Notes

test statistic = $\frac{\text{observed value - hypothesized value}}{\text{std dev of estimate (aka std error)}}$

e.g. **t-stat** of an estimator (such as of $\hat{\beta}_1$) for hypothesis testing = $\frac{\text{sample coefficient - hypothesized value}}{\text{std dev of coefficient (aka std error of coefficient)}}$

or **t-stat** of an estimator (such as of $\hat{\beta}_1$), from R regression output, (where hypothesized value = 0) = $\frac{\text{sample coefficient}}{\text{std dev of coefficient (aka std error of coefficient)}}$

p-value = P(Observed Test Statistic; given H_0 is true)

Example, using standard normal distribution:
z_0 = our Z test statistic, calculated using formula. The test statistic dependends on the particular type of test. p-value is calculated from the test statistic (use $\frac{\alpha}{2}$ for two-sided test).
z_α or $z_{\frac{\alpha}{2}}$ = our critical value (from table)

(1−α)% = confidence level
α = significance level = critical region = area under curve to the left or right of the **critical value** (for left-tailed or right-tailed, respectively) or both (for two-tailed test). See picture below.
p-value = area under curve to the left or right of the **test statistic** (for left-tailed or right-tailed, respectively) or both (for two-tailed test).

Hypothesis testing using z test statistic (standard normal)

	Left-tailed	Two-tailed	Right-tailed
Null Hypothesis	$H_0 = ?$	$H_0 = ?$	$H_0 = ?$
Alternative Hypothesis	$H_1 < ?$	$H_0 \neq ?$	$H_1 > ?$
Confidence level	(1−α)%	(1−α)%	(1−α)%
Traditional method: Uses **critical value**	...	...	...
critical value	$P(Z<z_\alpha) = \alpha$, find z_α	$P(Z<z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, find $z_{\frac{\alpha}{2}}$ qnorm(α)	$P(Z>z_\alpha) = \alpha$, find z_α qnorm(1 − α)
	reject H_0 if $z_0 < z_\alpha$	reject H_0 if $abs(z_0) > z_{\frac{\alpha}{2}}$	reject H_0 if $z_0 > z_\alpha$
p-value method: Uses **p-value**	...	...	...
p-value	$P(Z < z_0)$	$P(Z > \lvert z_0 \rvert) + P(Z < -\lvert z_0 \rvert)$ $= P(Z > \lvert z_0 \rvert) + P(Z < -\lvert z_0 \rvert)$ $= 2 * P(Z > \lvert z_0 \rvert)$	$P(Z > z_0)$
	pnorm(z0, lower.tail=TRUE)	2 * pnorm(abs(z0), lower.tail=FALSE)	pnorm(z0, lower.tail=FALSE)
	reject H_0 if p-val < α	reject H_0 if p-val < α	reject H_0 if p-val < α

	Traditional method	P-value method
reject H_0	If test statistic $z_0 \in$ critical region bounded by z_α (or $\pm z_{\frac{\alpha}{2}}$)	When p-val is low, H_0 must go [away]
fail to reject H_0	If test statistic $z_0 \notin$ critical region bounded by z_α (or $\pm z_{\frac{\alpha}{2}}$), fail to reject H_0	When p-val is high H_0 must fly [works]

In continuous RV: P(x<?) is same as P(x≤?)

Normal Distribution (same idea applies to T-distribution and χ^2-distribution, just different R functions):
$P(Z \leq z_0) = \text{pnorm}(z_0)$
$P(Z \geq z_0) = \text{pnorm}(z_0, \text{lower.tail=FALSE}) = 1 - P(Z \leq z_0) = 1 - \text{pnorm}(z_0)$
$P(Z > |z_0|) = P(Z \leq z_0) + P(Z \geq z_0) = 2 * P(Z \leq z_0)$ (or 2 * $P(Z \geq z_0)$) = 2^* pnorm(z_0)

Hypothesis testing using z test statistic (standard normal) - P-value method

```
z_0 = 0.7 #try z_0 = -0.7
# left-tailed test,  all commands below are identical
pnorm(z_0, lower.tail = TRUE)
#1 - pnorm(-z_0, lower.tail = TRUE) # note the negative sing
#1 - pnorm(z_0, lower.tail = FALSE) # note different tail

# Right-tailed test, all commands below are identical
1-pnorm(z_0, lower.tail= TRUE)
#pnorm(-z_0, lower.tail=TRUE) # note the negative sing
#pnorm(z_0, lower.tail=FALSE) # note different tail

# Two-tailed test, all commands below are identical
2 * pnorm(abs(z_0), lower.tail=FALSE)
# 2 * pnorm(-1 * abs(z_0), lower.tail=TRUE) # if z_0 > 0
# 2 * pnorm(z_0, lower.tail=TRUE) # if z_0 < 0
```

0.758036347776927

0.241963652223073

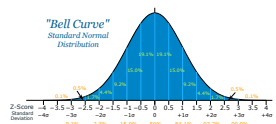
0.483927304446146

Hypothesis testing using t test statistic

	Left-tailed	Two-tailed	Right-tailed
Null Hypothesis	$H_0 = ?$	$H_0 = ?$	$H_0 = ?$
Alternative Hypothesis	$H_1 < ?$	$H_0 \neq ?$	$H_1 > ?$
Confidence level	(1−α)%	(1−α)%	(1−α)%
Traditional method: Uses **critical value**	...	...	...
critical value	$P(T<t_{\alpha,df}) = \alpha$, find $t_{\alpha,df}$ qt(α, df)	$P(T<t_{\frac{\alpha}{2},df}) = \frac{\alpha}{2}$, find $t_{\frac{\alpha}{2},df}$ qt(1 − $\frac{\alpha}{2}$, df)	$P(T>t_{\alpha}) = \alpha$, find $t_{\alpha,df}$ qt(1 − α, df)
	reject H_0 if $t_0 < t_{\alpha,df}$	reject H_0 if $abs(t_0) > t_{\frac{\alpha}{2},df}$	reject H_0 if $t_0 > t_{\alpha,df}$
p-value method: Uses **p-value**	...	...	...
p-value	$P(T < t_0)$	$P(T > \lvert t_0 \rvert) + P(T < -\lvert t_0 \rvert)$ $= P(T > \lvert t_0 \rvert) + P(T < -\lvert t_0 \rvert)$ $= 2 * P(T > \lvert t_0 \rvert)$	$P(T > t_0)$
	pt(t_0, df, lower.tail=TRUE)	2 * pt(abs(t_0), df, lower.tail=FALSE)	pt(t_0, df, lower.tail=FALSE)
	reject H_0 if p-val < α	reject H_0 if p-val < α	reject H_0 if p-val < α

	Traditional method	P-value method
reject H_0	If test statistic $t_0 \in$ critical region bounded by $t_{\alpha,df}$ (or $\pm t_{\frac{\alpha}{2},df}$)	When p-val is low, H_0 must go [away]
fail to reject H_0	If test statistic $t_0 \notin$ critical region bounded by $t_{\alpha,df}$ (or $\pm t_{\frac{\alpha}{2},df}$), fail to reject H_0	When p-val is high H_0 must fly [works]

Hypothesis testing using t test statistic - P-value method (same as z stat, just need df)



Hypothesis testing using F test statistic

	Right-tailed
Null Hypothesis	$H_0 = ?$
Alternative Hypothesis	H_1: At least one of the coefficients $\neq 0$
Confidence level	(1−α)%
Traditional method: Uses **critical value**	...
critical value	$P(F>F_\alpha) = \alpha$, find $F_{\alpha,df1,df2}$ qf(1 − α, df1, df2)
	reject H_0 if $F_0 > F_{\alpha,df1,df2}$
p-value method: Uses **p-value**	...
p-value	$P(F > F_0)$ pf(F_0, df1, df2, lower.tail=FALSE)
	reject H_0 if p-val < α

	Traditional method	P-value method
reject H_0	Reject H_0 if $F_{partial\ or\ overall} > F_{\alpha,df1,df2}$	When p-val is low, H_0 must go [away]
fail to reject H_0	Fail to reject H_0 if $F_{partial\ or\ overall} \leq F_{\alpha,df1,df2}$	When p-val is high H_0 must fly [works]

Hypothesis testing using F test statistic - Critical value, P-value methods (need df1 and df2, always right-tailed)

Example (partial F for a subset of coefficients, can also do overall F for all coefficients):
$y_{full} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
$y_{reduced} = \beta_0 + \beta_1 x_1 + \epsilon$
n = 22
restrictions = 2 = df1
k = 3
df2 = n-k-1 = 18

H_0: $\beta_2 = \beta_3 = 0$
H_1: At least one of the coefficients $\neq 0$

Critical value method: Reject H_0 if $F_{partial} > F_{\alpha,df1,df2}$
P-value method: Reject H_0 if p-value < α, where p-value = $P(F > F_{partial})$, right-tail shaded area, and $F \sim F_{df1,df2}$

```
In [10]: alpha = 0.01
         # F-test stat is always right-tail
         # see HW3 Self-Assessment example
         F_partial_2_18 = ((190.232 + 129.431)/2) / ((442.292)/18)
         F_alpha_2_18 = qf(1-alpha, df1=2, df2=18)
         p_val_F_partial_2_18 = pf(F_partial_2_18, df1=2, df2=18, lower.tail=FALSE)

         # Critical value method
         F_partial_2_18
         F_alpha_2_18

         # p-value method
         p_val_F_partial_2_18
         alpha
```

6.90467790509437

6.01290483480053

0.00748200247347742

0.01

R Functions and their defaults

x, q : vector of quantiles
p: vector of probabilities
df or df1, df2: degrees of freedom

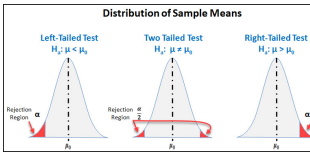
The Normal Distribution
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)

The Student t Distribution Functions in R
dt(x, df, ncp, log = FALSE) - density, yields density function value in a given point
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE) - probability, yields CDF, i.e. probability of returning number smaller than an argument to this function
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE) - quantile, inverse CDF, i.e. what value is at given quantile.
rt(n, df, ncp) - random generation for the t distribution with df degrees of freedom (and optional non-centrality parameter ncp).

The F Distribution
df(x, df1, df2, ncp, log = FALSE)
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
rf(n, df1, df2, ncp)

The (Non-Central) Chi-Squared Distribution
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)

Distribution of Sample Means



confidence interval = point estimate ± Probability Distribution pille * SD of point estimate = (?, ?)
statistically significant (0 ∈ interval)
statistically positive or statistically negative (only +ve or only -ve values ∈ interval)

prediction interval

The **prediction interval** should not be confused with a confidence interval for a fitted value, which will be narrower.

The prediction interval is used to provide an interval estimate for a prediction of y for one member of the population with a particular value of x_0:
The confidence interval is used to provide an interval estimate for the true average value of y for all members of the population with a particular value of x_0.

~ depending on the context: sigma can be of population, of linear regression error (aka variance of deviances ϵ), etc ~

SLR:
~ Theoretical model:
$y = \beta_0 + \beta_1 x + \epsilon$
$E[y|x] = \beta_0 + \beta_1 x$

~ Least squares line:
$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$

~ Residuals:
$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta_0} + \hat{\beta_1} x)$

ANOVA:
~ Residuals $\hat{\epsilon}_i = Y_{ij} - \hat{\mu}_i$, for i groups, j elements in each~

~error term can be from the theoretical model $y = \beta_0 + \beta_1 x + \epsilon$
or it can be from the fitted model (also called residuals or deviances) $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\epsilon}$, where $\hat{\epsilon}$ is also written as e (Roman letter) (notice all estimators of parameters have hats on them).

~ Note that $\hat{\sigma}^2$ is the estimate of the **variance of the error term** (i.e we may not know true ϵ and the its σ, we can only estimate them using $\hat{y}$ and its β_i~

Mean Squared Error MSE of regression $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-k} = \frac{\text{Sum of Squared Errors: SSE}}{n-k}$, where k:=number of variables in the regression due to which we lose degrees of freedom (such as intercept $\hat{\beta}_0$, $\hat{\beta}_1$). Note that while σ is one of the parameters of the regression, we do not lose a degree of freedom because of it.

~In a regression model, the predictor variables (aka X-variables, explanatory variables, covariates, etc.) are assumed to be fixed and known. They are not assumed to be random. All of the randomness in the model is assumed to be in the error term. Consider a simple linear regression model as standardly formulated: $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The error term, ϵ, is a random variable as well. But X is not assumed to be a random variable. (Of course, it might be a random variable in reality, but that is not assumed or reflected in the model.)~

~**sampling distribution**, which will yield the **confidence interval**, which is immediately analogous to the **test statistic**.~

In Regression Analysis Model (vs Optimization):

Response variable, aka dependent variable -is a Random Variable

- Varies with changes in predictors, along with random changes
- error term in regression is also a RV

Predicting variable, aka independent variable, aka predictor - Fixed Variable (even though in real life it's also RV)

- Does not change with response, but is set fixed before the response is measured.

Simple Linear Regression / Multiple Linear Regression - Assumptions Checks:

! note: In science, an empirical relationship or phenomenological relationship is a relationship or correlation that is supported by experiment and observation but not necessarily supported by theory. In regression modelling, we are looking for an empirical relationship between a response and one or more predictor variables ! note: SLR/MLR are your garden variety Ordinary Least Squares regression. OLS relies on constant variance assumption, if this is violated - see Weighted Least Squares.
! note: most plots are used in both SLR and ANOVA assumption checks. not sure if $E[\hat{\epsilon}_i] = 0$ or $E[\hat{\epsilon}_i^2] = 0$, similarly: not sure if $Var[\hat{\epsilon}_i] = \sigma^2$ or $Var[\hat{\epsilon}_i] = \sigma^2$
! etc. My guess is probably both, since $\hat{\epsilon}_i$ are "estimates" of ϵ_i
! note: fitted values vs residuals to identify nonlinearity, nonconstant variance, and presence of outliers
! note: possible variations: x vs y, x vs e_i, y vs e_i, $\hat{y}$ vs e_i

! note: plot x vs b in my notes means plot a on the x-axis and b on the y-axis.
! note: residuals = observed y - model-fitted $\hat{y} = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{\epsilon}_i$
! note: residuals vs fitted plot has same information as residuals vs predictor
! SLR: 1 intercept and 1 var, MLR: 1 intercept and many variables. ?? Best to plot $\hat{y}$ vs e instead of x vs e because there are many x
! note: SLR - marginal model, MLR -conditional model. Coefficients for the same variable in these 2 models can differ drastically in sign, magnitude, even significance.
! note: in lecture notes, Dr Serban uses residuals plot in SLR, but standardized residuals plot in MLR to assess constant variance assumption. In MLR, the rest of the assumptions diagnosed with plots of srmth vs residuals don't need to be using standardized ones.

SLR/MLR Objectives

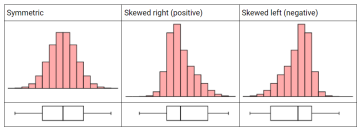
1. Prediction of response for new observation ("setting"),
2. Model the relationship between response var and predicting vars
3. Testing hypotheses (plural) on association relationships

Linearity Assumption (between x and y)

Means that $E[\epsilon_i] = 0$
Violation leads to difficulties in estimating $\hat{\beta}_0$ and that means that your model does not include a necessary systematic component.
(May need a transformation of just predicting variable x or both x and response y)

- Scatterplot of predicting vs predicted variable (x vs y)
  - A straight-ish line, going diagonally, either up or down
- **Scatterplot of predicting variable vs model residuals** (x vs e)
  - No pattern in the residuals with respect to predicting var, scattered around 0 (on y-axis)
  - If doing transformations, compare correlation coefficient before and after
- Scatterplot of fitted values vs residuals ($\hat{y}$ vs e) (PennState notes)
  - No patterns, scattered around 0 (on y-axis)

Linearity assumption of categorical (aka qualitative) variables vs response - only need/can to assess linearity of quantitative variables with respect to response. The means of response will vary per category, and whether they are statistically different can be checked by TukeyHSD command in R).



Variable types

- Controlling (control for selection bias)
  - Control variables are usually variables that you are not particularly interested in, but that are related to the dependent variables. You want to remove their effects from the equation. Consider, for example, you are interested in the difference in height of people from different countries. You could gather a sample of people from different countries and measure them and compare the heights. But you'd probably want to control for some other variables that are known to relate to height (e.g. gender). A controlling variable ??is a type confounding variables.
- Explanatory (explain variability)
  - explanatory variables aims to explain the variability in response with variability in predicting variables
- Predictive (minimize prediction error)
  - predictive variables aim to minimize the prediction error
- (NEW!) Confounding variables / omitted variables
  - They correlate with response and predictor(s). Need to be accounted for, if possible.

TL;DR: pecle f*ck up definitions of controlling and confounding variables.

There is a trade-off between bias and variance of the fitted model. With more predictors, the fitted model has smaller bias and larger variance, which implies although the mean estimated response is close to the ground truth, we are less confident (the confidence interval is wide) in a single prediction because of the large variance. On the other hand, less predictors result in larger bias with smaller variance. This time, the mean estimated response deviates farther from the ground truth, but we are more confident in the prediction.

Qualitative vs Quantitative:

Generally, can transform quantitative var into qualitative (e.g. when there is a non-linear relationship of that quant var with respect to response, e.g. when there are not a lot of observations for that variable, i.e. the year the movie was releases - if our data spans a small number of years, consider making years a categorical var)

Error Decompositions and F-tests

- Ways to test sihnificance of coefficient estimates:
  - ANOVA F-test on regression coefficient(s) - see below
  - t-test of the coefficient or its associated p-value

Constant Variance Assumption

Means that $Var[\epsilon_i] = \sigma^2$
Violations means estimates are not as efficient as they could be in estimating the true parameters, poor prediction intervals.
May need to do a transformation on response var (aka lambda transformation, aka Box-Cox transformation)

- *Scatter plot of the fitted values against the residuals* ($\hat{y}$ vs $\hat{\epsilon}_i$ (can also do y vs $\hat{\epsilon}$)
  - (NOT megaphone shape) the variance is not larger for larger fitted values
- **Scatterplot of predicting variable vs model residuals** (x vs e)
  - (NOT megaphone shape) the variance is not larger for larger predicting variable values

Independence assumption - CANNOT ASSESS WITH THIS PLOT

Means that $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ are independent Random Variables (RV).
Means that the deviances, or in fact the response variables y's, are independently drawn from the data-generating process.
Violation of this assumption can lead to very misleading assessments of the strength of the regression. This violation most often occurs in data that are ordered in time, like in time series data: auto-correlation.

Uncorrelated Errors - SETTLE FOR THIS

- *Scatter plot of the fitted values against the residuals* ($\hat{y}$ vs $\hat{\epsilon}_i$ (can also do y vs $\hat{\epsilon}$)
  - NO Clusters/groupings of the residuals
- **Scatterplot of predicting variable vs model residuals** (x vs e)

"Residual analysis does not check for the independence assumption. Remember, the assumption is independence, not uncorrelated errors. But all we can assess with residual analysis is uncorrelated errors. Independence is more complicated to evaluate. If the data are from a randomized trial, the independence is established. But most data you're going to apply regression on are from observational studies and thus independence does not hold. In those cases, we're going (residual analysis is going to) to assess uncorrelated errors, not independent errors.

SLR Error Decomposition

- Sum Square Errors SSE aka Sum Square Residuals (very confusing)
- n observations, k=2 coefficients, i.e. slope and intercept
- $y = \beta_0 + \beta_1 x_1 + \epsilon$

Sum Square Total SST = Sum Square Errors SSE (also SSResiduals) + Sum Square Regression S

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
$R^2 = 1 - \frac{SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = \frac{SSR}{SST}$
$MSE = \frac{SSE}{n-k} = \frac{SSE}{n-2} = \hat{\sigma}^2$
$MSR = \frac{SSR}{k-1} = \frac{SSR}{1}$
$MST = \frac{SST}{n-1}$

SLR F-test

H0: $\beta_1 = 0$, notice that β_0 is not a part of it.
H1: $\beta_1 \neq 0$
F-test test statistic $F_0 = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2}$
Critical value method: Reject H0 if $F_0 > F_{\alpha,k-1,n-k}$, a critical value depending on α, df1, df2
P-value method: Reject H0 if p-value < α, where p-value = $P(F > F_0)$, right-tail shaded area, and $F \sim F_{k-1,n-k}$

Normality Assumption Check

Means that $\epsilon_i \sim$ Normal, (thus $Y_i \sim$ Normal - lots of controversy on this one: $Y_i \sim$ Normal **given / conditional on values of X(s)**)
This is needed if we want to do any confidence or prediction intervals, or hypothesis test, which we usually do.
Violation makes hypothesis test and confidence and prediction intervals misleading
(May need to do a transformation on response var (aka lambda transformation, aka Box-Cox transformation)

- Q-Q plot
  - Roughly a diagonal line
- Histogram
  - Symmetric, no gaps, hopefully only 1 mode (if more - possibly selection bias, may need a control variable)
    - If there are multiple modes present, a categorical variable controlling for bias may be lacking

! note: A common misconception about linear regression is that it assumes that the outcome Y is normally distributed. Actually, linear regression assumes normality for the residual errors ϵ, which represent variation in Y which is not explained by the predictors. It may be the case that marginally (i.e. ignoring any predictors) Y is not normal, but after removing the effects of the predictors, the remaining variability, which is precisely what the residuals represent, are normal, or are more approximately normal. [Dr Serban: Y is normally distributed given X's.]
http://theobjectivestat.com/2013/08/07/assumptions-for-linear-regression/ (http://theobjectivestat.com/2013/08/07/assumptions-for-linear-regression/)

! note: It is reasonable for the residuals in a regression problem to be normally distributed, even though the response variable is not. Consider a univariate regression problem where $y \sim N(\beta_0 x, \sigma^2)$, so that the regression model is appropriate, and further assume that the true value of β_0 1. In this case, while the residuals of the true regression model are normal, the distribution of y depends on the distribution of x, as the conditional mean of y is a function of x. If the dataset has a lot of values of x that are close to zero and progressively fewer the higher the value of x, then the distribution of y will be skewed to the left. If values of x are distributed symmetrically, then y will be distributed symmetrically, and so forth. **For a regression problem, we only assume that response is normal conditioned on the value of x.**
https://stats.stackexchange.com/questions/12261/what-if-residuals-are-normally-distributed-but-y-is-not (https://stats.stackexchange.com/questions/12261/what-if-residuals-are-normally-distributed-but-y-is-not)

! note: predictors are not checked for normality, BUT if a predictor is strongly skewed, linearity with respect to response may not hold for it.

ANOVA - Assumptions Checks:

`anova (lm(y ~ x)) or anova (aov(y ~ x))` - answers the following question: are mean responses y of k groups within a qualitative variable x statistically different?
H_0 : $\mu_1 = \mu_2 = \ldots = \mu_k$
H_1 : At least one of the μ_i is different from the rest

`TukeyHSD(aov(y ~ x))` answers WHICH pairs of means are statistically different
For all categories k within a categorical variable x:
H_0 : $\mu_{ui} - \mu_{uj} = 0$
H_1 : $\mu_{ui} - \mu_{uj} \neq 0$

! note: most plots are used in both SLR and ANOVA assumption checks.
! note: residuals = observed Y - estimated mean for the groups = $Y_{ij} - \hat{\mu}_i = \hat{\epsilon}_{ij} = e_{ij}$
! note: plot residuals for each treatment group (note: aov() is a wrapper for the lm() function that produces an object that is basically an enhanced version of the model that would be produced by lm(). use either lm() or aov() to produce the model, and then pass that model to anova() to analyse it.

ANOVA objectives:

1. Compare variability within group to variability between groups
2. Testing for equal means ($\mu_1 = \mu_2 = \ldots = \mu_k$)
3. Estimation of simultaneous confidence intervals for differences of means ($\mu_i - \mu_j = 0$, for $i,j = 1..k$)

One-way ANOVA Error Decomposition

- k samples/groups with i=1 through k and d_i elements in each group with j=1 through x_i

$$\underbrace{y_{ij}}_{\text{jth element in ith group}} = \underbrace{\mu_i}_{\text{per sample/group mean}} + \underbrace{\epsilon_{ij}}_{\text{error}} \qquad \underbrace{\bar{y}}_{\text{overall mean}} = \underbrace{\mu}_{\text{overall mean}}$$

$$\hat{\mu}_i = \bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \qquad \bar{Y}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{N}$$

Sum Square Total SST = Sum Square Errors SSE + Sum Square due to Treatment SST_R

$\large \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y})^2$

- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$: total variability, last k degrees of freedom due to k group means
- $SST_R = \frac{SST}{k-1}$ = between-group variability
- $MST = \frac{SST}{N-1}$ - last 1 degree of freedom for substituting group mean with the overall mean

Variances in one-way ANOVA:

S^2_{pool} aka $\hat{\sigma}^2$ (general notation) aka MSE (concept) are estimators of σ^2
$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2}{N-k} = \frac{SSE}{N-k}$ = MSE, lost k degrees of freedom due to k group means
$s^2_0 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2}{N-1} = \frac{SST}{N-1}$ = MST, lost 1 degree of freedom for substituting group mean with the overall mean

One-way ANOVA F-test

- aka MLR (w/ intercept) F-test for all non-intercept regression coefficients
- k groups/samples/tabels - one factor
- N = total number of observations, n_i = number of records within each group
- $N = n_1 + n_2 + \ldots + n_k$

One-way ANOVA is an MLR with 1 categorical variable with k groups, transformed to dummy variables

- k-1 dummy vars in an intercept model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} + \epsilon$
- k dummy variables in a no-intercept model: $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$

ANOVA version:
H0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H1:At least one mean different from the rest

?!MLR version (regular, w/ intercept) / SLR with 1 categorical var converted to factor (aka dummy variables)
H0: $\beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$, notice that β_0 is not a part of it.
H1: At least one of the coefficients is different than 0.

F-test test statistic $F_0 = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{MST_R}{MSE} = \frac{SST_R/k-1}{SSE/N-k} \sim F_{k-1,N-k}$

Critical value method: Reject H0 if $F_0 > F_{\alpha,k-1,n-k}$, a critical value depending on α, df1, df2
P-value method: Reject H0 if p-value < α, where p-value = $P(F > F_0)$, right-tail shaded area, and $F \sim F_{k-1,N-k}$

TukeyHSD

Which of the means are statistically different?
TukeyHSD computes differences in all pairs of means and looks whether the (1−α)% confidence interval contains 0.
If it does, the difference between two given means is not statistically different than 0 (i.e. the difference is 0), thus the two given means are equal.
If it does not, the two means are statistically different from each other.

Two-way ANOVA - have not studied

- k groups/samples/tabels - one factor

MLR Error Decomposition

- Sum Square Errors SSE aka Sum Square Residuals (very confusing)
- k = p + 1, i.e. p variables and 1 intercept

Sum Square Total SST = Sum Square Errors SSE + Sum Square Regression SSR

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
$R^2 = 1 - \frac{SSE}{SST} = \frac{SST}{SST} - \frac{SSE}{SST} = \frac{SSR}{SST}$
$MSE = \frac{SSE}{n-k} = \frac{SSE}{n-p-1} = \hat{\sigma}^2$
$MSR = \frac{SSR}{k-1} = \frac{SSR}{p}$
$MST = \frac{SST}{n-1}$

MLR (w/ intercept) F-test on all non-intercept cofficients

- aka ANOVA for all non-intercept coefficients
- k= p + 1, i.e. p controlling factors, q explanatory factors, 1 intercept
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$

H0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, notice that β_0 is not a part of it.
H1: At least one of the coefficients is different than 0.

F-test test statistic $F_0 = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/n-p-1} \sim F_{p,n-p-1}$

Critical value method: Reject H0 if $F_0 > F_{\alpha,p,n-p-1}$, a critical value depending on α, df1, df2
P-value method: Reject H0 if p-value < α, where p-value = $P(F > F_0)$, right-tail shaded area, and $F \sim F_{p,n-p-1}$

MLR (w/ intercept) partial F-test

- aka ANOVA for a subset of regression coefficients, full model: controlling + other variables vs reduced model: just controlling variables
- k= p + q + 1, i.e. p controlling factors, q explanatory factors, 1 intercept
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + (\alpha_1 z_1 + \alpha_2 z_2 + \ldots + \alpha_q z_q) + \epsilon$

H0: $\alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$
H1: At least one of the coefficients for explanatory factors is different than 0

Partial F-test test statistic $F_{partial} = \frac{\frac{SSR_{full} - SSR_{reduced}}{df_{full} - df_{reduced}}}{\frac{SSE_{full}}{df_{full}}} = \frac{\frac{SSR_{full} - SSR_{reduced}}{q}}{\frac{SSE_{full}}{n-p-q-1}}$

Critical value method: Reject H0 if $F_{partial} > F_{\alpha,df_{full} - df_{reduced},df_{full}}$, a critical value depending on α, df1, df2
P-value method: Reject H0 if p-value < α, where p-value = $P(F > F_{partial})$, right-tail shaded area, and $F \sim F_{q,n-p-q-1}$

R commands: anova(fullmodel, reducedmodel)
R command: anova(fullmodel) - to get the values for the partial F-test stat (see lecture notes data example 3.4.1)

NO Linearity Assumption

Constant Variance Assumption

Means that $Var[\epsilon_i] = \sigma^2$
Violation makes inference on equality of means unreliable

- *Scatter plot of the fitted values against the residuals* ($\hat{y}$ vs $\hat{\epsilon}_i$)
  - (NOT megaphone shape) the variance is not wider **across groups** - this may be hard to achieve (e.g. response of control group vs some form of treatment group)

Independence assumption - CANNOT ASSESS WITH THIS PLOT

Means that $(\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ are independent Random Variables (RV).
Means that the deviances, or in fact the response variables y's, are independently drawn from the data-generating process.

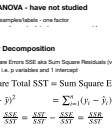
Uncorrelated Errors - SETTLE FOR THIS

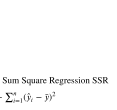
- Scatterplot of order of data collection (in time or space) vs residuals
  - Residuals randomly scattered around 0 (on y-axis), no patterns
- *Scatter plot of the fitted values against the residuals* ($\hat{y}$ vs $\hat{\epsilon}_i$)
  - NO clusters of errors **within groups**

Normality Assumption Check

Means that $\epsilon_i \sim$ Normal, (thus $Y_i \sim$ Normal)

- Q-Q plot
  - Roughly a diagonal line
- Histogram
  - Symmetric, no gaps





Intervals and Distributions

- note! In order to make statistical inference on the regression coefficients, we need to estimate the variance of the error terms (aka MSE, aka $\hat\sigma^2$), which gives us the standard deviation of the regression coefficient estimate (aka standard error); we also need normality assumption to hold to make inferences.
- note! $\hat\beta_i$ are a linear combination of normally distributed RV, so it's also normally distributed; but it's sampling distribution is a t-distribution (because we substitute σ with $\hat\sigma$)
- note! because $\hat\beta_i$ are normally distributed, so it $\hat y$; but it's sampling distribution is a t-distribution (because we substitute σ with $\hat\sigma$); if we do know σ, it's sampling distribution is normal.
  - variance of $\hat{\mathbf{y}}|\mathbf{x}^*$ is smallest if we look at regression line in the middle of the range of x, i.e. when $x^* = \bar x$. As x^* moves away from $\bar x$, in either direction, the variance increases (see formula to prove to self). Because variance is used to calculate confidence/prediction intervals, those too get wider as x^* moves away from $\bar x$.

SLR, k=2 (same applies to MLR, k > 2):
(1) $e_i = \hat\epsilon_i = y_i - \hat y_i = y_i = (\hat\beta_0 + \hat\beta_1 x_i)$
(2) $\hat\sigma^2 = \frac{\sum(\hat\epsilon_i)^2}{n-k}$
(3) Assumption $\hat\epsilon_i \sim \epsilon_i \sim N(0, \sigma^2)$
(4) Because we will never know β_0 and β_1, substitute $\hat\beta_0$ and $\hat\beta_1$ instead
(5) Then $\hat\sigma^2 \sim \chi^2_{n-k}$, loosing k=2 degrees of freedom for each substitution

$\hat\beta_1$
If (3) and (5) then $\hat\beta_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, but we may not know σ^2, must substitute with $\hat\sigma^2$
So $\frac{\hat\beta_1-\beta_1}{\sqrt{\frac{\hat\sigma^2}{S_{xx}}}} \sim t_{n-k}$ and thus Confidence Interval for $\hat\beta_1 = (\hat\beta_1 \pm t_{\frac{\alpha}{2},n-k}\sqrt{\frac{\hat\sigma^2}{S_{xx}}}) = (\hat\beta_1 \pm t_{\frac{\alpha}{2},n-k}\, \text{s.e.}(\hat\beta_1))$ (! assuming normality of residuals / response var **given x's**)

$\hat\beta_0$
$\hat\beta_0 = \bar y - \hat\beta_1 x_1$
Confidence Interval for $\hat\beta_0 = (\hat\beta_0 \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat\sigma^2(\frac{1}{n} + \frac{\bar x^2}{S_{xx}})}) = (\hat\beta_0 \pm t_{\frac{\alpha}{2},n-k}\, \text{s.e.}(\hat\beta_0))$ (! assuming normality of residuals / response var **given x's**)

$\hat{\mathbf{y}}|\mathbf{x}^*$
Confidence interval = $(\hat y|x^* \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat\sigma^2(\frac{1}{n} + \frac{(x^*-\bar x)^2}{S_{xx}})}) = (\hat y|x^* \pm t_{\frac{\alpha}{2},n-k}\text{s.e. }(\hat y|x^*))$ (! assuming normality of residuals / response var **given x's**)
Prediction interval = $(\hat y|x^* \pm t_{\frac{\alpha}{2},n-k}\sqrt{\hat\sigma^2(1 + \frac{1}{n} + \frac{(x^*-\bar x)^2}{S_{xx}})}) = (\hat y|x^* \pm t_{\frac{\alpha}{2},n-k}\text{s.e. }(\hat y|x^*))$ (! assuming normality of residuals / response var **given x's**)

One-way ANOVA
The individual sample variances for the k samples have a chi-square distribution because we assume that the data are normally distributed (see below: Sample Variance Estimator). An important property of the chi-square distribution is that, if we have independent chi-squared random variables, their sum is also a chi-square distribution.

Assume response var y (e.g. height) is normally distributed given predicting var x (e.g. voice pitch)
That is $Y_{1,i}, Y_{2,i}, \ldots, Y_{n,i} \sim N(\mu_i, \sigma^2)$
Then $s_1^2, s_2^2, \ldots, s_k^2$ are χ^2
Assume σ^2 is constant across all k samples/groups
$\frac{\text{SSE}}{\sigma^2} = \frac{n_1-s_1^2}{\sigma^2}+\ldots+\frac{n_k-s_k^2}{\sigma^2} \sim \chi^2_{N-k}$
$\text{MSE} = \hat\sigma^2 = \frac{\text{SSE}}{N-k} \sim \chi^2_{N-k}$

Assume response var y (e.g. height) is normally distributed given predicting var x (e.g. voice pitch)
That is $Y_{1,i}, Y_{2,i}, \ldots, Y_{n,i} \sim N(\mu_i, \sigma^2)$
Then $\hat\mu_i = \bar Y_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \sim N(\mu_i, \frac{\sigma^2}{n_i})$
However, we do not know σ^2, we replace it with $\hat\sigma^2$
So $\frac{\hat\mu_i-\mu_i}{\sqrt{\frac{\hat\sigma^2}{n_i}}} \sim t_{N-k}$ and thus Confidence Interval for means: $(\hat\mu_i \pm t_{\frac{\alpha}{2},N-k}\sqrt{\frac{\hat\sigma^2}{n_i}}) = (\hat\mu_i \pm t_{\frac{\alpha}{2},N-k}\text{s.e. }(\hat\mu_i))$

Summary:
sampling distribution of MSE = $\hat\sigma^2 = s^2_{pool} \sim \chi^2_{N-k}$
sampling distribution of $s_0^2 \sim \chi^2_{N-1}$
sampling distribution of $\frac{\hat\mu_i-\mu_i}{\sqrt{\frac{\hat\sigma^2}{n_i}}} \sim t_{N-k}$

One-way ANOVA: Pairs of means
Comparing all $\frac{k(k-1)}{2}$ pairs of treatments
q > t at any fixed α and degrees of freedom df, intervals are wider to compensate for the fact that we are making simultaneous (aka joint) comparisons (multiplicity correction)
$((\hat\mu_i - \hat\mu_j) \pm q_{\alpha,k,N-k}\sqrt{\frac{\hat\sigma^2}{2}(\frac{1}{n_i} + \frac{1}{n_j})}) = ((\hat\mu_i - \hat\mu_j) \pm q_{\alpha,k,N-k}\text{s.e.}((\hat\mu_i - \hat\mu_j)))$

MLR (w/ intercept), k = p + 1

- we are now dealing with vectors/matrices, not scalars

$\hat\beta_i$
$\hat\beta_i$ is a linear combination of $\{Y_1, Y_2, \ldots, Y_n\}$
Assumption $\hat\epsilon_i \sim \epsilon_i \sim N(0, \sigma^2)$
Then $\hat\beta_i \sim N(\beta, \Sigma)$
But since we do not know Σ, substitute with $\hat\sigma^2$
$\hat\sigma^2 \sim \chi^2_{n-p-1}$
So $\frac{\hat\beta_i-\beta_i}{\text{s.e.}(\hat\beta_i)} \sim t_{n-k}$ and thus Confidence Interval for $\hat\beta_1 = (\hat\beta_i \pm t_{\frac{\alpha}{2},n-p-1}\ \text{s.e.}(\hat\beta_i))$

$\hat{\mathbf{y}}|\mathbf{x}^*$

- note: $\hat y = \mathbf{x}^{*\intercal}\hat\beta$

Confidence Intervals for Regression Line - uncertainty due to estimation of coefficients
(1-α)% Confidence interval for the regression line (or mean respone) for **one** instance of predicting variables x
`$= (\hat y|x^ \pm t_{\tfrac{\alpha}{2}, n-p-1} \sqrt{\hat\sigma^2 \mathbf{x^}^\intercal (\mathbf{X^\intercal X})^{-1}\mathbf{x^}})$`

(1-α)% Confidence interval for the regression line (or mean respone) for **all** possible instances of predicting variables x*
The t critical point is replaced with the critical point based on the f-distribution which is meant to correct for the simultaneous inference across all xs.
`$= (\hat y|x^ \pm \sqrt{(p+1) F_{\alpha, p+1, n-p-1}} \sqrt{\hat\sigma^2 \mathbf{x^}^\intercal (\mathbf{X^\intercal X})^{-1}\mathbf{x^}})$`

Prediction Intervals for Regression Line - uncertainty due to estimation of coefficients, uncertainty due to new observations
(1-α)% Prediction interval for the regression line (or mean respone) for **one future** instance of response variable y* (at x*)
$= (\hat y|x^* \pm t_{\frac{\alpha}{2},n-p-1}\sqrt{1 + \hat\sigma^2\mathbf{x}^{*\intercal}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}^*})$

(1-α)% Prediction interval for the regression line (or mean respone) for **m future** instances of response variables y* (at x*)
The t critical point is replaced with the critical point based on the f-distribution which is meant to correct for the simultaneous inference across all xs.
`$= (\hat y|x^ \pm \sqrt{m F_{\alpha, m, n-p-1}} \sqrt{1 + \hat\sigma^2 \mathbf{x^}^\intercal (\mathbf{X^\intercal X})^{-1}\mathbf{x^}})$`

Summary Something is normally distributed. We may not know its true σ^2 so must use $\hat\sigma^2$ (which is chi-squared distributed). That someting now becomes t-distributed, so it's confidence interval uses t quantile.

Sample Variance Estimator (of any RV: response, error etc):
Assume $Z_1, Z_2, \ldots, Z_n \sim N(\mu, \sigma^2)$
$s^2 = \frac{\sum(Z_i-\bar Z)}{n-1}$, where $\bar Z = \frac{\sum Z_i}{n}$
$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
Loosing 1 degree of freedom when substituting μ with $\bar Z$

SLR/MLR Model Assessment

Things to evaluate for 1 model (or compare between 2 models):

1. Prediction/explanatory power: R-squared / Adjusted R-squared (for comparison) - if too high => may be overfitting data, cross-validation can tell us more about prediction
2. Goodness-of-fit: (regular or standardized) residual analysis done to assess (SLR or MLR, respectively) assumptions. Dr Serban: need to perform the assessment of the assumptions for goodness of fit. ~~overall F-test statistic p-value, p-values on predictors~~
3. Mean Square Error

Prediction Accuracy

In real life we most often don't have observed values right away to evaluate out predictions. But if we do, here are some measures:

Y_i = observed value
Y_i^* = predicted value
$\bar Y$ = mean of observed values used to test prediction
imdb.pred = predicted values
nimdb = observed values, used to test prediction

Mean Squared Prediction Error (MSPE) = $\frac{\sum_{i=1}^n (Y_i-Y_i^*)^2}{n}$ = mean((imdb.pred-nimdb)^2)

```
- OK for evaluating prediction accuracy of a linear regression
- depends on scale, susceptible to outliers
```

Mean Absolute Prediction Error (MAE) = $\frac{\sum_{i=1}^n |Y_i-Y_i^*|}{n}$ = mean(abs(imdb.pred-nimdb))

```
- NOT OK for evaluating prediction accuracy of a linear regression
- depends on scale, not susceptile to outliers
```

Mean Absolute Percentage Error (MAPE) = $\frac{\sum_{i=1}^n \frac{|Y_i-Y_i^*|}{Y_i}}{n}$ = mean(abs(imdb.pred-nimdb)/nimdb)

```
- NOT OK for evaluating prediction accuracy of a linear regression
- does not depend on scale, robust to outliers
```

Precision Measure (PM) = $\sum_{i=1}^n \frac{(Y_i-Y_i^*)^2}{(Y_i-\bar Y)^2} = \frac{\text{variability in prediction}}{\text{variability in new data}}$ = sum((imdb.pred-nimdb)^2)/(nimdb-mean(nimdb))^2)

```
- OK for evaluating prediction accuracy of a linear regression, best measure so far for lm
- similar to R^2 of a regression
- does not depend on scale
- PM -> 0 means better prediction
```

Last step: check whether predicted responses fall within prediction interval

Chi-square test - "goodness-of-fit"

- related: see testing the difference in two population proportions using Z test stastic (prop.test() i think)

Penn State:
Example, 2 categories (can be extended to k categories): Suppose the Penn State student population is 60% female and 40% male. Then, if a sample of 100 students yields 53 females and 47 males, can we conclude that the sample is (random and) representative of the population? That is, how "good" do the data "fit" the probability model
$H_0: p_F = 0.60$
$H_A: p_F \neq 0.60$

$Q_1 = \frac{(53-60)^2}{60} + \frac{(47-40)^2}{40} = 2.04$
Reject H_0 if $Q_1 \geq \chi^2_{0.05,1} (= 3.84)$

Chi-Square Test of Independence of Two Categorical/Qualitative Variables

Testing the independence of two categorical variables? Going forward, keep in mind that this Chi-square test, when significant, only provides statistical evidence of an association or relationship between the two categorical variables. Do NOT confuse this result with correlation which refers to a linear relationship.

chisq.test(rtdirector,awards)
Pearson's Chi-squared test
data: rtdirector and awards
X-squared = 26.192, df = 4, p-value = 2.895e-05

H_0 means rtdirector and awards are uncorrelated
H_1 means they correlated
Look at p-value to determine: p-value < alpha (=0.05), so the variables are correlated.

Review formulas

Population Parameters
population mean $\mu = \frac{\sum x_i}{N}$
population variance $\sigma^2 = \frac{\sum(x_i-\mu)^2}{N}$

Sample Statistics
sample mean $\hat\mu = \bar x = \frac{\sum x_i}{N}$
biased sample variance $s_n^2 = \frac{\sum(x_i-\bar x)^2}{n}$
unbiased sample variance $\hat\sigma^2 = s^2 = \frac{\sum(x_i-\bar x)^2}{n-1}$
Relationship between unbiased and biased sample variance $\hat\sigma^2 = (\frac{n}{n-1})s_n^2$

A different case (note the different denominator in $\hat\sigma^2$):
$\hat\sigma^2 = \frac{(x_i-\mu)^2}{n}$, where $\mu = \frac{\sum x_i}{N}$. But μ is often unknown, so use $\bar x$ instead of μ. This becomes $\bar s^2 = \frac{(x_i-\bar x)^2}{n}$

Standard Linear Regression (Simple Linear Regresion, Multiple Linear Regression) is a special case of Generalized Linear Regression

Dr Serban's diction

"Setting" in estimation vs prediction: instead of "setting", think "observation" (aka data point, aka row in dataframe)
"Confounding" and "controlling" variables are getting mixed up, but they are different things - we want controlling vars in the model, but not confounding (i think that is correct??)
"Inverse" is accidentally referred to as "indirect"

Outliers

Outlier - any point that is far from the majority of the data (x's and/or y)

- Leverage point - data point far from the mean of x
- Influential point - data point far from the mean of both x's and y

Outliers can be valid (perform analysis with and without them and look at differences) or errors in data entry/recording (need to be discarded)

Checking for outliers

Standardized residuals
$r_i^* = \frac{y_i-\hat y_i}{\sqrt{\text{MSE}}}$
Rule of thumb:

- if $|r_i^*| > 1$, then "large" outlier,
- if $|r_i^*| > 2$, then "extremely large"

Cook's distance
Measures how much all the values in the regression model change when the ith observation is removed.
Calculate Cook's distance D_i for each observation:
$D_i = \frac{(\hat Y_{(i)}-\hat Y)^\intercal(\hat Y_{(i)}-\hat Y)}{q\hat\sigma^2}$,
where $\hat Y_{(i)}$ is the model **without** the i-th observation, and $\hat Y$ are the fitted values from the model **with** the i-th observation (aka all observations included)

Rule of thumb:
$D_i > \frac{4}{n}$ or $D_i > 1$ or any "large" D_i should be investigated.

Multicollinearity

Multicollinearity (perfect collinearity, near collinearity) results in very large values of $\mathbf{X}^\intercal\mathbf{X}$, which results in very large standard error/variance/intervals.

Variance Inflation Factor VIF
VIF measures the proportional increase in the variance of $\hat\beta_j$ compared to what it would have been if the predicting variables had been completely uncorrelated.
What want to see is that the variance of $\hat\beta_j$ is not significantly larger when we have correlation among the predictive variables versus when we don't have correlation among the predictive variables, which means that multicollinearity will not cause a problem in the regression. So, we will compute VIF for every single predicting variables. If this condition holds for all the predicting variables, it means that the co-estimated coefficients are not likely to be unstable, so collinearity is not a problem. Again, it could be that the predicting variables to be correlated, but it doesn't necessarily mean that that will lead to a problem in the stability of the estimated regression coefficients.

Calculate VIF for each predicting variable:
$VIF_j = \frac{1}{1-R_j^2}$,
where R_j^2 is the coefficient of variation of the regression of the variable X_j on all other predicting variables.

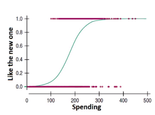
Rule of thumb:
Collinearity is NOT present if
$VIF_j < max(10, \frac{1}{1-R^2_{model}})$,
where R^2_{model} is the coefficient of variation of the regression model.

https://linareskevin.wordpress.com/2015/09/17/linear-regression-equation-in-latex-using-texmaths-under-libreoffice/
(https://linareskevin.wordpress.com/2015/09/17/linear-regression-equation-in-latex-using-texmaths-under-libreoffice/)

Here is the sample mean: \begin{document}
$$\bar y = \frac{1}{n}\sum_{i=1}^n y_i$$
Here is the sample variance:
$$\sigma^2 = \frac{\sum_{i=1}^n (y_i-\bar y)^2}{n-1}$$
Here is the sample standard deviation:
$$\sigma = \sqrt{\frac{\sum_{i=1}^n (y_i-\bar y)^2}{n-1}}$$
Regression Function OLS \begin{document}
Population regression line:
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$
Sample regression line:
$$\hat Y_i = \hat\beta_0 + \hat\beta_1 X_i + \hat\epsilon_i$$
Sample slope:
$$\hat\beta_1 = \frac{\sum(X_i-\bar X)(Y_i-\bar Y)}{\sum(X_i-\bar X)^2}$$
Sample intercept:
$$\hat\beta_0 = \bar Y - \hat\beta_1\bar X$$
Conditional Variance:
$$\hat\sigma^2 = \frac{\sum\hat\epsilon_i^2}{n-2} = \frac{\sum(Y_i-\hat Y_i)^2}{n-2}$$
Conditional Standard Deviation:
$$\hat\sigma = \sqrt{\frac{\sum(Y_i-\hat Y_i)^2}{n-2}}$$
Sample slope standard error:
$$\hat\sigma_{\hat\beta_1} = \frac{\hat\sigma}{\sqrt{\sum(X_i-\bar X)^2}}$$
Sample intercept standard error:
$$\hat\sigma_{\hat\beta_0} = \hat\sigma\sqrt{\frac{\sum(X_i)^2}{n\sum(X_i-\bar X)^2}}$$
\end{document}

Logistic Regression

Linear Regression = blue line



Logistic Regression = green line

@ spending >> 200: the probability of liking the new logo is so high that an additional point on the spending, adds little to the probability of liking the logo. Summary: the probability curve, as a function of spending, levels off for high values of spending, WHEN SPENDING INCREASES.
@ spending = 200: each additional dollar in the spending is associated with the fixed constant increase in the probability of liking the new logo.
@ spending << 200: the probability of liking the logo is so low, that one dollar lower on the spending subtracts little from the probability of liking the logo. Summary: he probability curve as a function of spending levels off for low values of spending, WHEN SPENDING DECREASES.

Model
response Y=0 or failure, Y=1 or success (though "success" does not necessarily mean desirable results)
linear model with p predictors and intercept: $p = P(Y = 1|x_1, \ldots, x_p)$ is the probability of success given predictors
(nonlinear) link functions link function $g(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ + ~~epsilon~~
logit(p) aka log odds ratio or log odds of success $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
rewritten as $\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$ => $p = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$

Logistic regression, with repeated trials: $Y_i \sim$ Binomial(n_i, p_i) with n_i>1
Logistic regression, without repeated trials: $Y_i \sim$ Binomial(n_i, p_i) aka Bernoulli(p_i)

- cannot perform calculation of residuals, hence cannot perform residual analysis for GoF.
- if all predictors are categorical, we can aggregate the data, thus allowing to find a subset of distinct (unique) combinations of predicting variables (i.e. $n_i > 1$).

Model Assumptions

Linearity Assumption - different than SLR/MLR

Linear relationship between predicting variables $x_1, \ldots, x_p$ and link function g, i.e. log odds success = $log(\frac{p}{1-p})$

```
- Plot: Predictors vs Residuals
- Plot: Predictors vs Logit of success rate
```

Independence Assumption - different than SLR/MLR

$Y_1, \ldots, Y_n$ are independent RV

```
- Plot: Predictors vs Residuals - cannot check independence, settle for uncorrelated errors
```

Link function - specific to logistic regression

Though logit function is most commonly used, sometimes the other ones may fit the data better.

- most common link function g: logit(p)=$log(\frac{p}{1-p}) = log(\frac{\text{probability of success}}{\text{probability of failure}}) = log(\text{odds ratio}) = log(\text{odds of success})$
  - advantages over the other 2 functions:
    - logit is a canonical link function, i.e. parameter estimates obtained under it are fully efficient and statistical tests on those parameters are better behaved for small samples
    - possibility of interpreting regression coefficient as log odds ratio or exp(regression coefficients) as odds ratio
- probit: inverse of cdf of std. normal, fits data with least heavy tails compared to other 2 link functions, so works well when probabilities are concentrated within a small range.
- complimentary log log. ##### Normality Assumption - CANNOT HAVE, response is binomial, HOWEVER need to check normality of residuals (Pearson residuals or deviance residuals - should follow approximately N(0,1) if to the model is a good fit)
- Check that residuals are **approximately** standard normal for GoF

```
- Plot: Q-Q plot
- Plot: Histogram
```

note! Logistic model has no error term!!
note! logistic regression/poisson regression relies on large sample size when testing (subsets) of coefficients, testing for statistical signficance on individual coefficients
note! careful with words like "response" and "significant". In SLR/MLR - response is the continuous variable we are modeling, in logistic regression - response is a binary variable, but we are modeling the probability of binary variable = "success", or 1 (doesn't necessarily mean that the outcome is desirable), which sometimes gets mislabled as response. "Significant" may mean statistically significant (e.g. the p-value is low), or "significant" may mean visible/large/enough.

Distributions

note!: In SLR/MLR we had an error term so we were able to distinguish between a response's confidence intervals and prediction intervals. In Poisson/logistic regression we have no error term! So we cannot really make that distinction. We do distinguish between in-sample predictions (using observations from training data to make a prediction), and out-of-sample predictions (using brand new, unused observations). We do still distinguish between a response interval and a coefficient interval though

$\hat\beta_j$ is approximately N(β, V)
(1-α)% Approximate Conf Interval (Wald test): $\hat\beta_j \pm z_{\frac{\alpha}{2}}\sqrt{V(\hat\beta_j)}$, as compared to SLR/MLR t-distribution (when variance unknown), or z-distribution (when variance is known)

Saturated model - when we estimate response, i.e. the probability of success, disregarding the predictors, that is we assume that estimated expected response is the observed response (in simple english: there is no model)
Fitted model - when we estimate response, i.e. the probability of success, using the predictors.

note ! [Residual] Deviance $\neq$ deviance residual d_i
note ! [Residual] Deviance = $\sum_{i=1}^n d_i^2$, sometimes it is used to calculate the test statistic, sometimes it is the test statistic.

Pearson residuals r_i = *std.zed (depends on definition) difference between observed response and expected/fitted response (cz observed and fitted responses have different variance)*
Deviance residuals d_i = *Signed square root of the log-likelihood evaluated at saturated model vs fitted model*
Deviance test stat (for GoF) $\sum_{i=1}^n d_i^2$ = *sum of deviance residuals*
OR Deviance test stat (for GoF) $\sum_{i=1}^n r_i^2$ = *sum of pearson residuals*
Deviance test stat (for all coefficients) = Null deviance - $\sum_{i=1}^n d_i^2$
Deviance test stat (for subset of coefficients) = Deviance of reduced model - Deviance of full model = $\sum_{i=1}^n d^2_{i,reduced} - \sum_{i=1}^n d^2_{i,full}$
Null deviance = hows how well the response variable is predicted by a model that includes only the intercept (grand mean).

Full model = model with all predictors
Reduced model = model with a subset of predictors removed (for testing on that subset of coefficients)
Null model = special case of reduced model, only intercept in the model

- deviance residuals `deviance_res = residuals(someLogisticModel,type="deviance")`
- pearson residuals `pearson_res = residuals(someLogisticModel,type="pearson")`
- Deviance using deviance residuals aka 'Residual deviance' aka difference between the model deviance and null deviance (per class TA) `deviance_using_deviance_res = sum(deviance_res^2)` (also `deviance(someLogisticModel)` or `someLogisticModel$deviance`)
- Deviance using pearson residuals `deviance_using_pearson_res = sum(pearson_res^2)`
- Null deviance `null_deviance = someLogisticModel$null.deviance`

For test on ALL regression coefficients (notice it's NOT an F test, it's Deviance test):
Model is $logit\ p(x_1, \ldots, x_p, z_1, \ldots, z_p) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$
$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$
H_1: at least one $\beta_i \neq 0$
P-value = P(χ^2_p > Deviance), reject H_0 if p-val < α

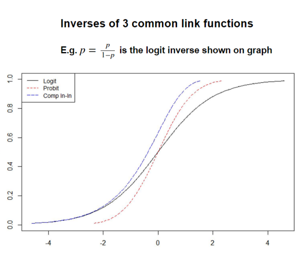
- `null_deviance` − Residual deviance `deviance_using_deviance_res = deviance_test_stat`
- `null_deviance` $\sim \chi^2_{n-1}$, Residual deviance `deviance_using_deviance_res` $\sim \chi^2_{n-p-1}$, `deviance_test_stat` $\sim \chi^2_p$
- Compute p-value for Deviance test statistic: `1 - pchisq(deviance_test_stat, df = p predictors)`

For test on SUBSET of regression coefficients(notice it's NOT an F test, it's Deviance test):
Full model is $logit\ p(x_1, \ldots, x_p, z_1, \ldots, z_p) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) + (\alpha_1 z_1 + \alpha_2 z_2 + \ldots + \alpha_q z_q)$
Reduced model is $logit\ p(x_1, \ldots, x_p, z_1, \ldots, z_p) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$
$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$
H_1: at least one $\alpha_i \neq 0$
P-value = P(χ^2_q > Deviance), reject H_0 if p-val < α

- Deviance of reduced model − Deviance of the full model= `deviance_test_stat`
- Deviance test statistic $\sim \chi^2_q$
- Compute p-value for Deviance test statistic: `1 - pchisq(deviance_test_stat, df= q predictors in subset)`

Somewhat related: Goodness of Fit: Hypothesis Testing (use Pearson or Deviance residuals for this Deviance test)
H_0: model fits data
H_1: model does NOT fit data
Deviance test stat (sum of deviance residuals) = $\sum_{i=1}^n d_i^2 \sim \chi^2_{n-p-1}$
OR Deviance test stat (sum of pearson residuals) = $\sum_{i=1}^n r_i^2 \sim \chi^2_{n-p-1}$
P-value = P(χ^2_{n-p-1} > Deviance), `1 - pchisq(deviance_using_deviance_res, df = n-p-1)`
OR P-value P(χ^2_{n-p-1} > Deviance), `1 - pchisq(deviance_using_pearson_res, df = n-p-1)`
Reject H_0 if p-val < α
THIS IS RARE WHEN WANT LARGE P-VALUES (SO THAT WE DO NOT REJECT H_0, MEANING MODEL FITS DATA)

For test of statistical significance on one predictor, given all other predictors in model: Wald test
See section Hypothesis testing using z test statistic for all 3 (left-tailed, 2-tailed, right-tailed)
$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$
Test stat: $\frac{\hat\beta_j-0}{\text{std error}(\hat\beta_j)}$, reject H_0 is |z-value| > $z_{\frac{\alpha}{2}}$



Cross-validation (recap from ISYE 6501)

1. Random subsampling (in 6501 there was train, test, AND validation)
```
for i = 1 to desired_repetitions:
  randomly split data into training set and testing set
  train, test and calculate classification error
  total_classification_error += classification_error
  i++
return mean(classification_error)
```
2. k-fold cross-validation
```
divide data into k chunks
for i = 1 to k:
  train model on all data except k subset
  test on k subset and calculate classification error
  total_classification_error += classification_error
  i++
return mean(classification_error)
```
3. leave-one-out cross-validation (n-fold cross-validation)
   - k is equal to n, number of observations

Predictive Power or Classification Error via Cross-Validation (at Different Thresholds)

For logistic regression, we need to pick a threshold t above which the fitted response > t gets rounded to 1, fitted response < t gets rounded to 0. Plot different tresholds t against their [cross-validated] classification errors at, so-called "elbow diagram".

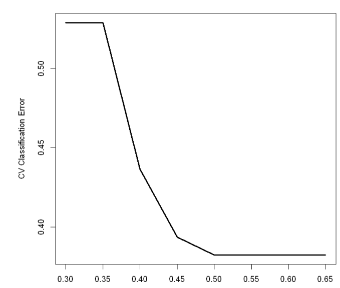
At different thresholds:
```
For t=0, 0.1, 0.2, ..., 0.9, 1 # do not make this a loop, so much harder to debug and understand
```

- compute the cost function:
```
cost_t = function(y, pi){
  ypred=rep(0,length(y))
  ypred[pi>t] = 1
  err = mean(abs(y-ypred)) # MAE
  return(err) # MAE
 }
```
- compute classification error for k-fold cross-validation (k is fixed here):
```
cv.err_t = cv.glm(obdata.fr,model,cost=cost_t, K=k)$delta[1] #cv.err_t is the mean of all MAEs in this case
t = t + 1/10 # can be 1/any_integer
```

Predictive accuracy (which is inversely proportional to classification error) - is comparing **classification error from cross-validation (i think it's the average of all MAEs) either at different thesholds (as above)** or for different models altogether.

- DO NOT FORGET TO COMPARE CLASSIFICATION ERROR OF A MODEL (I'm guessing at the threshold we picked) TO A CLASSIFICATION ERROR WHEN ALL RESPONSES ARE PREDICTED TO BE 1 OR 0, I.E. EQUAL TO A VALUE OF THE LARGEST CATEGORY. (I think this is equivalent to looking at cv.err_t when cost_0(y, pi) or cost_1(y, pi))
- If the classification error above 0.5 threshold is constant, bad news: our model has same or less predicting power than no model
- ```
  cv.err = c(cv.err0.35,cv.err0.35,cv.err0.4,cv.err0.45,cv.err0.5,cv.err0.55,cv.err0.6,cv.err0.65)
  plot(c(0.3, 0.35,0.4,0.45,0.5,0.55,0.6,0.65),cv.err,
    type="l",lwd=3,xlab="Threshold",ylab="CV Classification Error")
  ```

Cross-Validation Classification Errors at Different Thresholds



MLE

Maximum Likelihood Estimation (i.e. approach) or Maximum Likelihood Estimator (i.e. estimators obtained from the approach)

Poisson Regression

note! : Poisson, just like Logistic regression, has no error term
note! : Poisson, just like Logistic regression, relies on large sample size when testing (subsets) of coefficients, testing for statistical signficance on individual coefficients
note! : Most importantly for SLR/MLR under the assumption of normality, the statistical inference relies on the distribution that applies under both small and large samples. On the other hand, for logistic regression, the statistical inference based on the normal distribution applies only under large sample data.

note! : when the the number of counts per unit is small (i.e. a small rate, i.e. small response), the Poisson model will fit better than then SLR/MLR, because SLR/MLR assumes constant variance and this is not the case in Poisson. Notice it's the fit that's better, not statistical inference on the results. If we have larger counts per response, we can get away with SLR/MLR due to normality approximation

Model

Data: $\{(x_{11}, \ldots, x_{1p}), Y_1\}$, ... ,$\{(x_{n1}, \ldots, x_{np}), Y_n\}$, where Y_i is counts per unit
Model: log(E (y | $x_1, \ldots, x_p$) = $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ OR E (y | $x_1, \ldots, x_p$) = $e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$
When the "per" unit is not the same for all observations, cannot just scale response by dividing all counts by all units, use `glm(..., offset(some_units))` because R expects discrete counts, and division will cause non-discrete.

In poisson: E(Y) = V(Y) = λ = expected rate

Using SLR with log(response) will result in violations of the assumption of constant variance. We are to use Poisson Regression instead when data has small counts per unit. When data has large counts per unit, SLR could be used with transformed response (for stabilitation: $\sqrt{\mu + \frac{3}{8}}$) instead of log-transformed response

SLR w/ log(response)		Poisson Regression		
Expected(log(y) $	x_1, ..., x_p)=\beta_0 + \beta_1 x_1 + ... + \beta_p x_p$	log(Expected (y $	x_1, ..., x_p)=\beta_0 + \beta_1 x_1 + ... + \beta_p x_p< br > OR < /br >< br > Expected(y	x_1, ..., x_p)=e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}$</br>
Variance(log(y) $	x_1, ..., x_p$) = constant	log(Variance (y $	x_1, ..., x_p)=\beta_0 + \beta_1 x_1 + ... + \beta_p x_p< br > OR < /br >< br > Variance(y	x_1, ..., x_p)=e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}$</br>

Logistic regression	Poisson regression		
β_i = log of odds ratio = log of ratio of odds	β_i = log of rates ratio = log of ratio of rates		
$p(x_1, \ldots, x_p) = P(Y = 1\ x_1, \ldots, x_p) = \frac{e^{\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_1 x_1+\ldots+\beta_p x_p}}$	$log(\lambda_i(x)) = log(E(Y	x_1, ...,x_p)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p < br > OR < /br >< br >\lambda_i(x) = E(Y	x_1, ...,x_p) = \large e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}$</br>
link functions: logit, probit, complimentary log-log	link functions: log		
  ```

Model Assumptions

Linearity Assumption - different than SLR/MLR

Linear relationship between predicting variables $x_1, \ldots, x_p$ and log(Expected ly | $x_1, \ldots, x_p$)) aka log($\hat{u}(x)$) aka log(response)

- Plot: Predictors vs Residuals
- Plot: Predictors vs Logit of success rate

Independence Assumption - different than SLR/MLR

$Y_1, \ldots, Y_n$ are independent RV

- Plot: Predictors vs Residuals - cannot check independence, settle for uncorrelated errors

Variance Assumption

Expected ly | $x_1, \ldots, x_p$) = Variance ly | $x_1, \ldots, x_p$)= $e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$

Normality Assumption - CANNOT HAVE, response is binomial, HOWEVER need to check normality of residuals (Pearson residuals or deviance residuals - should follow approximately N(0,1) if to the model is a good fit)

- Check that residuals are **approximately** standard normal for GoF

- Plot: Q-Q plot
- Plot: Histogram

$Y_i | (x_{i1}, \ldots, x_{ip})^\top Poisson(\hat{u}(x_{i1}, \ldots, x_{ip}))$
Estimated rates $\lambda_i = \lambda_i(x_{i1}, \ldots, x_{ip}) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}$

Distributions - see Poisson section

For test on ALL regression coefficients (notice it's NOT an F test, it's Deviance test):
[Full] Model is: $log(E(Y|x_1, \ldots, x_p, z_1, \ldots, z_q)) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$
[Reduced] Model is $log(E(Y)) = \beta_0$
$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$
H_1: at least one $\beta_i \neq 0$
P-value = $P(\chi^2 > \text{Deviance})$, reject H_0 if p-val < α

- `null_deviance` = Residual deviance `deviance_using_deviance_res` = `deviance_test_stat` = null log-likelihood - full log-likelihood
- `null_deviance` $\sim \chi^2_{n-1}$ Residual deviance `deviance_using_deviance_res` $\sim \chi^2_{n-p-1}$, `deviance_test_stat` $\sim \chi^2_q$
- Compute p-value for Deviance test statistic: `1 - pchisq(deviance_test_stat, df = p predictors)`

For test on SUBSET of regression coefficients(notice it's NOT an F test, it's Deviance test):
Full model is $log(E(Y|x_1, \ldots, x_p, z_1, \ldots, z_q)) = \beta_0 + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) + (\alpha_1 z_1 + \alpha_2 z_2 + \ldots + \alpha_q z_q)$
Reduced model is $log(E(x_1, \ldots, x_p, z_1, \ldots, z_q)) = \beta_0 + (\beta_1 x_1 + \beta_2 c_2 + \ldots + \beta_p x_p)$
$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_q = 0$
H_1: at least one $\alpha_i \neq 0$
P-value = $P(\chi^2 > \text{Deviance})$, reject H_0 if p-val < α

- Deviance of reduced model — Deviance of the full model= `deviance_test_stat` = reduced log-likelihood - full log-likelihood
- Deviance test statistic $\sim \chi^2_q$
- Compute p-value for Deviance test statistic: `1 - pchisq(deviance_test_stat, df= q predictors in subset)`

Somewhat related: Goodness of Fit: Hypothesis Testing (see Pearson or Deviance residuals for this Deviance test)
H_0: model fits data
H_1: model does NOT fit data
Deviance test stat (sum of deviance residuals) = $\sum_{i=1}^{n} d_i^2 \sim \chi^2_{n-p-1}$
OR Deviance test stat (sum of pearson residuals) = $\sum_{i=1}^{n} r_i^2 \sim \chi^2_{n-p-1}$
P-value = $P(\chi^2_{n-p-1} > \text{Deviance})$, `1 - pchisq(deviance_using_deviance_res, df = n-p-1)`
OR P-value $P(\chi^2_{n-p-1} > \text{Deviance})$, `1 - pchisq(deviance_using_pearson_res, df = n-p-1)`
Reject H_0 if p-val < α
THIS IS INARE WHEN WANT LARGE P-VALUES (SO THAT WE DO NOT REJECT H_0, MEANING MODEL FITS DATA)

For test of statistical significance on one predictor, given all other predictors in model: Wald test
See section Hypothesis testing using z test statistic for all 3 (left-tailed, 2-tailed, right-tailed)
$H_0: \beta_i = 0$
$H_1: \beta_i \neq 0$
Test stat: $\frac{\hat{\beta}_i - 0}{\text{standard}(\hat{\beta}_i)}$, reject H_0 is |z-value| > $z_{\frac{\alpha}{2}}$
P-value: 2 * P(Z ≥ |z-value|), reject H_0 if p-val < α

Overdispersion in GLM - see Dr Serban notes

In this class

SLR/MLR:
goodness of fit assessment means analyzing residuals (from SLR - regular residuals, for MLR's constant variance - standardized [the lecture slides forgot to do that], for MLR's other assumptions - regular) to check if the 4 model assumptions hold,
R-squared is used for **predictive/explanatory power** (though cross-validation will tell us more on predictive power),
the overall F test, the partial F test, and the individual p-values of the coefficients is to determine statistical significance of all predictors excluding intercept, of a subset of predictors, and of an individual predictor, respectively;

Logistic Regression:
goodness of fit assessment means

- analyzing residuals (just deviance) to check if the 3 model assumptions hold (linearity, independence - more like uncorrelated errors, link function), AND checking those residuals are normally distributed (even though it's not in the original assumptions list)
- AND performing hypothesis testing on residuals (both Pearson and deviance), i.e. computing the p-value of the Deviance statistic D (where a LARGE p-value to NOT REJECT H0, meaning model fits data well),

predictive power means comparison of classification error from cross-validation at different thresholds

Poisson Regression:
goodness of fit assessment means

- analyzing residuals (just deviance) to check if the model assumptions hold (linearity, independence - more like uncorrelated errors, ?? variance assumption), AND checking those residuals are normally distributed (even though it's not in the original assumptions list)
- AND performing hypothesis testing on residuals (both Pearson and deviance), i.e. computing the p-value of the Deviance statistic D (where a LARGE p-value to NOT REJECT H0, meaning model fits data well),

predictive power means means comparison of classification error from cross-validation at different thresholds

From Avery Scott
Question on testing subsets of variables and model comparison:

- likelihood ratio tests compare models
  - Takes the form of partial F-test in multiple linear regression and simple linear regression
- Differences of the log of the differences in any generalized linear model
  - can compare any models with this concept

Model Selection

We cannot perform variable selection based on the statistical significance of the regression coefficients, statistical significance is only true in the context of that given model. Also, a better model can be found, even if the current one's variables are all statistically significant. It's possible to select a model to include variables that are not statistically significant, even though that model will provide the best prediction, for example, and vice versa.

Once the model selection yields a list of variables, re-fit model using glm() (or lm()) with the variables on their original scale

- Best Subset (not possible if p is large; 2^p subsets to check)
- Greedy algorithms (forward stepwise, backward stepwise, forward-backward stepwise)
  - Backward and forward stepwise regression will generally provide different sets of selected variables when p, the number of predicting variables, is large. Backward stepwise cannot be performed if p larger than n
  - Forward stepwise regression is preferable over backward stepwise regression because it starts with smaller models.
- Global algorithms - regularized regression (lasso, ridge (not for selection), elastic net = lasso + ridge)
  - Before fitting: Must standardize predicting variables (numerical ones), recommended to standardize response

Graphical Definition
We can create a graphical visualization of bias and variance using a bulls-eye diagram. Imagine that the center of the target is a model that perfectly predicts the correct values. As we move away from the bulls-eye, our predictions get worse and worse. Imagine we can repeat our entire model building process to get a number of separate hits on the target. Each hit represents an individual realization of our model, given the chance variability in the training data we gather. Sometimes we will get a good distribution of training data so we predict very well and we are close to the bulls-eye, while sometimes our training data might be full of outliers or non-standard values resulting in poorer predictions. These different realizations result in a scatter of hits on the target.

We can plot four different cases representing combinations of both high and low bias and variance.
(rina: how to remember: p - variance, variance - 1/bias)

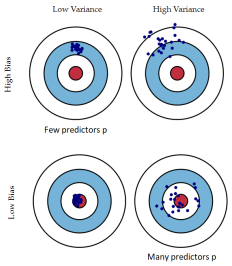


Fig. 1 Graphical illustration of bias and variance.

Notation

Given
S ⊂ {1, ⋯ , p} = a subset indices
{x_j for $j \in S$} = the subset of predicting vars with indices in S

for p predictor variables there are 2^p models to choose from

$\hat{\beta}(S)$ = estimated regression coefficients for the submodel with design matrix $X_s = \{x_j \text{ for } j \in S\}$ predicting variables
$\hat{Y}(S)$ = fitted value for submodel with {x_j for $j \in S$} predicting variables predicting variables
e.g. regression assuming normality $\hat{Y}(S) = X_s \hat{\beta}(S)$.
This will be referred to as S submodel

Prediction Risk Estimation (via Training Risk)

$R(S) = \frac{1}{n}\sum_{i=1}^{n} E(\overbrace{\hat{Y}_i(S)}^{\text{Prediction}} - \overbrace{Y_i^*}^{\text{Future observation}})^2 = \underbrace{\text{Variance of future observation } V(Y_i^*)}_{\text{irreducible error}} + \underbrace{\overbrace{\text{Bias Squared of prediction}}^{\text{Bias}^2(\hat{Y}_i(S))} + \overbrace{\text{Variance of prediction}}^{V(\hat{Y}_i(S))}}_{\text{Mean Squared Error, can be controlled}} =$ Prediction Risk for a submodel S

- Sometimes, it is possible to find a model with lower MSE than an unbiased model!
- It is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

For GLM, Prediction Risk = $-E[\log\text{-likelihood function}]$

Cannot obtain prediction risk at the time of prediction because we do not yet have the future observation.
To estimate prediction risk, substitute future observation Y_i^* with current observations Y_i:

$R_{train}(S) = \frac{1}{n}\sum_{i=1}^{n} E(\hat{Y}_i(S) - \overbrace{Y_i}^{\text{Current observation}})^2 =$ upward biased estimate of prediction risk since we used the data to both train the model AND estimate risk (data snooping).
$R_{train}(S)$ increases with the number of predictors p in the model, so **must correct for bias** by penalizing $R_{train}(S)$ so it doesn't automatically prefer a complex model aka a model with more predictors.
Thus,
$\hat{R}(S) = R_{train}(S) +$ Some Complexity Penalty CP:

Correcting Training Risk for bias in Linear Models:

Given |S| as number of predictors in the model and $\hat{R}(S) = R_{train}(S) +$ Some Complexity Penalty CP:

- using Mallow's CP: $\hat{R}(S) = R_{train}(S) + \frac{2|S|\hat{\sigma}^2_{\text{full model}}}{n}$, where $\hat{\sigma}^2_{\text{full model}}$ = estimated variance of full model (not always possible to estimate: e.g. when p > n)
- using Akaike IC CP: $\hat{R}(S) = R_{train}(S) + \frac{2|S|\sigma^2_{\text{full model}}}{n}$, where $\sigma^2_{\text{full model}}$ = true variance of full model, not estimated
  - Need to replace $\sigma^2_{\text{full model}}$ with $\hat{\sigma}^2_{\text{full model}}$ or $\hat{\sigma}^2_{\text{submodel}}(S)$.
  - Important! Most software replaces $\hat{\sigma}^2_{\text{full model}}$ with $\hat{\sigma}^2_{\text{submodel}}(S)$.
  - Akaike Information Criterion is an estimate for the prediction risk.
- using Bayesian IC CP: $\hat{R}(S) = R_{train}(S) + \frac{\log(n)|S|\sigma^2_{\text{full model}}}{n}$
  - Need to replace $\sigma^2_{\text{full model}}$ with $\hat{\sigma}^2_{\text{full model}}$ or $\hat{\sigma}^2_{\text{submodel}}(S)$
  - BIC penalizes complexity more than other approaches ⇒ preferred in model selection
- Leave-One-Out CV Approximation: $\hat{R}_{CV}(S) \approx R_{train}(S) + \frac{2|S|\hat{\sigma}^2_{\text{submodel}}(S)}{n}$
  - Because $\hat{\sigma}^2_{\text{submodel}}(S) \leq \hat{\sigma}^2_{\text{full model}}$, LOO CV Approximation CP penalizes complexity less than Mallow's CP

Leave-One-Out CV (a direct measurement of predictive power): $\hat{R}_{CV}(S) = \frac{1}{n}\sum_{i=1}(\hat{Y}_{(i)} - Y_i)^2$, where $\hat{Y}_{(i)}$ is $\hat{Y}$ estimated from a model fitted without observation i. LOO CV is approximately AIC when σ^2 is replaced by $\hat{\sigma}^2_{\text{submodel}}(S)$

Correcting Training Risk for bias in Generalized Linear Models (e.g. in Logistic or Poisson Regression):

Training risk $R_{train}(S)$ for a submodel S, fitted response for submodel S $\hat{Y}_i(S)$, future observation Y_i^*
$R_{train}(S) = \frac{1}{n}\sum_{i=1}^{n} D_i^* / \log(\frac{1}{f_i(S)}) + 2(\alpha_i - Y_i^*)\log(\frac{\alpha_i}{1-\alpha_i}) \overset{\alpha_i \to \hat{Y}_i(S)}{\to} =$ sum of square deviances of submodel S

- Akaike Information Criterion CP and Bayesian Information Criterion CP are commonly used for model selection in GLM since they are defined in terms of log-likelihood function

Penalizing the [Minimized] Sum of Squared Errors (aka Sum of Least Squares) in Regularized Regression

I think the $\hat{\beta}_i$ in my physical notes are meant to be $\hat{\beta}_{\lambda_i}$ so I'm fixing this
$\min Q(\beta_0, \ldots, \beta_p) = \sum_{i=1}(\beta_i - \hat{\gamma}_i)^2 + \lambda \cdot$ Some Penalty$(\hat{\beta}_1, \ldots, \hat{\beta}_p) = \sum_{i=1}(\beta_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}))^2 + \lambda \cdot$ Some Penalty$(\hat{\beta}_i$,

The penalty constant λ in penalized or regularized regression controls the trade-off between lack of fit and model complexity

Choices of Penalty:

- L_0 penalty: $||\beta||_0 = |\{i : \hat{\beta}_i \neq 0\}|$, provides best model given a selection criterion, but it requires fitting all possible submodels which may not be possible.
- L_1 penalty, Lasso: $||\beta||_1 = \sum_{i=1} |\hat{\beta}_i|$, measures sparsity
- L_2 penalty, Ridge: $||\beta||_2 = \sum_{i=1} \hat{\beta}_i^2$, easy to implement but it does not do variable selection, does not distinguish between sparse and non-sparse vectors

Note! For performing model selection using regularized regression, predicting variables MUST be ~~scaled~~ standardized, response is recommended to be ~~scaled~~ standardized. After selecting the "best" model, use the original scale when fitting the selected model for interpretation of the regression coefficients.
Note ! The L_1 penalty produces spare estimates, while the L_2 penalty does not.

$\frac{1}{n}\sum_{i=1} x_{ij} = 0, \quad \frac{1}{n}\sum_{i=1} x_{ij}^2 = 1$

$\frac{1}{n}\sum_{i=1} Y_i = 0, \quad \frac{1}{n}\sum_{i=1} Y_i^2 = 1$

Note on scaling vs standardization

Standardization is transforming your data so it has mean 0 and standard deviation 1, like a standard normal distribution: `x = (x - mean(x)) / sd(x)` . Confusingly enough, `scale()` in R will standardize data for you.
Scaling is transforming your data to a 0-1 range: `x = (x - min(x)) / (max(x) - min(x))`

Centering means substacting the mean of the random variable from the variables.
Scaling means dividing variable by its standard deviation.
Combination of the two is called standardization

Lasso	Ridge
Lasso performs variable selection	
Does not work when (number of predictors > number of observations)	Ridge does not perform variable selection
lasso can only select up to n vars.	
There is NO closed form regression for estimated regression coefficients, numerical algorithm required	There is a closed form regression for estimated regression coefficients
Estimated regression coefficients are less efficint than those from OLS: once model is selected using lasso, use OLS to estimate coefficients	
Lasso does not deal with multicollinearity: coefficients on most of the highly collinear predictors will be forced to zero, randomly (i.e. NOT based on theory)	Ridge deals with multicollinearity: a subset of highly collinear predictors with have very small coefficients.
Lasso forces some coefficients to be 0	Ridge shrinks coefficients towards 0, but does not force them to be 0

Elastic Net = Lasso + Ridge

$\sum_{i=1}(\beta_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}))^2 + \lambda_1 \cdot L_{1,\lambda_1 \text{Lasso}} + \lambda_2 \cdot L_{2,\lambda_2 \text{Ridge}} =$
$= \sum_{i=1}(\beta_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}))^2 + \lambda_1 \cdot \underbrace{\sum_{i=1} |\hat{\beta}_i|}_{L_{1,\lambda_1 \text{Lasso}}} + \underbrace{(1-\lambda_1)}_{\lambda_2} \cdot \underbrace{\sum_{i=1} \hat{\beta}_i^2}_{L_{2,\lambda_2 \text{Ridge}}}$

- `glmnet()` referent to the above formula:
  so when $\lambda_1 = 1 \Rightarrow$ lasso, when $\lambda_1 = 0 \Rightarrow$ ridge, when $0 < \lambda_1 < 1 \Rightarrow$ elastic net. (`glmnet()` refers to λ_1 as `alpha`)
- Elastic Net often outperforms Lasso in terms of prediction accuracy.
- λ gets derived from Cross-Validation: the selected λ yields minimal MSE in SLR/MLR or minimal Sum of Square Deviances in Poisson/Logistic

R libraries pertaining to Variable Selection Methods

`library(caret)` - Obtain Mallow's Cp, AIC, BIC criterion values for full model and submodel
`library(boot)` - used with `glm()` : 10-fold CV and leave one out CV
`library(leaps)` - stepwise model selection, searches over all possible 2^p submodels; cons: fitting all submodels may be impossible, does not have a AIC or BIC implementation, cannot force it to always include certain (controlling) variables in the model
`step(lm())` - stepwise model selection, more flexible: allows AIC and BIC, may specify to always include certain (controlling) variables in the model

- `forward.model = step(minimum, scope = list(lower=minimum, upper = full), direction = "forward")`
- `backward.model = step(full, scope = list(lower=minimum, upper = full), direction = "backward")`
- `both.min.model = step(minimum, scope = list(lower=minimum, upper = full), direction = "both")`
- `both.full.model = step(full, scope = list(lower=minimum, upper = full), direction = "both")`

`library(MASS)` - yet another implementation of Ridge
`library(lars)` - yet another implementation of Lasso
`library(glmnet)` - Elastic Net = Lasso + Ridge, Lasso, Ridge, Cross-Validation