The plan:

Next: NCI(T) = Irish MLE
 interval

\[ R = \text{Nominal Normal Search} \]

\[ R - \text{Gun or Depression} \]

\[ NCI(T) \]

Example: Real Estate

\[ t = 2.0 \]

\[ y = 57.7 \]

\[ \bar{y} = 58.7 \]

\[ \text{Confidence limits for mean} ]

\[ \bar{y} = 58.7 \pm 1.58 \times \frac{\text{S.E.}}{\sqrt{n}} \]

\[ \text{S.E.} = \frac{\sqrt{\sum (y_i - \bar{y})^2}}{n - 1} \]

\[ n = 9 \]

\[ \text{Mean} \]

\[ \text{S.E.} = \frac{\sqrt{1972}}{8} \]

\[ = 4.7 \]

\[ \text{Confidence limits for mean} ]

\[ \bar{y} \pm t_{\alpha/2} \times \text{S.E.} \]

\[ t_{0.025, 8} = 2.352 \]

\[ 58.7 \pm 2.352 \times \frac{4.7}{\sqrt{8}} \]

\[ = 58.7 \pm 1.58 \times \frac{4.7}{\sqrt{8}} \]

\[ = 58.7 \pm 1.58 \times 1.17 \]

\[ = 58.7 \pm 1.89 \]

\[ = (56.8, 60.6) \]

\[ y = 57.7 \]

\[ t = 2.0 \]

\[ \text{S.E.} = \frac{\sqrt{1972}}{8} \]

\[ = 4.7 \]

\[ \text{Confidence limits for mean} ]

\[ \bar{y} \pm t_{\alpha/2} \times \text{S.E.} \]

\[ t_{0.025, 8} = 2.352 \]

\[ 58.7 \pm 2.352 \times \frac{4.7}{\sqrt{8}} \]

\[ = 58.7 \pm 2.352 \times 1.17 \]

\[ = 58.7 \pm 2.75 \]

\[ = (56.0, 61.4) \]
Prediction Limits for new y

\[ 4.4 \pm (4.303)(.5657) \]

**Testing in MLR**

- **t tests for single coefficients**
- test \( H_0: \beta_j = 0 \) using
  \[ T = \frac{b_j - \hat{\beta}_j}{SE_{\beta_j}} \] to get p-value
- most common version is version with \( \hat{\beta}_j = 0 \)

Both \( H_0: \beta_1 = 0 \) and \( H_0: \beta_2 = 0 \) are implausible, both \( x_1 \) and \( x_2 \) are helpful in predicting/explaining \( y \)

**Example: Real Estate**

- \( H_0: \beta_0 = 0 \)
- \( T = \frac{b_1 - 0}{SE_{b_1}} = \frac{1.87}{0.76} = 24.56 \)
- \( \text{p-value} < .0001 \)

- \( H_0: \beta_2 = 0 \)
- \( T = \frac{b_2 - 0}{SE_{b_2}} = \frac{1.28}{0.144} = 8.86 \)
- \( \text{p-value} < .0001 \)

**Testing \( H_0: \beta_1 = 0 \) in MLR**

- \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \)
- asks "if I have both \( x_1 \) and \( x_2 \) to predict with, can I do without \( x_1 \)?"
- Note that if \( x_1, x_2 \) are themselves highly correlated, the answer to both questions might be NO!
- The answer to the second is YES!

(see a later discussion of "multicollinearity")
There are also a variety of possible F tests in MLR

"Overall F-test" / Model Utility Test

This is a test of \( H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \)

based on an overall ANOVA table

Model: \[ y = x_1 \beta_1 + x_2 \beta_2 + x_3 \beta_3 + \ldots + x_k \beta_k + \epsilon \]

Hypothesis: \[ y = x_1 \beta_1 + x_2 \beta_2 + \ldots + x_k \beta_k \]

So this is often interpreted as a test of whether someplace in \( x_1, x_2, \ldots, x_k \) there is

predicting power

p-values will come from the F tables with df: \( k \) and \( n-k-1 \)

Example: Real Estate

ANOVA (for testing \( H_0: \beta_1 = \beta_2 = 0 \))

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>819.3</td>
<td>2</td>
<td>409.7</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>8.2</td>
<td>7</td>
<td>1.2</td>
<td>X</td>
</tr>
<tr>
<td>Total</td>
<td>827.5</td>
<td>9</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Tiny p-value (consulting F dist with
\( df = 2, 7 \))

This is a way of attaching a p-value to \( R^2 = \frac{SSR}{SS_T} \) for MLR

ANOVA Table (for MLR)

\[ F = \frac{MSR}{MSE} \]

Source \( SS \) \( df \) \( MS \) \( MSE \) \( F \)
Regression \( SSR \) \( k \) \( \frac{SSR}{k} \) \( \frac{SSR}{k} \)
Error \( SSE \) \( n-k-1 \) \( \frac{SSE}{n-k-1} \) \( \frac{SSE}{n-k-1} \)
Total \( SS_T \) \( n-1 \) \( \frac{SS_T}{n-1} \) \( \frac{SS_T}{n-1} \)

from MLR

big observed \( F \) count as evidence
against \( H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \)

Caution: In \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \)

testing \( \{ H_0: \beta_1 = 0 \} \) separately is NOT

the same thing as testing \( H_0: \beta_1 = \beta_2 = 0 \)

It's possible to get

tiny p-value for \( H_0: \beta_1 = \beta_2 = 0 \)

and simultaneously

large p-value for \( H_0: \beta_1 = 0 \)

large p-value for \( H_0: \beta_2 = 0 \)
This kind of thing can happen when predictors are correlated.

That brings up the fact that interpretation of MLR is subtle when predictors are correlated — jargon for this circumstance is "multicollinearity."

Example \( x_1, x_2, y \)

note \( x_2 \approx 2x_1 \)

\[ y \approx 1 + 3x_1 \]

\[ y \approx 1 + 1.5x_2 \]

---

SLR \( x_1 \)

\[ \hat{y} = 1.15 + 2.896x_1 \]

\[ R^2 = .9372 \]

\[ T = 33 \]

for testing \( H_0: \beta_1 = 0 \)

SLR \( x_2 \)

\[ \hat{y} = .923 + 1.508x_2 \]

\[ R^2 = .9573 \]

\[ T = 38.4 \]

for testing \( H_0: \beta_2 = 0 \)

MLR \( x_1, x_2 \)

\[ \hat{y} = 1.03 + 1.239x_1 + .864x_2 \]

\[ R^2 = .9983 \]

\[ t's = 1.26, 1.68 \]

for \( x_1, x_2 \) are highly correlated

I can't really separate the effects on \( y \) (and need to be cautious in interpreting \( \beta \)'s) — because \( x \)'s are highly correlated, while I can pick out a line in 3 dimensions where \( (x_1, x_2, y) \) fall (I can predict \( y \) from \( x_1 \) or \( x_2 \) or \( (x_1, x_2) \))

I can't really pick out a plane, i.e. I can't separate the effects of \( x_1, x_2 \).
Exercise: For fake data

1) Find T statistics for testing
   \( H_0: \beta_1 = 0 \) \( H_0: \beta_2 = 0 \) on printout

2) Take previous hand work and make up ANOVA table and overall F
   for testing \( H_0: \beta_1 = \beta_2 = 0 \)

   ANOVA Table
   \[
   \begin{array}{lcccc}
     \text{Source} & \text{SS} & \text{df} & \text{MS} & \text{F} \\
     \text{Regression} & 15.6 & 2 & 7.8 & 33.0 \\
     \text{Error} & 4.2 & 2 & 2.1 & X \\
     \text{Total} & 16 & 4 & & X \\
   \end{array}
   \]

This is a way of asking "is the reduced model adequate, or is there something in the full model that is needed in addition to what's in the reduced model?"

Notice

\[ \text{SSR}_{\text{full}} \geq \text{SSR}_{\text{reduced}} \]
\[ R^2_{\text{full}} \geq R^2_{\text{reduced}} \]

and a "partial F" test is a way of attaching a p-value to the increase in \( \text{SSR} \) (or in \( R^2 \)) usually this is organized in a expanded ANOVA Table

There are other F tests associated with multiple linear regression.
"Partial F Tests" - MND + S do something equivalent in terms of comparing \( R^2 \) values - (see their page 661)
These are a way of comparing

Full Model: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \)
Reduced Model: \( y = \beta_0 + \beta_l x_1 + \ldots + \beta_l x_l + \epsilon \)
For \( l < k \)

\[
\begin{array}{lcccc}
   \text{Source} & \text{SS} & \text{df} & \text{MS} & \text{F} \\
   \text{Regression} & \text{SSR}_{\text{full}} & & & \\
   & \text{SSR}_{\text{red}} & & & \\
   & x_{l+1}, \ldots, x_n; x_1, \ldots, x_l & & & \\
   & x_{l+1}, \ldots, x_n & \text{SSR}_{\text{full}} - \text{SSR}_{\text{red}} (n-k-1) & & \\
   \text{Error} & \text{SSE}_{\text{full}} & n-k-1 & \text{MSSE}_{\text{full}} & \\
   \text{Total} & \text{SST} & n-1 & & \\
\end{array}
\]

E.G. \( k = 5 \), \( l = 2 \)

Full Model: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon \)
Reduced Model: \( y = \beta_0 + \beta_l x_1 + \beta_2 x_2 + \epsilon \)
Tests \( H_0: \beta_3 = \beta_4 = \beta_5 = 0 \)

60
Example: Real Estate

\[ R^2_{\text{full}} = 0.89 \quad R^2_{\text{SLR}} = 0.88 \]

I might want to ask whether this increase in \( R^2 \) is "statistically significant."

We've already used a t-test of \( H_0: \beta_2 = 0 \) but I could also use a partial F-test (when the full model has one more \( x \) than the reduced model, the partial F and t are equivalent).

Note that here (since the full model has 1 more predictor variable than the reduced one) the F-statistic is the square of the t-statistic for the coefficient potentially being dropped (\( \beta_2 \))

\[ 78.0 = (8.85)^2 \]

Exercise: Make the expanded ANOVA table for \( H_0: \beta_1 = 0 \) in Real Estate example

(SLR on "condition" gives SSR = 115)
"Model Building"
1. Making "new" predictors (and responses)
   - "interaction"
   - incorporating qualitative info
2. Searching
   - "algorithm (s)"
   - criteria
3. Model "checking" / Diagnosis
   - plots
   - statistics / measures

I can also take
\[ x_1 \rightarrow x' \]
\[ x_2 \]

\[ y' \rightarrow y \] (like in MLR homework)
model \( y' \), do inferences and then
\( y' \rightarrow y \) and state results in the original units — This possibility makes MLR
more flexible and widely applicable

\( z \rightarrow z' \) gives me things like
MLR homework, and gives me
Things like fitting
\[ y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \]

Example
\[ x_1 \succ x_2, y \]
\[ k=2 \]
\[ x_1 \rightarrow x' \]

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \]

average \( y \)

\[ x_2 = 4 \]

\[ x_2 = 3 \]

\[ x \]

\[ x_1 \]

No interaction effects
Example Real Estate

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \]

An alternative model would be:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \]

This leads to a plot "with interactions" / non-parallel faces:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 (1) + \beta_3 x_1 (1) \]
\[ = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 \]

How to incorporate "qualitative" information in a MLR model?

Use "dummy variables" / "indicator variables".

Hypothetical Real Estate Scenario

<table>
<thead>
<tr>
<th>( y = \text{price} )</th>
<th>#2</th>
<th>#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 = \text{size} )</td>
<td>NW</td>
<td>NE</td>
</tr>
<tr>
<td>( x_2 = \text{condition} )</td>
<td>SW</td>
<td>SE</td>
</tr>
</tbody>
</table>

I'd like to build "region" information into a MLR model.

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \]

But that is silly... it says that I get \( \beta_3 \) increase in price.

Region \#1 \( \rightarrow \) region \#2
Region \#2 \( \rightarrow \) region \#3
Region \#3 \( \rightarrow \) region \#4

I need to be more clever.
"More Clever" is using "Dummy Variables". If a qualitative factor $A$ has $I$ levels 1, 2, ..., $I$ ($A=$ region $I=4$) I define $I-1$ dummies:

$X_{A1} = \begin{cases} 1 & \text{if observation is from level 1 of } A \\ 0 & \text{otherwise} \end{cases}$

$X_{A2} = \begin{cases} 1 & \text{if observation is from level 2 of } A \\ 0 & \text{otherwise} \end{cases}$

$\vdots$

$X_{AI}$

A MLR involving $X_{A1}, X_{A2}, \ldots, X_{AI}$ will allow for a shift for each different level of $A$.

Example: Hypothetical Real Estate

$I=4$ quadrants define $I-1=3$ dummies

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_{A1}$</th>
<th>$X_{A2}$</th>
<th>$X_{A3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NE homes</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NW homes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SW homes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SE homes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

What does this do for me?

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{A1} + \beta_4 x_{A2} + \beta_5 x_{A3} + \epsilon$

Says that

NE homes are modeled

$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

NW homes

$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

SE homes

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

SW homes
SW homes
\[ y = (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \epsilon \]

SE homes
\[ y = \beta_0 + \beta_5 + \beta_3 x_3 + \epsilon \]

SE is a baseline and coefficients on dummy variables account for level I shifts up or down in price as we look at other quadrants.

Note as we've got it thus far, price/yr. is the same in each region, if I want it.

Model Searching (Mucking around in the possibilities looking for something that's plausible)

Qualitatively: one wants a model that reproduces y's faithfully and is small/simple for simple understanding and avoiding "overfitting" (where because it matches a lot of an equation describes data so well, but does a bad job for interpolating or extrapolating)

Change region to region, I could:

- Make up new predictors
  \[ x_1 x_3 \]
  \[ x_1 x_4 \]
  \[ x_1 x_3 \]

This allows price per square foot to change -- this would be called "interaction" between region and size.

JMP does something like this automatically (coding 1, 0, -1 rather than 1, 0) for "nominal" variables

Criteria (numerical) for comparing models?

- \( R^2 \) (generally wins big \( R^2 \))
- \( S \) (generally we like small \( S \))
- There are cases where \( R_i^2 > R_2^2 \), but \( S_2 < S_1 \)
- There are other possibilities
  - Mallows Cp
  - If I have 1k predictors possible

square foot would be called region and size like this automatically rather than 1, 0 variables

for comparing \( R^2 \), \( S \) (like small \( S \)) possibilities

predictors possible
\[ \ell_P = \frac{\text{SSE}_{\text{red}}}{\text{MSE}_{\text{full}}} + 2p - N \]

for a reduced model with \( p-1 \) predictors, this should be about \( p \)
(skip the rationale for why this is a happy circumstance)

\[ \text{AIC} = n \ln\left(\frac{\text{SSE}}{n}\right) + 2(p + 1) \]
(for a model with \( p \) predictors)
small AIC are desirable

JMP will do this under "Stepwise" "personality" for "Fit Model"
this gives a way to screen a huge # of possible models down a few to examine carefully (to see whether the normal MLR model is appropriate
we care because all of our inference formulas hang on the appropriateness of that model)

Model Checking / Diagnostics

**Basic Tool here is the idea of residuals**

"Search Algorithms"

Ancient History:
Forward Selection
Backward Elimination
These produce lists of candidate sets of predictors... at any given step they are not guaranteed to give best \( R^2 \) for a model of that size
Obsolete because there is an "all possible Rs" algorithm -

Recall
\[ e_i = y_i - \hat{y}_i = \text{residual for } i \text{th case} \]
These are empirical approximations for
\[ y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = e_i \]
and the model says these are random draws from a normal dsn with mean 0 and std dev \( \sigma \) —
so one would hope that residuals look like "noise" i.e. random variation
i.e., I expect/hope (if MLR model is sensible) that $e_i$:
1) have a bell-shaped histogram
2) are "patternless" when plotted against any sensible variable

This suggests I've missed something in modeling

I can plot $e_i$ vs $x_{1i}, x_{2i}, \ldots, x_{ki}, y_i$

This suggests that the "constant $\sigma$" part of the model is no good
Plotting against "other" variables not in a model

Example: Real Estate
\[ \hat{y} = \beta_0 + \beta_1 \text{size} + \beta_2 \text{condition} \]

\[ e_i = y_i - \hat{y}_i \quad \text{ordinary residuals} \]
\[ e_i^* = \frac{e_i}{\text{SE} e_i} \quad \text{standardized residuals} \]

East = side of town
West

suggests that "side of town" should be included in modeling

Deleted Residuals

\[ \hat{y}(i) = \text{value predicted by model for case } i \text{ when model is fit to only the other } n-1 \text{ cases} \]

\[ e(i) = y_i - \hat{y}(i) \quad \text{ith deleted residual} \]

and one hopes that \( e(i) \) are not much bigger than \( \text{SE} e_i \)

A way of measuring this is using

\[ \text{PRESS} = \sum e(i)^2 = \sum (y_i - \hat{y}(i))^2 \]

Recall \( \text{SSE} = \sum e_i^2 \)

\[ \text{PRESS} \geq \text{SSE} \]

(but we hope it is not too much bigger)

There are also "partial residuals" (that are origin of the JMP "leverage plots") put this on hold for a few minutes
Another idea (for identifying important cases in regression is the idea of "hats" (leverage values in non-SAS world) (JMP calls something else the leverage values)

**Fact:** In a given MLR there are n x n constants h_{ij} so that if \( y_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in} y_n \)

That value (h_{ii}) in some sense measures how important \( y_i \) is in predicting \( y_i \).

h_{ii} flag data points with \( z = 0 \) that put them on the "edge" of the data set as regards the predictors -- we might want a measure that also takes into account a data point's \( y_i \) -- "Cook's Distance"

\[
D_i = \frac{h_{ii}}{(k+1) \text{MSE}} \left( \frac{e_i}{1 - h_{ii}} \right)^2
\]

\[
\left( \frac{h_{ii}}{k+1} \right) \left( \frac{e_i}{s} \right)^2
\]

\( 0 < h_{ii} < 1 \)

\( \sum h_{ii} = k+1 \)

This means that \( h_{ii}'s \) average to \( \frac{k+1}{n} \)

A common rule of thumb is that

\( h_{ii} > 2 \frac{k+1}{n} \)

Flags a "high leverage" case, i.e. one that is influential in fitting potentially

To have a big \( D_i \) a data point must have both a big \( h_{ii} \) value (be near the "edge" of data set) and have a big deleted residual (be poorly predicted by a model fit without using it)