CSCI 3022 Intro to Data Science

Workflow for SLR so far:

1. Plot the data as a scatter plot
   1.1 Does **linearity** seem appropriate?
   1.2 Calculate $\hat{\beta}_0, \hat{\beta}_1$ and overlay the best-fit line $y = \beta_0 + \beta_1 X$.
2. Consider assumptions of SLR:
   2.1 Plot a histogram of the residuals: are they **normal**?
   2.2 Plot the residuals against x: are they changing?
3. Perform inference (on βs or on values of $Y|X$)

Announcements and Reminders

Sch 3 posted

HW 4 & 5, & Prac 1.

- Last HW due Friday!
- Check out Canvas for:
    1. Some (3) past final exams to study from. (Modules)
    2. The first half of your Final Practicum - second half to be posted this week.
    3. Another textbook link for linear regression if you want another reference (Chatterjee & Hadi *Regression Analysis by Example.*)
- Final weeks' schedule: Today: MLR theory. Wednesday: Regression Notebooks. Friday: ANOVA. Monday: Logistic Regression. Wednesday: Review session, any extra notebooks. Friday: No class, Zach will publish an *untested/optional* lecture on stochastic gradient optimization.
- Deadlines: HW7: this Friday. Pen-and-paper exam: Thursday Apr 29. Practicum: May 2.

Where we at?

Last time we talked about multiple linear regression. It's like simple linear regression, except now we can attempt to predict Y with a variety of things, including both different *features/predictors* X as well as transformations and augmentations of the original variables, like using $\underline{x^2}$.

sm.ols(y, X)

$$X = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \\ 1 & x_n \end{bmatrix}$$

↑ intercept

simple linear regression

Process: try to fix "problems" with assumptions. This means:

1. Plot the linear model
2. See if some predictors are redundant
3. Plot residuals of linear model, check for **normality, independence, structure.**
4. Hit model with a math-shaped stick to fix these problems.

n x 2 matrix → n x 3 matrix

$$\begin{bmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{bmatrix}$$

⇒ fit: ax^2+bx+c

Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another.

Definition: *Covariance:*

The covariance between two rv's X and Y is defined as:

$$E[\underbrace{(X-\mu_X)}_{\text{X versus its mean}} \overbrace{(Y-\mu_Y)}^{\text{Y versus its mean}}]$$

If both variables tend to deviate in the same direction (both go above their means or below their means at the same time), then the covariance will be positive.

If the opposite is true, the covariance will be negative.

If X and Y are not strongly related, the covariance will be near 0.

$$\frac{Cov(X,Y)}{SD(x) \cdot SD(Y)}$$

Definition: *Correlation*

The *correlation* coefficient of X and Y, denoted by $Corr[X,Y]$ or just $\boxed{\rho,}$ is the *unitless* measure of covariance defined by:

MLR

Example: Suppose y is the sale price of a house that we wish to predict. Then sensible predictors include
x_1 = the interior size of the house,
x_2 = the size of the lot on which the house sits,
x_3 = the number of bedrooms,
x_4 = the number of bathrooms, and
x_5 = the house's age.

Multiple Linear Regression

p: # of predictors
n data points

Definition: *Multiple Linear Regression*
The multiple regression model is one where we allow each data point to have multiple characteristics (features/predictors) that we use to predict y. So for each data point we have p different X's to predict y:

linear!

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_p X_{p,i} + \varepsilon_i$$

In matrix form:

$\boxed{\underline{Y}} := [Y_1, Y_2, \dots Y_n]^T$ vector (length n)

$\boxed{\underline{\beta}} := [\beta_0, \beta_1, \dots \beta_p]^T$ vector (length $p+1$)

$$\boldsymbol{X} := \begin{array}{cc} \begin{array}{cc} \beta_0 \text{ column} & \text{columns for each } X \text{ predictor} \\ \downarrow & \end{array} & \\ \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix} & \left.\vphantom{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}\right\} n \times (p+1) \text{ matrix} \end{array}$$

So our model is $\underline{Y} = \boxed{\boldsymbol{X}}\underline{\beta} + \underline{\varepsilon}$. $\boldsymbol{X}$ is called the **design** matrix. 4 (familiar) assumptions:

linearity, independence, **identical**, **normality**. And we find the same *optimum:* the β values that would minimize sum of squared deviations $\sum_i (Y_i - \hat{Y}_i)^2$.

Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

With 3 assumptions on $\underline{\varepsilon}$:

Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$

   With 3 assumptions on $\underline{\varepsilon}$:

Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$

   With 3 assumptions on $\underline{\varepsilon}$:

2. Cov$[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$
   Independence of errors

Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$

   With 3 assumptions on $\underline{\varepsilon}$:

2. Cov$[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$
   Independence of errors

3. $Var(\varepsilon_i) = \sigma^2 \qquad \forall i$
   Homoskedasticity of errors

Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate: $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$

   With 3 assumptions on $\underline{\varepsilon}$:

2. Cov$[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$
   Independence of errors

3. $Var(\varepsilon_i) = \sigma^2 \qquad \forall i$
   Homoskedasticity of errors

4. $\varepsilon_i \sim N(0, 1)$
   Distribution of errors

Multiple Linear Regression Estimators

depends on α.

CI for things: $\text{coef} \pm \boxed{t_{\text{crit}}} \cdot \text{s.e.}(\text{coef})$

dof: $n-(p+1)$

Our estimators for the regression coefficients β rely on these assumptions, and look similar to those in SLR.

sm.OLS(y,X).fit()

summary.table:

coef std err t stat pvalue CI

fits α

↑

↓

95%

row for each $\hat{\beta}$

$\hat{\beta}$

error of $(\hat{\beta})$

ε: error

Multiple Linear Regression Estimators

Our estimators for the regression coefficients β rely on these assumptions, and look similar to those in SLR.

$$\hat{\beta} = \begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p\end{pmatrix} = (X^T X)^{-1} \boxed{X^T \underline{Y}}$$

looks kinda like

$$\sum_i X_i Y_i$$

like

$\frac{1}{x^2}$? $\underline{\text{Covariance}}$ of X

Multiple Linear Regression Estimators

Our estimators for the regression coefficients β rely on these assumptions, and look similar to those in SLR.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \left(X^T X\right)^{-1} X^T \underline{Y}$$

The $\left(X^T X\right)^{-1}$ bit corresponds to the $1/\sum\left(X_i - \bar{X}\right)^2$ part from before, where the $X^T \underline{Y}$ part corresponds roughly to a covariance between X and Y.

SLR: $\dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

Spread of X

MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is $n-(p+1)$ because $p+1$ df are "lost" in estimating the $p+1$ coefficients. This leads to an estimate for the standard errors of

MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

$$SSE = \sum_{i=1}^{n} \left(\underset{\text{data}}{Y_i} - \overset{\text{estimate}}{\hat{Y}_i} \right)^2$$

It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is $n - (p+1)$ because $p+1$ df are "lost" in estimating the $p+1$ coefficients. This leads to an estimate for the standard errors of

MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is $n-(p+1)$ because $p+1$ df are "lost" in estimating the $p+1$ coefficients. This leads to an estimate for the standard errors of

$$\hat{\sigma}^2 = \frac{SSE}{n-(p+1)}$$

variance for each Y_i

MLR Coefficient of Determination

Just as before, the total sum of squares is:

And the regression sum of squares is:

Then the coefficient of multiple determination R^2 is:

It is interpreted in the same way as in SLR: it is the proportion of variability in Y captured by our linear model.

MLR Coefficient of Determination

SLR: SST = SSR + SSE

total error/ Y movement

regression movement

errors/ residuals

Just as before, the total sum of squares is:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y_i}\right)^2$$

own mean

And the regression sum of squares is:

regression

$$SSR = \sum_{i=1}^{n} \left(\hat{Y_i} - \bar{Y_i}\right)^2$$

to a "flat" surface.

Then the coefficient of multiple determination R^2 is:

It is interpreted in the same way as in SLR: it is the proportion of variability in Y captured by our linear model.

MLR Coefficient of Determination

Just as before, the total sum of squares is:

$$SST = \sum_{i=1}^{n} \left(Y_i - \bar{Y_i}\right)^2$$

And the regression sum of squares is:

$$SSR = \sum_{i=1}^{n} \left(\hat{Y_i} - \bar{Y_i}\right)^2$$

Then the coefficient of multiple determination R^2 is:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

It is interpreted in the same way as in SLR: it is the proportion of variability in Y captured by our linear model.

higher good: max=1 min=0.

MLR Coefficient of Determination

Unfortunately, there is a problem with R^2: Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous.

From our example of predicting house pricing y before, suppose we also add these predictors to the model:
$x_6 =$ the diameter of the doorknob on the coat closet,
$x_7 =$ the thickness of the cutting board in the kitchen,
$x_8 =$ the thickness of the patio slab.

Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed y variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust R^2 to take account of the size of the model:

Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed y variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = \frac{SSR/\overset{\text{\# of predictors}}{\boxed{(p+1)}}}{SST/\underset{\text{\# data}}{\boxed{(n-1)}}} \quad \text{vs} \quad \frac{SSR}{SST}$$

We can choose to use more/less columns

Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed y variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = \frac{SSR/(p+1)}{SST/(n-1)}$$

Idea: as p grows, we have to improve SSR proportionately just as fast, or it wasn't worth the new parameter.

MLR Errors

SSR is still the basis for estimating the remaining model parameter, σ^2:

SSE.

MLR Errors

SSR is still the basis for estimating the remaining model parameter, σ^2:

$$\hat{\sigma^2} = \frac{SSE}{n - (p + 1)}$$

MLR Errors

We can use Python to compute the standard errors of the regression coefficients. By hand, one would use the distribution of the least squares estimator to calculate the standard errors:

With the standard error, we can compute confidence intervals:

We can also conduct hypothesis tests:

MLR Errors

We can use Python to compute the standard errors of the regression coefficients. By hand, one would use the distribution of the least squares estimator to calculate the standard errors:

$$\underline{\hat{\beta}} \sim \boxed{N}(\underline{\beta}, \boxed{Var[\hat{\beta}]})$$

estimate and

→ s.e. from summary table

or $\hat{\beta}_j \sim N(\beta_j, (s.e.(\hat{\beta}_j))^2)$

1) Coef are normal

With the standard error, we can compute confidence intervals:

$$CI\,\beta_j : \quad \underline{\hat{\beta}_j} \pm \boxed{t_{\alpha/2,n-(p+1)}} \cdot \boxed{s.e.(\hat{\beta}_j)}$$

estimate ±

We can also conduct hypothesis tests:

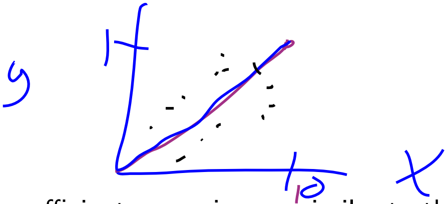
is slope for $\beta_j = 0$.

$$\boxed{H_0 : \beta_j = 0\ ;} \quad H_a : \beta_j \neq 0$$

"is this column currently useful"

for $j = 1, 2, \ldots p$, usually.

Collinearity



The $\left(X^TX\right)^{-1}$ term in our regression coefficient errors is very similar to the "spread of X" term in the SLR coefficients. This time, however, it's a little nastier: it's the spread of X across *all* p dimensions of the predictors. Example:

Suppose we have roughly linear data, and we decide to fit the data with the model $y = \beta_0 + \beta_1 x + \beta_2 \boxed{x^{1.000001}} + \varepsilon$

Collinearity

The predictors $x_1 = x$ and $x_2 = x^{1.000001}$ are very highly related. In particular, if the *true* value of the regression line is $y = x$, the following functions are visually nearly identical:

Collinearity

The predictors $x_1 = x$ and $x_2 = x^{1.000001}$ are very highly related. In particular, if the *true* value of the regression line is $y = x$, the following functions are visually nearly identical:

1. $y = x$
2. $y = x^{1.000001}$

Collinearity

The predictors $x_1 = x$ and $x_2 = x^{1.000001}$ are very highly related. In particular, if the *true* value of the regression line is $y = x$, the following functions are visually nearly identical:

1. $y = x$
2. $y = x^{1.000001}$
3. $y = .5x + .5x^{1.000001}$
4. $y = 2x^{1.000001} - x$

Collinearity

The predictors $x_1 = x$ and $x_2 = x^{1.000001}$ are very highly related. In particular, if the *true* value of the regression line is $y = x$, the following functions are visually nearly identical:

1. $y = x$
2. $y = x^{1.000001}$
3. $y = .5x + .5x^{1.000001}$
4. $y = 2x^{1.000001} - x$
5. $y = 10^6x - 999999x^{1.000001}$
6. $y = ax + bx^{1.000001}; \qquad \forall a + b = 1.$

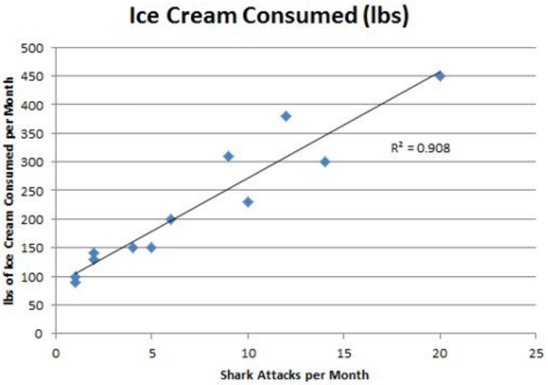
Collinearity

This is scary! The distribution of $\underline{\beta}$ has its own covariance, because the best choices for β_1 and β_2 may depend on each other. In the prior example, they would have a negative correlation of -1!.

In general, the interactions between coefficients is a function of the *linear independence* of the columns of the X matrix. In other words, we get a lot of negative effects if one predictor is describing one of the same things that we already have!

Correlations:

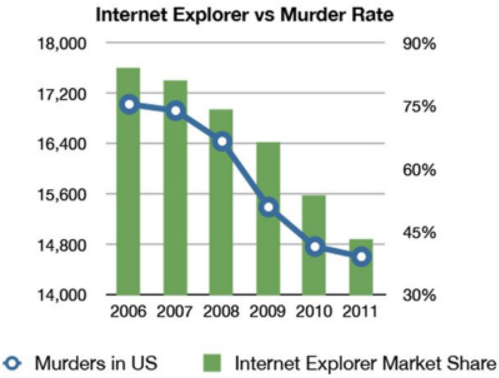
A SLR analysis of shark attacks vs ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.



MLR:

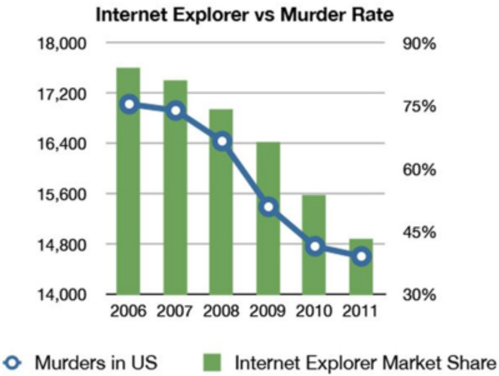
Suppose we included both **temperature** and shark attacks as features in our model of ice cream sales. What would happen? Which one should we probably exclude, and why?

Correlations:



Why is this correlated?

Correlations:



Why is this correlated?

MLR Variable Selection

"Full"

The selection process of variables for a multiple linear regression is complicated! Typically, we begin by *running the model with all parameters included.* Then - **one at a time** - we discard the least useful columns of the $\boldsymbol{X}$ matrix. The following are *all* possible criteria for excluding a column.

1. Check your X-values for *redundant*, non-independent information. Discard an offending column, then repeat. The measure for this is a *variance inflation factor*.

   statsmodels.variance_inflation_factor

2. The estimator $\hat{\beta}_i$ for predictor column i is *not significantly different from zero* per a $t-$test. This means that in the presence of the other features, it's not really helping us predict y!

   Check p-value for each coefficient.

3. The model with predictor i has lower *adjusted* R^2 than the model without it.

   "stepwise" optimization

4. The *variance captured* by the model with predictor i has lower *adjusted* R^2 than the model without it.

MLR Variable Selection

Typically we **always** include the first criteria, and then choose **one** of the other measures for improvement.

1. Discard x with high *variance inflation factors*.
2. Individual coefficient $t-$test significance.
3. Lower *adjusted* R^2 without a predictor.
4. The *variance captured* by the model with predictor i has lower *adjusted* R^2 than the model without it.

overall Compare SSE #1 to SSE #2.

Sometimes you may even want to include (non-significant predictors.) Why? This may help the model be more *predictive*, even if we can't statistically point to which factors matter the most. A smaller model does a better job at **suggesting** causal relationships, but we never actually get true causality, so sometimes a more accurate prediction is all we want!

$Pval > \alpha$

Special Cases: Variance

Definition: *Chi-Squared*
Let ν be a positive integer. The random variable X has a chi-squared distribution with parameter ν if the pdf of X is:

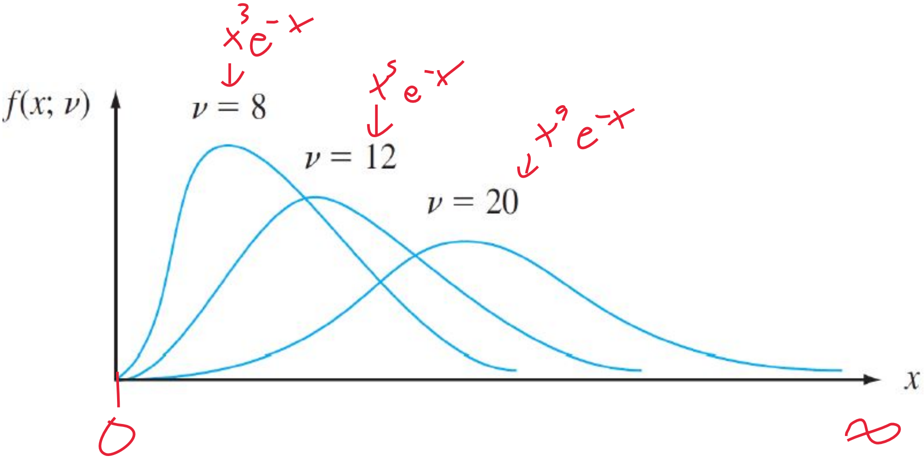
$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\underset{\uparrow}{\Gamma}(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The parameter ν is called the number of degrees of freedom (df) of X. The symbol $\boxed{\chi^2}$ is often used in place of "chi-squared."

χ^2: R.V. For Variances

Chi-Squared

Special Cases: Variance



Special Cases: Variance

Let $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then the random variable:

has a chi-squared (____) probability distribution with $n-1$ df.

(In this class, we don't consider the case where the data is not normally distributed.)

Special Cases: Variance

Let $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then the random variable:

$$\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

Sum of squared deviations

Sample variance

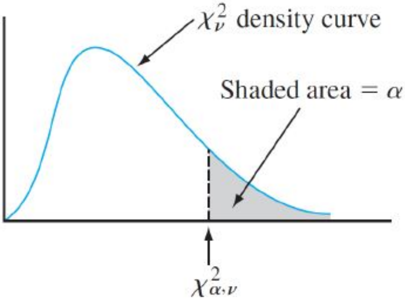
true variance

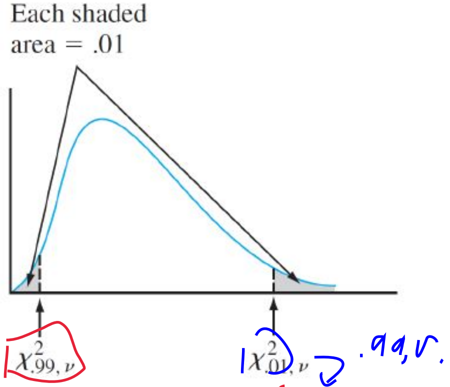
has a chi-squared ($\underline{\chi^2 \quad\quad}$) probability distribution with $n-1$ df.

(In this class, we don't consider the case where the data is not normally distributed.)

Special Cases: Variance

The chi-squared distribution is not symmetric, so these tables and functions contain values of _____ both for near 0 and 1.



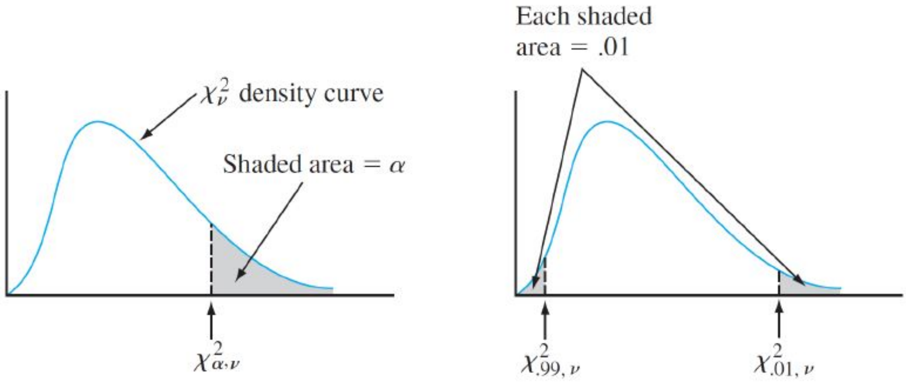


$.99, \nu$.

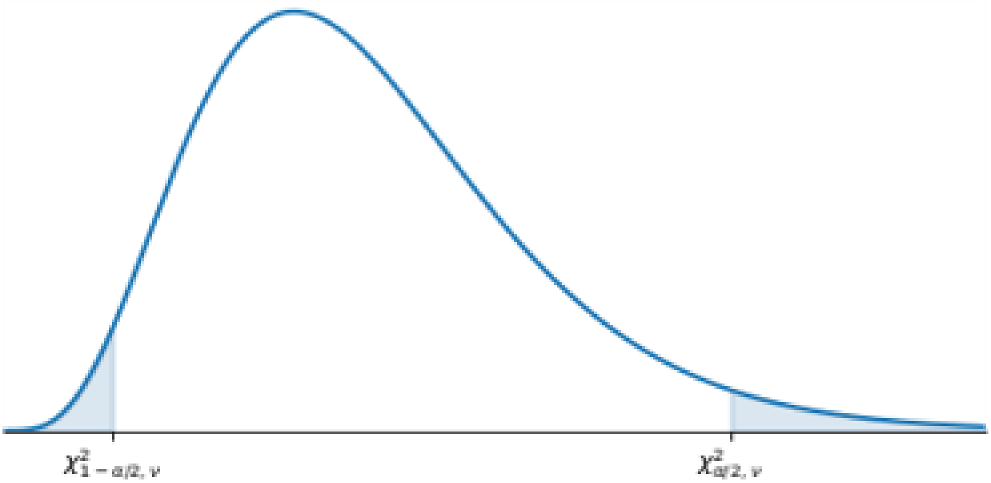
stats.chisq.ppf(.01 , ν)

Special Cases: Variance

The chi-squared distribution is not symmetric, so these tables and functions contain values of $\underline{\chi^2_\alpha\quad\quad}$ both for near 0 and 1.



Two tailed χ^2



Special Cases: Variance

As a consequence:

$$1 - \alpha = P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right)$$

Or, equivalently:

Thus we have a confidence interval for the variance. Taking square roots gives a CI for the standard deviation.

Special Cases: Variance

As a consequence:

$$1-\alpha = P\left(\chi^2_{1-\alpha/2,n-1} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}\right)$$

$$= P(1/\chi^2_{1-\alpha/2,n-1} \ge \frac{\sigma^2}{(n-1)s^2} \ge 1/\chi^2_{\alpha/2,n-1})$$

Or, equivalently:

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)$$

CI for variance

is a $100\%(1-\alpha)$ CI for σ^2.
Thus we have a confidence interval for the variance. Taking square roots gives a CI for the standard deviation.

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$$

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{.975,9} = \texttt{stats.chi2.ppf(0.025,9)} = 2.70$$

$$\chi^2_{\alpha/2,n-1} = \chi^2_{.025,9} = \texttt{stats.chi2.ppf(0.975,9)} = 19.02$$

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{.975,9} = \texttt{stats.chi2.ppf(0.025,9)} = 2.70$$

$$\chi^2_{\alpha/2,n-1} = \chi^2_{.025,9} = \texttt{stats.chi2.ppf(0.975,9)} = 19.02$$

$$\frac{(10-1)4.2}{19.02} < \sigma^2 \frac{(10-1)4.2}{2.70} \implies 1.99 < \sigma^2 < 14.0$$

$$\implies \sqrt{1.99} < \sigma^2 < \sqrt{14.0}$$

MLR

- Full Model vs. reduced model

Test for Equivalence of Variance

Compae 2 variances: - SSE

The F probability distribution has two parameters, denoted by ν_1 and ν_2. The parameter ν_1 is called the numerator degrees of freedom, and ν_2 is the denominator degrees of freedom.

A random variable that has an F distribution cannot assume a negative value. The density function is complicated and will not be used explicitly, so it's not shown.

There is an important connection between an F variable and chisquared variables.

Test for Equivalence of Variance

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

can be shown to have an F distribution.

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

Test for Equivalence of Variance

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

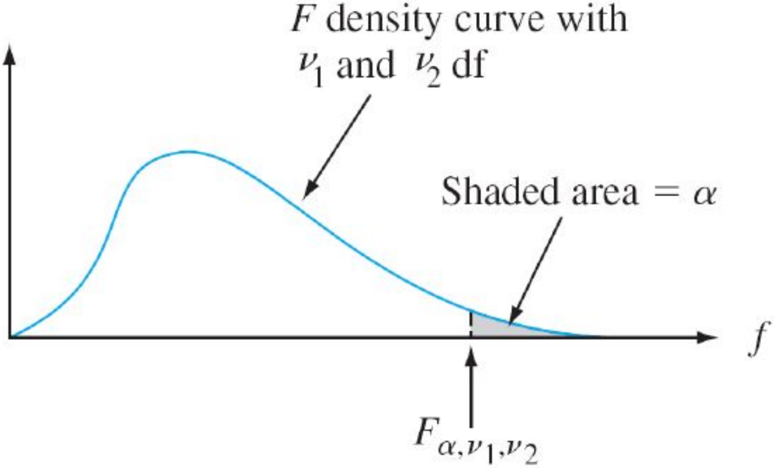
$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

can be shown to have an F distribution.

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

Test for Equivalence of Variance

Figure below illustrates a typical F density function.:



Test for Equivalence of Variance

We use F_{α,ν_1,ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,\nu_1,\nu_2} =$$

Test for Equivalence of Variance

We use F_{α,ν_1,ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_1,\nu_2}}$$

For example, $F_{.05,6,10} = 3.22$ and $F_{.95,10,6} = 0.31 = 1/3.22.$

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let $X_1, X_2, \dots, X_m$ be a random sample from a normal distribution with variance σ_1^2 let $Y_1, Y_2, \dots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let $X_1, X_2, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2 let $Y_1, Y_2, \ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

Test for Equivalence of Variance

This theorem results from combining the fact that the variables $\frac{(n-1)s_2^2}{\sigma_2^2}$ and $\frac{(m-1)s_1^2}{\sigma_1^2}$ each have a chi-squared distribution with $n-1$ and $m-1$ df, respectively.

Because F involves a ratio rather than a difference, the test statistic is the ratio of sample variances.

The claim that $\sigma_1^2 = \sigma_2^2$ is then rejected if the ratio s_1^2/s_2^2 differs by too much from 1.

Test for Equivalence of Variance

Null hypothesis: H_0 :

Test statistic value:

Alt Hypothesis   Rejection Region   p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2/s_2^2$$

Alt Hypothesis	Rejection Region	p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$		
$H_a : \sigma_1^2 < \sigma_2^2$		
$H_a : \sigma_1^2 \neq \sigma_2^2$		

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$	$F_{stat} > F_{\alpha,m-1,n-1}$	
$H_a : \sigma_1^2 < \sigma_2^2$	$F_{stat} < F_{1-\alpha,m-1,n-1}$	
$H_a : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} < F_{1-\alpha/2,m-1,n-1}$ OR $F_{stat} > F_{\alpha/2,m-1,n-1}$	

Test for Equivalence of Variance

Null hypothesis: H_0 :$\sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2/s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$	$F_{stat} > F_{\alpha,m-1,n-1}$	$P(F_{m-1,n-1} > F_{stat})$
$H_a : \sigma_1^2 < \sigma_2^2$	$F_{stat} < F_{1-\alpha,m-1,n-1}$	$P(F_{m-1,n-1} < F_{stat})$
$H_a : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} < F_{1-\alpha/2,m-1,n-1}$ OR $F_{stat} > F_{\alpha/2,m-1,n-1}$	(OR)

Test for Equivalence of Variance

Example: On the basis of data reported in the article “Serum Ferritin in an Elderly Population” (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha $=$ .01.

Test for Equivalence of Variance

Example: On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha = .01.

$$F_{27,25} = \frac{52.6}{84.2} = F_{stat}$$

Test for Equivalence of Variance

Example: On the basis of data reported in the article “Serum Ferritin in an Elderly Population” (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

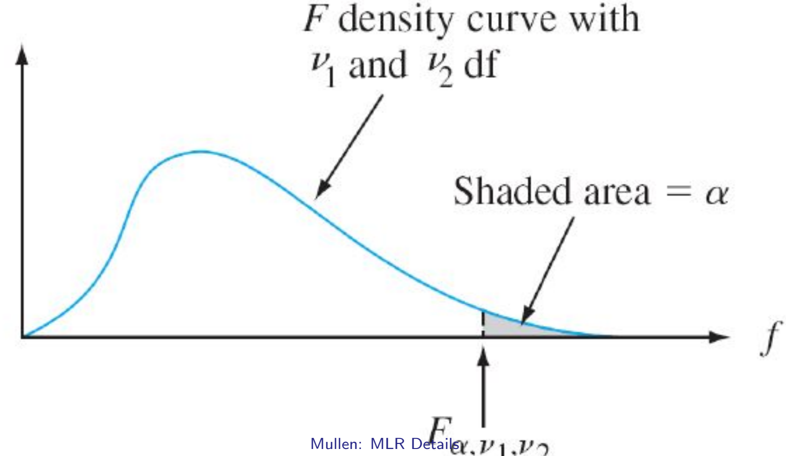
Does this data support the conclusion as applied to men? Use alpha $=$ .01.

$$F_{27,25} = \frac{52.6}{84.2} = F_{stat}$$

$$P(F_{27,25} \leq \frac{52.6}{84.2}) = \texttt{stats.f.cdf}(\frac{52.6}{84.2}, 27, 25) = .117 = p$$

Test for Equivalence of Variance

Recall: a typical F density function. When this thing took a value far from 1, we could conclude that the *ratio* being calculated had significantly different numerator from denominator. This is how we compared two variances.



The F-test

We use F statistics to compare variances. One way to compare linear models is to compare the variance in Y to the variance of your model: if your model is capturing a lot of the variance in Y, it's doing well!
In MLR we test the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

which says that there is no useful linear relationship between y and any of the p predictors. We test against:

$$H_a : \text{any of the } B_j's \text{ are nonzero.}$$

We could test each separately, but we would be commuting the multiple comparisons fallacy. A better test is a joint test, and is based on a statistic that has an F distribution when H_0 is true.

The Full F

Null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Alternative hypothesis:

$$H_a : \text{at least one } \beta_j \neq 0.$$

Test statistic value:

$$F = \frac{SSR/(p+1)}{SST/(n-p+1)}$$

Rejection region for a level test: $f \geq F_{\alpha,p+1,n-(p+1)}$

The Partial F

Comparing variances also gives us another way - besides just adjusted R^2 - to compare between models.

Idea: compare the amount of variance captured by the larger model to the smaller model. If they're significantly different, we know the larger model is "adding" lots of information!

As a hypothesis, this means testing that the parameters that are different between models are zero.

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:

Null:

Alternative:

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:
Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null:

Alternative:

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:
Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null: $H_0 : \beta_1 = \beta_3 = 0$

Alternative: Either/both of $\beta_1 \neq 0$ or $\beta_3 \neq 0$. Alternatively: the overall model captures significantly more variability in Y by including both β_1 and β_3.

The Partial F

Suppose we wanted to know whether *some* subset of model parameters were zero:
Full model: $\underline{Y} = \beta_0 + \beta_1 \underline{X_1} + \ldots \beta_4 \underline{X_4}$
Reduced Model: $\underline{Y} = \beta_0 + \beta_2 \underline{X_2} + \beta_4 \underline{X_4}$

Null: $H_0 : \beta_1 = \beta_3 = 0$

Alternative: Either/both of $\beta_1 \neq 0$ or $\beta_3 \neq 0$. Alternatively: the overall model captures significantly more variability in Y by including both β_1 and β_3.

In many situations, one first builds a model containing p predictors and then wishes to know whether any of the predictors in a particular subset provide useful information about Y.

The Partial F

The test is carried out by fitting both the full and reduced models.

Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model.

That is, if we let ________ be the sum of squared residuals for the full model and ________ be the corresponding sum for the reduced model, then ____________________

The Partial F

The test is carried out by fitting both the full and reduced models.

Because the full model contains not only the predictors of the reduced model but also some extra predictors, it should fit the data at least as well as the reduced model.

That is, if we let $\underline{SSE_{full}}$ be the sum of squared residuals for the full model and $\underline{SSE_{red}}$ be the corresponding sum for the reduced model, then $\underline{SSE_{full} < SSE_{red}}$

The Partial F

Intuitively, if ________ is a great deal smaller than ________, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction __________________ in unexplained variation.

Test statistic value:

Rejection region:

The Partial F

Intuitively, if $\underline{SSE_{full}}$ is a great deal smaller than $\underline{SSE_{red}}$, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction $\underline{SSE_{red} - SSE_{full}}$ in unexplained variation.

Test statistic value:

$$F = \frac{\left(SSE_{red} - SSE_{full}\right)/(p-k)}{SSE_{full}/(n-(p+1))}$$

Rejection region:

The Partial F

Intuitively, if $\underline{SSE_{full}}$ is a great deal smaller than $\underline{SSE_{red}}$, the full model provides a much better fit than the reduced model; the appropriate test statistic should then depend on the reduction $\underline{SSE_{red} - SSE_{full}}$ in unexplained variation.

Test statistic value:

$$F = \frac{\left(SSE_{red} - SSE_{full}\right)/(p-k)}{SSE_{full}/(n-(p+1))}$$

Rejection region: $f \geq F_{\alpha,p-k,n-(p+1)}$

Model Selection

So far, we have discussed a few of methods for finding the “best” model:

1. Comparison of adjusted R^2.
2. F-test for model utility and F-test for determining significance of a subset of predictors.

   There are other model selection techniques too:
3. Individual parameter t-tests.
4. Reduction of collinearity.
5. 'Best' transformations.
6. Forward/backward selection.

We will elaborate more on these and do some examples in the next day(s) of lecture.

Daily Recap

Today we learned

1. Multiple Linear Regression

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression! More predictor!