Homework 7

MTH 3270 Data Science

Due Mon., Apr. 13

Read These Chapters of the Book	**Then Do These Exercises**
7 Appendix E	No more exercises from this chapter E.3*, E.5, Problems 1 and 2 (below)

* For **Problem E.3**, after fitting the regression model using `lm()` and saving the result as, say, `my.reg`, the p-value for the `age` coefficient is labeled `Pr(>|t|)` in the output of:

```
summary(my.reg)
```

and the confidence interval for the `age` coefficient is obtained via:

```
confint(my.reg)
```

A p-value greater than 0.05 indicates that there **isn't** a statistically significant relationship between `wt` and `age`. A confidence interval that contains the value zero also indicates that `wt` and `age` **aren't** related.

1 The `Gestation` data set from the `"mdsr"` package (used in Exercise E.3) contains birth weight, date, and gestational period data collected as part of the Child Health and Development Studies. Information about the baby's parents – age, education, height, weight, and whether the mother smoked is also recorded.

For more information about the `Gestation` data set, see Exercise E.3 or type:

```
library(mdsr)
? Gestation
```

The goal is to model the weight of the infants (`wt`, in ounces) using variables including length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`ht`), and mother's pregnancy weight in pounds (`wt.1`).

a) Fit the multiple regression model to the data. Write out the **equation** of the fitted multiple regression model.

b) What is the **predicted** weight of an infant born after a gestation period of **280** days to a **27** year old, **64** inch tall, **130** pound mother?

c) By how much does the **weight** of an **infant** increase for each **1-day** increase in the **gestation** period (holding mother's age, height, and weight constant)?

d) The ***p-value*** for a coefficient (labeled `Pr(>|t|)` in the `summary()` output) is a measure of the strength of evidence that the explanatory variable is related to the response variable – a *smaller* p-value indicates *stronger* evidence.

Which of the four explanatory variables shows the *strongest* evidence for a relationship to infant weight? Which shows the *weakest* evidence?

e) This data set contains missing values – there are 1,236 observations (rows) in the data set, but some rows contain `NA`s. If a row contains `NA` for any one of the five variables used to build the model, `lm()` omits that entire row. How many rows were omitted? **Hint**: Look at the `summary()` output.

2 Refer to the `cdi` data set, described in Class Notes 6 and stored in the file **CDI.txt** on the course website.

a) For each **geographic region**, carry out a multiple regression analysis with response variable **number of serious crimes** (Y) and three explanatory variables: **population density**(X_1, total population divided by land area), **per capita personal income** (X_2), and **percent high school graduates** (X_3).

You can use `mutate()` from the `"dplyr"` package to create the **population density** variable.

Write out the **equations** of the four fitted multiple regression models.

Hint: You may want to use `group_by()` and `do()` (from `"dplyr"`) along with `lm()` followed by `summary()`.

b) Are the equations of the fitted models similar for the four regions? Discuss.

c) Obtain the $\sqrt{\textbf{MSE}}$ and $\boldsymbol{R^2}$ values for each region. Are these measures of model fit similar for the four regions? Discuss.

Hint: The **square root** of the **MSE** and the $\boldsymbol{R^2}$ value are reported in the output of `summary()` after passing it an *lm* object. They're labeled `Residual standard error` and `Multiple R-squared`, respectively.

d) Obtain the **residuals** for each fitted model and plot them in side-by-side boxplots. Interpret your plots and state your findings.