1. A psychiatrist wants to know whether the level of pathology (Y) in psychotic patients 6 months after treatment could be predicted with reasonable accuracy from knowledge of pretreatment symptom ratings of thinking disturbance (X_1) and hostile suspiciousness (X_2).

(a) The least squares estimation equation involving both independent variables is given by

\[ Y = -0.628 + 23.639(X_1) - 7.147(X_2) \]

Using this equation, determine the predicted level of pathology (Y) for a patient with pretreatment scores of 2.80 on thinking disturbance and 7.0 on hostile suspiciousness. How does the predicted value obtained compare with the actual value of 25 observed for this patient?

\[ \hat{Y} = -0.628 + 23.639 \cdot 2.80 - 7.147 \cdot 7.0 \Rightarrow \]

\[ \hat{Y} = 15.5322 \]

This value is lower than the observed value of 25.

(b) Using the analysis of variance tables below, carry out the overall regression F tests for models containing both X_1 and X_2, X_1 alone, and X_2 alone.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression on X_1</td>
<td>1</td>
<td>1546</td>
</tr>
<tr>
<td>Residual</td>
<td>51</td>
<td>12246</td>
</tr>
<tr>
<td>Source</td>
<td>DF</td>
<td>Sum of Squares</td>
</tr>
<tr>
<td>---------------</td>
<td>----</td>
<td>----------------</td>
</tr>
<tr>
<td>Regression on $X_2$</td>
<td>1</td>
<td>160</td>
</tr>
<tr>
<td>Residual</td>
<td>51</td>
<td>13632</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression on $X_1, X_2$</td>
<td>2</td>
<td>2784</td>
</tr>
<tr>
<td>Residual</td>
<td>50</td>
<td>11008</td>
</tr>
</tbody>
</table>

Model Containing $X_1$ and $X_2$

$$F = \left( \frac{SSQ \text{ regression on } X_1 \text{ and } X_2} {\text{Regression df}} \right) \div \left( \frac{SSQ \text{ residual} } {\text{Residual df}} \right)$$

$$= \left( \frac{2,784}{2} \right) \div \left( \frac{11,008}{50} \right) = 6.3227$$

on $DF=2,50$

$p$-value $= 0.00356$ $\Rightarrow$

Application of the null hypothesis model has led to an extremely unlikely result ($p$-value = 0.00356), prompting statistical rejection of the null hypothesis. The fitted linear model in $X_1$ and $X_2$ explains statistically significantly more of the variability in level of pathology ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.

Model Containing $X_1$ ALONE

$$F = \left( \frac{SSQ \text{ Regression on } X_1} {\text{DF Regression}} \right) \div \left( \frac{SSQ \text{ residual}} {\text{DF Residual}} \right)$$

$$= \left( \frac{1546}{1} \right) \div \left( \frac{12,246}{51} \right) = 6.4385$$

on $DF=1,51$

$p$-value $= 0.01427$

Here, too, application of the null hypothesis model has led to an extremely unlikely result ($p$-value = .014), prompting statistical rejection of the null hypothesis. The fitted linear model in $X_1$ explains statistically significantly more of the variability in level of pathology ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.

Model Containing $X_2$ ALONE

$$F = \left( \frac{SSQ \text{ Regression on } X_2} {\text{DF Regression}} \right) \div \left( \frac{SSQ \text{ residual/DF Residual}} {\text{DF Residual}} \right)$$

$$= \left( \frac{160}{1} \right) \div \left( \frac{13,632}{51} \right) = 0.5986$$

on $DF=1,51$

$p$-value $= 0.44268$

Here, application of the null hypothesis model has not led to an extremely unlikely result ($p$-value = .44). The null hypothesis is therefore not rejected. The fitted linear model in $X_2$ does not explain statistically significantly more of the variability in level of pathology ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.
(c) Based on your results in part (b), how would you rate the importance of the two variables in predicting Y?

\[ X_1 \text{ explains a significant proportion of the variability in } Y \text{ when modelled as a linear predictor.} \\
X_2 \text{ does not. (However, we don't know if a different functional form might have been important.)} \]

(d) What are the \( R^2 \) values for the three regressions referred to in part (b)?

Total SSQ = (Regression SSQ) + (Residual SSQ) is constant. Therefore total SSQ can be calculated from just one anova table:

\[
\text{Total (SSQ)} = 1,546 + 12,246 = 13,792 \\
\]

\[
R^2 (X_1 \text{ only}) = \frac{\text{Regression SSQ}}{\text{Total SSQ}} = \frac{1546}{13,792} = 0.1121 \\
R^2 (X_2 \text{ only}) = \frac{160}{13,792} = 0.0116 \\
R^2 (X_1 \text{ and } X_2) = \frac{2784}{13,792} = 0.2019
\]

(e) What is the best model involving either one or both of the two independent variables?

Eliminate from consideration model with \( X_2 \) only.
Compare model with \( X_1 \) alone versus \( X_1 \) and \( X_2 \) using partial F test.

\[
\text{Partial } F = \frac{\{(\text{SSQ Regression on } X_1,X_2) - (\text{SSQ Regression on } X_1)\}/\text{VDF}}{\text{SSQ Residual for model w } X_1,X_2/\text{Residual DF}} = \frac{2784 - 1546}{11,008}/50
\]

= 5.6263 on DF=1,50
P-value = 0.02162

Addition of \( X_2 \) to model containing \( X_1 \) is statistically significant (p-value = .02). \( \rightarrow \)

More appropriate model includes \( X_1 \) and \( X_2 \)
In an experiment to describe the toxic action of a certain chemical on silkworm larvae, the relationship of \( \log_{10}(\text{dose}) \) and \( \log_{10}(\text{larva weight}) \) to \( \log_{10}(\text{survival}) \) was sought. The data, obtained by feeding each larva a precisely measured dose of the chemical in an aqueous solution and then recording the survival time (ie time until death) are given in the table. Also given are relevant computer results and the analysis of variance table.

\[
\begin{align*}
\text{Larva} & \quad 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
Y = \log_{10}(\text{survival time}) & \quad 2.836 & 2.966 & 2.687 & 2.679 & 2.827 & 2.442 & 2.421 & 2.602 \\
X_1 = \log_{10}(\text{dose}) & \quad 0.150 & 0.214 & 0.487 & 0.509 & 0.570 & 0.640 & 0.781 & \\
X_2 = \log_{10}(\text{weight}) & \quad 0.425 & 0.439 & 0.301 & 0.325 & 0.371 & 0.093 & 0.140 & 0.406 \\
\text{Larva} & \quad 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
Y = \log_{10}(\text{survival time}) & \quad 2.556 & 2.441 & 2.420 & 2.439 & 2.385 & 2.452 & 2.351 & \\
X_1 = \log_{10}(\text{dose}) & \quad 0.739 & 0.832 & 0.865 & 0.904 & 0.942 & 1.090 & 1.194 & \\
X_2 = \log_{10}(\text{weight}) & \quad 0.364 & 0.156 & 0.247 & 0.278 & 0.141 & 0.289 & 0.193 &
\end{align*}
\]

\[
\begin{align*}
Y &= 2.952 - 0.550 (X_1) \\
Y &= 2.187 + 1.370 (X_2) \\
Y &= 2.593 - 0.381 (X_1) + 0.871 (X_2)
\end{align*}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression on ( X_1 )</td>
<td>1</td>
<td>0.3633</td>
</tr>
<tr>
<td>Residual</td>
<td>13</td>
<td>0.1480</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression on ( X_2 )</td>
<td>1</td>
<td>0.3367</td>
</tr>
<tr>
<td>Residual</td>
<td>13</td>
<td>0.1746</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression on ( X_1, X_2 )</td>
<td>2</td>
<td>0.4642</td>
</tr>
<tr>
<td>Residual</td>
<td>12</td>
<td>0.0471</td>
</tr>
</tbody>
</table>
(a) Test for the significance of the overall regression involving both independent variables $X_1$ and $X_2$.

\[
F = \frac{(SSQ \text{ regression on } X_1 \text{ and } X_2) / 2}{(SSQ \text{ Residual}) / 12} = \frac{0.4642}{0.0471} = 59.18 \text{ on } DF = 2,12
\]

\[P-value < 0.0001\]

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .0001), prompting statistical rejection of the null hypothesis. The fitted linear model in $X_1$ and $X_2$ explains statistically significantly more of the variability in $\log_{10}$ (survival time) ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.

(b) Test to see whether using $X_1$ alone significantly helps in predicting survival time.

\[
F = \frac{(SSQ \text{ Regression on } X_1) / 1}{(SSQ \text{ Residual}) / 13} = \frac{0.3633}{0.1480} = 31.9115 \text{ on } DF = 1,13
\]

\[P-value = 0.00008\]

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .00008), prompting statistical rejection of the null hypothesis. The fitted linear model in $X_1$ explains statistically significantly more of the variability in $\log_{10}$ (survival time) ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.

(c) Test to see whether using $X_2$ alone significantly helps in predicting survival time.

\[
F = \frac{(SSQ \text{ Regression on } X_2) / 1}{(SSQ \text{ Residual}) / 13} = \frac{0.3367}{0.1746} = 25.07 \text{ on } DF = 1,13
\]

\[P-value = 0.00027\]

Application of the null hypothesis model has led to an extremely unlikely result (p-value = .00027), prompting statistical rejection of the null hypothesis. The fitted linear model in $X_2$ explains statistically significantly more of the variability in $\log_{10}$ (survival time) ($Y$) than is explained by $\bar{Y}$ (the intercept model) alone.
(d) Compute $R^2$ for each of the three models.

TotalSSQ = 0.5113

$R^2 (X_1 \text{ and } X_2) = \frac{0.4642}{0.5113} = 0.9079$

$R^2 (X_1 \text{ alone}) = \frac{0.3633}{0.5113} = 0.7105$

$R^2 (X_2 \text{ alone}) = \frac{0.3367}{0.5113} = 0.6585$

(e) Which independent predictor do you consider to be the best single predictor of survival time?

Using just the criteria of the overall F test and comparison of $R^2$, the single predictor model containing $X_1$ is better.

(f) Which model involving one or both of the independent predictors do you prefer and why?

Partial F for comparing model with $X_1$ alone versus model with $X_1$ and $X_2$

$$= \frac{(\text{RegressionSSQ})/\text{Reg DF}}{(\text{ResidualSSQ-model w } X_1, X_2)/\text{Residual DF-model w } X_1, X_2} = \frac{(0.4642-0.3633)/(2-1)}{0.0471/12}$$

$$= 25.707 \text{ on } \text{DF=1,12}$$

P-value = 0.0003 Choose model with both $X_1$ and $X_2$
3. Using Stata, try your hand at reproducing the analysis of variance tables you worked with in problem #2.

* The following assumes you: a) downloaded larvae.dta; and b) opened it
* Stata commands describe and notes to get information on data set

```
. * describe

describe

Contains data from /Users/carolbigelow/Desktop/larvae.dta
obs: 15
vars: 4
size: 240

-------------------------------------------------------------------------------
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g                  larva id
y               float  %9.0g                  log10(survival)
x1              float  %9.0g                  log10(dose)
x2              float  %9.0g                  log10(weight)
-------------------------------------------------------------------------------
Sorted by:
. notes
dta:
  1. "Week 3 homework assignment exercises 2 and 3"

. ***** 2) Fit model containing x1 alone using command regress yvariable xvariable
. regress y x1

Source |       SS       df             MS
--------+--------------------------------------------------
Model   |  .36327405     1  .36327405             F(  1,    13) =   31.91
Residual|  .14799289    13  .01138406
        +--------------------------------------------------
R-squared     =  0.7105
Adj R-squared =  0.6883
Total      | .51126694    14  .03651907
--------+--------------------------------------------------

           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+--------------------------------------------------
x1 |  -.5498559   .0973376    -5.65   0.000    -0.7601409   -0.3395709
_cons |   2.952199   .0735550    40.14   0.000     2.793293    3.111105

. ***** 2) Model containing x2 alone
. regress y x2

Source |       SS       df             MS
--------+--------------------------------------------------
Model   |  .33667413     1  .33667413             F(  1,    13) =   25.07
Residual|  .17459282    13  .01343021
        +--------------------------------------------------
R-squared     =  0.6585
Adj R-squared =  0.6322
Total      | .51126694    14  .03651907
--------+--------------------------------------------------

           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+--------------------------------------------------
x2 |   1.375579   .2747401     5.01   0.000     .7820388    1.969119
_cons |   2.184706   .0819958    26.64   0.000     2.007565    2.361847
```
2. Regression and Correlation

2) Model containing both x1 and x2

```
Source |       SS       df       MS              Number of obs = 15
-------------+--------------------------------------------------
Model | .46420206     2   .23210103           Prob > F      =  0.0000
Residual | .04706488    12  .003922073           R-squared     =  0.9079
-------------+--------------------------------------------------
Total | .51126694    14  .036519067           Root MSE      =  .06263
-------------+--------------------------------------------------

                      y |  Coef.  Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+-----------------------------------------------
      x1 |  -.3784805   .0663741  -5.70   0.000   -.5230972  -.2338637
      x2 |   .8749749   .1724835   5.07   0.000    .4991656    1.250784
    _cons |   2.588996   .0836079  30.97   0.000    2.40683    2.771162
```

4. An educator examined the relationship between number of hours devoted to reading each week (Y) and the independent variables social class (X1), number of years school completed (X2), and reading speed measured by pages read per hour (X3). The analysis of variance table obtained from a stepwise regression analysis on data for a sample of 19 women over the age of 60 is shown.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (X3)</td>
<td>1</td>
<td>1058.628</td>
</tr>
<tr>
<td>(X2</td>
<td>X3)</td>
<td>1</td>
</tr>
<tr>
<td>(X1</td>
<td>X2,X3)</td>
<td>1</td>
</tr>
<tr>
<td>Residual, corrected</td>
<td>15</td>
<td>363.300</td>
</tr>
<tr>
<td>Total, corrected</td>
<td>18</td>
<td>1643.653</td>
</tr>
</tbody>
</table>

RECALL how to read this table.

**Regression (X3) Sum of Squares = 1058.628:**
This is the regression sum of squares for the model containing the one predictor X3.

**Regression (X2|X3) Sum of Squares = 183.743:**
The vertical line in this notation stands for “conditional on”
This is the extra regression sum of squares obtained by adding the predictor X2 to a model that already has the predictor X3. You can also think of this as the change in the model sum of squares accompanying the addition of the predictor X2, controlling for (conditional on X3).

**Regression (X1|X2,X3) Sum of Squares = 37.982:**
The vertical line in this notation stands for “conditional on”
This is the extra regression sum of squares obtained by adding the predictor X1 to a model that already has the predictors X2 and X3. You can also think of this as the change in the model sum of squares accompanying the addition of the predictor X1, controlling for (conditional on X2 and X3).
(a) Test the significance of each variable as it enters the model.

Total SSQ = 1643.653

**Step 1**  \( X_3 \) enters.

Regression SSQ = 1058.628 on DF=1  
Residual SSQ = (363.300) + (37.982) + (183.743) = 585.025 on DF=17  
Overall F = \( \frac{1058.628}{585.025} \) = 30.7622 on DF=1,17  
P-value = 0.00004  
**INTERPRETATION:** Model in \( X_3 \) is statistically significantly better than the “intercept alone” model.

**Step 2**  \( X_1 \) already in.  \( X_2 \) Enters

Additional Regression SSQ = 183.743 on DF=1  
Residual SSQ = (363.300) + (37.982) = 401.282 on DF=16  
Partial F = \( \frac{183.743}{401.282} \) = 7.3262 on DF=1,16  
P-value = 0.01556  
**INTERPRETATION:** The extra variability that is explained by \( X_2 \), beyond the variability that is explained by \( X_3 \) is statistically significant. Another way of saying this is the following. \( X_2 \) is significant, after adjustment for \( X_3 \).

**Step 3**  \( X_2 \) and \( X_3 \) already in.  \( X_1 \) Enters

Additional Regression SSQ = 37.982 on DF=1  
Residual SSQ = 363.300 on DF=15  
Partial F = \( \frac{37.982}{363.3} \) = 1.5682 on DF=1,15  
P-value=0.22964  
**INTERPRETATION:** The extra variability that is explained by \( X_1 \), beyond the variability that is explained by \( X_3 \) and \( X_2 \) is **not** statistically significant. Another way of saying this is the following. After adjustment for \( X_3 \) and \( X_2 \), \( X_1 \) is **not** statistically significant.
(b) Test \( H_0: \beta_1 = \beta_2 = 0 \) in the model \( Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E \).

Models to Compare: \( X_3 \) alone versus \( (X_1, X_2 \text{ and } X_3) \)

Note – The solution for the Partial F statistic looks different from the second formula that is on page 51 of the Unit 2 notes. Here, because we have the values of the extra sum of squares already, we can use them directly.

Partial \( F = \frac{\Delta \text{Regression SSQ}/\Delta \text{Reg DF}}{\text{Residual SSQ}/\text{Res DF}} \)

\[
= \frac{\left\{ \left[ \text{extra Regression SSQ w } X_2 \text{ given } X_3 \right] + \left[ \text{extra Regression SSQ w } X_1 \text{ given } X_3 \text{ and } X_2 \right] \right\} / (2)}{(\text{Residual SSQ for model with } X_3, X_2, \text{ and } X_1) / (\text{Residual DF for model with } X_3, X_2, \text{ and } X_1)}
\]

\[
= \frac{(183.743 + 37.982)/2}{363.300/15} = 4.5773 \text{ on DF } = 2, 15
\]

\( P \)-value= 0.02837

INTERPRETATION: The extra variability that is explained by \( X_1 \) and \( X_2 \), beyond the variability that is explained by \( X_3 \) is statistically significant. That is, \( X_1 \) and \( X_2 \) are significant, after adjustment for \( X_3 \).

(c) Why can’t we test \( H_0: \beta_1 = \beta_3 = 0 \) using the ANOVA table given? What formula would you use for this test?

The regression SSQ for the model containing \( X_2 \) alone is not available.

The formula for the required Partial F test is:

\[
\text{Partial F} = \frac{\left\{ \text{Regression SSQ(model w } X_1, X_2, X_3) - \text{ Regression SSQ(model w } X_2) \right\} / 2}{\text{Residual SSQ(model w } X_1, X_2, X_3)/15}
\]

(d) What is your overall evaluation concerning the appropriate model to use given the results in parts (a) and (b)?

The most appropriate model is the one with two predictors, \( X_1 \) and \( X_2 \) \((R^2 = 0.7559)\).

The additional predictive information in \( X_1 \) (change in \( R^2 = 0.0231 \)) is not statistically significant \((p=0.23)\).
5. Consider the following analysis of variance table.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression (X₁)</td>
<td>1</td>
<td>18,953.04</td>
</tr>
<tr>
<td>(X₃</td>
<td>X₁)</td>
<td>1</td>
</tr>
<tr>
<td>(X₂</td>
<td>X₁,X₃)</td>
<td>1</td>
</tr>
<tr>
<td>Residual</td>
<td>16</td>
<td>2,248.23</td>
</tr>
<tr>
<td>Total, corrected</td>
<td>19</td>
<td>28,222.23</td>
</tr>
</tbody>
</table>

Using a type I error of 0.05,

(a) Provide a test to compare the following two models:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E. \]  
VERSUS  
\[ Y = \beta_0 + \beta_1 X_1 + E. \]

\[
\text{Partial } F = \frac{\Delta \text{Regression SSQ}/\Delta \text{Reg DF}}{\text{Residual SSQ}/\text{Res DF}} \\
= \left( \frac{\text{extra Regression SSQ w X}_3 \text{ given } X_1}{} + \frac{\text{extra Regression SSQ w X}_2 \text{ given } X_1 \text{ and } X_3}{} \right) / (2) \\
\text{Residual SSQ for model with } X_1, X_3, \text{ and } X_2 \\
= \left( \frac{7010.03 + 10.93}{2248.23/16} \right) = 24.983 \text{ on } \text{DF} = 2, 16 \\
p\text{-value} < .0001
\]

INTERPRETATION: The extra variability that is explained by X₃ and X₂, beyond the variability that is explained by X₁ is statistically significant. Thus, X₃ and X₂ are statistically significant, after adjustment for X₁.
(b) Provide a test to compare the following two models:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E. \text{ VERSUS} \]

\[ Y = \beta_0 + E. \]

\[
\text{Overall } F = \frac{\text{Regression SSQ}(X_1, X_3)/2}{\text{Residual SSQ}(X_1, X_3)/17} = \frac{(18,953.04 + 7010.03)/2}{(2,248.23 + 10.93)/17}
\]

\[ = 97.685 \quad \text{on} \quad \text{DF}=2,17 \]

\[ P\text{-value}<0.00001 \]

**INTERPRETATION:** The fitted linear model in \(X_1, X_2, \text{and } X_3\) explains statistically significantly more of the variability in \(Y\) than is explained by \(\bar{Y}\) (the intercept model) alone.

(c) State which two models are being compared in computing:

\[ F = \frac{(18,953.04 + 7,010.03 + 10.93)/3}{(2248.23)/16} \]

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E \]

\[ \text{versus} \]

\[ Y = \beta_0 + E \]