This is a test of achieving a p-value from consideration of other predictions. It may drop predictor(s) from consideration... It's a way of asking whether each predictor is adequate as there are also a variety of F-tests associated with M.R. Overall = [test for Model Utility Test]

In M.R., testing $H_0: \beta = 0$ is a

The null hypothesis says $H_0: \beta_1 = \beta_2 = \ldots = 0$.

This hypothesis says that none of the predictors is related to the response variable. This test is often interpreted as a test of whether the model information from predictor(s) is explanatory.
P value very small

Caution: In

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

testing

\[ H_0 : \beta_1 = 0 \]
\[ H_0 : \beta_2 = 0 \]
separately

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 the ANOVA table and overall F for

\[ H_0 : \beta_1 = \beta_2 = 0 \]

ANOVA Table

<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>15.6</td>
<td>2</td>
<td>7.8</td>
<td>35.0</td>
</tr>
<tr>
<td>Error</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>X</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>4</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

This often happens when the predictors \( x_1, x_2 \) are highly correlated — I need only one to predict \( y \).

Thus are often F tests associated with MCR ... "Partial F tests" — (Section 4.4 of Diehlan)

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: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_l x_l + \epsilon \)

for \( l < k \)

This a way of asking "Is the reduced model adequate... or do I need the full model?"

\[ 
\begin{align*}
F & \text{ den with } 2,17 \text{ df:} \\
\text{P-value} & \text{very small}
\end{align*}
\]
Notice

\[ \text{SSR}_{\text{full}} \geq \text{SSR}_{\text{reduced}} \]

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

This "partial F" test is a way of attributing a p-value to the increase from \( R^2_{\text{reduced}} \) to \( R^2_{\text{full}} \)

calculations are based on an expanded ANOVA table

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

Full Model: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e \)

(One) Reduced Model: \( y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + e \)

I might test \( H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \)

in the full model - that's what the previous (expanded) table does for me

Example: Real Estate

\( R^2_{\text{null}} = .39 \quad R^2_{\text{null}} = .88 \)

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

- I've already used a t-test of \( H_0: \beta_3 = 0 \) but I could also do a partial F-test
(Expanded) ANOVA Table (For $H_0: \beta_2 = 0$)

<table>
<thead>
<tr>
<th>Source</th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression ($x_1, x_2$)</td>
<td>819.3</td>
<td>2</td>
<td>409.7</td>
</tr>
<tr>
<td>$x_1$</td>
<td>727.5</td>
<td>1</td>
<td>727.5</td>
</tr>
<tr>
<td>$x_2</td>
<td>x_1$</td>
<td>91.4</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>8.2</td>
<td>12</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>827.5</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

$\text{Variance}$ $= \frac{78.0}{15}$

Exercise: Make the expanded ANOVA table for testing $H_0: \beta = 0$ in the Real Estate problem ($\text{SLR on condition gives } SS_R = 115$)

<table>
<thead>
<tr>
<th>Source</th>
<th>$SS$</th>
<th>$df$</th>
<th>$MS$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression ($x_1, x_2$)</td>
<td>819.3</td>
<td>2</td>
<td>409.7</td>
</tr>
<tr>
<td>$x_2$</td>
<td>704.3</td>
<td>1</td>
<td>704.3</td>
</tr>
<tr>
<td>$x_1</td>
<td>x_2$</td>
<td>115</td>
<td>1</td>
</tr>
<tr>
<td>Error</td>
<td>8.2</td>
<td>12</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>827.5</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

$\text{Variance}$ $= \frac{60.3}{15}$

Section 6.3 of "Dielman: Beware!"

"Multicollinearity" (and difficulty of interpreting coefficients in MCR)

In practice, usually the predictors have correlations between them, so when that's the case interpreting individual $\beta$'s (b's) is dangerous as is trying to identify "most important" predictors.

Example: $x_1, x_2, y$ (handout)

**Note** $x_2 < 2x_1, y \approx 1 + 3x_1$

$R^2 = .987$

$\hat{y} = .55 + 1.65x_1 + 1.58x_2$

$\hat{y} = .923 + 1.508x_2$

$R^2 = .9572$

$t = 38.4$

for testing $H_0: \beta_1 = 0$

for testing $H_0: \beta_2 = 0$

MLR $x_1, x_2$

$R^2 = .958$

$t's: 1.26, 1.68$
since $x_1, x_2$ are so highly correlated, I can't really separate their effects on $y$... I need to be cautious in interpreting $\beta$'s (b's)

because $z$'s are highly correlated, I can't really pick out a plane (and thus I can't really separate the effects of $x_1$ and $x_2$)

this is in spite of the fact that I have a huge $R^2$ and can predict $y$ nearly perfectly....

Inference for Arbitrary Linear Combinations of $\beta$'s

We know how to infer $\beta_2$

$$\hat{y}_{ij|1,...,x_k} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

This is a special instance of

$$L = c_0 \hat{\beta}_0 + c_1 \hat{\beta}_1 + c_2 \hat{\beta}_2 + \cdots + c_k \hat{\beta}_k$$

where

$$c_0 = 1, c_1 = -x_1, c_2 = -x_2, \ldots, c_k = -x_k$$

It's no harder in theory to do inference for $L$ than for $y_{ij|1,...,x_k}$

In practice JMP enables it... and really we do care 😊

Example: Real Estate - houses of 2 types

- Type 1: $x_1 = 24, x_2 = 3$
- Type 2: $x_1 = 20, x_2 = 6$

It could be of interest to compare mean selling prices - e.g. I might wish to estimate

$$\mu_{y|x_1=24,x_2=3} - \mu_{y|x_1=20,x_2=6}$$
This difference is

\[ (\beta_0 + \beta_1(24) + \beta_2(3)) - (\beta_0 + \beta_1(20) + \beta_2(6)) \]

\[ = 4 \beta_1 - 3 \beta_2 = L \]

where \( c_0 = 0, c_1 = 4, c_2 = -3 \)

"Happily" I do inference for such \( L \)'s

\[ \hat{L} = c_0 \bar{b}_0 + c_1 \bar{b}_1 + \ldots + c_k \bar{b}_k \]

and confidence limits for \( L \) are

\[ \hat{L} \pm t (\text{std error for } \hat{L}) \]

There are 2 related prediction problems that we'll discuss

1) predicting the difference in 2 future \( y \)'s (e.g., the difference in actual prices at a home of type 1 and a home of type 2)

2) predicting the sum of several (\( m \)) future \( y \)'s (e.g., total revenue on actual sales of a home of type 1 and a home of type 2)

Example: Real Estate

\[ \hat{L} = 4 \bar{b}_1 - 3 \bar{b}_2 = 3.65 \]

\[ \text{std error for } \hat{L} = 2.54 \]

\[ \hat{L} \pm t (\text{std error for } \hat{L}) \]

99% confidence limits for \( L \) are

\[ 3.65 \pm 2.54 (\text{std error for } \hat{L}) \]

\[ 3.65 \pm 2.54 \]

Be upper 2.5% 5° with \( \frac{d}{a} = \frac{4}{5} \)

Prediction Limit's Turn out to be:

For a difference in 2 \( y \)'s

\[ \hat{L} \pm t \sqrt{2 s^2 + (\text{std error } \hat{L})^2} \]

For a sum of several future \( y \)'s

\[ \hat{L} \pm t \sqrt{m s^2 + (\text{std error } \hat{L})^2} \]

Example: Real Estate

I own both a Type 1 and a Type 2 home - sell them - predict both a price difference and a total
For the difference: 95% prediction limits are given as
\[ \hat{\epsilon} = 3.65 \]
\[ 3.65 \pm 2.365 \sqrt{\left(2 \cdot 1.081^2\right) + 0.54^2} \]
3.65 \pm 2.365

For sum:
\[ L = (\beta_0 + 2\beta_1 + 3\beta_2) + (\beta_0 + 2\beta_1 + 6\beta_2) \]
\[ = 2\beta_0 + 4\beta_1 + 9\beta_2 \]

End of Inference Formulas... What's Left?
- Time Series Ideas/Applications of MLR
- "Diagnose" and Model Building and Selection Ideas
  - Simple Time Series Ideas (Dielman integrates into his SLR + MLR chapter)
    - \[ y_1 = \text{response at time 1} \]
    - \[ y_2 = \text{response at time 2} \]

Main Ideas:
1. I might use time as predictor variable (fitting a trend)
   \[ x = t \]
   \[ y_t = \beta_0 + \beta_1 t + \epsilon_t \]
   Dielman 69a = 10

2. I might use "past" or "lagged" \( y \)'s as predictor variables i.e. I might consider fitting equations like
   \[ y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t \]
   \[ y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t \]
Chapter 5 at Diehlman

you can often build new predictor variables from old existing ones - an example is "polynomial regression" i.e. \( y = x \)

I can fit not only \( y = \beta_0 + \beta_1 x + \epsilon \)

I can also fit \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \)

BTW: This particular thing is automated in JMP. This particular thing is automated

Isn't it? How to examine various possibilities looking for one to use?... this pushes us to ch6 of Diehlman where he begins "Model Diagnostics"

Major Tool here is "residual analysis"

residual \( e_i = y_i - \hat{y}_i \)

\( = y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k) \)

is an empirical version of \( e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \ldots + \hat{\beta}_k x_i^k) \)

...and nothing stops us from starting with \( x_1 x_2 \)

and making up \( x_{1/2}, x_{1/4}, \ldots \)

\( \sqrt{y}, \frac{x}{y}, \log y, \ldots \) possibilities explode exponentially ...

The MLE model says the \( \epsilon \)'s are independent normal mean 0 std dev \( \sigma \) random variables

Since the \( \epsilon \)'s are approximations for \( e \)'s they ought to look roughly as if they were normal, mean 0 R.V.'s

So if plot them against any thing I can think of, the plot should be bell-shaped... if it is not I've problems with the model and seek another ...

a histogram of them should bell-shaped etc.
My Plan:

- Tentatively read Ch 6, Ch 7
- Test Questions, Ch 8, + ??
- Thrus. Exam, maybe 0-1 y's

Ch 6: Regression Diagnostics

- Hope to detect
  1) problems with a candidate model
  2) a few cases that are "unusual"
  and thereby influence model fit

Basic tool here is "residual analysis"

MLR Model: $e_i = \epsilon_i$ random draws from a normal universe

$$\hat{e}_i = y_i - \hat{y}_i \approx \epsilon_i$$
and Thus ought to look like "random draws from a normal dsm of $\epsilon$'s"

To judge/investigate this I can

1) make a histogram: when data sets are large this should look bell-shaped
e.g. with $(x,y)$

2) plot $e_i$'s against various other variables hoping to see only only "random scatter" - if instead I see a pattern that suggests that I've missed something in modeling

common plots of residuals are:

- $e_i$ vs $x_i$
- $e_i$ vs $z_i$
- $e_i$ vs $\hat{y}_i$
- $e_i$ vs time, etc.
response variability appears to increase with mean response ... that's a violation of MLE assumptions ... a standard trick in this kind of context is to replace y
with ỹ or log(y) and model this new response variable (as in the above).

etc.

There are a variety of embellishments on this basic idea —

Standardized/Studentized Residuals

\[ e_i = \frac{\hat{e}_i}{SE(e_i)} = \frac{y_i - \hat{y}_i}{\text{standard deviation}_y - \hat{y}_i} \]

This says I've missed the important "location" that I somehow need to include ... ??.

See Ch7 on Dummy Variables.

We expect these standardized residuals to typically be no more than 2 in absolute value ... when they are, the corresponding case gets flagged as being poorly fit by the model.

"Deleted Residuals" (and the "PRESS" statistic)

\[ \hat{y}_{(i)} = \text{value predicted for case } i \text{ when the model is fit not using case } i \]

\[ e_{(i)} = y_i - \hat{y}_{(i)} \] the "deleted residual" for case i.
I hope that the \( e_i \) aren't too much larger (aren't too different form) the \( e_i \) - a means of summarizing/assessing this is

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

"prediction sum of squares"

Recall that \( \text{SSE} = \sum e_i^2 = \sum (y_i - \bar{y})^2 \)

The fact is that \( \text{PRESS} \geq \text{SSE} \)

But we hope \( \text{PRESS} \) isn't too much larger than \( \text{SSE} \)

JMP plots these partial residuals for \( y \) (the response) vs partial residuals for the \( x_j \)'s themselves

\[
e_i(y_j) = \hat{y}_i - \text{prediction of } y_i \text{ from } x_j \text{ and all other } (k-1) \text{ other } x_i's
\]

This is what's unexplained in \( y_j \) using all other predictors

There are also studentized versions of these

Partial Residuals idea here is to drop a variable (a column from the data table) and predict using all others - e.g. if I drop variable \( j \) \( (x_j) \) I might regress \( y \) on all others and compute

\[
\hat{y}_i - y_i
\]

which is unexplained

\[
\hat{y}_i = \text{a prediction for case } i \text{ not using variable } x_j \text{ in fitting all else}
\]

\[
e_i(y_j) = e_i(y) \text{ (partial residual, residual for a reduced model not including } x_j)
\]

JMP leverage plots are plots of

\[
\hat{y} + \varepsilon_i(y) \text{ vs } \hat{y} + e_i(x_j)
\]

Example: Real Estate

\[
\hat{y} + (y - \hat{y}) \text{ from SSE of } y \text{ on } x_2
\]

\[
\hat{y} + (y - \hat{y}) \text{ from SSE of } y \text{ on } x_2
\]

\[
\hat{y} + (y - \hat{y}) \text{ from SSE of } y \text{ on } x_2
\]

For \( x_2 \)

For \( x_2 \)
The more left-to-right spread we see on our leverage plots, the more there is in $x_j$ that can’t be predicted from the other $x$’s.

I hope to see cases pretty much spread over the whole range (the range of the original $x_j$’s) — (If not I’d perceive multicollinearity.)

The more top-to-bottom spread on these plots, the more there is in $y$ that is unexplained by a reduced model that doesn’t include $x_j$.

JMP usage of “Leverage” is not standard terminology.

“Leverage values” $h_{ij}$ (“hats”)

In a given ULR model there are $nxn$ constants $h_{ij}$ so that

$$ y_i = h_{i1} y_1 + h_{i2} y_2 + \cdots + h_{in} y_n $$

That in some sense measures how important $y_i$ is in producing $\hat{y}_i$. This is true because

1. $0 < h_{ij} < 1$ and
2. $\sum h_{ij} = k_i$.

so $h_{ij}$’s are positive and average to $k_i$ as a rule of thumb people look for $h_{ij} > \frac{k_i}{n}$ as indicating a “high leverage” case i.e. one that is “influential” in determining a fitted equation.

Example: Real Estate — The hats draw my attention to cases 7 and 10 as influential...they’re “on the edge” of the $(x_1, x_2)$ data set.
The leverage values \( h_{ii} \) only depend on \( x \)'s — it might make sense to have a measure that also depends on \( y \).

\[
\begin{align*}
\text{Cook's Distance} & \quad \text{(a measure of "influence" of cases that depends not only on \( x \)'s but also on \( y \))} \\
D_i = \left( \frac{h_{ii}}{k_{ri}} \right) \left( \frac{e_i}{s_e} \right)^2
\end{align*}
\]

Remember that the \( h_{ii} \) average to \( \frac{p}{n} \) so \( \frac{h_{ii}}{k_{ri}} \) will be \( \frac{1}{n} \) on average — \( e_i \) should be "not more than 2 or 3 \( \frac{s_e}{2} \) in absolute value.

**Hypothetical Example: Real Estate**

\( x_1 = \text{size} \)
\( x_2 = \text{condition} \)
\( x_3 = \text{quadrant} \)
\( y = \text{price} \)

It would be stupid to enter \( x_3 \) as if it were "another predictor".

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

---

Dielman 7.1 Dummy/Indicator Variables (as predictors)

Motivation is to include qualitative information in MLR.
What can I do that's not silly?

I can use "dummy" or "indicator" variables - if a qualitative factor A has I different levels (or values), we will cook up I-1 dummies:

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

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

... 

\[ x_{A_{I-1}} = \begin{cases} 1 & \text{if observation is from level } I-1 \text{ of } A \\ 0 & \text{otherwise} \end{cases} \]

A MLR model including \( x_{A1}, x_{A2}, ..., x_{A_{I-1}} \) will allow different means for each level of A.

**Example: Hypothetical Real Estate**

\( I = 4 \) sections of town

define \( I-1 = 3 \) dummies

<table>
<thead>
<tr>
<th>( y )</th>
<th>( x_{A1} )</th>
<th>( x_{A2} )</th>
<th>( x_{A3} )</th>
<th>( x_{A4} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \ldots )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>

What does this produce? (for simplicity, let's ignore size + condition)

\[ y = \beta_0 + \beta_1 x_{A1} + \beta_2 x_{A2} + \beta_3 x_{A3} + \varepsilon \]

Says that:

- Mean price for NE homes is \( \beta_0 + \beta_1 \)
- Mean price for NW homes is \( \beta_1 + \beta_2 \)
- Mean price for SW homes is \( \beta_2 + \beta_3 \)
- Mean price for SE homes is \( \beta_3 \)
i.e. β's for levels 1 through 1 of factor A become increments in mean response over level I mean response

Example: If I add back in size

\[ y = \beta_0 + \beta_1 x_{A1} + \beta_2 x_{A2} + \beta_3 x_{A3} + \epsilon \]

To use this idea in JMP in exactly this form you'd have to create your own 0's and 1's. What JMP gives you automatically is close to this but not quite this. Instead JMP does this:

\[ x_{A1} = \begin{cases} 1 & \text{case is from level 1} \\ 0 & \text{otherwise} \end{cases} \]  
\[ x_{A2} = \begin{cases} -1 & \text{case is from level 2} \\ \text{otherwise} \end{cases} \]

Example: Hypothetical Real Estate (JMP)

<table>
<thead>
<tr>
<th>( )</th>
<th>( x_{A1} )</th>
<th>( x_{A2} )</th>
<th>( x_{A3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( )</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( )</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( )</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( )</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

What does this do?

\[ y = \beta_0 + \beta_1 x_{A1} + \beta_2 x_{A2} + \beta_3 x_{A3} + \epsilon \]

mean of \( \epsilon \) for the regions:
- mean for NE: \( \beta_0 + \beta_1 \)
- mean for NW: \( \beta_0 + \beta_2 \)
- mean for SW: \( \beta_0 + \beta_3 \)
- mean for SE: \( \beta_0 - \beta_1 - \beta_2 - \beta_3 \)

\( \beta_0 \) is the average response (averaged over the 4 regions) minus \( \beta \) increase for region 1, \( \beta_1 \) increase for region 2, \( \beta_2 \) increase for region 3, \( \beta_3 \) increase for region 4.