PSY 653 Module 9 ANSWER KEY: Categorical Predictors and Nonlinear Models
Gemma Wallace & Neil Yetz

Try It Yourself Activity

We will use global data on the global Covid-19 pandemic for this activity. This data is publicly available from Johns Hopkins University and includes data on confirmed cases, deaths, countries and regions, government restriction policies (e.g., lockdown policies, number of social distancing policies, etc.), country socioeconomic status, etc.

The following website describes a package for accessing and using the data: https://joachim-gassen.github.io/2020/03/meet-tidycovid19-yet-another-covid-19-related-r-package/

The variables that we are using in this activity include:

- **income**: countries are coded as one of four income category levels: Low income, Lower middle income, Upper middle income, High income
- **lockdown:** counties are coded from 0-8, with any number above 1 representing a country that has implemented a government-enforced lockdown procedure (more stringent than social distancing orders).
- **confirmed:** the number of confirmed cases confirmed in each country per day

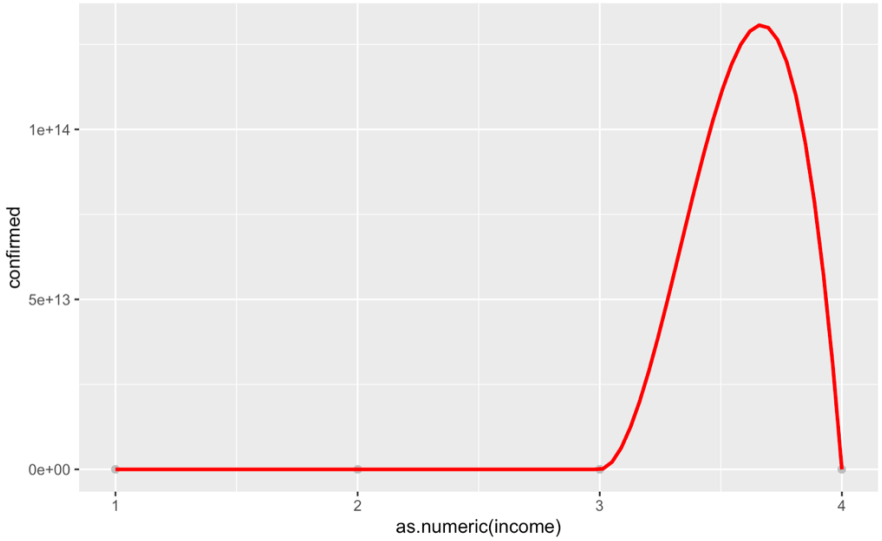
1) **Read in the datafile "covid.csv"**

```
## Read in covid data
```{r,message=FALSE}
covid <- read_csv("covid.csv")
```
```

2) **Transform the lockdown variable into a binary indicator where anything above 0 is transformed into 1. This creates a variable where any country that has implemented any form of formal lockdown is coded as 1, while countries without lockdown procedures (e.g., encouraging but not enforcing social distancing) is coded as 0.**

```
## Recode lockdown variable
```{r}
covid <- mutate(covid, lockdown = ifelse(lockdown > 0, 1, lockdown))
```
```

3) **Using dummy coding, use lockdown, income, and their interaction to predict confirmed cases (named "confirmed" in the dataframe).**

```
## Dummy code/factor income & lockdown
```{r}
covid <- mutate(covid,

               lowmid = ifelse(income == "Lower middle income", 1, 0),
               upmid  = ifelse(income == "Upper middle income", 1, 0),
               high   = ifelse(income == "High income"        , 1, 0),

               income = factor(income,
                               levels = c("Low income", "Lower middle income", "Upper middle
income", "High income"),
                               labels = c(1,2,3,4)),

               income = factor(income,
                               labels = c("Low income", "Lower middle income", "Upper middle
income", "High income"),
                               levels = c(1,2,3,4))
                               )
```
```

```
## Run dummy coding model
```{r}
cov_dum <- lm(confirmed ~  lockdown + lowmid + upmid + high + lockdown*lowmid + lockdown*upmid +
lockdown*high,
              data = covid)
ols_regress(cov_dum)
```
```

```
                         Model Summary
----------------------------------------------------------------
R                       0.264       RMSE                10392.575
R-Squared               0.069       Coef. Var             763.923
Adj. R-Squared          0.069       MSE                108005625.372
Pred R-Squared          0.068       MAE                   2103.602
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------------------
                    Sum of
                   Squares           DF        Mean Square       F          Sig.
--------------------------------------------------------------------------------
Regression    102859325608.697          7    14694189372.671     136.05    0.0000
Residual         1.378152e+12       12760      108005625.372
Total            1.481011e+12       12767
--------------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------------------
          model         Beta    Std. Error    Std. Beta      t        Sig          lower         upper
-----------------------------------------------------------------------------------------------------
    (Intercept)        5.574       230.889                0.024    0.981      -447.003       458.151
       lockdown       34.769      1054.603        0.138   0.033    0.974     -2032.412      2101.950
         lowmid       40.372       303.811        0.003   0.133    0.894      -555.143       635.886
          upmid     1566.199       293.698        0.069   5.333    0.000       990.507      2141.891
           high     1173.513       289.517        0.109   4.053    0.000       606.017      1741.009
lockdown:lowmid      266.898      1306.870        0.003   0.204    0.838     -2294.763      2828.560
 lockdown:upmid      805.222      1177.838        0.011   0.684    0.494     -1503.516      3113.961
  lockdown:high    15188.714      1185.263        0.194  12.815    0.000     12865.421     17512.007
-----------------------------------------------------------------------------------------------------
```

i) How well do these variables predict confirmed cases?

Overall, the model explains 6.9% of the variance in number of confirmed cases and is statistically significant at alpha <0.05.

ii) What information do the regression coefficients for lockdown & income give you about the differences in confirmed cases for countries of different lockdown status and income?

Interpretations of individual effects:

Intercept: The predicted number of confirmed cases when all X variables are zero, so rows with low income and no lockdown restrictions.

lockdown: This variable is involved in an interaction, so it's a simple slope. Specifically, it is the effect of having lockdown restrictions when lowmid, upmid, and high all = 0, so for rows with low income. It is the predicted difference in confirmed cases between rows with low income with versus without lockdown restrictions. The slope is positive, meaning that rows with lockdown restrictions tended to have more confirmed cases than rows without lockdown restrictions. However, the confidence interval for this effect contains 0 and is not statistically significant.

lowmid: This variable is involved in an interaction, so it is a simple slope. It is the effect of being in the lower middle income category (compared to the low income category) when lockdown = 0. It is the predicted difference in confirmed cases for rows with lower middle income without lockdown restrictions compared to rows with low income without lockdown restrictions. The slope is positive, meaning that rows with lower middle income tended to have more confirmed cases than rows with low income. However, the confidence interval for this effect contains 0 and is not statistically significant.

upmid: This variable is involved in an interaction, so it is a simple slope. It is the effect of being in the upper middle income category (compared to the low income category) when lockdown = 0. It is the predicted difference in confirmed cases for rows with upper middle income without lockdown restrictions compared to rows with low income without lockdown restrictions. The slope is positive, meaning that, on average, rows with upper middle income tended to have more confirmed cases than rows with low income. The confidence interval for this effect does not contain zero and it is statistically significant at $p<0.05$.

high: This variable is in an interaction, so it is a simple slope. It is the effect of being in the high income category (compared to the low income category) when lockdown = 0. It is the predicted difference in confirmed cases for rows with high income without lockdown restrictions compared to rows with low income without lockdown restrictions. The slope is positive, meaning that, on average, rows with high income tended to have more confirmed cases than rows with low income. The confidence interval for this effect does not contain zero and it is statistically significant at $p<0.05$.

lockdown:lowmid: The predicted differential effect of having lower middle income (compared to low income) for rows with versus without lockdown restrictions. This is not a statistically significant difference. The coefficient for this effect, 266.898, represents the difference in the effect of having lower middle income (compared to low income) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with

lockdown restrictions by taking the effect for having no lockdown restrictions (40.372) and adding the lockdown:lowmid interaction term (266.898). The effect of having low middle income compared to low income for rows with lockdown restrictions is 307.27.

lockdown:upmid: The predicted differential effect of having upper middle income compared to low income for rows with versus without lockdown restrictions. This is not a statistically significant difference. The coefficient for this effect, 805.222, represents the difference in the effect of having upper middle income (compared to low income) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with lockdown restrictions by taking the effect for having no lockdown restrictions (1566.199) and adding the lockdown:upmid interaction term (805.222). The effect of having upper middle income compared to low income for rows with lockdown restrictions is 2371.421.

lockdown:high: The predicted differential effect of having high income compared to low income for rows with versus without lockdown restrictions. This differential effect was significant at alpha <0.05. The coefficient for this effect, 15188.714, represents the difference in the effect of having high income (compared to low income) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with lockdown restrictions by taking the effect for having no lockdown restrictions (1173.513) and adding the lockdown:high interaction term (15188.714). The effect of having high income compared to low income for rows with lockdown restrictions is 16362.227.

iii) What conclusions do you reach about treatment lockdown, income, and their interaction?

Results indicate that, compared to having low income, countries with upper middle and high income had higher numbers of confirmed cases. Lockdown status did not appear to significantly impact number of confirmed cases among countries with low income. Model results also indicate that there was a significant differential effect of lockdown restrictions for countries with high income compared to those with low income; countries with high income (compared to low income) had significantly higher numbers of confirmed cases if they had implemented lockdown restrictions.

4) Redo this analysis using effect coding for income category

```
## Effect coding variables
```{r}
covid <- mutate(covid,

               lowmid = ifelse(income == "Lower middle income", 1, 0),
               upmid  = ifelse(income == "Upper middle income", 1, 0),
               high   = ifelse(income == "High income"        , 1, 0),

              lowmid = ifelse(income == "Low income", (-1), lowmid),
              upmid  = ifelse(income == "Low income", (-1), upmid ),
              high   = ifelse(income == "Low income", (-1), high  )
              )
```
```

```
## Run effect coded model
```{r}
cov_eff <- lm(confirmed ~  lockdown + lowmid + upmid + high + lockdown*lowmid + lockdown*upmid +
lockdown*high,
              data = covid)
ols_regress(cov_eff)
```
```

Model Summary

R	0.264	RMSE	10392.575
R-Squared	0.069	Coef. Var	763.923
Adj. R-Squared	0.069	MSE	108005625.372
Pred R-Squared	0.068	MAE	2103.602

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	102859325608.687	7	14694189372.670	136.05	0.0000
Residual	1.378152e+12	12760	108005625.372		
Total	1.481011e+12	12767			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	700.595	98.666		7.101	0.000	507.195	893.995
lockdown	4099.978	377.134	0.138	10.871	0.000	3360.738	4839.217
lowmid	-654.649	170.970	-0.058	-3.829	0.000	-989.775	-319.523
upmid	871.178	161.893	0.036	5.381	0.000	553.844	1188.512
high	478.492	158.082	0.093	3.027	0.002	168.627	788.357
lockdown:lowmid	-3798.310	663.393	-0.052	-5.726	0.000	-5098.661	-2497.960
lockdown:upmid	-3259.986	528.949	-0.056	-6.163	0.000	-4296.806	-2223.167
lockdown:high	11123.505	537.178	0.185	20.707	0.000	10070.556	12176.454

i) What information do the coefficients for income category variables give you about the differences in confirmed cases?

Overall, the model explains 6.9% of the variance in number of confirmed cases and is statistically significant at alpha <0.05.

<u>Interpretations of individual effects:</u>

Intercept: The grand mean of confirmed cases number among rows without lockdown restrictions (i.e., when lockdown = 0) across all four income categories.

lockdown: This variable is involved in an interaction, so it's a simple slope. It is the predicted difference in confirmed cases for rows with versus without lockdown restrictions across all income categories. This confidence interval for this effect does not contain zero and it is therefore significant at alpha <0.05. Rows with lockdown restrictions were predicted to have significantly more confirmed cases (4099.978) than rows without lockdown restrictions.

lowmid: This variable is involved in an interaction, so it is a simple slope. It is the effect of being in the lower middle income category (compared to the average of all income levels) when

lockdown = 0 (i.e., among rows without lockdown restrictions). It is the predicted difference in confirmed cases for rows with lower middle income without lockdown restrictions compared to the mean of confirmed cases for rows with no lockdown restrictions (the mean of the lower middle income rows was 654.649 lower than the mean). This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero).

upmid: This variable is involved in an interaction, so it is a simple slope. It is the effect of being in the upper middle income category (compared to the average of all income levels) when lockdown = 0 (i.e., among rows without lockdown restrictions). It is the predicted difference in confirmed cases for rows with upper middle income without lockdown restrictions compared to the mean of confirmed cases for rows with no lockdown restrictions (the mean of the upper middle income rows was 871.178 higher than the mean). This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero).

high: This variable is in an interaction, so it is a simple slope. It is the effect of being in the high income category (compared to the average of all income levels) when lockdown = 0 (i.e., among rows without lockdown restrictions). It is the predicted difference in confirmed cases for rows with high income without lockdown restrictions compared to the mean of confirmed cases for rows with no lockdown restrictions (the mean of the upper middle income rows was 478.492 higher than the mean). This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero).

lockdown:lowmid: The predicted differential effect of having lower middle income (compared to the average across income categories) for rows with versus without lockdown restrictions. This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero). The coefficient for this effect, -3798.310, represents the difference in the effect of having lower middle income (compared to the average across income categories) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with lockdown restrictions by taking the effect for having no lockdown restrictions (-654.649) and adding the lockdown:lowmid interaction term (-3798.310). The effect of having lower middle income compared to average in rows with lockdown restrictions is -4452.959.

lockdown:upmid: The predicted differential effect of having upper middle income (compared to the average across income categories) for rows with versus without lockdown restrictions. This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero). The coefficient for this effect, -3259.986, represents the difference in the effect of having upper middle income (compared to the average across income categories) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with lockdown restrictions by taking the effect for having no lockdown restrictions

(871.178) and adding the lockdown:upmid interaction term (-3259.986). The effect of having upper middle income compared to average in rows with lockdown restrictions is -4131.164.

lockdown:high: The predicted differential effect of having high income (compared to the average across income categories) for rows with versus without lockdown restrictions. This is a statistically significant difference (p-value is less than alpha and confidence interval does not contain zero). The coefficient for this effect, 11123.505, represents the difference in the effect of having high income (compared to the average across income categories) for rows with no lockdown restrictions compared to rows with lockdown restrictions. You can get the effect for rows with lockdown restrictions by taking the effect for having no lockdown restrictions (478.492) and adding the lockdown:high interaction term (11123.505). The effect of having upper middle income compared to average in rows with lockdown restrictions is 11601.997.

<u>Overall model interpretation:</u>
Results indicate that, compared to the average across all income levels, all four income levels significantly predicted number of confirmed cases. Countries with lower middle income were predicted to have significantly lower numbers of confirmed cases than average, while countries with upper middle and high income were predicted to have significantly higher numbers of confirmed cases than average. In addition, at average income level, countries that implemented lockdown restrictions were predicted to have significantly higher numbers of confirmed cases than countries that had not implemented lockdown restrictions.

Model results also indicated that there were significant differential effects of lockdown restrictions on number of confirmed cases for each income level compared to average. Specifically, countries with lower middle and upper middle income (compared to average) had significantly lower numbers of confirmed cases if they had implemented lockdown restrictions. In contrast, countries with high income (compared to average) had significantly more confirmed cases if they had implemented lockdown restrictions.

Plot the relationship between income and confirmed cases. What type of relationship do you think exists if any?

```
## Plot relationship between income and confirmed cases
```{r}
ggplot(covid, aes(x = as.numeric(income), y = confirmed)) +
  geom_point(color = "grey") +
  geom_smooth(method = "loess", se = FALSE, color = "red")
```
```

```
Removed 228 rows containing non-finite values (stat_smooth).
pseudoinverse used at 3
neighborhood radius 1
reciprocal condition number 0
Removed 228 rows containing missing values (geom_point).
```



There appears to be a strong curve to the data, with at least one inflection point. The relationship does not appear to be linear and instead looks quadratic.

5) Use the method of orthogonal polynomial coding to test hypotheses about the relationship (its form and strength) between the four different levels of income in predicting confirmed cases.
    i) Which type of relationship, if any, best fits the data for this research question?

Specify orthoganol polynomial contrasts

```{r}
covid <- mutate(covid,

              linear = ifelse(income == "Low income"         , -3, income ),
              linear = ifelse(income == "Lower middle income", -1, linear   ),
              linear = ifelse(income == "Upper middle income",  1, linear   ),
              linear = ifelse(income == "High income"        ,  3, linear   ),

              quadratic = ifelse(income == "Low income"         ,  1, income ),
              quadratic = ifelse(income == "Lower middle income", -1, quadratic),
              quadratic = ifelse(income == "Upper middle income", -1, quadratic),
              quadratic = ifelse(income == "High income"        ,  1, quadratic),

              cubic = ifelse(income == "Low income"         , -1, income),
              cubic = ifelse(income == "Lower middle income",  3, cubic   ),
              cubic = ifelse(income == "Upper middle income", -3, cubic   ),
              cubic = ifelse(income == "High income"        ,  1, cubic   )
              )

```

Run model with just the linear effect

```{r}
cov_linear <- lm(confirmed ~ linear, data = covid)
ols_regress(cov_linear)
```

```
                         Model Summary
----------------------------------------------------------------
R                       0.097       RMSE                10418.452
R-Squared               0.009       Coef. Var             809.254
Adj. R-Squared          0.009       MSE                108544139.098
Pred R-Squared          0.009       MAE                    2371.374
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------------------
                    Sum of
                   Squares           DF        Mean Square       F          Sig.
--------------------------------------------------------------------------------
Regression    14079197835.549          1    14079197835.549    129.709    0.0000
Residual          1.468168e+12     13526      108544139.098
Total             1.482247e+12     13527
--------------------------------------------------------------------------------

                            Parameter Estimates
----------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta      t        Sig       lower       upper
----------------------------------------------------------------------------------------
(Intercept)    1023.144        92.532                11.057    0.000    841.769    1204.519
     linear     480.001        42.146        0.097    11.389    0.000    397.389     562.613
----------------------------------------------------------------------------------------
```

These are the results of the liner model. The model testing the linear relationship between income level and confirmed explained 0.9% of the variance in confirmed cases, and the linear trend was statistically significant at alpha <0.001. Since we observed a curve when we plotted the data, there could be a better way to examine this relationship.

```
## Run model with the linear and quadratic effects
```{r}
cov_quadratic <- lm(confirmed ~ linear + quadratic, data = covid)
ols_regress(cov_quadratic)
```
```

```
                         Model Summary
-----------------------------------------------------------------
R                    0.099       RMSE                10417.225
R-Squared            0.010       Coef. Var             809.159
Adj. R-Squared       0.010       MSE                108518585.391
Pred R-Squared       0.009       MAE                   2288.473
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------------------
                    Sum of
                    Squares           DF       Mean Square      F          Sig.
-------------------------------------------------------------------------------
Regression    14533355870.521            2    7266677935.261    66.963    0.0000
Residual          1.467714e+12       13525     108518585.391
Total             1.482247e+12       13527
-------------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta      t        Sig       lower       upper
-----------------------------------------------------------------------------------------------
(Intercept)    1041.915        92.975                11.206    0.000    859.672    1224.158
     linear     461.230        43.128        0.094   10.694    0.000    376.692     545.768
  quadratic     187.708        91.756        0.018    2.046    0.041      7.855     367.562
-----------------------------------------------------------------------------------------------
```

These are the results of the quadratic model. The model testing the linear relationship between income level and confirmed cases explained 1% of the variance in confirmed cases, which is not meaningfully higher than the model that only tested the linear relationship. The quadratic term (named quadratic) is statistically significant at alpha <0.05, indicating that there is a curve to the relationship between X and Y (i.e., it's not linear). We need to maintain the quadratic term in the model.

```r
## Run model with the linear, quadratic, and cubic effects
```{r}
cov_cubic <- lm(confirmed ~ linear + quadratic + cubic, data = covid)
ols_regress(cov_cubic)
```
```

```
                         Model Summary
----------------------------------------------------------------
R                       0.101       RMSE                10415.157
R-Squared               0.010       Coef. Var             808.998
Adj. R-Squared          0.010       MSE                108475497.920
Pred R-Squared          0.010       MAE                     2245.172
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-------------------------------------------------------------------------------
                    Sum of
                    Squares           DF       Mean Square      F         Sig.
-------------------------------------------------------------------------------
Regression    15224589405.522            3    5074863135.174    46.783    0.0000
Residual        1.467023e+12        13524     108475497.920
Total           1.482247e+12        13527
-------------------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------------
      model        Beta    Std. Error    Std. Beta      t        Sig        lower       upper
-----------------------------------------------------------------------------------------------
(Intercept)    1044.954        92.964                11.240    0.000     862.732    1227.177
     linear     454.701        43.197        0.092   10.526    0.000     370.028     539.374
  quadratic     225.703        92.964        0.022    2.428    0.015      43.481     407.925
      cubic    -100.686        39.886       -0.022   -2.524    0.012    -178.868     -22.503
-----------------------------------------------------------------------------------------------
```

These are the results of the cubic model. We are testing the cubic term to determine if there is a second bend to the relationship between income level and confirmed. Since we have at least four levels of our categorical predictor, we can test the cubic trend using polynomial contrasts. This model explains the same amount of variance in score as the previous model that only included the linear and quadratic effects (i.e., adding the cubic effect does not increase the explanatory power of the model). The cubic term was significant, indicating that there was a second bend to the relationship. Therefore, of the three models we tested, this final model is the best fit for the data.

Note: with four levels of our categorial predictor, we can only evaluate linear, quadratic, and cubic effects (we cannot test polynomial degrees above 3).