Stats 203: Introduction to Regression Models and Analysis of Variance

Samuel Wong
Department of Statistics
Stanford University

Abstract

The course starts with the theoretical underpinnings of regression. The assumptions of linear regression are visited and the relationship between hypothesis testing and confidence intervals are explored. Distribution theory is examined. The course then transitions into diagnostics and methods for checking the assumptions for linear regression and whether or not they hold in practice. Continuous and discrete predictors are looked at and therefore ANOVA (one-way and two-way) are discussed. Lastly, discrete responses are examined in the framework of General Linear Models (Binomial and Poisson Regression).

Contents

1 Statistics Review

We want to compare two scenarios: one where we have data points (x_i, y_i), where the x is fixed, and a bivariate (X, Y), where the X is random. An example of when we would model using a fixed x, is if we were doing an experiment, and we controlled the values of x and wanted to observe what the outcome Y would be. An example of when we would model using a random variable X would be if we were observing or sampling data, and just measuring the X.

1.1 Formulas

1.1.1 Sample Summary Statistics

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- $s_x = \sqrt{var(x)}$
- $\bar{y}; s_y$ with analogous formulas to the x
- $r = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{s_x} * \frac{(y_i-\bar{y})}{s_y} \in [-1, 1]$. Note that this is an inner product between standardized x, y

1.1.2 Linear Regression

Least squares regression takes the formula $y = \hat{a} + \hat{b}x$ where

- $\hat{a} = \bar{y} - \hat{b}\bar{x}$ (note that the regression line runs through the point that is the average of x and average of y)
- $\hat{b} = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

The squared vertical distance between the points and the line is denoted $S(a, b) = \frac{1}{n}(y_i - a - bx_i)^2$. From a theorem by Gauss, least squares minimizes $S(a, b)$.

- Residuals (realized): $e_i = y_i - \hat{a} - \hat{b}x_i$
- MSE: $S(\hat{a}, \hat{b}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = (1 - r^2)var(y)$
- RMSE: $\sqrt{MSE}$ (used to keep original scale)

1.2 Models

If we model using a fixed x, we have the formula $Y_i = a + bx_i + \epsilon_i$, where the ϵ_i are iid with mean 0 and variance σ^2. Note that we observe x_i and Y_i, but do not observe a, b, ϵ_i. When we fit a model, our fitted model is $Y_i = \hat{a} + \hat{b}x_i + e_i$. Note that we observe all these values in the fitted model.

1.3 Normal Distribution

For a p-dimensional normal distribution, $X \sim N_p(\mu, \Sigma)$. We view

$$X_p = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

Our $\mu = E[X_p]$, and $\Sigma = Cov(X_p) = E[(X - \mu)(X - \mu)^T]$, which is a positive semi-definite matrix.

Specifically, when $p = 2$, the bivariate case, where we have

$$X_p = \begin{pmatrix} X \\ Y \end{pmatrix}; \mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}; \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_y \\ \rho\sigma_X\sigma_y & \sigma_Y^2 \end{pmatrix}$$

$\rho = Corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$

We have a theorem that says if X_p is normal with $p = 2$, then the conditional distribution of $Y|X = x$ is a univariate normal with

- $E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$
- $Var[Y|X = x] = \sigma_Y^2(1 - \rho^2)$

As we vary x, we can see that we get a line from $E[Y|X = x] = \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$. The slope of this line is $\beta = \rho\frac{\sigma_Y}{\sigma_X}$ and the intercept is $\alpha = \mu_Y - \beta\mu_X$, which is the same as what we got from least squares!

Note that we can get the same result assuming random variable X by minimizing $\mathcal{S}(a, b) = E[Y - a - bX]^2$. We obtain that $a^* = \alpha = \mu_Y - \beta\mu_X$ and $b^* = \beta = \rho\frac{\sigma_Y}{\sigma_X}$. Note that this minimization problem does not require the Gaussian assumption.

1.4 Regression Effect

Regression toward the mean can be understood as higher deviations from the mean can on average result in lower deviations from the mean on subsequent trials (simply because the probability of two large deviations is unlikely).

Mathematically, let $X = x = \mu_X + K\sigma_X$, where K is a constant. Then

$E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$
$= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(\mu_X + K\sigma_X - \mu_X)$
$= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(K\sigma_X)$
$= \mu_Y + \rho\sigma_Y K$

Since ρ is less than or equal to 1, then on average, Y is closer to its mean that X is to its mean. An example of this is with the height of fathers and sons. If a father is much taller compared to the average height of the fathers, then the son on average is less tall compared to the average height of the sons.

1.5 Square Root Law

$E[\bar{U}] = \mu$
$Var[\bar{U}] = \frac{\sigma^2}{n}$
$SE[\bar{U}] = \frac{\sigma}{\sqrt{n}}$

The square root law states that in order to decrease the Standard Error by X, we need to increase the sample size by X^2. For example, to reduce the standard error by $\frac{1}{2}$, we need to increase the sample size by 4.

2 Linear Regression

For a univariate linear regression, we can write it out as $Y_i = X_i\beta + \epsilon_i$, if we want the per observation view, or in matrix notation as $Y = X\beta + \epsilon$.

2.1 Least Squares

We find the optimal solution by solving the least squares problem. We want to minimize $||Y - X\beta||_2^2$, which is equivalent to minimizing $(Y - X\beta)^T(Y - X\beta) = \beta^T X^T X\beta - 2Y^T X\beta + Y^T Y$, which is a quadratic. By differentiaing with respect to β and setting it equal to 0, we get what are called the "normal equations" for the OLS estimator:

$X^T X\hat{\beta} = X^T Y$

Therefore, $\hat{\beta} = (X^T X)^{-1}X^T Y$. We also have that $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1}X^T Y = HY$. We call the matrix $X(X^T X)^{-1}X^T$ the "hat matrix." Note that H is idempotent and symmetric, so it is a projection matrix. In fact, it projects the values of y onto the column space of X.

Location Model: $Y = \mu + \epsilon$. This model has only a column of 1's as the X value. $\hat{\mu} = (X^T X)^{-1}(X^T Y)$ equals $\frac{1}{n}\sum_{i=1}^n Y_i = \bar{Y}$, the sample mean.

For univariate linear regression, we reparameterize $a + bx_i$ to $\beta_1 + \beta_2(x_i - \bar{x})$. After the orthogonalizing reparameterization, the design matrix X becomes

$$X = \begin{bmatrix} 1 & x_1 - \bar{x} \\ \vdots & \dots \\ 1 & x_n - \bar{x} \end{bmatrix}$$

Since the columns of X are now orthogonal, we have

$$X^T X = \begin{bmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2 \end{bmatrix}; X^T Y = \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n [(x_i - \bar{x})Y_i] \end{bmatrix}$$

Therefore, our OLS estimate is

$$\hat{\beta} = (X^T X)^{-1}X^T Y = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ \sum_{i=1}^n [(x_i - \bar{x})Y_i] \end{bmatrix}$$

Therefore, we have that $\hat{\beta_1} = \bar{Y}$ and $\hat{\beta_2} = \frac{\sum_{i=1}^n [(x_i - \bar{x})Y_i]}{\sum_{i=1}^n (x_i - \bar{x})^2}$

2.2 Geometric Interpretation

$X^T X\hat{\beta} = X^T Y$
$X^T(Y - X\hat{\beta}) = 0$
$X^T(e) = 0$

What we see is that the error vector e is orthogonal to all the columns of X. Therefore, using the model $Y = \hat{Y} + e$, we have that $||Y||_2^2 = ||\hat{Y}||_2^2 + ||e||_2^2$, due to the orthogonality shown above. This is called the sum of squares decomposition.

2.3 Variance (Explained vs Unexplained)

We claim that $var(y) = var(\hat{y}) + var(e)$ if there is an intercept term in the model. We see this as follows:

$var(y) = var(\hat{y} + e)$
$= var(\hat{y}) + var(e) + 2cov(\hat{y}, e)$

Looking at the $cov(\hat{y}, e)$ term, we have that $cov(\hat{y}, e) = \frac{1}{n}\sum_{i=1}^n (\hat{y_i} - \bar{\hat{y}})(e_i - \bar{e})$. We know that e is orthogonal to X, and namely orthogonal to the intercept column of X, the vector of all 1's. Therefore, $\sum_{i=1}^n e_i = 0$, so $\bar{e} = 0$. Therefore we have

$\frac{1}{n}\sum_{i=1}^n (\hat{y_i} - \bar{\hat{y}})(e_i - \bar{e})$
$= \frac{1}{n}\sum_{i=1}^n (\hat{y_i} - \bar{\hat{y}})(e_i)$
$= \frac{1}{n}\sum_{i=1}^n (\hat{y_i}e_i - \bar{\hat{y}}\bar{e})$
$= \frac{1}{n}\sum_{i=1}^n (\hat{y_i}e_i)$

$= \bar{e}\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i) = 0$

Therefore, we have shown that $var(y) = var(\hat{y}) + var(e)$. The term $var(\hat{y})$ is called the explained variance and $var(e)$ is called the unexplained variance. Intuitively, what we are doing is that we are decomposing the variance of y into what we are able to explain with the predictors that we have and what we cannot. A measure of how good the model performs is $R^2 = \frac{var(\hat{y})}{var(y)}$.

In this framework, $var(y)$ is called the TSS (total sum of squares) and $var(e)$ is called the RSS (residual sum of squares). $R^2 = \frac{var(\hat{y})}{var(y)} = corr^2(y, \hat{y}) = \frac{cov^2(y,\hat{y})}{var(y)var(\hat{y})}$. Note that in the simple regression case (1 predictor), that $R^2 = r^2 = corr^2(x, y) = \frac{cov^2(x,y)}{var(x)var(y)}$.

Note that we can also view $R^2 = 1 - \frac{var(e)}{var(y)} = 1 - \frac{RSS}{TSS}$.

Note that the R^2 can only increase as we add more variables to the model. Therefore, to measure model performance, we can use the adjusted-R^2. The adjusted R^2 is defined as $R_d^2 = 1 - \frac{\frac{RSS}{n-p}}{\frac{TSS}{n-1}} = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$.

2.4 Stochastic Assumptions

We assume the following:

- (LM) Model is $Y_i = X_i\beta + \epsilon_i$
- (E) ϵ_i are iid; $E[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$
- (RX) If X is random, assume X is independent of ϵ ($X \perp\!\!\!\perp \epsilon$)

Theorem: If we have the assumptions above and that $n > p$ and X is full rank, then

1. $E[\hat{\beta}|X] = \beta$
2. $Cov[\hat{\beta}|X] = \sigma^2(X^TX)^{-1}$
3. $E[\hat{\sigma^2}|X] = \sigma^2$

To prove this theorem we first have the following:

$\hat{\beta} = (X^TX)^{-1}X^TY$
$= (X^TX)^{-1}X^T(X\beta + \epsilon)$
$= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\epsilon$
$= \beta + (X^TX)^{-1}X^T\epsilon$
$= \beta + A_X\epsilon$, where we let $A_X = (X^TX)^{-1}X^T$

To prove 1., we have

$E[\hat{\beta}|X] = E[\beta + A_X\epsilon|X]$
$= E[\beta|X] + E[A_X\epsilon|X]$
$= \beta + A_XE[\epsilon|X]$, note that we can pull the A_X out since the matrix made of only X
$= \beta + A_XE[\epsilon]$, since by (RX) we know ϵ is independent of X
$= \beta + A_X * 0$, since by (E) we know $E[\epsilon] = 0$
$= \beta$

To prove 2., we have

$Cov[\hat{\beta}|X] = Cov[\beta + A_X\epsilon|X]$
$= Cov[A_X\epsilon|X]$
$= A_XCov[\epsilon|X]A_X^T$
$= A_X\sigma^2I_nA_X^T$
$= \sigma^2A_XA_X^T$
$= \sigma^2[(X^TX)^{-1}X^T][(X^TX)^{-1}X^T]^T$
$= \sigma^2[(X^TX)^{-1}X^TX(X^TX)^{-1}]$
$= \sigma^2(X^TX)^{-1}$

To prove 3., we first need to understand the Hat matrix H.

2.4.1 Hat Matrix

When we fit a model, we have $\hat{Y} = HY$, where $H = X(X^TX)^{-1}X^T$. We can then write the residual as $e = Y - \hat{Y} = (I - H)Y$.

We know the following about H:

- H is an orthogonal projection, which means
  - $H = H^T$ (symmetric)
  - $H^2 = H$ (idempotent)
  - $(I - H)$ is also an orthogonal projection
- $tr(H) = p$, the number of parameters of the model
  - This can be shown because $tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_p) = p$
  - $tr(I_n - H) = n - p$ (residual degrees of freedom)
- Residuals are less variable than the errors. If we denote $\tilde{H} = I - H$, we have:
  1. $e = \tilde{H}\epsilon$
     - Proof: $e = Y - \hat{Y} = (I - H)Y = (I - H)(X\beta + \epsilon) = (I - H)\epsilon = \tilde{H}\epsilon$
     - We can get from $(I - H)(X\beta + \epsilon)$ to $(I - H)\epsilon$, since we know $HX = X$, so $(I - H)X = 0$
  2. $||e||^2 = \epsilon^T\tilde{H}\epsilon$
     - Proof: $||e||^2 = e^Te = \epsilon^T\tilde{H}^T\tilde{H}\epsilon = \epsilon^T\tilde{H}\tilde{H}\epsilon = \epsilon^T\tilde{H}\epsilon$
     - For the proof, we used that fact that $\tilde{H}$ is a projection matrix (symmetric and idempotent)

We can see from our assumption (E) that $\sigma^2 = \frac{1}{n}E[\sum_{i=1}^n \epsilon_i^2] = \frac{1}{n}E[||\epsilon||^2]$. We claim that the unbiased estimate is $\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^n e_i^2$, also called the Residual Mean Square (RMS).

We can prove that $E[\hat{\sigma^2}] = \sigma^2$. It follows that $E[\hat{\sigma^2}|X] = \sigma^2$.

$||e||^2 = \epsilon^T\tilde{H}\epsilon$
$= tr(\epsilon^T\tilde{H}\epsilon)$, since it is a scalar
$= tr(\tilde{H}\epsilon\epsilon^T)$

Taking expectations of both sides,
$E[||e||_2^2] = E[tr(\tilde{H}\epsilon\epsilon^T)]$
$= tr(\tilde{H}E[\epsilon\epsilon^T])$
$= tr(\tilde{H}\sigma^2 I_n)$, since $E[\epsilon\epsilon^T] = \sigma^2 I_n$
$= \sigma^2 tr(\tilde{H})$
$= \sigma^2(n - p)$

Therefore, $\frac{E[||e||^2]}{n-p} = \sigma^2$, which proves that $E[\hat{\sigma^2}] = \sigma^2$.

3 Hypothesis Testing

In testing, such as the t-test or the F-test, on top of the previous assumptions we made (LM, E, RX), we make a further assumption that the errors are Gaussian (G). We assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

3.1 Standard Errors

To perform our tests, we need to calculate standard errors. We have that $SE(\hat{\beta}_i) = \sigma\sqrt{[(X^TX)^{-1}]_{ii}}$. In practice, we need to estimate this standard error, so the estimated standard error is $\hat{SE}(\hat{\beta}_i) = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{ii}}$

3.2 T-Testing

We first define what a Student's t distribution is. Assume we have iid standard normal distributions $Z, U_1, ..., U_d$. We define a Chi-square distribution with d degrees of freedom as $\chi^2_d = \sum_{i=1}^{d} U_i^2$. Denote $V = \chi^2_d$ for simplicity. $E[V] = d$. A t distribution with d degrees of freedom is defined as $t_d = \frac{Z}{\sqrt{\frac{V}{d}}}$.

As previously stated, we assume (LM), (E), (RX), and (G) for t-testing. In t-testing, we are testing whether ONE parameter and whether its coefficient is 0 or not (does that coefficient need to be in the model?) Therefore, our hypothesis for parameter k is

- $H_0 : \beta_k = 0$
- $H_A : \beta_k \neq 0$

Based off of our data, we calculate $t_k^{obs} = \frac{\hat{\beta}_k}{\hat{SE}(\hat{\beta}_k)}$. We reject the H_0 for large values of $|t_k^{obs}|$. We calculate the p-value (using a table). The p-value $= 2 * P(t_{n-p} > |t_k^{obs}|)$. What the p-value tells us is that if the null hypothesis were true, this value is the probability of us observing a value of $|t_k^{obs}|$ or more extreme.

Note that the t-test tests only one variable at a time, and that it is an added variable test. The variable being tested depends on the other variables in the model, so if two variables were correlated, both may not be significant when put together in the model, but if we remove either, then the other one becomes significant.

3.3 F-Testing

For F-testing, we are testing the significance of a group of variables at once. The hypotheses are:

- $H_\omega : \beta_i = 0;\ i = q+1, ..., p$
- H_Ω : some $\beta_i \neq 0;\ i = q+1, ..., p$

The three steps to perform the F-test are:

- Fit $Y = X\beta + \epsilon$ by minimizing $||Y - X\beta||^2$ (the full model). By solving this, we obtain these $\hat{\beta}$, and residuals $e_\Omega = Y - X\hat{\beta}$, and we have $n - p$ degrees of freedom
- Fit $Y = X_\omega\beta_\omega + \epsilon$ by minimizing $||Y - X_\omega\beta_\omega||^2$ (the smaller model). By solving this, we obtain these $\hat{\beta_\omega}$, and we have $p - q$ hypothesis degrees of freedom
- Calculate the F-statistic as $F_{p-q,n-p} = \frac{\frac{||X\hat{\beta} - X_\omega\hat{\beta_\omega}||^2}{p-q}}{\frac{||e||^2}{n-p}}$

We reject the null hypothesis when the F-statistic is too large. The F-distribution is only supported for non-negative values. Therefore, the p-value is $P(F_{p-q,n-p} \geq F_{obs})$ under the null model H_ω. Note that the F-distribution can also be described as a function of Chi-squared distributions. $F_{p-q,n-p} = \frac{\frac{U}{p-q}}{\frac{V}{n-p}}$ where $U \sim \chi^2_{p-q}$ and $V \sim \chi^2_{n-p}$ and U and V are independent.

3.4 Sum of Squares Decomposition

We first claim that the three vectors $e = Y - X\hat{\beta}, X\hat{\beta} - X\hat{\beta_\omega}, X\hat{\beta_\omega}$ are orthogonal.

We can prove this since we know that e is orthogonal to $X\gamma$ for any γ. This is because e is orthogonal to the column space of X. Therefore, e is orthogonal to the choices of $\gamma = \hat{\beta}$ and $\gamma = \hat{\beta_\omega}$. Therefore e is orthogonal to $X\hat{\beta} - X\hat{\beta_\omega}$ and $X\hat{\beta_\omega}$.

Now the only thing left to show is that $X\hat{\beta} - X\hat{\beta_\omega}$ and $X\hat{\beta_\omega}$ are orthogonal. We rewrite $X\hat{\beta} - X\hat{\beta_\omega} = (Y - X\hat{\beta_\omega}) - (Y - X\hat{\beta})$. When we fit the smaller model, by definition, $(Y - X\hat{\beta_\omega})$ is orthogonal to $X\hat{\beta_\omega}$. We just proved that $e = Y - X\hat{\beta}$ is orthogonal to $X\hat{\beta_\omega}$. Therefore, we conclude that $X\hat{\beta} - X\hat{\beta_\omega}$ and $X\hat{\beta_\omega}$ are orthogonal.

If we draw the triangle with the three vertices being $X\hat{\beta_\omega}$, $X\hat{\beta}$, and Y, using the Pythagorean Theorem, we have that

$||Y - X\hat{\beta_\omega}||^2 = ||Y - X\hat{\beta}||^2 + ||X\hat{\beta} - X\hat{\beta_\omega}||^2$
$RSS_\omega = RSS_\Omega + ||X\hat{\beta} - X\hat{\beta_\omega}||^2$
$RSS_\omega \geq RSS_\Omega$

We can rewrite our F-statistic as $F = \frac{\frac{RSS_\omega - RSS_\Omega}{p-q}}{\frac{RSS_\Omega}{n-p}}$.

As a corollary, we can express $Y = X\hat{\beta_\omega} + X\hat{\beta} - X\hat{\beta_\omega} + Y - X\hat{\beta}$. Therefore, due to orthogonality, the sum of squares decomposition is

$||Y||^2 = ||X\hat{\beta_\omega}||^2 + ||X\hat{\beta} - X\hat{\beta_\omega}||^2 + ||Y - X\hat{\beta}||^2$

4 Distribution Theory

We want to understand why we have the assumption that ϵ is normally distributed, and how that relates to the t-distributions and F-distributions. We need to first understand the properties of Multivariate Normals.

4.1 Multivariate Normal

Three facts about multivariate normal distribution. Let $X \sim N_p(\mu, \Sigma)$, so a p-dimensional vector.

1. For any matrix A (dimension $q \times p$) that is fixed, $AX \sim N_q(A\mu, A\Sigma A^T)$
2. We know that independence always implies orthogonality. However, if $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ has a joint normal distribution (multivariate), then the statement becomes an iff statement. $X_1 \perp\!\!\!\perp X_2 \Leftrightarrow Cov(X_1, X_2) = 0$
3. - If $Z \sim N_n(0, I)$ and H is an orthogonal projection, then $Z^T H Z \sim \chi^2_d$, with $d = tr(H)$
   - If $\tilde{H}$ is another orthogonal projection with $H\tilde{H} = 0$, then $Z^T \tilde{H} Z \sim \chi^2_{\tilde{d}}$, with $d = tr(\tilde{H})$. This $\chi^2_{\tilde{d}}$ is independent of χ^2_d

4.2 Distribution of $\hat{\beta}$ and e

We claim the following facts about $\hat{\beta}$ and e:

1. $\hat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$
2. $||e||^2 \sim \sigma^2 \chi^2_{n-p}$
3. $\hat{\beta} \perp\!\!\!\perp e$

To prove 1., we write $\hat{\beta} = A_X Y$, where $A_X = (X^T X)^{-1} X^T$. Note that A_X is fixed (only relies on the matrix X, which we assume is fixed). Y is a multivariate normal, since we assume the errors are normal and all else is fixed. We use the multivariate normal fact 1., and therefore conclude that $\hat{\beta}$ is also normally distributed. Previously, we showed that $E[\hat{\beta}] = \beta$, and $Cov[\hat{\beta}] = \sigma^2(X^T X)^{-1}$, so $\hat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$.

To prove 2., recall that $e = \tilde{H}\epsilon$. Let $Z \sim N_n(0, I)$, then $\epsilon = \sigma Z$. Therefore,

$||e||^2 = e^T e$
$= (\tilde{H}\epsilon)^T(\tilde{H}\epsilon)$
$= (\sigma\tilde{H}Z)^T(\sigma\tilde{H}Z)$
$= \sigma^2 Z^T \tilde{H}^T \tilde{H} Z$
$= \sigma^2 Z^T \tilde{H} Z$

We can now use the multivariate normal fact 3. $tr(\tilde{H}) = n - p$, so we have that $Z^T \tilde{H} Z \sim \chi^2_{n-p}$. Therefore, we conclude that $||e||^2 \sim \sigma^2 \chi^2_{n-p}$.

To prove 3., we write $\hat{Y} = HY$, and $e = \tilde{H}Y$. Both $\hat{Y}$ and e are normally distributed since we assume H and $\tilde{H}$ are fixed (since they only depend on X. Therefore, we have

$Cov(\hat{Y}, e) = Cov(HY, \tilde{H}Y)$
$= HCov(Y)\tilde{H}$
$= H\sigma^2 I_n \tilde{H}$
$= \sigma^2 H\tilde{H}$
$= \sigma^2 H(I - H) = 0$

Since $\hat{Y}$ and e are normally distributed, we can use the multivariate normal fact 2., and conclude that $\hat{Y}$ and e are independent. Since $\hat{Y}$ is only a function of $\hat{\beta}$, we conclude that $\hat{\beta}$ and e are independent.

4.3 T Distributions and F Distributions

In a t-test, our null hypothesis is that $H_0 : \beta_k = 0$. Using the first claim about the distribution of $\hat{\beta}$, we have that $\hat{\beta}_k \sim N_p(0, \sigma^2(X^T X)^{-1}_{kk})$, since $E[\hat{\beta}_k] = \beta_k = 0$ under the null hypothesis. Recall that we calculate $\hat{SE}(\hat{\beta}_k) = \hat{\sigma}\sqrt{(X^T X)^{-1}_{kk}}$.

Therefore, we have

$t_k = \frac{\hat{\beta}_k}{\hat{SE}(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{\hat{\sigma}\sqrt{(X^TX)^{-1}_{kk}}} = \frac{\frac{\hat{\beta}_k}{\sigma\sqrt{(X^TX)^{-1}_{kk}}}}{\frac{\hat{\sigma}\sqrt{(X^TX)^{-1}_{kk}}}{\sigma\sqrt{(X^TX)^{-1}_{kk}}}} \overset{D}{=} \frac{Z}{\sqrt{\frac{\hat{\sigma^2}}{\sigma^2}}}$, where $Z \sim N(0,1)$

We then note that

$\frac{\hat{\sigma^2}}{\sigma^2} = \frac{\frac{||e||^2}{n-p}}{\sigma^2} \overset{D}{=} \frac{V}{n-p}$, where $V \sim \chi^2_{n-p}$.

Therefore, $t_k \overset{D}{=} \frac{Z}{\sqrt{\frac{V}{n-p}}}$

Since Z is a function of $\hat{\beta}$ and V is a function of e, we know that they are independent by the third claim of the distribution of e. Therefore, under the null hypothesis, t_k has a t_{n-p} distribution.

Similarly, under the alternative hypothesis, where we have $H_A : \beta_k \neq 0$, call it δ_k, we would have $t_k \overset{D}{=} \frac{N(\delta_k,1)}{\sqrt{\frac{V}{n-p}}}$.

For the F-test, we will need to recall the orthogonality conditions. We can draw out a right triangle with points at $Y, \hat{Y} = HY$, and $\hat{Y}_\omega = H_\omega Y$. As a reminder, $H = X(X^TX)^{-1}X^T$ and $H_\omega = X_\omega(X_\omega^TX_\omega)^{-1}X_\omega^T$. We rewrite two of the sides of the triangle as $Y - \hat{Y} = \tilde{H}Y$, and $\hat{Y} - \hat{Y}_\omega = (H - H_\omega)Y = H_\Delta Y$, where $H_\Delta = H - H_\omega$.

Recalling the orthogonality of the F-test, we have that $Y = H_\omega Y + H_\Delta Y + \tilde{H}Y$. Therefore, we have three orthogonal projections, H_ω, H_Δ, and $\tilde{H}$, with corresponding traces of $q, p-q$, and $n-p$. We can also see that H_Δ is orthogonal to X_ω.

We recall that the F-distribution is $F_{p-q,n-p} = \frac{\frac{||X\hat{\beta} - X_\omega\hat{\beta_\omega}||^2}{p-q}}{\frac{||e||^2}{n-p}} = \frac{\frac{||\hat{Y}-\hat{Y}_\omega||^2}{p-q}}{\frac{||e||^2}{n-p}}$. Looking at the numerator, we want to understand $\hat{Y} - \hat{Y}_\omega$. We have that $\hat{Y} - \hat{Y}_\omega = H_\Delta Y = H_\Delta(X_\omega\beta_\omega + \epsilon)$. We know that H_Δ is orthogonal to X_ω, so this quantity equals $H_\Delta\epsilon$. Using fact 2, we have that $||\hat{Y} - \hat{Y}_\omega||^2 = ||H_\Delta\epsilon||^2 = \epsilon^T H_\Delta \epsilon \sim \sigma^2\chi^2_{n-p}$, since $tr(H_\Delta) = n - p$.

Looking at the denominator, we have that $||e||^2 \sim \sigma^2\chi^2_{n-p}$. We know that the numerator and the denominator are independent, so we have $F \overset{D}{=} \frac{\frac{\chi^2_{p-q}}{p-q}}{\frac{\chi^2_{n-p}}{n-p}}$, which is the definition of the F distribution.

5 Confidence Intervals

Confidence intervals use the same assumptions as for testing, namely the (LM), (E), (G), (RX). In fact, there is a duality between confidence intervals and testing.

A 2-sided, $100(1-\alpha)\%$ confidence interval for β_i is $\hat{\beta}_i \pm t_{n-p}^{(\frac{\alpha}{2})} \hat{SE}(\hat{\beta}_i)$. We can also denote this as $(\hat{\beta}_{\alpha,i}^L, \hat{\beta}_{\alpha,i}^U)$. This means that the $P(\hat{\beta}_{\alpha,i}^L, \hat{\beta}_{\alpha,i}^U)$ covers β_i is $1-\alpha$.

We can see that 0 not being in the CI for β_i has a one to one correspondence with a t-test for $\beta_i = 0$ rejecting. If 0 is not in the CI, then either $0 < \hat{\beta}_{\alpha,i}^L$ or $\hat{\beta}_{\alpha,i}^U < 0$. For the case where $0 < \hat{\beta}_{\alpha,i}^L$, we can rewrite this as $0 < \hat{\beta}_i - t_{n-p}^{(\frac{\alpha}{2})} \hat{SE}(\hat{\beta}_i)$, which simplifies to $\frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} > t_{n-p}^{(\frac{\alpha}{2})}$, which is exactly as we see in the t-test. For the upper bound, we can perform a symmetric analysis. In addition, $P(t_{n-p} > t^{obs}) < \frac{\alpha}{2}$ equals $2 * P(t_{n-p} > t^{obs}) < \alpha$, so the p-value is less than α, which means we reject the null for the t-test. The argument also goes the other way, by reversing all the implications hence the statements are equivalent.

Confidence intervals are for one predictor at a time. If we want to think about multiple predictors at once, we can form a Confidence Region, which is analogous to the F-test. The ellipsoid $\hat{E}_\alpha$, which is the shape of the Confidence Region is defined by $(\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \leq p\hat{\sigma^2} F_{p,n-p}^{(\alpha)}$.

6 Gauss-Markov Theorem

The Gauss-Markov Theorem states that the OLS (Ordinary Least Squares) estimates are BLUE (Best Linear Unbiased Estimators). We make weaker assumption here and this still holds. We do not assume that the errors are iid and normally distributed, only that $\epsilon \sim (0, \sigma^2 I_n)$. Therefore, the errors have mean 0 and standard deviation σ^2, but we have no guaratees of the same values of the moments after the 2nd moment. Therefore, these are called 2nd order assumptions.

We have $Y = X\beta + \epsilon$. Let a linear combination be $\gamma = c^T\beta$, where c is a $p \times 1$ vector. Our OLS estimate is $\hat{\gamma} = c^T\hat{\beta}$. Note that we can rewrite this as $\hat{\gamma} = c^T\hat{\beta} = c^T(X^TX)^{-1}X^TY = a_0^TY$, where $a_0 = X(X^TX)^{-1}c$. Notice that the OLS estimator as linear and unbiased, since $E[\hat{\gamma}] = E[c^T\hat{\beta}] = c^TE[\hat{\beta}] = c^T\beta$.

Let's compare our OLS estimator to all other linear unbiased estimators, $\tilde{\gamma}(Y) = a^TY$. We want to show that OLS is the "best" (lowest variance), and therefore $Var(\hat{\gamma}) \leq Var(\tilde{\gamma})$. We can calculate the variance as $Var(\tilde{\gamma}) = Var(a^TY) = a^TCov(Y)a = a^TCov(\epsilon)a = \sigma^2||a||^2$ (using the second order assumption). In particular, $Var(\hat{\gamma}) = \sigma^2||a_0||^2$. Therefore, to show OLS is the best, we need to show that $||a_0||^2 \leq ||a||^2$ for all unbiased a.

For $\tilde{\gamma}(Y)$ to be unbiased, $c^TY = E[\tilde{\gamma}(Y)] = E[a^TY] = a^TX\beta$. Therefore, $c^T = a^TX$ and $X^Ta = c$. In particular, for OLS, $X^Ta_0 = c$. Therefore,

$X^T(a - a_0) = 0$
$c^T(X^TX)^{-1}X^T(a - a_0) = 0$
$a_0^T(a - a_0) = 0$

Therefore, we have a is perpendicular to $a - a_0$, and thus we can write $||a||^2 = ||a_0||^2 + ||a - a_0||^2 \geq ||a_0||^2$. We have proved OLS is BLUE.

7 Generalized Least Squares

If we want to relax the second order assumptions, relaxing our assumption that the $Cov(\epsilon) = \sigma^2 I_n$, we can still use regression, but now we will move to Generalized Least Squares from Ordinary Least Squares. Our assumptions now are

- (LM) Model is $Y = X\beta + \epsilon$, full rank
- (E) $E[\epsilon] = 0$, $Cov[\epsilon] = G > 0$ (symmetric positive definite, not just positive semi-definite)

If we look at the OLS estimate $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y = A_X Y$, where $A_X = (X^T X)^{-1} X^T$, then the estimate is still unbiased, $E[\hat{\beta}_{OLS}] = \beta$. However, when we look at the covariance, we get $Cov[\hat{\beta}_{OLS}] = A_X Cov(Y) A_X^T = (X^T X)^{-1} X^T G[(X^T X)^{-1} X^T]^T$, which is not equal to $\sigma^2 (X^T X)^{-1}$ (unless $G = \sigma^2 I$). Therefore, we must find a new way.

We turn to Generalized Least Squares. In Generalized Least Squares we premultiply the linear model by $G^{-\frac{1}{2}}$.

What is $G^{-\frac{1}{2}}$? Since we assume G is SPD, then the eigendecomposition of G is $G = RDR^T$, where R is orthogonal and D is diagonal (eigenvalues). Then $G^{\frac{1}{2}} = RD^{\frac{1}{2}} R^T$ (we can see that $(G^{\frac{1}{2}})^2 = G$). Consequently, $G^{-\frac{1}{2}} = RD^{-\frac{1}{2}} R^T$, where $D^{-\frac{1}{2}}$ is the square root of the reciprocal of each of individual diagonal elements of G.

If we premultiply by $G^{-\frac{1}{2}}$ to the linear model, we have $G^{-\frac{1}{2}} Y = G^{-\frac{1}{2}} X\beta + G^{-\frac{1}{2}} \epsilon$. Let's rewrite this as $\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}$. Then we have $Cov(\tilde{\epsilon}) = G^{-\frac{1}{2}} Cov(\epsilon)(G^{-\frac{1}{2}})^T = G^{-\frac{1}{2}} G G^{-\frac{1}{2}} = I$.

Our new estimate $\hat{\beta}_{GLS}$ equals:

$\hat{\beta}_{GLS} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$
$= [(G^{-\frac{1}{2}} X)^T (G^{-\frac{1}{2}} X)]^{-1} [(G^{-\frac{1}{2}} X)^T (G^{-\frac{1}{2}} Y)]$
$= [X^T G^{-\frac{1}{2}} G^{-\frac{1}{2}} X]^{-1} [X^T G^{-\frac{1}{2}} G^{-\frac{1}{2}} Y]$, since $G^{-\frac{1}{2}} = [G^{-\frac{1}{2}}]^T$ since it is symmetric
$= (X^T G^{-1} X)^{-1} X^T G^{-1} Y$

We know that this estimate is BLUE (best linear unbiased estimator) since once we premultiply, it is simply an OLS problem. We can solve for the covariance matrix as follows:

$Cov(\hat{\beta}_{GLS}) = \tilde{A}_X Cov(\tilde{Y}) \tilde{A}_X^T$
$= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Cov(\tilde{Y}) [(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T]^T$
$= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T Cov(\tilde{\epsilon}) [(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T]^T$, since $Cov(\tilde{Y}) = Cov(\tilde{\epsilon})$
$= [(G^{-\frac{1}{2}} X)^T (G^{-\frac{1}{2}} X)]^{-1} (G^{-\frac{1}{2}} X)^T I [(G^{-\frac{1}{2}} X)^T (G^{-\frac{1}{2}} X)]^{-1} (G^{-\frac{1}{2}} X)^T]^T$
$= [X^T G^{-\frac{1}{2}} G^{-\frac{1}{2}} X]^{-1} X^T G^{-\frac{1}{2}} [(G^{-\frac{1}{2}} X)(X^T G^{-\frac{1}{2}})(G^{-\frac{1}{2}} X)]$, since $G^{-\frac{1}{2}} = [G^{-\frac{1}{2}}]^T$ since it is symmetric
$= [X^T G^{-1} X]^{-1} [X^T G^{-1} X] [X^T G^{-1} X]^{-1}$
$= (X^T G^{-1} X)^{-1}$, which is analogous to $\sigma^2 (X^T X)^{-1}$ if $G = \sigma^2 I$

7.1 Weighted Least Squares

When G is diagonal, the form of the Generalized Least Squares is called Weighted Least Squares. Since G is diagonal, we still make the assumption that the Y_i are uncorrelated. However, each Y_i may have a different variance. Therefore, we have $Var(Y_i) = Var(\epsilon_i) = \frac{\sigma^2}{w_i}$. We can view G as the matrix

$$G = \sigma^2 \begin{bmatrix} \frac{1}{w_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{w_n} \end{bmatrix}$$

Therefore, using premultiplying on the left by $G^{-\frac{1}{2}}$, we have that $\sqrt{w_i} Y_i = \sqrt{w_i} X_i \beta + \sqrt{w_i} \epsilon_i$. (We left out the σ^2 since it will not affect our least squares solution). To solve $\hat{\beta}_{GLS}$, we minimize

$\sum_{i=1}^n (\sqrt{w_i} Y_i - \sqrt{w_i} X_i \beta)^2$
$= \sum_{i=1}^n w_i (Y_i - X_i \beta)^2$

We can solve this, assuming we know the weights w_i. What if we don't know the weights beforehand? To solve this, we need to then estimate/model the variances. One way to model the variances is to assume a linear model $\sigma_i^2 = X_i \gamma$. We don't have σ_i^2, but we have a proxy for them from the residuals from our original model, e_i^2. We then fit $e_i^2 = X_i \gamma + \delta_i$ (this assumes that the δ_i

as homoscedastic). We then can estimate $\hat{\gamma}$ using least squares. From $\hat{\gamma}$ we can get $\hat{\sigma}^2 = X_i\hat{\gamma}$. From $\hat{\sigma}^2$ we can then construct $\hat{G}$.

Now, instead of Generalized Least Squares (GLS), since we estimate the variances, we have Feasible GLS. In FGLS, we have $\hat{\beta}_{FGLS} = (X^T\hat{G^{-1}}X)^{-1}X^T\hat{G^{-1}}Y$. The covariance matrix is $\hat{Cov}(\hat{\beta}_{FGLS}) \approx (X^T\hat{G^{-1}}X)^{-1}$. Note that $\hat{\beta}_{GLS}$ is linear in Y. However, $\hat{\beta}_{FGLS}$ may not be linear in Y because the Y is also influencing the e_i, which affects $\hat{\gamma}$ and therefore $\hat{G}$. It may be close to linear to Y, but may not be exactly linear in Y.

7.2 Iteratively Reweighted Least Squares

In iteratively reweighted least squares, we use a process like FGLS, but keep repeating it over and over. We start with $\beta^{(0)} = \hat{\beta}_{OLS}$, and we get the residual $e^{(0)} = Y - X\hat{\beta}^{(0)}$. Then we repeat the following:

- Use $e^{(n)}$ to obtain $\hat{G}^{(n)}$ (for example like in the linear way from FGLS)
- Calculate $\hat{\beta}^{(n)} = (X^T\hat{G^{-1}}^{(n)}X)^{-1}X^T\hat{G^{-1}}^{(n)}Y$
- Recalculate the new residuals from this $e^{(n)} = Y - X\hat{\beta}^{(n)}$

8 Data Snooping

A common process of building a model comes from taking a look at all the predictors on the target, filtering out all the predictors that are not statistically significant, and then building a final model with only those predictors that are significant. However, we need to do this with caution, especially how we interpret the model.

Take an example where the Y variable is randomly generated from a normal distribution (say 100 instances). We then generate 50 different X predictors randomly as well. We look at the t-statistic for these predictors at the 10% significance level. Due to random chance, we actually expect 5 of these predictors to be significant, and their corresponding p-values to be below 10%. Therefore, when filtering, we will keep these 5 predictors in the model. We then run our model using these 5 predictors, which will all be statistically significant. However, in reality, because we randomly generated these variables, we know that they have no predictive power.

In order to solve this problem, we can use a Bonferroni correction of a False Discovery Rate.

9 Diagnostics: Checking Error Assumptions

We may have the assumption in our model that $\epsilon \sim N(0, \sigma^2 I)$, assuming fixed X. We want to check to see if these assumptions hold. We want to check constant variance assumption, the normality assumption, and for correlation.

9.1 Checking Constant Variance

A common way to check for constant variance is to fit the model and check to see if the residuals are constant by plotting the fitted values, $\hat{Y}_i$, against the residuals, e_i, and seeing if they are constant for different values of $\hat{Y}_i$. If they are not constant, for example if they appear to be growing as the value of $\hat{Y}_i$ increases, then we need to perform variance stabilization.

Assume we have a variance that is not constant and depends on the mean of the Y_i. We have $Var(Y_i) = f(E[Y_i]) = f(\mu_i)$. Therefore, for variance stabilization, we seek a function h such that $Var[h(Y_i)] = c$, a constant. Using a first order Taylor expansion, near μ, we have that $h(y) \approx h(\mu) + (y - \mu)h'(\mu)$. Taking expectation of both sides, we have that $E[h(Y)] \approx E[h(\mu) + (y - \mu)h'(\mu)] \approx E[h(\mu)] = h(\mu)$ (since we are near μ, then $(y - \mu) \approx 0$). Looking at the variance, we have

$Var[h(Y)] = E[h(Y) - E[h(Y)]]^2$
$\approx E[(y - \mu)^2 (h'(\mu))^2]$
$= (h'(\mu))^2 Var(Y)$
$= (h'(\mu))^2 f(\mu) = c^2$

We want to set this value to a constant c.

$h'(\mu) = \frac{c}{\sqrt{f(\mu)}}$
$h(\mu) = \int_{\mu_0}^{\mu} \frac{c}{\sqrt{f(\kappa)}} d\kappa$

Two examples of how to apply this transformation. The first example is when the variance is a linear function of μ, such as the Poisson distribution where $Var(Y) = E(Y)$. In this case, we have $f(\mu) = \mu$. Using our formula above, we have that $h(\mu) = \int_{\mu_0}^{\mu} \frac{c}{\sqrt{\kappa}} d\kappa = 2c\sqrt{\mu} + d$, where d is another constant. This suggests that we should take a square root transformation of Y in order to stabilize the variance.

The second example is when the variance is a quadratic function of μ. In this case, we have $f(\mu) = \mu^2$. Using the formula, we have that $h(\mu) = \int_{\mu_0}^{\mu} \frac{c}{\kappa} d\kappa = clog(\mu) + d$, where d is another constant. This suggests that we should take a log transformation of Y in order to stabilize the variance.

One note is that we can use a Box-Cox analysis to test which transform we should use to stabilize the variance. In this analysis, we try different values of $\frac{y^\lambda - 1}{\lambda} = h_\lambda(y)$, and see which one works the best to stabilize the variance.

9.2 Checking Normality

To check the normality assumptions (typically of the errors), there are two common methods used: qqplots and the permutation test

9.2.1 QQ Plots

The idea behind a QQ Plot is that by looking at a plot, if the points are linear, then we know that they fulfill the normality assumption. How this would work in a regression setting is that the model would be fit, the $\hat{\beta}$ are obtained, and the residuals would be calculated: $e_1, e_2, ..., e_n$. These residuals are then ordered from smallest to largest: $e_{(1)} \leq e_{(2)} \leq ... \leq e_{(n)}$. We then calculate the expected normal order statistics $\gamma_{1n} \leq \gamma_{2n} \leq ... \leq \gamma_{nn}$, where $\gamma_{in} = \Phi^{-1}(\frac{i - \frac{1}{2}}{n})$. We then plot $(\gamma_{in}, e_{(i)})$, with the residuals on the y-axis and the expected normal values on the x-axis and see if it's linear.

Some brief intuition as to why this works. Assume we draw n elements from a Uniform (0, 1) distribution and then order them, so we have $U_{(1)}, U_{(2)}, ..., U_{(n)}$. Our expected value at each observation is $E[U_{(i)}] = \frac{i - \frac{1}{2}}{n}$. We adjust for $\frac{1}{2}$ as without it, the nth observation would have expected value 1, which we can clearly see is a bit too high. Assume we want to simulate standard normal variables using these uniform draws and the inversion method. Then we generate $Z_{(i)} = \Phi^{-1}(U_{(i)})$. Clearly the order is preserved since the CDF is a monotone function. We then have $E[Z_{(i)}] = E[\Phi^{-1}(U_{(i)})] \approx \Phi^{-1}(E[U_{(i)}]) = \Phi^{-1}(\frac{i - \frac{1}{2}}{n}) = \gamma_{in}$.

Note that the argument above was for a standard normal, so the simulated values have mean 0 and variance 1. If $X_i \sim N(\mu, \sigma^2)$, then $E[X_{(i)}] = \mu + \sigma\gamma_{in}$. Therefore we now have slope σ and intercept μ; however, we should still see a straight line. For

example, typically, when we have $Y = X\beta + \epsilon$, our residuals should be $e \sim N(0, \sigma^2(I - H))$, so we may not have slope 1, but we should still see linearity in the QQ Plot if our assumption holds.

9.2.2 Permutation Tests

If we are performing a t-test or an F-test and want to verify the normality of the errors, or we look at a QQ Plot, and it is a bit unclear if the errors are normal or not, we can use permutation tests. The idea is that we find the t or F-statistic from the observed data, and then permute the data and run the t or F-test many times. If the original t or F-statistic is more extreme than these permuted t or F statistics a lot of the time, then our original test holds.

For example, with the F-test (assuming we were testing all the variables together for significance), we would first run the F-test with our current data and get the F-statistics, call it F^{obs}. We then randomly permute the Y's while keeping the X's the same. We then fit this permutation and run the F-test and obtain a new F-statistic. We repeat this process where we permute, fit, and obtain new F-statistics of the permutation many times. We count the number of permuted F-statistics that exceed F^{obs}, divide by the number of times we repeated this permutation, and that is our p-value. $P = \frac{\#F_i > F^{obs}}{N}$. It can be shown that the histogram of these permuted F-statistics actually matches the density $F_{p,n-p}$. Therefore, we are able to perform an F-test without necessarily the Gaussian error assumption. However, the tradeoff is that we have to fit the model many times, which is computationally expensive.

With the permutation test on the t-statistic, we follow similar steps to the F-statistic mentioned above, with one key difference, which is that we permute the predictor of interest, rather than the target. This is because we need Y to stay the same as we are regressing the other variables on it as well, and we are only testing the significance of one single predictor variable, so we permute that one instead. In this case, the p-value is $P = 2\frac{\#|T_i| > |T^{obs}|}{N}$, since it is a two-sided test. Similarly, the histogram of these permuted t-statistics actually matches the density t_{n-p}. Again the tradeoff for not relying on the Gaussian error assumption is that we have to fit the model many times, which is computationally expensive.

9.3 Checking Correlation

Omitting a variable can generate observed correlation in errors. In this case, the source of correlation is that the errors in the fitted model reflect an important variable that is not included in the model.

To test this, we can use a Durbin-Watson test. The Durbin-Watson statistic is $d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$.

10 Prediction Intervals

While we can use confidence intervals when we talk about the CI of an estimator $\hat{\beta}$, or a CI of an average, when we make a prediction about a single point that was not in the training set, we use a prediction interval (one single point). This is to be contrasted with predicting a mean response (averaging over many observations), where we use a confidence interval.

Say we have fit a model with the parameters $\hat{\beta}$. We want to predict the value of the mean response (average of many observations), x_0. The prediction is unbiased, so we have $\hat{y_0} = x_0^T \hat{\beta}$. When we want to form a confidence interval around this mean response, we have

$Var(\hat{Y_0}) = Var(x_0^T \hat{\beta})$
$= x_0^T Cov(\hat{\beta}) x_0$
$= \sigma^2 x_0^T (X^T X)^{-1} x_0$

Therefore, our estimate of the standard error is $\hat{SE}_{mr} = \hat{\sigma}\sqrt{x_0^T (X^T X)^{-1} x_0}$. Thus, a 95% Confidence Interval for the mean response is $x_0^T \hat{\beta} \pm t_{n-p}^{\frac{\alpha}{2}} \hat{SE}_{mr}$.

If we want to make a prediction for a single point, then our best estimate is again $\hat{y_0} = x_0^T \hat{\beta}$. However, our variance is larger, because we have to account for the variability ϵ_0 as well for that fluctuation in the individual observation. Therefore, we have

$Var(\hat{Y_0} - Y_0) = E[(\hat{Y_0} - Y_0)^2] - E[(\hat{Y_0} - Y_0)]^2$, but we know that $E[\hat{Y_0}] = E[Y_0]$ because unbiased
$= E[(\hat{Y_0} - Y_0)^2]$
$= E[(x_0^T \hat{\beta} - x_0^T \beta - \epsilon_0)^2]$
$= Var(x_0^T \hat{\beta}) + Var(\epsilon_0)$
$= \sigma^2 x_0^T (X^T X)^{-1} x_0 + \sigma^2$
$= \sigma^2 (1 + x_0^T (X^T X)^{-1} x_0)$

Therefore, our estimate of the standard error is $\hat{SE}_{fv} = \hat{\sigma}\sqrt{1 + X_0^T (X^T X)^{-1} X_0}$. Thus a 95% Confidence Interval for a single future observation is $X_0^T \hat{\beta} \pm t_{n-p}^{\frac{\alpha}{2}} \hat{SE}_{fv}$.

Let us mean correct the columns of X besides the intercept column, such that we have

$$\tilde{X} = \begin{bmatrix} x_1^T - \bar{x}^T \\ \vdots \\ x_n^T - \bar{x}^T \end{bmatrix}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Let us also denote $x_0^T = (1 \;\; \tilde{x_0}^T)$. Then it can be shown that $x_0^T (X^T X)^{-1} x_0 = \frac{1}{n} + (\tilde{x_0} - \bar{x})(\tilde{X}^T \tilde{X})^{-1} (\tilde{x_0} - \bar{x})$. Therefore, we can see that the variance of our prediction grows quadratically as x_0 moves away from the average of the predictors in the design matrix.

11 Finding Unusual Observations

We may find certain values that are unusual (large or small) compared to the majority of the points in our data set. How far the independent variable values of an observation are from those of the other observations is called leverage. For example, if most data points have x-values between 5 and 10, and this one data point have an x-value of 50, we say it has high leverage.

To measure the leverage of an observation, we look at the hat matrix. Specifically, to find the leverage of the ith point, we look at the H_{ii} entry, let us denote that h_i. Recall that $Var(e) = \sigma^2(I - H)$, therefore $Var(e_i) = \sigma^2(1 - h_i)$. Note that for large values of h_i (high leverage), the $Var(e_i)$ is close to 0. What that means is for points with large leverage, the control the entire regression line and pull it close to themselves such that the residual is small.

As a rule of thumb we say a point has large leverage if $h_i > \frac{2p}{n}$. This is because $\frac{1}{n}\sum_{i=1}^{n} h_i = \frac{1}{n}tr(H) = \frac{p}{n}$. This is the average leverage of a point. Therefore, if we exceed twice this, we say the point is high leverage.

Another way to check for points with high leverage is to plot them. h_i must be between 0 and 1, so we plot these h_i against a half normal plot (which is non-negative). A half normal plot is just the positive part of a normal distribution, but the area at each point is doubled so the total area is still 1. The steps to plot are as follows:

- Sort h_i. $h_{(1)} \leq h_{(2)}, ..., h_{(n)}$
- Compute scores $\delta_i = \Phi^{-1}(\frac{n+i}{2n+1})$
- Plot $h_{(1)}$ vs. δ_i
- Look for outlying points (but not for a straight line, since there is no reason for h_i to be Gaussian)

How do we find outliers? The initial feeling may be to look for points with large residuals. However, if the outlier also has high leverage, it may pull the regression line close to itself and therefore have a small residual. That is why, we need to combine the idea of residuals along with leverage in order to determine the outliers.

For each point that we want to test is an outlier, we fit a model with and without that point and the compare the results. More precisely, we can

- Exclude the point we want to test (x_i, y_i) and fit a model to obtain estimates $\hat{\beta}_{(i)}$
- Predict the value of y_i using this model for the excluded point $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$
- If $\hat{y}_{(i)}$ is close to y_i, then the point is not an outlier. If it is far, then the point is an outlier.
  - We measure this by $Var(y_i - \hat{y}_{(i)}) = \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$
  - Note that this is the same standard error as used for the prediction interval
- Calculate the t-statistic: $t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{SE}(y_i - \hat{y}_{(i)})} = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}[1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i]^{\frac{1}{2}}}$
  - Here, $\hat{\sigma}_{(i)}$ is the unbiased estimate of σ^2 in the $(n-1)$ observation model, thus we have $n-p-1$ degrees of freedom

From first glance, it appears like a lot of computation to have to fit a model to exclude each point that we want to test one by one and then run a t-test on it. As it turns out, it can be proved that $t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i(\frac{n-p-1}{n-p-r_i^2})^{\frac{1}{2}}$. Therefore, we can save lots of computation.

If we are running many t-tests, we would assume that some would come up false positives. Therefore, we need to use a test correction. One of the more conservative approaches is to use the Bonferroni correction.

To figure out which points are influencial observations, we measure the change in fit due to deleting observation i. We can use Cook's distance as a metric. Cook's distance is calculated by:

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

D_i near 1 (or larger) should attract attention e.g. refit the model without that observation (or those observations).

12 Diagnostics: Checking Predictors

12.1 Added Variable (Partial Regression) Plots

If we want to understand the structure between Y and X (for example to see if Y has a linear relationship with X_i), we use an Added Variable plot. Assume that our design matrix X has p columns and we want to understand the relationship between Y and the jth predictor $X^{(j)}$. We cannot simply plot Y vs. $X^{(j)}$ since that would take into account the effect of all the other predictors. Therefore, we first remove the effects of all the other predictors on $X^{(j)}$ as well as Y before we look at the results.

Let us create the hat matrix H_{w_j}, which is the the hat matrix with all the predictors except for $X^{(j)}$. Then we calculate the following:

- $Y_{.w_j} = Y - H_{w_j}Y$ (adjusted Y)
- $X^{(j)}_{.w_j} = X^{(j)} - H_{w_j}X^{(j)}$ (adjusted $X_{(j)}$)

We then plot the adjusted Y_i against the adjusted $X_i^{(j)}$. Note that we use no intercept as that was already removed as part of the adjustment. Therefore our plot is $Y_{.w_j} = \hat{\beta}_j^{AV} X^{(j)}_{.w_j}$. Also, it is a fact that this $\hat{\beta}_j^{AV}$ will be equivalent to the OLS estimate $\hat{\beta}_j$ from the full model.

12.2 Measurement Error in Predictors

What happens to the least squares estimate if the predictors are measured with error? Consider the following model, where we only are able to observe the Y_i and X_i.

- $Y_i = \beta_0 + \beta_1 W_i + \epsilon_i$
- $X_i = W_i + \delta_i$

In this scenario, there is a true linear relation between W_i and Y_i, but we only observe X_i and Y_i, and X_i differs from W_i by a (measurement) error. If we substitute $W_i = X_i - \delta_i$ into the first equation, we get that

$Y_i = \beta_0 + \beta_1(X_i - \delta_i) + \epsilon_i$
$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i - \beta_1\delta_i$
$Y_i = \beta_0 + \beta_1 X_i + \tilde{\epsilon_i}$, where $\tilde{\epsilon_i} = \epsilon_i - \beta_1\delta_i$

The issue here is that we violate our typical (RX) assumption since X_i and $\tilde{\epsilon_i}$ actually are correlated since both are a function of δ_i. The result of this is that our OLS estimate $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}$ actually is biased! We can show this as follows.

Under a reasonable set of assumptions, assume that $Cov(W_i, \epsilon_i) = Cov(W_i, \delta_i) = Cov(\epsilon_i, \delta_i) = 0$. Also lets assume that $Var(W_i) = \sigma_W^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and $Var(\delta_i) = \sigma_\delta^2$. Let's also assume these are uncorrelated over i (the observations). We also know that $E[\hat{\beta}_1] \approx \frac{Cov(X,Y)}{Var(X)}$ (it is approximate since it will be a sample estimate, not population). Then we have $Cov(X,Y) = Cov(W + \delta, \beta_0 + \beta_1 W + \epsilon) = \beta_1\sigma_W^2$. We also have $Var(X) = Cov(W + \delta, W + \delta) = \sigma_W^2 + \sigma_\delta^2$.

Therefore, we have $E[\hat{\beta}_1] \approx \beta_1 \frac{\sigma_W^2}{\sigma_W^2 + \sigma_\delta^2} = \frac{\beta_1}{1+\theta}$, where $\theta = \frac{\sigma_\delta^2}{\sigma_W^2} \geq 0$. Therefore, since $\theta \geq 0$, then $\frac{\beta_1}{1+\theta} \leq \beta_1$, so the estimate will be biased downward.

There is a method called simulation-extrapolation (SIMEX) that can improve our OLS estimate in this scenario.

12.3 Collinearity

Collinearity occurs when one (or more) of the predictors is (nearly) a linear combination of (some of) the others. We do not want this. Collinearity leads to instability in the estimates of coefficients (higher variances).

In a univariate regression of Y on $X^{(j)}$ (with intercept), the variance of the slope estimate $Var(\hat{\beta}_j^{OLS}) = \frac{\sigma^2}{s_j^2}$, where $s_j^2 = \sum_{i=1}^n (X_i^{(j)} - \bar{X}^{(j)})^2$.

If we regress $X^{(j)}$ on all the other predictors $X^{(k)}$, $k \neq j$, then $R_j^2 = corr^2[X^{(j)}, \hat{X}^{(j)}]$.

It can be proven that $Var(\hat{\beta}_j) = \sigma^2(X^TX)^{-1}_{jj} = \frac{1}{1-R_j^2} * \frac{\sigma^2}{s_j^2}$. The quantity $\frac{1}{1-R_j^2}$ is called the Variance Inflation Factor. What this factor means is the how much the variance of $\hat{\beta}_j$ is inflated relative to the variance of the univariate regression.

13 Variable Selection

We may not want to always keep every single predictor in our model. Therefore, we want to find a subset C of the predictors from $1, ..., p$ for our model that fits well. Thus, our linear model looks like $Y = X_C\beta_C + \epsilon$.

There is no single correct way to approach variable selection, and especially for datasets with many variables, there is typically not an obvious single "best" choice of model. There is a considerable danger of being overwhelmed by large amounts of computer output from automated selection algorithms, often to no good purpose. One should bring to bear one's knowledge of the context of the data set in thinking about competing models.

It is important to keep in mind the purpose of the analysis in the context of the data: are "simple" (typically smaller) models desirable for purposes of (attempted) interpretation, or is effective prediction(most likely from a larger model) the main goal? Also, remember that that the p-values for coefficients that are retained in selected models are likely to be exaggerated.

13.1 Testing Strategies

- Backward Elimination
  - Start with the full model.
  - Remove the variable with the highest p-value $> \alpha_{delete}$
  - Repeat until there are no more such variables to delete.
  - Typical α_{delete} may be 0.05
- Forward Selection
  - Start with the null model (or minimal model with forced inclusions).
  - Add the variable with the smallest p-value $< \alpha_{enter}$
  - Repeat until there are no more such variables to add.
- Stepwise
  - Start with model m.
  - Add variable with smallest p-value $< \alpha_{enter}$ and/or remove variable with highest p-value $> \alpha_{delete}$
  - Repeat until there are no more such variables.

13.2 Rules of Thumb for Variable Selection

- If we have an interaction term $x_1 * x_2$ in the model, we should also keep x_1 and x_2 in the model
- If we have a higher order term x_1^k in the model, then we should also keep all the lower order terms of x_1^j, where $j < k$ in the model
- If we have a factor variable, we should have all the levels in the model or none of the levels (do not just have some of the levels in the model)

13.3 Variable Selection Criteria

We have different criteria that we can use to help us determine which variables to keep in our model. The following are criteria that help us optimize over all the 2^p candidate models X_C that are a subset of the full model. Each of these criteria differ in the way they trade of model size, number of variables in C (p_C), quality of fit, and the RSS_C. We define $RSS_C = ||Y - X_C\beta_C||^2$.

13.3.1 Akaike Information Criterion (AIC)

$\text{AIC}(C) = nlog(\frac{RSS_C}{n}) + 2p_C$

As p_C increases, RSS_C decreases. However, the penalty term $2p_C$ increase with model size (complexity). The optimal model is the one minimizing the AIC criterion.

13.3.2 Bayes Information Criterion (BIC)

$\text{BIC}(C) = nlog(\frac{RSS_C}{n}) + p_C(logn)$

The penalty term is larger than AIC. It penalizes each new variable by $logn$, not 2. Minimizing BIC therefore leads to smaller models than AIC.

13.3.3 Mallows' C_p

Mallows' C_p is a prediction criterion. Assume we fit the smaller model with C predictors and make predictions on all the x's. Therefore we have $\hat{y}_{i,C} = x_{i,C}^T \hat{\beta}_C$. The prediction error for all n cases is equal to $\frac{1}{\sigma^2} \sum_{i=1}^n E[(\hat{y}_{i,C} - E[y_i])^2]$.

It can be shown that a (nearly) unbiased estimate of this prediction error is

$C_p = \frac{RSS_C}{\hat{\sigma^2}} + 2p_C - n$

where $\hat{\sigma^2}$ is the residual variance estimate from the full model (p predictors). To find the optimal model, we want to minimize C_p.

13.3.4 Adjusted R^2

$R^2_{adj} = 1 - \frac{\frac{RSS_C}{n-p_C}}{\frac{TSS}{n-1}}$

As opposed to R^2, which can only increase are more variables are added, the $n - p_C$ term in Adjusted R^2 adds a penalty for larger models. We want to maximize the Adjusted R^2 to find the optimal model.

14 Shrinkage Methods: Ridge and Lasso

14.1 Ridge Regression

- $\hat{\beta}^{ridge} = argmin_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$
- L_2 penalty: a penalized residual sum of squares
- λ is a regularization: the larger the λ, the more shrinkage toward 0
- Note that the intercept coefficient β_0 is not shrunk (not in penalty), so that fit doesn't depend on origin chosen for y; therefore $\hat{\beta}_0 = \bar{y}$ still if x is centered
- The X_{ij} are usually scaled to mean 0 and variance 1 before fitting Ridge. This makes the ridge coefficient estimates $\hat{\beta}^{ridge}$ equivalent under scaling of the input variables. $X_{ij} \to a + bX_{ij}$
- An equivalent way to write ridge regression is $\hat{\beta}^{ridge} = argmin_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$ subject to $\sum_{j=1}^p \beta_j^2 \leq t$

In matrix form, we are able to find, in closed form, the solution to the ridge regression. We want to minimize $RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$. Using calculus to take the derivate and set it equal to 0, we find that the minimizer is

$\hat{\beta}^{ridge} = (X^TX + \lambda I_p)^{-1}X^Ty$

Note that the solution shows that $\hat{\beta}^{ridge}$ is a linear function of y, and that even if X^TX is not invertible, $X^TX + \lambda I_p$ is.

In OLS, our hat matrix have trace = p, and thus there are p degrees of freedom. In ridge regression our "hat matrix" is called the "smoother matrix" and denoted $H_\lambda = X(X^TX + \lambda I_p)^{-1}X^T$. Note that H_λ is not a projection. We can find the effective degrees of freedom by looking and its trace, by using the SVD decomposition.

$H_\lambda = X(X^TX + \lambda I_p)^{-1}X^T = UDV^T[(UDV^T)^T(UDV^T) + \lambda I_p]^{-1}(UDV^T)^T = UD(D^2 + \lambda I_p)^{-1}DU^T$. Therefore we find that the trace equals $tr(H_\lambda) = tr[(D^2 + \lambda I_p)D^2]$. Expanding it out, this equals $\sum_{j=1}^p \frac{d_j^2}{d_j^2+\lambda}$. Therefore, we can now see that if $\lambda = 0$, our effective degrees of freedom is p, which is our OLS case. Also, as $\lambda \to \infty$, our effective degrees of freedom $\to 0$.

In the special case where X is an orthogonal matrix, $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty = (1 + \lambda)^{-1}X^Ty = \frac{\hat{\beta}^{LS}}{1+\lambda}$. Therefore, we can see that as λ increases, our LS estimate is shrunk. More generally, the trace shows that our predictors are shrunk towards 0.

14.2 Lasso Regression

- $\hat{\beta}^{lasso} = argmin_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$
- L_1 penalty: a penalized residual sum of squares
- The solution is now non-linear in y; no closed form expression, must be found numerically. Quadratic programming problem with very efficient algorithms now available
- Again, the intercept coefficient β_0 is not penalized, so $\hat{\beta} = \bar{y}$ if x is centered
- An equivalent way to write Lasso regression is $\hat{\beta}^{lasso} = argmin_\beta \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$ subject to $\sum_{j=1}^p |\beta_j| \leq t$
- If $t > \sum_{j=1}^p |\hat{\beta}_j^{LS}|$, then $\hat{\beta}^{lasso} = \hat{\beta}^{LS}$
- As t is reduced, coefficients are shrunk in a non-linear way
- In the special case where X is an orthogonal matrix, $\hat{\beta}^{lasso}$ is obtained by "soft thresholding." In this special case, $\hat{\beta}_j^{lasso} = sign(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)_+$
- We can see that in the special case, the estimates are moved towards 0 by λ, soft thresholded at 0, and then the original sign of the estimate is returned

For both Ridge and Lasso, we can check the quality of the choice of λ by cross validation. We can use a specific type called Generalized cross validation. To do this, we minimize $GCS(\lambda) = \frac{||Y - X\hat{\beta}(\lambda)||^2}{(n - df(\lambda))^2}$.

15 ANOVA - One Factor

ANOVA is a special case of regression still with a numeric Y variable, but with categorical X variables. In the framework of ANOVA, the "predictors" are called "factors," and the "coefficients" are called "effects."

In the One-Factor model, we are only looking at different levels for one factor. The model and the assumptions are below:

- Model: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- n_i observations at level $i = 1, ..., I$; $j = 1, ..., n_i$ (can be different number of observations for different levels)
- $n = \sum_{i=1}^{I} n_i$ (total number of observations)
- Gaussian Error Assumption (used for testing): $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
- μ is the "overall mean"
- α_i is the mean of the level

The one-factor model as shown above has one issue, which is that it is not identifiable. For example we could write μ as $\mu + c$ and then α_i as $\alpha_i - c$. Therefore, we need to add more constraints to the model, which we can do in 3 different ways. The three ways are:

1. $\mu = 0$. In this case, our $\hat{\alpha_i} = \bar{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$
2. $\alpha_1 = 0$. In this case, our $\hat{\mu} = \bar{y_i}$. Our $\hat{a_i} = \bar{y_i} - \hat{\mu}$. This method is called "treatment contrasts."
3. $\sum_{i=1}^{I} \alpha_i = 0$. In this case, our $\hat{\mu} = \frac{1}{I} \sum_{i=1}^{I} \bar{y_i}$. Our $\hat{a_i} = \bar{y_i} - \hat{\mu}$. This method is called "sum contrasts."

For methods 2. and 3., when we perform F-tests, our null hypothesis is $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_I = 0$, which is the same as saying that $\mu_1 = \mu_2 = ... = \mu_I$. In this F-Test, our degrees of freedom are $(I - 1, n - I)$. For method 1. (less commonly used), when we perform F-tests, our null hypothesis is $H_0 : \mu_1 = \mu_2 = ... = \mu_I = 0$. In this F-Test, our degrees of freedom are $(I, n - I)$.

15.1 Coding of Factors

How do we actually code up our design matrix X? The example below should help us see what to do. Assume we have a factor that takes on levels A, B, and C, and a Y variable that takes on some continuous values. Assume we have 7 values: 3 from the first level, 2 from the second level, and 2 from the third level. Recall, we are modeling $E[Y_{ij}] = \mu + \alpha_i$. Then using the first method where $\mu = 0$, we would write:

$$E[Y] = X\beta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

If we were using the second method, the "treatment constants" where $\alpha_1 = 0$, we would have:

$$E[Y] = X\beta = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

If we were using the third method, the "sum constants" where $\sum_{i=1}^{I} \alpha_i = 0$, we would have:

$$E[Y] = X\beta = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

15.2 Error Variance

Recall from Least Squares theory that $\hat{\sigma^2} = \frac{||e||^2}{n-p}$. From before, we defined $e = y - \hat{y}$. Now in ANOVA, since we have different levels, we can write each residual as $e_{ij} = y_{ij} - \bar{y_i}$. Therefore, we have $\hat{\sigma}^2 = \frac{\sum_i^I \sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2}{n-I}$.

From distribution theory, we have the $\hat{\sigma^2} \sim \frac{\sigma^2 \chi^2_\nu}{\nu}$, where $\nu = n - I$, which is the degrees of freedom. In addition, $\hat{\sigma^2}$ is independent of $\{\bar{y_i}\}_{i=1}^I$.

15.3 Diagnostics

To check the normality of the residuals, we can similarly use a QQ-plot.

To check whether or not the errors are similarly distributed in each level, that $\epsilon_{ij} \sim N(0, \sigma_i^2)$, we can use Levene's test. We will use medians rather than means (for robustness) for this test. Since the only thing in the model that varies is the error term, we can write our model as

$med_j y_{ij} = \mu + \alpha_i + med_j \epsilon_{ij}$
$|y_{ij} - med_j y_{ij}| = |\epsilon_{ij} - med_j \epsilon_{ij}|$, denote this z_{ij}

Looking at the right hand side, $|\epsilon_{ij} - med_j \epsilon_{ij}| \approx \sigma_i Z$, where $Z = N(0, 1)$. This is because the $med_j \epsilon_{ij}$ should be 0, and the ϵ_{ij} should be normally distributed. Therefore, we have $E[z_{ij}] \approx c\sigma_i$. Looking at the equation, we can see that we now have a similar ANOVA setup, where the different levels are the σ_i, the standard deviation for each level. Therefore, we can use $\{z_{ij}\}$ in an ANOVA test where our null hypothesis is $H_0 : \sigma_1^2 = ... = \sigma_I^2$.

As a rule of thumb, most tests and CIs are relatively insensitive to non-constant variance, so we don't need to be concerned with whether the errors are the same in each level unless the Levene's test shows significance at a p-value of more than 0.01.

15.4 Pairwise Comparison of Means

Looking at the F-test, we can see whether or not there is a difference between the levels as a whole. However, many times we want to see pairwise comparisons of 2 levels at a time to see which are statistically significantly different from each other. As usual, we can view this as a Confidence Interval or a t-test.

From a confidence interval viewpoint, if we are comparing level i to level j we can write $\hat{\alpha_i} - \hat{\alpha_j} \pm t_\nu^{\frac{\alpha}{2}} \hat{SE}(\hat{\alpha_i} - \hat{\alpha_j})$, where $\hat{SE}(\hat{\alpha_i} - \hat{\alpha_j}) = \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$.

To perform all the different pairwise comparisons between the I levels, we would need to perform $\binom{I}{2} = \frac{I(I-1)}{2}$ different pairwise tests. Therefore, we run into the problem where if we make many comparisons, some may appear statistically significant just by chance. Therefore we need to make adjustments (which we can do in several different ways).

15.5 Multiplicity Adjustments

15.5.1 Tukey HSD (Honestly Significant Difference)

When making a pairwise comparison using the t-statistic, we have $\hat{\alpha_i} - \hat{\alpha_j} \pm t_\nu^{\frac{\alpha}{2}} \hat{SE}(\hat{\alpha_i} - \hat{\alpha_j})$, where $\hat{SE}(\hat{\alpha_i} - \hat{\alpha_j}) = \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$. However, due to the many comparisons we are making, there is a chance that something could appear outside of the CI bound due to random chance and appear statistically significant even though it is not. For example, for a 95% CI, we would expect on average that if we ran 20 non-significant pairwise comparisons, that one would be outside the CI due to chance. Thus, we make an adjustment to make the CI larger. For Tukey HSD, we replace $t_\nu^{\frac{\alpha}{2}}$ with $\frac{1}{\sqrt{2}} q_{I,\nu}^\alpha$. Thus, our CI using Tukey's HSD is:

$\hat{\alpha_i} - \hat{\alpha_j} \pm \frac{1}{\sqrt{2}} q_{I,\nu}^\alpha \hat{SE}(\hat{\alpha_i} - \hat{\alpha_j})$

For example, if $t_\nu^{\frac{\alpha}{2}}$ was around 2, then $\frac{1}{\sqrt{2}} q_{I,\nu}^\alpha$ might be around 3. Let us define more clearly where the $q_{I,\nu}^\alpha$ comes from. $q_{I,\nu}^\alpha$ is a critical point in the studentized range distribution.

What is the studentized range distribution? It is a distribution that is made from a studentized range. The studentized range is the difference between the largest and smallest datapoints in a sample divided by the standard deviation. Let $Y_1, ..., Y_I$ be the means of the various levels and $s^2 = \frac{\chi^2_\nu}{\nu}$, where ν is the degrees of freedom. Then our studentized range is $max_{i \neq j} \frac{Y_i - Y_j}{s}$.

Therefore, for a given significance level α using the studentized distribution, we have that $P(max_{i \neq j} \frac{|Y_i - Y_j|}{s} \geq q^{\alpha}_{I,\nu}) = \alpha$.

Let's look at the similarity between the t-statistic and the Tukey statistic when we make only one comparison. Note that the Tukey statistic is $\frac{|Y_1 - Y_2|}{s}$. The numerator $(Y_1 - Y_2) \sim N(0, 2)$, and the denominator is $\sqrt{\frac{\chi^2_\nu}{\nu}}$. This is equivalent to $\sqrt{2}t_\nu$! Therefore, we have the following expression:

$q^{\alpha}_{I,\nu} \geq q^{\alpha}_{2,\nu} = \sqrt{2}t^{\frac{\alpha}{2}}_{\nu}$

Thus, we can now understand the formula for the CI using Tukey's HSD.

When we create a CI for all the comparisons, what we are saying is that $P(\alpha_i - \alpha_j \in CI_{i,j} \text{ for all } i \neq j) \geq 1 - \alpha$. It actually equals $1 - \alpha$ in the case where the levels are balances (equal number of observations for each level (n_i are the same)). If the classes are very unbalanced, then the Tukey CI may be overly conservative.

15.5.2 Bonferroni Bounds

Bonferroni Bound is considered the most conservative. This simple method simply looks at the number of comparisons we are making, call it N, and we divide our significance level by it. Basically, if we make N comparisons, then our t-statistic is $t^{\frac{\alpha}{2N}}_{\nu}$ instead of the original $t^{\frac{\alpha}{2}}_{\nu}$.

15.5.3 Scheffe Bounds

Scheffe Bounds can be used if we are making comparisons not just of one vs. one, but possibly a linear combination of the means. For example, if we wanted to see if $\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4 + \alpha_5$, then we would use a Scheffe bound. The Scheffe statistic is $\sqrt{(I-1)F^{\alpha}_{I-1,\nu}}$. We pay for a larger test statistic since we are doing "more comparisons," not just one vs. one, but all the different combinations.

16 ANOVA - Two Factor

In two-factor (or two-way) ANOVA, we pull the ideas from the one-factor ANOVA, but we now have two different factors that we want to test. We also want to examine the interaction term between them. The model and assumptions in two-factor ANOVA are:

- Model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$
- n_{ij} observations at level $i = 1, ..., I$ for α and level $j = 1, ..., J$ for β; $k = 1, ..., n_{ij}$
- $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ (total number of observations)
- Gaussian Error Assumption (used for testing): $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$
- μ is the "overall mean"
- α_i is the mean of level i (for the first factor)
- β_j is the mean of level j (for the second factor)
- $(\alpha\beta)_{ij}$ is the mean of the interaction of level i and level j

Similar to one-factor ANOVA, the model above is not identifiable until we put some constraints on it. Again we can use two different methodologies, the treatment contrasts and the sum contrasts.

1. $\alpha_1 = 0$; $\beta_1 = 0$; $(\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$. This method is called "treatment contrasts."
2. $\sum_{i=1}^{I} \alpha_i = 0$; $\sum_{j=1}^{J} \beta_j = 0$; $\sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0$. This method is called "sum contrasts."

In two-way ANOVA, we want to test for interactions between the two factors. We can view this as a plot by plotting $\bar{y_{ij}}$ on the y-axis against either α_i or β_j on the x-axis. If we see a set of parallel lines, we can see there is no interaction. Otherwise there is interaction.

We can also test for interactions using an F-test. We test the full model (above) against the "additive model," which is simply $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$. Keep in mind that in two-way ANOVA, we have multiple tests that we are running: we want to see if the factor α is significant, the factor β is significant, and if the interaction between them is significant. Hence, when we run ANOVA, we obtain a table below:

Source	df	SS	Mean Sq.	F
α	$I-1$	SS_A	$MS_A = SS_A/(I-1)$	MS_A/MS_e
β	$I-1$	SS_B	$MS_B = SS_B/(J-1)$	MS_B/MS_e
$\alpha\beta$	$(I-1)(J-1)$	SS_{AB}	$MS_{AB} = SS_{AB}/((I-1)(J-1))$	MS_{AB}/MS_e
Resid	$n-IJ$	SS_e	$MS_e = SS_e/(n-IJ)$	

So how do we get the SS column, the sum of squares? Well, for each of these lines we have a small model ω and a large model Ω. The sum of squares is equal to $SS = ||\hat{Y}_\Omega - \hat{Y}_\omega||^2$, where $\hat{Y}_\Omega = H_\Omega Y$ and $\hat{Y}_\omega = H_\omega Y$. For SS_A the small model is just the column of 1's (the first column of the design matrix) and the large model is the column of 1's and all the columns corresponding to α, call them X_α. Therefore in the calculation of SS_A, to find $\hat{Y}_\Omega$, it is the projection onto $[1\ X_A]$. $\hat{Y}_\omega$ is the projection onto $[1]$. We can do the same to find SS_B. For SS_{AB}, $\hat{Y}_\Omega$ is the projection onto $[1\ X_A\ X_B\ X_{AB}]$. Then, $\hat{Y}_\omega$ is the projection onto $[1\ X_A\ X_B]$.

One note about two-factor ANOVA is that if $n_{ij} = 1$, basically if there is only one observation in a particular cell, then $n = IJ$ and we cannot test the interactions. There is a way called Tukey's "one degree of freedom" test for non-additivity that we can use.

Another note is that in this analysis we treat the levels of factors α and β as fixed. If the levels α_i (or β_j) are chosen in some random way from a larger population (so that we might model $\alpha_i \sim N(0, \sigma_\alpha^2)$ say), then a different, "random effects" model is used.

If instead of two factors, we have one continuous predictor and one factor, we use ANCOVA (analysis of covariance) instead of ANOVA, which has its own modeling and interpretation.

17 Generalized Linear Models

So far, the statistical analysis done has been assuming that $Y_i \overset{indep}{\sim} N(x_i^T\beta, \sigma^2)$, since we assumed Gaussian errors distributed $N(0, \sigma^2)$. This means that Y_i is a continuous random variable. But what happens if we want to model Y_i as discrete (say we want to look at counts). In that case, we need to use Generalized Linear Models (GLM). In a GLM, we have a linear regression similar to before $\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \eta_i$, and this is our "linear predictor." However, instead of just finishing there, we use a "link function," which links this result to what we are trying to model. Let's call what we are trying to model θ_i. Then the link function g links η_i to θ_i with the relationship $\eta_i = g(\theta_i)$. Therefore, we can expand our ideas from linear regression to the general setting by changing the link function to whatever fits the situation appropriately.

17.1 Binomial Regression

In Binomial Regression, we assume the $Y_i \overset{indep}{\sim} Bin(m_i, \theta_i)$. Therefore, the m_i are how many trials and θ_i is the success probability (as seen in the Binomial distribution). The link function for binomial regression is usually:

- Logit: $\eta = log(\frac{\theta}{1-\theta})$; Inverse Logit: $\theta = \frac{e^\eta}{1+e^\eta}$
- Probit: $\eta = \Phi^{-1}(\theta)$; Inverse Probit: $\theta = \Phi(\eta)$

One special case of binomial regression that is commonly used is logistic regression. The logistic regression is binomial regression when all the $m_i = 1$. It can be thought of as a yes/no or a 0/1 response.

17.2 Poisson Regression

In Poisson Regression, we are modeling counts. Similar to the Binomial Regression, we assume our Y_i's are non-negative. However, now we assume they take the Poisson distribution: $Y_i \overset{indep}{\sim} Poisson(\mu_i)$. The link function for Poisson Regression is usually:

- $\eta = log(\mu)$; Inverse: $\mu = e^\eta$

17.3 MLE Estimates

We use MLE (Maximum Likelihood Estimation) in order to solve for the $\hat{\beta}$ in GLMs. We will see that in the Gaussian case, this equates to Least Squares.

17.3.1 Gaussian

We assume that $Y_i \overset{indep}{\sim} N(x_i^T\beta, \sigma^2)$. Recall, the PDF of a Gaussian is $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Using MLE, the log-likelihood is

$logf(y|\beta, \sigma^2) = \sum_{i=1}^n -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2}\sigma^2(y_i - x_i^T\beta)^2$

We can see that the first term does not depend on β, therefore, to maximize this expression using β, we need to simply maximize $(y_i - x_i^T\beta)^2$, which is the same as least squares! Therefore, $\hat{\beta}_{MLE} = \hat{\beta}_{LS}$. If we then solve for the value of σ^2 that maximizes this likelihood, we get $\sigma^2_{MLE} = \frac{||y - X\hat{\beta}||^2}{n}$.

We can see that the estimates are the same as Least Squares!

17.3.2 One Binomial

Let's assume we only have one binomial, that is that $Y \sim Bin(m, \theta)$. Recall the PDF of a binomial is $P(X = k|m, \theta) = \binom{m}{k}\theta^k(1-\theta)^{m-k}$. Therefore our log-likelihood that we are trying to maximize is

$logf(y|\theta) = ylog\theta + (m - y)log(1 - \theta) + c(y)$, where $c(y)$ is a constant

Taking the derivative and setting it equal to 0, we get that

$\nabla l(\theta) = \frac{y}{\theta} - \frac{m-y}{1-\theta} = 0$
$\hat{\theta}_{MLE} = \frac{y}{m}$

17.3.3 Many Binomials

When we have many binomials, our log likelihood function becomes

$logf(y|\beta,\theta) = \sum_{i=1}^{n} y_i log\theta_i + (m_i - y_i)log(1-\theta_i) + c(y_i)$
$= \sum_{i=1}^{n} y_i log(\frac{\theta_i}{1-\theta_i}) + m_i(log(1-\theta_i)) + c(y_i)$
$= \sum_{i=1}^{n} y_i x_i^T \beta - m_i log(1 + e^{x_i^T \beta}) + c(y_i)$

If we take the derivative and set it equal to 0, we get that

$\frac{dlogf}{d\beta} = X^T[y - \mu(\beta)]$, where $\mu_i(\beta) = E[Y_i]$

In the Binomial case, $E[Y_i] = m_i\theta_i = m_i * g^{-1}(x_i^T\beta)$.

Note that when we set the derivative equal to 0 and rearrange, we get the expression $X^T\mu(\hat{\beta}) = X^Ty$. In the Gaussian case, $E[Y_i] = X\hat{\beta}$, so we have $X^TX\hat{\beta} = X^Ty$, so we obtain our normal equations.

The equation $X^T\mu(\hat{\beta}) = X^Ty$ is true for all the exponential family. Where they differ by distribution is how $\mu(\hat{\beta})$ is calculated. Similarly, asymptotically, we have that $\hat{\beta} \overset{D}{\approx} N(\beta, Var(\hat{\beta}))$. We compute $Var(\hat{\beta}) = (X^TV_{\hat{\beta}}X)^{-1}$. How $V_{\hat{\beta}}$ is calculated will differ by distribution.

For the Binomial distribution, $V_\beta = diag(m_i\theta_i(1-\theta_i))$. Again $\theta_i = g^{-1}(x_i^T\beta)$, where $g^{-1}(x) = log(\frac{x}{1-x})$. For the Poisson distribution, $V_\beta = diag(\mu_i)$. Here we have $\mu_i = g^{-1}(x_i^T\beta)$, where $g^{-1}(x) = e^x$.

17.4 Confidence Intervals

For GLMs, we can find confidence intervals as follows. We can asympotically assume normally for the estimates $\hat{\beta}$. Therefore, to produce a CI for β_i, we have $\hat{\beta}_i \pm z_{\frac{\alpha}{2}}\hat{SE}(\hat{\beta}_i)$.

For predictions intervals, we also use confidence intervals. Say we have an observation x_0, then our prediction using the linear model is $\hat{\eta}_0 = x_0^T\hat{\beta}$. We then calculate the variance of this estimator as $Var(\hat{\eta}_0) = x_0^TVar(\hat{\beta})x_0$. We then construct the CI for the linear part of the GLM as $\hat{\eta}_0 \pm z_{\frac{\alpha}{2}}\hat{SE}(\hat{\eta}_0)$. We then use the link function to obtain the confidence interval for the parameter of interest, say θ in the Binomial regression or μ in the Poisson regression. Our final confidence interval for parameter θ_0 is $\hat{\theta}_0 \pm g^{-1}(\hat{\eta}_0)$, where g is the inverse link function.

17.5 Likelihood Ratio Tests

When we fit models by using Maximum Likelihood Estimation, instead of comparing them using Sum of Squares, we use a metric called deviance. Deviance is to sum of squares what MLE is to least squares. It is a generalization of the idea of using the sum of squares of residuals in ordinary least squares to cases where model-fitting is achieved by maximum likelihood.

Assume we have fit a model by MLE and we want to test a small model against a larger model (similar to an F-test with OLS fitted model). Denote the smaller model as $\beta \in B \subset \Omega$. Then what we calculate is the Likelihood Ratio Test where

$LR = \frac{max_{\omega\in\Omega}L(\omega|y)}{max_{\beta\in B}L(\beta|y)}$

We are taking the ratio of the likelihoods of the larger model over the smaller model. If we take logs and write $logL(\beta|y)$ as $l(\beta)$ and similarly $l(\omega)$, then we have

$2log(LR) = 2[l(\hat{\omega}) - l(\hat{\beta})] \sim \chi^2_{dim(\Omega)-dim(B)}$

This quantity is called the residual deviance and has the approximate null distribution of the LR test where $dim(\Omega) - dim(B)$ is the difference between the number of parameters being estimated under Ω and B, the large and small model.

17.5.1 Under Gaussian Assumption

Under the Gaussian Assumption, we have that $Y_i \sim N(x_i^T\beta, \sigma^2)$. Let the small model be $\beta \in B_0$, call the estimates $\hat{\beta}_0$. Let the large model be $\beta \in B$, call the estimates $\hat{\beta}$. Say the design matrix X's block design looks like $[X_0\ X_1]$ and we are testing the null hypothesis that $\beta_1 = 0$ (the coefficients of the predictors that are the columns X_1).

In this scenario, using the likelihood of the Gaussian distribution, our residual deviance is

$2log(LR) = \frac{1}{\sigma^2}[||Y - X_0\hat{\beta}_0||^2 - ||Y - X\hat{\beta}||^2]$

Under the null hypothesis, this is distributed $\chi^2_{dim(B)-dim(B_0)}$. Therefore, we can check the value we get against the χ^2 table to see if we should reject or accept the null. Note that this distribution becomes the same as the F when $n \to \infty$.

17.5.2 Under Binomial Assumption

Under the Binomial assumption, where $Y_i \overset{indep}{\sim} Bin(m_i, \theta_i)$, our likelihood function is $\sum_{i=1}^n y_i log\theta_i + (m_i - y_i)log(1 - \theta_i)$. In that case, the residual deviance is

$G^2 = 2log(LR) = 2[l(\hat{\theta}) - l(\hat{\theta}_0)] = 2\sum_{i=1}^n ylog(\frac{y_i}{\hat{y}_i}) + (m_i - y_i)log(\frac{m_i - y_i}{m_i - \hat{y}_i})$, where $\hat{y}_i = m_i\hat{\theta}_i = m_i * g^{-1}(x_i^T\hat{\beta})$

Assuming our large model is the full model (saturated model), there are n df there. Assuming the small model has $p - 1$ df, then the residual deviance $G^2 \sim \chi^2_{n-p-1}$.

17.5.3 Analysis of Deviance

Similar to how we have seen ANOVA tables, in the LRT scenario we have a generalization of that, called that Analysis of Deviance table. Assume we have

- Two models: Ω_0 and Ω_1
- Dimensions: $p_0 < p_1$
- Predictors: $X_0\beta_0$ and $X\beta$

Our design matrix X in block form looks like $X = [X_0\ X_1]$. Then our Analysis of Deviance table looks like

Model	Resid df	Resid Dev	Term	Δ df	Δ Dev
null	$N - 1$				
Ω_0	$N - p_0$	$G^2_{\Omega_0}$	"Ω_0"		
Ω_1	$N - p_1$	$G^2_{\Omega_1}$	"Ω_1 adjusted for Ω_0"	$p_1 - p_0$	$G^2_{\Omega_0} - G^2_{\Omega_1}$

Therefore we have:

- Interpret sequentially, starting from the top and going to the bottom - each line adjusts for terms above and ignores those below.
- To assess the "significance of Δ Deviance $= G^2_{\Omega_0} - G^2_{\Omega_1}$, we use the nominal asymptotic approximation under H_0 true, we have $G^2_{\Omega_0} - G^2_{\Omega_1} \overset{H_0}{\sim} \chi^2_{p_1-p_0}$.
- The quality of this chi^2 approximation is often OK for the difference of deviances (Δ Dev above) especially since the number of parameters $p_1 - p_0$ is often small. However, more care is needed for goodness of fit tests.