1ZM31  Multivariate Data Analysis(2017/2018)

## Group Assignment 2

## Report

Multivariate Data Analysis

Dr. Sarah Gelper

# Multiple and Logistics Regression Analysis

| | |
|---|---|
| Group | 23 |

| | |
|---|---|
| A.W.H (Sander) Berkers | 0658745 |
| Bhoomica Mysore Nataraja | 1282832 |
| D.F (Diego Fernando) Barreto Trujillo | 1280341 |
| Nitish Singh | 1283901 |

| | |
|---|---|
| Date | <05/10/2017> |

# Modelling R&D Cooperation

## 1. What is the proportion of enterprises that co-operate on innovation activities with other enterprises or institutions?

27.08% of the companies in the MIP sample cooperate with others on innovation activities.

## 2. Build a model with as dependent variable whether or not the enterprise co-operates on innovations with other enterprises or institutions. Make sure that the model accounts for company size and the factor score obtained from the factor analysis on "innovations with environmental benefits" (for now, only include these variables). Write down the model equation.

In this case, the dependent variable is non-metric nominal value (1 means YES; 0 means NO) so the model must be a logistic regression (Hair, et.al, 2014). Variables "Size" (Ratio) and "InnovEnvironBenefits" (EFA Result) are to be included. The variable "InnovEnvironBenefits" is a result from an Explanatory Factor Analysis, the interpretation is not easy if it is not standardized first in a multiple linear regression, but in a logistic regression it is not necessary. Finally, there is no error in the model ($\varepsilon_i$) since the model is a logistic regression (Hair, et.al, 2014).

Thus, the first proposed model equation is:

$$CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

A logistic regression does not need a linear regression between independent and dependent variables and there is no need to look for normality in the independent variables. However, a logarithm or quadratic transformation was going to be evaluated to know if the model would improve or not; for this, added value plots were used for the first model.

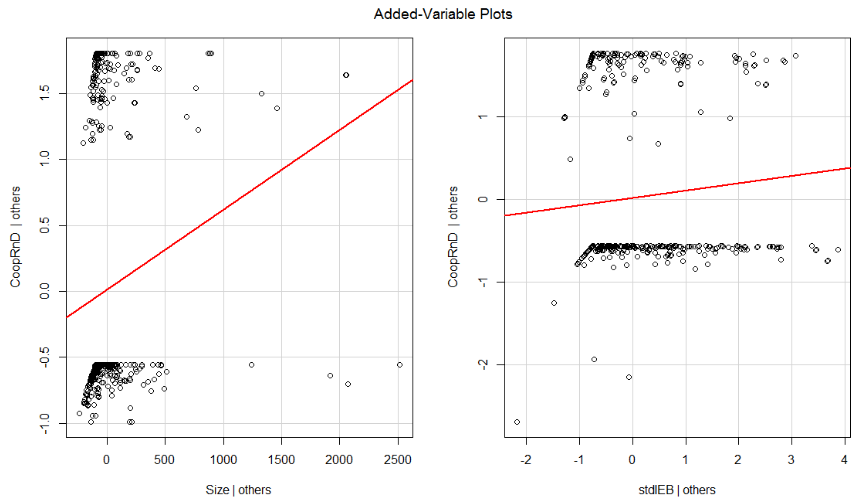


*Figure 1. Added Variable Plots for first model*

According to Figure 1, it seems to be good to log transform "Size" variable since the point are right-skewed and this leads to a second model:

$$CoopRnD_i = \alpha + \beta_1 \ln(Size_i) + \beta_2 InnovEnvironBenefits_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

The second model seems logical because it will spread the added variable plot for the size as shown in Figure 2.

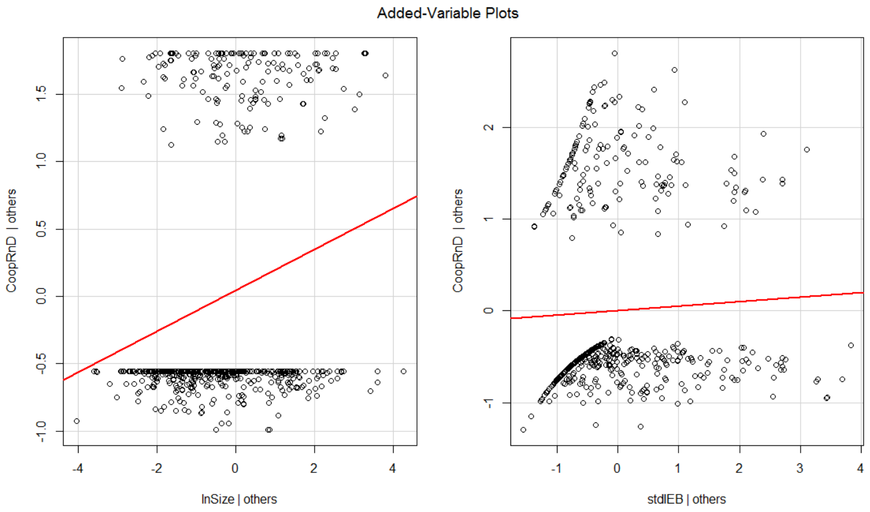


*Figure 2. Added Variable Plots for second model where logarithm is applied to Size*

The added variable plots are more spread for the second model, but when fitting the two models the results of the second model shows that $\beta_2$ is not significant (p-value is greater than 0.05 proving null hypothesis of $\beta_2 = 0$) as shown in Table 1. The second model was dropped off because the question indicates: "*Make sure that the model accounts for company size and the factor score obtained from the factor analysis on innovations with environmental benefits*", although the AIC and HIT rate were better for the model 2 than the model 1.

| | Model 1 | | | | | Model 2 | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Exp (Estimate) | p-value | | | Estimate | Exp (Estimate) | p-value |
| Intercept | -1.1317 | 0.3225 | <0.001 | | Intercept | -2.2801 | 0.1023 | <0.001 |
| Size | 0.0011 | 1.0011 | 0.0028 | | ln(Size) | 0.3444 | 1.4111 | 2.95E-07 |
| InnovEnvironBenefits | 0.2220 | 1.2486 | 0.0203 | | InnovEnvironBenefits | 0.1199 | 1.1274 | **0.2300** |
| | | | | | | | | |
| AIC | 660.87 | | | | AIC | 634.65 | | |
| HIT Rate | 0.7292 | | | | HIT Rate | 0.7361 | | |

*Table 1. Comparison between models 1 and 2, leading to drop model 2 because beta2 was not significant*

Quadratic transformation is not an option in this case because the data points of added variable plots do not show that behavior. Since the logarithm transformation of size seems logical and Innovation Environment variable should be in the model, it leads to a third model that includes logarithm of size and look for a transformation to have a significant $\beta_2$:

$$CoopRnD_i = \alpha + \beta_1 \ln(Size_i) + \beta_2 \ln(InnovEnvironBenefits_i)$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

. The added-variable plots for the third model are the following and do not visually change.

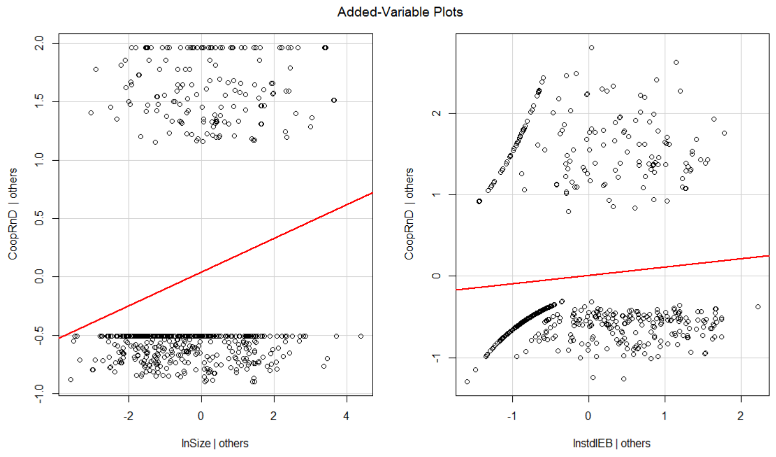
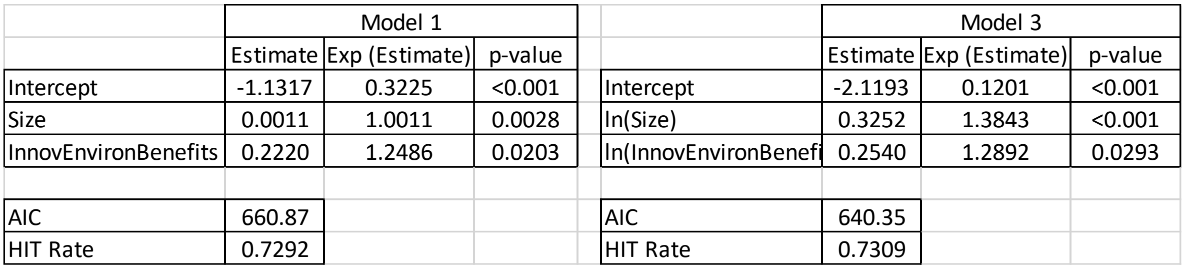


*Figure 3. Added-variable plots for third model*

A comparison was made between model 1 and model 3 to look which one was better; the results of the model fit are shown on Table 2.



| | Model 1 | | | | | Model 3 | | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | Exp (Estimate) | p-value | | | Estimate | Exp (Estimate) | p-value |
| Intercept | -1.1317 | 0.3225 | <0.001 | | Intercept | -2.1193 | 0.1201 | <0.001 |
| Size | 0.0011 | 1.0011 | 0.0028 | | ln(Size) | 0.3252 | 1.3843 | <0.001 |
| InnovEnvironBenefits | 0.2220 | 1.2486 | 0.0203 | | ln(InnovEnvironBenefi | 0.2540 | 1.2892 | 0.0293 |
| | | | | | | | | |
| AIC | 660.87 | | | | AIC | 640.35 | | |
| HIT Rate | 0.7292 | | | | HIT Rate | 0.7309 | | |

*Table 2. Comparison between models 1 and 3, suggesting that model three is better according to AIC and HIT rate*

Looking at the comparison between model 1 (no log transformation) and 3 (log transformation on both variables) it is known that $\beta_1$ and $\beta_2$ are significant (different from 0) because their p-value is less than 0.05. Also, there is no multicollinearity as VIF are less than 2 in both models.

**Model 3 seems to be a more appropriate model to use because AIC and HIT rate are better. However, interpretation of ln(size) and ln(InnovEnvironBenefits) is harder and the values of AIC and Hit Rate are really close to the ones of model 1 (the difference is close to 3%). For these reasons, the selected model to work with is Model 1:**

$$CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1-\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

## 3. Fit your model and report the estimation output.

Considering the previous model, it was fitted and the estimation output is in Table 3.

| DV = CoopRnD | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | z value | p-value | Exp (Estimate) |
| Intercept | -1.1317 | 0.1054 | -10.7440 | < 0.001 | 0.3225 |
| Size | 0.0011 | 0.0004 | 2.9870 | 0.0028 | 1.0011 |
| InnovEnvironBenefits | 0.2220 | 0.0957 | 2.3210 | 0.0203 | 1.2486 |
| AIC | 660.87 | | | | |

*Table 3. Estimation output of the fitted model*

As established before, both, Size and InnovEnvironBenefits, variables are significant to the logit model (p-values are less than 0.05) with estimates that must be recalculated for a better an easier interpretation. This leads to have the following model:

$$CoopRnD_i = -1.1317 + 0.0011 \times Size_i + 0.2220 \times InnovEnvironBenefits_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

## 4. Based on your results from question 2, interpret the regression coefficient of company size.

According to results on Table 3, holding the innovations with environmental benefits constant, for every additional employee (on average) in the company, the odds of being in favor of cooperating with other companies in innovation increase by a factor of 1.001. The coefficient of “Size” is significant to cooperation with other companies because the p value is 0.0028, less than 0.05, rejecting the null hypothesis of $\beta_1 = 0$

## 5. Based on your results from question 2, interpret the regression coefficient of the factor scores of the factor analysis on “innovations with environmental benefits".

According to Table 3, holding constant the average of number of employees in the company, for every unit increased in the standardized factor score in Innovation Environment Benefits, the odds of being in favor of cooperating with other companies in innovation increase by a factor of 1.249. The coefficient of “InnovEnvironment” is significant to cooperation with other companies because the p value is 0.0203, less than 0.05, rejecting the null hypothesis of $\beta_2 = 0$

## 6. Report and interpret the hit-rate, the true-positive rate and the true-negative rate.

To compute the HIT rate in this case with dependent nominal variable (1 means YES and 0 means NO) for the prediction of the model, the fitted values with value equal or greater than 0.5 ($\geq 0.5$), will be rounded to 1 to show that the company cooperates with others in innovation and viceversa.

The HIT rate is 0.7292; this means that for 72.92% of the companies in the MIP data set, the model predicted in the correct way whether the enterprises cooperates with other companies in innovation or not.

The true-positive rate was 0.0385; this means that only 3.85% of the enterprises that cooperate with other enterprises in innovation are identified as such. Thus, it is very difficult to predict if the company will cooperate with others for innovation with this model.

The true-negative rate was 0.9857; this means that 98.57% of the enterprises that do not cooperate with other enterprises in innovation are identified as such. Thus, with this model is easy to predict which companies will not cooperate for innovation compared to the ones that cooperate.

## 7. If you were to include one more explanatory variable from the data set, which would it be? Explain your choice and write down the model equation.

The variable that should be included in the model is "ExternalRnD" that represents the expenditures in million Euros on innovation activities contracted out to companies or institutes. To get to the answer for the more appropriate variable to include, each variable was added to the model that contained "Size" and "InnovEnvironBenefits" and the HIT Rates and AIC were obtained to make the comparison. The models evaluated were:

$$(7.1) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 Turnover_i$$

$$(7.2) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 InhouseRnD_i$$

$$(7.3) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 ExternalRnD_i$$

$$(7.4) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 Machinery_i$$

$$(7.5) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 ExternalKnow_i$$

$$(7.6) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 MarketSpend_i$$

$$(7.7) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 ProcessInnov_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

With the models defined, a direct comparison was made and the results among different indicators are summarized in the following table.

| | Variable to Add | | | | | | |
|---|---|---|---|---|---|---|---|
| P-values | Turnover | InhouseRnD | ExternalRnD | Machinery | ExternalKnow | MarketSpend | ProcessInnov |
| Intercept | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 | 2.00E-16 |
| Beta 1 (Size) | 8.38E-01 | 2.51E-01 | 6.41E-01 | 4.58E-02 | 1.10E-02 | 2.17E-01 | 3.98E-02 |
| Beta 2 (IEB) | 2.22E-02 | 3.05E-02 | 1.86E-01 | 7.12E-02 | 2.69E-02 | 1.37E-02 | 7.52E-01 |
| Beta 3 (Variable to Add) | 7.04E-02 | 3.61E-05 | 8.82E-10 | 1.36E-02 | 2.33E-02 | 7.06E-02 | 2.48E-12 |
| | | | | | | | |
| AIC | 659.42 | 633.89 | 581.27 | 649.12 | 657.69 | 658.79 | 610.39 |
| | | | | | | | |
| HIT Rate | 0.731 | 0.759 | 0.786 | 0.729 | 0.726 | 0.733 | 0.738 |
| True Positive | 0.058 | 0.160 | 0.282 | 0.071 | 0.058 | 0.051 | 0.090 |
| True Negative | 0.981 | 0.981 | 0.974 | 0.974 | 0.974 | 0.986 | 0.979 |

*Table 4. Results on p-values, AIC and HIT rates of the different models with 3 variables (keeping size and InnovEnvironBenefits)*

According to the table 4, the model that has a better AIC is the one with "ExternalRnD" with 581.27. The model that predicts better according to the data set is also the one with "ExternalRnD" with a HIT Rate of 78.6% showing that the model predicted in the correct way whether the enterprises cooperates with other companies in innovation or not; moreover, compared to the other models, it is by far the best one predicting true positive cases (although it is only 28.2%, it is almost 5 times the percentage of the other models) and also keeping a good prediction for true negatives. It is logical to have such a variable since companies that invest in innovation probably want to collaborate with others.

No transformation was made to the "ExternalRnD" since the question 2 showed that it is hard to interpret the outcome of the result. To sum, the selected model was:

$$(7.3) CoopRnD_i = \alpha + \beta_1 Size_i + \beta_2 InnovEnvironBenefits_i + \beta_3 ExternalRnD_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

## 8. Fit the model specified in question 7 and report the estimation output.

First of all, looking at the data, the greatest value of "ExternalRnD" is almost 880,000 Euros so the analysis cannot be made in millions. This is because the interpretation will be: "an additional million of euros spent in externalRnD will increase the odds in cooperating with other companies by a factor of X" and X will tend to infinite because it surpass the maximum value spent in externalRnD. Considering this, the model was fitted with "ExternalRnD" in thousands euros. The estimation output is in Table 5.

| DV = CoopRnD | | | | | |
|---|---|---|---|---|---|
| | Estimate | Std. Error | z value | p-value | Exp (Estimate) |
| Intercept | -1.4119 | 0.1187 | -11.8920 | <0.001 | 0.2437 |
| Size | 0.0002 | 0.0004 | 0.4670 | 0.6410 | 1.0002 |
| InnovEnvironBenefits | 0.1414 | 0.1070 | 1.3220 | 0.1860 | 1.1519 |
| ExternalRnD | 0.0274 | 0.0045 | 6.1300 | <0.001 | 1.0278 |
| AIC | 581.27 | | | | |

*Table 5. Estimation output of the fitted model*

In this case, "Size" and "InnovEnvironment" are not significant to the logit model (p-values are greater than 0.05) but ExternalRnD is significant (p-value is less than 0.05) with estimates that must be recalculated for a better an easier interpretation. This leads to the following model equation:

$$(7.3) CoopRnD_i = -1.4119 + 0.0002 \times Size_i + 0.1414 \times InnovEnvironBenefits_i + 0.0274 \times ExternalRnD_i$$

$$Where \qquad CoopRnD_i = \ln\left(\frac{\Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}{1 - \Pr(Company\ i\ cooperates\ with\ others\ on\ innovation)}\right)$$

## 9. Compare the AIC of the model in question 3 to the AIC of the model in question 8. What do you conclude?

According to Table 3 and Table 5, the AIC for the model with variables “Size” and “InnovEnvironBenefits” is 660.87 while the second model with the variables “Size”, “InnovEnvironBenefits” and “ExternalRnD” has an AIC of 581.27. This means that the model from the question 8 has a better fit and is not penalized for adding other variable since it increases the likelihood.

## 10. Based on your results from question 8, interpret the regression coefficient of company size. Does your answer differ from your answer to question 4? If so, why is the interpretation different?

According to results on Table 5, holding constant the innovations with environmental benefits and external RnD, for every additional employee (on average) in the company, the odds of being in favor of cooperating with other companies in innovation increase by a factor of 1.0002 if the variable is significant and here is not the case; the coefficient of “Size” is not significant to cooperation with other companies because the p value is 0.641, more than 0.05, not rejecting the null hypothesis of $\beta_1 = 0$. In conclusion answer for question 4 and 8 differ not only because the factor changes (from 1.001 in Q4 to 1.002 in Q8) but because in Q8 the “Size” variable is not significant.

## 11. Based on your results from question 8, interpret the regression coefficient of the factor scores of the factor analysis on “innovations with environmental benefits". Does your answer differ from your answer to question 4? If so, why is the interpretation different?

According to Table 3, holding constant the average of number of employees in the company, for every unit increased in the standardized factor score in Innovation Environment Benefits, the odds of being in favor of cooperating with other companies in innovation increase by a factor of 1.249.

According to results on Table 5, holding constant the average of number of employees in the company and external RnD, for every unit increased in the standardized factor score in Innovation Environment Benefits, the odds of being in favor of cooperating with other companies in innovation increase by a factor of 1.143 if the variable is significant and here is not the case; the coefficient of "InnovEnvironBenefits" is not significant to cooperation with other companies because the p value is 0.186, more than 0.05, not rejecting the null hypothesis of $\beta_2 = 0$. In conclusion answer for question 4 and 8 differ not only because the factor changes (from 1.249 in Q4 to 1.143 in Q8) but because in Q8 the "InnovEnvironBenefits" variable is not significant.

## 12. Report and interpret the hit-rate, the true-positive rate and the true-negative rate. Compare these numbers to your answer to question 6. What do you conclude?

To compute the HIT rate in this case with dependent nominal variable (1 means YES and 0 means NO) for the prediction of the model with value equal or greater than 0.5 ($\geq 0.5$), the value will be rounded to 1 to show that the company cooperates with others in innovation and viceversa. Table 4 summarize the values of HIT Rate, true positives and true negatives.

The HIT rate is 0.7865; this means that for 78.65% of the companies in the MIP data set, the model predicted in the correct way whether the enterprises cooperates with other companies in innovation or not. This is almost 6 pp more than the Q6 (78.65% on Q12 vs 72.92% Q8) which shows that the last model predicts in general better whether the enterprises cooperates with other companies in innovation or not.

The true-positive rate was 0.2821; this means that only 28.21% of the enterprises that cooperate with other enterprises in innovation are identified as such. Thus, it is not easy to predict if the company will cooperate with others for innovation with this model, but it is easier than with the model of Q6 (28.21% on Q12 vs 3.85% Q8) which shows that the last model predicts better if the company will certainly cooperate with other in innovation.

The true-negative rate was 0.9738; this means that 97.38% of the enterprises that do not cooperate with other enterprises in innovation are identified as such. Thus, with this model is easy to predict which companies will not cooperate for innovation compared to the ones that cooperate, but it is easier with the model of Q6 (97.38% on Q12 vs 98.57% Q8). Although the model of Q6 was better to the one in Q12 in true-negative rate, it is only a bit better and probably not significantly different. Both models are good to predict which companies will not cooperate with others in innovation.

## Modelling Firm Turnover

In this part of the assignment, you are interested in the following two questions:

Q1: Do firms that co-operate on their innovations with external partners have, on average, a higher turnover {controlling for other factors that might affect turnover?

Q2: What is the elasticity of turnover with respect to in-house R&D spending {controlling for other factors that might affect turnover?

For questions 13 to 18, use only the variables Size, CoopRnD and the variables related to innovation expenditures in the model (i.e. InhouseRnD, ExternalRnD, Machinery and ExternalKnow).

### 13. Write down the model specification that allows you to answer the above two questions. Explain all your modelling choices (e.g., if you apply any data transformations, explain why).

The model was initially regressed with all the innovation relevant independent variables available for modelling – CoopRnd, Size, InhouseRnD, ExternalRnD, Machinery and ExternalKnow and the dependent variable was Turnover.

Our initial analysis with the above mentioned variables produced the following result:-

**Model 2.1 :** $Turnover_{\mathrm{i}} = \alpha + \beta_1 Size_{\mathrm{i}} + \beta_2 CoopRnD_{\mathrm{i}} + \beta_3 InhouseRnD_{\mathrm{i}} + \beta_4 ExternalRnD_{\mathrm{i}} + \beta_5 Machinery_{\mathrm{i}} + \beta_6 ExternalKnow_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$

| **Coefficients** | **Estimate** | **Std. Error** | **t value** | **Pr(>\|t\|)** |
|---|---|---|---|---|
| (Intercept) | 1.55E+02 | 1.46E+03 | 0.106 | 0.9158 |
| Size | 1.55E+02 | 5.09E+00 | 30.364 | <2.00e-16 |
| CoopRnD | -3.53E+03 | 2.89E+03 | -1.224 | 0.2215 |
| InhouseRnD | 4.96E+03 | 1.46E+03 | 3.405 | 7.07e-4 |
| ExternalRnD | 1.11E+05 | 1.79E+04 | 6.214 | 9.98e-10 |
| Machinery | 5.15E+03 | 1.12E+03 | 4.601 | 5.18e-06 |
| ExternalKnow | -2.03E+05 | 1.14E+05 | -1.786 | 0.0746 |
| *Residual Standard Error: 28.5 on 569 degrees of Freedom Multiple R-squared: 0.7508,* ***Adjusted R-squared: 0.7482*** | | | | |

*Table 6: Ordinary Least Squares Regression Output from Model 2.1*

As a means to reduce the skewness of the data distribution, the new model contains log transformations (natural base e) of monetary variables which were diminishing in nature, such as – InhouseRnD, Machinery, ExternalRnD, ExternalKnow and Turnover. In addition, Size was log-transformed (natural base e) to achieve better spread of values and to attain normality, the results of this model can be found in table 7. The model thus obtained had better adjusted R-squared values (0.8626) in comparison to the previous model (0.7508) and had small p-values indicating that all the variables were significant in predicting the outcome variable, apart from ExternalKnow, the p-value of this variable is 0.92602, which is far greater than 0.05. However, the significance of the coefficient is not a valid reason to drop it.

**Model 2.2:**

$$\log(Turnover_{\mathrm{i}}) = \alpha + \beta_1 \log(Size_{\mathrm{i}}) + \beta_2 CoopRnD_{\mathrm{i}} + \beta_3 \log(InhouseRnD_{\mathrm{i}}) \\ + \beta_4 \log(ExternalRnD_{\mathrm{i}}) + \beta_5 \log(Machinery_i) + \beta_6 \log(ExternalKnow_i) + \varepsilon_{\mathrm{i}}$$

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.0159 | 0.0577 | -17.593 | < 2e-16 |
| log(Size) | 0.8003 | 0.0162 | 49.314 | < 2e-16 |
| CoopRnD | -0.1215 | 0.0540 | -2.251 | 0.0248 |
| log(InhouseRnD) | 0.2369 | 0.0774 | 3.060 | 0.0023 |
| log(ExternalRnD) | 1.2200 | 0.4034 | 3.024 | 0.0026 |
| log(Machinery) | 0.2505 | 0.0826 | 3.031 | 0.0026 |
| log(ExternalKnow) | 0.1996 | 2.1489 | 0.093 | 0.9260 |
| *Residual standard error: 0.5183 on 569 degrees of freedom* | | | | |
| *Multiple R-squared: 0.8626,* | | ***Adjusted R-squared: 0.8612*** | | |

*Table 7: Ordinary Least Squares Regression Output from Model 2.2*

## 14. Estimate the model you suggested by ordinary least squares and report the results.

The regression model equation (Model 2.2) after estimation is:

$$\log(Turnover_{\mathrm{i}}) = \text{-}1.0159 + \mathit{0.8003}\, x \log(Size_{\mathrm{i}}) - 0.1215\, x\, CoopRnD_{\mathrm{i}} \\ + 0.2369\, x \log(InhouseRnD_{\mathrm{i}}) + 1.2200\, x \log(ExternalRnD_{\mathrm{i}}) \\ + \mathit{0.2505}\, x \log(Machinery_i) + \mathit{0.1996}\, x \log(ExternalKnow_i) + \varepsilon_{\mathrm{i}}$$

The result of the Ordinary Least Squares is reported in the Table 8.

| Coefficients | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -1.0159 | 0.0577 | -17.593 | < 2e-16 |
| log(Size) | 0.8003 | 0.0162 | 49.314 | < 2e-16 |
| CoopRnD | -0.1215 | 0.0540 | -2.251 | 0.0248 |
| log(InhouseRnD) | 0.2369 | 0.0774 | 3.060 | 0.0023 |
| log(ExternalRnD) | 1.2200 | 0.4034 | 3.024 | 0.0026 |
| log(Machinery) | 0.2505 | 0.0826 | 3.031 | 0.0026 |
| log(ExternalKnow) | 0.1996 | 2.1489 | 0.093 | 0.9260 |
| *Residual standard error: 0.5183 on 569 degrees of freedom*<br>*Multiple R-squared: 0.8626,* ***Adjusted R-squared: 0.8612***<br>*F-statistic: 595.5 on 6 and 569 DF, p-value: < 2.2e-16* | | | | |

*Table 8: Ordinary Least Squares Regression Output of Model 2.2*

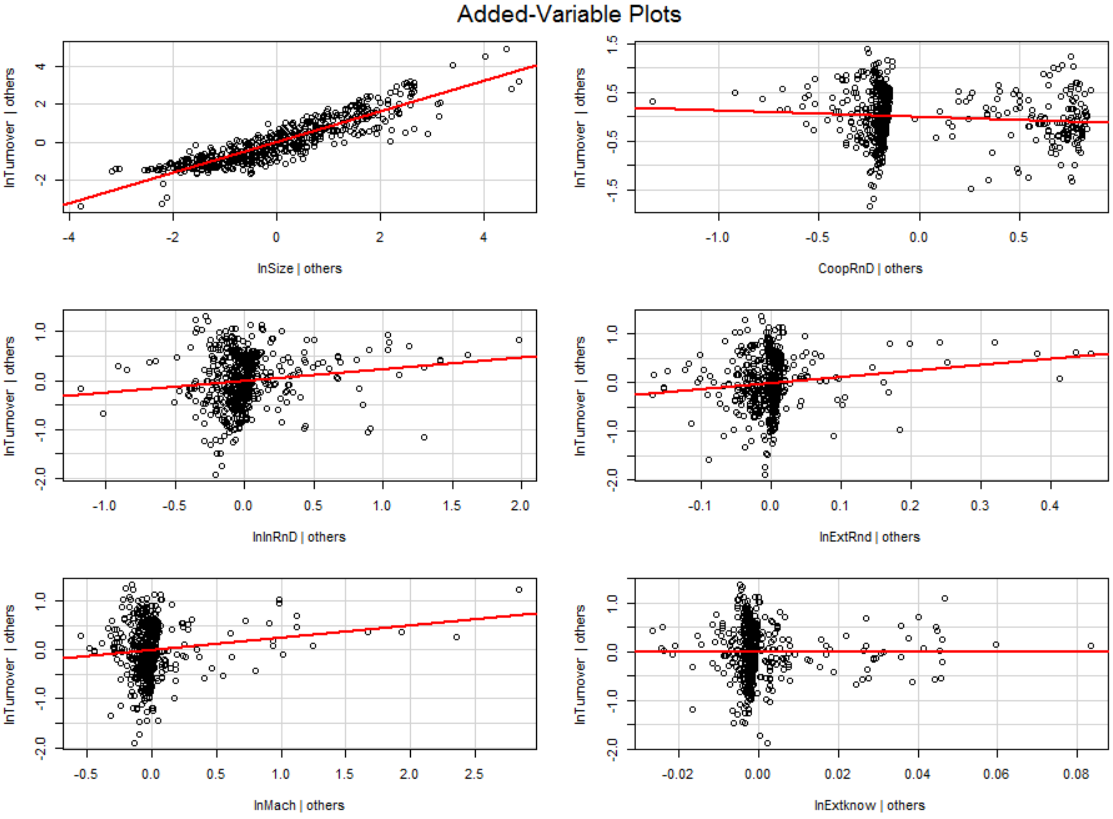


*Figure 4: Added Variable plots of Model 2.2*

## 15. Interpret the $R^2$ value.

R-squared value, also known as the coefficient of determination, is a measure of predictive accuracy of the regression model. It also represents the amount of variance in the dependent variable explained by the independent variables. R-squared value and the adjusted R-squared value are reported below, the prediction accuracy of the regression model is 86%.

*Multiple R-squared: 0.8626, Adjusted R-squared: 0.8612*

## 16. Do you need to apply a correction to the standard errors? Why or why not? Support your answer by a plot and statistical test. Explain why you use that specific plot and test.

To determine whether a correction is required to the standard errors, we examine the plot of the residuals against the fitted values. If the model is heteroskedastic then the plot has a structure, which, from the graph below, can be seen. Hence, we need to check further for heteroskedasticity.

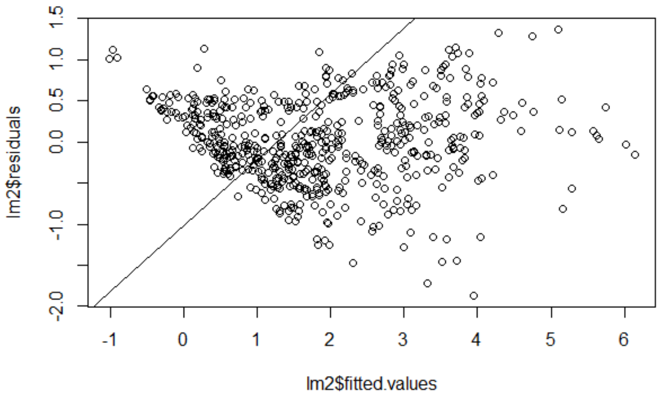


*Figure 5: Heteroskedasticity plot*

Further diagnostics is carried out by using the Breusch-Pagan test. The Breusch-Pagan test statistic is 54.397 with a p-value <0.001, which rejects the null hypothesis and thus indicates heteroskedasticity. This would mean that the estimated regression coefficients are correct, but the associated standard errors, t-statistics and p-values are not.

Breusch-Pagan test

BP = 54.397, df = 6, p-value = 6.134e-10

Null Hypothesis ($H_0$): Homoskedasticity is present.

## 17. If you have to apply a correction to the standard errors, apply the correction and report the new results. What has changed? If you do not apply a correction, you can simply skip this question.

Yes, a correction to the standard errors has been applied by using the Heteroskedasticity-Corrected Covariance Matrix of the regression parameters. The standard errors of the regression parameters, the t-statistics and the p-value are now different as reported in Table. However, the substantial conclusions from the model remain unchanged.

The column 'Before' in the table below indicates the values of standard errors, t value and p-value before the application of correction and the column 'After' shows the values obtained based on the corrected co-variance matrix.

| **Coefficients** | **Std. Error** | | **t value** | | **P value** | |
|---|---|---|---|---|---|---|
| | Before | After | Before | After | Before | After |
| Intercept | 0.0577 | 0.0638 | -17.593 | -15.9190 | < 2e-16 | < 2.2e-16 |
| log(Size) | 0.0162 | 0.0215 | 49.314 | 37.1519 | < 2e-16 | < 2.2e-16 |
| CoopRnD | 0.0540 | 0.0517 | -2.251 | -2.3509 | 0.02475 | 0.0191 |
| log(InhouseRnD) | 0.0774 | 0.0870 | 3.060 | 2.7232 | 0.00232 | 0.0067 |
| log(ExternalRnD) | 0.4034 | 0.3353 | 3.024 | 3.6381 | 0.00261 | 0.0003 |
| log(Machinery) | 0.0826 | 0.0762 | 3.031 | 3.2867 | 0.00255 | 0.0011 |
| log(ExternalKnow) | 2.1489 | 1.8770 | 0.093 | 0.1064 | 0.92602 | 0.9153 |

*Table 9: Values of Std. Error, t value and p values before and after correction*

Furthermore, a plot of residuals shows a normal trend on visual inspection and satisfies our assumption that the error terms are normally distributed, as shown in the graph below.

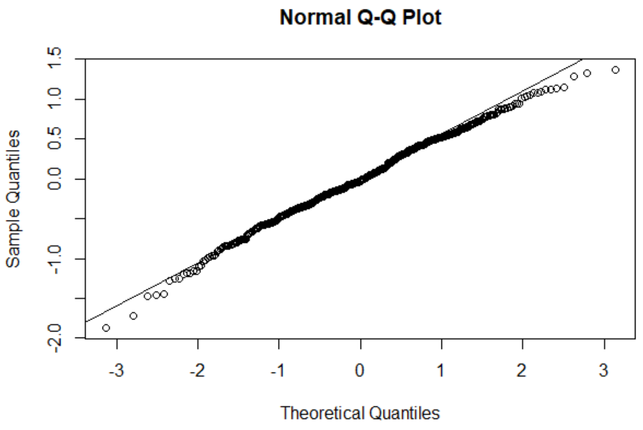


*Figure 6: Check for normality of the error terms*

## 18. Interpret the regression parameters.

The final regression model is as shown below. According to this model, Size of a company, InhouseRND expenditures, ExternalRnD expenditures and Machinery expenditures have a positive effect on the turnover of a company while cooperation on innovation activities with other companies tends to have a negative effect on turnover of that organization.

$$
\begin{aligned}
log(Turnover_i) = -1.0159 + 0.8003\, x\, log(Size_i) &- 0.1215\, x\, CoopRnD_i \\
&+ 0.2369\, x\, log(InhouseRnD_i) + 1.2200\, x\, log(ExternalRnD_i) \\
&+ 0.2505\, x\, log(Machinery_i) + 0.1996\, x\, log(ExternalKnow_i) + \varepsilon_i
\end{aligned}
$$

- The alpha value (-1.01590) is the intercept.
- Simply put, holding Size and other expenditures constant, a company that Cooperates on its innovations with other companies would generate 0.12% less in turnover.
- A company that co-operates on its innovation activities with other enterprises/institutions is expected to generate
    - 0.8% more in turnover when the company size is increased by 1%, holding InhouseRnD, ExternalRnD, Machinery and ExternalKnow constant.
    - 0.24% more in turnover when the company invests 1% more in InhouseRnD, holding Company Size, ExternalRnD, Machinery and ExternalKnow constant.
    - 1.23% more in turnover when the company invests 1% more in ExternalRnD, holding Company Size, InhouseRnD, Machinery and ExternalKnow constant.
    - 0.25% more in turnover when the company invests 1% more in Machinery, holding Company Size, Inhouse RnD, ExternalRnD and ExternalKnow constant.
- ExternalKnow is not a significant variable to the model, since the p-value of its coefficient is greater than 0.05.

## 19. Which changes do you suggest to the turnover model to answer questions Q1 and Q2? Write down the model equation.

To make our regression model more robust we begin investigating the effects of other independent variables that were left out before. In the previous model, the coefficient of 'ExternalKnow' was not significant, in order to try to improve the model, this variable was exchanged for the other available variables in the data. 'MarketSpend' has a significant effect and also increases the prediction accuracy of the model upon inclusion. The variables were log-transformed to achieve a better spread and because of their inherent diminishing nature. No quadratic effect was observed. The final model equation is as follows:

$$\text{Model 2.3 } \log(Turnover_{\mathrm{i}}) = \alpha + \beta_1 \log(Size_{\mathrm{i}}) + \beta_2 CoopRnD_{\mathrm{i}} + \beta_3 \log(InhouseRnD_{\mathrm{i}}) + \beta_4 \log(ExternalRnD_{\mathrm{i}}) + \beta_5 \log(Machinery_{j}) + \beta_6 \log(MarketSpend_{j}) + \varepsilon_{\mathrm{i}}$$

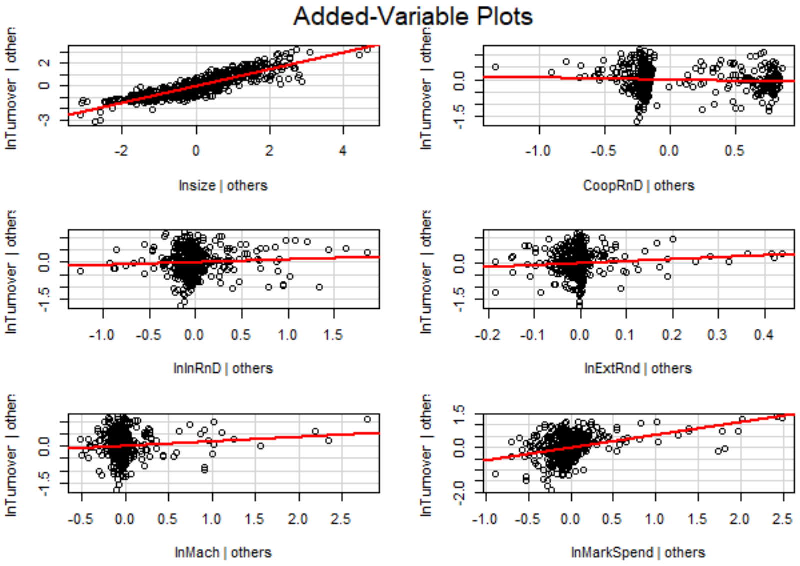


*Figure 7: Added Variable plots of Model 2.3*

## 20. Estimate and report the results. Apply a correction to the standard errors if needed.

The table below shows the results of the new model before correction to standard errors:

| **Dependent Variable: log(Turnover)** | | | | |
|---|---|---|---|---|
| **Coefficients** | **Estimate** | **Std. Error** | **T value** | **P value** |
| (Intercept) | -0.8779 | 0.0565 | -15.544 | <2e-16 |
| log(Size) | 0.7469 | 0.0164 | 45.590 | <2e-16 |
| CoopRnD | -0.1121 | 0.0507 | -2.212 | 0.0273 |
| log(InhouseRnD) | 0.1126 | 0.0741 | 1.520 | 0.1290 |
| log(ExternalRnD) | 0.7831 | 0.3762 | 2.082 | 0.0378 |
| log(Machinery) | 0.1762 | 0.0769 | 2.289 | 0.0224 |
| log(MarketSpend) | 0.5567 | 0.0638 | 8.728 | <2e-16 |
| Multiple R-squared: 0.8789, | | Adjusted R-squared: 0.8776 | | |

*Table 10: Ordinary Least Squares output of Model 2.3*

The plot of residuals against fitted values displayed a structure and a Breusch-Pagan test confirmed heteroskedasticity.

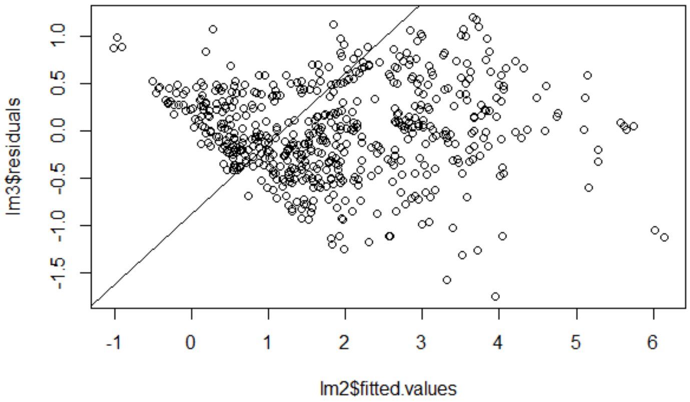


*Figure 8: Heteroskedasticity plot of Model 2.3*

Breusch-Pagan test had test statistics of 46.633 with p-value<0.001 and confirmed heteroskedasticity.

Breusch-Pagan test

BP = 46.633, df = 6, p-value = 2.214e-08

Null Hypothesis ($H_0$): Homoskedasticity is present.

The model had slight changes in the Standard errors of the regression parameters, t statistics and p-value after using the Heteroskedasticity-Corrected Covariance Matrix. The changes are not significant and further corrective actions are not required.

After standard error correction the output is:

| **Dependent Variable: log(Turnover)** | | | | |
|---|---|---|---|---|
| **Coefficients** | **Estimate** | **Std. Error** | **t value** | **P value** |
| (Intercept) | -0.8780 | 0.0571 | -15.3858 | <2.2e-16 |
| log(Size) | 0.7469 | 0.0201 | 37.2486 | <2.2e-16 |
| CoopRnD | -0.1121 | 0.0476 | -2.3554 | 0.0188 |
| log(InhouseRnD) | 0.1126 | 0.0879 | 1.2818 | 0.2004 |
| log(ExternalRnD) | 0.7831 | 0.3053 | 2.5651 | 0.0106 |
| log(Machinery) | 0.1762 | 0.0858 | 2.0533 | 0.0405 |
| log(MarketSpend) | 0.5567 | 0.0731 | 7.6194 | 1,07E-13 |

*Table 11: Reporting Results of Model 2.3 after Correction of Standard Errors*

Final model equation:

$$\begin{aligned}\log(Turnover_\mathrm{i}) = &\text{-}0.8780 + 0.7468\ x \log(Size_\mathrm{i}) - 0.1121\ x\, CoopRnD_\mathrm{i} \\ &+ 0.1126\ x \log(InhouseRnD_\mathrm{i}) + 0.7834\ x \log(ExternalRnD_\mathrm{i}) \\ &+ 0.1762\ x \log(Machinery_i)\ + 0.5567\ x \log(MarketSpend_i) + \varepsilon_\mathrm{i}\end{aligned}$$

## 21. Do you have to worry about multicollinearity?

### (a) Explain how you know whether you have to worry.

### (b) If multicollinearity is an issue, re-specify the model. Make very clear how you decide on the re-specification of your model.

A good indicator of multicollinearity in the model is the VIF(Variance Inflation Fator) scores. A generally well regarded rule is to reconsider variables if VIF>2 and to completely drop them off if VIF>10. VIF scores in our model do not exceed 2 for any variable and hence multicollinearity is not a problem.

| log(Size) | CoopRnD | log(InhouseRnD) | log(ExternalRnD) | log(Machinery) | log(MarketSpend) |
|---|---|---|---|---|---|
| 1.4980 | 1.2331 | 1.6686 | 1.4620 | 1.1858 | 1.5485 |

*Table 12:VIF output*

## 22. Compare the model fit of to the turnover model of question 14. What do you conclude?

Model from question 14:

$$\begin{aligned}\log(Turnover_\mathrm{i}) = &\text{-}1.0159 + 0.8003\ x \log(Size_\mathrm{i}) - 0.1215\ x\, CoopRnD_\mathrm{i} \\ &+ 0.2369\ x \log(InhouseRnD_\mathrm{i}) + 1.2200\ x \log(ExternalRnD_\mathrm{i}) \\ &+ 0.2505\ x \log(Machinery_i) + 0.1996\ x \log(ExternalKnow_i) + \varepsilon_\mathrm{i}\end{aligned}$$

Current model:

$$\begin{aligned}\log(Turnover_\mathrm{i}) = &\text{-}0.8779 + 0.7468\ x \log(Size_\mathrm{i}) - 0.1121\ x\, CoopRnD_\mathrm{i} \\ &+ 0.1126\ x \log(InhouseRnD_\mathrm{i}) + 0.7834\ x \log(ExternalRnD_\mathrm{i}) \\ &+ 0.1762\ x \log(Machinery_i)\ + 0.5567\ x \log(MarketSpend_i) + \varepsilon_\mathrm{i}\end{aligned}$$

| Dependent Variable: log(Turnover) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Model (Question 14) | | | | Model (Question 21) | | | |
| **Coefficients** | **Estimate** | **Std. Error** | **t-value** | **p-value** | **Estimate** | **Std. Error** | **t-value** | **p-value** |
| (Intercept) | -1.0159 | 0.0577 | -17.593 | < 2e-16 | -0.8779 | 0.0571 | -15.3858 | <2.2e-16 |
| log(Size) | 0.8003 | 0.0162 | 49.314 | < 2e-16 | 0.7469 | 0.0201 | 37.2486 | <2.2e-16 |
| CoopRnD | -0.1215 | 0.0540 | -2.251 | 0.0248 | -0.1121 | 0.0476 | -2.3554 | 0.0188 |
| log(InhouseRnD) | 0.2369 | 0.0774 | 3.060 | 0.0023 | 0.1126 | 0.0879 | 1.2818 | 0.2004 |
| log(ExternalRnD) | 1.2200 | 0.4034 | 3.024 | 0.0026 | 0.7831 | 0.3053 | 2.5651 | 0.0106 |
| log(Machinery) | 0.2505 | 0.0826 | 3.031 | 0.0026 | 0.1762 | 0.0858 | 2.0533 | 0.0405 |
| log(ExternalKnow) | 0.1996 | 2.1489 | 0.093 | 0.9260 | | | | |
| log(MarketSpend) | | | | | 0.5567 | 0.0731 | 7.6194 | 1.07e-13 |
| Adjusted R-squared | 0.8612 | | | | 0.8776 | | | |

*Table 13: Comparison of Models from Q14 and Q21*

As showcased in the table above, there are 2 main differences:

- **InHouseRnD** – In the model from question 14, InhouseRnD has a significant effect in predicting the outcome variable, however, in our current model its significance has diminished (characterized by a large p-value, Null Hypothesis: Independent variable is significant) due to addition of the predictor variable – MarketSpend. This is because MarketSpend has a lot more significance in predicting the turnover of a company than InhouseRnD investment and predictive effect of MarketSpend overshadows that of InhouseRnD

- **Adjusted R-squared**- Adjusted R-squared, which is used to compare models with the same dependent variable but different number of explanatory variables, in the current model shows a statistic of 87% in comparison to the model from question 14 which had a statistic of 86%. The difference, though quite small, is an indication that the current model has a higher predictive accuracy.

## 23. Q1 Do firms that co-operate on their innovations with external partners have, on average, a higher turnover {controlling for other factors that might affect turnover?

No, the estimate coefficient value predicts that the enterprises that co-operate on their innovations with external partners, on average, generate lesser turnover (negative coefficient) holding the other predictors constant. This is evident from the p-value and coefficient of the predictor CoopRnD in the current model which are 0.0188 and -0.1121 respectively. The p-value is obtained based on corrected covariance matrix and thus CoopRnD can be considered significant.

## 24. Q2 What is the elasticity of turnover with respect to in-house R&D spending {controlling for other factors that might affect turnover?

The relation between R&D spending and Turnover is linear holding the other factors constant. The parameter therefore represents the elasticity:

$$\frac{\partial Turnover_{\mathrm{i}}}{Turnover_{\mathrm{i}}} = \beta_3 * \frac{\partial InhouseRnD_{\mathrm{i}}}{InhouseRnD_{\mathrm{i}}}$$

The elasticity of turnover with respect to in-house R&D spending equals 0.1126 and is insignificant (p-value = 0.20044 which is greater than 0.05).