Transformations

Box-Cox works well for monotonic / simple relationships

Box-Cox: $f(x) = \begin{cases} \frac{x^p - 1}{p}, & p \neq 0 \\ \ln(x), & p = 0 \end{cases}$ * $\lim_{p\to 0} \frac{x^p-1}{p} = \lim_{p\to 0} \frac{(\ln x)x^p}{1}$ ← L'Hôpital's Rule

ex) Box-Cox$(x, p, \lambda) = \frac{x^p - 1}{p} \to \frac{(x-\mu)^p - 1}{p}$

x	x^{-1}	$x^{-1/2}$
1	1	1
2	1/2	$-1/\sqrt{2}$
3	1/3	$-1/\sqrt{3}$

- what if the reason for dividing by p?
  - → preserves the direction of x, which would be reversed when $p<0$
  - → For $p \to 0$, $\lim_{p\to 0} \frac{x^p-1}{p} = \ln(x)$ so Box-Cox is continuous in p.
- x shift by 1 ensures that all transformations go through $(0,1)$.

[Metric for symmetry]: $S = \frac{(\text{upper quartile}) - \text{median}}{\text{median} - (\text{lower quartile})}$

$S > 1$: skewed right

- for monotonically increasing transformations, $\text{IQR}(f(x)) = f(\text{IQR}(x))$, median$(f(x)) = f(\text{median})$ for all quartiles.

Q. When is power transformations effective?

only when

1) all X values are positive
   - → $\sqrt{\ }$, log, inverse are unaffected for negative or zero values
   - → power transformations are non-monotone (eg. x^2)

x	-2	-1	0	1	2
x^2	4	1	0	1	4

Solution:

If data domain includes 0 or is negative or data are all large or close together, then **shift values by some u**

2) Ratio of the largest data to smallest data is sufficiently large. If the ratio ≈ 1, then the power transformation is roughly linear, and ineffective at bending the data. **Shifting by u solves the problem.**

x	$\log_{10} x$		x	$\log_{10}(x - 2010)$
2011	3.30341		2011	0
2012	3.30363	0.00022	2012	0.301
2013	3.30384	0.00021	2013	0.477
2014	3.30406	0.00022	2014	0.602
2015	3.30428	0.00022	2015	0.699

		0.176	
		0.125	
		0.097	

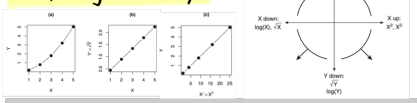
Power transformations preserve the order of the data only when all values are positive and are effective only when the ratio of the largest to the smallest data values is itself large. When these conditions do not hold, we can impose them by adding a positive or negative start to all the data values.

Descending the ladder of powers (e.g., to $\log X$) tends to correct a positive skew; ascending the ladder of powers (e.g., to X^2) tends to correct a negative skew.

positive skew — log transformation pulls in the right tail.

x	$\log x$
1	0
10	1
100	2
1000	3

Transforming Nonlinearity



Simple monotone nonlinearity can often be corrected by a power transformation of X, of Y, or of both variables. Mosteller and Tukey's bulging rule assists in selecting linearizing transformations.

Transforming Proportions

① logit $p \to \text{logit}(p) = \log_e \frac{p}{1-p}$

- remove the upper & lower boundaries of the scale, spreading out the tail distribution and making resulting quantiles symmetric about 0.

② probit $p \to \text{probit}(p) = \Phi^{-1}(p)$

- logit $\approx (\pi/\sqrt{3}) \times$ probit

Linear Algebra Review

- Norm: $\|\vec{x}\|_p = (\sum |x_i|^p)^{1/p}$
- Dot product $\langle \vec{x}, \vec{y} \rangle = \vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \sum x_i y_i$

  $\vec{x} \cdot \vec{y} = \|\vec{x}\|_2 \|\vec{y}\|_2 \cos\theta$ $\quad \|\vec{x}\|_2^2 = \vec{x} \cdot \vec{x}$
- Orthogonality = dot-product $\langle \vec{x}, \vec{y} \rangle = 0$
- $\|\vec{x}+\vec{y}\|_2 = (\vec{x}+\vec{y})^T(\vec{x}+\vec{y}) = \vec{x}^T\vec{x} + \vec{x}^T\vec{y} \cdot 2 + \vec{y}^T\vec{y}$
- Orthogonal basis: If $\vec{u} \in V$ and $\{\vec{v}_1, \dots, \vec{v}_p\}$ is orthobasis of V, then $\vec{u}$ is a unique linear combinations of $\vec{v}_i$'s.

  $\vec{u} = \sum c_i \vec{v}_i$ where $c_i = \frac{\vec{u} \cdot \vec{v}_i}{\vec{v}_i \cdot \vec{v}_i}$
- Projections: $\text{argmin}_c \|\vec{v} - c\vec{u}\|_2^2 = \frac{\vec{u}^T\vec{v}}{\vec{u}^T\vec{u}} := c^*$
- projecting a vector onto the basis is the minimizer.

  ex) project $\vec{y}$ onto $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$: $c^* = \frac{\vec{1}^T\vec{y}}{\vec{1}^T\vec{1}} = \bar{y}$

Gram Schmidt for generic $\{\vec{v}_1, \vec{v}_2, \vec{v}_3\}$

① $\vec{u}_1 = \vec{v}_1$

② $\vec{u}_2 = \vec{v}_2 - \text{proj}_{\vec{u}_1}(\vec{v}_2)$

③ $\vec{u}_3 = \vec{v}_3 - \text{proj}_{\vec{u}_1}(\vec{v}_3) - \text{proj}_{\vec{u}_2}(\vec{v}_3)$

ex) $\vec{1}, \vec{x}$: ① $\vec{u}_1 = \vec{1}$ ② $\vec{u}_2 = \vec{x} - \text{proj}_{\vec{1}}(\vec{x}) = \vec{x} - \bar{x}\vec{1}$

For square matrix A, $\text{trace}(A)$ is the sum of diagonal elements.

- $\text{tr}(A-B) = \text{tr}(A) - \text{tr}(B)$

Multiple Regression

$y_i = \alpha + B_1 x_{i1} + B_2 x_{i2} + \dots + B_k x_{ik} + \varepsilon_i$

$\vec{y} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & & \\ 1 & x_{n1} & & x_{np} \end{bmatrix} \begin{bmatrix} B_0 \\ \vdots \\ B_p \end{bmatrix} + \vec{\varepsilon} = X\vec{B} + \vec{\varepsilon}$

- Assume X is full rank $\Rightarrow X^TX$ invertible
- optimal $\vec{B}$

$\vec{y} = \text{proj}_{\text{span}(X)}(\vec{y}) + (\vec{y} - \text{proj}_{(X)}(\vec{y})) = X\vec{B} + \vec{\varepsilon}$

$X^T\vec{y} = X^TX\vec{B} + X^T\vec{\varepsilon}$, $X^T\vec{\varepsilon} = 0$ $(\vec{x} \perp \vec{\varepsilon})$

$\hat{B} = (X^TX)^{-1}X^T\vec{y}$

Probabilistic viewpoint

- consider vector valued random variable $V = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$
- For scalar v: $\text{var}(v) = \mathbb{E}[(v - \mathbb{E}(v))^2]$

For vector V: $\mathbb{E}(V) = \begin{bmatrix} \mathbb{E}v_1 \\ \vdots \\ \mathbb{E}v_n \end{bmatrix}$

$\text{var}(V) := \mathbb{E}[(V - \mathbb{E}(V))(V - \mathbb{E}(V))^T] = \begin{bmatrix} \text{var}v_1 & \text{cov}(v_1, v_2) & \cdots \\ \text{cov}(v_2, v_1) & & \\ \vdots & & \text{var}v_n \end{bmatrix} := \Sigma_{V,V}$

(Quadratic Form) $\forall^T A \forall$: $\forall \in \mathbb{R}^n$ RV, $A \in \mathbb{R}^{n \times n}$ constant, square matrix

i) $\forall^T A \forall \in \mathbb{R}$

ii) $\forall^T A \forall = \forall^T \begin{bmatrix} \sum a_{ij} v_j \\ \vdots \\ \sum a_{nj} v_j \end{bmatrix} = \sum_{i=1}^n v_i \sum_{j=1}^n a_{ij} v_j = \sum_{i=1}^n \sum_{j=1}^n a_{ij} v_i v_j$

Some facts about quadratic form

① [Quadratic Expectation] $\mathbb{E}[\forall^T A \forall] = \text{trace}(A\Sigma_{\forall\forall}) + (\mathbb{E}\forall)^T A \mathbb{E}\forall$

② If $\vec{u} := A\vec{v}$, $\vec{w} := B\vec{v}$ with A, B constant, and $\vec{v}$ random,

$\Sigma_{\vec{u},\vec{w}} = \mathbb{E}[(\vec{u} - \mathbb{E}\vec{u})(\vec{w} - \mathbb{E}\vec{w})^T] = A\Sigma_{\vec{v}\vec{v}} B^T$

Model $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{B} \in \mathbb{R}^p$, $\vec{\varepsilon} \in \mathbb{R}^n$

$\vec{y} = X\vec{B} + \vec{\varepsilon}$

- $\mathbb{E}[\vec{y}|X] = \mathbb{E}[X\vec{B} + \vec{\varepsilon}|X] = \mathbb{E}[X\vec{B}|X] + \mathbb{E}[\vec{\varepsilon}|X] = X\mathbb{E}[\vec{B}|X] = X\vec{B}$
- $\Sigma_{\vec{y},\vec{y}|X} = \mathbb{E}[(\vec{y} - \mathbb{E}[\vec{y}|X])(\vec{y} - \mathbb{E}[\vec{y}|X])^T|X]$

  $= \mathbb{E}[(X\vec{B} + \vec{\varepsilon} - X\vec{B})(X\vec{B} + \vec{\varepsilon} - X\vec{B})^T] = \mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T|X] = \Sigma_{\vec{\varepsilon},\vec{\varepsilon}|X} = \sigma_\varepsilon^2 \mathbb{I}_n$
- $\mathbb{E}[\hat{\vec{B}}|X] = \mathbb{E}[(X^TX)^{-1}X^T\vec{y}|X] = (X^TX)^{-1}X^T\mathbb{E}[\vec{y}|X] = (X^TX)^{-1}X^TX\vec{B} = \vec{B}$
- $\Sigma_{\hat{\vec{B}},\hat{\vec{B}}|X} = \mathbb{E}\left[(\hat{\vec{B}} - \mathbb{E}[\hat{\vec{B}}|X])(\hat{\vec{B}} - \mathbb{E}[\hat{\vec{B}}|X])^T\right]$

  $= \mathbb{E}[((X^TX)^{-1}X^T\vec{\varepsilon})((X^TX)^{-1}X^T\vec{\varepsilon})^T] = \mathbb{E}[(X^TX)^{-1}X^T\vec{\varepsilon}\vec{\varepsilon}^TX(X^TX)^{-1}]$

  $= (X^TX)^{-1}X^T\mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T|X](X^TX)^{-1} = \sigma_\varepsilon^2(X^TX)^{-1}$ ($\mathbb{E}[\vec{\varepsilon}\vec{\varepsilon}^T|X] = \sigma^2\mathbb{I}_n$)
- $\hat{B}$ under ground truth model: $N(B, \Sigma_{\hat{B},\hat{B}|X})$
- In practice, we estimate σ^2 as $s_\varepsilon^2 = \frac{\sum e_i^2}{n-p-1}$ where p is # of covariate. ($\hat{\sigma}^2_{\varepsilon_{ii}}$: variance of regression)

The standard error of the regression, $S_E = \sqrt{\sum E_i^2/(n-k-1)}$, gives the "average" size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS}/\text{TSS}$, indicates the proportion of the variation in Y that is captured by its linear regression on the X's.

- $\hat{B}$ (from projection) is also MLE for B under the normal model.

(Gauss Markov Thm) Among all linear unbiased estimator $\hat{B}$ for $\vec{y} = X\vec{B} + \vec{\varepsilon}$, where ε_i are uncorrelated, have equal variance, and have mean zero, $\hat{B}_{OLS}$ has smallest sampling variance.

R_j^2: squared multiple correlation from the regression of x_j on all other x_s.

VIF

$V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$

Variance Inflation Factor (VIF)

- large when x_j is strongly correlated with other explanatory variables ($\vec{x}_j$) → How collinearity affects variance of parameter estimates.
- VIF $= \infty$ if one explanatory variable is perfectly collinear with others.

- Two reasons why we should be careful about selecting only necessary features
  - ① $\hat{\sigma}_\varepsilon^2 = \frac{1}{n-(p+1)}\sum e_i^2$, $\hat{\sigma}_\varepsilon^2 \uparrow$ as $p \uparrow$
  - ② $SE(\hat{B}_j)$ increase (collinearity).

Confidence Interval & Hypothesis Testing

$s_\varepsilon^2 = \frac{\sum E_i^2}{n-k-1}$ (unbiased estimator of σ_ε^2)

Single slope $B_k = 0$

$\text{SE}(B_j) = \frac{1}{\sqrt{1-R_j^2}} \times \frac{S_E}{\sqrt{\sum (x_{ij} - \bar{x}_j)^2}}$

Subset of slopes

$H_0: B_1 = B_2 = \dots = B_q = 0$

$F_0 = \frac{(\text{RegSS}_1 - \text{RegSS}_0)/q}{\text{RSS}_1/(n-k-1)} = \frac{n-k-1}{q} \times \frac{R_1^2 - R_0^2}{1 - R_1^2}$ $\sim F_{q, n-k-1}$

① RegSS_1: result from full

② RSS_0, RegSS_0: result from model w/ 0 $B_1, \dots, B_q$

All slopes

$H_0: B_1 = B_2 = \dots = B_k = 0$

$F_0 = \frac{\text{RegSS}/k}{\text{RSS}/(n-k-1)} = \frac{\rho}{k} \times \frac{R^2}{1-R^2}$ F-distribution $(\rho, n-\rho-1)$

relating $SE(\hat{B}_j)$ with FWL thm

Proof of FWL

Start with OLS:

$\vec{y} = X_1\hat{B}_1 + X_2\hat{B}_2 + \vec{e}_n$

$\Rightarrow M_{X_1}\vec{y} = M_{X_1}(X_1\hat{B}_1 + X_2\hat{B}_2 + \vec{e}_n)$

$= 0 + M_{X_1}(X_2\hat{B}_2) + M_{X_1}\vec{e}_n$

$(M_{X_1}\vec{e}_n = \vec{e}_n = M_X\vec{y}) = M_{X_1}(X_2\hat{B}_2) + \vec{e}_n$

Let $M_{X_1} = \mathbb{I}_n - X_1(X_1^TX_1)^{-1}X_1^T$ be the orthogonal complement of X_1 projection matrix.

Then for OLS regression:

$M_{X_1}\vec{y} = (M_{X_1}X_2)\hat{B}_2 + \vec{e}_n$ where $\hat{B}_2 = \hat{B}_2$

→ B_2 is applied to the part of X_2 that is uncorrelated with X_1

Terms $M_{X_1}\vec{y}$: residuals of $\vec{y}$ on X_1

$M_{X_1}X_2$: residuals of X_2 on X_1.

$M_{X_j}\vec{y} = (M_{X_j}X_j)\hat{B}_j + \vec{e}_n$ [simple regression]

$SE^2(\hat{B}_j) = \frac{\sigma_\varepsilon^2}{\sum[(M_{X_j}\vec{x}_j)\cdot(M_{X_j}\vec{x}_j)]^2} = \frac{\sigma_\varepsilon^2}{\text{RSS}_{\vec{x}_j, X_{-j}}} = \frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j, X_{-j}} - \text{RegSS}_{\vec{x}_j, X_{-j}}}$

$= \frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j}(1-R_j^2)} = \frac{1}{1-R_j^2}\frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j}}$

- $\hat{B}_j$ is essentially partial correlation with size/variance of $X_j\vec{y}$ into account.

Degrees of freedom dimension of the subspace to which the vector is confined

$\vec{e}_n = \vec{y} - \vec{\hat{y}} = \vec{y} - X(X^TX)^{-1}X^T\vec{y} = (\mathbb{I}_n - X(X^TX)^{-1}X^T)\vec{y} := Q\vec{y}$

$\text{rank}(Q) = n-(p+1)$

$\dim\{\vec{y} - \hat{\vec{y}} \mid \vec{y} \in \mathbb{R}^n\} = \dim\{Q\vec{y} \mid \vec{y} \in \mathbb{R}^n\} = \text{rank}(Q) = n-(p+1) = df_{\vec{e}}$

properties of Q

- idempotent & Symmetric $\Rightarrow \text{rank}(Q) = \text{tr}(Q)$
- $\text{tr}(Q) = \text{tr}(\mathbb{I}_n - X(X^TX)^{-1}X^T) = \text{tr}(\mathbb{I}_n) - \text{tr}((X^TX)^{-1}X^TX) = n - \text{tr}(\mathbb{I}_{p+1})$

  $= n-(p+1) = \text{rank}(Q)$

$\vec{e}_n \sim N(0, \sigma_\varepsilon^2 Q)$

- $\text{var}(\vec{e}_n) = \text{var}(Q\vec{y}) = Q\text{var}(\vec{y})Q^T = Q\sigma_\varepsilon^2\mathbb{I}_nQ^T = \sigma_\varepsilon^2 QQ^T = \sigma_\varepsilon^2 Q$

Simple Linear Regression

In simple linear regression

$$Y_i = A + BX_i + E_i$$

the least-squares coefficients are given by $A = \bar{Y} - B\bar{X}$ and $B = \sum(X_i - \bar{X})(Y_i - \bar{Y}) / \sum(X_i - \bar{X})^2$. The slope coefficient B represents the average change in Y associated with a one-unit increase in X. The intercept A is the fitted value of Y when $X = 0$.

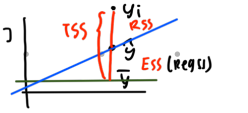
- SE (residual standard error) $SE^2(B_{simple}) = \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2}$

$s_E^2 = \frac{\sum E_i^2}{n-2}$

R squared ← total variation of Y that is captured by its linear regression on X.

- $\text{cov}(X,y) = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

$R^2 = \frac{\text{RegSS}}{\text{TotSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$ $\in [0,1]$



- RSS (residual sum of squares) $\sum E_i^2 = \sum(y_i - \hat{y}_i)^2$
- TSS (total sum of squares) $\sum E_i'^2 = \sum(y_i - \bar{y})^2$
- RegSS (regression sum of squares) $\text{RegSS} = \text{TSS} - \text{RSS}$

$SD(\vec{e}) \approx \sqrt{1-r^2}\, SD(\vec{y})$

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

Facts about R^2 1) R^2 is unitless.

r	trend
1	perfect positive relation
-1	perfect negative relation

R can be viewed under 2 lights:
① regression of y on x
② correlation $\rho = \sigma_{xy}/\sigma_x\sigma_y$ between 2 RV.

(sample covariance) between x and y: $S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})\sum(y_i - \bar{y})^2}}$

$\|\vec{y} - \bar{y}\vec{1}\|_2^2 = \|\vec{y} - \hat{\vec{y}}\|^2 + \|\hat{\vec{y}} - \bar{y}\vec{1}\|^2$

TSS = RSS + RegSS

For our simple regression, $R^2 = \rho_{x,y}^2$

Adjusted R squared

$\text{Adj } R^2 = 1 - \left(\frac{n-1}{n-(p+1)}\right)\frac{\text{RSS}}{\text{TSS}} = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$

Omitted Variable Bias

OVB: $\vec{y} = X\beta + \vec{z}\delta + \vec{\varepsilon}$; $X \in \mathbb{R}^{n \times p}$, $\vec{z} \in \mathbb{R}^n$

(confounder) $\vec{z}$ is a confounder if i) it determines $\vec{y}$

ii) correlates with other variables of interest X.

OLS without $\vec{z}$: $\hat{\beta} = (X^TX)^{-1}X^T\vec{y}$

$= (X^TX)^{-1}X^T(X\beta + \vec{z}\delta + \varepsilon)$

$= \beta + (X^TX)^{-1}X^T\vec{z}\delta + (X^TX)^{-1}X^T\vec{\varepsilon}$

$\mathbb{E}[\hat{\beta}|X] = \beta + (X^TX)^{-1}\mathbb{E}[X^T\vec{z}]\delta + 0$

omitted variable BIAS (OVB)

- If X and $\vec{z}$ have higher correlation then $\mathbb{E}[X^T\vec{z}|X]$ increases
- If $\vec{y}$ and $\vec{z}$ have higher dependence, then δ increases
  - → Bias from omitting $\vec{z}$ is larger.

$MSE := \frac{1}{n}\|\vec{y} - \hat{\vec{y}}\|^2$

partial correlation: linear relationship between $\vec{y}$ and $\vec{x}_i$ in the space orthogonal to X_2.

The partial correlation between $\vec{y}$ and x_i given X_2 is computed as follows:

① Regress $\vec{y}$ on X_2, yielding $\vec{e}_y$

② Regress each column of X_1 on X_2, yielding $\vec{e}_{\vec{x}_1}, \dots, \vec{e}_{\vec{x}_p}$

③ The correlation between $\vec{e}_y$ and $\vec{e}_{\vec{x}_1}, \dots, \vec{e}_{\vec{x}_p}$ are partial correlation.

- Multiple regression yield parameters that directly correspond to the partial correlations.

FWL Theorem

- (use) reducing multivariate regression to univariate ones.
- Model: $y_i = B_1x_{i1} + B_2x_{i2} + \varepsilon_i$

  B_1 are equivalent to:
  ① OLS estimator obtained by regressing y on x_1 and x_2
  ② OLS estimator obtained by regressing y on the residual from regression of x_1 on x_2
  ③ OLS estimator obtained by regressing residual from the regression y on x_2 on x_1.

$\vec{y} = X_1\hat{B}_1 + X_2\hat{B}_2 + \vec{e}_n$

Let $M_{X_1} = \mathbb{I}_n - X_1(X_1^TX_1)^{-1}X_1^T$ be the orthogonal complement of X_1 projection matrix.

Then for OLS regression:

$M_{X_1}\vec{y} = (M_{X_1}X_2)\hat{B}_2 + \vec{e}_n$ where $\hat{B}_2 = \hat{B}_2$

→ B_2 is applied to the part of X_2 that is uncorrelated with X_1

Terms $M_{X_1}\vec{y}$: residuals of $\vec{y}$ on X_1

$M_{X_1}X_2$: residuals of X_2 on X_1.

$M_{X_j}\vec{y} = (M_{X_j}X_j)\hat{B}_j + \vec{e}_n$ [simple regression]

$SE^2(\hat{B}_j) = \frac{\sigma_\varepsilon^2}{\sum[(M_{X_j}\vec{x}_j)\cdot(M_{X_j}\vec{x}_j)]^2} = \frac{\sigma_\varepsilon^2}{\text{RSS}_{\vec{x}_j, X_{-j}}} = \frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j, X_{-j}} - \text{RegSS}_{\vec{x}_j, X_{-j}}}$

$= \frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j}(1-R_j^2)} = \frac{1}{1-R_j^2}\frac{\sigma_\varepsilon^2}{\text{TSS}_{\vec{x}_j}}$

- $\hat{B}_j$ is essentially partial correlation with size/variance of $X_j\vec{y}$ into account.

- Note that Q's diagonal elements are less than 1. Hence, $\mathbb{E}[\frac{1}{n}\sum e_i^2] \neq \sigma_\varepsilon^2$ as the e_i's have deflated variance to σ_ε^2 ($= V(e_i) \leq \sigma_\varepsilon^2$), since there are only $n-(p+1)$ free components in $\vec{e}_n$, this suggests that we can represent $\vec{e}_n$ as an $n-(p+1)$ size vector.
- → use GRAM-SCHMIDT to transform $\vec{e}_n$ into lower dimension (Gram schmidt procedure for e_i)

Let $V = \text{col}(X)$

① Find basis of $V^\perp = \text{Null}(X^T)$

→ since $\text{rank}(X) = p+1$, $G \in \mathbb{R}^{n \times (n-p-1)}$

② Make G orthogonal using Gram Schmidt

③ Normalize $\begin{bmatrix} \vec{g}_1 \\ \overline{g_1} \end{bmatrix} \cdots \begin{bmatrix} \vec{g}_{n-p-1} \\ \overline{g_{n-p-1}} \end{bmatrix}$ s.t. $\|\vec{g}_i\|^2 = 1$

properties of G matrix

i) $G^TX = 0$

ii) $G^TG = \mathbb{I}_{n-p-1}$

iii) For $\vec{z}_{n-p-1} := G^T\vec{e}_n$ ($\vec{e}_n \sim N(0, \sigma_\varepsilon^2\mathbb{I}_n)$)

Table 9.1 Comparison Between Simple Regression Using Scalars and Multiple Regression Using Matrices

	Simple Regression	*Multiple Regression*
Model	$Y = \alpha + \beta x + \varepsilon$	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Least-squares estimator	$B = \frac{\sum x^* Y^*}{\sum x^{*2}}$ $= (\sum x^{*2})^{-1} \sum x^* Y^*$	$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Sampling variance	$V(B) = \frac{\sigma_\varepsilon^2}{\sum x^{*2}}$ $= \sigma_\varepsilon^2 (\sum x^{*2})^{-1}$	$V(\mathbf{b}) = \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}$
Distribution	$B \sim N[\beta, \sigma_\varepsilon^2 (\sum x^{*2})^{-1}]$	$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2 (\mathbf{X}'\mathbf{X})^{-1}]$

NOTE: Subscripts are suppressed in this table; in particular, $x^* \equiv x_i - \bar{x}$ and $Y^* \equiv Y_i - \bar{Y}$.

Maximum Likelihood Estimation

- under the linear model, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}_n)$, b is MLE of $\boldsymbol{\beta}$.
- For the ith observation, $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_\varepsilon^2)$ where $\mathbf{x}_i'$ is ith row of model matrix $\mathbf{X}$.
- pdf of observation i is:

$$p(y_i) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} \exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right]$$

Because the n observations are independent, their joint probability density is the product of their marginal densities:

$$p(\mathbf{y}) = \frac{1}{(\sigma_\varepsilon\sqrt{2\pi})^n} \exp\left[-\frac{\sum (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma_\varepsilon^2}\right] = \frac{1}{(2\pi\sigma_\varepsilon^2)^{n/2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\varepsilon^2}\right] \quad (9.12)$$

- log likelihood is:

$$\log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- partial derivatives is:

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma_\varepsilon^2}(2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y})$$

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma_\varepsilon^2)}{\partial \sigma_\varepsilon^2} = -\frac{n}{2}\left(\frac{1}{\sigma_\varepsilon^2}\right) + \frac{1}{2\sigma_\varepsilon^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- der = 0: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, (MLE $\hat{\boldsymbol{\beta}}$ = OLS $\hat{\boldsymbol{\beta}}$)

$$\hat{\sigma}_\varepsilon^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n} \quad \text{(biased)}$$

↳ MLE's variance $\hat{\sigma}_\varepsilon^2$ is biased. So, we prefer unbiased estimator $S_E^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$ to $\hat{\sigma}_\varepsilon^2$. MLE is still consistent b/c as $n \to \infty$, $\hat{\sigma}_\varepsilon^2 \to 0$.

Inference for Individual Coefficients

We saw that the least-squares estimator $\mathbf{b}$ follows a normal distribution with expectation $\boldsymbol{\beta}$ and covariance matrix $\sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}$.[33] Consequently, an individual coefficient B_j is normally distributed with expectation β_j and sampling variance $\sigma_\varepsilon^2 v_{jj}$, where v_{jj} is the jth diagonal entry of $(\mathbf{X}'\mathbf{X})^{-1}$.[34] The ratio $(B_j - \beta_j)/\sigma_\varepsilon\sqrt{v_{jj}}$, therefore, follows the unit-normal distribution $N(0, 1)$, and to test the hypothesis $H_0\colon \beta_j = \beta_j^{(0)}$, we can calculate the test statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\sigma_\varepsilon\sqrt{v_{jj}}}$$

unknown

σ_ε^2 is unknown but $S_E^2 = \frac{\mathbf{e}'\mathbf{e}}{n-p-1}$ can estimate it.

$\hat{V}(b) = S_E^2(X'X)^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n-p-1}(X'X)^{-1}$

- Testing Hypothesis $H_0: \beta_j = \beta_j^{(0)}$, we calculate: $t_0 = \frac{B_j - B_j^{(0)}}{SE(B_j)}$
- 100(1-a)% CI for β_j: $\beta_j = B_j \pm t_{a/2, n-p-1} SE(B_j)$

The estimated covariance matrix of the least-squares coefficients is $\widehat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1}$. The standard errors of the regression coefficients are the square-root diagonal entries of this matrix. Under the assumptions of the model, $(B_j - \beta_j)/\mathrm{SE}(B_j) \sim t_{n-k-1}$, providing a basis for hypothesis tests and confidence intervals for individual coefficients.

Inference for Multiple Coefficients

$H_0: \beta_1 = \cdots = \beta_q = 0$

(LRT) $\lambda_{LRT} = \frac{\mathcal{L}_{H_0}(\vec{y})}{\mathcal{L}_{H_1}(\vec{y})}$ where $L_{H_0}(\vec{y})$ is max likelihood of the data under the null model,

and the test is rejects H_0 if $\lambda_{LRT} < c$.

Thm (Neyman-Pearson Lemma) The LRT is the most powerful test for any given hypothesis. Highest power for a given Type I error rate.
[reject H_0 when H_A is true].

$L_{H_1}(\vec{y}) = \frac{1}{(2\pi\hat{\sigma}_\varepsilon^2)^{n/2}} \exp\left(\frac{-(\vec{y} - X\hat{\beta}_{MLE})^T(\vec{y} - X\hat{\beta}_{MLE})}{2\hat{\sigma}_\varepsilon^2}\right)$ $\frac{\vec{e}'\vec{e}}{n} = \hat{\sigma}_\varepsilon^2$ for MLE.

$= (2\pi)^{-n/2}(\frac{1}{n}\vec{e}'\vec{e})^{-n/2} \exp\left(\frac{-\vec{e}'\vec{e}}{2\frac{1}{n}\vec{e}'\vec{e}}\right) = (2\pi)^{-n/2}\exp(-n/2)(\frac{1}{n}\vec{e}'\vec{e})^{-n/2}$ (c)

$L_{H_0}(\vec{y}) = c(\frac{1}{n}\vec{e}_q'\vec{e}_q)^{-n/2} = c(\frac{1}{n}RSS_q)^{-n/2}$

LRT is thus, $\lambda_{LRT} = \frac{\mathcal{L}_{H_0}(\vec{y})}{\mathcal{L}_{H_A}(\vec{y})} = \left(\frac{\vec{e}'\vec{e}}{\vec{e}_q'\vec{e}_q}\right)^{n/2}$ ($\vec{e}'\vec{e}$ = RSS, $\vec{e}_q'\vec{e}_q$ = RSS_q) $= (\lambda_{LRT})^{2/n}$

* Fact: If λ is an LRT, and $f(\lambda)$ is monotone then $f(\lambda)$ is an LRT.

→ use the fact above to know that F is an LRT for the H_0.

Recall that for F-test: $F_{q,n-p-1} = \frac{(RSS_q - RSS)/q}{RSS/n-p-1}$

$\Rightarrow \frac{q}{n-p-1} F_{q,n-p-1} = \frac{RSS_q}{RSS} - \frac{RSS}{RSS}$

$\Rightarrow \left(1 + \frac{q}{n-p-1}F_{q,n-p-1}\right)^{-1} = \frac{RSS}{RSS_q} = \lambda_{LRT}$

$\Rightarrow$ (invert) $\frac{n-p-1}{q}\left[(\lambda_{LRT,2})^{-1} - 1\right] = F_{q,n-p-1}$

Thus, $F_{q,n-p-1}$ is an LRT for this hypothesis.

	H_0 is True	H_A is True
Reject H_0	Type I Error False Positive α (alpha)	Correct True Positive 1-β (power)
Fail to reject H_0	Correct True Negative 1-α	Type II Error False Negative β (beta)

- Benefit of F-test over LRT
  - we know exact dist of F_q $\{F(q, n-p-1)\}$
  - λ_{LRT} has asymptotic dist. $\{-2\ln\lambda_{LRT} \xrightarrow{n\to\infty} \chi^2_q\}$

General Linear Hypothesis

Even more generally, we can test the linear hypothesis

$$H_0\colon \underset{(q\times k+1)}{\mathbf{L}}\ \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$

where $\mathbf{L}$ and $\mathbf{c}$ contain prespecified constants, and the *hypothesis matrix* $\mathbf{L}$ is of full row rank $q \le k+1$. The resulting F-statistic,

$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\mathbf{b} - \mathbf{c})}{qS_E^2} \quad (9.17)$$

follows an F-distribution with q and $n-k-1$ degrees of freedom if H_0 is true.

To understand the structure of Equation 9.17, recall that $\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}]$. As a consequence,

$$\mathbf{L}\mathbf{b} \sim N_q[\mathbf{L}\boldsymbol{\beta}, \sigma_\varepsilon^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']$$

Under H_0, $\mathbf{L}\boldsymbol{\beta} = \mathbf{c}$, and thus

$$(\mathbf{L}\mathbf{b} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{L}\mathbf{b} - \mathbf{c})/\sigma_\varepsilon^2 \sim \chi^2_q$$

ex) $L: \begin{bmatrix}0&1&0\\0&0&1\end{bmatrix}$; $C = \begin{bmatrix}0\\0\end{bmatrix}$ tests $B_1 = B_2 = 0$

$L = [0\ 1\ 1]$; $C = [0]$ tests $B_1 + B_2 = 0$

Joint Confidence Region

- inversion of F test:

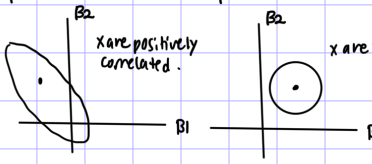
1) Let $\vec{\beta}_q := [\beta_1, \ldots, \beta_q]^T$ be our params of interest. If $\vec{\beta}_q^{(0)}$ is in CI, then we should be able to reject $H_0: \vec{\beta}_q = \vec{\beta}_q^{(0)}$ at level α.

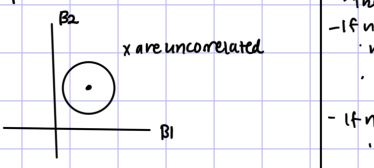
2) If we apply general F test for $H_0: \vec{\beta}_q = \vec{\beta}_q^{(0)}$, we get

$$F_{(0)} = \frac{(\hat{\vec{\beta}}_q - \vec{\beta}_q^{(0)})^T V_q^{-1}(\hat{\vec{\beta}}_q - \vec{\beta}_q^{(0)})/q}{RSS/(n-p-1)}$$

where V_q (not inverted) are rows/columns of $(X^TX)^{-1}$ corresponding to X_q.

3) Let F_α be the critical value at α for the test in 2). If $F_{(0)} \le F_\alpha$, then $\vec{\beta}_q^{(0)}$ is in joint CI for $\vec{\beta}_q$.





Dummy Variable Regression

$$X = \begin{pmatrix}1&0\\1&0\\1&0\\0&1\\0&1\end{pmatrix} \quad X^TX = \begin{pmatrix}n_S&0\\0&n_L\end{pmatrix} \quad (X^TX)^{-1} = \begin{pmatrix}1/n_S&0\\0&1/n_L\end{pmatrix} \quad X^Ty = \begin{pmatrix}\sum_S y\\ \sum_L y\end{pmatrix}$$

Then, $\hat{\beta} = \begin{pmatrix}1/n_S&0\\0&1/n_L\end{pmatrix}\begin{pmatrix}\sum_S y\\ \sum_L y\end{pmatrix} = \begin{pmatrix}\bar{y}_S\\ \bar{y}_L\end{pmatrix}$

An intercept and single dummy gives us the same info.

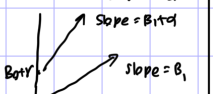
$$X = \begin{pmatrix}1&0\\1&0\\1&0\\1&1\\1&1\end{pmatrix} \quad X^TX = \begin{pmatrix}n&n_L\\n_L&n_L\end{pmatrix} \quad (X^TX)^{-1} = \begin{bmatrix}1/n_S&-1/n_S\\-1/n_S&1/n_S+1/n_L\end{bmatrix} \quad X^Ty = \begin{bmatrix}\sum y\\ \sum_L y\end{bmatrix}$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{bmatrix}1/n_S&-1/n_S\\-1/n_S&1/n_S+1/n_L\end{bmatrix}\begin{bmatrix}\sum y\\ \sum_L y\end{bmatrix} = \begin{bmatrix}\bar{y}_S\\ \bar{y}_L - \bar{y}_S\end{bmatrix} = \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}$$

Interaction terms

$\vec{y} = \beta_0 + \beta_1\vec{x} + \gamma\vec{d} + \delta(\vec{x} * \vec{d}) + \varepsilon$

→ interaction allows different slope with respect to the continuous covariate.



Categorical Variables

$$X = \begin{pmatrix}E1&E2&E3\\1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\end{pmatrix}$$

These columns are orthogonal to each other.

$\hat{y} = \sum_{j=1}^{p} \vec{x}_j \frac{\vec{y}'\vec{x}_j}{\|\vec{x}_j\|^2}$

$\hat{y}_i = \sum_{\text{energy}=2} y_i/n_{\text{energy}=2}$

with intercept:

$$X = \begin{pmatrix}\text{intercept}&E2&E3\\1&0&0\\1&0&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\end{pmatrix}$$

(Hyp. testing)

$\hat{\beta}_0 = \bar{y}_{\text{energy}1}$

$\hat{\beta}_1 = \bar{y}_{\text{energy}2} - \bar{y}_{\text{energy}1}$

$\hat{\beta}_2 = \bar{y}_{\text{energy}3} - \bar{y}_{\text{energy}1}$

If we want to test $\bar{y}_1 = \bar{y}_2 = \bar{y}_3$, then we run subset F test against intercept only model.

= principle of marginality

```
tapply(canada_occ$prestige, canada_occ$type0, mean)

##       bc       wc     prof
## 35.52727 42.24348 67.84839
```

When we fit these means with lm, the

```
lm.t = lm(prestige ~ type, data = canada_occ)
summary(lm.t)

##
## Call:
## lm(formula = prestige ~ type, data = canada_
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2273  -7.1773  -0.0854   6.1174  25.2565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.527      1.432  24.810  < 2e-16 ***
## typeprof      32.321      2.227  14.511  < 2e-16 ***
## typewc         6.716      2.444   2.748  0.00718 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.499 on 95 degrees of freedom
## Multiple R-squared:  0.6976,  Adjusted R-squared:  0.6913
## F-statistic: 109.6 on 2 and 95 DF,  p-value: < 2.2e-16
```

difference from the avg of BC

job

→ In categorical variable, if you want to test whether the variable should be in the model or not, then we examine several coefficients. Thus, incremental F-test is useful.

① income by type interaction.

```
## Analysis of Variance Table
##
## Model 1: prestige ~ income + education + type + education * type
## Model 2: prestige ~ income + education + type + income * type + education *
##     type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     91 4504.6
## 2     89 3552.9  2    951.77 11.921 2.588e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

② income by type interaction.

```
## Analysis of Variance Table
##
## Model 1: prestige ~ income + education + type + income * type
## Model 2: prestige ~ income + education + type + income * type + education *
##     type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     91 3791.3
## 2     89 3552.9  2    238.4 2.9859 0.05557 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bootstrapping

① Nonparametric for $\vec{\beta}_j$ (↳ call this bootstrap sample b.)

- resample our actual data randomly with replacement.
- Compute $\hat{\beta}_{j,b}$ [$\hat{\beta}_j$ for sample b]
- repeat B times, yielding $\{\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,B}\}$

$\{\hat{\beta}_{j,1}, \ldots, \hat{\beta}_{j,B}\}$ is empirical distribution of the sample statistics $\hat{\beta}_j^*$.

→ Based on the assumption that the sample has approximately the same distribution as the population.

② parametric Bootstrap [assumes 1) linearity in X, 2) Normal error]

① run OLS on original data, yielding $\hat{\beta}$. (If model is correct then $\vec{y} \sim N(X\hat{\beta}, \hat{\sigma}_\varepsilon^2 I_n)$)

② Generate n random errors from $N(0, \hat{\sigma}_\varepsilon^2 I_n) := \vec{\varepsilon}_b$

③ $\vec{y}_b = X\hat{\beta} + \vec{\varepsilon}_b$

④ Regress $\vec{y}_b$ on X to yield $\hat{\beta}_{j,b}^*$

⑤ Repeat B times to yield $\{\hat{\beta}_{j,1}^*, \ldots, \hat{\beta}_{j,B}^*\}$

③ semiparametric Bootstrap [doesn't assume that errors are normal.]

1) original model yields residuals and predictions: $\vec{e}$ and $\hat{y}$.

2) Sample n residuals (with replacement) from $\vec{e}$, yielding $\vec{e}_b$

3) $\vec{y}_b := \hat{y} + \vec{e}_b$

4) regress $\vec{y}_b$ on X to yield $\hat{\beta}_{j,b}^*$

5) repeat B times.

In General, for inference on $\vec{\beta}_j$:

- If n is small:
  - nonparametric bootstrap is bad
  - theory based inference is bad if errors aren't normal
- If n is moderate: (10X predictors)
  - nonparametric is generally ok.
  - If assumptions are correct, then theory/parametric bootstraps are more powerful.
- If n is large:
  - Doesn't matter
  - CLT tells us that $\hat{\beta}_j$ is approximately normal.

If linear model is misspecified, then:

- X is missing confounders
- $\vec{y}$ is not linear in X.

Then, inference is always bad!

Bootstrap F tests

Consider a general setting where a vector of observations $\mathbf{y}$ is sampled iid from some population with an underlying parameter θ, and suppose we have an estimate of this parameter $\hat{\theta}$. To conduct a one-sided hypothesis test using the bootstrap, we use the following recipe:

1. Specify a null hypothesis which sets some parameter in your model to a known value θ_0.
2. Choose a test statistic $f(\hat{\theta}; \theta_0)$ that describes (qualitatively) how far your estimate "deviates" from the null value θ_0, where larger values indicate larger deviation from the null.
3. Draw bootstrap samples $\mathbf{y}_1^*, \ldots, \mathbf{y}_B^*$ from the sample with replacement.
4. From these bootstrap samples compute bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
5. Compute bootstrap test statistics $f(\hat{\theta}_1^*; \hat{\theta}), \ldots, f(\hat{\theta}_B^*; \hat{\theta})$. Here we have followed the bootstrap principle and substituted hat-star quantities for hat quantities and hat quantities for parameters.
6. Compare the observed value of the test statistic $f(\hat{\theta}; \theta_0)$ to the bootstrap distribution of the test statistic $f(\hat{\theta}_1^*; \hat{\theta}), \ldots, f(\hat{\theta}_B^*; \hat{\theta})$. P-values can be computed directly by counting the proportion of bootstrap values that exceed the observed value.

The most common place you'll want to use this recipe in linear modeling is to conduct incremental F-tests comparing larger and smaller models, i.e. with null hypotheses of the form $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ where $k \le p$. Here the parameter θ is just β, and the test statistic is the usual incremental F-statistic:

$$F = \frac{\frac{1}{k}(RegSS_{full} - RegSS_{small})}{\frac{1}{n-p-1}ErrSS_{full}} \quad (1)$$

We can instead set this up as a test of the general linear hypothesis $L\beta = \beta_0$ where β_0 is a $k \times 1$ vector of 0s and L is a $k \times (p+1)$ matrix with form:

$$L = \begin{pmatrix}\mathbf{0}_{k\times 1} & \mathbf{I}_{k\times k} & \mathbf{0}_{k\times(p-k+1)}\end{pmatrix}$$

Remember, this is just a matrix way to write the condition that the first k coefficients (excluding the intercept) are all equal to zero). This gives us an alternate formula for the test statistic (1) as follows:

$$F = \frac{\frac{1}{k}\left\{(L\hat{\beta} - \beta_0)^T\left[L(X^TX)^{-1}L\right]^{-1}(L\hat{\beta} - \beta_0)\right\}}{\frac{1}{n-p-1}ErrSS_{full}} \quad (2)$$

next pg.

ANOVA: exploring the significance of terms in categorical variables.

→ anova(·) two models yields incremental F test results between two models

→ Anova(·) on the full model returns a Type II ANOVA

(where for each categorical variable in the model, consider incremental F-test of removing only that variable.)

Lastly, as an example, if we want to test whether the main effect for type is 0, then the model contains no interaction terms with type. That is, according to the principle of marginality, we compare the following two models:

* keep all related main effects if interaction is significant.

```
lm.no.type = lm(prestige ~ income + education,
                data = canada_occ[!is.na(canada_occ$type), ])
lm.main = lm(prestige ~ income + education + type, data = canada_occ)
anova(lm.no.type, lm.main)
```

p-value = 0.00396

→ rejects the test that the coefficients for occupation type is 0.

One last comment, notice the code `data = canada_occ[!is.na(canada_occ$type), ]`. Here we are dropping the rows that have missing values for type from the first model (which does not include type). If we do not drop these rows then we get an error because the models are fitted with differing numbers of observations.

Rule of Thumb: with only 98 observations, we want to be careful to not overfit. A rule of thumb is to keep the number of variables to be below 1/10 of the number of observations. Another check is to examine the standard errors of the coefficients:

With these things in mind, let's review the bootstrap hypothesis testing recipe for the case of the F-test.

1. Specify $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ or equivalently $L\beta = \beta_0 = \mathbf{0}_k$.
2. Choose a test statistic

$$F(\widehat{\beta}; \beta_0) = \frac{\frac{1}{k}\left\{(L\widehat{\beta} - \beta_0)^T \left[L(X^T X)^{-1} L\right]^{-1} (L\widehat{\beta} - \beta_0)\right\}}{\frac{1}{n-p-1} ErrSS_{full}}$$

that is large when the observed data deviates far from what is expected under H_0.

3. Draw bootstrap samples $\mathbf{y}_1^*, \ldots, \mathbf{y}_B^*$ from the sample with replacement. You can do this either by bootstrapping pairs (in which case you will also get bootstrapped design matrices X^*) or by bootstrapping residuals.
4. Fit the regression in each bootstrap sample and obtain bootstrap coefficient vectors $\widehat{\beta}_1^*, \ldots, \widehat{\beta}_B^*$.
5. Compute bootstrap F-statistics:

$$F(\widehat{\beta}_j^*; \widehat{\beta}) = \frac{\frac{1}{k}\left\{(L\widehat{\beta}_j^* - \widehat{\beta})^T \left[L(X^{*T} X^*)^{-1} L\right]^{-1} (L\widehat{\beta}_j^* - \widehat{\beta})\right\}}{\frac{1}{n-p-1} ErrSS^*_{full,\ j}} \quad (3)$$

for bootstrap samples $j = 1, \ldots, B$. Notice that we have followed the bootstrap principle and substituted hat-star quantities for hat quantities (as well as X^* for X if bootstrapping pairs) and hat quantities for parameters; this results in a test statistic that looks like it is testing the general linear hypothesis $L\widehat{\beta}_j^* = \widehat{\beta}$, although it's really just the bootstrap version of the usual incremental F-test..

6. Compare the observed value of the test statistic F to the bootstrap distribution of the test statistic $F_1^*, \ldots, F_B^*$ computed in the previous step. P-values can be computed directly by counting the proportion of bootstrap values that exceed the observed value. Note that although we're using F-statistics for convenience we never look at an F-distribution but rely on the bootstrap to get the sampling distribution.

Solution: *To test for whether the observation is an outlier, we use the fact that the studentized residuals of a linear regression model are t-distributed under the standard normal error assumptions. First we construct the studentized residual for observation 42.*

$$e^*_{42} = \frac{e_i}{s_{e_{(-42)}}\sqrt{1 - h_{42}}}$$

The raw residual e_i is given by the output from the first regression model; clearly it is the largest negative residual, which is listed as -18.4788. Plugging in the leverage $h_{42} \approx 0.68$ and the residual standard error from the regression not involving observation 42 — 5.406 from the second regression output — we have

$$e^*_{42} = -18.7488/(5.406\sqrt{0.32}) \approx -6.1309$$

This studentized residual will be t-distributed with $n-p-2 = 192$ degrees of freedom under the null hypothesis (the same as the error degrees of freedom in the second regression model). Letting t_{192} be a random variable with this distribution, we could report the two-sided p-value $2 \cdot P(t_{192} > |-6.1309|)$. However, this would not be quite right since we have selected this residual because of its unusually extreme value. To avoid invalid testing due to this selection step, we use a Bonferroni correction and multiply the p-value by the number of data points in the original regression. Since the model has 7 parameters and 193 error degrees of freedom, there are 200 observations. So the correct p-value for this test is approximately $400 \cdot P(t_{192} > 6.1309)$.

weighted least squares

$X \in \mathbb{R}^{n \times (p+1)}$, $\vec{y} \in \mathbb{R}^n$

$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$ where $\vec{\varepsilon} \sim N(\vec{0}, \sigma_\varepsilon^2 \begin{bmatrix} 1/w_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/w_n^2 \end{bmatrix}$

and let $W^{-1} = \begin{bmatrix} 1/w_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/w_n^2 \end{bmatrix} = \text{diag}\left(\frac{1}{w_1^2}, \ldots, \frac{1}{w_n^2}\right)$

$\Rightarrow W = \text{diag}(w_1^2, \ldots, w_n^2) \Rightarrow W^{1/2} = \text{diag}(w_1, \ldots, w_n)$

$W^{1/2}\vec{y} = W^{1/2}(X\vec{\beta} + \vec{\varepsilon})$

Then, $\hat{\beta}_{WLS} = \underset{\vec{\beta}}{\text{argmin}} \| W^{1/2}\vec{y} - W^{1/2}X\vec{\beta}\|_2^2 = (X^TWX)^{-1}X^TW\vec{y}$

- $\hat{\beta}_{WLS}$ is more efficient than $\hat{\beta}_{OLS}$. (has lower MSE)
- In practice, we don't know the weights. so, we fit the parametric model to weights themselves. → FGLS

It is possible that W is unrelated with X but,

- we cannot estimate w_i^2 [variance] just from a single residual, e_i^2
- we only assume $\text{var}(\vec{\varepsilon}) \perp X$ in order to compute $\text{var}(\hat{\beta})$ b/c. it doesn't affect the validity of point estimate or its unbiased property.

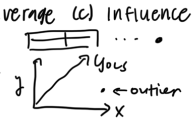
3. Regress $\vec{y}$ on X, get $\vec{e}_n$.
2. Regress $\log(\vec{e}_n^2)$ on X to yield fitted values, which we call $\hat{g}_1, \ldots, \hat{g}_n$.
3. $\hat{h}_i := \exp(\hat{g}_i)$; $\hat{W}^{-1} = \begin{bmatrix} \hat{h}_1 & & \\ & \ddots & \\ & & \hat{h}_n \end{bmatrix}$, and analogously, $\hat{W} = \begin{bmatrix} 1/\hat{h}_1 & & \\ & \ddots & \\ & & 1/\hat{h}_n \end{bmatrix}$. This is because $\hat{h}_i$ is the estimator for $Var(\epsilon_i) = \sigma_\varepsilon^2 / w_i^2$.
4. $\hat{\beta}_{FGLS} = (X^\top \hat{W} X)^{-1} X^\top \hat{W} \vec{y}$. For inference, we will use the plug-in estimator of the standard error, $se(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 \left(X^\top \hat{W} X\right)^{-1}$. Since we have incorporated all of $\hat{h}_i$ into each $\hat{w}_i^2$, it can also suffice to just set $\hat{\sigma}_\varepsilon^2 = 1$.

↳ double usage of X in the above estimation is problematic for the bias of $\hat{\beta}_{FGLS}$, but the estimator is still consistent and asymptotically efficient if in fact the errors have different variances as function of X.

model diagnostics

(a) outliers (b) leverage (c) influence

Outlier: univariate, multivariate



response y_i is conditionally unusual given x_i.

- residuals don't have equal variances
  ↳ $V(E_i) = \sigma_\varepsilon^2(1 - h_i)$
  $\Sigma_{\vec{e},\vec{e}} = \sigma_\varepsilon^2 Q$ ← values along the diagonal are not equal.

↓ Fix: standardized residuals: $E_i' = \frac{E_i}{S_E\sqrt{1-h_i}}$

issue: e_i and $\hat{\sigma}_\varepsilon$ are not independent, preventing E_i' from following t-distribution.

$\tilde{e}_i := \frac{e_i}{\hat{\sigma}_\varepsilon\sqrt{(1-h)_{ii}}} = \frac{e_i}{\hat{\sigma}_\varepsilon\sqrt{1-h_{ii}}}$

↳ Fix: studentized residuals.

Suppose, however, that we refit the model deleting the ith observation, obtaining an estimate $S_{E(-i)}$ of σ_ε that is based on the *remaining* $n-1$ observations. Then the *studentized residual*

$$E_i^* \equiv \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}} = \tilde{e}_i\sqrt{\frac{n-(p+1)-1}{n-(p+1)-\tilde{e}_i^2}} \quad (11.1)$$

has an independent numerator and denominator and follows a t-distribution with $n-k-2$ degrees of freedom.

$\tilde{e}_i$: standardized residuals.

procedure:

① refit mean-shift model n times, once for each observation, producing studentized residuals $E_1^*, \ldots, E_n^*$.

② usually, our interest focuses on the E^*_{max}. However, because we have picked the biggest of n test statistics, it is not legitimate to find p-value of E^*_{max}.

③ To address issue ②, we do Bonferroni adjustment to the p-value for the largest absolute E_i.

- If we have only one residual, we set up: $p_{t_{n-p-2}}(|e_i^*| > |e^*_{obs}|)$
- If we have n residuals, then compute $p_{t_{n-p-2}}(|e_i^*| > |e^*_{max}|)$ compare to $\frac{\alpha}{n} = \frac{0.05}{n}$

Leverage

idea: is $\vec{x}_i^T$ (ith row of X) unusual compared to other $\vec{x}_j^T$'s?

leverage $(\vec{x}_i^T) = h_{ii}$ ← ith diagonal of H: $X(X^TX)^{-1}X^T$

properties of H: ① symmetric ② idempotent ③ rank = p+1

④ $\frac{1}{n} \le h_{ii} \le 1$ ⑤ $\frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{p+1}{n}$

- why is leverage a useful Quantity

$\hat{\vec{y}} = H\vec{y} \Rightarrow \hat{y}_i = \sum_{j=1}^n h_{ij}y_j$ * h_{ij}: weights for how much y_j determines $\hat{y}_i$.

→ if h_{ii} is large, then the line near y_i is almost entirely determined by y_i.

In simple regression: * tells us if x_i is far from $\bar{x}$ relative to the full data.

$w_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$ → $n h_{ii} = 1 + \frac{(x_i - \bar{x})^2}{\frac{1}{n}\sum_{j=1}^n (x_j - \bar{x})^2}$

Influence we want to measure how much our estimates (β or $\hat{\sigma}_\varepsilon^2$) would change if we remove datapoint i.

$\hat{\beta}_{(-i)}$: OLS estimate with ith data point removed.

Q. why not we just do influence $_i = \|\hat{\beta} - \hat{\beta}_{(-i)}\|^2$?

A. This is treating β as in the same scale but each β may have different units.

[Solution] Cook's distance.

Let $\mathbf{b}_{(-i)}$ denote the vector of least-squares regression coefficients calculated with the ith observation omitted. Then, $\mathbf{d}_i \equiv \mathbf{b} - \mathbf{b}_{(-i)}$ represents the influence of observation i on the regression coefficients. The influence vector $\mathbf{d}_i$ can be calculated efficiently as[39]

$$\mathbf{d}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' \frac{E_i}{1 - h_i} \quad (11.8)$$

where $\mathbf{x}_i'$ is the ith row of the model matrix $\mathbf{X}$ (and $\mathbf{x}_i$ is the ith row written as a column vector).

Cook's D_i is the F-statistic for testing the "hypothesis" that $\beta = \mathbf{b}_{(-i)}$:

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{(-i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(-i)})}{(k+1)S_E^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{(k+1)S_E^2}$$

An alternative interpretation of D_i, therefore, is that it measures the aggregate influence of observation i on the fitted values $\hat{\mathbf{y}}$. This is why Belsley et al. (1980) call their similar statistic "DFFITS." Using Equation 11.8,

$$D_i = \frac{E_i^2}{S_E^2(k+1)} \times \frac{h_i}{(1-h_i)^2} = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

which is the formula for Cook's D given in Section 11.4.

- cook's distance is considered high if it is greater than 0.5 and extreme if it is greater than 1.

An influence measure proposed by Belsley et al. (1980) closely approximates the squared ratio of volumes of the deleted and full-data confidence regions for the regression coefficients:[18]

$$\text{COVRATIO}_i = \frac{1}{(1-h_i)\left(\frac{n-k-2+E_i'^2}{n-k-1}\right)^{k+1}}$$

Observations that *increase* the precision of estimation have values of COVRATIO that are larger than 1; those that *decrease* the precision of estimation have values smaller than 1. Look for values of COVRATIO, therefore, that differ considerably from 1.

- covratio > 1 : i is not extreme
- covratio < 1 : i is extreme enough that $\hat{\sigma}_\varepsilon^2$ shrinks when it is excluded.

Assumption diagnostics

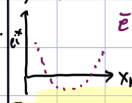
our full model reads $\vec{y} \sim N(X\beta, \sigma_\varepsilon^2 I_n)$

Assumptions are 1) linearity $E[\vec{\varepsilon}] = 0$ 3) Homoscedasticity 2) independence of ε_i 4) normality

① Mean zero errors

check ✓ studentized residuals plotted against:

1) columns of X 3) order of data / time of collection
2) variables not in the model



$\vec{e} = 0$ but the assumption is $E[\varepsilon_i | X_i] = 0$ regardless of X.

③ Nonconstant variance (Heteroscedasticity)

$\text{var}(\varepsilon_i | X) = \sigma_\varepsilon^2 \ \forall i$

check ✓ plot residuals vs fitted values (or vs covariate)

Q why not plot against $\vec{y}$?

$\vec{y} = \hat{\vec{y}} + \vec{e}$, and $\text{corr}(\hat{y}_i, \vec{e}) = \sqrt{1-R^2}$

→ It will be odd to study the correlation between $\vec{y}$ and $\vec{e}$ that are inherently correlated to each other. But, by construction, $\hat{\vec{y}}$ and $\vec{e}$ are orthogonal.

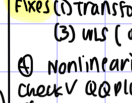
Q. why prefer plot studentized over e_i?

A the residuals have unequal variance $\{\sigma_\varepsilon^2 Q\}$ even though the random errors might not. In a sense, we are removing the variance just due to Q matrix.

Fixes (1) transformation (2) (Non parametric) Bootstrap (3) WLS (or other standard error procedure)

④ Nonlinearity $\Phi^{-1}\left(\frac{\text{rank}(|e_i^*|)}{n}\right)$

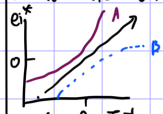
Check ✓ QQ plot



$\cdot \Phi^{-1}(\cdot) : [0,1] \to \mathbb{R}$

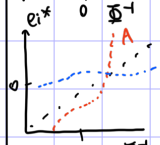
$\cdot \text{rank}|e_i^*| = 1$ if $|e_i^*|$ is smallest.

Problematic QQ plots



A: right skewed: residuals larger than expected on the right tail

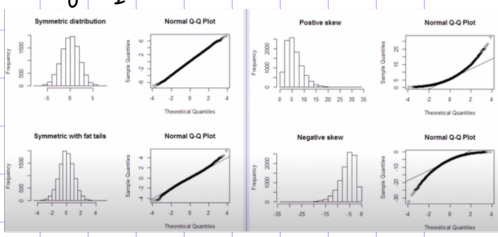
B: Left skewed: residual larger than expected on the left tail



A: large tailed: more data in the extremes.

B: small tailed: more data in the center

* things will be flipped if Φ^{-1} is in an X axis.



Fixes 1) transformations 2) Bootstrap 3) Assume new error distribution

Model selection

Let's say we have p numeric covariates.

How many parameters for including:

(i) squared covariates: p more
(ii) pairwise interactions: $\binom{p}{2}$ more
(iii) three way interactions: $\binom{p}{3}$ more

Simple / parsimonious is good → interpretable → limit future data collection → overfitting concerns.

Issues with ANOVA F test

1. Assumption based
2. significance based tool ≠ predictive importance
3. paired model comparisons
4. multiple testing problems

goal: estimate $E[(y_{new} - \hat{y}_{new})^2]$

Model → criterion → Quality score

Common criteria 1) Adj R^2, 2) cross validation Errors, 3) AIC/BIC, 4) C_p

each of these criteria help manage the bias variance trade off.

1) Adjusted R squared

Ideally, we want to know the error our model will have in predicting unseen data. Adjusted R^2 heuristically penalizes the additional variance that results from extra covariates

$\text{Adj } R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\frac{RSS}{TSS} = 1 - \frac{n-1}{n-p-1}(1-R^2)$

* But, adjusted R^2 is not an actual theoretical approximation of $E[(y_{new} - \hat{y}_{new})^2]$. It is just a heuristic.

2) cross validation

- cycles through the entire dataset and ensure that every element of $\vec{y}$ is used as hold out data exactly once.

(i) Leave one out method $\widehat{MSE}(\hat{y}_{new}, y_{new}) = \frac{1}{n}\sum(y_i - x_i^T\hat{\beta}_{(-i)})^2 = \frac{1}{n}\sum\left(\frac{e_i}{(-h_{ii})}\right)^2$ for OLS only.

(ii) K Folds: hold out random 1/K fraction of data at a time.

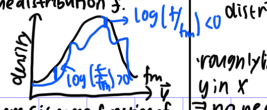
③ AIC / BIC

- AIC/BIC based on KL-divergence.

* $D_{KL}(f, f_{(m)}) = \int \log\left(\frac{f(y)}{f_m(y)}\right) f(y)\,dy$

true density of $(X, \vec{y})$; density under model m.

D_{KL} is a principled way of measuring distance from the distribution f.



Fact: $D_{KL}(f, f_{(m)}) = C - \mathbb{E}[\log(f_m(y))]$ where C is some function of $f(y)$

pf $D_{KL}(f, f_{(m)}) = \int \log\left(\frac{f(y)}{f_m(y)}\right) f(y)\,dy = \int \log(f(y)) f(y)\,dy - \int \log(f_m(y)) f(y)\,dy$

So if we can estimate $\mathbb{E}[\log(f_m(y))] = C - \mathbb{E}[\log(f_m(y))]$ then we can compare diff models.

- $\mathbb{E}[\log(f_m(y))] \approx \frac{1}{n}\sum \log(f_m(y_i)) = \frac{1}{n}\log \pi(f_m(y_i)) = \frac{1}{n}\log L(\hat{\beta}_{(m)} | y_1, \dots, y_n) \overset{n\to\infty}{\approx} \frac{1}{n}\log L(\hat{\beta}_{(m)} | y_1, \dots, y_n)$

↳ 2 approximations.

we've used the same data to approximate the initial expectations & model likelihood. This overstates the expected log-likelihood of f_m under f.

[Suggests that min KL ≈ max likelihood]

- posinve bias of $\log L(\hat{\beta}_{(m)} | y_1, \dots y_n)$ is $p_{(m)} + 1$.
- Thus, $\underset{m}{\text{argmin}} \{D_{KL}(f, f_{(m)})\} = \underset{m}{\text{argmin}} \{C - \mathbb{E}[\log(f_m(y))]\}$

$= \underset{m}{\text{argmin}} \{-\mathbb{E}[\log(f_m(y))]\} \approx \underset{m}{\text{argmin}} \{-\log L(\hat{\beta}_{(m)} | \vec{y}) + (p_{(m)} + 1)\}$

$:= \frac{1}{2} AIC(m)$

→ $AIC = -2\log \mathcal{L}(\hat{\beta}_{(m)} | \vec{y}) + 2(p_{(m)} + 1)$

→ $BIC = -2\log \mathcal{L}(\hat{\beta}_{(m)} | \vec{y}) + (\log(n))(p_{(m)} + 1)$ → parsimonious.

- Lower AIC, BIC means better fit.

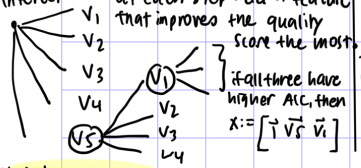
Tradeoffs of using AIC and/or BIC (vs. cross validation)

- only train each model once
- use all the data
- are based on a likelihood model and can be difficult to compare across model types.

"All subsets" Approach computes 2^k models (if k total features.)

- one way to lower the total number of models tested is to use Forward / Backward selection

Forward



- at each step add a feature that improves the quality score the most.
- if all three have higher AIC, then $X := [\vec{1}\ \vec{v}_5\ \vec{v}_i]$

Forward & Backward is greedy search. It doesn't gaurantee we pick an optimal model.
we can lower variance by inducing some bias.

Shrinkage methods:

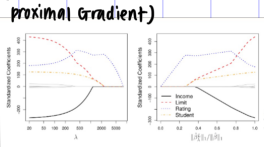
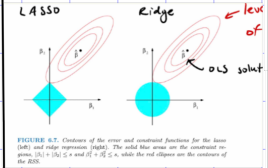
① Preprocessing : is needed b/c
1) we exclude intercept from variable selection
2) the variables are on different scales.

1) Standardization: $X \to Z = \left[\frac{\vec{x}_1 - \bar{x}_1\vec{1}}{s_{x_1}} \dots \frac{\vec{x}_p - \bar{x}_p\vec{1}}{s_{x_p}}\right]$

$\vec{y} \to \vec{y} - \bar{y}\vec{1} := \vec{y}^*$

2) Now regress $\vec{y}^*$ on Z.

3) To shrink the parameters, we place constraints on the least squares optimization

Ridge	Lasso
$\min_{\vec{\beta}} \|\vec{y}^* - Z\vec{\beta}\|^2 + \lambda\sum_{i=1}^{p} \|\beta_i\|$	$\min_{\vec{\beta}} \|\vec{y}^* - Z\vec{\beta}\|^2 + \lambda\|\vec{\beta}\|^2$
1) Has closed form solution	1) Lasso Doesn't have closed form → but, since it is still convex, there is a single local optimum. There are fast numerical solvers (coordinate descent, iterative WLS, proximal Gradient)
$f(\vec{\beta}) = \|\vec{y}^* - Z\vec{\beta}\|_2^2 + \lambda\|\vec{\beta}\|_2^2$ $= (\vec{y} - Z\vec{\beta})^T(\vec{y} - Z\vec{\beta}) + \lambda\vec{\beta}^T\vec{\beta}$ $= \vec{y}^T\vec{y} - 2\vec{y}^TZ\vec{\beta} + \vec{\beta}^TZ^TZ\vec{\beta} + \vec{\beta}^T\lambda I_p\vec{\beta}$ $= \vec{y}^T\vec{y} - 2\vec{y}^TZ\vec{\beta} + \vec{\beta}^T(Z^TZ + \lambda I_p)\vec{\beta}$ $\frac{\partial f}{\partial \vec{\beta}} = 0 - 2Z^T\vec{y}^* + 2(Z^TZ + \lambda I_p)\vec{\beta}$ $\Rightarrow 2Z^T\vec{y}^* = 2(Z^TZ + \lambda I_p)\vec{\beta}$ $\Rightarrow \hat{\beta}_\lambda = (Z^TZ + \lambda I_p)^{-1}Z^T\vec{y}^*$	 FIGURE 6.6. The standardized lasso coefficients on the Credit data set are shown as a function of λ and $\|\hat{\beta}_\lambda\|/\|\hat{\beta}\|$.
2) Larger $\lambda \Rightarrow$ Smaller $\hat{\beta}$ $\Rightarrow$ smaller $\text{var}(\hat{\beta})$	→ Lasso coefficients shrink to 0.
• $\text{var}(\hat{\beta}_\lambda) = (Z^TZ + \lambda I_p)^{-1}Z^T\text{var}(y)[A]$ $= \sigma_\varepsilon^2(Z^TZ + \lambda I_p)^{-1}Z^TZ(Z^TZ + \lambda I_p)^{-1}$	 FIGURE 6.7. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $\|\beta_1\| + \|\beta_2\| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.
• $\text{Bias}(\hat{\beta}_\lambda) = \mathbb{E}[\hat{\beta}_\lambda] - \vec{\beta}$ $= (Z^TZ + \lambda I_p)^{-1}Z^T\mathbb{E}[\vec{y}^*] - \vec{\beta}$ $= (Z^TZ + \lambda I_p)^{-1}Z^TZ\vec{\beta} - \vec{\beta}$ diverges from $\vec{0}$ as λ gets larger.	
* there is ALWAYS some value $\lambda > 0$ s.t. the MSE of Ridge regression is lower than OLS. CA is unknown however	
PRO • analytical solution for each λ. • theoretically can always outperform OLS	PRO • variable selection • is convex/computationally feasible
CON • No explicit variable selection • No clear inference approach → OLS is preferable for causal inference since ridge coefficients have bias and don't come with CI.	CON • No linear solution to analyze • still no clear non-Bayesian inference
	* use CV to find the optimal λ.

# of regressors (excluding intercept) with nonzero coefficients in the fitted model associated with corresponding λ.

Generalized Linear Models (GLM)

- Framework that allow $\vec{y}$ to follow various parametric distribution:

Bernoulli	y_i is binary
Poisson, NB	$y \in \mathbb{Z}^+$ (y is count)
Exponential, Gamma	$y \in \mathbb{R}^+$
multinomial	y: count of occurence of k different outcomes

- roughly linear y in x
- ∄ no need of link $f(x)$.

- GLM recipe

① specify the conditional distribution of $y_i | \vec{x}_i^T$ (e.g. $y_i \sim$ Bernoulli(π_i))

② consider parameter of distribution in ①, their relation to $\mathbb{E}[y_i | x_i^T]$, and their domain (e.g. $y_i | \vec{x}_i^T \sim \text{Ber}(\pi_i)$, $\pi_i = \mathbb{E}[y_i | \vec{x}_i^T]$; $\pi_i \in (0,1)$)

③ Find link function $g(\cdot)$: that transforms $\mathbb{E}[y_i | x_i^T]$ to the range: $(-\infty, \infty)$.
(e.g. $y_i | \vec{x}_i^T \sim \text{Ber}(\pi_i)$, let $g(\cdot)$ be logit function $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
Then, $g: (0,1) \to (-\infty, \infty)$, as desired

④ Assume that $g(\pi_i)$ can be modeled linearly. $g(\pi_i) = \vec{x}_i^T\vec{\beta}$.
we then define inverse link function $g^{-1}(\cdot)$
(e.g. $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \vec{x}_i^T\vec{\beta}$

$\Rightarrow \frac{\pi_i}{1-\pi_i} = \exp(\vec{x}_i^T\vec{\beta})$ $\Rightarrow \pi_i = \exp(\vec{x}_i^T\vec{\beta}) - \pi_i\exp(\vec{x}_i^T\vec{\beta})$

$\Rightarrow \pi_i = \frac{\exp(\vec{x}_i^T\vec{\beta})}{1 + \exp(\vec{x}_i^T\vec{\beta})} = \frac{1}{1 + \exp(-\vec{x}_i^T\vec{\beta})}$

$:= g^{-1}(\vec{x}_i^T\vec{\beta}) := g^{-1}(\eta_i)$

Table 15.2 Canonical Link, Response Range, and Conditional Variance Function for Exponential Families

Family	*Canonical Link*	*Range of Y_i*	*$V(Y_i|\eta_i)$*
Gaussian	Identity	$(-\infty, +\infty)$	ϕ
Binomial	Logit	$\frac{0, 1, \dots, n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	Log	$0, 1, 2, \dots$	μ_i
Gamma	Inverse	$(0, \infty)$	$\phi\mu_i^2$
Inverse-Gaussian	Inverse-square	$(0, \infty)$	$\phi\mu_i^3$

NOTE: ϕ is the dispersion parameter, η_i is the linear predictor, and μ_i is the expectation of Y_i (the response). In the binomial family, n_i is the number of trials.

- Compare OLS with GLM:

① OLS: $\mathbb{E}[g(y_i)] = \vec{x}_i^T\vec{\beta}$
② GLM: $g(\mathbb{E}[y_i]) = \vec{x}_i^T\vec{\beta}$

- Now, the variance of y_i is often tied to its mean so it can fluctuate.
ex) Ber(π_i)'s variance: $\pi_i(1-\pi_i)$
Pois(λ_i)'s var: λ_i

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e\mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e\frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response, η_i is the linear predictor, and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

A generalized linear model (or GLM) consists of three components:

1. A random component, specifying the conditional distribution of the response variable, Y_i (for the ith of n independently sampled observations), given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of Y_i was a member of an exponential family, such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions.
2. A linear predictor—that is, a linear function of regressors,
$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$
3. A smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_i = E(Y_i)$, to the linear predictor:
$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

Summary of advantages of GLMs over traditional (OLS) regression

- We do not need to transform the response to have a normal distribution.
- The choice of link is separate from the choice of random component, giving us more flexibility in modeling.
- The models are fitted via maximum likelihood estimation, so likelihood functions and parameter estimates benefit from asymptotic normal and chi-square distributions.
- All the inference tools and model checking that we will discuss for logistic and Poisson regression models apply for other GLMs too; e.g., Wald and Likelihood ratio tests, deviance, residuals, confidence intervals, and overdispersion.
- There is often one procedure in a software package to capture all the models listed above, e.g. PROC GENMOD in SAS or glm() in R, etc., with options to vary the three components.

- The data $Y_1, Y_2, \dots, Y_n$ are independently distributed, i.e., cases are independent.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.).
- A GLM does NOT assume a linear relationship between the response variable and the explanatory variables, but it does assume a linear relationship between the transformed expected response in terms of the link function and the explanatory variables; e.g., for binary logistic regression $\text{logit}(\pi) = \beta_0 + \beta_1 x$.
- Explanatory variables can be nonlinear transformations of some original variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure.
- Errors need to be independent but NOT normally distributed.
- Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS).

Estimation & the data model for Bernoulli is now $y_i \sim \text{Ber}(g^{-1}(\vec{x}_i\vec{\beta}))$. This no longer is least squares problem (so it doesn't have closed form solutions.) But we can still write the likelihood and $\log L(\vec{\beta}, \vec{y}, X)$ is convex for many GLM and we can solve it. If dispersion (ϕ) must be estimated, we can do alternating MLE.

① Init: $t = 0$, $\hat{\phi}_0$
while not converged:
② $\hat{\beta}_{t+1} = \hat{\beta}_{MLE} \mid \hat{\phi}_t$
③ $\hat{\phi}_{t+1} = \hat{\phi}(\hat{\beta}_{t+1})$
④ $t = t+1$

prediction once we optimize the GLM for $\hat{\beta}$, prediction is straightforward.

$\hat{y}_i = \mathbb{E}[y_i | \vec{x}_i^T] = g^{-1}(\vec{x}_i^T\hat{\beta})$ [eg logistic regression] $\hat{y}_i = \frac{1}{1+\exp(-\vec{x}_i^T\hat{\beta})}$

Inference

- Model performance:

① residual Deviance (D_m): model with one parameter for each observation.

$D_m := 2(\log \mathcal{L}_S - \log \mathcal{L}_m)$

$\mathcal{L}_S$: likelihood of saturated model
$\mathcal{L}_m$: likelihood of our model.

* Fact: For logistic regression, $D_m = -2\log \mathcal{L}_m$

pf: In saturated model, we can fit the training logits perfectly. (i.e. if $y_i = 1$, then $g^{-1}(\vec{x}_i^T\hat{\beta}_{MLE}) = 1$)

$\mathcal{L}_S = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ ← Bernoulli likelihood = 1 ⇒ log-likelihood = 0.

$= 1^n = 1 \Rightarrow \log \mathcal{L}_S = 0 \Rightarrow D_m = -2\log \mathcal{L}_m$

- $D_m \overset{n\to\infty}{\sim} \chi^2(n-(p+1))$ ← params of null model; ↑ params of saturated model

→ D_m has $n-p-1$ df

- comparing D_m to Null Deviance (D_0):

→ D_0 is the residual deviance for an intercept only model

$\Rightarrow D_0 \underset{n\to\infty}{\sim} \chi^2_{(n-1)}$ (df)

- In general, with any main model w/ deviance D_1 and null model with a subset of parameters $\{\beta_1 = \dots = \beta_k = 0\}$ and deviance D_0, under the null:

$H_0 \Rightarrow \beta_1 = \beta_2 = \dots = \beta_k = 0$

we have $D_0 - D_1 \underset{n\to\infty}{\sim} \chi^2_{(p-k)}$ (unless ϕ is fixed.)

Natural Exponential Family. [not negative binom; gaussian, Binom, poisson, Gamma]
& pdf can be written:

$p(y_i; \theta_i, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$

- $\theta := g_c(\mathbb{E}[y])$ is some function of $\mathbb{E}[y]$ called canonical link

ex) binomial

$p(y; \theta; p) = \exp\left[\frac{y\theta - \log(1+\exp\theta)}{1/n} + \log\binom{n}{ny}\right]$ ($b(\theta)$, $c(y,\phi)$, $a(\phi)$)

Where $\theta = \log\left(\frac{p}{1-p}\right)$, n is known, ϕ is irrelevant.
→ canonical link is $\log\left(\frac{p}{1-p}\right)$.

① $\mu = b'(\theta)$, $\text{Var}(y) = a(\phi)b''(\theta)$

② log likelihood of data point y_i is:

$l_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \phi)$

③ $\eta_i := \vec{x}_i^T\vec{\beta} \Rightarrow g(\mu) = \eta_i$ ⑤ $\mathbb{E}[y_i] = \mu_i$

④ $\theta_i := g_c(\mu_i) := g_c(g^{-1}(\vec{x}_i^T\vec{\beta}))$ (canonical link; g is selected link fn.)

to find $\hat{\beta}_{MLE}$, we need to find the critical pt of the log likelihood:

$\frac{\partial l_i}{\partial \beta_j} = \frac{x_{ij}}{g'(\mu_i)} \frac{y_i - \mu_i}{a_i(\phi)b''(\theta_i)}$ for generically selected link function $g(\cdot)$.

Full model likelihood is:
$\log(\prod p(y_i)) = \sum_{i=1} l_i$

Thus, log likelihood gradient (score):

$S(\vec{\beta}) = \left[\sum_{i=1}^{n}\frac{\partial l_i}{\partial \beta_0}, \dots, \sum_{i=1}^{n}\frac{\partial l_i}{\partial \beta_p}\right] = X^TD(\vec{y} - \vec{\mu})$

where $D = \text{diag}\left(\left\{[g'(\mu_i)a_i(\phi)b''(\theta_i)]^{-1}, i = 1, \dots, n\right\}\right)$