STAT 410 Notes

1 Simple Linear Regression

Interested in linear relationship between 2+ variables

Two fitted lines - one perfectly representing the population(exists but is impossible to get), and one from given data which looks to best represent the line fitting the population

Summarized by (Pearson) correlation coefficient

1. Single number
2. Measure of linear association between variables
3. Inter-class correlation

Given vector of numbers - assign bijective vector which shows actual position - Spearman correlation does regression on ranks instead of values - both measure correlation, not causation

When mapping two variables via scatterplot - an elliptical shape shows signs of a linear relationship - not perfectly linear due to randomness in real life - need to verify assumptions of linear models

Standard deviation lines(look into this) - where change in σ_x would effect change in σ_y and vice versa, not part of regression, points on a scatter plot that are equal number of standard deviations away from the average in each dimension - line which passes through center of mass / averages, with slope equal to the rate of standard deviations($\frac{SD_x}{SD_y}$)

Pearson Correlation, where $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ are n sample pairs

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the averages of $(x_1, x_2, ...x_n)$ and $(y_1, y_2, ...y_n)$ respectively

Denominators - $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is similar to the sample variance

Numerator - signs show what quadrant the values are - sum shows how strongly the data shows up in each quadrant - covariance

Essentially

$$r = \frac{Cov[x, y]}{SD[x]SD[y]} \tag{2}$$

If result is close to 0 - little association - if result is close to 1 - strong positive association and vice versa

Simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, ..., n \tag{3}$$

Deterministic part($\beta_0 + \beta_1 x_1, ...$) and stochastic part($\epsilon_i$), where result($Y_i$) is stochastic(sum of deterministic + stochastic) - where ϵ_i - the collection of errors is seen as white noise - $y = \beta_0 + \beta_1 x$ is the population regression line

Could also be seen as(not used anymore)

$$Y_i|_{X_i = x_i} = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4}$$

Assumptions:

- mean: $E[\epsilon_i] = 0$
- variance: $Var[\epsilon_i] = \sigma^2 < \infty$ - homogeneous
- Independence between errors

Conditional mean of Y_i

$$E[Y_i|X_i = x_i] = \beta_0 + \beta_1 x_i \tag{5}$$

Expected value of Y_i given that $X_i = x_i$ is the deterministic part of a simple linear regression

as, since $\beta_0 + \beta_1 x_i$ is just a constant and $E[\epsilon_i] = 0$,

$$E[Y_i|X_i = x_i] = E[\beta_0 + \beta_1 x_i + \epsilon_i] = \beta_0 + \beta_1 x_i + E[\epsilon_i] = \beta_0 + \beta_1 x_i \tag{6}$$

Therefore $Y_i = E[Y_i|X_i = x_i] + \epsilon_i$ - conditional expectation is deterministic part

In the simple linear regression case:

$y = E[Y|X = x] = \beta_0 + \beta_1 x$

and in general(non-linear), $y = E[Y|X = x]$

$Var(Y_i|X_i = x_i) = \sigma^2$ - where variance of Y_i is homogeneous or homoscedastic

Since B_0 and $B_1 x_i$ are constants and $E[\epsilon_i] = \sigma^2$ by assumption

$$Var[Y_i|X_i = x_i] = Var[B_0 + B_1 x_i + \epsilon_i] = Var[\epsilon_i] = \sigma^2 \tag{7}$$

Interpreting β_1 and β_0

- β_1: change in mean of distribution of response Y_i produced by a unit change in x_i - slope
- β_0: mean of the distribution Y_i when $x_i = 0$

Under stochastic assumption - errors are normally distributed

$$\epsilon_i \sim (iid)N(0, \sigma^2) \tag{8}$$

Conditional distribution of Y_i $Y_i \sim (independent)N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$E[Y_i|X_i = x_i] = \beta_0 + \beta_1 \tag{9}$$
$$Var[Y_i|X_i = x_i] = \sigma^2 \tag{10}$$
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{11}$$

Uncorrelated + Normality = Independence, If independent, the covariance is 0, but not vice-versa

Given model with stochastic assumptions: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Data to obtain estimates(estimators) of the parameters, where Y_i must be independent

$$\{(x_1, Y_1), ..., (x_n, Y_n)(estimator); (x_1, y_1), ...(x_n, y_n)(estimates)\} \implies \hat{\beta_0} + \hat{\beta_1} + \hat{\sigma^2} \tag{12}$$

Estimators - general procedure; estimates - value obtained from specific set of data

To calculate these parameters, we use maximum likelihood estimation(MLE)

Asymptotic properties of ML estimators:

- Efficiency - exceeds Cramer-Rao lower bound(variance), Fisher information($n \to \infty$, becomes asymptotically efficient)
- Consistent - estimation converges in probability to parameter estimated
- Normal - converge to normal distribution($n \to \infty$, sample distribution of estimator approachs normal)
- Functional invariance - $E(estimator) = E(function)$

Joint likelihood of the errors ϵ_i

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(Y_i - B_0 - B_1 x_i)^2\} \tag{13}$$

where $Y_i - B_0 - B_1 x_i$ is the error(ϵ_i), $e_i \sim N(0, \sigma^2), i = 1, ..., n$.

Essentially, we want to find the parameters which maximize the likelihood of getting the data

Maximizing the likelihood function:

$$(\hat{\beta_0^{MLE}}, \hat{\beta_1^{MLE}}, \hat{\sigma^{2MLE}}) = argmax(L(\beta_0, \beta_1, \sigma^2)) \tag{14}$$

- Find $L()$
- Calculate log $L()$
- $\frac{\partial}{\partial \theta_i} logL() = 0)$
- Solve for $\hat{\theta_i}$ - theoretically need to find if max, min, or saddle point but not needed in this class

$$logL(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 + constant \tag{15}$$

Find partials for $\beta_0, \beta_1, \sigma^2$ - then set to 0 and solve for corresponding values

Normal equations and MLEs - solving for β_1 and β_1 respectively for their partial derivates

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i \tag{16}$$

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \tag{17}$$

$$\hat{\beta}_1{}^{MLE} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{18}$$

$$\hat{\beta}_0{}^{MLE} = \bar{Y} - \hat{\beta}_1{}^{MLE} \bar{x} \tag{19}$$

$$\hat{\sigma^2_{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - (\hat{\beta_0^{MLE}} + \hat{\beta_1^{MLE}} x_i)\}^2 \tag{20}$$

Since Y is a random variable, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma^2}$ are also random variables with properties(distribution, variances, etc.)

Notation for summations:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{21}$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y}) Y_i = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \tag{22}$$

$$S_{xY} = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \tag{23}$$

Fitted values

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{24}$$

Cannot have more datapoints than predictors - leads to infinite solutions - need to find a line of best fit minimizing SSE

If we use data, x_i, y_i are constants, if we use general procedure, x_i, y_1 are random variables - error component is unobservable since from population regression(which is unobtainable)

Can measure how well the line fits using $e_i = (Y_i - \hat{y_i})$ - the residuals / prediction errors not actual error

Alternatively,

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} \tag{25}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{26}$$

$$\hat{\sigma^2_{MLE}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 (residual) \tag{27}$$

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{YY}}} \tag{28}$$

Ordinary least Squares(OLS)

$$(\hat{\beta_0^{LS}}, \hat{\beta_1^{LS}}) = argmin_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{29}$$

Where we can find $argmin$ using calculus, and the betas from OLS are the exact same as MLE

Unbiased estimator of σ^2

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} \{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\}^2 \tag{30}$$

Where we use $n-2$ as a denominator since we lose 2 degrees of freedom due to using $\hat{\beta}_0$ and $\hat{\beta}_1$

2 Inference with SLR Estimators

Random Sample($X = (X_1, X_2, ..., X_n)$ - collection of iid random variables

Probability density function of $X_i (f(x_i|\theta)$

Statistic / Estimator - ($\hat{\theta} = \delta(X) = \delta(X_i, ..., X_n)$ - function applied to a random sample(min, max, median, mean, variance etc.) - sometimes used to estimate parameters

Two desired properties of estimator($\hat{\theta}$)

- Overall Accuracy: $E[\hat{\theta}|\theta]$ is close to θ $\forall\theta$
- Overall Precision: $Var[\hat{\theta}|\theta]$ is small $\forall\theta$

An unbiased estimator - overall accurate($E[\hat{\theta}|\theta] = \theta$ $\forall\theta$. If repeated multiple times - in the average, will be accurate

Bias: $Bias[\hat{\theta}|\theta] = E[\hat{\theta}|\theta] - \theta$

Loss function and MSE

- Squared Error Loss: $L(a, \theta) = (a - \theta)^2$
- Mean Squared Error: $MSE[\hat{\theta}|\theta] = E[(\hat{\theta} - \theta)^2|\theta]$ - different from variance - MSE = variance for unbiased estimators

$MSE[\hat{\theta}|\theta] = E[(\hat{\theta} - \theta)^2|\theta] = Var[\hat{\theta}|\theta] + Bias^2[\hat{\theta}|\theta]$

Properties of SLR Estimators

- $E[\hat{\beta}_1] = \beta_1$
- $E[\hat{\beta}_0] = \beta_0$
- $Var[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$
- $Var[\hat{\beta}_0] = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$ - **all should be proven as an exercise**

LS estimators are BLUE(Best Linear Unbiased Estimator that exists)

Some properties:

- $\sum_{i=1}^n (Y_i - \hat{y}_i) = \sum_{i=1}^n e_i = 0$
- $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{y}_i$
- $\sum_{i=1}^n x_i e_i = 0$
- $\sum_{i=1}^n \hat{y}_i e_i = 0$

Properties of $\hat{\sigma^2}$

- $SS_{residuals} = SSE = \sum_{i=1}^n e_i^2$
- $\hat{\sigma^2} = \frac{SS_{residuals}}{n-2} = MS_{residuals} = MSE$

$\hat{\sigma^2}$ is unbiased

Gaussian Assumption

$$\epsilon_i \sim (iid)N(0, \sigma^2) \tag{31}$$

Inferences about parameters:

- Hypothesis testing
- Interval estimation or confidence intervals

Predictions about future values of Y. **Derive these as an exercise**

$$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma_2}{S_x x} \tag{32}$$

$$\hat{\beta_0} \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})) \tag{33}$$

$$\frac{(n-2)\hat{\sigma^2}}{\sigma^2} = \frac{SS_{residual}}{\sigma^2} \sim \chi^2_{n-2} \tag{34}$$

Since the estimator is distributed as a normal whose mean is the thing to be estimated(circular), must use pivoting sampling distribution

$$\frac{\hat{\beta_1} - \beta_1}{se(\hat{\beta_1}} \sim N(0,1), se(\hat{\beta_1}) = \frac{\sigma}{\sqrt{S_{xx}}} \tag{35}$$

$$\frac{\hat{\beta_0} - \beta_0}{se(\hat{\beta_0}} \sim N(0,1), se(\hat{\beta_0}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{36}$$

$$\frac{(n-2)\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2} \tag{37}$$

Where se is the standard error(standard deviation of a statistics) - need to find σ

Pivots with unknown σ^2

$$\frac{\hat{\beta_1} - \beta_1}{se(\hat{\beta_1})} \sim t_{n-2}, \hat{se}(\hat{\beta_1}) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \tag{38}$$

$$\frac{\hat{\beta_0} - \beta_0}{\hat{se}(\hat{\beta_0})} \sim t_{n-2}, \hat{se}(\hat{\beta_0}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \tag{39}$$

If σ is known, use $se(\hat{\beta}) or \hat{\beta_0}$ and distribution $N(0,1)$.

If we don't know β_1, we

- Propose β_1 with hypothesis test
- See where it lands on t distribution, if close to center, you feel good

Null Hypothesis: $H_0 : \beta_1 = \beta_{10}$ vs Alternative Hypothesis: $H_1 : \beta_1 \neq \beta_{10}$ - in a case of infinite data, we reject the null

Test-statistic:

$$T = \frac{\hat{\beta_1} - \beta_{10}}{\hat{se}(\hat{\beta_1})} \tag{40}$$

$$\hat{se}(\hat{\beta_1}) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \tag{41}$$

Where $T \sim t_{n-2}$ and we reject H_0 if $|T| > t_{1-\alpha/2,n/2}$, where $t_{1-\alpha/2,n/2}$ is the $1-\alpha/2$ quantile of t_{n-2}

$$p - value = 2(1 - F_{t_{n-2}}(|T|)) \tag{42}$$

Where $F_{t_{n-2}}$ is the CDF of t_{n-2}. If $p > \alpha$, fail to reject, and if $p < \alpha$, we reject.

Type 1 Errors - Reject True, Type 2 - do not reject a false

To find $\beta_1 0$:

- Want to reject $\beta_1 0 = 0$ - if $\beta_1 = 0$, there is no relation

Confidence Intervals - we want to look for interval, for slope β_1 and similarly for β_0

$$\hat{\beta}_1 - t_{1-\alpha/2,n-2}\hat{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2,n-2}\hat{se}(\hat{\beta}_1) \tag{43}$$

For variance of errors σ^2

$$\frac{(n-2)\hat{\sigma^2}}{\chi^2_{1-\alpha/2,n-2}} \leq \sigma^2 \leq \frac{(n-2\hat{\sigma^2}}{\chi^2_{\alpha/2,n-2}} \tag{44}$$

$P(L \leq \mu \leq u) = 1 - \alpha$

$$P(t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{p_1}} \leq t_{1-\alpha/2,n-2})) = 1 - \alpha \tag{45}$$

$$-t_{1-\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)} \leq t_{1-\alpha,n-2} \tag{46}$$

We are $1-\alpha\%$ confident that β_1 within range - if we do this infinite times, $1-\alpha\%$ of times capture

CI of Mean Response Parameter $E[Y_0|X_0 = x_0] = \beta_0 + \beta_1 x_0$, make confidence bands with $1-\alpha\%$ of containing true line

Parameter: $E[Y_0|X_0 = x_0] = E[Y_0|x_0] = \mu_{Y_0|x_0} = \beta_0 + \beta_1 x_0$

Point Estimator: $\hat{E}[Y_0|X_0 = x_0] = \hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$$Var(\hat{y_0}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = Var(\bar{Y} + \hat{\beta}_1(x_0 - \bar{x})) = \tag{47}$$

$$Var[\hat{Y}] + Var(\hat{\beta}_1(x_0 - \bar{x}) = \tag{48}$$

$$\frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2\sigma^2}{S_{xy}} = \sigma^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}) \tag{49}$$

Should prove that $Cov[\bar{Y}, \hat{\beta}_1] = 0$ as exercise, and since $\bar{Y}, \hat{\beta}_1$ are normally distributed, are independent.

$$se(\hat{y_0}) = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{50}$$

$$\hat{se}(\hat{y_0}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{51}$$

$$\hat{y_0} - t_{1-\alpha,n-2}\hat{se}(\hat{y_0}) \leq \mu_{Y_0|x_0} \leq \hat{y_0} + t_{1-\alpha,n-2}\hat{se}(\hat{y_0}) \tag{52}$$

Prediction interval of future Y_0

$$Y_0 = \hat{y_0} + \epsilon_0, \hat{y_0} = \hat{\beta_0} + \hat{\beta_1}x_0 \tag{53}$$

$$E[Y_0 - y_0] = 0, Var[Y_0 - \hat{y_0}] = \sigma^2\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{54}$$

$$\hat{se}(Y_0 - \hat{y_0}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{55}$$

$$\hat{y_0} - t_{1-\alpha/2,n-2}\hat{se}(Y_0 - \hat{y_0}) \leq Y_0 \leq \hat{y_0} + t_{1-\alpha/2,n-2}\hat{se}(Y_0 - \hat{y_0}) \tag{56}$$

Graph of averages - average of points in a strip(interval in x) of data - combination of all these points is the graph of averages - if points seem linear - regression might be appropriate

As correlation approaches 1, the regression line and SD line converge

3 Multiple Linear Regression

Matrix version of SLR

$$Y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1 \tag{57}$$

$$Y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2 \tag{58}$$

$$\vdots \tag{59}$$

$$Y_n = \beta_0 + \beta_1 x_n + \epsilon_n \tag{60}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_3 \end{bmatrix} \tag{61}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_3 \end{bmatrix} \tag{62}$$

Giving us $Y = x\beta + \epsilon$

For multiple predictors$(y_i; x_{i1}, ...x_{in})$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + ... + \beta_1 x_{in} + \epsilon_i \tag{63}$$

Samee as SLR, but x matrix is multidimensional, each column is a predictor(**insert here!!**)

$$\hat{\beta} = argmin_\beta\{(Y - x\beta)'(Y - x\beta)\} = argmin_\beta|Y - x\beta|^2 \tag{64}$$

Where $Y - x\beta = \epsilon$ - essentially squaring the epsilons and then getting magnitude

Loss function(scalar)

$$S(\beta) = (Y - x\beta)'(Y - x\beta) = \tag{65}$$
$$Y'Y - 2\beta'x'Y + \beta'x'x\beta \tag{66}$$
$$\frac{\partial}{\partial\beta}S(\beta) = -2x'Y + 2x'x\beta \tag{67}$$
$$-2x'Y + 2x'x\hat{\beta} = 0 \tag{68}$$
$$x'x\hat{\beta} = x'Y \tag{69}$$
$$\hat{\beta} = (x'x)^{-1}x'Y \tag{70}$$

Where (69) is considered the normal equations

Prove that $x'x$ is invertible(Exercise)

$\hat{y} = x\hat{\beta} = x(x'x)^{-1}x'Y$, where $H = x(x'x)^{-1}x'$ is the hat matrix which is a projection matrix which projects vectors onto the span of the x vectors

$$e = Y - \hat{y} = Y - HY = (I - H)Y \tag{71}$$

Where residuals are perpandicular to $\hat{y}$ going to y Properties of the H matrix(Prove as an exercise)

- H is symmetric: $H' = H$
- H is idempotent: $H^2 = H$
- I - H is symmetric: $(I - H)' = I - H$
- I - H is idempotent: $(I - H)^2 = I - H$

Let Z be a random vector and A a matrix of constants

- $E[A] = A$
- $E[W] = E[AZ = AE[Z]$
- $Cov[W] = \begin{bmatrix} \sigma_1^2 & \gamma_{12} & ... & \gamma_{1p} \\ \gamma_{21} & \sigma_2^2 & ... & \gamma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{p1} & \gamma_{p2} & ... & \sigma_p^2 \end{bmatrix} = ACov[Z]A^t$, where $\gamma_{ij} = \gamma_{ji} = Cov[W_i, W_j], \sigma_i^2 = Cov(W_i, W_i) = Var(W_i)$ - essentially must a matrix of covariances

$Cov[\epsilon] = Cov[Y] = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 I$, since independent, so other is 0

$Cov = 0$, Normally distributed $\implies$ *Independence*

Prove as exercise

$$Cov[e] = \sigma^2(I - H) \tag{72}$$

Properties of LS estimator(**Prove first two as exercise**)

- Unbiased: $E[\hat{\beta}] = \beta$
- Variance: $Var[\hat{\beta}] = Cov[\hat{\beta}] = \sigma^2(x'x)^{-1}$
- BLUE(Gauss-Markov theorem)

Gaussian Assumption, if Errors are normal:

- MLE same as LS estimator
- $\hat{\beta}$ follows MN $(\beta, Var[\hat{\beta}]$

Estimator of σ^2

$$\hat{\sigma^2} = \frac{SS_{res}}{n - (k+1)} = MS_{res} \tag{73}$$

See geometric interpretation too

4 Inference in Multiple Linear Regression

Random Vectors

$$Y = \begin{bmatrix} Y_1 & Y_2 & Y_3 & Y_4 \end{bmatrix}' \tag{74}$$

$$\mu = E[Y] = \begin{bmatrix} E[Y_1] & E[Y_2] & E[Y_3] & E[Y_4] \end{bmatrix} \tag{75}$$

$$Var/Covar = \sum = E[(Y - \mu)(Y - \mu)'] \tag{76}$$

Variance / Covariance Matrix

$$\Sigma = \begin{bmatrix} Var[Y_1] & Cov[Y_1, Y_2] & ... & Cov[Y_1, Y_p] \\ Cov[Y_2, Y_1] & Var[Y_2] & ... & Cov[Y_2, Y_p] \\ \vdots & \vdots & \vdots & \vdots \\ Cov[Y_p, Y_1] & Cov[Y_p, Y_2] & ... & Var[Y_p] \end{bmatrix} \tag{77}$$

Σ is both symmetric and positive semidefinite($xMx^T \geq 0$)

For constant $m \times p$ matrix A

$$E[AY] = A\mu, Var[AY] = A\sum A' \tag{78}$$

For constant $p \times p$ matrix A

$$E[Y'AY] = \mu' A\mu + trace[A\sum] \tag{79}$$

Multivariate Normal

$$Y = [Y_1, Y_2, ..., Y_p]' \sim MVN(\mu, \sum) \tag{80}$$

PDF:

$$f(Y) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} exp\{-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)\} \tag{81}$$

For any non-singular matrx A and a vector b, if $Y \sim MVN(\mu, \Sigma)$, where Y is a $p \times 1$ random matrix:

$$AY + b \sim MVN(A\mu + b, A\Sigma A') \tag{82}$$

Corollaries:

1. Linear Combination of Y is MVN
2. $Y_i (1 \leq i \leq p)$ are $N(\mu_i, \Sigma_{ii})$, μ is a vector of p elements
3. $(Y-\mu)'\Sigma^{-1}(Y-\mu) \sim \chi^2_p$(like a Z-Score, standardizes)

Assumption: $\epsilon \sim MVN(0, \sigma^2 I)$

Implies: $Y \sim MVN(x\beta, \sigma^2 I)$

Using same data for multiple comparisons - problematic as repeating tests - splitting data with test and train sets

Joint Sampling Distribution

$$\hat{\beta} \sim MVN(\beta, Var[\hat{\beta}] \tag{83}$$

$$where\ Var[\hat{\beta}] = Cov[\hat{\beta}] = \sigma^2(x'x)^{-1} \tag{84}$$

$$In\ particular,\ Var(\hat{\beta}_i) = \sigma^2 C_{ii}, Cov[\hat{\beta}_i, \hat{\beta}_j] = \sigma^2 C_{ij} \tag{85}$$

and C_{ij} is the ijth entry of $(x'x)^{-1}$

Marginal Sampling Distribution

Known σ^2

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j} \sim N(0,1), se(\hat{\beta}_j) = \sigma\sqrt{C_{jj}} \tag{86}$$

Unknown σ^2

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j} \sim t_{n-k-1}, se(\hat{\beta}_j) = \hat{\sigma}\sqrt{C_{jj}} \tag{87}$$

Marginal(beware of multiple comparisons) test of hypothesis of β_j

Hypotheses:

$$H_0 : \beta_j = 0, H_1 : \beta_j \neq 0 \tag{88}$$

Test statistic(pivot)

$$T = \frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j}, \hat{se}(\hat{\beta}_j = \hat{\sigma}\sqrt{C_{jj}} \tag{89}$$

Sampling distribution under H_0 - $T \sim t_{n-k-1}$

Rule: reject H_0 if $|T| > t_{1-\alpha/2,n-k-1}$

Where $t_{1-\alpha/2,n-k-1}$ is the $1-\alpha/2$ quantile of t_{n-k-1}. p-value $= 2(1 - F_{t_{n-k-1}}(|T|))$, where $F_{t_{n-k-1}}$ is the CDF of t_{n-k-1}

For marginal confidence intervals

For the regression coefficient β_j

$$\hat{\beta}_j - t_{1-\alpha/2,n-k-1}\hat{se}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2,n-k-1}\hat{se}(\hat{\beta}_j) \tag{90}$$

For variance of the errors σ^2

$$\frac{(n-k-1)\hat{\sigma^2}}{\chi^2_{1-\alpha/2,n-k-1}} \leq \sigma^2 \leq \frac{(n-k-1)\hat{\sigma^2}}{\chi^2_{\alpha/2,n-k-1}} \tag{91}$$

CI of mean response:

Parameter: $E[Y_0|X_0 = x_0] = E[Y_0|x_0] = \mu_{Y_0|x_0} = x_0'\beta$

Point estimator: $\hat{E}[Y_0|X_0 = x_0] = \mu_{\hat{Y_0}|x_0} = \hat{y}_0 = x_0'\hat{\beta}$, where $x_0 = [1, x_{01}, ..., x_{0k}]'$

$$Var[\hat{y}_0] = \tag{92}$$

$$Var[x_0'\hat{\beta}] = \tag{93}$$

$$\sigma^2 x_0'(x'x)^{-1}x_0 \tag{94}$$

$$se(\hat{y}_0) = \sigma\sqrt{x_0'(x'x)^{-1}x_0} \tag{95}$$

$$= \hat{se}(\hat{y}_0) = \hat{\sigma}\sqrt{x_0'(x'x)^{-1}x_0} \tag{96}$$

Confidence interval of the mean response

$$\hat{y}_0 - t_{1-\alpha/2,n-k-1}\hat{se}(\hat{y}_0) \leq \mu Y_0|x_0 \leq \hat{y}_0 + t_{1-\alpha/2,n-k-1}\hat{se}(\hat{y}_0) \tag{97}$$

Prediction Interval of Future Y_0 $Y_0 = \hat{y}_0 + \epsilon_0$. where $\hat{y}_0 = x_0'\hat{\beta}$

$E[Y_0 - \hat{y_0} = 0, Var[Y_0 - \hat{y_0}] = \sigma^2(1 + x_0'(x'x)^{-1}x_0)$

$\hat{se}(Y_0 - \hat{y_0}) = \hat{\sigma}\sqrt{1 + x_0'(x'x)^{-1}x_0}$

Prediction Interval

$$\hat{y_0} - t_{1-\alpha/2,n-k-1}\hat{se}(Y_0 - \hat{y}_0) \leq Y_0 \leq \hat{y_0} + t_{1-\alpha/2,n-k-1}\hat{se}(Y_0 - \hat{y}_0) \tag{98}$$

Bonferroni Method - reliance on alpha - limits alpha - can be too conservative

$\beta_j \in [\hat{\beta}_j - t_{1-\alpha*/2,n-k-1}\hat{se}(\hat{\beta}_j), \hat{\beta}_j + t_{1-\alpha*/2,n-k-1}\hat{se}(\hat{\beta}_j)$, where $\alpha* = \frac{\alpha}{m}$ and m is the number of hypothesis(k+1 for MLR)

General Linear Hypothesis(overall tests, not marginal - mitigates multiple comparisons problem - not practical at times so sometimes marginal is used‘):

$H_0 : T\beta = c, H_1 : T\beta \neq c$, where T is an $r \times (k+1)$ matrix of constants, where rows are linearly independent($rank(T) = r$), T is the design matrix of experiments - defines family of tests wanted to run simultaneously - as many rows as βs

Under $H_0 : T\hat{\beta} - c \sim MVN(0, \sigma^2 T(x'x)^{-1}T')$

Then $\frac{(T\hat{\beta}-c)'[T(x'x)^{-1}T')]^{-1}(T\hat{\beta}-c)}{\sigma^2} \sim \chi^2_r$

We already know:

$$\frac{SS_{res}}{\sigma^2} = \frac{(n-k-1)MS_{res}}{\sigma^2} \sim \chi^2_{n-k-1} \tag{99}$$

$$\tag{100}$$

$$MS_{res} = \frac{SS_{res}}{n-k-1} \tag{101}$$

$$F_0 = \frac{(T\hat{\beta} - c)'[T(x'x)^{-1}T']^{-1}(T\hat{\beta} - c)/r}{MS_{res}} \tag{102}$$

Under $H_0 : F_0 \sim F_{r,n-k-1}$, where a F distribution describes the division of two chi-square variables with the df of the chi-squares divided as the parameters

Reject H_0 if $F_0 \geq F_{1-\alpha,r,n-k-1}$

p-value $= 1 - P(F_{r,n-k-1} \leq F_0)$

Confidence region:

$$c = T\beta \tag{103}$$

$$\{T\beta : \frac{(T\hat{\beta} - T\beta)'[T(x'x)^{-1}T']^{-1}(T\hat{\beta} - T\beta)/r}{MS_{res}} \leq F_{1-\alpha,r,n-k-1}\} \tag{104}$$

$$\tag{105}$$

For $T = I_{k+1}$

$$\{\beta : \frac{(\hat{\beta} - \beta)'(x'x)(\hat{\beta} - \beta)/r}{MS_{res}}\} \tag{106}$$

For SLR:

$$\frac{n(\hat{\beta}_0 - \beta_0)^2 + 2\Sigma x_i(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \Sigma x_i^2(\hat{\beta}_1 - \beta_1)^2}{2MS_{res}} \leq F_{1-\alpha,2,n-2} \tag{107}$$

, where β_0, β_1 are random variables - find which variables give a valid region

5 ANOVA and R2 in MLR

ANOVA allows testing over three(or more) populations

ANOVA: 3 different means - wanting to minimize the difference between $Y_i, \hat{y}_i$

Partitioning the distance

$$Y_i - \bar{Y} = (Y_i - \hat{y}_i) + (\hat{y}_i - \bar{Y}) \tag{108}$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}[(Y_i - \hat{y}_i) + (\hat{y}_i - \bar{Y})^2] \tag{109}$$

$$= \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{y}_i)(\hat{y}_i - \bar{Y}) \tag{110}$$

Where $2\sum_{i=1}^{n}(Y_i - \hat{y}_i)(\hat{y}_i - \bar{Y}) = 0$**Prove as an exercise**

Sums of Squares: $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$, where we want to maximize $Y_i - \hat{y}_i$

- $SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$(total sum of squares)
- $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2$(sum of model regressors)
- $SS_{res} = \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2$
- $SS_T = SS_R + SS_{res}$

All estimators of standard deviations are biased - there exists for variance but not standard deviation

Want to maxmimize SS_R - larger SS_R mean model addresses more variability

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T} \tag{111}$$

$$0 \leq R^2 \leq 1 \tag{112}$$

In SLR: $R^2 = r^2$

Source of Variation	**Sum of Squares**	**df**	**Mean Square(MS)**	**E[MS]**
Regression	SS_R	1	$MS_R = \frac{SS_R}{1}$	$\sigma^2 + \beta_1^2 S_{xx}$
Residual	SS_{res}	n-2	$MS_{res} = \frac{SS_{res}}{n-2}$	σ^2
Total	SS_T	n-1		

where $SS_R = \hat{\beta_1}^2 S_{xx} = \hat{\beta}_1 S_{xy}$ - prove $E[MS]$ as exercise.

$E[MS]_{Regression} = E[MS]_{Residual}$ when $\beta_1 = 0$

F-Test:

$$H_0 : B_1 = 0 vs. H_1 : B_1 \neq 0 \tag{113}$$

$$F_0 = \frac{MS}{MS_{res}} = \frac{SS_R/1}{SS_{res}/(n-2)} \sim F(1, n-2) \tag{114}$$

$$\tag{115}$$

Reject H_0 if $F_0 > F_{1-\alpha,1,n-2}$

Going back to the t-test for SLR

$$T_0 = \frac{\hat{\beta}_0}{\hat{se}[\hat{\beta}_1]} = \frac{\hat{\beta}_1}{\sqrt{MS_{res}/S_{xx}}} \tag{116}$$

$$T_0^2 = \frac{\hat{\beta_1}^2}{MS_{res}/S_{xx}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{res}} = \frac{MS_R}{MS_{Res}} = F_0 \tag{117}$$

Source of Variation	**Sum of Squares**	**df**	**Mean Square(MS)**	**E[MS]**
Regression	SS_R	k	$MS_R = \frac{SS_R}{1}$	$\sigma^2 + \frac{\beta * x_C' x_c \beta *}{k\sigma^2}$
Residual	SS_{res}	n-k-1	$MS_{res} = \frac{SS_{res}}{n-k-1}$	σ^2
Total	SS_T	n-1		

Where $\beta * = (\beta_1, ...\beta_k)'$ excluding the intercept β_0, **Also put x_c here - essentially all elements of x subtracted by its column mean**

$E[MS]_{regression} = E[MS]_{residual}$ when $\beta = 0$

F-test of significance of regression

$$H_0 : \beta_1 = ...\beta_k = 0, H_1 : notH_0 \tag{118}$$

$$F_0 = \frac{MS_R}{MS_{res}} = \frac{SS_R/k}{SS_{res}/(n-k-1)} \sim F(k, n-k-1) \tag{119}$$

We reject H_0 if $F_0 > F_{1-\alpha,k,n-k-1}$

$$SS_{res} = Y'(I-H)Y \tag{120}$$

$$SS_{res} = e'Y = Y'Y - \hat{\beta}' x'Y \tag{121}$$

$$SS_T = Y'(I - \frac{1}{n}11')Y \tag{122}$$

$$SS_R = Y'(H - \frac{1}{n}11') \tag{123}$$

Prove as exercise

With General Model

$$H_0 = T\beta = c, H_1 = T\beta \neq c \tag{124}$$

To connect between GLH and ANOVA:

$$T = \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & ... & 1 \end{bmatrix} \tag{125}$$

$$c = 0 \tag{126}$$

Essentially, we are testing $\beta_1 = \beta_2 = ...\beta_k = 0$

Extra sum of squares method

- Particular case of GLH - test one or more β equal to zero that can be solved in a simple way

- Assesses the increase in the Regression sum of squares that result by adding the removed regressors to a reduced model containing the rest of the regressors
- Differences are done in terms of regression sum of squares - results are the same as doing it with the residual sum of squares, reversing the order
- $SS_{res(H_0} = SS_{res(Red)} - SS_{res(Full)} = SS_{R(Full)} - SS_{R(Red}$

Partial F-Test

- Test conducted is a partial F-test or partial significance test - measures contribution of the missing regressors given that the rest of the regressors are in the model
- If we reject the test - means at least one of the removed regression coefficients significantly contribute to the model - should be included
- If we fail to reject - no evidence that supports that the reduced model is inadequate
- Useful in model building - many regressors are available and we would like to find the best set of regressors for model in use
- Can show that this partial F test in this case is equivalent to the t test for $H_0 : \beta_j = 0, H_1 : \beta_j \neq 1$
- When applied to more than one - multiple comparisons

H_0 : reduced model is sufficient - H_1 : not H_0

$$F_0 = \frac{SS_{res(H_0)}/r}{SS_{res(Full}/(n-k-1)} \tag{127}$$

where r is the number of removed coefficients(which coincides with the number of restrictions of T) - reject H_0 if $F_0 > F_{1-\alpha,r,n-k-1}$ - so if we are testing $\beta_1 = \beta_3$ - we only remove one coefficient, not two

6 Categorical Predictors

Can code qualitative(and binary) data using

- 0 if observation is type A
- 1 if observation is type B

Assume MLR model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_2 \tag{128}$$

When $x_{i2} = 0$, the model is $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$, and when $x_{i2} = 1$, the model is $(\beta_0 + \beta_2) + \beta_1 x_{i1}$ - essentially creating two simultaneous SLR - same slope different intercept

- Model implies that the two types lead to two parallel regression lines with slope β_1
- β_0 is the intercept corresponding to Type A
- β_2 is the difference in mean type resulting from changing from type A to type B
- $\beta_0 + \beta_2$ is the intercept corresponding to type B

If we do not want to enforce a common slope - can use an interaction term, where x_1x_2 is the interaction between x_1 and x_2

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i \tag{129}$$

- β_0: intercept for Type A
- β_1 : slope for Type A
- β_2 : difference in the intercept resulting from changing from Type A to Type B
- β_3 : difference in slope resulting from changing from Type A to Type B

If additive is not significant but multiplicative is, still keep additive factors.

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$

- When $x_{i2} = 0$, the model is $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
- When $x_{i2} = 1$, the model is $Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} + \epsilon_i$

If we have K levels(types) - need $K - 1$ indicator(dummy) variables - for three types:

- Type A: 0, 0
- Type B: 1, 0
- Type C: 0, 1

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$

- When Type A: $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
- When Type B: $Y_i = (\beta_0 + \beta_2) + \beta_1 x_{i1} + \epsilon_i$
- When Type C: $Y_i = (\beta_0 + \beta_3) + \beta_1 x_{i1} + \epsilon_i$

Analysis of Covariance

- All regression models seen are known as cases of analysis of covariance
- Analysis of covariance is a type of linear model where there is a combination of a quantitative factor with a qualitative factor
- Consider regression of Y_i on F where F is a factor with K levels and $K \geq 2$

- Need to create $K-1$ indicator variables and put them into the MLR model
- Suppose F has 3 levels(A, B,C)
- – A has data $y_{11}, y_{12}, y_{13}, y_{14}$ with sample mean $\bar{y}_{1+}$
  – B has data $y_{21}, y_{22}, y_{23}, y_{24}$ with sample mean $\bar{y}_{2+}$
  – C has data $y_{31}, y_{32}, y_{33}, y_{34}$ with sample mean $\bar{y}_{3+}$
- Should create 2 dummy variables x_1, x_2 such that $x_1 = 1$ if B and $x_2 = 1$ if C, where x_1 and x_2 cannot be both 1

Model: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

$$E[Y|F = A] = E[Y|x_1 = 0, x_2 = 0] = \beta_0 \tag{130}$$
$$E[Y|F = B] = E[Y|x_1 = 1, x_2 = 0] = \beta_0 + \beta_1 \tag{131}$$
$$E[Y|F = C] = E[Y|x_1 = 0, x_2 = 1] = \beta_0 + \beta_2 \tag{132}$$

β_0 - baseline mean, β_1, β_2 - difference in mean to β_0 One Way Anova:

- In prior example - coding of x_1, x_2 is arbitrary
- All coding schemes are based on the model: $Y_{ij} = \mu_i + \epsilon_{ij}, e_{ij} \sim N(0, \sigma^2)$, where $i = 1, ...K$ and $j = 1, ...m$
- μ_i is the population mean of group i and there are K of them. All treatments(levels) have m responses(balanced) - being unbalanced means losing power
- $\mu_1 = \beta_0, \mu_2 = \beta_1 + \beta_0, \mu_3 = \beta_2 + \beta_0, ...$

$$n = K \cdot m \tag{133}$$
$$\bar{Y_{i+}} = \frac{1}{m} \sum_{j=1}^{m} Y_{ij} \tag{134}$$
$$\bar{Y_{++}} = \frac{1}{K} \sum_{i=1}^{K} \bar{Y_{i+}} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{m} Y_{ij} \tag{135}$$

$K = k + 1$ where k is the number of MLR predictors

$$SS_T = TSS = \sum_{i=1}^{K} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_{++})^2, df = n - 1 \tag{136}$$
$$SS_R = BSS = \sum_{i=1}^{K} \sum_{j=1}^{m} (\bar{Y}_{i+} - \bar{Y}_{++})^2, df = K - 1 \tag{137}$$
$$SS_{res} = WSS = \sum_{i=1}^{K} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_{i+})^2, df = n - K \tag{138}$$

Source	SS	df	MS
Between	SS_R	K - 1	$SS_R/(K-1)$
Within	SS_{res}	n - K	$SS_{res}/(n-K)$
Total	SS_T	n - 1	

Want to maximize SS_R

Test of hypothesis:

$H_0 : \mu_1 = ...\mu_K$ vs $H_1 : not\ H_0$ - not setting $\mu_i = 0$

F-test(right tail): $F = \frac{MS_R}{MS_Res} \sim F(K-1, n-K)$

Alternative approach:

$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, ..., K, j = 1, ...m$

- Y_{ij} is the jth observation for the ith treatment or factor level
- μ is a parameter common to all K treatments - usually called the grand mean
- τ_i - parameter that represents the effect of the ith treatment
- For simplicity all treatments have m experimental units

$$\mu = \beta_0 \tag{139}$$
$$\tau_1 = \beta_1 \tag{140}$$
$$... \tag{141}$$
$$\tau_K = \beta_k \tag{142}$$

Hypothesis Test: $H_0 : \tau_1 = ... = \tau_K = 0$, $H_1 : Not\ H_0$

F-test: $F = MS_R/MS_{Res} \sum F(K-1, n-K)$

Special case of ANOVA with categorical variables

7 Residual Analysis for Model Validation

Assume that Y_i are independent - assessed by context - need to conclude if it is reasonable to assume independence in context

Even if Y_i are independent, but take a sample of convenience - Y_is of sample may be correlated and not representative of the population under study. Can work other way - problem of dependence can vanish.

Suppose that right regression model is $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, but we instead decide to use SLRM incorrectly: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Calculating residuals, we get

$$e_i = Y_i - \hat{y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{143}$$
$$E[e_i] = E[Y_i] - E[\beta_0 + \beta_1 x_i] \tag{144}$$
$$= E[Y_i] - (\beta_0 + \beta_1 x_i) = E[\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i] - (\beta_0 + \beta_1 x_i) \tag{145}$$
$$= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 - (\beta_0 + \beta_1 x_i) \tag{146}$$
$$= \beta_2 x_i^2 \tag{147}$$

Where the real $E[Y_i] = E[\beta_0 + \beta_1 x_i + \beta_2 x_i^2]$. Residuals provide hints about model misspecification, or incorrect model specification - expected of residual is not 0 - problem - show misrepresentation about model.

If residual plot has a pattern and is not random - problem with model misspecification

Residual Analysis - used to check model adequacy - residuals are proxies of the errors:

$$e = (I - H)Y \tag{148}$$
$$e_i = Y_i - \hat{y}_i \tag{149}$$

Residuals are on the original units of the errors

To remove the scale(semi-standardize / semi-studentize):

$$s_i = \frac{e_i}{\hat{\sigma}} = \frac{e_i}{\sqrt{MS_{res}}} \tag{150}$$

Plain residuals / studentized residuals and their plots are useful for model assumptions

- Histogram - do residuals look normal
- QQ Plot - Is the QQ Plot on a straight line
- Test of normality(Shapiro / KS) - test reject for large n
- Residuals vs fitted values - expect to see random scatter of points around horizontal access - under MLR $e'\hat{y} = 0$, and e and $\hat{y}$ are independent from normal assumption
- Residuals vs Regressors

Autocorrelation - Correlation of variables between two time intervals(use time series to account for this)

Modified Levene Test - use medians instead of means

- Divide data in half using fitted values, obtaining a left and right side, compare the spread on the left with the one on the right(not needed that $n_1 = n_2$, as long as quantities are close)
- Find x median, separate raw residuals $\hat{e}_i$ in left and right groups, depending on where, relative to the median, the x coordinate lies

- Calculate medians of both groups of residuals, calling them $\tilde{e}_L and \tilde{e}_R$
- For left side residuals, calculate $d_i = |e_i - \tilde{e}_L$, while for right side residuals, calculate $d_j = |e_j - \tilde{e}_R|$
- So our data will be, for the left, $d_1, ..., d_{nL}$ and for the right, $d_1, ..., d_{nR}$
- Do a t-test on the d's, defining μ_L the mean of the left side of the d's and μ_R, the mean of he right side of the d's

$$\frac{\tilde{d_L} - \tilde{d_R}}{SE[\tilde{d_L} - \tilde{d_R}]} \sim t_{n-2} \tag{151}$$

$$\tilde{d_L} = \frac{1}{n_L}\sum_{i=1}^{n_L} d_i, \tilde{d_R} = \frac{1}{n_R}\sum_{i=1}^{n_R} d_i \tag{152}$$

$$Var[\tilde{d_L} - \tilde{d_R}] = (\frac{1}{n_L} + \frac{1}{n_R})\hat{\sigma^2}) \tag{153}$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{n_L}(d_i - \tilde{d_L})^2 + \sum_{j=1}^{n_R}(d_j - \tilde{d_R})^2}{n-2} \tag{154}$$

An MLR model is a valid model if

- The conditional mean of Y given X = x is a linear function of x(relationship must be linear): $E[Y|X = x] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$
- The conditional variance of Y given $X = x$ is constant(homogeneous): $Var[Y|X = x] = \sigma^2$
- Any plot of standarized resiuals against any predictor should show
  - A random scatter of points around the horizontal axis
  - Homogeneity
- Any observed pattern may indicate the fitted model is invalid

To be able to use the residual plots to diagnose the way where the model is misspecified, two conditions must hold:

- The conditional expectation of Y given X = x has to be a function of the linear regression function: $E[Y|X = x] = g(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$
- Linearity Condition: any x_i, regressed on another $x_j (i \neq j)$ has to be approximately linear: $E[X_i|X_j = x_j] \approx \alpha_0 + \alpha_1 x_j$
- If either condition does not hold, a pattern in a residual plot indicates that an incorrect model has been fit, but the pattern itself does not provide direct information on how the model is misspecified
- If conditions (1) and (2) hold, then the plot of Y against fitted values $\hat{Y}$ provides direct information about g

8 Outliers, Leverage Points, and Multicollinearity

Outlier - point(x_i, Y_i) that does not follow the bulk of the data in the y axis from the perspective of the model

- For small(< 30) to moderate($30 - 100$) sized datasets, outliers can have significant influence
- For large sized datasets - influence tends to dissipate

Leverage Point - point (x_i, Y_i) that may influence the model fit - points that have x_i far from $\bar{x}$

- A good leverage point - point that has x_i far from $\bar{x}$ but is not an outlier
- A bad leverage point - influences fit of the data that is also an outlier

Leverage - exposes the potential role of an individual data point:

$$\text{Leverage} = \frac{d\hat{y}_i}{dY_i} \tag{155}$$

- Idea(in SLR) - Let us perturb Y_i, a little bit at fixed x_i. How much do we expect $\hat{y}_i$ to change?
- If $\hat{y}_i$ changes as much as Y_i, then (x_i, Y_i) has the potential to drive the regression - so (x_i, Y_i) is leveraged
- if $\hat{y}_i$ hardly changes then clearly (x_i, Y_i) has no chance of driving the regression

Since $\hat{y} = HY$, and we let h_{ij} be the (i, j) entry of $H = X(X'X)^{-1}X'$. Then h_{ii} is the leverage of (x_i, Y_i)(elements on prinicipal diagonal

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} Y_j \tag{156}$$

$$h_{ii} = \frac{d\hat{y}_i}{dY_i} \tag{157}$$

Where x_i is the ith row of X:

$$h_{ii} = x_i'(X'X)^{-1}x_i \tag{158}$$

- Leverage of a point only depends on x_i. It is a standardized measure of the distance of x_i from the center of x space

For SLR**(Prove as exercise)**:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \tag{159}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \tag{160}$$

Properties of leverage**(Prove 2-4 as exercise)**:

- $0 \leq h_{ii} \leq 1$
- For fixed i: $\sum_{j=1}^{n} h_{ij} = 1$
- $\bar{h_{ii}} = \text{Ave}[h_{ii}] = \sum_{i=1}^{n} \frac{h_{ii}}{n} = (k+1)/n$
- For fixed $i : h_{ii} = \sum_{j=1}^{n} h_{ij}^2$
- It follows that $h_{ii} \approx 1 \implies h_{ij} \approx 0 \implies \hat{y_i} = y_i$. Thus if $h_{ii} \approx 1$, the fitted regression line(almost) goes through (x_i, y_i)
- A leverage point is considered one when $h_{ii} \geq 2\bar{h_i}i = \frac{2(k+1)}{n}$

Outliers - residuals are used for identifying problems with normality, constnacy of variance, linearity

- Semi-standardized(Semi-studentized) residuals allow the residuals to be compared on the "standard scale"
- However the residuals $e_i = y_i - \hat{y_i}$(and also semi-standardized residuals s_i) will be influenced if y_i is really leveraged as it will drag the regression line toward it
- Solution: use the idea of leave-one-out(or jackknifing) and calculate $e_i = y_{(i)} - \hat{y_{(i)}}$ where $\hat{y_{(i)}}$ is the fitted value at x_i excluding y_i.
- It can be proved that

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \tag{161}$$

- If we further standardize $e_{(i)}$ by removing its scale:

$$\frac{e_i}{\sigma^2(1 - h_{ii}} \tag{162}$$

- Note that $\sigma^2(1 - h_{ii})$ is actually the variance of e_i
- This leads to the (internally) studentized residuals

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}} \tag{163}$$

- The externally studentized residuals use the $\hat{\sigma}$ obtained by excluding the point i
- $r_i = \frac{e_{(1)}}{\hat{se}(e_i)} = \frac{e_i}{\hat{se}(e_i)}$
- $|r_i| > \begin{cases} 2 & \text{unusual} \\ 3 & \text{very unusual} \\ 4 & \text{point extremly unusual} \end{cases}$

- Many other tools to detect influence(Cook's DFFITS, DFBETAS)

Removing outliers - points that are outliers might not need to be removed - need to do a careful investigation

- If clerical error - can correct(or discard)
- If we do not have justification to remove it, we should leave it
- Can report results both leaving and removing it

Multicollinearity - existence of linear associations between covariates

- Simple - high correlation between two covariates
- Multivariate - high correlation between one covariate and a linear combination of other covariates

Consequences of multicollinearity:

- Magnitude of $\hat{\beta}$ will be biased(too big or too small) - difficult to detect unless we have prior expectation of the values of the coefficients
- $\hat{\beta}s$ can be of the opposite side to the one expected(positive when we expected negative and vice-versa)
- Variances of $\hat{\beta}$ will be too big, making the estimation unreliable - can test with variance inflation factor
- System becomes very sensitive to the inclusion of new data

Alternate approach - standardized multiple regression model - needed to obtain VIFs:

1. Helps to reduce unstable $(x^t x)^{-1}$, in case $|x^t x|$ is close to zero
2. Improves round-off errors
3. Allows for comparison of estimated regression coefficients in common units

- Standardization allow to uncover which regressors have a greater effect on Y_i, irrespective of the different units of measure used
- Standardized coefficients indicate how many standard deviations a dependent variable will change per standard deviation increase in a predictor variable(after accounting for the change in the rest of the predictors)

Correlation transformation:

- Transformation needed to obtain the standardized regression model
- Simple modification of the usual standardization of a variable

Standardizing involves:

- centering - different between variable and its mean
- scaling - dividing by its standard deviation

$$\frac{Y_i - \bar{Y}}{S_Y}, S_Y = \sqrt{\frac{1}{n-1}\sum (Y_i - \bar{Y})^2} \tag{164}$$

$$\frac{x_{ij} - \bar{x_j}}{S_{x_j}}, S_{x_j} = \sqrt{\frac{1}{n-1}\sum (x_{ij} - \bar{x_j})^2} \tag{165}$$

$$\tilde{Y_i} = \frac{1}{\sqrt{n-1}}\frac{Y_i - \bar{Y}}{S_Y} \tag{166}$$

$$\tilde{x_{ij}} = \frac{1}{\sqrt{n-1}}\frac{x_{ij} - \bar{x_j}}{S_{x_j}} \tag{167}$$

$$\tilde{Y_i} = \sum_{j=1}^{k} \tilde{\beta_j}\tilde{x_{ij}} + \tilde{\epsilon_i}, \tilde{\epsilon_i} \sim (iid)N(0, \tilde{\sigma^2} \tag{168}$$

1. $\tilde{\beta_0} = 0$
2. $\tilde{x}$ is $n \times k$
3. $\beta_j = \frac{S_Y}{S_{x_j}}\tilde{\beta_j}$
4. $\beta_0 = \hat{Y} - \sum_{j=1}^{k} \beta_j \bar{x_j}$

Let r_{xx} be the correlation matrix of the x variables. This matrix has as its elements the coefficients of simple correlation between all pairs of the x variables - if determinant is close to 0, very likely that model is bad: $\begin{bmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & ... & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & ... & 1 \end{bmatrix}$, where $r_{hj} = r(x_h, x_j)$

r_{xx} is symmetric$(r_{hj} = r_{jh})h : 1, ..., k, j : 1, ...k$

Let r_{Yx} be a vector containing the coefficientso f simple correlation between Y and each x variable: $\begin{bmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & ... & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & ... & 1 \end{bmatrix}$ Properties(**Show as exercise**:

1. $\tilde{x^t}\tilde{x} = r_{xx}$
2. $\tilde{x^t}\tilde{Y} = r_{Yx}$

, where $\tilde{x}$ and $\tilde{Y}$ are made of $\tilde{x_{ij}}$ and $\tilde{Y_{ij}}$ respectively. Then using the normal equations:

$$\tilde{x}^t\tilde{x}\hat{\tilde{\beta}} = \tilde{x^t}\hat{Y} \tag{169}$$

$$r_{xx}\hat{\tilde{\beta}} = r_{Yx} \tag{170}$$

$$\hat{\tilde{\beta}} = r_{xx}^{-1}r_{Yx} \tag{171}$$

- Elements of $\hat{\tilde{\beta}}$ are called the estimators of the standardized regression coefficients
- The estimators of the coefficients of the untransformed regression can be obtained as:

$$\hat{\beta}_j = \frac{S_Y}{S_{x_j}}\hat{\tilde{\beta}}_j, j = 1, ..., k \tag{172}$$

$$\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^{k}\hat{\beta}_j\bar{x_j} \tag{173}$$

Since $Cov[\hat{\beta}] = \sigma^2(x^tx)^{-1}$, $Cov[\hat{\tilde{\beta}}] = \tilde{\sigma^2}r_{xx}^{-1}$

For $k = 2 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, where $|r_{xx}| = 1 - r_{12}^2$ and $r_{xx} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$ and $r_Y = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix}$

As $|r_{12}|$ increases, the variance of the $\hat{\tilde{\beta}}_k$ get inflated by the factor $\frac{1}{1-r_{12}^2}$, found in the principal diagonal of r_{xx}^{-1} and it is called a Variance inflation factor(VIF)

Generalizing for $k > 2$ predictors, the elements of the principal diagonal of r_{xx}^{-1} are $\frac{1}{1-R_j^2}$(extra credit), where R_j^2 is the coefficient of multiple determination when x_j is regressed on the $k-1$ other x variables in the model - the elementso f the principal diagonal of r_{xx}^{-1} are called Variance Inflation Factors, or VIF:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{174}$$

- If $VIF_j = 1$, x_j is not linearly related to the other $x's$ and the variance is not inflated
- In practice, $VIF_j > 5$ is considered an important level of multicollinearity
- If the maximum VIF_j of a given model exceeds 10 it is an indication that multicollinearity may be adversely influencing the least squares estimates
- There will always be some level of inflation - the importance is establishing the level of it - multicollinearity is related to the actual values of the regressors
- Uncorrelated variables may have correlated data

9 Model Selection

- $R^2 = \frac{SS_R}{SS_T}$, increasea as p increases, whether or not the added predictors are linearly associated with Y or not
- When $p = n$, then $R^2 = 1$, where p is the number of predictors
- Do not use R^2 to compare models with different number of parameters, use R^2_{adj} instead

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-k-1)}{SS_T/(n-1)} = 1 - \frac{MS_{res}}{MS_T} \tag{175}$$

- Suitable to compare models with different number of predictors, as the inclusion of new parameters, that tends to increase R^2 is compensated by penalizing such addition
- R^2_{adj} has the mathematical oddity that it may result in a negative number - only for very bad models

$$AIC = [-2lnL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \sigma^2|Y) + 2K] \tag{176}$$

- L is likelihood, K is the penalization term to account for the addition of parameters - for MLE, one of the estimated parameters is σ^2, $K = k + 2$, where k is the number of parameters, for MLR it reduces to:

$$AIC = nln\frac{SS_{res}}{n} + 2k + constant \tag{177}$$

- The smaller the AIC, the better

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \tag{178}$$

- AIC_c is recommended over AIC - correct bias incurred by AIC - AIC_c penalizes models with more parameters heavily than AIC
- Comparing differences between removing and adding parameters
- Bayesian Information Criterion:

$$BIC = [-2lnL(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p, \sigma^2|Y) + Kln(n)] \tag{179}$$

- Mallow's C_p - criterion is related to the total MSE of the n fitted values for each subset regression model
- Total Mean Squared Error(TMSE):

$$TMSE = \sum_{i=1}^{n} MSE[y_i] = \sum_{i=1}^{n} E[(\hat{y}_i - \mu_i)^2] \tag{180}$$

- $\mu_i = E[Y_i|X = x_i]$
- The criterion is defined as:

$$\Gamma_p = \frac{TMSE}{\sigma^2} \tag{181}$$

$$C_p = \frac{SS_{resp}}{\hat{\sigma^2_{full}}} - (n - 2(k+1)) \tag{182}$$

- SS_{resp} corresponds to the model with the subset of k while $\hat{\sigma^2_{full}}$ corresponds to the model with all available covariates - p corresponds to the number of parameters(k)
- Variant: $C_p = \frac{SS_{resp}}{\hat{\sigma^2_{full}}} + 2k$. The one on the previous slides centers the C_p value along zero for the full model, simplifying interpretation
- Want C_p to be as small as possible - good models are ones where C_p is close to $p = k$
- C_p/p indiates a poor model(leaving out an important variable) - should correspond to rejecting a partial F test
- C_p relies on the full model - if an incorrect model, will provide incorrect assessments
- If $k > n$, full model cannot be specified, the C_p cannot be obtained

Determinant of Correlation, matrix of covariates: $\begin{bmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & ... & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & ... & 1 \end{bmatrix}$, $r_jh = r(x_j, x_h), r_xx$ is symmetric

- DET is the determinant of the correlation matrix: $DET = |rxx|$ - models with $DET < 0.1$ are usually considered poor ones - do not use DET to compare models, only for discarding poor ones
- Leave out one data-point, fit corresponding model, use x coordinate to find the predicted Y, and find the error of the prediction - coincides with the already calculated deleted resisuals $e_{(i)}$ - squaring each $e_{(i)}$ and adding them gives you the PRESS:

$$PRESS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\frac{e_i}{1 - h_{ii}})^2 \tag{183}$$

- PRESS measures the predicted power of the model - could also use k-fold validation instead
- All possible subsets: Approach considers all $2^k - 1$ possible regression models(2^p if also considering the model with only β_0) with the aim of finding the model that is better according to some criterion

- For fixed number of terms - $R^2_{adj}, AIC, AIC_c, BIC$ will agree that the best choice is the model with smallest SS_{res}

Forward Selection:

- Step 1:
  - Fit a SLR for each of the k potential x variables
  - For each fit a p-value(ort statistic) is obtained
  - Select the variable that has lower p-value(or higher t statistic) if it is below a specified cutting p-value(or above a specified t statistic)
  - If no variable can be selected, stop
- Step 2:
  - If a variable has been selected(e.g. x_7), fit now all the possible $k-1$ two variable models where one is always x_7
  - Repeat the selection as in Step 1 by looking only at the p-values(or t statistics) of the added variables
  - Continue adding variables which cutting values are met, or stop otherwise
- Forward Stepwise regression:
  - After step 2, once you have added a new variable, check if other variables now should be dropped for being above the cutting p-value(or below the cutting t statistic)
- Stepwise regression algorithm allows for an x variable brought into the model at an earlier stage to be dropped subsequently if it is no longer helpful in conjunction with variables added at later stages

Backwards Elimination:

- Start with the model with al lvariables and eliminate the ones with the higher p-value one at a time, until no non-significant variables are left
- Backwards Stepwise Regression - as backwards elimination but allow already eliminated variables to enter the model again
- For small and moderate number of variables, some statisticians argue for backwards stepwise instead of forward stepwise, based on the MSE tending to be inflated during the initial steps

Inference After variable selection:

- Selection process changes the properties of the estimators(biased) and the tests(lower p-values than real) and confidence intervals(smaller than real), as it happens in cases of multiple comparisons - we are "fishing" for a model