H, I-H$ symmetric, (iii) $H, I-H$ idempotent ($H^2 = H$) (iv) $HX = X$ (v) $\hat{e} \perp X$ (vi) $(I-H)X = 0$ (vii) $(I-H)H = H(I-H) = 0$ (viii) $\forall a \in \mathbb{R}^n, Ha \perp (I-H)a$ (ix) H only has eigen values 0,1 because $Hx = x$ if x in span H.
Variance Estimate: $\text{E}(\|\hat{e}\|) = \text{E}(\hat{e}^T(I-H)\hat{e}) = \text{E}(tr(\hat{e}\hat{e}^T(I-H))) = n\sigma^2(n-p)$, so

$$\hat{\sigma}^2 = \frac{\hat{e}^T\hat{e}}{n-p} = \frac{RSS}{n-p}$$

Variance $\hat{\beta}$: $\text{Var}(\hat{\beta}) = (X^TX)^{-1}X^T\text{Var}(Y)X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$
Gauss Markov: If a^TY is an unbiased estimator of $c^T\beta$, then $\text{Var}(c^T\hat{\beta}) \leq \text{Var}(a^TY)$. Proof: first note $c^T\beta = \text{E}(a^TY) = a^TX\beta \to c^T = a^TX$ Thus,

$$\begin{aligned}
\text{Var}(a^TY) - \text{Var}(c^T\hat{\beta}) &= \text{Var}(a^T(X\beta + e)) - \text{Var}(a^TX\beta)\\
&= \text{Var}(a^Te) - \text{Var}(a^THY)\\
&= a^T\text{Var}(e)a - \text{Var}(a^THX\beta + a^THe)\\
&= \sigma^2\|a\|^2 - \text{Var}(a^THe)\\
&= \sigma^2\|a\|^2 - Ha\text{Var}(e)a^TH\\
&= \sigma^2\|a\|^2 - \sigma^2\|Ha\|^2
\end{aligned}$$

R-squared: $R^2 = 1 - \frac{RSS}{SYY} = \text{Corr}(\hat{y}, y)$

0.2 Inference

ANOVA Table:

	df	ss	ms	F
Reg	p	$\sum(\hat{Y} - \bar{Y})$	SS/p	$F = \frac{SS/p}{\hat{\sigma}^2}$
Res	n-p	RSS	$\hat{\sigma}^2 = \frac{RSS}{n-p}$	

Distribution Estimators: $\hat{\beta}$ and $\hat{\sigma}^2$ independent under least squares, $\hat{\beta} \sim N_p(\beta, \sigma^2(X^TX)^{-1})$, and $\frac{\hat{\sigma}^2}{\sigma^2}(n-p) \sim \chi^2_{n-p}$.. Distribution of $\hat{\beta}$ follows from it being a linear transformation of Y and variance as said earlier.
Proof: Since $(I-H)$ symmetric, for P orthogonal matrix of eigenvalues and D matrix with eigenvalues on diagonal, $I - H = PDP^T$. All eigenvalues are 0 or 1, so get

$$I - H = PDP^T = [P_1P_2]\begin{bmatrix} I_{n-p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}[P_1P_2]^T = P_1P_1^T$$

So, $\text{Var}(P_1^T\hat{e}) = \text{E}(P_1^T\hat{e}\hat{e}^TP_1) - \text{E}(P_1^T\hat{e})^2 = \sigma^2P_1^TP_1 = \sigma^2I_{n-p}$. This gives us that $\frac{1}{\sigma^2}\hat{e}^T\hat{e} = \frac{1}{\sigma^2}\hat{e}^TP_1P_1^T\hat{e} \sim \chi^2_{n-p}$
Distribution Standardized Estimators: $\hat{\beta}_i \sim t_{n-p}$. Proof: $\text{Var}(\hat{\beta}_i) = \sigma^2(X^TX)^{-1}_{ii}$ so $SE(\hat{\beta}_i) = \hat{\sigma}\sqrt{(X^TX)^{-1}}$. Thus

$$\frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2(\hat{\beta}_i)}}\sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} \sim N(0,1)\sqrt{\frac{n-p}{\chi^2_{n-p}}} \sim t_{n-p}$$

t-test: $2\text{P}[t_{n-p} > \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)}]$
Prediction Interval:

$$\text{P}\left(\hat{Y}_* \in (x_*^T\hat{\beta} \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{x_*^T(X^TX)^{-1}x_*})\right) = 1 - \alpha$$

$$\text{P}\left(Y_* \in (x_*^T\hat{\beta} \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{1 + x_*^T(X^TX)^{-1}x_*})\right) = 1 - \alpha$$

F-test: If you have two models where one is a subset of the other ($span(H_1) \subset span(H_2)$), then if $rank(H_1) = q, rank(H_2) = p$,

$$\frac{\frac{1}{p-q}(\|\hat{e}_1\|^2 - \|\hat{e}_2\|^2)}{\frac{1}{n-p}\|\hat{e}_2\|^2} \sim F_{p-q,n-p}$$

This is a one sided test. Good for testing sets of parameters.
Joint Confidence Interval: A $1-\alpha$ confidence region for β is

$$\frac{\frac{1}{p}(\hat{\beta} - \beta)^T(X^TX)(\hat{\beta} - \beta)}{\hat{\sigma}^2} \leq p\hat{\sigma}^2 f_{p,n-p,\alpha}$$

If $R\beta$ has rank q, a $1-\alpha$ confidence region for $R\beta$ is

$$\frac{\frac{1}{p}(R\hat{\beta} - R\beta)^T(R(X^TX)^{-1}R^T)^{-1}(R\hat{\beta} - R\beta)}{\hat{\sigma}^2} \leq p\hat{\sigma}^2 f_{q,n-p,\alpha}$$

0.3 Numerical Techniques

Condition Number: This is something to do with the effect of a small change in Y on β. With

$$cos(\theta) = \frac{\|\hat{Y}\|}{\|Y\|} = \frac{\|X\hat{\beta}\|}{\|Y\|}$$

$$\frac{\|\Delta\hat{\beta}\|}{\|\hat{\beta}\|} \leq cond(X)\frac{1}{cos(\theta)}\frac{\|\Delta Y\|}{\|Y\|}$$

Cholesky Factorization: If X has rank n, X^TX has full rank, and has Cholesky factorization LL^T. Thus, $X^TX\hat{\beta} = X^TY$, which can be solved in stages $Lz = X^TY$ and then $L^T\hat{\beta} = z$.
QR Factorization: $\exists Q \in \mathbb{R}^{n\times n}, R \in \mathbb{R}^{p\times p}$ where Q is orthogonal and R is upper triangular such that

$$X = \begin{bmatrix} R \\ \mathbf{0}\end{bmatrix}$$

so we get

$$Q^TX\hat{\beta} = \begin{bmatrix} R \\ \mathbf{0}\end{bmatrix}\hat{\beta} \cong \begin{bmatrix} f \\ r\end{bmatrix} = Q^TY$$

This gives us $RSS = \|y - X\hat{\beta}\|^2 = \|Q^Ty - Q^TX\hat{\beta}\|^2 = \|f - R\hat{\beta}\|^2 + \|r\|^2$, which is minimized by $f = R\hat{\beta}$.

0.4 Resampling

Permutation Sampling: Test significance of set of predictors by shuffling them over outcomes and other predictors a number of times. If F statistic original model higher than all but α of shuffles, significant.
Bootstrap: Get confidence interval of statistic (possibly θ) by drawing with replacement a number of times and calculating statistic.

0.5 Designed Experiment

Orthogonal Predictor: If X_1, X_2 orthogonal, then

$$\begin{aligned}\beta = (X^TX)^{-1}X^TY &= \begin{bmatrix} X_1^TX_1 & 0 \\ 0 & X_2^TX_2 \end{bmatrix}^{-1} X^TY \\ &= \begin{bmatrix} (X_1^TX_1)^{-1}X_1^TY \\ (X_2^TX_2)^{-1}X_2^TY \end{bmatrix}\end{aligned}$$

Estimates don't change if X_1 or X_w removed, both less dependent other non-orthogonal vars.
Randomization: If Z can't be included in regression, in an experiment, by randomly assigning it to observations, $\text{Cov}(X, Z)$ should be 0, so effect Z part of error.
Lurking Variable If Z correlated with X, then,

$$\text{E}(Y|x, z) = X\beta + \delta z$$

$$\text{E}(Z|x) = X\gamma$$

so

$$\text{E}(Y|x) = X(\beta + \gamma)$$

0.6 Diagnostics

Non-Constant Variance: Can Regress $|\hat{e}|$ on $\hat{Y}$ if. **Transform:** Transform non-linear/non-constant residual data.

$$h(Y) = \log(Y + \delta), h(Y) = \sqrt(Y)$$

Not Normal: QQplot, Shapiro-Wilk
Correlated Error: Durbin-Watson, where ρ autocorrelation:

$$d = \frac{\sum_{i=2}^{n}(\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n}\hat{e}_i^2} \sim 2(1-\rho)$$

Leverage: $h_i = H_ii = x_i^T(X^TX)^{-1}x_i$. How strongly effects model.
Outlier Test: $\hat{y}_{(i)}$ excludes ith observation.

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}_{(i)}^2(x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i + 1)}}$$

Where $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}$ (studentized residuals), this gives us

$$t_i = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}} \sim t_{n-p-1}$$

Bonferroni Correction: reject only if $t_{n-p-1,\alpha/n} > t_i$.
Cook Statistic: Indicates influential point, whose removal effects fit.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T(X^TX)(\hat{\beta} - \hat{\beta}_{(i)})}{p\hat{\sigma}^2} = \frac{1}{p}r_i^2\frac{h_i}{1-h_i}$$

Partial Residual Plots Fit models, where $X_{(i)}$ excludes column i,

$$Y = X_{(i)}\beta_{(i)} + q_i, X_i = X_{(i)}\gamma + s_i$$

Plot q_i in terms s_i. Can see leverage of points on β_i

0.7 Distributions

Normal Distribution: $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, joint normal vars independent iff $\text{Cov}(Z_1, Z_2) = 0$.
Multivariate Normal Distribution: if $X \sim N(\mu, \Sigma)$ and Σ positive definite,

$$f_X(x) = \frac{1}{(2\pi)^{k/2}|\Sigma||^{1/2}}exp\left(-\frac{1}{2}(x-mu)^T\Sigma^{-1}(x-\mu)\right)$$

Chi Square: k degree of freedom, then $\sum_{i=1}^k Z_i^2 \sim \chi_k^2$ if Z_is independent standard normals
Student's T: $t_\nu \sim Z\sqrt{\frac{\nu}{\chi_\nu^2}}$ if Z standard normal independent of χ_ν^2.
F: $F_{d_1,d_2} \sim \frac{\chi_{d_1}^2/d_1}{\chi_{d_2}^2/d_2}$ if $\chi_{d_1}^2$ and $\chi_{d_2}^2$ independent.

0.8 Linear Algebra

Cauchy Scwarz: $|\langle x, y\rangle| \leq \|x\|\|y\|$. Equality iff linearly independent
Triangle Inequality: $\|x+y\|^{\leq}\|x\| + \|y\|$
Rank: Number linearly independent columns/rows. $rk(A_{m\times n}) \leq \min(m,n), rk(AB) \leq \min(rk(A), rk(B)), rk(A+B) \leq rk(A) + rk(B), Rk(AA^T = rk(A))$
Orthogonal: $A^TA = I$. Columns A orthonormal basis R^n, rotate/reflect vector. $\langle Ax, Ax\rangle = \langle x, x\rangle$.
Idempotent: $AA = A$. Projection matrix. If $x \in span(A), Ax = x$.
Determinant: $|AB| = |A||B|$, if A orthogonal, $|A| = \pm 1, |A^TBA| = |B|$.
Trace: Sum diagonal entries. $tr(A) = tr(A^T), tr(A+B) = tr(A) + tr(B), tr(ABC) = tr(CAB)$, if A idempotent, $rk(A) = tr(A)$, if A nonsingular, $tr(A^{-1}BA) = tr(B)$.
Eigenvalues: If A idempotent, $\lambda = 1, 0$. If Orthogonal λ has modulus 1 (radius in complex plane less than 1). Symmetric matrix has real eigenvalues.
Positive (semi) definite: $x^TAx > (\geq)0\forall x$. BB^T always positive semi definite.
Eigen Decomposition: If A symmetric, $A = P^TDP$, where P matrix eigenvectors and D diagonal matrix of eigenvalues. If AB=BA and symmetric, $B = P^TD_BP$ for same P.
Diagonally Dominant: If each diagonal greatest entry in column, A positive semi definite if A symmetric and diagonals positive. Strictly diagonally dominant matrix nonsingular.
Cholesky Decomposition: If A symmetric and positive definite, $\exists L$ unique lower diagonal with positive diagonal entries such that $A = LL^T$. If A only positive semi-definite, L may not be unique and may have 0 diagonal entries. **Underdetermined Linear System:** Smallest norm solution to $Ax = y$ is $x = A^T(AA^T)^{-1}y$.