Risto B. Rushford
GSCM 451-Spr 2019
HW04

Concepts to Understand

Briefly comment on each of the following concepts, usually just several sentences.
Topics that can appear on the short-answer part of the Final.

1. Positive and negative correlations.

A correlation is when a data set changes value in a way corresponding to another data set. In terms of a positive correlation, when one data set increases, the other is also observed to increase (but not always at the same rate). And a negative correlation is observed when as one data set increases in value, the other data set decreases in value.

2. The relation between the amount of scattering in a scatter plot and the correlation coefficient.

The statistic which reflects the strength of a linear relationship, which itself correlates on a scale between -1 and positive 1. A -1 correlation coefficient indicates a perfect negative correlation, while 1 represents a perfect positive correlation. A 0 indicates no correlation at all.

3. Relation of an ellipse to the scatter in a scatter plot.

The .95 data ellipse is a tool used in scatterplots to indicate which data points are within two standard deviations of the regression line, with the wide axis of the ellipse centered on the regression line.

4. Null and alternative hypotheses for testing a correlation coefficient.

The null hypothesis is that there is no relationship between two data sets (that is, you begin with the assumption that there is no relationship until proven otherwise). An alternative hypothesis is that the two data sets will have some correlation, whether positive or negative. These hypotheses are being used to find how well the sample statistic corresponds to the population.

5. The confidence interval of a correlation coefficient and what it estimates.

The confidence interval is a range of values that are considered likely to contain the true population correlation (ρ). So for a given estimate of ρ, the confidence interval provides an estimate for ρ given a certain level of confidence.

6. The two goals of regression analysis are:
    - To forecast the unknown value of the response variable (Y) from one or more variables (X) that are used to make the prediction.
    - To analyze the relationship among a set of X-variables to each other and to the response variable Y.

7. Graph of X with Y vs the graph of X with ___.

If X and Y are related to each other, then their graph can be used to forecast the conditional mean which shows the linear relationship between the values.

If X and Y are not related, then X will be plotted as an ordered sequence or time series.

8. Meaning of the slope coefficient in $= b_0 + b_1X_1$.

The slope coefficient (b_1) represents the delta-Yhat for each increase in X, but only from within the particular data set from which it was derived.

9. Meaning and interpretation of the hypothesis test of the slope coefficient.

The hypothesis test of the slope coefficient uses the null hypothesis that $b_1=0$. If the test results with $p > .05$ then there is no statistically significant correlation. If $p <$ then there is a correlation, either positive or negative.

10. Meaning and interpretation of the confidence interval of the slope coefficient.

The confidence interval of the slope coefficient consists of the range of likely possible values for β_1. If 0 is in the interval, then there is no relationship detected. If 0 is not on the interval, then if the interval has all positive values that indicate a positive relationship or a negative interval indicates a negative relationship.

11. Unconditional vs conditional mean:
    - Unconditional mean - the mean of all of the data, *m.*
    - Conditional mean - the mean of Y for just the data with a specific value of X

12. Criterion of ordinary least squares regression to obtain the estimated model

The OLS criterion chooses the sample coefficients (b_0 and b_1) which provide the minimal possible value of the sum of the squared residuals for the specific sample being considered.

13. Training data vs testing data

Training data is used to estimate a model which utilizes known data. All "forecasts" made using this model are fitted values. Any error in this model is modeling error, not forecasting error.

Once the model is developed, it can be fed new data, called *testing data* to make new forecasts. The standard error for these forecasts now include both modeling and forecasting error.

14. Meaning of the residual variable e within the context of the …

a. training data

e in the context of training data represents the Yhat residuals of modeling error.

b. test data or new data

In the context of test data or new data, *e* represents residuals from errors in both modeling and forecasting. To tease out the forecasting error from modeling error, calculate the residuals of Yhat from testing data and not from training data.

15. Standard error of forecast

The standard error of forecast is inclusive of the error from both modeling and forecasting. The residuals that it is derived from will be larger than those from the training data and they cannot be known until the new Y data are measured.

16. Prediction interval

The 95% Prediction Interval is the range of values containing 95% of all future values of Y (the response variable) given the predictor variable X. The size of a prediction interval is dependent upon the extent of training error and sampling variability.

Worked Problems

The analysis questions for a regression analysis are presented at the end of this homework. Use those questions for the analysis of both regression problems.
If there are no seasonal effects (or cycles), we can use linear regression to forecast a trend directly from the data (otherwise we first deseasonalize the data).

Regression with an Explanatory Variable

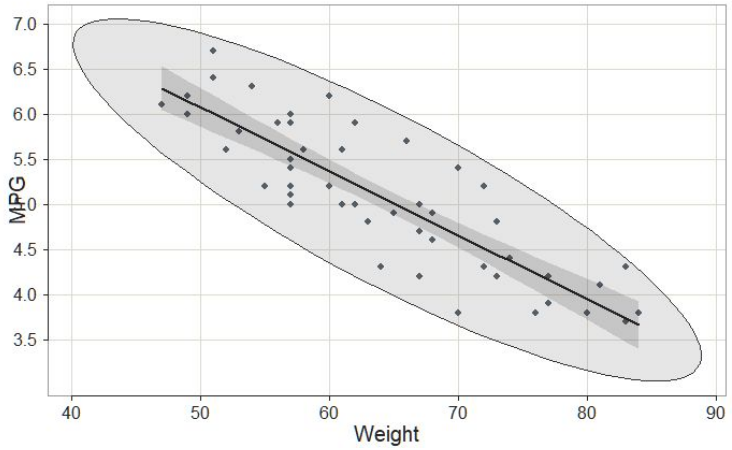
Minimizing, or at least understanding, transportation costs is a concern of supply chain analysis. One important consideration in the estimate of the cost of a trip is fuel mileage. For any truck transportation, the more the weight of the cargo (in thousands of lbs), the worse the fuel mileage (per gallon). Consider the following data regarding this relationship.

Data: http://web.pdx.edu/~gerbing/451/Data/HW4_1.xlsx

Analyze this model with the questions at the end of this homework. The given values of X to forecast from are 59 and 60 thousand pounds of weight.

One-Predictor Regression Analysis Questions:

A. Show the scatter plot of Y and X with the regression line and 0.95 confidence ellipse.



B. What is the correlation of X and Y?

There is a negative correlation.

C. Does there appear to be a relationship between the two variables? Describe.

Yes, as the weight (X, the independent variable) increases, the MPG decreases correspondingly.

Forecasts from the Estimated Model and

D. Write the estimated regression model.

MPG = 9.616 - 0.071(Weight)

E. For the given values of X to forecast, what are the forecasted Y's?

a. Show the form of the computation for the forecasted value of Y applied to the specified values of X. (No computations.)

Yhat = 9.616 - 0.071(59), Yhat = 9.616 - 0.07(60)

b. Have the regression program calculate.

```
reg(MPG ~ Weight, X1.new=c(59, 60))
```

c. What are the prediction intervals for these forecasts?

For weight = 59, pi = 5.434 +/- 0.861
weight = 60, pi = 5.363 +/- 0.86

d. Show the calculation for the prediction intervals given the standard error of forecast (sf on the output). You can solve for the t-cutoff from the information on the output, but feel free to just use 2.
[Note: You will encounter rounding error with only three decimal digits in the output. You can use the digits.d parameter to add more precision if you wish, but not required because this question is about understanding the concept. We let the computer do the calculating.]

sf = 0.428, 2*0.428 = 0.856

e. Interpret the prediction intervals.

The prediction intervals tell me that a 59k lb truckload will have approximately 5.434 MPG, with the variance between 4.574 MPG at the low end and 6.295 at the high end. Similarly, a 60k lb truckload will have approximately 5.363 MPG, with the variance between 4.504 MPG at the low end and 6.223 MPG at the high end.

Model Relationships

F. Specify the null hypothesis and alternative hypothesis for the hypothesis test of the slope coefficient of no relation. [Sec 8.2, #35]

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

G. Show the calculation of the t-statistic of the slope coefficient with the specific numerical values in this analysis. [Sec 8.2, #36]

```
Estimated Model
             Estimate Std Err t-value p-value Lower95% Upper95%
(Intercept)   9.616    0.396   24.288  0.000   8.820    10.412
     Weight  -0.071    0.006  -11.676  0.000  -0.083    -0.059
```

$t_{b1} = (b_1 - 0)/s_{b1} = (-0.071 - 0)/0.006 = -11.8333$

H. Specify the decision rule and outcome of the hypothesis test. [Sec 8.2, #36]

p-value = 0.000 < α = 0.05, so reject the null hypothesis. Since $b_1 < 0$, I conclude that $\beta_1 < 0$.

I. Interpret this hypothesis test (jargon free). [Sec 8.2, #35,36]

As the truckload weight increases, on average, the MPG decreases.

J. What is the value to estimate with the confidence interval for the slope coefficient for the predictor variable? [Sec 8.2, #37]

The confidence interval for the slope coefficient allows for an estimate of β_1 by providing the range of possible values.

K. Show the calculation of the 95% confidence interval for the slope coefficient for the predictor variable listed in the preceding regression output with the specific numerical values in this analysis. [Sec 8.2, #38]

```
            Lower 95%   Upper 95%
(Intercept)  8.820       10.412
Weight      -0.083       -0.059
```

L. Interpret this confidence interval (jargon free). [Sec 8.2, #38

At the 95% level of confidence, for each 1k lbs of weight, on average, the truck's fuel economy will likely decrease from between .059 to 0.083 MPG.

M. Are the results from the confidence interval and hypothesis test consistent with each other? Why or why not? Which analysis provides more information and why? (Focus on the value of the null hypothesis, 0. How does that value relate to the HT and the CI?)

These results are consistent with each other. The hypothesis test assumed that as the weight increased, there would be a corresponding decrease in fuel economy. I believe that the confidence interval provides the more useful test, by indicating the range of how much to expect fuel economy to change for each unit weight change.

Model Fit

N. What is the standard error of estimate? What are some representative values of the standard error of forecast?

The standard error of estimate is 0.006.

O. Compare the fit indices? Which one is larger? Why?

The larger fit index is R-squared (0.740 vs 0.734 adjusted or 0.721 PRESS). R-squared is larger because it is calculated inclusive of the predicted variables. Adjusted divides the sumes of the squares by the degrees of freedom to account for this.

P. What is the value of R2? Does this value indicate reasonable fit??

All of the R^2 variants are greater than 0.7, with 1 being a perfect fit, and 0.6 considered an "excellent" fit.

Q. What is your assessment of overall fit?

I agree with the assessment. There seems to be a very limited variability about the mean, and there is a clear correlation between X and Y in the analyses.

Regression with Time Series

Currency conversions from the dollar are important to the global supply chain. Consider the following data that shows the value of the Euro and the Pound to the USD from the end of October, 2018 into April of 2019.

Data: http://web.pdx.edu/~gerbing/451/Data/HW4_2.xlsx

Analyze this model for the Pound with the questions at the end of this homework. The given values of X to forecast from are Day 181, that is, April 22, 2019, and Day 188, April 29, 2019.

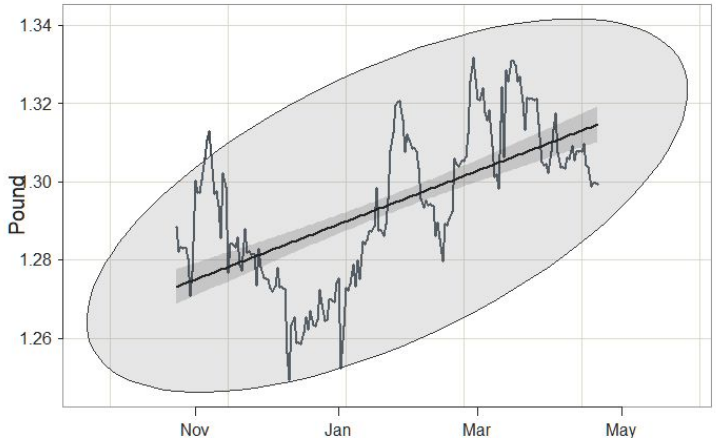
Note: My Plot() function nicely handles dates on the x-axis. The regression program does not, as it works only with the Indexes, the ordinal numbers from 1 to the last day. So the plot is better looking from the Plot() function.

Note: Current version of Regression() does not handle dates for forecasting new variables. Instead of using Day (a date) as the predictor variable, create the count from 1 to the number of data values, the number of rows in the d data frame: nrow(d).

d$DayCount <- 1:nrow(d)

One-Predictor Regression Analysis Questions:

A. Show the scatter plot of Y and X with the regression line and 0.95 confidence ellipse.



B. What is the correlation of X and Y?

The correlation, if any, is positive.

C. Does there appear to be a relationship between the two variables? Describe.

There may be a relationship between the variables. If so, it has a seasonal variation with an upward trend.

Forecasts from the Estimated Model and

D. Write the estimated regression model.

Pound = 0.00023(Day) - 2.86768 USD

E. For the given values of X to forecast, what are the forecasted Y's?

a. Show the form of the computation for the forecasted value of Y applied to the specified values of X. (No computations.)

Yhat = 0.00023(89) - 2.86768 USD; 0.00023(90) - 2.86768 USD

b. Have the regression program calculate.

```
Regression(Pound ~ DayCount, X1.new=c(89, 90))
```

c. What are the prediction intervals for these forecasts?

For: Day = 89, pi = 1.29355 +/- 0.03033 (width of 0.06066)

Day = 90, pi = 1.29378 +/- 0.03033

d. Show the calculation for the prediction intervals given the standard error of forecast (sf on the output). You can solve for the t-cutoff from the information on the output, but feel free to just use 2.
[Note: You will encounter a rounding error with only three decimal digits in the output. You can use the digits.d parameter to add more precision if you wish, but not required because this question is about understanding the concept. We let the computer do the calculating.]

sf = 0.01537, 2*2*0.01537 = .06148

e. Interpret the prediction intervals.

The prediction intervals tell me that on day 89, the Pound will be worth 1.29322 USD, with the variance between 1.26322 USD at the low end and 1.32388 USD at the high end. Similarly, on day 90 the Pound will be worth 1.29378 USD, with the variance between 1.26345 USD at the low end and 1.32411 USD at the high end.

Model Relationships

N. Specify the null hypothesis and alternative hypothesis for the hypothesis test of the slope coefficient of no relation. [Sec 8.2, #35]

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

O. Show the calculation of the t-statistic of the slope coefficient with the specific numerical values in this analysis. [Sec 8.2, #36]

```
Estimated Model
            Estimate Std Err t-value p-value Lower95% Upper95%
(Intercept) 1.27288  0.00229 554.761 0.000   1.26835  1.27741
  DayCount  0.00023  0.00002 10.564  0.000   0.00019  0.00028
```

$t_{b1} = (b_1 - 0)/s_{b1} = (0.00023 - 0)/0.00002 = 11.5$ <- *rounding error?*

P. Specify the decision rule and outcome of the hypothesis test. [Sec 8.2, #36]

p-value = 0.000 < α = 0.05, so reject the null hypothesis. Since $b_1 > 0$, I conclude that $\beta_1 > 0$.

Q. Interpret this hypothesis test (jargon free). [Sec 8.2, #35,36]

As the days progress, on average, the Pound appreciates slightly in value compared to the US Dollar.

R. What is the value to estimate with the confidence interval for the slope coefficient for the predictor variable? [Sec 8.2, #37]

The confidence interval for the slope coefficient allows for an estimate of β_1 by providing the range of possible values.

S. Show the calculation of the 95% confidence interval for the slope coefficient for the predictor variable listed in the preceding regression output with the specific numerical values in this analysis. [Sec 8.2, #38]

```
            Lower 95%   Upper 95%
(Intercept) 1.26835     1.27741
DayCount    0.00019     0.00028
```

U. Interpret this confidence interval (jargon-free). [Sec 8.2, #38

At the 95% level of confidence, for each day that goes by, on average, the Pound will likely appreciate in value from between .00019 USD to 0.00028 USD.

V. Are the results from the confidence interval and hypothesis test consistent with each other? Why or why not? Which analysis provides more information and why? (Focus on the value of the null hypothesis, 0. How does that value relate to the HT and the CI?)

These results are consistent with each other. The hypothesis test assumed that as the days go by, there would be a very slight corresponding increase in value of the Pound. Because the rate of change is so slight, the confidence interval is still the more useful but to limited effect. Granted, there may be seasonal effects that I haven't accounted for.

Model Fit

R. What is the standard error of estimate? What are some representative values of the standard error of forecast?

The standard error of estimate is 0.00002.

S. Compare the fit indices? Which one is larger? Why?

The larger fit index is R-squared (0.385 vs 0.382 adjusted or 0.372 PRESS). R-squared is larger because it is calculated as inclusive of the predicted variables. Adjusted divides the sums of the squares by the degrees of freedom to account for this.

T. What is the value of R2? Does this value indicate reasonable fit??

All of the R^2 variants are slightly greater than 0.3, meaning that this regression model provides an "adequate" fit.

U. What is your assessment of overall fit?

I agree with the assessment in terms of a long time horizon, without getting into the granularity of seasonal adjustment.