Math 610:Homework 2

Neal Kuperman

January 13, 2026

ISLP 3.10

This question should be answered using the Carseats data set.

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

(f) How well do the models in (a) and (e) fit the data?

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

(h) Is there evidence of outliers or high leverage observations in the model from (e)?

Note(s)

The data is a simulated data set containing sales of child car seats at 400 different stores. Information on the data set can be found on the ISLP documentation.

Variable	Description
Sales	Unit sales (in thousands) at each location
CompPrice	Price charged by competitor at each location
Income	Community income level (in thousands of dollars)
Advertising	Local advertising budget for company at each location (in thousands of dollars)
Population	Population size in region (in thousands)
Price	Price company charges for car seats at each site
ShelveLoc	Factor with levels Bad, Good, Medium — quality of shelving location
Age	Average age of the local population
Education	Education level at each location
Urban	Factor (No/Yes) — whether store is in urban or rural location
US	Factor (No/Yes) — whether store is in the US or not

Table 1: Carseats Dataset Variables

Solution

ISLP 3.10 (a)

Table 2 shows the results of a linear model to predict Sales using Price, Urban, and US. The table was generated using the `statsmodels` library in Python.

Dep. Variable:	Sales	**R-squared:**	0.239
Model:	OLS	**Adj. R-squared:**	0.234
Method:	Least Squares	**F-statistic:**	41.52
Date:	Tue, 13 Jan 2026	**Prob (F-statistic):**	2.39e-23
Time:	18:41:11	**Log-Likelihood:**	-927.66
No. Observations:	400	**AIC:**	1863.
Df Residuals:	396	**BIC:**	1879.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	13.0435	0.651	20.036	0.000	11.764	14.323
Price	-0.0545	0.005	-10.389	0.000	-0.065	-0.044
Urban_Yes	-0.0219	0.272	-0.081	0.936	-0.556	0.512
US_Yes	1.2006	0.259	4.635	0.000	0.691	1.710

Omnibus:	0.676	**Durbin-Watson:**	1.912
Prob(Omnibus):	0.713	**Jarque-Bera (JB):**	0.758
Skew:	0.093	**Prob(JB):**	0.684
Kurtosis:	2.897	**Cond. No.**	628.

Table 2: OLS Regression Results

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

ISLP 3.14

This problem focuses on the *collinearity* problem.

(a) Perform the following commands in Python:

```
rng = np.random.default_rng(10)
x1 = rng.uniform(0, 1, size=100)
x2 = 0.5 * x1 + rng.normal(size=100) / 10
y = 2 + 2 * x1 + 0.3 * x2 + rng.normal(size=100)
```

The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$? How do these relate to the true β_1, β_2 and β_3? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

(f) Do the results obtained in (c)-(e) contradict each other? Explain your answer.

(g) Suppose we obtain one additional observation, which was unfortunately mismeasured. We use the function np.concatenate() to add this additional observation to each of x1, x2 and y.

```
x1 = np.concatenate([x1, [0.1]])
x2 = np.concatenate([x2, [0.8]])
y = np.concatenate([y, [6]])
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Solution

ISLP 3.15 (a, b, d)

This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \tag{1}$$

Solution

ESL 3.17

Repeat the analysis of Table 3.3 on the spam data discussed in Chapter 1. Include LS, Best Subset, Ridge regression, and Lasso. (skip PCR and PLS)

Solution