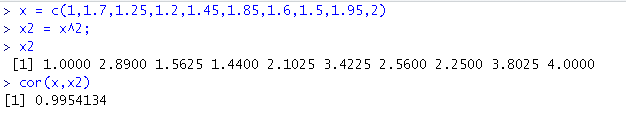
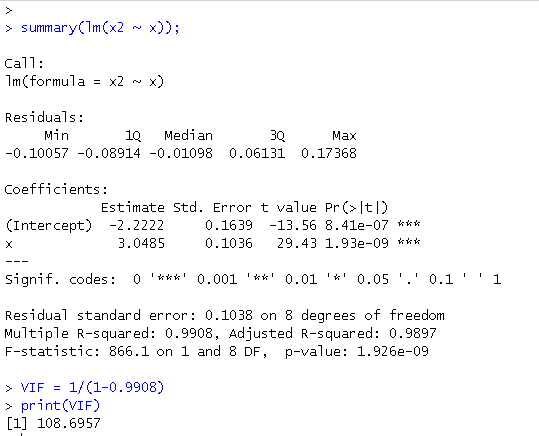
STAT 5532– final excercises

**Thanh Doan – Student ID 0159701**

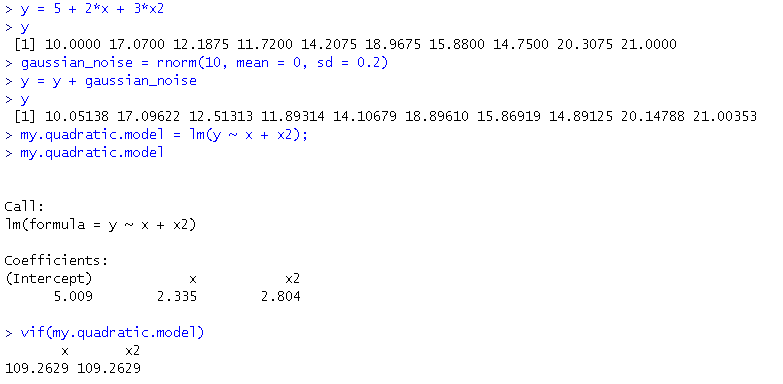
1. Exercise 7.1



* The correlation between **x** and **x2** is 0.9954
* From the correlation between x and **x2** we can see the potential difficulty in fitting a second-order model because of the multi-co-linearity problem.
* To show the multi-co-linearity problem, I regress **x2** on **x** and compute the VIF

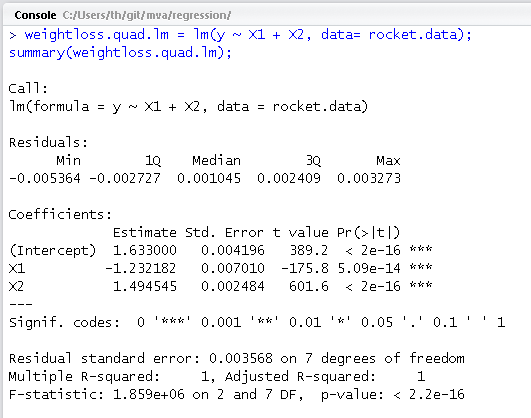
****

* I also created a response vector, **y**, and regress **y** on **x** and **x2** then use the **vif**() function to compute the variance inflation factors.



* The VIF values in this step are slightly different than the ones in previous steps because of a round-off error, computed, in the previous step.
* The VIF values of 109.26 are high and suggest serious multi-co-linearity problem.

1. **Exercise 7.2**
   1. Fit a second-order polynomial that expresses weight loss as a function of the number of months since production

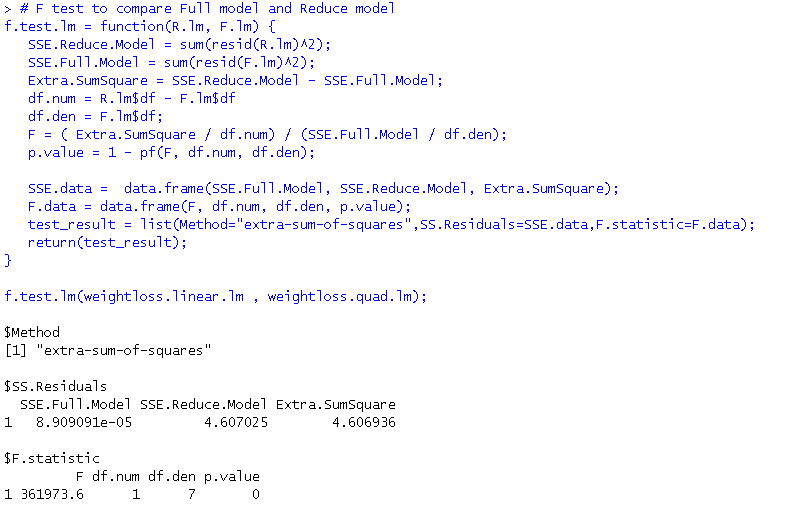


The fitted model is

* 1. Test for significance of the regression

The statistic. Thus the regression is significant

* 1. Test the hypothesis

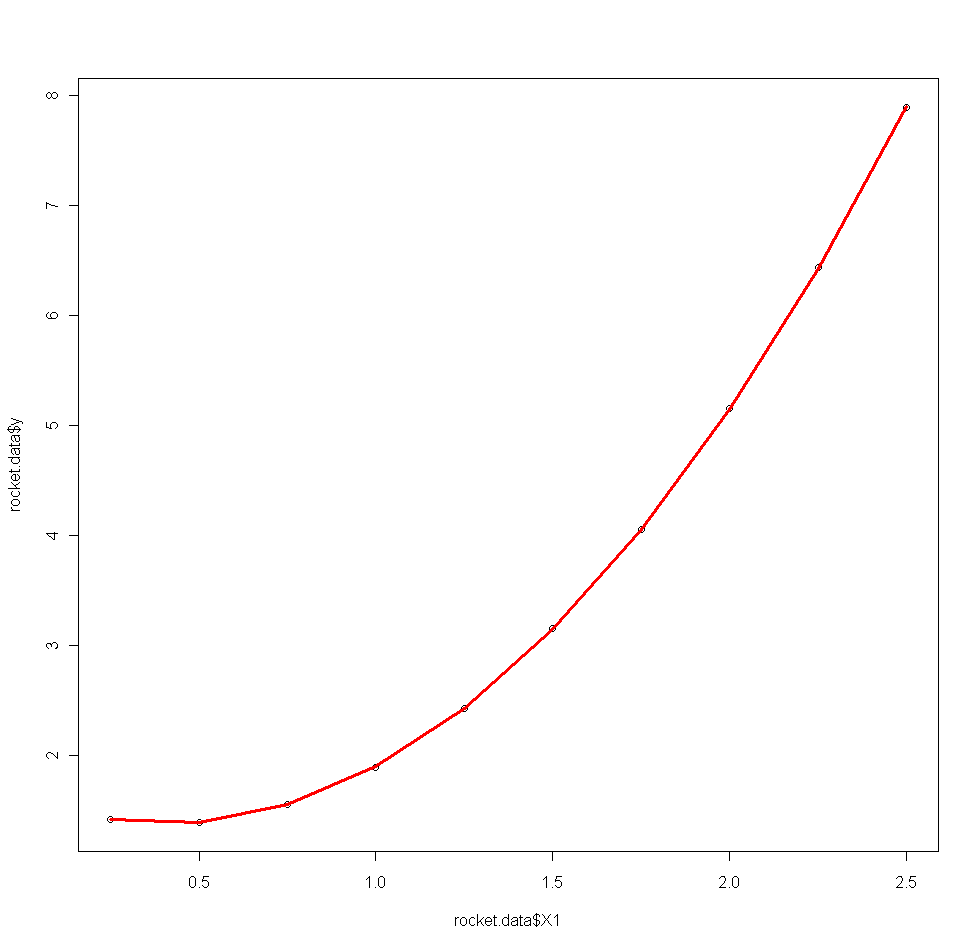


Using extra-sum-of-squares method to compare the full model (with x2 term in)

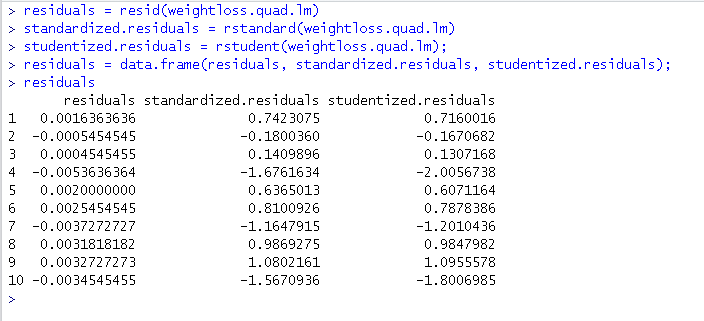
with reduce model we have F = 361973.6 and p-value less than 0.0001

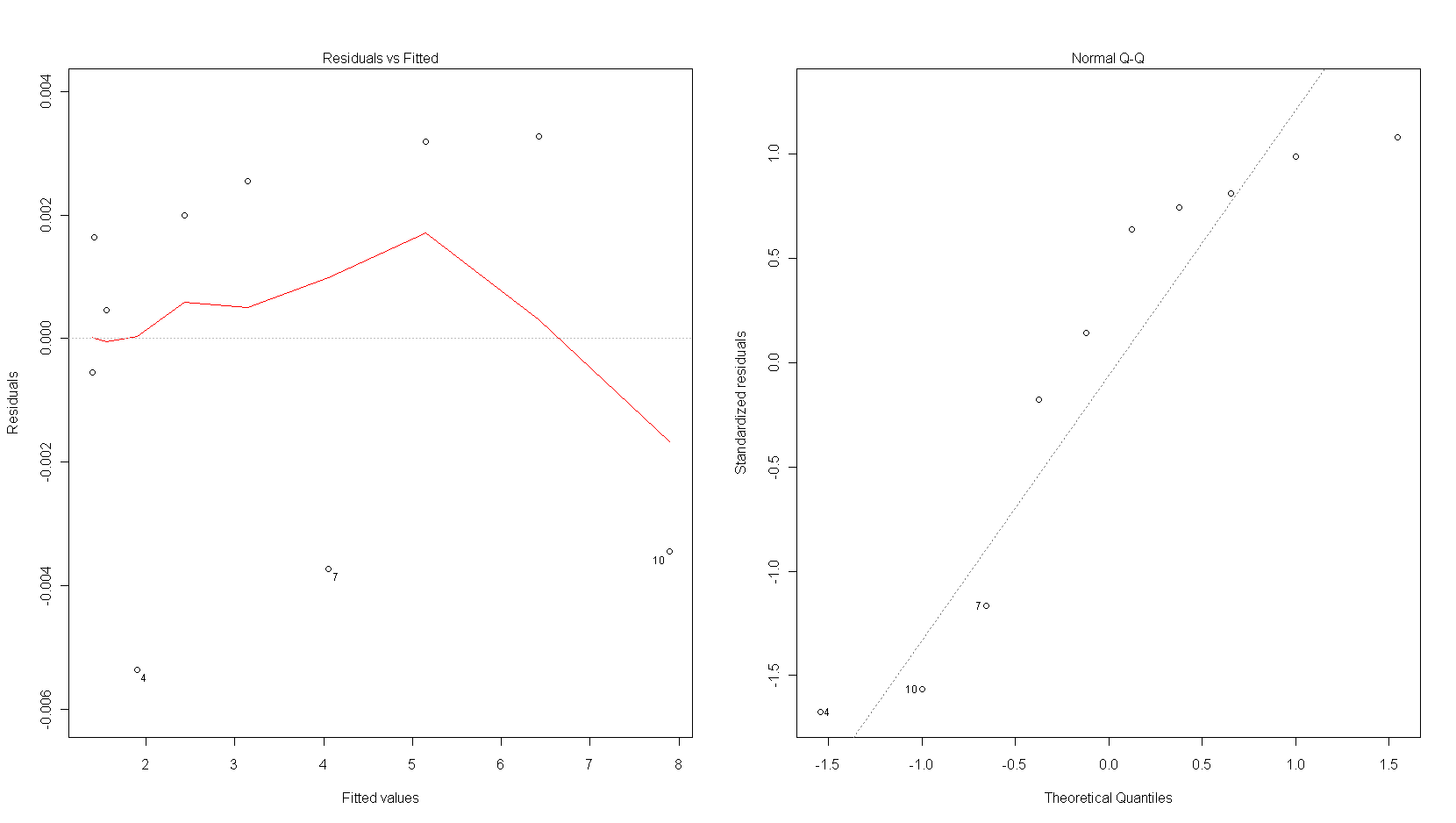
We can reject the hypothesis

* 1. Yes. This second-order polynomial model, while fitting the given data below very well, can be potential hazards in extrapolating as most quadratic models fitting in a small value range of x.



1. **Exercise 7.3**

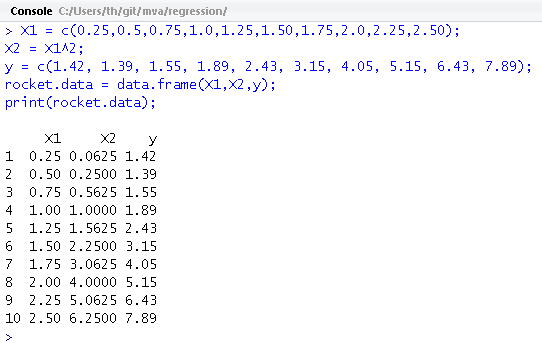




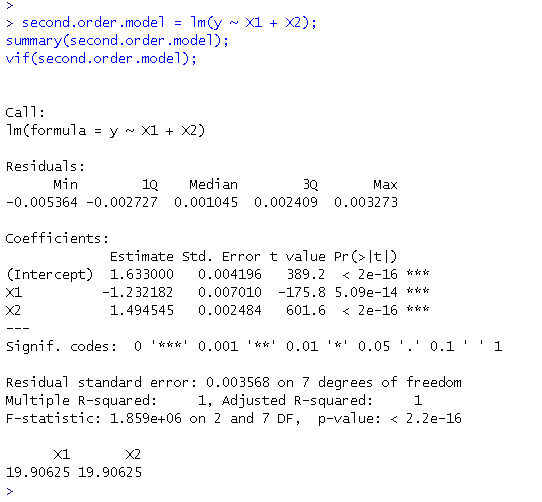
* From the computed residuals above… there is no suggestion of outlier.
* There is a problem of normality, but the size of the sample is small so there is no need to read in too much
* The residuals seem to indicate that the quadratic model is adequate.

1. **Exercise 7.14**

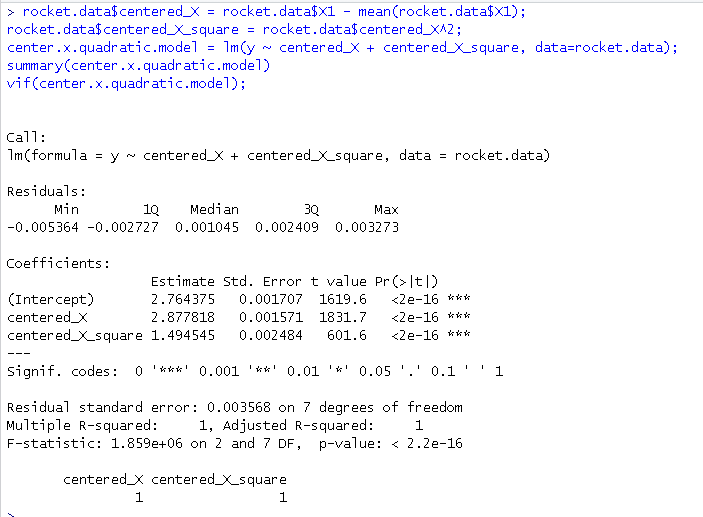
Data, x2 is square of x1.



* 1. Fit a second-order model to the data and evaluate the VIF.

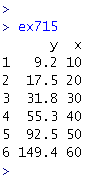


* VIF = 19.9 is rather high.
  1. Fit a second-order model to the data and evaluate the VIF.

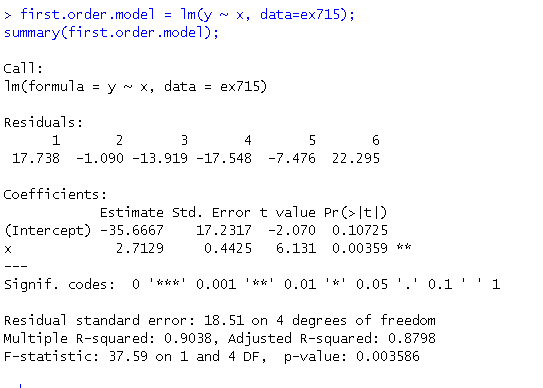


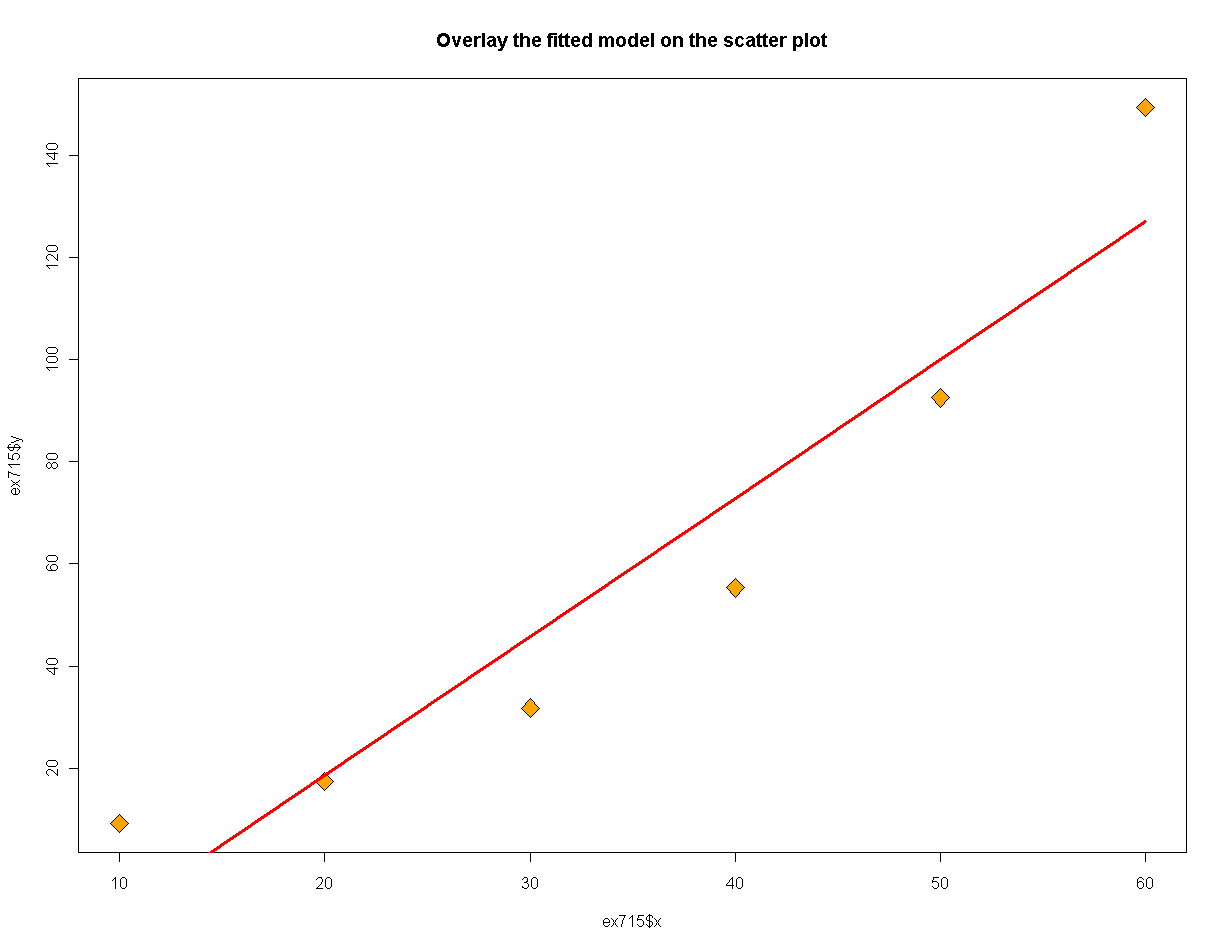
* The fitted model is
* The VIF values are 1.
  1. The impact of centering the x’s in a polynomial model on multicolinearity is good. By centering the x’s the VIF values reduce from 19 to 1 and this technique remove the ill-conditioning of the X’X matrix.

1. **Exercise 7.15**

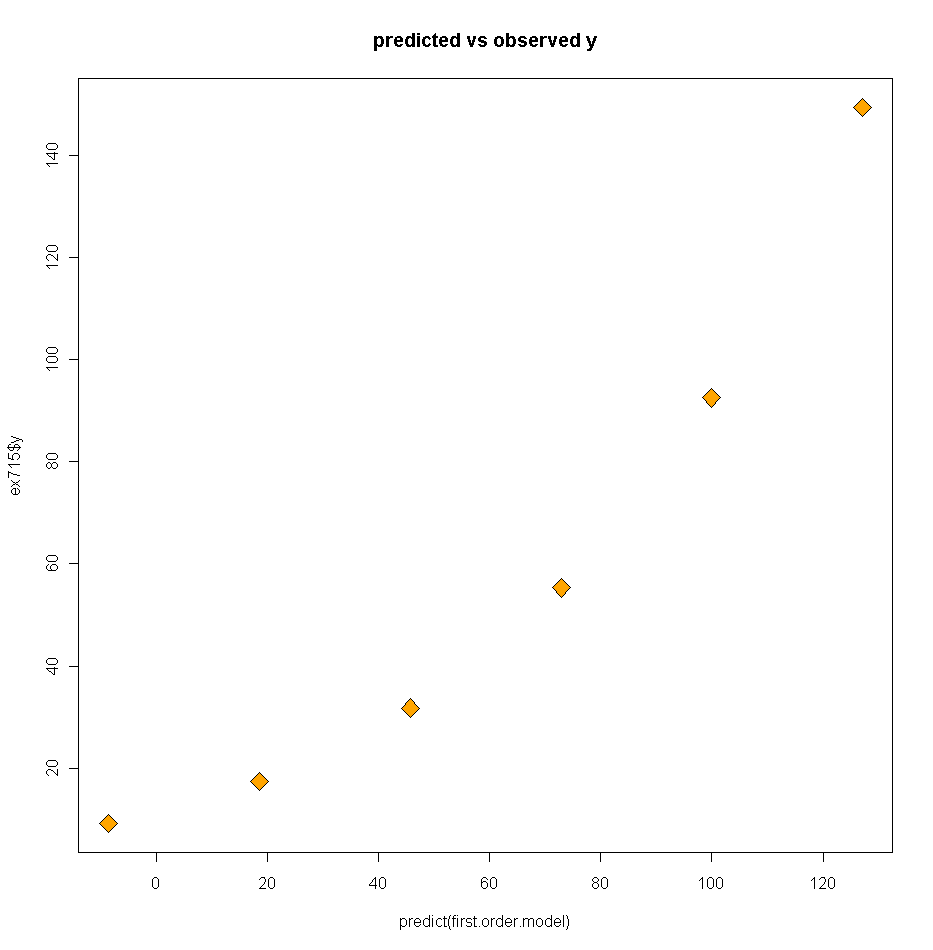


* 1. Fit a first-order model to the data

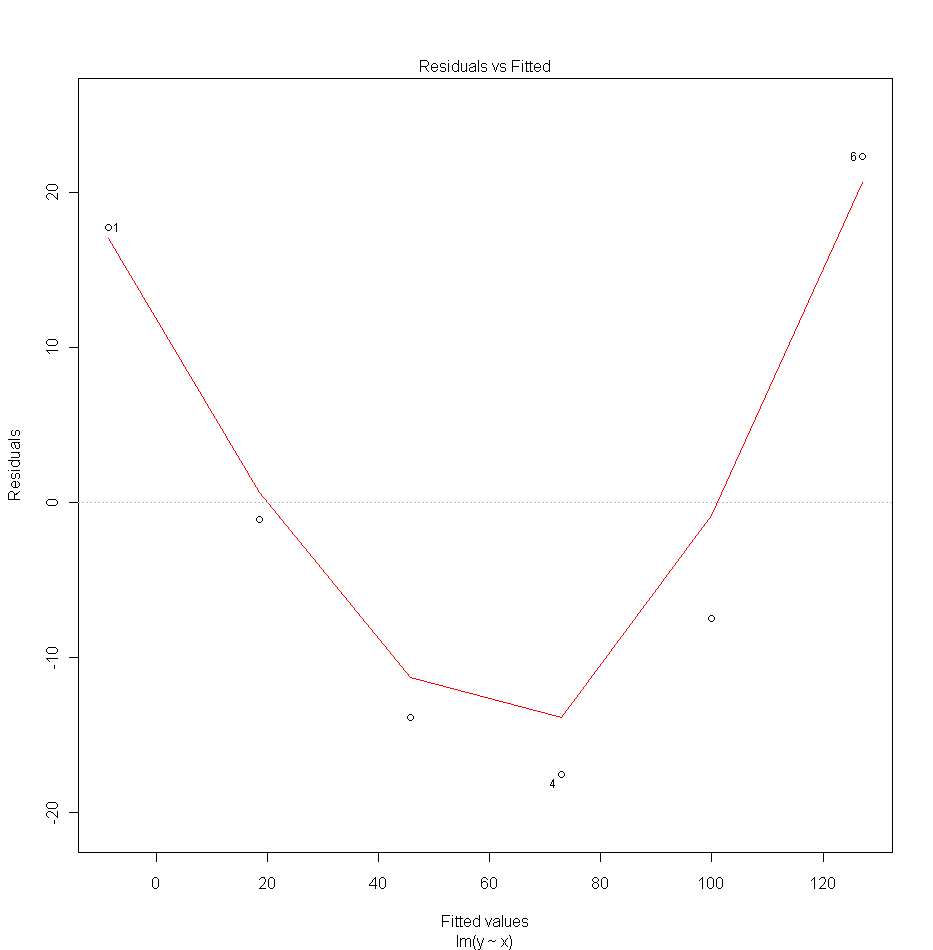




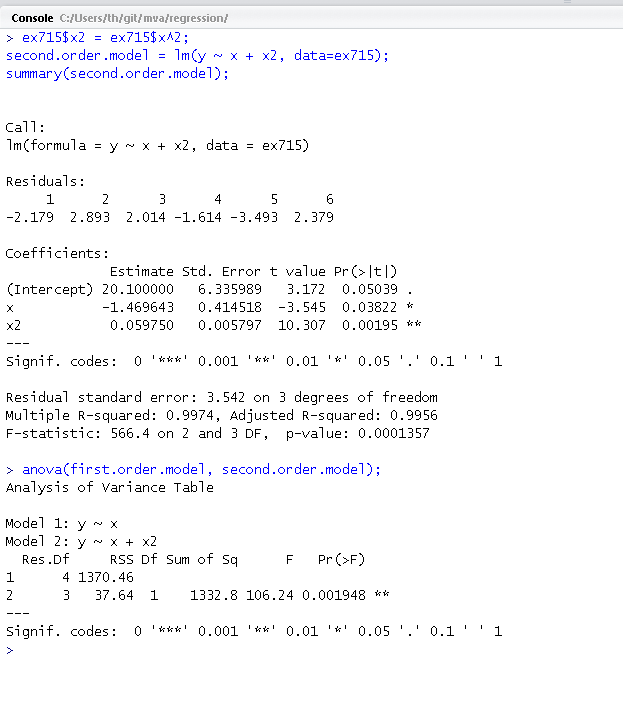
* The fitted model is has high adjusted R square and the test of regression (F.statistic = 37.59) is significant
* However the overlay of fitted model does not appear to follow the data well. The data appear to be nonlinear.
  1. Scatter plot of predicted y versus observed y.



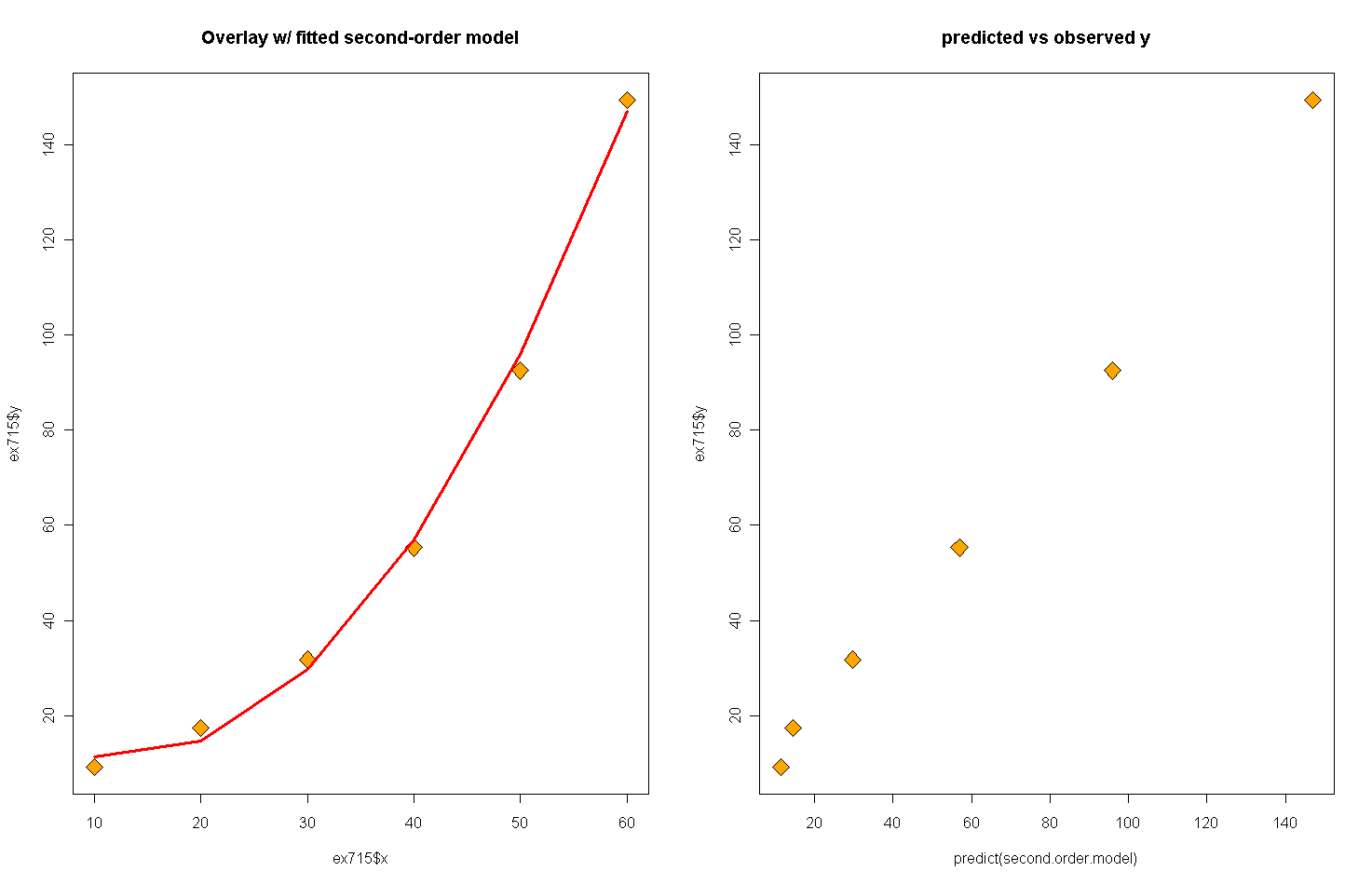
* The scatter plot of predicted y.hat vs. observed y suggests the first order model does not fit the data very well.
  1. Plot the residuals vs. the fitted y.

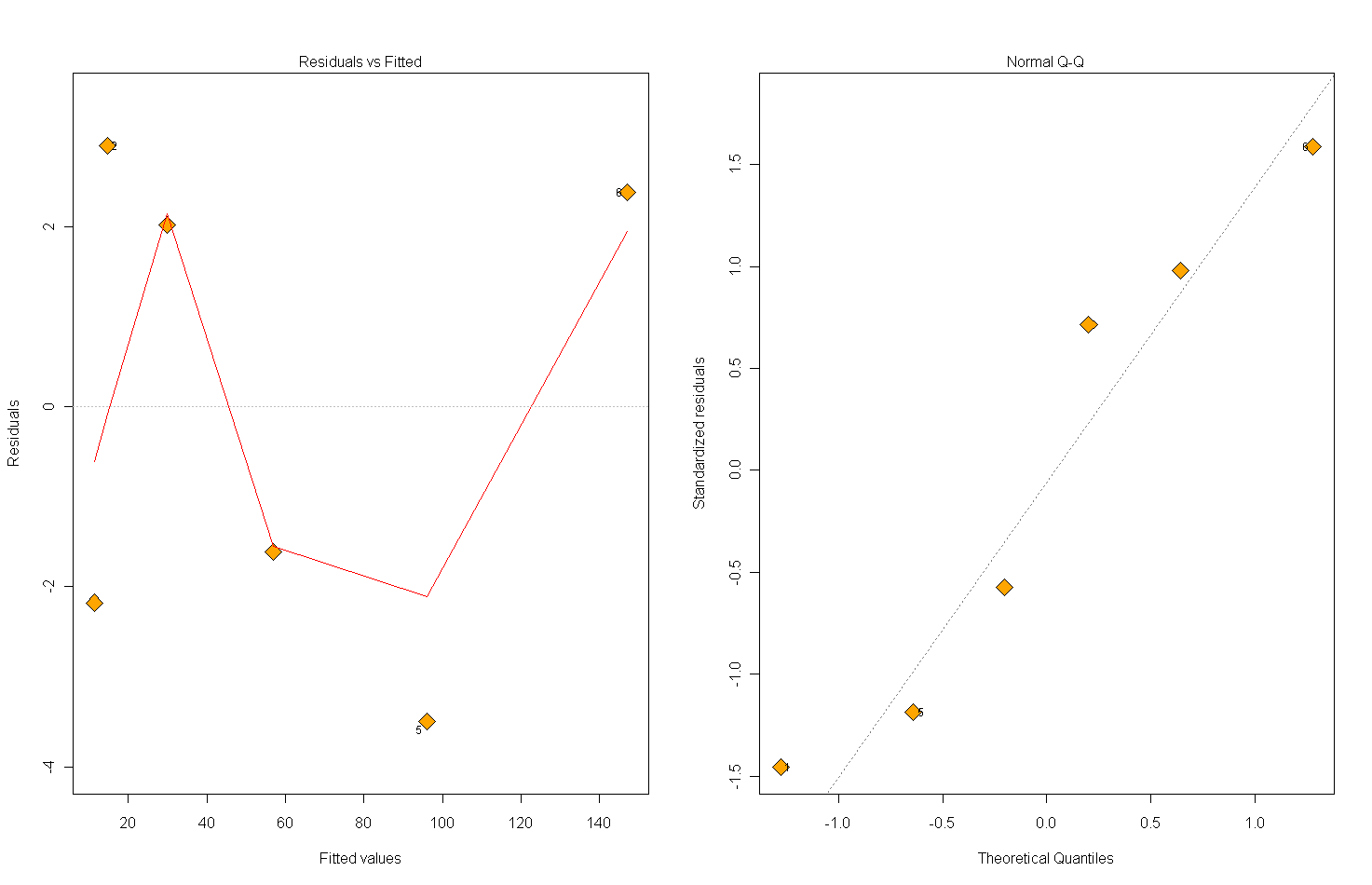


* The residuals vs. fitted y plot suggest the model is inadequate.
  1. Fit a second-order model to the data



* The fitted second-order model is
* Using anova (similar to extra-sum-of-squares method) to compare the two models and the result suggest the quadratic term is significant (F=106.24, p-value=0.0019).
  1. Overlay second-order model on scatter plot. Plot predicted vs. observed y and residuals vs. fitted y.



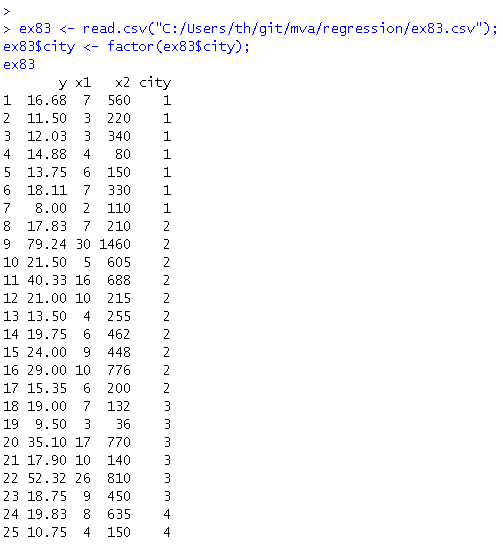


* The overlay of fitted second-order model on the scatter plot suggest this model fit the data well
* The predicted vs. observed y suggest predicted values do not deviate much from the observed ones.
* The residuals vs. fitted values plot suggest the residuals spread more evenly comparing to the first model. This plot, together with the probability plot suggests the there is no serious problem with model adequacy.

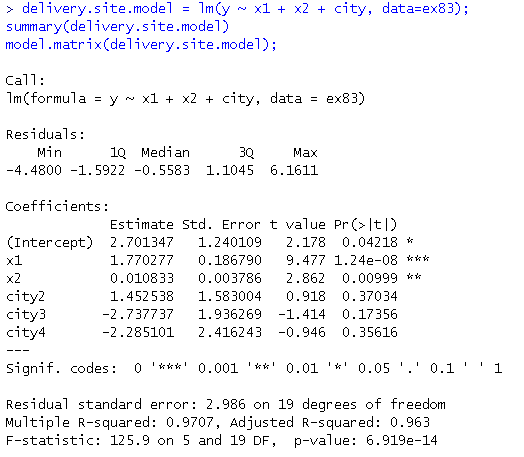
1. **Exercise 8.3**

Dataset:

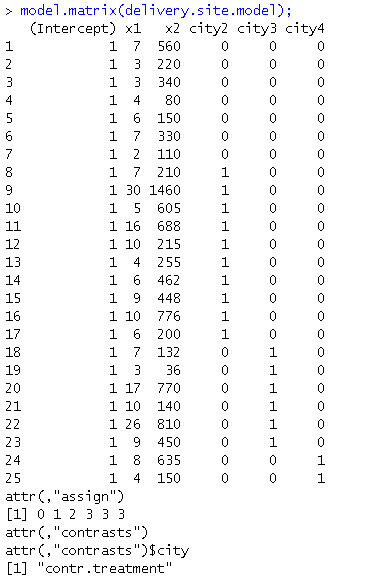
* city=1 (San Diego)
* city=2 (Boston)
* city=3 (Austin)
* city=4 (Minneapolis)



* 1. A model that relate delivery time y to cases x1, distance x2 and city



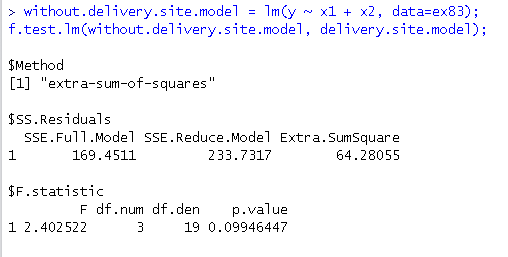
* Since city is a factor/categorical variable… R internally created the design matrix as follow



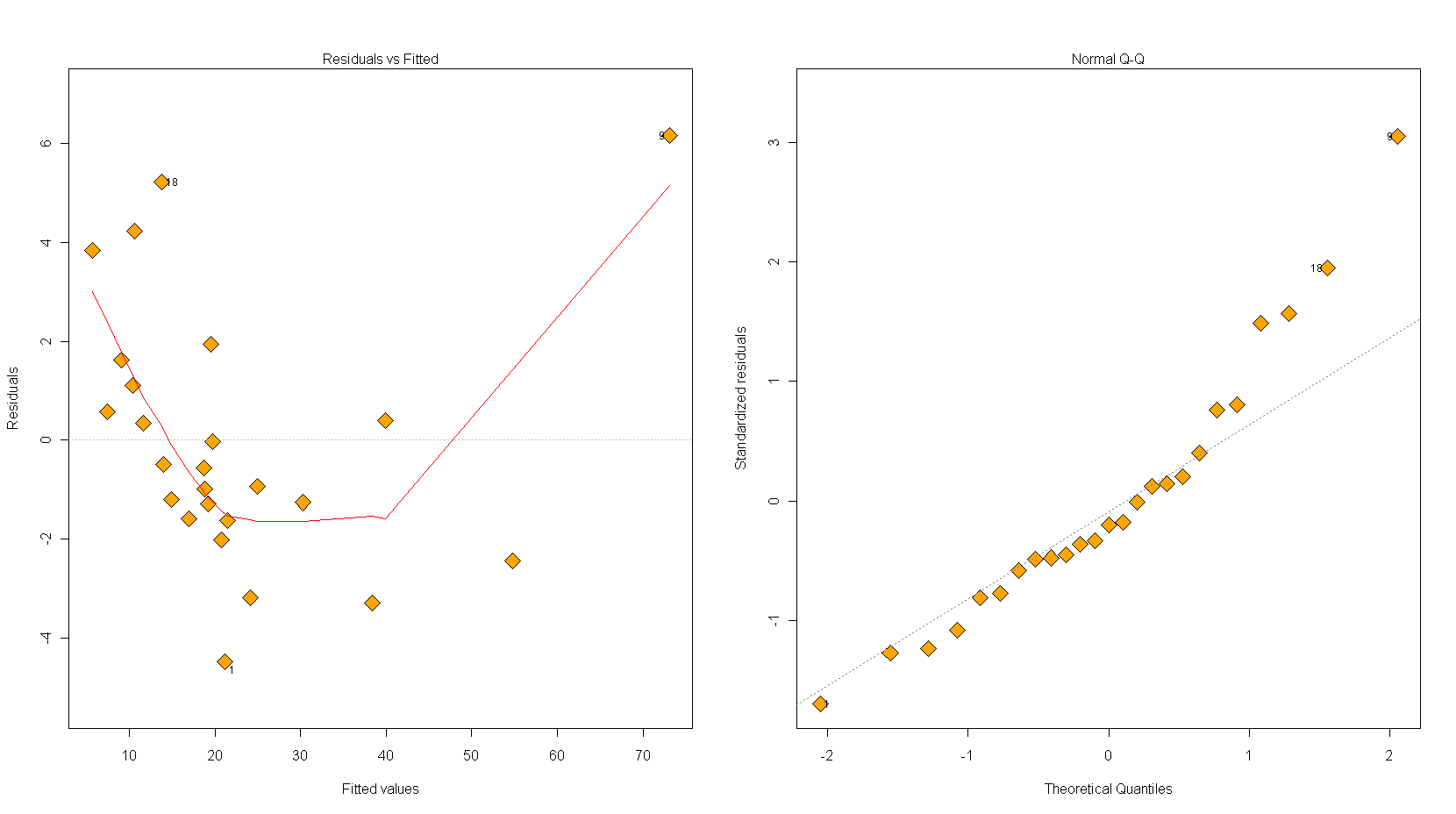
|  |  |  |  |
| --- | --- | --- | --- |
| **Indicator coded variables** | | | |
| **city2** | **city3** | **city4** | **Interpretation** |
| 0 | 0 | 0 | Observation from San Diego |
| 1 | 0 | 0 | Observation from Boston |
| 0 | 1 | 0 | Observation from Austin |
| 0 | 0 | 1 | Observation from Minneapolis |

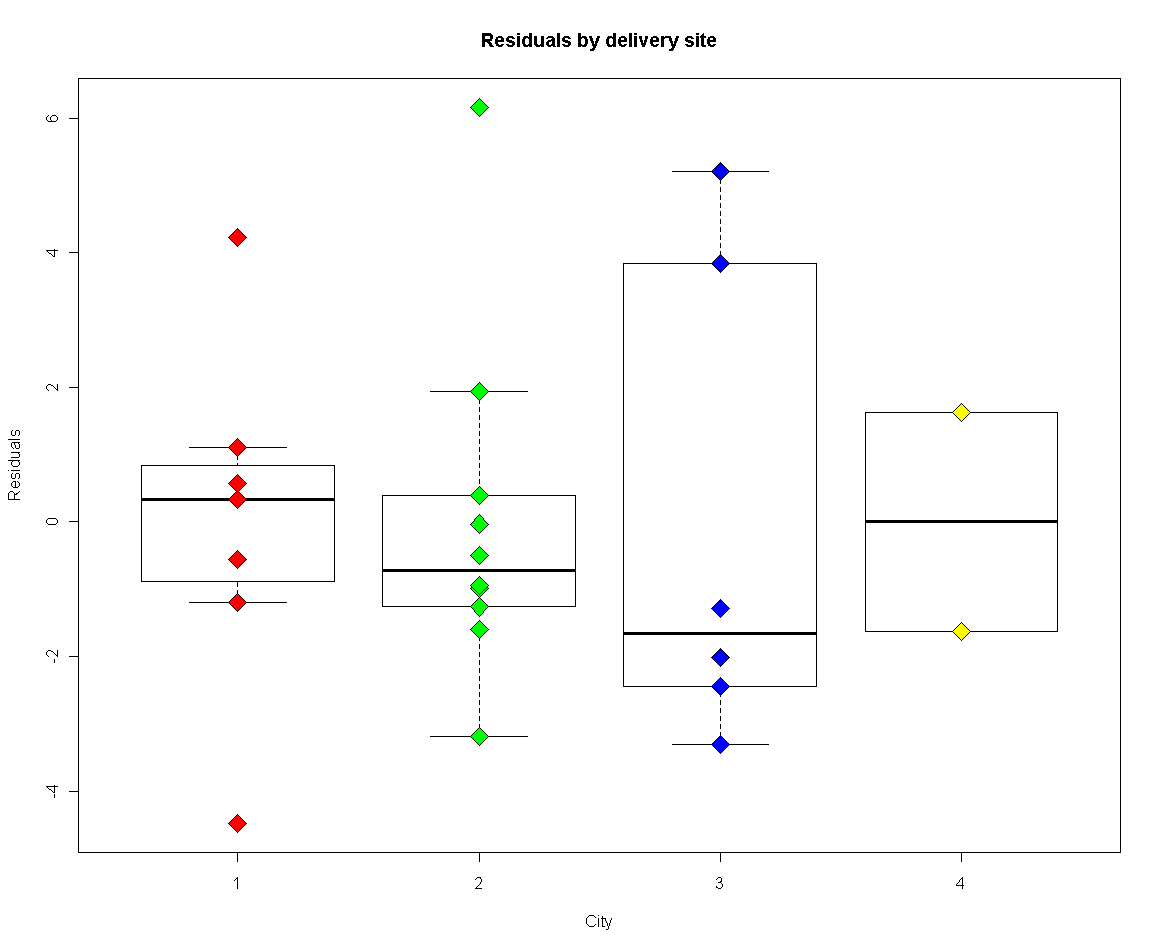
* The estimated parameters of the model is
  1. Is delivery site (city) is an important variable?

Use extra-sum-of-squares method to compare ***delivery.site.model*** model with a reduced model that removes delivery site as an explained variable.



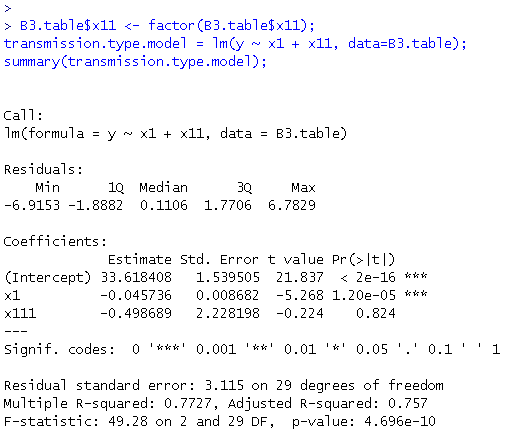
* F = 2.4, p-value = 0.099 suggest delivery site is not an important variable
  1. Analyze the residuals



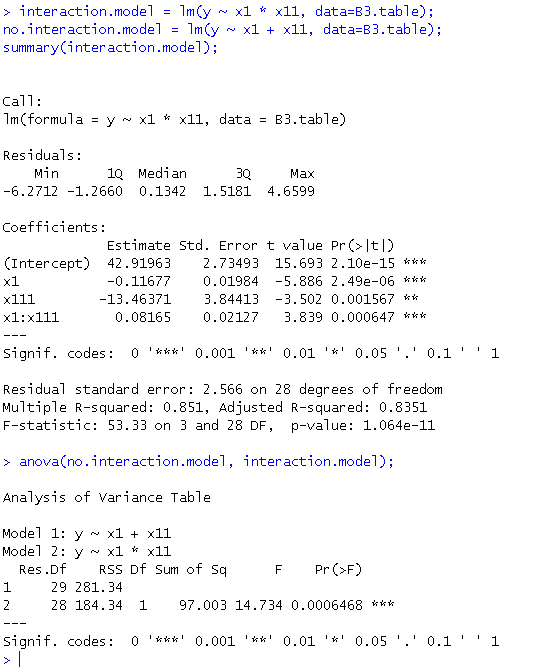


* Observation 9 has large residual

1. **Exercise 8.4**
   1. Fit a linear model relating y to x1 and x11



* The fitted model is
* The ***t*** statistic = -0.22 with p-value = 0.824 suggest the type of transmission does NOT significantly effect the mileage performance.
  1. Fit an interaction model



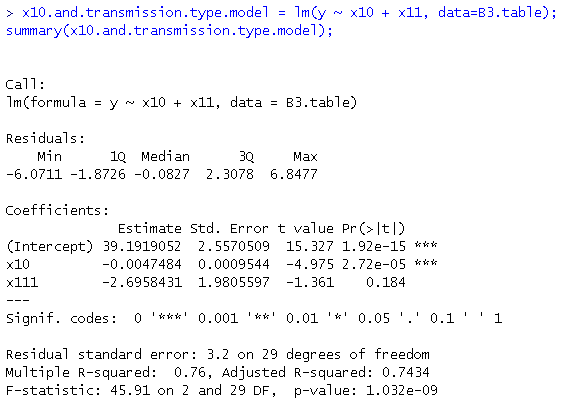
* The fitted model is
* There is significant interaction between engine displacement and the type of transmission
* When transmission is automatic (x11=1),

This suggest that… on average… for one cubic inch increase in engine displacement, miles per gallon decreases by 0.035

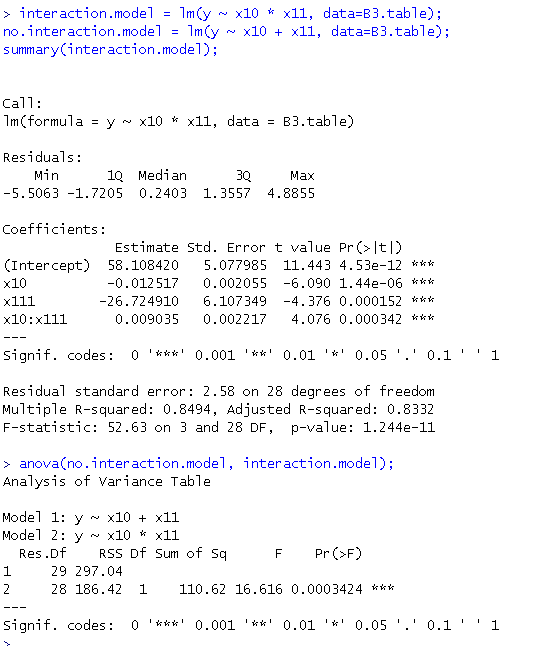
* When transmission is automatic (x11=0),

This suggest that… on average… for one cubic inch increase in engine displacement, miles per gallon decreases by 0.117

1. **Exercise 8.5**
   1. Fit a linear model relating y to x10 and x11



* The fitted model is
* The ***t*** statistic = -1.36 with p-value = 0.184 suggest the type of transmission does NOT significantly effect the mileage performance.
  1. Fit an interaction model



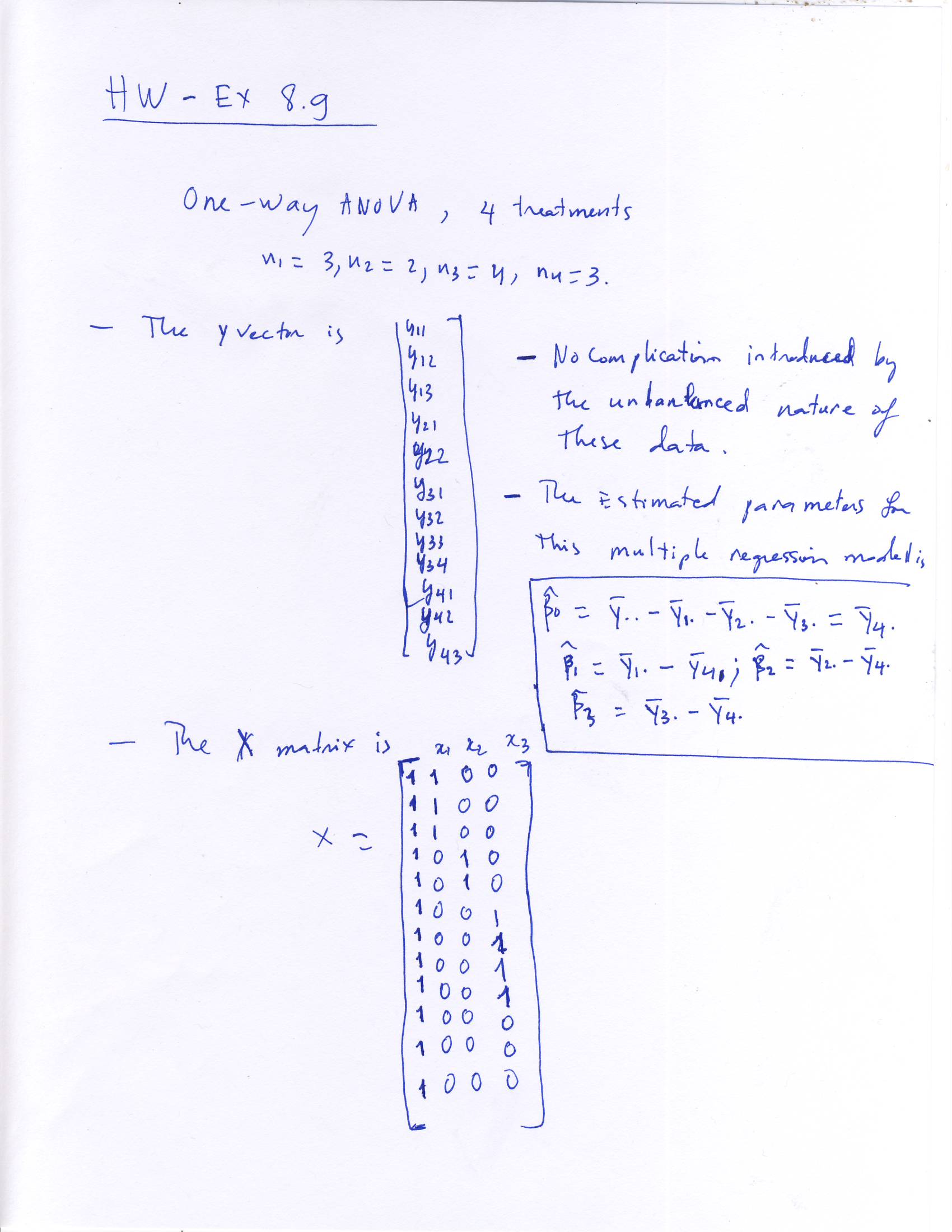
* The fitted model is
* There is significant interaction between vehicle weight and the type of transmission
* When transmission is automatic (x11=1),

Which indicate that… on average… for one lb increase in vehicle weight, miles per gallon decreases by 0.0035.

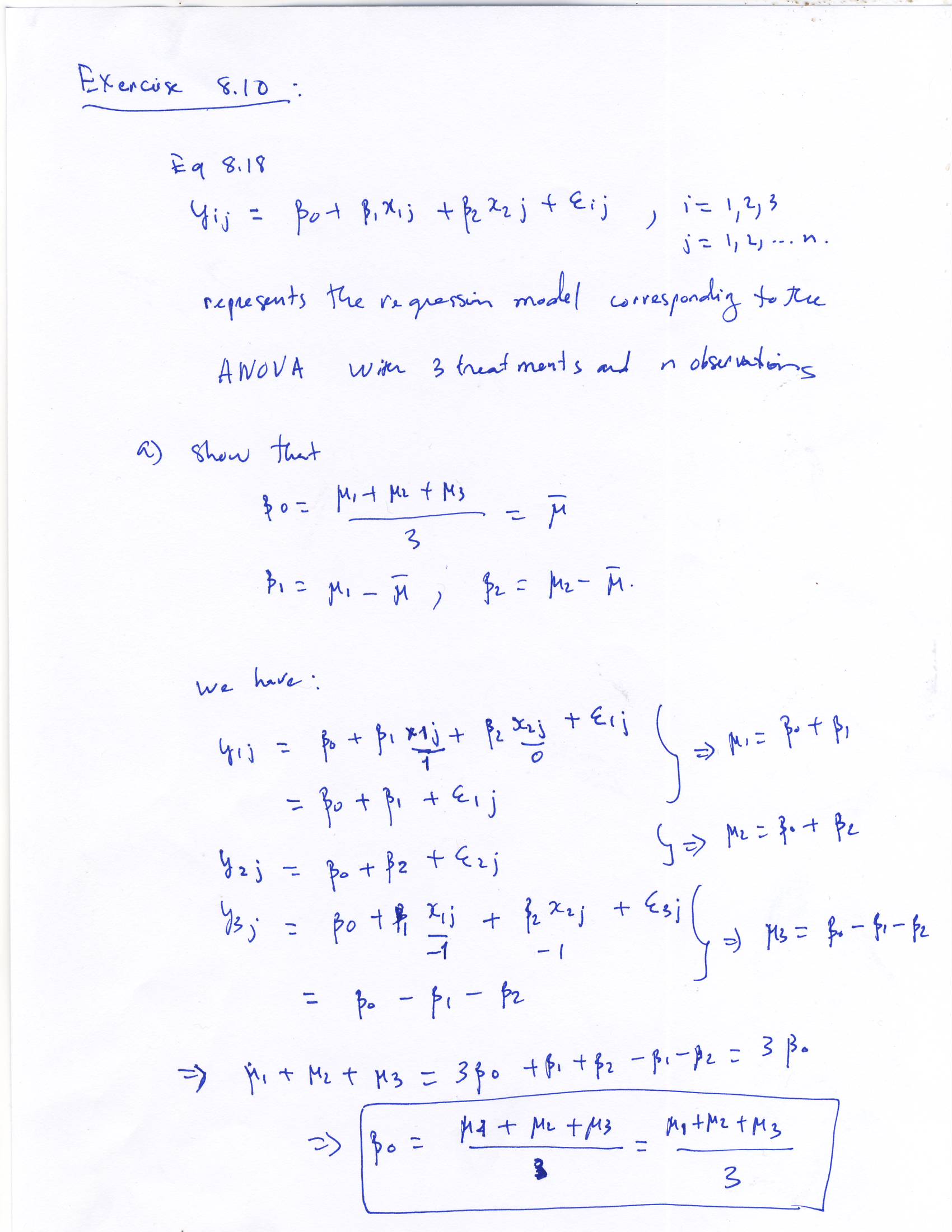
* When transmission is automatic (x11=0),

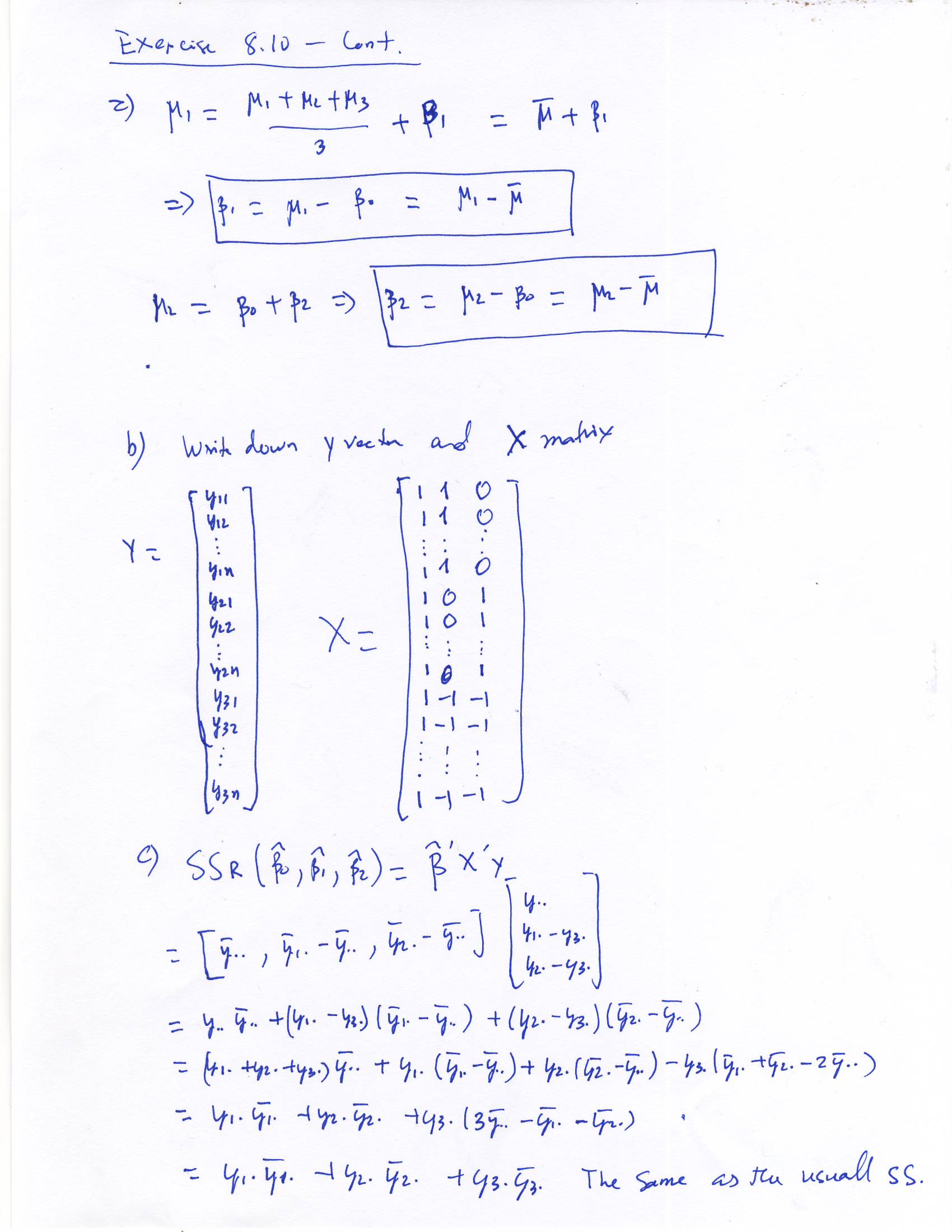
This suggest that… on average… for one lb increase in vehicle weight, miles per gallon decreases by 0.0125

1. **Exercise 8.9**

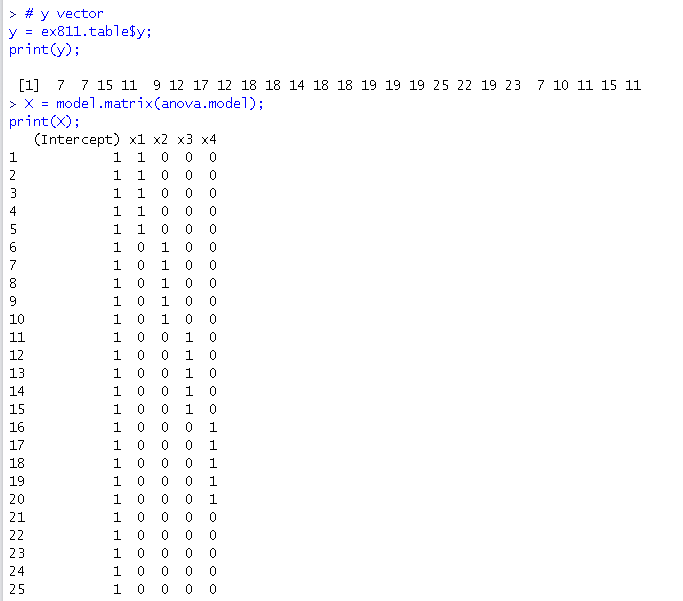


1. **Exercise 8.10**

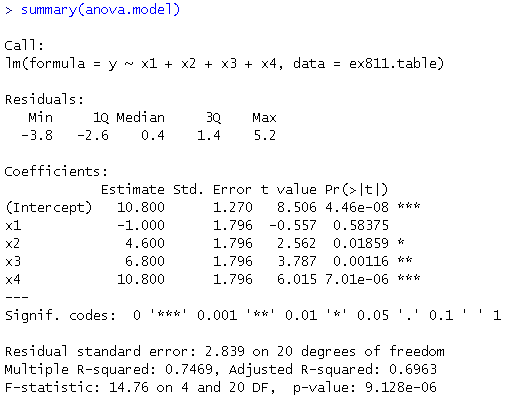




1. **Exercise 8.11**
   1. Write down y vector and X matrix for the corresponding regression model

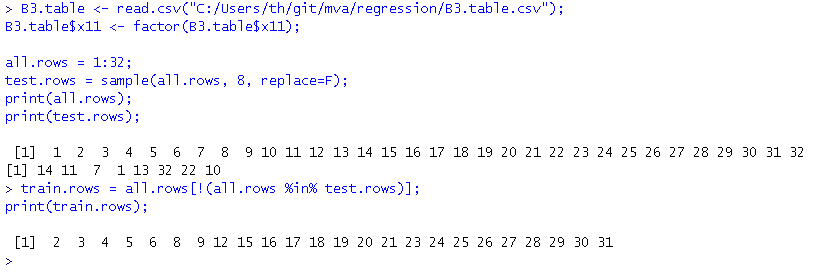


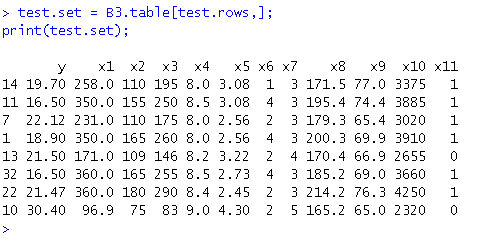
* 1. Find the estimates of the model parameters

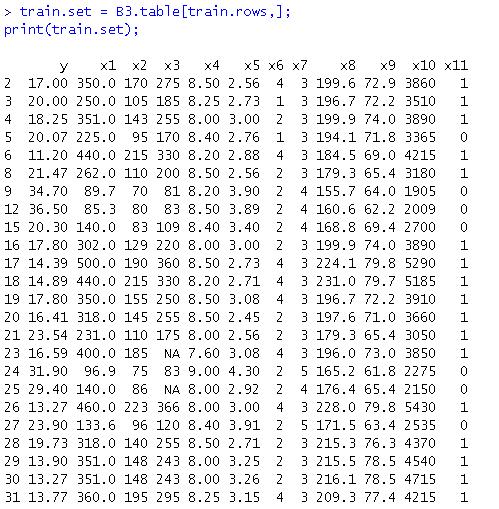


* The estimated
* The estimated
* The estimated
* The estimated
* The estimated
  1. Find point estimate of the difference in mean strength between 15% and 25% cotton
* The estimated difference is
* The estimated difference is
  1. The F statistic F=14.76 with p-value < 0.0001 indicates that the mean tensile strength is not the same for all cotton percentages

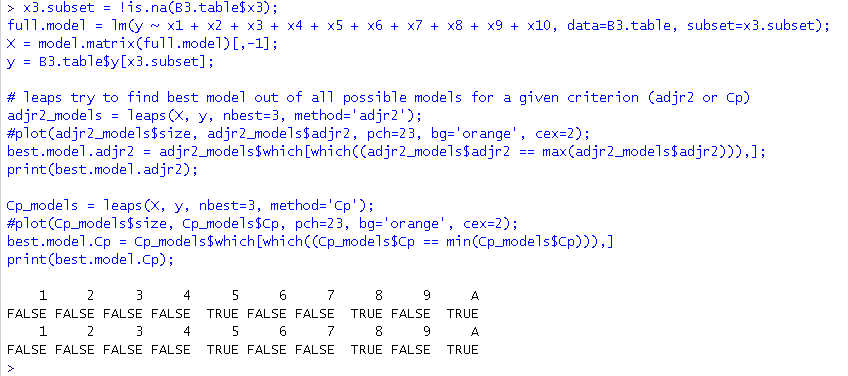
1. **Exercise 10.9**
   1. Sample 8 random rows and call them as **test.set.** The rest of is put into a **training.set**



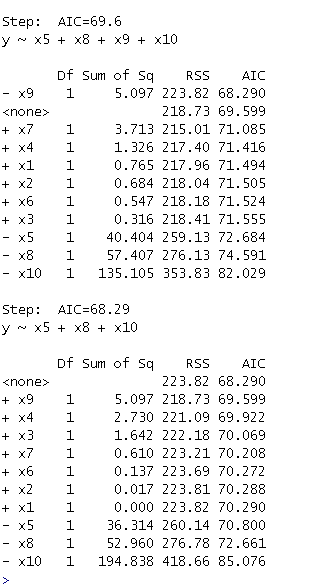




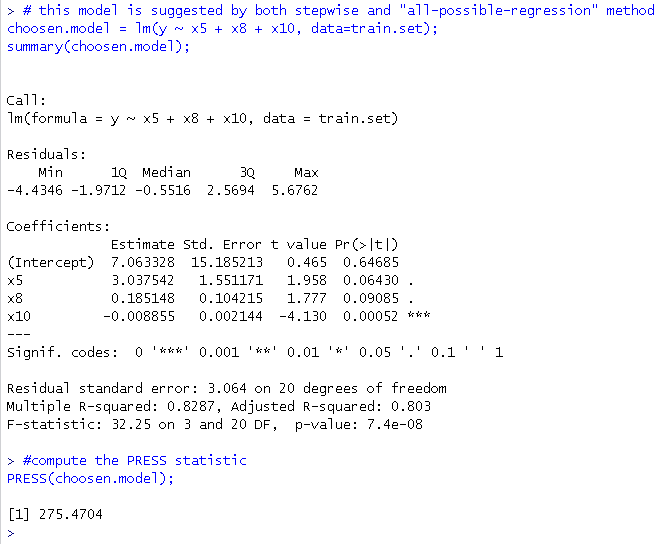
* 1. Find a appropriate regression model
     1. Fit initial (full) model
     2. Use **All Possible Regressions** (leaps package in R) with **adjusted R2** and **Cp** criteria to find the best subset of explanatory variables



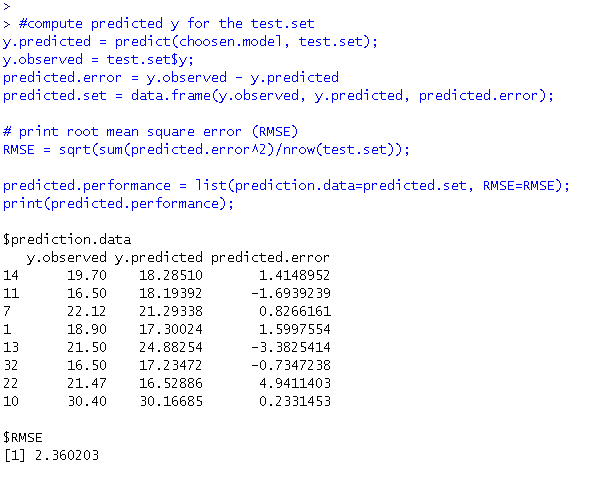
* The best subset in terms of **adjusted R2** is
* The best subset in terms of Mallows **Cp statistic** is
  + 1. Use **stepwise** regression to select explanatory variables



* **stepwise also suggest** 
  1. Fit the model to the training data (train.set) to estimate the parameters



* The fitted model is
* The PRESS statistic of this model is 275.47
  1. Use the fitted model to predict 8 withheld observations in the test.set
     1. Use the training model to predict the unseen data in the test.set
     2. Compute the prediction error
     3. Compute the average prediction error (Root Means Square Error) as a way to assess the model predictive power

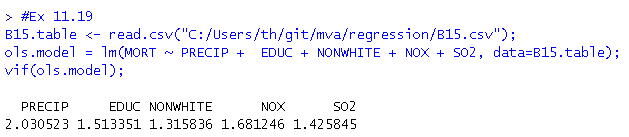


* The model is is predicting pretty well

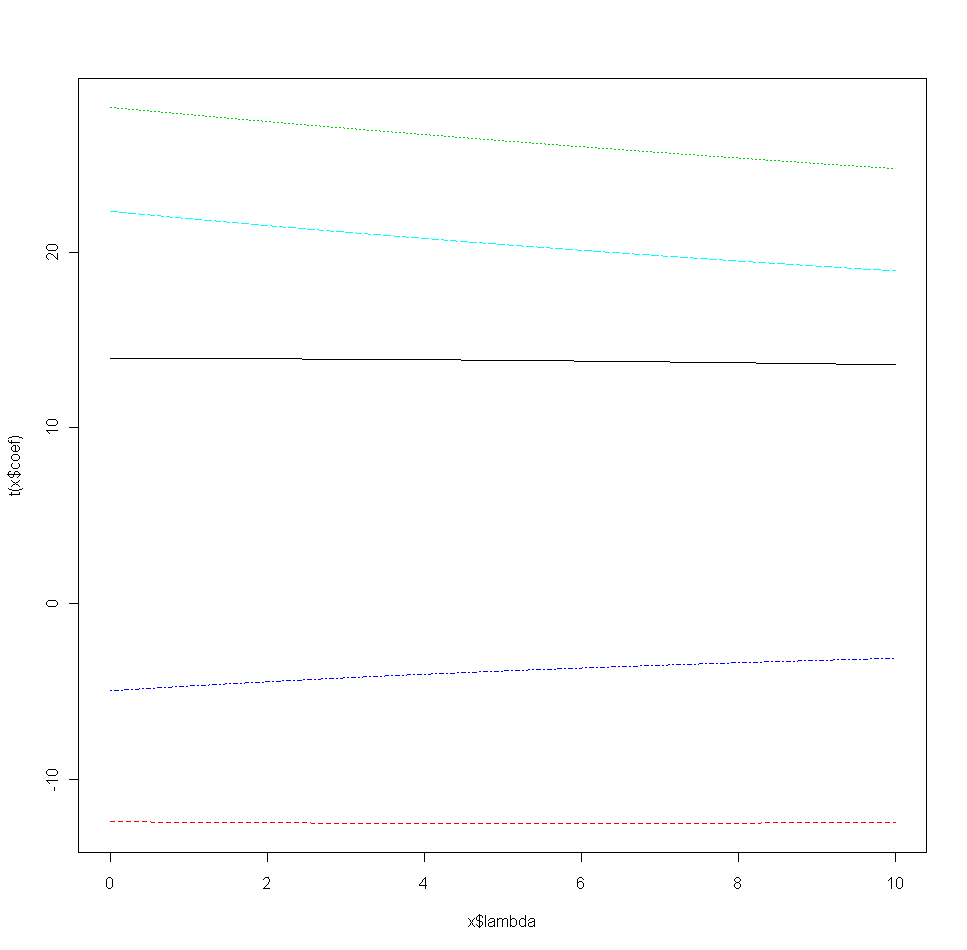
1. **Exercise 11.12**

* There is one large condition index (. Thus there is one dependence in the column of the design matrix X
* The variance decomposition proportions all exceed 0.5, indicating that the following regressors are involved in multicolinearity relationsip

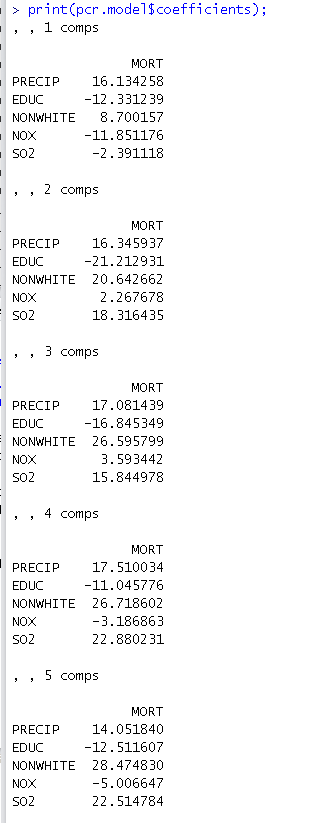
1. **Exercise 11.19**
   1. There is no evidence of multicolinearity in this dataset
      1. Why: Compute the VIFs and see all of these values are under 2.0

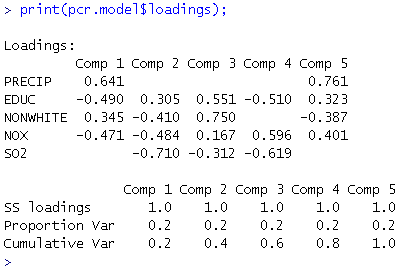


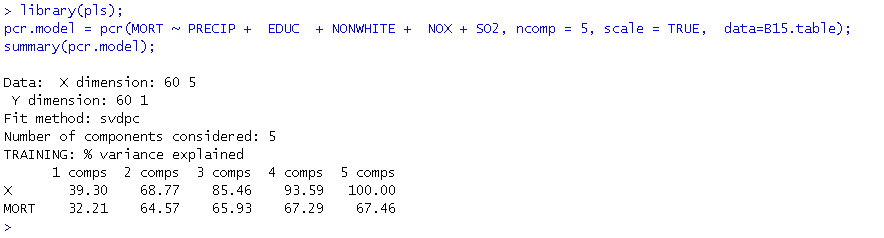
* 1. Perform the ridge trace on the data



* The ridge trace shows flat lines
  1. The ridge trace suggest k=0, hence the estimates of the coefficients for ridge and ordinary Least Square are the same.
  2. Principal component regression gives



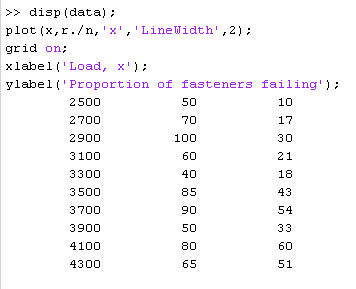




* The principal components regression account for 85.46% of the variation with 3 variables (components)
* While the ordinary least square accounts for only 67.5% of the variation in the model with all 5 variables

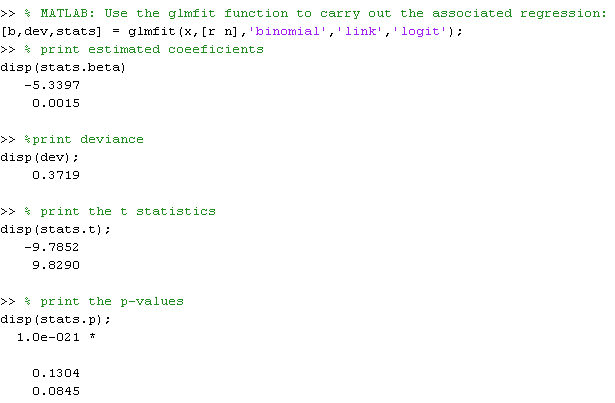
1. **Exercise 14.3**

Data:

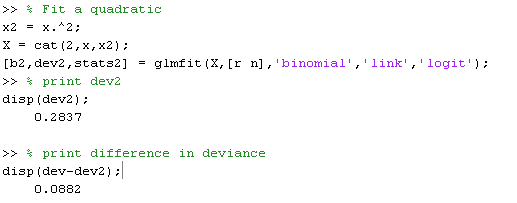




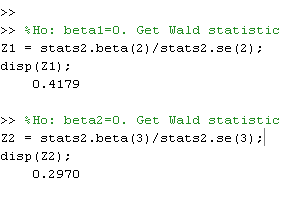
* 1. Fit a logistic model



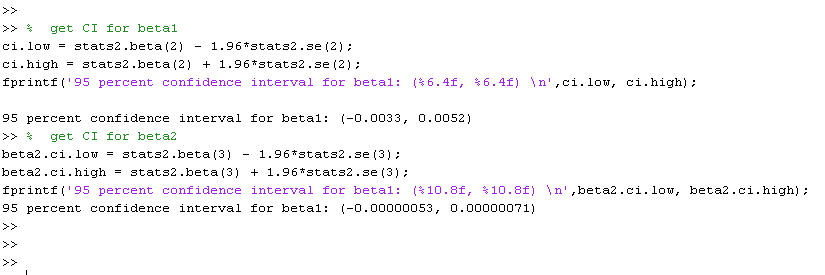
* Thus, the fitted model is
  1. Check the adequacy of the model
* The deviance = 0.3719
* The model is adequate
  1. The difference in deviances is 0.372 – 0.284 = 0.088



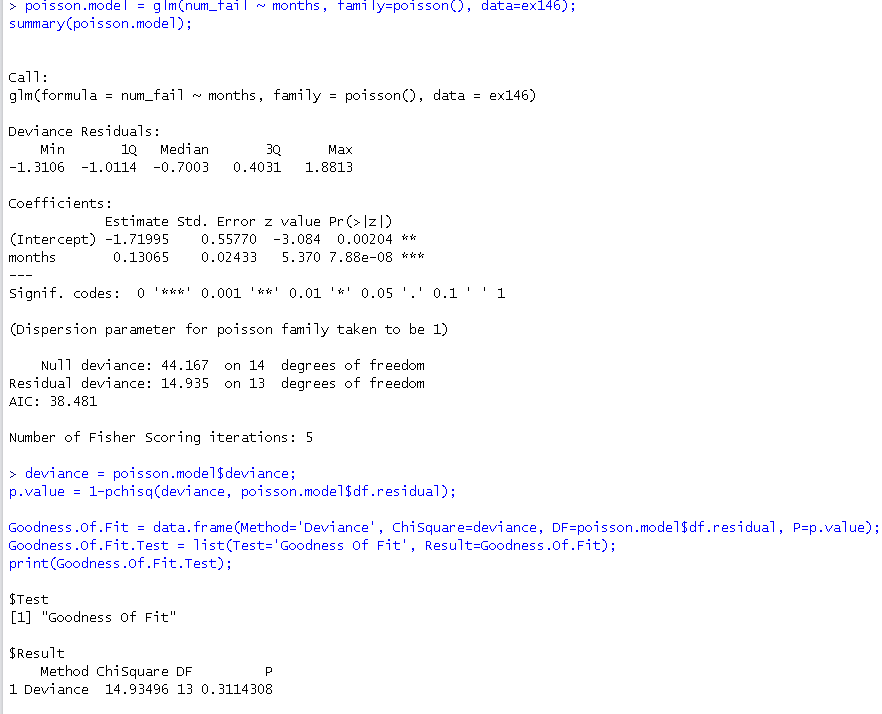
* This small difference in deviances, when comparing to Chi-square 1 d.f. indicate that there is no need for the quadratic term.
  1. Find Wald statistics for each individual parameters for the quadratic model
     1. For H0: β1=0, the Wald statistic Z = 0.42 which is not significant
     2. For H0: β2=0, the Wald statistic Z = 0.30 which is not significant



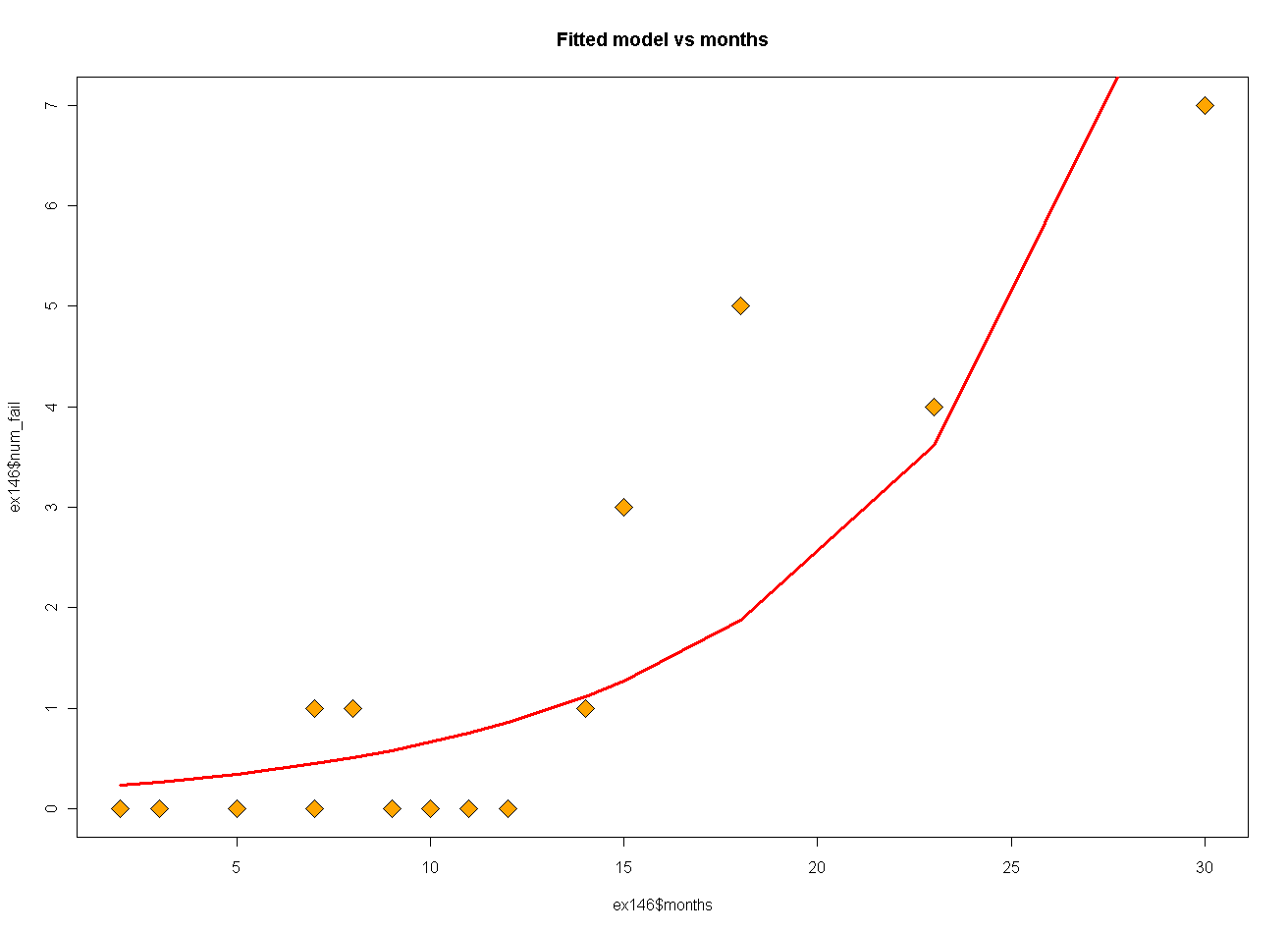
* 1. Find approximate 95% CIs on the model parameters for the model in part C
     1. For β1:, the CI is [-0.0033, 0.0052]
     2. For β2:, the CI is [-0.00000053, 0.00000071]



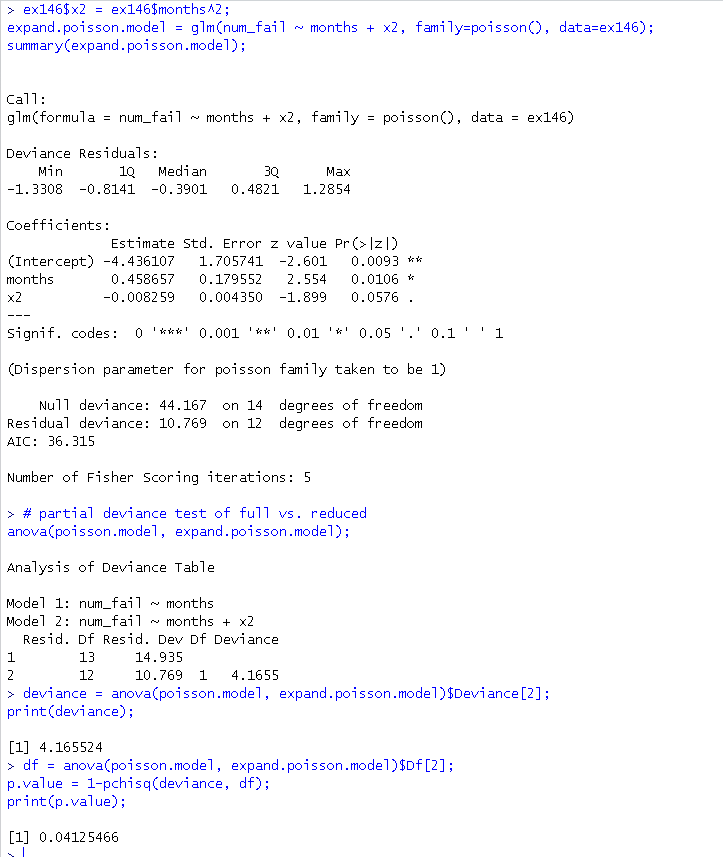
1. **Exercise 14.6**
   1. Fit a logistic model



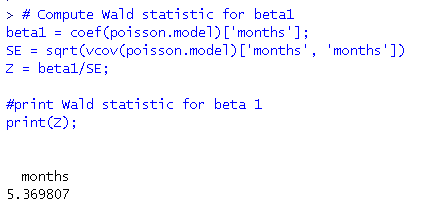
* The fitted model is
  1. The deviance = 14.935, with p-value =0.311 indicate the model is adequate.
  2. Construct graph overlay by fitted model



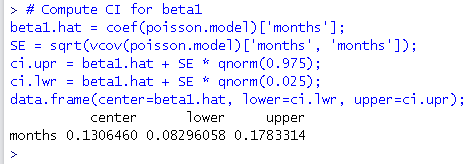
* 1. Expand the model in part a to include a quadratic term



* The expanded model is
* The deviance difference is 4.165 and p-value = 0.04. Thus the quadratic term is significant at the level of alpha = 0.05.
  1. Find the Wald statistics for each individual parameter in the model developed in part A
* For beta1, the Wald statistic Z = 5.37



* 1. Find approximate 95% confidence interval on beta1
* The 95% confidence interval for β1 is (0.083, 0.178)



1. **Exercise 14.12**

**Reconsider the model for aircraft fastener from problem 14.3**

* Construct plots of the deviance residuals from model



* Plot of the deviance residuals from the logistic model developed from problem 14.3
* The normal probability plot of deviance residuals indicate no severe problem with normality
* If time permit, should also plot the deviance residuals vs.

1. **Exercise 14.21**

