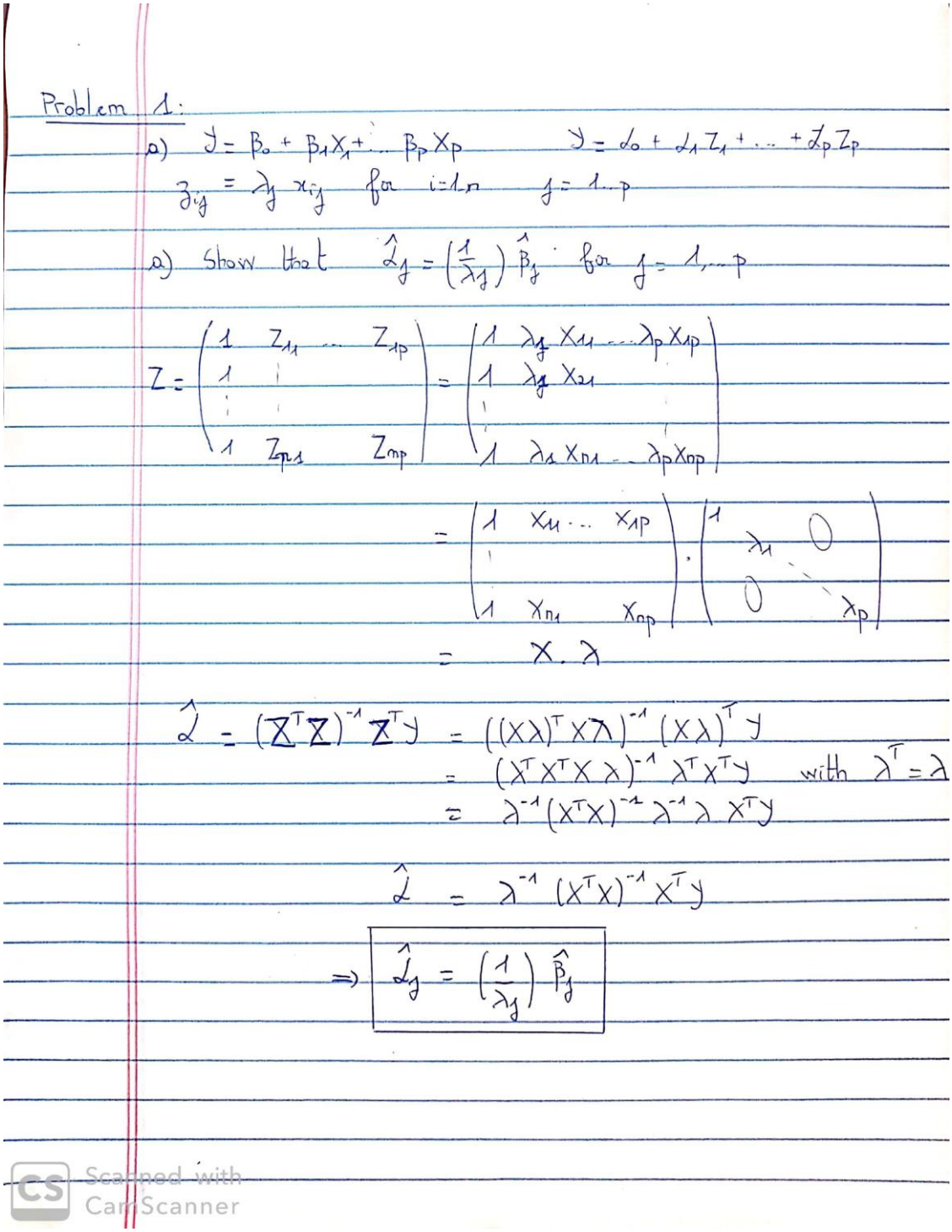
Homework 1



Problem 1:

a) $y = \beta_0 + \beta_1 X_1 + \dots \beta_p X_p \qquad y = \alpha_0 + \alpha_1 Z_1 + \dots + \alpha_p Z_p$

$z_{ij} = \lambda_j x_{ij}$ for $i = 1..n \qquad j = 1..p$

a) Show that $\hat{\alpha}_j = \left(\frac{1}{\lambda_j}\right) \hat{\beta}_j$ for $j = 1, \dots p$

$$Z = \begin{pmatrix} 1 & Z_{11} & \dots & Z_{1p} \\ 1 & & & \\ \vdots & & & \\ 1 & Z_{n1} & & Z_{np} \end{pmatrix} = \begin{pmatrix} 1 & \lambda_1 X_{11} & \dots & \lambda_p X_{1p} \\ 1 & \lambda_1 X_{21} & & \\ & & & \\ 1 & \lambda_1 X_{n1} & & \lambda_p X_{np} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ & & & \\ 1 & X_{n1} & & X_{np} \end{pmatrix} \cdot \begin{pmatrix} 1 & & & \\ & \lambda_1 & & 0 \\ & & \ddots & \\ & 0 & & \lambda_p \end{pmatrix}$$

$$= X \cdot \lambda$$

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T y = ((X\lambda)^T X\lambda)^{-1} (X\lambda)^T y$$

$$= (\lambda^T X^T X \lambda)^{-1} \lambda^T X^T y \quad \text{with } \lambda^T = \lambda$$

$$= \lambda^{-1} (X^T X)^{-1} \lambda^{-1} \lambda X^T y$$

$$\hat{\alpha} = \lambda^{-1} (X^T X)^{-1} X^T y$$

$$\Rightarrow \boxed{\hat{\alpha}_j = \left(\frac{1}{\lambda_j}\right) \hat{\beta}_j}$$

b) In the case of this linear regression, a change in a size of the features does not change the prediction. There estimator would be “rescaled” by dividing by the multiplicator coefficient of the feature as shown in a). Sometimes large sizes of features can slow down gradient descent which is not use in this problem.

Problem 2:

Problem 2

$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$ $\quad w_i = y_i - \bar{y}$ $\quad z_{ij} = x_{ij} - \bar{x}_j$

$W = \alpha_0 + \alpha_1 Z_1 + \dots + \alpha_p Z_p + \epsilon$

a) Show $\hat{\alpha}_0 = 0$

$$\frac{\partial}{\partial \alpha} \|W - Z\alpha\|_2^2 = 0 \Rightarrow 2Z^T(W - Z\hat{\alpha}) = 0 \qquad W - Z\alpha$$

$$Z = \begin{pmatrix} 1 & z_{11} & \dots & z_{1p} \\ \vdots & & & \\ 1 & z_{m1} & \dots & z_{mp} \end{pmatrix} \qquad Z^T = \begin{pmatrix} 1 & \dots & \dots & 1 \\ z_{11} & & & z_{m1} \\ \vdots & & & \vdots \\ z_{1p} & & & z_{mp} \end{pmatrix} \begin{pmatrix} w_1 - (\alpha_0 + \sum_{i=1}^{p} z_{1i}\alpha_i) \\ \vdots \\ w_n - (\alpha_0 + \sum_{i=1}^{p} z_{mi}\alpha_i) \end{pmatrix}$$

first line $Z^T \cdot (W - Z\alpha) = \sum_{i=1}^{n} W - n\hat{\alpha}_0 - \sum_{j=1}^{p} \alpha_j \left(\sum_{i=1}^{m} z_{ij} \right) = 0$

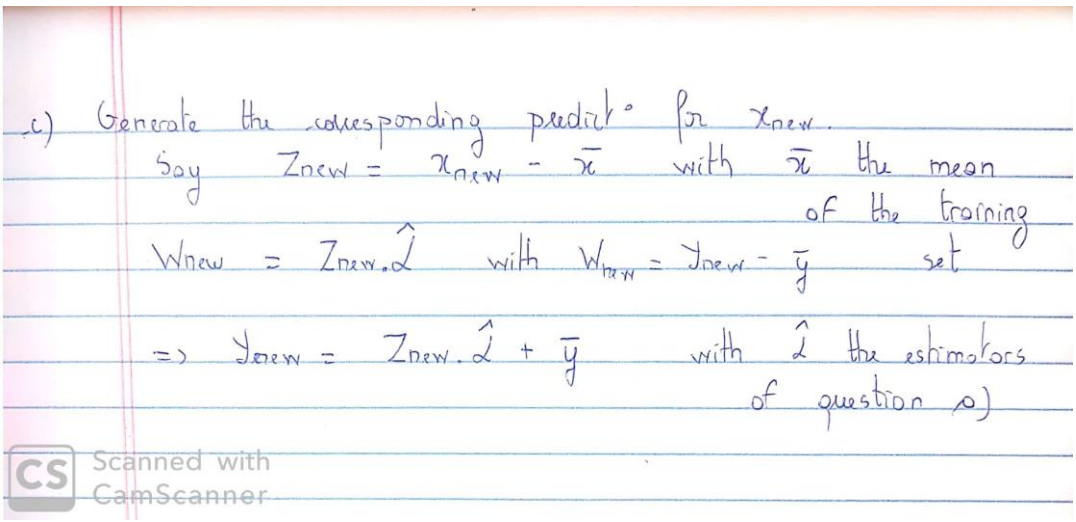
with $\bar{W} = 0 \Rightarrow \sum_{i=1}^{n} W = 0$

$\bar{z}_j = 0 \Rightarrow \sum_{i=1}^{n} z_{ij} = 0$

$\Rightarrow \quad -n\hat{\alpha}_0 = 0$

$\Rightarrow \quad \hat{\alpha}_0 = 0$

b) I see the regression estimates as the slope of the regression plane surface. Thus in the direction of the independent variable. Thus, it is constant independently of the values of the variable



Problem 3: Forecasting Jeep Wrangler Sales

a) Building the regression model

On firsthand, we consider building a regression model based on four estimators (Unemployment = X_1, WranglerQueries= X_2, CPI.Energy= X_3 and CPI.All= X_4) to predict WranglerSales =Y.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad with\ \beta = (X^t X)^{-1} X^t Y$$

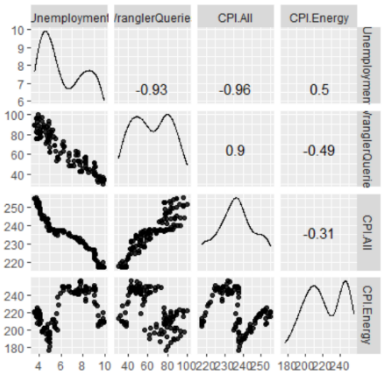
	Estimate	Pr(>|t|)	
(Intercept)	-45958.06	0.442	β_0
Unemployment	897.97	0.427	
WranglerQueries	270.23	2.87e-11 ***	β_1
CPI.Energy	-11.06	0.727	
CPI.All	174.57	0.493	

Interpretation:

$\beta_1 = 897.97$ means an additional percent of Unemployment is expected to result in an additional 898 units of WranglerSales which is not consistent since an increase in unemployment leads to a decrease in purchasing power.

$\beta_3 = -11.06$ means an additional unit of CPI.Energy is expected to result in a decrease of 11 units of WranglerSales.

In addition, except from WranglerQueries, the P-values of the estimators are high, so we will address **multicollinearity.**



According to the chart the are collinearity between several parameters. We will use the Variance inflation factors (VIF) (Good is VIF < 5) and their p-value to eliminate these coefficients from our model.

Finally, after some iterations, the only significant estimator is WranglerQueries

```
                Estimate Pr(>|t|)
(Intercept)     -952.18     0.27
WranglerQueries  257.86   <2e-16 ***
```

$$Y = -952.18 + 257.86 * WranglerQueries$$

We use the correlation factor to assess our model's performance on the trainning set

The model has a good performance at least on our training set. $\boldsymbol{R^2 = 0.79}$

b) Use the categorical variable **MonthFactor** to improve our model

i) When we add seasonality in our model, the regression equation becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \sum_{i=5}^{15} \beta_i X_i \quad with\ \beta = (X^t X)^{-1} X^t Y$$

With X_i the Month going from January (5) to December (15). We should note that we use 11 dummy variables to work with 12 categories, to avoid redundancy. We can consider that the last month goes into the intercept β_0. For a sale concerning the month i, the variable X_i take the value 1, 0 otherwise.

After computing we have the new estimators:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -69628.03   54007.71  -1.289  0.20262
Unemployment           845.80    1001.10   0.845  0.40178
WranglerQueries        175.69      61.67   2.849  0.00613 **
CPI.Energy             -25.28      28.71  -0.880  0.38240
CPI.All                317.32     236.31   1.343  0.18475
MonthFactorAugust      -62.76     944.94  -0.066  0.94728
MonthFactorDecember   -175.82    1060.77  -0.166  0.86896
MonthFactorFebruary  -1078.14     906.32  -1.190  0.23923
MonthFactorJanuary   -3262.64     953.93  -3.420  0.00117 **
MonthFactorJuly       -176.09    1003.11  -0.176  0.86128
MonthFactorJune        313.29     948.52   0.330  0.74241
MonthFactorMarch      -173.75     887.58  -0.196  0.84551
MonthFactorMay         1894.71    910.63   2.081  0.04205 *
MonthFactorNovember  -1660.69    1008.11  -1.647  0.10509
MonthFactorOctober     -776.15     986.73  -0.787  0.43484
MonthFactorSeptember   -945.17     890.99  -1.061  0.29333
```

- Interpretation:

For MonthFactorAugust, if a sale was made on August X = 1. If we consider the other parameters constant, a sale on August decrease the WranglerSales by 62.76.

For MonthFactorMay, if a sale was made on May X = 1. If we consider the other parameters constant, a sale on May increase the WranglerSales by 1894.71

With this model, we can predict the actual price for a given month.

ii) For this new model $\mathbf{R^2 = 0.87}$. The model has a good performance on the training set. The variables **WranglerQuerries**, **MonthFactorJanuary** and **MonthFactorMay** are significant.

iii) Adding the independent variable MonthFactor improve our model in a sense that the demand can be specific for a period of the year; for instance, for this model of jeep we can imagine that we will have an increase of sales for Summer an holidays (estimate of monthFactorMay positive). While it is the opposite for the beginning of the year (MonthFactorJanuary <0).

We also note that the seasonality increases our $R^2\ of\ 10\%$.

iv) We can a observe the sales or the others independent variables on a sequence of continuous discrete-time (Months) and try to identify patterns to make predictions. In can improve our model since there is a natural ordering in our observation.

c) Model selection

To build a more accurate model, we progressively delete the less significant variables by analyzing the p values and by comparing the VIF (delete if >5) to identify collinearity.

We finally have a model with only the variables **WranglerQuerries** and **MonthFactor**

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	105.57	1012.37	0.104	0.91730	
WranglerQueries	245.55	15.27	16.079	< 2e-16	***
MonthFactorAugust	-320.44	894.71	-0.358	0.72151	
MonthFactorDecember	488.42	891.27	0.548	0.58576	
MonthFactorFebruary	-911.47	890.09	-1.024	0.31001	
MonthFactorJanuary	-2894.73	894.71	-3.235	0.00199	**
MonthFactorJuly	-670.71	898.96	-0.746	0.45857	
MonthFactorJune	-64.07	893.52	-0.072	0.94308	
MonthFactorMarch	-141.03	888.17	-0.159	0.87438	
MonthFactorMay	1672.39	889.92	1.879	0.06515	.
MonthFactorNovember	-1048.56	889.92	-1.178	0.24342	
MonthFactorOctober	-256.15	889.76	-0.288	0.77444	
MonthFactorSeptember	-821.42	888.36	-0.925	0.35892	

$$WanglerSales = 105.57 + 245.55 * WranglerQueries + \sum_{2}^{12} \beta_i X_i$$

The training set $\mathbf{R^2 = 0.86}$ and the $\mathbf{OSR^2 = 0.54}$ which is not enough relevant. Thus far, the model cannot be useful to Jeep/FCA since it doesn't have a good performance on a different sample than the training sample.

d) Improve the model

We can hypothesize that the price of an oil barrel can be related to Wrangler Sales. An increase in the price of the barrel could means a decrease of the WranglerSales (the data was provided by https://fred.stlouisfed.org/series/MCOILWTICO). Thus, we expect the estimated coefficient associated to this variable negative.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-839.470	1693.110	-0.496	0.62190	
WranglerQueries	250.505	16.901	14.822	< 2e-16	***
MonthFactorAugust	-302.243	899.002	-0.336	0.73793	
MonthFactorDecember	608.926	911.661	0.668	0.50683	
MonthFactorFebruary	-851.589	898.089	-0.948	0.34695	
MonthFactorJanuary	-2807.176	907.334	-3.094	0.00304	**
MonthFactorJuly	-684.640	903.111	-0.758	0.45147	
MonthFactorJune	-57.991	897.473	-0.065	0.94870	
MonthFactorMarch	-117.735	892.679	-0.132	0.89553	
MonthFactorMay	1679.119	893.863	1.878	0.06534	.
MonthFactorNovember	-946.867	905.605	-1.046	0.30010	
MonthFactorOctober	-166.408	902.847	-0.184	0.85441	
MonthFactorSeptember	-772.450	894.995	-0.863	0.39165	
Barrel	7.558	10.827	0.698	0.48792	

Unfortunately, this variable is not significant since its P value is not < 5% and its estimate is not negative as expected. However, we have a better $\mathbf{OSR^2 = 0.66}$, on overall it improves the performance of the combined collection of predictors. However, the OSR^2 is still to low to use the model as a predictor of Wrangler Sales.