Question 1

1.

```
. use "/Users/taryar/Downloads/prod.dta"

. summarize cost, detail

                            cost
-------------------------------------------------------------
      Percentiles      Smallest
 1%         .213          .1304
 5%        .7606           .213
10%       2.2587          .3158       Obs                 158
25%      10.1902          .4887       Sum of wgt.         158

50%      25.5454                      Mean           53.26996
                        Largest       Std. dev.      87.05933
75%      55.3624       282.2479
90%     119.3736       282.9401       Variance       7579.326
95%     240.4858       469.1852       Skewness       4.410934
99%     469.1852       737.4088       Kurtosis       29.37466
```

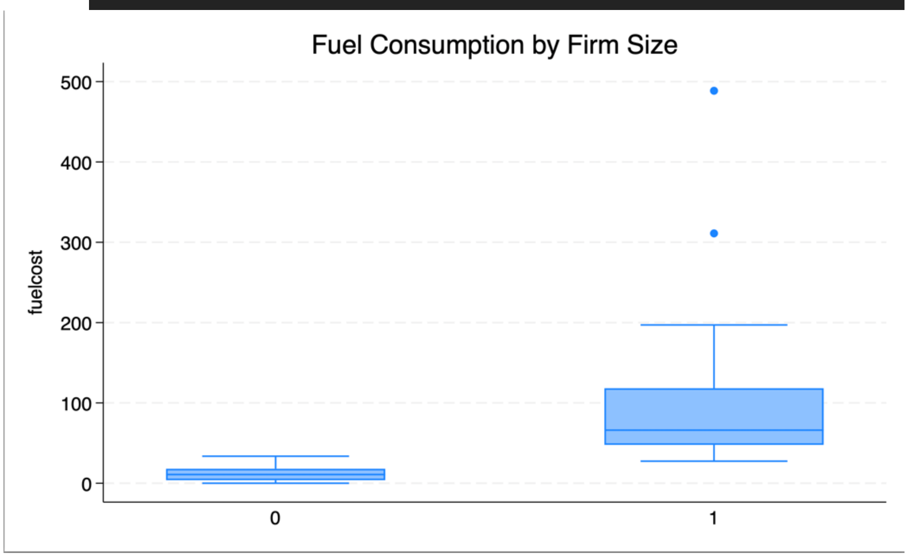
The mean of the cost of production (cost) for rms in the data is 53.26996.

The median of the cost of production (cost) for rms in the data is 25.5454.

The skewness of the cost of production for rms in the data is 4.410934, it has a positive skewness, and it shows that it would be the right tailed skewness, which means few firms are having much higher production costs than the majority. This can also indicate that larger firms or those with high output levels have disproportionately higher costs.

2.

```
. generate fuelcost = cost*sf

. bysort large: summarize fuelcost

-> large = 0

    Variable │        Obs        Mean    Std. dev.       Min
>        Max
             │
> ─────────
    fuelcost │        113    11.81878    8.801215   .0535056
>   33.66621

-> large = 1

    Variable │        Obs        Mean    Std. dev.       Min
>        Max
             │
> ─────────
    fuelcost │         45    93.11083    83.40882    27.4943
>   488.6071


. graph box fuelcost, over(large)
```



In the box plot, the median line for large firms is higher, indicating a higher median fuel consumption compared to small firms. The interquartile range (IQR) is also larger for large firms, showing greater variability in fuel consumption. The whiskers extend further for large firms, and the presence of outliers suggests that some large firms have exceptionally high fuel consumption.

3.

```
. summarize sl, detail

                             sl
-------------------------------------------------------------
      Percentiles      Smallest
 1%        .0527          .0459
 5%        .0714          .0527
10%        .0845          .0576       Obs                 158
25%        .0997          .0651       Sum of wgt.         158

50%        .1231                      Mean           .1389715
                        Largest       Std. dev.      .0547353
75%          .17          .2855
90%        .2144          .2963       Variance       .0029959
95%        .2555          .2994       Skewness         1.1039
99%        .2994          .3291       Kurtosis       3.992884

. summarize sf, detail

                             sf
-------------------------------------------------------------
      Percentiles      Smallest
 1%        .2512          .2435
 5%        .5072          .2512
10%        .5507          .3631       Obs                 158
25%        .5896          .3879       Sum of wgt.         158

50%         .645                      Mean           .6323551
                        Largest       Std. dev.      .0833241
75%         .687          .7668
90%        .7147          .7669       Variance       .0069429
95%        .7265          .7988       Skewness      -1.650708
99%        .7988          .8136       Kurtosis        8.35923
```

```
. count if large == 0 & sl > .1231 & sf >.645
  15
```

There are 15 small firms that spend more than the median cost share on both labour (sl) and fuel (sf).

4.

```
. generate size = .
(158 missing values generated)
```

```
. replace size = 1 if q <= 1961
(0 real changes made)

. replace size = 2 if q > 1961 & q <= 12542
(79 real changes made)

. replace size = 3 if q > 12542
(0 real changes made)
```

The average price of the labour for the size 1 is 7729.713.

The average price of the labour for the size 2 is 8148.077.

The average price of the labour for the size 3 is 7984.814.

5.

```
. bysort size : summarize pl

-> size = 1

    Variable |        Obs        Mean    Std. dev.       Min
>        Max
-------------+
> ---------
          pl |         40    7729.713    1317.769    5063.49
>    10963.9

-> size = 2

    Variable |        Obs        Mean    Std. dev.       Min
>        Max
-------------+
> ---------
          pl |         79    8148.077     1460.03    5879.51
>      13044

-> size = 3

    Variable |        Obs        Mean    Std. dev.       Min
>        Max
-------------+
> ---------
          pl |         39    7984.814    1343.116    5571.05
>    9914.36
```

On average, the labour cost share (sl) is smaller for large firms than for small firms, suggesting that larger firms may have more diversified or efficient cost structures.

Question 2

1.

```
. reg cost q pk pf pl

      Source |       SS           df       MS      Number of obs   =       158
-------------+----------------------------------   F(4, 153)       =    737.26
       Model |  1131262.92         4  282815.729   Prob > F        =    0.0000
    Residual |  58691.3183       153  383.603388   R-squared       =    0.9507
-------------+----------------------------------   Adj R-squared   =    0.9494
       Total |  1189954.23       157  7579.32634   Root MSE        =    19.586

------------------------------------------------------------------------------
        cost | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           q |   .0054712   .0001036    52.82   0.000     .0052666    .0056758
          pk |   .3034786   .1369908     2.22   0.028     .0328409    .5741164
          pf |   1.071896   .2063884     5.19   0.000     .6641575    1.479635
          pl |   .0032385   .0012065     2.68   0.008     .0008551     .005622
       _cons |  -84.55572   15.11173    -5.60   0.000    -114.4103   -54.70114
------------------------------------------------------------------------------
```

The **R-squared (R^2)** value is **0.950**, which means that 95% of the variation in production cost (cost) is explained by the independent variables: quantity of output (q), price of capital (pk), price of fuel (pf), and price of labor (pl). This indicates a strong model fit, as the independent variables collectively explain most of the variance in production costs.

The intercept, $\boldsymbol{\beta_0}$ **= -84.55572**, represents the estimated production cost when all independent variables (q, pk, pf, and pl) are zero. While it's unlikely for all these variables to be zero, this value serves as the baseline level of production cost in the model.

The coefficient for pf (price of fuel) is $\boldsymbol{\beta_3}$ **= 1.071896**. This means that for each one-unit increase in the price of fuel, the production cost is expected to increase by approximately 1.072 units, assuming other variables remain constant. This suggests that fuel prices have a significant impact on production costs, as higher fuel prices increase the overall cost.

2.

```
. margins, at(q=20000 pk=80 pf=40 pl=9000) level(95)

Adjusted predictions                                    Number of obs = 158
Model VCE: OLS

Expression: Linear prediction, predict()
At: q  = 20000
    pk =    80
    pf =    40
    pl =  9000

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   121.1695   2.860453    42.36   0.000     115.5184    126.8206
------------------------------------------------------------------------------
```

The predicted cost to produce the 20,000 units is 121.1695, and around the 95% confidence interval, the predicted cost is from 115.5184 to 126.820.

3.

```
. generate q2 = q^2

. generate q_pk = q * pk

.
. regress cost q pk pf pl q2 q_pk

      Source |       SS           df       MS      Number of obs   =       158
-------------+----------------------------------   F(6, 151)       =    945.30
       Model |  1159095.66         6  193182.61    Prob > F        =    0.0000
    Residual |  30858.5748       151  204.361422   R-squared       =    0.9741
-------------+----------------------------------   Adj R-squared   =    0.9730
       Total |  1189954.23       157  7579.32634   Root MSE        =    14.296

------------------------------------------------------------------------------
        cost | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           q |  -.0016007   .0007083    -2.26   0.025    -.0030001   -.0002012
          pk |  -.1619575   .1175686    -1.38   0.170    -.3942493    .0703343
          pf |   .6759468   .1547045     4.37   0.000     .3702818    .9816118
          pl |   .0036264   .0008818     4.11   0.000     .0018841    .0053687
          q2 |   1.34e-08   2.00e-09     6.67   0.000     9.41e-09    1.73e-08
        q_pk |   .0000838   9.75e-06     8.60   0.000     .0000645     .000103
       _cons |  -36.44648   12.21131    -2.98   0.003    -60.57357   -12.3194
------------------------------------------------------------------------------
```

a) the marginal effect of the quantity of output(q) on the cost of production(cost) is -.0016007 and this means that every 1 unit of output increase will decrease the amount of cost by -0.0016007.

b) the coefficient of q_pk (the variable q*pk) is 0.0000838. this indicates that as the price of capital(pk) increases, the effect of increasing output(q) on production cost(cost) becomes slightly larger. So, when capital prices are higher, producing more output leads to a slightly higher increase in costs, perhaps due to greater reliance on capital-intensive production processes as output scales up.

```
. margins, dydx(q) at(q=9000 pk=60)

Average marginal effects                            Number of obs = 158
Model VCE: OLS

Expression: Linear prediction, predict()
dy/dx wrt:  q
At: q  = 9000
    pk =   60

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           q |  -.0016007   .0007083    -2.26   0.025    -.0030001   -.0002012
------------------------------------------------------------------------------
```

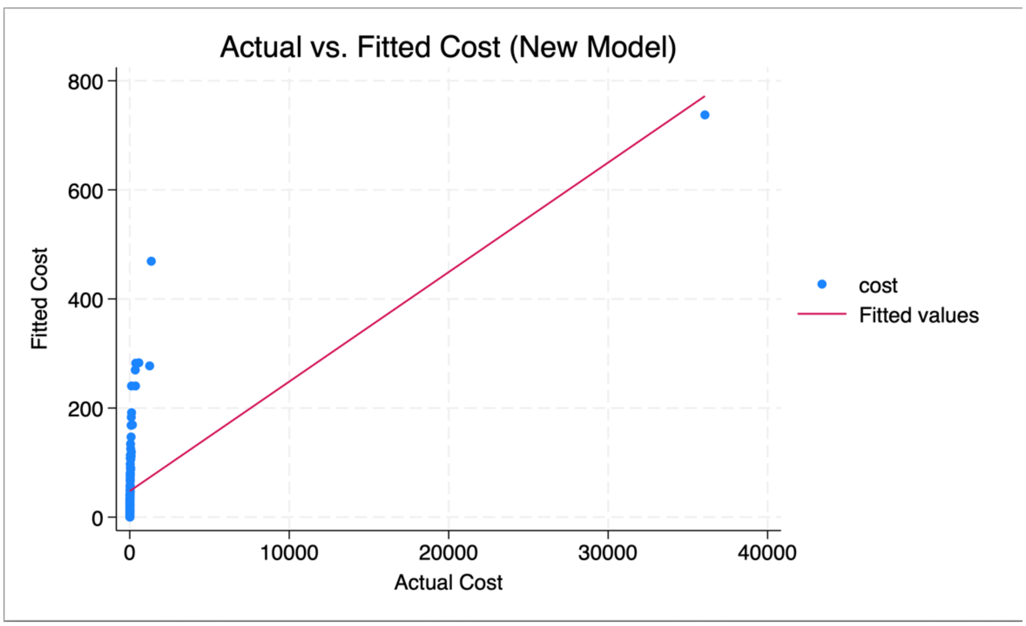
c) The adjusted R squared for the equation 1 is 0.949. and the adjusted r squared for the new regression equation is 0.973. The adjusted r squared for the new regression increased but not significantly. It means adding the additional variables q2 and q_pk doesn’t add much value to explain the variability of the cost.

4.

```
. gen log_cost = log(cost)

. regress log_cost q pk pf pl

      Source |       SS           df       MS      Number of obs   =       158
-------------+----------------------------------   F(4, 153)       =     39.44
       Model |  189.697058         4  47.4242645   Prob > F        =    0.0000
    Residual |  183.967507       153    1.202402   R-squared       =    0.5077
-------------+----------------------------------   Adj R-squared   =    0.4948
       Total |  373.664565       157  2.38002907   Root MSE        =    1.0965

------------------------------------------------------------------------------
    log_cost | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           q |   .0000696   5.80e-06    12.01   0.000     .0000582    .0000811
          pk |   .0110828   .0076696     1.45   0.150    -.0040693    .0262349
          pf |   -.001812    .011555    -0.16   0.876    -.0246399    .0210159
          pl |    .000149   .0000675     2.21   0.029     .0000156    .0002825
       _cons |   .4144068   .8460532     0.49   0.625    -1.257048    2.085861
------------------------------------------------------------------------------
```

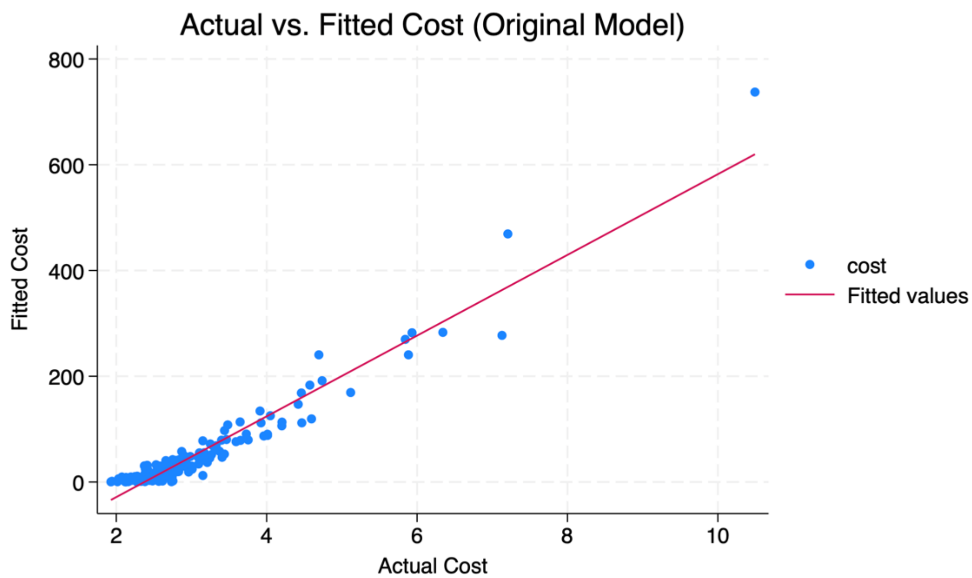
```
. predict fitted_log_cost, xb

. gen fitted_cost_new = exp(fitted_log_cost)
```

```
. twoway (scatter cost fitted_cost_new) (lfit cost fitted_cost_new),title("Actual vs. Fitted Cost (New Model)")xtitle("Actual Cost") ytitle
> ("Fitted Cost")
```



```
. predict fitted_cost_1, xb
```

```
. twoway (scatter cost fitted_cost_1) (lfit cost fitted_cost_1),title("Actual vs. Fitted Cost (Original Model)")xtitle("Actual Cost") ytitl
> e("Fitted Cost")
```



The scatter plot of the original model shows a strong alignment between the actual cost and fitted values. The points follow the fitted line closely, indicating a good fit between predicted and actual values.

The scatter plot for the new model shows a significant discrepancy between actual costs and fitted values after re-transforming. The actual costs are spread out, while the fitted values appear concentrated and follow a different scale, which suggests that the log transformation is not aligning well when converted back to the original scale.

Based on the scatter plot comparison, the original model provides a better fit for predicting cost than the new model with log(cost) as the dependent variable. The fitted values from the original model align closely with the actual cost values, while the new model's fitted values, when transformed back to the cost scale, do not match the actual costs well. Thus, the original model is the better choice for predicting cost.

5. The relationship between the cost of production (cost) and the price of fuel (pf) observed in the regression analysis is correlational, not causal. This is because the data is observational and lacks the experimental or quasi-experimental structure needed to isolate the effect of pf on cost. Unobserved factors may be affecting both pf and cost, resulting in a correlation that does not imply causation. Establishing causality would require a setup in which changes in pf occur independently of other determinants of cost, such as through an instrumental variable or experimental design. In the absence of such a framework, the coefficient on pfrepresents an association, capturing the average relationship between pf and cost in the data, but it does not imply that changes in pf will cause changes in cost.

Commands:

Question 1

summarize cost, detail

generate fuelcost = cost*sf

bysort large: summarize fuelcost

graph box fuelcost, over(large) title("Fuel Consumption by Firm Size")ytitle(fuelcost)

summarize sl, detail

summarize sf, detail

count if large == 0 & sl > .1231 & sf >.645

generate size = .

replace size = 1 if q <= 1961

replace size = 2 if q > 1961 & q <= 12542

replace size = 3 if q > 12542

bysort size : summarize pl

Question 2

reg cost q pk pf pl

margins, at(q=20000 pk=80 pf=40 pl=9000) level(95)

generate q2 = q^2

generate q_pk = q * pk

regress cost q pk pf pl q2 q_pk

margins, dydx(q) at(q=9000 pk=60)

gen log_cost = log(cost) regress log_cost q pk pf pl

```
predict fitted_log_cost, xb
gen fitted_cost_new = exp(fitted_log_cost)

twoway (scatter cost fitted_cost_new) (lfit cost fitted_cost_new),title("Actual vs. Fitted Cost (New Model)")xtitle("Actual Cost")ytitle("Fitted Cost")
predict fitted_cost_1, xb
twoway (scatter cost fitted_cost_1) (lfit cost fitted_cost_1),title("Actual vs. Fitted Cost (Original Model)") xtitle("Actual Cost") ytitle("Fitted Cost")
```