Homework #11

Math 455, Spring 2026

Due Friday, May 1

Problem 1: In the March 2016 Current Population Survey, there are n = 47233 observations of the variables Y, hourly earnings, C, which is 1 if the subject has a college degree, and 0 otherwise, and F, which is 1 if the subject is female and 0 otherwise. You fit 3 models to the data:

$$\text{(model 1)} : \log Y = \beta_0 + \beta_1 C$$

$$\text{(model 2)} : \log Y = \beta_0 + \beta_1 C + \beta_2 F$$

$$\text{(model 3)} : \log Y = \beta_0 + \beta_1 C + \beta_2 F + \beta_3 CF$$

(a) For model 1, you obtain the coefficient estimate $\hat{\beta}_1 = 0.1056$. Write a sentence interpreting this result, understandable to the non-statistician. *Hints: for small x, $e^x \approx 1 + x$. It is better to say, "A is 17% more than B" than to say "A differs from B by a factor of 1.17".*

(b) Does the coefficient estimate in (a) for model 1 represent the returns to getting a college degree? Why or why not?

(c) Using the coefficient estimates $\hat{\beta}_1 = 0.1089$, $\hat{\beta}_2 = -0.2520$ model 2, describe the difference in hourly earnings between men having a college degree and men not having a college degree.

(d) Using the coefficient estimates $\hat{\beta}_1 = 0.1089$, $\hat{\beta}_2 = -0.2520$ model 2, describe the difference in hourly earnings between women having a college degree and women not having a college degree.

(e) Using the coefficient estimates $\hat{\beta}_1 = 0.1089$, $\hat{\beta}_2 = -0.2520$ model 2, describe the difference in hourly earnings between women having a college degree and men having a college degree.

(f) Do these coefficient estimates allow us to conclude that discrimination is the cause of the difference in earnings?

(g) For model 3, you obtain the coefficient estimates $\hat{\beta}_1 = 0.1063$, $\hat{\beta}_2 = -0.3420$, $\hat{\beta}_3 = 0.0063$, with standard errors 0.0018, 0.0260, and 0.0018 respectively. Use this to test the null hypothesis that the change in hourly earnings associated with having a college degree is the same for females and males. *Hint: The "change in hourly earnings associated with having a college degree for males" is the difference in average earnings between males with a college degree and males without a college degree. Similarly for females. You*

are asked to test the hypothesis that the difference between these differences is zero. This boils down to looking at one regression coefficient.

Answer the following multiple choice questions and give a short reason for each of your answers:

Problem 2: In a simple linear regression model, the predicted value of y, when x has a value x^*, is

A. $\alpha + \beta x^* + \epsilon$;

B. $\alpha + \beta x^*$;

C. $\hat{\alpha} + \hat{\beta} x^* + \epsilon$;

D. $\hat{\alpha} + \hat{\beta} x^*$;

Problem 3: In a linear regression analysis with the usual assumptions, which one of the following quantities is the same for all individual units in the analysis?

A. Leverage h_{ii};

B. $s.e.\{Y_i\}$;

C. $s.e.\{e_i\}$;

D. $s.e.\{\hat{Y}_i\}$;

Problem 4: A regression line is used for all of the following except one. Which one is not a valid use of a regression line?

A. to estimate the average value of Y at a specified value of X;

B. to predict the value of Y for an individual, given that individual's X-value;

C. to estimate the change in Y for a one-unit change in X;

D. to determine if a change in X causes a change in Y .

Problem 5: Which choice is **not** an appropriate description of $\hat{Y}$ in a regression equation?

A. Estimated response;

B. Predicted response;

C. Estimated average response;

D. Observed response.

Problem 6: Which of the following is the **best** way to determine whether or not there is a statistically significant linear relationship between two quantitative variables?

A. Compute a regression line from a sample and see if the sample slope is 0;

B. Compute the correlation coefficient and see if it is greater than 0.5 or less than -0.5;

C. Conduct a test of the null hypothesis that the population slope is 0;

D. Conduct a test of the null hypothesis that the population intercept is 0.

Problem 7: Which of the following methods is the most appropriate for testing H_0: $\beta_k = 0$ versus H_a: $\beta_k > 0$?

A. A t-test;

B. An F-test;

C. A test of a full versus reduced model;

D. All of the above are equally good.

Problem 8: Which of the following is **not** a valid null hypothesis?

A. $H_0 : \beta_1 = 0$;

B. $H_0 : \beta_1 = \beta_2$;

C. $H_0 : \hat{\beta}_1 = \hat{\beta}_2 = 0$;

D. All of the above *are* valid null hypotheses.

Problem 9: Which of the following can never be 0 (unless the population standard deviation $\sigma = 0$)?

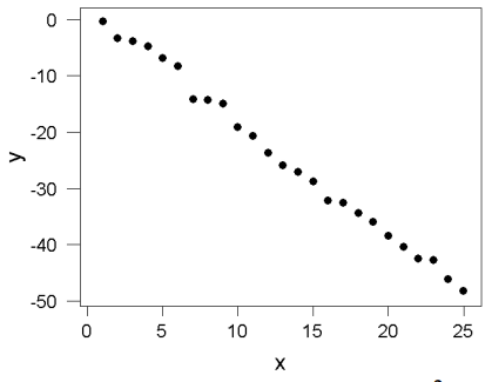
A. The estimated intercept, $\hat{\beta}_0$;

B. A studentized deleted residual, t_i;

C. The variance of the prediction error, $\sigma^2\{pred\}$;

D. The estimate of $E[Y_h]$, $\hat{Y}_h$

Problem 10: Shown below is a scatterplot of Y versus X Which choice is



most likely to be the approximate value of R^2?

A. -99.5%;

B. 2%;

C. 50.0%;

D. 99.5%.

Problem 11: Suppose you have four possible predictor variables (X_1, X_2, X_3, and X_4) that could be used in a regression analysis.

You run a forward selection procedure, and the variables are added as follows: Step 1: X_2 Step 2: X_4 Step 3: X_1 Step 4: X_3 In other words, after Step 1, the model is $E\{Y\} = \beta_0 + \beta_1 X_2$ After Step 2, the model is $E\{Y\} = \beta_0 + \beta_1 X_2 + \beta_2 X_4$, and so on.

You also run an all subsets regression analysis using R^2 as the criterion for the "best" model for each possible number of predictors (1, 2, 3, 4). Would the same models result from this analysis as from the forward stepwise procedure? In other words, would "all subsets regression" *definitely* identify the following as the best models for 1, 2, 3, and 4 variables?

A. β_0 + 1 variable, best model would be $E\{Y\} = \beta_0 + \beta_1 X_2$

B. β_0 + 2 variable, best model would be $E\{Y\} = \beta_0 + \beta_1 X_2 + \beta_2 X_4$

C. β_0 + 3 variable, best model would be $E\{Y\} = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1$

D. β_0 + 4 variable, best model would be $E\{Y\} = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1 + \beta_4 X_3$

For each of the cases A-D, you should indicate whether the answer is yes or no, and give a short explanation of why that answer is correct.

Problem 12: The following code runs all subsets regression on some data. Using the output given below, write down the top 2 models according to R^2, adjusted R^2, and BIC. These models should be in terms of y and $x1$ to $x8$, and your answer should be given in Wilkinson-Rogers notation.

```
SumX = summary(regsubsets(y ~x1+x2+x3+x4+x5+x6+x7+x8, data=SurgicalUnit,nbest=2))
outputMat = cbind(SumX$which, SumX$rsq,SumX$adjr2, SumX$bic)
colnames(outputMat) <- c(colnames(SumX$which), "rsq","adjr2","bic")
View(outputMat)
```

	(Intercept)	x1	x2	x3	x4	x5	x6	x7	x8	rsq	adjr2	bic
1	1	0	0	1	0	0	0	0	0	0.3343453	0.3215443	-13.99918
1	1	0	0	0	1	0	0	0	0	0.4545389	0.4440492	-24.75271
2	1	0	0	0	1	0	0	0	1	0.5667409	0.5497504	-33.19970
2	1	0	0	1	1	0	0	0	0	0.5615945	0.5444021	-32.56204
3	1	0	1	1	0	0	0	0	1	0.6734734	0.6538818	-44.48324
3	1	1	1	1	0	0	0	0	0	0.6841275	0.6651752	-46.27456
4	1	1	1	1	0	0	0	0	1	0.7713565	0.7526917	-59.73701
4	1	0	1	1	1	0	0	0	1	0.7501457	0.7297495	-54.94646
5	1	1	1	1	0	1	0	0	1	0.7738184	0.7502578	-56.33261
5	1	1	1	1	1	0	0	0	1	0.7809054	0.7580831	-58.05170
6	1	1	1	1	1	0	1	0	1	0.7811874	0.7532539	-54.13226
6	1	1	1	1	1	1	0	0	1	0.7814169	0.7535127	-54.18892
7	1	1	1	1	1	1	0	1	1	0.7814814	0.7482286	-50.21587
7	1	1	1	1	1	1	1	0	1	0.7817703	0.7485615	-50.28731
8	1	1	1	1	1	1	1	1	1	0.7818387	0.7430544	-46.31525

A. Top 2 according to R^2

B. Top 2 according to adjusted R^2

C. Top 2 according to BIC

Problem 13: The following code shows the last steps of forward regression, backward, and stepwise regression. Answer the two questions following.

Forward

```
Step:  AIC=587.31
y ~ x4 + x8 + x3 + x2 + x1

       Df Sum of Sq     RSS    AIC
<none>              1833716 587.31
+ x5    1    4280.8 1829436 591.17
+ x6    1    2360.3 1831356 591.23
+ x7    1     217.0 1833499 591.29

Call:
lm(formula = y ~ x4 + x8 + x3 + x2 + x1, data = SurgicalUnit)

Coefficients:
(Intercept)          x4          x8          x3          x2          x1
  -1178.330       58.064      317.848       9.748       8.924      59.864
```

Backward

```
Step:  AIC=585.62
y ~ x1 + x2 + x3 + x8

       Df Sum of Sq     RSS    AIC
<none>              1913636 585.62
- x8    1    730065 2643701 599.09
- x1    1    819235 2732871 600.88
- x2    1   1523696 3437332 613.26
- x3    1   2885539 4799175 631.28

Call:
lm(formula = y ~ x1 + x2 + x3 + x8, data = SurgicalUnit)

Coefficients:
(Intercept)           x1           x2           x3           x8
   -1334.42        81.44        10.13        11.24       312.78
```

Stepwise

```
Step:  AIC=585.62
y ~ x8 + x3 + x2 + x1

       Df Sum of Sq     RSS    AIC
<none>              1913636 585.62
+ x4    1     79920 1833716 587.31
+ x5    1     20604 1893032 589.03
+ x6    1     14420 1899216 589.20
+ x7    1       527 1913109 589.60
- x8    1    730065 2643701 599.09
- x1    1    819235 2732871 600.88
- x2    1   1523696 3437332 613.26
- x3    1   2885539 4799175 631.28

Call:
lm(formula = y ~ x8 + x3 + x2 + x1, data = SurgicalUnit)

Coefficients:
(Intercept)           x8           x3           x2           x1
   -1334.42       312.78        11.24        10.13        81.44
```

A. Explain why the forward regression model has $x4$ while backward and stepwise do not.

B. Say which of the models should be preferred, and explain briefly why.