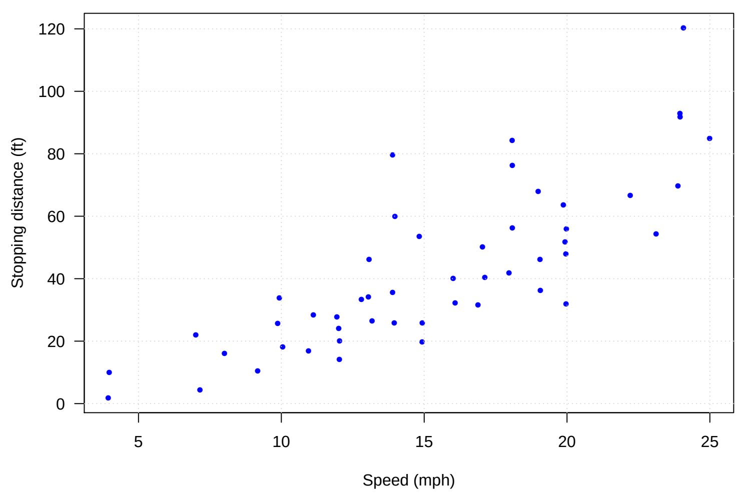
STAT 8020 Statistical Methods II
Practice Exam I

Instructor: Whitney Huang (wkhuang@clemson.edu)

September 21, 2019

Problem 1

A researcher is interested in the relationship between the speed of cars (`speed`) and the distances taken to stop (`dist`). She performed an experimental study (way back in 1920) and the data set is presented in the scatterplot below.



1. Let's use X to denote `speed` and use Y to denote `dist`. Write the form of the corresponding simple linear regression.

2. Use the fact that $\sum_{i=1}^{n=50}(X_i - \bar{X})(Y_i - \bar{Y}) = 5384.40$, $\sum_{i=1}^{n=50}(X_i - \bar{X})^2 = 1370.51$, $\bar{X} = 15.40$, and $\bar{Y} = 42.99$ to compute the estimated slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$

3. Write down the least squares regression line and compute the fitted value with `speed` $= 15$mph.

4. Using the information SSE $= \sum_{i=1}^{50}(Y_i - \hat{Y}_i)^2 = 11362.39$ to compute $\hat{\sigma}$

5. Construct the 95% confidence interval (using $t(0.975, df = 48) = 2.01$) for β_1

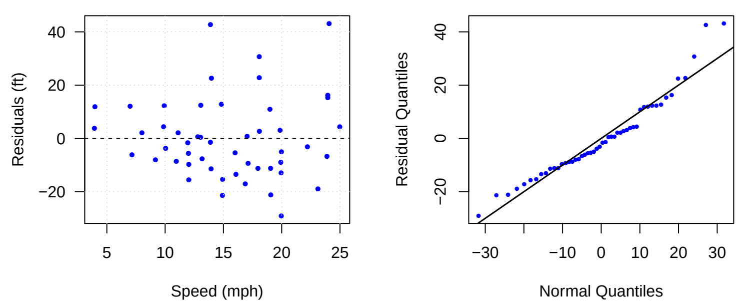
6. Test the following hypothesis: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ with $\alpha = 0.05$. You may use the confidence interval from (6). State your conclusion in plain language in the present context.

7. Construct the 90% prediction interval for a future observation of `dist` with `speed` = 20mph.

8. Fill in the missing values in the ANOVA table below and compute the R^2, the coefficient of determination.

Source	df	SS	MS	F
Model	?	SSR =?	MSR = ?	?
Error	?	SSE = 11362.39	MSE = ?	
Total	?	SST = 32516.40		

9. Do the residual plot and the Normal Q-Q plot below suggest any regression assumptions may be violated? Explain your answer.



10. Is that a good idea to predict `dist` given `speed` = 40mph? Explain your answer.

Problem 2

Suppose the researcher who performed the experiment in problem 1 wants to model the relationship between `dist` and `speed` using a 3rd polynomial regression (`CubeModel`) and to compare with a simple linear regression (`LinearModel`).

1. Suggest two different approaches to choose between `LinearModel` and `CubeModel`.

2. Perform a general linear test using the R output below:

```
Analysis of Variance Table

Model 1: dist_new ~ speed_new
Model 2: dist_new ~ poly(speed_new, 3)
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     48 11362
2     46 10616  2    746.46 1.6172 0.2095
```

Problem 3

The dean of a college in a University would like to monitor salary differences between male and female faculty members and she performed a multiple linear regression where the response variable `salary` is regressed on `sex` (male, female), `yrs.service` (years of service), `discipline` (A: "theoretical" departments, B: applied departments), and `rank` (Assistant, Associate, Full Professor). Use the R output below to answer the following questions:

```
Call:
lm(formula = salary ~ sex * yrs.service + discipline + rank,
    data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-64141 -14219  -1491  10684  99213

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         67732.17    6294.32  10.761  < 2e-16 ***
sexMale              5496.74    6464.52   0.850 0.395683
yrs.service           -29.40     437.54  -0.067 0.946460
disciplineB         13459.60    2320.48   5.800 1.37e-08 ***
rankAssocProf       14484.08    4139.34   3.499 0.000521 ***
rankProf            49072.60    3889.05  12.618  < 2e-16 ***
sexMale:yrs.service   -60.84     433.42  -0.140 0.888441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22680 on 390 degrees of freedom
Multiple R-squared:  0.4478,    Adjusted R-squared:  0.4393
F-statistic: 52.71 on 6 and 390 DF,  p-value: < 2.2e-16
```

1. Identify the dummy variables.
2. Write down the regression equation for each sex/discpline/rank combination (e.g., female/applied departments/Full Professor).