Risto B. Rushford
GSCM 451-Spr 2019
HW06

Concepts to Understand

Briefly comment on each of the following concepts, usually just several sentences.

Data Analysis

1. The two goals of regression analysis are:

   Purpose 1) Forecast the value of Y

   Purpose 2) Explain the value of Y with supporting data such as slope coefficients with all other predictor variables held constant.

2. What is $\hat{y}$?

   $\hat{y}$ is the average predicted y value for a given predictor variable x, as found using regression.

3. Graph of X with Y vs the graph of X with $\hat{y}$

   The graph of X with Y is made from pure numbers and shows the direct relationship of Y to X. The graph of X to $\hat{y}$ is the graph of the regression line which is made from the average actual measurements or instances of Y for each instance of X.

4. Meaning of the slope coefficient in $\hat{y} = b_0 + b_1X_1$

   The slope coefficient, b_1, is the amount of change in $\hat{y}$ for each unit increase of X_1.

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

   The test of the slope coefficient is to determine if there is a correlation between $\hat{y}$ and the predictor variable. A null hypothesis is that there is no relationship.

   The interpretation of this test is to write out a statement about it: A relationship (or no relationship) of $\hat{y}$ exists for each X.

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

The confidence interval determines how confident we can be that the value of $\hat{y}$ will fall within a range of numbers.

The interpretation is written as: With a __% confidence, b_1 is between __ and __.

7. Meaning of the residual variable ϵ

   ϵ is representative of the random error affecting the fit of the model. $e + R^2$ represents the total variation of the model, with R^2 being the explainable variation.

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

   $\epsilon_{m,b} = \sum(y - \hat{y})^2$, which again is the difference between the estimated model and the actual correlation between the variables, if it exists. The smaller the value for ϵ, the better the case for the model being considered a good fit.

   A good estimated model has a nearly flat OLS regression.

9. Model Fit: The standard deviation of the residuals to interpret model fit

   Residuals are a quantification of the unexplained variations of the changes of Y for each unit increase of X. The smaller the standard deviation, the better of a fit the model is to explain the data.

10. Model Fit: R-squared as a relative index of fit

    R^2 is a measure of the explainable variation of the data. R^2 will always be between 0 and 1, and represents the percentage of the variation that is predicted by the regression line. The closer R^2 is to 1, the better the regression line predicts Y.

11. How multiple regression enhances the two primary purposes of regression analysis

    Purpose 1) Forecast the value of Y. Multiple regression enhances this effort by considering additional predictor variables that could determine or affect Y's value.

    Purpose 2) Explain the value of Y with supporting data such as slope coefficients with all other predictor variables held constant. Multiple regression enhances this effort by giving a more complete breakdown of the supporting data, increasing the value of R^2 and decreasing ϵ.

12. Understand how b_1 differs in $\hat{y} = b_0 + b_1X_1$ vs. $\hat{y} = b_0 + b_1X_1 + b_2X_2$

    In $\hat{y} = b_0 + b_1X_1$, b_1 represents the entire correlation between X and $\hat{y}$. In $\hat{y} = b_0 + b_1X_1 + b_2X_2$, b_1 and b_2 represent partial slope coefficients, where each value of X_i has its own a limited but quantifiable effect, b_i on $\hat{y}$.

13. Regression models and causality

    Regression models may or may not provide an explanation for the cause of an event. It is important to understand that just because two variables show a strong correlation, that doesn't mean that one is actually causing the other. It could actually be that both variables are being affected by another unaccounted for variable that is partially determining their values. Or, it could be that there is no shared causality at all.

14. Purpose of model selection

    The more predictor variables identified that have very little to no influence on each other, the better the model will be. The lesson slides for this section uses the example of using both height and age to examine their relationships to weight. Using the two variables increases R^2 and simultaneously reduces ϵ.

15. Criteria that a potential predictor variable should satisfy before added to a model

    New information: the proposed predictor variables should have little correlation to the other predictor variables already being used.

    Relevant information: the proposed predictor variable should correlate to Y.

16. All possible subsets regression

    All possible models are compared using the specified predictor variables and then the best-fitting models are displayed showing one predictor, two predictors, and so on. This is what the scatterplot matrix in the lesson slides depict visually on slide 25.

17. Define and detect collinearity

    Predictor variables that are found to be collinear have some amount of correlation. This makes the model produce poor estimations. If collinearity is detected between predictor variables, then those variables should be removed from the model to keep it as lean and accurate as possible.

Worked Problem

Fuel mileage

HW #4 contained a problem in which miles per gallon, MPG, was regressed on cargo Weight (in thousands of lbs) to examine the impact of weight on fuel mileage. Here a second variable is added to the model: cetane. Similar to the octane number rating that is applied to gasoline to rate its ignition stability, cetane number is the rating assigned to diesel fuel to rate its combustion quality. Depending on the engine, the higher the cetane rating, the more efficiently the engine runs.

Find the data at: http://web.pdx.edu/~gerbing/451/Data/HW8_1.xlsx

We will analyze this multiple regression model, and also compare to the corresponding single-predictor regression model from HW #4.

Forecasts from the Model

a.

i. Write the estimated regression model from this analysis. [Sec 9.1, #17 shows the R output which does not write an equation but gives a table of numbers, #18 shows the estimated linear model as written from that output]

$\hat{y}_{MPG} = 2.057 + (-0.040X_{Weight}) + (0.111X_{Cetane})$

ii. Have the regression program calculate the forecasted values for X_{Weight} = 59 and X_{Cetane} = 50 and 55. [Sec 9.1, #26]

For X_{Weight} = 59 (held constant), $\hat{y}$ = 5.233 for X_{Cetane} = 50 and $\hat{y}$ =5.788 for X_{Cetane} = 55.

iii. Show the calculation for the residual for X_{Weight} = 59 and X_{Cetane} = 50.

$\hat{y}_{MPG} = 2.057 + (-0.040 * 59) + (0.111 * 50) = 5.233$

iv. What are the prediction intervals for these two forecasts?

With a 95% confidence level, we predict that when X_{Weight} = 59 and X_{Cetane} = 50, then $\hat{y}$ = between 4.618 and 5.848.

With a 95% confidence level, we predict that when X_{Weight} = 59 and X_{Cetane} = 55, then $\hat{y}$ = between 5.168 and 6.409.

v. Show the calculation for the prediction interval given the standard error of forecast (sf on the output) for $X_{Weight} = 59$ and $X_{Cetane} = 50$. You can solve for the *t*-cutoff from the information on the output, but feel free to just use 2.

Prediction Interval Lower: $\hat{y}$ - (t.cut*sf) ≈ 5.233 - (2*0.306) ≈ 4.618
Prediction Interval Upper: $\hat{y}$ - (t.cut*sf) ≈ 5.233 + (2*0.306) ≈ 5.848

vi. Interpret the prediction intervals.

For a weight of 59,000 lbs and a fuel Cetane value of 50, there is a 95% probability that the MPG will range from 4.618 to 5.848 MPG.
For a weight of 59,000 lbs and a fuel Cetane value of 55, there is a 95% probability that the MPG will range from 5.168 to 6.409 MPG.

vii. Compare the size of the predictor intervals for the one-predictor and two-predictor models. Explain the differences.

Using the one-predictor model for $X_{Weight} = 59$ produces a prediction with 95% probability that MPG will range from 4.574 to 6.295 MPG with a width of 1.722. Adding the second predictor reduced the prediction interval to a width of 1.230. Having an additional predictor variable allowed a more constrained model with more of the variance accounted for by the included variables, thus reducing random variance in the model.

Hypothesis Test: Apply to predictor variable Cetane

b. Specify the null hypothesis and alternative hypothesis for the hypothesis test of the slope coefficient of no relation. [Answer with respect to the specifics of this analysis, e.g., not Predictor 1 but the actual name of each predictor in this specific analysis; Sec 8.2, #35; Sec 9.1, #19]

Weight:

Null hypothesis: $\beta_{Weight} = 0$, as the weight population changes, there will be no corresponding change in the value of MPG.

Alt hypothesis: $\beta_{Weight} \neq 0$, as the weight population changes, on average MPG will either consistently increase or decrease.

Cetane:

Null hypothesis: $\beta_{Cetane} = 0$, as the cetane population changes, there will be no corresponding change in the value of MPG.
Alt hypothesis: $\beta_{Cetane} \neq 0$, as the cetane population changes, on average MPG will

either consistently increase or decrease.

c. Show the calculation of how many standard errors the estimated slope coefficient, b, is from the hypothesized population value, Beta=0.
[show the definition of the concept by applying the relevant numbers of this specific analysis, with or without a formula; Sec 8.2, #36]

	Estimate	Std Err	t-value	p-value	Lower 95%	Upper 95%
(Intercept)	2.057	1.126	1.827	0.074	-0.208	4.322
Weight	-0.040	0.006	-6.520	0.000	-0.053	-0.028
Cetane	0.111	0.016	6.933	0.000	0.079	0.143

$$t_{b\text{Weight}} = \frac{b_1 - 0}{s_{b\text{Weight}}} = \frac{-0.040}{0.006} = -6.520$$

The standard error for the slope coefficient for Weight is -6.520

$$t_{b\text{Cetane}} = \frac{b_1 - 0}{s_{b\text{Cetane}}} = \frac{0.111}{0.016} = 6.933$$

The standard error for the slope coefficient for Cetane is 6.933

d. *p*-value [apply the definition of the p-value to the relevant numbers in this specific analysis; Sec 9.1, #17-20]

$p_{b\text{Weight}} = 0.000$

$p_{b\text{Cetane}} = 0.000$

e. The basis for the statistical decision and the resulting statistical conclusion.
[specific with the numbers from this analysis as to the evaluation of the null hypothesis; Sec 9.1, #19,20]

For Weight, the *p*-value = 0.000 < α = 0.05, reject the null hypothesis.

For Cetane, the *p*-value = 0.000 < α = 0.05, reject the null hypothesis.

f. Interpret this hypothesis test. [applied to the relevant numbers of this specific analysis to generalize the results to the population, with *no* jargon like p-value or t-value; Sec 9.1, #19,20]

A relation of Weight and MPG is detected for truckloads with fuel of the same cetane rating.

A relation of Cetane and MPG is detected for truckloads with the same Weight.

Confidence Interval: Apply to predictor variable Cetane

g. The value the confidence interval estimates. [do *not* provide the confidence interval, which is the estimate *not* the value estimated; Sec 8.2, #37]

The confidence interval is the range of plausible values of the slope coefficient of the population, β.

h. Margin of error. [show the definition of the concept by applying the relevant numbers of this specific analysis, with or without a formula; Sec 9.1, #20]

ME_{Cetane} = t-$value_{Cetane}$ * Std Err_{Cetane} = 2 * 0.016 = 0.032

i. Confidence interval [show the definition of the concept by applying the relevant numbers of this specific analysis, with or without a formula; Sec 9.1, #20]

The 95% CI for β_{Cetane} is from 0.111-.032 = 0.079 to 0.111+.032 = 0.143

j. CI: Interpretation [no jargon, which includes the phrase “slope coefficient”, nothing about hypothesis tests];

At the 95% level of confidence, for each unit increase in Cetane, on average, the MPG likely increases somewhere from 0.079 MPG to 0.143 MPG.

k. Consistency of the Confidence Interval and Hypothesis Test [Comparison includes the specifics of the numbers for this specific analysis for both inferential results]

The results are consistent. The null hypothesis that $\beta_{Cetane} = 0$ was rejected, meaning that there is some correlation between Cetane and MPG. The confidence interval

values are all > 0, concluding that β_{Cetane} is positive and ranges from 0.079 to 0.143

l. Compare the slope coefficient of Weight for the two models.

The slope coefficient for the model considering weight alone is -0.071, with a 95% confidence that it is in the range from -0.083 to -0.059; but when evaluated along with cetane, the model values the slope coefficient for weight lower, at -0.040 with a 95% confidence that it is in the range from -0.053 to -0.028.

The multiple regression model shows that weight has a decreased overall correlation to MPG when other factors are considered.

Consider All the Predictor Variables and Model Fit

m. Evaluate fit with the standard deviation of residuals. [Sec 8.3, #8,9; Sec 9.1, #21]

```
Standard deviation of residuals:  0.300 for 47 degrees of
freedom
```

The standard deviation of residuals is 0.300 for 47 degrees of freedom. If normal, the approximate 95% range of residuals about each fitted value is 2*t-cutoff*0.300, with a 95% interval t-cutoff of 2.

95% range of variation: 1.2

The standard deviation of residuals for the single predictor model was 0.423 for 48 degrees of freedom.

Descriptive result: For a sample of data in which the model minimizes the sum of squared residuals of MPG about the fitted value for specific values of Weight, Cetane, and MPG, 95% of the fitted MPG are estimated to span a range of 1.2 MPG, which seems reasonable.

n. Evaluate fit with R-squared. [Sec 8.3, #15,18; Sec 9.1, #22]

```
R-squared:  0.871    Adjusted R-squared:  0.866    PRESS
R-squared:  0.845
```

R^2 = 0.871 for the MRM, as opposed to 0.740 for the single predictor model. So we see a reduction in training error moving from the single predictor model to the MRM.

o. Compare the standard deviation of residuals and R-squared across the two models. What do you conclude about the efficacy of including Cetane in the model? [Sec 9.1, #23, compares R^2 across two models to see if adding a predictor variable is useful]

The Standard deviation of residuals was decreased from 0.423 in the single predictor model down to 0.300 for the MRM. Meanwhile, we see an inversely corresponding increase in R^2 from 0.740 to 0.871. Seeing the increase in the level of fit along with a decrease in model error leads me to conclude that the MRM method increases the efficacy of the model and should be used when feasible.