Homework #3 (Solutions)

STAT-UB.0003: Regression and Forecasting Models

1. MBS, Ex. 11.27

> **Solution:** The dataset in graph b would have the smallest variance; the points deviate least from the linear trend.

2. MBS, Ex. 11.58. The problem asks for evidence of a "decrease" (one-sided alternative). We did not cover one-sided alternatives in class. Use a two-sided alternative instead (testing for evidence of "change".)

Solution: Here is the Minitab linear regression fit of `MASS` vs. `TIME`:

```
  Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.857257  85.33%     84.64%      80.83%

Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant    5.221    0.296    17.64    0.000
TIME      -0.1140   0.0103   -11.05    0.000  1.00

Regression Equation

MASS = 5.221 - 0.1140 TIME
```

To test the null hypothesis of "no change", $H_0 : \beta_1 = 0$, we check the p-value for the coefficient of `TIME`. Minitab reports this value as 0.000, which means that it is below 0.001. There is extremely strong evidence that expected mass changes as time increases.

A 95% confidence interval for β_1, the amount that expected mass changes for every unit increase in time is given by

$$\begin{aligned}\hat{\beta}_1 \pm t_{.025,n-2}\text{SE}(\hat{\beta}_1) &= (-0.1140) \pm (2.080)(0.0103) \\ &= -0.1140 \pm 0.0422 \\ &= (-0.1562, -0.0718).\end{aligned}$$

Note: you get full credit if you use the approximation $t_{.025,n-2} \approx 2$.

3. Financial institutions charge, in general, different interest rates on their loans.

A financial analyst was interested in the relationship between Y and X, where

Y = the default rate per 1000 loans (i. e. the number of loans that default per 1000 loans)

X = the interest rate (%) on a loan.

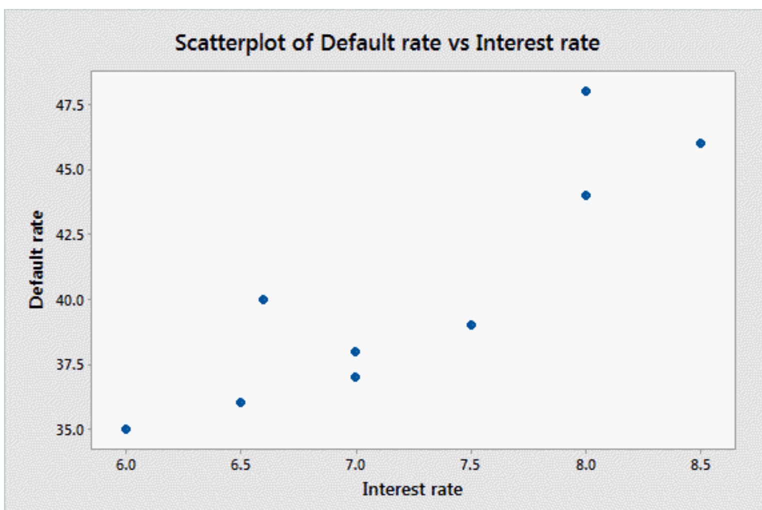
She collected data on a random sample of financial institutions, that is for each institution in the sample she recorded the interest rate charged by that institution and the number of defaults per 1000 loans given by that institution. Answer the following questions using your own Minitab output. The data are posted on the course website in the `DefaultInt` data file.

(a) What is a reasonable population for this dataset?

Solution: The interest and default rates of all financial institutions.

(b) Use Minitab to make a scatter plot of the data (*Graph* ⇒ *Scatterplot*; choose "Simple"; then set the X and Y variables in row 1 of the table). Does the scatter plot indicate an approximate linear relationship between Y and X?

Solution: The scatterplot indicates an approximate linear relationship:



(c) Fit a regression model of Y on X (*Stat* ⇒ *Regression* ⇒ *Regression* ⇒ *Fit Regression Model*; use the "Responses" and the "Continuous Predictors" boxes to select that Y and X variables). Interpret the coefficients in the fitted model.

Solution: Minitab gives the following fit:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
```

```
2.31404  77.95%     74.80%      66.47%

Coefficients

Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant        4.57     7.23     0.63    0.548
Interest rate  4.945    0.994     4.97    0.002  1.00

Regression Equation

Default rate = 4.57 + 4.945 InterestÂărate
```

An increase in "Interest rate" by one unit (1%) is associated with a decrease in expected "Default rate" by 4.945 units (4.945 defaults per thousand loans).

The intercept (4.57) does not have a direct interpretation since "Interest rate $= 0$" is outside the range of the data.

(d) Explain the meanings of the fitted coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the true coefficients (β_0 and β_1). How are they different?

Solution: The fitted coefficients are the least squares fits to the sample; these are estimates of the true coefficients, which are the least squares fits to the population.

(e) Form a 95% confidence interval for β_1.

Solution:

$$\begin{aligned}\hat{\beta}_1 \pm t_{.025,n-2}\mathrm{SE}(\hat{\beta}_1) &= (4.945) \pm (2.365)(0.994) \\ &= 4.945 \pm 2.351 \\ &= (2.594, 7.296)\end{aligned}$$

Note: you get full credit if you use the approximation $t_{.025,n-2} \approx 2$.

(f) Find the standard error of regression (s) and interpret it.

Solution: The standard error of regression is $s = 2.31404$. Roughly 95% of the Y values are within $2s$ of the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$;

(g) Use *Stat* ⇒ *Basic Statistics* ⇒ *Display Descriptive Statistics*, then enter the response variable in the "Variables" box. Report the sample standard deviation of the default rates (s_Y). What are the interpretations of s and s_Y? How are they different?

> **Solution:** Here are the descriptive statistics from Minitab:
>
> ```
> Variable      N   Mean  SE Mean  StDev
> Default rate  9  40.33     1.54   4.61
> ```
>
> The standard deviation of the default rate is $s_Y = 4.61$. Roughly 95% of the Y values are within $2s$ of the mean $\bar{y} = 40.33$. The values s and s_Y are both standard deviations; s is the standard deviation *after adjusting for* X; s_Y is the standard deviation without adjusting for X (for a randomly chosen Y value from the sample).

(h) Is there a statistically significant linear relationship between Default rate and Interest rate? Test at level $\alpha = 0.01$? (State H_0 and H_a, and your conclusion.)

> **Solution:** The null and alternative hypotheses are
>
> $$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)}$$
> $$H_a : \beta_1 \neq 0 \quad \text{(linear relationship)}$$
>
> The p-value is $p = 0.002$. Since $p < \alpha$, we reject the null hypothesis. There is a statistically significant linear relationship.

(i) What is the meaning of the p-value associated with $\hat{\beta}_1$? (Provide a one-sentence description.)

> **Solution:** If there were no linear relationship between default rate and interest rate, then the chance of seeing data at least as extreme as observed would be 0.002.