A Few Scraps Regarding Ch 2

- computation of \( r \)
  1) note: \( r = \frac{1}{n} \sqrt{r^2} \)
  2) or use JMP "multivariate" function
  3) or use a calculator
  - Cautions / Caveats about Interpretation

1) The least squares line and \( r \) (or \( R^2 \)) are highly sensitive to a few "extreme" data points.

2) \( r, R^2 \) measure only linear association.

3) There is a perfect x-y relationship here, but it's not a straight line relationship.

4) correlation in not necessarily causation.

On to Ch 10 and Inference in SLR.
To support this, a model is needed. The most convenient and commonly used such model is the Normal Simple Linear Regression Model.

In words: The relationship between $x$ and average $y$ is linear ($My|x = \beta_0 + \beta_1x$) and for a given $x$, $y$ is normal with standard deviation $\sigma$.

BTW, this really is not much different from what we did in Ch6 or Ch7 except that now we let the mean response change with $x$ (the predictor/explanatory variable).

1st issue if I'm going to use such a model is simple number estimates of model parameters ($\beta_0, \beta_1$). Unknown values in model:

This means no variable $y$ is for fixed $x$.

Positive

Picture:

$My|x = \beta_0 + \beta_1x$

In Symbols:

$y = (\beta_0 + \beta_1x) + \epsilon$

accounts for deviation of $y$ above or below the line.

takes care of setting $y$ change with $x$.

To estimate model parameters $\beta_0, \beta_1$, we’ll be and $b_1$, least squares intercept and slope from Ch2.

For $\sigma$ (a measure of variability in $y$ for a given $x$) a measure of how much $y$ misses the line $My|x = \beta_0 + \beta_1x$... the standard deviation of $\epsilon$ in model equation.

Well, do the following.
observed values from least squares line

\[ \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = S_{\text{SLR}} \]

JMP
root mean square error

Example Real Estate - For the SLR model I'll estimate \( \beta_0 \) with \( b_0 = 16.008 \), \( \beta_1 \), with \( 1.9001 = b_1 \), and \( \sigma \) with \( s = \text{RMSE} = 3.53 \)

Confidence limits for \( \sigma \)
These can be based on a \( z^2 \) distribution with \( df = n-2 \)

\[ (s \sqrt{\frac{n-2}{u}}, s \sqrt{\frac{n-2}{L}}) \]

Example Real Estate 95% confidence limits for \( \sigma \)

\[ (3.53 \sqrt{\frac{10-2}{17.535}}, 3.53 \sqrt{\frac{10-2}{2.180}}) \]

2.38 6.76

interpretation of the last of these is that I'm estimating that the std dev of price for any fixed home size is about 3.53 (\$1000)

BTW

\[ S_{\text{SLR}} = \sqrt{\frac{\text{SSE}}{n-2}} \]

confidence limits? (I want to quantify how precise \( b_0, b_1, \sigma \) are as estimates of \( \beta_0, \beta_1, \sigma \))

Confidence limits for \( \beta_1 \)
(\( \beta_1 \) is the rate of change of mean \( y \) wrt \( x \), ... change in mean \( y \) that accompanies a unit change in \( x \))

\[ b_1 \pm t \cdot \text{SE}_{b_1} \]

slope of l.s. line : n-2 = df

\[ s \sqrt{\frac{\sum (x_i - \text{mean})^2}{n-1}} \]

\[ n \cdot \frac{S_{\text{SLR}}^2}{(n-1) \cdot \sigma^2} \]
Example Real Estate

95% confidence limits for the increase in average price that accompanies a unit increase in size ($\beta_1$)

$$b_1 = \pm \frac{s}{\sqrt{\sum (x-x)^2}} SE_{b_1}$$

$$1.5001 \pm 2.306 \sqrt{\frac{3.53}{20.6}}$$

SE$_{b_1}$ = 2.486

Confidence limits for $\beta_0$ (average $y$ for a particular $x$)

(e.g. This could be average price per hour of size 11 (10 ft$^2$))

$$d.f. = n-2$$

Value at which I'm estimating from the $n$ data points in hand

Exercise For "by hand" problem

1) Make 95% confidence limits for $\sigma$

2) Make 95% confidence limits for $\beta_1$

$$S = \sqrt{\frac{\sum (y - \bar{y})^2}{n-2}} = \sqrt{1.60} = .7303$$

Limits for $\sigma$:

$$(.7303 \sqrt{\frac{5-2}{2.348}}, .7303 \sqrt{\frac{5-2}{2.166}})$$

Limits for $\beta_1$:

$$-1.2 \pm 3.162 \cdot .7303$$

Prediction limits for $y_{new}$ at $x$

$$\hat{y} \pm t \cdot SE_{\hat{y}}$$

$$b_0 + b_1 x \pm S \sqrt{\frac{1}{n} + \frac{(x-x)^2}{2 \sum (x-x)^2}}$$

$$d.f. = n-2$$

$$(SE_{\hat{y}})^2$$

$$(\frac{1}{n})$$

45
Example Real Estate Problem

\[ x = 20 \quad \text{(2000 sq. ft. home)} \]

- 95% CI for average price
- 95% PI for next price

\[ \bar{y} = 16.0081 + 1.5001(20) = 54.010 \]

CI for an average price (for houses of this size)

\[ 54.010 \pm 2.306(3.53) \sqrt{\frac{1}{10} + \frac{(20-18.8)^2}{201.6}} \]

upper 2.5% of t₈

Exercise: \[ \bar{y} = 2.0 -1.2(1) = .8 \]

Confidence limits:

\[ .8 \pm 3.182 (.73) \sqrt{\frac{1}{5} + \frac{(1-0)^2}{10}} \]

.8 ± 1.27 ≤

Prediction limits:

\[ .8 \pm 3.182 (.73) \sqrt{1 + \frac{1}{5} + \frac{(1-0)^2}{10}} \]

.8 ± 2.65

JMP CE+PI's read off graphs!!

\[ 54.010 \pm 2.664 \]

PI for an additional price if \( x = 20 \)

\[ 54.010 \pm 2.306(3.53) \sqrt{1 + \frac{1}{16} + \frac{(20-18.8)^2}{201.6}} \]

54.010 ± 8.565

Exercise: For "by hand" data at \( x = 1 \)

make 95% confidence limits for many 95% prediction limits for your new

Note that the CI's for \( y_1 | x \) and PI's for \( y_{new} \) are narrowest at \( x = \bar{x} \) ... That makes sense

Also (a comment I should have made earlier about estimating \( \beta_1 \)) because

\[ SE_{\hat{b}_1} = \frac{s}{\sqrt{2(\bar{x}_i - \bar{x})^2}} \]

The more spread out the data pairs are horizontally, the more precision I have determining the slope.
2. Study a value is 0.138 (rounded up)

\[ Y_{10} \text{ SE}_b = \frac{1.38}{1.2} = -1.2 \]

\[ t = \frac{b}{SE_b} = \frac{b}{1.2} = -1.2 \]

H_0: \beta = 0 vs. H_a: \beta \neq 0

Exercise: For two linear data set this

\[ \text{ changes with } \]

\[ \text{ the mean } \]

\[ \text{ which is to be able to see.} \]

\[ \text{ Their } \]

\[ \text{ we have evidence enough} \]

\[ \text{ to } \]

\[ \text{ Ho: } \beta = 0 \]

\[ \text{ P-value: } \text{Ha: } \beta \neq 0 \]

\[ \text{ i.e. } \]

\[ \text{ Calculating the } \]

\[ \text{ SE}_b \]

\[ \frac{\text{ SE}_b}{b} = \frac{1.2}{1.2} = 1.0 \]

\[ \text{ Statistic becomes } \]  

\[ \text{ The test } \]

\[ \text{ N. I. E. } \text{ Ho: } \beta = 0 \text{ The test } \]

\[ \text{ that } \text{ is no effect or influence only } \]

\[ \text{ To mean } \]

\[ \text{ i.e. because } \text{ doesn't change with } \]

\[ \text{ if } \beta = 0 \text{ then } \text{ no change with } \]

\[ \text{ that } \]

\[ \text{ REA and } \text{ are common. } \]

\[ \text{ Testing and } \text{ SLR} \]

\[ \text{ using } \frac{\text{ SE}_b}{\beta} \]

\[ \text{ And I can test } \text{ Ho: } \beta = 0 \text{ with } \text{ if } n = 2 \]

\[ \text{ using } \frac{\text{ SE}_b}{\beta} = \frac{b}{1.2} = -1.2 \]

\[ \text{ I can test } \text{ Ho: } \beta = 0 \text{ if } \text{ or } \]
There is a 2nd method of testing $H_0: \beta_1 = 0$ (that for SLR turns out to be equivalent to $t$ testing) and is built on the sums of squares we cooked up when defining $R^2$ - computations for this are usually organized in an ANOVA table (or Analysis of Variance table). The particular ANOVA table for SLR is:

\[
\begin{array}{c|c|c|c|c}
\text{Source} & \text{SS} & \text{df} & \text{MS} & \frac{F}{F_{\alpha}} = F \\
\hline
\text{Regression} & \text{SSR} & 1 & \text{MSR} & \text{SSR/MSR} = F \\
\text{Error} & \text{SSE} & n-2 & \text{MSE} & \text{SSE/MSE} = F \\
\hline
\text{Total} & \text{SSTot} & n-1 & & \\
\end{array}
\]

Exempyle Real Estate (test $H_0: \beta_1 = 0$ using "ANOVA table")

\[
\begin{array}{c|c|c|c|c}
\text{Source} & \text{SS} & \text{df} & \text{MS} & F \text{ ratio} \\
\hline
\text{Regression (size)} & 727.85 & 1 & 727.85 & 58.43 \\
\text{Error} & 827.50 & 8 & 124.43 & X \\
\hline
\text{Total} & 827.50 & n-1 & & \\
\end{array}
\]

Intuitively, it's reasonable to say that large observed $F$ will count as evidence against $H_0: \beta_1 = 0$ in favor of $H_a: \beta_1 \neq 0$ i.e. $X$ is important in describing $y$ ... What is "large"?

We need to know a new probability fact and need to introduce a new set of distributions -

\[
F = \frac{\text{MSS}/\text{df}}{\text{MSE}}
\]
These look like:

See Tables beginning on T-12 for upper % pts. For this application we want numerator d.f. = 1

denominator d.f. = n - 2.

We therefore have evidence that size is an important predictor of home price.

It looks like we have 2 different ways of testing $H_0: \beta = 0$ (t-test
and F test) — but as it turns out

$\left( \frac{\text{value of } T}{\text{stat for } H_0: \beta = 0} \right)^2 = \frac{\text{value of } F \text{ statistic for } H_0: \beta = 0}{F_{1, n-2}}$

Also, values in 1st column of F table are squares of values in t table.

Examples: Real Estate

Use F dist with 1, 8 d.f.

so we see that the observed F is way beyond the upper .001 pt of this dist

$\text{F}_{1, 8}$

50.43

p-value = very small

For example

$\left( \frac{\text{upptl 5%}}{\text{pt of t}_8} \right)^2 = 3.46 = \text{upper 1% pt of } F_{1, 8}$

$(1.86)^2$

So why introduce both tests? Answer: in SLR case they are the same, but do different jobs in MLR

Exercise: Take the "by hand" data, Make an ANOVA table and compute F
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>14.4</td>
<td>1</td>
<td>14.4</td>
<td>27.0</td>
</tr>
<tr>
<td>Error</td>
<td>1.6</td>
<td>3</td>
<td>0.533</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: $T^2 = \left( \frac{1.2}{\frac{173}{110}} \right)^2 = 27.0$

In those cases where I do wish to do inference for $\beta_0$, I simply use my formula for the mean value of $y$

$\hat{y}_x = \beta_0 + \beta_1 x$

with the choice $x = 0$ -- so, for example,

$SE_{\beta_0} = \sqrt{\frac{SE_{\beta_0}^2}{n}}$ where I use $x = 0$

$= \sqrt{\frac{1}{n} + \frac{(0-\bar{x})^2}{\sum(x-\bar{x})^2}}$

A bit of house cleaning regarding SLR -- it's not usually of interest to make inferences for $\beta_0$.

$y$

Because $\beta_0$ is the $y$-value for $x=0$ and that's typically an extrapolation.

Multiple Linear Regression Ch II

Real problems usually have more than 1 predictor variable:

$x_1, x_2, \ldots, x_k$ > use all to predict/explain $y$

Generalize from

SLR $y = \beta_0 + \beta_1 x + \epsilon$

to

MLR $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$

and end up doing everyting from SLR plus more....
Example Real Estate

\[ y = \text{price} \]
\[ x_1 = \text{size} \]
\[ x_2 = \text{condition (10 = best)} \]

and we'll try to model/explain price
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \]

Fitted equation
\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \]

I choose \( b_0, b_1, b_2, \ldots, b_k \) to minimize
\[ \sum (y - \hat{y})^2 \]

This a calculus problem that has a solution that can conveniently be written only using matrices - we'll use JMP and trust that SAS knows how to do least squares

**NOTICE:** SLR formulas for coefficients do NOT work here.

Conceptually here's a \( k=2 \) picture

\[ \hat{y} = 9.78 + 1.87 x_1 + 1.28 x_2 \]

Not the same as for SLR

1.87 \( \rightarrow \) \$18.70/ft\(^2\) increase in price holding condition fixed

1.28 \( \rightarrow \) \$1,280/condition \# increase holding size fixed
How to measure the "goodness" of an equation fit to the data? Use $R^2$. 

$R^2$ is interpreted as a "fraction of raw variability in response (y) accounted for by $x_1, x_2, \ldots, x_k$".

**Example: Real Estate**

$y = \text{price}$ $x_1 = \text{size}$ $x_2 = \text{condition}$

SLR (on size) $R^2 = .88$

MLR (on size + condition) $R^2 = .99$

Adding "condition" improves my ability to explain/predict price.

BTW SLR (on condition) $R^2 = .14$.

Use $R^2 = \frac{SSR}{SS_{Tot}}$ as before $\sum(y - \bar{y})^2$

As in SLR, I can make up data vectors into $\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$,

$$\text{SSE} = \sum (y - \hat{y})^2 \quad \text{I'll then define}\quad SSR = \text{Tot-SSE}$$

(Where the $\hat{y}$'s? I have to plug data vectors into $\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k$).

Note: $.99 \neq .88 + .14$

BTW $R^2$ is a squared correlation between $y$ and $\hat{y}$

**Example: "by hand" fake data**

Now with an $x_2$ in addition to $x_1$.)

From JMP

$\hat{y} = .8 - 1.9 x_1 + 2 x_2$

Find: $\hat{y}$ values, $\text{SSE}, R^2$
\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \]

**Picture:** \((k = 2)\)

**In symbols:**

- **To estimate \( \beta \)'s I use \( b \)'s**
- **Least squares coefficients (from JMP)**

**Supplementary Notes:**

- **Simple - Multiple Linear Regression (MLR)**
- **Inference in MLR**
- **To do this we need a model... the (normal) multiple linear regression model**
- **In Words:** The mean of \( y \) depends linearly on \( x_1, x_2, \ldots, x_k \) and for fixed \( x_1, x_2, \ldots, x_k \), \( y \)'s are normal with a std dev that doesn't depend on \( x \)'s

**Recall:** SSTot = 16 so

\[ \text{SSE} = 16 - 4 = 12.6 \]

**In symbols:**

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

**Normal, mean 0, std dev \( \sigma \), variable**

**To estimate \( \beta \)'s I use \( b \)'s**

**Least squares coefficients (from JMP)**

**Supplementary Notes:**

- **To estimate \( \beta \)'s I use **
- **Least squares coefficients (from JMP)**

**Supplementary Notes:**

- **To estimate \( \beta \)'s I use**
- **Least squares coefficients (from JMP)**

\[ s_{\text{MLR}} = \sqrt{\frac{\text{SSE}}{n - k - 1}} \]

**Supplementary Notes:**

- **I will soon drop this**

**Supplementary Notes:**

- **I will soon drop this**
Example: Real Estate MLR

handout: $s = 1.08 \quad (+1000$)

"Root mean square error" in JMP

This is my estimated standard deviation of
price for any fixed combination of
size ($x_1$) and condition ($x_2$)

\[ \frac{1.08}{\text{MLR}} \leq \frac{3.53}{\text{SLR}} \]

This makes sense

(there are no simple formulas for
SE_b_i) (do not try to carry over SLR
formulas)

Example: Real Estate

95% confidence limits for $r$

\[ (1.08 \frac{7}{16.01}, 1.08 \frac{7}{1.08}) \]

upper 2.5% pt of $X^2$ with 7 df.

lower 2.5% pt of $X^2$ with 7 df.

Confidence Intervals:

For $r$: We can here use $X^2$ tables
and formula

\[ (s_{\text{MLR}} \frac{n-k-1}{\chi^2}, s_{\text{MLR}} \frac{n-k-1}{\chi^2}) \]

% pts of $X^2$ dfn with df $= n-k-1$

For $\beta_i$ use JMP report and
df $= n-k-1$

$\hat{b}_i \pm t SE_{b_i}$

95% confidence limits for $\beta_1, \beta_2$:

For $\beta_1$:

\[ b_1 \pm \pm 0.076 \]

1.87 \pm 2.365 (.076)

For $\beta_2$:

\[ b_2 \pm \pm 0.14 \]

1.28 \pm 2.365 (.14)

(0.71, 2.19)
(b's are estimated increases in average y for a unit increase in xj all other x's held fixed)

Exercise  For fake data, find s, make 95% confidence limits for \( \beta \) and use std errors on JMP report to make 95% confidence limits for \( \beta_1, \beta_2 \)

\[
s = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{4}{5-2-1}} = \sqrt{2}
\]

\[
= .4472
\]

Confidence Limits for \( \hat{\beta} \mid x_1, x_2, ..., x_k \)

\[
\uparrow \pm t \left( \text{SE}_{\hat{\beta}} \right)
\]

has to come from JMP report - changes with \( x \)'s

Prediction Limits for \( y \) now for a given set of predictors \( x_1, x_2, ..., x_k \)

\[
\uparrow \pm t \left( \text{SE}_y \right)
\]

\[
\text{SE}_y = \sqrt{s^2 + (\text{SE}_{\hat{\beta}})^2}
\]

95% confidence limits use
d.f. = 5-2-1 = 2

\[
\left( s\sqrt{\frac{2}{u}}, s\sqrt{\frac{2}{l}} \right)
\]

\[
\left( .4472\sqrt{\frac{2}{7.373}}, .4472\sqrt{\frac{2}{0.816}} \right)
\]

For \( \beta_1 \): 
\[
-1.8 \pm 4.303(1.283)
\]

upper 2.5% pt 1.22
lower 2.5% pt 1.22

For \( \beta_2 \): 
\[
\hat{\beta}_2 = 2 \pm 4.303 \sqrt{0.816} (3.51)
\]