

3.7 Exercises

Conceptual

- Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of **sales**, **TV**, **radio**, and **newspaper**, rather than in terms of the coefficients of the linear model.
- Carefully explain the differences between the KNN classifier and KNN regression methods.
- Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.
 - Which answer is correct, and why?
 - For a fixed value of IQ and GPA, males earn more on average than females.
 - For a fixed value of IQ and GPA, females earn more on average than males.
 - For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
 - For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
 - Predict the salary of a female with IQ of 110 and a GPA of 4.0.
 - True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.
- I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.
 - Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

- (b) Answer (a) using test rather than training RSS.
 - (c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
 - (d) Answer (c) using test rather than training RSS.
5. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the i th fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta},$$

where

$$\hat{\beta} = \left(\sum_{i=1}^n x_i y_i \right) / \left(\sum_{i'=1}^n x_{i'}^2 \right). \quad (3.38)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^n a_{i'} y_{i'}.$$

What is $a_{i'}$?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

- 6. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point (\bar{x}, \bar{y}) .
- 7. It is claimed in the text that in the case of simple linear regression of Y onto X , the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.



Applied

- 8. This question involves the use of simple linear regression on the **Auto** data set.
 - (a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

- i. Is there a relationship between the predictor and the response?
 - ii. How strong is the relationship between the predictor and the response?
 - iii. Is the relationship between the predictor and the response positive or negative?
 - iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95 % confidence and prediction intervals?
 - (b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.
 - (c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.
9. This question involves the use of multiple linear regression on the `Auto` data set.
- (a) Produce a scatterplot matrix which includes all of the variables in the data set.
 - (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative. `cor()`
 - (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:
 - i. Is there a relationship between the predictors and the response?
 - ii. Which predictors appear to have a statistically significant relationship to the response?
 - iii. What does the coefficient for the `year` variable suggest?
 - (d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
 - (e) Use the `*` and `:` symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?
 - (f) Try a few different transformations of the variables, such as $\log(X)$, \sqrt{X} , X^2 . Comment on your findings.

10. This question should be answered using the `Carseats` data set.
- Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.
 - Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!
 - Write out the model in equation form, being careful to handle the qualitative variables properly.
 - For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?
 - 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.
 - How well do the models in (a) and (e) fit the data?
 - Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
 - Is there evidence of outliers or high leverage observations in the model from (e)?
11. In this problem we will investigate the t-statistic for the null hypothesis $H_0 : \beta = 0$ in simple linear regression without an intercept. To begin, we generate a predictor `x` and a response `y` as follows.

```
> set.seed(1)
> x=rnorm(100)
> y=2*x+rnorm(100)
```

- Perform a simple linear regression of `y` onto `x`, *without* an intercept. Report the coefficient estimate $\hat{\beta}$, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results. (You can perform regression without an intercept using the command `lm(y~x+0)`.)
- Now perform a simple linear regression of `x` onto `y` without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis $H_0 : \beta = 0$. Comment on these results.
- What is the relationship between the results obtained in (a) and (b)?
- For the regression of Y onto X without an intercept, the t-statistic for $H_0 : \beta = 0$ takes the form $\hat{\beta}/\text{SE}(\hat{\beta})$, where $\hat{\beta}$ is given by (3.38), and where

$$\text{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i \hat{\beta})^2}{(n-1) \sum_{i'=1}^n x_{i'}^2}}.$$

(These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in **R**, that the t-statistic can be written as

$$\frac{(\sqrt{n-1}) \sum_{i=1}^n x_i y_i}{\sqrt{(\sum_{i=1}^n x_i^2)(\sum_{i'=1}^n y_{i'}^2) - (\sum_{i'=1}^n x_{i'} y_{i'})^2}}.$$

- (e) Using the results from (d), argue that the t-statistic for the regression of **y** onto **x** is the same as the t-statistic for the regression of **x** onto **y**.
 - (f) In **R**, show that when regression is performed *with* an intercept, the t-statistic for $H_0 : \beta_1 = 0$ is the same for the regression of **y** onto **x** as it is for the regression of **x** onto **y**.
12. This problem involves simple linear regression without an intercept.
- (a) Recall that the coefficient estimate $\hat{\beta}$ for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X ?
 - (b) Generate an example in **R** with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is *different from* the coefficient estimate for the regression of Y onto X .
 - (c) Generate an example in **R** with $n = 100$ observations in which the coefficient estimate for the regression of X onto Y is *the same as* the coefficient estimate for the regression of Y onto X .
13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.
- (a) Using the `rnorm()` function, create a vector, **x**, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X .
 - (b) Using the `rnorm()` function, create a vector, **eps**, containing 100 observations drawn from a $N(0, 0.25)$ distribution i.e. a normal distribution with mean zero and variance 0.25.
 - (c) Using **x** and **eps**, generate a vector **y** according to the model

$$Y = -1 + 0.5X + \epsilon. \quad (3.39)$$

What is the length of the vector **y**? What are the values of β_0 and β_1 in this linear model?

- (d) Create a scatterplot displaying the relationship between \mathbf{x} and \mathbf{y} . Comment on what you observe.
 - (e) Fit a least squares linear model to predict \mathbf{y} using \mathbf{x} . Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1 ?
 - (f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.
 - (g) Now fit a polynomial regression model that predicts \mathbf{y} using \mathbf{x} and \mathbf{x}^2 . Is there evidence that the quadratic term improves the model fit? Explain your answer.
 - (h) Repeat (a)–(f) after modifying the data generation process in such a way that there is *less* noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.
 - (i) Repeat (a)–(f) after modifying the data generation process in such a way that there is *more* noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.
 - (j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.
14. This problem focuses on the *collinearity* problem.
- (a) Perform the following commands in **R**:


```
> set.seed(1)
> x1=runif(100)
> x2=0.5*x1+rnorm(100)/10
> y=2+2*x1+0.3*x2+rnorm(100)
```

The last line corresponds to creating a linear model in which \mathbf{y} is a function of $\mathbf{x1}$ and $\mathbf{x2}$. Write out the form of the linear model. What are the regression coefficients?
 - (b) What is the correlation between $\mathbf{x1}$ and $\mathbf{x2}$? Create a scatterplot displaying the relationship between the variables.
 - (c) Using this data, fit a least squares regression to predict \mathbf{y} using $\mathbf{x1}$ and $\mathbf{x2}$. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0 , β_1 , and β_2 ? 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 x_1 . 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 x_2 . Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?
- (f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.
- (g) Now suppose we obtain one additional observation, which was unfortunately mis-measured.

```
> x1=c(x1, 0.1)
> x2=c(x2, 0.8)
> y=c(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.

15. 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 $H_0 : \beta_j = 0$?
 - (c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x -axis, and the multiple regression coefficients from (b) on the y -axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x -axis, and its coefficient estimate in the multiple linear regression model is shown on the y -axis.
 - (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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$