OLS Regression - Worksheet

This assignment consists of 5 parts:

- Compute regression coefficients
- Compute R^2
- Compute F statistic and perform hypothesis test
- Using regression model for prediction
- Run a regression analysis using Python

Table 1 shows the raw data where P is pharmacy, x = % ingredients purchased directly, y = sales volume (in \$1000).

Part 1: Compute regression coefficients

- Find the least squares estimate for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- Plot the x, y data and the prediction equation, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

<u>**Getting Started**</u>

a) Compute $\bar{x}$ and $\bar{y}$

b) Complete section 1 in **Table 1** using $\bar{x}$ and $\bar{y}$

c) Confirm the following (by filling in the table):

$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 3407.60$, $\quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 6714.60$

d) Confirm the following computation of the regression coefficients:

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6714.60}{3407.60} = 1.9704778$ rounded to 1.97

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 71.3 - 1.9704778(33.8) = 4.6978519$, rounded to 4.70

e) Write out the regression line:

$y =$

Part II: Compute R squared

R^2 is also called the coefficient of determination. It is a number that indicates how well data fit a statistical model.

We will compute the R^2 for this set of data by completing the rest of **Table 1**.

Getting Started

a) First, we need to compute $\hat{y}$ for each row of data. $\hat{y}$ is the predicted value, computed from our given regression line: $\hat{y} = 4.70 + 1.97x$

b) Compute the residual for each data point. $residual = y - \hat{y}$

c) Compute SS_{res} which is Residuals Sum of Squares (aka SS_e = Error Sum of Squares).

d) Compute SS_{expl} which is Explained Sum of Squares (aka SS_{reg} = Regression Sum of Squares).

e) Compute SS_{yy} (also called SS_{tot}) which is Total Sum of Squares.

f) Compute $R^2 = SS_{expl}/SS_{tot} =$

Part III: Compute F statistic and perform hypothesis test

Recall our equation for predicted line is : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

a) State the null and alternative hypothesis:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

b) Compute the degrees of freedom.

In this numerical example:

$n = 10$ and $p = 2$

$DF_{expl} = p - 1 = 2 - 1 = 1$ $(Note : p_1 = \beta_0$ and $p_2 = \beta_1)$

$DF_{res} = n - p = 10 - 2 = 8$

c) Compute the F test statistic.

$F_{(test\ statistic)} = \frac{SS_{expl}}{DF_{expl}} / \frac{SS_{res}}{DF_{res}}$

$F_{(test\ statistic)} = \frac{MS_{expl}}{MS_{res}}$

$F_{(test\ statistic)} = \frac{explained\_variance}{unexplained\_variance}$

$F = \frac{\frac{13224.6}{1}}{\frac{651.13}{8}}$

$F_{(test\ statistic)} = xxx.x$ with num, den degrees of freedom

d) Compute F critical.

$F_{(critical)} = F_{(0.05,1,8)} = 5.32$

e) Test the hypothesis. Compare F test statistic to F critical and draw a conclusion.

$\therefore$ Since $(F_{(test\ statistic)} = xxx.x) > (F_{(critical)} = 5.32)$ reject the null hypothesis that the slope is 0.

Note: you should replace "xxx.x" with the computed F.

Part IV: Using regression model for prediction

Given a new x value, we can predict the y value using the regression model. Let's predict the following:

- Sales volume for a pharmacy that purchases 15 percent of its prescriptions directly from the supplier
- 95% confidence interval for the prediction $\hat{y}$ when $x = 15$

Getting Started

If the regression line is: $\hat{y} = 4.7 + 1.97x$ then,

Predicted y is: $\hat{y} = 4.7 + 1.97 * (15) = 34.25$

The 95% confidence interval for the forecasted value $\hat{y}$ is: $\hat{y} \pm t_{crit} * s.e.$

where $s.e. = \hat{\sigma} = \sqrt{SSE/(n-2)}$, where $n-2$ is the degrees of freedom (we lose 2 degrees of freedom when we estimate β_0 and β_1)

and

SSE = Error Sum of Squares (which is same as Residuals Sum of Squares)

and

where $t^*_{crit(df=8,\alpha=0.05)} = 2.306$

A confidence interval for $E(y|x^*)$, the average (expected) value of y for a given x^*, is

$\hat{y} \ \pm \ t^* \ MS_{res}\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{(n-1)s_x^2}}$, where: $MS_{res} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{(n-2)}} = \sqrt{\frac{SS_{res}}{(n-2)}}$

$\hat{y} \ \pm 2.306 * \sqrt{81.3912} * \sqrt{\frac{1}{10} + \frac{(15-33.8)^2}{(10-1)(3407.60)}}$

$\hat{y} \ \pm$ _______ (fill in this quantity and finish computing prediction intervals)

Answer: The 95% CI for $\hat{y} = 34.25$ when $x = 15$ is: (,)

Source: Linear Regression Example from Duke University

Part V: Examining the programming output

Look at Listing 1, the regression output generated by R. Run the regression, this time, using Python.

Can you place all the statistics we computed in this worksheet to the computer generated output?

Notice the following:

- In R regression output t = 12.750. The F-statistic is 162. Did you know that in a simple linear regression, $t^2 = F$?
- Do the prediction intervals you computed in Part IV match the regression output?
- What's the Sum Squares of Regression?
- What's the Sum Squares of Error?
- What's the residual standard error? The degrees of freedom?
- Does the R squared you computed match the regression output?
- Look at the summary of residuals (minimum and maximum). Do they match the min and max that you see in the hand-calculated regression table?

Regression Code in R

```
# get current working directory
getwd()

# set working directory
setwd("/Users/reshamashaikh/ds/metis/mypractice/_regression")
getwd()

# output directed to myfile.txt in cwd. output is appended
# to existing file. output also send to terminal.
sink("pharmacy_regr_output.txt", append=TRUE, split=TRUE)

x=c(10, 18, 25, 40, 50, 63, 42, 30, 5, 55)
y=c(25, 55, 50, 75, 110, 138, 90, 60, 10, 100)

sprintf("Plot")
plot(x,y)

sprintf("------------------------------------------------------------")
sprintf("Regression␣Model")
model = lm(y ~ x)
model

sprintf("------------------------------------------------------------")
sprintf("ANOVA")
anova(model)

sprintf("------------------------------------------------------------")
sprintf("Coefficients␣and␣other␣regression␣output")
summary(model)

# add regression line
abline(model)

# predictions
sprintf("------------------------------------------------------------")
sprintf("Predicting␣y␣for␣x=15")
predict(model, newdata=data.frame(x=15))

sprintf("------------------------------------------------------------")
sprintf("Prediction␣Intervals␣for␣x=15")
predict(model, newdata=data.frame(x=15), interval = "pred")

sprintf("------------------------------------------------------------")
sprintf("Confidence␣Intervals␣for␣x=15")
predict(model, newdata=data.frame(x=15), interval = "confidence")
```

Notes

OLS = Ordinary Least Squares

Source: Statistical Thinking for Managers (Hildebrand, Ott & Gray, 2nd Edition, page 510)

Created by: Reshama

Date updated: 20-Jan-2016

Listing 1: R output

```
[1] "Plot"
[1] "--------------------------------------------------------"
[1] "Regression Model"

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      4.698        1.970

[1] "--------------------------------------------------------"
[1] "ANOVA"
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 13231.0 13231.0  162.56 1.349e-06 ***
Residuals  8   651.1    81.4
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
[1] "--------------------------------------------------------"
[1] "Coefficients and other regression output"

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-13.074  -4.403  -1.607   5.719  14.834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6979     5.9520   0.789    0.453
x             1.9705     0.1545  12.750 1.35e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 9.022 on 8 degrees of freedom
Multiple R-squared:  0.9531,    Adjusted R-squared:  0.9472
F-statistic: 162.6 on 1 and 8 DF,  p-value: 1.349e-06

[1] "--------------------------------------------------------"
[1] "Predicting y for x=15"
       1
34.25502
[1] "--------------------------------------------------------"
[1] "Prediction Intervals for x=15"
       fit      lwr      upr
1 34.25502 11.42996 57.08008
[1] "--------------------------------------------------------"
[1] "Confidence Intervals for x=15"
       fit      lwr      upr
1 34.25502 24.86499 43.64505
```

Table 1: Regression Computation

Given			To Compute Regression Coefficients				To Compute R^2				
1	2	3	4	5	6	7	8	9	10	11	12
P	SVol ($1000)	Ing			SS_{xy}	SS_{xx}	$Pred.$	$residual$	SS_{res}	SS_{expl}	$SS_{yy} = SS_{tot}$
i	y	x	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})(x_i - \bar{x})$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})^2$
1	25	10		-23.8		566.44	24.4	0.6	0.36	2199.61	2143.69
2	55	18			257.54						
3	50	25									
4	75	40	3.7								
5	110	50									
6	138	63									
7	90	42									
8	60	30									
9	10	5									
10	100	55	28.7	21.2	608.44	449.44	113.05	-13.05	170.30	1743.06	823.69
Total	$\sum_{i=1}^{n} y_i =$	$\sum_{i=1}^{n} x_i =$	0	0	6714.60	3407.60			651.13		13882.10
Mean	$\bar{y} =$	$\bar{x} =$									