MA684 Homework from Class 4

Multiple linear regression

This material is covered in Chapter 8 and 9 of KKM and N. Sections 8.1 – 8.5 give background information on the theory of multiple regression. Sections 8.6 and 8.7 give more practical examples of multiple regressions. Chapter 9, sections 9.1 – 9.3 discuss the overall F-test for multiple regression and the partial F-test for a single variable.

1. (Calculations intended to be done by hand) To develop a prediction equation for home prices in a particular town, a real estate agent collects data on the last 25 houses that were sold in the town. The analysis below focuses on predicting the selling price (in thousands of dollars) from the number of bedrooms in the home and the age of the home (in years).

Some results:

Descriptive data:

|  |  |  |
| --- | --- | --- |
| Variable | Mean | sd |
| Selling price  Number bedrooms  Age | 132.76  3.64  11.56 | 28.61  0.64  7.51 |

ANOVA for the regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | df | SS | MS | F |
| Model  Residual |  | 7068.3  12580.0 |  |  |
| Total |  |  |  |  |

Regression equation predicting job satisfaction:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Parameter  Estimate | Standard  Error | t-statistic | p-value | 95% CI |
| Intercept  Number bedrooms  Age | 73.13  20.32  -1.24 | ---  7.76  0.66 | 2.62  ?? | 0.016  ?? | 4.23, 36.41  ?? |

1A. Complete the ANOVA table for the regression.

* Find R2 ,and give an interpretation of R2.
* Find the value of the F-statistic, and the p-value from the F-statistic.
* What null hypothesis is being tested by the F-statistic?
* Find the adjusted R2.

1B. From the table of slopes and standard errors:

* Give an interpretation for the slope for the number of bedrooms in this model. How is the interpretation of this slope different from the interpretation of a slope in a simple regression?
* Find the t-statistic for the slope for age, and the df and p-value for this t-statistic.
* Find the 95% confidence interval for the slope for age.

1C. To describe the relative importance of the two variables in the regression equation, find the standardized slopes for the number of bedrooms and Age (the standardized slopes can be calculated from the slopes and the standard deviations of the relevant variables). What can you conclude from these standardized slopes?

2. (All calculations intended to be done through R) Using the quality of life data set from Homework 3 (the data set is re-posted with this assignment) fit a multiple linear regression predicting quality of life from age, sex, education, income, and marital status.

Cut-and-paste (or reproduce) the table of slopes, standard errors, t-statistics, and p-values from the regression (this is the standard presentation of the results of a multiple regression analysis).

Based on the results of this regression:

2A. Report and interpret the R2 from this regression. Report the F-statistic from the Anova for this regression, the degrees freedom and the p-value for the F-statistic. What can you conclude from this p-value?

2B. In this regression model, which of the 5 independent variables are significantly associated with quality of life?

2C. Give an interpretation for the slope for age (a continuous predictor) from this regression model. (Not to be graded - give an interpretation for the slope for marital status (a categorical predictor) from this regression model.) Give a confidence interval for the slope of age in this model (do this directly through R, don’t calculate by hand).

2D. To describe the relative importance of the independent variables in this regression model, find the standardized slopes (do this directly through R, don’t calculate by hand). What variable is most strongly associated with quality of life?

2E. (Using R, do not calculate by hand) What is the predicted quality of life for a 30 year old, single male with a college education (16 years of education), and an income of 50 thousand dollars? Give a prediction interval for this quality of life.

What is the predicted quality of life for a 60 year old married female with a college education (16 years) and an income of 50 thousand dollars? Give a confidence interval for the mean quality of life for all such women.

3. (All calculations intended to be done through R) The purpose of this question is to illustrate using multiple regression as an adjusted analysis to control for confounding. We will use a subset of the Framingham Heart Study data from the first exam (back in the 1950s) which is attached with this homework. We are interested in the association between current smoking (CURSMOKE, coded 0 for non-smokers, 1 for smokers) and systolic blood pressure (SYSBP, in mm Hg; higher blood pressure is a risk for heart problems).

3A. Run a t-test to compare the mean systolic blood pressure between smokers and non-smokers (use the equal-variance version of the t-test). Summarize the results in the following table:

Systolic blood pressure in smokers and non-smokers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Non-smokers | Smokers | t-value | p-value |
| n  mean  standard deviation |  |  |  |  |

What is your conclusion from this (unadjusted) analysis – does mean blood pressure differ between smokers and non-smokers?

3B. As a second unadjusted analysis, run a regression predicting systolic blood pressure from the CURSMOKE variable. Summarize the results of the regression in the following table (OK to cut-ant-paste the table from R):

Regression predicting systolic blood pressure from smoking status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Slope | Std Error | t-value | p-value |
| Intercept  Current Smoking |  |  |  |  |

Does this analysis show a significant association between smoking and systolic blood pressure? How does the t-statistic from this analysis compare to the t-statistic from 3A above? Calculate the predicted systolic blood pressures for a non-smoker and a smoker – how do these predicted values compare to the mean blood pressures in 3A? (I think the slope from the regression is the difference in mean systolic blood pressure for smokers vs. non-smokers.)

3C. Smokers are younger and have lower BMI in this data set, and so the differences in systolic blood pressure between smokers and non-smokers may in part be due to differences on these variables (there may be confounding). As an adjusted analysis, run a multiple linear regression predicting SYSBP from AGE, SEX (coded 1 for males, 2 for females) BMI, and CURSMOKE. Summarize the results in the following table (it’s OK to cut-and-paste the table from R):

Regression predicting systolic blood pressure from smoking status

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Slope | Std Error | t-value | p-value |
| Intercept  Age  Sex  BMI  Current Smoking |  |  |  |  |

Based on this analysis, is there a significant association between current smoking and systolic blood pressure? Explain.

3D. Why are the slope and p-value for Current Smoking different in the multiple regression analysis in 3C and the simple regression analysis in 3B? Which analysis do you think is more appropriate?