Important limiting results:

Central Limit Theorem

- If $X_1, X_2, X_3 \dots X_n$ is an iid sample with ANY distribution w/ mean μ and standard deviation σ, then:

$$\bar{X} \xrightarrow{D} N\left(\mu, \frac{\sigma^2}{n}\right), \text{ as } n \to \infty$$

("as the sample size increases, the sample mean behaves as if it is from a Normal distribution with mean μ and standard deviation $(\sigma/\sqrt{n})$")

Maximum Likelyhood Estimate (MLE)

- Likelyhood function, $lik(\theta)$, corresponds to the probability of observing the particular data in our sample for various values θ, assuming the data came from some distribution w/ parameter θ.

Our goal is to find the value of θ that maximizes the likelyhood function.

The likelyhood function

$$\begin{aligned} lik(\theta) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n) \\ &\quad * \text{cause } X_i \text{ are iid} \\ &= \prod_{i=1}^{n} f_\theta(x_i) \\ &\quad * \text{since each } X_i \text{ have density } f_\theta) \end{aligned}$$

Computing MLE

- our goal is to compute the format of the parameter θ that maximizes the likelyhood function.

1. Find likelyhood $lik(\theta) = \prod_i^n f_\theta(X_i)$
2. Calculate the log of the likelyhood function (aka log likelyhood)
$$\ell(\theta) = \log(\prod f_\theta(x_i)) = \sum \log(f_\theta(x_i))$$
3. Differentiate in respect to θ, set to zero, and solve for θ.

Consistency:

- An estimate of $\hat{\theta}_n$, of θ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$

Consistency of MLE:

Theorem: the MLE $\hat{\theta}_{MLE,n}$, is a consistent estimator of the parameter, θ, that it is estimating, which means that:

$$\hat{\theta}_{MLE,\theta} \xrightarrow{p} \theta \text{ as } n \to \infty$$

"actual value of the parameter estimate approaches the true value of the parameter"

Asymptomatic unbiasedness:

$$E[\hat{\theta}_{MLE,n}] \to \theta \text{ as } n \to \infty$$

"the expected value of the parameter estimate approaches the true value of the parameter"

Asymptotic normality of the MLE:

- the MLE is asymptotically normal. If $\hat{\theta}_{MLE,n}$ is the ML estimate of a parameter θ whos true value is θ_0, then as $n \to \infty$, we have that:

$$\hat{\theta}_{ML,n} \xrightarrow{D} N\left(\theta_0, \frac{1}{nI(\theta_0)}\right) \xrightarrow{as} n \to \infty$$

$I(\theta_0)$ is the fisher infomation

$$\hookrightarrow -E\left[\frac{d^2}{d^2\theta}\log(f_\theta(X))\Big|_{\theta_0}\right]$$

$I(\theta_0)$ measures how "peaked" $\ell(\theta)$ is around θ_0. If this value is large, then it is easier to detect θ_0. $\therefore$ lower variance and vice versa (and should result the same distribution as CLT)

Cramer Rao Lower Bound

Theorem: if X_i are iid from a distribution with density f_θ, under smoothness conditions on f_θ, we have that:

If $\hat{\theta}$ is an unbiased estimator for θ, then: $\text{var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$ ← aka variance of MLE

This means the MLE has the LOWEST possible variance among unbiased estimators.

Efficiency:

Given two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is:

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\tilde{\theta})}{\text{var}(\hat{\theta})}$$

If $\text{eff}(\hat{\theta}, \tilde{\theta}) \leq 1$, then $\text{var}(\hat{\theta}) \geq \text{var}(\tilde{\theta})$, which implies that $\tilde{\theta}$ is less efficient than $\hat{\theta}$. Vice versa applies.

Efficient Estimators:

An unbiased estimator that achieves the cramer Rao lower bound is called efficient. $\text{var}(\hat{\theta}) = \frac{1}{nI(\theta)}$ ← cannot do any better than this

- MLE is asymptotically efficient, but not really for finite samples.

Confidence Intervals:

- 95% confidence intervals computed from iid random samples contain the true population parameter 95% of the time.

Making CI

- if it follows CLT or MLE results:

$$P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$$

where α is our level 5% or 1% are most common.

Confidence Interval of MLE

$$\left[\hat{\theta}_{MLE} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta}_{MLE})}}\right]$$

CI w/unknown pop variance

- If the X_i are iid and normally distributed with unknown population variance σ_i^2 then an unbiased estimate of σ^2 is:

$$\sigma^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2 \quad \leftarrow \text{(sample variance)}$$

And it turns out that:

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

- t distribution with n-1 degrees of freedom

→ even if the data isn't normal, for large sample sizes, it still holds.

thus the CI is as follows:

$$\left[\bar{X} \pm \frac{t_{n-1,\alpha/2}\,\hat{\sigma}}{\sqrt{n}}\right]$$

where $t_{n-1,\alpha}$ is the value such that $P(T \leq t_{n-1,\alpha}) = 1-\alpha$

Hypothesis Test

Set up:

H_0 : accepted hypothesis

H_1 : alternate hypothesis

p-val : value to test at our significance level

null	alternate	p-value
$\mu = \mu_0$	$\mu > \mu_0$	$P(\hat{z} \geq z \mid H_0) = 1 - \Phi(z)$
	$\mu < \mu_0$	$P(\hat{z} \leq z \mid H_0) = \Phi(z)$
	$\mu \neq \mu_0$	$P(\hat{z} \geq \lvert z \rvert \mid H_0) = 2(1 - \Phi(\lvert z \rvert))$

One-Sample z tests w/ known variance

$$z = \frac{\text{estimated val} - \text{null value}}{\text{SD of estimate}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

if the z value looks unlikely to have come from a $N(0,1)$ distribution, that is evidence against the null hypothesis.

↳ due to known properties of the normal curve, we can find probability of the z value happening as per the conditions of the alternate hypothesis.

* The p-value is not the probability that the null is false or is it the probability the null is true.

↳ just tells us the probability of the test statistic happening in the null distribution.

CLT and how it relates to testing:

- If a 95% CI for μ contains μ_0, then we would not reject H_0, at the $\alpha = 0.05$ level.

One Sample T-test w/unknown variance

- if we dont have σ, we can use the sample SD w/ t-test instead.
- if the X_i are iid $N(\mu, \sigma^2)$, then:

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

- assumption data is normal but tends to hold at $n > 30$

Two Sample z-test (var known)

- $X_1 \dots X_n$ IID from pop w/unknown μ_1 and known σ_1^2
- $Y_1 \dots Y_m$ IID from pop w/unknown μ_2 and known σ_2^2

$H_0 : \mu_1 = \mu_2 \longrightarrow$ gives us $\mu_1 - \mu_2 = 0$

$$z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{H_0}{\sim} N(0,1)$$

Two Sample T-test (unknown and unequal σ^2)

- AKA Welches test
- same X and Y from above, but with unknown variance

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \sim t_{df} \qquad df = \frac{\left(\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)^2}{\frac{\hat{\sigma}_1^4}{n^2(n-1)} + \frac{\hat{\sigma}_2^4}{m^2(m-1)}}$$

Two-Sample t-test: var unknown but equal

- AKA pooled variance
- same conditions as above but $\sigma_1^2 = \sigma_2^2$

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\hat{\sigma}_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{m+n-2}$$

and we use pooled variance

$$\hat{\sigma}_p^2 = \frac{(n-1)\sigma_1^2 + (m-1)\sigma_2^2}{n+m-2} = \frac{\sum_1^n (X_i - \bar{X})^2 + \sum_1^m (Y_i - \bar{Y})^2}{n+m-2}$$

Paired two sample test (z)

- A paired test can be reduced to a single sample test that the mean difference is 0!

Define $D_i = X_i - Y_i$

$H_0 : \mu_0 = 0 \qquad H_1 : \mu_0 \neq 0$

$$Z = \frac{\bar{D}}{\sigma_0/\sqrt{n}}$$

$$\therefore \text{p val} = 2\left(1 - \Phi\left(\left|\frac{\bar{d}}{\sigma_d/\sqrt{n}}\right|\right)\right)$$

Paired two sample test (T)

$$T = \frac{\bar{D}}{\hat{\sigma}_0/\sqrt{n}} \sim t_{n-1}$$

but replace w tstat not Φ

Two sample test for Proportions

$X_1 \dots X_n \sim \text{Bernoulli}(p_1)$ $\quad \hat{p}_1$ and $\hat{p}_2$ are estimate p from the sample

$Y_1 \dots Y_m \sim \text{Bernoulli}(p_2)$

$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim N(0,1)$$

where $\hat{p}$ is the pooled estimate of variance

$$\hat{p} := \frac{\sum_1^n X_i + \sum_1^m Y_i}{n+m}$$

* can use this test to test for independence between the data

All test above are PARAMETRIC tests

Mann Whitney Test

Non-normal data assuption

$X_1 \dots X_n$ are IID with unknown dist. F

$Y_1 \dots Y_m$ are IID with unknown dist. G

→ checks to see if there is a difference in the ranks of the two samples

$H_0 : F = G \qquad H_1 : F \neq G$

First step is to rank all data, w/1 the lowest, increasing from there. Ties are split by giving .5 on the rank it would have been.

Second is to sum the ranks from each sample.

U-test statistic:

- computes the amount of overlap in the ranks of each sample
- Smaller U means bigger difference
- Smaller U test stat is MORE significant.
- Using the R that is lower, you find the # of values smaller in the other sample, sample by sample, then sum. The ranks that were split are given .5.

ex.

		Rank	
A	6.25	5	$R_A = 5+4+1 = 10$
A	5.50	4	$R_B = 3+2 = 5$ ← choose R_B
A	4.75	1	$U = 1+1 = 2$
B	5.25	3	
B	5.00	2	

But also a formula:

$$U_1 = mn + \frac{n(n+1)}{2} - R_1$$

$$U_2 = mn + \frac{m(m+1)}{2} - R_2$$

$$U = \min(R_1, R_2)$$

We can use the normal curve to approximate $E(U) = \frac{nm}{2}$ $\quad Var(U) = \frac{nm(n+m+1)}{12}$

$$z = \frac{U - E(U)}{\sqrt{\text{var}(U)}} \sim N(0,1)$$

Signed Test

- typically on before/after data
- X IID from some dist F
- Y IID from some dist G

$H_0 : F = G \qquad H_1 : F > G$

$W \sim \text{Binomial}(n, .5)$

Use normal approximation

$E(W) = np \qquad Var(W) = np(1-p)$

$$z = \frac{W - E(W)}{\sqrt{\text{var}(W)}} \sim N(0,1)$$

→ this test has generally low power

Wilcoxon Signed Rank test

General Procedure

1. Remove observations w/no difference
2. Compute R_i = rank of $|D_i|$
3. Calculate $W_i = R_i \times \text{sign}(D_i)$
4. Compute test statistic

$$W_+ = \sum_{i:W_i>0} W_i$$

in other words, sum the positive rank values

If H_0 is true, then:

$$E(W_+) = \frac{n(n+1)}{4}$$

$$Var(W_+) = \frac{n(n+1)(2n+1)}{24}$$

} use normal dist to calc p-val

$$z = \frac{W_+ - E(W_+)}{\sqrt{\text{var}(W_+)}}$$

Example:

Before	After	diff	|d|	rank |d|	signed rank
125	118	7	7	3	3
132	134	-2	2	1	-1
138	130	8	8	4	4
120	124	-4	4	2	-2

In this case, $W_+ = 3+4 = 7$

$E(W_+) = 4(4+1)/4 = 5$ $\quad Var(W_+) = \frac{4(4+1)(2(4)+1)}{24} = \frac{45}{6}$

$$z = \frac{7-5}{\sqrt{\frac{45}{6}}} = \frac{2}{3/\sqrt{6}} = \frac{2\sqrt{6}}{7} = 0.69 \leftarrow \text{test stat}$$

Bonferroni Correction

- for multiple testing

If we are conducting k-test w/ $\alpha = 0.05$

$H_0 : M_X^{(1)} = M_Y^{(1)}$
$H_0 : M_X^{(2)} = M_Y^{(2)}$
$H_0 : M_X^{(3)} = M_Y^{(3)}$
$\vdots$
$H_0 : M_X^{(k)} = M_Y^{(k)}$

Assume null is true in ALL cases

prob of rejection of rejecting ANY test = 5%. but prob of at least one incorrect rejection of the null

The family wise error rate (FWER) is the probability of at least one incorrect rejection among k test.

Our goal is to set 0.05, and currently its 99.4% if k = 100.

→ stricter and lower significance level for each test.

If you have k hypothesis tests, reject H_0 if the p-value is less than α/k, rather than α

our new α becomes $0.05/100 = 0.0005$

Bias:

- An estimate is considered unbiased if the $E(\hat{\theta}) - \theta = 0$ that is:

$$E(\hat{\theta}) = \theta$$

Analysis of Variance (ANOVA)

Suppose we have a sample J observations from each of I populations (groups)

G_1	G_2	⋯	G_I
Y_{11}	Y_{21}	⋯	Y_{I1}
Y_{12}	Y_{22}	⋯	Y_{I2}
⋮	⋮	⋮	⋮
Y_{1J}	Y_{2J}	⋯	Y_{IJ}

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$

H_1 : at least one has different mean

Assumptions:

1. The observations in each sample are independants
2. Each sample is Normally distributed
3. Populations have common variance σ^2

Sums of Squares:

We can capture different types of var using sum of squares

Total Sum of Squares : compares each observation to overall mean

$$SS_T = \sum_i^I \sum_j^J (Y_{ij} - \bar{Y})^2$$

Between sum of Squares (variance between groups): compares each group mean to the overall mean

$$SS_B = J\sum_i^I (\bar{Y}_i - \bar{Y})^2$$

Within Sums of Squares (variance within groups): compares observation to its corresponding group mean

$$SS_W = \sum_i^I \sum_j^J (Y_{ij} - \bar{Y}_i)^2$$

Source	df	SS	MS	F
Between	I-1	SS_B	MS_B	MS_B/MS_W
Within	I(J-1)	SS_W	MS_W	
Total	IJ-1	SS_T		

where:

$MS_B = SS_B/(I-1)$

$MS_W = SS_W/(I(J-1))$

are the mean sum of squares

Test statistic

$$F = \frac{SS_B/(I-1)}{SS_W/(I(J-1))} \overset{H_0}{\sim} F_{I-1,I(J-1)}$$

and the corresponding p-val $P(F \geq f)$

Note: a more extreme test statistic is ALWAYS larger for this test.

Pearson χ^2 test

- goodness of fit test

H_0 : the data comes from specified dist

H_1 : the data doesnt come from the dist

χ^2 test statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-c-1}$$

where:

O is observed value of data

E is expected value of data under the fit.

n is # of independant "cells"/"counts"

c is # of parameters estimated from the data

Fisher's Exact Test

- an exact test for proportions

	A_1	A_2	totals
B_1	N_{11}	N_{12}	$n_{1\cdot}$
B_2	N_{21}	N_{12}	$n_{2\cdot}$
Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	n

under $H_0 : p_1 = p_2$

$N \sim \text{Hypergeo}(n, n_{1\cdot}, n_{\cdot 1})$

$$\therefore P(N_{11} = n_{11}) = \frac{\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

$$E(N_{11}) = \frac{n_{\cdot 1} n_{1\cdot}}{n_{\cdot\cdot}}$$

Let N_{11} be our test statistic. Reject H_0 for extreme values of N_{11}

Our p value for test of independence is $P(|N_{11} - E(N_{11})| \geq |n_{11} - E(N_{11})|)$

Pearson χ^2 test for homogeneity

$H_0 : \pi_{1i} = \pi_{2i} = \dots = \pi_{Ii}$, $i = 1,2,\dots,I$

value (I total)	group (J total) A	B	C	D
a	56	104	80	62
b	12	1	15	12
c	32	32	45	15

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

$$\text{p-value} = P(X^2 \geq x^2)$$

Pearson χ^2 test for Independance

- same as above but
H_0 : the row variable and column are independant

Linear Regression

Linear Relationships

- We assume that each observation can be represented as:

$$y_i = \beta_0 + \beta_1 x_1 + \epsilon_i$$

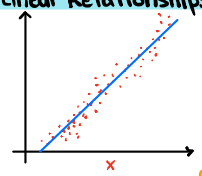
A general linear relationship that describes the relationship between x and y across all observations

An "error" term that allows for variation from linear trend for individual x values

Where $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ for all i

Linear Regression (Continued)

Linear Relationships



This fitted line can be written as:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{\beta}_0$: predicted response; $\hat{\beta}_1$: coefficients of the linear fit; x: covariate/feature/predictor

The "hat", e.g. $\hat{y}, \hat{\beta}$, means that the quantity is a number computed (estimated) from the data.

Least Absolute Deviation (LAD)

- estimating a linear relationship by taking the line that has the lowest average/total prediction error.
- choosing a line based on the absolute value error is called LAD

L1 Loss function

$$\hat{\beta}_0 = \underset{\beta_0}{\operatorname{argmin}} \sum |y_i - (\beta_0 + \beta_1 x_i)|$$

$$\hat{\beta}_1 = \underset{\beta_1}{\operatorname{argmin}} \sum |y_i - (\beta_0 + \beta_1 x_i)|$$

- L1 loss is not differentiable at the minimum

i.e. the LAD is the line such that the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen as so to minimize the absolute difference between the observed and predicted response

Least Squares Algorithm

- the L2 loss is the sum of the areas of the squares whos edges are the vertical distances from the point to the line

"Squared" Prediction Err $= (y - \hat{y})^2$

L2 loss function

$\sum (y - \hat{y})^2$ = loss function

$$\hat{\beta}_0 = \underset{\beta_0}{\operatorname{argmin}} \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\hat{\beta}_1 = \underset{\beta_1}{\operatorname{argmin}} \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

note L2 loss is differentiable at the minimum

ie: the LS line is the line such that the values of $\hat{\beta}_0$ and $\hat{\beta}$ are chosen so as to minimize the squared difference between the observed and predicted responses

- good thing about least squares is that we can derive a formula for the values that minimize loss for β_0 and β_1

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}$$

Comparing LAD w/ LS

- outliers - LS is more sensitive than LAD

Training and Test Set

- before you begin, you need to split your data into:

Training Set: data to train your algorithm
Test Set: data to evaluate your algorithm
Commonly Split 60/40, 70/30, 80/20

Mean Squared Error (MSE)

- value of L2 loss when you plug in the estimate parameters

$$MSE = \frac{1}{n}\sum (y_i - \hat{y}_i)^2 = \frac{1}{n}\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Mean Absolute Deviation (MAD)

- value of the L1 loss when you plug in the estimated parameters

$$MAD = \frac{1}{n}\sum |y_i - \hat{y}| = \frac{1}{n}\sum |y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)|$$

R^2 value (Coefficient of Determination)

- R^2 value is squared correlation between observed and predicted response

$R^2 = \rho^2_{y,\hat{y}} \qquad 0 < R^2 < 1$

- corresponds to the proportion of variability in the observed response that can be "explained" by the covariate

$$R^2 = \frac{\hat{\sigma}^2_y - \hat{\sigma}^2_e}{\hat{\sigma}^2_y} \quad \text{where } \hat{e}_i = y_i - \hat{y}_i$$

- If $\hat{\sigma}^2_{\hat{e}}$ is very small, then $\hat{y}_i$ is very similar to y_i on average, which means x conveys a lot of information about y

Inference for β_0 and β_1

- Linear relationships assumptions
  - ϵ_i is random thus y_i is random
  - ϵ_i are IID
  - $E(\epsilon_i) = 0$, $\operatorname{var}(\epsilon_i) = \sigma^2$ for all i
- x_i is fixed
- β_0, β_1 are parameters that we need to estimate

Bias and Variance of LS estimates of β_0 and β_1

- $\hat{\beta}_1$ is unbiased, that is, $E(\hat{\beta}_1) = \beta_1$,

proof:

$$E(\hat{\beta}_1) = E\left[\frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right] = \frac{\sum (x_i - \bar{x}) E(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$E(y_i - \bar{y}) = E(\beta_1 x_i + \epsilon_i - (\beta_0 + \beta_1 \bar{x} + \bar{\epsilon})) = \beta_1 (x_i - \bar{x}) + E(\epsilon_i - \bar{\epsilon}) = \beta_1 (x_i - \bar{x})$

plug in

$$E(\hat{\beta}_1) = \frac{\beta_1 \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})} = \beta_1$$

$\therefore \hat{\beta}_1$ is unbiased

- $\hat{\beta}_0$ is also unbiased!

Proof:

$$E(\hat{\beta}_0) = E[\bar{y} - \hat{\beta}_1 \bar{x}] = E[(\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}) - \beta_1 \bar{x}] = \beta_0 + \beta_1 \bar{x} + E(\bar{\epsilon}) - \beta_1 \bar{x} = \beta_0$$

$\therefore \beta_0$ is unbiased!

Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

- The variance and covariance of the LS estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ are given by:

$$\operatorname{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\operatorname{var}(\hat{\beta}_1) = \frac{n\sigma^2}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum x_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are asymptotically normal

- since the LS estimates (w/ normality assumption) are the MLE, this means that they are asymptotically normal

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{n\sigma^2}{n\sum x_i^2 - (\sum x_i)^2}\right)$$

Hypothesis testing and CI for β_1 and β_0

$H_0: \beta_j = 0 \qquad H_1: \beta_j \neq 0$

Test Statistic: $t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ where $T \sim t_{n-p}$; p is # of parameters

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sigma^2 \sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}} \qquad \hat{\sigma}_{\beta_1} = \sqrt{\frac{n\sigma^2}{n\sum x_i^2 - (\sum x_i)^2}}$$

and $\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{1}{n-p}\sum r_i^2$

Confidence Interval:

$$\left[\hat{\beta}_j - t_{n-p,\alpha/2}\hat{\sigma}_{\hat{\beta}_j},\ \hat{\beta}_j + t_{n-p,\alpha/2}\hat{\sigma}_{\hat{\beta}_j}\right]$$

Residual Plots for assessing inference assumptions

Evaluating homoskedasticity

- Homoskedasticity - variance of error with each observation is identical and does not depend on x
- Heteroskedasticity - the variance of error associated with each observation is different and may depend on x

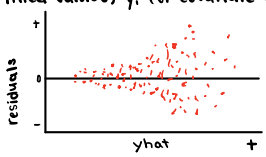
We dont observe ϵ, but perhaps we can get a sense for whether these assumptions are reasonable using the residuals

$r_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i$

Visualizing Predictive Performance

residual plot

- plots the residuals $(y_i - \hat{y}_i)$ against the fitted values, $\hat{y}_i$ (or covariate values x_i)



← the residuals are more variable for larger predicted responses - heteroskedacity

Matrix notation of Multiple Linear Regression

Matrix notation

Obs 1: $y_1 = \beta_0 \times 1 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1$

Obs 2: $y_2 = \beta_0 \times 1 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \cdots + \beta_{p-1} x_{2,p-1} + \epsilon_2$

$\vdots$

Obs n: $y_n = \beta_0 \times 1 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \cdots + \beta_{p-1} x_{n,p-1} + \epsilon_n$

Also can be written as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$Y = X \cdot \beta + \epsilon$

Requirements on error vector:

$$E(\epsilon) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad Cov(\epsilon) = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

So $\therefore \operatorname{cov}(y) = \operatorname{cov}(X\beta + \epsilon) = \operatorname{cov}(\epsilon) = \sigma I$

To generate a prediction, we need to estimate the vector β, and then compute $\hat{y} = X\hat{\beta}$

LS estimate of Multiple LR

$S(\beta_0, \beta_1, \ldots \beta_{p-1}) = \sum (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2$ is the same as finding the vector that minimizes $S(\beta) = \|Y - X\beta\|_2^2 = (Y - X\beta)^T (Y - X\beta)$

this is just a #!

$\|X\|_2 = \sqrt{X^T X}$

LS Estimate of β is:

$\hat{\beta} = (X^T X)^{-1} X^T y$

$\hat{\beta}$ is unbiased

proof:

$$E(\hat{\beta}) = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T E(X\beta + \epsilon) = (X^T X)^{-1} X^T X \beta = \beta$$

$\hat{\beta}$ variance - covariance matrix

- LS estimate, $\hat{\beta}$ has var/cov matrix:

$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

Proof:

$$\begin{aligned} Cov(\hat{\beta}) &= \operatorname{cov}((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T \operatorname{cov}(Y) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \operatorname{cov}(X\beta + \epsilon) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \operatorname{cov}(\epsilon) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T (\sigma^2 I) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Theoretical Distribution of $\hat{\beta}$

$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$

Each element is $\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj})$

where $C = (X^T X)^{-1}$

* this assumes we know $\hat{\sigma}^2$

Hypothesis Testing for β_j

$H_0: \beta_j = 0$ against $H_1: \beta_j \neq 0$

Test statistic: $t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ Where $T \sim t_{n-p}$ (p is the # of parameters)

Confidence Interval for β_j

$$\left[\hat{\beta}_j \pm t_{n-p,\alpha/2}\hat{\sigma}_{\hat{\beta}_j}\right]$$

where $\hat{\sigma}^2_{\hat{\beta}_j} = \hat{\sigma}^2 c_{jj}$

and $\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\|r\|_2^2}{n-p} = \frac{\|Y - X\beta\|_2^2}{n-p}$

Correlation + Covariance

- correlation is "scaled" covariance
- $-1 \leq \operatorname{corr}(x,y) \leq 1$

population:

$$\operatorname{corr}(x,y) = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \cdot \sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$

data sample:

$$\operatorname{corr}(x,y) = \frac{\sum^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\hat{\sigma}_x \cdot \hat{\sigma}_y} = \frac{\sum^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

OLS Proof:

LS coefficient proof $(\hat{\beta}_0)$ $\quad S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$

In order to find the values of β_0 and β_1 that minimize $S(\beta_0, \beta_1)$, we need to set $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to zero, and solve for β_0 and β_1, respectively.

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

So $\frac{\partial S}{\partial \beta_0} = 0$ implies that $0 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$$\Rightarrow n\hat{\beta}_0 = \sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \Rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

LS coefficient proof $(\hat{\beta}_1)$ $\quad S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$

In order to find the values of β_0 and β_1 that minimize $S(\beta_0, \beta_1)$, we need to set $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to zero, and solve for β_0 and β_1, respectively.

Next, $\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i)$

So $\frac{\partial S}{\partial \beta_1} = 0$ implies that $0 = \sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2$

Substituting $\hat{\beta}_0$ with $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$: $0 = \sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2$

Rearranging to make $\hat{\beta}_1$ the subject: $\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

$S(\beta_0, \beta_1) = \sum^n (y_i - \beta_1 x_i - \beta_0)^2$

→ find $\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ then set = 0 to get proof. Use β_0 in β_1 proof

Remembering Linear Algebra

if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then $A^{-1} = \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$\det(A) = ad - bc = |A|$

If all of these are matrixes:

- $A + B = B + A$
- $A + (B + C) = (A + B) + C$
- $A + 0 = A$ where 0 is identity matrix
- $A + (-A) = 0$
- $A(BC) = (AB)C$
- $A(B+C) = AB + AC$
- $(A+B)C = AC + BC$
- $IA = AI = A$
- Let r and s be scalars
- $r(sA) = (rs)A$
- $(r+s)A = rA + sA$
- $r(A+B) = rA + rB$
- $A(rB) = r(AB) = (rA)B$
- $(A^T)^T = A$
- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$
- $(rA)^T = rA^T$

Probability Properties

- $0 \leq P(A) \leq 1$
- $P(A) = 1 - P(A') \rightarrow 1 = P(A) + P(A')$
- if A and B are mutually exclusive $P(A \cap B) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Expectation Properties

- $E(c) = c$
- $E(cX) = cE(X)$
- $E(X+Y) = E(X) + E(Y)$
- $E(\sum f(x)) = \sum E(f(x))$
- $E = \int x f(x) dx$ → continous
- $E = \sum x \cdot P(x=x)$ → discrete

Variance Properties:

- $\operatorname{var}(X) = E(X^2) - (E(X))^2$
- $sd(X) = \sqrt{\operatorname{var}(x)}$
- $\operatorname{var}(aX+b) = a^2 \operatorname{var}(X)$

Covariance Properties

- $\operatorname{cov}(aX, Y) = a\operatorname{cov}(X,Y)$
- $\operatorname{cov}(X, aY) = a\operatorname{cov}(X,Y)$
- $\operatorname{cov}(aX, bY) = a \cdot b \cdot \operatorname{cov}(x,y)$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(X, c) = 0$
- $\operatorname{cov}(X,X) = \operatorname{var}(X)$
- $\operatorname{cov}(X+Y, Z) = \operatorname{cov}(X,Z) + \operatorname{cov}(Y,Z)$
- $\operatorname{cov}(X, Y+Z) = \operatorname{cov}(X,Y) + \operatorname{cov}(X,Z)$

Known Distributions

- **Uniform (a,b)**
  - $\mu = \frac{a+b}{2}$
  - $\sigma^2 = (b-a+1)^2 - 1/12$
- **Bernoulli (p)**
  - $\mu = p$
  - $\sigma^2 = p(1-p)$
- **Binomial (np)**
  - $\mu = np$
  - $\sigma^2 = np(1-p)$
- **Poisson (μ)**
  - $\mu = \mu$
  - $\sigma^2 = \mu$
- **Hypergeometric (n, N, G)**
  - $\mu = nG/N$
  - $\sigma^2 = n\left(\frac{G}{N}\right)\left(\frac{N-G}{N}\right)\left(\frac{N-n}{N-1}\right)$
- **Geometric (p)** start @1
  - $\mu = 1/p$
  - $\sigma^2 = 1-p/p^2$
- **Geometric (p)** start @0
  - $\mu = 1-p/p$
  - $\sigma^2 = 1-p/p^2$
- **Neg Binomial (r, p)**
  - $\mu = r(1-p)/p$
  - $\sigma^2 = r(1-p)/p^2$
- **Normal (0,1)**
- **Normal (μ, σ^2)**
  - $\mu = \mu$
  - $\sigma^2 = \sigma^2$
- **exponential (λ) or Gamma (1, λ)**
  - $\mu = 1/\lambda$
  - $\sigma^2 = 1/\lambda^2$
- **gamma (r, λ)**
  - $\mu = r/\lambda$
  - $\sigma^2 = r/\lambda^2$
- **chi-squared (n) or Gamma ($\frac{n}{2}$, $\frac{1}{2}$)**
  - $\mu = n$
  - $\sigma^2 = 2n$
- **Rayleigh**
  - $\mu = \sqrt{\pi/2}$
  - $\sigma^2 = 4-\pi/2$
- **beta (r, s)**
  - $\mu = r/r+s$
  - $\sigma^2 = rs/(r+s)^2(r+s+1)$
- **arcsin = beta (1/2, 1/2)**
  - $\mu = 1/2$
  - $\sigma^2 = 1/8$