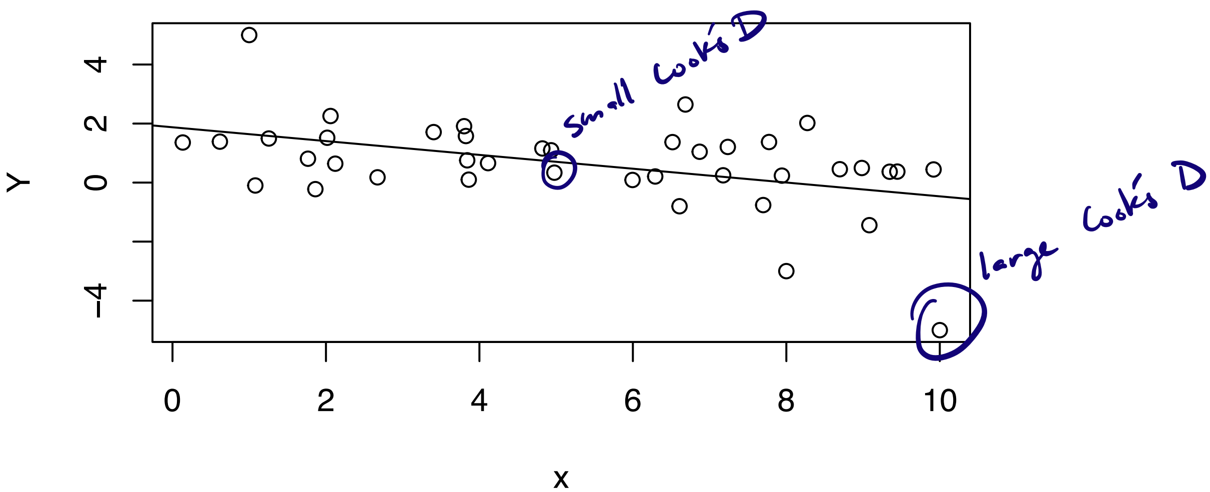
STAT 516 sp 2024 exam 01

75 minutes, no calculators or notes allowed

1. Simple linear regression

Below is a scatterplot of $n = 40$ data points $(x_1, Y_1), \ldots, (x_{40}, Y_{40})$ with the least squares line overlaid.

```
plot(Y~x)
abline(lm(Y~x))
```




```
summary(lm(Y~x))
```

```
Call:
lm(formula = Y ~ x)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-4.5336 -0.6735  0.0845  0.7369  3.3597

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.87442    0.45380   4.131 0.000191 ***
x           -0.23409    0.07469  -3.134 0.003315 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.389 on 38 degrees of freedom
Multiple R-squared:  0.2054,    Adjusted R-squared:  0.1845
F-statistic: 9.824 on 1 and 38 DF,  p-value: 0.003315
```

Some of the following questions refer to the above R output; some are general questions that you can answer without referring to the R output.

(a) What do we call the quantity $\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$ and what does it represent?

This is the total sum of squares.
It represents the total amount of variability in the values of Y.

(b) What do we call the quantity $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2$ and what does it represent?

This is the regression sum of squares.
It represents the amount of variability in Y accounted for by considering the predictor x.

(c) Give the value shown in the R output for $\dfrac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}$. Interpret the value.

This $R^2 = \dfrac{SS_{Reg}}{SS_{Tot}} = 0.2050.$ The predictor x is able to explain 20.5 % of the variability in Y.

(d) Obtain the value of $\dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$ from the R output. What does it estimate?

This is $(1.389)^2$. It estimates the error term variance σ^2, while 1.389 estimates the error term standard deviation σ.

(e) Give the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the R output. Give an interpretation of $\hat{\beta}_1$.

We have $\hat{\beta}_0 = 1.874$ and $\hat{\beta}_1 = -0.234$.

If x is increased by 1 unit, the expected value of Y is estimated to decrease by 0.234.

(f) Confidence intervals for $\beta_0 + \beta_1 x_{\text{new}}$ as well as prediction intervals for Y_{new} at new values of the predictor $x_{\text{new}} = 5$ and $x_{\text{new}} = 9$ are given below. For each interval, indicate whether it is a CI or a PI and indicate to which value of x_{new} it corresponds.

i. (-3.13, 2.67) Widest, so PI for Y_{new} at $x_{new} = 9$

ii. (0.26, 1.15) Narrowest, so CI for $\beta_0+\beta_1 x_{new}$ at $x_{new} = 5$

iii. (-0.94, 0.48) Second narrowest, so CI for $\beta_0+\beta_1 x_{new}$ at $x_{new} = 9$

iv. (-2.14, 3.55) Second widest, so PI for Y_{new} at $x_{new} = 5$

C.I.

$$\begin{array}{r} 2.67 \\ 3.13 \\ \hline 5.80 \end{array} \qquad \begin{array}{r} 2.14 \\ 3.55 \\ \hline 5.69 \end{array}$$

(g) Circle a data point on the scatterplot which would have a large value of Cook's D. Explain your choice of data point.

This data point is far from the mean of the x values (so it has high leverage) and it is also far from the other data points vertically. It will exert a strong influence on the least squares line.

(h) Circle a data point on the scatterplot which would have a small value of Cook's D. Explain your choice of data point.

This data point reinforces the pattern formed by the majority of the data points. Moreover it has small leverage since it is close to the mean.

(i) There is a p-value which appears twice in the R output. Explain why the same p-value appears twice.

The p-value 0.003315 is the p-value for testing

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0.$$

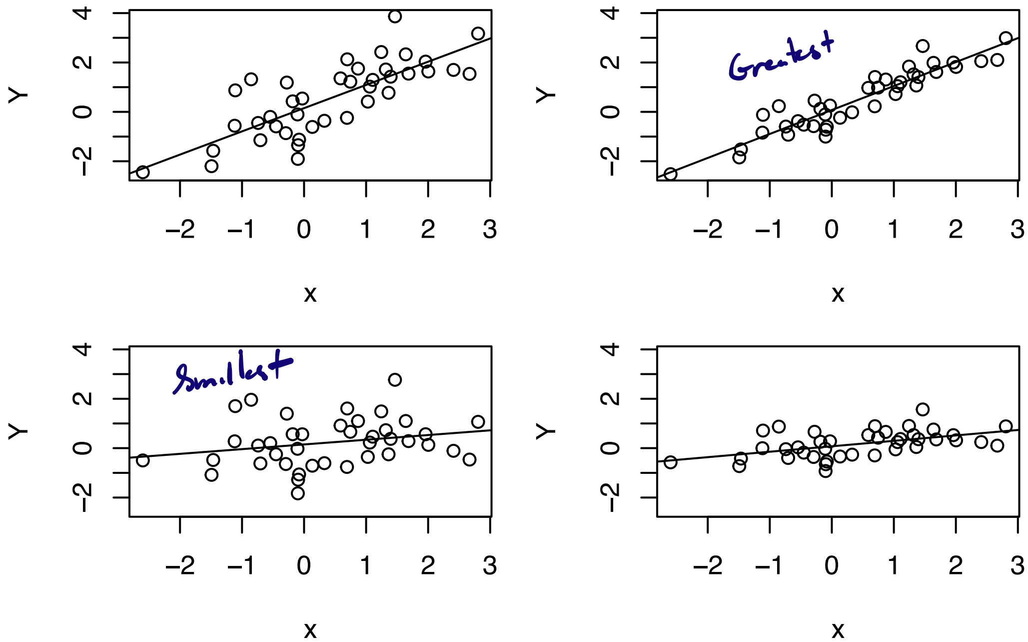
Since there is only a single predictor variable, the overall F-test for significance tests the same null and alternate hypotheses. So the p-value is the same.

(j) Scatterplots of four different data sets are shown below. Indicate for which data set the value of $F_{\text{stat}} = \dfrac{\text{MS}_{\text{Reg}}}{\text{MS}_{\text{Error}}}$ would be

a. the greatest.
b. the smallest.

* Spread of points gives denominator

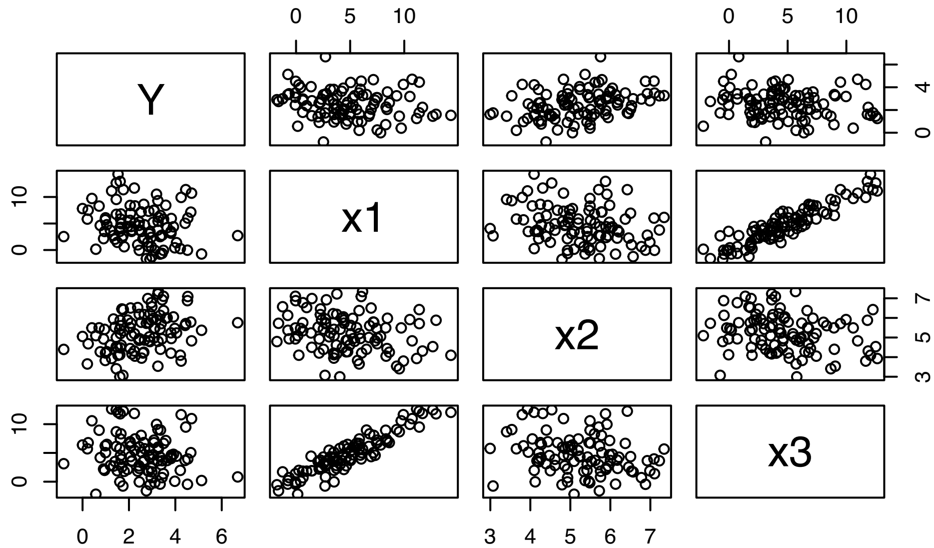
* Steepness of line gives numerator.



2. Multiple linear regression

The plot below shows scatterplots between all pairs of variables in a data set. Following that is some regression output.

```
plot(data)
```




```r
lm1 <- lm(Y ~ x1 + x2 + x3, data = data)
summary(lm1)
```

```
Call:
lm(formula = Y ~ x1 + x2 + x3, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0473 -0.8223 -0.0535  0.6444  3.9421 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.43876    0.74379   0.590  0.55664   
x1          -0.05382    0.08834  -0.609  0.54385   
x2           0.42276    0.12841   3.292  0.00139 **
x3           0.02437    0.09137   0.267  0.79025   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.194 on 96 degrees of freedom
Multiple R-squared:  0.1277,    Adjusted R-squared:  0.1004 
F-statistic: 4.684 on 3 and 96 DF,  p-value: 0.00426
```

```r
lm2 <- lm(Y ~ x2, data = data)
summary(lm2)
```

```
Call:
lm(formula = Y ~ x2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9448 -0.7886  0.0424  0.6243  3.9408 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.1536     0.6588   0.233 0.816083    
x2            0.4506     0.1237   3.643 0.000433 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.188 on 98 degrees of freedom
Multiple R-squared:  0.1193,	Adjusted R-squared:  0.1103 
F-statistic: 13.27 on 1 and 98 DF,  p-value: 0.0004333
```

Use the above R output to answer the following questions.

(a) For the model with all three predictors, give the value of each entry in the ANOVA table:

Source	Df	SS	MS	F value	p-value
Regression	i.	ii.	iii.	iv.	v.
Error	vi.	vii.	viii.		
Total	ix.	x.			

Since you may not use a calculator, give expressions that could be evaluated in order to obtain the right numbers!

i. $3 = df_{SS}$

ii. $(4.684)^* (1.194)^2 * 3 = SS_{Reg}$

or $(1.194)^2 * 96 \cdot \left(\frac{0.1297}{1-0.1297}\right) = SS_{Reg}$

$MS_{Reg} = \frac{SS_{Reg}}{p}$

$\Leftrightarrow SS_{Reg} = MS_{Reg} * p$

iii. $(4.684) * (1.194)^2 = MS_{Reg}$ $\qquad F_{stat} = \frac{MS_{Reg}}{MS_{Error}} \iff MS_{Reg} = F_{stat} \cdot MS_{Error}$

iv. $4.684 = F_{stat}$

v. $0.00426 = $ p-val

vi. $96 = df_{Error}$

vii. $(1.194)^2 * 96 = SS_{Error}$ $\qquad MS_{Error} = \frac{SS_{Error}}{n-(p+1)} \iff SS_{Error} = (n-(p+1))\, MS_{Error}$

viii. $(1.194)^2 = MS_{Error}$

ix. $99 = df_{Tot}$

x. $(4.684) * (1.194)^2 \cdot 3 / 0.1193$ $\qquad R^2 = \frac{SS_{Reg}}{SS_{Tot}} \iff SS_{Tot} = \frac{SS_{Reg}}{R^2}$

(b) Which two predictor variables will have the highest variance inflation factors? How can you tell?

The variables x_1 and x_3, since they each have high correlations with the other predictors.

(c) For the model with all three predictors, give the null and alternate hypotheses for the overall F-test of significance.

$H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$ vs $H_1:$ at least one $\beta_j \neq 0$, $j = 1, 2, 3$.

(d) Suppose we wish to test simultaneously the significance of `x1` and `x2`. Write down the relevant null and alternate hypotheses.

$H_0: \beta_1 = 0, \beta_2 = 0$ vs $H_1: \beta_1$ and β_2 are not both 0.

(e) Give the value of s needed to compute the test statistic

$$F_{\text{stat}} = \frac{(\text{SS}_{\text{Error}}(\text{Reduced}) - \text{SS}_{\text{Error}}(\text{Full}))/s}{\text{SS}_{\text{Error}}(\text{Full})/(n-(p+1))}$$

of the full-reduced model F-test.

Since the null hypothesis sets 2 slope coefficients equal to 0, we use $s = 2$.

(f) The value of the test statistic F_{stat} for the full-reduced model F-test is 0.464. Moreover, $F_{2,96,0.05} = 3.091$. What do we conclude about the significance of `x1` and `x2`?

Since $F_{stat} < F_{2,96,0.05}$, we fail to reject H_0.

So it is safe to regard x_1 and x_2 as having no important contribution to the value of Y.

3. Inference on the mean of a Normal distribution

Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$ and suppose we wish to test H_0: $\mu = 1$ versus H_1: $\mu \neq 1$. Let

$$T_{\text{stat}} = \frac{\bar{X}_n - 1}{S_n/\sqrt{n}}$$

and suppose we reject H_0 when $|T_{\text{stat}}| > t_{n-1,\alpha/2}$ for some significance level α. Answer the following questions about the probability $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$, which is the probability of rejecting H_0, also called the *power* of the test.

(a) Suppose μ is truly equal to 1. Then give $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$

This is equal to α.

(b) What happens to $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$ as μ moves away from 1?

It increases from α, limiting to 1 as μ moves further from 1 in either direction.

(c) Suppose μ is not equal to 1. What happens to $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$ if the sample size is increased?

It increases.

(d) Suppose μ is not equal to 1. What is the effect on $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$ of a larger variance σ^2?

Larger variance will decrease the power when $\mu \neq 1$.

(d) Suppose μ is truly equal to 1. What is the effect on $P(|T_{\text{stat}}| > t_{n-1,\alpha/2})$ of a larger sample size n?

It will have no effect, since the critical value $t_{n-1,\alpha/2}$ is calibrated based on the sample size and α such that for any sample size n, the Type I error rate is exactly α.